

Д-р. У. Снедекор

СТАТИСТИЧЕСКИЕ
МЕТОДЫ

В ПРИМЕНЕНИИ
К ИССЛЕДОВАНИЯМ
В СЕЛЬСКОМ ХОЗЯЙСТВЕ
И БИОЛОГИИ

Дж. У. С Н Е Д Е К О Р

СТАТИСТИЧЕСКИЕ МЕТОДЫ

В ПРИМЕНЕНИИ
К ИССЛЕДОВАНИЯМ
В СЕЛЬСКОМ ХОЗЯЙСТВЕ
И БИОЛОГИИ

*

Перевод с английского
В. Н. ПЕРЕГУДОВА

ИЗДАТЕЛЬСТВО
СЕЛЬСКОХОЗЯЙСТВЕННОЙ ЛИТЕРАТУРЫ,
ЖУРНАЛОВ И ПЛАКАТОВ
МОСКВА - 1961

О Т И З Д А Т Е Л Ъ С Т В А

В книге описаны современные методы математической статистики для использования их при обработке результатов сельскохозяйственных опытов и данных биологических экспериментов. Книга рассчитана одновременно и на читателя, мало знакомого с методами математической статистики, и на экспериментатора, желающего получить представление о более усовершенствованных методах обработки материала. Подробно изложен дисперсионный анализ многофакторных опытов, в которых на испытуемые объекты действует не один, а несколько факторов.

Книга имеет ясно выраженный характер практического руководства. В ней нет математического обоснования и вывода рекомендуемых формул. Но читатель имеет возможность экспериментальным путем получить довольно полное представление об основных принципах современной математической статистики. Автор дает достаточную характеристику каждого метода, определяет его целевую установку, приводит примеры вычислений, доводя их до конечного результата и выводов, и, наконец, дает задачи для упражнения.

Материал, даже наиболее трудный, изложен вполне доступно для понимания и не требует знакомства с дифференциальным и интегральным исчислением. В книге приведено много подсобных таблиц, необходимых для оценки достоверности выводов, полученных в результате статистической обработки экспериментальных данных или данных наблюдения.

Книга представляет интерес для научных работников, преподавателей, аспирантов и студентов сельскохозяйственных вузов, биологических факультетов университетов и педагогических институтов, работников опытных станций и полей, сортоиспытательных участков, а также агрономов, ведущих опытную работу в колхозах и совхозах.

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Предлагаемая читателю книга Дж. У. Снедекора является одной из наиболее известных работ по прикладной математической статистике и пользуется широким признанием в кругах исследователей — агрономов, зоотехников и биологов, в связи с чем она неоднократно переиздавалась. Почти в каждой работе или статье в области биологии и агрономии авторы, использующие статистические методы анализа экспериментальных данных, в обязательном порядке ссылаются на книги Р. А. Фишера¹ и Дж. У. Снедекора. Первый автор, являющийся творцом многих современных статистических методов, применяемых в биологии, упоминается чаще всего как классик, т. е. в порядке признания его заслуг и авторитета. Сама же книга Р. А. Фишера хотя и составлена с намерением дать популярное изложение статистических методов, все же в ряде мест малодоступна для исследователя — биолога или агронома, не имеющего специальной математической подготовки, и написана довольно трудным и несколько абстрактным языком. Книга Дж. У. Снедекора, менее глубокая по своему теоретическому содержанию, чем книга Р. А. Фишера, имеет по сравнению с последней то преимущество, что охватывает более широкий круг практических вопросов и, что самое важное, является действительно популярным изложением статистических методов исследования в биологии. Пожалуй, следует считать, что эти две книги в известной мере дополняют друг друга.

Предлагаемая книга Дж. У. Снедекора во многих отношениях оригинальна и значительно отличается от других работ, излагающих статистические методы исследования. Именно этот особый подход к построению книги и обусловил признание и широкую известность ее среди агрономов и биологов-экспериментаторов.

Прежде всего следует отметить, что автор хорошо понимает, скажем, психологию экспериментатора и стремится так изложить и обосновать предлагаемые им статистические методы, чтобы они органически вошли в круг тех методологических представлений, которыми привык пользоваться экспериментатор. Основная масса фактов в биологии добыта опытным путем, и даже биологические теории и положения, имеющие вначале умозрительный характер, как правило, подвергаются доступной экспериментальной проверке; во всяком случае, такая проверка придает им особо убедительную силу. Биолог, привыкнув руководствоваться логикой эксперимента, в тех случаях, когда он обращается за консультацией к руководству по математической статистике, встречается с иной логикой исследования: в основе рекомендаций статистики лежат некоторые абстрактные положения, из

¹ Р. А. Фишер. Статистические методы для исследователей. Госстатиздат, Москва, 1958. Перевод В. Н. Чергудова.

которых чисто дедуктивным путем выводятся соответствующие следствия, оформленные в виде правил и схем разработки конкретных данных. Исследователь-биолог принужден в этих случаях брать рекомендации статистики на веру и всецело полагаться на авторитет математиков. Конечно, он понимает, что это обстоятельство обусловлено особенностями математики, которая не принадлежит к числу экспериментальных наук и основным методом которой является дедукция. Однако некоторая неудовлетворенность у него все же остается: в то время как результаты эксперимента или своих наблюдений он ощущает непосредственно, статистическая разработка соответствующих данных как бы уходит от его контроля.

Дж. У. Снедекор, учитывая все это, стремится встать на точку зрения экспериментатора и помочь ему ознакомиться с основными статистическими положениями при помощи привычного для него экспериментального метода. Уже в первой главе он опытным путем строит довольно сложное и в то же время имеющее большое значение в статистике распределение «хи-квадрат». Тут же он предлагает читателю самостоятельно построить такое же распределение, используя приведенные в книге указания. (Здесь, кстати, следует сказать, что подобного рода рекомендации автора, рассеянные по всей книге, и, в частности, требования обязательного самостоятельного решения всех приводимых им задач и упражнений всецело поддерживаются нами. Автор прав, указывая, что изучение математической статистики без соответствующей практики невозможно.)

Принятый в книге экспериментальный подход к установлению важнейших статистических закономерностей дает читателю возможность, во-первых, получить непосредственное представление об этих закономерностях, обычно облакаемых математиками в форму готовых таблиц и формул, и, во-вторых, убедиться в том, что каждая такая закономерность (например, распределение того или иного статистического показателя), будучи установлена по ограниченному количеству данных, теряет свою строгую форму и в какой-то мере искажается. Тем самым читатель воочию убеждается в том, что всякий эксперимент, в том числе и эксперимент, направленный на установление статистических закономерностей, всегда связан с теми или иными случайными погрешностями. Когда после этого Дж. У. Снедекор приводит уже строгие, установленные теоретическим путем формы этих закономерностей, то читатель, с одной стороны, получает удовлетворение от того, что он и без сложных математических выкладок достигает понимания сущности их, и, с другой стороны, он проникается уважением к математике, которая способна эти закономерности представить в их чистом виде, освобожденными от влияния погрешностей, что, как он только что убедился, невозможно сделать на основе только экспериментального метода. В целом этот педагогический прием автора следует признать весьма удачным и облегчающим агрономам и биологам-экспериментаторам изучение основ статистической теории.

Другая особенность книги Дж. У. Снедекора состоит в том, что при ее написании автор сознательно нарушил некоторые традиции, ставшие почти обязательными для каждого популярного изложения основ статистики. Как правило, соответствующее руководство начинается с изложения некоторых вопросов теории вероятностей, далее даются главы о средних, о показателях варьирования, о распределениях и т. д. Порядок тем бывает различным, но традиция требует, чтобы каждая из этих тем в соответствующей главе излагалась со всей доступной для данного курса полнотой. Такой способ изложения обычно характерен для учебников и направлен на то, чтобы полное представление учащегося о предмете сложилось после последовательного изучения всего курса. Дж. У. Снедекор, по-видимому, имел в виду иного читателя — в основном уже сложившегося исследователя-биолога, который стремится в кратчайшее время определить практическое значение и полезность для него статистических методов анализа и которого поэтому необходимо сразу ввести в круг наиболее важ-

ных теоретико-статистических понятий. Автор разрешает эту задачу мастерски. Приходится восхищаться тем, как он в первой главе, содержащей каких-нибудь 30 страниц, знакомит читателя почти со всеми теоретическими положениями математической статистики: здесь даются понятия о выборочном методе, о распределениях и, в частности, о распределении «хи-квадрат», занимающем в теории статистики особое положение и являющемся исходным для построения целого ряда важнейших критериев, о параметрах распределений и их статистических оценках, о доверительной вероятности и доверительном интервале, о проверке гипотез, об ошибках первого и второго рода и т. д., причем все это сопровождается иллюстрациями на конкретном биологическом материале. Читатель после прочтения только одной этой главы почти полностью введен в круг основных статистических понятий; далее они будут уточняться, углубляться и все более и более конкретизироваться, но уже здесь читателю становится понятным, в чем суть и своеобразие статистических методов исследования.

Правда, такое построение книги создает впечатление некоторой разбросанности и неупорядоченности изложения: одна и та же тема рассматривается в разных местах книги, вследствие чего она теряет свою цельность. Однако именно благодаря этому Дж. У. Снедекору удалось в одной книге по существу совместить несколько руководств, начиная с первоначального краткого курса для начинающего исследователя и кончая полным курсом или полным изложением определенной темы для исследователя, уже имеющего известный опыт статистической работы. Рассеянные по всей книге указания автора, выделенные курсивом, отсылают читателя, начинающего изучать статистику, к следующим разделам элементарного курса и советуют ему оставить пока в стороне более сложные вопросы. В этих же указаниях читатель, интересующийся частной темой, найдет рекомендацию автора обратиться к соответствующему месту книги.

Литературный стиль автора также своеобразен. Он часто непосредственно обращается к читателю, задает ему вопросы, осведомляется, как тот смотрит на предмет обсуждения, высказывает свое личное мнение и даже иногда прибегает к шутливому тону.

Все это придает изложению характер собеседования автора с читателем. Это во многом облегчает чтение отнюдь не столь простой и легкой книги.

Отмеченные здесь особенности в построении книги Дж. У. Снедекора являются как бы уступкой математика читателю — агроному или биологу, сделанной для того, чтобы в максимальной степени облегчить изучение столь сложной науки, какой является математическая статистика. Можно было бы ожидать, что такое отступление от обычных норм изложения в какой-то мере скажется отрицательно на глубине и строгости самого изложения. Однако автор оказался столь искусным педагогом, что он вполне благополучно миновал эту опасность. Логика предмета, определения понятий, формулировка выводов вполне строгая, точная и безукоризненная в той мере, в какой это вообще доступно для популярного руководства. Изложение всех затронутых вопросов находится на уровне современных знаний. В частности, следует отметить, что в книге уделено достаточно большое внимание новому, быстро развивающемуся разделу, излагающему непараметрические методы оценок, который в обычные курсы статистики пока еще не включается.

Конечно, предлагаемая книга не является всеобъемлющей, и в нее не вошел целый ряд специальных методов, применяемых в биологии и агрономии. Но это не вина автора, ибо в настоящее время ведется успешная разработка статистических методов во многих, самых различных направлениях, объединить которые в силу их специфичности, а иногда и в силу большого объема полученных результатов не представляется возможным. Автор, например, ограничивается рассмотрением только простейших схем полевых опытов и даже не упоминает такие широко распространенные в настоящее

время методы построения этих опытов, как метод «смешивания», метод «решетки» и др. Он не мог касаться этих вопросов, так как теория построения и обработки результатов полевых опытов в данный момент приобрела столь обширные размеры, что вышедшая недавно книга У. Т. Федерера «Планирование опытов» (Federer W. T. Experimental design), посвященная этому вопросу, по своему объему не уступает всей книге Снедекора.

Советский читатель будет удивлен тем, что нигде в книге и в списке литературы не упоминаются ни советские, ни прежние русские математики, способствовавшие развитию теории вероятностей и математической статистики. Между тем достижения русской и советской математической школы в этой области широко известны и пользуются мировой славой: достаточно упомянуть работы Чебышева, Ляпунова, Маркова, Колмогорова, Романовского и др. Но замалчивание этих достижений, по-видимому, стало традицией англо-американских книг по статистике. Например, метод параболического выравнивания, разработанный еще в прошлом веке академиком Чебышевым, Снедекор приписывает Р. А. Фишеру. В своих примерах и упражнениях автор иногда приводит данные опытов, основанные на концепциях биологического характера, не согласующихся с данными советской биологической науки. Однако поскольку не они составляют предмет изложения, нет необходимости специально оговаривать те места книги, с которыми советский читатель-биолог не может согласиться.

Несколько слов относительно терминологии и символики. К сожалению, в разных изданиях одни и те же английские термины переводятся по-разному, и пока еще нет полностью стандартизированной символики. Например, термин *analysis of variance* переводится то как анализ рассеяния, то как вариантный анализ, то как дисперсионный анализ. При переводе данной книги мы стремились придерживаться наиболее употребительной в русском языке терминологии и полностью воспроизвели символику автора. Например, указанный выше английский термин переводился как дисперсионный анализ. Здесь следует отметить только два особых случая. Ранее переводчик в своих самостоятельных работах и в переводах с английского применял в свое время широко распространенный русский термин «среднее квадратическое отклонение». Однако, принимая во внимание, что дословный перевод термина *standard deviation* — стандартное отклонение в настоящее время приобретает более широкое применение в нашей математической литературе, переводчик решил отказаться от прежнего термина и использовать этот дословный перевод. Это подсказывается и сутью дела: стандартное отклонение, при вычислении которого производится деление суммы квадратов отклонений на число степеней свободы, а не на число слагаемых, уже не является средней величиной в обычном понимании этого слова. Термин *signification* в математической литературе часто переводится дословно — «значимость», однако мы переводили его как «существенность». Это сделано, во-первых, потому, что последний термин более близок биологу и агроному, часто говорящим, хотя и в другом смысле, о существенности или несущественности сортовых, породных и других различий, и, во-вторых, в связи с тем, что, приняв термин «значимость», нам часто пришлось бы при переводе этой книги говорить о «значимом значении» того или иного статистического показателя. По нашему мнению, термин «существенность» для русского читателя звучит более привычно и мягко, чем слишком жесткий термин «значимость» и тем более производное от него слово «значимо».

Иногда вместо термина «значимость» применяют термин «достоверность» (ранее применял этот термин и автор этих строк). Однако как этимологически, так и в обычном словоупотреблении «достоверность» означает нечто обязательное и неукоснительное; оно придает утверждению категорический характер полной истинности. В теории вероятности достоверным считается то, что обладает вероятностью, равной единице, означающей полную обязательность. Между тем статистические утверждения обладают только определенной степенью вероятности, характеризующейся тем или иным уровнем доверия.

Нам представляется, что термин «существенность» в этом отношении более гибок, так как позволяет говорить о различных градациях существенности, что в отношении термина «достоверность» вряд ли было бы уместным.

* * *

Книга Дж. У. Снедекора написана в целом столь хорошо, что не требует каких-либо дополнений и пояснений. Однако один вопрос для советского читателя может показаться спорным и не разъясненным в книге до конца. Речь идет об обязательном случайном отборе наблюдений, или, как иногда говорят, о рендомизации. Автор, уделяя большое внимание технике случайного отбора наблюдений, в частности при помощи таблицы случайных чисел, все же не приводит в достаточной мере убедительных аргументов в обоснование обязательности рендомизации. В нашей стране в биологических исследованиях рендомизация, как правило, не применяется, но рекомендуемые автором приемы статистической обработки данных покоятся на допущении, что исследователь строго придерживается описанных в книге правил случайного отбора наблюдений. В связи с этим у читателя может возникнуть вопрос: допустимо ли применение описываемых автором методов, если рендомизация не была осуществлена при проведении эксперимента или наблюдения, или же, наоборот, требование рендомизации не является столь жестким и обязательным, как это следует из текста автора. Поэтому мы считали необходимым выделить этот вопрос и изложить его здесь более подробно, чем это делает в своей книге автор.

Английские и американские авторы руководств по применению статистики к биологии и агрономии, как и Дж. У. Снедекор, мало уделяют внимания разъяснению вопроса о рендомизации, выдвигая ее в качестве обязательного условия при организации эксперимента или наблюдений. Это обстоятельство, по-видимому, объясняется тем, что авторитет Р. А. Фишера, особенно настаивавшего на этом требовании, среди исследователей-биологов Великобритании и США весьма высок и его рекомендации пользуются полным доверием. Вместе с тем в этих странах в биологии и опытной агрономии работает большое число лиц со специальным математическим образованием (большинство авторов руководств — математики). Для математика же, знакомого с предпосылками, на основе которых строится вся теория математической статистики и определяются практические критерии существенности, собственно говоря, не возникает никакого сомнения в отношении обязательности рендомизации. Такой общий фронт авторов-математиков, считающих рендомизацию обязательной, заставляет исследователей-биологов этих стран не сомневаться в том, что это условие является необходимым.

Исходным пунктом для всех математико-статистических построений является предпосылка о наличии случайной выборки. В той мере, в какой эта предпосылка соответствует действительности, все выводы статистики будут правильными; в случае же нарушения этого принципа случайного отбора законность и реальность последующих выводов становятся неопределенными и недостаточно обоснованными. Конечно, нельзя ожидать, что абстрактные и в известной мере идеализированные условия, на которых строится математическая теория, могут быть полностью осуществлены на практике. Однако приближение к этим условиям в той мере, в какой оно достижимо, безусловно, обязательно. Рендомизация является именно тем средством, которое в достижимой степени приближает нас к абстрактной случайной выборке и которое дает известные гарантии в правильности практических выводов, полученных в результате статистической разработки конкретных экспериментальных данных или данных наблюдения.

Если у математика этот вопрос не вызывает никаких сомнений и не требует какого-либо специального обоснования, то для биолога-экспериментатора, организующего свой эксперимент по заранее обдуманному и обоснованному плану, сознательное привлечение к участию в эксперименте столь неожиданного помощника, каким является «случайность», представляется

неуместным, если не вредным. Экспериментатор, вооруженный достаточно совершенной техникой эксперимента, знающий и учитывающий внешние условия проведения опыта, часто уверен в том, что все нити у него в руках и что он способен определить и учесть все существенные элементы и составные части эксперимента. Однако если он думает так, то глубоко ошибается, ибо необходимость и случайность, как известно, составляют диалектическое единство, и поэтому участие случайности во всяком сколь угодно точном и хорошо продуманном эксперименте неизбежно. Поэтому правильная организация эксперимента возможна только тогда, когда, кроме планирования и учета контролируемых элементов эксперимента, учитываются и по-своему «планируются» и элементы эксперимента, связанные со случайностью. Рендомизация является этим «планированием» эксперимента в части, касающейся элементов опыта, имеющих случайный характер, введением этих элементов в определенные рамки. Следовательно, рендомизация — это отнюдь не подчинение экспериментатора воле случая, а, наоборот, подчинение случая воле экспериментатора, при котором случайность, оставаясь сама собой, т. е. обусловленной неконтролируемой причинностью, становится в то же время в известном смысле контролируемой.

Для конкретизации этих общих положений можно привести ряд примеров. Прежде всего рассмотрим вопрос об отборе так называемого типичного образца. Экспериментатору очень часто представляется, что он сможет более точно охарактеризовать совокупность изучаемых объектов, если он вместо образца, составленного из объектов, отобранных случайным порядком, возьмет в состав этого образца объекты, которые представляются ему наиболее типичными для данной совокупности. Экспериментатор стремится здесь заменить «слепой случай» сознательным подходом к образованию образца, причем обычно он всегда уверен в том, что при таком отборе полностью сохранена объективность и что этот отбор более совершенен, чем случайный отбор образца. Однако чаще всего в этом случае объективность является мнимой. Интересный пример приводит английский статистик Ф. Нейте в работе «Application of the Sampling Technique to Crop Estimation and Forecasting» (Manchester Statistical Society, 1936 г.). На Ротамстедской опытной станции был проведен такой опыт. На стол было положено 1200 камней различного размера и веса, и 12 лицам предложено отобрать по три серии, каждая из 20 камней, по-возможности, среднего веса. Конечно, после каждого испытания камни возвращались обратно и перемешивались, так что условия для каждого наблюдателя были одинаковы. Результаты опыта представлены в приводимой ниже таблице.

Средний вес камня (в унциях) при отборе серий по 20 камней

Серии	Наблюдатели											
	1	2	3	4	5	6	7	8	9	10	11	12
1-я	1,9	2,4	2,4	1,9	2,2	2,8	2,4	1,6	2,2	2,6	2,4	2,4
2-я	1,8	3,0	2,4	2,0	2,7	2,6	2,6	2,0	2,2	2,2	2,4	3,0
3-я	1,7	2,4	2,1	2,0	3,1	2,8	2,5	2,0	2,2	3,1	1,8	2,4
Средняя . . .	1,8	2,6	2,3	2,0	2,7	2,7	2,5	1,9	2,2	2,6	2,2	2,6

Средняя всех серий 2,34 унции, точная же средняя для всех 1200 камней 1,91 унции; таким образом, наблюдатели в целом преувеличили средний вес камня примерно на 25%. Далее мы видим, что подавляющее число лиц, участвовавших в эксперименте, по каким-то причинам отбирали более тяжелые камни; только двое из двенадцати в среднем отбрали более легкие камни (1-й и 8-й наблюдатели). Из общего числа 36 серий только в 6 сериях средний вес камня оказался ниже или равен фактической средней (1,91 унции).

Вместе с тем замечается определенная склонность отдельных наблюдателей делать примерно одинаковую ошибку в том или ином направлении. Так, первый наблюдатель постоянно берет более легкие камни, в то время как второй — более тяжелые, четвертый держится на уровне среднего веса и т. д. Ошибки наблюдателей таковы, что преуменьшение веса незначительно (максимальное отклонение в эту сторону равно 0,31 унци), в то время как преувеличение веса встречается чаще и по своему размеру больше (максимальное отклонение — 1,09 унци). Этот эксперимент весьма наглядно показывает, сколь относительной и несовершенной является способность человека отобрать типичное, и убеждает в том, что систематические ошибки, вызванные субъективными заблуждениями исследователя, могут достигать значительной величины и быть устойчивыми. Случайный отбор исключает возможность этих систематических субъективных ошибок.

Приведенный пример является несколько отвлеченным и, возможно, поэтому не вполне убедительным для биолога. Тот же автор приводит другие данные, уже непосредственно относящиеся к обсуждаемому предмету. По заданию Британского агрометеорологического комитета производилось широкое обследование высоты растений яровой пшеницы в целом ряде пунктов. По правилам следовало взять 32 пробы от каждого сорта, причем каждая проба состояла из четырех отрезков рядков в 0,25 м, размещенных по концам рядков. Измерялись стебли на концах каждого рядка (8 на пробу). В некоторых случаях отбор можно было произвести только на трех рядках, и, следовательно, представлялась возможность учета только 6 концов. Чтобы не нарушать общей системы 8 измерений, наблюдателям в этих случаях предлагалось недостающие 2 пробы отбирать по своему усмотрению, ориентируясь, конечно, на среднее состояние растений. Записи велись так, что группу наблюдений, произведенных по принципу отбора типичного образца, можно было выделить и сопоставить с данными остальных наблюдений, которые в силу большей численности этих наблюдений были более достоверными. Промеры велись в три срока: 31 мая, 14 июня и 28 июня. Цель наблюдения — установить интенсивность роста растений. Ниже приводятся результаты наблюдений (средняя высота растений в см):

Дата измерения	Отбор по системе	Отбор типичного образца	Отклонения
31 мая	47,49	50,82	+ 3,33
14 июня	76,56	78,29	+ 1,73
28 июня	118,84	116,12	- 2,72

Как видим, наблюдатели в первые сроки при низком росте растений произвольно стараются их «подтянуть»; в последний же срок, когда растения достигли большой высоты, наблюдатели не доверяют этому факту и стремятся отобрать более низкорослые растения. В то время как фактический средний прирост растений за этот период составляет 74,35 см, наблюдатели на основе типичных образцов устанавливают прирост только в 65,30 см, т. е. преуменьшают его почти на 10%.

Два приведенных здесь примера показывают, что каким бы добросовестным и осторожным ни был наблюдатель, мы не можем доверять объективности его оценки типичного образца. Самое опасное в этой оценке не то, что она не дает точных результатов, а то, что она тенденциозна. Оценка на основе случайного отбора (или механического отбора по определенной системе, что часто равнозначно случайному отбору) тоже не может дать точных результатов, но она в отличие от первой не будет зависеть от субъективных впечатлений наблюдателя и будет полностью свободна от той или иной тенденциозности. Следовательно, применяя рандомизацию (или меха-

нический отбор), мы вводим в план опыта или наблюдения своего рода заслон против проникновения в результаты опыта субъективных представлений экспериментатора.

Иногда, применяя случайный или механический отбор объектов, ограничивают его некоторыми условиями, которые, как думают, обеспечивают получение более правильных результатов, т. е. проводят определенную корректировку результатов случайного отбора. Например, многие инструкции и указания по отбору так называемых растительных проб в полевом опыте (по этим пробам устанавливаются для вариантов опыта густота стояния растений, число колосов на одно растение и другие биологические характеристики), рекомендуя определенную систему расположения этих проб на опытных делянках (по диагонали, шахматное и другие расположения проб), в то же время предоставляют право наблюдателю, если он обнаружил, что проба случайно попала в особо неблагоприятное или особо благоприятное место, заменить ее на более типичную. Здесь, таким образом, наблюдателю предоставляется право корректировать результаты случайного отбора.

Для определения того, каким может быть результат такого случайного отбора под контролем наблюдателя, автором настоящих строк на одном из сортоучастков Государственной комиссии по сортоиспытанию сельскохозяйственных культур в 1941 г. было проведено специальное исследование. На делянках опыта с 13 сортами яровой пшеницы (78 делянок) было размещено повышенное по сравнению с обычным число пробных площадок (18 площадок с площадью в $\frac{1}{6}$ кв. метра на делянку). Каждая делянка делилась поперек на 6 парцелл, на каждой из которых в определенной системе было выделено по 3 площадки. На этих площадках производился подсчет растений с последующим пересчетом на 1 кв. метр. При проведении этой работы персоналу сортоучастка предлагалось произвести выбор из трех проб каждой парцеллы той пробы, которая ему представляется типичной для парцеллы. Если среди трех проб не оказывалось ни одной, которую можно было бы безоговорочно назвать типичной для данной парцеллы, то выбиралась та, которая более всего подходила к типичной и которая в этом случае называлась условно типичной. На основе этих данных была определена густота стояния по каждому из 13 сортов: по всем пробам (108 на каждый сорт), по типичным и по условно типичным. Ниже приводятся данные для трех групп сортов (по 4 сорта): сорта с малой густотой стояния, сорта со средней густотой стояния и сорта с большой густотой стояния, а также для стандартного сорта:

Сорта	Число растений на 1 кв. м		
	по всем пробам	по условно типичным	по типичным
Стандарт	286	316	324
С малой густотой стояния	244	253	266
» средней »	284	285	279
» большой »	304	299	282

Эти данные со всей очевидностью показывают, сколь далек от объективности отбор типичного образца, произведенный персоналом сортоучастка, в добросовестности которого сомневаться не приходится. Прежде всего видна «осторожность» наблюдателей, по-видимому, боявшихся преувеличить преимущество сортов перед стандартным сортом, вследствие чего они произвольно впали в другую крайность — неосознанное «подтягивание» стандарта привело к тому, что этот сорт, относящийся фактически к группе сортов со средней густотой стояния (по всем пробам), на основе условно

типичных и в особенности на основе типичных проб охарактеризован как сорт с наибольшей густотой стояния. Эта же «осторожность» привела к преувеличенной оценке сортов с малой густотой стояния и к преуменьшенной оценке сортов с большой густотой стояния растений. Если судить по типичным пробам, то заметных различий по данному признаку между сортами, кроме стандарта, практически нет (различие между крайними группами составляет около 6%), хотя фактически эти различия довольно существенны (крайние группы различаются примерно на 25%). Следует также отметить, что субъективность оценки особенно проявилась при использовании типичных проб и в меньшей степени, когда были взяты условно типичные пробы, т. е. когда субъективизм наблюдателя был в известной мере ограничен.

Приведенные данные, собственно говоря, не дают прямого ответа на поставленный вопрос о том, каким будет результат корректировки случайного отбора. Все зависит от того, как часто наблюдатель будет браковать случайные пробы и заменять их «типичными». Чем реже он будет это делать, тем более правильными будут результаты наблюдений; наоборот, при частой браковке эти результаты будут ближе к тем, которые получены на основе типичных проб, т. е. более искаженными. Если теперь вспомнить данные первого нашего примера, в котором установлено, что существуют явные различия в оценке, определяемые индивидуальностью наблюдателей, то становится очевидным, что представление права браковки случайных проб и замены их «типичными» большому числу наблюдателей чревато большой опасностью получить данные, отражающие личные впечатления этих наблюдателей, а не объективную действительность. Случайный отбор или механический отбор там, где они равнозначны (см. последнюю главу настоящей книги), дают определенную гарантию того, что субъективность наблюдателей будет исключена.

Замена неудачной случайной пробы на «типичную» обычно обосновывается тем, что эта неудачная проба, если ее оставить, может в значительной мере исказить характеристику данного сорта или варианта опыта (так как число проб для сорта или варианта, как правило, небольшое). Следовательно, эта замена делается ради спасения сорта или варианта опыта, чтобы не приносить его в жертву случаю. Как бы ни был внешне убедителен этот довод, приведенные выше примеры говорят против него: спасая один сорт или вариант и легализуя корректировку случайного отбора, мы рискуем исказить результаты по всем сортам или вариантам опыта. Теперь в жертву хотя и не случайности, а субъективным взглядам наблюдателей приносится не отдельный сорт или вариант, а весь опыт в целом. Следует вместе с этим заметить, что случаи значительных искажений характеристик отдельных сортов или вариантов опыта статистикой определенным образом учитываются, так как все выводы строятся на основе выбранной доверительной вероятности, т. е. эти относительно резкие искажения входят в те 5% или 1% ошибок первого рода, которые заранее положены в основу критерия оценки.

Пока мы рассматривали вопрос о механическом способе отбора объектов, который часто, но отнюдь не всегда равнозначен случайному отбору. В одном частном, но весьма важном случае приходится ставить под сомнение применимость механического способа и требовать рендомизацию в ее чистом виде. Речь идет о порядке размещения вариантов полевого опыта. Общепринятая у нас в СССР методика полевого опыта предусматривает систематическое размещение вариантов с сохранением определенного порядка их в каждом повторении. Если в первом повторении для шести сортов установлен порядок *ABCDEF*, то этот же порядок сохраняется во втором, третьем и т. д. повторении; при расположении опыта в два яруса в повторениях второго яруса производится сдвиг этого порядка и берется последовательность вариантов *DEFABC*. Применяются сдвиги и других видов, но характерная особенность этих расположений остается: порядок вариантов устанавливается не по истребию, а по той или иной системе. Это расположение является по существу механическим способом отбора: вариант *A* в каждом повторении

первого яруса занимает первое место, а во втором ярусе — четвертое; вариант *B* в первом ярусе занимает каждую вторую деланку, а во втором — каждую пятую и т. д.

Ротамстедская опытная станция в Англии, где Р. А. Фишер долгое время руководил статистической лабораторией, уже более трех десятилетий применяет рендомизированный способ размещения вариантов по деланкам каждого повторения, т. е. такого размещения, при котором место каждого варианта определяется жребием (жеребьевка может быть заменена использованием таблицы случайных чисел). Этот метод постепенно завоевал признание в самой Англии, потом в США, а также в ряде других стран. Однако в Советском Союзе, в некоторых странах Центральной Европы и Скандинавии рендомизация не была принята безоговорочно и является предметом дискуссии.

Рендомизация вариантов полевого опыта на многих работников опытной агрономии производит прежде всего впечатление какой-то детской забавы, а не серьезного научного мероприятия; на память приходит жеребьевка, которую проводят дети перед началом некоторых игр. Однако установление порядка на основе жеребьевки производится не только в детских играх, но и в таких ответственных случаях, как спортивные соревнования. Это означает, что в тех случаях, когда для участников игры или соревнования не могут быть созданы абсолютно тождественные условия, единственно справедливым и, можно сказать, научным способом распределения этих условий является рендомизация. Научность и справедливость этого способа состоит в том, что хотя участники не могут пользоваться одинаковыми условиями соревнования, однако они обладают одинаковой вероятностью попасть в те или иные благоприятные или неблагоприятные условия.

Дж. Уишарт и Г. Сандерс в своей книге «Основы методики полевого опыта» (Издательство иностранной литературы, 1958 г.) приводят в качестве одного из аргументов в пользу рендомизации историческую справку: этот прием применялся, когда при общинном владении землей участки различного качества делились на части по числу членов общества, и каждому члену общества по жребию отводилась деланка с каждого участка. Этот способ рендомизации применялся на заре цивилизации, но спрашивается, могли бы мы в настоящее время в этих условиях предложить какой-либо более совершенный и научный метод справедливого распределения участков земли? Какие бы современные методы изучения плодородия почвы мы ни применили и какой бы способ справедливого распределения земли между участниками мы ни предложили, все равно нам не избежать споров и различных претензий со стороны участников. Они уснокоятся только тогда, когда мы поставим их в равные условия перед случаем, т. е. когда применим все ту же доисторическую жеребьевку. Оставим прошлые времена и обратимся к современности. Допустим, что в конкурсном сортоиспытании принимают участие авторы сортов. Каждый из этих авторов, ознакомившись с опытным участком и его отдельными деланками, не решится из чувства справедливости претендовать на лучшие деланки, но все же будет категорически протестовать против размещения его сорта на худших деланках. После более или менее длительного спора и обсуждения различных «справедливых», но для кого-нибудь все же неприемлемых вариантов размещения сортов авторы этих сортов наверно придут к согласию только тогда, когда примут жеребьевку.

Иногда против рендомизации возражают по техническим соображениям. Когда применяется систематическое размещение вариантов, то экспериментатор, запомнив порядок вариантов опыта, легко находит нужную ему деланку в каждом повторении для того, чтобы провести на ней соответствующие агротехнические мероприятия (посев, внесение удобрений и пр.) или наблюдения (определение густоты стояния растений, фаз развития и пр.). При рендомизации, когда в каждом повторении порядок размещения вариантов различен, экспериментатор все время должен иметь перед глазами план опы-

та и сопоставлять натуру с этим планом, что, с одной стороны, затрудняет работу и, с другой стороны, увеличивает, как некоторым кажется, вероятность перепутать деланки и допустить тем самым грубые ошибки, что в особенности возможно, когда работа поручается недостаточно квалифицированному техническому персоналу. Следует прежде всего заметить, что если то или иное мероприятие, направленное на совершенствование опыта, признано необходимым, то оно неукоснительно проводится, сколько бы хлопот или трудов оно ни стоило. Постороннему наблюдателю множество мероприятий, проводимых при закладке полевых опытов, таких, как тщательная регулировка сеялки, точный обмер деланок, посев по шнуру и пр., может показаться излишним или, по крайней мере, проводимым с ненужной тщательностью. Однако эти мероприятия наукой считаются обязательными для того, чтобы опыт признавался доброкачественным. Надо сказать, что затруднения, вызываемые несистематическим расположением вариантов, чаще всего преувеличены и, вообще говоря, при желании легко устранимы.

Вопрос о будто бы повышенной вероятности спутать деланки при несистематическом расположении вариантов оказался не столь серьезным. На Ротамстедской опытной станции был проведен специальный учет таких ошибок, и оказалось, что они при рендомизации встречаются реже, чем при систематическом размещении вариантов. Это понятно: исследователь или технический работник при рендомизации опыта не может полагаться на свою память, которая, как известно, всегда может изменить. Он принужден все время обращаться к плану опыта и соблюдать необходимую осторожность. Ясно, что он будет реже ошибаться.

Кроме указанных здесь возражений общего характера, против рендомизации возражают иногда и по мотивам математическим. На основе разработки данных условных опытов, т. е. таких, которые накладываются на однородный дробный учет урожая участка и в которых варианты фактически не отличаются друг от друга, установлено, что при систематическом размещении вариантов варьирование средних по вариантам оказывается меньшим, чем варьирование этих средних при случайном размещении вариантов опыта. Иными словами, получается так, что при систематическом размещении точность опыта выше, чем при рендомизации. Правда, при обработке данных по методу дисперсионного анализа остаточное варьирование, которое кладется в основу оценки опыта, при систематическом расположении оказывается более высоким. Таким образом, при разработке данных фактического, а не условного опыта при систематическом расположении вариантов, когда точность опыта определяется по остаточному варьированию, получается так, что исчисленная ошибка опыта будет в той или иной мере преувеличена. При случайном размещении вариантов фактическая и исчисленная ошибки опыта практически совпадают, и поэтому последняя является правильной оценкой точности опыта.

Н. Ф. Деревицкий («Научные записки по сахарной промышленности» № 5—6, 1935, стр. 80—81) приводит такие данные разработки результатов условного опыта.

	Систематическое расположение	Рендомизированное расположение
Дисперсия:		
между сортами	2,041	3,536
остаточная	3,851	3,685

Следует иметь в виду, что когда обрабатываются данные фактического, а не условного опыта, то показатели по строке «между сортами» остаются

неизвестными, так как к ним примешивается различие сортов. Поэтому оценка опыта производится на основе остаточной дисперсии. Р. А. Фишер, обсуждая этот вопрос в своей книге «Статистические методы для исследователей», указывает, что большая точность опыта в случае систематического расположения для нас не имеет никакого значения, так как она остается неуловимой и оценивается остаточной дисперсией неправильно. Н. Ф. Деревицкий, а в последнее время и некоторые другие исследователи считают, что лучше иметь более точные данные опыта, даже если правильная оценка точности опыта невозможна, чем иметь опыт, вполне точно оцениваемый, но более грубый сам по себе. Приверженцы последней точки зрения упускают из виду, что если опыт точен, но для оценки его результатов мы будем применять завышенный показатель ошибки опыта, то многие фактически существенные различия между сортами или вариантами будут отнесены нами к числу несущественных, т. е. будут полностью уничтожены все преимущества этого точного опыта.

При обсуждении этого вопроса, кажется, пока еще никто не обращал внимания на одно обстоятельство, которое также говорит не в пользу систематического расположения делянок. Н. Ф. Деревицкий и другие исследователи, придерживающиеся его точки зрения, считают, что уменьшение дисперсии по строке «между сортами» при систематическом размещении сортов вполне определенно говорит о том, что точность опыта в этом случае выше, чем при рендомизации. Однако при этом упускается из виду, что дисперсия или производное от нее стандартное отклонение является мерой точности только в условиях нормального распределения. Между тем можно теоретически доказать, что при систематическом расположении вариантов распределение средних уже не будет нормальным, и поэтому стандартное отклонение уже не будет достаточно надежной мерой точности опыта.

Следует иметь в виду, что стандартное отклонение в нормальном распределении является мерой точности опыта только потому, что с ним непосредственно связана вероятность появления отклонения того или иного размера. Так, мы знаем, что в пределах $\pm 2\sigma$ содержится примерно 95% всех ошибок, а вне этого интервала 5% ошибок, в пределах $\pm 3\sigma$ содержится 99,7% ошибок и т. д. Если же распределение не относится к числу нормальных, то непосредственной связи вероятности отклонений со стандартным отклонением уже нет. В этом случае может оказаться так, что ненормальное распределение, имеющее меньшее стандартное отклонение σ , чем нормальное распределение, будет в то же время иметь большее число крупных ошибок, превосходящих 2σ . Именно такое положение и создается, когда применяется систематическое расположение вариантов.

Не имея возможности дать здесь подробное изложение вопроса, приведем данные нашей обработки результатов двух серий условных опытов, наложенных на данные одного и того же дробного учета. Одна серия состояла из 32 опытов с 6 условными сортами в 6-кратной повторности при систематическом размещении (по схеме так называемого датского квадрата) сортов; вторая серия отличалась от первой только тем, что было применено рендомизированное размещение сортов (по схеме латинского квадрата). Стандартное отклонение (среднее) первой серии $s_1=29,6$, а второй серии $s_2=31,0$. Судя по этим показателям, точность опытов первой серии при систематическом размещении сортов, как и следовало ожидать, выше, чем точность опытов второй серии. Но рассмотрим два приводимых ниже распределения ошибок опыта для первой и второй серии опытов.

Ошибки	< -100	-80	-60	-40	-20	0	20	40	60	80	
1-я серия	2	—	6	24	32	55	48	18	6	1	$s_1=29,6$
2-я серия	—	1	5	26	48	40	36	27	6	3	$s_2=31,0$

Как видим, первая серия имеет распределение островершинной формы, и в то же время оно более разбросано вширь; оно не является нормальным.

Если определить количество ошибок, выходящих за пределы $\pm 2s$, т. е. за пределы $\pm 59,2$ в первой серии и $\pm 62,0$ во второй серии, то получим в первом случае 15 отклонений, а во втором только 13. Таким образом, «более точная» первая серия опытов, хотя она и имеет меньшее стандартное отклонение, дает большее число крупных ошибок и поэтому, конечно, не является более точной. Следовательно, рендомизированное расположение вариантов опыта, вопреки мнению Н. Ф. Деревницкого и других, не приводит к потере точности опыта и в то же время позволяет дать вполне обоснованную и правильную оценку точности опыта.

Можно было бы привести еще целый ряд соображений и данных, говорящих в пользу рендомизации, однако рамки настоящего предисловия не позволяют углублять этот вопрос. Все же можно надеяться, что читатель, ознакомившись с приведенными здесь положениями, убедится, что вопрос о рендомизации достаточно серьезен и что он не простая выдумка незнакомого с практикой математика. Конечно, нельзя рассматривать рендомизацию как жесткое и обязательное требование ко всякому эксперименту. Могут быть условия (например, при проведении некоторых агротехнических опытов), когда рендомизация технически невозможна. Точно так же нельзя считать, что опыт без рендомизации теряет всю свою ценность; отсутствие рендомизации наносит только некоторый ущерб чистоте опыта и в какой-то мере понижает его безукоризненность. Если исследователь дорожит этой чистотой и если к тому есть все условия, то он, невзирая на некоторые дополнительные хлопоты, должен ради придания своим результатам большей объективности пойти на применение в своих опытах рендомизации.

В. Н. Перегудов

ПРЕДИСЛОВИЕ АВТОРА

В последние годы наблюдается поразительное развитие статистики; непрерывно возникают новые теории и новые практические приложения ее; не прекращается спрос на разработку новых экспериментальных схем. Для удовлетворения запросов лиц, применяющих статистику в биологических исследованиях, в это, пятое, издание я включил целый ряд новейших методов, которые обещают быть весьма полезными в экспериментальной работе.

Как и в предыдущих изданиях, книга рассчитана на две категории читателей: на начинающих и на опытных исследователей-биологов. Для первой группы читателей начальные части всех глав вполне доступны, по некоторым вопросам даны подробные объяснения. Руководящие указания, вставленные в текст, отсылают этих читателей к следующим элементарно изложенным частям книги; кроме того, они могут воспользоваться оглавлением такого краткого курса, приведенным перед основным оглавлением. Имея в виду потребности второй группы читателей, я при подборе методов руководствовался своей практикой в качестве консультанта по применению статистики в опытных делах. При этом особое внимание уделялось смысловому содержанию и ограничениям различных схем обработки данных. Сами же расчеты изложены так, что они могут быть выполнены простым вычислителем. Для более углубленного знакомства с предметом в конце каждой главы приводится список литературы, который иногда по необходимости является сокращенным.

Наиболее заметным изменением текста в настоящем издании является введение главы 12, посвященной многофакторным опытам. Здесь я собрал соответствующие методы, прежде рассеянные по всей книге, и несколько дополнил их. Я надеюсь, что это будет признано полезным многими исследователями, которые в своих опытах применяют факториальное расположение вариантов, но не могут извлечь всю ту информацию, которая содержится в результатах этих опытов.

Сложные расчеты могут отвлекать внимание читателя от самого содержания статистической задачи, поэтому я пытался, особенно в начальной части каждой главы, облегчить тяготу вычислительной работы и стремился центрировать внимание на биологической сущности вопросов.

Нечто новое состоит также в том, что центр тяжести вероятностных суждений переносится с критериев существенности на оценки значения и интервала. В какой мере эта тенденция продолжится, будет видно из дальнейшего. Однако в настоящее время вполне очевидно, что такие оценки часто дают большую информацию, чем критерии существенности.

Для ознакомления читателя с многочисленной группой новых непараметрических критериев в книге приведено описание различных приложений некоторых из этих методов и произведено сравнение результатов, которые они дают, с результатами более привычных методов.

Представлялось желательным ввести некоторые изменения в обозначения и терминологию с целью облегчить учащемуся пользование текущей журнальной литературой. Большинство этих изменений направлены на подчеркивание различий между параметрами совокупности и статистическими характеристиками выборки. Но я везде ставил перед собой задачу по возможности уменьшить зависимость читателя-нематематика от алгебраической символики. обстоятельные объяснения и указания по поводу системы расчетов в большинстве случаев достаточны для того, чтобы обойтись без формул.

Я счастлив тем, что имею возможность включить в книгу новую главу 17 под названием «Планирование и анализ выборочных наблюдений», написанную профессором Уильямом Дж. Кокраном, являющимся признанным авторитетом в этой быстро развивающейся области. Подобранный профессором Кокраном иллюстративный материал находится в полном соответствии с биологической ориентацией данной книги.

За любезно предоставленное мне разрешение воспроизвести опубликованные ими таблицы я приношу свою признательность сэру Рональду Фишеру и его издателям Оливеру и Бойду из Эдинбурга, Мэксин Меррингтон, Кэтрин М. Томсон, Джойс М. Мей и Э. Лорду, чьи таблицы были опубликованы редактором «Биометрика», доктором Э. С. Пирсоном, а также Бернарду Остлу и его издателю — «Издательство колледжа штата Айова».

Без того искреннего интереса, который был проявлен моими коллегами из среды статистиков и биологов, и без сотрудничества с ними эта книга не могла бы быть написана. Статистики с готовностью дали ряд советов и конструктивных критических замечаний. Экспериментаторы внесли свою долю, предоставив мне консультацию по их специальности, и разрешили мне пользоваться их цифровым материалом. Если мне удалось правильно изложить их указания, то накопленный ими опыт будет полезным для тех лиц, которые только вступают в область исследования.

Джордж У. Снедекор

Статистическая лаборатория
Колледжа штата Айова
Июнь 1956 г.

ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ КАЧЕСТВЕННЫХ ПРИЗНАКОВ. БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

1. Выборочное наблюдение. В настоящее время каждый человек так или иначе соприкасается с выборочным наблюдением. Газеты публикуют и обсуждают отклики общественного мнения по тому или другому вопросу. Статистический контроль качества, проводимый выборочным методом, находит широкое применение в промышленности в целях сохранения однородности выпускаемой продукции. Наблюдения над объемом потребления в одинаковой мере необходимы как для производства, так и для торговли. Статистическое бюро имеет хорошо разработанную программу выборочных наблюдений, которые во многом заменяют общие переписи, происходящие через каждые 10 лет.

В меньшей степени известны, но не менее важны другие многочисленные случаи практического применения выборочных наблюдений. Например, наблюдение над дюжиной крыс может открыть много важных фактов из жизни больших масс этих животных, достигающих миллионов особей. Партия угля может быть принята или отвергнута на основе изучения его качества по пробе, составляющей только несколько фунтов. Врач делает свои заключения о крови пациента на основе анализа одной капли ее. Эта практика выборочных наблюдений столь распространена и имеет такое большое значение, что заслуживает гораздо большего внимания, чем простое и популярное упоминание о ней.

Выборочное наблюдение может проходить в широких территориальных границах, как, например, при попытке предсказать исход ближайших выборов, или оно может ограничиться площадью небольшого числа опытных делянок, или даже только полем зрения микроскопа. В этой книге, кроме ее последней главы, мы будем иметь дело с выборочным наблюдением (средним по размеру), применяемым в экспериментальных работах, в которых различные сопутствующие условия более или менее контролируются.

2. Две основные задачи выборочного наблюдения. *Выборка* состоит из небольшого количества единиц, взятого из некоторого более обширного собрания объектов, в отношении которого мы желаем получить ту или иную информацию. Эта выборка исследуется, и полученные этим путем факты изучаются. Задача выборочного метода состоит в том, чтобы, основываясь на этих фактах, сделать правильные выводы относительно всего собрания объектов, или *совокупности* их. Таким образом, имеется выборка, которую мы наблюдаем, и в то же время имеется совокупность, с которой мы стремимся познакомиться при помощи этой выборки.

Данная задача не возникает, когда нет никакого варьирования объектов. Если все объекты подобны друг другу, то выборка, состоящая только из одного из них, дает полную информацию о всей совокупности. В действительности всегда имеется бесконечное разнообразие как самих объектов, так и окружающих их условий. Вследствие этого внутри ряда последователь-

ных выборок всегда существует некоторое различие. Поэтому факты, наблюдаемые на основе выборки, не могут быть безоговорочно отнесены ко всей совокупности. В этом случае наша задача состоит в том, чтобы, несмотря на выборочное варьирование, все же сделать максимально обоснованные выводы относительно данной совокупности.

Но не каждая выборка содержит в себе информацию относительно той совокупности, из которой она взята. Допустим, что предметом выборочного наблюдения является изучаемая в опыте скорость роста совокупности молодых мышей на новом рационе. Для проведения этого эксперимента 10 животных посажены в клетку. Но клетка с ними помещена на холодном сквозняке или в темном углу; или может быть так, что среди мышей в этой клетке распространится не замеченная вначале инфекция. Если что-либо подобное случится, то скорость роста, установленная по выборке животных, не даст сколько-нибудь заслуживающей внимания информации относительно скорости роста совокупности нормальных мышей. Еще пример: допустим, что некоторый информатор об общественном мнении выбирает для опроса только семьи своих друзей, для которых, как он думает, его визит будет приятен. Такая выборка не будет представлять всех мнений общества. Все это приводит нас ко второй задаче: отбирать выборку так, чтобы в ней содержалась бы искомая информация.

Итак, мы ставим перед исследователем двойную задачу: необходимо планировать выборочное наблюдение так, чтобы оно представляло соответствующую совокупность; после чего следует строить на основе изучения выборки правильные выводы относительно совокупности, из которой взята данная выборка.

3. Выборка данных о фермах. Оценки значения и интервала. В 1950 г. отдел изучения насекомых — вредителей зерновых и кормовых культур Министерства сельского хозяйства США совместно с сельскохозяйственной опытной станцией штата Айова провели широкое выборочное наблюдение в округе Бун (Айова) в целях изучения мероприятий по борьбе с европейским кукурузным мотыльком [16]. Один из вопросов программы требовал определения размеров площади, на которой необходимо провести опрыскивание или опыливание для уничтожения этого вредителя. Для решения этого вопроса была взята случайная выборка из 100 фермеров, которые и были опрошены; 23 из них сказали, что они применяют на своих полях меры борьбы с вредителем. Таковы факты, которые дала нам выборка.

Какие на основе этого можно сделать выводы по отношению 2300 фермеров округа Бун? Может быть два вида таких выводов. Первому из них приписывается наименование *оценка значения*, в то время как второй называется *оценкой интервала*.

а) *Оценкой значения* для доли фермеров, применяющих опрыскивание, является 23 %, т. е. отношение внутри выборки; это означает, что, по нашей оценке, 23 % фермеров округа Бун в 1950 г. производили опрыскивание своих полей. Данная величина может рассматриваться как среднее число фермеров, применяющих опрыскивание, из каждых 100 фермеров. На основе фактического подсчета числа фермеров, опрыскивавших посевы кукурузы, в одной группе из 100 фермеров делается вывод, что среднее число таких фермеров во всех возможных выборках по 100 составляет 23.

Такой переход от выборки к совокупности обычно считается вполне допустимым. Многие, не задумываясь, идут непосредственно от фактов, установленных по выборке, к таким же выводам, относящимся ко всей совокупности. Однако логически между тем и другим существует различие. Действительно, может быть так, что указанный вывод будет неправильным, хотя выборка была произведена безукоризненно. Поэтому будет разумным, прежде чем приписывать всей совокупности процент, установленный по выборке, исследовать сам процесс выборочного наблюдения.

б) *Оценка интервала* для данного пункта производится при помощи таблицы 1. В первой части этой таблицы, имеющей в заголовке «95%-ный

95%-ный доверительный интервал (в %) для биномиального распределения

Число наблюдений f	Размер выборки n						Доля признака в выборке f/n	Размер выборки	
	10	15	20	30	50	100		250	1000
0	0 31	0 22	0 17	0 12	0 07	0 4	0,00	0 1	0 0
1	0 45	0 32	0 25	0 17	0 11	0 5	0,01	0 4	0 2
2	3 56	2 40	1 31	1 22	0 14	0 7	0,02	1 5	1 3
3	7 65	4 48	3 38	2 27	1 17	1 8	0,03	1 6	2 4
4	12 74	8 55	6 44	4 31	2 19	1 10	0,04	2 7	3 5
5	19 81	12 62	9 49	6 35	3 22	2 11	0,05	3 9	4 7
6	26 88	16 68	12 54	8 39	5 24	2 12	0,06	3 10	5 8
7	35 93	21 73	15 59	10 43	6 27	3 14	0,07	4 11	6 9
8	44 97	27 79	19 64	12 46	7 29	4 15	0,08	5 12	6 10
9	55 100	32 84	23 68	15 50	9 31	4 16	0,09	6 13	7 11
10	69 100	38 88	27 73	17 53	10 34	5 18	0,10	7 14	8 12
11		45 92	32 77	20 56	12 36	5 19	0,11	7 16	9 13
12		52 96	36 81	23 60	13 38	6 20	0,12	8 17	10 14
13		60 98	41 85	25 63	15 41	7 21	0,13	9 18	11 15
14		68 100	46 88	28 66	16 43	8 22	0,14	10 19	12 16
15		78 100	51 91	31 69	18 44	9 24	0,15	10 20	13 17
16			56 94	34 72	20 46	9 25	0,16	11 21	14 18
17			62 97	37 75	21 48	10 26	0,17	12 22	15 19
18			69 99	40 77	23 50	11 27	0,18	13 23	16 21
19			75 100	44 80	25 53	12 28	0,19	14 24	17 22
20			83 100	47 83	27 55	13 29	0,20	15 26	18 23
21				50 85	28 57	14 30	0,21	16 27	19 24
22				54 88	30 59	14 31	0,22	17 28	19 25
23				57 90	32 61	15 32	0,23	18 29	20 26
24				61 92	34 63	16 33	0,24	19 30	21 27
25				65 94	36 64	17 35	0,25	20 31	22 28
26				69 96	37 66	18 36	0,26	20 32	23 29
27				73 98	39 68	19 37	0,27	21 33	24 30
28				78 99	41 70	19 38	0,28	22 34	25 31
29				83 100	43 72	20 39	0,29	23 35	26 32
30				88 100	45 73	21 40	0,30	24 36	27 33
31					47 75	22 41	0,31	25 37	28 34
32					50 77	23 42	0,32	26 38	29 35
33					52 79	24 43	0,33	27 39	30 36
34					54 80	25 44	0,34	28 40	31 37
35					56 82	26 45	0,35	29 41	32 38
36					57 84	27 46	0,36	30 42	33 39
37					59 85	28 47	0,37	31 43	34 40
38					62 87	28 48	0,38	32 44	35 41
39					64 88	29 49	0,39	33 45	36 42
40					66 90	30 50	0,40	34 46	37 43
41					69 91	31 51	0,41	35 47	38 44
42					71 93	32 52	0,42	36 48	39 45
43					73 94	33 53	0,43	37 49	40 46
44					76 95	34 54	0,44	38 50	41 47
45					78 97	35 55	0,45	39 51	42 48
46					81 98	36 56	0,46	40 52	43 49
47					83 99	37 57	0,47	41 53	44 50
48					86 100	38 58	0,48	42 54	45 51
49					89 100	39 59	0,49	43 55	46 52
50					93 100	40 60	0,50	44 56	47 53

¹ Если f превосходит 50, то следует брать за наблюдаемое число $100-f$ и вычитать каждый доверительный предел из 100.

² Если f/n превосходит 0,50, то за наблюдаемую долю следует брать $1,00-f/n$ и вычитать каждый доверительный предел из 100.

99%-ный доверительный интервал (в %) для биномиального распределения

Число наблюдений f	Размер выборки n						Доля признака в выборке f/n	Размер выборки	
	10	15	20	30	50	100		250	1000
0	0 41	0 30	0 23	0 16	0 10	0 5	0,00	0 2	0 1
1	0 54	0 40	0 32	0 22	0 14	0 7	0,01	0 5	0 2
2	1 65	1 49	1 39	0 28	0 17	0 9	0,02	1 6	1 3
3	4 74	2 56	2 45	1 32	1 20	0 10	0,03	1 7	2 4
4	8 81	5 63	4 51	3 36	1 23	1 12	0,04	2 9	3 6
5	13 87	8 69	6 56	4 40	2 26	1 13	0,05	2 10	3 7
6	19 92	12 74	8 61	6 44	3 29	2 14	0,06	3 11	4 8
7	26 96	16 79	11 66	8 48	4 31	2 16	0,07	3 13	5 9
8	35 99	21 84	15 70	10 52	6 33	3 17	0,08	4 14	6 10
9	46 100	26 88	18 74	12 55	7 36	3 18	0,09	5 15	7 12
10	59 100	31 92	22 78	14 58	8 38	4 19	0,10	6 16	8 13
11		37 95	26 82	16 62	10 40	4 20	0,11	6 17	9 14
12		44 98	30 85	18 65	11 43	5 21	0,12	7 18	9 15
13		51 99	34 89	21 68	12 45	6 23	0,13	8 19	10 16
14		0 100	39 92	24 71	14 47	6 24	0,14	9 20	11 17
15		70 100	44 94	26 74	15 49	7 26	0,15	9 22	12 18
16			49 96	29 76	17 51	8 27	0,16	10 23	13 19
17			55 98	32 79	18 53	9 29	0,17	11 24	14 20
18			61 99	35 82	20 55	9 30	0,18	12 25	15 21
19			68 100	38 84	21 57	10 31	0,19	13 26	16 22
20			77 100	42 86	23 59	11 32	0,20	14 27	17 23
21				45 88	24 61	12 33	0,21	15 28	18 24
22				48 90	26 63	12 34	0,22	16 30	19 26
23				52 92	28 65	13 35	0,23	17 31	20 27
24				56 94	29 67	14 36	0,24	18 32	21 28
25				60 96	31 69	15 38	0,25	18 33	22 29
26				64 97	33 71	16 39	0,26	19 34	22 30
27				68 99	35 72	16 40	0,27	20 35	23 31
28				72 100	37 74	17 41	0,28	21 36	24 32
29				78 100	39 76	18 42	0,29	22 37	25 33
30				84 100	41 77	19 43	0,30	23 38	26 34
31					43 79	20 44	0,31	24 39	27 35
32					45 80	21 45	0,32	25 40	28 36
33					47 82	21 46	0,33	26 41	29 37
34					49 83	22 47	0,34	26 42	30 38
35					51 85	23 48	0,35	27 43	31 39
36					53 86	24 49	0,36	28 44	32 40
37					55 88	25 50	0,37	29 45	33 41
38					57 89	26 51	0,38	30 46	34 42
39					60 90	27 52	0,39	31 47	35 43
40					62 92	28 53	0,40	32 48	36 44
41					64 93	29 54	0,41	33 50	37 45
42					67 94	29 55	0,42	34 51	38 46
43					69 96	30 56	0,43	35 52	39 47
44					71 97	31 57	0,44	36 53	40 48
45					74 98	32 58	0,45	37 54	41 49
46					77 99	33 59	0,46	38 55	42 50
47					80 99	34 60	0,47	39 55	43 51
48					83 100	35 61	0,48	40 56	44 52
49					86 100	36 62	0,49	41 57	45 53
50					90 100	37 63	0,50	42 58	46 54

1

2

2

¹ Если f превосходит 50, то следует брать за наблюдаемое число $100-f$ и вычитать каждый доверительный предел из 100.

² Если f/n превосходит 0,50, то за наблюдаемую долю следует брать $1,00-f/n$ и вычитать каждый доверительный предел из 100.

доверительный интервал», находим в верхней части размер выборки 100, а в левой колонке наблюдаемое число (или частота) — 23 фермера, применяющих опрыскивание. На пересечении соответствующих ряда и колонки находим числа 15 и 32. Смысл их таков: можно быть уверенным в том, что фактический процент фермеров, применяющих опрыскивание, в совокупности, из которой взята выборка, находится в интервале от 15 до 32%. Этот интервал называется *доверительным интервалом*. Сущность нашей уверенности, о которой здесь идет речь, будет установлена позже.

В итоге, основываясь на случайной выборке, мы прежде всего устанавливаем, что наша оценка процента фермеров, применяющих опрыскивание в округе Бун, дает 23%, но мы не можем указать на величину возможной ошибки этой оценки. Далее мы высказываем уверенность, что действительный процент этих фермеров отклоняется от нашей оценки — 23% — не более чем на 8% в сторону уменьшения и на 9% в сторону увеличения.

В связи с тем, что изложенная здесь концепция лежит в основе теории выборочного метода, изложим ее по-другому. Представим себе мешок, наполненный фасолью белого цвета и других цветов; семена тщательно перемешаны. Погружая наугад в этот мешок щуп и подсчитывая число семян фасоли каждого цвета, находим, что белые семена составляют, положим, 40%. Это число является не только процентом белых семян в выборке, но и оценкой доли этих семян во всем мешке. Каково содержание этой оценки? Это выясняется, когда мы переходим ко второй части выводов. Если в щуп попало 250 семян, то, справляясь в таблице 1, мы можем высказать уверенность, что процент белых семян фасоли в этом мешке находится в интервале между 34 и 46%.

Пока мы не касались меры той уверенности, с которой делается эта вторая часть наших выводов. Заголовок таблицы «95%-ный доверительный интервал» указывает на степень этой уверенности, что может быть выражено следующим образом. Если такое выборочное наблюдение продолжить до бесконечности, то каждая выборка даст свой особый доверительный интервал (т. е. новую оценку интервала); в этом случае 95% интервалов, построенных на основе выборок, покроет собой фактическое значение процента в совокупности. Можно сказать еще и так: если производится указанный процесс отбора и если для каждой выборки устанавливается, что процент в совокупности находится внутри соответствующего доверительного интервала, то около 95% наших утверждений будут правильными. Другие, более краткие определения будут даны позднее. Обычно достаточно просто указать, что вы вполне уверены в том, что оценка интервала покрывает собой оцениваемую долю совокупности.

Если вам кажется, что формулировка этого второго вывода при 5%-ном шансе сделать неправильное утверждение недостаточна и опасна, то вы можете использовать вторую часть таблицы 1, озаглавленную «99%-ный доверительный интервал». Для выборочного наблюдения над фермами округа Бун этот интервал расширяется до 13—35%. Если кто-нибудь скажет, что процент, вычисленный по всей совокупности, находится внутри этих пределов, то он будет, вообще говоря, прав, если не считаться с одним шансом из 100 в пользу обратного случая, возможного при выборочном наблюдении.

В тех случаях, когда размер совокупности известен, как это имеет место в обследовании фермеров округа Бун, можно от процентного выражения оценок значения и интервала перейти к оценкам, дающим численности объектов. Так, мы можем определить число фермеров округа Бун, применявших в 1950 г. опрыскивание посевов, путем умножения: $0,23 \times 2300 = 529$ фермеров. Этим же способом для 95%-ного доверительного интервала находим: от 345 до 736 фермеров.

В этом кратком введении я выставил на показ добрую часть того товара, который предлагается статистикой: взятие выборки из совокупности, изучение фактов, сообщаемых выборкой, и формулировка на основе этих фактов

выводов относительно совокупности, из которой взята выборка. Прежде чем идти дальше, вы можете закрепить эти свои знания, проделав несколько упражнений.

Примеры составляют существенную часть моего изложения статистики. Каждая группа примеров построена так, чтобы вы могли начать с наиболее легких. Следует отметить, что некоторое число примеров в каждой группе можно решить сразу после первого ознакомления с текстом, оставая более трудные из них до тех пор, пока не будет накоплен некоторый опыт. Я твердо убежден в том, что статистика не может быть хорошо изучена без таких или подобных им примеров.

Пример 1. На основе выборки в 100 семян, взятой случайным порядком из мешка с однородной смесью семян мятлики, при помощи стандартной методики установлена всхожесть семян в 92%. Контролер желает сделать заключение относительно доли семян в мешке, которые будут всхожими при тех же условиях, причем он может идти на риск дать ошибочное заключение только в одном случае из 100. Что он может сказать при этих условиях? *Ответ:* всхожесть между 83 и 97%. Указание: при использовании таблицы следует взять $100 - 92 = 8\%$.

Пример 2. Если контролер семян предыдущего примера возьмет для испытания 1000 семян вместо 100 и найдет среди них только 8% невсхожих, какое заключение он может сделать при шансах 99 против 1? *Ответ:* всхожесть между 90 и 94%.

Пример 3. Нский исследователь в процессе широкого национального обследования опросил 115 женщин примерно 40-летнего возраста из сельских местностей средне-западных штатов, имеющих достаток ниже среднего уровня. Срок шесть из них в течение предыдущего месяца три или большее число раз слушали по радио ту или иную программу. Предполагая выборку случайной и используя 99%-ный доверительный интервал, определить, какого рода утверждение можно сделать относительно доли слушателей радио во всей совокупности такой категории женщин. *Ответ:* приблизительно между 28,4 и 52,5% слушателей. Эти значения вы должны будете получить путем интерполяции данных таблицы.

Пример 4. Лицо, собирающее сведения об общественном мнении, попросило 50 мужчин сказать, кого они предпочитают из двух кандидатов — А или В. Двадцать из них предпочли кандидата А. Считая эту выборку случайной из совокупности в 5000 человек, это лицо утверждает, что из этого числа 1150—2950 человек отдадут предпочтение кандидату А. Каким доверительным интервалом это лицо в данном случае пользовалось? *Ответ:* 99%-ным доверительным интервалом.

Пример 5. При случайном отборе из некоторой совокупности фермеров 86% из числа опрошенных фермеров отозвались одобительно в отношении некоторого агротехнического мероприятия. Основываясь на этой выборке, статистик утверждает, что имеется не больше как один шанс из 20 против того, что во всей совокупности фермеров, из которой взята выборка, процент фермеров, отдающих предпочтение данному мероприятию, находится между 81 и 90%. Каким был размер выборки? *Ответ:* 250.

Пример 6. Если у вас есть основание считать, что в некоторой совокупности домашних хозяйств от 25 до 75% их являются собственниками определенного хозяйственного приспособления, и если вы желаете из нее взять выборку так, чтобы она при 95%-ном уровне доверия давала бы отклонение от фактического процента этих собственников не больше чем на 6%, то спрашивается: какого размера выборку вы должны взять? *Ответ:* 250.

Пример 7. Если вы надлежащим образом произвели некоторую выборку, после чего сделали заключение по поводу доли признака в совокупности, то это заключение может оказаться как правильным, так и неправильным; причем вы не знаете, будет ли оно в данном случае тем или другим. Представляете ли вы ясно, с каким риском власть в ошибку сопряжено это ваше утверждение?

Пример 8. При обследовании некоторой общины было установлено [12], что среди 978 детей до 16-летнего возраста было 54,4% девочек. Определить 95%-ный доверительный интервал, заключающий в себе действительное отношение полов среди данной категории детей.

Пример 9. Если в выборке 100 фермеров округа Бун не окажется ни одного, который применял бы опрыскивание посевов, то какое 95%-ное доверительное утверждение вы сделаете относительно фермеров этого округа? *Ответ:* число фермеров, применяющих опрыскивание, содержится в интервале между 0 и 4%. Теперь допустим, что все фермеры, попавшие в выборку, применяли опрыскивание. Каким будет 99%-ный доверительный интервал? *Ответ:* 95—100%.

4. Определения и обсуждение некоторых вопросов. Размер выборки. Я так стремительно провел вас через три первых параграфа с тем, чтобы доставить вам удовольствие сразу войти в круг идей и приложений статистики. Однако теперь ряд положений нуждается в значительном уточнении. Некоторые из них будут обсуждены в настоящем параграфе, другие — в последующем изложении.

Часть заголовка настоящей главы содержит в себе фразу: «Выборочное наблюдение качественных признаков». Ранее мы имели дело с такими качественными признаками, как опрыскивание посевов кукурузы, всхожесть семян, одобрение агротехнического мероприятия и интерес к радиовещанию. В связи с наличием или отсутствием качественного признака образуются два класса объектов, составляющих изучаемую нами совокупность. В результате выборки устанавливаются количества объектов, у которых наблюдается или нет исследуемый признак. Я назову эти результаты *данными подсчета* в отличие от результатов измерений таких переменных величин, как рост или возраст. Теория и методы обработки таких *данных измерения* будут изложены в дальнейшем, начиная с главы 2. Для того чтобы выборка содержала в себе достаточно надежную информацию о совокупности, необходимо, чтобы каждый член выборки был отобран *случайно*. Случайный отбор предполагает, что каждый член совокупности имеет одну и ту же определенную вероятность попасть в выборку. Это обстоятельство не должно упускаться из виду лицом, производящим выборку; если это лицо употребит некоторый нарочитый отбор, то указанная вероятность в процессе такого отбора становится неопределенной, и вся теория теряет свою силу. Это положение относительно случайности отбора приводит к более строгому определению понятия *совокупность*. Теперь можно сказать, что совокупность является таким собранием объектов, каждый член которого имеет известную вероятность попасть в выборку.

Для того чтобы достигнуть случайности отбора, существует два основных способа: или сама совокупность может быть полностью перемешана, или в основу отбора членов выборки может быть положен некоторый механический процесс, который не находится под контролем лица, производящего выборку. Равномерно перемешанная совокупность, например, получается при растирании и размешивании кормов или каких-либо химических продуктов перед производством отбора образцов или создается естественным перемешиванием, как это имеет место в потоке крови. Второй способ обеспечения случайности имеет более широкое применение. Например, при производстве выборки в округе Бун был использован план округа по районам, в котором указано местоположение каждой фермы. Фермы были занумерованы, и случайная выборка сводилась к отбору 100 из этих номеров наугад при размешивании их в ящике. Фактически в данном случае была применена схема, известная под названием *последовательная случайная выборка*. Из ферм каждого района была взята случайная выборка, величина которой была пропорциональна числу ферм в этом районе. Главное преимущество этого способа состоит в том, что выборка распределяется более равномерно по всему округу, сохраняя принцип случайности внутри каждого района. Статистические методы обработки данных последовательных выборок даны в главе 17. Соответствующие выводы только немного изменяются по сравнению с заключениями, основанными на полностью случайной выборке.

Полная случайность отбора на практике, возможно, никогда не осуществима. Однако она является той математической моделью, на которую должна опираться теория статистики; в связи с тем, что эта теория лежит в основе выводов, которые делаются при повседневном проведении наблюдений, в интересах самого исследователя по возможности приблизить практику отбора к этим идеальным условиям. Чем лучше будет это приближение, тем более правильными будут извлекаемые из выборочных данных выводы.

Наиболее частым вопросом, который задают экспериментаторы и другие исследователи, пользующиеся выборочными наблюдениями, является: «Сколь большую выборку я должен взять?». Ответы будут различными в зависимости от различных условий и всегда будут содержать в себе некоторый элемент риска. Однако в дальнейшем мы постараемся внести ясность в эти сложные вопросы. Вы, вероятно, обратили внимание на особенности постановки вопроса в примере 6, где считались данными как доверительный интервал, так и предполагаемый процент признака в совокупности. Возможно,

вы также заметили, что эти же данные требуются и при использовании правой и левой части таблицы 1; здесь не только происходит быстрое изменение доверительных пределов в зависимости от изменения неизвестной доли признака f , но и ширина самого интервала не остается постоянной при различных значениях этой доли. Несмотря на эти особенности таблицы, она может служить отправным пунктом приближенных решений, вполне достаточных для практических целей. Всякая попытка абсолютно точно определить размер выборки бесполезна, так как неизбежное варьирование между выборками уничтожит самые хорошие планы. В данный момент я рекомендую обращаться к этой таблице с наиболее точными данными, которые вам доступны, т. е. иметь выборку такого большого размера, какую только вы способны произвести. Если же вы не можете взять столь большую выборку и в то же время не можете примириться со слишком большим доверительным интервалом, то вам ничего не остается, как отказаться от выборочного наблюдения и тем самым воздержаться от ненужных затрат времени и денег.

По поводу оценки интервала следует особо подчеркнуть два момента. Во-первых, доверительное утверждение является утверждением об отношении признаков внутри совокупности, а не об отношении их в других возможных выборках. Во-вторых, некоторая неопределенность выводов обусловлена самим процессом образования выборки. Каждая выборка дает свой особый интервал оценки. Будет или не будет этот интервал заключать в себе фактическую долю признака в совокупности, — все дело случая. Мы только теоретически знаем, что интервал при уровне доверия 0,95 определен так, что 95 % таких интервалов будут покрывать фактическое значение доли признака.

Отсюда следует правило: перед тем как производить выборку, каждый должен определить *вероятность* правильности своего будущего доверительного утверждения. Он должен сказать себе: «Я желаю произвести случайную выборку и дать на основе ее оценку интервала, при этом беру вероятность 0,95 того, что интервал покроет долю признака в совокупности». После того как выборка произведена и правильность или неправильность доверительного утверждения стала осуществившимся фактом, бессмысленны какие-либо новые заявления по поводу этой вероятности.

Иногда совокупность может состоять из объектов не двух, а большего числа видов. Примером этого может служить выборочное изучение общественного мнения, когда обычно вместо двух видов ответов предлагается три категории их: «да», «нет» и «не знаю». Точно так же, при изучении наследственных признаков обычно приходится иметь дело с совокупностями, состоящими из нескольких видов потомков. В этих условиях для формулировки доверительных утверждений по поводу каждого отдельного вида признака необходимо основываться уже не на биномиальном, а на мультиномиальном распределении, которое в этой книге не рассматривается. Однако такой полный ряд доверительных утверждений не всегда бывает необходимым. Например, можно при опросе ограничиться только одним видом ответа, положим «да». И оставить задачу построения доверительного интервала только для него. В этом случае такой интервал оценивает «да», отделенное от всех других ответов, в результате чего образуется две категории: «да» и «не да». Оценка интервала здесь будет произведена на основе общего размера выборки и числа ответов «да».

5. Упражнение по проведению выборки. Всякий раз, когда мне приходится производить выборку из определенной совокупности, я удивляюсь капризности той последовательности, с которой идут друг за другом члены выборки. Экспериментатору или лицу, производящему подсчеты, полезно иногда познакомиться с законами изменения таких рядов, чтобы не слишком доверять своим выборочным данным. Классическими способами получения выборки являются: извлечение шаров из урны, карт из колоды, подбрасывание монеты и выкидывание костей. В прежние времена я моим слушателям рекомендовал производить выборки путем извлечения из мешка помеченных или замурованных семян фасоли, но теперь можно применить более

Десять тысяч случайных цифр

	00—04	05—09	10—14	15—19	20—24	25—29	30—34	35—39	40—44	45—49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	46756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59344	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	96801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89372	45375	85436
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	70046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39035	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	63706	61203	53634	22557
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716

	50—54	55—59	60—64	65—69	70—74	75—79	80—84	85—89	90—94	95—99
00	59391	58030	52098	82718	87024	82848	04190	93574	90464	29065
01	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	86859	19558	64432	16706	99612	59798	32803	67703	15297	28612
04	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	95068	88628	35911	14530	33020	80428	39936	31855	34334	64835
06	54463	47237	73800	91017	36239	71824	83671	39892	60518	37092
07	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	22518	55576	98215	82068	10798	86211	36584	67466	69373	40054
17	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	80327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	57430	82270	10421	05540	43648	75888	66049	21511	47676	33444
21	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	78388	16638	09134	59880	63806	48472	39318	35434	24057	74739
24	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
25	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	76970	80876	10237	39515	79152	74798	39357	09054	73579	92359
27	37074	65198	44785	68624	98336	84481	97610	78735	46703	93265
28	83712	06514	30101	78295	54653	85417	43189	60048	72781	72606
29	20287	56862	69727	94443	64936	08366	27227	05158	50326	59566
30	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	64081	49863	08478	96001	18888	14810	70545	89755	59034	07210
32	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	65988	72850	48737	54719	52056	01596	03845	35067	03134	70322
35	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	56760	10909	98147	34736	33863	95253	12731	66598	50771	83665
37	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	77888	38100	03052	58103	47961	83841	25878	23746	55903	44115
39	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873

	00—04	05—09	10—14	15—19	20—24	25—29	30—34	35—39	40—44	45—49
40	79152	53829	77250	20199	56535	18760	69942	77448	33278	48805
41	44560	38750	83635	56540	64930	42912	13953	79149	18710	68618
42	68328	83378	63369	71381	39534	05615	42451	64559	97501	65747
43	43939	38689	58625	08342	30459	85863	20781	09284	26333	91777
44	83544	83141	15707	96253	23068	13782	08467	89469	93842	55349
45	91621	03881	04900	54224	43177	55309	17852	27491	89415	23466
46	91896	67123	04151	03795	59077	11848	12630	98375	52068	60142
47	55751	62515	21108	80830	02233	29303	37204	96926	30506	09808
48	85156	87689	95493	88842	00634	55017	55539	17771	69448	87530
49	07521	53898	12236	60277	39102	62315	12239	07105	11844	01117
50	64249	63664	39652	40646	97303	31741	07294	84149	46797	82487
51	26538	44249	04050	48174	65570	44072	40192	51153	11397	58212
52	05845	00512	78630	55328	18116	69296	91705	86224	29503	57071
53	74897	68373	67359	51014	33510	83048	17056	72506	82949	54600
54	20872	54570	35017	88132	25730	22626	86723	91691	13191	77212
55	31432	96156	89177	75541	81355	24480	77243	76690	42507	84362
56	66890	61505	01240	00660	05873	13568	76082	79172	57913	93448
57	48194	57790	79970	33106	86994	48119	52503	24130	72824	21627
58	11303	87118	81471	52936	08555	28420	49416	44448	04269	27029
59	54374	57325	16947	45356	78371	10563	97191	53798	12693	27928
60	64852	34421	61046	90849	13966	39810	42699	21753	76192	10508
61	16309	20384	09491	91588	97720	89846	30376	76970	23063	35894
62	42587	37065	24526	72602	57589	98131	37292	05967	26002	51945
63	49177	98590	97161	41682	84533	67588	62036	49967	01990	72308
64	82309	76128	93965	26743	24141	04838	40254	26065	07938	76236
65	79788	68243	59732	04257	27084	14743	17520	95401	55811	76099
66	40538	79000	89559	25026	42274	23489	34502	75508	06059	86682
67	64916	73598	18609	73150	62463	33102	45205	87440	96767	67042
68	49767	12691	17903	93871	99721	79109	09425	26904	07419	76013
69	76974	55108	29795	08404	82684	00497	51126	79935	57450	55671
70	23854	08480	85983	96025	50117	64610	99425	62291	86943	21541
71	68973	70551	25098	78033	98573	79848	31778	29555	61446	23037
72	36444	93600	63350	14971	25325	00427	52073	64280	18847	24768
73	03003	87800	07391	11594	21196	00781	32550	57158	58887	73041
74	17540	26188	36647	78386	04558	61463	57842	90382	77019	24210
75	38916	55809	47982	41968	69760	79422	80154	91486	19180	15100
76	64288	19843	69122	42502	48508	28820	59933	72998	99942	10515
77	86809	51564	38040	39418	49915	19000	58050	16899	79952	57849
78	99800	99566	14742	05028	36033	94889	53381	23656	75787	59223
79	92345	31890	95712	08279	91794	94068	49337	88674	35355	12267

	50—54	55—59	60—64	65—69	70—74	75—79	80—84	85—89	90—94	95—99
40	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	47606	93410	16359	89033	89696	47231	64498	31778	05383	39902
42	52639	45030	98279	14709	52372	87832	02735	50803	72744	83208
43	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	75033	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	13636	93593	23377	51133	95126	61493	42474	45141	46660	42338
50	32847	31282	03345	89593	69214	70381	78285	20054	91018	16742
51	16916	00041	30236	55023	14253	76582	12092	86532	92426	37655
52	66176	34047	21005	27137	03191	48970	64625	22394	39622	79085
53	46299	13335	12180	16861	38043	59292	62675	63631	37020	78195
54	22847	47839	45385	23289	47526	54098	45683	55849	51575	64689
55	41851	54160	92320	69936	34803	92479	33299	71160	64777	83378
56	28444	59497	91586	95917	68553	28639	06455	34174	11130	91994
57	47520	62378	98855	83174	13088	16561	68559	26679	06238	51254
58	34978	63271	13142	82681	05271	08822	06490	44984	49307	62717
59	37404	80416	69035	92980	49486	74378	75610	74976	70056	15478
60	32400	65482	52099	53676	74648	94148	65095	69597	52771	71551
61	89262	86332	51718	70663	11623	29834	79820	73002	84886	03591
62	86866	09127	98021	03871	27789	58444	44832	36505	40672	30180
63	90814	14833	08759	74645	05046	94056	99094	65091	32663	73040
64	19192	82756	20553	58446	55376	88914	75096	26119	83898	43816
65	77585	52593	56612	95766	10019	29531	73064	20953	53523	58136
66	23757	16364	05096	03192	62386	45389	85332	18877	55710	96459
67	45989	96257	23850	26216	23309	21526	07425	50254	19455	29315
68	92970	94243	07316	41467	64837	52406	25225	51553	31220	14032
69	74346	59596	40088	98176	17896	86900	20249	77753	19099	48885
70	87646	41309	27636	45153	29988	94770	07255	70908	05340	99751
71	50099	71038	45146	06146	55211	99429	43169	66259	97786	59180
72	10127	46900	64984	75348	04115	33624	68774	60013	35515	62556
73	67995	81977	18984	64091	02785	27762	42529	97144	80407	64524
74	26304	80217	84934	82657	69291	35297	98714	35104	08187	48109
75	81994	41670	56642	64091	31229	02595	13513	45148	78722	30144
76	59537	34632	79631	89403	65212	09975	06118	86197	58208	16162
77	51228	10937	62396	81460	47331	91403	95007	06047	16846	64809
78	31689	37995	29577	07828	42272	54016	21950	86192	99046	84864
79	38207	97938	93459	75174	79460	55436	57206	87644	21296	43395

	00—04	05—09	10—14	15—19	20—24	25—29	30—34	35—39	40—44	45—49
80	90363	65162	32245	82279	79256	80834	03088	99462	56705	06118
81	64437	32242	48431	04835	39070	59702	31508	60935	22390	52246
82	91714	53662	28373	34333	55791	74758	51144	18827	10704	76803
83	20902	17646	31391	31459	33315	03444	55743	74701	58851	27427
84	12217	86007	70371	52281	14510	76094	96579	54853	78339	20839
85	45177	02863	42307	53571	22532	74921	17735	42201	80540	54721
86	28325	90814	08804	52746	47913	54577	47525	77705	95330	21866
87	29019	28776	56116	54791	64604	08815	46049	71486	34650	14994
88	84979	81353	56219	67062	26146	82567	33122	14124	46240	92973
89	50371	26347	48513	63915	11158	25563	91915	18431	92978	11591
90	53422	06825	69711	67950	64716	18003	49581	45378	99878	61130
91	67453	35651	89316	41620	32048	70225	47597	33137	31443	51445
92	07294	85353	74819	23445	68237	07202	99515	62282	53809	26685
93	79544	00302	45338	16015	66613	88968	14595	63836	77716	79596
94	64144	85442	82060	46471	24162	39500	87351	36637	42833	71875
95	90919	11883	58318	00042	52402	28210	34075	33272	00840	73268
96	06670	57353	86275	92276	77591	46924	60839	55437	03183	13191
97	36634	93976	52062	83678	41256	60948	18685	48992	19462	96062
98	75101	72891	85745	67106	26010	62107	60885	37503	55461	71213
99	05112	71222	72654	51583	05228	62056	57390	42746	39272	96659

современный способ, используя таблицу случайных чисел, подобную таблице 2. Эта таблица содержит в себе 10 000 цифр, тщательно перемешанных и расположенных в случайном порядке; блоки из этих цифр 5×5 образованы просто для удобства пользования. Здесь имеется 100 рядов и 100 колонок, занумерованных от 00 до 99. Вместо того, чтобы изучать, будут ли горошины морщинистыми или гладкими, или вместо того, чтобы задавать населению вопросы: «да» или «нет», вы можете взять в качестве упражнения случайную выборку цифр из этой таблицы и определить, будет ли каждая из них нечетной или четной. Поскольку можно считать, что цифры полностью перемешаны и не группируются в каком-либо особом порядке, любая их последовательность может быть принята в качестве случайной; поэтому можно начать с любого места таблицы и прямо подсчитывать число нечетных цифр в выборке заранее установленного размера. Например, допустим, что вы взяли 50 за объем выборки и избрали за начало 31-й ряд и 17-ю колонку. Вы находите: 752 54450 19031 ... Продолжаете до конца ряда, после чего переходите к началу 32-го ряда и идете дальше, пока не наберете все 50 цифр, т. е. пока не закончите цифрами 33851 44. В этой частной выборке вы находите 23 нечетных числа; значит, в выборке из 50 цифр оказалось 46% нечетных.

Чтобы уменьшить вероятность повторения данной выборки, в других выборках следует основываться на таких трех элементах процесса отбора. 1) Каждую выборку следует начинать со случайно выбранного пункта. Удобный способ выбора начального пункта состоит в том, что вы наугад попадаете своим карандашом в какую-либо цифру таблицы и используете ее и три следующие за ней цифры как указание на начальный пункт вашей первой выборки. Допустим, например, что ваш карандаш натолкнулся на цифру 2 в 80-м ряду и в 84-й колонке. Эта цифра с последующими тремя дает 2061, что определяет пересечение 20-го ряда и 61-й колонки в качестве исходного пункта выборки. Вторую выборку можно начинать с пункта, определяемого

	50—54	55—59	60—64	65—69	70—74	75—79	80—84	85—89	90—94	95—99
80	88666	31142	09474	89712	63153	62333	22124	06140	42594	43671
81	53365	56134	67582	92557	89520	33452	05134	70628	27612	33738
82	89807	74530	38004	90102	11693	90257	05500	79920	62700	43325
83	18682	81038	85662	90915	91631	22223	91588	80774	07716	12548
84	63571	32579	63942	25371	09234	94592	98475	76884	37635	33608
85	68927	56492	67799	95398	77642	54913	91853	08424	81450	76229
86	56401	63186	39389	88798	31356	89235	97036	32341	33292	73757
87	24333	95603	02359	72942	46287	95382	08452	62862	97869	71775
88	17025	84202	95199	62272	06366	16175	97577	99304	41587	03686
89	02804	08253	52133	20224	68034	50865	57868	22343	55111	03607
90	08298	03879	20995	19850	73090	13191	18963	82244	78479	99121
91	59883	01785	82403	96062	03785	03488	12970	64896	38336	30030
92	46982	06682	62864	91837	74021	89094	39952	64158	79614	78235
93	31121	47266	07061	02051	67599	24471	69843	83696	71402	76287
94	97867	56641	63416	17577	30161	87320	37752	73276	48969	41915
95	57364	86746	08415	14621	49430	22311	15836	72492	49372	44103
96	09559	26263	69511	28064	75999	44540	13337	10918	79846	54809
97	53873	55571	00608	42661	91332	63956	74087	59008	47493	99581
98	35531	19162	86406	05299	77511	24311	57257	22826	77555	05941
99	28229	88629	25695	94932	30721	16197	78742	34974	97528	45447

четырьмя первыми (или четырьмя последними, или четырьмя какими-либо другими удобными для вас) цифрами первой выборки, и т. д. 2) Менять направление, по которому вы двигаетесь по таблице: то вправо, то влево, то вниз, то по диагонали. 3) Изменять размер выборки, производя тем же случайным способом выборки различного объема, например: 10, 15 ... 250, указанные в заголовке таблицы 1 (1000 — слишком велико для упражнения).

В отношении каждой выборки отмечается следующее: 1) начальный пункт; 2) направление, в котором идет подсчет; 3) размер выборки и 4) частота появления нечетных цифр. Если вы можете уделить время на производство указанных способом 100 или более выборок, при соблюдении тщательности записей, получите хороший материал для дальнейшего использования.

Проделав все это, вы имеете возможность экспериментально получить то, что математик-статистик получает теоретически, изучая распределение случайных выборок, взятых из определенной совокупности; вы тем самым можете проверить результаты, которые он получает из своих уравнений. Рассмотрим более подробно эти результаты.

6. Распределение частот и его графическое изображение. Одна группа лиц, изучающих статистику, получила 200 выборок при $n = 10$. Результаты сведены в таблицу *распределения частот* (табл. 3).

Каждое число нечетных цифр в выборке будет называться *классом*. Число выборок, попавших в некоторый класс, является *численностью класса* (частота); сумма численностей всех классов является общим количеством выборок. Вопросу о построении статистических таблиц посвящено целое учение; о подробностях смотри в литературе [13]. Я думаю, что большинство изучающих статистику легко освоит приведенную здесь таблицу распределения частот, пользуясь краткими указаниями или даже без них.

Одна из заметных особенностей выборочного распределения состоит в концентрации частот вблизи середины таблицы. Наибольшая числен-

Распределение количества нечетных цифр в 200 выборках при $n=10$

Класс (число нечетных цифр)	Численность класса	Теоретическая численность класса
0	1	0,2
1	1	1,8
2	8	8,2
3	25	22,3
4	39	40,0
5	45	49,2
6	36	42,0
7	25	24,6
8	16	9,4
9	4	2,1
10	0	0,2
Общая численность	200	200,0

ность соответствует классу при 5 нечетных цифрах, т. е. случаю, когда в выборке половина четных и половина нечетных цифр. Три срединных класса 4- 5- и 6-й содержат в себе $39 + 45 + 36 = 120$ выборков, т. е. более половины общего числа их. Эта центростремительная тенденция замечательна тем, что она внушает нам уверенность в отношении выборочного метода: большинство выборок будут давать довольно близкие оценки для доли нечетных цифр в совокупности. Данное положение является известным противовесом тому, может быть, обескураживающему факту, что некоторые из выборок имеют довольно значительные отклонения.

Другой интересной особенностью является симметричность распределения: численность в центре достигает наибольшего значения и последовательно снижается к обоим концам. Это происходит потому, что в данном случае доля нечетных цифр в совокупности около 50%; если бы доля признака в совокупности была около нуля или 100%, то численности сосредоточивались бы на одном конце или вблизи от него.

Регулярность, наблюдаемая в этом распределении, указывает на то, что случайные явления подчиняются определенному закону. Появление нечетных чисел при том случайном процессе, который здесь рассматривается, может показаться весьма неустойчивым фактом: будет ли следующая цифра четной или нечетной — это чисто случайное явление. Однако объединение многих таких явлений создает некоторую форму распределения, которую можно предугадать (исключая выборочное варьирование). Такое распределение, определенное теоретически и представленное в последней колонке таблицы 3, называется *биномиальным распределением*; оно определяет собой вид совокупности, из которой мы произвели выборочный отбор и которая имеет точно установленную форму.

Вычисление теоретического распределения для нашего случая, когда доля нечетных цифр 0,506 (так как в таблице содержится всего 5060 этих цифр) и объем выборки $n=10$, произведено способом, с которым мы познакомимся в главе 16. Я думаю, что вы вполне удовлетворены той согласованностью, которая имеет место между фактическими и теоретическими численностями.

Графическое изображение на рисунке 1 дает наглядное представление о биномиальном распределении. На горизонтальной оси нанесены интервалы, соответствующие классам данного распределения частот. Над каждым из них построен прямоугольник, высота которого пропорциональна соответствующей численности класса; шкала для этих численностей указана слева. Такой вид графика, представляющий распределение частот, называет

ся *гистограммой*. Применению графиков в статистике посвящено много работ и книг; работа [1], указанная в списке литературы, дает достаточно подробное описание техники графического изображения.

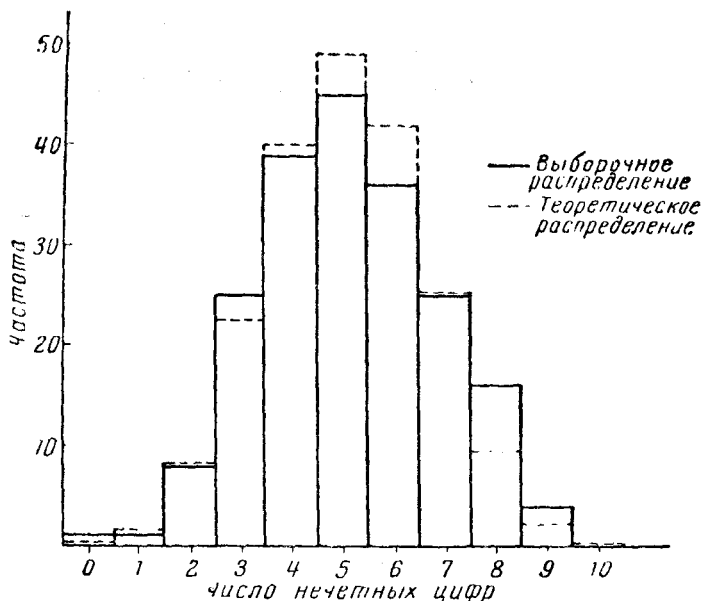


Рис. 1. Распределение нечетных цифр в 200 выборках по 10. Пунктирная линия изображает теоретическое биномиальное распределение, из которых взяты эти выборки.

7. Доверительный интервал; проверка теории. Выборки, которые вы произвели, можно использовать для проверки теории доверительного интервала. По данным таблицы 1 для каждой выборки определите 95 %-ный интервал и установите, будет ли им покрываться известная вам доля нечетных цифр в совокупности. Так, в выборке в 50 цифр, которая была произведена ранее, оказалось 23 нечетных цифры. На основе этого можно утверждать, что процент нечетных цифр во всей совокупности находится между 32 и 61%; это заключение будет правильным, так как нам известно, что во всей таблице содержится 50,6 % нечетных цифр. Но теперь допустим, что ваша выборка в 250 цифр начинается с 85-го ряда и 23-й колонки и что вы производите подсчет, двигаясь вниз от этого пункта и переходя к соседним колонкам, тогда окажется, что в этой выборке будет только 101 нечетная цифра, т. е. 40,4%; вы в этом случае будете утверждать, что фактическая доля нечетных цифр находится в интервале между 34 и 46 %. Теперь вы сделаете ошибочное заключение, несмотря на то, что эта выборка взята случайным порядком из той же совокупности, из которой взяты и другие выборки. Данная выборка чисто случайно оказалась сильно отклоняющейся. Вы встретите около 5 выборок на 100 таких, которые приведут вас к неправильным заключениям, но не следует удивляться и тому, что таких выборок будет только 3 или даже 7. При проверке интервалов для $P=0,99$ вы должны ожидать (учитывая, как всегда, выборочную вариацию), что около 99 из 100 этих интервалов будут содержать в себе фактическую долю признаков 50,6%. Я надеюсь, что ваши собственные результаты будут достаточно согласованными с теорией, вследствие чего у вас появится доверие к этой последней. Однако в одном я полностью уверен: вы теперь более осведомлены о причудах выборочного наблюдения, чем это было вначале, и знаете, что этот факт может быть одним из объектов экспериментального исследования. Другой урок, полученный нами, состоит в том, что при малых выборках доверительные

интервалы будут широкими и что даже при такой ширине выводы могут быть ошибочными.

Наконец, можно видеть, что по мере увеличения объема выборки интервал сужается, но это происходит довольно медленно. Данное положение можно сделать наглядным, если объединить все 2000 наблюдений таблицы 3 в одну случайную выборку. Число нечетных цифр будет:

$$1 \times 1 + 2 \times 8 + 3 \times 25 + \dots 9 \times 4 = 1028.$$

По этой большой выборке для доли нечетных цифр в совокупности находим оценку $1028/2000 = 0,5056$. Применяя методы, описанные в главе 16, можно установить, что 95%-ный доверительный интервал простирается от 0,4837 до 0,5275. Сравните его с интервалом для $n = 1000$.

Я особенно подчеркиваю то положение, что при попытке получить точные выводы мы всегда будем в небольшой части выборок терпеть неудачу и что это не является помехой, которую можно было бы устранить, а служит неизбежным ограничивающим условием применения выборочного метода. Ошибочность лежит в самой природе данных, привлекаемых здесь для доказательства. Лицо, производящее выборку, может только предпринять надлежащие предосторожности, устанавливая свою норму для доли возможных ошибок. Но в этом он не одинок. Журналист, судья, банкир, лицо, дающее прогноз погоды, — все они, как и многие другие из нас, сталкиваются с законами случая, и каждый из них определяет для себя некоторую норму ошибочных предвидений. Статистик находится даже в лучшем положении: при соответствующих обстоятельствах он может точно знать вероятность этой ошибки.

Пример 10. Если вы желаете установить при помощи выборочного наблюдения, какая часть лиц, проживающих в каком-либо районе и имеющих радиоприемники, прослушала некоторую программу, получите ли вы надежную выборку, если произведете опрос по телефону? Будут ли все владельцы радиоприемников иметь одинаковый шанс попасть в число опрошенных?

Пример 11. Лицо, изучающее общественное мнение, на основе выборочного опроса некоторого количества избирателей штата дает оценку шансов определенного кандидата в губернаторы. Допустим, что в момент проведения опроса его оценка была очень близкой к фактическим шансам; могут ли результаты голосования в день выборов быть совсем иными?

Пример 12. Почему опрос при помощи рассылаемой по почте анкеты часто приводит к неправильным оценкам?

Пример 13. Объедините все ваши выборки, какого бы размера они ни были, в одну большую выборку. Правильное значение процента нечетных цифр может быть определено двумя способами:

- 1) путем суммирования количеств нечетных цифр по всем выборкам и деления на общее число цифр в них;
- 2) путем вычисления взвешенной средней из процентов при различных размерах выборок на основе формулы:

$$\text{Взвешенная средняя} = \frac{W_1 p_1 + W_2 p_2 + W_3 p_3 + \dots}{W_1 + W_2 + W_3 + \dots}$$

где p_1 — процент нечетных цифр в выборках по 10 и W_1 — общее число цифр, вошедших в эти выборки; p_2 и W_2 означают те же данные для выборок по 15 и т. д.

Пример 14. Основываясь на объединенной выборке предыдущего примера, определите 99%-ный доверительный интервал. Покроет ли он известную вам долю признака в совокупности?

Пример 15. При выборочном опросе глав семей штата Айова, проживающих в пределах обслуживания телецентра WOJ-TV, в 1951 г. оказалось, что из 425 опрошенных у 106 имеется действующая телевизионная установка. Произведите 95%-ную оценку интервала для процента владельцев телевизоров в данном районе. *Ответ:* 24—30%.

Пример 16. Было установлено, что внутри района обслуживания телецентра проживает 31 025 семей. Произведите оценку значения и интервала для числа телевизионных установок в этом районе. *Ответ:* 7750 установок; между 6500 и 9300.

Пример 17. Пять игральные кости были брошены 100 раз, причем получены результаты, приведенные в таблице на стр. 37.

Построить гистограмму выборочного распределения и теоретического биномиального распределения для числа появления двух очков. Как вы думаете, были кости правыми и честно ли они выбрасывались? Смотрите пример 30.

Число появления двух очков на один выброс	Число случаев	Теоретическая численность
5	2	0,013
4	3	0,322
3	3	3,214
2	18	16,075
1	42	40,188
0	32	40,188
Всего . . .	100	100 000

8. Гипотезы относительно совокупности. В противоположность рассмотренному выше случаю оценки доли признака в совокупности, исследователь часто имеет в виду определенную гипотезу относительно этой доли; задача выборочного наблюдения в этом случае состоит в том, чтобы доставить исследователю данные, касающиеся его гипотезы. Так, генетик, изучающий наследственность на томатах, имеет основание ожидать, что в генерации F_1 соотношение плодов с красной и желтой мякотью будет 3 : 1. Подсчет плодов с красной мякотью дает 310 из 400 вместо ожидаемых 300. Зная из предыдущего о выборочном варьировании, воспримете ли вы эти данные как подтверждение или, наоборот, как опровержение гипотезы? Еще пример: лицо, изучающее общественное мнение, пытается предвидеть результаты ближайших выборов в некотором штате. Он выдвигает гипотезу о равном числе голосов в пользу двух главных партий; фактическое разделение 10 000 избирателей, которых он имел в виду, оказалось таким: 5100 против 4900. Должен ли он считать, что все это полностью подтверждает или, наоборот, не подтверждает его гипотезу о ничейном результате голосования?

Для того чтобы ответить на такие вопросы, необходимо иметь две величины, одна из которых должна измерять отклонение доли признака, наблюдаемой в выборке, от гипотетической доли в совокупности, и вторая должна давать возможность судить, будет ли это отклонение таким, которое является обычным при выборочном наблюдении, или, наоборот, оно столь велико, что вызывает сомнение в правильности гипотезы. Оба эти показателя были введены в статистику в 1899 г. Карлом Пирсоном [8]. Он построил специальный показатель рассеяния, обозначенный χ^2 (хи-квадрат), и составил таблицу вероятных значений его при выборочном наблюдении¹. Сначала мы познакомимся с показателем рассеяния.

9. Хи-квадрат — показатель рассеяния. Естественно, что этот показатель строится на основе отклонений наблюдаемых значений от значений, определяемых гипотезой. При проведении выборочного наблюдения в примере предыдущего параграфа, согласно гипотезе, ожидалось, что голоса избирателей разобьются поровну, т. е. по 5000. Отклонения наблюдаемых значений здесь оказались:

$$5100 - 5000 = 100$$

и

$$4900 - 5000 = -100;$$

сумма их равна нулю. Значение хи-квадрат здесь будет:

$$\chi^2 = \frac{(100)^2}{5000} + \frac{(-100)^2}{5000} = 4.$$

Эта величина строится так: каждое отклонение возводится в квадрат, каждый квадрат делится на гипотетическое значение и частные складываются. Гипотетические значения поставлены в знаменателях, в связи с чем в вели-

¹ Дж. У. Юл и Р. А. Фишер обнаружили ошибку в правилах использования этой таблицы и устранили ее [15,2].

чину χ^2 вводится размер выборки; эта величина является *относительным* значением отклонения, что особенно важно. У вас может вызвать недоумение факт возведения отклонений в квадрат. Из дальнейшего вы увидите, что это широко практикуется в статистике. Помимо прочего, возведение в квадрат освобождает нас от знаков отклонений, так как квадрат отрицательного числа тот же самый, что и соответствующего положительного. Очевидно, если все выборочные численности будут равны гипотетическим, то хи-квадрат обращается в нуль, и при увеличении отклонений выборочных значений от гипотетических величина хи-квадрат увеличивается. В этом отношении величина χ^2 является вполне подходящей мерой варьирования. Но не совсем ясно, следует ли наше значение хи-квадрат = 4 считать большим, умеренным или малым. Наша следующая задача состоит в том, чтобы дать вам основу для необходимого в этом случае сравнения. Пирсон пришел к своим результатам путем решения довольно сложного уравнения и табулирования этого решения, но для нематематиков более привычным путем при знакомстве с предметом является эксперимент. Однако прежде чем перейти к этому, следует дать две формулы, сопроводив их несколькими примерами для закрепления их в памяти.

10. Формула хи-квадрат. Обозначим для удобства численности объектов выборки, обладающих и не обладающих изучаемым признаком, через f_1 и f_2 и соответствующие гипотетические частоты через F_1 и F_2 . В этом случае отклонения будут $f_1 - F_1$ и $f_2 - F_2$, что дает для хи-квадрат формулу:

$$\chi^2 = (f_1 - F_1)^2 / F_1 + (f_2 - F_2)^2 / F_2.$$

Эта формула может быть написана более сжато и в более общем виде:

$$\chi^2 = \sum (f - F)^2 / F,$$

где Σ — символ суммирования. На словах это значит: «Хи-квадрат является суммой отношений

(квадрат отклонения): (гипотетическая численность)».

Как будет видно из дальнейшего, суммирование может производиться по большему, чем два, числу отклонений.

Применим эту формулу к примеру подсчета красных и желтых плодов томатов, который был приведен в параграфе 8. Здесь $f_1 = 310$, $f_2 = 400 - 310 = 90$, F_1 составляет $3/4$ от 400, т. е. 300, и F_2 является $1/4$ от 400, т. е. 100. Отсюда

$$\chi^2 = \frac{(310 - 300)^2}{300} + \frac{(90 - 100)^2}{100} = 1,33.$$

Из этого следует, что единственной основой для деления выборки на две части, в соответствии с гипотетическими численностями, здесь является гипотеза, высказанная в отношении совокупности. Поэтому исследователь может интересоваться показателем хи-квадрат только в том случае, если он намерен произвести испытание некоторой гипотезы.

Пример 18. При скрещивании двух сортов гороха исследователь на основании генетической теории ожидает в половине случаев морщинистые семена и в половине — гладкие. В выборке из 800 семян он нашел 440 морщинистых. Каким будет значение хи-квадрат? *Ответ:* 8.

Пример 19. Если бы в условиях предшествующего примера в выборке из 400 семян было обнаружено 220 морщинистых, то будет ли значение хи-квадрат также в два раза меньше первоначального?

Пример 20. В приведенном в тексте примере о численности томатов с красной и желтой мякотью отклонение от ожидаемого значения было равно 10. Если такое же отклонение встретится в выборке, размер которой в два раза больше (т. е. в выборке из 800 плодов), каким будет значение хи-квадрат? *Ответ:* 0,67, т. е. половина первоначального значения.

Пример 21. В процессе предвыборной кампании некоторый кандидат считает, что 60% избирателей будут голосовать за него. Наблюдатель общественного мнения опросил 1000 зарегистрированных избирателей в отношении их намерения голосовать за этого кандидата и в 55% случаев получил ответ «да». Определите значение хи-квадрат при предположении, что оценка своих шансов кандидатом произведена правильно. *Ответ:* 10,42.

Пример 22. Какое утверждение может сделать наблюдатель общественного мнения при тех же, что и в предыдущем примере, данных, но на основе 99%-ного доверительного интервала? *Ответ:* в пользу данного кандидата будут голосовать от 51 до 59% избирателей.

11. Экспериментальное изучение выборочного распределения хи-квадрат. Теперь вы получили некоторую практику по вычислению хи-квадрат. Следует думать, что этот показатель воспринимается вами как вполне обоснованная и имеющая определенное содержание мера для отклонения наблюдаемой в выборке численности от численности, ожидаемой на основе некоторой гипотезы. Однако он выполняет и другую роль, а именно дает нам возможность судить о том, в какой степени значительно отклоняется отношение признаков внутри выборки от отношения их в гипотетической совокупности. Для того чтобы использовать хи-квадрат в этих целях, необходимо дать ответ на поставленный ранее вопрос: какая величина хи-квадрат указывает на слишком необычное отклонение и что следует считать нормальным выборочным варьированием. Наш экспериментальный метод отыскания ответа на этот вопрос будет состоять в том, что для каждой выборки, взятой из таблицы случайных чисел, будет вычисляться хи-квадрат, после чего будет установлено, какие значения хи-квадрат относятся к наиболее отклоняющимся от нормы выборкам. Взяв большое число выборок различного размера и вычислив для каждой из них величину хи-квадрат, мы можем установить распределение этой величины.

Покажем на примере одной выборки систему расчетов. Допустим, что объем выборки составляет 100; так как во всей таблице содержится 50,6% нечетных цифр, то $F_1 = 0,506 \times 100 = 50,6$ и $F_2 = 49,4$. Если в данной выборке оказалось 44 нечетные цифры, то значение хи-квадрат будет:

$$\chi^2 = \frac{(44 - 50,6)^2}{50,6} + \frac{(56 - 49,4)^2}{49,4} = 1,74.$$

Если имеется несколько выборок одинакового размера, то путем преобразования данной формулы можно упростить эти расчеты. Обратим внимание на то, что оба отклонения имеют одно и то же абсолютное значение (т. е. без учета их знаков); в данном случае оно равно 6,6 для того и другого отклонения. Поэтому хи-квадрат можно вычислить так:

$$\begin{aligned} \chi^2 &= (f_1 - F_1)^2 \times (1/F_1 + 1/F_2) = 6,6^2 (1/50,6 + 1/49,4) = \\ &= 43,56 (0,019763 + 0,020243) = 43,56 \times 0,040006 = 1,74, \end{aligned}$$

т. е. то же, что и ранее, число. Преимущество этого способа расчета состоит в том, что, вычислив один раз сумму обратных значений для ожидаемых численностей, эту сумму можно использовать во всех случаях, в которых размер выборки и доля признака в совокупности остаются одними и теми же, для чего следует только отклонение от ожидаемой численности возвести в квадрат и результат умножить на 0,040006. Конечно, для каждого размера выборки должен быть вычислен свой самостоятельный множитель.

Те, кто не имеет надлежащих средств для облегчения этих расчетов (таблицы квадратов, логарифмической линейки, счетной машины), могут, жертвуя несколько точностью, значительно сократить время на проведение этого эксперимента. Можно считать, что в таблице нечетные цифры составляют ровно 50%; тогда $F_1 = F_2 = n/2$ будет для большинства выборок целым числом. Для $n = 100$ приведенная выше формула теперь будет такой:

$$\chi^2 = 0,04 (f - 50)^2.$$

Это в известной мере скажется на распределении частот, но общая его форма останется без изменения.

Произведите этот расчет величины хи-квадрат для каждой вашей выборки. Если взято 100 или более выборок, то результаты вычислений могут быть обобщены в виде выборочного распределения по образцу таблицы 4.

Выборочное распределение 230 значений хи-квадрат, вычисленных по выборкам из таблицы 2

(Размеры выборок: 10, 15, 20, 30, 50, 100 и 250 цифр)

Классовый интервал	Частота	Классовый интервал	Частота
0,00—0,49	116	6,00—6,49	0
0,50—0,99	39	6,50—6,99	1
1,00—1,49	18	7,00—7,49	0
1,50—1,99	22	7,50—7,99	0
2,00—2,49	12	8,00—8,49	0
2,50—2,99	5	8,50—8,99	1
3,00—3,49	5	9,00—9,49	0
3,50—3,99	6	9,50—9,99	0
4,00—4,49	1	10,00—10,49	1
4,50—4,99	2	10,50—10,99	0
5,00—5,49	0	11,00—11,49	1
5,50—5,99	0		

Здесь наблюдается концентрация выборочных значений хи-квадрат в первых классах (класс, или *классовый интервал*, содержит в себе не одно, а несколько значений переменной); фактически половина значений хи-квадрат оказалась меньше 0,5. Мы уже обращали внимание на то, что в процессе выборочных наблюдений доминируют малые отклонения (и поэтому малые значения хи-квадрат); этот факт является основой нашего доверия к выборочному методу.

Однако, придерживаясь менее оптимистического взгляда, можно найти здесь выборки, имеющие большие отклонения и значения хи-квадрат. Возможность получения таких выборок, как эти последние, призывает к осторожности в отношении выводов, сделанных на основании выборки. В данном

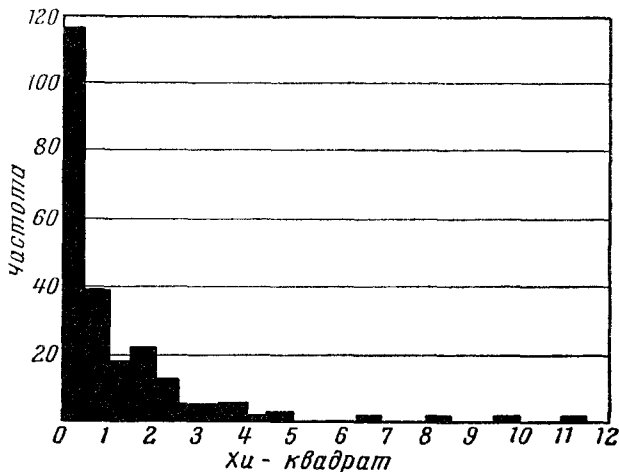


Рис. 2. Гистограмма, изображающая распределение 230 выборочных значений хи-квадрат по данным таблицы 4.

случае нам известно отношение признаков внутри совокупности, и поэтому различия между выборками не вводят нас в заблуждение, но в фактических исследованиях, когда не известно, является ли выдвигаемая гипотеза правильной, возможность получения большого значения хи-квадрат создает определенную дилемму: должны ли мы в этом случае считать, что оно означает только нетипичную для гипотетической совокупности выборку, или мы

должны прийти к выводу, что данная гипотеза не соответствует фактическому соотношению признаков в совокупности? На этот вопрос теория статистики не может дать точного ответа, вместо этого она *определяет вероятность возможных отклонений выборки от данной гипотетической совокупности*. Если значение хи-квадрат велико, то тем самым исследователь предупреждается о том, что данная выборка в условиях его гипотезы относится к числу маловероятных. Эти сведения добавляются к тем, которыми он уже обладает, и все это становится основой для его решений. Далее, через один параграф, будет дано более точное определение этой вероятности.

На рисунке 2 дано графическое изображение нашего распределения хи-квадрат. Этот график дает наглядное представление как о концентрации малых значений хи-квадрат в его левой части, так и о небольшом числе относительно крупных значений этого показателя, разбросанных в правой его части. График показывает, что $\chi^2=4$, полученное при выборочном опросе избирателей в параграфе 9, не выходит из пределов данной шкалы, но в то же время это значение превосходит большинство значений хи-квадрат в этом распределении. Если избиратели действительно разбиваются на две равные группы, то это значение следует считать слишком большим для величины хи-квадрат.

12. Распределение частот при неравных классовых интервалах. В отношении нашего распределения хи-квадрат необходимо исследовать еще два вопроса: 1) каким образом сравнить его с соответствующим теоретическим распределением и 2) как можно определить более точно вероятности для различных значений хи-квадрат? Оба эти вопроса разрешаются сами собой, если произвести преобразование интервалов группировки по образцу первой графы таблицы 5. Эти теоретически установленные интервалы, начинаясь с очень узких, последовательно расширяются так, что образуется симметрическое распределение хи-квадрат, представленное в графе 4 этой таблицы. Более того, эти интервалы подобраны так, что каждый из них содержит в себе целое число процентов от общего числа значений хи-квадрат, в связи с чем вероятности тех или иных значений χ^2 находятся весьма просто. Например, последняя строка графы 4 указывает на то, что при случайной выборке из гипотетической совокупности появление значений хи-квадрат, больших, чем 6,635, имеет вероятность 0,01.

Наше распределение 230 выборочных значений хи-квадрат, представленное в графах 2 и 3, теперь уже становится сравнимым с теоретическим распределением в графе 4. Легко видеть, что различия между ними не слишком

ТАБЛИЦА 5

Распределение выборочных значений хи-квадрат при неодинаковых интервалах группировки

Классовый интервал	Дискретная переменная		Непрерывная переменная	
	распределение 230 выборочных значений	распределение (в %)	теоретическое распределение (в %)	распределение в виде накопленной суммы (в %)
—0,000157	0	0,00	1	100
0,00157—0,00393	11	4,78	4	99
0,00393—0,0158	20	8,70	5	95
0,0158—0,1015	26	11,31	15	90
0,1015—0,455	59	25,65	25	75
0,455—1,323	62	26,96	25	50
1,323—2,706	32	13,91	15	25
2,706—3,841	14	6,09	5	10
3,841—6,635	3	1,30	4	5
6,635—∞	3	1,30	1	1
Итого . . .	230	100,00	100	—

большие. Эти различия частично обусловлены выборочной вариацией, но в основном тем, что теоретические распределения в графах 4 и 5 относятся к «непрерывной» переменной, которая может принимать все численные значения, в то время как распределения в графах 2 и 3 получаются в результате подсчета численностей «прерывной», или «дискретной», переменной, которая переходит от одного целочисленного значения к другому скачком, минуя промежуточные (дробные и иррациональные) значения. Например, вы можете заметить, что при тех размерах выборки, которые были взяты, ни одна не может дать значения хи-квадрат, соответствующего первому интервалу, но если взять выборки в 500 наблюдений или более, то в этом случае могут встретиться даже и такие малые значения хи-квадрат.

Графа 5 дает *распределение накопленных численностей*, выраженных в процентах. Каждое число графы 5 представляет сумму всех предыдущих данных графы 4, начиная снизу. Данные этой графы имеют такой смысл: например, третья цифра снизу означает, что 10% всех выборок, в соответствии с теоретическим распределением, имеют значение хи-квадрат, превосходящее 2,706, а пятая цифра снизу указывает, что последнее число может рассматриваться как среднее значение хи-квадрат, так как выборочные значения хи-квадрат могут быть одинаково часто как больше, так и меньше этого значения. Наконец, значения хи-квадрат, большие, чем 6,635, являются довольно редкими, встречающимися только в одной из 100 выборок. Таким образом, это распределение выборочных хи-квадрат дает нам возможность определить на вероятностной основе, какие значения хи-квадрат следует считать малыми и какие большими. Теперь рассмотрим вопрос о практическом применении этой возможности.

13. Проверка нулевой гипотезы, или критерий существенности. В параграфе 8 указывалось, что объект исследования часто позволяет сформулировать некоторую гипотезу относительно соответствующих экспериментальных данных. Как вы помните, генетик, зная, что на основе определенной схемы наследования следует ожидать отношения признаков 3 : 1, выдвигает гипотезу, что это отношение имеет место для численности красных и желтых плодов в исследуемой им совокупности томатов. Такая гипотеза называется *нулевой гипотезой* в том смысле, что между гипотетическим отношением признаков и фактическим их отношением в совокупности нет никакого различия. Если нулевая гипотеза верна, то отношение признаков в выборках объема n будет распределено по биномиальному закону, а вычисленное по этим выборкам хи-квадрат будет иметь распределение, представленное в таблице 5. Для проверки этой гипотезы производится выборка, и по ее данным вычисляется величина хи-квадрат. В приведенном ранее примере величина хи-квадрат была равна 1,33; обращаясь к таблице, мы видим, что если нулевая гипотеза верна, то значение хи-квадрат, равное 1,33, не является чем-либо необычным, так как вероятность получить большие значения равна 0,25, т. е. достаточно велика. При таком результате проверки у генетика не будет оснований для того, чтобы отвергнуть эту гипотезу. Конечно, он учитывает, что может впасть в ошибку, т. е. что в совокупности отношение между количествами красных и желтых плодов томатов не будет точно равно 3 : 1. Но такое отклонение от этого отношения, если оно и имеет место в данном случае, так мало, что выборка не может дать в пользу него сколько-нибудь убедительных доказательств.

В противоположность опыту по генетике, выборочное наблюдение по поводу выборов привело к $\chi^2 = 4$. Если бы нулевая гипотеза (о равенстве числа избирателей, голосующих за ту и другую партию) была правильной, то вероятность больших, чем 4, значений хи-квадрат была бы только около 0,05. Это позволяет считать, что данная гипотеза неправильна, и поэтому исследователь имеет возможность ее отвергнуть. Но, как и ранее, он может впасть в ошибку, так как его выборка может оказаться именно той из 5 выборок на 100, в которых хи-квадрат превосходит 3,841, даже в случае отбора наблюдений из совокупности, разделенной на две равные части. Отвергая свою нуле-

вую гипотезу, исследователь может оказаться неправым. Такого рода риск неизбежен, когда производится проверка тех или иных гипотез и на основе этой проверки делаются определенные выводы.

Из приведенных примеров видно, что при проверке гипотезы встречаются *ошибки двух видов*. Если выборка приводит к опровержению гипотезы, когда она на самом деле верна, то можно сказать, что мы совершаем *ошибку первого рода*, или ошибку типа I. Если же, наоборот, мы принимаем гипотезу, когда она неправильна, то наша ошибка будет *ошибкой второго рода*, или ошибкой типа II. Соотношение между ошибками этих типов со всей подробностью рассматривается в теории испытания гипотез Неймана—Пирсона. О современном состоянии этой теории можно узнать из литературы [6, 7, 14].

В биологических исследованиях принято отвергать нулевую гипотезу в том случае, когда хи-квадрат превосходит 3,841; это означает, что значения хи-квадрат, *большие, чем 3,841*, образуют *область отказа* от гипотезы. Выборочное значение хи-квадрат, попавшее в эту область, называется *существенным*. Как видно из последней графы предыдущей таблицы, при этих условиях вероятность совершить ошибку типа I, когда на самом деле нулевая гипотеза верна, равна только 0,05. В случае нашей экспериментальной проверки эта вероятность находит не столь уже плохое подтверждение, так как она дает $3 + 3 = 6$ значений хи-квадрат, превосходящих 3,841, что составляет $6/230 = 0,026$ вместо теоретической 0,05.

Следует отметить, что в случае, когда нулевая гипотеза H_0 отвергается, должна быть некоторая альтернативная гипотеза H_A , которая будет принята. В тех условиях, с которыми мы пока имели дело, эта альтернатива ясна сама по себе; если гипотеза об отсутствии различия отвергнута, то принимается альтернатива: некоторое различие имеет место. В дальнейшем мы встретимся с такими условиями, когда результат проверки гипотезы будет зависеть от выбора альтернативы.

Необходимо полное понимание того, что биолог редко или даже никогда не основывает свои выводы только на результате проверки гипотезы. Имея в виду, что выборка снабжает нас только данными, но не доказательством, эти данные может присоединить к другим, уже накопленным в процессе его исследования или исследований других авторов. Эти данные представляют собой дополнительные сведения об изучаемом предмете и постепенно накапливаются в процессе совершенствования эксперимента. Наконец, если эксперимент дал достаточно ценные результаты, то это приводит к ряду новых гипотез, которые в порядке их исследования должны быть проверены или путем серии новых экспериментов, или путем их сопоставления с известными положениями данной науки. На обязанности исследователя лежит обобщение всех таких данных и получение на основе этого того или иного решения вопроса. Никакой ссылкой на величину хи-квадрат он не может снять с себя ответственность за это решение. Вероятность того, что он, пользуясь обобщенными результатами, придет к неправильным выводам, будет наверное значительно меньше вероятности его ошибок первого рода.

Пример 23. В примере 18 $\chi^2 = 8$. По всей вероятности, исследователь отвергнет гипотезу о равном числе морщинистых и гладких горошин и примет противоположную гипотезу о неравном их числе. Если это так, то какой будет вероятность ошибки? *Ответ:* $P < 0,01$ (более точно $P = 0,0046$). О способе определения вероятности, выходящей за пределы нашей таблицы, смотри параграф 8 главы 8 п. в.

Пример 24. В одном опыте по наследственности томатов Мак-Артур [5] обнаружил 529 плодов с красной мякотью и 1176 с желтой. В данном случае было взято поколение F_2 , в котором теоретическое отношение признаков равно 3 : 1. Вычислите $\chi^2 = 0,71$. Мак-Артур сделал заключение, что «различие между наблюдаемым и ожидаемым отношениями признаков не является существенным».

Пример 25. В той же серии экспериментов Мак-Артур нашел 671 растение с зелеными листьями и 569 с желтой. В данном случае было произведено обратное скрещивание, при котором теоретическое отношение признаков 1 : 1. Хи-квадрат равно 8,39 и явно свидетельствует против гипотезы об отношении 1 : 1, «что без сомнения говорит о пониженной жизнеспособности рецессивных форм».

Пример 26. При обследовании в 1943 г. ферм штата Южная Дакота из 1000 взятых образцов фермеров 480 были отнесены к числу владельцев (или частичных владельцев)

ную гипотезу, исследователь может оказаться неправым. Такого рода риск неизбежен, когда производится проверка тех или иных гипотез и на основе этой проверки делаются определенные выводы.

Из приведенных примеров видно, что при проверке гипотезы встречаются *ошибки двух видов*. Если выборка приводит к опровержению гипотезы, когда она на самом деле верна, то можно сказать, что мы совершаем *ошибку первого рода*, или ошибку типа I. Если же, наоборот, мы принимаем гипотезу, когда она неправильна, то наша ошибка будет *ошибкой второго рода*, или ошибкой типа II. Соотношение между ошибками этих типов со всей подробностью рассматривается в теории испытания гипотез Пеймана—Нирсона. О современном состоянии этой теории можно узнать из литературы [6, 7, 14].

В биологических исследованиях принято отвергать нулевую гипотезу в том случае, когда хи-квадрат превосходит 3,841; это означает, что значения хи-квадрат, большие, чем 3,841, образуют *область отказа* от гипотезы. Выборочное значение хи-квадрат, попавшее в эту область, называется *существенным*. Как видно из последней графы предыдущей таблицы, при этих условиях вероятность совершить ошибку типа I, когда на самом деле нулевая гипотеза верна, равна только 0,05. В случае нашей экспериментальной проверки эта вероятность находит не столь уже плохое подтверждение, так как она дает $3 + 3 = 6$ значений хи-квадрат, превосходящих 3,841, что составляет $6/230 = 0,026$ вместо теоретической 0,05.

Следует отметить, что в случае, когда нулевая гипотеза H_0 отвергается, должна быть некоторая альтернативная гипотеза H_A , которая будет принята. В тех условиях, с которыми мы пока имели дело, эта альтернатива ясна сама по себе; если гипотеза об отсутствии различия отвергнута, то принимается альтернатива: некоторое различие имеет место. В дальнейшем мы встретимся с такими условиями, когда результат проверки гипотезы будет зависеть от выбора альтернативы.

Необходимо полное понимание того, что биолог редко или даже никогда не основывает свои выводы только на результате проверки гипотезы. Имея в виду, что выборка снабжает нас только данными, но не доказательством, эти данные может присоединить к другим, уже накопленным в процессе его исследования или исследований других авторов. Эти данные представляют собой дополнительные сведения об изучаемом предмете и постепенно накапливаются в процессе совершенствования эксперимента. Наконец, если эксперимент дал достаточно ценные результаты, то это приводит к ряду новых гипотез, которые в порядке их исследования должны быть проверены или путем серии новых экспериментов, или путем их сопоставления с известными положениями данной науки. На обязанности исследователя лежит обобщение всех таких данных и получение на основе этого того или иного решения вопроса. Никакой ссылкой на величину хи-квадрат он не может снять с себя ответственность за это решение. Вероятность того, что он, пользуясь обобщенными результатами, придет к неправильным выводам, будет наверное значительно меньше вероятности его ошибок первого рода.

Пример 23. В примере 18 $\chi^2 = 8$. По всей вероятности, исследователь отвергнет гипотезу о равном числе морщинистых и гладких горошин и примет противоположную гипотезу о неравном их числе. Если это так, то какой будет вероятность ошибки? *Ответ:* $P < 0,01$ (более точно $P = 0,0046$). О способе определения вероятности, выходящей за пределы нашей таблицы, смотри параграф 8 главы 8 и. »

Пример 24. В одном опыте по наследственности томатов Мак-Артур [5] обнаружил 3629 плодов с красной мякотью и 1176 с желтой. В данном случае было взято поколение F_2 , в котором теоретическое отношение признаков равно 3 : 1. Вычислите $\chi^2 = 0,74$. Мак-Артур сделал заключение, что «различие между наблюдаемым и ожидаемым отношениями признаков не является существенным».

Пример 25. В той же серии экспериментов Мак-Артур нашел 671 растение с зеленой листвой и 569 с желтой. В данном случае было произведено обратное скрещивание, при котором теоретическое отношение признаков 1 : 1. Хи-квадрат равно 8,39 и явно свидетельствует против гипотезы об отношении 1 : 1, «что без сомнения говорит о повышенной жизнеспособности рецессивных форм».

Пример 26. При обследовании в 1943 г. ферм штата Южная Дакота из 1000 взятых на учет фермеров 480 были отнесены к числу владельцев (или частичных владельцев)

ферм, а остальные 520 к числу арендаторов. Известно, что 47% из 700 фермеров этого района являются собственниками. Принимая эту величину за процент собственников в совокупности, определите значение хи-квадрат для выборки в 1000 фермеров. *Ответ:* $\chi^2=0,41$. Увеличит ли этот результат вашу уверенность в том, что выборка была случайной? Такого рода косвенное доказательство применяется довольно часто. В основе его лежит предположение, что если выборка дает представление о некотором свойстве, то возможно, что она репрезентирует и другое подлежащее исследованию свойство, если оба они находятся в определенной связи между собой.

Пример 27. Джеймс Снедекор [10] исследовал влияние инъекции в куриные яйца женского полового гормона. В одном случае вылупились из яиц 2 нормальных цыпленка мужского пола и 19 цыплят, которые были отнесены или к группе нормальных цыплят женского пола, или к группе со значительно выраженными женскими признаками. Какова вероятность получить в выборке из совокупности с равным числом особей обоих полов отношение полов 2 : 19 или еще более резкое? *Ответ:* $\chi^2=13,76$; $P=0,0002$ (см. пример 24 главы 8).

Пример 28. В таблице 5 имеется $62+32+14+3+3=114$ выборов, имеющих значение хи-квадрат, большее, чем 0,455, в то время как теоретически ожидается таких выборов 50% от 230. Какова вероятность получить большее различие, если отбор выборов производится строго случайным порядком? *Ответ* $\chi^2=0,0174$; $P=0,90$. Какова вероятность тех отклонений от разделения выборов половина на половину, которые будут большими, чем отклонения ваших выборов, взятых в соответствии с указаниями параграфа 11 этой главы.

14. Случай, когда число признаков в совокупности больше двух. Выше (параграф 4) уже указывалось, что часто внутри совокупности может быть не два, а несколько видов объектов. Если исследователь имеет какую-либо гипотезу относительно доли каждого из этих видов в совокупности, то он может произвести проверку ее при помощи хи-квадрат. Но теперь выборочное распределение хи-квадрат будет иным, чем в таблице 5. Форма этого распределения зависит от числа степеней свободы, соответствующего данному выборочному наблюдению. Ранее мы имели дело с проверкой нулевой гипотезы, относящейся к двум видам объектов, и в этом случае имелась только одна степень свободы (хотя пока ничего не было сказано о том, что это такое). Для того чтобы определить это понятие, рассмотрим выборку фиксированного объема, положим, в 100 объектов. Если 30 из них обладают интересующим нас признаком, то уже одно это полностью определяет значение хи-квадрат, так как число объектов, имеющих противоположный признак, находится простым вычитанием: $100-30$. Единичная степень свободы связана со случайным появлением в выборке объектов определенного вида. Число этих объектов может колебаться от 0 до 100, но как только это число оказалось установленным, то уже не остается никакой свободы для возможных численностей второй, альтернативной, категории.

Если же имеется несколько видов объектов, то число степеней свободы возрастает. В настоящее время мы будем иметь дело с такими условиями, при которых число степеней на единицу меньше, чем число видов; аргументация этого остается той же, что и при двух видах объектов.

В качестве примера возьмем данные Лидстрема [4], произведшего скрещивание двух рецессивных форм кукурузы — золотистой и зелено-полосатой. «В поколении F_2 были получены четыре различных типа растений. Два из них принадлежали к родительским формам (один зелено-полосатый и другой — золотистый), один — к форме гибрида F_1 (зеленый) и один был совершенно новым типом, являясь сочетанием двух рецессивных типов, а именно золотисто-зелено-полосатый». Распределение 1301 растения по типам было таким:

$$\begin{aligned} f_1 &= 773 \text{ зеленых} \\ f_2 &= 231 \text{ золотистое} \\ f_3 &= 238 \text{ зелено-полосатых} \\ f_4 &= 59 \text{ золотисто-зелено-полосатых} \\ \hline &1301 \end{aligned}$$

В соответствии с менделевским законом наследования соотношение типов должно быть 9 : 3 : 3 : 1. Это соотношение мы возьмем в качестве нулевой

гипотезы. Соответствующие гипотетические численности типов будут:

$$\begin{aligned} F_1 &= \frac{9}{16} \times 1301 = 731,9 \\ F_2 &= \frac{3}{16} \times 1301 = 243,9 \\ F_3 &= \frac{3}{16} \times 1301 = 243,9 \\ F_4 &= \frac{1}{16} \times 1301 = 81,3 \\ &\hline &1301,0 \end{aligned}$$

Подставляя эти данные в формулу хи-квадрат (параграф 10 этой главы)

$$\chi^2 = \Sigma (j - F)^2 / F,$$

получим:

$$\begin{aligned} \chi^2 &= \frac{(773 - 731,9)^2}{731,9} + \frac{(231 - 243,9)^2}{243,9} + \frac{(238 - 243,9)^2}{243,9} + \frac{(59 - 81,3)^2}{81,3} = \\ &= \frac{(41,1)^2}{731,9} + \frac{(-12,9)^2}{243,9} + \frac{(-5,9)^2}{243,9} + \frac{(-22,3)^2}{81,3} = \\ &= 2,31 + 0,68 + 0,14 + 6,12 = 9,25 \end{aligned}$$

при $4 - 1 = 3$ степеням свободы. Отметим, что сумма четырех отклонений $41,1 - 12,9 - 5,9 - 22,3$ равна нулю. Это может служить контролем предыдущих расчетов. В то же время этот факт является также проверкой того положения, что здесь имеется три степени свободы, так как только три отклонения могут быть какими угодно, четвертое же принимает определенное значение, будучи равно нулю минус сумма первых трех отклонений.

Можно ли считать $\chi^2 = 9,25$ при трех степенях свободы случаем обычным в условиях выборки из совокупности, в которой соотношение видов объектов, согласно нулевой гипотезе, $9 : 3 : 3 : 1$, или же это редкий случай? Для ответа на этот вопрос следует обратиться к таблице 6, взяв строку для трех степеней свободы. Здесь вы найдете, что $9,25$ выше 5%-ного уровня и близко к 2,5%-ному уровню. Основываясь на одном этом статистическом показателе, данную гипотезу следует отвергнуть. Все же Линдстром приходит к выводу: «Наблюденные данные не приближаются достаточно близко к отношению $9 : 3 : 3 : 1$, но эти отклонения здесь во многом обусловлены физиологическими причинами, так как в связи с уменьшением против нормы содержания хлорофилла условия существования для растений трех последних классов ухудшились. Последний класс (золотисто-зелено-полосатый) не принадлежит к числу очень жизнеспособных». Этот последний класс внес в величину хи-квадрат часть, составляющую почти две трети ее. Таким образом, здесь, несмотря на большое значение хи-квадрат, нулевая гипотеза все же не отвергнута.

Пример 29. Составьте вероятности, стоящие в заголовке таблицы 6, с вероятностями, соответствующими графе 1 таблицы 5.

Пример 30. В примере 17 дано распределение частот появления 2 очков при выбрасывании 5 игральных костей. Теоретическое распределение правильных показаний костей при надлежащем способе их бросания дано в правой колонке таблицы. При применении величины хи-квадрат для проверки гипотез рекомендуется образовывать классы так, чтобы теоретические численности их были не меньше 5. В соответствии с этим объединим три первых класса вместе и округлим данные до одной десятой:

Число появлений 2 очков	f	F
5, 4, 3	8	3,5
2	18	16,1
1	42	40,2
0	32	40,2

Вычислите хи-квадрат. *Ответ:* 7,76, число степеней свободы = 3. Следовательно, гипотеза случайной выборки из этого теоретического распределения не может считаться опровергнутой, но все же вызывает некоторое подозрение слишком большое число появлений 2 очков в первых трех классах.

Пример 31. По поводу выборок таблицы 3 известно, что они получены случайным порядком из соответствующего теоретического распределения. В виде практики вычислите величину хи-квадрат, предварительно объединив три первых и три последних класса. *Ответ:* 7,47; число степеней свободы = 6. Вероятность получить в выборках из данной гео-

Распределение хи-квадрат¹

Число степеней свободы	Вероятность большего значения												
	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	0,02	0,10	0,45	1,32	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,27	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	13,68	17,34	21,60	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	16,34	20,34	24,93	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	17,24	21,34	26,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	18,14	22,34	27,14	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	19,04	23,34	28,24	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	19,94	24,34	29,34	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	20,84	25,34	30,43	35,53	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	21,75	26,34	31,53	36,74	40,11	43,19	46,93	49,64
28	12,46	13,56	15,31	16,93	18,94	22,66	27,34	32,62	37,92	41,34	44,43	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	23,57	28,34	33,71	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	24,48	29,34	34,80	40,26	43,77	46,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	33,66	39,34	45,62	51,80	55,78	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	42,94	49,33	53,33	63,17	67,59	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,43	52,29	59,33	66,98	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	61,70	69,33	77,58	85,53	90,53	95,02	100,42	104,22
80	51,17	53,54	57,15	60,39	64,28	71,14	79,33	88,13	96,58	101,88	106,63	112,33	116,36
90	59,20	61,75	65,65	69,13	73,29	80,62	89,33	98,64	107,53	113,14	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	82,36	90,13	99,33	109,14	118,50	124,34	129,56	135,81	140,17

¹ Настоящая таблица является сокращением шестизначной таблицы, составленной К. М. Томпсон; печатается с разрешения редактора «Биометрикс» [11].

регической совокупности значения хи-квадрат больше, чем 7,47, равна примерно 0,28. Но вместе с этим закрадывается сомнение — не были ли допущены ошибки при подсчете цифр в некоторых выборках класса 8.

В этой начальной главе читатель введен в курс основных принципов теоретической статистики. Он здесь ознакомился с проблемами изучения совокупности посредством выборки, взятой из этой совокупности. Он узнал, что выводы относительно совокупности имеют силу только в том случае, если сведения, извлекаемые из выборочных данных, получены при помощи надлежащим образом спланированного и умело проведенного выборочного наблюдения. Он научился производить оценки значения и интервала, а также

проверку гипотез. Наконец, он получил представление о тех неопределенностях, которые входят в состав заключений, основанных на результатах выборочного наблюдения. Дальнейшие главы этой книги посвящены последовательно усложняющимся схемам получения статистических данных и соответствующим методам извлечения из них необходимых сведений.

Те из читателей, кто в первую очередь интересуется статистической обработкой данных измерения, а не подсчета, могут непосредственно перейти к чтению второй главы, опустив оставшиеся параграфы настоящей главы. Необходимая для этого подготовка уже имеется. Те же, кто предпочитает не прерывать изучение статистической обработки данных подсчета, найдут продолжение темы 1-й главы в главе 9-й.

15. Хи-квадрат и размер выборки. Из формулы

$$\chi^2 = \Sigma (f - F)^2 / F$$

следует, что при постоянных отклонениях $f - F$ величина хи-квадрат уменьшается при увеличении размера выборки, так как при некоторой частной гипотезе величина F меняется пропорционально этому размеру. Первый пример, приведенный в таблице 7, дает этому числовую иллюстрацию. Увеличение размера выборки n в 4 раза привело здесь к уменьшению величины хи-квадрат до одной четвертой ее первоначального значения. Это объясняется тем, что отклонение, равное 10, является исключительным при выборке в 100 единиц, но оно становится уже обычным при выборке большего размера, так как при большой выборке имеет место и большой размах выборочного варьирования.

ТАБЛИЦА 7

Три примера, показывающие изменения хи-квадрат, возникающие при увеличении первоначального размера выборки в 4 раза

	Первоначальная случайная выборка	Пример 1 Отклонение не меняется	Пример 2 Отклонение увеличивается пропорционально n	Пример 3 Отклонение увеличивается пропорционально \sqrt{n}
Размер выборки	100	400	400	400
Число объектов, обладающих признаком	60	240	240	220
Ожидаемое число при гипотезе 1:1	50	200	200	200
Отклонение	10	10	40	20
Хи-квадрат	4	1	16	4

Во втором примере той же таблицы одновременно увеличены в 4 раза как отклонение, так и размер выборки; в результате величина хи-квадрат также возросла в 4 раза. Отклонение, равное 40, в большей выборке менее возможно, чем отклонение 10 в исходной выборке. Это указывает на то, что размах выборочного варьирования не возрастает пропорционально размеру выборки.

Отклонение третьего примера при увеличении исходного размера выборки в 4 раза возросло только в 2 раза; в этом случае величина хи-квадрат осталась неизменной, что означает, что отклонение в 20 единиц при выборке в 400 столь же возможно, как и отклонение в 10 единиц при выборке в 4 раза меньшего размера. Таким образом, чтобы отклонения имели одну и ту же вероятность появиться в выборках различного размера, необходимо, чтобы они изменялись пропорционально квадратному корню из n .

Необходимо ясное понимание установленных здесь положений, так как они позволяют уяснить сущность показателя хи-квадрат. Одно из применений этих положений состоит в следующем. Довольно часто данные представляются не в виде чисел первоначального подсчета объектов, облада-

ющих изучаемым признаком, а в виде процентов их к общему числу объектов, т. е. производится пересчет на 100. Ясно, что эти проценты не могут быть непосредственно использованы для вычисления величины хи-квадрат, исключая разве случай, когда размер выборки точно равен 100. Во всех других случаях, прежде чем вычислять хи-квадрат, необходимо проценты перевести в натуральные числа объектов, имеющих данный признак, или же применить в конце расчетов соответствующий метод приведения их к окончательному виду.

Другим применением этих положений является использование их для планирования размера выборки. Рассмотрим такой случай. Генетик, изучающий наследственность у томатов, производит наблюдение над 100 проростками и находит среди них 46 с желтой листвой. Теоретически ожидается, что половина растений должна иметь зеленую листву, а половина — желтую, но исследователь допускает возможность пониженной жизнеспособности семян с рецессивным геном желтой листвы. Отклонение числа желтых растений от ожидаемых 50 составляет 4%, но в этом случае хи-квадрат будет равно только 0,64, что отнюдь не является свидетельством против отношения 1 : 1. Между тем предположение о пониженной жизнеспособности желтых растений должно быть проверено, в связи с чем возникает вопрос: «Если отклонение в 4% будет появляться и в выборках большего размера, то спрашивается, какое число растений следует взять для того, чтобы величина хи-квадрат возросла в 10 раз по сравнению с ее значением в данной выборке»? В таблице 7 этому случаю соответствует второй пример, в котором отклонение, составляющее у нас 4% от общего числа растений, пропорционально n . Отсюда следует, что и хи-квадрат будет возрастать по этому же закону, и поэтому для получения $\chi^2 = 6,40$ требуется выборка, в 10 раз большая, чем 100 растений. Приведенные в примере 25 данные Мак-Артура могли быть получены в результате именно таких соображений, предусматривающих проверку предполагаемых особенностей гена. С другой стороны, если увеличение выборки даст отклонение от отношения 1 : 1, не превосходящее норму, то предположение об этих особенностях не будет оправданным. В том и другом случае исследователь получает желаемую информацию по интересующему его вопросу.

16. Другая формула хи-квадрат. Эта новая формула дает те же результаты, что и приведенная ранее более общая формула, но она часто облегчает вычисления:

$$\chi^2 = \frac{(a - rb)^2}{r(a + b)}.$$

Здесь a и b — численности двух классов и r — гипотетическое соотношение соответствующих численностей в совокупности. Применяя ее к примеру с томатами параграфа 8, находим:

$$a = \text{число плодов с красной мякотью} = 310,$$

$$b = \text{число плодов с желтой мякотью} = 90,$$

$$a + b = 400,$$

$$r = 3 : 1 = 3.$$

$$\chi^2 = \frac{(310 - 3 \times 90)^2}{3 \times 400} = 1,33.$$

т. е. то же число, что и ранее.

Пример 32. Дженкинс и Белл [3] сообщают о гене, определяющем желтую окраску проростков кукурузы и являющемся простым рецессивом нормальной зеленой окраски. У 9717 растений было установлено отношение зеленых к желтым, как 78,95% к 21,05%: в то же время предполагаемое отношение их равно 3 : 1. Подтверждают ли выборочные данные это гипотетическое соотношение? *Ответ:* $\chi^2 = 81,04$. Названные исследователи при помощи последующих опытов показали, что этот ген «был сцеплен с летальными или полублетальными факторами, которые и явились причиной недостаточного числа растений в классе с желтой окраской. Требуемое отношение 3 : 1 не было получено до тех пор, пока эти культуры не были подвергнуты неродственному скрещиванию и пока отрицательные факторы не были элиминированы».

17. Отношения, относительные коэффициенты и процентные отношения.

Эти показатели имеют широкое применение в статистике, так как, будучи величинами, приведенными к одному знаменателю, обладают непосредственно сравнимыми числителями. В таблице 8 приведены соответствующие примеры. В качестве знаменателя процентного количества самцов везде взято 100 морских свинок; приведенные в таблице два ряда отношений имеют в качестве знаменателей соответственно *одну* или 100 самок. Иногда применяются и другие знаменатели отношений. Например, коэффициент смертности обычно выражается в пересчете на тысячу населения или в случае редких причин смерти даже на 100 000 населения.

Удобство сравнения таких отношений достигается за счет утраты некоторых сведений о предмете. Отвлекаясь от фактических знаменателей

ТАБЛИЦА 8

Средние числа самцов и самок, рожденных в группе морских свинок, по месяцам.

Данные Шотта и Ламберта [9]

(Различные формы выражения для отношения полов)

	Январь	Февраль	Март	Апрель	Май	Июнь	
Самцы	65	64	65	41	72	80	
Самки	49	58	81	48	62	80	
Всего	114	122	146	89	134	160	
Процент самцов	57,0	52,5	44,5	46,1	53,7	50,0	
Отношение: число самцов к числу самок	1,327	1,103	0,802	0,854	1,161	1,000	
Отношение: число самцов на 100 самок	132,7	110,3	80,2	85,4	116,1	100,0	
	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь	Всего
Самцы	88	114	80	129	112	86	996
Самки	95	118	94	104	144	85	1018
Всего	183	232	174	233	256	171	2014
Процент самцов	48,1	49,1	46,0	55,4	43,8	50,3	49,45
Отношение: числа самцов к числу самок	0,926	0,966	0,851	1,240	0,778	1,012	0,9784
Отношение: числа самцов на 100 самок	92,6	96,6	85,1	124,0	77,8	101,2	97,84

отношений, а иногда и совсем теряя их из виду, мы можем сделать недопустимую ошибку. Некоторые расчеты требуют обратного перехода к первоначальным данным; например, доля самцов по итоговой колонке будет определяться следующим образом:

$$49,45\% = 100 \times \frac{996}{2014} \text{ и т. д.}$$

Напомним уже отмеченный ранее факт (параграф 15), что при вычислении хи-квадрат проценты не могут быть использованы, пока не будет применен тот или иной метод приведения данных к первоначальной базе и, таким образом, учтен размер выборки.

В ряде случаев наблюдаются незначительные различия между величинами, положенными в основу вычисления процентов; в этих случаях можно без большой ошибки произвести вычисление средней из самих процентов. Так, по процентам самцов морских свинок мы находим среднюю:

$$\frac{57,0+52,5+\dots+50,3}{12} = \frac{596,5}{12} = 49,71\%,$$

которая весьма близка к точному значению 49,45%. Однако, если требуется точное определение средней из процентов, имеющих различные базы, то необходимо перейти к первоначальным отношениям, сложить их числители и отдельно знаменатели и после этого разделить первую сумму на вторую.

Пример 33. При обследовании трех стад молочного скота были получены такие данные о реакции на туберкулез:

число коров в стаде . . .	40	100	10
процент коров, давших реакцию на пробу . . .	5	2	60

Определите средний процент (6,7%). Как вы думаете, это более правильная средняя, чем $(5+2+60):3=22,3\%$?

Пример 34. Процент сорняков в двух образцах семян тимфеевки составляет 0,01 и 0,05. Если каждый образец состоит из 10 000 семян, то чему будет равен средний процент сорняков? *Ответ:* 0,03%.

Пример 35. Каким был бы средний процент, если бы образцы предыдущего примера содержали в себе соответственно 80 000 и 20 000 семян. *Ответ:* 0,018%, что очень близко к проценту в большем образце.

Пример 36. Шотт и Ламберт указывают, что данные таблицы 8 являются средними за 7 лет, так что общее число самцов было 6972, самок 7126 и отношение полов составляло 97,84 самца на 100 самок. Проверьте гипотезу, что в совокупности отношение полов равно 100 на 100. *Ответ:* $\chi^2=1,68$. *Примечание:* если бы при этих расчетах были использованы средние за 7 лет, то хи-квадрат был бы равен 0,24, т. е. только одной седьмой части правильного значения этой величины.

Л И Т Е Р А Т У Р А

1. Arkin Herbert, Colton Raymond R., Graphs: How to Make and Use Them. Harper and Brothers, New York, Revised Edition, 1940.
2. Fisher R. A., Journal of the Royal Statistical Society, 85, 87, 1922.
3. Jenkins Merle T., Bell Martin A., Genetics, 15, 253, 1930.
4. Lindstrom E. W., Cornell University Agricultural Experiment Station Memoir, 13, 1918.
5. MacArthur John W., Transactions of the Royal Canadian Institute, 18, 1, 1931.
6. Mood A. M. Introduction to the Theory of Statistics. McGraw-Hill Book Company, Inc., New York, 1950.
7. Neyman J., First Course in Probability and Statistics. Henry Holt and Company, New York, 1950.
8. Pearson Karl, Philosophical Magazine, Series 5, 50, 157, 1899.
9. Schott R. G., Lambert W. V., Iowa State College Journal of Science, 4, 343, 1930.
10. Snedecor J. G., Journal of Experimental Zoology, 110, 205, 1949.
11. Thompson Catherine M., Biometrika, 32, 187, 1941.
12. Wakeley Ray E., Iowa State College Research Bulletin 249, 1938.
13. Walker Helen M., Durost N., Statistical Tables; Their Structure and Use. Teachers College, Columbia University, New York, 1936.
14. Wilks S. S., Mathematical Statistics. Princeton University Press, Princeton, New Jersey, 1946.
15. Yule G. U., Journal of the Royal Statistical Society, 85, 95, 1922.
16. Данные любезно представлены Т. А. Бриндли.

Глава 2

ВЫБОРОЧНЫЕ НАБЛЮДЕНИЯ ИЗ НОРМАЛЬНОЙ СОВОКУПНОСТИ

1. Нормальная совокупность. Рассмотренные в главе 1 выборочные наблюдения производились главным образом из совокупности объектов с двумя противоположными признаками: четная или нечетная цифра, живой или мертвый, зараженный или незараженный организм и пр. Случайные выборки объема n из такой совокупности образуют *биномиальное распределение*. Здесь переменная величина — число появления признака — является дискретной. Теперь мы перейдем к другому виду совокупностей, члены

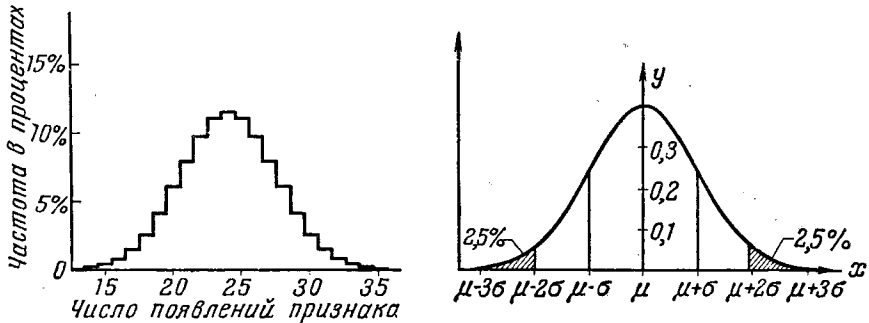


Рис. 3. Слева — биномиальное распределение для числа появлений признака в выборках в 48 наблюдений, взятых из совокупности с соотношением двух признаков 1:1. Справа — нормальное распределение со средней μ и стандартным отклонением σ (см. табл. 71); заштрихованная часть составляет 5% всей площади.

которой имеют такие признаки, как высота, урожай, доход и пр. В этом случае переменная величина переходит от одного значения к последующему без скачка и является непрерывной с неограниченным числом возможных значений. Непрерывные переменные могут быть распределены самым различным образом, но мы здесь займемся главным образом *нормальным распределением*.

Нормальное распределение было открыто давно, вскоре после биномиального распределения. В 1733 г., через 20 лет после того, как Бернулли дал подробное описание биномиального распределения, Де Муавр опубликовал формулу нормального распределения. Эти два распределения находятся в определенном соотноении друг с другом, что можно видеть из рисунка 3.

В левой части дан график симметричного биномиального распределения, подобного тому, который был представлен на рисунке 1. На этом новом рисунке представлен случай, когда объем выборки равен 48, и совокупность, из которой берутся выборки, имеет одинаковое число объектов с двумя противоположными признаками. Число выборок здесь считалось неопределенно

большим, и поэтому частоты выражены в процентах к итогу. Хотя в некоторых случаях наблюдалось появление признака меньше чем 13 и больше чем 35 раз, однако относительные численности этих случаев столь малы, что они не могли быть изображены на этом графике.

Теперь представим себе, что размер выборки беспрестанно увеличивается и соответственно этому уменьшается ширина интервалов по горизонтальной оси. Ступени гистограммы будут становиться все меньше и меньше, и сама диаграмма постепенно будет приобретать вид непрерывной кривой правой части рисунка. Этим и определяется соотношение между нормальным и биномиальными распределениями. Дискретная переменная становится *непрерывной*, а частоты переходят одна в другую уже без скачка.

Нормальное распределение полностью определяется двумя константами, или *параметрами*. Во-первых, таким параметром является *средняя* μ , которая находится в центре распределения. Во-вторых, *стандартное отклонение* σ , измеряющее разброс, или вариацию, отдельных наблюдений; величина σ является фактически *масштабом* (единицей измерения) изменчивости, подчиненной закону нормального распределения.

Рассматривая рисунок, можно видеть, что в пределах одной сигмы, отложенной по ту и другую сторону от μ , частоты уменьшаются с большей быстротой, чем в точках, лежащих вне этой области, где они уменьшаются в непрерывно сокращающемся темпе. Когда переменная X выходит за пределы $\pm 3\sigma$, частоты в процентах становятся весьма малыми. Хотя теоретически частоты нигде на всей оси совсем не исчезают, однако они по мере беспрестанного увеличения X последовательно приближаются к нулю. Концентрация наблюдений около μ находит свое выражение в том, что примерно $\frac{2}{3}$ всего количества наблюдений лежит в пределах $\mu \pm \sigma$, около 95% их находится в интервале $\mu \pm 2\sigma$; вне интервала $\pm 3\sigma$ находятся только 0,26% всех наблюдений.

Вы, возможно, выразите свое удивление по поводу того, что здесь принимается некоторая искусственная модель, которая, вообще говоря, не может быть приписана никакой реальной совокупности. Вызывает изумление и то обстоятельство, что нормальное распределение занимает в статистической практике такое же доминирующее положение, как и в теории. Некоторые из причин этого положения будут рассмотрены в соответствующем месте (параграфы 4-й главы 3, 6-й и 7-й главы 5), но три момента могут быть отмечены уже здесь. Во-первых, имеется множество переменных величин биологического характера, распределения которых близки к нормальному: например, рост человека, или длина початка кукурузы, или убойный выход свинины и пр. Во-вторых, как это следует из теории и опыта, выводы, к которым мы приходим исходя из экспериментальных выборочных распределений, почти не меняются под влиянием обычно встречаемых отклонений от нормальностей формы распределения. В-третьих, математические преобразования уравнения нормального распределения, к счастью, оказываются относительно простыми и позволяют построить широко развитую теорию, имеющую множество практических приложений. Более подробное описание нормального распределения будет дано в главе 8.

2. *Оценки μ и σ* . Параметры μ и σ редко бывают известны, и поэтому обычно приходится определять их приближенные значения на основе выборочных данных. В качестве иллюстрации такой приближенной оценки этих параметров рассмотрим данные, представленные в таблице 9. В 1936 г. Совет по питанию при Американской медицинской ассоциации провел выборочное исследование содержания витамина С в консервированном томатном соке, поступающем в продажу; при исследовании было взято по одному образцу от каждой партии, представленной в Совет для апробации [3]. Полученные данные приведены во втором столбце таблицы.

При условии, что выборка является случайной и взята из нормальной совокупности, приближенное определение параметра μ производится при помощи показателя, называемого *средней выборки*, или, что более правильно,

Содержание витамина С в 17 образцах консервированного томатного сока, 1936 г.¹

Номер наблюдения	Содержание витамина С (мг на 100 г)	Отклонение от средней	Квадрат отклонения
n	X	$x = X - \bar{x}$	x^2
1	16	-4	16
2	22	+2	4
3	21	+1	1
4	20	0	0
5	23	+3	9
6	21	+1	1
7	19	-1	1
8	15	-5	25
9	13	-7	49
10	23	+3	9
11	17	-3	9
12	20	0	0
13	29	+9	81
14	18	-2	4
15	22	+2	4
16	16	-4	16
17	25	+5	25
Итого . . .	340	-26 +26	254

$$\bar{x} = 340/17 = 20 \text{ мг на } 100 \text{ г;}$$

$$s^2 = \Sigma x^2 / (n - 1) = 254/16 = 15,88; \quad s = 3,98 \text{ мг/100 г.}$$

$$s_x^2 = s^2/n = 15,88/17 = 0,934; \quad s_{\bar{x}} = s/\sqrt{17} = 0,965 \text{ мг/100 г.}$$

¹ Данные для облегчения вычислений, как это будет принято и в дальнейшем, несколько изменены. Выводы при этом остаются теми же; фактические данные см. в примере 36 этой главы.

выборочной средней. Вычисление ее производится обычным способом, т. е. делением суммы наблюдений X на их число. Обозначая выборочную среднюю через \bar{x} , имеем:

$$\bar{x} = 340/17 = 20 \text{ мг на } 100 \text{ г сока.}$$

Символ \bar{x} обычно читается « x с чертой», или « x -черта». Мы будем говорить, что выборочная средняя является оценкой μ , или что μ оценивается этой средней.

Что же касается стандартного отклонения σ , то простейшей оценкой его является *размах варьирования* выборочных наблюдений, который представляет собой разность между наибольшим и наименьшим наблюдениями. Для наших данных имеем:

$$\text{размах} = 29 - 13 = 16 \text{ мг/100 г.}$$

Переход от размаха варьирования к оценке σ производится путем умножения на коэффициент, зависящий от размера выборки; см. таблицу 10 [13, 18]. Для $n=17$, стоящего посередине между 16 и 18, этот коэффициент равен 0,279. Поэтому σ оценивается величиной $0,279 \times 16 = 4,46$ мг/100 г.

Теперь нетрудно определить *оценку значения* для каждого из параметров нормальной совокупности: эти оценки представляют собой обобщения той информации о параметрах, которая содержится в данной выборке. Выборочная средняя в качестве оценки μ не может быть улучшена, но для σ в дальнейшем мы определим более эффективную оценку. Точно так же в последующем будут рассмотрены и вопросы об оценке интервалов и о проверке гипотез. Но прежде чем перейти к этим вопросам, необходимо более подробно изучить нашу выборку.

Первый вопрос, подлежащий выяснению, состоит в следующем. Какую совокупность репрезентирует выборка из 17 наблюдений о количестве витамина С в томатном соке? Я ставлю этот вопрос с некоторым запозданием, так как он является первым из всего того, что должно быть решено при планировании любой выборки. Прежде всего ясно, что выборка характеризует не все партии продукции, а только те семнадцать, которые были представлены на апробацию в Совет. Временем консервирования в большинстве случаев были август и сентябрь 1936 г., т. е. примерно за год до анализа продукции. Совет в своем отчете указывает, что содержание витамина «зависит от сорта томата, от условий его выращивания, от степени зрелости плодов и от других факторов». Таким образом, все, что может быть в данном случае установлено, — это только то, что выборочная совокупность имеет годовую давность, и то содержание витамина, которое сохранилось в 17 взятых образцах за это время.

В следующих параграфах будут рассмотрены некоторые другие подробности.

ТАБЛИЦА 10

Отношение σ к размаху в выборках из нормальной совокупности в зависимости от размера выборки n . Эффективность размаха в качестве оценки σ . Число наблюдений, необходимых для оценки σ при помощи размаха, если за 100 принято число наблюдений, необходимых для оценки при помощи s

n	σ Размах	Относительная эффективность	Численность на 100	n	σ Размах	Относительная эффективность	Численность на 100
2	0,886	1,000	100	12	0,307	0,815	123
3	0,591	0,992	101	14	0,294	0,783	128
4	0,486	0,975	103	16	0,283	0,753	133
5	0,430	0,955	105	18	0,275	0,726	138
6	0,395	0,933	107	20	0,268	0,700	143
7	0,370	0,912	110	30	0,245	0,604	166
8	0,351	0,890	112	40	0,231	0,536	186
9	0,337	0,869	115	50	0,222	0,490	204
10	0,325	0,850	118				

3. Ранжированный ряд и его графическое изображение. Для удобства обозрения выборку можно представить в виде *ранжированного ряда*, расположив наблюдения в порядке возрастания, начиная с наименьшего и кончая наибольшим. Так, ранжированный ряд данных о содержании витамина будет следующим: 13, 15, 16, 16, 17, 18, 19, 20, 20, 21, 21, 22, 22, 23, 23, 25, 29.

При малом числе наблюдений ранжированный ряд выполняет то же назначение, которое отводится распределению частот при большой выборке.

В нашем ранжированном ряду размах варьирования от 13 до 29 является сам собой; вместе с этим легко установить и такие факты, как концентрация наблюдений около центра ряда и их меньшая густота на концах его. В этом находит свое отражение указанная выше особенность распределения нормальной совокупности, из которой взята данная выборка. Однако при очень малой выборке такое отражение характерных черт распределения будет крайне неустойчивым.

При просмотре данных о содержании витамина С прежде всего бросается в глаза варьирование этих данных. Каковы причины такого варьирования? По всей вероятности, оно возникает в связи с различными способами производства продукта, а также в связи с различными источниками сырья. Вместе с этим несомненно и то обстоятельство, что исследуемые образцы сами представляют выборки из соответствующих партий продукции и поэтому в какой-то мере отличаются от средних внутри этих партий. Наконец, и методы химического анализа никогда не могут быть абсолютно точными. Варьирование

вообще является наиболее существенной особенностью статистического материала.

На рисунке 4 дано графическое изображение приведенного выше ранжированного ряда 17 показателей содержания витамина. Каждая точка соответствует отдельному наблюдению; расстояние точки от вертикальной оси пропорционально содержанию аскорбиновой кислоты (витамина С) в данном образце, отсчитываемому по горизонтальной оси, в миллиграммах на 100 г.

Этот график дает ясное представление не только о варьировании и концентрации выборочных наблюдений, но и о двух других характерных особенностях наших данных: 1) о довольно хорошо выраженном симметричном размещении наблюдений по отношению к средней и 2) о том, что основная масса данных располагается вблизи среднего уровня и что имеется только

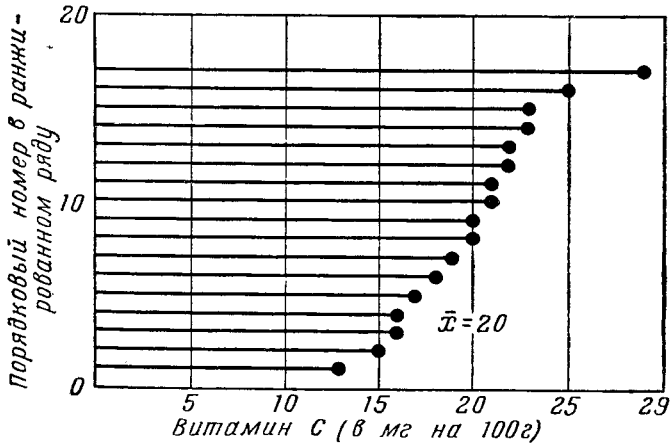


Рис. 4. Графическое изображение ранжированного ряда. Данные о содержании витамина С.

небольшое число наблюдений с очень малым или очень большим содержанием витамина С. Эти особенности опять указывают на существование особого рода устойчивости, проявляющейся в выборках из нормальных совокупностей. Средний уровень и размах варьирования являются специфическими показателями для многих признаков у живых организмов, отражающими в себе то, что с наибольшей последовательностью проявляется в массе наблюдений. Эти нормы остаются более или менее постоянными, несмотря на то, что отдельные индивидуумы развиваются в различных условиях и в той или иной мере различны. Многие из наших представлений основываются на предпосылке о существовании таких устойчивых показателей. Слова «порода скота», «вид растения», «человек» приобретают более глубокий смысл, когда получают количественную характеристику путем обобщения множества данных. Трудно себе представить возможность какого-либо прогресса в наших знаниях, если данные об отдельных индивидуальностях не будут сведены в такие понятия, как средний уровень и размах варьирования.

4. Символические обозначения. Отдельные наблюдения будем обозначать:

$$X_1, X_2, X_3, \dots, X_n,$$

где подстрочные указатели 1, 2, ..., n определяют местоположение наблюдения в ряду (не обязательно ранжированном). Следующие друг за другом точки в данном случае читаются «и так далее». Связывая эти символы с данными третьей графы таблицы 9, имеем:

$$X_1 = 16, X_2 = 22 \dots, X_{17} = 25 \text{ мг/100 г.}$$

Выборочную среднюю будем обозначать \bar{x} ; таким образом:

$$\bar{x} = (X_1 + X_2 + \dots + X_n)/n.$$

Это можно представить в более компактной форме:

$$\bar{x} = (\Sigma X)/n,$$

где подразумевается, что X последовательно принимает значения всего ряда наблюдений. Символ ΣX читается, как «сумма X ». Применяя указанную формулу к значениям X из таблицы 9, находим:

$$\Sigma X = 340 \text{ и } \bar{x} = 340/17 = 20 \text{ мг/100 г.}$$

5. Отклонения от выборочной средней. Варьирование отдельных значений в каком-либо ряду наблюдений лучше всего характеризуется *отклонениями* этих значений от некоторого центрального значения, например от выборочной средней. Так, для первого значения X в таблице 9 отклонение от средней равно $16 - 20 = -4$ мг на 100 г, т. е. это наблюдение меньше средней \bar{x} на 4 мг/100 г. Однако особый интерес представляет весь ряд отклонений, например ряд отклонений, вычисленных для ранжированного ряда данных в параграфе 3 этой главы:

$-7, -5, -4, -4, -3, -2, -1, 0, 0, 1, 1, 2, 2, 3, 3, 5, 9.$

На рисунке 4 этим отклонениям соответствуют расстояния представленных там точек от вертикальной прямой, проходящей через значение выборочной средней.

В нашем мышлении представление об отклонениях играет почти такую же роль, как и представление о среднем, или нормальном, уровне. «Как кит среди свиней» — таково в английском языке метафорическое выражение для слишком большого отклонения от того, что говорящий считает нормальным. Сплетни и пересуды кумушек касаются обычно отклонений от установленной нормы поведения. Забавно то, что интерес в этом случае часто сосредоточивается именно на отклонениях от нормы, так что сама норма отходит как бы на второй план, по отношению к которому отклонения выступают вперед. Причуды являются причудами только потому, что, говоря статистическим языком, они являются заметными отклонениями.

Для обозначения отклонений будем применять соответствующие строчные буквы:

$$\begin{aligned} x_1 &= X_1 - \bar{x} \\ x_2 &= X_2 - \bar{x} \\ &\vdots \\ x_n &= X_n - \bar{x}. \end{aligned}$$

Подобно тому, как X означает любое из ряда наблюдений или все эти наблюдения, взятые в их последовательности, символ x означает отклонения этих наблюдений от их средней, т. е. вообще

$$x = X - \bar{x}.$$

Легко установить, что сумма всех отклонений от средней равна нулю, или $\Sigma x = 0$. Ряд отклонений, представленный в таблице 9, в сумме дает нуль, так как сумма положительных отклонений равна сумме отрицательных отклонений. Это свойство отклонений от средней обычно используется для проверки вычислений: если отклонения вычислены правильно, должен получиться нуль. Из этого свойства следует, что *среднее* отклонение всегда равно нулю; данное положение в дальнейшем найдет себе применение.

Пример 1. Взвешивание 12 растений конопля в начале апреля, проведенное на опытной станции колледжа в Техасе [16], дало следующие округленные результаты:

13, 11, 16, 5, 3, 18, 9, 9, 8, 6, 27 и 7 г.

Построить по этим данным ранжированный ряд и представить его графически. Вычислить выборочную среднюю (11 г) и отклонения от нее. Проверить равенство $\Sigma x = 0$. Показать, что оценка σ равна 7,4 г.

Пример 2. Рост 11 мужчин равен 64, 70, 65, 69, 68, 67, 68, 67, 66, 72 и 61 дюйм. Вычислить выборочную среднюю и произвести проверку вычислений путем суммирования отклонений. Будет ли число положительных и отрицательных отклонений одинаковым, или это относится только к их суммам?

Пример 3. Урожай люцерны на 10 делянках равны 0,8; 1,3; 1,5; 1,7; 1,7; 1,8; 2,0; 2,0; 2,0 и 2,2 т на 1 акр. Сколько здесь положительных и отрицательных отклонений? Равна ли их сумма нулю? Оценить σ . *Ответ:* 0,46 т на 1 акр.

Пример 4. Взвешивание 11 сорокалетних мужчин дало 148, 154, 158, 160, 161, 162, 166, 170, 182, 195 и 236 фунтов. Сопоставить графическое изображение этого ранжированного ряда с изображением ряда предыдущего примера. Следует отметить тот факт, что только 3 из этих наблюдений превышают выборочную среднюю. Можно ли на основе этого ожидать нормальное распределение веса мужчин? В параграфе 5 главы 8 будет дан специальный критерий для определения симметричности распределения.

Пример 5. Ниже приводятся урожай двух сортов овса за 5 последовательных лет (в бушелях на 1 акр):

Сорта	Годы				
	1	2	3	4	5
A	34	30	41	25	45
B	30	17	33	25	25

Определить 5 разностей между сортами А и В. Можно ли считать, что эти разности составляют выборку из нормальной совокупности. Допустив это, произвести оценку μ и σ . *Ответ:* 9,0 и 8,6 бушеля на 1 акр.

Пример 6. Приводимые ниже данные взяты из исследования Редди [14] о влиянии на урожай дезинфекции семян кукурузы, зараженных диплоидозом. Урожай даны в бушелях на 1 акр.

	Парные делянки в 1933 г.					
	1	2	3	4	5	6
Обработанные	68,1	74,6	64,4	69,2	61,8	57,9
Необработанные	64,7	62,5	66,8	69,2	53,9	58,5
Разности	3,4	12,1	-2,4	0,0	7,9	-0,6

Парные делянки в 1934 г.

1	2	3	4	5	6	7	8	9	10
18,0	24,0	18,8	17,8	18,5	27,2	23,6	23,9	20,3	11,9
10,9	24,4	15,1	16,8	13,2	21,6	13,7	17,5	16,3	15,5
7,1	-0,4	3,7	1,0	5,3	5,6	9,9	6,4	4,0	-3,6

Совершенно очевидно, что урожай этих двух лет нельзя считать двумя выборками из одной и той же совокупности, но в отношении разностей это вполне допустимо. Дайте графическое изображение ранжированного ряда этих разностей. Произведите оценку μ и σ совокупности таких разностей. *Ответ:* 3,7 и 4,4 бушеля на 1 акр.

Пример 7. Если в предыдущем примере вы сложите все отклонения от 3,7 бушеля на 1 акр, то нуль не получится. Почему? Если вы к 3,7 прибавите среднюю из отклонений, то не даст ли это точную среднюю разность?

Пример 8. Если по данным упражнения 6 вы вычислите средний урожай всех 16 делаяк с необработанными семенами, то будет ли эта средняя служить оценкой параметра некоторой реальной совокушности?

Пример 9. Допустим, что вы желаете определить урожай с участка кукурузы, состоящего из 300 рядков; фактически же вы убрали урожай с 10 рядков и получили в среднем 5 бушелей на 1 рядок. Можете ли вы категорически утверждать, что урожай всего участка составляет ровно 1500 бушелей? В этом случае вы предполагаете, что средняя для всего участка остается той же самой, что и у 10 убранных рядков, причем расчет ведется на основе равенства $\Sigma X = n\bar{x}$.

Пример 10. Если у вас есть навык в отношении алгебраических преобразований, то докажите равенство $\Sigma x = 0$. В данном случае следует начать с соотношения $x = X - \bar{x}$, потом просуммировать обе части его и произвести подстановку $\Sigma X = n\bar{x}$.

Пример 11. Вы имеете два ряда попарно связанных данных, таких, как в примере 5, и вычислили соответствующий ряд разностей; докажите, что средняя разность равна разности средних этих рядов. Проверить это положение по данным примера 5.

6. Другая возможность оценки σ . Стандартное отклонение выборки. Размах, поскольку он зависит от значений двух крайних наблюдений выборки, имеет выборочное распределение, варьирующее в более широких пределах, чем распределение оценки, основанной на всем ряде отклонений от средней, а не только на наименьшем и наибольшем из этих отклонений. Такой ряд из 17 отклонений был приведен в таблице 9 и повторно дан в виде ранжированного ряда в параграфе 5 этой главы. Спрашивается, какого рода среднюю следует применить при обобщении этих отклонений для того, чтобы получить оценку σ , обладающую наименьшим выборочным варьированием?

Очевидно, что простая средняя из этих отклонений не может быть применена в качестве такой оценки, ибо она всегда равна нулю. В связи с этим сама собой возникает мысль — не обращать внимания на знаки и вычислить среднюю из абсолютных значений этих отклонений. В прошлом такой способ измерения варьирования при помощи *среднего абсолютного отклонения* имел широкое применение. Однако в настоящее время имеются другие оценки, которые более эффективны и более удобны для вычислений.

Одной из самых эффективных оценок является *выборочное стандартное отклонение*, которое мы будем обозначать буквой s . В крайней правой графе таблицы 9 дан расчет этого показателя. Сначала каждое из отклонений возводится в квадрат, после чего *сумма квадратов* Σx^2 делится на число *степеней свободы*, т. е. на численность выборки, уменьшенную на единицу. В результате получается *средний квадрат* s^2 . Наконец, путем извлечения квадратного корня получается s , причем происходит возвращение к исходным единицам измерения (в нашем примере в миллиграммах на 100 г.). Прежде чем перейти к обсуждению вопросов, связанных с этим показателем, следует на небольшом числе примеров закрепить представление об этой величине.

Пример 12. Пять разностей А—В в примере 5 соответственно равны 4, 13, 8, 0 и 20 бушелей на 1 акр. Вычислите s . Ответ: 7,8 бушеля на 1 акр. Сравните этот результат с оценкой, вычисленной на основе размаха.

Пример 13. Вычислите выборочное стандартное отклонение по данным примера 1. Ответ: 6,7 г. Сравните этот результат с прежней оценкой σ .

Пример 14. Вычислите s по данным об урожаях люцерны в примере 3. Ответ: 0,41 т на 1 акр.

Может показаться неожиданным то обстоятельство, что для определения среднего квадрата в данном случае применяется делитель $n-1$, тогда как обычно для вычисления средней берется делитель n — число наблюдений. Необходимость при вычислении s^2 производить деление на число степеней свободы $n-1$ объясняется стремлением избежать введения систематической ошибки в оценку σ^2 . Деление на n также дает оценку σ^2 , но это будет *смещенная оценка*. В кругу рассматриваемых здесь вопросов нам придется иметь дело только с несмещенными оценками.

Теперь вы имеете две оценки σ ; хотя одна из них вычисляется проще, однако она менее эффективна. Вы желаете знать, что вкладывается в слова

«менее эффективна» и чем следует руководствоваться при выборе между этими оценками. О том и другом можно получить некоторые сведения из четвертой графы таблицы 10. Например, если $n=10$, то эффективность оценки, вычисленной на основе размаха варьирования, составляет только 85 % от эффективности оценки при помощи s ; это значит, что для достижения одной и той же точности оценки σ при использовании s выборка из 10 наблюдений эквивалентна выборке из $10/0,85=12$ наблюдений, по которой оценка σ произведена на основе размаха варьирования. Данное положение станет более ясным, если учесть, что в указанной таблице расчет эффективности произведен при прочих равных условиях, и если в дополнение к этому произвести сопоставление затрат на более сложные расчеты показателя s с затратами на увеличение числа наблюдений, например с 10 до 12. В настоящее время довольно часто наблюдения производятся для целей, не имеющих прямого отношения к статистике, и если они в дальнейшем используются статистикой, то не требуют никаких дополнительных затрат, так как для оценки σ требуются затраты труда только на их выписку. В этих условиях само собой напрашивается применение для оценки σ размаха варьирования. Но возьмем теперь опыт с люцерной, приведенный в примере 14. Во что обходятся исследователю две дополнительные делянки? Это будет стоимость земли и амортизации орудий обработки почвы, заработная плата персонала и, наконец, рыночная стоимость семян люцерны. Очевидно, стоимость этих дополнительных делянок не уравновешивается тем небольшим дополнительным числом пенсов или минут, которые необходимы для вычисления оценки s . Я думаю, что если относиться бережливо к той информации, которую дает эксперимент, то, безусловно, необходимо производить оценку σ при помощи s . В этом случае мы извлечем максимум информации из имеющихся у нас данных. Однако преимущество размаха варьирования состоит в том, что он дает возможность быстро произвести предварительную оценку σ . Поэтому, учитывая, что вычислитель в своих расчетах иногда может допустить ошибки, в качестве известного контроля за правильностью вычисления s полезно применять оценку, основанную на размахе варьирования. С этой целью рекомендуется запомнить следующие отношения σ к размаху.

Если n примерно равно 5, 10, 25, 100, то грубая оценка σ получится от деления размаха варьирования на 2, 3, 4, 5.

В обычной статистической практике такая грубая и в то же время быстрая оценка σ является вполне приемлемым способом оценки, но в тех условиях, в каких получают экспериментальные данные, становятся оправданными большие затраты труда на вычисление s .

Если вы знакомы с математической статистикой, то слышали о принципе наименьших квадратов. Согласно этому принципу, средняя и отклонения от нее находятся в следующем соотношении: если взять отклонения от средней, то сумма их квадратов будет минимальной. Отсюда, в частности, следует, что если в таблице 9 взять отклонения от какого-либо числа, отличающегося от средней, равной 20, то сумма квадратов этих отклонений будет больше 254 (здесь указанное выше положение взято в своей обратной формулировке). Проверьте это, взяв отклонения, положим, от 19 или от 22.

Можно отметить и такую характерную черту варьирования: объекты, имеющие большие размеры, обычно варьируют сильнее, малые — меньше. Поэтому часто представляет известные удобства выражение стандартного отклонения в долях средней; такой производный показатель называется *относительным стандартным отклонением*, или *коэффициентом вариации*, и обозначается C . Например, установлено [1], что средний рост у однолетних и восемнадцатилетних девочек соответственно равен 74,4 и 161,0 см, а стандартные отклонения — 2,64 и 6,12 см. Коэффициенты вариации

$$C_1 = 2,64/74,4 = 0,036$$

$$C_{18} = 6,12/161,0 = 0,038$$

почти одинаковы. Обычно S выражается в процентах; например, S_1 составляет 3,6%. Более подробному рассмотрению вопроса об этой характеристике будет посвящен параграф 16 этой главы.

Пример 15. Дан вес 20 только что родившихся морских свинок, взятых по две из каждого помета: 30, 30, 26, 32, 30, 23, 29, 31, 36, 30, 25, 34, 32, 24, 28, 27, 38, 31, 34, 30 г. Произвести оценку σ тремя способами: 1) грубо приближенно, взяв одну четвертую часть от размаха варьирования. *Ответ:* 3,8 г; 2) используя коэффициент 0,268, взятый из таблицы 10. *Ответ:* 4,0 г; 3) путем вычисления s . *Ответ:* 3,85 г. *N.B.* Определить время, употребленное на вычисление s .

Пример 16. По данным предыдущего примера установить, сколько наблюдений необходимо для того, чтобы на основе размаха варьирования получить результат с такой же точностью, как и в случае определения s . *Ответ:* 29 наблюдений.

Пример 17. Положим, что на процесс взвешивания каждой морской свинки (извлечение из клетки, возврат в нее, взвешивание и запись результата) уходит 5 минут и еще требуется 2 минуты на расчеты размаха варьирования; получите ли вы экономию или, наоборот, потерю времени, если произведете оценку при помощи s .

Пример 18. Допустим, что 16 восемнадцатилетних новобранцев расположены в ряд соответственно их росту, после чего произведено измерение роста самого низкого из них, что дало 64 дюйма, и самого высокого — 72 дюйма. Можно ли в качестве грубо приближенной оценки μ принять середину интервала $(64+72):2=68$ дюйма, а в качестве такой же упрощенной оценки σ величину $8/3=2,7$ дюйма?

Пример 19. Средний урожай сена люцерны, убранный с 15 делянок, равен 2,2 т на 1 акр при $s=0,35$ т на 1 акр. Определить приближенно размах варьирования, приняв $\sigma=0,35$ т на 1 акр. *Ответ:* 1,2 т на 1 акр. Можно ли считать, что наивысший урожай с делянки в этом случае составляет около 2,8 т на 1 акр?

7. Распределение t Стьюдента. Теперь после того, как установлены оценки значения для μ и σ , следует рассмотреть вопрос об *оценках интервала и критериях при проверке гипотез*.

Прежде всего следует познакомиться с одним выборочным распределением, аналогичным распределению хи-квадрат. Оно известно под названием *распределения t Стьюдента*, открыто В. С. Госсетом в 1908 г. [15] и уточнено Р. А. Фишером в 1924 г. [6]. Это распределение произвело революцию в теории малых выборок. В следующей главе будет рассмотрен вопрос о проверке этого распределения по примеру проверки, ранее примененной по отношению распределения хи-квадрат; именно этот вид выборочного наблюдения рассматривался Госсетом, когда он впервые изучал это распределение.

Величина t определяется равенством:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Это значит, что t является отклонением выборочной средней от средней совокупности, выраженным в долях величины s/\sqrt{n} , принятой за единицу. Как \bar{x} , так и s определяются по выборке, состоящей из n наблюдений, причем предполагается, что выборка случайна и взята из нормальной совокупности. Вообще говоря, значение μ нам не известно, однако относительно этого значения может быть выдвинута та или иная гипотеза. Хотя t не может быть вычислено, если μ не известно, однако это не является препятствием для установления распределения t при наличии гипотезы относительно μ .

Делитель s/\sqrt{n} является показателем, оценивающим величину σ/\sqrt{n} , или *стандартную ошибку*. Мы будем s/\sqrt{n} называть стандартной ошибкой выборки и обозначим $s_{\bar{x}}$. Более подробные сведения об этой величине будут даны в главе 3.

По данным о содержании витамина С в таблице 9 находим: $s_{\bar{x}} = 3,98 : \sqrt{17} = 0,965$ мг/100 г.

Распределение t представлено в таблице 11. Для больших выборок это распределение практически совпадает с нормальным при $\mu=0$ и $\sigma=1$. Только при выборках, численность которых меньше 30, различие между этими распределениями становится достаточно явным.

Как и нормальное распределение, распределение t симметрично относительно средней. В связи с этим таблицы составлены так, что вероятности

Распределение t^1

Степени свободы	Вероятность превзойти данное значение по абсолютной величине								
	0,500	0,400	0,200	0,100	0,050	0,025	0,010	0,005	0,001
1	1,000	1,376	3,078	6,314	12,706	25,452	63,657	—	—
2	0,816	1,061	1,886	2,920	4,303	6,205	9,925	14,089	31,598
3	0,765	0,978	1,638	2,353	3,182	4,176	5,841	7,453	12,941
4	0,741	0,941	1,533	2,132	2,776	3,495	4,604	5,598	8,610
5	0,727	0,920	1,476	2,015	2,571	3,163	4,032	4,773	6,859
6	0,718	0,906	1,440	1,943	2,447	2,969	3,707	4,317	5,959
7	0,711	0,896	1,415	1,895	2,365	2,841	3,499	4,029	5,405
8	0,706	0,889	1,397	1,860	2,303	2,752	3,355	3,832	5,041
9	0,703	0,883	1,383	1,833	2,262	2,685	3,250	3,690	4,781
10	0,700	0,879	1,372	1,812	2,228	2,634	3,169	3,581	4,587
11	0,697	0,876	1,363	1,796	2,201	2,593	3,106	3,497	4,437
12	0,695	0,873	1,356	1,782	2,179	2,560	3,055	3,428	4,318
13	0,694	0,870	1,350	1,771	2,160	2,533	3,012	3,372	4,221
14	0,692	0,868	1,345	1,761	2,145	2,510	2,977	3,326	4,140
15	0,691	0,866	1,341	1,753	2,131	2,490	2,947	3,286	4,073
16	0,690	0,865	1,337	1,746	2,120	2,473	2,921	3,252	4,015
17	0,689	0,863	1,333	1,740	2,110	2,458	2,898	3,222	3,965
18	0,688	0,862	1,330	1,734	2,101	2,445	2,878	3,197	3,922
19	0,688	0,861	1,328	1,729	2,093	2,433	2,861	3,174	3,883
20	0,687	0,860	1,325	1,725	2,086	2,423	2,845	3,153	3,850
21	0,686	0,859	1,323	1,721	2,080	2,414	2,831	3,135	3,819
22	0,686	0,858	1,321	1,717	2,074	2,406	2,819	3,119	3,792
23	0,685	0,858	1,319	1,714	2,069	2,398	2,807	3,104	3,767
24	0,685	0,857	1,318	1,711	2,064	2,391	2,797	3,090	3,745
25	0,684	0,856	1,316	1,708	2,060	2,385	2,787	3,078	3,725
26	0,684	0,856	1,315	1,706	2,056	2,379	2,779	3,067	3,707
27	0,684	0,855	1,314	1,703	2,052	2,373	2,771	3,056	3,690
28	0,683	0,855	1,313	1,701	2,048	2,368	2,763	3,047	3,674
29	0,683	0,854	1,311	1,699	2,045	2,364	2,756	3,038	3,659
30	0,683	0,854	1,310	1,697	2,042	2,360	2,750	3,030	3,646
35	0,682	0,852	1,306	1,690	2,030	2,342	2,724	2,996	3,591
40	0,681	0,851	1,303	1,684	2,021	2,329	2,704	2,971	3,551
45	0,680	0,850	1,301	1,680	2,014	2,319	2,690	2,952	3,520
50	0,680	0,849	1,299	1,676	2,008	2,310	2,678	2,937	3,496
55	0,679	0,849	1,297	1,673	2,004	2,304	2,669	2,925	3,476
60	0,679	0,848	1,296	1,671	2,000	2,299	2,660	2,915	3,460
70	0,678	0,847	1,294	1,667	1,994	2,290	2,648	2,899	3,435
80	0,678	0,847	1,293	1,665	1,989	2,284	2,638	2,887	3,416
90	0,678	0,846	1,291	1,662	1,986	2,279	2,631	2,878	3,402
100	0,677	0,846	1,290	1,661	1,982	2,276	2,625	2,871	3,390
120	0,677	0,845	1,289	1,658	1,980	2,270	2,617	2,860	3,373
∞	0,6745	0,8416	1,2816	1,6448	1,9600	2,2414	2,5758	2,8070	3,2905

¹ Эта таблица по частям с разрешения авторов взята у Р. А. Фишера: «Statistical Methods for Research Workers», издание Oliver and Boyd, Edinburgh (1925—1950); у М. Меррингтон: «Table of Percentage Points of the t-Distribution», Biometrika, 32, 300 (1942), и у Бернарда Остла: «Statistics in Research», Iowa State College Press (1954).

определяются для значений t , превосходящих некоторое абсолютное число, т. е. без учета знака. Рассмотрим в качестве примера значение $t=1,96$ при бесконечном (∞) числе степеней свободы, т. е. в нормальном распределении. В данном случае вероятность равна 0,05. Это означает, что из общего числа выборок большого размера, взятых случайным порядком из нормальной совокупности, 5% выборок будут иметь или $t > 1,96$, или $t < -1,96$. На рисунке 5 значения t показаны заштрихованными площадями; 2,5% из них находится в одном конце кривой и 2,5% — в другом.

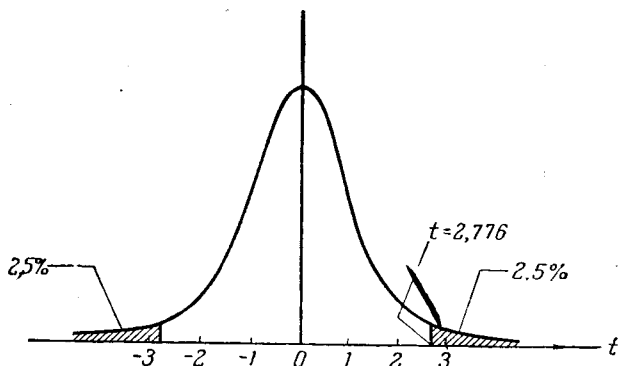


Рис. 5. Распределение t при 4 степенях свободы. Заштрихованные площади заключают в себе 5% всей площади. Распределение имеет более острую вершину и более приподнятые концы по сравнению с нормальным распределением.

Таким образом, таблица дает совмещение двух равных и симметрично расположенных частей графика, составляющих вместе всю заштрихованную площадь (сумму вероятностей). Следовательно, вероятность, определяемая двумя концами распределения t , соответствует вероятности одного конца в распределении хи-квадрат, приведенном в таблице 6. Основание для такого построения таблицы будет указано ниже.

Пример 20. Имея данные таблицы 9 о содержании витамина С, принять гипотезу о том, что $\mu=17,954$ мг/100 г. Вычислите t . *Ответ:* 2,12.

Пример 21. В выборке, характеризующей содержание витамина С, число степеней свободы (являющееся знаменателем частного при определении s^2) равно $17-1=16$. Найти по таблице 11 вероятность появления значения t , превосходящего по своей абсолютной величине 2,12. *Ответ:* 0,05. Это означает, что среди случайных выборок из нормальной совокупности, имеющих объем $n=17$, из общего их числа 5% будут иметь t меньше $-2,12$ или больше $+2,12$.

Пример 22. Имеется ряд выборок объема $n=17$, взятых случайным порядком из нормальной совокупности; по каждой из них определена величина t . Какова вероятность того, что t содержится между $-2,12$ и $+2,12$? *Ответ:* 0,95.

Пример 23. Случайная выборка объема $n=17$ взята из нормальной совокупности; какова вероятность того, что t превосходит 2,12? *Ответ:* 0,025.

Пример 24. При каком размере случайной выборки из нормальной совокупности в 5% всех случаев будет наблюдаться $t > |2|$? *Ответ:* 61. (Напомним, что символ «абсолютное значение» — две вертикальные черты — означает, что знак величины опускается.)

Пример 25. Какое значение может превзойти t в 2,5% всех случаев, если размер выборки очень большой (число степеней свободы = ∞)? *Ответ:* 1,96.

8. Оценка интервала для μ . Доверительный интервал. Обоснование оценки интервала для μ несколько сложно, и поэтому я сначала покажу, как строится этот интервал, и уже потом дам соответствующие объяснения. Для конкретности вернемся к таблице 9 с ее данными о содержании витамина С; в этом случае $n=17$, $\bar{x}=20$ и $s_{\bar{x}}=0,965$ мг/100 г. Найдем доверительный интервал (оценку интервала) при вероятности в 95%.

1. На пересечении строки таблицы 11 для числа степеней свободы $=17-1=16$ и столбца, озаглавленного 0,05, находим $t_{0,05}=2,12$.

2. Вычисляем величину:

$$t_{0,05} \times s_{\bar{x}} = 2,12 \times 0,965 = 2,05 \text{ мг/100 г.}$$

3. Доверительный интервал в этом случае будет от $20 - 2,05 = 17,95$ до $20 + 2,05 = 22,05$ мг/100 г.

Если вы будете утверждать, что интервал от 17,95 до 22,05 мг/100 г покрывает μ , то шанс в процессе производства выборок оказаться неправым составит 1 из 20.

Оценка интервала для μ при вероятности 95 % может быть представлена в более компактной форме: $20 \pm 2,05$ мг/100 г (или в общем виде при помощи формулы $\bar{x} \pm t_{0,05} \times s_{\bar{x}}$).

Для применения этой схемы расчетов необходимо выбрать надлежащее значение t из соответствующей таблицы распределения t . Если найдено табличное значение $t_{0,05}$ для числа степеней свободы $n-1$, то можно ожидать, что это значение t по абсолютной своей величине будет превзойдено в 5 % всех выборок при условии случайного отбора из нормальной совокупности. Это же положение может быть представлено в такой измененной форме: в 95 % всех выборок t окажется между $-t_{0,05}$ и $+t_{0,05}$, т. е. вероятность того, что $-t_{0,05} \leq t \leq t_{0,05}$, равна 0,95. Если же произвести подстановку вместо t указанного ранее его выражения, то можно сказать, что вероятность того,

что $-t_{0,05} \leq \frac{\bar{x} - \mu}{s_{\bar{x}}} \leq t_{0,05}$, равна 0,95. Умножая каждую из частей этих неравенств на $s_{\bar{x}}$, при той же вероятности 0,95 получим $-t_{0,05} \times s_{\bar{x}} \leq \bar{x} - \mu \leq t_{0,05} \times s_{\bar{x}}$. Переносим \bar{x} , меняя знаки на обратные и перемещая крайние члены этого выражения, окончательно получим:

$$\bar{x} - t_{0,05} \times s_{\bar{x}} \leq \mu \leq \bar{x} + t_{0,05} \times s_{\bar{x}}.$$

Этот результат интерпретируется так: до получения выборки взята вероятность 0,95 того, что указанный выше интервал содержит в себе неизвестное μ . После того как выборка произведена и в формулу интервала подставлены значения \bar{x} , s и n , можно с определенной уверенностью утверждать, что этот интервал содержит в себе μ ; это положение может оказаться в процессе производства выборок неверным в одном случае из 20.

Об оценке интервала для σ будет сказано в параграфе 14.

Пример 26. По данным об урожаях люцерны в примерах 3 и 14 найдено $n=10$, $\bar{x}=1,70$ и $s=0,41$ т на 1 акр. Определить 95%-ные доверительные пределы для средней той совокупности, из которой взята эта случайная выборка. *Ответ:* 1,41 и 1,99 т на 1 акр.

Пример 27. Пять разностей из примеров 5 и 12 характеризуются показателями $\bar{d}=9,0$, $s_D=7,8$ бушелей на 1 акр. Определить 99%-ную оценку интервала для μ . *Ответ:* от $-7,0$ до $25,0$ бушелей на 1 акр. Возможно ли, что разность в совокупности равна нулю?

Пример 28. В одном исследовании, в котором изучался рост школьников 8 частных школ [7], средний рост 265 мальчиков в возрасте от $13\frac{1}{2}$ до $14\frac{1}{2}$ лет оказался 63,84 дюйма при стандартном отклонении 3,08 дюйма. Каким будет 95%-ный доверительный интервал для μ ? *Ответ:* от 63,5 до 64,2 дюйма. Определить C . *Ответ:* $C=4,8\%$.

9. **Оценки и критерии существенности для разностей.** Чаще всего опыты предназначаются не столько для установления самих эффектов, сколько для установления и оценки разностей между эффектами, например разностей между урожаями, полученными при использовании различных удобрений, или разности между приростами в весе животных, полученными при различных рационах. Одна из наиболее простых форм такого рода экспериментов предусматривает сопоставление эффектов двух вариантов. В этом случае образуются попарные наблюдения, одно из которых относится к первому варианту, а другое — ко второму. Такими парными наблюдениями могут быть опытные делянки, или свиньи, или семьи пчел и т. д. Если имеется только одна такая пара наблюдений, то нельзя установить, будет ли соответствующая разность обусловлена действием этих двух вариантов, или есте-

ственным варьированием наблюдений, или тем и другим вместе. Следовательно, необходимо, чтобы было две или несколько таких пар наблюдений, т. е. *посторений*, построенных так, чтобы один из членов каждой пары, выбранный в случайном порядке, предназначался для испытания первого варианта и другой — для испытания второго. Разности между показаниями двух таких пар образуют выборку, на основе которой мы делаем свои выводы. Если бы случайное варьирование полностью отсутствовало, то все разности были бы равны друг другу, однако случайное варьирование всегда имеет место, вследствие чего варьируют и разности.

В условиях большинства экспериментов вполне допустимо считать, что разности образуют случайную выборку из нормальной совокупности. В этом случае возникает задача об определении значения средней в этой совокупности μ , в частности, вопрос, можно ли считать эту среднюю отклоняющейся от нуля.

Рассмотрим пример такого опыта.

Юден и Бил [19] поставили задачу определить, имеется ли различие в действии на листья табака двух препаратов вируса табачной мозаики. Примененный ими метод состоял в том, что одна половина листа натиралась куском марли, смоченным в одном препарате экстракта вируса, а вторая половина листа таким же образом обрабатывалась другим препаратом. Сила действия препарата характеризовалась числом мест поражения вирусом на соответствующей половине листа; эти данные рассматривались в качестве значений непрерывной переменной величины.

Данные, приведенные в таблице 12, относятся ко второму листу на каждом из 8 растений. Разности $D = X_1 - X_2$ четвертого столбца таблицы представляют собой интересующую экспериментатора случайную выборку. Вопрос, на который должен дать ответ эксперимент, состоит в следующем: оказывает ли свое влияние различие двух препаратов вируса на число пораженных мест? В статистических терминах это звучит так: можно ли считать, что средняя совокупности, из которой взята выборка, равна нулю или она отличается от нуля?

На этот вопрос мы можем ответить, определив доверительные интервалы для средней разности μ_D . Имея $\bar{d} = 4$ точкам поражения, $s_{\bar{d}} = 1,52$ поражения и $t_{0,05} = 2,365$, мы можем сказать, что нарушение неравенства

$$4 - 2,365 \times 1,52 \leq \mu_D \leq 4 + 2,365 \times 1,52$$

имеет только 1 шанс из 20. Это означает, что μ_D ожидается больше $4 - 2,365 \times 1,52 = 0,4$ поражения и меньше 7,6 поражения. Так как нуль не содержится в этом интервале, то можно сделать вывод, что μ_D — число положительное, т. е. что второй препарат дает меньшее число пораженных мест, чем первый.

Если бы доверительный интервал содержал в себе нуль, то следовало бы сделать вывод, что μ_D может быть равно нулю, и, следовательно, возможно, что оба препарата обладают одинаковой силой действия.

Читатель может высказать удивление по поводу того, что доверительные границы выражаются дробным числом мест поражения. Дело в том, что эти явно дискретные данные мы обработали в предположении, что они взяты из непрерывной совокупности. Но возникает вопрос: можно ли данный метод обработки с одинаковым правом применять как к результатам подсчета, т. е. дискретным величинам, так и к результатам измерения, т. е. непрерывным величинам; иными словами, можно ли в том и другом случае вводить допущение о нормальном распределении? На этот вопрос не всегда можно категорически ответить «да». Целочисленные результаты подсчета могут дать распределение, близкое к нормальному, и в этих случаях вполне допустимо применить описанный здесь метод обработки, однако из данных настоящей выборки отнюдь не следует, что эти условия имеют место и в настоящем случае. С другой стороны, здесь нельзя говорить о каких-либо теоретических численностях, и поэтому методы главы 1 не могут быть применены.

Число пораженных мест на половине 8 листьев табака; у каждого из 8 растений взят второй лист. Числом поражений измеряется действие двух препаратов вируса мозаичной болезни табака¹

Номер растения	Число поражений на половине листа		Разность $D = X_1 - X_2$	Отклонение $d = D - \bar{d}$	Квадрат отклонения d^2
	препарат 1 X_1	препарат 2 X_2			
1	9	10	-1	-5	25
2	17	11	6	2	4
3	31	18	13	9	81
4	18	14	4	0	0
5	7	6	1	-3	9
6	8	7	1	-3	9
7	20	17	3	-1	1
8	10	5	5	1	1
Итого	120	88	32	0	130
Средняя	15	11	$\bar{d} = 4$ поражениям		$s_D^2 = 18,57$

$$s_D = \sqrt{18,57} = 4,31 \text{ поражения;}$$

$$s_{\bar{d}}^2 = 18,57/8 = 2,32;$$

$$s_{\bar{d}} = 1,52$$

¹ Данные несколько изменены в целях упрощения расчетов.

Вместе с тем следует учесть и то обстоятельство, что теория нормального распределения имеет широкое применение в связи с тем, что она дает хорошие результаты даже тогда, когда выборки берутся из совокупностей, значительно отклоняющихся от нормальной. В дальнейшем (пример 23 главы 11) будут рассмотрены и некоторые другие возможности, но приведенные выше выводы останутся без изменения.

Другим способом заключений относительно μ_D является классический метод применения критерия *существенности*. По этому вопросу отсылаем читателя к параграфу 13 главы 1. Обычная форма нулевой гипотезы состоит в допущении $\mu_D = 0$. Если выборки взяты из нормальной совокупности, то критерий такой гипотезы формируется на основе распределения t . Для примера с мозаичной болезнью табака мы имеем гипотезу H_0 , согласно которой $\mu_D = 0$, а также выборочные характеристики $\bar{d} = 4$ поражениям и $s_{\bar{d}} = 1,52$ поражения; делая подстановку всех этих значений в формулу t , получим:

$$t = \frac{\bar{d} - \mu}{s_{\bar{d}}} = \frac{4 - 0}{1,52} = 2,63; \text{ число степеней свободы} = n - 1 = 7.$$

Является ли значение t обычным для выборок из нормальной совокупности со средней $\mu_D = 0$ или, наоборот, оно столь необычно, что заставляет отвергнуть данную нулевую гипотезу? По таблице 11 находим, что при 7 степенях свободы вероятность получить в выборках при $\mu_D = 0$ значение t , превосходящее 2,63, несколько меньше 0,05 и примерно равна 0,04; очевидно, что настоящая выборка дает основания отвергнуть гипотезу H_0 . В этом случае разность $\bar{d} = 4$ мест поражения может быть названа *существенной*. Отсюда следует заключение, что сравниваемые два препарата дают определенно различное число поражений на вторых листьях растений табака.

Таким образом, мы видим, что два метода исследования, касающиеся μ_D , приводят к одному и тому же выводу. Доверительный интервал дает оценку

того, что μ_D находится вне (или внутри) этого интервала. Критерий же существенности дает вероятность появления в выборках из совокупности при $\mu_D=0$ значений, больших наблюдаемого t . Из этих методов можно применить тот или другой, или оба вместе.

В опытах, в которых, подобно рассмотренному, образуются пары наблюдений, величина s_D^2 (или иногда s_D) называется *экспериментальной ошибкой*. Она является оценкой σ_D^2 , которая характеризует варьирование совокупности выборочных разностей. При планировании опыта весьма важно обеспечить правильную оценку действительной ошибки опыта. В дальнейшем мы еще не раз будем говорить об оценке экспериментальной ошибки.

Пример 29. Л. С. Гров [8] определил среднее число соцветий гладнолуса сорта Превосходный на 7 парах делянок; на одной делянке из каждой пары были высажены крупные (одногодичные) клубни, на другой — мелкие (двухлетние и более старые) клубни. Средние по отдельным делянкам таковы:

	Клубни		Число соцветий						
	Крупные	Мелкие	1	2	3	4	5	6	7
Крупные	11,2	13,3	12,8	13,7	12,2	11,9	12,1		
Мелкие	14,6	12,6	15,0	15,6	12,7	12,0	13,1		

Вычислить среднюю разность. *Ответ* 1,2 соцветия. Проверить относящуюся к совокупности таких разностей гипотезу $H_0: \mu_D=0$. *Ответ*: приблизительно $P=0,06$. Строго говоря, гипотезу H_0 нельзя считать отвергнутой, но все же следует вспомнить замечания последнего абзаца в параграфе 13 главы 1.

Пример 30. Митчелл, Беррос и Видлс [10] определили биологическую ценность протеина, полученного из сырых (P) и поджаренных (R) семян арахиса, в эксперименте с 10 парами крыс. Парные значения P и R получены следующие: 61 и 55, 60 и 54, 56 и 47, 63 и 59, 56 и 51, 63 и 61, 59 и 57, 56 и 54, 44 и 63, 61 и 58. Определить выборочную среднюю разность (*Ответ*: 2,0) и стандартное отклонение разностей (7,72 единицы). Так как здесь $t=0,82$, то почти в половине таких выборок из совокупности, средняя которой $\mu_D=0$, следует ожидать значения t , превосходящие данное значение.

Отметим, что 9 из 10 разностей $P-R$ положительны. Это говорит о наличии некоторых особенностей у пары 44 и 62. Здесь первое число кажется явно ненормальным. Столь необычные наблюдения, как данное, времяами появляются и в опытах, проведенных с достаточной тщательностью; причины их возникновения всегда подлежат специальному изучению. Несомненно, что здесь были произведены поиски ошибок в записях и расчетах, но они не были обнаружены. Как следует поступать в отношении таких искаженных наблюдений — вопрос спорный; во всяком случае их появление уменьшает доверие к опыту. Этот вопрос будет обсуждаться в дальнейшем, в параграфе 15 этой главы.

Пример 31. Кремpton [4] сообщает данные о приросте в весе в пересчете на 100 фунтов кормов у 10 пар свиней, выбранных по жребию в группы I и III (опыт по кормлению свиней в колледже Макдональда № 33 Б). При подборе пар животных стремились, чтобы исходный вес их по возможности был одинаковым. Приросты в весе по парам получились следующие (данные о животных группы I поставлены первыми): 17 и 20; 22 и 21; 22 и 21; 15 и 24; 24 и 24; 22 и 22; 21 и 23; 21 и 22; 17 и 22; 21 и 23 фунта прироста на 100 фунтов кормов. Определить среднюю разность. *Ответ*: 2,0 фунта на 100 фунтов кормов. Проверить гипотезу, что оба рациона дают одинаковый прирост веса. *Ответ*: приблизительно $P=0,07$; гипотеза H_0 , пожалуй, не может быть отвергнута. Вычислить 95%-ный доверительный интервал для μ . *Ответ*: от $-0,2$ до 4,2 фунта на 100 фунтов кормов.

Пример 32. Профессор Х. Х. Лав из Корнельского университета может считаться первым из тех, кто содействовал в Америке введению теории малых выборок Стьюдента. В качестве одной из иллюстраций применения этой теории он проводит [9] сравнение урожаев двух сортов овса: Грейт Норсерн (Great Northern)— G и Биг Фур (Big Four)— B за девятилетие 1912—1920 гг. Получены такие разности $G-B$: 16,3; 13,4; 3,8; 7,9; 2,6; 2,5; 9,6; 7,2 и 3,3 бушеля на 1 акр. Вычислено $\bar{x}=7,4$ бушеля на 1 акр. Проверить нулевую гипотезу о том, что урожай этих двух сортов был одинаковым. *Ответ*: $P < 0,01$. К какой совокупности принадлежит эта случайная выборка? Построить 95%-ный доверительный интервал для совокупности разностей между урожаями сортов. *Ответ*: от 3,6 до 11,2 бушеля на 1 акр.

10. Основания и условия парного сравнения. Читатель уже заметил, что при проведении опытов, подобных описанному выше опыту Юдена и Биля, большое значение имеет предварительное знакомство с особенностями подопытного материала. Так, в опыте, в котором наблюдения проводятся на

каждой половине листа, особенно эффективным оказался метод парного сравнения. В этом случае принято во внимание то, что наблюдается большое различие в реакции листьев на вирус, прежде всего в зависимости от высоты прикрепления их к стеблю. Но даже если произвести случайный отбор вторых листьев, но использовать их полностью, а не их половины, то разности $D = X_1 - X_2$ окажутся гораздо большими, чем те, которые были получены в таблице 12. В этом возрастании варьирования можно убедиться на основе данных второй и третьей граф этой таблицы; если по этим данным образовать случайные пары, то размах варьирования может достигнуть значения 37 (от $7 - 18 = -11$ до $31 - 5 = 26$). Это даст оценку σ , примерно равную $37/3 = 12$ поражениям вместо фактической оценки в опыте $s = 4,3$ поражения. Конечно, возможность появления такого крайнего сочетания пар представляет маловероятный случай. Как будет дальше показано, для достижения той же точности, которая получена в опыте Юдена и Била при 8 парах, при случайном отборе пар необходимо иметь 40 парных наблюдений. Следовательно, если варианты опыта располагать на целых листьях второго яруса, то стоимость опыта возрастет в 5 раз.

Возникает вопрос: какие предварительные сведения мы должны иметь, чтобы применить парный метод? Независимо от того, будет ли применен этот метод или не будет, мы всегда должны быть уверены в том, что при тождественности вариантов опыта соответствующие наблюдения (если оставить в стороне случайное варьирование) должны быть тождественными. В противном случае эксперимент не дает определенного ответа: мы не сможем решить, будут ли наблюдаемые различия связаны с различием вариантов или они происходят от других причин. Парный метод применим в том случае, если будут образованы такие пары, что различия внутри них будут меньше, чем различия между парами. Например, сходство двоен является естественным условием, позволяющим образовать такие пары; поэтому в опытах с животными такие пары обычно создаются из животных одного и того же пола, принадлежащих к одному помету. В этом случае при тождественности вариантов опыта сходство животных, составляющих пару, больше, чем животных, находящихся в менее родственных отношениях. Два наблюдения над одним и тем же объектом также естественным образом образуют пару связанных между собой наблюдений: например, результаты взвешивания участника футбольной игры до и после игры или результаты анализа на сахар крови у одного и того же лица до и после инъекции инсулина. Наконец, очень часто образованию парных наблюдений содействует изменчивость, обусловленная сходством окружающей обстановки. Так, два варианта на опытном поле лучше всего располагать рядом друг с другом, а в теплице — так, чтобы устранить влияние возможных различий в почве, влажности, температуре и т. п. Две делянки или два сосуда, расположенные близко друг к другу, обеспечивают большее сходство условий, чем делянки или сосуды, удаленные друг от друга. Животные, предназначенные для двух вариантов опыта, должны находиться или в одном загоне, или в смежных загонах.

Метод парного сравнения должен применяться везде, где для этого имеются подходящие условия. Если перед вами стоит задача изучения различий между эффектами вариантов, то наиболее экономным является испытание этих вариантов на парных объектах, по возможности близких друг к другу во всех других отношениях.

Если бы было известно, что два взятых для опыта объекта абсолютно тождественны друг другу, то различие между вариантами могло бы быть установлено путем сравнения одной пары объектов. Чем больше различий между объектами, тем больше потребуются таких парных наблюдений для того, чтобы сбалансировать различия случайного характера и получить достаточно явные результаты. Такое сбалансирование случайностей обеспечивается при помощи распределения вариантов между членами каждой пары случайным порядком, по жребию. Наиболее выгодно образовывать пары из схожих объектов.

В главе 4 будут рассмотрены специальные статистические методы для таких случаев, когда нет оснований для применения парного метода.

Прежде чем перейти к некоторым второстепенным вопросам, резюмируем изложенное выше. Сначала мы познакомились с нормальным распределением, имеющим два параметра: среднюю μ и стандартное отклонение σ (или дисперсию σ^2). Далее мы ввели некоторые показатели — оценки, получаемые на основе данных случайной выборки.

1. Выборочная средняя \bar{x} , являющаяся оценкой μ .
2. Выборочное стандартное отклонение s и размах варьирования, умноженный на соответствующий коэффициент, в качестве оценки σ .
3. Средний квадрат s^2 в качестве оценки σ^2 .

После того как было установлено распределение критерия t , произведена оценка интервала для μ и проверка гипотезы $\mu=0$ при альтернативной гипотезе $H_A: \mu \neq 0$. Наконец, мы рассмотрели вопрос об оценке различия в эффекте двух вариантов на основе эксперимента с повторными парами сходных друг с другом объектов, который представляет собой случайную выборку соответствующих разностей.

Дальше в этой главе будут затронуты вопросы о нулевых гипотезах иного характера, о другой форме альтернативной гипотезы относительно μ при построении интервальных оценок и о критериях гипотез относительно σ^2 . Но прежде чем перейти к рассмотрению этих вопросов, полезно обратить внимание на упрощение вычислительной работы.

11. Вычисление s без счетной машины. Пока мы имели дело с такими примерами, у которых ΣX точно делилась на n . Это упрощало работу по вычислению квадратов отклонений. На практике, когда такое точное деление невозможно, вычисление средней производится с одним или двумя добавочными знаками по отношению к исходным данным; в связи с этим вычисление отклонений и их квадратов становится уже более трудоемким. Эти вычисления могут быть упрощены, если брать отклонения от некоторого произвольно выбранного числа G ; оно обычно берется где-нибудь поблизости от средней. Сумма отклонений от G уже не будет равна нулю, но она позволяет ввести поправку, которая дает возможность вычислить \bar{x} с любой точностью. Сумма квадратов отклонений от G больше суммы $\Sigma x^2 = \Sigma (X - \bar{x})^2$, так как последняя является наименьшей суммой квадратов, но опять сумма $\Sigma (X - G)$ дает возможность ввести поправку, которая позволяет получить Σx^2 с требуемой точностью.

Для показа этого способа возьмем данные о весе при рождении 20 морских свинок, представленные в левом столбце таблицы 13 (см. пример 15). Не вычисляя средней \bar{x} , просматриваем ряд данных и выбираем условное начало, положим $G=30$ г, близкое к среднему значению ряда. Далее будет показано, что выбор, положим, 29 или 40 г или какого-либо другого произвольно взятого числа все равно приведет после введения поправок к одним и тем же результатам. Но вычисления значительно проще, если G взять вблизи от \bar{x} .

Сумма отклонений от G не равна нулю, ибо 30 г не является средней. Но среднее отклонение 0,25 г как раз и будет поправкой для получения точной средней; для получения последней следует к G прибавить это среднее отклонение, что дает $\bar{x}=30,25$ г.

В последней графе таблицы даны квадраты отклонений от G и их сумма 247. Теперь нам следует определить Σx^2 , т. е. сумму квадратов отклонений от средней, обычно называемую просто *суммой квадратов*. Чтобы получить Σx^2 , соответствующую поправку всегда следует вычитать из $\Sigma (X - G)^2$. Полученный таким образом результат 245,75 может быть проверен путем вычисления всех 20 отклонений от \bar{x} ($30 - 30,25 = -0,25$ и т. д.) и суммирования их квадратов. Ниже приводится пример 33, требующий более простых расчетов.

Иногда лица, не имеющие вычислительных навыков, смешивают две величины, входящие в приведенные выше расчеты. Во-первых, сумму квадратов отклонений от G , т. е. $\Sigma(X-G)^2=247$, и, во-вторых, квадрат суммы этих же отклонений [$\Sigma(X-G)$] $^2=25$. В первом случае возведение в квадрат производится до суммирования, во втором же случае порядок этих действий меняется. Содержание этих величин и соответствующие результаты столь различны, что требуется только небольшое внимание для того, чтобы их не спутать.

ТАБЛИЦА 13

Вычисление стандартного отклонения при помощи способа условного начала отсчета
Вес 20 морских свинок при рождении, по два из помета

Вес при рождении (в г)	Отклонение от 30 г		Квадрат отклонений
X	$X-G$		$(X-G)^2$
30		0	0
30		0	0
26	—4		16
32		2	4
30		0	0
23	—7		49
29	—1		1
31		1	1
36		6	36
30		0	0
25	—5		25
34		4	16
32		2	4
29	—1		1
28	—2		4
27	—3		9
38		8	64
31		1	1
34		4	16
30		0	0
Итого	—23	28	247

$$\begin{aligned} \Sigma(X-G) &= 28-23=5 & \Sigma(X-G)^2 &= 247 \\ [\Sigma(X-G)]/n &= 5/20=0,25 & [\Sigma(X-G)]^2/n &= 5^2/20=1,25 \\ x &= G + [\Sigma(X-G)]/n = & \Sigma x^2 &= 245,75 \\ &= 30+0,25= & s^2 &= 245,75/19=12,93 \\ &= 30,25 \text{ г} & s &= 3,60 \text{ г} \end{aligned}$$

Пример 33. Следующие данные взяты такими, чтобы облегчить вычисления: 15, 12, 10, 10, 10, 8, 7, 7, 4, 4, 1.

По примеру таблицы 9 следует вычислить $\bar{x}=8$ и $s=4$. После этого примените различные условные начала, например 5, 10 и 1. Продолжите вычисления до тех пор, пока не убедитесь, что ответы $\bar{x}=8$ и $s=4$ получаются одними и теми же при любом выбранном значении, и пока не приобретете навыка в проведении этих расчетов вполне уверенно. Наконец, примите $G=0$.

Пример 34. Если в таблице 13 взять $G=23$ или, что более удобно, 20, то отрицательных отклонений уже не будет. Имеются ли какие-либо преимущества в том, что в вычислениях участвуют только положительные отклонения?

Пример 35. Тот, кто имеет вкус к алгебраическим преобразованиям, может самостоятельно вывести формулу: $\Sigma x^2 = \Sigma(X-G)^2 - [\Sigma(X-G)]^2/n$. Начните с равенства $x = X - \bar{x} = (X-G) - (\bar{x}-G) = (X-G) - \Sigma(X-G)/n$. После переноса члена $\Sigma(X-G)/n$ возведите в квадрат и просуммируйте получаемые таким образом уравнения.

12. Вычисление s при помощи счетной машины. Описанные выше методы были разработаны с той целью, чтобы производить вычисления, пользуясь по возможности простыми числами. Но если имеется возможность воспользоваться счетной машиной, то оперирование большими числами становится столь же легким, как и работа с небольшими числами. Поэтому теперь можно

преследовать другие цели. Во-первых, освободиться от отрицательных чисел и, во-вторых, исключить большинство действий вычитания. Обе эти цели легко достижимы, если применить метод, специально приспособленный к работе на счетной машине. Этот метод таков.

Вернемся к данным о весе морских свинок в первой графе таблицы 13; их сумма $\Sigma X = 605$ и сумма квадратов $\Sigma X^2 = 18547$ быстро определяются на любой счетной машине, хотя способы вычислений могут быть различными в зависимости от типа машины. Обычно обе эти суммы могут вычисляться одновременно, в соответствии с указаниями, прилагаемыми каждой фирмой, выпускающей счетные машины. При работе на машине нет необходимости выписывать квадраты каждого отдельного наблюдения; здесь в процессе работы производится накопление соответствующей суммы. Только после окончания работы выписываются результаты по схеме, приведенной ниже. В схеме показан метод расчета средней и суммы квадратов при работе на счетной машине на примере данных таблицы 13 о весе 20 морских свинок при рождении:

$$\begin{array}{rcl} n = 20 & \Sigma X^2 = 18547 & \\ \Sigma X = 605 & (\Sigma X)^2/n = 18301,25 & \\ \bar{x} = 30,25 \text{ г} & \Sigma x^2 = 245,75 & \end{array}$$

Средняя в нашем случае будет $(\Sigma X)/n = 605/20 = 30,25$ г. Поправка для суммы квадратов $(\Sigma X)^2/n = 605^2/20 = 18301,25$. Вычитая эту величину из ΣX^2 , получаем прежнюю исправленную сумму квадратов 245,75.

Этот метод вычислений на счетной машине можно рассматривать как частный случай общего метода, описанного в параграфе 11 этой главы. Если считать, что $G=0$, то условное начало является не чем иным, как нулевой точкой на шкале веса. Отклонение $X-G=X-0=X$ является просто самим наблюдаемым значением X . Поправка, необходимая для перехода к точной сумме квадратов, в этом случае будет $[\Sigma(X-G)]^2/n = [\Sigma(X-0)]^2/n = (\Sigma X)^2/n$. Этот метод вычислений на машине не только сокращает время работы, но, исключая действие вычитания при вычислении отклонений, тем самым уменьшает возможность случайных просчетов.

При первом чтении оставшаяся часть этой главы и следующая, третья глава могут быть опущены. В главе четвертой будет рассмотрен вопрос о сравнении двух вариантов при таких условиях, когда нет основы для применения парного метода.

Пример 36. Данные о содержании витамина С в томатном соке в таблице 9 приведены в несколько измененном виде, фактически они были такими:

16, 22, 24, 20, 23, 22, 17, 15, 13, 22, 17, 18, 29, 17
22, 16, 23

Вычислить на счетной машине $\Sigma x = 333$, $\Sigma X^2 = 6773$ и по этим результатам найти $\bar{x} = 19,6$ и $s = 3,95$ мг/100 г.

Пример 37. На основе результатов, полученных в предыдущем примере, построить доверительный интервал для μ при $P = 0,95$. *Ответ:* 17,6—21,6 мг/100 г. Сравнить этот интервал с интервалом в параграфе 8 этой главы.

Пример 38. По данным примера 6 этой главы вычислить $\bar{d} = 3,71$ и $s_d = 4,46$ бушеля на 1 акр. Сравнить этот результат с ранее произведенными оценками. В этом случае при одновременном накоплении ΣX и ΣX^2 отрицательные отклонения могут внести путаницу. Лучшее всего выписать отдельно результаты для положительных и отрицательных отклонений.

	ΣX	ΣX^2
Положительные	66,4	499,90
Отрицательные	-7,0	49,24

Итого 59,4 519,14

Пример 39. Установить 99%-ные доверительные границы для средней разности в совокушности урожайных данных в примерах 6 и 38. *Ответ:* 0,42 и 7,00 бушелей на 1 акр.

Замечание: так как доверительный интервал не включает в себя нуль, то критерий для гипотезы $H_0: \mu=0$ должен отвергнуть ее на 1%-ном уровне. Проверить это.

Пример 40. Убедитесь в справедливости формулы $\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n$. Начните с соотношения $x + \bar{x} = X$. Покажите также, что $\Sigma x^2 = \Sigma X^2 - \Sigma X\bar{x} = \Sigma X^2 - n\bar{x}^2$.

13. Критерий для других видов нулевой гипотезы о значении μ . Нулевая гипотеза $\mu_D = 0$ не является единственно возможной и интересующей практику гипотезой; точно так же *альтернативная* гипотеза $\mu_D \neq 0$ не всегда будет иметь реальный смысл. Пример этого мы найдем в параграфе 3 главы 1.

На 14 фермах была установлена эффективность опрыскивания путем определения урожая при опрыскивании и без опрыскивания на каждом отдельном участке. Соответствующие данные приведены в таблице 14. Средняя разность равна 4,7 бушеля на 1 акр при $s_D = 6,48$ и $s_{\bar{d}} = 6,48/\sqrt{14} = 1,73$ бушеля на 1 акр.

ТАБЛИЦА 14

Урожай кукурузы (в бушелях на 1 акр) при опрыскивании и без опрыскивания, полученный на 14 участках округа Бун, Айова, 1950 г.

С опрыскиванием	64,3	78,1	93,0	80,7	89,0	79,9	90,6
Без опрыскивания	70,0	74,4	86,6	79,2	84,7	75,1	87,3
Разность	-5,7	3,7	6,4	1,5	4,3	4,8	3,3
С опрыскиванием	102,4	70,7	106,1	107,4	74,0	72,6	69,5
Без опрыскивания	98,8	70,2	101,1	83,4	65,2	68,1	68,4
Разность	3,6	0,5	5,0	24,0	8,8	4,5	1,1

Первый случай: гипотеза $H_A: \mu_D > 0$. Ранее было установлено, что опрыскивание при той концентрации, которая была взята, не должно давать снижение урожая. Хотя на первом участке и наблюдается уменьшение урожая, но это следует отнести не за счет опрыскивания, а за счет каких-то других причин, т. е. случайного варьирования. Следовательно, если установлено, что μ_D не равно нулю, то оно может быть только больше нуля. В соответствии с этим в данном опыте будет проверяться гипотеза $H_0: \mu_D = 0$ при альтернативе $H_A: \mu_D > 0$. Как и ранее, имеем:

$$t = \frac{4,7 - 0}{1,73} = 2,72; \text{ число степеней свободы} = 13.$$

Проверка нулевой гипотезы при наличии такого рода альтернативы называется *односторонней* оценкой, так как здесь исключается из рассмотрения левая половина распределения t , содержащая отрицательные значения. Поэтому в таблице 11 надо отбросить половину указанной там вероятности, и, следовательно, у вероятностей, стоящих в заголовке таблицы, например 0,05, теперь надо брать вторую, оставшуюся половину, т. е. 0,025. Отсюда следует такое практическое правило: *чтобы произвести одностороннюю оценку, следует отыскать в таблице 11 выборочное значение t и взять половину указанной там вероятности.*

Применяя это правило к установленному выше $t = 2,72$, находим, что P несколько меньше, чем 0,02/2; следовательно, нулевая гипотеза отвергнута при $P < 0,01$. Очевидно, опрыскивание, уменьшая повреждения от кукурузного мотылька, привело в 1950 г. к увеличению урожая кукурузы в округе Бун.

Второй случай: $\mu_D \neq 0$. Этот же самый опыт в округе Бун можно использовать для проверки нулевой гипотезы, отличающейся от гипотезы $\mu_D = 0$. Целью этого эксперимента может быть проверка гипотезы такого рода:

«Стоимость опрыскивания равна прибыли от увеличения урожая под влиянием этого опрыскивания». При определении стоимости установлено, что опрыскивание по рыночной цене обходится в 3 долл. на 1 акр и что в 1950 г. цена на зерно составляет примерно 1,50 долл. за 1 бушель. Таким образом, 2 бушеля прибавки урожая на 1 акр окупают опрыскивание. Здесь проверке будет подвергнута гипотеза $H_0: \mu_D = 2$ бушелям на 1 акр, при альтернативе $H_A: \mu_D \neq 2$; в результате получаем:

$$t = \frac{4,7 - 2,0}{1,73} = 1,56; \text{ число степеней свободы} = 13.$$

Соответствующая вероятность будет около $P = 0,15$, и поэтому данную гипотезу, пожалуй, не следует считать отвергнутой. Фермеры округа Бун могут считать, что при прочих равных условиях в целом они не понесут убытка от применения опрыскивания, конечно, при условии, что все они будут применять его. Это вряд ли может быть стимулом для внедрения опрыскивания, если не учитывать, что опрыскивание может привести к уменьшению повреждений кукурузным мотыльком в следующем году.

Третий случай: альтернативная гипотеза $H_A: \mu_D > \mu_0$. Возможен такой случай, когда гипотеза $H_0: \mu_D = 2$ бушелям на 1 акр проверяется при альтернативе $H_A: \mu_D > 2$ бушелей на 1 акр; это означает, что альтернативная гипотеза может быть выражена словами: «Опрыскивание выгодно». Если взять эту гипотезу, то при $t = 1,56$ будет $P = 0,15/2 = 0,075$, что означает несущественность выгоды. Следует обратить внимание, что при такой односторонней оценке гипотезы она принимается без учета того, что выборочная средняя может оказаться меньше 2 бушелей на 1 акр. Я думаю, что такая постановка вопроса не соответствует реальным условиям. Действительно, никто не пойдет на то, чтобы принять $\mu_D = 2$, если выборочная средняя будет, например, —2 или —3 бушеля на 1 акр. Скорее всего отвергнут гипотезу $\mu_D = 2$ и будут считать, что опрыскивание не оправдывает себя, т. е. что $\mu_D < 2$ бушелей на 1 акр. Это означает, что в данном случае пригодна только двухсторонняя оценка гипотезы.

На это положение следует обратить особое внимание, так как в последнее время появилась тенденция слишком широкого использования односторонней оценки, основывающаяся на таких высказываниях: «Я не интересуюсь другой альтернативой». Если основываться на такой посылке, то оценка становится явно субъективной, так как результаты опыта будут различными лицами интерпретироваться по-разному. Это в известной мере равносильно тому, когда исследователь применяет для отклонения гипотезы $P < 0,01$ вместо $P < 0,05$. Правильная формулировка вопроса при односторонней оценке состоит в следующем: H_0 будет приемлема при всех выборочных результатах, исключая те из них, которые значительно выше μ_D . Такого же рода положение относится и к случаю альтернативы $\mu_D < \mu_0$.

Пример 41. В одном исследовании [2] влияния на рост свиной витамин B_{12} при норме 10 мг на 1 фунт кормов было образовано 8 парных групп, по 6 животных одного и того же помета. Эти пары различались между собой тем, что получали различные дозы ауреомицина — антибиотика, который не взаимодействует с витамином B_{12} ; поэтому различия в отношении витамина не зависят от ауреомицина. Средние приросты за сутки (к начальному уровню около 200 фунтов живого веса) обобщены в виде следующей таблицы.

Рацион	Парные группы							
	1	2	3	4	5	6	7	8
с B_{12}	1,60	1,68	1,75	1,64	1,75	1,79	1,78	1,77
без B_{12}	1,56	1,52	1,52	1,49	1,59	1,56	1,60	1,56
Разность D	0,04	0,16	0,23	0,15	0,16	0,23	0,18	0,21

Определить по этим разностям $d=0,170$ фунта в день и $s_d=0,0217$ фунта в день.

Пример 42. Известно, что добавление небольших количеств витамина B_{12} не может вызвать уменьшение скорости роста. Хотя в данном случае совершенно очевидно, что d будет существенно отклоняться от нуля, так как все разности положительны и, при одном исключении, достаточно устойчивы (см. параграф 8 главы 5), все же может представлять интерес значение величины t . (Ответ: 7,83, что лежит много ниже уровня 0,01.) Это относится к случаю альтернативной гипотезы $\mu > 0$.

Пример 43. Влияние витамина B_{12} состоит в стимулировании обмена веществ в связи с улучшением аппетита. Свиньи съедали больше корма и росли быстрее. В этом опыте стоимость дополнительного количества кормов, включая и стоимость витамина, была равна стоимости привеса примерно 0,130 фунта в день. Проверить гипотезу, что чистый доход, полученный от применения витамина B_{12} , равен нулю. Ответ: $t=1,84$; $P=0,11$ (при двухсторонней альтернативе).

Пример 44. При производстве предметов питания и лекарств часто осуществляется контроль над определенными специфическими компонентами путем установления некоторых предельных их значений, например содержание жира в масле не должно быть ниже 80%. Было взято 25 проб масла производства одного маслобойного завода и получены такие показатели жирности: $\bar{x}=80,16\%$ и $s=0,316\%$. С точки зрения официального контроля гипотеза $\mu_0=80\%$ не будет отвергнута при любой жирности выше 80% и будет отклонена при более низком проценте. Если этот закон будет строго предписан, а пока этого нет, то официальный контроль будет проверять гипотезу $\mu=80\%$ при альтернативе $H_A: \mu < 80\%$. Вычислить t и определить вероятность больших, чем это t , значений при $\mu=80\%$. Ответ: $P=0,01$.

14. Оценка интервала и критерии для σ^2 . Выборки из нормальной совокупности позволяют построить оценки интервалов для σ^2 [11]. Например, 95 %-ный доверительный интервал может быть приближенно определен, исходя из соотношения

$$\frac{\Sigma x^2}{\chi_{0,025}^2} \leq \sigma^2 \leq \frac{\Sigma x^2}{\chi_{0,975}^2}.$$

Для иллюстрации возьмем данные о содержании витамина С из параграфа 2 этой главы. Из таблицы 9 берем $\Sigma \bar{x}^2=254$; число степеней свободы=16, в то же время по таблице 6 находим $\chi_{0,975}^2=6,91$ и $\chi_{0,025}^2=28,8$. После подстановки этих значений получаем:

$$\frac{254}{28,8} \leq \sigma^2 \leq \frac{254}{6,91},$$

или

$$8,82 \leq \sigma^2 \leq 36,76.$$

Это и будет доверительный интервал для σ^2 . Отсюда следует, что имеется только 1 шанс из 20 против того, что σ находится между 2,97 и 6,06 мг/100 г. Заметим, что $s=3,98$ не находится в середине этого интервала; распределение s не является симметричным (см. параграф 5 главы 3).

Нулевая гипотеза о том, что дисперсия нормальной совокупности имеет некоторое определенное значение σ_0^2 , может быть проверена на основе случайной выборки объема n , взятой из этой совокупности. Альтернативной гипотезой может быть как $\sigma^2 \neq \sigma_0^2$, так и $\sigma^2 > \sigma_0^2$ (или $\sigma^2 < \sigma_0^2$). Для этого следует просто вычислить

$$\chi^2 = \frac{\Sigma x^2}{\sigma_0^2}$$

и сравнить этот результат с табличным значением χ^2 при $n-1$ степенях свободы. Для иллюстрации возьмем такой пример. В отношении опытной группы нормальных и здоровых крыс-самцов (в возрасте от 56 до 84 дней) сделано предположение, что стандартное отклонение веса равно $\sigma_0=26$ г. Предполагается, что введение нового рациона может увеличить стандартное отклонение, но ни в коем случае не уменьшить его. Опыт с 20 крысами дал $\Sigma x^2=23000$. Отсюда

$$\chi^2 = \frac{23000}{(26)^2} = 34,02; \text{ число степеней свободы} = 19.$$

Из таблицы 6 видно, что $\chi^2 > \chi_{0,05}^2$, и, следовательно, нулевая гипотеза $\sigma_0^2=676$ отвергнута.

Если же альтернатива H_A будет $\sigma^2 \neq \sigma_0^2$, то область, в которой нулевая гипотеза отвергается, будет $\chi^2 < \chi_{0,975}^2$ и $\chi^2 > \chi_{0,025}^2$. Для $H_A: \sigma^2 < \sigma_0^2$ гипотеза отвергается, если $\chi^2 < \chi_{0,975}^2$.

Случай, когда возникает необходимость проверки гипотезы о том, что два выборочных средних квадрата относятся к общей дисперсии σ^2 , будет рассмотрен в параграфе 8 главы 4.

15. Размер выборки. Установление числа повторных наблюдений также является вопросом, требующим обсуждения. Эксперимент со слишком малым числом наблюдений не сможет обнаружить интересующие нас различия; неоправданно большой опыт приведет к ненужным затратам времени и средств.

Для решения этого вопроса необходимы сведения двойного рода. Во-первых, необходимо знать величину σ той совокупности, откуда берется выборка. Эти сведения обычно берутся из данных предыдущих опытов или на основе некоторого представления о размахе случайного варьирования. Во-вторых, необходимо примерно определить тот наибольший доверительный интервал, который позволял бы уловить нормальные или даже несколько меньшие из средних разностей, представляющих практический интерес.

Проблема размера выборки в целом представляет известные трудности, и поэтому мы сначала рассмотрим более простой вопрос. Допустим, что предварительно фиксировано число повторных наблюдений, например оно не может превышать n парных наблюдений. При этих условиях экспериментатору необходимо узнать, сколь малы будут те разности, которые удастся обнаружить и признать существенными, положим, при 5%-ном уровне. Необходимая для этого формула такова:

$$\delta = \frac{s_D t_{0,05}}{\sqrt{n}},$$

где δ — искомая разность, s_D — выборочное стандартное отклонение разности и n — предполагаемое число повторных наблюдений.

Для иллюстрации возьмем данные примера 30. Там было определено, что $s_D = 7,72$ единицы при выборке $n = 10$ пар. Так как число степеней свободы = 9, то $t_{0,05} = 2,262$, отсюда:

$$\delta = \frac{7,72 \times 2,262}{\sqrt{10}} = 5,52 \text{ единицы.}$$

Если действительная разность будет равна 5,52 единицы, то можно ожидать, что соответствующая разность, наблюдаемая в опыте, будет существенной при 5%-ном уровне. Другими словами, этот опыт достаточен для того, чтобы в нем были обнаружены разности в 5,52 единицы или еще большие, если они существуют.

Это положение требует уточнения. Не может быть никакой гарантии в том, что эксперимент окажется точно таким, каким он ожидается. Приведенная выше формула дает только примерную наметку. Более точное выражение рассматриваемого здесь положения таково: если действительная разность равна 5,52 единицы, то вероятность того, что эта разность будет признана существенной на уровне 5%, будет равна 0,5.

Если понадобится повышение уровня уверенности или если экспериментатор пожелает обнаруживать более мелкие существенные различия, то для этого придется увеличить размер выборки (параграф 18 главы 10).

Как ранее указывалось, одна из пар наблюдений этого опыта вызывает сомнения. Если бы была доказана ошибочность этой из ряда вон выходящей пары (44 и 63) и на этом основании она была бы исключена, то получились бы такие показатели: $n = 9$, $\bar{d} = 4,3$, $s_D = 2,40$ единицы и $t_{0,05} = 2,306$. При этих условиях разность, которая будет обнаружена в опыте при соотношении шансов 50 : 50, равна

$$\delta = \frac{2,40 \times 2,306}{\sqrt{9}} = 1,84 \text{ единицы.}$$

Эта разность составляет только $1,84/57,3=3,2\%$ от средней всех 18 наблюдений, что указывает на большую чувствительность этого эксперимента по сравнению с первоначальным, в котором могли быть обнаружены разности, не меньшие, чем $5,52/56,9=9,7\%$.

Вопрос о том, каким должно быть δ , чтобы эксперимент считать удачным, сводится к вопросу: «Какой должна быть повторность опыта, чтобы обнаружить имеющую практическое значение разность?» Например, в рассматриваемом биологическом опыте можно считать достаточным, чтобы была обнаружена действительная разность, составляющая 2% от средней, т. е. $0,02 \times 57,3 = 1,15$ единицы. Для ответа на поставленный вопрос мы воспользуемся приведенной выше формулой, но несколько иначе, чем раньше. Это уравнение нельзя непосредственно решить относительно n , так как $t_{0,05}$ зависит от n . Поэтому приходится производить подбор n до тех пор, пока δ не получит заданного значения. Сначала ориентировочно возьмем $n=16$; в этом случае число степеней свободы $=15$, $t_{0,05}=2,131$ и при $s_D=2,40$ получим

$$\delta = \frac{2,40 \times 2,131}{\sqrt{16}} = 1,28.$$

Так как это число больше заданного, 1,15 единицы, то попробуем увеличить n , положим, до $n=20$. В этом случае

$$\delta = \frac{2,40 \times 2,093}{\sqrt{20}} = 1,12.$$

Теперь получено δ несколько меньше, чем нужно, поэтому попробуем уменьшить n . Можно найти, что $n=19$ дает почти точный результат. Этот подбор n упрощается, если учесть, что при не слишком малых выборках t приблизительно равно 2.

В данном случае опять существует соотношение шансов 1 : 1 в пользу того, что в таком опыте, состоящем из 19 парных наблюдений, будет обнаружена на 5%-ном уровне существенная разность, равная 1,15. В тех случаях, когда необходимо увеличить уровень уверенности при данном δ , то следует применить метод параграфа 18 главы 10. Так, можно установить, что если взять соотношение шансов 3 : 1 (т. е. $P=0,75$), то потребуется уже не 19, а 28 наблюдений.

Если такой же вопрос возникает в отношении доверительного интервала, то применяется та же формула для δ , но только теперь δ представляет собой половину ширины интервала, т. е. величину, которую надо прибавить или отнять от \bar{d} .

Часто величину δ для удобства выражают в процентах от средней. Соответствующая формула такова:

$$\frac{\delta}{\bar{d}} = \frac{C_D \cdot t_{0,05}}{\sqrt{n}},$$

где C_D — коэффициент вариации этой разности (параграф 6 этой главы).

Пример 45. Если взять данные примера 30, где $s_D=7,72$ и задаться $\delta=1,15$ единицы при 5%-ном уровне и при вероятности 0,5, то каким должен быть размер выборки? *Ответ:* 176 пар.

Пример 46. Исключите сомнительную пару (44, 63) и после этого определите размер выборки так, чтобы можно было обнаружить $\delta=1,15$ единицы при 1%-ном уровне и при вероятности 0,5. *Ответ:* 33 пары.

Пример 47. По таблице 9, где приведены данные о содержании витамина С в томатном соке, $s=3,98$ мг/100 г, так что $s_D=3,98 \times \sqrt{2}=5,63$ (параграф 9 главы 3). Дайте план опыта, в котором можно было бы обнаружить потерю витамина С в томатном соке при варке. Содержание витамина в каждой банке будет определяться перед и после варки. Желательно обнаружить при 5%-ном уровне такую потерю витамина С, которая составляет не меньше 10% от средней, т. е. не меньше 2 мг/100 г. Определите n . *Ответ:* 33 банки.

Пример 48. По данным таблицы 13, относящимся к случайной выборке веса морских свинок при рождении, имеем $\bar{x}=30,25$ г и $s=3,60$ г. Требуется взять выборку, в которой оценка μ производилась бы в пределах 3%. Возьмем $\delta=0,03 \times 30,25=0,91$ г и вычислите n . *Ответ:* 63. Подсказываю: берите ε , но не s_D .

16. Относительная изменчивость. Коэффициент вариации. В параграфе 6 коэффициент вариации C был определен как отношение s/x . Как указывалось, применение его основано на том факте, что во многих случаях ряды средних и стандартных отклонений имеют более или менее согласованное изменение. Это, в частности, иллюстрируется графиком на рисунке 6, где представлено изменение среднего роста и соответствующего стандартного отклонения у девочек от 1- до 18-летнего возраста. До 12-летнего возраста стандартное отклонение возрастает несколько в более быстром темпе, чем средний рост, вследствие чего коэффициент вариации тоже растет, но к 17-летнему возрасту C возвращается к своему начальному значению. Без сколько-

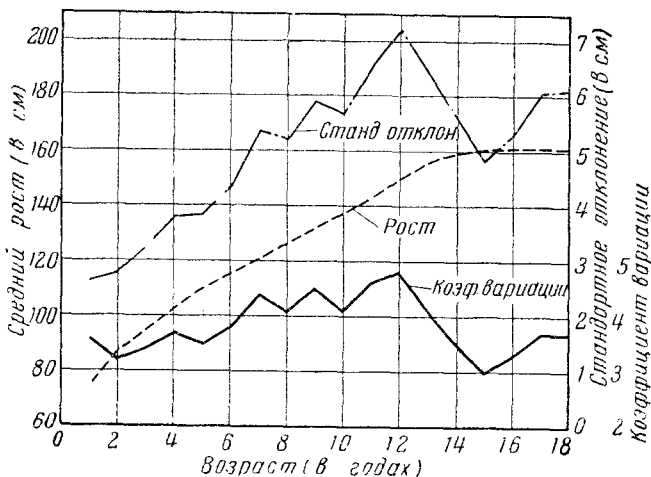


Рис. 6. График 3 рядов: роста, стандартного отклонения и коэффициента вариации для роста у девочек от 1 до 18 лет. См. список литературы [1].

нибудь серьезной ошибки можно считать, что относительное значение стандартного отклонения примерно $C=3,75\%$. Вернее, коэффициент вариации постепенно возрастает в период от младенчества до момента половой зрелости, далее в период формирования он резко падает, после чего уже устанавливается его постоянный уровень — около $3,75\%$.

При планировании опытов и обработке данных предварительное знание относительной изменчивости материала представляет известную ценность. Так, в предыдущем параграфе был приведен пример использования C при определении размера выборки. При обобщении этих показателей для определенного вида опытов появляется возможность на основе C осуществлять известный контроль за качеством опытов. Например, в опытах по сортоиспытанию зерновых культур при большом варьировании в зависимости от местных и метеорологических условий как среднего урожая, так и стандартного отклонения коэффициент вариации все же обладает известной устойчивостью и очень часто находится между 5 и 15%. Если исследователь получит значение C , выходящее из этого интервала, то он должен проверить, не является ли это результатом ошибки в вычислениях или наличия некоторых необычных обстоятельств, снижающих ценность эксперимента. Подобным же образом каждый исследователь, пользующийся выборочным методом, может установить значение C , ожидаемое в условиях проводимых им наблюдений, и взять под подозрение некоторые слишком большие отклонения.

Коэффициент вариации имеет и другое достаточно частое, но уже второстепенное применение. Так как коэффициент вариации C есть отношение двух средних, измеренных в одних и тех же единицах, то он является отвлеченным числом. Так, он останется тем же независимо от того, в чем будет измерен рост — в дюймах, футах или сантиметрах. На том же основании

коэффициент вариации для урожая сена сравним с коэффициентом вариации для урожая зерна.

Подопытные животные также характеризуются особыми коэффициентами вариации, и поэтому появляется возможность их сравнения, несмотря на разнообразие объектов измерения. Например, живой вес белых крыс-самцов в возрасте от 90 до 243 дней имеет C , равное в среднем около 14% [5], а годовая яйценоскость курицы породы полосатый плимутрок имеет C около 32% [12]. Такого рода сведения представляют большую ценность при планировании опытов.

Коэффициент вариации, как и многие другие относительные показатели, столь удобен для практики, что некоторые исследователи, пользуясь им в широких масштабах, неоправданно абстрагируются от информации, которая содержится в первоначальных данных, положенных в основу этого показателя. Нетрудно представить себе, как мы ограничим истолкование коэффициента вариации для роста девочек, если оно будет проводиться в отрыве от \bar{x} и s . Мы не сможем установить, происходит ли изменение C под влиянием увеличения s или уменьшения \bar{x} и не возникает ли зубчатая форма кривой C в результате нерегулярного изменения одного или обоих показателей. Коэффициент вариации является полезным и ценным показателем только при наличии \bar{x} и s , но, абстрагируясь от этих последних, мы рискуем иногда впасть в заблуждение.

Пример 49. При проведении опытов по определению содержания хлорофилла в листьях ананаса [17] возник вопрос о методе, позволяющем получить достаточно точные результаты. Было испробовано три способа определения, каждый на 12 листьях растения, после чего вычислены приведенные ниже показатели. На основании коэффициентов вариации было установлено, что все три метода одинаково точны и что выбор из них того, который технически наиболее удобен, может быть произведен без всякого ущерба для точности.

Статистические показатели при определении содержания хлорофилла на 12 листьях ананаса тремя способами

Показатели	Определение содержания хлорофилла на		
	100 г зеленой массы	100 г сухой массы	100 кв. см поверхности листьев
Выборочная средняя (в мг)	61,4	337	13,71
Выборочное стандартное отклонение (в мг)	5,22	31,2	1,20
Коэффициент вариации (в %)	8,5	9,3	8,8

Пример 50. В тщательно проводимых опытах с молодыми свиньями стандартное отклонение привеса животных составляет 10% от средней. Если известно, что группа свиней за некоторый промежуток времени дала привес 150 фунтов, можно ли ожидать, что стандартное отклонение привеса будет около 15 фунтов? Будет ли правильным считать, что при среднем суточном привесе в 1,5 фунта соответствующее выборочное стандартное отклонение будет примерно 0,15 фунта в день?

Пример 51. В одной лаборатории имеется группа подопытных крыс; коэффициент вариации для веса самцов в возрасте от 56 до 84 дней составляет около 13%. Определите стандартное отклонение веса этих крыс, если средний вес их равен 200 г. *Ответ:* 26 г.

Пример 52. Допустим, что коэффициент вариации урожая пшеницы в полевых опытах обычно составляет около 5%. Будете ли вы удивлены тем, что в некотором опыте при среднем урожае в 25 бушелей на 1 акр стандартное отклонение поделяночных урожаев составляет 0,5 бушеля на 1 акр?

ЛИТЕРАТУРА

1. B o u n t o n Bernice, University of Iowa Studies in Child Welfare, Vol. 12, No. 4, 1936.
2. C a t r o n D. V., M a d d o c k H. M., S p e e r V. C., V o h s R. L., Antibiotics and Chemotherapy, 1, 31, 1951.
3. Council on Foods, Journal of the American Medical Association, 110, 651, 1938.
4. C r a m p t o n Earl Wilcox, Journal of Nutrition, 7, 305, 1934.
5. D o n a l d s o n Henry H., The Rat. The Wistar Institute of Anatomy and Biology, No. 6. Philadelphia, 1924.

6. Fisher R. A., Proceedings of the International Mathematical Congress, Toronto, 805, 1924.
7. Gray Horace, Ayres J. C., Growth in Private School Children. The University of Chicago Press, 1931.
8. Grove L. C., Iowa Agricultural Experiment Station Bulletin 253, 1939.
9. Love H. H., Journal of the American Society of Agronomy, 16, 64, 1924.
10. Mitchell H. H., Burroughs Wise, Beadles Jesse R., Journal of Nutrition, 11, 257, 1936.
11. Mood Alexander McFarlane, Introduction to the Theory of Statistics. McGraw-Hill Book Company, Inc., New York, 1950.
12. Pearl Raymond, Surface F. M., U.S.D.A., Bureau of Animal Industry Bulletin No. 110, Part I, 1909.
13. Pearson E. S., Biometrika, 24, 416, 1932.
14. Reddy Chas. S., Iowa State College Journal of Science, 9, 530, 1935.
15. «Student», Biometrika, 6, 1, 1908.
16. Talley Paul J., Plant Physiology, 9, 737, 1934.
17. Tam R. K., Magistad O. C., Plant Physiology, 10, 161, 1935.
18. Tippett L. H. C., Biometrika, 17, 386, 1925.
19. Youden W. J., Beale Helen Purdy, Contribution from Boyce Thompson Institute, 6, 437, 1934.

Глава 3

ВЫБОРОЧНЫЕ НАБЛЮДЕНИЯ ИЗ НОРМАЛЬНОЙ СОВОКУПНОСТИ. ПОСТРОЕНИЕ ВЫБОРОЧНЫХ РАСПРЕДЕЛЕНИЙ

1. Введение. Положения, относящиеся к доверительным интервалам в главе 1, были подвергнуты проверке при помощи выборочных наблюдений; этим же путем было установлено распределение хи-квадрат, определяющее вероятности появления тех или иных случаев. Но приведенные в главе 2 величины s_x и t были приняты пока без такой экспериментальной проверки. Теперь мы обратимся к изучению опытным путем распределений этих величин, а также распределений \bar{x} , s^2 и s , что позволит нам лучше понять содержание установленных ранее положений, относящихся к этим величинам.

Для изучения выборочных свойств статистических показателей можно применить два способа. Первый является методом, применяемым математиком, который приходит к своим результатам при помощи логики, облеченной в символы. Эти результаты он дает в виде формул и таблиц, что придает статистическим методам точность и делает удобным их применение. Второй способ состоит в том, что производятся фактические выборки. Он не требует математических знаний и в то же время дает возможность убедиться в правильности многих выводов теории вероятностей. Что особенно важно, такая опытная проверка выборочного метода дает возможность не специалисту-математику уяснить себе те основы, на которых строится математическая теория статистики.

Как математик, так и экспериментатор стремятся иметь дело с выборками, взятыми из фактических совокупностей, которые более или менее близки к некоторым математическим моделям. Например, можно взять данные об урожайности, положим, тысячи малых делянок какой-либо культуры [1,5,7]. В этом случае *территориальное распределение* переменной — урожайности — отождествляется с *распределением частот*. В то время как последнее распределение можно сопоставить с нормальным распределением, которое обычно считается источником выборок, территориальное распределение может служить основой для изучения эффективности различных схем полевого опыта.

В данном случае мы, как и прежде, возьмем некоторое конкретное распределение и при помощи таблицы случайных чисел произведем из него серию выборок. Но теперь вместо биномиального распределения мы должны взять нормальное распределение, описанное в параграфе 1 главы 2 и в главе 8. В этом распределении, взятом в качестве модели, число различных значений переменной безгранично, но практически деления шкалы переменной не могут быть более частыми, чем это возможно при наименьшей единице измерения. Поэтому совокупность, которую мы предполагаем взять в качестве имитации нормального распределения, отличается от этой модели в двух отношениях: она ограничена по своему объему и размаху вместо того, чтобы быть безграничной, и относится к *прерывной* переменной вместо того, чтобы относиться к *непрерывной* переменной, положенной в основу теории нор-

мального распределения. Влияние этих отклонений от нормы вряд ли будет заметным, так как оно слишком мало по сравнению с выборочной вариацией.

2. Конечная совокупность, имитирующая нормальную совокупность. В таблице 15 приведены данные о привесе 100 свиней; эти данные, взятые из опыта, несколько изменены с целью получить более близкую к нормальной форму распределения. Данные занумерованы от номера 00 до 99, что позволяет их отождествить с соответствующими числами, взятыми из таблицы случайных цифр. Наглядное представление об этом распределении можно

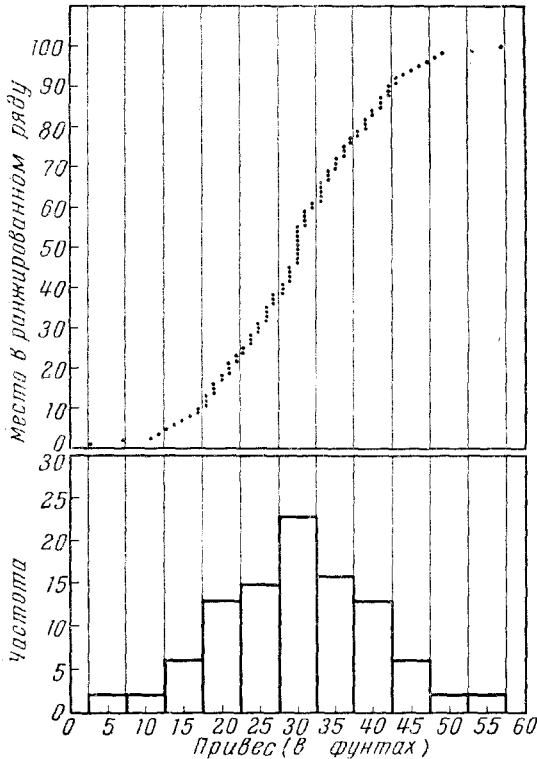


Рис. 7. Верхняя часть: графическое изображение ранжированного ряда 100 нормально распределенных привесов. Нижняя часть: гистограмма тех же привесов. Высота прямоугольника гистограммы пропорциональна числу точек в графике ранжированного ряда, которые заключены между вертикальными линиями.

получить из рисунка 7. Здесь данные, сгруппированные наиболее густо в середине ранжированного ряда, начинают постепенно и симметрично становиться более редкими, сначала медленно, а потом все быстрее и быстрее; в результате две трети всего числа наблюдений сосредоточивается в интервале 30 ± 10 фунтов, т. е. в интервале двух стандартных отклонений, центром которого является средняя. В действительной совокупности, состоящей из неопределенно большого числа объектов, несомненно, будут встречаться и большие отклонения от средней, но в данном случае это не имеет никакого значения.

Соответствие между графиком ранжированного ряда и гистограммой очевидно. После того как установлены границы между классами, остается подсчитать число точек между двумя параллельными линиями и построить прямоугольник, высота которого пропорциональна числу наблюдений. На горизонтальной шкале привесов в каждом интервале отмечено центральное, или срединное, значение класса.

В таблице 16 дано распределение численностей, представленное графически на рисунке 7, причем здесь в первой строке указаны только срединные значения классов; соответствующие же классовые интервалы будут: от 2,5 до 7,5, от 7,5 до 12,5 и т. д. Этот способ дает возможность представить в краткой форме данные больших выборок.

3. Случайные выборки из нормальной совокупности. Наиболее легким способом получения случайных выборок из таблицы привесов свиней является последовательный отбор чисел из таблицы случайных цифр 2, после чего эти числа с помощью номеров от 00 до 99 ассоциируются с соответствующими данными привесов таблицы 15. Чтобы избежать дублирования выборок, следует начинать не с начала таблицы случайных чисел, а с какого-нибудь случайно выбранного пункта и идти от него вверх, вниз или по диагонали. Допустим, что, применяя описанный в параграфе 5 первой главы способ, вы случайно наткнулись на цифру 8 в ряду 71 и колонке 29. Эта цифра и следующая за ней цифра 3 дают номер 83; в таблице 15 свинья под этим номером имеет привес 40 фунтов. Таким образом, первое наблюдение выборки дает

Ранжированный ряд привесов (в фунтах) 100 свиней, полученных за 20 дней
 Распределение привесов близко к нормальному распределению с $\mu=30$ фунтам
 и $\sigma=10$ фунтам

Номер наблюдения	Привес	Номер наблюдения	Привес	Номер наблюдения	Привес	Номер наблюдения	Привес
00	3	25	24	50	30	75	37
01	7	26	24	51	30	76	37
02	11	27	24	52	30	77	38
03	12	28	25	53	30	78	38
04	13	29	25	54	30	79	39
05	14	30	25	55	31	80	39
06	15	31	26	56	31	81	39
07	16	32	26	57	31	82	40
08	17	33	26	58	31	83	40
09	17	34	26	59	32	84	41
10	18	35	27	60	32	85	41
11	18	36	27	61	33	86	41
12	18	37	27	62	33	87	42
13	19	38	28	63	33	88	42
14	19	39	28	64	33	89	42
15	19	40	28	65	33	90	43
16	20	41	29	66	34	91	43
17	20	42	29	67	34	92	44
18	21	43	29	68	34	93	45
19	21	44	29	69	35	94	46
20	21	45	30	70	35	95	47
21	22	46	30	71	35	96	48
22	22	47	30	72	36	97	49
23	23	48	30	73	36	98	53
24	23	49	30	74	36	99	57

ТАБЛИЦА 16

Распределение численностей для привесов 100 свиней
 Ограниченная совокупность, близкая к нормальной

Серединные значения классов (фунты)	5	10	15	20	25	30	35	40	45	50	55
Численности	2	2	6	13	15	23	16	13	6	2	2

40 фунтов. Двигаясь вверх по колонке случайных чисел, вы находите последовательно 09, 75, 90 и т. д. и выписываете соответствующие привесы из таблицы 15: 17, 37, 43 и т. д. фунтов. Продолжая этот процесс, вы можете получить столько наблюдений и столько выборок, сколько было намечено.

Остановимся на выборке в 10 наблюдений. В противоположность тому, что делалось в главе 1, теперь все выборки должны быть одинакового размера, так как распределения их характеристик зависят от n . Рекомендуется запись наблюдений вести столбиком, оставляя 6 нижних строк для записи результатов последующих расчетов. Для примера в таблице 17 приведены 4 выборки. Расчетные результаты, показанные внизу таблицы, будут объяснены в дальнейшем. Возьмите столько выборок, сколько вы можете обработать за время, имеющееся в вашем распоряжении. Хорошо, если несколько человек работают вместе, так как каждый из них может использовать весь обобщенный материал. Записи следует бережно сохранить, так как вы ими в дальнейшем будете пользоваться еще не раз.

Можно отметить, что каждый из наблюдаемых привесов свиньи может появиться в выборке столько раз, сколько раз соответствующий номер

попадает в таблице случайных цифр, т. е. это наблюдение не выбывает из совокупности после первого его появления. Таким образом, выборочные наблюдения все время берутся из одной и той же совокупности, и вероятность появления какого-либо частного значения остается постоянной в течение всего процесса.

ТАБЛИЦА 17

Четыре выборки по 10 наблюдений, взятых случайным порядком из данных о привесе свиней в таблице 15. О показателях, помещенных внизу, будет сказано в параграфах 4—8 главы 3

Номер наблюдения и формулы	Номер выборки			
	1	2	3	4
1	33	32	39	17
2	53	31	34	22
3	34	11	33	20
4	29	30	33	19
5	39	19	33	3
6	57	24	39	21
7	12	53	36	3
8	24	44	32	25
9	39	19	32	40
10	36	30	30	21
\bar{x}	35,60	29,30	34,10	19,40
s^2	169,80	151,60	9,00	112,30
s	13,00	12,31	3,00	10,60
$s_{\bar{x}}$	4,11	3,89	0,95	3,35
t	1,36	-0,18	4,32	-3,25
$t_{0,05} s_{\bar{x}}$	9,30	8,80	2,20	7,60
$\bar{x} \pm t_{0,05} s_{\bar{x}}$	26,3—44,9	20,5—38,1	31,9—36,3	11,5—26,7

Пример 1. Определите размах варьирования в каждой из ваших выборок объема $n=10$. Средняя из этих размахов оценивает величину $\sigma/0,325$ (табл. 10), т. е. $10/0,325=30,8$. Сколь близкой к этой величине оказалась ваша оценка?

4. Распределение выборочной средней. Суммируя данные по каждой выборке и деля эти суммы на 10, выпишите внизу выборочные средние \bar{x} . Хотя каждая из этих средних является оценкой величины $\mu=30$ фунтам, все же между ними наблюдается довольно большое различие. Расположите в ранжированный ряд ваши выборочные средние. Если этих средних у вас достаточно много, то образуйте из них распределение частот по образцу таблицы 18.

В нашей лаборатории получен ряд средних, колеблющихся от 19 до 39 фунтов, что, может быть, является весьма неожиданным для новичка в этом деле. Чтобы понять, к чему приводит такая колеблемость средних, представьте себе, что вы имеете один эксперимент, давший вместо точного значения средней (30 фунтов) — одну из наиболее уклоняющихся средних. Если, помимо тех сведений о совокупности, которые дает выборка, вы ничего не имеете иного, то вы неизбежно в этом случае впадете в заблуждение, причем нет никакой возможности избежать этого.

Одна из задач предложенного вам опыта по проведению выборочных наблюдений состоит в том, чтобы познакомить вас с тем риском, который сопутствует всем выводам, основывающимся на небольшой доле совокупности. Исследователь весьма редко знает точное значение параметров совокупности, из которой берутся выборки, и может довольствоваться только выборочными оценками этих параметров. Он должен научиться смотреть на свои экспериментальные данные в свете того факта, что они сопровождаются ошибками выборки. Поэтому его суждения должны учитывать не только факты, которые составляет его выборка, но и всю ту информацию о предмете, которая накоплена им и другими исследователями.

Однако более обнадеживающим является то обстоятельство, что имеется тенденция к получению достаточно большого числа выборочных средних, близких к центру распределения совокупности. Если бы этого не было, то выборочный метод не был бы столь популярным и полезным. Практически почти полная невозможность получить совсем неправильные оценки придает нам определенную уверенность, когда мы на основе выборки делаем те или иные выводы. И это очень хорошо. Опытный исследователь, найдя правильное равновесие между уверенностью и осторожностью, только в редких случаях станет жертвой капризов выборочного метода.

Рассматривая распределение выборочных средних, приведенное в таблице 18, следует обратить внимание на следующие основные его свойства. Во-первых, само это распределение близко к нормальному, в чем вы можете убедиться, сравнивая фактические и рассчитанные теоретически для нормального распределения частоты. Теория доказывает, что средние выборок, взятых из нормальной совокупности, имеют подобное же нормальное распределение. Более того, можно утверждать, что по мере увеличения размера выборки распределение выборочных средних стремится к нормальному даже в том случае, когда исходная совокупность не является нормальной; это имеет большое практическое значение, так как обычно точная форма совокупности, из которой берется выборка, остается неизвестной. Эта *центральная предельная теорема* «является наиболее важной как с теоретической, так и практической точки зрения теоремой статистики» [6].

Второе положение, которое следует отметить, состоит в том, что размах варьирования средних, каким бы большим он ни казался, все же составляет только около одной трети размаха основной совокупности в таблице 15. Как мы уже указывали, средние варьируют в меньшей степени, чем первоначальные данные. Обсуждение этого вопроса будет проведено в параграфе 6 главы 3.

ТАБЛИЦА 18

Распределение 511 средних, вычисленных по выборкам в 10 наблюдений, случайно отобранных из таблицы 15 (привес свиней). Для сравнения приведены теоретические частоты нормального распределения

Серединное значение класса (фунты)	Частоты	Теоретические частоты (способ вычисления см. примеры 22 и 23 главы 8)
19	1	0,20
20	1	0,41
21	0	1,18
22	7	2,71
23	5	5,62
24	10	10,78
25	19	18,60
26	30	29,02
27	41	41,14
28	48	52,07
29	66	61,63
30	72	64,23
31	56	61,63
32	46	52,07
33	45	41,14
34	22	29,02
35	24	18,60
36	12	10,78
37	5	5,62
38	0	2,71
39	1	1,84 ¹
Итого	511	511,0

¹ Включая численности всех последующих классов.

Наконец, третья характерная особенность этого распределения относится к его средней. Последняя может быть вычислена путем суммирования всех средних и делением результата на 511; в параграфе главы 8 будет рассмотрен более простой способ вычисления. Эта общая средняя, составляющая 29,87 фунта, является оценкой величины $\mu=30$ фунтам. Очевидно, имея в своем распоряжении такую большую выборку, как такая объединенная выборка, мы можем уже не бояться сильного влияния выборочной вариации.

5. Выборочные распределения s^2 и s . Эти оценки вычисляются так, как показано в параграфе 12 главы 2. Рассматривая таблицу 1, вы можете видеть, что три средних квадрата s^2 превышают $\sigma^2=100$, в то время как четвертый очень мал. Изучите более внимательно те из ваших выборок, у которых встречаются такие необычные s^2 , и установите, какие имеются в этих рядах особенности. Например, наша выделяющаяся из других выборка № 3 имеет размах варьирования, равный только $39-30=9$ фунтам, и не содержит в себе ни одного наблюдения, меньшего μ . Можно ли встретить выборку с размахом варьирования, равным размаху основной совокупности?

В таблице 19 дано распределение s^2 в наших 511 выборках. Следует отметить *скошенность* этого распределения, которая выражается в том, что ниже среднего значения s^2 размещены скученно, а выше средней они образуют длинный «хвост»; это отчасти напоминает распределение хи-квадрат главы 1, хотя данная особенность здесь менее выражена. Несмотря на это, средняя из этих средних квадратов 101,5 достаточно близка к дисперсии совокупности 100, что служит подтверждением того факта, что выборочный средний квадрат является несмещенной оценкой σ^2 .

Вы, возможно, чувствуете, что теперь мы подходим к вопросу о доверительных пределах и к проверке гипотез на основе распределения s или s^2 . Величина s^2 является мерой рассеяния, подобной хи-квадрат, и ей можно было бы пользоваться так же, как и этой последней, если бы не было установлено, что для этих целей более подходит величина t . Этот вопрос будет рассмотрен в параграфе 7 этой главы.

ТАБЛИЦА 19

Распределение 511 средних квадратов выборок объема $n=10$, взятых из нормальной совокупности таблицы 15

Серединное значение класса	0	20	40	60	80	100	120	140	160	180	200	220	240	260	280	300	320	340
Частота	1	11	47	92	93	72	73	42	29	26	11	8	2	1	0	1	1	1

Распределение s , т. е. квадрата корня из s^2 , представленное в таблице 20, вполне согласуется с теорией. Оно немного скошено (в меньшей степени, чем распределение s^2) и имеет очень небольшое смещение средней 9,8 фунта, что только немногим меньше $\sigma=10$ фунтам. Таким образом, даже при столь малой выборке, как $n=10$, выборочное стандартное отклонение имеет смещение, которым можно пренебречь. Но просмотрите снова параграф 14 главы 2.

ТАБЛИЦА 20

Распределение 511 выборочных стандартных отклонений, соответствующих средним квадратам таблицы 19

Серединное значение класса	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Частота	1	2	9	18	58	77	80	71	79	44	41	17	8	3	1	2

Пример 2. Теоретически 50% средних таблицы 18 должны быть меньше 30 фунтов. Полагая, что из всех средних, попавших в интервал, срединное значение которого 30, ровно половина имеет значения, меньше 30, находим, что общее число выборочных средних, меньших 30 фунтов, равно 264, а остальные 247 средних больше 30 фунтов. Вычислите хи-квадрат в предположении, что в совокупности эти численности относятся как 1:1. *Ответ:* 0,57.

Пример 3. Считая, что половина выборочных стандартных отклонений таблицы 20 должна быть меньше 10 фунтов, вычислите $\chi^2=4,89$. Это явно говорит против симметричности распределения.

Пример 4. В таблице 19 нашло известное отражение теоретическое распределение s^2 в случайных выборках объема n из нормальной совокупности, имеющей дисперсию σ^2 . Теоретическое распределение имеет среднюю $\mu_{s^2}=\sigma^2$ и дисперсию $2\sigma^4/(n-1)$. Соответственно для фактического распределения средняя из 511 значений s^2 , равная $\bar{s}^2=101,5$, оценивает $\mu_{s^2}=100$, а средний квадрат для 511 значений s^2 дает оценку величины $2 \times 100^2/9=2222$. Этот средний квадрат, вычисленный по данным таблицы 68, равен 2424.

Пример 5. На основе s^2 можно произвести оценку величины $2\sigma^4/(n-1)$. Эта оценка будет $2s^4/(n+1)$. Примеры: выборка № 1 в таблице 17 имеет $s^2=169,8$ и дает оценку для величины 2222, равную $2 \times 169,8^2/11=5,242$, но в выборке № 4, которая ближе к середине, оценкой величины 2222 будет $2 \times 112,3^2/11=2,293$.

6. Оценка стандартной ошибки $\sigma\sqrt{n}$. Эта оценка является наиболее важной в нашем статистическом инструментарии; она вычисляется по формуле:

$$s_{\bar{x}} = s/\sqrt{n}, \text{ или } s_{\bar{x}} = \sqrt{s^2/n},$$

что более удобно. Значение $s_{\bar{x}}$ каждой из наших выборок объема 10 является оценкой:

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{10} = 3,162 \text{ фунта.}$$

В ваших выборках вы встретите обычную выборочную вариацию, но в среднем ваши выборочные стандартные ошибки должны дать примерно 3,162 фунта. Величина $s_{\bar{x}}$, подобно величине s , имеет небольшое смещение, поэтому средняя из $s_{\bar{x}}$, если она получена простым суммированием выборочных оценок $s_{\bar{x}}$, будет все еще более или менее смещенной. Для получения более точной оценки следует сложить ваши несмещенные средние квадраты

$$s_{\bar{x}}^2 = s^2/n$$

и уже после этого извлечь квадратный корень из их средней. Например, для наших 511 выборок эта средняя легко находится путем вычисления средней из средних квадратов 101,5, деления на 10 и извлечения корня $\sqrt{10,15} = 3,186$ фунта. Это весьма хорошее приближение к значению показателя в совокупности служит свидетельством того, что $s_{\bar{x}}$ является оценкой $\sigma_{\bar{x}}$. Но спрашивается, какова природа величины $\sigma_{\bar{x}}$, которая оценивается величиной $s_{\bar{x}}$? Как было показано в параграфе 7 главы 2, $\sigma_{\bar{x}}$ является стандартным отклонением средних, вычисленных по выборкам объема n , взятым случайным образом из нормальной совокупности со стандартным отклонением σ . Подтверждение этого положения можно найти в распределении 511 выборочных средних, которое дано в таблице 18. Выборочное стандартное отклонение этого распределения равно 3,178 фунта (в параграфе 2 главы 8 будет дан способ расчета, более простой, чем тот, который приведен в таблице 9), что является достаточно хорошим приближением к теоретическому значению 3,162 фунта. Если теперь суммировать все сказанное, то вы придете к такому положению, что по данным каждой случайной выборки можно определить $s_{\bar{x}}$, являющуюся оценкой стандартного отклонения средних или стандартной ошибки $\sigma_{\bar{x}}$. Обычно из совокупности берется только одна выборка. Изучая свои малые выборки, вы увидите, что оценка такой единичной выборки может иногда довольно значительно отклоняться от средней, но вместе с тем мы теперь установили, что эта оценка имеет только очень небольшое смещение даже при столь малой выборке, как выборка в 10 наблюдений.

Пример 6. Начало цветения у 400 растений одного из гибридов ржи [3] наблюдалось через 70,5 дня после посева; стандартное отклонение этого показателя составляло

6,9 дня. Определите стандартное отклонение средних в таких выборках. *Ответ:* $s_{\bar{x}}=0,34$ дня.

Пример 7. Какой параметр совокупности будет оценивать величина $s_{\bar{x}}$, если из данных о привесе свиней таблицы 15 взята выборка в 40 наблюдений? *Ответ:* 1,58 фунта. Это положение вы можете легко проверить, производя соединение 4 ваших выборок по 10 и поступая в дальнейшем так, как вы делали в отношении ваших первоначальных выборок.

Пример 8. Имея 10 или более выборок по 10 наблюдений, вы легко можете образовать выборку в 100 наблюдений. Для быстрого вычисления Σx^2 такой объединенной выборки в 100 наблюдений вы можете применить способ параграфа 12 главы 2, соединяя 10 значений ΣX и ΣX^2 . Какой параметр будет оценивать величина $s_{\bar{x}}$, вычисленная для этой объединенной выборки? *Ответ:* 1 фунт.

Пример 9. В печати было опубликовано, что в одной группе свиней с ежедневным привесом в 1,4 фунта получено $s_{\bar{x}}=0,02$ фунта в день. Автор публикации не дал сведений о числе свиней в группе. Если вы желаете получить представление о количестве свиней в группе, то для этого вы можете использовать соотношение $s_{\bar{x}}=s/\sqrt{n}$ и ваше полученное из опыта знание, что при откорме свиней наблюдается коэффициент вариации, равный 10%. Это последнее позволяет найти приближенное значение s , которое будет равно 10% от 1,4 фунта, т. е. 0,14 фунта в день. После этого: $\sqrt{n}=s/s_{\bar{x}}=0,14/0,02=7$ и, наконец, $n=49$ свиней. Значение коэффициента вариации в 10% при откорме свиней является уже сравнительно небольшим, но при помощи специальных эффективных методов экспериментирования можно добиться дальнейшего его снижения.

7. Распределение t . Вы уже имели возможность познакомиться с этим показателем в параграфах 7, 8 и 9 главы 2, но более подробное его изучение было отложено нами до настоящего времени. Формула этого показателя:

$$t = (\bar{x} - \mu) / s_{\bar{x}}$$

Для произведенных вами выборочных наблюдений величина μ известна, она равна 30 фунтам. Эта формула отличается от той, которая была дана в параграфе 9 главы 2, где, согласно гипотезе, было принято $\mu=0$ и где, следовательно, t было равно $\bar{x}/s_{\bar{x}}$, т. е. полностью определялось выборочными данными. Второе различие наблюдается с формулой параграфа 8 главы 2, где значение μ считалось неизвестным. Здесь доверительные пределы были определены как:

$$\bar{x} - t_{0,05} \cdot s_{\bar{x}} \quad \text{и} \quad \bar{x} + t_{0,05} \cdot s_{\bar{x}},$$

причем $t_{0,05}$ являлось положительным числом, взятым из соответствующей таблицы. Теперь мы должны более подробно изучить эту весьма важную величину t .

Так как значения \bar{x} и $s_{\bar{x}}$ уже определены для всех ваших выборок в 10 наблюдений, то можно найти выборочные значения t , подставляя в приведенную ранее формулу $\mu=30$, что даст

$$t = (\bar{x} - 30) / s_{\bar{x}}$$

В этом случае значение t может быть как положительным, так и отрицательным, в зависимости от того, будет ли \bar{x} больше или меньше 30 фунтов. При данных условиях появление в выборке этих двух знаков одинаково вероятно, и поэтому можно ожидать появления того и другого знака в половине всех выборок. При такой симметричности распределения вы должны получить среднюю из всех значений t , близкую к нулю.

Выборки таблицы 17 были подобраны так, чтобы можно было видеть, при каких условиях в процессе выборочного наблюдения возникают большие, малые и промежуточные значения t . Небольшое отклонение $(\bar{x} - \mu)$ и (или) большое значение выборочной стандартной ошибки приводит к тому, что t становится малым. В этой таблице даны некоторые крайние комбинации этих величин; вы, несомненно, найдете подобные случаи и среди своих выборок.

В таблице 21 представлено распределение выборочных t , полученное в нашей лаборатории. Здесь, как и в таблице 5, были взяты неравные интервалы классов с той целью, чтобы представить распределение в виде ряда легко улаживаемых глазом и удобных для расчета вероятностей. Для сравнения вместе с выборочным распределением в таблице приведены теоретические частоты, выраженные в процентах. В двух последних колонках таблицы приведены накопленные частоты в процентах, что дает возможность использовать таблицу для построения доверительных утверждений и для проверки гипотез. Из данных этой таблицы следует, что 2,5% всех значений, полученных в выборках по 10 наблюдений, будет превосходить 2,262 и в то же время другие 2,5% значений будут меньше — 2,262. Объединяя оба крыла распределения, как это сделано в последней колонке, мы получим 5% значений t , отдаленных от центра распределения больше чем на $|2,262|$; это последнее значение называется поэтому 5%-ным уровнем. Заметим, что этот 5%-ный уровень лежит выше 5%-ного уровня каждого отдельного крыла распределения ($\pm 1,833$). Применение таких уровней будет рассмотрено в дальнейшем. А пока постройте распределение ваших выборочных значений t и сравните его с приведенным в нашей таблице распределением.

ТАБЛИЦА 21

*Выборочное и теоретическое распределение t . Выборки по 10 наблюдений.
9 степеней свободы.*

Интервалы t		Выборочное распределение		Теоретическое распределение		
от	до	частоты	частоты (в %)	частоты (в %)	накопленные частоты	
					в одном направлении	в двух направлениях
—	—3,250	3	0,6	0,5	100,0	
—3,250	—2,821	4	0,8	0,5	99,5	
—2,821	—2,262	5	1,0	1,5	99,0	
—2,262	—1,833	16	3,1	2,5	97,5	
—1,833	—1,383	31	6,1	5,0	95,0	
—1,383	—1,100	25	4,9	5,0	90,0	
—1,100	—0,703	52	10,2	10,0	85,0	
—0,703	0,0	132	25,8	25,0	75,0	
0,0	0,703	126	24,6	25,0	50,0	100,0
0,703	1,100	41	8,0	10,0	25,0	50,0
1,100	1,383	32	6,3	5,0	15,0	30,0
1,383	1,833	18	3,5	5,0	10,0	20,0
1,833	2,262	13	2,5	2,5	5,0	10,0
2,262	2,821	8	1,6	1,5	2,5	5,0
2,821	3,250	2	0,4	0,5	1,0	2,0
3,250	—	3	0,6	0,5	0,5	1,0
		511	100,0	100,0		

В таблице 11 были даны распределения t в виде накопленных частот для выборок различного размера. В связи с симметричностью этой таблицы даны только положительные значения t , т. е. взята только часть, соответствующая нижней половине таблицы 21. В верхней строке таблицы приведены накопленные вероятности для обеих сторон распределения, а в левой колонке — число степеней свободы $n-1$. Пограничные значения классовых интервалов таблицы 21 (0,703; 1,383; 1,833 и т. д.) в таблице 11 можно найти в строке для 9 степеней свободы.

8. Интервальная оценка μ . Доверительный интервал. Теперь при помощи ваших выборок можно произвести проверку теории доверительного интервала. Каждая выборка дает интервал $\bar{x} \pm t_{0,05} \cdot s_{\bar{x}}$, предназначенный для покрытия μ . Для его определения подставьте в формулу оценки каждой выборки \bar{x}

и s_x и значение $t_{0,05}=2,262$, соответствующее уровню 0,05 и 9 степеням свободы. В результате вы получите данные, подобные тем, которые приведены в последней строке таблицы 17. Если вы теперь на основе данных какой-либо частной выборки будете утверждать, что такой интервал заключает в себе μ , то вы будете или правы, или неправы; в данном случае легко установить, случится ли то или другое, так как вам известно, что $\mu=30$ фунтам. Теория будет подтверждена, если около 95% таких ваших утверждений будут правильными и около 5% — неправильными.

Возвращаясь опять к таблице 17, можно видеть, что выборка № 1 подтверждает то положение, что средняя совокупности находится между 26,3 и 44,9 фунта, так как в данном случае нам известно, что этот интервал действительно содержит μ . Наоборот, выборки № 3 и 4 подобраны так, чтобы показать случаи, которые приводят к неправильным утверждениям; в первом случае это связано с необычно большим отклонением выборочной средней, в другом же — со слишком малым значением выборочного стандартного отклонения. Из 511 полученных у нас в лаборатории выборок 486 привели к правильным утверждениям, что составляет 95,1% всех случаев. Процент ошибочных утверждений — 4,9% весьма близок к теоретическим — 5%. Следует всегда иметь в виду оговорку, включенную в каждое доверительное утверждение при 5%-ном уровне, а именно: это утверждение правильно, если не учитывать 1 шанс из 20 в пользу противоположного.

Практическое применение этой теории лицами, имеющими дело с опытами и другими выборочными наблюдениями, проводится в условиях, когда параметры совокупности остаются неизвестными. Поэтому, когда они высказывают доверительные утверждения, то не знают, будут ли они в том или ином отдельном случае правы или неправы; они знают только выбранную ими доверительную вероятность. Например, возможно, хотя это и маловероятно, что первые пять взятые вами выборки одна вслед за другой приведут к неправильным утверждениям, и не по причине неискusstного проведения этой операции, а просто в силу неудачно сложившейся случайности. Но будем надеяться, что мудрый совет более опытного работника позволит вам избежать подобного рода ловушек выборочного метода.

Я думаю, что теперь вы имеете ясное представление о сущности 95% ной интервальной оценки и что вы сможете ею воспользоваться, разумно сочетая осторожность и доверие. В следующих параграфах мы рассмотрим вопрос о выводах другого характера, касающихся некоторых гипотез относительно совокупностей.

Пример 10. Имея частоты таблицы 21, проверьте гипотезу (установлено, что она соответствует действительности), что распределение t симметрично в том смысле, что половина всех частот больше, а половина меньше нуля. *Ответ:* $\chi^2=1,22$.

Пример 11. Из таблицы 21 видно, что $3+4+5+8+2+3=25$ выборок имеют $|t| > 2,262$. Проверьте гипотезу, что 5% значений t в совокупности больше, чем $|2,262|$. *Ответ:* $\chi^2=0,0124$.

Пример 12. Определите накопленные суммы частот обеих половин распределения таблицы 21 и, выразив их в процентах, сравните с данными последней колонки этой таблицы.

Пример 13. В течение осени 1943 г. примерно одна из 1000 городских семей штата Айова (по определению — городом считается населенный пункт, имеющий 2500 и более жителей) была опрошена относительно количества кварт употребленных ею консервированных продуктов. В среднем для 300 семей получено 165 кварт при стандартном отклонении 153 кварталы. Определите 95%-ные доверительные границы. *Ответ:* 165 ± 47 кварт.

Пример 14. По данным переписи 1940 г., в городах штата Айова было 312 000 отдельных хозяйств (условно говоря, семей). Используя показатели предыдущего примера, определите число кварт консервированных продуктов, потребленных в городах штата Айова в 1943 г. *Ответ:* 51 500 000 кварт; при 95%-ных доверительных границах 46 200 000 и 58 800 000 кварт.

Пример 15. Определите по данным примера 13 99%-ные доверительные границы средней. *Ответ:* 165 ± 23 кварталы.

9. Разности. Проверка нулевых гипотез. В параграфах 13 главы 1 и 9 главы 2 вы познакомились с вопросом проверки гипотез, относящихся к совокупности, из которых взяты выборки. Теперь, после ознакомления

с техникой отбора выборок из гипотетической совокупности, представляется возможным изучить тему последнего из указанных двух параграфов более подробно. На практике нулевая гипотеза чаще всего выдвигается тогда, когда допускается отсутствие различий между двумя совокупностями, из которых взяты выборки, т. е. допускается, что наиболее контрастные члены этих выборок получены все же из одного источника. В данном случае удобной статистической моделью будет представление о нормальной совокупности разностей, средняя которой равна нулю. Образец такой совокупности можно легко создать, если воспользоваться такой теоремой: если из наблюдений, случайно отобранных из нормальной совокупности, будут случайным же порядком составлены пары, то разности таких парных наблюдений будут распределены нормально со средней, равной нулю. Проще всего сделать так: образуйте из ваших выборок пары, как это показано в таблице 22, отнимите из каждого наблюдения первой выборки соответствующее наблюдение второй выборки и рассматривайте эти разности в качестве членов новых выборок. Разумеется, что выборки в этом случае должны браться такими, какими они получены в процессе первоначального отбора, без какого-либо предвзятого назначения одной из них первой, а другой — второй и без какой-либо перестановки наблюдений внутри них. Две такие пары отобраны из наших 511 выборок для иллюстрации обсуждаемых далее положений. На третьем же специально подобранном примере будет установлено положение, которое понадобится нам в дальнейшем.

ТАБЛИЦА 22

Две случайные выборки разностей, взятые из таблицы 15,
и одна специально подобранная выборка

$$\mu_d = 0, \quad s_d^2 = 200$$

1. Случайная выборка			2. Случайная выборка			Специально подобранная выборка		
парные наблюдения		разности	парные наблюдения		разности	парные наблюдения		разности
X ₁	X ₂		X ₁	X ₂		X ₁	X ₂	
39	17	22	19	33	-14	32	22	10
34	29	5	14	30	-16	31	17	14
22	30	-8	57	41	16	28	34	-6
27	36	-9	34	42	-8	24	24	0
33	41	-8	39	33	6	44	12	32
42	30	12	34	21	13	53	29	24
36	3	33	39	36	3	9	23	-14
24	23	1	13	33	-20	35	37	-2
25	38	-13	39	33	6	33	21	12
29	30	-1	23	22	1	31	11	20
\bar{d}			-1,3			9		
s_d^2			155,6			207,4		
s_d^2			15,56			20,74		
$s_{\bar{d}}$			3,94			4,55		
t			-0,33			1,98		

Сразу можно отметить следующие два факта: 1) почти равномерное распределение разностей, взятых по их численности или по их величине, между областями положительных и отрицательных чисел, в результате чего выборочные средние с одинаковой вероятностью будут как больше, так и меньше нуля, и 2) варьирование разностей много больше варьирования первоначальных данных. Причину возникновения этого второго факта легко установить, если вы учтете, что в нашем примере таблицы 15 возможный размах варьирования разностей простирается от 3—57 = -54 фунтам до 57—3 = 54 фунтам.

т. е. в два раза больше первоначального размаха в привесе животных. Теперь для вас не будут неожиданными такие теоремы: 1) дисперсия случайных разностей σ_D^2 равна удвоенной дисперсии исходной совокупности и 2) средний квадрат выборочных разностей s_D^2 является несмещенной оценкой величины $2\sigma^2=200$. Так, если вы найдете среднюю из средних квадратов по всем вашим выборкам разностей, то она не должна слишком отличаться от 200, т. е. от удвоенной дисперсии привеса свиней. Я имел 144 таких средних квадрата и получил среднюю из них — 207,9. Производя вычисления s_D^2 по своим выборкам, вы, несомненно, обратите внимание и, может быть, будете удивлены большим варьированием этой величины.

Теперь можно сформулировать следующую теорему: как разности, так и выборочные средние этих разностей распределены нормально около $\mu=0$; дисперсия средних разностей будет $\sigma_a^2=2\sigma/n$. Как и ранее, средний квадрат s_a^2 каждой выборки является несмещенной оценкой σ_a^2 . Для проверки этого я определил среднюю из моих 144 выборочных средних разностей и получил 0,51 фунта при среднем квадрате, равном 22,2; ожидаемые значения $\mu_a=0$ и $\sigma_a^2=20$.

То обстоятельство, что дисперсии разностей и средней разности в два раза больше соответствующих дисперсий исходной совокупности, может показаться странным тем, кто намеревается проводить опыты, в которых основное значение имеют разности вариантов. Возникает вопрос: что в условиях столь большого варьирования можно предпринять, чтобы получить такие доверительные интервалы для средних разностей, которые были бы с практической точки зрения достаточно узкими? Обычный путь — это увеличение объема выборок, но это не всегда бывает возможным. Иногда вопрос может быть разрешен путем уменьшения экспериментальных ошибок, а также путем ограничения, в пределах возможного, случайности при образовании парных наблюдений. В следующем параграфе будут рассмотрены обе эти возможности. В настоящий же момент нам следует обратить внимание на значения t в наших выборках разностей. Так как μ равно нулю, то формула

$$t = \bar{d}/s_a$$

будет совпадать с той, которая дана в параграфе 9 главы 2. Расчеты по этой формуле приведены в таблице 22. Здесь наблюдается один новый и интересный факт. Так как разности имеют большее варьирование, чем первоначальные данные о привесе животных, то как \bar{d} , так и s_a будут иметь более широкий размах варьирования, но отношение их, t , распределено точно так же, как и t в таблице 21, доказательством чему служит распределение таблицы 23. Очевидно, поэтому 144 значения t могут быть объединены с прежними 511 значениями, в результате чего получится выборка из 655 случайных и независимых значений t , взятая из одной и той же совокупности t . Данное положение позволяет использовать одну и ту же таблицу значений t как в случае выборок первоначальных привесов животных, так и в случае выборок разностей между ними. При некотором определенном размере выборки распределение t зависит только от нормальности совокупности, из которой взята выборка, безотносительно к тому, каковы значения μ и σ . В этом заключается одна из причин широкого применения этого статистического показателя.

Будут ли в ваших выборках разностей такие значения t , которые больше (по своей абсолютной величине) 5%-ного уровня, т. е. большие, чем 2,262? Среди своих 144 выборочных t я нашел 10 таких значений вместо ожидаемых 7. Одно значение было даже больше 3,250, т. е. 1%-ного уровня. Вы уже осведомлены о том, что если бы эти данные были получены в результате экспериментов, то нулевая гипотеза была бы в этих 10 случаях отвергнута, хотя она фактически была бы верна.

Необходимо еще отметить следующее: если в таблице 22 X_1 будут случайно взяты из $N(\mu_1, \sigma_1^2)$, что символически означает: «нормальное распределение со средней μ_1 и дисперсией σ_1^2 », а X_2 случайно взяты из $N(\mu_2, \sigma_2^2)$, то разности

Распределение 144 выборочных значений t для разностей, полученных случайным образом из таблицы 15; $n=10$

Классовые интервалы 1		Частота	Частоты (в %)	Теоретические частоты (в %)
—	—2,821	0	0,00	1
—2,821	—1,833	6	4,17	4
—1,833	—1,100	11	7,64	10
—1,100	0,000	48	33,33	35
0,000	1,100	55	38,20	35
1,100	1,833	13	9,03	10
1,833	1,821	10	6,94	4
2,821	—	1	0,69	1
		144	100,00	100

¹ В связи с тем, что данная выборка t мала, произведемо попарное объединение интервалов таблицы 21.

$X_1 - X_2$ будут иметь нормальное распределение со средней $\mu_1 - \mu_2$ и дисперсией $\sigma_1^2 + \sigma_2^2$. Здесь мы имеем дело с более общей формулой, а приведенная ранее является частным случаем ее при условии $\mu_1 = \mu_2$ и $\sigma_1^2 = \sigma_2^2$.

Пример 16. Возьмите ваши выборки из данных таблицы 15 и расположите в одну колонку первые 30 наблюдений, а во вторую — следующие 30 наблюдений. Полученные между ними разности образуют случайную выборку из нормальной совокупности при $\mu_D = 0$ и $\sigma_d^2 = 200$. Вычислите по данным этой вашей выборки оценки \bar{d} и s_d^2 .

Вспомнив изложенное в параграфе 14 главы 2, вы можете на основе этих выборочных показателей построить интервальную оценку для σ_d^2 , а также проверить гипотезу $\sigma_d^2 = 200$. Получилась ли ваша выборка обычной или, наоборот, она оказалась необычного вида?

Пример 17. В параграфе 10 главы 2 было обращено внимание на то, что в таблице 12 осуществлено специальное, а не случайное образование парных наблюдений над действием вируса табачной мозаики. Теперь на основе варьирования разностей вы можете определить эффективность такого образования пар. Допустим, что наблюдения X_1 случайно соединены в пары с наблюдениями X_2 . В этом случае необходимо установить оценку дисперсии полученной этим путем совокупности случайных разностей. Чтобы подойти к решению этого вопроса, можно с некоторым основанием допустить, что в обеих совокупностях, относящихся к двум препаратам вируса, имеется одна и та же величина σ^2 . Используя приведенные ранее данные о числе мест поражения X_1 и X_2 , находим:

Препарат	Σx^2	Число степеней свободы
1	468	7
2	172	7
Итого . . .	640	14

Отсюда $s^2 = 640/14 = 45,71$, что является оценкой дисперсии, общей для обеих совокупностей, не зависящей от значений двух средних. Дисперсия разностей будет в данном случае оцениваться величиной $2 \times 45,71 = 91,42$. В соответствии с тем, что говорилось по этому поводу в параграфе 10 главы 2, полученная здесь величина примерно в 5 раз больше фактической оценки $s^2 = 18,75$, полученной в таблице 12. Условия проведения опыта с препаратами вируса были выбраны так, чтобы устранить слишком большое варьирование, возникающее при случайном образовании парных наблюдений; это построение опыта основывалось на том факте, что две половины одного и того же листа в отношении их реакции на вирус находятся в более сходных условиях, чем случайно выбранные половины у разных листьев.

Пример 18. Коллинс, Флинт и Мак-Лейн [2] изучали влияние слабого электрического тока на рост всходов кукурузы. Проростки были выращены попарно в ящиках, причем один из них был подвергнут воздействию тока, а другой не подвергался этому воздействию. Разности высоты растений (под воздействием тока — без воздействия его) были

получены следующие: 6,0; 1,3; 10,2; 23,9; 3,1; 6,8; -1,5; -14,7; -3,3 и 11,1 мм. Хотя и так очевидно, что нулевая гипотеза $\mu=0$ не будет отвергнута, однако для практики все же проверьте гипотезу $H_0: \mu=0$ и $H_A: \mu \neq 0$. Значение $t=1,33$. К какому вы придете заключению относительно влияния электрического тока на рост всходов кукурузы?

Пример 19. Фишер в главе 3 книги «Планирование опытов» [4] приводит интересный случай об одном из наиболее ранних применений парного метода при оценке результатов эксперимента. Из того, что там сказано, следует, что даже Дарвин и Гальтон были недостаточно осведомлены о принципах, лежащих в основе такого рода экспериментов.

Пример 20. Доказать, что средний квадрат величины aX определяется формулой

$$s_{aX}^2 = a^2 s_X^2.$$

Пример 21. Доказать:

$$s_{aX+bY}^2 = a^2 s_X^2 + b^2 s_Y^2.$$

10. Снижение ошибки выборочного наблюдения. Наблюдая повышенное варьирование случайных разностей, мы предвидели, что в экспериментальных работах иногда может возникнуть необходимость устранения этого отрицательного явления. Увеличением выборки до большого числа парных наблюдений можно уменьшить $s_{\bar{x}}$ до любого заданного значения, но этот путь будет слишком дорогим способом. Более того, так как на практике выборки часто берутся из совокупностей, размер которых ограничен, то увеличение n до любого значения становится просто невозможным. Следовательно, для повышения точности необходимо искать какие-то другие пути.

В общем случае можно указать на следующие три возможности уменьшения выборочных ошибок: 1) выбирать или образовывать такие совокупности, в отношении которых известно, что варьирование внутри них незначительно; 2) подразделять совокупности на отдельные группы (strata) схожих друг с другом объектов, и 3) использовать сведения о сопутствующих условиях. Эти способы уменьшения ошибки могут быть комбинированы друг с другом. Некоторые сведения о них будут даны здесь, но общее рассмотрение вопроса распределено по всем следующим частям этой книги.

В некоторых случаях практически допустимо основываться на более ограниченных совокупностях, имеющих внутри себя небольшое варьирование. Например, при сборе сведений о слушателях определенной радиопрограммы можно пойти на ограничение опрашиваемого круга лиц только женщинами-домохозяйками некоторого города, женщинами, имеющими одного ребенка или более, или семьи которых имеют доход, положим, между 3000 и 4000 долл. Переменная величина, такая, как, например, расход на туалетное мыло, внутри такой ограниченной совокупности может иметь более узкий размах варьирования. Ясно, что этим путем мы не получаем никакой информации относительно других частей общей совокупности, но сведения о совокупности, из которой берется выборка, будут более точными. Еще пример: при изучении некоторых вопросов питания на крысах представляется достаточным взять в качестве выборки генетически однородную группу животных, у которых размах варьирования веса в 28-дневном возрасте, положим, в два раза меньший, чем у животных, полученных в результате случайного отбора. Стандартные ошибки средних для n таких животных будут составлять половину ошибок в совокупности животных смешанного происхождения, что приведет к сокращению, в этом же отношении доверительных интервалов и к удвоению выборочных значений t .

Установленные здесь факты могут быть проверены при помощи выборок из нормальной совокупности, приведенной в таблице 24. Эта совокупность имеет $\sigma=5$ фунтов вместо $\sigma=10$ фунтов совокупности из таблицы 15; размах варьирования ее также составляет половину прежнего значения. Так как указанные факты очевидны, то вряд ли вам следует для проверки их отбирать серию выборок из этой таблицы; она будет использована нами в дальнейшем для других целей. Пожалуй, вполне достаточно привести полученное в нашей лаборатории распределение 400 средних из 10 наблюдений, представленное в таблице 25. Можно видеть, что размах варьирования здесь составляет только половину размаха таблицы 18. Средняя из выборочных

Ранжированный ряд привесов (в фунтах) свиней, имеющий размах варьирования и стандартное отклонение, в два раза меньшие, чем соответствующие показатели таблицы 15. Распределение приблизительно нормальное при $\mu=30$ фунтам и $\sigma=5$ фунтам

№ наблюдения	Привес	№ наблюдения	Привес	№ наблюдения	Привес	№ наблюдения	Привес
00	17	25	27	50	30	75	33
01	19	26	27	51	30	76	34
02	20	27	27	52	30	77	34
03	21	28	27	53	30	78	34
04	22	29	27	54	31	79	34
05	22	30	27	55	31	80	34
06	22	31	28	56	31	81	34
07	23	32	28	57	31	82	35
08	23	33	28	58	31	83	35
09	23	34	28	59	31	84	35
10	23	35	28	60	31	85	35
11	24	36	28	61	31	86	35
12	24	37	29	62	32	87	36
13	24	38	29	63	32	88	36
14	25	39	29	64	32	89	36
15	25	40	29	65	32	90	36
16	25	41	29	66	32	91	37
17	25	42	29	67	32	92	37
18	26	43	29	68	33	93	37
19	26	44	29	69	33	94	38
20	26	45	30	70	33	95	38
21	26	46	30	71	33	96	39
22	26	47	30	72	33	97	40
23	26	48	30	73	33	98	41
24	27	49	30	74	33	99	44

средних таблицы 25 равна 30 фунтам, а стандартное отклонение средних составляет 1,57 фунта, в то время как ожидаемое значение его равно

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 5/\sqrt{10} = 1,58 \text{ фунта.}$$

Очевидно, что этот способ уменьшения выборочного варьирования, сводящийся к отбору выборок из совокупностей с малым размахом варьирования, является достаточно эффективным. К сожалению, далеко не всегда имеются условия, при которых этот метод может быть применен. Обычно требуются сведения относительно таких фактических совокупностей, варьирование которых не находится под контролем лица, производящего выборки; вместе с тем часто менее варьирующая совокупность не может репрезентировать ту, другую, которая нас интересует.

Второй способ уменьшения варьирования ошибок путем подразделения совокупности на относительно однородные части находит себе применение при самых различных обстоятельствах. Он применяется в большинстве случаев выборочного наблюдения и экспериментальных схем самостоятельно или совместно с другими способами. Читатель по мере чтения настоящей книги много раз встретится с этим методом, после чего он познакомится со специальным обсуждением данного вопроса. Примером этого, второго способа уменьшения варьирования может служить описанный в параграфе 9 главы 2 случай проведения парных наблюдений на двух половинах табачных листьев, на которых изучалась концентрация вируса мозаики табака. Площадки листовой поверхности, отведенные для двух препаратов, были взяты в соседстве друг с другом потому, что есть основание считать, что их реакция на вирус более схожа, чем реакция площадок на различных листьях. Поэтому, если на одну половину листа будет нанесен препарат № 1, а на другую — препарат № 2, то следует ожидать, что разности в степени поражения их будут варьировать меньше, чем в том случае, когда эти препараты будут

нанесены на половины листьев, выбранных в случайном порядке на различных растениях.

Следует еще раз подчеркнуть, что эффективность этого метода расслоения совокупности зависит от наших предварительных знаний относительно особенностей подопытного материала; эти знания могут быть приобретены только на основе накопленного опыта при работе с данным материалом. Нельзя ожидать преимущества этого метода, если при отсутствии различия между вариантами не будет обнаружено сходства членов, составляющих пару. Если различие между этими членами столь же велико, как и различие между парами наблюдений, то метод расслоения совокупности не будет лучшим по сравнению с методом случайного образования парных наблюдений.

ТАБЛИЦА 25

Распределение средних по выборкам из 10 наблюдений, случайно отобранных из нормальной совокупности таблицы 24

Середина класса (фунты)	26	27	28	29	30	31	32	33	34
Частоты	5	20	43	80	99	85	48	18	2

Является ли способ расслоения совокупности отступлением от основного принципа рендомизации? Во все нет. Он является просто способом, направленным на использование наших знаний о возможном результате исключения из оценки ошибки той вариации, которая нами предугадывается. В эксперименте с вирусом табачной мозаики на основании предварительных наблюдений было установлено два источника варьирования: 1) различия при переходе от растения к растению и 2) различия между листьями одного и того же растения. Эта вариация была исключена из оценки экспериментальной ошибки путем проверки действия обоих препаратов на двух половинах одного и того же листа. Здесь рендомизация была осуществлена путем размещения по жребью двух препаратов вируса на правой и левой половинах листа; сама жеребьевка была проведена при помощи некоторого механического способа, как, например, подбрасывание монеты или выбор случайных чисел. В этом случае варьирование, включенное в оценку ошибки опыта, обусловлено только случайными различиями в реакции двух участков одного и того же листа.

О третьем методе контролирования ошибки, основанном на использовании дополнительных сведений, которые дают наблюдения над сопутствующими условиями, будет сказано в главе 6.

Пример 22. В примере 30 главы 2 были приведены данные о биологической ценности сырых и поджаренных семян арахиса, полученные в опыте на 10 парах крыс. Выборочное стандартное отклонение разностей равно 7,72, поэтому выборочная дисперсия в пересчете на одно животное равна $7,72^2/2=29,8$. Определите среднюю из средних квадратов, вычисленных по каждой из двух групп животных. *Ответ:* 26,9. Это указывает на то, что при определении биологической ценности изучаемого корма парный метод оказался менее эффективным, чем способ случайного отбора.

Пример 23. Если в примере 30 главы 2 не включать в обработку девятую пару наблюдений (44, 63), то вычисления показывают, что при парном методе средний квадрат в пересчете на одно животное будет примерно в 4 раза меньше средней из средних квадратов для двух групп животных без образования пар. Так как столь сильное изменение в соотношении двух методов зависит только от одной пары крыс, то исследователь должен проявить известную осторожность в своих выводах.

Пример 24. В таблице 22 для первой выборки разностей $s_D^2=222,9$. Произведите оценку σ^2 той совокупности, из которой взяты две выборки X_1 и X_2 . *Ответ:* σ^2 оценивается значением 111,45.

Пример 25. Если из двух нормальных совокупностей, имеющих средние μ_1 и μ_2 и дисперсии σ_1^2 и σ_2^2 , взято по одной выборке объема n и если из этих наблюдений составлены случайные пары, распределение попарных разностей будет нормальным со средней $\mu_D=\mu_1-\mu_2$ и дисперсией $\sigma_D^2=\sigma_1^2+\sigma_2^2$

Если вы возьмете из таблицы 15 выборку в 25 наблюдений и такую же выборку из таблицы 24, то какой параметр будет оценивать средняя \bar{d} из разностей случайно образованных пар? *Ответ:* $\mu_D=0$. Какой параметр будет оценивать соответствующее стандартное отклонение? *Ответ:* $\sigma_D=\sqrt{125}$, или 11,18 фунта (параграф 9 главы 3, последний абзац). Проверьте это.

Л И Т Е Р А Т У Р А

1. B r y a n A. A., Iowa Agricultural Experiment Station Research Bulletin 163, 1933.
2. C o l l i n s G. N., F l i n t L. H., M c L a n e J. W., Journal of Agricultural Research, 38, 585, 1929.
3. D a v i d Pedro A., Iowa State College Journal of Science, 5, 285, 1931.
4. F i s h e r R. A., Design of Experiments. Oliver and Boyd, London, 1942.
5. M e r c e r W. B., H a l l A. D., Journal of Agricultural Science, 4, 107, 1911.
6. M o o d Alexander McFarlane, Introduction to the Theory of Statistics. McGraw-Hill Book Company, Inc., New York, 1950.
7. W i e b e Gustav A., Journal of Agricultural Research, 50, 331, 1935.

СРАВНЕНИЕ ДВУХ СЛУЧАЙНЫХ ГРУПП

1. Введение. Образование из сходных объектов парных наблюдений является, как мы видели, довольно эффективным методом построения экспериментальных схем. Но у нас может не хватать сведений об особенностях экспериментального материала для правильного выбора таких пар, и тогда парный метод может оказаться невыгодным. В этом случае проще всего произвести случайное распределение объектов между двумя группами или партиями и предоставить каждую из них одному из вариантов опыта. Теперь будет производиться сравнение уже только двух средних, а не отдельных наблюдений. Соответствующая ошибка опыта в этом случае определяется обобщенным варьированием объектов внутри групп.

2. Построение опыта с двумя случайными группами или партиями. С особенностями такого опыта можно познакомиться, например, по двум случайным выборкам из данных о привесе свиней, приведенным в главе 3. Рассмотрим выборки № 1 и 2 таблицы 17. Допустим, что первая выборка является группой свиней, получающих рацион, который способен дать средний привес совокупности в μ_1 фунтов. Рацион второй группы может дать μ_2 фунтов привеса. На самом же деле мы имеем групповые средние $\bar{x}_1=35,6$ и $\bar{x}_2=29,3$ фунта. Знакомая уже нам постановка вопроса будет такой: обусловлена ли разность $\bar{x}_1-\bar{x}_2=6,3$ фунта наличием реальной разности $\mu_1-\mu_2$ или она есть результат случайных отклонений от средней одной и той же совокупности?

Для решения этого вопроса мы вводим нулевую гипотезу: $\mu_1-\mu_2=0$. Распределение t дает нам критерий:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}},$$

который при $H_0: \mu_1 - \mu_2 = 0$ обращается в такую форму:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}}.$$

Здесь остается только определить знаменатель.

Каждая из этих групп дает оценку общей для обеих совокупностей дисперсии σ^2 . Оценка $s_1^2=169,8$ и оценка $s_2^2=151,6$. Средняя из них будет $(169,8 + 151,6) : 2 = 160,7$. По этой величине мы можем найти дисперсию каждой из средних:

$$s_x^2 = \frac{s^2}{n} = \frac{160,7}{10} = 16,07.$$

Но нам дана разность этих двух средних, и необходимо найти оценку дисперсии этой разности. Разности между средними двух выборок из нормальной

совокупности подчиняются тому же закону, что и разности между отдельными наблюдениями, так как они, подобно этим последним, распределены нормально. Следовательно, мы имеем (параграф 9 главы 3):

$$s_{x_1 - x_2}^2 = 2s_x^2 = 2 \times 16,07 = 32,14.$$

Подставляя в формулу t разность между групповыми средними 6,3 фунта и ее стандартное отклонение $\sqrt{32,14} = 5,67$, получим:

$$t = \frac{6,3}{5,67} = 1,11.$$

Так как s_1 соответствует $n-1=9$ степеням свободы и s_2 тоже 9 степеням свободы, то в целом будет $2(n-1)=18$ степеням свободы; здесь n — численность каждой группы.

Сравнение $t=1,11$, при числе степеней свободы = 18 со значением из таблицы 11 указывает на то, что в данном случае t имеет небольшое значение. Нулевая гипотеза остается непровергнутой, что и должно быть в случае нашего условного опыта.

Теперь посмотрим, что будет, если изучаемые рационы приводят к фактическому изменению средних. Добавим к каждому наблюдению первой группы по 10 фунтов. В результате получим $\mu_1=40$ фунтам вместо прежних 30 и $\bar{x}_1=45,6$ фунта вместо 35,6. Выборочное стандартное отклонение останется без изменения. Подстановка новых данных в формулу t при прежней гипотезе $H_0: \mu_1 - \mu_2 = 0$ дает:

$$t = \frac{45,6 - 29,3}{5,67} = 2,87; \text{ число степеней свободы} = 18.$$

Это указывает на необходимость отвергнуть гипотезу H_0 и приводит к заключению, что μ_1 и μ_2 различны.

Теперь рассмотрим особенности данной схемы эксперимента. Две группы объектов получены из совокупностей, которые относятся к разным вариантам и могут иметь различные средние, но отнюдь не различные дисперсии. Разность между групповыми средними является оценкой разности средних в совокупностях $\mu_1 - \mu_2$. Эта разность испытывается при помощи нулевой гипотезы $\mu_1 - \mu_2 = 0$, т. е. равенства $\mu_1 = \mu_2$. Дисперсия ошибки σ^2 одинакова для обеих совокупностей. В связи с этим дисперсия разности между средними $2\sigma^2/n$ оценивается величиной

$$\frac{2}{n} \cdot \frac{s_1^2 + s_2^2}{2} = \frac{2}{n} \cdot \frac{\sum x_1^2 + \sum x_2^2}{2(n-1)}.$$

Теперь можно обратиться уже к фактическому опыту этого типа.

3. Опыт, в котором производится сравнение двух групп одинакового размера. Бренеман [2] сопоставлял 15-дневный средний вес гребешка у молодых петушков, разбитых на две группы, одна из которых получала половой гормон А (тестостерон), а вторая — гормон С (дегидроандростерон). Для каждого из вариантов опыта были выделены в случайном порядке две группы однодневных цыплят, по 11 штук в каждой. Обе группы содержали совместно, а для различения их применяли мечение голов в красный и пурпурный цвет. В таблице 26 приведены результаты взвешивания гребешков. При такой постановке опыта нет никаких оснований для сравнения веса гребешка у какого-нибудь петушка, получавшего гормон А, с весом гребешка у какого-то определенного петушка, получавшего гормон С; здесь имеет смысл сравнивать только средние.

Расчеты удобнее всего располагать по схеме таблицы 27. По сравнению с предыдущей схемой здесь введено одно изменение: вместо вычисления двух средних квадратов и выведения из них средней произведено сложение, т. е. объединение скорректированных сумм квадратов, и полученная сумма разделена на общее число степеней свободы.

Вес гребешков¹ (за 15 дней) у петушков, получавших гормоны А и С

Гор- мон	Число петушков	Вес гребешка (мг)											Сред- няя
		57	120	101	137	119	117	104	73	53	68	118	
А	11	57	120	101	137	119	117	104	73	53	68	118	97
С	11	89	30	82	50	39	22	57	32	96	31	88	56

¹ Для упрощения вычислений произведено незначительное изменение в одном из первоначальных показателей.

Сводные показатели при сравнении двух групп одинакового размера
(По данным таблицы 26)

Гор- мон	Число петушков	Число степеней свободы	Средний вес гребешка (в мг)	Сумма квадратов
А	11	10	97	8472
С	11	10	56	7748
Сумма = 20			Разность $\bar{x}_1 - \bar{x}_2 = 41$	Сумма = $\Sigma x^2 = 16220$

$$\text{Обобщенный средний квадрат} = s^2 = \frac{\Sigma x^2}{2(n-1)} = 16220/20 = 811.$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{2s^2/n} = \sqrt{2 \times 811/11} = \sqrt{147,5} = 12,14 \text{ мг.}$$

$$t = (\bar{x}_1 - \bar{x}_2) / s_{\bar{x}_1 - \bar{x}_2} = 41/12,14 = 3,38; \quad p < 1\%.$$

Эти данные убеждают в том, что гормоны приводят к образованию двух совокупностей веса петушинных гребешков с различными средними значениями; это означает, что разность в 41 мг существенна.

Отметим, что здесь в основу положено допущение о равенстве дисперсий двух совокупностей. Одного взгляда на приведенные выше суммы квадратов достаточно для того, чтобы считать это допущение обоснованным. Обычно варианты подобного рода не изменяют дисперсию, а если и изменяют, то так незначительно, что это в опыте не обнаруживается. Но в некоторых других опытах нельзя идти так легко на такое допущение. Если имеются основания ожидать, что под влиянием изучаемых вариантов σ будет меняться, то необходимо применять метод совсем иной, чем описанный здесь. Смотрите параграф 9 этой главы.

После того как читатель усвоил предпосылки, положенные в основу таблицы 27, он может легко использовать ее данные для получения окончательных выводов при помощи формулы

$$t = (\bar{x}_1 - \bar{x}_2) \sqrt{\frac{n(n-1)}{\Sigma x^2}},$$

в которой $\bar{x}_1 - \bar{x}_2$ означает разность двух групповых средних, а Σx^2 является объединенной суммой квадратов. Подстановка данных из опыта с цыплятами $\bar{x}_1 - \bar{x}_2 = 97 - 56 = 41$ мг, $n = 11$ и $\Sigma x^2 = 8472 + 7748 = 16220$ дает

$$t = 41 \times \sqrt{\frac{11 \times 10}{16220}} = 3,38,$$

т. е. то же значение, что и ранее.

Придя к выводу о том, что в данном случае имеются не одна, а две различные совокупности, можно определить доверительные пределы для разности их средних, которая в качестве несмещенной оценки имеет наблюдаемую

разность в 41 мг. Так как для 20 степеней свободы $t_{0,05} = 2,086$, то 95 %-ные пределы будут

$$41 - 2,086 \times 12,1 \quad \text{и} \quad 41 + 2,086 \times 12,1,$$

т. е. 16 и 66 мг. Если экспериментатор будет утверждать, что разность между средними двух совокупностей находится между этими пределами, то он имеет один шанс из двадцати оказаться неправым, когда разность фактически выйдет из этих пределов.

Заметим, что этот доверительный интервал можно использовать для заключения по поводу гипотезы $H_0: \mu_1 - \mu_2 = 0$. Так как он не содержит в себе нуль, то H_0 отвергается.

Пример 1. Две группы пчел, по 10 штук в каждой, подкармливали сиропом 20- и 65%-ной концентрации; кормушка находилась на расстоянии полмили от улья [7]. После прилета пчел к улью их взяток извлекали и устанавливали концентрацию в нем сахара. Во всех случаях наблюдалось уменьшение концентрации сахара по сравнению с той, которая была в кормушке. Уменьшение для 20%-ного сиропа составляло: 0,7; 0,5; 0,4; 0,7; 0,5; 0,4; 0,7; 0,4; 0,2 и 0,5% и для 65%-ного — 1,7; 2,8; 2,2; 1,4; 1,3; 2,1; 0,8; 3,4; 1,9 и 1,4%. Здесь каждое наблюдение второй выборки больше любого наблюдения первой выборки, поэтому скорей всего $\mu_1 < \mu_2$. Показать, что при $\mu_1 - \mu_2 = 0$ критерий $t = 5,6$. Можно почти не сомневаться, что в данных условиях опыта в течение перелета происходит большее снижение концентрации в случае подкормки 65%-ным сиропом. Но как обстоит дело в отношении однородности дисперсий? Обсуждение этого вопроса смотри в параграфах 8 и 9 этой главы.

Пример 2. Произведено по 4 определения рН в одном образце суглинка с помощью двух типов стеклянных электродов [5]. При использовании электродов первого типа были получены показания: 5,78; 5,74; 5,84 и 5,80, а второго — 5,82; 5,87; 5,96 и 5,89. Взяв гипотезу $\mu_1 - \mu_2 = 0$, определите $t = 2,66$.

Пример 3. 15 зрелых кукурузных зерен были испытаны на сопротивление раздавливанию. Измерение этого сопротивления в фунтах дало: 50, 36, 34, 45, 56, 42, 53, 25, 65, 33, 40, 42, 39, 43, 42. Другая проба из 15 зерен была испытана при уборке кукурузы в стадии восковой спелости и дала: 43, 44, 51, 40, 29, 49, 39, 59, 43, 48, 67, 44, 46, 54, 64. Проверьте существенность разности между этими двумя средними. *Ответ:* $t = 1,38$.

Пример 4. При случайном подборе годовалых бычков для опыта по откорму их предшествующая история так мало известна, что предварительная оценка продуктивности становится просто невозможной. В этом случае как раз и следует применить способ случайных групп. Кальбертсон и Хэммонд [4] приводят средний привес за день двух групп бычков, одна из которых получала дополнительно к корму льняной жмых, а вторая — небольшую порцию (1 фунт в день на бычка) семян сои. Скорость роста была: у первой группы — 1,95; 2,17; 2,06; 2,11; 2,24; 2,52; 2,04 и 1,95 и у второй — 1,82; 1,85; 1,87; 1,74; 2,04; 1,78; 1,76 и 1,86 фунта в день. Определите разность средних 0,29 фунта в день и $t = 3,92$.

Пример 5. Вы, возможно, удивлены тем, что в опытах по методу группового сравнения не производится случайного образования попарных наблюдений и не применяется обработка данных по способу параграфа 9 главы 2 и таблицы 22. Вообще это можно было бы сделать. Но в этом случае при парном методе возникают некоторые дополнительные осложнения, которые лучше избегать. Дело в том, что возможно множество различных соединений в пары, каждая из которых будет приводить к своему особому результату. В этом аспекте приведенные в таблице 27 данные представляют собой как бы средние результаты всех возможных случаев образования пар.

Две первые выборки таблицы 17, если их наблюдения взять попарно в том виде, как они получены, образуют случайные пары. Соответствующие разности будут $33 - 32 = 1$, $53 - 31 = 22$ и т. д., причем некоторые из них будут отрицательными. Если вы в этом случае произведете оценку существенности разности 6,3 фунта, то получите $t = 0,88$ вместо значения, полученного в параграфе 2 этой главы. Другие случайные попарные комбинации будут приводить к иным значениям t ; как вы увидите из дальнейшего (параграф 6 этой главы), некоторые из них случайно будут казаться даже существенными.

Пример 6. При ознакомлении с научной литературой иногда возникает желание произвести оценку существенности различий, которую автор почему-то не счел нужным сделать. Например, Смит [10] дает средние урожаи и стандартные ошибки для двух гибридов кукурузы: $8,84 \pm 0,39$ и $7,00 \pm 0,18$. Каждая средняя была получена при 5 повторностях. Мы желаем установить, является ли эта разность существенной. В этом случае необходимо восстановить первоначальные результаты, дойдя до каждой отдельной суммы квадратов. Начиная с величины $s_x = 0,39$, получаем последовательно: $s_x^2 = 0,1521$, $s^2 = 0,7605$ и $\Sigma x^2 = 3,042$. Подобно этому, для второго гибрида $\Sigma x^2 = 0,648$. Объединяя эти две суммы квадратов, можно определить значение t при 8 степенях свободы. *Ответ:* $t = 4,29$.

Пример 7. Выведите формулу $t = (\bar{x}_1 - \bar{x}_2) \cdot \sqrt{\frac{n(n-1)}{\Sigma x^2}}$; начать следует с расчетных формул, приведенных в конце таблицы 27.

4. Затруднения, встречающиеся в опытах с групповым сравнением. Описанный выше метод проверки нулевой гипотезы $\mu_1 - \mu_2 = 0$ основан на допущении, что обе совокупности имеют нормальное распределение и что они имеют одну и ту же дисперсию.

Как правило, можно считать, что даже значительные отклонения от нормального распределения оказывают небольшое влияние на результаты. Вопрос о дисперсии совокупности будет рассмотрен в параграфе 8 этой главы. В данный же момент мы рассмотрим вопрос, в какой мере оценка дисперсии s^2 правильно определяет ошибки, которым подвержены экспериментальные средние. В примере, рассмотренном ранее, принимались меры, чтобы подобрать цыплят одного и того же происхождения, пола и возраста и поместить их в одинаковые условия. Предположительно считается, что нет никаких внешних сопутствующих условий, создающих различие между группами, кроме вариантов гормона, обусловивших различие между средними совокупностей. Однако во многих случаях такие схемы опытов с групповыми сравнениями могут вызывать сомнения. Если, например, группы размещены в различных помещениях или если на эти группы оказывает влияние их размещение внутри одного помещения, то наблюдаемые различия между совокупностями могут возникнуть как от действия вариантов опыта, так и под влиянием других сопутствующих условий, и поэтому s^2 не будет правильной оценкой ошибки опыта. В этом случае различия, обусловленные вариантами, как говорят, *смешаны* с другими различиями, причем не представляется возможным отделить одно от другого. При таких обстоятельствах констатация того, что средние совокупностей различны, имеет весьма малое отношение к вопросу о возможном влиянии вариантов опыта. Хотя в этом случае статистические показатели могут правильно отвергнуть нулевую гипотезу и хотя сама гипотеза может быть фактически неверной, все же варианты опыта сами по себе здесь могут быть не при чем и не оказывать никакого влияния или даже могут быть источником различий, противоположных наблюдаемым в опыте. Далее, так как нет никакой возможности произвести отдельную оценку величины и направления двух видов влияния со стороны вариантов и внешних сопутствующих условий, то остается неизвестным, в какой пропорции войдут эти два влияния в совокупность разностей, т. е. совокупность их остается неопределенной. Если опыт будет проведен так, чтобы избежать все эти неприятности, то σ^2 будет правильно оцениваться величиной s^2 и реальное влияние вариантов опыта может дать вполне достоверные разности.

5. Группы с разным количеством объектов. Вообще говоря, лучше всего иметь группы одинакового размера, но во многих экспериментальных работах неудобно или даже невозможно образовывать группы с равным числом объектов. Например, две группы цыплят, вылупившихся из двух групп яиц, составляющих варианты опыта, будут почти всегда различны и по численности. Более того, и в течение жизни цыплят надеж и различного рода случайности могут затронуть эти группы неодинаково. Все это не может, конечно, изменить статистическую теорию обработки данных, но некоторое видоизменение методов все же необходимо.

В качестве иллюстрации возьмем данные таблицы 28 о привесе двух групп крыс, получавших различное количество протеина. Если n_1 и n_2 —

ТАБЛИЦА 28

*Привесы двух групп крыс (самок) при различных рационах.
Привесы относятся к возрасту от 28 до 84 дней*

Рацион	Число крыс	Привес (в г)											
		134	146	104	119	124	161	107	83	113	129	97	123
Много протеина	12	134	146	104	119	124	161	107	83	113	129	97	123
Мало протеина	7	70	118	101	85	107	132	94					

количества наблюдений в двух группах, то соответствующие степени свободы будут $n_1 - 1$ и $n_2 - 1$, а в целом $n_1 + n_2 - 2$. Средние квадраты для двух групповых средних теперь различны — s^2/n_1 и s^2/n_2 , и поэтому средний квадрат разности между средними теперь будет равен не $2s^2/n$, а сумме этих средних квадратов:

$$s^2/n_1 + s^2/n_2 = s^2(1/n_1 + 1/n_2) = s^2 \frac{n_1 + n_2}{n_1 n_2},$$

откуда

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 \frac{n_1 + n_2}{n_1 n_2}} = s \sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

Если $n_1 = n_2$, то данная формула обращается, как это и должно быть, в формулу таблицы 27. Соответствующие расчеты приведены в таблице 29.

ТАБЛИЦА 29

Сводные показатели при сравнении двух групп неодинакового размера (по данным таблицы 28)

Рацион	Число крыс	Число степеней свободы	Средний привес (в г)	Сумма квадратов
Много протеина .	12	11	120	5032
Мало протеина . .	7	6	101	2552
		Сумма 17	Разность $\bar{x}_1 - \bar{x}_2 = 19$	Сумма $= \Sigma x^2 = 7584$

Обобщенный средний квадрат $= s^2 = 7584/17 = 446,12$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s^2 (n_1 + n_2)/n_1 n_2} = \sqrt{446,12 \times (12 + 7)/(12 \times 7)} = 10,04 \text{ г}$$

$t = 19/10,04 = 1,89$; число степеней свободы $= 17$; $P = 0,08$

Так как $P > 0,05$, то нулевая гипотеза не может быть отвергнута. В той мере, в какой имеющееся небольшое число наблюдений является показательным, следует считать, что различие в содержании протеина в пище не привело к расслоению совокупности привесов на две отдельные совокупности.

Чтобы убедиться в том, что дисперсии одного и того же порядка, сравните $s_1^2 = 5032 : 11 = 457$ и $s_2^2 = 2552 : 6 = 425$.

Описанный выше процесс оценки различий двух групп неодинакового размера может быть сведен в одну формулу:

$$t = (\bar{x}_1 - \bar{x}_2) \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{(n_1 + n_2) \Sigma x^2}},$$

где $\bar{x}_1 - \bar{x}_2$, как и прежде, — разность между групповыми средними и Σx^2 — объединенная сумма квадратов. Применяя эту формулу к данным о привесе крыс, найдем: $\bar{x}_1 - \bar{x}_2 = 120 - 101 = 19$ г; $n_1 n_2 = 12 \times 7 = 84$; $n_1 + n_2 = 19$; $n_1 + n_2 - 2 = 17$ и $\Sigma x^2 = 5032 + 2552 = 7584$. При подстановке в формулу получим, как и в таблице 29, $t = 1,89$.

Если исследователя больше интересует не проверка существенности разности, а ее доверительная оценка, то он может непосредственно определить доверительный интервал для оцениваемой разности. В нашем случае он может установить, что разность оценивается значением $\bar{x}_1 - \bar{x}_2 = 19$ г при доверительном интервале от $-2,2$ до $40,2$ г.

Пример 8. Ниже приводятся данные о скорости диффузии углекислоты в двух почвах различной свжажности [9]. Через легкую почву (f): 20, 31, 18, 23, 23, 28, 23, 26, 27, 26, 12, 17, 25; через тяжелую почву (c): 19, 30, 32, 28, 15, 26, 35, 18, 25, 27, 35, 34. Показать, что объединенная оценка $s^2 = 35,83$; $s_{\bar{x}_1 - \bar{x}_2} = 2,40$; число степеней свободы $= 23$ и $t = 1,67$. Следовательно, разность не будет существенной.

Пример 9. У нормальных альбиносных крыс в возрасте 37 и 180 дней было определено содержание общего азота в плазме крови [11]. Результаты выражены в граммах на 100 куб. см плазмы. Девять крыс в возрасте 37 дней имели: 0,98; 0,83; 0,99; 0,86; 0,90; 0,81; 0,94; 0,92 и 0,87, а 8 крыс в возрасте 180 дней имели: 1,20; 1,18; 1,33; 1,21; 1,20; 1,07; 1,13 и 1,12 г на 100 куб. см. Так как существенность различия здесь очевидна, то непосредственно определите 95%-ный доверительный интервал для разности средних в совокупности. *Ответ:* от 0,21 до 0,35 г на 100 куб. см.

Пример 10. Пирсон и Кетчпол [8] приводят средние и стандартные ошибки содержания неорганического фосфора на 100 мл сыворотки крови, определенного у 26 лошадей породы першерон и 18 лошадей шетлендской породы: $3,29 \pm 0,27$ и $3,96 \pm 0,40$ мг; здесь числа, стоящие со знаками \pm , означают выборочные стандартные ошибки. Чтобы определить существенность разности между средними 0,67 мг, вычислите объединенную сумму квадратов 96,345, ошибку разности $s_{x_1 - x_2} = 0,464$ и $t = 1,44$. Разность не является существенной, и вполне может быть, что она есть следствие выборочного варьирования.

Пример 11. Установлено, что средний процент бекона в тушах 15 свиной польско-китайской породы составлял 6,59 при стандартной ошибке 0,300%. Если вы здесь допустите нормальное распределение, то можете определить по этой выборке верхний и нижний пределы: 4,6 и 8,6%. Имеете ли вы какое-либо основание для сомнений по поводу гипотезы о нормальности?

6. Противоречие между индивидуальными и групповыми сравнениями.

Вопрос о том, следует ли опыт ориентировать на индивидуальные или групповые сравнения, имеет большое практическое значение. Решение его может даже предопределить удачу или неуспех самого эксперимента. Хотя нет никакой формулы для определения преимущества того или другого способа, все же имеются некоторые условия, на основе которых может быть найдено соответствующее решение. Многие из этих условий уже обсуждались в параграфах 10 главы 2 и 10 главы 3. Теперь с некоторым повторением предыдущего, на которое, я думаю, стоит затратить время, попытаемся рассмотреть этот вопрос в целом, привлекая соответствующую аргументацию.

Парный метод сравнения часто подсаживается самим экспериментальным материалом. Например, урожайность любой культуры, наблюдаемая за ряд лет, варьирует в такой степени, что вряд ли кому-нибудь придет мысль сравнивать вариант опыта прошлого года с другим вариантом данного года. Очевидно, что оба варианта должны изучаться в условиях того и другого года и давать соответствующие парные наблюдения урожая. То же самое следует сказать и в отношении территориального размещения вариантов при изучении их в одном и том же году.

С другой стороны, в некоторых опытах образование пар не имеет практического смысла. Так, в опыте Бренемана с цыплятами (параграф 3) нет никаких оснований для подбора пар, и поэтому здесь цыплята назначаются для вариантов по жребью. Когда производится отбор случайных бычков для опыта по откорму, то так мало известно об их предыдущей истории, что всякая предварительная оценка продуктивности становится сомнительной. Кроме того, если условия заставляют работать с различным числом объектов, предназначенных для разных вариантов, то парный метод просто невозможен.

Оставляя в стороне эти условия, когда решение вопроса очевидно, следует сказать, что построение опыта часто определяется наличием у исследователя сведений относительно подопытного материала и сопутствующих условий. Например, ему известно, что растения в сосудах, размещенных на различных стеллажах вегетационного домика, в одном и том же варианте опыта развиваются одинаково. В этом случае он может для одного варианта опыта взять сосуд, помещенный на одном стеллаже, а для другого — сосуд, находящийся на другом стеллаже, и производить сравнение этих вариантов по методу, описанному в этой главе. Однако гораздо чаще исследователь осведомлен о том, что условия произрастания растений на двух стеллажах и даже на соседних участках одного и того же стеллажа в какой-то мере различны. В этом случае он во избежание влияния этих варьирующих условий будет располагать сосуды с различными вариантами попарно, рядом друг с другом и получит повторения из этих пар. Соответствующий статистический метод обработки опытных данных будет таким, как в главе 2.

По мере того как в той или иной области в результате большей изученности особенностей подопытного материала происходит уточнение экспериментальных исследований, метод группового сравнения все чаще уступает место методу сравнения отдельных объектов. В опытах по откорму свиней обычно составляются пары животных, принадлежащих к одному помету и имеющих один и тот же пол и первоначальный вес, так как установлено, что при одинаковом рационе привес у таких животных более выравнен, чем у животных, разбитых на две группы случайным порядком. Поэтому, если один из членов каждой пары, выбранный случайно, получит один рацион, а другой член — другой, то различия в привесах будут примерно одинаковыми. Отсюда следует, что ошибка опыта, определенная по правилам, приведенным в главе 2, по всей вероятности, будет меньшей, чем при групповом сравнении, когда свиньи по изучаемым рационам распределяются по жребию.

В таких опытах весь успех парного метода зависит от предвидения результатов. Но от применения этого метода нельзя получить никакой выгоды, если экспериментатор не может с известной надеждой на успех предсказать, что члены одной пары имеют некоторые особенности, отличающие их от членов другой пары.

Характерной чертой метода индивидуальных сравнений является то, что он дает возможность исследователю использовать все имеющиеся у него сведения о поведении подопытного материала при данных условиях. При отсутствии таких сведений о материале или о соответствующих соотношениях внутри него более подходящим будет метод группового сравнения.

Чтобы научиться обходить эти затруднения, начинающему читателю полезно познакомиться с результатами применения неправильного статистического метода к обработке опытных данных. Допустим, что в отношении данных о вирусе табачной мозаики (параграф 9 главы 2) ошибочно применен метод группового сравнения. Результаты в этом случае будут такими:

Препарат вируса	Число половин листа	Число степеней свободы	Среднее число поражений	Сумма квадратов
1	8	7	15	468
2	8	7	11	172
Сумма = 14			$\bar{x}_1 - \bar{x}_2 = 4$	$\Sigma x^2 = 640$

$$s^2 = 45,71; \quad t = 4 \sqrt{(8 \times 7) / 640} = 1,18; \quad P = 0,26 \text{ приблизительно}$$

Исследователь должен был бы прийти к выводу, что нулевая гипотеза остается неотвергнутой, в то время как он обязательно отвергнет ее, если будет основываться на правильных расчетах показателей таблицы 12. Рекомендуется сопоставить это обсуждение ошибочного применения статистических методов с тем, что было сказано в параграфе 10 главы 2; где рассматривались вопросы о назначениях парного метода.

В равной мере неожиданные результаты получаются и при произвольном образовании парных наблюдений в опыте, предназначенном для группового сравнения. Если вы вспомните, что последний способ применим только в том случае, когда исследователь не может предвидеть каких-либо преимуществ попарного размещения исследуемых объектов, то для вас будет очевидной ошибочность расстановки наблюдений в пары по жребию после того, как опыт уже проведен. Хорошей иллюстрацией этого могут служить две выборки, рассмотренные в параграфе 2 этой главы. Выборочная разность 6,3 фунта и средний квадрат $2 \times 160,7 = 321,4$ являются несмещенными оцен-

ками величин $\mu_1 - \mu_2 = 0$ и $\sigma_D^2 = 200$. Далее, в соответствии с тем, что ожидается в 95 % таких сравнений, значение t оказалось меньше 5 %-ного уровня. Однако произведем некоторую перестановку данных в этих двух выборках, соединяя в пары два самых меньших наблюдения, потом следующих по величине и так далее, вплоть до самых больших, т. е. так, как показано в левой части таблицы 30, или, наоборот, комбинируя по парам малые наблюдения с большими, т. е. так, как показано в правой части этой таблицы. В то время как средние выборок и разностей остались без изменения, средние квадраты двух рядов разностей приводят к явной несогласованности, так как ни один из них не будет несмещенной оценкой дисперсии основной совокупности $\sigma^2 = 200$. В первом случае большое значение t будет определено приводить к отказу от гипотезы о том, что разность равна нулю. Во втором же случае величина t недооценивает фактическое выборочное значение $t = 1,11$. Из этих примеров видно, что статистический метод обработки, определяемый схемой эксперимента, не может меняться по произволу.

ТАБЛИЦА 30

Два способа образования парных наблюдений в случайных выборках. Здесь показываются те ошибочные выводы, к которым можно прийти на основе таких парных наблюдений

	Способ I			Способ II		
	выборка 1	выборка 2	разность	выборка 1	выборка 2	разность
	12	11	1	12	53	-41
	24	19	5	24	44	-20
	29	19	10	29	32	-3
	33	24	9	33	31	2
	34	30	4	34	30	4
	36	30	6	36	30	6
	39	31	8	39	24	15
	39	32	7	39	19	20
	53	44	9	53	19	34
	57	53	4	57	11	46
ΣX	356	293	63	356	293	63
\bar{x}	35,6	29,3	$\bar{x}_1 - \bar{x}_2 = 6,3$	35,6	29,3	$\bar{x}_1 - \bar{x}_2 = 6,3$
s^2			8,01			627,34
t			7,04			0,80

Читатель, желающий получить первое, предварительное представление о настоящем курсе, может считать себя подготовленным для непосредственного перехода к главе 6, посвященной регрессии, или к главе 10, излагающей метод дисперсионного анализа. В этом случае оставшаяся часть данной главы может быть опущена.

7. Размер выборки. Методы определения размера выборки в данном случае только немногим отличаются от тех, которые были даны в параграфе 15 главы 2. Основой для оценки σ являются такие данные, которые приведены в таблицах 27 и 29. Так как теперь встает вопрос об оценке и проверке существенности разностей между средними, то в соответствующие формулы необходимо ввести множитель $\sqrt{2}$. В тех случаях, когда нужно определить наибольший из допустимых доверительный интервал или наименьшее значение

средней разности, которую надлежит установить, расчет производится по формуле

$$\delta = \frac{\sqrt{2} \cdot st}{\sqrt{n}},$$

где n — численность каждой из двух групп и число степеней свободы $= 2(n-1)$.

Возьмем для иллюстрации данные примера 4. Здесь исследователи ограничились наблюдением над 8 животными в каждой из двух групп и ставили перед собой задачу определить размер разности, которую они могут считать существенной на 5%-ном уровне и при отношении шансов 1 : 1. В предыдущих опытах они обычно встречались с выборочными стандартными отклонениями от 0,12 до 0,18 фунта в день, поэтому они выбирают $s=0,15$ и приписывают этой величине 60 степеней свободы. По этим данным они определяют

$$\delta = \frac{\sqrt{2} \times 0,15 \times 2}{\sqrt{8}} = 0,15 \text{ фунта в день}$$

в качестве той минимальной разности, которую можно обнаружить при этих условиях. (Отметим, что если оставить в стороне малые выборки, то выбор числа степеней свободы для s не имеет никакого значения, так как всегда $t=2$.) Как оказалось, выборочная разность средних составляла 0,29 фунта в день при 95 %-ном доверительном интервале от 0,13 до 0,45 фунта.

Если бы исследователи пожелали составить план будущего опыта, который давал бы возможность обнаружить разность, составляющую 5 % от общего среднего привеса животных, т. е. разность $0,05 \times 1,985 = 0,1$ фунта в день, то они могли бы на основе данных, взятых из первого своего опыта, где $s=0,148$ фунта в день, определить необходимую для этого численность групп. Они получают уравнение:

$$0,1 = \frac{\sqrt{2} \times 0,148 t_{0,05}}{\sqrt{n}},$$

при решении которого по методу, описанному в параграфе 15 главы 2, получается $n=18$.

Можно поставить вопрос о большей степени доверия, чем 0,5 (т. е. чем соотношение шансов 1 : 1), например, взять 0,75 (т. е. отношение шансов 3 : 1). В данном случае, применяя метод из параграфа 18 главы 10, можно найти, что необходимая численность групп должна составлять 25.

Пример 12. В примере 3 средняя из всех данных о сопротивлении раздавливанию зерен кукурузы равна 45,5 фунта. Признано желательным, чтобы опыт позволял обнаружить фактическую разность в 10% от средней. Сколько наблюдений следует взять для каждой стадии спелости кукурузы, чтобы половина доверительного интервала была бы не больше 4,55 фунта? *Ответ:* 38 зерен.

Пример 13. В примере 1 обобщенный средний квадрат был 0,308 при числе степеней свободы $=18$. Какую в этом опыте можно обнаружить разность средних на 5%-ном уровне существенности? *Ответ:* 0,52%.

Пример 14. В том же примере 1, сколько нужно наблюдений для обнаружения разности $x_1 - x_2 = 1,4\%$ при 5%-ном уровне? *Ответ:* 3 пчелы в группе. Заметим: при повышении шанса обнаружения до 3:1 будет необходимо 5 пчел в группе.

8. Критерий однородности дисперсий. В тех случаях, когда предполагается, что изучаемые в опыте варианты могут изменять дисперсию совокупности, этот вопрос должен быть решен до проведения опыта. Если все дело сводится к выборкам, суммы квадратов которых значительно отличаются друг от друга, то подозрение падает на выборочное варьирование, и в этом случае суммы квадратов должны, как обычно, быть объединены. Однако возможны случаи, когда возникает необходимость произвести проверку того, в какой мере варианты опыта воздействуют на дисперсию. У экспериментатора может возникнуть желание произвести проверку гипотезы $H_0: \sigma_1^2 = \sigma_2^2$, и с этой целью он может планировать специальный эксперимент.

При проверке однородности дисперсии используется новое выборочное распределение величины F , а не распределение χ^2 , о котором была речь в параграфе 14 главы 2. Этот вопрос более подробно будет рассмотрен в параграфе 5 главы 10. В настоящее же время обратимся к таблице 31, которая в более развернутом виде будет дана в главе 10.

ТАБЛИЦА 31

5-%-ный уровень (двухсторонний) распределения F

f_2 = число степеней свободы для меньшего среднего квадрата	f_1 = число степеней свободы для большего среднего квадрата									
	2	4	6	8	10	12	15	20	30	∞
2	39,00	39,25	39,33	39,37	39,40	39,42	39,43	39,45	39,46	39,50
3	16,04	15,10	14,74	14,54	14,42	14,34	14,25	14,17	14,08	13,90
4	10,65	9,60	9,20	8,98	8,84	8,75	8,66	8,56	8,46	8,26
5	8,43	7,39	6,98	6,76	6,62	6,52	6,43	6,33	6,23	6,02
6	7,26	6,23	5,82	5,60	5,46	5,37	5,27	5,17	5,07	4,85
7	6,54	5,52	5,12	4,90	4,76	4,67	4,57	4,47	4,36	4,14
8	6,06	5,05	4,65	4,43	4,30	4,20	4,10	4,00	3,89	3,67
9	5,71	4,72	4,32	4,10	3,96	3,87	3,77	3,67	3,56	3,33
10	5,46	4,47	4,07	3,85	3,72	3,62	3,52	3,42	3,31	3,08
12	5,10	4,12	3,73	3,51	3,37	3,28	3,18	3,07	2,96	2,72
15	4,76	3,80	3,41	3,20	3,06	2,96	2,86	2,76	2,64	2,40
20	4,46	3,51	3,13	2,91	2,77	2,68	2,57	2,46	2,35	2,09
30	4,18	3,25	2,87	2,65	2,51	2,41	2,31	2,20	2,07	1,79
∞	3,69	2,79	2,41	2,19	2,05	1,94	1,83	1,71	1,57	1,00

Чтобы показать, как применять эту таблицу, возьмем данные примера 1. Хотя различие между дисперсиями здесь не вызывает никаких сомнений, все же мы произведем расчеты, необходимые для оценки этого различия между средними квадратами. Для 20 %-ного спреда $s^2=0,027$, а для 65 %-ного — $s^2=0,589$. Каждый из этих средних квадратов имеет 9 степеней свободы. Величина F является отношением большего среднего квадрата к меньшему:

$$F = 0,589 : 0,027 = 22,1.$$

Для проверки нулевой гипотезы $\sigma_1^2 = \sigma_2^2$ следует сравнить данное выборочное значение F с табличным значением 4,03, которое находится на пересечении ряда, соответствующего 9 степеням свободы, и столбца для 9 степеней свободы (последний находится путем интерполяции между столбцами для 8 и 10 степеней свободы). Так как выборочное F значительно превосходит табличное значение для 5 %-ного уровня, гипотеза H_0 , как обычно, будет отвергнута. Раз это так, то нет никакой общей дисперсии σ^2 , которую оценивает объединенный средний квадрат, и результат оценки примера 1 остается под сомнением. Вопрос о том, как в этом случае поступить, будет рассмотрен в следующем параграфе.

Пример 15. Произвести по данным таблицы 28 проверку гипотезы о том, что два рациона, содержащие протеин, не влияют на дисперсию привеса у крыс. *Ответ:* $F=1,08$ по сравнению с табличным $F_{0,05}=5,41$. Этот результат не препятствует объединению сумм квадратов.

Пример 16. По данным примера 10 произвести проверку гипотезы о том, что дисперсия содержания неорганического фосфора на 100 мл сыворотки крови у першеронов и шетлендских лошадей одинакова. *Ответ:* $F=1,52$ по сравнению с табличным $F_{0,05}=2,37$.

9. Способ оценки при $\sigma_1^2 \neq \sigma_2^2$. Описанный в настоящей главе метод оценки нельзя применять в тех случаях, когда есть основания считать, что стандартные отклонения нормальных совокупностей, из которых берутся выборки, неодинаковы. Для проведения соответствующей оценки в этом случае применяется метод Беренса — Фишера, для чего требуются специальные таблицы

[1, 6, 12]. В порядке известного приближения (3) можно использовать более общие таблицы, которые для наших целей достаточно точны.

Случай 1: $\sigma_1 \neq \sigma_2$, но $n_1 = n_2$. Если группы одинакового размера, то вычисления и оценка производятся так же, как в параграфе 3 этой главы, за исключением того, что табличное значение t берется для $n-1$ степеней свободы, а не для $2(n-1)$. Пример: из двух распределений, приведенных в таблицах 15 и 24, я взял выборки таблицы 32. У этих совокупностей средние одинаковы, но стандартные отклонения различны. Если взять правильное число степеней свободы, то нулевая гипотеза относительно равенства средних принимается, как это и должно быть. Но если основываться на обычном числе степеней свободы $2(n-1)=18$, то нулевая гипотеза будет ошибочно отвергнута. Эти специально подобранные выборки (они не были случайными) показывают, что при одинаковом размере выборок применение обычного неправильного при $\sigma_1 \neq \sigma_2$ способа оценки будет приводить к тому, что верная нулевая гипотеза будет отвергаться чаще, чем в 5% случаев.

Случай 2: $\sigma_1 \neq \sigma_2$ и $n_1 \neq n_2$. Если выборки различного размера, то уже нужна некоторая модификация вычислений. Этот случай показан на примере двух выборок, отобранных из тех же двух совокупностей, что и ранее. Как показано в таблице 33, в этом случае суммы квадратов не должны объединяться вместе. Теперь оценка определяется как $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$. Но результирующее отношение $t' = (\bar{x}_1 - \bar{x}_2) / s_{\bar{x}_1 - \bar{x}_2}^2$ уже не распределено по закону t Стьюдента.

ТАБЛИЦА 32

Две выборки одинакового размера из совокупностей с различными стандартными отклонениями; проверка гипотезы $H_0: \mu_1 = \mu_2; n_1 = n_2 = 10$

Выборка 1 при $\sigma = 10$, табл. 15	32	23	48	41	20	29	53	39	30	43
Выборка 2 при $\sigma = 5$, табл. 24	27	30	32	26	31	27	23	29	35	20
Выборка	Размер		Число степеней свободы		Средняя			Сумма квадратов		
1	10		9		35,8			1041,6		
2	10		9		28,0			174,0		
	Сумма=18				$\bar{x}_1 - \bar{x}_2 = 7,8$			$\Sigma x^2 = 1215,6$		

$$t = (\bar{x}_1 - \bar{x}_2) \cdot \sqrt{\frac{n(n-1)}{\Sigma x^2}} = 7,8 \cdot \sqrt{\frac{10 \times 9}{1215,6}} = 2,122; \text{ степеней свободы} = 9; P > 0,05$$

Кокран показал, что t' при уровне 0,05 приближенно можно считать взвешенной средней из двух значений, взятых из обычной таблицы 11; t_1 соответствует $n_1 - 1$ степеням свободы и $t_2 - n_2 - 1$ степеням свободы. За веса следует брать средние квадраты двух выборочных средних s_1^2/n_1 и s_2^2/n_2 . Все расчеты даны в таблице 33.

Так как $t' = 2,316$ превосходит 5%-ный уровень 2,280, то H_0 отвергается; здесь мы имеем дело с 1 шансом из 20 допустить в заключениях ошибку типа 1. Однако если бы мы ошибочно применили метод параграфа 5, то имели бы $\Sigma x^2 = 1062,31 + 173,71 = 1236,02$ и

$$t = 7,66 \sqrt{\frac{91 \times 18}{20 \times 1236,02}} = 1,792,$$

Две выборки неодинакового размера из совокупностей с различными стандартными отклонениями; проверка гипотезы $H_0: \mu_1 = \mu_2$; $n_1 = 13$, $n_2 = 7$

Выборка 1 при $\sigma = 10$	39	42	19	29	49	38	32	39	25	45	39	35	53
Выборка 2 при $\sigma = 5$	30	24	26	33	38	31	28						
Выборка	Размер	Число степеней свободы		$t_{0,05}$	Выборочные средние		s^2	$s^2_{\bar{x}} = s^2/n$					
1	13	12		2,179	37,23		88,53	6,810					
2	7	6		2,447	29,57		28,95	4,136					
		18			$\bar{x}_1 - \bar{x}_2 = 7,66$		$s^2_{\bar{x}_1 - \bar{x}_2} = 10,946$						

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{10,946} = 3,308; \quad t' = (\bar{x}_1 - \bar{x}_2) / s_{\bar{x}_1 - \bar{x}_2} = 7,66 / 3,308 = 2,316$$

$$5\% \text{-ный уровень} = \frac{\sum s^2_{\bar{x}} \cdot t_{0,05}}{\sum s^2_{\bar{x}}} = \frac{6,810 \times 2,179 + 4,136 \times 2,447}{6,810 + 4,136} = 2,280$$

что соответствует $P = 0,07$; гипотеза H_0 не будет отвергнута. Такой противоречивый результат не является типичным; он может получиться, если выборка большего размера имеет большую дисперсию, как в таблице 33. Типичным же является то, что, незаконно применяя обычный критерий t , мы будем отвергать гипотезу H_0 , когда она верна чаще, чем в 5% случаев.

В обоих этих случаях, в которых обычный критерий t приводит к более частому, чем следует, опровержению гипотезы, имеет место смещение критерия в связи с неравенством $\sigma_1 \neq \sigma_2$. Однако эти стандартные отклонения совокупностей редко бывают известны. В связи с этим, если между средними квадратами выборок наблюдается значительное различие, то возникает вопрос о том, как следует поступить в таком случае. Решение этого вопроса затруднительно.

Довольно упрощенный способ решения этого вопроса сводится к правилу, согласно которому в том случае, когда средние квадраты не отличаются друг от друга существенно, применяется обычный критерий t ; в противном же случае применяется критерий Беренса — Фишера или описанный приближенный способ. Но при таком решении вопроса мы будем иметь дело с двумя видами ошибок; первая из них связана с проверкой гипотезы $H_0: \sigma_1 = \sigma_2$, а вторая — с проверкой гипотезы $H_0: \mu_1 = \mu_2$. Влияние на окончательный критерий комбинации этих двух ошибок столь сложно, что задача пока не имеет общего решения. Поэтому применение указанного здесь упрощенного правила является нежелательным.

Мои рекомендации сводятся к следующему. Если большие различия между средними квадратами проявляются со всей очевидностью, то пересмотрите построение опыта с целью найти какое-либо вероятное объяснение этого факта. Может оказаться так, что вы найдете некоторые основания для оправдания того, что стандартные отклонения различны, если даже соответствующие причины в процессе первоначального планирования опыта не были вами учтены. В этом случае вполне уместно использовать критерий Беренса — Фишера или приближенный критерий Кокрана. Если же никаких причин для различия между стандартными отклонениями не будет найдено, то остается применение обычного критерия t , но с большей осторожностью, при формулировке решений относительно средних, так как выборочные оценки σ^2 в этом случае сомнительны.

Как же быть в отношении опыта с пчелами из примера 1 и параграфа 8? Если бы 20%-ный сироп был схож по своей концентрации с обычно встре-

чающимся нектаром, то можно было бы считать, что пчелы берут его с меньшей вариацией, потому что для них это привычное дело; но фактически они привыкли к более высокой концентрации; причиной различия в варьировании может быть и то, что при более низкой концентрации изменения меньше, т. е. небольшое варьирование связано с небольшим значением средней. В этом случае нет никакой необходимости искать критерий для чего-либо другого, кроме этих средних. Доктор Парк не нашел никакого рационального объяснения этого различия и поэтому применил обычный критерий. Конечно, в данном опыте различие между средними столь велико, что гипотеза $H_0: \mu_1 = \mu_2$ будет отвергнута при любом критерии, каким бы он ни был.

Прежде чем никогда не удастся выбрать подходящий для данных результатов метод оценки, если опусываться только на этих результатах. Правильность метода зависит от гипотезы относительно параметра σ .

Пример 17. Янг [13], изучая основной обмен у 26 студенток, получил для двух групп $n_1=15$ и $n_2=11$, $\bar{x}_1=34,45$ и $\bar{x}_2=33,57$ кал. на 1 кв. м в час; $\Sigma x_1^2=69,36$; $\Sigma x_2^2=13,66$. Проверьте гипотезу $H_0: \sigma_1 = \sigma_2$. Ответ: $F=3,62$ по сравнению с табличным $F_{0,05}=3,55$.

Общий обмен у 26 студенток
(калорий на 1 кв. м в час)

7 или более часов сна				6 или менее часов сна			
1	35,3	9	33,3	1	32,5	7	34,6
2	35,9	10	33,6	2	34,0	8	33,5
3	37,2	11	37,9	3	34,4	9	33,6
4	33,0	12	35,6	4	31,8	10	31,5
5	31,9	13	29,0	5	35,0	11	33,8
6	33,7	14	33,7	6	34,6		
7	36,0	15	35,7				
8	35,0						
$\Sigma X_1=516,8$				$\Sigma X_2=369,3$			
$\bar{x}_1=34,45$ кал. на 1 кв. м в час				$\bar{x}_2=33,57$ кал. на 1 кв. м в час			

Пример 18. В опыте, приведенном в предыдущем примере, различие между групповыми средними небольшое, и поэтому представляется невозможным, чтобы различие между стандартными отклонениями было обусловлено столь незначительным различием между вариантами опыта. Более правдоподобным объяснением будет допущение, что в большей выборке случайно появились некоторые слишком нетипичные показатели — 29,0; 37,2; 37,9, которые привели к тому, что s^2 дает завышенную оценку σ^2 . Однако, несмотря на это, допустите, что для гипотезы $\sigma_1 \neq \sigma_2$ было найдено некоторое рациональное объяснение, и проверьте гипотезу $H_0: \mu_1 = \mu_2$. Ответ: $t'=1,31$; $t_{0,05}=2,17$. Сравните это с показателями $t=1,19$ и $t_{0,05}=2,048$, которые получаются при гипотезе $\sigma_1 = \sigma_2$. Таким образом, эта гипотеза в отношении стандартных отклонений не вносит каких-либо изменений в наши выводы.

10. Статистика и эксперимент. В процессе изложения материала будет все больше и больше выясняться, что статистик и экспериментатор (которые часто соединены в одном лице) для достижения определенного успеха в исследовании должны образовать между собой тесное сотрудничество. Экспериментатор определяет условия, в которых должен проводиться опыт, подопытный материал, его происхождение, изучаемые варианты опыта и различные сопутствующие условия, а также те признаки, которые подлежат измерению. Статистик выбирает или создает план эксперимента (схему опыта), который позволит избежать смещенных или смешанных с другими оценкам при сохранении надлежащей точности их. Экспериментатор имеет непосредственное отношение к лабораторным или полевым работам, прилагая все усилия к тому, чтобы по возможности исключить всякого рода грубые ошибки.

Если он в этом деле добьется успеха, то полученные в опыте данные будут содержать в себе те сведения, ради которых был поставлен эксперимент. Статистик применяет соответствующие методы для извлечения всех сведений, заключенных в этих данных. Наконец, экспериментатор интерпретирует полученные таким образом сведения в свете существующих в данной науке положений.

Иногда наблюдается непонимание того, что статистические методы способны выявлять только ту информацию, которая содержится в опытных данных благодаря тщательному проведению опыта и правильному использованию выборочного метода. Статистическая обработка никак не исключает, а наоборот требует тщательности экспериментальных работ. Многие выводы будут бесполезными, если они появились в результате обработки небрежно собранных данных. Равным образом часто упускают из виду, что если план опыта имеет серьезные недостатки, то даже многочисленные и добросовестно полученные данные могут содержать в себе такие сведения, которые совсем не оправдывают затраченного на них времени. Только при одновременном наличии надлежащего плана опыта, умелого проведения его и подходящих статистических методов обработки опытных данных исследователь может быть уверен в надежности результатов, на которых он строит свои выводы.

Л И Т Е Р А Т У Р А

1. Behrens W. U., *Landwirtschaftliche Jahrbücher*, 68, 807, 1929.
2. Breneman W. R., *Личное сообщение*.
3. Cochran W. G., Cox Gertrude M., *Experimental Designs*. John Wiley and Sons, Inc., New York, 1950.
4. Culbertson C. C., Hammond W. E., *Iowa Agricultural Experiment Station Animal Husbandry Leaflet*, 144, 1933.
5. Dean Harold L., Walker R. H., *Journal of the American Society of Agronomy*, 27, 433, 1935.
6. Fisher R. A., *Annals of Eugenics*, 11, 141, 1941.
7. Park O. W., *Iowa Agricultural Experiment Station Research Bulletin*, 151, 1932.
8. Pearson P. B., Catchpole H. R., *The American Journal of Physiology*, 115, 90, 1936.
9. Smith F. B., Brown P. E., *Soil Science*, 35, 413, 1933.
10. Smith Stewart N., *Journal of the American Society of Agronomy*, 26, 792, 1934.
11. Swanson Pearl P., Smith Arthur H., *The Journal of Biological Chemistry*, 97, 745, 1932.
12. Sukhatme P. V., *Sankhya*, 4, 39, 1938.
13. Young Charlotte M., *Thesis submitted for the degree, Ph. D., Iowa State College*, 1940.

СОКРАЩЕННЫЕ И ПРИБЛИЖЕННЫЕ СПОСОБЫ. МЕТОДЫ НЕПОЛНОЙ ЭФФЕКТИВНОСТИ И НЕПАРАМЕТРИЧЕСКИЕ

1. Введение. При изложении предыдущих глав были приняты меры, чтобы необходимые расчеты не были слишком сложны. Мы брали, например, слегка измененные данные, чтобы средние получались в виде целых чисел. Далее, в параграфах 11 и 12 главы 2, были описаны некоторые способы сокращенных вычислений и вычислений на счетной машине. Но имеются и другие способы, уменьшающие сложность расчетов. Некоторые из них будут рассмотрены в настоящей главе, другие же будут описаны в последующих главах.

В последние годы большое внимание было обращено на разработку быстрых и легких способов расчета, относящихся к выборкам из нормальных совокупностей. Было установлено, что при малых выборках размах варьирования, рассматриваемый как замена выборочного стандартного отклонения, обладает удивительно большой эффективностью по отношению к s . Грубо говоря, размах, определенный по 10 наблюдениям, дает примерно столько же сведений о варьировании, сколько и выборочное стандартное отклонение, вычисленное по 9 наблюдениям. Поэтому, если стоимость такого дополнительного наблюдения меньше стоимости расчетов s , то размах будет вполне оправданной заменой s . Но даже если наблюдения более дороги, то быстрая предварительная оценка при помощи размаха может иногда указать на то, что вычислять s нет никакой необходимости, так как выводы становятся очевидными. В настоящей главе будет дано применение этого метода к обработке данных двух выборок при парных и непарных сравнениях.

В связи с этим нормальное распределение здесь рассматривается в качестве источника, из которого получают большинство наших выборочных наблюдений. Описываемые здесь статистические методы при условии нормального распределения обладают полной эффективностью. К счастью, эти же методы сохраняют высокую эффективность и в случае выборок из распределений, умеренно отклоняющихся от нормальной формы. Но это еще не все. По мере увеличения размера выборок эти же методы приближаются к полной эффективности даже в случае выборок из любых совокупностей, имеющих конечное значение стандартного отклонения. Во многих случаях это приближение к полной эффективности происходит столь быстро, что уже в выборках при $n=3$ или $n=4$ оно становится достаточно близким, и только при чрезмерных отклонениях от нормальной формы требуются выборки в 100 и более наблюдений. Вместе с тем при малых выборках из совокупностей с резко выраженными отклонениями от нормальности применимы некоторые непараметрические методы. Наиболее известным из них является метод оценки, основанный на знаках плюс и минус, но имеются и другие удобные методы, основанные на рангах. Об этом будет сказано в последней части настоящей главы и в последующих главах. Вместе с тем будет показано, что непараметрические методы могут быть также использованы для

«быстрой прикидки» в отношении выборочных данных из нормальных совокупностей.

2. **Линейное преобразование, или кодирование.** В параграфе 12 главы 2 было установлено, что вычисления на счетной машине позволяют избежать расчетов отклонений от \bar{x} или G . Но даже при применении счетной машины иногда полезно брать отклонения от некоторого произвольного начала, которое легко вычисляется в уме. Например, возьмем вес 10 плодов томатов: 206, 217, 224, 227, 228, 231, 236, 241, 245, 258 г. Из этих данных легко вычесть 200 г и, не выписывая ряд полученных результатов, прямо на машине способом нарастающего итога получить для этих отклонений $\Sigma X = 6 + 17 + \dots + 58 = 313$ и $\Sigma X^2 = 6^2 + \dots + 58^2 = 11\,741$. В этом случае средняя будет $200 + 313/10 = 231,3$ г и сумма квадратов $11\,741 - 313^2/10 = 1944,1$. Заметим, что для перехода к действительному значению средней необходимо к среднему отклонению прибавить число, которое ранее было вычтено, т. е. 200. Для суммы квадратов такая поправка не нужна, так как при вычитании произвольного начала произошло простое передвижение всего ряда наблюдений вниз по шкале, по которой отложены измерения, причем при этом движении вниз не происходит никакого изменения размаха варьирования или стандартного отклонения. Этот способ заслуживает внимания, так как позволяет уменьшать размер чисел, с которыми приходится работать, при очень небольшой вероятности допустить ошибку при вычитании.

Здесь показан один из простых видов кодирования. Числа кода получаются в результате линейных преобразований наблюдаемых данных; здесь все операции ограничиваются действиями сложения и умножения и их обратными действиями вычитания и деления.

Несколько более усложненный пример построения кода показан в таблице 34. Для наших целей приведенные в таблице данные о концентрации ионов водорода (рН) пока не нужны. Для данных второго столбца код составлен следующим образом: 1) умножением на 10 000 и 2) вычитанием 200. Расчет \bar{x} и s произведен в условных единицах кода. Поэтому средняя сначала должна быть увеличена на 200, а потом разделена на 10 000, т. е. здесь производятся обратные действия и в обратном к операции кодирования порядке. Вы уже знаете, что на s не влияет вычитание 200, но, подобно тому, как это было в случае средней, здесь для компенсации операции кодирования необходимо умножение на 10 000. Построение кода путем умножения или деления может рассматриваться как изменение единицы измерения. Если,

ТАБЛИЦА 34

Применение чисел кода при вычислении средней и стандартного отклонения.
Цианогенный азот из древесины персикового дерева [2]

Начальное значение рН	Количество граммов на 50 г древесины X'	Числа кода $X = 10\,000X' - 200$
5,1	0,0238	38
5,0	0,0238	38
5,3	0,0241	41
5,3	0,0241	41
5,7	0,0250	50
5,7	0,0247	47
6,0	0,0241	41
6,1	0,0238	38
6,6	0,0226	26
6,5	0,0232	32

$n=10$
 $\Sigma X=392$
 $\bar{X}=39,2$ единицы кода
 $=0,02392$ г

$\Sigma X^2=15784$
 $(\Sigma X)^2/n=15366,4$
 $\Sigma x^2=417,6$
 $s^2=46,4$

$s=6,81$ единицы кода
 $=0,000681$ г

как в нашем случае, мы получаем большое облегчение вычислений при условной единице, равной 0,0001 г, то безусловно следует пойти на крайне незначительный риск ввести какую-нибудь ошибку в расчете в связи с таким изменением единицы измерения. В конце вычислений переход к первоначальной единице в данном случае производится путем простого переноса запятой, отделяющей десятичные знаки.

3. Округление и кодирование. Одна из возможностей применения кода возникает в условиях, в которых довольно часто оказываются исследователи. Эти последние как в целях предосторожности, так и для удобства стремятся выражать собираемые ими данные в единицах, соответствующих их практическому смыслу, но не всегда удобных для статистической обработки. В качестве примера возьмем данные обследования 145 фермерских семей штата Айова, собранные для решения вопроса о питании этих семей [9]. Фактически данные о годовом потреблении мяса были выражены в фунтах; так, для первых 20 семей мы имеем: 726, 296, 928, 668, 287, 1206, 517, 1638, 2414, 610, 494, 2489, 1198, 676, 1302, 440, 1247, 1053, 1029, 218 фунтов. Размах варьирования здесь составляет $2489 - 218 = 2271$ единица. Для большинства статистических исследований обычно бывает вполне достаточным иметь размах в 20—40 единиц. Исходя из этого, можно заключить, что в данном случае более удобной единицей для расчета будет 100 фунтов; соответствующие числа кода получаются делением всех данных на 100 и округлением до единицы: 7, 3, 9, 7, 3, 12, 5, 16, 24, 6, 5, 25, 12, 7, 13, 4, 12, 11, 10 и 2 ц. Характеристиками этой кодированной выборки являются $\bar{x} = 9,65$ и $s = 6,36$ ц. Конечно, эти данные можно умножить на 100 и перейти к первоначальным единицам измерения, что даст $\bar{x} = 965$ фунтам и $s = 636$ фунтам.

При округлении такие числа, как 13,51, увеличиваются до 14, а такие, как 13,49, уменьшаются до 13. Но куда отнести, если оно будет встречено, такое пограничное число, как 13,50? Чтобы обеспечить примерно одинаковое влияние преувеличений и преуменьшений, возникающих при округлении, рекомендуется взять за правило пограничные числа округлять до четных чисел. Так, 13,50 и 14,50 будут округляться до 14.

Вы немедленно пожелаете узнать, не будет ли при таком округлении потеряна какая-нибудь ценная информация. Показатели первоначальной выборки таковы: $\bar{x} = 972$ и $s = 637$ фунтам. Не вызывает ли это различие какое-либо беспокойство? Посмотрите на стандартную ошибку средней $\bar{s}_x = 637/\sqrt{20} = 142$ фунтам.

О применении кодирования и округлений при построении и вычислении распределений численностей см. в параграфе 2 главы 8.

Пример 1. Сухой вес отрезков стебля, взятых с 5 растений мятлика [5], был таким: 1,2; 0,9; 1,0; 1,4; 0,9 г. После вычисления $\bar{x} = 1,08$ и $s = 0,217$ г умножьте мысленно каждое наблюдение на 10, т. е. откиньте запятую у первоначальных данных. Точные результаты получаются от деления показателей \bar{x} и s , исчисленных по коду, на 10.

Пример 2. Какое название вы дадите величинам $\sum X^2$ и $\sum(X-G)^2$? Вы не можете применить термин «сумма квадратов», так как он по условию предназначен для $\sum x^2$. Вы должны говорить: «сумма квадратов наблюдений» и «сумма квадратов отклонений от G ».

Ниже приводятся 3 примера для тех читателей, кто интересуется алгебраическими соотношениями, встречающимися в предыдущих и последующих параграфах.

Пример 3. Если $Y = A + X$, где A — константа, то докажите, что $\bar{y} = A + \bar{x}$ и $\Sigma y^2 = \Sigma x^2$.

Пример 4. Если $Y = bX$ (b — другая константа), то докажите, что $\bar{y} = b\bar{x}$ и $\Sigma y^2 = b^2 \Sigma x^2$. Вы не должны упускать из виду, что b может быть и дробным числом.

Пример 5. Если $Y = A + bX$, то докажите, что $\bar{y} = A + b\bar{x}$ и $\Sigma y^2 = b^2 \Sigma x^2$.

4. Правила и предосторожности в отношении использования чисел кода. Цель введения кода состоит в повышении точности расчетов и в экономии труда. Конечно, этих преимуществ достаточно, чтобы пренебречь необходимостью затрат времени на проведение кодирования и риском допустить какую-нибудь ошибку.

Каждое наблюдение ряда должно быть подвергнуто одной и той же операции. Если вы станете прибавлять к некоторым данным одно число, а к другим — другое, то такие результаты не будут иметь никакого смысла.

Средняя зависит от каждой операции кодирования. Например, если из первоначальных данных при введении кода было сначала вычтено 100, а потом результат разделен на 10, то средняя кода должна сначала умножаться на 10, а потом к ней должно быть прибавлено 100, т. е. обратные действия должны применяться в обратном порядке.

Стандартное отклонение зависит только от действий умножения и деления. Следует ясно себе представлять, что действия сложения и вычитания не влияют на показатели варьирования, какими являются размах и s , так как такое кодирование просто переносит начало отсчета, без какого-либо сжатия или расширения единицы измерения. С другой стороны, умножение и деление меняют единицу измерения, что требует компенсации при помощи применения к сводным показателям кода обратных действий.

Точность окончательных результатов не зависит от примененных действий сложения, вычитания, умножения и деления, но на нее влияют округления и опускаемые цифры.

Пример 6. В примере 33 главы 6 приведены данные процентного содержания воды в ветвях яблони и соответствующие показатели удельной теплоемкости. Если вы умеете вычислять на счетной машине, то определите \bar{x} и s для каждого из этих признаков. Обратите особое внимание на то, что вы не имеете права округлять последние цифры как в том, так и в другом столбце. Если вы это сделаете, то размах округленных данных, например во втором столбце, будет только $65 - 49 = 16$, что меньше требуемого. Но, конечно, если вы желаете, то можете в процентных данных отбросить запятую, определяющую десятичный знак, и отнять после этого 400. Однако при наличии счетной машины остается сомнительным выигрыш во времени, если вы не имеете достаточных навыков проведения всех операций по кодированию непосредственно на машине и без переписки преобразованных данных. Проверьте это.

5. Значение цифры. В параграфе 3 было указано, что желательно, чтобы размах варьирования составлял 20—40 единиц. Это положение относится только к числам кода, которые обычно выражаются в целых числах, а не в десятичных дробях. Более общее положение состоит в том, что размах варьирования, выраженный в *значащих цифрах*, должен быть между 20 и 40. Значение цифры (этот термин не имеет никакого отношения к тем случайным цифрам, с которыми мы имели дело в параграфе 13 главы 1, параграфе 9 главы 2 и параграфе 3 главы 4) не зависит от запятой, отделяющей десятичные знаки; они будут одинаковыми у чисел 647; 0.0647 и 6,47. Если размах варьирования одного ряда будет $0,067 - 0,031 = 0,036$ г, а другого — $23,1 - 19,5 = 3,6$ фунта, то размах, выраженный в значащих цифрах, будет одним и тем же, т. е. 36. Когда для расчетов оставлено определенное число цифр, то запятую, отделяющую десятичные знаки, можно опустить и принять ее во внимание только во время выписки результатов в принятую схему записей.

«Сколько десятичных знаков необходимо взять?» — такой вопрос можно часто слышать. В такой форме на него нельзя дать определенного ответа. Когда же вопрос ставится в другой форме: «Сколько следует взять значащих цифр?», то ответ возможен, хотя он будет в различных случаях разным. Различие между этими вопросами станет ясным из такого примера. Взвешивание на весах непосредственно производится в пределах одной пятой грамма, но глазомерно доводится до одной десятой. Пусть вес равен 25,3 г. Но этот результат может быть записан и как 25 300 мг и как 0,0253 кг. При всех трех формах записи значащими цифрами будут 253; они совсем не зависят от принятой единицы измерения. Однако даже в том случае, когда указанный выше вопрос относится исключительно к значащим цифрам, имеется целый ряд соображений, определяющих необходимое число цифр; по-видимому, нет никакой возможности подвести этот выбор под какие-либо общие правила. В этой работе мы везде старались данные, с которыми приходится иметь дело, представить в наиболее удобной форме. В особых случаях даются специальные объяснения. В данном параграфе мы попытаемся вкратце коснуться того, что

заставит работать вашу мысль в этом направлении и позволит приобрести некоторый опыт в расчетах.

Допустим, что наблюдение 25,3 г—это сухой вес некоторого кормового растения, выращенного в вегетационном сосуде; в этом случае может возникнуть необходимость произвести пересчет на 1 акр. Переводной коэффициент, положим, равен 94,327. Сколько значащих цифр следует взять у произведения? Учитывая, что 25,3 г фактически является заменой какого-то веса между 25,25 и 25,35, мы заключаем, что вес в фунтах на 1 акр находится где-то между:

$$25,25 \times 94,327 = 2381,8 \text{ фунта на 1 акр}$$

и

$$25,35 \times 94,327 = 2391,2 \text{ фунта на 1 акр.}$$

Средняя из этих двух произведений равна $23,5 \times 94,327 = 2386,5$. Очевидны следующие три, располагающиеся в некоторой последовательности положения: 1) даже третья значащая цифра может быть неправильной, но 2) четыре значащие цифры 2386 уже дают вполне удовлетворительную оценку произведения и 3) множитель, имеющий меньшее число знаков, определяет точность произведения. Легко проверить, что произведение двух трехзначных чисел будет менее точным, чем установленный выше результат, и что четвертая цифра все еще имеет значение для среднего произведения. Отсюда вытекает правило: в произведении следует брать на одну цифру больше, чем их имеется у меньшего из множителей.

На основе небольшого опыта вы можете установить, что многое из того, что сказано в отношении произведения, относится и к частному. Но рассмотрим совместное применение вычитания и деления на примере процентного уменьшения содержания амидного азота в растении капусты в процессе дистилляции с окисью кальция [8]. Перед дистилляцией было 14,3 мг, после дистилляции—10,8 мг; следовательно, уменьшение составляло 24,5%. Крайние значения этого уменьшения находятся в пределах:

$$14,35 - 10,75 = 3,60 \text{ мг}$$

и

$$14,25 - 10,85 = 3,40 \text{ мг.}$$

Поэтому процент уменьшения азота будет между

$$3,60 : 14,35 = 25,1\%$$

и

$$3,40 : 14,25 = 23,9\%.$$

Таким образом, здесь погрешность вкралась даже во вторую значащую цифру, и все цифры за третьей не будут иметь никакого смысла.

При вычислении суммы некоторого числа данных, подобных нашему весу в 25,3 г, в связи с тем, что некоторые из таких данных будут завышенными, а другие заниженными, возникает необходимость исключения этих погрешностей. Поэтому среднюю ради осторожности следует определять числом знаков, на один или на два большим, чем число знаков у исходных данных.

Установленные здесь правила относятся к указанным выше расчетам, но они теряют свое значение, как только речь идет о целой серии вычислений, как, например, в случае расчета стандартной ошибки средней. Постоянное выполнение этих правил приведет к накоплению погрешностей. В этих случаях из предосторожности следует брать две или три лишние цифры, в особенности тогда, когда имеется под руками счетная машина. После этого окончательные данные могут быть сокращены до требуемого числа значащих цифр.

Но что такое требуемое число значащих цифр? Рассмотренные выше ошибки измерения становятся относительно малозначащими, когда они сравниваются с выборочным варьированием. Например, средняя большой выборки 55,957 мг определена с точностью до пяти значащих цифр, а ее стандартная

ошибка равна 1,754 мг. Но так как половина 95%-ного доверительного интервала ($t = \infty$) равна $1,960 \times 1,754 = 3,4$ мг, то для определения этой оценки достаточно указать пределы $56,0 \pm 3,4$ мг. Однако и в этом случае возможно некоторое усложнение вопроса. Если мы намерены по этим показателям воспроизвести выборку, на основе которой они определены (как это было, положим, в примере 11 главы 4), то в этом случае выгодно иметь у показателя s_x^2 число значащих цифр по крайней мере такое же, как и у средней \bar{x} ; но это число знаков отчасти зависит и от размера выборки (так как s_x^2 получается делением на \sqrt{n}).

Теперь вам ясно, почему нельзя руководствоваться только каким-то определенным набором правил в отношении значащих цифр. В случае простых произведений или отношений величин вопрос о том, как привести результаты в соответствие с точностью первоначальных данных, ясен. Если же вы производите более или менее длинный ряд последовательных расчетов, то для первого раза возьмите несколько большее число знаков, чем то, которое вы намерены сохранить в окончательном виде. При оформлении ваших заключительных результатов ограничьте числа знаков у своих данных в соответствии с их точностью и назначением.

Пример 7. В одной газете, датированной 2 ноября 1952 г., выставлено требование для мирового рекорда по плаванию под водой — опуститься без кислородной маски на глубину 127,957 фута в бухте Неаполя. Что вы скажете по поводу последних тысячных долей фута в этой величине?

6. Выводы, основанные на размахе варьирования в выборках из нормальной совокупностей. В параграфе 6 главы 2 указывалось, что, несмотря на то, что размах варьирования позволяет весьма просто получить оценку σ , все же этот показатель иногда бывает малоэффективным по сравнению с s . В особенности в выборках при $n > 10$. Но в случае интервальных оценок эффективность размаха при $n > 20$ устанавливается уже на уровне 90% [10], поэтому размах варьирования приобретает широкое значение в качестве сокращенного способа оценки в тех случаях, когда имеют дело с малыми выборками из нормальной совокупности.

Для удобства построения интервальных оценок и проверки гипотез Лордом [6] составлены таблицы, аналогичные таблицам значений t . В левой части таблицы 35 дана таблица для единичных выборок, подобная той, которая была приведена в главе 2. В ней приведены величины $\frac{x - \mu}{\omega}$, где ω обозначает размах варьирования в выборке. Это отношение играет роль величины t в параграфе 7 главы 2, и поэтому я называю ее t_w .

Чтобы показать, как на основе таблицы 35 строятся доверительные интервалы, вернемся к данным о содержании витамина С, приведенным в таблице 9. Здесь $\bar{x} = 20$, $\omega = 29 - 13 = 16$ мг/100 г и $n = 17$. По новой таблице в колонке, озаглавленной 0,05, и по строке $n = 17$ находим число 0,144. Это означает, что в случайных выборках из нормальной совокупности вероятность $t_w \leq 0,144$ равна 0,95. В соответствии с параграфом 8 главы 2 95%-ный доверительный интервал определяется неравенством

$$\bar{x} - t_w \cdot \omega \leq \mu \leq \bar{x} + t_w \cdot \omega.$$

Подставляя наши данные, получим:

$$20 - 0,144 \times 16 \leq \mu \leq 20 + 0,144 \times 16$$

$$17,7 \leq \mu \leq 22,3 \text{ мг/100 г.}$$

Сравним этот результат с интервалом, основанным на s :

$$17,95 \leq \mu \leq 22,05.$$

Проверку гипотезы при помощи t_w покажем на примере данных о вилусе мозаики табака, приведенных в таблице 12. Здесь $\bar{d} = 4$ и $\omega = 14$ местам

поражений, в то время как $n=8$. Если H_0 состоит в том, что $\mu=0$, то имеем

$$\frac{4-0}{14} = t_w = 0,286,$$

что практически находится на 5%-ном уровне. Вывод будет тот же, что и ранее. В тех случаях, когда t_w находится вблизи от критического уровня или когда необходим более строгий критерий, следует пользоваться более эффективным показателем t .

ТАБЛИЦА 35

Два уровня величин, аналогичных t , для построения выводов относительно выборок из нормальной совокупности на основе размаха варьирования¹

Двусторонняя оценка

Единичные выборки. Значения t_w			Две выборки одинакового размера. Значения t_w		
размер выборки	вероятность		размер каждой выборки	вероятность	
	0,05	0,01		0,05	0,01
2	6,353	31,828	2	3,427	7,916
3	1,304	3,008	3	1,272	2,093
4	0,717	1,316	4	0,813	1,237
5	0,507	0,843	5	0,613	0,896
6	0,399	0,628	6	0,499	0,714
7	0,333	0,507	7	0,426	0,600
8	0,288	0,429	8	0,373	0,521
9	0,255	0,374	9	0,334	0,464
10	0,230	0,333	10	0,304	0,419
11	0,210	0,302	11	0,280	0,384
12	0,194	0,277	12	0,260	0,355
13	0,181	0,256	13	0,243	0,331
14	0,170	0,239	14	0,228	0,311
15	0,160	0,224	15	0,216	0,293
16	0,151	0,212	16	0,205	0,278
17	0,144	0,201	17	0,195	0,264
18	0,137	0,191	18	0,187	0,252
19	0,131	0,182	19	0,179	0,242
20	0,126	0,175	20	0,172	0,232

¹ Данные взяты из более полной таблицы с разрешения Е. Лорда и редакции «Биометрики».

Пример 8. В примере 3 главы 2 для урожая люцерны мы имели $\bar{x}=1,70$ и $w=1,4$ т на 1 акр при $n=10$. Постройте 95%-ный доверительный интервал. *Ответ:* от 1,38 до 2,02 т на 1 акр. Сравните этот результат с данными примера 26 главы 2.

Пример 9. Используя показатели примера 26 главы 2, постройте 99%-ный доверительный интервал при помощи w и данных примеров 5 и 12 главы 2. *Ответ:* $9 \pm 16,9$ бушеля на 1 акр. Учитывая, что этот доверительный интервал содержит в себе нуль, мы не можем считать нулевую гипотезу отвергнутой.

Теперь мы перейдем к применению размаха варьирования при обработке данных таких опытов, с которыми мы имели дело в главе 4. Для оценки существенности разности между средними выборок одинакового размера по n наблюдений в каждой применяется правая часть таблицы 35. В этом случае вместо w берется средний размах $\bar{w}=(w_1+w_2)/2$ и d заменяется на $\bar{x}_1-\bar{x}_2$.

В качестве примера возьмем опять данные опыта с петушками, приведенные в таблице 26. Здесь $n=11$, $\bar{x}_1-\bar{x}_2=41$ мг, $w_1=137-53=84$ мг, $w^2=96-22=74$ мг и $\bar{w}=(84+74):2=79$ мг. Отсюда $t_w=(\bar{x}_1-\bar{x}_2)/\bar{w}=41/79=0,519$, что превосходит табличное 0,384 для уровня существенности 0,01. Когда, как в данном случае, заключение является вполне определенным, нет никакой необходимости обращаться к точному критерию t .

Если интерес сосредоточивается на интервальной оценке разности между средними, то в первую очередь может быть произведен расчет доверительного интервала, причем он может служить заменой расчета для критерия существенности. В случае опыта с петушками доверительный интервал при вероятности 0,99 будет:

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 - t_{\bar{w}} \cdot \bar{w} &\leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\bar{w}} \cdot \bar{w} \\ 41 - 0,384 \times 79 &\leq \mu_1 - \mu_2 \leq 41 + 0,384 \times 79 \\ 10,7 &\leq \mu_1 - \mu_2 \leq 51,3 \text{ мг.} \end{aligned}$$

Так как этот 99%-ный доверительный интервал не содержит в себе нуль, то отсюда следует, что гипотеза $H_0: \mu_1 - \mu_2 = 0$ отвергнута на 1%-ном уровне доверия.

Размах варьирования выборок из нормальных совокупностей является весьма удобной заменой показателя s , если можно допустить потерю 5—10% информации о варьировании, содержащейся в выборке. Но нормальность распределения не принадлежит к числу обязательных условий, в которых применяется выборочный метод. Известно, что на выводы, основанные на критерии t , не оказывают сколько-нибудь значительного влияния умеренные отклонения от нормальности распределения и даже большие отклонения при больших значениях n . Но выводы, основанные на размахе варьирования, более чувствительны к отклонениям от нормальности, и в особенности к скошенности распределения. Поэтому в тех случаях, когда есть основание ожидать скошенности распределения, не следует возлагать больших надежд на w как заместителя s .

Эффективность размаха варьирования убывает по мере возрастания n . В некоторых приложениях наибольшей эффективностью обладает w при $n=8$. При n от 12 до 22 рекомендуется разделить выборку случайным порядком на две равные части и воспользоваться для критериев и оценок средней из соответствующих двух размахов. Соответствующие таблицы можно найти в работе Лорда.

В случае двух выборок разного размера применение размаха варьирования оказывается столь же трудоемким, как и вычисление Σx^2 .

Пример 10. Примените критерий, основанный на размахе, к данным определения рН в примере 2 главы 4. *Ответ:* $t_{\bar{w}}=0,792$, $P > 0,05$ по сравнению с $t=2,66$, $P=0,04$.

7. Непараметрические методы: медиана и квартили. В тех случаях, когда совершенно очевидно, что совокупность, из которой взята выборка, столь сильно отклоняется от нормальной формы, что методы, разработанные для нормальной совокупности, теряют всякую основу, могут применяться различные *непараметрические* методы [11]. Они не зависят от формы распределения совокупности и не требуют предварительного знания функции плотности вероятности, подобной функции для нормального распределения. В данном и следующих параграфах рассмотрены некоторые из этих методов.

Наиболее известной непараметрической оценкой значения является *медиана*. Она представляет собой срединное наблюдение в ранжированном ряду данных; если n нечетное число, то медианой будет наблюдение за номером $(n+1)/2$; если же n — четное число, то медианой будет средняя из наблюдений за номерами $n/2$ и $(n+2)/2$. Так, в примере 2 главы 2 медианой веса 11 мужчин будет вес мужчины за номером $(11+1)/2=6$, что в ранжированном ряду дает 162 фунта. В примере 6 главы 2 медиана для ряда из 16 разностей равна средней из разностей за номерами $n/2=8$ и $(n+2)/2=9$, т. е. она равна $(3,7+4,0)/2=3,85$ бушеля на 1 акр.

Вместе с медианой в выборках из совокупностей, отклоняющихся от нормальной формы, часто применяются и другие оценки значения. Например, после того как установлено местоположение медианы, каждая из двух половин выборки снова может быть подразделена на две равные части при помощи *квартилей*. Меньшая квартиль, обычно называемая первой квартилью, является наблюдением, которое разделяет четвертую часть данных,

меньших этой квартили, от остальных трех четвертей более высоких наблюдений. Первая квартиль — это наблюдение за номером $(n+1)/4$. Если взять ряд разностей урожаев кукурузы (пример 6 главы 2), то этой квартилью будет разность за номером $(16+1)/4=4,25$. В таких случаях чаще всего вполне допустимо считать, что первой квартилью будет четвертая разность — 4,0 бушеля на 1 акр. Если же необходимо более точное определение первой квартили, то следует взять взвешенную среднюю из четвертого и пятого наблюдений, приняв за веса соответственно 3 и 1:

$$Q_1 = \frac{3 \times (-4) + 1 \times 0}{3+1} = -3 \text{ бушелям на 1 акр.}$$

Здесь Q_1 обозначает нижнюю квартиль. Подобно этому, верхней квартилью будет разность за номером 3 $(n+1)/4=12,75$. Тринадцатая разность равна 7,1 бушеля на 1 акр. Более точное значение будет

$$Q_3 = \frac{1 \times (6,4) + 3 \times (7,1)}{1+3} = 6,92 \text{ бушеля на 1 акр.}$$

Вам ясно, что медиана может быть названа второй квартилью Q_2 .

При наличии большой выборки или конечной совокупности этот процесс дробления ряда данных может быть продолжен: *децилы* разделяют ряд наблюдений на 10 равных частей, *проценти* — на 100 и *промиле* — на 1000.

Квартили могут дать оценку варьирования. *Междуквартильный интервал* $Q_3 - Q_1$ включает в себя половину всех наблюдаемых значений. В результате этого можно дать некоторое приближенное описание выборки (или совокупности), указав, например, что медиана разностей урожаев кукурузы равна 3,85 бушеля на 1 акр при междуквартильном размахе:

$$Q_3 - Q_1 = 6,92 - (-3) = 9,92 \text{ бушеля на 1 акр,}$$

или, попросту можно сказать, что половина разностей содержится между —3 и 7 бушелями на 1 акр.

Основное свое применение медиана и квартили находят при характеристике больших выборок из совокупностей, отклоняющихся от нормальной формы. В прошлом эти показатели широко использовались для проведения приближенной оценки выборок из нормальной совокупности, но теперь при наличии счетных машин вычисление средней является более легким делом, чем вычисление медианы, а описанный ранее упрощенный метод среднего размаха оказывается более эффективным. Но и для распределений, отклоняющихся от нормального, мы теперь имеем лучшие методы. Некоторые из них будут рассмотрены в следующем параграфе, другие же — в последующем по мере того, как в этом будет необходимость. В биологии потребность применения схемы медиана — квартили, по-видимому, крайне незначительна.

Для иллюстрации этого метода возьмем данные таблицы 36 [1]. В целях упрощения я допустил, что в опыте было 179 коров, хотя в действительности большая часть данных является повторными наблюдениями после каждого последующего отела. Это, конечно, вносит известные сомнения в отношении приводимых ниже заключений.

ТАБЛИЦА 36

Распределение периодов между отелом и первой последующей половой охотой у нормальных коров. Стадо голштин-фризской породы

Средний класс (дни)	10,5	30,5	50,5	70,5	90,5	110,5	130,5	150,5	170,5	190,5	210,5
Частоты	8	33	50	32	15	20	11	6	2	1	1
Накопленные частоты		41	91	123	138	158	169	175	177	178	179

Эти частоты образуют вершину около 50 дней. Число дней, соответствующее наибольшей численности, будет называться, по Карлу Пирсону, *модой*. В данном случае имеется вторичная мода около 110 дней. Такого рода *бимодальность* кривой вместе со скошенностью ее явно указывает на то, что распределение не является нормальным.

Средняя равна 69,9 дня и находится примерно в центре следующего за модальным классом (способ вычисления средней и стандартного отклонения см. в параграфе 3 главы 8). Эта средняя не репрезентирует типичский период между отелом и половой охотой в том смысле, как это делает мода. Каково же значение средней в распределении, отличающемся от нормального? Основным свойством средней является то, что $\Sigma x = 0$, т. е. что сумма положительных отклонений равна сумме отрицательных отклонений. В этом смысле она является центром тяжести распределения. Если вы вычертите на картоне гистограмму этого распределения и вырежете ее, то она должна находиться в сбалансированном состоянии на острие полки, если это острие проходит по вертикали точки $X = 69,9$ дня.

Квартили вычисляются по формуле:

$$Q = x_L + \frac{(n_Q - f_L)I}{f},$$

где x_L = значение X , являющееся нижней границей класса, содержащего число наблюдений искомой квартили;

n_Q = номер наблюдения, относящегося к искомой квартили;

f_L = накопленная частота, предшествующая классу, в котором содержится n_Q ;

I = классовый интервал;

f = численность класса, содержащего n_Q .

В случае нашего примера номера наблюдений для трех искомых квартилей равны: $n_{Q_1} = (179 + 1)/4 = 45$, $n_{Q_2} = 90$ и $n_{Q_3} = 135$.

Для расчета медианы имеем: $n_{Q_2} = 90$ содержится в классе, нижняя граница которого $x_L = 40,5$ дня, численность этого класса $f = 50$ коровам, а предшествующая накопленная сумма — 41 корове. Отсюда

$$Q_2 = \text{медиана} = 40,5 + \frac{(90 - 41) \times 20}{50} = 60 \text{ дням.}$$

В данном случае медиана находится около верхней границы модального класса. Это означает, что периодов, меньших 60 дней, столько, сколько периодов, превышающих 60 дней. Отметим, что медиана является центром численности периодов, в то время как средняя — центром длительности периодов. Ни та, ни другая из этих величин не совпадает с точкой максимальной частоты.

Для двух других квартилей имеем:

$$Q_1 = 40,5 + \frac{(45 - 41) \times 20}{50} = 42 \text{ дням.}$$

$$Q_3 = 80,5 + \frac{(135 - 123) \times 20}{50} = 96 \text{ дням.}$$

Вы сразу замечаете, что медиана 60 дней не находится в середине междуквартильного интервала.

Теперь можно дать такое описание распределения изучаемых периодов: это распределение имеет моды 50 и 110 дней, причем 50 дней является более выраженной модой; оно имеет медиану 60 дней при нижней квартили 42 и верхней 96 дней; междуквартильный интервал составляет $96 - 42 = 54$ дней; в этом интервале содержится половина всех периодов. Все это служит известным дополнением к описанию совокупности, которое дается распределением частот и его графическим изображением.

В случае малых выборок (примерно $n < 20$), взятых из совокупности с таким видом распределения, как в таблице 36, может возникнуть вопрос о правомерности применения описанных ранее методов, относящихся к нормаль-

ному распределению. В следующих параграфах будут рассмотрены соответствующие этому случаю способы.

Если наши 179 периодов можно рассматривать в качестве случайной выборки из некоторой обширной совокупности, положим, из совокупности изучаемых периодов у всех коров голштин-фризской породы, то к такой большой выборке я без колебаний применю центральную предельную теорему. Я вычислю $s_{\bar{x}}=2,9$ и построю 95 %-ный доверительный интервал для $\mu: 69,9 - 1,96 \times 2,9 \leq \mu \leq 69,9 + 1,96 \times 2,9$, где $t_{0,05}=1,96$ для числа степеней свободы $=\infty$.

Точно так же в том случае, если наши данные представляют случайную выборку, может быть построен доверительный интервал и для медианы основной совокупности [8]. Для этого определяются две численности путем прибавления к n_{Q_2} и вычитания из n_{Q_2} величины $t\sqrt{n/2}$, которая при $t_{0,05}=1,96$ примерно равна \sqrt{n} . Так, имеем $n_{Q_2} + \sqrt{n} = 90 \pm \sqrt{179}$, что дает 77 и 103. Далее определяются периоды, соответствующие этим численностям, по формуле для Q :

$$\text{для } 77 \quad 40,5 + \frac{(77-41) \times 20}{50} = 55 \text{ дням,}$$

$$\text{для } 103 \quad 60,5 + \frac{(103-91) \times 20}{32} = 68 \text{ дням.}$$

Таким образом, медиана совокупности содержится между 55 и 68 днями, причем значения, выходящие за эти пределы, могут встретиться один раз из 20 испытаний.

С другой стороны, может быть так, что данные таблицы 36 не являются выборкой; они могут представлять собой всю совокупность интересующих нас периодов, относящихся к определенному стаду. В этом случае квартили, средняя и стандартное отклонение дают точную характеристику этой совокупности и не подвержены выборочному варьированию. Если это так, то концепции оценок критериев существенности и доверительных интервалов к данному случаю не имеют никакого отношения, так как теперь нет никаких других коров вне данного стада, которые имели бы возможность быть включенными в этот эксперимент.

Конечно, владелец молочной фермы может не согласиться с такой узкой трактовкой. Ему, может быть, известно, что способы разведения и содержания скота в этом стаде не оказывают никакого влияния на период между отелом и первой после этого половой охотой у коров голштин-фризской породы. Это будет свидетельствовать о том, что данное стадо репрезентирует более обширную совокупность коров голштин-фризской породы, вроде того, как это делает и случайная выборка. Но такого рода положение возникает из знакомства с особенностями данной породы коров и не имеет никакого отношения к статистической теории выборочного метода. Теперь вспомним и о другой особенности наших данных; здесь представлено меньше чем 179 коров, так как многие данные являются повторными наблюдениями над одними и теми же животными. Эта выборка может репрезентировать только такую совокупность, в которой доли одинарных, двойных, тройных и прочих наблюдений те же, что и в этой выборке.

8. Непараметрические методы; ранжирование двух вариантов. Часто для измерения признака нет какой-либо количественной шкалы, и все же имеется возможность распознавать различные степени свойств или достоинств объектов. Например, владелец животных дает им оценку на основе формы их тела, располагая отдельных особей от наилучших к наихудшим, т. е. строя *ранжированный ряд* 1, 2, ..., n . Тем же самым путем эксперт по продуктам питания располагает в ряд образцы этих продуктов в соответствии с их запахом и вкусом. Если подобного рода ранжирования ряда объектов или вариантов эксперимента будут произведены при помощи случайной выборки из числа экспертов, то возникает возможность сделать выводы о ранжиро-

ванном ряде в совокупности, из которой взята эта выборка экспертов, несмотря на то, что этому распределению нельзя приписать никаких параметров.

Рассмотрим сначала ранжирование двух продуктов m дегустаторами. Например, $m=8$ дегустаторов располагают в порядок по качеству пирожки, начиненные мясом, которые хранились в домашнем холодильнике в течение 8 месяцев при двух температурах [4]. В основу ранжирования был положен запах мяса. Восемь пирожков, по одному для каждого дегустатора, были выдержаны при 0° Фаренгейта; вторая выборка из 8 пирожков была взята из холодильника, температура которого колебалась между 0 и 15° Фаренгейта. Результаты ранжирования представлены в таблице 37.

ТАБЛИЦА 37

Ранжирование по запаху пирожков, начиненных мясом.

Восемь дегустаторов. Ранг 1—лучшее качество,
2—худшее качество

Дегустаторы	Выборка 1-я. $0^\circ F$	Выборка 2-я. Колеблющаяся температура
A	1	2
B	1	2
C	2	1
D	1	2
E	1	2
F	1	2
G	1	2
H	1	2

Нулевая гипотеза, подлежащая проверке, состоит в том, что нет никакого различия между рангами запаха у пирожков, взятых из этих двух холодильников. В данном случае мы видим, что результаты одного ранжирования находятся в противоречии с семью другими результатами. Может ли это произойти в силу выборочного варьирования или же следует отвергнуть нашу гипотезу H_0 ?

Проверку гипотезы 1 : 1 произведем при помощи хи-квадрат. В данной выборке два различных порядка ранжирования находятся в отношении 7 : 1. Наиболее простой формулой хи-квадрат будет формула параграфа 16 главы 1 при $r=1$, $a=7$ и $c=1$. При такой малой выборке, как наша, рекомендуется введение некоторой поправки, указанной в параграфе 5 главы 9. С этой поправкой формула хи квадрат будет такой:

$$\chi^2 = \frac{(a-b-1)^2}{n}$$

где $n = a + b$;

$$\chi^2 = \frac{(7-1-1)^2}{8} = 3,12; \quad P = 0,08.$$

Это указывает на несущественность различий, хотя такое решение не является окончательным; все же остается под вопросом, не оказывает ли колебание температуры влияния на различия в запахе.

Замечание: этот критерий хи-квадрат с его поправкой, примененный к m ранжированным рядам из двух предметов, эквивалентен «критерию знаков» Диксона и Муда [31]. Так как здесь применяется обычное вычисление хи-квадрат, то нет никакой потребности в специальных таблицах для критерия знаков. Действительно, если иметь в виду, что $\chi_{0,05}^2 = 3,841$ и $\chi_{0,01}^2 = 6,635$, то для требуемой оценки хи-квадрат нет никакой нужды обращаться к какой-либо таблице.

Пример 11. Сигареты производства двух фабрикантов были оформлены одинаковым образом и предложены в случайном порядке 6 случайно выбранным курильщикам. Пятеро из них поставили на первое место по аромату сигареты фабриканта А. Если вы

отвергнете гипотезу о том, что нет никакого различия в аромате этих сигарет, то какова вероятность того, что вы ошибаетесь? *Ответ:* $\chi^2=2,67$; число степеней свободы=1, $P=0,1$. К какой совокупности относится это заключение?

Пример 12. Два вида сливочного мороженого были выработаны с различным ароматом, но без каких-либо других внешних различий, вроде цвета и пр. Шесть экспертов молочной промышленности все подряд отдали предпочтение аромату А. Может ли это быть статистическим доказательством того, что и потребители предпочитают этот аромат?

9. Непараметрические методы; ранжирование разностей между количественными показателями. Ранговый метод можно применить к выборкам количественных показателей, которые не оправдывают допущений, лежащих в основе критерия t . При применении этого критерия в случае парных наблюдений допускается, что разности образуют выборку из $N(\mu_D, \sigma_D)$. Отклонения от этой модели могут быть двух видов: разности могут не быть нормально распределены и дисперсии этих разностей могут быть различными. Отклонение от модели второго рода может быть таким, как в примере 6 главы 2. Обычно считается, что дисперсия высоких урожаев выше, чем низких урожаев, и что коэффициент вариации имеет тенденцию находиться примерно на одном уровне при различной урожайности. Если это так, то σ^2 в 1933 г. будет больше, чем в 1934 г., вследствие чего критерий t будет недействительным.

Соответствующий непараметрический критерий предложен Уилкоксоном [13]. Сначала ранжируются абсолютные значения разностей, после чего порядковым номерам приписываются знаки соответствующих разностей.

Данные урожайности кукурузы (пример 6 главы 2) для удобства, хотя это и не является обязательным, представлены ниже в виде ранжированного ряда абсолютных разностей с вынесенными отдельно их знаками:

Абсолютные разности	0,0	0,4	0,6	1,0	2,4	3,4	3,6	3,7
Знак разности	+	-	-	+	-	+	-	+
Ранг со знаком разности	1	-2	-3	4	-5	6	-7	8

Здесь ранги абсолютных значений выписаны с теми знаками, которые были вначале отброшены. Следующие, более высокие ранги здесь не представлены, так как между ними нет отрицательных.

Сумма отрицательных рангов равна $-2-3-5-7=-17$. Это число, отбрасывая его знак, следует сравнить с табличным, взятым из таблицы 38. При 16 парных наблюдениях сумма рангов ≤ 19 приводит к опровержению гипотезы на уровне $P=0,01$. Так как наша сумма меньше этого табличного значения, то можно вполне уверенно отвергнуть гипотезу $H_0: \mu_D=0$; вероятность ошибки типа 1 меньше 0,01.

ТАБЛИЦА 38

Сумма рангов при двух уровнях P^1 .

Данные или меньшие числа приводят к опровержению гипотезы.
Двусторонний критерий

Число пар	0,05	0,01	Число пар	0,05	0,01
7	2	0	12	14	7
8	2	0	13	17	10
9	6	2	14	21	13
10	8	3	15	25	16
11	11	5	16	29	19

¹ Составлена на основе статьи Уилкоксона. Более точные значения вероятностей даны в его статье.

Если окажется, что сумма положительных рангов будет меньше суммы отрицательных, то с табличным числом следует сравнивать эту первую сумму; вообще всегда следует брать меньшую сумму.

Сопоставление этого результата с результатом примера 39 главы 2 показывает, что оба критерия приводят к одному и тому же выводу. Это указывает на то, что если и существуют в данном эксперименте различия между дисперсиями, то они оказывают весьма небольшое влияние на критерий t . Мы уже имели случай отмечать ранее (и должны повторить это еще раз), что критерий t довольно инертен в отношении даже значительных отклонений от теоретических предположек, на которых он основан.

В тех случаях, когда две или больше число разностей равны, то часто вполне достаточно обозначить каждый член такой группы средним рангом, рассчитанным из рангов, предназначенных для этой группы. Так, если две равные разности попадают на пятое и шестое место ранжированного ряда, то каждой из них приписывается ранг $5\frac{1}{2}$.

Пример 13. Из двух J -образных совокупностей, схожих с распределением хи-квадрат при числе степеней свободы = 1 (см. рис. 2), я взял две выборки при $n=10$ и составил из них случайные пары.

Выборка 1-я	1,98	3,30	5,91	4,05	1,01	1,44	3,42	2,17	1,37	1,13
Выборка 2-я	0,33	0,11	0,04	0,24	1,56	0,42	0,00	0,22	0,82	2,54
Разность	1,65	3,19	5,87	0,81	-0,55	1,02	3,42	1,95	0,55	-1,41
Ранг	6	8	10	3	-1,5	4	9	7	1,5	-5

Разность между средними совокупностей была равна 1, а в выборке $\bar{x}_1 - \bar{x}_2 = 1,65$. Две самые малые абсолютные разности равны между собой, поэтому каждой из них приписывается ранг $(1+2):2=1,5$. Сумма отрицательных рангов равна 6,5 и содержится между критическими суммами 3 и 8 таблицы 38. H_0 отвергается примерно при $P=0,04$.

Пример 14. Если вам не известно, что разности предыдущего примера были получены из совокупностей, резко отличающихся от нормальных, то вы, вероятно, примените критерий t . Придете ли вы в этом случае к иному выводу? *Ответ:* $t=2,48$, $P=0,04$.

Пример 15. Если вы примените метод параграфа 8 этой главы (эквивалентный критерию знаков) к данным примера 13, то получите $\chi^2=(8-2-1)^2:5=5$ и $P=0,03$.

Пример 16. Примените критерий знаков к данным урожаям кукурузы, о которых говорилось в тексте этого параграфа. *Ответ:* $\chi^2=3,06$, $P=0,08$. В критерии знаков численное значение отрицательного ранга оставляется без внимания, в то время как в критерии Уилкоксона малое значение ранга с отрицательным знаком играет совсем иную роль, чем большое значение ранга.

10. Непараметрические методы; ранжирование непарных количественных показателей. Возвращаясь к вопросам об оценке данных двух выборок, о которых говорилось в главе 4, мы рассмотрим ранжирование в качестве непараметрического метода в приложениях к случайным выборкам количественных показателей в тех случаях, когда эти выборки не согласуются с обычными моделями нормального распределения. Метод, предложенный Уайтом [12], в равной мере приложим как в случае одинакового, так и в случае неодинакового размера выборок. При этом методе все наблюдения обеих групп располагаются в один ранжированный ряд, но при этом принимаются меры, чтобы данные каждой группы были отличимы. В этом случае ранги относятся к составному ранжированному ряду. После этого в целях определения существенности меньшая сумма рангов T сравнивается с данными таблицы 41.

В качестве примера возьмем наблюдения над кукурузным мотыльком в округе Бун штата Айова (параграф 3 главы 1). Было установлено, что яйца мотылька бывают отложены чаще на высоких растениях, чем на низких. Я взял данные о числе яиц, обнаруженных на 20 растениях довольно однородного поля. Растения были взяты из двух случайно отобранных участков, по 10 растений с каждого. Таблица 39 дает результаты подсчета яиц.

Число яиц кукурузного мотылька на растениях кукурузы.
Округ Бун, Айова, 1950 г.

Высота растений	Число яиц									
	0	14	18	0	31	0	0	0	11	0
Менее 23 дюймов	0	14	18	0	31	0	0	0	11	0
Более 23 дюймов	37	42	12	32	105	84	15	47	51	65

Распределение численности яиц мотылька в 1950 г. во многом напоминает распределение хи-квадрат при одной степени свободы (рис. 2). Большинство растений имеют небольшое число яиц, и в то же время небольшое число их оказываются сильно зараженными. Это распределение имеет *J*-образную форму и резко отличается от нормального. Более того, здесь имеется тенденция к увеличению стандартного отклонения по мере увеличения числа яиц. Поэтому теория нормального распределения не может быть основой для правильных выводов из такой малой выборки.

Для удобства я произвел перестановку 20 данных в ранжированный ряд таблицы 40. Количество яиц на высоких растениях здесь выделено жирным шрифтом, и все последующие данные опущены, так как они относятся к высоким растениям.

ТАБЛИЦА 40

Ранжированный ряд данных 20 подсчетов яиц и соответствующие ранги.

Жирный шрифт относится к подсчетам на растениях высотой в 23 дюйма и более.

Ранжированный ряд	0	0	0	0	0	0	11	12	14	15	18	31
Ранг	3 ^{1/2}	3 ^{1/2}	3 ^{1/2}	3 ^{1/2}	3 ^{1/2}	3 ^{1/2}	7	8	9	10	11	12

В соответствии с указанным ранее правилом группе из 6 одинаковых наблюдений приписывается средний ранг:

$$\frac{1+2+3+4+5+6}{6} = 3\frac{1}{2}.$$

Так как весь этот набор 6 рангов относится к одной группе, то в расчете средней нет никакой потребности; сумма 6 рангов равна 21 и будет одна и та же, возьмем ли мы в каждом случае средний ранг или отдельные ранги, 1, ...6. Но если этот набор рангов относится к разным группам, то вычисление среднего ранга необходимо.

Следующий этап состоит в суммировании n_1 рангов группы, имеющей меньшую сумму их; у нас это будут ранги растений ниже 23 дюймов:

$$T = 6 \times 3\frac{1}{2} + 7 + 9 + 11 + 12 = 60.$$

Эта сумма сравнивается с табличным значением T при $n_1 = n_2 = 10$ (табл. 41). Так как выборочное T меньше, чем $T_{0,01} = 71$, то нулевая гипотеза отвергнута при $P < 0,01$. Соответствующий вывод состоит в том, что число отложенных яиц зависит от высоты растения.

Проверяемая в данном случае нулевая гипотеза состоит в том, что сумма T представляет собой сумму n_1 рангов, выбранных случайным порядком из $n_1 + n_2$ рангов в объединенной их группе. Общая сумма рангов равна:

$$n_1(n_1 + n_2 + 1),$$

что в наших наблюдениях над кукурузным мотыльком дает $10 \times 21 = 210$.

Может случиться так, что ранги, взятые для суммирования, дают сумму, большую половины общей суммы. Если это произойдет, то отнимите эту

Значения T для двух уровней.

Двусторонний критерий. Данные или меньшие значения отвергают гипотезу.
Следует брать $n_1 \leq n_2$ ¹

		Для уровня 0,05														
$n_2 \backslash n_1$		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
4				10												
5			6	11	17											
6			7	12	18	26										
7			7	13	20	27	36									
8	3		8	14	21	29	38	49								
9	3		8	15	22	31	40	51	63							
10	3		9	15	23	32	42	53	65	78						
11	4		9	16	24	34	44	55	68	81	96					
12	4		10	17	26	35	46	58	71	85	99	115				
13	4		10	18	27	37	48	60	73	88	103	119	137			
14	4		11	19	28	38	50	63	76	91	106	123	141	160		
15	4		11	20	29	40	52	65	79	94	110	127	145	164	185	
16	4		12	21	31	42	54	67	82	97	114	131	150	169		
17	5		12	21	32	43	56	70	84	100	117	135	154			
18	5		13	22	33	45	58	72	87	103	121	139				
19	5		13	23	34	46	60	74	90	107	124					
20	5		14	24	35	48	62	77	93	110						
21	6		14	25	37	50	64	79	95							
22	6		15	26	38	51	66	82								
23	6		15	27	39	53	68									
24	6		16	28	40	55										
25	6		16	28	42											
26	7		17	29												
27	7		17													
28	7		17													
		Для уровня 0,01														
$n_2 \backslash n_1$		2	3	4	5	6	7	8	9	10	11	12	13	14	15	
5				10	15											
6				10	16	23										
7				10	17	24	32									
8				11	17	25	34	43								
9			6	11	18	26	35	45	56							
10			6	12	19	27	37	47	58	71						
11			6	12	20	28	38	49	61	74	87					
12			7	13	21	30	40	51	63	76	90	106				
13			7	14	22	31	41	53	65	79	93	109	125			
14			7	14	22	32	43	54	67	81	96	112	129	147		
15			8	15	23	33	44	56	70	84	99	115	133	151	171	
16			8	15	24	34	46	58	72	86	102	119	137	155		
17			8	16	25	36	47	60	74	89	105	122	140			
18			8	16	26	37	49	62	76	92	108	125				
19	3		9	17	27	38	50	64	78	94	111					
20	3		9	18	28	39	52	66	81	97						
21	3		9	18	29	40	53	68	83							
22	3		10	19	29	42	55	70								
23	3		10	19	30	43	57									
24	3		10	20	31	44										
25	3		11	20	32											
26	3		11	21												
27	4		11													
28	4		11													

¹ n_1 и n_2 —число наблюдений в двух группах. Если группы неравные, то за n_1 следует брать меньшее число. Таблица взята из работы Уайта [12], который подверг дальнейшей разработке метод Уилкоксона.

сумму из общей $n_1(n_1+n_2+1)$ и вы узнаете значение T , подлежащее сравнению с табличным T ; всегда берите меньшую сумму. Эту меньшую сумму можно найти и без использования указанной формулы, если начать нумерацию рангов с противоположного конца ранжированного ряда.

Пример 17. В статье, помещенной в «Биометрике», Уайт, ссылаясь на Райта [14], приводит данные о времени сохранения жизненной способности в анаэробных условиях перонеальных нервов у кошек и кроликов. Это время для нервов 4 кошек было 25, 45, 33 и 43 минуты, а у 14 кроликов: 28, 15, 35, 28, 35, 22, 23, 22, 17, 20, 30 и 16 минут. Проверьте гипотезу H_0 : время выживания у этих двух видов животных одинаково. *Ответ:* если 1 означает меньший ранг, то $T=58$, что больше половины $n_1(n_1+n_2+1)=76$. Поэтому берем $T=76-58=18$, что легко получить, обозначая рангом 1 наибольшую величину. $P=0,04$.

Пример 18. Нет никаких оснований считать, что время выживания в предыдущем примере не распределено нормально. Проверьте H_0 при помощи критерия t . *Ответ:* $t=2,94$; $P=0,01$.

Пример 19. В примерах 17 и 18 главы 4 некоторые выборочные наблюдения указывают на неравенство дисперсий в совокупностях, из которых взяты выборки. Примените непараметрический ранговый критерий. *Ответ:* $T=121,5$; $T_{0,05}=110$.

11. Непараметрические методы в качестве сокращенных способов обработки выборок из нормальной совокупности. Весьма полезным является знакомство с непараметрическими методами как с методами более быстрого получения обычных выводов по выборкам из нормальных совокупностей. Вопрос сводится к соотношению стоимости операций при том и другом методе. Ответ зависит от затрат на одно наблюдение и относительной эффективности критериев. Кроме того, так как на практике встречаются и распределения, отклоняющиеся от нормального, то при решении вопроса необходимо специально учитывать аномальности совокупности, из которой берется выборка, и возникающую в связи с этим потерю эффективности критерия t . В связи с этим нельзя предложить какой-либо универсальный принцип, все же некоторые общие положения могут быть высказаны.

Когда ранговые критерии Уилкоксона были применены к выборкам из нормальной совокупности, то оказалось, что они имеют эффективность примерно 95% при условии, что размер выборки близок к бесконечности [7]. Допустим, что при обычных условиях их эффективность 90%; это означает, что выборка при $n=9$ из нормальной совокупности при использовании исходных количественных данных будет содержать в себе столько же информации, как и выборка при $n=10$, в которой произведена замена на ранги. Если стоимость расчета критерия t равна 30 центам, а рангового критерия — 15 центам, то затраты при обоих методах будут одинаковыми при стоимости одного наблюдения около 15 центов. В области биологического эксперимента найдется только немного столь дешевых наблюдений. С другой стороны, известно, что эффективность критерия t при аномальной совокупности уменьшается. Если известно, что совокупность по своей форме далека от нормальной, то, наверное, при помощи рангового критерия можно получить больше информации и при меньших затратах. Для относительно дорогих наблюдений из близких к нормальной распределений я буду настаивать на критерии t .

Прежде чем решить. Решение вопроса о том, какой критерий можно использовать в том или ином случае, должен решаться без ссылки на материал, имеющийся на руках. Если вы примените одновременно оба метода и после этого выберете тот из них, который отвечает вашим взглядам, то тем самым вы введете в свои заключения преднамеренность. Конечно, если оба метода приводят к одним и тем же выводам, то вы получите уверенность в том, что это решение не обусловлено формой совокупности.

Ранговые критерии иногда служат в качестве предварительных критериев, требующих последующего применения критерия t в случае, если решение остается неопределенным. Этот прием иногда оправдывает затраты времени, особенно для лиц, не имеющих опыта. Если же решение очевидно, то, конечно, нет необходимости затрачивать на это время. Но работа по ранжированию, особенно в случае непарных сравнений, немного меньше, чем работа по вычислению t . Если оба метода будут применяться к значительной

части опытов (в большинстве опытов решение не очевидно), то произойдет неэкономная трата времени.

В защиту критерия знаков (последний абзац параграфа 8), примененного к выборке нормально распределенных разностей, мало что можно сказать, разве только то, что он легок и быстр. Его эффективность по отношению к t около 65%. В качестве сокращенного метода для нормально распределенных разностей следует предпочесть ранговый критерий Уилкоксона.

Л И Т Е Р А Т У Р А

1. Chapman A. B., Casida L. E., *Journal of Agricultural Research*, 54, 417, 1937.
2. Davidson O. W., Shive J. W., *Plant Physiology*, 10, 73, 1935.
3. Dixon W. J., Mood A. M., *Journal of the American Statistical Association*, 41, 557, 1946.
4. Ehrenkrantz Florence, Roberts Harriett, *Journal of Home Economics*, 44, 441, 1952.
5. Harrison Carter M., *Plant Physiology*, 9, 94, 1934.
6. Lord E., *Biometrika*, 34, 56, 1947.
7. Mood A. M., *Annals of Mathematical Statistics*, 25, 514, 1954.
8. Mood A. M., *Introduction to the Theory of Statistics*. McGraw-Hill Book Co., Inc., p. 388, 1950.
9. Nelson P. Mabel, Hoyt Elizabeth E., McLaughlin Laura, Morgan Ethel Cessna, *Iowa Agricultural Experiment Station Bulletin* 337, 1935.
10. Pallai R. C. S., *Annals of Mathematical Statistics*, 22, 469, 1951.
11. Savage Richard, *Journal of the American Statistical Association*, 48, 844, 1953.
12. White Colin, *Biometrics*, 8, 33, 1952.
13. Wilcoxon Frank, *Biometrics (Bulletin)*, 1, 80, 1945.
14. Wright E. B., *American Journal of Physiology*, 147, 78, 1946.

ЛИНЕЙНАЯ РЕГРЕССИЯ

1. Введение. Рассмотренные в предыдущих главах вопросы касались главным образом таких проблем, когда у каждого объекта брался только один признак. Таким признаком могло быть содержание витамина или вес, когда переменная непрерывна, или же цвет и общественное мнение, когда переменная дискретна. Такие наблюдения подвергаются единичному или групповому сравнению, но характерным тут является то, что у каждого объекта берется одно качество или один вид количественного признака. Теперь наше внимание будет направлено на выводы, основывающиеся на двух или более признаках каждого члена выборки. Например, можно сделать более правильные суждения по поводу 120-дневной прибавки в весе свиней, если известны такие сопутствующие показатели, как начальный вес животных, их возраст или количество съеденного корма; увеличение дохода фермы за некоторый период времени может, в частности, определяться изменением количества мясной продукции; решение вопроса об урожайности нескольких сортов кукурузы может быть связано с числом растений на опытных делянках.

В этой главе основное внимание будет уделено зависимости одной переменной Y от другой *независимой* переменной X . На математическом языке Y называется *функцией* X , но в биологической статистике со времен Гальтона широко применяется более образный термин — *регрессия*. Возрастная кривая роста рассматривается в качестве регрессии роста на возраст; в токсикологии смертельное действие яда характеризуется регрессией процента смертельных случаев на концентрацию яда. Происхождение термина регрессия будет объяснено в параграфе 12 этой главы.

Применение регрессии самое разнообразное. Возможно, она нужна только для того, чтобы установить, зависит ли Y от X ; если это так, то стоит задача получения меры этого соотношения. Но может быть и так, что целью является предугадывание Y по X . Иногда желательно определить форму кривой регрессии. В других случаях, как это показано в параграфе 10 главы 3, внимание сосредоточивается на фактической ошибке, содержащейся в экспериментальных данных, после исключения влияния сопутствующей переменной. Реже приходится иметь дело с некоторыми теориями относительно причинности и эффекта связи переменных и с применением регрессии для проверки таких гипотез. Для удовлетворения всех этих потребностей мы дадим широкий обзор соответствующих методов, предоставив каждому читателю выбирать из них те, которые его больше всего интересуют. Как обычно, для знакомства с предметом мы прежде всего обратимся к примеру.

2. Регрессия кровяного давления человека на возраст. Сельскохозяйственные опытные станции 9 средне-западных штатов провели исследование уровня питания разных групп населения. Из опубликованных материалов

я выбрал данные о систолическом (верхнем) давлении крови у 58 женщин старше 30-летнего возраста; эта была случайная выборка из жителей одного района вблизи Эймса, штат Айова [18]. Для целей, стоящих перед нами, была произведена группировка возрастов по 10-летним классам и для каждого класса было вычислено среднее кровяное давление. Результаты представлены в первых двух колонках таблицы 42.

ТАБЛИЦА 42

Среднее систолическое давление крови у 58 женщин при 10-летних возрастных классах

Середина возрастного класса X	Среднее давление крови Y	Отклонения от средних		Квадраты		Произведе- дения xy
		x	y	x^2	y^2	
35	114	-20	-27	400	729	540
45	124	-10	-17	100	289	170
55	143	0	2	0	4	0
65	158	10	17	100	289	170
75	166	20	25	400	625	500
Сумма 275 Средняя 55	715 141	0	0	1000	1936	1380

Выборочный коэффициент регрессии $b = \frac{\sum xy}{\sum x^2} = \frac{1380}{1000} = 1,38$ единицы давления на год.

В большинстве случаев при изучении вопросов о регрессии в первую очередь рекомендуется построить график по образцу рисунка 8. Независимая

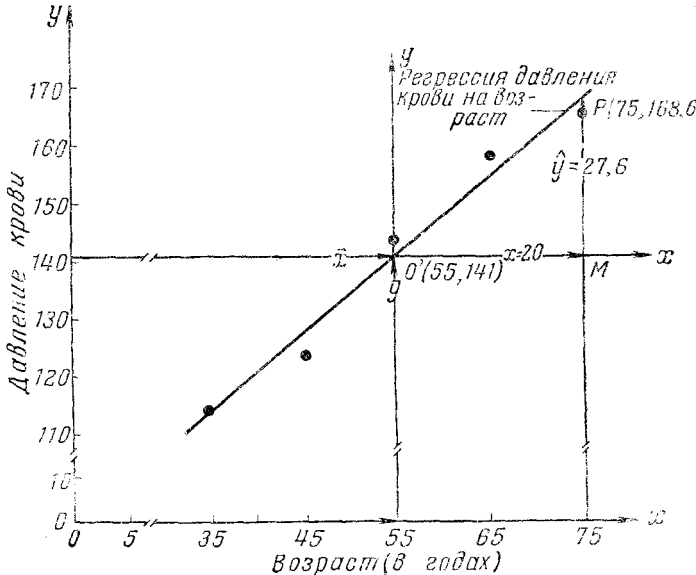


Рис. 8. Выборочная регрессия давления крови на возраст. Разрыв осей координат указывает на то, что нижние части соответствующих шкал опущены в целях более крупного изображения той части графика, где сосредоточены данные.

переменная X располагается по горизонтальной оси. Каждое наблюдаемое значение зависимой переменной Y изображается точкой, расположенной над соответствующим значением X . Легко видеть, что изменение давления крови с возрастом направлено вверх почти по прямой линии.

Эта прямая линия, показанная на рисунке, называется *выборочною регрессией Y на X*. Она проходит через точку $O'(\bar{x}, \bar{y})$, которая в нашем случае соответствует значениям (55, 141), и имеет наклон, определяемый b единицами Y на одну единицу X , где b — *выборочный коэффициент регрессии*, вычисляемый, как показано в таблице 42. В это вычисление входит операция, с которой мы ранее не встречались, — расчет *суммы произведений* из отклонений x и y . Сумма произведений может быть как положительной, так и отрицательной, в зависимости от того, куда направлена прямая — вверх или вниз. Во взятом нами примере $b=1,38$ означает, что кровяное давление увеличивается в среднем на 1,38 единицы на каждый год возраста.

Теперь можно написать *уравнение выборочной регрессии Y на X*:

$$\hat{y} = bx,$$

где \hat{y} является *оценкой* отклонения Y , соответствующего некоторому отклонению x . Например, если $x=20$ лет, то $\hat{y}=1,38 \times 20=27,6$ единицы давления крови.

Уравнение выборочной регрессии дает нам полную возможность построить выборочную линию регрессии рисунка 8. Откладываем вправо от точки O' расстояние $O'M=20$ лет, после чего восстанавливаем перпендикуляр $MP=27,6$ единицы давления крови. Линия $O'P$ в этом случае имеет наклон 1,38 единицы давления крови на 1 год возраста.

В первоначальных единицах уравнение выборочной регрессии будет таким:

$$\hat{Y} - \bar{y} = b(X - \bar{x}).$$

Для давления крови это будет

$$\hat{Y} - 141 = 1,38(x - 55),$$

или

$$\hat{Y} = 141 + 1,38(X - 55) = 65,1 + 1,38X.$$

Если в это уравнение подставить $X=75$, то \hat{Y} будет $65,1 + 1,38 \times 75 = 168,6$ единицы давления крови. Соответствующая точка (75; 168,6) на нашем рисунке обозначена P .

Теперь для того, чтобы установить степень согласия линии регрессии с данными, мы можем произвести сравнение фактических точек выборки с соответствующими вычисленными \hat{Y} . Для этого подставляем в уравнение регрессии каждое значение X и вычисляем \hat{Y} . Результаты пяти таких расчетов даны в таблице 43. *Отклонения от регрессии* $Y - \hat{Y} = d_{y \cdot x}$ дают меру несоответствия этой линии с наблюдаемыми данными. Обратите внимание на то, что в данном случае у 45-летних женщин среднее давление крови ниже линии регрессии, а у 65-летних — выше нее.

ТАБЛИЦА 43

Вычисления \hat{Y} и отклонений от регрессии $d_{y \cdot x} = Y - \hat{Y}$.
Данные о давлении крови

Середина возрастного класса X	Среднее давление крови Y	Ожидаемое давление крови \hat{Y}	Отклонение от регрессии $Y - \hat{Y} = d_{y \cdot x}$	Квадрат отклонения $d_{y \cdot x}^2$
35	114	113,4	0,6	0,36
45	124	127,2	-3,2	10,24
55	143	141,0	2,0	4,00
65	158	154,8	3,2	10,24
75	166	168,6	-2,6	6,76
Сумма			$\sum d_{y \cdot x} = 0$	$\sum d_{y \cdot x}^2 = 31,60$

Сумма квадратов отклонений $\Sigma d_{y \cdot x}^2 = 31,60$ является основой для оценки ошибки, возникающей при подборе линии регрессии. Соответствующее число степеней свободы равно $n - 2 = 3$, так как в этих расчетах были использованы две средние величины \bar{y} и b . Поэтому мы имеем

$$s_{y \cdot x}^2 = \Sigma d_{y \cdot x}^2 / (n - 2) = 10,53,$$

где $s_{y \cdot x}^2$ является средним квадратом отклонений от регрессии. Отсюда выборочное стандартное отклонение от регрессии будет $s_{y \cdot x} = \sqrt{s_{y \cdot x}^2} = 3,24$ единицы давления крови, что соответствует показателю s в задачах с одной переменной. В частности, эта величина позволяет определить выборочное стандартное отклонение коэффициента регрессии:

$$s_b = s_{y \cdot x} / \sqrt{\Sigma x^2}.$$

Так, мы имеем: $3,24 / \sqrt{1000} = 0,102$ единицы давления крови.

Критерий существенности для b определяется формулой

$$t = b/s_b \text{ при числе степеней свободы} = n - 2.$$

Прилагая это к данным о давлении крови, имеем:

$$t = 1,38 / 0,102 = 13,5^{**}; \text{ число степеней свободы} = 3.$$

Примечание. Для удобства часто довольствуются обозначением существенности при помощи звездочек. Одна звездочка указывает, что вероятность содержится между 0,05 и 0,01, а две — что вероятность равна или меньше 0,01.

Довольно часто отдельные значения $d_{y \cdot x}$ также, как в таблице 43, не представляют никакого интереса. В этом случае $\Sigma d_{y \cdot x}^2$ может быть вычислено непосредственно по формуле:

$$\Sigma d_{y \cdot x}^2 = \Sigma y^2 - (\Sigma xy)^2 / (\Sigma x^2).$$

Подставляя сюда данные нашего примера из таблицы 42, имеем:

$$\Sigma d_{y \cdot x}^2 = 1,936 - 1,380^2 / 1000 = 31,60,$$

т. е. тот же результат, что и ранее.

Пример 1. Ниже приводятся результаты измерения высоты растений на одном участке посева сои, отобранных случайным порядком через каждую неделю [21]:

Возраст (недели)	1	2	3	4	5	6	7
Высота (в см)	5	13	16	23	33	38	40

Проверьте такие результаты расчетов: $\bar{x} = 4$ недели, $\bar{y} = 24$ см, $\Sigma x^2 = 28$, $\Sigma y^2 = 1080$, $\Sigma xy = 172$. Определите выборочную регрессию $\hat{Y} = 6,143X - 0,572$ см.

Пример 2. Нанесите на график выборочные точки, соответствующие данным о высоте растений сои, после чего проведите выборочную линию регрессии. Расположатся ли точки примерно поровну выше и ниже этой линии?

Пример 3. Вычислите $s_b = 0,409$ см в неделю. Определите 95%-ный доверительный интервал для этой регрессии в совокупности. Ответ: 5,09—7,20 см в неделю.

Пример 4. По данным о высоте растений сои постройте кривую роста. Можете ли вы считать, что изменение роста в совокупности действительно идет по прямой? Как вы спланировали эксперимент для определения кривой роста давления крови у женщин штата Айова?

Пример 5. Восемнадцать образцов почвы были смешаны с различными количествами (X) неорганического фосфора. Растения кукурузы, выращенные на каждой из этих почв, были срезаны по прошествии 38 дней и подвергнуты анализу на содержание в них фосфора. На основании этого было установлено количество доступного для растений фосфора в почве (параграф 2 главы 14). Девять из этих наблюдений, отобранных так, чтобы облегчить вычисления, приведены в нижеследующей таблице:

Неорганический фосфор в почве (в мг на 1 кг) \bar{X}	1	4	5	9	13	11	23	23	28
Доступный для растения фосфор (в мг на 1 кг) \bar{Y}	64	71	54	81	93	76	77	95	109

Вычислите: $b = 1,417$; $s_b = 0,395$; $t = 3,59^{**}$.

Далее идет краткое изложение более общих методов регрессии; читатель, если он желает, может прямо перейти к дисперсионному анализу, излагаемому в главах 10 и 11.

3. Модель I совокупностей, из которых берутся выборки: фиксированные значения X . Перед тем как идти далее, необходимо дать определение совокупностей, из которых берутся выборки. Они имеют следующие 3 характерные особенности.

1. Каждому отдельно взятому X соответствует нормальное распределение Y , из которого случайно отбирается выборочное \bar{Y} . Если нужно, то из каждого такого распределения может быть взято несколько значений Y (см. параграф 13 этой главы).

2. Средние μ всех таких совокупностей лежат на прямой линии регрессии.

3. Все эти совокупности, откуда берутся выборки, имеют нормальное распределение с общей для них σ . Эта модель I определяется в сжатой форме уравнением

$$Y = \alpha + \beta x + \varepsilon,$$

где Y — некоторое значение зависимой переменной; α и β — параметры (что будет далее объяснено); x — отклонение $X - \bar{x}$ и ε — случайная переменная, полученная из $N(0, \sigma)$.

Эта модель Y состоит из случайной части ε и части, фиксированной значением x . Эта фиксированная часть в соответствии с пунктом 2 определяет средние совокупностей, по одной средней для каждого отдельного x . Последовательность этих средних располагается на прямой линии, определяемой уравнением $\mu = \alpha + \beta x$, т. е. *линией регрессии в совокупности*. Параметр α является средней этой совокупности и соответствует $x=0$; в этом смысле α определяет высоту линии регрессии в совокупности над осью X . Величина β является *наклоном*, или *скоростью роста*, линии регрессии, т. е. *изменением Y на единицу изменения x* . Что же касается переменной части Y , т. е. ε , полученной случайно из $N(0, \sigma)$, то она *не зависит* от x .

Нематематику эту модель лучше всего можно объяснить при помощи некоторой арифметической конструкции. Для этого дадим величине X значения 0, 2, 3, 7, 8, 10, приведенные в таблице 44. Эти значения выбраны произвольно; то, каким образом фиксируются X , не имеет никакого отношения к данной иллюстрации.

Сначала вычисляем \bar{x} и в колонке 2 отклонения $x = X - \bar{x}$.

Теперь берем $\beta = 0,5$; это означает, что средние частных совокупностей увеличиваются на половину единицы, когда x меняется на единицу. На основе этого вычисляются данные колонки 3.

Далее возьмем $\alpha = 4$, означающее, что при $x=0$ регрессия в совокупности приподнята над осью X на 4 единицы.

Беря отдельные значения X и указанные α и β , определяем в колонке 4 последовательность средних. На рисунке 9 эти средние представлены пустыми кружками, лежащими на линии регрессии в совокупности. Таким образом, все эти величины фиксированы пока без участия выборочного варьирования.

Переходим теперь к переменной части Y ; значения ε выбираются случайно с помощью следующего процесса.

1. Обращаясь к таблице случайных чисел (глава 1), я случайным образом выбираю исходную позицию — 61-й ряд и колонки 55—56 и, начиная

Конструирование выборки для модели I;
 $Y = \alpha + \beta x + \varepsilon$, при $\alpha = 4$, $\beta = 0,5$ и ε , взятой из $N(0,1)$

X	x	$\beta x = 0,5x$	$\alpha + \beta x = 4 + 0,5x$	ε	$Y = \alpha + \beta x + \varepsilon$
(1)	(2)	(3)	(4)	(5)	(6)
0	-5	-2,5	1,5	1,1	2,6
2	-3	-1,5	2,5	-1,3	1,2
3	-2	-1,0	3,0	-1,1	1,9
7	2	1,0	5,0	1,0	6,0
8	3	1,5	5,5	0,0	5,5
10	5	2,5	6,5	-1,0	5,5

Вычисление оценок для выборочной регрессии Y на X

$\Sigma X = 30$	$\Sigma XY = 149,1$	$\Sigma Y = 22,7$
$\bar{x} = 5$	$\bar{y} = 3,78$	
$\Sigma X^2 = 226$	$C = (\Sigma X \cdot \Sigma Y) / n = 113,5$	$\Sigma Y^2 = 108,31$
$C = 150$	$\Sigma xy = 35,6$	$C = 85,88$
$\Sigma x^2 = 76$		$\Sigma y^2 = 22,43$

$$b = \Sigma xy / \Sigma x^2 = 35,6 : 76 = 0,468.$$

$$\hat{Y} = 3,78 + 0,468(X - 5) = 1,44 + 0,468X.$$

$$\Sigma d_{y,x}^2 = \Sigma y^2 - (\Sigma y)^2 / \Sigma x^2 = 22,43 - 35,6^2 / 76 = 5,75.$$

$$s_{y,x}^2 = \Sigma d_{y,x}^2 / (n - 2) = 5,75 : 4 = 1,44; s_{y,x} = \sqrt{1,44} = 1,20.$$

отсюда, получаю числа 86, 09, 14, 82, 52 и 16. Сопоставляя эти номера с нормально распределенными наблюдениями таблицы 15, я получаю 41, 17, 19, 40, 30 и 20.

2. Для того чтобы сделать среднюю равной нулю, я из всех этих данных вычитаю 30 и перехожу к отклонениям 11, -13, -11, 10, 0 и -10.

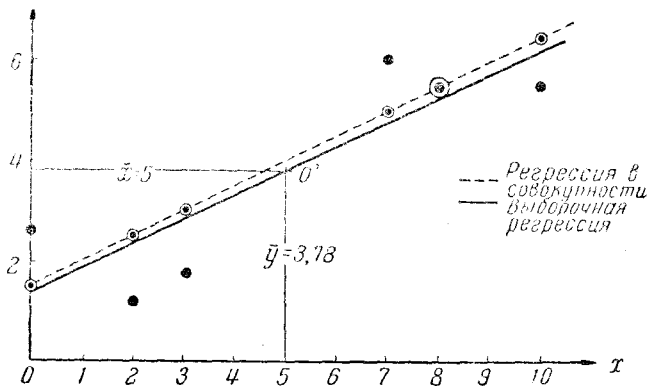


Рис. 9. Регрессия в совокупности $\mu = 4 + 0,5x$. Выборочная регрессия $\hat{Y} = 3,78 + 0,468x$.

3. Для того чтобы перейти от вариации $\sigma = 10$ к заданной $\sigma = 1$, я делю каждое отклонение на 10, получая тем самым стандартное отклонение, равное 1. Результаты приведены в колонке 5 и представляют собой случайную выборку из $N(0,1)$.

Последняя колонка таблицы содержит в себе *выборочные Y*; здесь каждое число является суммой фиксированной части колонки 4 и соответствующей случайной части колонки 5. Установленные таким образом *выборочные точки*

на рисунке представлены в виде черных кружков. Все последующие расчеты приведены в нижней части таблицы.

Пример 6. В таблице 44 $b=0,468$. Вычислите 6 отклонений от регрессии $d_{y \cdot x}$ и отождествите их с расстоянием соответствующих точек от выборочной линии регрессии. Сумма этих отклонений должна быть равна нулю, а сумма квадратов их — около 5,75.

Пример 7. Сконструируйте выборку при $\alpha=6$ и $\beta=-1$. Отрицательное β означает, что линия регрессии наклонена в правой стороне вниз. Возьмите $X=1, 2, \dots, 9$ при $\bar{x}=5$. Значения ϵ возьмите случайным образом из $N(0,5)$ таблицы 24. Постройте таблицу для расчета выборки значений Y . Нанесите на график регрессию совокупности и выборочные точки. Все это сохраните для использования в будущем.

4. Оценки значений α и β . Как было показано, $\alpha=4$ оценивается средней $\bar{y}=3,78$. Поэтому *выборочная линия* регрессии проходит через точку (\bar{x}, \bar{y}) ; у нас на рисунке через точку $(5; 3,78)$.

Наклон $\beta=0,5$ оценивается $b=0,468$ (см. таблицу 44). Выборочная линия регрессии почти параллельна регрессии совокупности, но лежит несколько ниже последней, так как была получена несколько преуменьшенная оценка α . Различие между этими линиями полностью обусловлено случайным выборочным варьированием.

Константа 1,44 в уравнении выборочной регрессии является точкой *пересечения* оси Y при $X=0$.

Отметим, что здесь нет случайного варьирования X ; средняя \bar{x} не зависит от ϵ . В этом и состоит основная особенность модели I.

Пример 8. По вашей выборке примера 7 вычислите \bar{y} и b , после чего нанесите на ваш график выборочную линию регрессии. Вычислите отклонения $d_{y \cdot x}$ и сравните их с соответствующими ϵ . Равенство $\sum d_{y \cdot x} = 0$ является частичным контролем точности этих расчетов.

5. \hat{Y} в качестве оценки $\mu = \alpha + \beta x$. Приведенные значения Y . При любом x выборочное значение Y является оценкой соответствующей средней $\mu = \alpha + \beta x$. Например, вы уже видели, что при $x=0$ (т. е. при $X=5$) $Y_5 = \bar{y}$, т. е. оценивает $\mu_5 = \alpha$. Другой пример: при $x=2$, когда $X=7$, $Y_7 = 1,44 + 0,468 \times 7 = 4,72$, что является оценкой $\mu_7 = 4 + 0,5 \times 2 = 5$.

Разность $\mu - \hat{Y}$ имеет два источника возникновения, причем оба они обусловлены случайностью ϵ . Один из них является различием между уровнями линий регрессий в выборке и в совокупности, т. е. $\alpha - \bar{y}$; второй — различием в углах наклона этих двух линий $\beta - b$.

Оценка значений μ часто не ограничивается теми совокупностями, которые характеризуются значениями X , наблюдаемыми при выборке, а относится к таким значениям X , которые лежат между фиксированными X . Например, при $X=4$

$$\hat{Y}_4 = 1,44 + 0,468 \times 4 = 3,31$$

находится на выборочной линии регрессии и на перпендикуляре, восстановленном из точки $X=4$. Очевидно, что в данном случае мы оцениваем μ такой совокупности, которая не участвовала в нашей выборке. Этой оценке не соответствует какое-либо выборочное наблюдение; она основана на наличии у исследователя убеждения, что промежуточная совокупность имеет то же значение σ , что и у тех, которые участвовали в выборке, и имеет μ на линии регрессии $\alpha + \beta x$.

На основе такой же аргументации можно произвести оценку μ для X , выходящего из интервала фиксированных в выборке X . Так, для $X=12$ имеем

$$\hat{Y}_{12} = 1,44 + 0,468 \times 12 = 7,06.$$

Такого рода экстраполяция связана с большим риском ввиду того, что регрессия в совокупности может быть криволинейной, причем эта криволинейность

становится более выраженной именно вне интервала фиксированных в выборках X .

Если брать выборку, как она есть, то значения \hat{Y} дают нам возможность установить, будет ли то или иное наблюдение значения Y выше или ниже соответствующей средней, что особенно ясно видно на графике. Возьмем, например, первую точку слева. Здесь $Y_0=2,6$ по сравнению с $\hat{Y}_0=1,44$. Положительная разность $Y_0-\hat{Y}_0=1,16$ указывает на то, что Y_0 превосходит свою оценку на 1,16 единицы. Это может означать, что наблюдаемое Y_0 , случайно взятое из первой нашей совокупности, больше средней этой совокупности. Это фактически подтверждается в нашей сконструированной выборке: $Y_0=2,6$ лежит выше $\mu_0=1,5$, что обусловлено ошибкой $\varepsilon_0=1,1$.

Иногда желательно сравнить такие отклонения от регрессии для двух значений X . Например, для $X=10$ отклонение $Y_{10}-\hat{Y}_{10}=5,5-(1,44+0,468 \times 10)=-0,62$; соответствующая точка лежит на 0,62 ниже выборочной регрессии. По отношению к выборочной регрессии это шестое наблюдение Y находится ниже первого на $1,16-(-0,62)=1,78$ единицы.

Такое сравнение отдельных случаев более удобно производить при помощи приведенных значений Y . Эти последние определяются путем прибавления к \bar{y} тех или иных отклонений от регрессии. Например,

$$Y_{A,0} = \bar{y} + (Y_0 - \hat{Y}_0) = 3,78 + 1,16 = 4,94,$$

где $Y_{A,0}$ означает приведенное значение Y при $X=0$. Точно так же

$$Y_{A,10} = 3,78 + (-0,62) = 3,16.$$

Отметим, что разность между приведенными Y та же самая, что и между соответствующими отклонениями:

$$Y_{A,0} - Y_{A,10} = 4,94 - 3,16 = 1,78.$$

Графическое истолкование смысла Y_A дано на рисунке 10. Оно является тем значением Y , которое получится, если соответствующее ему X передвинется в точку \bar{x} и если при этом Y будет двигаться параллельно выборочной линии регрессии. Этот процесс содержит в себе определенные предпосылки, которые в биологической практике могут быть оправданными или неоправданными; данный вопрос должен решить сам исследователь. Дальнейшее обсуждение вопроса будет проведено в главе 13.

Пример 9. Используя данные о давлении крови в параграфе 2, произведите оценку μ для 30-летнего возраста. *Ответ:* 106,5 единицы.

Пример 10. На основе формулы $Y-bx$ определите приведенные Y для каждой возрастной группы. Проверьте ваши результаты на основе равенства сумм $\Sigma Y_A = \Sigma Y$. Выскажите свои суждения по поводу различий между приведенными Y .

Пример 11. Г-жа А, 65 лет, имеет давление крови 135 единиц и г-жа В, 35 лет, имеет такое же давление. Разность между приведенными давлениями крови составляет 41,4 единицы, на которые давление крови у г-жи В выше. Объясните смысл этого факта. Будет ли такое сравнение разумным?

Пример 12. По данным о содержании доступного для растений фосфора примера 5 вычислите Y_A для 9-го образца почвы. *Ответ:* 87,7 мг на 1 кг. Имеет ли этот результат какой-нибудь биологический смысл?

Пример 13. В примере 1 этой главы дана кривая роста растений сои. Вычислите Y_A для 7-недельных растений. *Ответ:* 21,6 см, на 2,4 см выше y . Можете ли вы дать этому реальное истолкование?

Пример 14. Докажите, что $\bar{y} + Y - \hat{Y} = Y - bx$.

Пример 15. Докажите, что $\mu - \hat{Y} = (a - \bar{y}) + x(\beta - b)$.

6. Оценка σ^2 . Отклонения от выборочной регрессии $d_{y,x}$ аналогичны ε . Так как ε получены путем случайного отбора из $N(0, \sigma)$, то естественно

оценку σ^2 строить на основе $d_{y,x}$. Это и было сделано в таблице 43, где такие отклонения были возведены в квадрат, просуммированы и после этого было произведено деление на число степеней свободы $n - 2$. Теперь ясно, что $s_{y,x}^2$ является оценкой σ^2 ; в частности, 1,44 является оценкой $\sigma^2 = 1$ (или $1,20 =$ оценкой $\sigma = 1$).

Если бы нам ничего не было известно относительно X , то выборка значений Y нам дала бы $s_y^2 = 22,43 : 5 = 4,49$ в качестве оценки экспериментальной ошибки. Но использование наших сведений об X снижает ее до $s_{y,x}^2 = 1,44$. Не обращая пока внимания на изменение числа степеней свободы, можно считать, что эффективность регрессии составляет $4,49 : 1,44 = 3,12$, или 312 %.

Пример 16. По вашей выборке, которую вы произвели в соответствии с примером 7, вычислите $s_{y,x}^2$ и сравните ее с $\sigma^2 = 25$.

Пример 17. По данным примера 5 произведите оценку σ совокупности, из которой взята выборка. *Ответ:* $s_{y,x} = 10,69$ мг на 1 кг почвы. Сравните ее с $s_y = 16,86$ мг на 1 кг почвы.

7. Модель II совокупности, из которой берется выборка; нормальное распределение двух переменных. Исследователь не всегда может сознательно и преднамеренно выбирать значения переменной X ; часто он берет из некоторой совокупности какой-либо объект и производит измерение двух признаков, соотношение которых его в данном случае интересует. Эти две переменные образуют распределение совокупности, *варьирующей в двух направлениях*, о чем более подробно будет сказано в следующей главе. Здесь мы будем иметь дело только с *нормальным распределением двух переменных*. Те же из читателей, кто желает получить более подробные сведения о предмете, должны обратиться к соответствующей литературе [1, 13, 14].

К счастью, интересующие нас методы находят полное отражение в двух моделях. Выборка, взятая нами для показа модели II, представляет интерес и в другом отношении; она будет относиться уже не к положительной, а отрицательной регрессии.

Вообще есть основание думать, что размеры повреждений, наносимых личинкой яблонной моли, больше на яблонях с меньшим урожаем. По-видимому, численность бабочек моли при массовом лете в основном везде одинакова, в связи с чем при небольшом числе плодов на дереве вероятность поражения отдельного плода увеличивается. Данные таблицы 45 специально отобраны из результатов опыта [10], проведенного для проверки этого явления. Двенадцать деревьев при опадении лепестков были подвергнуты опрыскиванию мышьяковокислым свинцом с последующим пятикратным общим опрыскиванием 3 фунтами мышьяковокислого марганца и 1 квартой рыбьего жира на 100 галлонов. Здесь обнаруживается определенная тенденция, особенно хорошо выраженная на рисунке 11, к уменьшению процента поврежденных плодов при увеличении числа плодов на дереве. В случае данной группы деревьев связь между этими двумя переменными даже более тесная, чем это обычно наблюдается.

Новым в данной обработке результатов наблюдения является то, что большинство произведений xy отрицательны, что обусловлено тенденцией к соответствию малых значений Y большим значениям X . Выборочный коэффициент регрессии указывает, что процент поврежденных плодов уменьшается (о чем говорит знак минус у этого коэффициента) на 1,013 при увеличении урожая на 100 плодов. Линия регрессии и, конечно, вместе с этим процент поврежденных плодов снижается от точки $O'(\bar{x}, \bar{y})$ на 1,013 для каждой единицы урожая сверх 19 сотен.

График регрессии в наглядной форме дает представление об отклонениях от подвижной средней, образующей эту линию, т. е. отклонениях, которые дают меру погрешностей при попытке объяснить величину урожая плодов только варьированием степени повреждения. Деревья за номерами 4, 9 и 11 имеют значительные отклонения процента поврежденных плодов от ожидаемого их процента, в то время как на деревьях за номерами 2 и 5

Регрессия процента плодов яблони, поврежденных личинкой моли, на урожай яблок

Номер дерева	Урожай (деревя в сотнях плодов)	Процент поврежденных плодов	Отклонение от средней		Квадраты отклонений		Произведение отклонений	Оценка \hat{Y}	Отклонение от регрессии $Y - \hat{Y} = d_{Y..}$	Квадрат $(Y - \hat{Y})^2 = d_{Y..}^2$
	X	Y	x	y	x ²	y ²				
1	8	59	-11	14	121	196	-154	56,14	2,86	8,18
2	6	38	-13	13	169	169	-169	58,17	-0,17	0,03
3	11	56	-8	11	64	121	-88	53,10	2,90	8,41
4	22	53	3	8	9	64	24	41,96	11,04	121,88
5	14	50	-5	5	25	25	-25	50,06	-0,06	0,00
6	17	45	-2	0	4	0	0	47,03	-2,03	4,12
7	18	43	-1	-2	1	4	2	46,01	-3,01	9,03
8	24	42	5	-3	25	9	-15	39,94	2,06	4,24
9	19	39	0	-6	0	36	0	45,00	-6,00	36,00
10	23	38	4	-7	16	49	-28	40,95	-2,95	8,70
11	26	30	7	-15	49	225	-105	37,91	-7,91	62,57
12	40	27	21	-18	441	324	-378	23,73	3,27	10,69
Сумма	228	540	0	0	924	1222	-936	540,00	0,00	273,88
Средняя	19	45						45,00		

$$b = \frac{\sum xy - 936}{\sum x^2 - 924} = -1,013\% \text{ на } 100 \text{ плодов.}$$

$$s_{y..x}^2 = \sum d_{y..x}^2 / (n-2) = 273,88 : 10 = 27,388.$$

$$\hat{Y} = \bar{y} + b(X - \bar{x}) = 45 - 1,013(X - 19) = 64,247 - 1,013X \text{ сотен плодов.}$$

получен ожидаемый процент поврежденных плодов. В соответствии с указанной ранее моделью эти отклонения являются случайными отклонениями от средних (определяемых регрессией) значений, но более тщательные наблюдения над деревьями во время полета яблонной моли могут

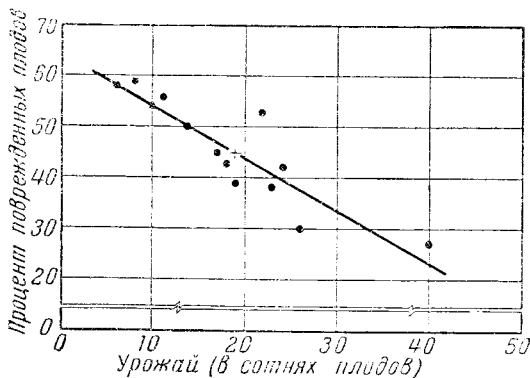


Рис. 11. Выборочная регрессия процента поврежденных плодов на размер урожая яблок. Крестик показывает начало отсчета отклонений.

открыть нам некоторые особенности данного явления. Например, дерево № 4 могло оказаться на той стороне, откуда происходил полет моли, или может быть так, что это дерево имело такую форму или положение, которые привели к недостаточно интенсивному опрыскиванию инсектицидом. Деревья же № 9 и 11 могли иметь некоторые особенности в форме листьев, которые служат для них защитой. Тщательное изучение деревьев № 2 и 5 также может пролить свет на их особенности или их местоположения, которые могут объяснить нормальный характер повреждения плодов на этих деревьях. Такого рода *изучение причинности* обычно не меняет нашего отношения к выборочным показателям, но оно может значительно расширить знакомство экспериментатора с подопытным материалом и указать средства

для усовершенствования будущих опытов. Так, например, в результате такого изучения может быть установлено, что некоторые причины варьирования материала могут быть заранее учтены и исключены из ошибки опыта при помощи специального построения схемы опыта (см. главу 11).

Вы, несомненно, обратите внимание на тот факт, что в условиях регрессии констатация того, что «в фруктовом саду, средняя поврежденность плодов в котором составляет 45%, имеется дерево, на котором только 27% поврежденных плодов», является не вполне определенной. Вы сразу спросите: «А каков урожай этого дерева?» Если вы учтете, что урожай этого дерева составлял 40 бушелей, то убедитесь в том, что 27% указывает на более сильную степень повреждения плодов, чем следовало бы ожидать у дерева со столь высоким урожаем плодов. В данном случае нет какой-то общей для всего сада средней, с которой следовало бы сравнивать процент повреждения плодов каждого дерева; для этой цели для каждой отдельной величины урожая имеется свой ожидаемый процент поврежденности. Фактически теперь вместо средней базой сравнения становится регрессия. Эта новая база не остается одной и той же для всех деревьев, а меняется в зависимости от величины плодоношения.

Как и в упомянутом в параграфе 5 случае, по отдельным деревьям следует производить сравнение не непосредственно степени повреждения плодов, а различий их отклонений от регрессии. Степени повреждения плодов двух деревьев непосредственно сравнимы только в том случае, если их урожай одинаковы. Другими словами, эти отклонения являются ошибками тех оценок, которые подлежат сравнению. Из таблицы 45 видно, что в то время как деревья № 1 и 4 по фактическому проценту поврежденных плодов различаются на $59 - 53 = 6\%$, при сравнении их на основе величины урожая разность между ними уже будет $2,86 - 11,04 = -8,18\%$; это означает, что если бы эти деревья дали одно и то же количество плодов, то процент поврежденных плодов дерева № 1 по нашей оценке был бы на 8,18% меньше, чем у дерева № 4. В данном случае мы имеем другой способ сравнений, описанных в параграфе 5, при котором вместо сравнения приведенных значений Y используются отклонения от регрессии.

В нашем отношении к экспериментальным данным могут быть две крайности, которые следует избегать. Одна из них состоит в том, что уделяется слишком большое внимание мелким деталям выборочного варьирования и не используется возможность обобщения данных с последующими выводами относительно совокупности. Вторая крайность характеризуется пренебрежительным отношением к исходным данным и стремлением поскорей перейти к средним и к другим обобщенным показателям. Оба эти направления приводят к потере информации, содержащейся в экспериментальных данных. Осмотрительный исследователь всегда найдет время изучить каждый отдельный объект и результат его измерения. Он, конечно, сделает попытку отделить нормальное для данного случая варьирование материала от резко уклоняющихся наблюдений. Только после этого он приступит к обобщениям и к формулировке своих суждений о совокупности, в результате чего его выводы будут отнесены к четкому фону наблюдаемых фактов.

Пример 18. Другая группа из 12 деревьев, обследованных Ханберри и Ричардсоном, была подвергнута в течение всего сезона опрыскиванию мышьяковоокислым свинцом. В дополнение к этому было применено четырех- и пятикратное общее опрыскивание 1%-ной эмульсией минеральных масел и никотина сульфата при норме 1 пинта на 100 галлонов. Данные наблюдения приведены ниже. Предлагается проверить следующие результаты: $\Sigma X = 240$, $\Sigma Y = 384$, $\Sigma x^2 = 808$, $\Sigma y^2 = 1428$, $\Sigma xy = -582$; коэффициент регрессии $= -0,7203$; $\hat{Y} = 46,41 - 0,7203 X$; $Y - \hat{Y}$ для первого дерева $= 16,40\%$.

Урожай плодов в сотнях . . .	15	15	12	26	18	12	8	38	26	19	29	22
Процент поврежденных плодов	52	46	38	37	37	37	34	25	22	22	20	14

Пример 19. При помощи формулы параграфа 2 вычислите по данным таблицы 45 сумму $\Sigma d_{y \cdot x}^2 = 273,88$.

Пример 20. Следующие данные об общем весе и весе гребней 15-дневных петушков породы белый леггорн взяты из работы Снедекора и Бренемана [16].

Номер петушка	1	2	3	4	5	6	7	8	9	10
Общий вес (в г)	83	72	69	90	90	95	95	91	75	70
Вес гребня (в мг)	56	42	18	84	56	107	90	68	31	48

Определите выборочное уравнение регрессии $\hat{Y} = 60 + 2,302(X - 83)$.

Пример 21. Постройте по приведенным выше данным о петушках график, располагая общий вес по горизонтальной оси. Проведите линию регрессии.

8. Оценки интервалов и проверка нулевых гипотез. Познакомившись с оценкой значений, которые имеют параметры регрессии в совокупности, обратимся к вопросам об интервальных оценках и проверке гипотез относительно этих параметров. Эти вопросы решаются одинаково как для модели I, так и для модели II.

В соответствии с большим практическим значением этого показателя сначала обратимся к выборочному коэффициенту регрессии b , являющемуся оценкой β . Как мы знаем из параграфа 2 этой главы, при проведении случайных наблюдений дисперсия b оценивается величиной:

$$s_b^2 = s_{y \cdot x}^2 / \Sigma x^2.$$

Так, при наблюдениях над повреждениями плодов яблони (таблица 45)

$$s_b^2 = 27,388 / 924 = 0,0296; \quad s_b = 0,172\%.$$

Далее, так как величина $(b - \beta) / s_b$ подчинена закону распределения t при $n - 2$ степенях свободы, то с определенной уверенностью можно утверждать, что

$$b - t_{0,05} s_b \leq \beta \leq b + t_{0,05} s_b.$$

Для примера с яблонями имеем:

$$\text{число степеней свободы} = 10, \quad t_{0,05}^{\text{кр}} = 2,228,$$

$$t_{0,05} s_b = 2,228 \times 0,172 = 0,383;$$

$$b - t_{0,05} s_b = -1,013 - 0,383 = -1,396\% \text{ на } 100 \text{ плодов};$$

$$b + t_{0,05} s_b = -1,013 + 0,383 = -0,630\% \text{ на } 100 \text{ плодов и, наконец,} \\ -1,396 \leq \beta \leq -0,630.$$

Если мы скажем, что коэффициент регрессии совокупности находится внутри этих пределов, то такое утверждение будет правильным, если не считаться с выборками, противоречащими этому, которые возможны в одном случае из 20 испытаний.

Характерно то, что эти доверительные границы относятся к двум линиям регрессии с наклонами $-1,396$ и $-0,630$, которые перекрещиваются друг с другом и с выборочной линией регрессии в точке $O'(\bar{x}, \bar{y})$. Мы в данном случае уверены, что линия регрессии совокупности наклонена сверху вниз и что она имеет наклон между этими пределами. Но мы пока еще не установили пределов для уровня этой линии.

Пример 22. На графике рисунка 11 нанесите две линии, соответствующие доверительным пределам β .

Вместо интервальной оценки β нас может интересовать проверка некоторой нулевой гипотезы. Так как в нашем примере совершенно очевидно, что

$H_0: \beta=0$ должна быть отвергнута, то и возьмем в качестве иллюстрации именно этот случай; если будет испытываться какое-либо другое гипотетическое значение β , то вместо нуля следует в формуле взять это значение. Так как величина $(b-\beta)/s_b$ подчиняется закону распределения t , мы находим:

$$t = \frac{b-\beta}{s_b} = \frac{-1,013-0}{0,172} = -5,89; \text{ число степеней свободы} = n-2 = 10.$$

Знак t опускается, так как таблица значений t содержит в себе обе половины этого распределения. H_0 — отвергнута. Отсюда заключим, что в совокупности, из которой взята выборка, существует определенная регрессия процента поврежденных плодов на величину урожая; значение коэффициента регрессии, вероятно, содержится между $-0,630$ и $-1,396\%$ на 100 плодов.

Следующее по важности место занимают заключения относительно $\mu = \alpha + \beta x$. Наиболее простым случаем будет тот, когда $x=0$ и $\mu(=\alpha)$ оценивается непосредственно величиной \bar{y} . Так как σ^2 оценивается величиной $s_{\bar{y}.x}^2$, то дисперсия средней будет иметь оценку $s_{\bar{y}.x}^2/n$. Для нашего примера, в котором $\bar{y}=45\%$, имеем:

$$s_{\bar{y}.x} = s_{\bar{y}.x}^2/n = 27,388/12 = 2,282; \quad s_{\bar{y}.x} = 1,51\%.$$

Доверительный интервал μ при $x=0$ строится на основе $S_{\bar{y}.x}$ с $n-2$ степенями свободы. Мы имеем

$$\bar{y} - t_{0,05} s_{\bar{y}.x} \leq \alpha \leq \bar{y} + t_{0,05} s_{\bar{y}.x},$$

что для нашего примера дает

$$45 - 2,228 \times 1,51 \leq \alpha \leq 45 + 2,228 \times 1,51;$$

$$41,6 \leq \alpha \leq 48,4\%.$$

Вы можете удивиться тому, что средний квадрат \bar{y} не равен $s_{\bar{y}}^2/n$, как это должно быть, если основываться на материалах главы 3. За счет чего произошло уменьшение среднего квадрата до $s_{\bar{y}.x}^2/n$? В порядке первого и грубого объяснения можно сказать, что использование X обуславливает уменьшение этой оценки. Более строго $s_{\bar{y}.x}^2/n$ является оценкой σ^2/n , т. е. дисперсии выборочной средней из нашей нормальной совокупности при фиксированных X (модель I) или из нормальной совокупности двух переменных (модель II). Сюда включается и тот случай, который рассматривается нами, а именно $X=\bar{x}$. Причина уменьшения дисперсии будет ясной, если вы вспомните, что в модели I различные значения X фиксированы. При производстве выборок из этой модели мы все время имеем дело с одним и тем же рядом значений X . Поэтому Y имеет только варьирование, которое обусловлено ε . Так, если в наблюдениях над давлением крови у женщин был бы взят какой-либо другой, не вошедший в выборку возраст, например менее 30 лет, то выборочная средняя могла бы иметь значительно большее варьирование, чем то, которое обнаружилось в данной выборке. Ограничение, наложенное на выбор значений X , и является ценой, за которую покупается уменьшение среднего квадрата. В главе 7 будет показано, что и модель II приводит к той же самой оценке $\sigma_{\bar{y}.x}^2$.

В том случае, когда имеется некоторая гипотеза относительно α , она может быть проверена при помощи $s_{\bar{y}.x}$. Могло бы быть так, что в некотором большом районе производства яблок убыль плодов от повреждений молью в течение данного сезона была в среднем 50%. В целях исследования относительной степени повреждения сада, из которого взята наша выборка, можно было бы выставить гипотезу, что при нашем ряде значений X величина $\alpha=50\%$. В этом случае:

$$t = \frac{\bar{y}-\alpha}{s_{\bar{y}.x}} = \frac{45-50}{1,51} = -3,31; \text{ число степеней свободы} = 10.$$

H_0 будет отвергнута. Вывод мог бы быть таким: 1) опрыскивание дает больший эффект, чем обычная практика ведения садоводства в районе, или

2) данный сад по неизвестным причинам подвергся меньшему нападению вредителя, чем район в среднем, или 3) в этом саду было большее число плодов на 1 дерево, чем в среднем по району.

Если $X \neq \bar{x}$, то \bar{y} испытывает влияние второй причины варьирования. В данном случае может иметь ошибку не только выборочная средняя \bar{y} , но также и выборочный коэффициент регрессии b . Так как $\hat{Y} = \bar{y} + bx$, то средний квадрат этой величины является суммой двух независимых средних квадратов; для \bar{y} это будет $s_{\bar{y},x}^2/n$, а для b

$$\frac{s_{y \cdot x}^2 \cdot x^2}{\sum x^2}.$$

Первый средний квадрат один и тот же для всех X , но второй изменяется пропорционально x^2 . В результате мы имеем:

$$s_{\hat{Y}}^2 = \frac{s_{y \cdot x}^2}{n} + \frac{s_{y \cdot x}^2 \cdot x^2}{\sum x^2} = s_{y \cdot x}^2 \left(\frac{1}{n} + \frac{x^2}{\sum x^2} \right),$$

отсюда

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{1/n + x^2/\sum x^2}.$$

Для нашего примера $s_{y \cdot x} = \sqrt{27,388}$ и

$$s_{\hat{Y}} = \sqrt{27,388} \cdot \sqrt{1/12 + x^2/924} = \sqrt{2,282 + 0,02964x^2}.$$

Так, для деревьев, урожай которых примерно такой, как у дерева № 12, имеем $x=21$ и $s_{\hat{Y}}=3,92\%$, что заметно больше, чем $s_{\bar{y},x}=1,51\%$ при $x=0$.

Второй член выражения $s_{\hat{Y}}^2$ часто бывает относительно небольшим, если берется интервал изменения X выборки, но в случае выхода из этого интервала этот член, зависящий от x^2 , может приобрести преобладающее значение.

Причина того, что $s_{\hat{Y}}$ увеличивается по мере того, как X удаляется от \bar{x} , станет вполне ясной, если вы вдумаетесь в сущность доверительного интервала для β . Два предельных наклона определяют собой две предельные линии, проходящие через начало координат и расходящиеся по мере увеличения $|x|$. Это расхождение доверительных границ, относящееся к β , и является причиной расширения доверительного интервала \hat{Y} .

Соответственно любому значению \hat{Y} , являющемуся оценкой значения μ , можно определить оценку интервала для μ :

$$\hat{Y} - t_{0,05} s_{\hat{Y}} \leq \mu \leq \hat{Y} + t_{0,05} s_{\hat{Y}}.$$

Так может возникнуть потребность в оценке среднего процента поврежденных плодов яблони μ в точке $X=30$ сотням плодов с дерева. В этом случае

$$x = X - \bar{x} = 30 - 19 = 11 \text{ сотням плодов}$$

$$Y = \bar{y} + bx = 45 - 1,013 \times 11 = 33,86\%$$

$$t_{0,05} s_{\hat{Y}} = 2,228 \times \sqrt{2,282 + 0,02964 \times 11^2} = 5,40\%$$

$$33,86 - 5,40 \leq \mu \leq 33,86 + 5,40$$

и, наконец,

$$28,46\% \leq \mu \leq 39,26\%.$$

В этом случае можно сказать, что при $X=30$ сотням плодов средняя совокупности оценивается как 33,86% поврежденных плодов при 95%-ных доверительных пределах от 28,46 до 39,26%. Этот доверительный интервал на рисунке 12 представлен отрезком AB .

Если провести подобные этому расчеты для различных значений и отложить соответствующие доверительные пределы выше и ниже выборочной линии регрессии, то можно получить доверительный пояс, или зону, окаймлен-

ную кривыми линиями, как это представлено на рисунке 12. Эти кривые являются ветвями одной гиперболы. Мы уверены с вероятностью 0,95 в том, что для любого X величина μ_y находится внутри этого пояса. Этот рисунок дает возможность ясно видеть, что по мере отхода X от \bar{x} неточность предсказаний значения μ увеличивается.

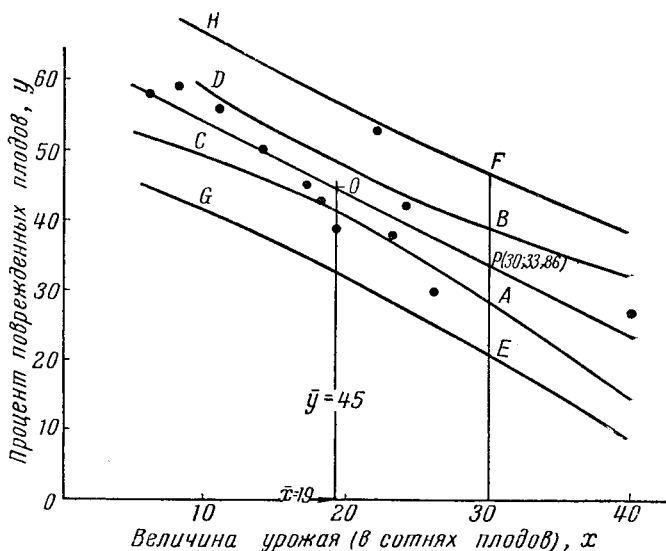


Рис. 12. Доверительная зона для μ — $ABCD$ и для Y — $EFGH$.
Данные об урожаях яблок.

Когда кто-либо прибегает к подобного рода предсказаниям, то он обычно имеет в виду предугадывание отдельного наблюдения, но не значения $\mu = \alpha + \beta x$. В этом случае возникает дополнительный источник неопределенности—случайный элемент ϵ . В результате средний квадрат Y приобретает еще один член и становится таким:

$$s_Y^2 = s_{y \cdot x}^2 + \frac{s_y^2 \cdot x^2}{n} + \frac{s_y^2 \cdot x^2 \cdot x^2}{\Sigma x^2},$$

откуда

$$s_Y = s_{y \cdot x} \cdot \sqrt{1 + 1/n + x^2/\Sigma x^2}.$$

Если вы, например, желаете предугадать процент поврежденных плодов на дереве, урожайность которого 30 сотен плодов, то вы имеете

$$t_{0,05} s_Y = 2,228 \sqrt{27,338} \sqrt{1 + 1/12 - 11^2/924} = 12,35\%.$$

При этом результате и при $\hat{Y} = 33,86\%$ соответствующий доверительный интервал будет

$$33,86 - 12,85 \leq \mu \leq 33,86 + 12,85,$$

или

$$21,01\% \leq \mu \leq 46,71\%,$$

что на рисунке 12 показано в виде отрезка EF . В данном случае мы делаем вывод, что для деревьев, приносящих 3000 плодов, процент поврежденных плодов находится между 21,01 и 46,71, если не считаться с одним из 20 случаев, который может встретиться при выборочном наблюдении, когда этот процент выйдет за указанные пределы.

Продолжая этот вычислительный процесс, можно определить и нанести на график доверительный пояс для таких Y . Из рисунка видно, что все наблюдаемые выборочные точки оказались внутри этого пояса. Вообще же следует ожидать, что около 5% их может быть и вне этого пояса.

Так как $(\hat{Y}-\mu)/s_{\hat{Y}}$ и $(Y-\mu)/s_Y$ подчинены закону распределения t , то на основе последнего может быть проверена любая имеющая смысл гипотеза относительно μ .

В целях иллюстрации все вышеприведенные вероятностные суждения были сделаны так, как будто они не зависят одно от другого. Положения относительно \bar{y} и b независимы, но, с другой стороны, ясно, что вероятности, относящиеся к \hat{Y} и Y , определены частично на основе тех же наблюдений данной выборки, которые определили \bar{y} и b . Это обстоятельство обуславливает сложное переплетение всех таких вероятностных суждений. Достаточно, например, указать, что если будет установлено, что средняя совокупности оказалась вне доверительных пределов, основанных на выборках, то вероятность того, что значение Y , соответствующее регрессии в совокупности, окажется вне доверительных пределов, основанных на \hat{Y} , может быть много больше, чем 0,05.

Пример 23. При определении регрессии веса гребня на общий вес петушка в примере 20 получено: $n=10$, $\bar{x}=83$ г, $\bar{y}=60$ мг, $\Sigma x^2=1000$, $\Sigma y^2=6854$ и $\Sigma xy=2302$. Найдите 5%-ные доверительные пределы для \bar{y} , считая, что ряд общих весов остается тем же самым. Ответ: 49,8—70,2 мг.

Пример 24. Для данных о петушках $b=2,302$. Проверьте гипотезу, что $\beta=0$. Ответ: $t=5,22$; $P < 0,01$.

Пример 25. Так как совершенно очевидно, что в совокупности существует определенная регрессия веса гребня на общий вес петушка, то установите доверительные пределы для соответствующего коэффициента регрессии. Ответ: 1,28—3,32 мг на 1 г.

Пример 26. Предскажите средний вес гребня в совокупности для 100-граммовых цыплят: Ответ: 99,1 мг; при 95%-ных пределах—79—119,2 мг.

Пример 27. Определите 95%-ные доверительные пределы для предсказания веса гребня у случайно выбранного 100-граммового цыпленка. Ответ: 61,3—136,9 мг.

9. Сокращенные методы вычисления регрессии. Поскольку это относится к примерам настоящей главы, вычисления пока не представляли для нас большого труда. На практике же расчеты регрессии часто становятся столь утомительными, что необходимо использовать счетную машину или применить в той или иной форме кодирование. Сокращенные способы вычисления регрессии остаются в основном теми же самыми, что и описанные в главе 5, но только с введением нового момента. Данные таблицы 46 послужат нам для иллюстрации этих способов.

Когда имеется возможность применить счетную машину, методы параграфа 2 главы 5 могут быть легко распространены и на случай вычисления регрессии. На большинстве счетных машин можно получить сумму для каждой переменной в процессе вычисления ее суммы квадратов; таким образом, в конце одной серии операций на месте итоговых показателей одновременно получаются ΣX и ΣX^2 . Точно так же сумма произведений $50 \times 128 + 64 \times 159 + \dots + 65 \times 104 = 108\,530$ может быть взята с соответствующего счетчика после всей последовательности умножений без выписки каждого отдельного произведения. Соответствующие указания вы найдете или в руководствах, или у ближайшего инструктора.

ТАБЛИЦА 46

Начальные веса и привесы (в г) 15 самок крыс в возрасте от 28 до 81 дней при рационе, насыщенном протеином

	Номер крысы														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Начальный вес	50	64	76	64	74	60	69	68	56	48	57	59	46	45	65
Привес	128	159	158	119	133	112	96	126	132	118	107	103	82	103	104

На некоторых счетных машинах может быть совмещено в один процесс вычисление сразу двух сумм, двух сумм квадратов и суммы произведений. Если

есть возможность использовать перфорационные машины, то это представляет большие удобства при обработке больших выборок (200 или более рядов наблюдений) [2]. Результаты вычислений на какой-либо из таких машин выписываются по следующей форме. Ниже приведены вычисления регрессии при помощи счетной машины, произведенные на основе данных таблицы 46.

$$\begin{array}{lll}
 \Sigma x = 901 & \Sigma Y = 1783 & n = 15 \\
 \bar{x} = 60,07 \text{ г} & \bar{y} = 118,87 \text{ г} & \\
 \Sigma X^2 = 55465 & \Sigma Y^2 = 218297 & \Sigma XY = 108530 \\
 (\Sigma X)^2/n = \underline{54120,07} & (\Sigma y)^2/n = \underline{211939,27} & (\Sigma X)(\Sigma Y)/n = \underline{107098,87} \\
 \Sigma x^2 = 1344,93 & \Sigma y^2 = 6357,73 & \Sigma xy = 1431,13
 \end{array}$$

Одно из преимуществ счетных машин заключается в возможности избежать выписки большого количества чисел. Это служит важным дополнением к точности и скорости вычислений на машине. В приведенную схему вычислений выписываются только те данные, которые будут необходимы в дальнейшем. Отметим один новый момент, а именно — поправочный член суммы произведений. Он равен произведению ΣX и ΣY , а не квадрату того или иного из них и является совершенно обоснованным способом корректирования произведений. Не следует упускать из виду то положение, что этот поправочный член может быть и больше ΣXY , в результате чего Σxy будет отрицательной. Это будет указывать на то, что линия регрессии имеет наклон книзу.

Шесть величин

$$n, \bar{x}, \bar{y}, \Sigma x^2, \Sigma y^2 \text{ и } \Sigma xy$$

содержат в себе всю информацию, необходимую для полного определения регрессии. Мы, например, имеем:

$$b = \Sigma xy / \Sigma x^2 = 1431,13 / 1344,93 = 1,0641.$$

Таким образом, привес таких же, как в нашем опыте, самок крыс может быть частично предсказан в связи с тем, что каждый грамм увеличения начального веса соответствует 1,0641 г привеса. Однако

$$\Sigma d_{y \cdot x}^2 = \Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2 = 6357,73 - (1431,13)^2 / 1344,93 = 4834,88$$

$$s_{y \cdot x}^2 = \Sigma d_{y \cdot x}^2 / (n - 2) = 4834,88 / 13 = 371,91$$

$$s_b^2 = s_{y \cdot x}^2 / \Sigma x^2 = 371,91 / 1344,93 = 0,2765$$

$$s_b = \sqrt{0,2765} = 0,526$$

$$t = b / s_b = 1,0641 / 0,526 = 2,02.$$

Построив по этим данным график, подобный графику примера 22, вы можете убедиться в том, что в данном случае s_b велико относительно b . Так как t несколько не достигает 5%-ного уровня, то у нас нет большой уверенности в отношении реального существования регрессии в совокупности. Однако нельзя и игнорировать данную выборочную оценку, так как мы имеем здесь дело со свидетельством в пользу определенного соотношения между начальным весом и привесом. В случае данного примера мы имеем некоторые соображения относительно плана эксперимента и поэтому должны будем вернуться к этим результатам в следующем параграфе.

Теперь вы подготовлены к чтению глав 10 и 11, излагающих дисперсионный анализ, но если хотите, вы можете углубиться и в изучение глав 7 и 9.

В тех случаях, когда для вычислений не удастся использовать счетную машину, полезно применить метод кодирования, описанный в главе 5. Каждую переменную следует привести к ряду целых чисел, имеющих цен-

тром нуль и размах варьирования, выраженный в 20—50 единицах. При этом не только достигается относительная простота вычислений, но также получаются результаты с достаточной для обычных условий точностью. Процесс расчетов показан в таблице 47. Эти данные были взяты из исследования относительно тягового усилия, необходимого для тяги плугов с обычными для трактора скоростями [3]. Особенности почвы и глубина вспашки в той мере, в какой это возможно в эксперименте, контролировались. При этих ограничениях опыта, как показало графическое изображение данных, тяговое усилие было только примерно пропорциональным скорости.

ТАБЛИЦА 47

Тяговое усилие и скорость движения прицепных плугов
Применение кодирования для вычисления регрессии

Скорость (миль в час) V	Тяговое усилие (фунты) D	Числа кода		Квадраты чисел кода		Произведения чисел кода XU
		$(10V-30)/2$ (округлено) X	$(D/10)-50$ (округлено) U	X^2	U^2	
0,9	425	-10	-8	100	64	80
1,3	420	-8	-8	64	64	64
2,0	480	-5	-2	25	4	10
2,7	495	-2	0	4	0	0
3,4	540	2	4	4	16	8
3,4	530	2	3	4	9	6
4,1	590	6	9	36	81	54
5,2	610	11	11	121	121	121
5,5	690	12	19	144	361	228
6,0	680	15	18	225	324	270
34,5	5460	48	64	727	1044	841
		-25	-18			
		23	46			

$$\Sigma X = 23$$

$$\Sigma Y = 46$$

$$\bar{x} = 2,3 \text{ единицы кода}$$

$$\bar{y} = 4,6 \text{ единицы кода}$$

$$\bar{v} = \frac{2 \times 2,3 + 30}{10} = 3,46 \text{ мили в час}$$

$$\bar{d} = (4,6 + 50) \times 10 = 546 \text{ фунтам}$$

$$\Sigma X^2 = 727$$

$$\Sigma Y^2 = 1044$$

$$(\Sigma X)^2/n = 52,9$$

$$(\Sigma Y)^2/n = 211,6$$

$$\Sigma x^2 = 674,1$$

$$\Sigma y^2 = 832,4$$

$$\Sigma v^2 = 674,1 \times (2/10)^2 = 26,964$$

$$\Sigma d^2 = 832,4 \times 10^2 = 83\,240$$

$$\Sigma XY = 841$$

$$(\Sigma X)(\Sigma Y)/n = 105,8$$

$$\Sigma xy = 735,2$$

$$\Sigma vd = 735,2 \times 2/10 \times 10 = 1470,4$$

Во избежание больших чисел и десятичных знаков в данном случае кодирование было доведено до крайних пределов. Несомненно, кое-кто предпочтет работать с десятичными знаками и большими цифрами; аргумент, выставляемый против кодирования, состоит в том, что при проведении этого процесса возможны ошибки. Однако хорошее ознакомление с этим способом делает его уже не столь непривлекательным, и имеются такие данные, кодирование которых, безусловно, выгодно.

При обратном переходе от кода возникают некоторые новые моменты. Суммы квадратов при этом переходе должны определяться при помощи квадратов отношений, входящих в код, а сумма произведений — при помощи их произведения. Если вам это не представляется достаточно обоснованным, произведите проверку, извлекая квадратный корень из суммы квадратов, делая переход от кода обычным порядком и производя после этого возведение в квадрат этих результатов.

Коэффициент регрессии $b = \Sigma vd / \Sigma v^2 = 1470,4 / 26,964 = 54,53$ указывает на то, что тяга на крюке увеличивается в среднем на 54,53 фунта на каждую милю в час возрастающей скорости. Формула для оценки тяги при различных скоростях такая:

$$\hat{D} = 546 + 54,53 (V - 3,46) = 54,53 V + 357,3 \text{ фунта.}$$

Так как применение этой формулы регрессии может сводиться к прогнозу тяги на крюке для отдельных испытаний, в этом случае уместно определить дисперсию s_D^2 для случайно выбранного D :

$$\Sigma (D - \hat{D})^2 = \Sigma d^2 - (\Sigma vd)^2 / \Sigma v^2 = 83240 - 1470,4^2 / 26,964 = 3056$$

$$s_{d.v}^2 = \Sigma (D - \hat{D})^2 / (n - 2) = 3056 / 8 = 382,0$$

$$s_D^2 = s_{d.v}^2 (1 + 1/n + v^2 / \Sigma v^2) = 382,0 (1 + 0,1 + v^2 / 26,964) = 420,2 + 14,167 v^2.$$

В качестве примера произведем оценку тяги на крюке трактора, делающего 5 миль в час:

$$\hat{D} = 54,53 \times 5 + 357,3 = 630 \text{ фунтам.}$$

$$v = 5 - 3,46 = 1,54 \text{ мили в час.}$$

$$s_D^2 = 420,2 + 14,167 \times 1,54^2 = 453,8.$$

Стандартная ошибка = 21,3 фунта: $t_{0,05} = 2,306$. Доверительные пределы будут

$$630 - 21,3 \times 2,306 \leq \mu \leq 630 + 21,3 \times 2,306$$

$$581 \leq \mu \leq 679.$$

Пример 28. Скоростные рекорды (миль в час), достигнутые на автомобильных гонках в честь традиционного праздничного дня в Индианополисе за период 1911—1955 гг., были следующие:

Год	1911	1912	1913	1914	1915	1916	1917	1918
Скорость	74,6	78,7	75,9	82,5	89,8	83,3	(гонки не было)	
	1919	1920	1921	1922	1923	1924	1925	1926
	88,1	88,6	89,6	94,5	91,0	98,2	101,1	95,9
	1927	1928	1929	1930	1931	1932	1933	1934
	97,5	99,5	97,6	100,4	96,6	104,1	104,2	104,9
	1935	1936	1937	1938	1939	1940	1941	1942
	106,2	109,1	113,6	117,2	115,0	114,3	115,1	(гонки не было)
	1943	1944	1945	1946	1947	1948	1949	1950
	(гонки не было)			114,8	116,3	119,8	121,3	124,0
	1951	1952	1953	1954	1955			
	126,2	128,9	128,7	130,8	128,2			

Вычислите для 29 гонок, прошедших до второй мировой войны, т. е. за период 1911—1941 гг., величины $\Sigma x^2 = 2325,03$, $\Sigma y^2 = 4039,81$, $\Sigma xy = 2971,23$; $\hat{Y} = 1,278X + 63,51$ миль в час (введите для годов код, вычитая из каждого 1900).

Пример 29. Произведите оценку скоростного рекорда в 1946 и 1952 гг., имея в виду, что здесь требуется предугадать индивидуальные рекорды, а не \hat{Y} . *Ответ:* в 1946 г. $Y = 122,3$ мили в час при 95%-ном доверительном интервале 118,9—125,7; в 1952 г. $Y = 130,0$; 126,0—134,0 мили в час. Заметьте, что в последнем случае доверительный интервал покрывает наблюдаемый рекорд.

Пример 30. На графике, на осях которого взяты года и рекордные скорости, можно провести кривые, соответствующие доверительным интервалам ежегодных рекордов. Для гонок довоенного периода эти кривые будут наиболее близко подходить к линии регрессии в точке $\bar{x} = 26,59$ (т. е. без дублирования в точке 1926, 59 г) и будут все больше и больше отклоняться от линии регрессии по мере того, как X будет отходить от своей средней в ту и другую сторону. Эти кривые будут представлять две ветви гиперболы. Если взять $t_{0,05}$, то при условии случайной выборки около 95% рекордов должно попасть в область между этими кривыми.

Пример 31. Для *временных рядов*, к числу которых принадлежит наша последовательность рекордов скоростей, модель I не совсем подходит; здесь ϵ не подбираются случайно, и каждый из них может не быть независимым от ближайших. Изменения в правилах соревнования, новые технические открытия, завоевание рекорда одним и тем же лицом в ряде следующих друг за другом гонок, отсутствие гонок во время войн — все это ставит под сомнение правомерность модели I. В этой книге мы не будем касаться сложных проблем временных рядов.

Пример 32. Возникла необходимость предсказать урожай 29 двойных и тройных гибридов кукурузы, подвергшихся сильному поражению со стороны второго поколения клопа-черепашки [11]. В основу прогноза были положены данные о средних урожаях гибридов сортов с родительскими линиями, выращенных в одинаковых условиях. Была использована формула, установленная Дженкинсом [12]. Предсказанные урожай X и фактические экспериментально установленные урожай Y , определенные нами по точечной диаграмме, были следующие: 20,20; 24,29; 25,34; 26,35; 29,43; 31,11; 33,24; 34,35; 34,45; 35,34; 36,34; 36,36; 37,35; 37,46; 37,42; 41,41; 42,43; 43,43; 43,55; 50,53; 50,55; 51,54; 51,55; 52,42; 55,55; 56,54; 56,51; 57,50; 71,61 бушеля на акр. Восстановите точечную диаграмму. Считаете ли вы, что регрессия прямолинейна? Допуская, что это именно так, определите выборочную регрессию Y на X , $\hat{Y} = 0,79X + 9,4$.

Пример 33. В процессе исследования вопроса морозостойчивости яблони Старк [17] намеревался определить удельную теплоемкость Y ветвей этого дерева. Лабораторная техника для определения этого трудна и неточна. Так как более половины веса ветвей падает на воду и так как теплоемкость воды большая, то он решил проверить, нельзя ли определить Y по проценту влаги в ветвях X , имея в виду, что последняя может быть установлена довольно просто и точно. Результаты 21 пары измерений Старка следующие: 49,4, 0,646; 50,1, 0,644; 50,8, 0,665; 51,2, 0,670; 51,5, 0,666; 51,9, 0,653; 52,5, 0,669; 52,7, 0,657; 53,1, 0,689; 53,6, 0,669; 55,7, 0,685; 56,3, 0,696; 57,0, 0,700; 58,0, 0,690; 58,5, 0,711; 59,2, 0,704; 59,7, 0,696; 61,3, 0,713; 62,0, 0,719; 63,1, 0,731; 64,9, 0,731. Вычислите: $\bar{x} = 55,83\%$; $\Sigma x^2 = 431,31$; $\bar{y} = 0,6859$; $\Sigma y^2 = 0,0142$; $\Sigma xy = 2,3536$; $\hat{Y} = 0,005457X + 0,3812$. Так как $s_{y^2} = 0,027$ и $s_{y^2, X} = 0,0086$, то примерно третья часть варьирования Y не находится ни в какой связи с X . Так как X определяется с весьма небольшой ошибкой, Старк сделал вывод, что для определения Y с достаточной точностью следует сначала установить процент влаги в ветвях яблони, после чего уже можно произвести оценку по уравнению регрессии удельной теплоемкости.

Пример 34. Выведите формулу $\Sigma xy = \Sigma XY - (\Sigma X \cdot \Sigma Y)/n$. Указание: начните с равенств $X = x + \bar{x}$ и $Y = y + \bar{y}$. Умножьте и сложите, учитывая, что $\Sigma x = \Sigma y = 0$.

10. Использование регрессии при планировании опытов. Статистический контроль. Как уже указывалось в параграфе 10 главы 3, регрессия является одним из способов уменьшения ошибки опыта, не поддающейся непосредственному контролю. Часто варьирование материала, которое в процессе проведения эксперимента не может быть уменьшено, связано с некоторой количественной переменной; измерение этой переменной дает исследователю возможность использовать регрессию в целях увеличения получаемой им информации. Примером этого могут служить данные таблицы 46 о привесе крыс. Хотя и было хорошо известно, что увеличение веса обычно связано с начальным весом животного, все же не представлялось возможным поставить под контроль этот внешний фактор, т. е. нельзя было подобрать крыс, имеющих одинаковый исходный вес. Однако очевидно, что можно было бы произвести измерение этой переменной X , т. е. учесть начальный вес и определить регрессию привеса на нее. Какой выигрыш при этом был бы получен? Без учета этого фактора ошибка опыта будет основываться на среднем квадрате:

$$s^2 = \Sigma y^2 / (n - 1) = 6357,73 / 14 = 454,12.$$

Но при использовании сведений относительно начального веса животных средний квадрат ошибки снижается до

$$s_{y,x}^2 = 371,91.$$

Продолжая аргументацию параграфа 10 главы 3, можно сказать, что информация в пересчете на одно животное увеличилась в отношении $454,12/371,91 = 122\%$. Другими словами, можно сказать, что для достижения одной и той же надежности результатов при неиспользовании данных о начальном весе крыс этих животных понадобится на 22% больше. Проведение такого не требующего больших затрат предварительного измерения начального веса и его использование в регрессии увеличивают информацию, доставляемую экспериментом, на 22%.

Даже в тех случаях, когда контроль, т. е. выравнивание некоторых условий в опыте возможен, он все же не всегда будет желательным. Так, в опыте с тяговым усилием, конечно, нет необходимости ограничивать условия опыта выбором только одной определенной скорости трактора.

В то же время из опыта будет получена очень незначительная информация, если в качестве экспериментальной ошибки взята величину

$$s^2 = \Sigma d^2 / (n - 1) = 83240 / 9 = 9249.$$

Этот эксперимент предназначен для использования регрессии, и соответствующая этому оценка случайной ошибки опыта будет

$$s_{d.v}^2 = 382,0.$$

Этот опыт не только дает более широкую основу для практического использования его результатов, но и увеличивает получаемую информацию в отношении $9249/382,0 = 2421\%$, т. е. информация на одно испытание в этом случае на 2321% больше, чем она была бы, если бы данные о скоростях не были использованы.

Теперь ясно, что во многих случаях, когда не представляется удобным или желательным фактическое выравнивание (контроль) некоторой переменной, связанной с той переменной, которая подлежит изучению, можно ввести *статистический контроль* [19]. Путем измерения величины, относящейся к сопутствующему условию, и использования регрессии можно получить такой конечный результат, который был бы получен при сохранении в опыте этой величины на постоянном уровне. По всей вероятности, ошибка $s_{y,x}$ не будет больше, чем s_y , которую можно было бы получить, если бы удалось фактически поставить выборку под контроль, сделав X константным. Становится совершенно очевидным, что во многих случаях статистический контроль является более чем желательным; в этом случае вместо искусственно созданных условий изучаются реальные условия, наблюдения охватывают более широкий круг условий, в связи с чем и выводы имеют более широкую основу, и, наконец, получаются сведения об изменчивости двух величин вместо одной, а также о соотношении между ними. В каждой области исследования необходимо на основе опыта изучить вопрос о наиболее пригодном для данного случая виде контроля и вопрос о том, как наиболее быстро и с меньшими затратами можно получить некоторую добавочную информацию.

Пример 34 а. Данные о цыплятах примера 20 были взяты из первоначальных наблюдений над 13 петушками, у которых общий вес (X г) и вес гребня (Y мг) был таким: 83; 54; 72; 42; 69; 29; 78; 37; 69; 18; 90; 84; 90; 56; 95; 107; 95; 90; 91; 68; 75; 31; 70; 48; 76; 41. Из этих данных следует: $n = 13$, $\bar{x} = 81,0$, $\bar{y} = 54,2$ мг, $\Sigma x^2 = 1218,00$; $\Sigma y^2 = 8292,31$; $\Sigma xy = 2834,00$. Условившись решение вопроса базировать на весе гребня, определите увеличение информации при использовании общего веса в качестве статистического контроля. *Ответ:* 348%.

Пример 35. По данным о повреждении плодов яблони молью, приведенным в таблице 45, произведите оценку снижения точности (информации), которое получится в результате отказа от использования величины урожая в качестве статистического контроля. *Ответ:* 73,5%.

Пример 36. В отчете Отдела определения урожайности для 1594 ферм восточной части штата Южная Дакота за 1942 г. приведены данные об X = акров земли на ферме и Y = акров посевов кукурузы. Средний квадрат для числа акров посевов кукурузы был $s_y^2 = 1859$, в то время как средний квадрат для отклонений от регрессии был $s_{y \cdot x}^2 = 100$. Если вы возьмете случайную выборку в 100 ферм, по которой проведете оценку среднего числа акров посевов кукурузы, то $s_{\bar{y}}^2$ будет 18,59.

Сколько ферм нужно взять в выборку, чтобы, используя данные о размере ферм в качестве статистического контроля, получить такое же количество информации?
Ответ: 54 фермы.

11. Разложение суммы квадратов зависимой переменной. Расчеты регрессии становятся особенно наглядными, если произвести подразделение величины ΣY^2 на 3 части, каждая из которых имеет свое содержание и применение. Вы уже получили навык в разложении ΣY^2 на 2 части: на $(\Sigma Y)^2/n$ и остаток Σy^2 ; после этого производится разделение Σy^2 на $(\Sigma xy)^2/\Sigma x^2$ и $\Sigma d_{y \cdot x}^2$. Это означает, что вы производите разложение ΣY^2 на 3 части:

$$\Sigma Y^2 = (\Sigma Y)^2/n + (\Sigma xy)^2/\Sigma x^2 + \Sigma d_{y \cdot x}^2.$$

Каждая из этих частей может рассматриваться в полном соответствии с отдельными частями ординат Y . Для иллюстрации возьмем выборку, представленную в таблице 48, и график рисунка 13.

ТАБЛИЦА 48

Данные, взятые для иллюстрации разложения ΣY^2 на части

X	2	4	6	8	10	12	14	$\Sigma X = 56$
Y	4	2	5	9	3	11	8	$\Sigma Y = 42$
$n = 7$	$\bar{x} = 8$	$\bar{y} = 6$	$\Sigma x^2 = 412$	$\Sigma y^2 = 68$	$\Sigma xy = 56$			

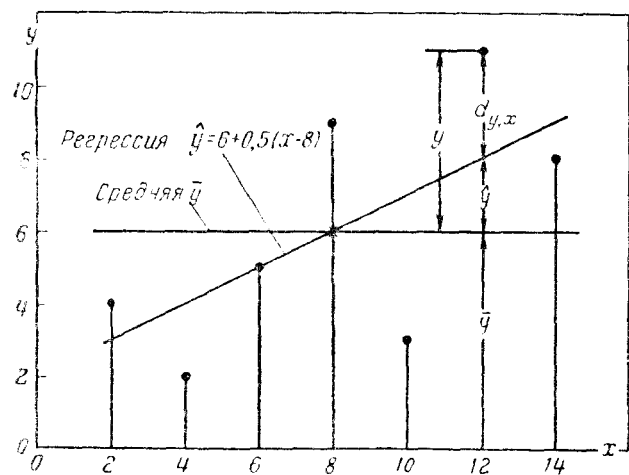


Рис. 13. График по данным таблицы 48. Ордината для $X=12$ разделена на две части: $\bar{y}=6$ и $y=5$. После этого y делится на $\hat{y}=2$ и $d_{y \cdot x}=3$. Таким образом, $Y = \bar{y} + \hat{y} + d_{y \cdot x} = 6 + 2 + 3 = 11$.

На примере шестой ординаты показано деление ее на 3 отрезка:

$$Y = \bar{y} + \hat{y} + d_{y \cdot x}.$$

Каждая из всех остальных ординат может быть разделена подобным же образом, хотя введение отрицательных отрезков делает геометрическую интерпретацию менее наглядной. Длины всех таких отрезков даны в таблице 49.

а на рисунке 14 они для наглядности представлены раздельно. Легко видеть, что по каждой строке таблицы (включая и две итоговые строки) сумма последних 3 чисел дает число в колонке Y .

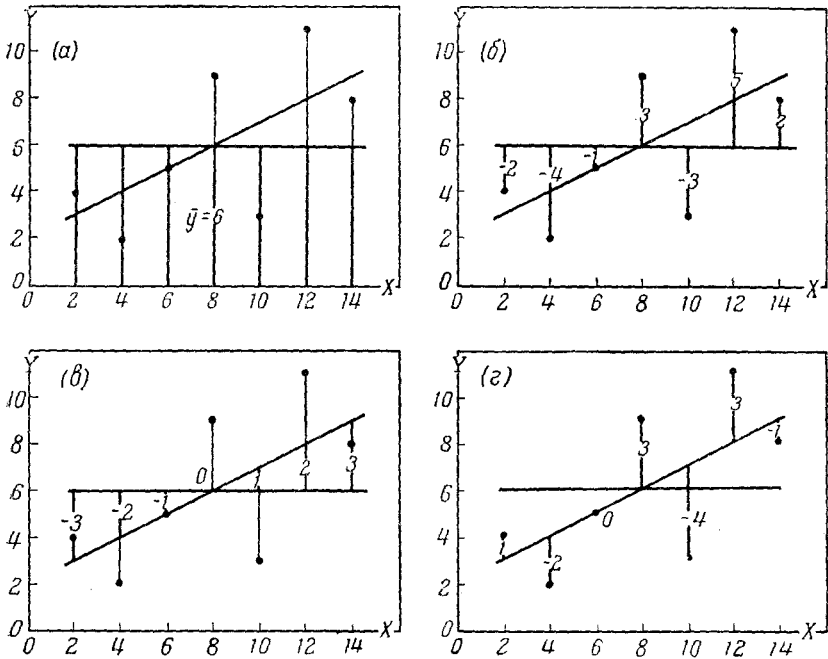


Рис. 14. *a*—здесь даны отрезки, равные средней \bar{y} , вычитание которых из Y приводит к отклонениям y от этой средней, представленным на графике *б*. Сумме квадратов этих отрезков $7 \times 6^2 = 252$ соответствует 1 степень свободы, относящаяся к поправке $(\Sigma Y)^2/n = 42^2/7 = 252$; *б*—здесь даны отклонения от средней y , которые в дальнейшем на графиках *в* и *г* разлагаются на части; *в*—жирные линии относятся к отклонениям от средней y значений \hat{Y} , вычисленным по регрессии. Эти значения полностью определяются на основе \bar{y} , b и ряда фиксированных значений x . Сумма квадратов $\Sigma \hat{y}^2 = 28$ является частью ΣY^2 , приписываемой регрессии; *г*—случайные отклонения от регрессии $d_{y.x}$. Каждое $d_{y.x}$, сложенное с соответствующими \bar{y} и \hat{y} графиков *а* и *в*, образует ординату Y рисунка 13. Сумма квадратов этих отклонений $\Sigma d_{y.x}^2 = 40$; ей соответствует 5 степеней свободы, что дает оценку σ^2 в виде $s_{y.x}^2 = 40/5 = 8$.

ТАБЛИЦА 49

Длины ординат таблицы 48 и отрезки, на которые они подразделяются

Номер парного наблюдения	Ордината Y	Средняя y	Отклонение \hat{y}	Отклонение от регрессии $d_{y.x}$
1	4	6	-3	1
2	2	6	-2	-2
3	5	6	-1	0
4	9	6	0	3
5	3	6	1	-4
6	11	6	2	3
7	8	6	3	-1
	42	42	0	0
Сумма квадратов	320	252	28	40

Замечательно то, что хотя

$$Y = \bar{y} + \hat{y} + d_{y \cdot x},$$

все же сумма квадратов будет

$$\Sigma Y^2 = \Sigma \bar{y}^2 + \Sigma \hat{y}^2 + \Sigma d_{y \cdot x}^2,$$

так как 3 члена, являющиеся суммами удвоенных произведений, равны нулю (пример 41). Суммы квадратов ординат $\Sigma Y^2 = 320$ и отклонений от регрессии $\Sigma d_{y \cdot x}^2 = 40$ нам уже знакомы. Остается теперь только доказать тождественность $(\Sigma Y)^2/n$ с $\Sigma \bar{y}^2$ и $(\Sigma xy)^2/\Sigma x^2$ с $\Sigma \hat{y}^2$.

Во-первых:

$$(\Sigma Y)^2/n = n \cdot \frac{(\Sigma Y)^2}{n^2} = n\bar{y}^2 = \Sigma \bar{y}^2.$$

Таким образом, поправка на среднюю $(\Sigma Y)^2/n$ является не чем иным, как n -кратным квадратом средней. Причина того, что этот факт не используется при проведении вычислений, состоит в том, что ошибки округления \bar{y} оказывают более заметное влияние при возведении в квадрат и при умножении на n .

Во-вторых:

$$\frac{(\Sigma xy)^2}{\Sigma x^2} = \frac{(\Sigma xy)^2}{(\Sigma x^2)^2} \cdot \Sigma x^2 = b^2 \Sigma x^2 = \Sigma b^2 x^2 = \Sigma \hat{y}^2.$$

Таким образом, сумма квадратов, приписываемая регрессии, сводится к сумме квадратов n отклонений \hat{Y} от их средней \bar{y} . Конечно, непосредственное вычисление $\Sigma \hat{y}^2$ каждый раз, когда нужно определить сумму квадратов регрессии, будет являться довольно утомительной работой, в связи с чем вам теперь понятно, в чем ценность сокращенного способа вычислений этой суммы по формуле $(\Sigma xy)^2/\Sigma x^2$.

Я надеюсь, что вы оценили все изящество таких расчетов регрессии. В этом случае каждая формула приобретает смысловое содержание и не является просто нагромождением букв и цифр. Сопоставьте каждую из них с данными отдельных колонок таблицы и с линейными отрезками рисунка, а также с формулировками определений, которые даны были ранее. Знакомство со всем этим является не только полезным, но и приводит в восхищение стройностью теории.

Разложение ΣY^2 наводит на мысль произвести соответствующее разложение общего числа степеней свободы на 3 части, относящиеся к каждой части ΣY^2 (табл. 50). Такая связь отдельных сумм квадратов с соответствующими степенями свободы является одним из самых важных предложений профессора Р. А. Фишера [8]. Этот метод показан в таблице 50, называемой таблицей дисперсионного анализа; с содержанием данного термина вы более близко познакомитесь в главах 10 и 11.

Обычно первая строка такой таблицы опускается; чаще встречающаяся и более удобная форма дисперсионного анализа представлена в таблице 51.

ТАБЛИЦА 50

Дисперсионный анализ для Y из таблицы 48

Источник варьирования	Символ	Число степеней свободы	Сумма квадратов	Средний квадрат
Средняя	\bar{y}	1	$(\Sigma Y)^2/n = 252$	$s^2_{y \cdot x} = 8$
Регрессия	b	1	$(\Sigma xy)^2/\Sigma x^2 = 28$	
Отклонение от регрессии	$d_{y \cdot x}$	$n - 2 = 5$	$\Sigma d_{y \cdot x}^2 = 40$	
Итого	Y	$n = 7$	$\Sigma Y^2 = 320$	

$$\Sigma y^2 = 28 + 40 = 68; \text{ число степеней свободы} = n - 1 = 6$$

Дисперсионный анализ для Y из таблицы 48

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Регрессия	1	28	—
Отклонения от регрессии	5	40	8
Отклонения от средней	6	68	11,3

Возвращаясь к обсуждению рисунка 14, можно видеть, что сумма отклонений от горизонтальной прямой в графиках (а) и (б), как и от линии регрессии в графике (з), равна нулю. Это положение будет иметь место при выборе любой линии, проходящей через точку (\bar{x}, \bar{y}) , являющуюся началом координат для отклонений. Сумма квадратов отклонений в графике (б) будет меньше соответствующей суммы от любой другой *горизонтальной* прямой. Но если вы будете производить вращение этой прямой в направлении против часовой стрелки около центра (x, y) , то сумма квадратов будет становиться все меньшей и меньшей до тех пор, пока эта прямая не достигнет положения линии регрессии. Именно введение второй переменной, такой, как величина урожая в случае нашего примера с повреждением плодов яблони, придает определенное содержание этому вращению. До тех пор пока рассматривается факт повреждения плодов яблони, взятый сам по себе и без учета различий в урожае, познавательное значение имеет только горизонтальная прямая, выражающая средний уровень явления. Действительно, в этом случае все точки могут быть размещены на какой-либо прямой, подобной горизонтальной оси OY на рисунке 8. В данном случае отклонения от начала их отсчета будут не чем иным, как отклонениями от средней, рассмотренными в предыдущих главах. Но добавление сведений об урожайности деревьев служит основой для включения в круг наших идей представления о регрессии. Большая часть варьирования процента поврежденных плодов находит свое объяснение в величине урожая. И только оставшаяся часть варьирования, т. е. та, которая ассоциируется с отклонениями от регрессии, является, согласно нашим взглядам, выборочным варьированием повреждения яблок. Выбор такой прямой, для которой сумма отклонений равна нулю, а сумма квадратов отклонений минимальна, является простым результатом логического продолжения тех же идей, которые приводят к выбору средней в качестве соответствующего центра концентрации наблюдений. Положением, связывающим наши идеи об осереднении при единичной переменной и о регрессии при двух переменных, является принцип наименьших квадратов.

Пример 37. Дос [5] определил «плотность» содержания пигмента меланина в коже 24 самцов лягушек и их вес. «Некоторые из этих 24 самцов... были взяты с весьма темной или бледной окраской для того, чтобы получить достаточно большую ширину варьирования». Следовательно, так как отбор был ориентирован на эту плотность, данная переменная уже не будет случайной и не может оцениваться на основе выборки (параграф 3). Если же допустить случайность веса, то он может быть взят за переменную Y . Фактические данные следующие:

Плотность X	0,13	0,15	0,28	0,58	0,68	0,31	0,35	0,58
Вес Y	13	18	18	18	18	19	21	22
Плотность X	0,03	0,69	0,38	0,54	1,00	0,73	0,77	0,82
Вес Y	22	24	25	25	25	27	27	27
Плотность X	1,29	0,70	0,38	0,54	1,08	0,86	0,40	1,67
Вес Y	28	29	30	30	35	37	39	42

Вычислите $\bar{x}=0,6225$ единицы, $\bar{y}=25,79$, $\Sigma x^2=3,3276$, $\Sigma y^2=1214,96$, $\Sigma xy=40,022$.

Пример 38. По данным примера 37 проверьте гипотезу $\beta=0$. *Ответ:* $t=3,81$; $P < 0,01$.

Пример 39. Дисперсионный анализ для веса лягушек будет таким:

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Средняя	1	15 965,04	
Регрессия	1	481,36	
Отклонения	22	730,60	33,21
Итого	24	17 177,00	

Пример 40. В какой мере измерение плотности меланина X свободно от ошибки: После приготовления вытяжки из кожи лягушек при помощи колориметра была определена интенсивность окраски, после чего эти данные графически путем приведены к средним плотностям. Эти результаты являлись средними из 3—6 наблюдений. Ошибка такого рода измерений обычно довольно значительна, в связи с чем оценка регрессии может оказаться смещенной в сторону снижения. Если бы исследователь не ставил перед собой задачу изучения крайних границ плотности, то регрессия плотности на вес могла бы быть не только несмещенной, но и содержала бы в себе большую информацию.

Пример 41. Докажите, что все суммы удвоенных произведений в разложении ΣU равны нулю. Из них только для одной суммы возникают затруднения, а именно для суммы произведений $\Sigma \hat{y}d_{y,x}$; в этом случае возьмите $\hat{y}=bx$ и $d_{y,x}=y-\hat{y}=y-bx$. Перемножьте и просуммируйте, учитывая, что $\Sigma xy=b\Sigma x^2$.

12. Применение термина «регрессия» Гальтоном. В своих работах о наследственности Гальтон определил понятие о регрессии. Формулируя «Закон универсальной регрессии», он [9] говорит: «Каждая особенность человека находит свое частичное выражение и у его кровных потомков, но в *среднем* в меньшей степени». Его друг и последователь Карл Пирсон [15] произвел более тысячи наблюдений над ростом, длиной руки до локтя и стопы у лиц, связанных семейным родством. Рисунок 15 дает регрессию роста сыновей на рост отцов. Хотя видна определенная тенденция, по которой высокие отцы имеют высоких сыновей, однако средний рост сыновей у группы высоких отцов все же меньше, чем рост этих отцов. Здесь имеет место *регрессия*, или тенденция к возврату роста сыновей к среднему росту всех мужчин.

В качестве одного из доказательств Гальтон приводит результаты опыта по наследованию такого признака, как величина семян душистого горошка. Для посева он отобрал по 100 семян для каждой из 7 групп по их размеру и после этого установил среднюю величину семян у потомства. Он приводит следующие данные о диаметре семян родительских растений душистого горошка и средний диаметр семян дочерних растений: 21 и 17,5; 20 и 17,3; 19 и 16,0; 18 и 16,3; 17 и 15,6; 16 и 16,0; 15 и 15,3; все эти данные приведены в сотых дюйма.

Коэффициент регрессии величины семян дочерних растений на величину семян родительских растений, т. е. Y на X , здесь равен 0,3. Гальтон думал, что регрессия в таких случаях вообще должна быть около одной третьей, так как в среднем потомки отклоняются от общей средней примерно только в одном из трех случаев, когда отклонения данного признака у родителей случайно оказались одного и того же направления.

Если вы изучали генетику и, в частности, учение о гетерозиготности, доминантности и пр., то вы знаете, что данное явление не столь просто. Дополнительное усложнение вопроса о росте возникает в связи с тенденцией вступления в брак мужчины и женщины примерно одинакового относительного роста. Представляется, что указанный закон универсальной регрессии объединяется с подобными рода тенденциями у видов сохранять свой характер-

ные черты. Если высокий отец производит более высоких дочерей, а высокие дочери рожают более высоких сыновей, то не будет никакой регрессии, т. е. возврата к среднему уровню, и в этом случае не будет иметь место ни то, ни другое из двух противоположных положений, относящихся к сохранению особенностей и к возврату к среднему уровню. Средний уровень является мерой той характеристики, к которой направлена регрессия, в то время как средняя изменчивость устанавливает константность при переходе от поколения к поколению при этой регрессии.

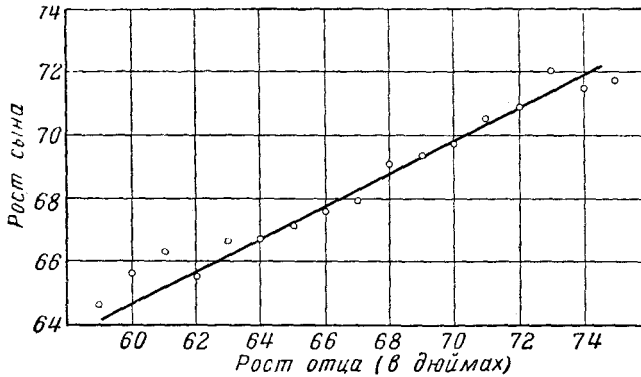


Рис. 15. Регрессия роста сына на рост отца [12].

$$\hat{Y} = 0,516X + 33,73. \text{ По данным 1078 семей.}$$

Пример 42. Уоррен [20] исследовал величину регрессии в партеногенетических поколениях тли. Он измерил фронтальную ширину насекомого в единицах, составляющих 0,221 мм; результаты измерения родителей (X) и потомков (Y) были следующими: 14,8; 13,7; 18,4; 14,6; 17,5; 14,5; 16,3; 14,3; 16,9; 14,7; 15,8; 14,2; 15,4; 14,8; 15,2; 15,0; 15,9; 15,8; 16,1; 15,7; 16,3; 15,1; 16,7; 15,6; 17,0; 15,5; 17,8; 15,9; 18,1; 15,4; 18,5; 15,8; 18,9; 15,2; 15,8; 16,2; 16,0; 16,0; 16,2; 16,7; 16,6; 16,6; 16,8; 16,9; 17,5; 16,4; 18,2; 16,3; 18,8; 16,7; 19,0; 16,9; 15,7; 17,2; 16,1; 17,0; 16,5; 17,6; 16,8; 17,4; 17,1; 17,8; 17,7; 17,5; 18,0; 17,7; 18,8; 17,9; 19,5; 17,6; 16,0; 18,1; 17,5; 18,8; 19,0; 18,4; 18,9; 19,5. Вычислите коэффициент регрессии 0,4469.

13. Регрессия модель IA : σ пропорциональна X и $\alpha = 0$. Принцип наименьших квадратов приводит к различным моделям, частично изменяющимся в связи со сделанным допущением относительно ϵ . В моделях I и II величина ϵ появляется в результате случайного и независимого отбора при условии общей σ . Но не так редко бывает, что σ меняется при изменении X . В модели, которую мы будем сейчас рассматривать, σ прямо пропорциональна X , т. е. $\sigma = kX$, или $\sigma/X = k$, где k — некоторое постоянное отношение. Хотя эта модель и сама по себе имеет некоторые применения, все же главной целью введения ее здесь является стремление осветить вопрос о широко распространенной практике применения долей и отношений взамен регрессии.

Модель IA отличается от модели I только в двух отношениях: тем, что $\sigma = kX$, и тем, что линия регрессии обязательно проходит через начало координат $O(0, 0)$. Эта точка относится как к линии регрессии в совокупности, так и к линии в выборке. Поэтому данная модель будет такой:

$$Y_{ij} = \beta X_i + E_{ij}; \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m,$$

где подстрочные указатели означают принадлежность Y к тем величинам X , у которых имеются соответствующие указатели, и где ϵ_{ij} взяты случайно из $N(0, \sigma_i)$ при $\sigma_i = kX_i$. Второй указатель указывает на то, что для каждого из n значений X взято m значений Y . Данные таблицы 52 относятся к определению процента земель под посевами кукурузы на фермах восточной части штата Южная Дакота в 1942 г. Были выбраны 5 наиболее часто встречающихся размеров ферм, после чего для каждого из них случайным порядком было взято 5 наблюдений.

Число акров под посевами кукурузы на 25 фермах штата Южная Дакота (1942 г.)
при выбранных размерах фермы

Размер фермы (в акрах) X	Число акров под посевами кукурузы Y	Размах варьирова- ния	Стандартное отклонение s_y	Отношение s_y/X	Отношение Y/X
80	25	22	8,05	1,01	0,312
	10				0,125
	20				0,250
	32				0,400
	20				0,250
160	60	40	14,58	0,91	0,375
	35				0,219
	20				0,125
	45				0,281
	40				0,250
240	65	55	21,51	0,90	0,271
	80				0,333
	65				0,271
	85				0,354
	30				0,125
320	70	80	29,15	0,91	0,219
	110				0,344
	30				0,094
	55				0,172
	60				0,188
400	75	105	39,21	0,98	0,188
	35				0,088
	140				0,350
	90				0,225
	110				0,275
Средняя	56,28				0,243

$N = mn = 25$; $b = \frac{\sum(Y/X)}{N} = 0,243$ акра посевов кукурузы на 1 акр всех посевов фермы.

В порядке дальнейшего объяснения символики, принятой в этой модели, укажем, что X_i является последовательно $X_1=80$, $X_2=160$ и т. д.; i пробегает значения от 1 до 5. Для каждого X_i указатель j также принимает значения от 1 до 5. Например, $Y_{11}=25$, $Y_{12}=10$, $Y_{21}=60$, $Y_{34}=85$.

Размахи варьирования отдельных групп Y указывают на то, что σ увеличивается при увеличении X . На этот же факт указывает и рисунок 16. Для более полного изучения этого вопроса произведено вычисление для каждой группы s_y и найдено отношение s_y к X . Эти отношения столь близки к константности, что позволяют допустить, что в совокупности отношение σ/X константно. Вместе с этим представляется обоснованным предположить, что точка $O(0, 0)$ лежит на этой линии регрессии. Две эти характерные особенности определяют собой модель IA при условии, что регрессия прямолинейна, и совокупности, из которых взяты выборки, нормальны.

Рассматривая b как основанную на принципе наименьших квадратов оценку β , получаем уравнение выборочной регрессии $\hat{Y}=bX$.

Способ вычисления b оказывается неожиданно совсем иным, чем тот, который применялся прежде; b является просто средним из выборочных отношений Y/X , приведенных в последней колонке таблицы. Это будет

$$b = \frac{\sum (Y/X)}{N} = 0,243,$$

где N — общее число наблюдений, т. е. $N=mn$. Уравнение выборочной регрессии будет $\hat{Y}=0,243X$. Здесь выборочный коэффициент регрессии характеризует средний рост площади под посевами кукурузы при увеличении площади всех посевов фермы. В соответствии с этим содержанием $Y/X=0,243$ является среднее из отношений Y/X для отдельных ферм. Отметим, что эти

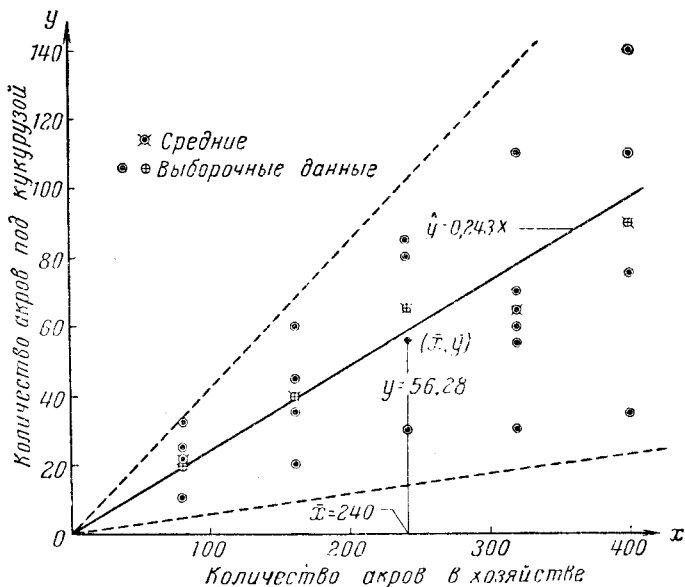


Рис. 16. Регрессия площади посевов кукурузы на общую площадь фермы. Модель IA. Пунктирные линии дают 95%-ный доверительный интервал для Y ; их наклон $b \pm t_{0,05} s_{Y/X}$.

отношения довольно стабильны и варьируют только от 0,088 до 0,400. Такая стабильность обусловлена тем, что числитель Y и его стандартное отклонение σ примерно одинаково увеличиваются по мере увеличения X .

Характерной особенностью модели IA является и то, что точка (\bar{x}, \bar{y}) может не находиться на линии регрессии. Вы, например, видите из нашего рисунка, что точка $(240; 56,28)$ оказалась ниже линии регрессии. Это объясняется тем, что \bar{y}/\bar{x} , вообще говоря, не одно и то же, что и средняя из Y/X .

Доверительный интервал для β и проверка какой-либо подходящей гипотезы относительно этой величины могут быть произведены при помощи стандартного отклонения отношений Y/X . Соответствующие выводы не будут вполне точными, так как обычно такие отношения не имеют нормального распределения, но это не должно нас слишком беспокоить. Средний квадрат из 25 отношений вычисляется обычным путем, что приводит к $s_{Y/X}^2 = 0,008069$. Отсюда средний квадрат b будет:

$$s_b^2 = \frac{s_{Y/X}^2}{N} = \frac{0,008069}{25} = 0,0003228,$$

$$s_b = 0,0180; \text{ число степеней свободы} = N - 1 = 24.$$

5%-ная интервальная оценка для β строится обычным способом:

$$b - t_{0,05} s_b \leq \beta \leq b + t_{0,05} s_b,$$

что у нас дает $0,206 \leq \beta \leq 0,280$. В модели IA это будет являться также и интервальной оценкой среднего отношения Y к X . Если будет взята некоторая гипотеза относительно β , то она может быть проверена при помощи вычислений:

$$t = \frac{b - \beta}{s_b}; \text{ число степеней свободы} = N - 1.$$

Другие критерии и оценки усложняются в связи с изменением значений σ . Здесь мы не будем рассматривать соответствующие методы, но их можно найти в [6].

Пример 43. Данган [7] изучал влияние на урожай зерна уменьшения доли листовой поверхности у кукурузы в штате Иллинойс. В опыте было произведено удаление листьев в фазе молочной спелости; в 6 группах растений общая площадь листовой поверхности была соответственно снижена на $16^2/3$; $33^1/3$; 50; $66^2/3$; $88^1/3$ и 100%. Соответствующее снижение урожая зерна было 13, 17, 28, 22 и 50%. Считая, что линия регрессии проходит через начало координат и что σ пропорциональна площади листовой поверхности, определите выборочную регрессию процента снижения урожая на процент уменьшения листовой поверхности. *Ответ:* $\hat{Y} = 0,52X$; $s_b = 0,06$.

14. О распространенных злоупотреблениях долями и отношениями.

В связи с весьма широким применением долей и отношений следует

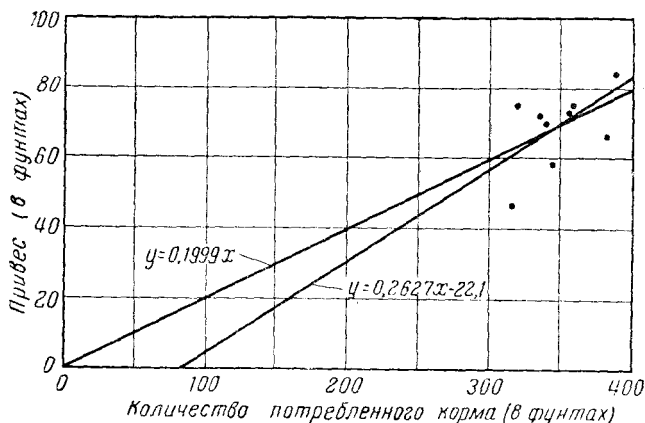


Рис. 17. График регрессии привеса на потребленный корм. Данные Кремптона о свиньях группы 1.

рассмотреть положение, когда подобного рода практика становится сомнительной. Применяя соответствующие методы регрессии, Кремpton [4] наблюдал изменение в весе (Y) свиней при скармливании им определенных количеств корма (X). У группы I (табл. 53) $n=10$; $\bar{x} = 353,5$ фунта кормов; $\bar{y} = 70,8$ фунта привеса; $\Sigma x^2 = 7334,5$, $\Sigma y^2 = 1207,5$ и $\Sigma xy = 1927,0$. Рисунок 17 дает точечную диаграмму и соответствующую линию регрессии $\hat{Y} = 0,2627X - 22,1$. Кремpton обращает внимание на тот факт, что эта регрессия не проходит через начало координат в связи с тем, что для поддержания жизни животного требуется минимальное количество корма, которое не может дать никакого привеса. Более того, выборочные данные согласуются с гипотезой о постоянстве дисперсии Y во всем интервале изменения X ; вместе с тем здесь отсутствует сколько-нибудь заметная погрешность при измерении X . Поэтому представляется вполне правомерным принять здесь для регрессии модель I.

И все же многие экспериментаторы предпочитают иметь дело с отношением Y/X , игнорируя тот факт, что они вступают в этом случае в противоречие с условиями применения модели IA. Они берут среднюю $b=0,1999$ фунта привеса на фунт потребленных кормов, не учитывая того, что тем самым они недооценивают преимуществ обильного кормления свиней и переоценивают оплату низших норм кормления. Они упускают из виду то обстоятельство,

что регрессия при модели IA проходит через начало координат, что явно противоречит процессу образования привеса животных.

ТАБЛИЦА 53

Данные о свиньях группы I в опыте Кремптона

Потребленные корма (в фунтах) X	Привес (в фунтах) Y	Отношение Y/X	Вычисленный привес	
			модель I $\hat{Y} = 0,2627X - 22,1$	модель IA $\hat{Y} = 0,1999X$
382	66	0,1728	78,3	76,4
335	72	0,2149	65,9	67,0
388	84	0,2165	79,8	77,6
316	47	0,1487	60,9	63,2
319	75	0,2351	61,7	63,8
399	87	0,2180	82,7	79,8
358	75	0,2095	72,0	71,6
355	73	0,2056	71,2	71,0
344	59	0,1715	68,3	68,8
339	70	0,2065	67,0	67,8
Средняя 353,5	70,8	0,1999	70,8	70,7

Упускается из виду и другое следствие неправильного применения модели IA, относящееся к изменению величины Y/X . В этой модели предполагается, что Y/X константно для всех X , но в основной модели I отношение Y/X увеличивается при увеличении X . Это можно установить, вычислив в нашем случае уравнение выборочной регрессии $\hat{Y} = 0,2627X - 22,1$ и разделив его на X :

$$\frac{\hat{Y}}{X} = 0,2627 - \frac{22,1}{X}.$$

Ясно, что \hat{Y}/X увеличивается, если X возрастает, что означает увеличение относительного привеса при более обильном кормлении свиней. Применение при подобных обстоятельствах модели IA приводит к потере некоторой части информации, содержащейся в опытных данных.

До тех пор пока нет веских оснований считать, что уравнение регрессии должно проходить через начало координат и что оно прямо пропорционально X , отношение Y/X имеет сомнительную ценность. Его можно применять в работах, имеющих характер первого приближения, но в экспериментальных работах, требующих осторожности, следует применять более эффективный статистический метод.

15. Итоги. При имеющих наибольшее применение моделях I и II вся информация о регрессии сосредоточена в 6 выборочных величинах n , \bar{x} , Σx^2 , Σy^2 и Σxy :

1. Коэффициент регрессии Y на X : $b = \Sigma xy / \Sigma x^2$.
2. Уравнение выборочной регрессии Y на X : $\hat{y} = a + bx$.
3. Приведенное значение Y для X . Приведенное $Y = Y - bx$.
4. Сумма квадратов, относящаяся к регрессии:

$$(\Sigma xy)^2 / \Sigma x^2 = \Sigma y^2.$$

5. Сумма квадратов отклонений от регрессии:

$$\Sigma y^2 - (\Sigma xy)^2 / \Sigma x^2 = \Sigma d_{y \cdot x}^2.$$

6. Средний квадрат отклонения от регрессии:

$$\Sigma d_{y \cdot x}^2 / (n - 2) = s_{y \cdot x}^2.$$

7. Выборочная стандартная ошибка Y для X :

$$s_{\hat{y} \cdot x} = s_{y \cdot x} / \sqrt{n}.$$

8. Выборочная стандартная ошибка коэффициента регрессии:

$$s_b = s_{y \cdot x} / \sqrt{\sum x^2}.$$

9. Выборочная стандартная ошибка \hat{Y} :

$$s_{\hat{Y}} = s_{y \cdot x} \sqrt{1/n + x^2/\sum x^2}.$$

10. Выборочное стандартное отклонение Y :

$$s_Y = s_{y \cdot x} \sqrt{1 + 1/n + x^2/\sum x^2}.$$

Л И Т Е Р А Т У Р А

1. Anderson R. L., Bancroft T. A., Statistical Theory in Research. McGraw-Hill Book Company, Inc., New York, 1952.
2. Brandt A. E., Journal of the American Statistical Association, 23, 291, 1928.
3. Collins E. V., Transactions, American Society of Agricultural Engineers, 14, 39, 1920.
4. Crampton Earle Wilcox, Journal of Nutrition, 7, 305, 1934.
5. Dawes Ben, Journal of Experimental Biology, 18, 26, 1941.
6. Deming W. Edwards, Statistical Adjustment of Data. John Wiley and Sons, Inc., New York, 1943.
7. Dungan George H., Plant Physiology, 9, 749, 1934.
8. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925—1950.
9. Galton Frances, Natural Inheritance. Macmillan and Company, London, 1889.
10. Hansberry T. Roy, Richardson Charles H., Iowa State College Journal of Science, 10, 27, 1935.
11. Holbert J. R., Flint W. P., Bigger J. H., Dungan G. H., Iowa State College Journal of Science, 9, 413, 1935.
12. Jenkins M. T., Journal of the American Society of Agronomy, 1934.
13. Kendall M. G., The Advanced Theory of Statistics. Charles Griffin and Company, Ltd., London, 1947.
14. Mood Alexander McFarlane, Introduction to the Theory of Statistics. McGraw-Hill Book Company, Inc., New York, 1950.
15. Pearson Karl, Lee Alice, Biometrika, 2, 357, 1903.
16. Snedecor George W., Breneman W. R., Iowa State College Journal of Science, 19, 333, 1945.
17. Stark Arvil L., Plant Physiology, 8, 168, 1933.
18. Swanson Pearl P., Leverton Ruth, Gram Mary, Roberts Harriett, Pesek Isabel, Journal of Gerontology, 10, 41, 1955.
19. Wallace H. A., Snedecor George W., Correlation and Machine Calculation. Iowa State College Official Publication 30, No. 4, 1931.
20. Warren Ernest, Biometrika, 1, 129, 1902.
21. Wentz J. B., Stewart R. T., Journal of the American Society of Agronomy, 16, 534, 1924.

Глава 7

КОРРЕЛЯЦИЯ

1. Введение. Совокупности с двумя переменными часто представляют большой интерес и не столько потому, что в этом случае одна величина предопределяется другой, сколько потому, что здесь проявляется взаимосвязь между этими двумя количественными величинами. Эта взаимосвязь нуждается в определении даже в тех случаях, когда нет оснований предполагать зависимость одной переменной от другой. Выдвижение Гальтоном этой проблемы возникло в связи с его учением о наследственных признаках — учением, привлекавшим внимание научных кругов в течение последней половины 19-го столетия. Развитие этой теории и ее приложений шло под знаком соревнования между биологами и математиками. В результате оказалось, что предмет *корреляции* входит в одну группу с концепциями регрессии, в связи с чем многие авторы совсем не делают различий между ними.

Пирсоном и Ли [17], которые собрали данные более чем о тысячи британских семей, были опубликованы данные о корреляции между ростом, состоянием между концами растянутых большого и указательного пальцев и длиной предплечья. В качестве одной из иллюстраций собранных ими данных служит таблица 54, где приведены результаты измерения роста братьев и сестер. Так как здесь нет определенной подчиненности одной переменной другой, то эти переменные вместо прежних обозначений X и Y будут обозначаться X_1 и X_2 . Вам не доставит никаких затруднений произвести проверку расчетов сумм квадратов и произведений.

Здесь вводится новый показатель — *выборочный коэффициент корреляции*, общепринятое обозначение которого r . Он является численным выражением общей схожести по наблюдаемому признаку детей одних и тех же родителей, т. е. тенденции более высоким сестрам иметь и более высоких братьев. Это можно представить графически по образцу рисунка 18.

ТАБЛИЦА 54

Рост (в дюймах) братьев и сестер.
Данные взяты из выборки 1401 семьи, произведенной Пирсоном и Ли

№ семьи	1	2	3	4	5	6	7	8	9	10	11
рост брата X_1	71	68	66	67	70	71	70	73	72	65	66
рост сестры X_2	69	64	65	63	65	62	65	64	66	59	62

$$n = 11; \bar{x}_1 = 69; \bar{x}_2 = 64; \Sigma x_1^2 = 74; \Sigma x_2^2 = 66; \Sigma x_1 x_2 = 39.$$

$$r = \Sigma x_1 x_2 / \sqrt{\Sigma x_1^2 \cdot \Sigma x_2^2} = 39 / \sqrt{74 \times 66} = 0,558. \text{ У Пирсона и Ли } r = 0,553.$$

Значение $r=0,558$ отражает тот факт, что точки, изображающие наблюдения, составляют группу, направленную из нижнего левого угла к верхнему правому углу графика, а не разбросаны случайно по всей площади этого рисунка. Точки в целом образуют эллиптическую фигуру с главной осью, направленной вверх и вправо. Биологически $r=0,558$ характеризует общее наследование роста у этих родственных лиц.

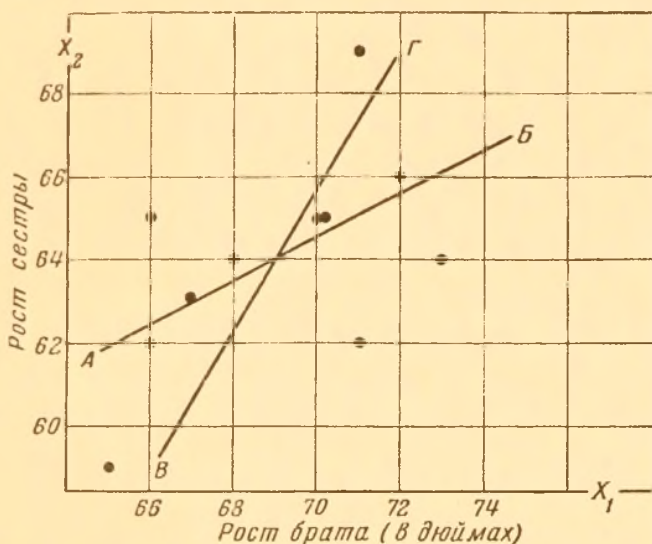


Рис. 18. Точечная диаграмма роста 11 пар братьев и сестер; $r=0,558$.

Пример 1. Вычислите r для следующих парных наблюдений:

$$X_1: 1, 2, 3, 4, 5;$$

$$X_2: 3, 5, 7, 9, 11.$$

Представьте эти данные графически по образцу рисунка 18.

Пример 2. Проверьте $r=0,91$ при таких данных:

$$X_1: 2, 5, 6, 8, 10, 12, 14, 15, 18, 20;$$

$$X_2: 1, 2, 2, 3, 2, 4, 3, 4, 4, 5.$$

Постройте эллиптическую фигуру из этих точек.

Пример 3. Показать, что $r=0,20$ при таких данных:

$$X_1: 3, 5, 8, 11, 12, 12, 17;$$

$$X_2: 11, 5, 6, 8, 7, 18, 9.$$

Присмотритесь к разбросу этих точек на диаграмме.

Пример 4. По данным об урожае яблонь в таблице 45 $\Sigma x^2=924$; $\Sigma y^2=1222$; $\Sigma xy=-936$. Вычислите $r=-0,88$.

2. Выборочный коэффициент корреляции, r . Выражение $\Sigma x_1 x_2 / \sqrt{\Sigma x_1^2 \Sigma x_2^2}$ может принимать значения между -1 и $+1$ в зависимости от того, какой является связь между переменными в совокупности, из которой взята выборка. Отрицательное значение r , подобно такому же значению b , указывает на наклон эллипса точек вниз и вправо, т. е. на то, что большим значениям одной переменной соответствуют малые значения другой переменной, и наоборот. Для того чтобы вы смогли приобрести некоторое представление о свойствах коэффициента корреляции r , ниже на рисунке 19 приводятся несколько простых таблиц с соответствующими графиками. В каждой из этих таблиц $n=9$, $\bar{x}_1=12$, $\bar{x}_2=6$, $\Sigma x_1^2=576$, $\Sigma x_2^2=144$, и только $\Sigma x_1 x_2$ меняется в соответствии со значением r . Так как $\sqrt{\Sigma x_1^2 \Sigma x_2^2} = \sqrt{576 \times 144} = 288$, то в каждой из этих таблиц коэффициент корреляции легко определяется путем вычисления $\Sigma x_1 x_2$ с последующим делением ее на 288 (или при наличии счетной машины умножением на $1/288 = 0,0034722\dots$).

В случае *A* 9 точек расположились строго по прямой линии; при этих условиях $r=1$. Эта линия является «вырожденным» эллипсом — она имеет длину, но не имеет ширины. В этом случае соотношение между двумя переменными подчиняется строгой закономерности: любое изменение одной из них будет сопровождаться пропорциональным изменением другой. В случае *B* имеет место некоторое отклонение от линейности; здесь эллипс имеет удли-

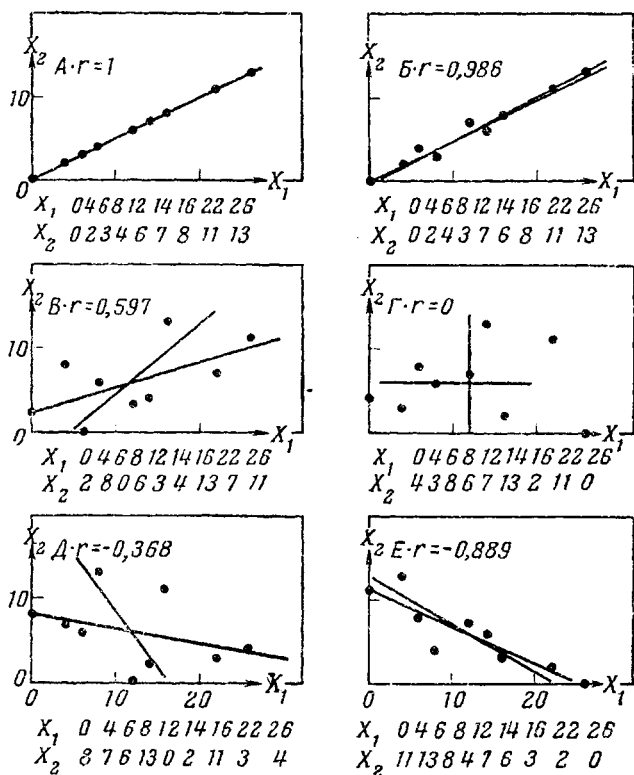


Рис. 19. Точечные диаграммы и корреляции, изменяющиеся в пределах от 1 до $-0,889$.

ненную и узкую форму и r несколько меньше 1. В случае *B* эллипс расширяется, после чего он достигает круговой формы в случае *Г*, где $r=0$. Это означает отсутствие какой-либо связи между двумя переменными. Случай *D* и *E* представляют отрицательные корреляции, стремящиеся к -1 . Подытоживая, можно сказать, что тонина эллипса точек выражает размер r , в то время как наклон оси эллипса сверху вниз или снизу вверх определяет его знак. Следует обратить внимание на то, что наклон оси эллипса зависит от шкал, выбранных для переменных и отлагаемых по осям координат графика; поэтому он не может служить непосредственным измерителем величины r . Такой мерой является концентрация точек вблизи оси эллипса, что означает высокую коррелированность, в особенности если при этом имеется и высокая регрессия.

Я думаю, вы согласитесь с тем, что высокие корреляции, будут ли они положительными или отрицательными, легко устанавливаются непосредственно из графика. Не так легко сделать глазомерную оценку корреляции, если абсолютное значение r около 0,5; при r между $-0,3$ и $+0,3$ даже направление наклона эллипса может быть для вас неуловимым. При таких малых выборках, как наши, даже единичная точка может быть причиной заметных различий. Если, например, в случае *Г* точка (26,0) изменила бы свое положение на (26,9), то r увеличился бы до 0,505. Этот факт подчеркивает то положение, что выборочный коэффициент корреляции из совокуп-

ности двух переменных, корреляция в которой ρ (греческое «ро» обозначает параметр, соответствующий r), при малом n очень изменчив. Выборочное распределение r будет рассмотрено в параграфе 5 этой главы. Для того чтобы сделать примерную оценку значения r по таблице, выберите некоторые крайние значения одной переменной и установите, не будут ли они соответствовать крайним значениям другой переменной. Если такой тенденции не будет обнаружено, то r , по всей вероятности, довольно мал.

Полная корреляция ($r = 1$) в биологических исследованиях, как правило, не встречается, хотя столь высокие значения r , как 0,99, отнюдь не являются невозможным случаем. В каждой области исследования встречается обычно некоторый определенный интервал для корреляции. Наследственные признаки, подобные росту, обычно имеют корреляции между 0,35 и 0,55. В работах по улучшению пород животных r в среднем около 0,35 [3]. Пирсон и Ли определили «органические корреляции», что означает корреляции между такими признаками, как рост и расстояние между концами растянутых большого и указательного пальцев у одного и того же человека; они находятся в пределах от 0,60 до 0,83. Брандт [2] определил выборочную корреляцию, равную 0,986, между живым и убойным весом у 533 свиней. Эвард и др. [6] нашли $r = -0,68$ между средним суточным привесом свиньи и количеством кормов, требуемых для получения 1 фунта привеса.

Очевидно, что суждение о достаточной или недостаточной величине корреляции должно строиться на основе сравнения со значениями корреляции, встречаемыми в данной области исследования, и при учете теоретических границ ± 1 .

3. Нормальная совокупность двух переменных: модель регрессии II. Предполагается, что r определяется по выборке из некоторой совокупности двух переменных, но это мало что говорит об ее структуре (параграф 7 главы 6). Более полное представление об этой структуре можно получить, рассматривая большую выборку, например приведенную в таблице 55. Эти данные взяты из статьи Гальтона [10], опубликованной в 1888 г., когда впервые было введено понятие о «соотношении» («corelation»).

При рассмотрении этой таблицы можно заметить 5 ее особенностей. 1. Каждый ряд и каждый столбец внутри таблицы представляет собой распределение частот. Точно так же столбец справа, озаглавленный f_2 , является итоговым распределением частот X_2 — длины предплечья, а третий снизу ряд дает итоговое распределение частот X_1 — роста. 2. Частоты концентрируются внутри эллиптической площади с главной осью, направленной вправо снизу вверх. Это означает, что нет очень низких мужчин с длинным предплечьем и нет очень высоких мужчин с коротким предплечьем. 3. Частоты сгущены вблизи главной оси эллипса, достигая своего максимума около центра распределения. Они уменьшаются к краям распределения, исчезая совсем за границами эллипса. 4. Центр таблицы находится в точке $\bar{x}_1 = 67,5$ дюйма и $\bar{x}_2 = 18,1$ дюйма. Эта точка по счастливой случайности попала в клетку, содержащую наибольшую частоту — 28 человек. 5. Гистограмму для распределения двух переменных можно построить путем установки над каждой клеткой таблицы, высота столбиков которой пропорциональна частотам соответствующих клеток. Столбики с наибольшими величинами роста будут в центре окружены более низкими столбиками. Высоты столбиков будут постепенно снижаться по направлению к периметру эллипса и совсем исчезнут за границами его. Гребень из высоких столбиков будет вытянут примерно по главной оси эллипса.

Если вы вообразите беспредельное увеличение общей численности при соответствующем уменьшении площади каждой клетки таблицы, то получите ясное представление о форме распределения двух переменных. Оно будет представлять собой гладкую поверхность, покрывающую эту таблицу, имеющую свою наивысшую высоту в центре (μ_1, μ_2) и постепенно снижающуюся по мере того, как расстояние от центра по осям X и Y будет уходить в бесконечность.

Частота парных изменений роста и длины предплечья. Данные Гальтона при том распределении по классам, как оно представлено в таблице 9 у Пирсона и Ли

Длина предплечья X_2	Рост в дюймах X_1																f_2	$\bar{x}_{1.2}$	$s_{1.2}$
	59-60	60-61	61-62	62-63	63-64	64-65	65-66	66-67	67-68	68-69	69-70	70-71	71-72	72-73	73-74	74-75			
21,0-21,5																4	1	74,4	0,00
20,5-21,0													1	1			2	72,0	0,71
20,0-20,5													1	1			1	72,4	0,00
19,5-20,0										2			1		2		5	71,0	2,51
19,0-19,5									2	4	6	11	8	4	2	1	38	70,6	1,62
18,5-19,0					1		2	6	8	7	15	13	2	1			55	68,8	1,77
18,0-18,5						3	8	15	28	14	25	5	2	2			102	68,0	1,67
17,5-18,0				2	4	2	12	18	15	7	2	1	1				61	66,7	1,62
17,0-17,5			1	3	6	11	10	7	7	3	1						49	65,4	1,80
16,5-17,0			1	5	6	5	4	1	1	1	1						25	64,4	1,97
16,0-16,5	1	1	1	3	2												8	62,0	1,41
15,5-16,0		1															1	60,4	0,00
Частота f_1	1	2	3	13	16	24	35	47	61	38	50	30	15	9	4	2	348		
Средняя $\bar{x}_{2.1}$	16,2	16,0	16,8	16,9	17,1	17,3	17,7	17,9	18,1	18,3	18,4	18,8	19,1	20,2	19,5	20,2	18,1		
Стандартное отклонение $s_{2.1}$	0,0	0,36	0,50	0,52	0,60	0,48	0,54	0,42	0,51	0,68	0,48	0,41	0,70	0,83	0,29	1,41	0,905		

Некоторые особенности этой новой модели таковы. 1. Каждое сечение, перпендикулярное оси X_1 , дает нормальное распределение; то же самое относится и к каждому сечению, перпендикулярному оси X_2 . Это означает, что каждый столбец и каждый ряд таблицы 55 являются выборками из нормального распределения. 2. Распределения частот, перпендикулярные оси X_1 , имеют все одно и то же стандартное отклонение $\sigma_{2.1}$, а все их средние находятся на прямой линии регрессии $\mu_{2.1} = \alpha_2 + \beta_{2.1}X_1$. Это является свойством модели I главы 6. Соответствующие выборочные средние и стандартные отклонения у нас выписаны в последних двух строках таблицы. Хотя имеется довольно значительная изменчивость значений $s_{2.1}$, все же каждое из них является оценкой общего параметра $\sigma_{2.1}$. 3. Распределения частот, перпендикулярные оси X_2 , имеют общее стандартное отклонение $\sigma_{1.2}$ (его оценки приведены в столбце с правой стороны таблицы), и их средние лежат уже на второй линии регрессии $\mu_{1.2} = \alpha_1 + \beta_{1.2}X_2$. Аналогичного свойства в модели I главы 6 не будет. Эти линии регрессии представлены на рисунках 18 и 19. 4. Каждое итоговое распределение частот нормально. То, которое находится справа, является $N(\mu_2, \sigma_2)$, в то время как находящееся внизу таблицы будет $N(\mu_1, \sigma_1)$. 5. Распределение численностей двух переменных определяется произведением $1/2 \pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}$ на e в степени:

$$\frac{-1}{2(1-\rho^2)} [(X_1 - \mu_1)^2 / \sigma_1^2 - 2\rho (X_1 - \mu_1)(X_2 - \mu_2) / \sigma_1 \sigma_2 + (X_2 - \mu_2)^2 / \sigma_2^2]$$

Это распределение имеет 5 параметров. Четыре из них — знакомые нам характеристики: μ_1 , μ_2 , σ_1 и σ_2 . Пятым является коэффициент корреляции ρ , оценкой которого служит r .

Параметр ρ измеряет тесноту связи между X_1 и X_2 . Это означает, что он определяет удлиненность или суженность эллипса, содержащего в себе большую часть наблюдений; проявление этого в выборке было показано на рисунке 19. В таблице для роста и длины предплечья ρ оценивается величиной $r = 0.7$. Способ вычисления таких оценок на основе таблиц, подобных таблице 55, будет описан в параграфе 9 главы 8.

Вычисления же для малой выборки были уже даны. Следует к этому добавить только несколько слов. Если произведено кодирование данных (параграф 4 главы 5), то коэффициент корреляции может быть вычислен непосредственно по суммам квадратов и произведений кодированных данных. В противоположность коэффициенту регрессии r не требует обратного перехода от кода к исходным данным. Например, по данным таблицы 47 находим $r = 735,2 / \sqrt{674,1 \times 832,4} = 0,98$. Тот же результат будет получен и при использовании первоначальных данных $r = 1470,4 / \sqrt{26,96 \times 83240}$.

Пример 5. Постройте график для $\bar{x}_{2.1}$, взятых из предпоследней строки таблицы 55. Значения X_1 здесь являются серединами классов, указанных в верхней части столбцов. Серединой первого класса следует считать 59,5 дюйма. Сравните свой график с рисунком 15.

Пример 6. Разместите график для $\bar{x}_{1.2}$ на том же месте, где построили график для $\bar{x}_{2.1}$, причем срединные значения классов для X_2 откладывайте по вертикальной оси. Серединой первого класса будет 21,25 дюйма. Если вы будете удивлены тем, что эти две линии регрессии различны, то вспомните, что $\bar{x}_{2.1}$ являются средними по столбцам, а $\bar{x}_{1.2}$ — средними по рядам. Каждая пара таких средних имеет только по одному общему показателю, находящемуся в клетке, которая стоит на пересечении данного столбца и данного ряда.

Пример 7. Постройте график $s_{2.1}$ в соответствии с X_1 . Вы увидите, что здесь нет никакой определенной тенденции к изменению; это указывает на то, что все $s_{2.1}$ могут относиться к случайным выборкам при общей $\sigma_{2.1}$.

Пример 8. Данные примера 20 главы 6 можно рассматривать в качестве случайной выборки из нормальной совокупности двух переменных. Вы в этом случае имеете $\bar{x} = 83$ г, $\bar{y} = 60$ мг, $\Sigma x^2 = 1000$, $\Sigma y^2 = 6854$, $\Sigma xy = 2302$. Вычислите регрессию общего веса X на вес грешня Y . *Ответ:* $\hat{X} = 83 + 0,336(Y - 60)$ г. Постройте график этой прямой вместе с прямой примера 21 главы 6. Обратите внимание, что угол, тангенс которого 0,336, отмеряется от этой оси Y .

Пример 9. По данным опыта с выклятами произведите оценку $\sigma_{y \cdot x}$. Ответ: $\sigma_{y \cdot x} = 13,9$ мг. Оцените также $\sigma_{x \cdot y}$. Ответ: $\sigma_{x \cdot y} = 15,1$ г. Отклонения от регрессии, вошедшие в $s_{x \cdot y}$, отмеряются в горизонтальном направлении.

Пример 10. По данным опыта с выклятами оцените r . Ответ: $r = 0,88$.

Пример 11. Граут [11] измерил длину крыльев и длину хоботка у 44 пчел (в мм):

Крыло	9,68	9,81	9,59	9,68	9,84	9,59	9,61	9,55	9,25	9,68	9,70
Хоботок	6,53	6,71	6,70	6,69	6,70	6,62	6,59	6,55	6,35	6,25	6,61
	9,60	9,50	9,74	9,72	9,64	9,73	9,77	9,72	9,54	9,65	
	6,51	6,55	6,74	6,75	6,45	6,75	6,70	6,65	6,68	6,77	
	9,74	9,59	9,71	9,56	9,61	9,61	9,55	9,78	9,74	9,48	
	6,44	6,54	6,64	6,55	6,57	6,61	6,64	6,64	6,63	6,62	
	9,71	9,20	9,53	9,74	9,67	9,58	9,49	9,64	9,45	9,52	
	6,55	6,22	6,43	6,67	6,68	6,62	6,71	6,70	6,50	6,41	
	9,58	9,60	9,68								
	6,50	6,62	6,69								

Кодрируйте эти данные путем вычитания из данных о длине крыла 9 и из данных о длине хоботка 6. Вычислите $r = 0,731$.

Пример 12. Тридцатью студентами получены следующие отметки на двух испытаниях по математике:

I	73	41	83	71	39	60	51	41	85	88	44	71	52	74	50
II	29	24	34	27	24	26	35	18	33	39	27	35	25	29	13
	42	85	53	85	44	65	60	33	43	76	51	57	35	40	76
	13	40	23	40	22	25	24	26	19	29	25	19	17	17	35

Вычислите $r = 0,774$.

Пример 13. После просмотра примеров 3, 4 и 5 главы 5 докажите, что если $Y = a + bX$ и $U = c + dz$, то $r_{YU} = \pm r_{xz}$.

4. Связь между выборочными коэффициентами корреляции и регрессии. Если X_2 будет означать зависимую переменную, то ее регрессия на X_1 будет $b_{21} = \Sigma x_1 x_2 / \Sigma x_1^2$; если же X_1 будет зависимой переменной, то регрессия X_1 на X_2 будет $b_{12} = \Sigma x_1 x_2 / \Sigma x_2^2$. Эти выражения отличаются только знаменателями. Линии регрессии, относящиеся к этим двум условиям, представлены на некоторых диаграммах рисунка 19. Если r близок к ± 1 , то эти линии близки друг к другу, сливаясь вместе в случае А; если $r = 0$, они перпендикулярны друг к другу. Отметим, что если X_2 является независимой переменной, то наклон b_{12} отмеряется от оси X_2 , т. е. в вертикальном направлении. Регрессия X_1 на X_2 всегда является той прямой, которая образует с вертикальной осью меньший угол.

Теперь легко видеть, что

$$b_{21} \cdot b_{12} = \frac{\Sigma x_1 x_2}{\Sigma x_1^2} \cdot \frac{\Sigma x_1 x_2}{\Sigma x_2^2} = \frac{(\Sigma x_1 x_2)^2}{\Sigma x_1^2 \cdot \Sigma x_2^2} = r^2.$$

Таким образом,

$$r = \sqrt{b_{21} \cdot b_{12}}.$$

Другими словами, r является геометрической средней из двух коэффициентов регрессии; коэффициент корреляции r является промежуточной величиной между ними. Покажем это на примере.

Если нас интересует определение роста сестер по росту их братьев, то мы имеем (табл. 54):

$b_{21} = 39/74 = 0,527$ дюйма роста сестры на 1 дюйм роста брата, что графически на рисунке 18 изображено прямой АВ. С другой стороны, если наш интерес будет заключаться в оценке роста братьев, то:

$b_{12} = 39/66 = 0,591$ дюйма роста брата на 1 дюйм роста сестры; этому соответствует линия регрессии CD. Однако обычно интерес представляет

взаимное соотношение, которое привлекает внимание в связи с тем, что он репрезентирует наследование роста; таким образом, мы находим

$$r = \sqrt{0,527 \times 0,591} = 0,558,$$

т. е. среднюю из двух коэффициентов регрессии.

Это положение может служить основой для выяснения отношения между коэффициентами корреляции и регрессии. Последний является коэффициентом, соответствующим случаю, когда одна переменная Y может рассматриваться в качестве зависимой от другой переменной X . Величина Y в этом случае может частично контролироваться или находиться в причинной связи с X , как это будет, когда надлежащее количество секрета некоторой железы является причиной различий в размерах организмов. Y также может оказывать влияние на X , как в случае привеса животных в опытах по кормлению, зависящего от величины начального веса. В таких случаях регрессия Y на X является показателем, который обычно дает нам всю желаемую информацию. Этот случай соответствует тому, когда требуется произвести оценку значения Y на основе известного значения X . С другой стороны, корреляция является соответствующей мерой связи между двумя переменными, подобными росту братьев и сестер. Относительно этих двух величин известно, что они через сложный механизм наследственности находятся в определенном соответствии друг с другом, но ни одна из них не может считаться следствием другой; они обе являются следствиями каких-то общих элементов. В этом смысле корреляция является средним соотношением в двух направлениях, в то время как регрессия — единично направленным соотношением. Конечно, существует множество переменных величин, связь между которыми может быть изучена как при помощи корреляции, так и регрессии или при помощи той и другой вместе. В этих случаях необходимо только иметь ясное представление о характере рассматриваемой связи.

При делении числителя и знаменателя отношения, определяющего r , на $n-1$ можно установить еще одну особенность r :

$$r = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \cdot \sum x_2^2}} = \frac{\frac{\sum x_1 x_2}{n-1}}{\sqrt{\frac{\sum x_1^2}{n-1} \cdot \frac{\sum x_2^2}{n-1}}} = \frac{\text{ковариация}}{\sqrt{s_1^2 \cdot s_2^2}} = \frac{\text{ковариация}}{\text{геометрическая средняя из дисперсий}}.$$

Как можно видеть, корреляция является частным от деления двух средних показателей варьирования: один из них — ковариация двух количественных признаков X_1 и X_2 , второй — средняя из двух дисперсий (т. е. средних квадратов). Коэффициент корреляции является отвлеченным числом, измеряющим сопряженное варьирование (covariation); если размер каждого из сопоставляемых признаков варьирует, то их корреляция является мерой такой величины, которая входит в оба эти варьирования.

Подобно этому и b приобретает новое истолкование, если произвести приводимые ниже преобразования:

$$b_{12} = \frac{\sum xy}{\sum x^2} = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \cdot \frac{\sqrt{\sum y^2}}{\sqrt{\sum x^2}} = r \frac{\sqrt{\sum y^2/(n-1)}}{\sqrt{\sum x^2/(n-1)}} = r \frac{s_y}{s_x}.$$

В этом виде b находится в явно выраженной связи с r . Каждый из них может быть вычислен на основе другого, если, конечно, известны выборочные стандартные отклонения этих двух переменных.

Часто представляется удобным взять стандартное отклонение в качестве единицы для измерения отклонения $x = X - \bar{x}$. Нормированное отклонение (standard deviate) $x' = x/s$ можно рассматривать как наблюдение, случайно взятое из $N(0,1)$. Например, допустим, что применяются два теста для оценки способностей, один — к ученикам четвертого класса, а второй — шестого

класса. Обозначим эти тесты через IV и VI. Далее допустим, что IV имеет $\bar{x}=40$ и $s=10$, в то время как VI имеет $\bar{x}=60$ и $s=20$. Если ученик A получил по IV отметку 60, а его брат B отмечен 80 по VI, то их нормированные отклонения (или стандартизированные показатели) будут

$$A: x' = \frac{60-40}{10} = 2 \quad \text{и} \quad B: x' = \frac{80-60}{20} = 1,$$

что указывает относительные места, занятые ими при этом испытании.

Очевидно, что для каждой выборки нормированных отклонений средняя $\bar{x}'=0$ и $s'=1$. Возьмем в качестве примера ряд

$$X: 6, 6, 2, 8, 6, 8, 6.$$

Так как средняя равна 6, то отклонения будут

$$x: 0, 0, -4, 2, 0, 2, 0.$$

Здесь $s=2$, и поэтому нормированные отклонения составят ряд

$$x': 0, 0, -2, 1, 0, 1, 0,$$

где $\bar{x}'=0$ и $s'=1$.

Может представить интерес выражение в этой же форме регрессии параграфа 4 главы 6:

$$\hat{Y} - \bar{y} = b(X - \bar{x}).$$

Если мы возьмем $\hat{y} = \hat{Y} - \bar{y}$ и $x = X - \bar{x}$, то получим

$$\hat{y} = bx = r \frac{s_y}{s_x} \cdot x.$$

Подставляя сюда обозначения нормированных отклонений, получим простое уравнение:

$$\hat{y} = rx',$$

так как $s_y = s_x = 1$, поэтому здесь r является не чем иным, как b . Таким образом, r , выраженное в нормированных отклонениях, является коэффициентом регрессии, т. е. средним изменением y' при изменении x' на величину стандартного отклонения. Мы уже указывали, что r не зависит от единиц измерения X и Y . Отсюда следует, что если мы применим нормирование переменных, результаты которого также не зависят от единиц измерения признаков, то различие между коэффициентами корреляции и регрессии просто исчезает.

5. Выборочное варьирование коэффициента корреляции. Одним из удобных методов получения выборок из нормальной совокупности двух переменных является старый способ, называемый способом *общих элементов* [7]. Вам следует вернуться к схеме случайных выборок параграфа 3 главы 3 или к своим выборкам, которые вы ранее произвели из данных таблицы 15. В новую таблицу, такую, как таблица 56, выписывается какое-нибудь удобное число, положим 3, наблюдений над привесом свиней. Эти привесы, или элементы, выписываются в таблицу дважды. После этого прибавляется, например, одно следующее случайно взятое наблюдение к левому столбцу и два следующих — к правому столбцу. Соответствующие суммы образуют парные значения X_1 и X_2 . В нашей таблице приведено образование трех таких пар. Очевидно, что между двумя суммами каждой пары имеется определенная связь. Если, положим, все три общих элемента оказались большими, то как X_1 , так и X_2 будут, по всей вероятности, большими, какими бы ни были остальные элементы, содержащиеся в них. Именно благодаря наличию этих необщих элементов связь не будет полной. Если вы продолжите этот процесс образования парных наблюдений, создав 100 или более пар, и после этого рассчитаете коэффициент корреляции, то вы получите значение r , кото-

Определение трех парных значений переменных X_1 и X_2 ,
имеющих общие элементы

Этими элементами здесь являются данные о прорвесе свиней таблицы 15.

Пары	Элементы	
1	23 } 44 } 43 }	Общие { 23 44 43
	37 }	Разные { 30 33
	$X_1 = 147$	$173 = X_2$
2	40 } 16 } 19 }	Общие { 40 16 19
	30 }	Разные { 29 13
	$X_1 = 105$	$117 = X_2$
3	23 } 38 } 37 }	Общие { 23 38 37
	30 }	Разные { 31 41
	$X_1 = 128$	$170 = X_2$

рое не будет слишком отличаться от значения этого показателя в совокупности:

$$q = 3/\sqrt{4 \times 5} = 0,67.$$

Числитель этой дроби является числом одинаковых элементов, в то время как знаменатель есть среднее геометрическое общего числа элементов в каждой из двух сумм X_1 и X_2 . Для параметра q можно построить и общую формулу. Если n_{12} означает число общих элементов, а n_1 и n_2 — полное число элементов, входящих в каждую из двух сумм, то теоретически корреляция между такими суммами будет

$$q = n_{12}/\sqrt{n_1 \cdot n_2}.$$

Конечно, в величины, рассчитанные таким путем, будет входить и выборочная вариация. Поэтому вам удастся получить более или менее хорошую проверку данного положения только при наличии 10 или даже 20 парных сумм. При 50 парах мы обычно получали коэффициент корреляции, отличающийся от ожидаемого параметра на несколько сотых, но однажды нами был встречен и коэффициент 0,28, когда он в совокупности был равен

$$n_{12}/\sqrt{n_1 n_2} = 6/\sqrt{9 \times 16} = 0,5.$$

Если в X_1 и X_2 включить одинаковое число элементов, то $n_1 = n_2$. Обозначая это число элементов через n , получим параметр q в таком виде

$$q = n_{12}/n,$$

т. е. как отношение числа общих элементов ко всему числу элементов в каждой сумме. В этом особом случае коэффициент корреляции является просто долей тех элементов, которые общие для обеих сумм. На основе этого можно

сделать такую примерную интерпретацию корреляции между ростом братьев и ростом сестер (табл. 54), которая обычно приблизительно равна 0,5: в среднем 50% генов, определяющих рост, общи для братьев и сестер.

В качестве другой иллюстрации этого особого случая можно взять определение какой-либо физической или химической константы при помощи двух методов. Например, рассмотрим случай определения содержания калия в соке стебля кукурузы, произведенного двумя методами: колориметрическим и гравиметрическим [15]. Из одного и того же материала отбирались две пробы, каждая из которых исследовалась одним из этих двух методов. Общие элементы, входящие в результаты определения, находятся в связи с действительным количеством калия в паре выборок. Лишние, или необщие, элементы могут быть приписаны различной точности, с которой определяется содержание калия по этим двум методам, и случайному варьированию внутри пары проб. Выборочный коэффициент корреляции при 24 парных определениях оказался равным 0,87. При допущении, что каждое такое определение является суммой некоторого постоянного числа элементов, взятых случайно из однородной совокупности, можно утверждать, что 87% этих элементов у двух определений общие. Таким образом, 87% из числа элементов связывается с фактическим содержанием калия в соке растения, остальные же 13% обусловлены ошибками измерений и выборочной вариацией.

Менее наглядно истолкование корреляции $r=0,70$, полученной между весом стержня початка и зерна у 250 початков сахарной кукурузы сорта Каунтри Дженгльмен [12]. Если мы введем постулат об одинаковом числе элементов, определяющих вес стержня и вес зерна, тогда можно считать, что 70% из них общие. Но что они собой представляют? Некоторые из них, вероятно, наследственного характера, а некоторые другие связаны с сопутствующими условиями.

Читатель не встретит никаких затруднений при продолжении этого списка примеров. Однако при этом необходимо иметь в виду следующие ограничения. Элементы, входящие в две переменные, должны быть не только равны численно, но и обязательно должны быть выбраны из нормально распределенной совокупности. Более того, совместное влияние этих элементов должно найти свое отражение в виде суммы, а не какой-либо иной функции. Если один элемент, включаемый по жребию, будет систематически дублировать эффекты других элементов, то наша интерпретация корреляции не будет правильной. Однако, несмотря на эти ограничения, осторожный и вдумчивый исследователь может при помощи этого представления об общих элементах значительно обогатить свое понятие о корреляции.

В тех случаях, когда X_1 и X_2 не имеют одну и ту же единицу измерения, возможность применения теории общих элементов становится уже ограниченной. Например, в исследовании Дангана о повреждениях кукурузы градом (пример 43 главы 6) хотя обе переменные и были выражены в процентах, все же в одном случае в основе лежали площади, а в другом — веса. Здесь будет уже трудно определить природу общих элементов, хотя они, несомненно, имеются.

После того как вы проведете один или несколько расчетов r с общими элементами, у вас сложится достаточно полное представление о выборочном варьировании этого показателя. Однако будет слишком утомительной работой собрать этим путем достаточное количество коэффициентов корреляции даже для выявления, хотя бы грубого, представления о кривой распределения. Это дело должно быть предоставлено математикам, чтобы они дали решение, основанное на теоретических соображениях. На рисунке 20 изображены кривые для выборок в 8 парных наблюдениях из совокупностей, в которых корреляции 0 и 0,8. Даже первая из них не является нормальной кривой. Нетрудно догадаться и о причине явно выраженной скошенности у второй кривой. Так как в этом случае параметр 0,8, то выборочные значения могут превзойти эту величину только на 0,2, в сторону уменьшения они могут

отстоять от нее даже на 1,8. В связи с тем что вариация этого показателя на одном конце шкалы имеет определенный предел (+1), а на другом конце его практически нет, кривая распределения становится явно асимметричной. Конечно, при увеличении размера выборки появляется тенденция к исчезновению этой скошенности. Выборки в 300 пар, взятые из совокупности, имеющей даже столь высокую корреляцию, как 0,8, дадут очень небольшое число отклонений от параметра, превосходящих 0,05 по ту и другую сторону от него. Следовательно, в этом случае верхний предел — единица не будет служить ограничением, и распределение будет почти нормальным.

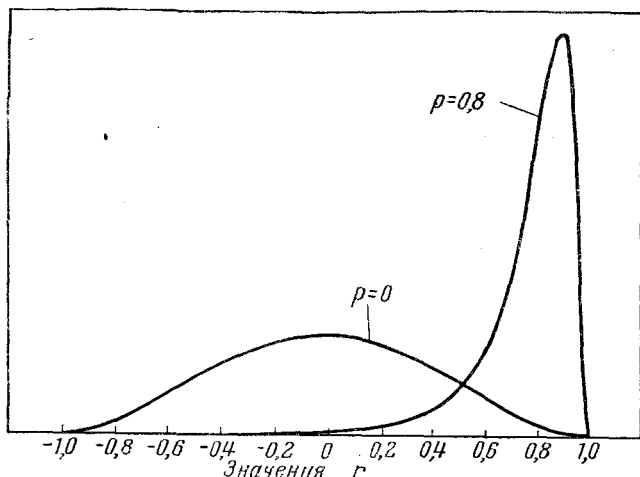


Рис. 20. Распределение выборочных коэффициентов корреляции в выборках из 8 парных наблюдений, взятых из двух нормально распределенных совокупностей с двумя переменными, имеющих указанные значения ρ .

Пример 14. Допустим, что все элементы X_1 входят в X_2 . В этом случае $n_{12} = n_1$ и $\rho = \sqrt{n_1/n_2}$. Это будет связь, которая существует между целым X_2 и его частью X_1 . Примером может служить корреляция между живым и убойным весом свиней [2]. У 533 изученных животных средний живой вес 222,6 фунта, а средний вес туши 185,1 фунта. Если считать, что каждый фунт убойного веса является общим элементом, тогда r можно определить так: $\sqrt{185,1/222,6} = 0,91$. Сравните этот результат с выборочным $r = 0,986$; обсуждение вопроса см. в параграфе 11 главы 7.

Пример 15. Урожай 16 делянок одной плантации чайного куста [5], собранные за 14-недельный период, были коррелированы с урожаями тех же делянок за следующий период такой же длительности. Коэффициент корреляции был равен 0,91. Можете ли вы интерпретировать этот результат в терминах общих элементов?

6. Оценка коэффициента корреляции; доверительные утверждения и проверка гипотез. Если из нормальной совокупности двух переменных, корреляция в которой ρ , берутся случайные выборки, то показатель r является оценкой ρ . Ошибки X_1 и X_2 будут источником смещения этого коэффициента. В противоположность выборочной регрессии в этом случае ни значения X_1 , ни значения X_2 не могут отбираться произвольно. Так, вы видите, что в таких выборках, как измерение высоты растений сор в примере 1 главы 6, вычисленный коэффициент корреляции не является оценкой ρ , потому что значения X_1 специально подобраны. В этом заключается заметное различие между выборками, взятыми для оценки β и ρ .

Фишер [8] установил, что для проверки нулевой гипотезы $\rho = 0$ можно применить критерий t , выборочное значение которого определяется по формуле:

$$t = r \sqrt{(n-2)/(1-r^2)}; \text{ число степеней свободы} = n-2.$$

Для примера допустим, что $r = 0,597$ получен в случае B рисунка 19 при указанных выше условиях. Какова вероятность появления в выборке значений,

больших $|r|$ при $\varrho=0$? Для ответа находим:

$$t = 0,597 \sqrt{(9-2)/(1-0,597^2)} = 1,969; \text{ число степеней свободы} = 7.$$

Интерполяция в таблице значений t приводит к вероятности больших значений, приблизительно равной 0,09. Следовательно, поскольку в выборках при $\varrho=0$ значения корреляции, большие $|r|$, могут встретиться около 9 на 100, нулевая гипотеза не может быть отвергнута. Этот факт проливает свет на затруднения (параграф 2 этой главы), возникающие при графическом определении небольших корреляций, в особенности если число степеней свободы незначительно; такая корреляция может быть не более чем случайностью выборочного наблюдения (см. пример 25 этой главы).

Проверка $H_0: \varrho=0$, $H_A: \varrho \neq 0$ может быть произведена без расчетов, а непосредственно по таблице 57. Взяв строку, соответствующую числу степеней свободы = 7, просто производим сравнение выборочного r с табличным значением. Наш $r = 0,597$ много меньше значения для 5 %-ного уровня — 0,666, что приводит к тому же, что и ранее, заключению. Эта оценка производится без учета знака r . Рассматривая следующую табличку, обратите, в частности, внимание на то, какое оказывают влияние на выводы размер выборки и величина r .

Число пар	Число степеней свободы	r	Вывод относительно гипотезы $\varrho = 0$
20	18	0,60	Отвергается на 1%-ном уровне
100	98	0,21	Отвергается на 5%-ном уровне
10	8	0,60	Не отвергается
15	13	-0,50	Не отвергается
500	498	-0,15	Отвергается на 1%-ном уровне

ТАБЛИЦА 57

Коэффициенты корреляции на 5%-ном и 1%-ном уровне существенности

Число степеней свободы	5%		Число степеней свободы	1%		Число степеней свободы	1%	
	5%	1%		5%	1%		5%	1%
1	0,997	1,000	16	0,468	0,590	35	0,325	0,418
2	0,950	0,990	17	0,456	0,575	40	0,304	0,393
3	0,878	0,959	18	0,444	0,561	45	0,288	0,372
4	0,811	0,917	19	0,433	0,549	50	0,273	0,354
5	0,754	0,874	20	0,423	0,537	60	0,250	0,325
6	0,707	0,834	21	0,413	0,526	70	0,232	0,302
7	0,666	0,798	22	0,404	0,515	80	0,217	0,283
8	0,632	0,765	23	0,396	0,505	90	0,205	0,267
9	0,602	0,735	24	0,388	0,496	100	0,195	0,254
10	0,576	0,708	25	0,381	0,487	125	0,174	0,228
11	0,553	0,684	26	0,374	0,478	150	0,159	0,208
12	0,532	0,661	27	0,367	0,470	200	0,138	0,181
13	0,514	0,641	28	0,361	0,463	300	0,113	0,148
14	0,497	0,623	29	0,355	0,456	400	0,098	0,128
15	0,482	0,603	30	0,349	0,449	500	0,088	0,115
						1000	0,062	0,081

Примечание. Частично настоящая таблица взята из таблицы VA в «Statistical Methods for Research Workers» с разрешения профессора Р. А. Фишера и его издателей Оливера и Бойда.

Те, кто желает при первом чтении книги сразу получить представление о главных положениях статистики, могут опустить остальную часть настоящей главы и всю следующую главу. Переходите к главе 9, если вы желаете

изучить в первую очередь статистику качественных признаков, или в главе 10, если намерены познакомиться с дисперсионным анализом.

На рисунке 20 было представлено асимметричное распределение значений r при малых выборках из совокупности с большим значением ρ . Использование критерия t для оценки существенности r имеет законную силу только при гипотезе $\rho=0$. Он совсем не подходит к проверке каких-либо других гипотез, например $\rho=0,05$ или $\rho_1-\rho_2=0$. Равным образом критерий t нельзя применять для построения доверительных утверждений относительно корреляций, установленных для малых выборок. Удобное и достаточно точное решение этих задач было предложено Фишером (9), который ввел преобразование r в величину z , имеющую почти нормальное распределение с дисперсией

$$\sigma_z^2 = \frac{1}{n-3}$$

и «практически не зависящую от значения корреляции в совокупности, из которой взята выборка». Связь z и r определяется формулой

$$z = 1/2 [\log_e (1+r) - \log_e (1-r)],$$

но мы можем применить схему рисунка 21, дающую возможность перехода от одной величины к другой с достаточной в обычных условиях точностью. Ниже дается несколько примеров применения величины z .

а) Требуется определить доверительные пределы для значения ρ совокупности, из которой получено выборочное r . В качестве примера возьмем $r=-0,889$, основанный на 9 парных наблюдениях рисунка 19Е.

Из рисунка 21 следует, что $r=0,889$ соответствует $z=1,417$. Так как $n=9$, то $s_z = 1/\sqrt{9-3}=0,408$. Теперь, так как z распределено почти нормально при любом размере выборки, $t_{0,01} = 2,576$ при числе степеней свободы $=\infty$. Это значение t может быть взято из таблицы 11 или приблизительно из таблицы 72 вероятностей для нормального распределения. Для $P=0,99$ доверительные пределы z будут:

$$1,417 - 2,576 \times 0,408 \leq z \leq 1,417 + 2,576 \times 0,408;$$

$$0,366 \leq z \leq 2,468.$$

Определяя по тому же рисунку в обратном порядке значения r и восстанавливая знак корреляции, получим доверительные границы для ρ при $P=0,99$:

$$-0,986 \leq \rho \leq -0,350.$$

Подчеркнем два вытекающих отсюда факта: 1) при малых выборках оценка r не очень надежна и 2) пределы не находятся на одинаковом расстоянии по обе стороны от r ; следовательно, его распределение скошено.

б) Иногда есть основание для проверки гипотезы о том, что ρ в совокупности имеет некоторое частное значение, отличное от нуля ($\rho=0$, как вы помните, проверяется при помощи таблицы 57). Пример возьмем из параграфа 5 этой главы, где $r=0,28$ был определен по выборке из 50 парных наблюдений и при $\rho=0,5$. Какова вероятность еще больших отклонений? Для $r=0,28$ по рисунку 21 находим $z=0,288$, а для $\rho=0,5$ имеем $z=0,549$. Разность $0,549-0,288=0,261$ имеет стандартную ошибку $1/\sqrt{n-3}=1/\sqrt{47}=0,1459$. Следовательно, $t=0,261/0,1459=1,80$ при числе степеней свободы $=\infty$, что не достигает 5%-ного уровня; наша выборка при шансах 1 : 20 не является чем-то необычным.

в) Для проверки гипотезы, что два выборочных значения r получены случайно из одной и той же совокупности, следует определить соответствующее z и после этого установить существенность разности между двумя z . В таблице 58 приведены данные о корреляции между привесом и количеством потребленных кормов в двух группах свиней.

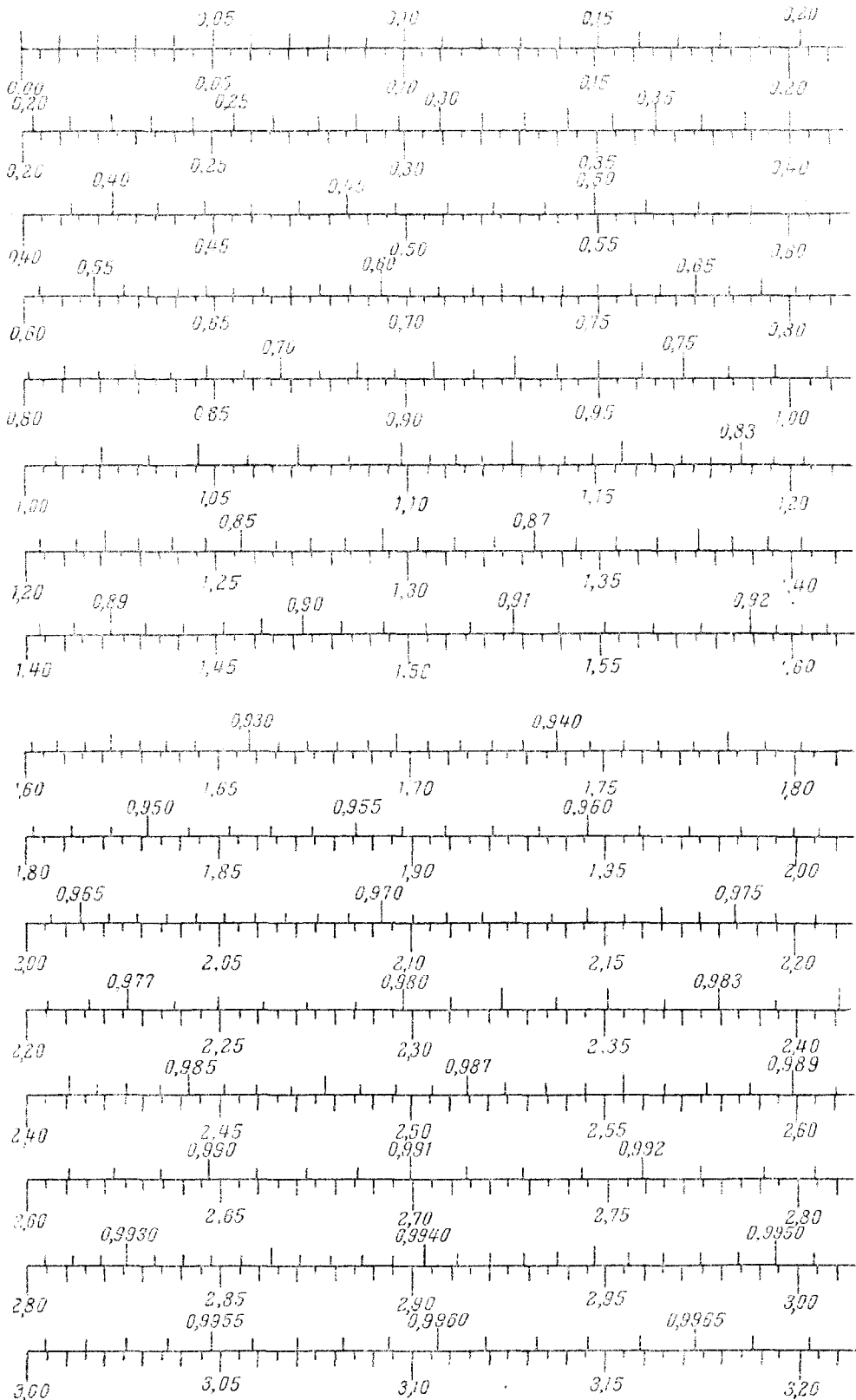


Рис. 21. Соответствующие друг другу значения τ и z .

Проверка существенности различия двух корреляций между привесом свиней и количеством потребленных кормов

Группа	Число свиней	r	z	$1/(n-3)$
1	5	0,870	1,333	0,500
2	12	0,560	0,633	0,111

Разность = 0,700. Сумма = 0,611

$$s_{z_1 - z_2} = \sqrt{0,611} = 0,782; t = 0,700/0,782 = 0,895; \text{число степеней свободы} = \infty; P = 0,37$$

Разности между значениями z , равной 0,700, соответствует средний квадрат:

$$\frac{1}{n_1-3} + \frac{1}{n_2-3} = \frac{1}{2} + \frac{1}{9} = 0,611; \text{число степеней свободы} = \infty.$$

Проверка производится обычным способом, т. е. путем вычисления t как отношения разности значений z к стандартной ошибке этой разности. При $P=0,37$, очевидно, нет никаких оснований для опровержения гипотезы о том, что оба z получены из одной и той же совокупности, и, следовательно, о том, что оба r относятся к одной общей совокупности.

г) Проверить гипотезу о том, что несколько r получены при одном и том же значении ρ , а также произвести объединение r для получения оценки ρ . Довольно часто можно предполагать, что несколько выборочных коэффициентов корреляции получены при одном и том же общем значении ρ . Если такая нулевая гипотеза не будет отвергнута, то выгодно объединить все r в одну общую оценку ρ , которая будет более надежной, чем любая оценка при отдельных значениях r . Лаш [14] интересовался средней корреляцией между начальным весом и привесом в 6 группах бычков. Соответствующие расчеты приведены в таблице 59. Каждое z взвешивалось (умножалось) по количествам, обратным средним квадратам, благодаря чему меньшим выборкам

ТАБЛИЦА 59

Проверка гипотезы об общем ρ и оценка этого ρ . Корреляции между начальным весом и привесом бычков

Выборки	Число бычков n	Число степеней свободы n	r	z	Взвешенное $z = (n-3)z$	Взвешенный квадрат = $(n-3)z^2$	Поправка z
1927 г. герефорды	4	1	0,929	1,651	1,651	2,726	1,589
1927 » зебу	13	10	0,570	0,648	6,480	4,199	0,633
1927 » помеси	9	6	0,455	0,491	2,946	1,446	0,468
1928 » герефорды	6	3	-0,092	-0,092	-0,276	0,025	-0,055
1928 » зебу	11	8	0,123	0,124	0,992	0,123	0,106
1928 » помеси	14	11	0,323	0,335	3,685	1,234	0,321
	57	39			15,478	9,753	14,941
Среднее $z = 0,397$						6,145	$z = 0,383$
Средний $r = 0,377$					$\chi^2 = 3,608$	$r = 0,365$	

придан и меньший удельный вес. Сумма взвешенных значений z , равная 15,478, делится на сумму весов 39, что дает среднее $z=0,397$. В следующем столбце приведены расчеты для проверки гипотезы о том, что все 6 выборочных корреляций получены из одной общей совокупности. Из суммы величин $(n-3)z^2$, которые получены умножением каждого z на соответствующее $(n-3)z$, вычитается поправка:

$$[\Sigma (n-3)z] \times (\text{среднее } z) = 15,478 \times 0,397 = 6,145.$$

Этот результат является значением хи-квадрат. Это значение у нас 3,608 и имеет 5 степеней свободы, что на 1 меньше числа групп. Из таблицы 6 находим $P=0,61$, т. е. H_0 остается неотвергнутой.

Поскольку все 6 выборочных корреляций можно считать полученными из одной и той же совокупности, вполне уместно рассматривать их среднее значение в качестве оценки общего ρ . Пользуясь таблицей 57 для среднего z , находим усредненную корреляцию 0,377 (при доверительном интервале 0,079—0,615). Не упустите случая заметить широкое варьирование этих корреляций, вычисленных по малым выборкам. Фишер указал, что у величины z имеется небольшое смещение, не превосходящее величину

$$\frac{\rho}{2(n-1)}.$$

Вы видите, что во всех предыдущих случаях смещение крайне незначительно. Однако оно может быть более или менее серьезным, если производится объединение большого числа корреляций, так как смещения накапливаются, складываясь одно с другим при каждом новом z . Ясно, что для получения ρ в совокупности такие смещения должны быть вычтены из каждого z . Если бы возникла необходимость увеличения точности расчетов в таблице 59, то следует вместо ρ подставить $r=0,377$; после этого можно из каждого z вычесть приблизительное значение смещения и повторить расчет среднего z . Так как эти поправки приведут к уменьшению оценки r , то ρ следует считать несколько меньше первого значения среднего r . Так, например, если взять $\rho=0,37$, то поправка для первого z будет $0,37/2 \times (4-1) = 0,062$ и исправленное z будет $1,651 - 0,062 = 1,589$. Остальные исправленные z приведены в последнем столбце таблицы. Сумма произведений

$$\Sigma [(n-3) \times (\text{исправленное } z)] = 14,941$$

получается непосредственно на счетной машине без промежуточных записей и при делении на 39 дает исправленное среднее значение $z = 0,383$. Соответствующий коэффициент корреляции будет 0,365. Применяя этот способ уточнения, следует помнить о двух моментах: а) значение ρ нам известно только приближенно и б) значения r и z по графику устанавливаются только с точностью до третьего десятичного знака. Таблицы распределения r можно найти в [4].

Пример 16. Исходя из того положения, что нарочитый отбор парных наблюдений влияет на величину корреляции, проверьте это, взяв 5 наименьших отметок при II испытании студентов (пример 12) и 6 наивысших. Корреляция между этими 11 отметками и соответствующими отметками I испытания равна 0,89, что указывает на искажение, вводимое в r при случайном отборе данных.

Пример 17. Определите 95%-ные доверительные пределы для корреляции между живым и убойным весом свиней 0,986 при $n=533$. *Ответ:* 0,983—0,988.

Какими бы были доверительные пределы, если бы число свиней было 25? *Ответ:* 0,968—0,994.

Пример 18. Корреляция между длиной крыла и длиной хоботка пчел (пример 11), равная 0,731, являлась одним из 4 подобных показателей, вычисленных Граутом; другие 3 корреляции, вычисленные каждая по 44 данным, были 0,354; 0,690 и 0,740. Проверьте гипотезу, что все эти 4 выборки взяты из совокупности с общей корреляцией ρ . *Ответ:* $\chi^2=9,164$, число степеней свободы=3 и $P=0,03$. Только в 3 испытаниях из 100 вы можете ожидать такую несогласованность между 4 корреляциями, взятыми из одной общей совокупности. Прежде чем делать какие-либо выводы, необходимо иметь более подробные сведения по поводу несогласованности с другими коэффициента корреляции 0,354.

Пример 19. Произведите оценку ρ в совокупности, из которой взяты 3 корреляции указанных признаков у пчел: 0,731; 0,690; 0,740. *Ответ:* 0,721.

Пример 20. Постройте 99%-ые доверительные пределы для корреляции предыдущего примера. Заметьте, что $r=0,721$ основывается на $3 \times 41 = 123$ степеням свободы. Поэтому он равносильен единичному r , вычисленному по выборке $123 + 3 = 126$ пчел. Доверительные пределы $0,590 - 0,815$.

Пример 21. Хотя у нас имеются свидетельства против гипотезы о том, что 4 корреляции, о которых шла речь выше, взяты из совокупности с общим ρ , это, однако, еще не говорит определенно за то, что меньшая корреляция существенно отклоняется от средней из трех больших корреляций. Для иллюстрации произведем проверку гипотезы о том, что $r=0,354$, который получен при некоторых особых обстоятельствах, возник из того же самого ρ , как и средняя других трех корреляций, которые, положим, относятся к другому улью. Вы уже нашли, что эти три наибольшие корреляции дают в среднем $0,721$. Покажите, что эта средняя существенно отклоняется от меньшей корреляции. *Ответ:* $\chi^2=8,97$; число степеней свободы=1; $P < 0,01$.

7. Корреляция и регрессия. Каждая пара столбцов в данных о повреждении плодов яблонь в таблице 45 дает определенный коэффициент корреляции; некоторые из этих коэффициентов представляют значительный интерес. Рассмотрим сначала ряд равных корреляций:

$$r_{XY} = r_{xy} = r_{Y\hat{Y}} = r_{Y\hat{X}} = r_{\hat{Y}\hat{X}}.$$

Каждая из соответствующих этим корреляциям пар столбцов содержит в себе кодированные значения другой пары столбцов, и поэтому все эти корреляции одинаковы (параграф 3 этой главы). Например: 1) отклонения x и y являются кодированными значениями X и Y , полученными путем вычитания \bar{x} из каждого X и \bar{y} из каждого Y ; следовательно, $r_{xy} = r_{XY}$; 2) так как $\hat{Y} = a + bX$, то \hat{Y} является кодированным X , причем каждое X умножено на b и к произведению прибавлено a ; следовательно, корреляция между X и Y равна корреляции между \hat{Y} и Y .

Эта корреляция между Y и \hat{Y} имеет особый смысл. Вскрывая его, мы можем сказать, что корреляция между Y и \hat{Y} — его оценкой при помощи регрессии одинакова с корреляцией между Y и X . Это означает, что если Y все в одинаковой мере близки к линии регрессии, то $r_{Y\hat{Y}}$ и равный ему r_{XY} близки к ± 1 . Таким образом, в этом смысле r_{XY} является мерой приближения при оценке Y с помощью регрессии.

Другую заслуживающую упоминания корреляцию вы можете определить из столбцов X и \hat{Y} . Так как \hat{Y} является кодированным значением X , то корреляция между X и его собственным кодом будет ± 1 , причем знак берется тот, который стоит у b . Путем проверки этого равенства вы можете предупредить проникновение грубых просчетов. Смысл этой корреляции состоит в том, что \hat{Y} меняются в строгой пропорциональности с изменением X ; последовательность всех точек (X, \hat{Y}) лежит точно на линии регрессии.

Точно так же могут быть вычислены и соответствующим образом интерпретированы корреляции $d_{y,x}$ с x и y . Первый из них, $r_{xd_{y,x}} = 0$, подсказывается допущениями, сделанными при подборе линии регрессии: стандартное отклонение $d_{y,x}$ в совокупности одинаково по всему интервалу изменения x , и поэтому $d_{y,x}$ не коррелировано с x . Смысл второй корреляции, $r_{yd_{y,x}} = \sqrt{1 - r_{xy}^2}$, может быть установлен следующим образом: если регрессия равна нулю, то линия регрессии горизонтальна и $d_{y,x}$ совпадают с y , т. е. $r_{yd_{y,x}}$ равен единице, когда $r_{xy} = 0$. В противоположном крайнем случае, по мере того как точки все более и более приближаются к линии регрессии, r_{xy} приближается к 1, а все отклонения $d_{y,x}$ приближаются к нулю, т. е. не остается почти ничего, в соответствии с чем изменялось бы y , и поэтому $r_{yd_{y,x}}$ приближается к нулю.

Можно видеть, что в настоящем параграфе r рассматривался просто как некоторая расчетная величина, а не в аспекте учения об оценках и заключениях вероятностного характера. Формулы, определяющие связь корреляции и регрессии, будут в одинаковой мере правильными как в случае дан-

ных о сое (пример 1 главы 6), так и в случае данных о повреждении плодов яблонь, несмотря на то что при специальном подборе значений X корреляция между временем и ростом не является оценкой какого-либо параметра совокупности. Необходимо соблюдать осторожность и проводить различие между r в качестве оценки ρ и r в качестве простого и удобно вычисляемого показателя. Так же как в настоящем параграфе, в трех последующих параграфах r будет взят просто как один из частных случаев общих концепций регрессии. Вопрос о том, можно или нельзя рассматривать r в качестве оценки ρ , зависит от того, в какой мере имеются налицо условия выборочного наблюдения, перечисленные в параграфе 3 этой главы.

Пример 22. Докажите, что корреляция $d_{y \cdot x}$ с x равна нулю. Указание: используйте форму отклонения $d_{y \cdot x} = y - bx$. После этого сумма произведений будет $\sum x d_{y \cdot x} = \sum xy - b \sum x^2$. Так как $b = \sum xy / \sum x^2$, то в результате получается нуль.

Пример 23. Докажите, что корреляция $d_{y \cdot x}$ с y равна $\sqrt{1 - r^2}$.

Пример 24. Покажите, что $b = r \sqrt{\sum y^2 / \sum x^2} = r s_y / s_x$.

Пример 25. Покажите, что проверка гипотезы $H_0: \beta = 0$ равнозначна проверке $H_0: \rho = 0$. Указание: докажите, что значение t при испытании b равно t при испытании r .

8. Корреляция и разложение $\sum y^2$. Сумма квадратов, относящаяся к регрессии (параграф 11 главы 6), может легко быть преобразована следующим образом:

$$\frac{(\sum xy)^2}{\sum x^2} = \frac{(\sum xy)^2}{\sum x^2 \cdot \sum y^2} \cdot \sum y^2 = r^2 \sum y^2.$$

Значит, r^2 определяет часть $\sum y^2$, относящуюся к регрессии. Остаток будет таким:

$$\sum d_{y \cdot x}^2 = \sum y^2 - r^2 \sum y^2 = (1 - r^2) \sum y^2.$$

Вы видите здесь указание на определенную связь между r и разложением $\sum y^2$ при определении регрессии. Величина r^2 является долей $\sum y^2$, обусловленной согласованным изменением Y и X , в то время как $(1 - r^2)$ является остаточной частью, совсем не зависящей от X .

Эти факты обобщены в таблице 60 дисперсионного анализа, которая подобна таблице 51. Ранее приводимые исходные данные представлены здесь в последней строке. Отметим, что $r^2 = (\sum xy)^2 / (\sum x^2 \cdot \sum y^2) = 56^2 / (112 \times 68) = 7/17$.

ТАБЛИЦА 60

Дисперсионный анализ y по данным таблицы 51

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Регрессия	1	$r^2 \sum y^2 = 28$	
Отклонения от регрессии	5	$(1 - r^2) \sum y^2 = 40$	8
Отклонения от средней	6	$\sum y^2 = 68$	11,3

$$\sum x^2 = 112, \quad \sum y^2 = 68, \quad \sum xy = 56, \quad r^2 = 7/17 = 0,41, \quad r = 0,64$$

Опять можно отметить, что здесь r участвует не как оценка определенного параметра; он является в данном случае просто алгебраическим отношением, связанным с определением линейной регрессии как при модели I, так и II. В той мере, в какой это относится к вычислениям, нет никакого различия в том, будет ли выборка взята из той или другой модели или ни из одной из них, лишь бы регрессия была линейна. Эти символы менее сложны, чем символы, относящиеся к регрессии, поэтому их применение становится почти универсальным. Следует предостеречь начинающего читателя против

неправильного представления о том, что будто бы r всегда применяется только как оценка некоторого ρ .

Если вы будете руководствоваться этими положениями и рассмотрите их в аспекте данных рисунка 11 и таблицы 45, то это прольет свет на то, каким образом данный комплекс факторов варьирования влияет на величину r . Возьмем, например, крайний случай, когда $r=0$ и, следовательно, $1-r^2=1$. В этом случае не будет наблюдаться никакого варьирования процента поврежденных плодов яблонь, обусловленного различиями в размере урожая плодов. Линия регрессии будет горизонтальной и пересечет вертикальную шкалу в точке $\bar{y}=45\%$. Процент поврежденных плодов, определенный по регрессии, в этой таблице всюду будет одинаковым и равен \bar{y} . Отклонения от регрессии будут просто являться отклонениями от средней. Расчет $s_{y \cdot x}$ будет тот же самый, что и расчет s_y , с учетом их числа степеней свободы.

Противоположный крайний случай будет тогда, когда $r=\pm 1$ и $1-r^2=0$. Теперь все варьирование процента поврежденных яблок будет относиться к регрессии; каждая точка будет точно находиться на линии регрессии; проценты поврежденных плодов, вычисленные по регрессии, будут совпадать с соответствующими наблюдаемыми значениями, а все отклонения от регрессии и стандартное отклонение от регрессии обратятся в нуль.

9. Корреляция, общие элементы и регрессия. Представление об общих элементах позволило выявить некоторые факты, относящиеся к регрессии. Для полного понимания этого вопроса необходимо сделать небольшое дополнение к положениям, изложенным в параграфе 5 этой главы. Можно мысленно себе представить, что в таблице 56 произведен отбор четырех элементов, сумма которых составляет X_1 . Далее три из них, взятых случайно, переносятся в X_2 , после чего к ним добавляется еще два элемента. В результате этого значение X_2 в известной мере контролируется, т. е. определяется этими элементами, перенесенными из X_1 . В этом смысле X_1 является одной из причин изменения X_2 . В то время как расчет r не зависит от этой новой точки зрения, она вносит изменение в отношении регрессии. Было установлено [7], что при использовании отклонений от средней x_1 и x_2 регрессия x_2 на x_1 будет $\hat{x}_2=(n_{12}/n_1) \cdot x_1$, т. е. коэффициент регрессии оказывается отношением числа перенесенных элементов к общему числу их в x_1 ¹. Рассмотрим в качестве примера наследственность в отношении роста. В среднем половина генов, определяющих рост, переходит от отца к сыну. Если при этом не будет влияния каких-либо привходящих обстоятельств, то следует ожидать, что регрессия роста сына на рост отца будет $\hat{x}_2=0,5 x_1$, т. е. ожидаемые отклонения величины роста сыновей от их средней будут составлять в среднем половину отклонений величин роста у отцов от их средней. Если вы знакомы с генетикой, то знаете, что коэффициент регрессии часто ниже этого значения.

Уравнение $\hat{x}_2=(n_{12}/n_1) \cdot x_1$ подчеркивает тот факт, что оцениваемое значение \hat{x}_2 не зависит от числа элементов в x_2 . Это оцениваемое значение всецело зависит от двух величин: 1) от доли элементов, переносимых из x_1 в x_2 , и 2) от значения суммы x_1 . Ни число элементов в x_2 , ни их сумма не влияют на \hat{x}_2 . Величины x_2 и x_1 коррелированы (параграф 7 этой главы) только потому, что они содержат общие элементы n_{12} .

10. Корреляция и разности. Любой ряд парных наблюдений может представлять интерес одновременно с двух точек зрения или только с одной из них: с точки зрения учения о корреляции и регрессии, которому посвящены главы 6 и 7, или с точки зрения учения о разностях, изложенного в главах 2, 3 и 4. Фактически эти две точки зрения находятся в определенных отношениях друг с другом; некоторые из них будут рассмотрены здесь.

¹ Вместо X_1 и X_2 удобнее в качестве переменных взять x_1 и x_2 . Эти новые переменные являются суммами, которые вы получите в таблице 56 путем вычитания $\bar{x}=30$ из каждого привеса свиньи в таблице 45.

В параграфе 9 главы 3 вы познакомились с разностями между случайными парными наблюдениями из нормальной совокупности при $\sigma^2=100$. Эти разности, центром которых является нуль, имеют удвоенную дисперсию, т. е. $\sigma_D^2=2\sigma^2$. Данное положение относится и к разностям между средними; эти разности сгущены около нуля с дисперсией $2\sigma^2/n$. Теперь необходимо обратиться к рассмотрению разностей более общего вида. Представим себе, что в разности $X_D=X_1-X_2$ величина X_1 взята из совокупности со средней μ_1 и дисперсией σ_1^2 , а величина X_2 — из совокупности при μ_2 и σ_2^2 . Для получения выборок из такой совокупности разностей вы можете взять X_1 из таблицы 15 привесов свиней и X_2 из тех же привесов таблицы 24, вычитая для удобства 10 фунтов из каждого последнего показателя, что даст $\mu_2=20$ фунтам. Такие разности будут иметь нормальное распределение при

$$\begin{aligned}\mu_D &= \mu_1 - \mu_2 = 30 - 20 = 10 \text{ фунтам,} \\ \sigma_D^2 &= \sigma_1^2 + \sigma_2^2 = 100 + 25 = 125.\end{aligned}$$

В таблице 61 даны три выборки этого рода. Конечно, в этих выборках принимает участие и выборочное варьирование средней и дисперсии, что можно видеть из сравнения этих величин с соответствующими параметрами. Выборочное варьирование приводит к различию между суммами квадратов для выборки и совокупности. Для выборки легко установить (пример 31), что

$$\Sigma x_D^2 = \Sigma x_1^2 + \Sigma x_2^2 - 2\Sigma x_1x_2.$$

Проверим это по данным первой выборки: $918,4=408,9+268,1-2(-120,7)$. То же самое и по другим выборкам. В выборке 3 оказалось $\Sigma x_1x_2=0$, и поэтому $1352=1206+146$.

Если вместо Σx_1x_2 подставить равную ей величину $r_{12} \sqrt{\Sigma x_1^2 \cdot \Sigma x_2^2}$ и после этого разделить обе части уравнения на $n-1$, то в результате получим:

$$s_D^2 = s_1^2 + s_2^2 - 2r_{12}s_1s_2.$$

Это уравнение будет вполне точным для любого ряда разностей, безотносительно к выборочному варьированию. Почему же в формуле σ_D^2 для совокупности нет соответствующего последнего члена? Причина этого такова: при данном способе образования случайных парных наблюдений X_1 и X_2 корреляция этих пар в совокупности равна нулю. При этом капризы случая могут создать в выборках некоторую небольшую корреляцию, но знак ее в одинаковой мере может быть как положительным, так и отрицательным; поэтому сумма ряда таких корреляций будет при увеличении их числа стремиться к нулю.

Если вы сравните коэффициенты корреляции r выборок 1 и 2 с данными таблицы 57, то убедитесь, что вычисленные значения могут считаться выборочными значениями при $\rho=0$. И действительно, если вы продолжите этот процесс конструирования выборок, то придете к тому, что $|r| > 0,632$ будет встречаться примерно один раз из 20 случаев.

Мы уже обсуждали вопрос о разностях из некоррелированных совокупностей. Что же получится, если X_1 и X_2 будут выбраны из совокупности, в которой между ними существует корреляция ρ ? Такую совокупность можно получить при помощи общих элементов параграфа 5 этой главы, а выборки из нее можно получить по способу, показанному в таблице 56. Эта совокупность, как вы помните, имеет корреляцию $\rho_{12}=0,67$. Если для каждой пары наблюдений вычислить $X_D=X_1-X_2$, то

$$\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2$$

будет формулой, по образцу которой построена и формула для разностей из некоррелированных совокупностей. В той мере, к какой это относится к выборке, между данными формулами нет никакого различия. Выборки разностей как из коррелированных, так и некоррелированных совокупно-

стей будут иметь дисперсию

$$s_D^2 = s_1^2 + s_2^2 - 2r_{12}s_1s_2.$$

Конечно, в случае, когда выборка берется из коррелированной совокупности, вы можете ожидать существенного значения r , в особенности при q и n достаточно больших.

ТАБЛИЦА 61

Три выборки по 10 разностей $X_D = X_1 - X_2$

Значения X_1 взяты из таблицы 15 и X_2 из таблицы 24 при уменьшении каждого наблюдения на 10 фунтов

Номер наблюдения	1-я выборка			2-я выборка			3-я выборка		
	X_1	X_2	X_D	X_1	X_2	X_D	X_1	X_2	X_D
1	39	13	26	19	22	-3	32	18	14
2	34	19	15	14	20	-6	31	13	18
3	22	20	2	57	25	32	28	22	6
4	27	23	4	34	26	8	24	17	7
5	33	25	8	39	22	17	44	11	33
6	42	20	22	34	16	18	53	19	34
7	36	7	29	39	23	16	9	16	-7
8	24	16	8	13	22	-9	35	23	12
9	25	24	1	39	22	17	33	23	10
10	29	20	9	23	16	7	31	18	13
Средняя	31,1	18,7	12,4	31,1	21,4	9,7	32	18	14
Сумма квадратов . .	408,9	238,1	918,4	1706,9	98,4	1480,1	1206	146	1352
Сумма произведений	-120,7			162,6			0		
Средний квадрат . .	45,43	29,79	102,04	189,66	10,93	164,46	134,00	16,22	150,22
Ковариация	-13,41			18,07			0		
Корреляция	-0,365			0,102			0		

Эта формула для дисперсии разности дает объяснение результатов произвольного составления парных наблюдений, рассмотренных в параграфе 6 главы 4. Разности случайного составленных парных наблюдений первых двух колонок таблицы 17 будут: 1, 22, 23, -1, 20, 33, -41, -20, 20 и 6, а их дисперсия $s_D^2 = 513,79$. В первом столбце таблицы 62 дана проверка этой дисперсии при помощи приведенной ранее формулы. Вы можете отметить, что в этом случае корреляция отрицательна; поэтому третий член формулы здесь прибавляется к первым двум, что дает $s_D^2 = 513,79$. Теперь рассмотрим произвольное составление парных наблюдений по методу 1 в таблице 30. Здесь перестановка данных приводит к изменению случайной корреляции -0,59962 на весьма высокое положительное значение 0,97664, вследствие чего дисперсия разности уменьшается до 8,01 (средний столбец таблицы 62). Построение пар по методу II, наоборот, усиливает отрицательное значение случайной корреляции и увеличивает дисперсию разности до $s_D^2 = 627,35$.

Выгода образования парных наблюдений в опыте зависит от того, в какой мере удастся использовать положительную корреляцию. Если исследователь знает свой подопытный материал в такой степени, чтобы быть уверенным в одинаковой реакции членов каждой пары объектов, то он может поставить целью образование разностей, подобных разностям по методу 1, и получить небольшую дисперсию их. Но если он приступит к образованию парных наблюдений без надлежащего знакомства с подопытным материалом, то может прийти к отрицательной корреляции, что даст большую ошибку опыта, подобную ошибке по методу II. В этом случае ему лучше иметь дело с групповыми сравнениями главы 4.

Иногда полезно иметь в виду, что корреляция в выборке парных наблюдений может быть вычислена на основе приведенной выше формулы путем такого ее преобразования:

$$r_{12} = \frac{s_1^2 + s_2^2 - s_D^2}{2s_1s_2} = \frac{\Sigma x_1^2 + \Sigma x_2^2 - \Sigma x_D^2}{2\sqrt{\Sigma x_1^2 \cdot \Sigma x_2^2}},$$

Иллюстрация влияния корреляции на стандартное отклонение разностей
Данные таблицы 17 и 30

Из первых двух колонок таблицы 17	Метод I из таблицы 30	Метод II из таблицы 30
r_{12} $-0,599621$ s_1^2 $169,82$ s_2^2 $151,57$	0,97664	-0,95353
$s_1^2 + s_2^2$ $321,39$ $2r_{12}s_1s_2$ $-192,40$	321,39 313,38	321,39 -305,96
s_D^2 513,79	8,01	627,35

¹ Такое большое число десятичных знаков у r необходимо для точности арифметических расчетов.

где $x_D = x_1 - x_2$, т. е. отклонение X_D от ее средней. В качестве примера можно взять данные о поражении табака вирусной болезнью из таблицы 12, где $\Sigma x_D^2 = 130$ уже вычислена. Легко определить суммы квадратов числа поражений двух рядов наблюдений над половинами листа: $\Sigma x_1^2 = 468$ и $\Sigma x_2^2 = 172$. Подстановка в формулу дает:

$$r_{12} = \frac{468 + 172 - 130}{2\sqrt{468 \times 172}} = 0,90.$$

Эффективность этого опыта, о чем говорилось в параграфе 10 главы 3, возникла в связи со схожестью реакции половин листьев на прививку; эта схожесть и нашла свое отражение в столь высокой корреляции — 0,90.

Основной практический интерес обычно представляют разности, однако полезно также знать и формулу для дисперсии суммы $X_S = X_1 + X_2$:

$$s_S^2 = s_1^2 + s_2^2 + 2r_{12}s_1s_2.$$

Пример 26. Взяв данные таблицы 54, вычтите из величины роста братьев величины роста сестер, после чего определите скорректированную сумму квадратов этих разностей. На основе этой суммы и суммы $\Sigma x_1^2 = 74$ и $\Sigma x_2^2 = 66$ произведите при помощи приведенной выше формулы проверку равенства $r_{12} = 0,558$.

Пример 27. Обратите внимание на то, что если последняя из приведенных выше формул обращается в $s_D^2 = s_1^2 + s_2^2$, как это имеет место в выборке 3 таблицы 61, то $r_{12} = 0$; если $s_D^2 < s_1^2 + s_2^2$, как в выборке 2, то r_{12} положителен.

Пример 28. Хейбер [12] приводит следующие сведения о 300 початках сахарной кукурузы:

Показатели	Средняя (г)	Стандартное отклонение (г)
Вес початка E	106,55	27,68
Вес стержня початка C	15,59	
Вес зерна G	90,96	24,62

Здесь наблюдается равенство $E = C + G$, а также $r_{eg} = 0,9940$.
Определите $s_e = 4,19$ г и $r_{eg} = 0,6906$.

Пример 29. Если $r_{12} = 1$, то покажите, что $s_D = s_1 - s_2$.

Пример 30. Если $r_{12} = -1$, то покажите, что $s_D = s_1 + s_2$.

Пример 31. Допуская, что корреляция между X_1 и X_2 последовательно равна сначала нулю, а потом r_{12} , определите соответствующие формулы для Σx_D^2 и Σx_S^2 .

11. Корреляция между суммами и отношениями. Корреляции, возникающие от общих причин. Часто представляют интерес корреляции пере-

менной X_1 с суммой ее и другой переменной X_2 , т. е. X_1 с $X_S = X_1 + X_2$. Например, корреляция 0,986 (приведенная в параграфе 2 этой главы) между убойным весом X_1 и живым весом X_S свиней, где переменной X_2 служит вес отходов. Точно так же можно получить некоторую информацию при коррелировании X_1 с $X_P = X_1/X_2$ или с $X_Q = X_2/X_1$ [18]. Примером корреляции r_{1P} может служить корреляция между стандартным отклонением и коэффициентом вариации, т. е. между s и s/\bar{x} . В качестве иллюстрации было взято 50 пар этих статистических показателей, относящихся к случайным выборкам из привесов свиней параграфа 3 главы 3. Средняя из 50 стандартных отклонений равна 9,87, а из средних — 29,76. Корреляция между s и \bar{x} оказалась равной 0,140, что является оценкой $\rho = 0$, так как эти показатели в такого рода выборках теоретически некоррелированы. Но корреляция между s и S оказалась равной 0,932. Примером r_{1Q} может служить корреляция между живым весом X_1 и процентом выхода продукции X_Q :

$$X_Q = \frac{\text{вес туши } X_2}{\text{живой вес } X_1}.$$

Еще более сложная связь, определяемая корреляцией — 0,68, существует между средним суточным привесом свиньи и потребленными кормами на один фунт привеса [6], где привес является числителем первого отношения и в то же время знаменателем второго отношения.

Встречаясь со случаями неоправданных истолкований такой корреляции, Карл Пирсон назвал ее «ложной» [16]; это не совсем уместное название привело к общему недоверию в отношении этих корреляций. Конечно, такая слишком смелая интерпретация сама может быть ложной, так как корреляции подобного рода имеют под собой такое же твердое основание, как и другие: они являются надлежащими оценками параметров, если соблюдаются условия, выведенные в параграфе 3 этой главы [1]. При интерпретации ряда явлений для понимания корреляции между X_1 и суммами или отношениями, включающими эту переменную, несомненно требуются знания r_{12} .

Один особый случай выборки, когда $r_{12} = 0$, показан в таблице 63. Корреляция $r_{1,S} = 0,9445$ подобна той, которая существует между убойным весом (X_1) и живым весом (X_S) при условии, что между двумя частями живого веса нет никакой корреляции. Конечно, должна представлять интерес и корреляция между убойным весом и весом отходов.

ТАБЛИЦА 63

Корреляция X_1 с $X_S = X_1 + X_2$ и $X_Q = X_2/X_1$. Здесь X_1 и X_2 взяты из специально подобранной выборки 3 из таблицы 61, в которой $r_{12} = 0$

	X_1	X_2	X_1	X_S	X_1	X_Q
	32	18	32	50	32	0,562
	31	13	31	44	31	0,419
	28	22	28	50	28	0,786
	24	17	24	41	24	0,708
	44	11	44	55	44	0,250
	53	19	53	72	53	0,358
	9	16	9	25	9	1,778
	35	23	35	58	35	0,657
	33	23	33	56	33	0,697
	31	18	31	49	31	0,581
Сумма	320	180	320	500	320	6,796
Сумма квадратов	1206	146	1206	1352	1206	1,59887
Сумма произведений	0		1206			-37,516
Корреляция	0		0,9445			-0,8543

Корреляция для части с целым по исходным выборочным данным может быть вычислена при помощи формулы:

$$r_{1S} = \frac{s_1 - r_{12}s_2}{\sqrt{s_1^2 + 2r_{12}s_1s_2 + s_2^2}}$$

Если $r_{12} = 0$, как это наблюдается в таблице 63, то формула упрощается:

$$r_{1S} = \frac{s_1}{\sqrt{s_1^2 + s_2^2}} = \frac{1}{\sqrt{1 + s_2^2/s_1^2}} = \frac{1}{\sqrt{1 + \Sigma x_2^2/\Sigma x_1^2}}$$

Вычисленная в таблицах корреляция части с целым может быть проверена путем подстановки данных первой выборки:

$$r_{1S} = \frac{1}{\sqrt{1 + 146/1206}} = 0,9445.$$

Получено то же число, что и ранее. Если в дополнение к условию $r_{12} = 0$ также имеется и равенство $s_2 = s_1$, то

$$r_{1S} = 1/\sqrt{2} = 0,707.$$

Это дает нам представление о размере ложных корреляций в таких особых случаях и показывает, что наряду с корреляцией между этими суммируемыми переменными в нее входят и оба стандартных отклонения.

В этом специальном случае корреляция части с целым может рассматриваться в аспекте учения об общих элементах. Здесь X_1 является общей величиной для X_1 и X_S , в то время как X_2 является частью, на которую они отличаются друг от друга. Если X_1 и X_2 взяты из одной и той же совокупности, то можно положить $n_{12} = 1$, $n_1 = 1$ и $n_2 = 2$; в этом случае каждый выборочный коэффициент корреляции будет являться оценкой корреляции

$$\rho = n_{12}/\sqrt{n_1 \cdot n_2} = 1/\sqrt{2} = 0,707,$$

как это и было при особом случае r_{1S} .

Корреляция $r_{1Q} = -0,8543$ отрицательна отчасти потому, что X_1 входит в качестве знаменателя в дробь X_Q . Эта корреляция может быть также установлена по непосредственным данным выборки. Соответствующая приближенная формула имеет вид:

$$r_{1Q} = \frac{r_{12}C_2 - C_1}{\sqrt{C_1^2 + 2r_{12}C_1C_2 + C_2^2}}$$

где C — коэффициент вариации s/x , выраженный в долях (но не в процентах). По данным первых двух столбцов таблицы 63 $r_{12} = 0$, $C_1^2 = 0,13090$, $C_2^2 = 0,05007$ и $C_1 = 0,3618$. Подставляя эти данные, получаем:

$$r_{1Q} = \frac{-0,3618}{\sqrt{0,05007 + 0,13090}} = -0,8505.$$

Небольшое различие между этой величиной и значением, вычисленным в таблице, объясняется приближенным характером приведенной формулы. В эту корреляцию отношений наряду со стандартными отклонениями первоначальных выборок входят и соответствующие средние.

Корреляция между двумя переменными может полностью или частично возникнуть из их общей связи с одним или несколькими другими факторами. В качестве примеров нами уже рассматривались корреляции между признаками одного и того же организма. Крупное животное является крупным во всех своих частях, и поэтому любые две части тела будут, вероятно, коррелированы в связи с тем, что они входят в общий организм данного размера. Точно так же, две величины, меняющиеся во времени, могут дать высокую корреляцию между собой. Например, имеется высокая корреляция, $-0,98$, между коэффициентом рождаемости в Великобритании в период с 1875 г. по 1920 г. и производством чугуна в Соединенных Штатах. Это положение привело Юла к вопросу: «Почему мы иногда приходим к бессмысленным корреляциям между временными рядами?» [20]. Среди предложенных

ответов на этот вопрос только один соответствует существу дела: парные наблюдения, взятые в порядке времени, не являются случайно отобранными в обычном смысле и могут не относиться к какой-либо нормальной совокупности двух переменных. Следовательно, они не согласуются с условиями, положенными нами в основу оценок и заключений вероятностного характера. (О другом ответе на этот вопрос см. в разделе «Частная корреляция», гл. 14.)

12. Непараметрические методы. Ранговая корреляция. Часто нет оснований считать, что совокупность двух переменных имеет нормальную форму распределения. В этом случае нет никакого параметра ρ , подлежащего оценке. Но в то же время можно предположить, что эти две переменные не независимы друг от друга, и в этом смысле про них можно сказать, что они коррелированы.

Наиболее известную «совокупность» этого рода представляет случай, когда переменные не являются результатами измерения, а просто образуют расположения объектов в соответствии с некоторым принятым критерием. Так, группа свиней может быть расположена в ряд, от высшего к низшему наблюдению, на основе жирности хребта. Каждая из свиней получит свой ранг: первый, второй и т. д. Составление двух ранжированных рядов, имеющих такой или примерно такой смысл, привело Спирмана [19] к построению формулы для ранговой корреляции:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)},$$

которая применена у нас для вычислений в таблице 64. Подобно r , ранговый коэффициент корреляции в выборках может принимать значение от -1 (полное противоположное соответствие) до $+1$ (полное прямое соответствие).

ТАБЛИЦА 64

Ранжирование 7 крыс двумя наблюдателями в условиях 3-недельного содержания при неполноценном рационе

Номер крысы	Ранжирование, произведенное		Разность d	d^2
	1-м наблюдателем	2-м наблюдателем		
1	4	4	0	0
2	1	2	-1	1
3	6	5	1	1
4	5	6	-1	1
5	3	1	2	4
6	2	3	-1	1
7	7	7	0	0
			$\sum d = 0$	$\sum d^2 = 8$

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 8}{7 \times (49-1)} = 0,857$$

Распределение r_s известно только для случая $\rho=0$. Это означает, что оценка интервала, вообще говоря, не может быть произведена, если не проверяется гипотеза $H_0: \rho=0$. При выборках с числом рангов, превосходящим 8, в качестве приближения используют нормальную форму, как это указано при описании таблицы 57; если же n равно 8 или меньше, то применяют таблицу 65. Из этой таблицы следует, что при 5%-ном уровне указанная нулевая гипотеза вообще не может быть отвергнута при выборке в 4 или меньше наблюдений. Корреляция $r_s=0,857$, вычисленная в таблице 64, существенна на 5%-ном, но не на 1%-ном уровне, т. е. имеется около 2 шансов из 100 получить ранговую корреляцию, большую 0,857, даже в том

Уровни существенности рангового коэффициента корреляции в малых выборках при гипотезе $\rho=0$

Размер выборки	5%-ный уровень	1%-ный уровень
4 или менее	Нет	Нет
5	1	Нет
6	0,886	1
7	0,750	0,893
8	0,714	0,857
9 или более	Применяется таблица 57	

случае, если бы все крысы находились в одних и тех же условиях, так что суждения наших наблюдателей не имели бы никакого основания.

Если распределение двух непрерывных переменных значительно отличается от нормального распределения, то некоторое представление о взаимосвязи между ними по выборочным данным может быть составлено следующим образом. Наблюденные значения каждой переменной располагаются в порядке их размера от большего к меньшему. Нумеруются последовательные значения каждой переменной 1, 2 ... n ; назовем эти номера *рангами* (для полного знакомства с теорией этого вопроса см. [13] параграф 16.25). Первоначально наблюдаемые значения заменяются на эти ранговые числа, после чего вычисляется r_s . Проверка нулевой гипотезы $\rho=0$ может быть произведена при помощи таблицы 65.

ТАБЛИЦА 66

Обычные и ранговые корреляции в 10 выборках из нормальной совокупности двух переменных при $\rho=0,559$

Выборка	r	r_s	Выборка	r	r_s
1	0,663	0,505	6	0,712	0,680
2	0,708	0,738	7	0,627	0,474
3	0,505	0,584	8	0,444	0,352
4	0,374	0,311	9	0,529	0,398
5	0,536	0,562	10	0,546	0,484
В среднем				0,564	0,512

Иногда r_s применяется в качестве сокращенного способа оценки по выборке из нормальной совокупности двух переменных. Эта процедура не совсем оправдана. Верно, что, например, при $n=30$ или больше получается небольшая экономия времени и значения r и r_s в среднем не слишком различаются друг от друга. Последнее можно показать следующим образом. Используя общие элементы (параграф 5 этой главы), я построил 10 выборок, каждая при $n=20$, относящихся к нормальной совокупности двух переменных при

$$\rho = \frac{n_{12}}{\sqrt{n_1 \cdot n_2}} = \frac{10}{\sqrt{20 \times 16}} = 0,559,$$

после чего вычислил r и r_s по каждой выборке. Результаты представлены в таблице 66. Обе средние, указанные в последней строке таблицы, получены при помощи z (параграф 6 этой главы), несмотря на то, что теория не оправдывает применения этого преобразования r_s . В этой небольшой выборке коэффициентов корреляции r_s в среднем немного меньше r и несколько больше варьирует. Корреляция между 10 парными значениями r

и r_S равна 0,64. Если ваши данные не представляют большой ценности, то сокращенный способ вычисления корреляции может быть вполне удовлетворительным, но если эти данные получены с затратой больших средств и времени, обычная корреляция, вычисленная по методу моментов, заслуживает того, чтобы и на нее затратить некоторое время.

Л И Т Е Р А Т У Р А

1. B a k e r G. A., Journal of the American Statistical Association, 37, 537, 1942.
2. B r a n d t A. E., Thesis submitted for the degree Doctor of Philosophy, Iowa State College, 1932.
3. C r a t h o r n e A. R., Reorganisation of Mathematics in Secondary Education. Mathematical Association of America, Inc., page 105, 1923.
4. D a v i d F. W., Tables of the correlation coefficient. Cambridge University Press, 1938.
5. E d e n T., Journal of Agricultural Science, 21, 547, 1931.
6. E v v a r d John M., S n e l l M. G., C u l b e r t s o n C. C., S n e d e c o r George W., Proceedings of the American Society of Animal Production, 1927, page 2.
7. F i s c h e r Carl H., The Annals of Mathematical Statistics, 4, 403, 1933.
8. F i s h e r R. A., Biometrika, 10, 507, 1915.
9. F i s h e r R. A., Metron, 1, 3, 1921.
10. G a l t o n Francis, Proceedings of the Royal Society of London, 45, 135, 1888.
11. G r o u t R. A., Iowa Agricultural Experiment Station Research Bulletin No. 218, 1937.
12. H a b e r E. S., Данные Айовской с.-х. опытной станции.
13. K e n d a l l Maurice G., The advanced theory of statistics, Vol. I. Charles Griffin and Company, Limited, London, 1943.
14. L u s h Jay L., Journal of Agricultural Research, 42, 853, 1931.
15. M o r r i s V. H., G e r d e l R. W., Plant Physiology, 8, 315, 1933.
16. P e a r s o n Karl, Proceedings of the Royal Society, A, 60, 489, 1897.
17. P e a r s o n Karl, L e e Alice, Biometrika, 2, 357, 1902—1903.
18. R e e d Lowell J., Journal of the Washington Academy of Science, 11, 449, 1921.
19. S p e a r m a n C., American Journal of Psychology, 15, 88, 1904.
20. Y u l e G. U d n y, Journal of the Royal Statistical Society, 89, 1, 1926.

МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ВЫБОРОК

1. **Введение.** Большие выборки, взятые случайно из одной совокупности, встречаются редко. При большом n , положим 500, объекты наблюдения обычно относятся к нескольким совокупностям. Статистик, изучающий такие выборки, вынужден затрачивать значительные усилия для разделения наблюдений на ряд более или менее однородных групп. Он попытается произвести сравнение некоторых частей выборки или вскрыть определенные соотношения между компонующими частями и постарается дифференцировать свои заключения по отношению к этим различным совокупностям. Его задача может состоять именно в такой дифференциации, но он может интересоваться и элементами, общими для всех таких совокупностей.

Большие случайные выборки из единичной совокупности способны достаточно верно отображать форму и параметры этой совокупности. Выборочные статистические характеристики довольно близки к параметрам совокупности. Оценки интервалов и проверка гипотез в этом случае не имеют такого существенного значения, как при малых выборках.

Первое представление о большой выборке почти всегда возникает на основе распределения частот. Нашим первым шагом будет изучение некоторых соответствующих данному случаю способов вычислений. После этого мы будем иметь дело с различными методами, относящимися к большим выборкам из единичной нормальной совокупности. В главе 17 будут рассмотрены некоторые вопросы планирования больших выборок из нескольких совокупностей.

2. **Вычисление средней и стандартного отклонения на основе распределения частот.** В параграфе 4 главы 3 вам был обещан сокращенный способ вычисления указанных статистических показателей по распределению частот. Те же 511 выборочных средних, которые рассматривались там, перенесены в первые две колонки таблицы 67. Фактические выборочные средние были сгруппированы по классам с середиными их значениями 19, 20, ... фунтов; число средних в каждом классе указано во второй колонке. Для удобства расчетов за произвольное начало G выбрано 29 фунтов в связи с тем, что это число занимает срединное место в колонке, а также потому, что данному классу соответствует одна из наибольших численностей. Числами кода являются просто отклонения срединных значений классов от G . В четвертой колонке выписаны суммы отклонений для каждого отдельного класса. Например, в четвертом классе, срединное значение которого 22, отклонение равно -7 и таких отклонений всего 7, поэтому соответствующая сумма будет $-7 \times 7 = -49$. В нижнем конце таблицы показаны отдельно суммы положительных и отрицательных отклонений и их общая сумма 446.

Вычисленные в последнем столбце таблицы суммы квадратов отклонений получены наиболее удобным способом — путем умножения fX на X .

Возьмем в качестве примера снова середину четвертого класса 22; здесь умножение $(-7) \times (-49) = 343$ легче, чем вычисление $(-7)^2 \times 7$. Следует просто производить умножение на каждое число кода дважды, сначала f на X , что даст fX , после чего fX на X , что даст fX^2 . Вычисления, приведенные внизу таблицы, не представляют собой чего-либо нового.

Заметили ли вы идентичность этих чисел кода с теми, о которых говорилось в главе 5? В данном случае подразумевается, что каждая средняя кодируется сначала вычитанием 29, после чего производится округление. Возьмем среднюю, которая вычислена в параграфе 4 главы 3 и равна 33,2 фунта. Соответствующий код будет $33,2 - 29 = 4,2$, что при округлении дает $X = 4$. В действительности же в процессе построения распределения частот, выполняемого путем группировки, округление производится на первом этапе кодирования. Например, средняя 33,2 фунта попадает в класс, срединное значение которого 33 фунта. Вычитание произвольного начала теперь завершает кодирование. Легко убедиться, что числами кода, например, таких средних, как 31,6, 25,1 и 26,8 фунта, будут 3, -4 и -2. Но как быть со средней 30,5? Вспомните правило: округлять до ближайшего четного числа. Следовательно, $30,5 - 29 = 1,5$ при округлении дает 2. Но и $31,5 - 29 = 2,5$ также дает при округлении 2. В результате получается небольшое смещение численностей в сторону четночисленных классов. Эти классы оказываются несколько шире других.

ТАБЛИЦА 67

Распределение частот 511 средних в выборках по 10 наблюдений (табл. 18).
Вычисление средней и стандартного отклонения. $G = 29$ фунтов

Середина класса (фунты)	Частота f	Числа кода X	Сумма чисел кода fX	Сумма квадратов чисел кода fX^2
19	1	-10	-10	100
20	1	-9	-9	81
21	0	-8	0	0
22	7	-7	-49	343
23	5	-6	-30	180
24	10	-5	-50	250
25	19	-4	-76	304
26	30	-3	-90	270
27	41	-2	-82	164
28	48	-1	-48	48
29	66	0	0	0
30	72	1	72	72
31	56	2	112	224
32	46	3	138	414
33	45	4	180	720
34	22	5	110	550
35	24	6	144	864
36	12	7	84	588
37	5	8	40	320
38	0	9	0	0
39	1	10	10	100

$$n = \sum f = 511$$

$$-444$$

$$\sum fX^2 = 5592$$

$$\frac{890}{446}$$

$$\sum fX = 446$$

$$\sum fX^2 = 5592$$

$$\sum fX/n = 446/511 = 0,87$$

$$(\sum fX)^2/n = (446)^2/511 = 389,27$$

$$\bar{x} = G + \sum fX/n = 29 + 0,87 = 29,87 \text{ фунта}$$

$$\sum x^2 = 5202,73$$

$$s_x = \sqrt{s^2/n} = \sqrt{10,2014/511} = 0,141 \text{ фунта}$$

$$s^2 = \sum x^2/(n-1) = 5202,73/510 = 10,2014$$

$$s = 3,194 \text{ фунта}$$

Второе правило главы 5, на которое следует обратить внимание, состоит в требовании, чтобы размах чисел кода был не меньше 20. В случае, который здесь рассматривается, для точности обработки необходимо иметь не менее 20 классов; предпочтительнее иметь несколько большее число их. После проведения вычислений можно применить еще такой критерий: выборочное стандартное отклонение должно быть по крайней мере в 4 раза больше классового интервала. Вы можете заметить, что это требование не полностью удовлетворяется в таблице 67: значение s только на немного превосходит трехкратное значение интервала, который у нас равен 1 фунту. Было бы лучше взять интервал в половину или в три четверти фунта.

Следует иметь в виду, что такого рода потери точности обусловлены только применением чисел кода, получаемых при группировке и округлении. Все это обычно ограничивается случаями, когда переменная непрерывна и когда наблюдения могут попадать на любую точку данной шкалы. Ни о какой жертве в точности вычислений не может быть речи, когда переменная выражается такими целочисленными значениями, как в следующем примере.

Пример 1. Эдуардс [3] на каждой из 183 агаровых пластинок размещал по 20 семян сои и определял после этого число проростков. Распределение частот было таким:

Число проростков на каждой пластинке X . . .	6	7	8	9	10	11	12	13	14	15	16
Число пластинок f	5	9	8	19	26	34	26	22	21	10	3

Определите среднее число проросших семян на пластинке (11,2) и выборочную стандартную ошибку (0,17 семени). Возникает ли какая-нибудь неточность в этих расчетах от применения произвольного начала? Обратите внимание на использование здесь этих данных, полученных в результате подсчета так, как если бы они были результатами измерений. Это часто представляется удобным. Если выборка велика и распределение является симметричным биномиальным распределением, то разработку данных можно без всякого опасения вести так, как если бы они были данными о нормально распределенной непрерывной переменной.

3. Вычисление средней и стандартного отклонения на основе распределения частот (продолжение). В предыдущем параграфе единица измерения — 1 фунт — являлась удобной для применения ее в качестве классового интервала; в распределении же таблицы 68 для получения достаточного количества классов взят интервал в 10 фунтов. Интервал первого класса простирается от 75 до 85 фунтов (но не включает это последнее число); середина интервала приходится на 80 фунтов. За произвольное начало G взята середина класса с наибольшей численностью, несмотря на то что этот класс не является серединой всего ряда классов. (Величина поправки 11 фунтов указывает на то, что лучше было бы взять G равным 180 фунтам.) Значение классового интервала 10 фунтов при обратном переходе от кода должно быть введено в расчеты средней и дисперсии.

Вы можете заметить, что и здесь мы все еще не достигли достаточного количества классов. Стандартное отклонение оказалось несколько меньше четвертого интервала, но это различие не столь серьезно.

Пример 2. Вычислите выборочную среднюю и стандартное отклонение для данных о живом весе таблицы 68, взяв в качестве произвольного начала $G=180$ фунтам.

Пример 3. Уилб [11] определил урожай (в граммах) растений 1499 рядков посева пшеницы. По этим данным составлена следующая таблица:

Середина класса	Частота	Середина класса	Частота	Середина класса	Частота
375	3	600	127	825	10
400	13	625	140	850	10
425	41	650	122	875	4
450	99	675	94	900	4
475	97	700	64	925	2
500	118	725	49	950	3
525	138	750	31	975	1
550	146	775	26	1000	1
575	136	800	20		
				Итого	1499

Вычислите $\bar{x}=587,74$ г и $s=100,55$ г. Имеется ли в этом распределении достаточное число классов?

ТАБЛИЦА 68

Распределение частот живого веса 533 свиней. Вычисление средней и стандартного отклонения при классовой интервале, отличающемся от единицы.
 $I=10$ фунтам, $G=170$ фунтам

Середина класса (фунты)	Частота f	Числа кода X	Сумма чисел кода fX	Квадраты fX^2
80	1	-9	-9	81
90	0	-8	0	0
100	0	-7	0	0
110	7	-6	-42	252
120	18	-5	-90	450
130	21	-4	-84	336
140	22	-3	-66	198
150	44	-2	-88	176
160	67	-1	-67	67
170	76	0	0	0
180	55	1	55	55
190	57	2	114	228
200	47	3	141	423
210	33	4	132	528
220	30	5	150	750
230	23	6	138	828
240	11	7	77	539
250	5	8	40	320
260	5	9	45	405
270	4	10	40	400
280	5	11	55	605
290	2	12	24	288
$n=533$			-446 1011 $\Sigma fX=565$	$\Sigma fX^2=6929$

$$\Sigma fX = 565$$

$$I \times (\Sigma fX) / n = 10 (565 / 533) = 11 \text{ фунтам}$$

$$\bar{x} = G + I \times (\Sigma fX) / n = 170 + 11 = 181 \text{ фунты}$$

$$s_x = \sqrt{s^2 / n} = \sqrt{1189,86 / 533} = 1,49 \text{ фунта}$$

$$\Sigma fX^2 = 6929$$

$$(\Sigma fX)^2 / n = 565^2 / 533 = 598,92$$

$$\text{Для чисел кода } \Sigma x^2 = 6330,08$$

$$s^2 = I^2 \times (\Sigma x^2) / (n-1) = 10^2 \times 6330,08 / 532 = 1189,86$$

$$s = 34,5 \text{ фунта}$$

Пример 4. Линдстром [6] подсчитал число рядков, образуемых зернами в 327 початках растений кукурузы поколения F_2 от скрещивания инбредных линий:

Число рядков зерен	12	14	16	18	20	22
Число початков	25	75	133	68	21	5

Выборочная средняя равна 16,00 рядков, выборочное стандартное отклонение 2,136 рядка. Возникает ли здесь какая-нибудь неточность в расчетах в связи с небольшим числом классов?

4. Различные замечания по поводу распределения частот. В отношении числа классов в распределении частот существует некоторое противоречие. В данном случае следует считаться с двумя различными целями: а) придание наблюдениям обобщенной формы, обычно связанное с графическим изображением распределения, и б) вычисление выборочных статистических характеристик. Требование иметь 20 или более классов относится только к последним условиям. Для построения же таблиц и графиков обычно желательно иметь число классов меньше 10.

При определении середины класса в качестве центра его интервала нужна некоторая осторожность. Допустим, что вес учитывается в целых фунтах, без дробных его значений. Определение таких весов на практике делается с точностью до половины фунта, т. е. цифра 25 фунтов включает в себя все возможные веса от $24\frac{1}{2}$ до $25\frac{1}{2}$ фунтов. Если в качестве классовых интервалов взяты 20—29, 30—39 и т. д., то первый класс фактически простирается от 19,5 до 29,5 фунта. В этом случае средней класса будет $(19,5 + 29,5) : 2 = 24,5$, а не 25. Поэтому сначала фиксируйте с возможной точностью границы класса и уже после этого определяйте точку середины класса.

При группировке предполагается, что наблюдения, появившиеся в один класс, распределены внутри интервала до некоторой степени равномерно. Если это условие осуществляется в действительности, то средняя из этих наблюдений будет совсем незначительно отклоняться от середины интервала. По преимуществу это возможно в интервалах с большими численностями, играющими наибольшую роль в расчетах. Если же внутри интервала имеется некоторая особая группировка данных, то необходима определенная осторожность. Например, производилось наблюдение степени концентрации семян томата в гнезде плода, причем учет ограничивался целыми и половинными гнездами. Оказалось, что только небольшое число плодов имели половинные гнезда. Сначала классовые интервалы были взяты размером от 2 до 3, не включая это последнее число, и т. д., в результате чего серединами интервалов оказались $2\frac{1}{2}$, $3\frac{1}{2}$ и т. д. Фактически же средние числа гнезд в классах почти совпадали с нижними границами их 2, 3 и т. д. Такая систематическая ошибка приводит к преувеличению средней почти на половину гнезда.

В целях очищения оценки стандартного отклонения непрерывной переменной от такого рода погрешностей можно применить поправку на группировку Шешарда [10]. Она сводится к вычитанию из кодированного среднего квадрата одной двенадцатой части единицы. В таблице 67 эта поправка приводит к $10,2014 - 0,0833 = 10,1181$ и $s = 3,181$. Однако такая исправленная величина s не должна применяться при оценке существования [5].

Пример 5. Примените в таблице 68 поправку Шешарда для определения оценки $s = 34,37$. После исправления кодированный средний квадрат будет 11,8986.

Пример 6. Введите в примере 3 поправку к значению $s = 100,55$ г.

Пример 7. Имеется ли в примерах 1 и 4 этой главы, в которых переменные дискретны, какая-нибудь ошибка группировки, требующая введения поправки?

Пример 8. Две группы мышей, зараженных различными штаммами тифа (см. параграф 16 главы 12), оставались в живых в течение различного числа дней, что характеризуется следующим распределением частот:

Число дней до смерти	1	2	3	4	5	6	7	8	9	10	11	12
Штамм 9D	7	5	2	16	11	7	2	5	2	2	0	0
Штамм DSC1	1	0	0	1	6	8	4	6	3	3	1	3

Оцените существенность разности между средними числами дней до смерти (глава 4). Суммы квадратов равны 316,88 и 209,64; разность между выборочными средними равна 2,628 дня, $t=5,22$ и число степеней свободы=93.

5. Критерий симметричности. Ранее изложенные положения, которыми вы могли бы руководствоваться при своем суждении о нормальности тех или иных рядов данных, относятся только к требованию симметричности и определенной доли наблюдений, содержащихся в интервале $x \pm s$. Более строгие критерии нормальности будут изложены в этом и последующих параграфах 14). Мы рассмотрим два вида отклонений от нормальности. При одном из них распределение данных по своей форме ассиметричное или скошенное; в этом случае средняя и медиана не совпадают друг с другом. Другой вид отклонений от нормальности встречается в симметричных рядах и характерен некоторым излишком или недостатком наблюдений, сконцентрированных в центре ряда. В целях упрощения изложения мы сначала на примере малой выборки покажем только критерий симметричности.

В примере 4 главы 2 в качестве образца ассиметричного ряда были приведены данные о весе 11 мужчин. Произведем измерение этой скошенности и оценим существенность ее. Соответствующие вычисления приведены в таблице 69. Первое, что нам ранее не встречалось, это колонка кубов с их сум-

ТАБЛИЦА 69

Критерий симметричности в малой выборке
Вес (в фунтах) 11 мужчин

Вес X	Отклонение x	Квадрат x ²	Куб x ³
148	-24	576	-13824
154	-18	324	-5832
158	-14	196	-2744
160	-12	144	-1728
161	-11	121	-1331
162	-10	100	-1000
166	-6	36	-216
170	-2	4	-8
182	10	100	1000
195	23	529	12167
236	64	4096	262144
$\bar{x} = 172$	$S_1 = 0$	$S_2 = 6223$	$S_3 = 248628$
$n = 11$	$k_2 = S_2 / (n-1) =$ $= 6226 / 10 =$ $= 622,6$	$k_3 = nS_3 / (n-1)(n-2) =$ $= 11 \times 248628 / (10 \times 9) =$ $= 2734908 / 90 = 30387,9$	
$g_1 = k_3 / \sqrt{k_2^3} = k_3 / k_2 \sqrt{k_2} = 30387,9 / (622,6 \times$ $\times \sqrt{622,6}) = 30387,9 / 15535 = 1,96$		$s_{g1}^2 = 6n(n-1) / (n-2)(n+1)(n+3) =$ $= (6 \times 11 \times 10) / (9 \times 12 \times 14) = 660 / 1512 =$ $= 0,437$	
$s_{g1} = 0,661$		$t = g_1 / s_{g1} = 1,96 / 0,661 = 2,96$; число степеней свободы = ∞	

мой $S_3=248\ 628$. Следующее, что мы замечаем, это две характеристики k , из них k_2 является известным нам средним квадратом, а k_3 — соответственно средней из третьих степеней отклонений от средней. Мерой скошенности является $g_1=1,96$. Очевидно, что S_3 , и поэтому g_1 могут быть положительными, отрицательными или равными нулю. При g_1 , равном нулю, симметричность выборки очевидна. Положительное g_1 , как это имеет место в нашем примере, указывает на перевес в численности наблюдений, которые меньше средней. Если будет доказана существенность отклонения g_1 от нуля, то такой излишек малых значений может быть приписан асимметричности совокупности, из которой взята данная выборка. Стандартная ошибка g_1 у нас оказалась равной 0,661, так что $t=2,93$. При этой оценке существенности табличное t берется при бесконечном числе степеней свободы; для 1%-ного уровня табличное t равно 2,576.

Чтобы помочь вам следить за этими вычислениями, не отвлекая внимания на второстепенные вопросы, мы привели их в таблице 69, оставляя в стороне правила для установления необходимого числа значащих цифр. Вы можете для приобретения практических навыков вычислить те же показатели, взяв не более четырех значащих цифр у каждой величины.

В качестве второго примера произведем оценку симметричности распределения количества нейтрального жира в плазме крови у 64 нормальных мужчин [8]: 419, 162, 149, 219, 248, 313, 211, 169, 91, 281, 264, 172, 124, 235, 94, 62, 224, 58, 92, 205, 132, 145, 305, 285, 174, 107, 240, 269, 396, 416, 662, 703, 249, 179, 136, 157, 198, 95, 100, 178, 145, 199, 54, 407, 166, 94, 248, 235, 66, 120, 239, 128, 560, 233, 80, 557, 217, 542, 252, 175, 103, 165, 351, 107 мг на 100 куб. см.

Эти данные будем обрабатывать при помощи счетной машины. Накапливаем сумму 64 данных $s_1=14\ 361$, сумму их квадратов $s_2=4\ 518\ 741$ и сумму их кубов $s_3=1\ 878\ 426\ 855$. Наименее просто последняя сумма кубов находится при помощи одной из таблиц, содержащих первые три степени целых чисел [1], [2], [7].

По s_2 обычным порядком определяется сумма квадратов отклонений от средней:

$$S_2 = s_2 - s_1^2/n = 4\ 518\ 741 - 14\ 361^2/64 = 1\ 296\ 267,$$

где s_1 является суммой X . Для определения суммы кубов отклонений от средней S_3 существует соответствующая формула:

$$S_3 = s_3 - 3s_1s_2/n + 2s_1^3/n^2 = 1\ 878\ 426\ 855 - 3 \times 14\ 361 \times 4\ 518\ 741/64 + 2 \times 14\ 361^3/64^2 = 282\ 723\ 308.$$

Имея эти результаты, можно провести полностью всю обработку, указанную в таблице 69, что приведет к $g_1=1,5696$ и $t=5,24$. Конечно, здесь нет никакого сомнения в асимметричности совокупности, из которой взята эта выборка. Положительное значение g_1 указывает на то, что средняя больше медианы.

6. Критерии нормальности при большой выборке. Возможна оценка двух видов отклонения распределений от нормальной формы, а именно по признаку скошенности и по признаку *крутизны* распределения. Последняя измеряется показателем g_2 , основанным на сумме четвертых степеней отклонений от средней. Если g_2 равно нулю, то у признака, характеризуемого этой величиной, нет никакого отклонения от нормального распределения. Положительное значение g_2 указывает на излишек наблюдений около средней и в то же время на дальнем расстоянии от нее при соответствующей недостатке наблюдений в промежуточных, боковых частях распределения. Именно в этом отношении распределение t отличается от нормального. Отрицательное значение g_2 возникает при плосковершинности кривых распределения.

Соответствующие вычисления приведены в таблице 70. Первые 5 колонок соответствуют таким же колонкам таблиц 67 и 68. Числа последних

5 столбцов получаются путем повторного умножения на X . Так, например, по 3-й строке получаем: $4 \times (-6) = -24$, $(-24) \times (-6) = 144$, $144 \times (-6) = -864$ и $(-864) \times (-6) = 5184$. Вы уже знакомы с системой остальных расчетов, сохраняющей силу и при расчете S_4 , k_4 , g_2 и g_2 . Соответствующие формулы приведены в таблице, но при вычислении по ним требуется известная осторожность.

Вы можете заметить, что g_1 и g_2 очень малы, каждая из них меньше своей стандартной ошибки. Отрицательное значение g_1 указывает на небольшую асимметричность при несколько большей численности наблюдений, превосходящих среднюю, что приводит к смещению вершины кривой распределения вправо. Отрицательное значение g_2 указывает на умеренную крутизну, придающую кривой распределения несколько плоскообразную форму вблизи его центра при несколько повышенном числе умеренных отклонений. Однако, так как ни g_1 , ни g_2 не являются существенными, у нас мало оснований считать данное распределение отклоняющимся от нормальной формы.

ТАБЛИЦА 70

Критерий нормальности распределений. Диаметр (в мм) початков желтозерной кукурузы в поколении F_2 (Линдстрем)

Середина класса	Частота f	Код X	fX	fX^2	fX^3	fX^4
36	1	-8	-8	64	-512	4096
37	0	-7	0	0	0	0
38	4	-6	-24	144	-864	5184
39	7	-5	-35	175	-875	4375
40	18	-4	-72	288	-1152	4608
41	26	-3	-78	234	-702	2106
42	28	-2	-56	112	-224	448
43	51	-1	-51	51	-51	51
44	47	0	0	0	0	0
45	49	1	49	49	49	49
46	31	2	62	124	248	496
47	33	3	99	297	891	2673
48	19	4	76	304	1216	4864
49	4	5	20	100	500	2500
50	8	6	48	288	1728	10368
51	1	7	7	49	343	2401
	327		$s_1=37$	$s_2=2279$	$s_3=595$	$s_4=44219$

$$S_2 = s_2 - s_1^2/n = 2274,81$$

$$S_3 = s_3 - 3s_1s_2/n + 2s_1^3/n^2 = -177,659$$

$$S_4 = s_4 - 4s_1s_3/n + 6s_1^2s_2/n^2 - 3s_1^4/n^3 = 44124,6$$

$$k_1 = s_1/n = 0,1131$$

$$k_2 = S_2/(n-1) = 6,9779$$

$$k_3 = nS_3/(n-1)(n-2) = -0,5483$$

$$k_4 = n[(n+1)S_4 - 3(n-1)S_2^2/n]/(n-1)(n-2)(n-3) = -9,563$$

$$g_1 = k_3/(k_2\sqrt{k_2}) = -0,0297 \quad g_2 = k_4/k_2^2 = -0,1964$$

$$s^2_{g_1} = 6n(n-1)/(n-2)(n+1)(n+3) = 0,0182$$

$$s^2_{g_2} = 24n(n-1)^2/(n-3)(n-2)(n+3)(n+5) = 0,0723$$

Стандартные отклонения: для $g_1 = 0,135$; для $g_2 = 0,269$. Значения t : для $g_1 = 0,22$, для $g_2 = 0,73$; число степеней свободы $= \infty$.

Пример 9. По данным таблицы 70 определите средний диаметр початка и его стандартную ошибку, $44,11 \pm 0,146$ мм.

Пример 10. По данным таблицы 16 вычислите $g_1 = -0,0139$ и $g_2 = 0,0460$, показав тем самым, что распределение практически может считаться нормальным.

Пример 11. В таблице 20 приведено выборочное распределение 511 стандартных отклонений. Вычислите $g_1 = 0,3074$ при стандартной ошибке 0,108. Это указывает на то, что ожидаемое в этом случае отклонение от нормальной формы существенно. Величина g_1 здесь положительна, так как мода лежит слева от средней.

Пример 12. Ниже приводятся 511 значений t , о которых говорилось в параграфе 7 главы 3. Весьма существенное значение g_2 (0,5340) говорит о том, что численности около

Середина класса	f	Середина класса	f	Середина класса	f	Середина класса	f
-3,13	3	-1,13	29	0,87	31	2,87	1
-2,88	5	-0,88	35	1,12	23	3,12	1
-2,63	1	-0,63	38	1,37	17	3,37	2
-2,38	3	-0,38	40	1,62	11	3,62	0
-2,13	6	-0,13	52	1,87	8	3,87	0
-1,88	12	0,12	57	2,12	10	4,12	0
-1,63	21	0,37	43	2,37	6	4,37	1
-1,38	16	0,62	37	2,62	2		
						Итого	511

моды и на концах распределения выше тех, которые должны быть в нормальном распределении, в то время как по бокам распределения они меньше. Этого и следовало ожидать. Но $g_1 = 0,1356$ не существенно, что также следовало ожидать, так как теоретическое распределение t симметрично.

7. Построение графика нормального распределения. Вы впервые познакомились с нормальным распределением в параграфе 1 главы 2; графическое изображение его было дано на рисунке 3. Уравнение, определяющее это распределение, таково:

$$y = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(X+\mu)^2}{2\sigma^2}}$$

На рисунке 22 представлено распределение, у которого μ равно нулю и σ равна 1, т. е. этот график дает $N(0,1)$. Представляется удобным положить $\tau = (X - \mu)/\sigma$, где τ является нормированным отклонением, откладываемым по абсциссе. В этом случае уравнение приобретает более простую форму:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\tau^2/2}$$

В таблице 71 даны ординаты этой кривой для различных значений τ . Например, при $\tau = 1$ ордината равна 0,2420, как это видно и из рисунка 22. Та же самая ордината относится и к $\tau = -1$. Другой пример: при $\tau = 2,46$ значение ординаты находится в столбце, озаглавленном 0,06, против значения $\tau = 2,4$, стоящего в левом столбце, что дает 0,0194. Вы можете заметить, что по мере того, как кривая отходит от центра, она прогрессивно снижается, пока не достигнет ординаты в точке $\tau = 1$; после этой точки уменьшение ординаты происходит медленнее.

Не представляет труда привести в соответствие ординаты любого такого распределения, как распределение диаметра початков кукурузы в таблице 70, с теоретическими ординатами таблицы 71. Вы уже знакомы с построением гистограммы для выборки. Теперь, вводя в дисперсию k_2 таблицы 70 поправку Шеппарда, вы получите $s = \sqrt{k_2 - 0,0833} = \sqrt{6,8946} = 2,626$ мм

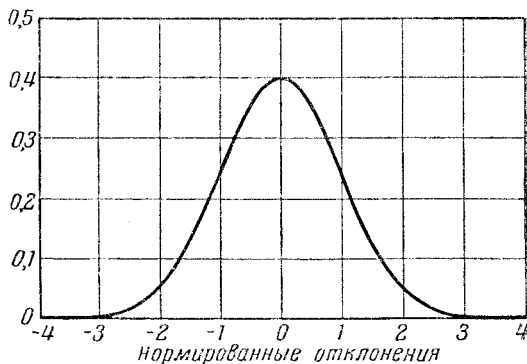


Рис. 22. Нормальная кривая распределения. На абсциссе отложены нормированные отклонения. Ординаты взяты из таблицы 71.

Ординаты нормальной кривой (умножение на 10 000)

Второй десятичный знак τ										
τ	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3725	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	989	973	957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046

Первый десятичный знак τ

τ	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
3	0044	0033	0024	0017	0012	0009	0006	0004	0003	0002
4	0001	0001	0001	0000	0000	0000	0000	0000	0000	0000

в качестве наиболее подходящей оценки параметра σ . Остается только определить постоянный множитель для перехода от ординат таблицы 71 к соответствующим ординатам выборки. Этот множитель определяется как отношение

$$\frac{n}{s} = \frac{327}{2,626} = 124,5.$$

Начнем с ординаты, соответствующей средней 44,11 мм. Ее значение $124,5 \times 0,3989 = 49,66$ початка; здесь второй множитель взят из таблицы 71 для $\tau=0$. Эта ордината откладывается на шкале частот вашей гистограммы. Далее вычислите ординату для одной из середины класса, положим для 39 мм. Нормированное отклонение этой точки будет $\tau = (39 - 44,11) / 2,626 = -1,95$. Соответствующая табличная ордината будет 0,0596; ордината,

которую следует отложить на графике, будет $124,5 \times 0,0596 = 7,42$ початка. В качестве последней иллюстрации вычислим ординату для одного из пограничных значений наших классов, положим для 47,5 мм. Вы находите $\tau = 1,29$, откуда искомая ордината будет 21,61 початка. Продолжая этот процесс, вы можете вычислить столько ординат, сколько окажется необходимым для проведения плавной нормальной кривой. В результате ваш график будет иметь форму, представленную в общем виде на рисунке 22.

8. Нормальное распределение накопленных частот. Таблица 72 отличается от обычного распределения частот в одном основном и в нескольких второстепенных отношениях. Главным различием является то, что здесь численности накоплены (и выражаются в процентах или площадях). В качестве иллюстрации рассмотрим число, стоящее во втором столбце против $\tau = 1,0$, где вы находите число 3413. Смысл этого числа тот, что 34,13% из общей численности в 10 000 падает на значения τ , лежащие между нулем и 1,00. Это соответствует тому факту, что 34,13% площади кривой рисунка 22 содержится между ординатами $\tau = 0$ и $\tau = 1$. Число 34,13% составляет половину знакомого вам числа 68,27%. Точно так же против 2,0 вы находите число 4772, имеющее тот смысл, что 47,72% частот лежит между $\tau = 0$ и $\tau = 2$. Удваивая этот процент, вы получите 95,44% численностей внутри интервала $\tau = \pm 2$. Остальные 4,56% численностей находятся вне этих пределов. Вообще числа этой таблицы соответствуют площадям рисунка 22 между ординатами 0 и τ .

Более второстепенные различия между распределением таблицы 72 и обычным распределением частот состоят в следующем.

1. Так как τ нормированное отклонение, то ее значения измеряются через стандартное отклонение совокупности σ , принятое в качестве единицы. Так, $\tau = 2$ означает $x = 2\sigma$ или $x/\sigma = 2$. 2. Значения τ не являются серединами классовых интервалов, а определяются расстояниями от центра до внешних границ интервалов. 3. В таблице 72 даны только положительные значения τ , соответствующие правой половине рисунка 22. 4. Значения τ даны с двумя десятичными знаками; вторая цифра дана в заголовке таблицы. В качестве иллюстрации последнего положения возьмем число 4535, стоящее в столбце, озаглавленном 0,08, против 1,6 первого столбца. Оно истолковывается так, что 45,35% частот расположены между $\tau = 0$ и $\tau = 1,68$.

В последующих абзацах даются различные применения таблицы нормального распределения частот, или, как теперь часто говорят, функции плотности нормального распределения.

а) У р о в н и τ . Это такие значения $|\tau|$, которые превышаются определенными процентами частот. Например, определим 5%-ный уровень, обозначаемый $\tau_{0,05}$. С положительной стороны 2,5% численностей имеют значения, большие $\tau_{0,05}$, так что 47,5% их расположены между 0 и $\tau_{0,05}$. По таблице значение τ , соответствующее 0,4750, равно 1,96. Выше 1,96 находится 2,5% общего количества частот, в то время как ниже $-1,96$ находятся другие 2,5%. Поэтому 5%-ный уровень $\tau_{0,05} = 1,96$. Это то же самое, что и значение 5%-ного уровня t в таблице 11 для бесконечного числа степеней свободы. Это значит, что по мере увеличения n до бесконечности распределение t стремится к нормальному распределению τ . При большой выборке (положим, $n > 50$) 5%-ный уровень t , так же как и τ , обычно округляется до 2.

б) В ы б о р о ч н а я с т а н д а р т н а я о ш и б к а р а з н о с т и д в у х с р е д н и х. При большой выборке обычно не делается различия между числом наблюдений n и числом степеней свободы. Вместо обобщенной суммы квадратов, как это делалось в главе 4, применяются формулы параграфа 10 главы 7. Если сравниваемые средние определены по случайным выборкам, то они теоретически некоррелированы, и поэтому

$$s_{x_1 - x_2}^2 = s_1^2/n_1 + s_2^2/n_2 = s_{x_1}^2 + s_{x_2}^2.$$

Нормальное распределение накопленных частот
 τ — нормированное отклонение; $n=10\,000$

τ	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4655	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4935
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3,0	4987	4987	4987	4988	4988	4989	4989	4989	4990	4990
3,1	4990	4991	4991	4991	4992	4992	4992	4992	4993	4993
3,2	4993	4993	4994	4994	4994	4994	4994	4995	4995	4995
3,3	4995	4995	4995	4996	4996	4996	4996	4996	4996	4997
3,4	4997	4997	4997	4997	4997	4997	4997	4997	4997	4998
3,6	4998	4998	4999	4999	4999	4999	4999	4999	4999	4999
3,9	5000									

Например, установлено, что у 841 13-летнего мальчика средний рост был $\bar{x}_1=57,30$ дюйма при $s_1=2,76$ дюйма, а у 784 девочек этого же возраста $\bar{x}_2=58,60$ дюйма при $s_2=2,44$ дюйма. Поэтому разность средних составит 1,30 дюйма со стандартной ошибкой

$$\sqrt{2,76^2/841 + 2,44^2/784} = 0,129 \text{ дюйма.}$$

95 %-ный доверительный интервал будет:

$$1,30 - 1,96 \times 0,129 \leq \mu_1 - \mu_2 \leq 1,30 + 1,96 \times 0,129.$$

Можно считать, что разность средних в совокупности находится между 1,05 и 1,55 дюйма; больший рост девочек по сравнению с мальчиками можно считать явно выраженной и характерной особенностью этого 13-летнего

возраста. То же заключение можно сделать в результате проверки гипотезы $H_0: \mu_1 - \mu_0 = 0$. Здесь $t = 1,30/0,129 = 10$, что намного больше 1%-ного уровня $t = 2,576$.

в) Распределение хи-квадрат с единичной степенью свободы. Это распределение то же самое, что распределение t при бесконечном числе степеней свободы. Следовательно, вероятность необычно больших значений хи-квадрат, лежащих вне пределов таблицы 5 (которая представляет собой первую строку таблицы 6), может быть определена по таблице 72, но с учетом того, что 5%-ный уровень t , с которым мы имели дело ранее, то же самое, что 2,5%-ная точка для t . Иллюстрация: в примере 23 главы 1 вас отсылали к настоящему параграфу для определения вероятности $\chi^2 > 8$. В этом случае мы имеем $t_{\text{степеней свободы}=\infty} = \chi_{\text{степеней свободы}=1}^2 = \sqrt{8} = 2,83$. Из таблицы видно, что вероятность для $t > 2,83$ равна $0,5 - 0,4977 = 0,0023$, откуда вероятность $|t| > 2,83$ равна $0,0046$, что и является вероятностью $\chi^2 > 8$. Другие случаи применения таблицы 72 будут даны в нижеприведенных примерах.

Пример 13. Пользуясь таблицей 72, показать, что 92,16% наблюдений, взятых из нормально распределенной совокупности, находится между $-1,76\sigma$ и $1,76\sigma$.

Пример 14. Показать, что 65,24% наблюдений, взятых из нормальной совокупности, находится в пределах между $\tau = -1,4\sigma$ и $\tau = 0,8\sigma$.

Пример 15. Показать, что 13,59% наблюдений попадает между $\tau = \sigma$ и $\tau = 2\sigma$.

Пример 16. Пользуясь таблицей 72, показать, что приблизительно 50% наблюдений находится в интервале от $-0,6745\sigma$ и до $0,6745\sigma$. Отклонение $0,6745\sigma$ называется вероятным отклонением, а $0,6745\sigma/\sqrt{n}$ — вероятной ошибкой.

Пример 17. Показать, что 1%-ный уровень равен $x/\sigma = 2,575$. В таблице 11 при $n = \infty$ вы найдете несколько более точное значение этого уровня.

Пример 18. Проверьте положение, что интервал $\bar{x} \pm 1,645\sigma$ содержит 90% численностей нормально распределенных.

Пример 19. Было установлено, что из 1000 мальчиков 13-летнего возраста у 390 рост не отклонялся от среднего роста (57,3 дюйма) более чем на 1,4 дюйма. Показать, что при допущении нормального распределения $s = 2,75$ дюйма.

Пример 20. Средняя из 1000 нормально распределенных подесяточных урожаев кукурузы составляла 20,50 фунта на делянку. Было установлено, что 200 делянок дали урожай 17,75 фунта или меньше. Определите стандартное отклонение урожая. *Ответ:* 3,27 фунта на делянку.

Пример 21. Средняя и стандартная ошибка процента убойного веса у одной большой группы свиней были $81,7 \pm 0,46\%$, а у другой $80,0 \pm 0,37\%$. Показать, что при оценке существенности этой разности $t = 2,88$, что указывает на высокую существенность разности.

Пример 22. По данным таблицы 70 вычислите ожидаемую численность класса, середина которого 42, т. е. интервал которого 41,5—42,5. Указание: в форме нормированных отклонений эти пограничные значения будут

$$\frac{41,50 - 44,11}{2,626} = -0,994 \quad \text{и} \quad \frac{42,5 - 44,11}{2,626} = -0,613.$$

Из таблицы 72 накопленные относительные численности будут 0,3399 и 0,2301. Следовательно, относительные численности внутри этого интервала составят $0,3399 - 0,2301 = 0,1098$. Наконец, ожидаемая численность будет $0,1098 \times 327 = 35,9$, этот результат можно сравнить с фактической численностью — 28 початков.

Пример 23. Вычислите для таблицы 70 ожидаемую численность для класса, середина которого равна 44. *Ответ:* 49,2 початка. Продолжая этот процесс, можно вычислить ожидаемые численности для всех интервалов этой таблицы и даже для концов распределения, в которых не было никаких наблюдений. Этот процесс рассматривается как градуирование распределения, которое состоит в подборе численностей классов, подчиняющихся определенному закону распределения, в данном случае закону нормального распределения. Сумма относительных численностей должна составлять 1, а сумма ожидаемых численностей должна дать 327. Полученное нормальное распределение имеет в качестве своих параметров статистические показатели $\mu = 44,11$ початка и $\sigma = 2,626$ початка, но оно все же является только приближением к $N(\mu, \sigma)$, из которого данная выборка початков была взята.

Степень согласия наблюдаемого распределения с вычисленным нормальным распределением может быть оценена при помощи хи-квадрат, число степеней свободы которого на 3 меньше, чем число классов. В параграфе 14 главы 1 мы брали число степеней свободы только на единицу меньше, чем число классов, но теперь мы ввели дополнительные ограничения, вычисляя по выборочным данным показатели \bar{x} и s . При каждом дополнительном ограничении должна вычитаться одна степень свободы.

Вычисление выборочного коэффициента корреляции по таблице частот с двумя входами.
Численности початков кукурузы с определенным диаметром и весом

Вес (в г)	Диаметр (в м)																f_y	Код Y	ΣXf_x	Произве- дение Y · ΣXf_x
	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51				
320													1				12	4	48	
310												1					11	3	33	
300													3				10	12	120	
290											1		1				9	6	54	
280										1	2	1	1		4		8	8	32	
270								1	1	1	1		2	1		1	7	7	23	
260							1	2	2	7		3	3				19	6	23	
250								3	1		3	2	2	1			11	5	24	
240								3	1	3	5	6	2		1		23	4	30	
230						1	1	1	4	4	3	4		2		20	3	26		
220						1	2	6	7	5	3	1	1		1	23	2	5		
210						4	1	7	4	4	3	3				28	1	4		
200				1	1	3	2	6	6	4	3	1	2			29	0	-7		
190						1	2	5	5	11	2	5	1			32	-1	22		
180				2		4	4	6	3	2	2	4	2			28	-2	-14		
170			1	1		2	5	5	1	4	2	2	2			21	-3	-20		
160				1	5	1	1	2	2	4	2	3			1	19	-4	-15		
150			1		1	2	1	4	1	2	1	1				14	-5	-15		
140					3	3	2	1		1			1			11	-6	-22		
130						2	1									3	-7	-8		
120			1	1	5			2								9	-8	-33		
110						2				1		1				3	-9	-4		
100					1			1								2	-10	-5		
90			1				1		2							4	-11	-8		
80				1	1	1	1									3	-12	-9		
70				1												1	-13	-5		
60					1											1	-14	-4		
50	1															1	-15	-8		
f_x	1	0	4	7	18	26	28	51	47	49	31	33	19	4	8	1	327	37	2318	
Код X	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7				

Пример 24. Проверьте вероятность 0,0002, которая была приведена в ответе примера 27 главы 1.

9. Вычисление r по случайной выборке из нормальной совокупности двух переменных. Таблица с двумя входами 73 содержит в компактной форме данные о двух количественных признаках 327 початков кукурузы [6]. Диаметры X были сгруппированы, как и в таблице 70, в классы через 1 мм; соответствующие частоты и числа кода приведены в нижних рядах таблицы. Классы по признаку веса Y взяты через 10 г; частоты и числа кода указаны справа. Корреляция этих данных проявляется в тенденции к расположению высоких численностей по диагонали таблицы и в том, что два угла таблицы остаются пустыми — здесь нет очень тяжелых початков с малым диаметром. В каждом строке (столбце) для отдельных диаметров имеет место заметное варьирование веса, но оно отнюдь не столь велико, как полный размах варьирования, представленный в столбце слева. Регрессия веса на диаметр будет представлена линией, проходящей снизу вверх по направлению расположения данных. Эта линия будет определена по методу наименьших квадратов для средних, относящихся к последовательным строкам диаметра.

Вычисления проводятся по уже известному образцу. Поэтому Σx^2 можно взять из соответствующей части таблицы 70, которая сама построена по образцу таблицы 67. Соответствующие вычисления для веса проводятся на основе данных столбцов f_y и Y . Результаты сведены в приведенной в конце этого параграфа схеме.

Мы пока еще не объяснили способ получения ΣXY , для которого ответены два последних столбца таблицы 73. Каждое число столбца ΣXf_x вычисляется по численности соответствующей строки и по числам кода в нижней строке. Например:

1) в 3-й строке: $3 \times 4 = 12$;

2) в 4-й строке: $1 \times 2 + 1 \times 4 = 6$;

3) в 5-й строке: $1 \times 1 + 2 \times 2 + 1 \times 3 + 4 \times 6 = 32$;

4) в 7-й строке: $1 \times (-2) + 3 \times (-1) + 7 \times 1 + 3 \times 3 + 3 \times 4 = 23$.

Произведения последнего столбца получаются из двух предшествующих столбцов. Сумма в конце этого столбца и является искомой ΣXY . Вы помните, что для вычисления r обратный переход от кода не является обязательным. Между прочим, сумма столбца ΣXf_x проверяется величиной s_1 из таблицы 70.

Вообще имеются различные способы обработки таблицы с двумя входами для определения r . Все они приводят к одним и тем же результатам и требуют примерно одинакового количества времени. Если представляется возможным применить перфорационные карточки, то при $n > 300$ (примерно) экономнее произвести расчеты сумм, сумм квадратов и сумм произведений нарастающим итогом на табуляторе.

Большое количество времени может быть потрачено впустую, если коэффициент корреляции вычисляется не по случайно отобранной выборке или выборкам, отобранным не из нормальной совокупности двух переменных (см. параграф 3 главы 7). В таких случаях r не оценивает никакого параметра. Верно, что r^2 является частью Σy^2 , которая обусловлена линейной регрессией, — это простой результат арифметических операций, но в этом случае мы имеем дело с регрессией, а не с корреляцией. Даже регрессия может не иметь смысла, если предполагается ее линейность, когда на самом деле она явным образом нелинейна. Перед началом какой-либо обработки большой выборки внимательно изучите данные, чтобы убедиться в том, что они согласуются с описанной в параграфе 3 главы 7 моделью. Вместе с этим рекомендуется вычертить обе регрессии двух переменных с целью узнать, будут ли они определены линейны. В главе 10 вы познакомитесь с критериями однородности дисперсий, а в главе 15 — с критериями линейности регрессии.

Схема вычисления коэффициента корреляции по данным таблицы 73

$$\begin{array}{r}
 \Sigma X f_x = 37 \\
 \Sigma X^2 f_x = 2279 \\
 (\Sigma X f_x)^2 / n = 4,19 \\
 \hline
 \Sigma x^2 = 2274,81
 \end{array}
 \qquad
 \begin{array}{r}
 \Sigma Y f_y = -46 \\
 \Sigma Y^2 f_y = 7264 \\
 (\Sigma Y f_y)^2 / n = 6,47 \\
 \hline
 \Sigma y^2 = 7257,53
 \end{array}
 \qquad
 \begin{array}{r}
 \Sigma XY = 2318 \\
 (\Sigma X f_x)(\Sigma Y f_y) / n = -5,20 \\
 \hline
 \Sigma xy = 2323,20
 \end{array}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{2323,20}{\sqrt{2274,81 \times 7257,53}} = 0,5718.$$

Пример 25. По данным столбцов f_y и Y таблицы 73 вычислите $\Sigma y^2 = 7257,53$, а также выборочную среднюю и ее стандартную ошибку $198,6 \pm 2,61$.

Пример 26. Вычислите выборочную среднюю 44,1 и стандартное отклонение 2,64 для весов в строе, который относится к 42-миллиметровому диаметру.

Пример 27. Вычислите для диаметров в строе, который относится к 200-граммовому весу, $\bar{x} = 198,6$ и $s = 47,18$.

Пример 28. Определите выборочный коэффициент регрессии веса на диаметр 1,0213 и уравнение регрессии $\hat{Y} = 1,0213X + 154,81$.

Пример 29. Вычислите для каждого из 16 строев диаметра таблицы 73 средний вес початка. Представьте эти средние в виде графика, у которого ордината — значение средней, а абсцисса — величина диаметра. На этом графике проведите линию регрессии. Получили ли вы хорошую согласованность ряда средних с регрессией? Не замечается ли здесь некоторая кривизна регрессии средних?

Пример 30. Вычислите для каждого из 28 строев веса средний диаметр. Расположите на графике эти средние в соответствии с серединами классов по весу. Не видна ли здесь сколько-нибудь выраженная криволинейность регрессии этих средних диаметров на вес? Можете вы написать уравнение регрессии, определяющее диаметр початка для каждого его веса?

Пример 31. Если вы интересуетесь математикой, то, конечно, сделаете попытку вывести формулу для ΣXY таблицы 73. Сама по себе эта таблица выдвигает много проблем, заслуживающих внимания. Превосходное изложение этого вопроса вы найдете в [9], где дана и соответствующая библиография.

Л И Т Е Р А Т У Р А

1. Barlow Peter, Tables. E. and F. N. Spon, Ltd., London, 1921.
2. Davenport C. B., E. K. S. Merle P., Statistical Methods in Biology, Medicine and Psychology. John Wiley and Sons, Inc., New York, 1936.
3. Edwards T. I., Plant Physiology, 9, 8, 1934.
4. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh. См. раздел 13 и приложение Д к гл. III.
5. Fisher R. A., Biometrics, 11, 237, 1955.
6. Lindstrom E. W., The American Naturalist, 49, 311, 1935.
7. Mathematical Tables. Chemical Rubber Publishing Co., Cleveland.
8. Page Irvine H., Kirk Esben, Lewis William H., Thompson William, Van Slyke Donald D., The Journal of Biological Chemistry, 111, 613, 1935.
9. Rietz H. L., Editor-in-Chief. Handbook of Mathematical Statistics. Houghton Mifflin Co., New York, 1924.
10. Sheppard W. F., Proceedings of the London Mathematical Society, 29, 353, 1898.
11. Wiebe Gustav A., Journal of Agricultural Research, 50, 331, 1935.

РЕЗУЛЬТАТЫ ПОДСЧЕТА ЧИСЛЕННОСТИ ПРИ ЧИСЛЕ СТЕПЕНЕЙ СВОБОДЫ, ПРЕВОСХОДЯЩЕМ ЕДИНИЦУ

1. Введение. Основная задача, поставленная перед собой автором в 1-й главе, состояла в том, чтобы в наиболее простой форме дать читателю представление о выводах, основанных на выборках и относящихся к некоторым совокупностям. Оценки значения и интервала, а также вопросы проверки и гипотез были даны там в столь сжатой форме, как только представлялось это возможным. Дальнейшее обсуждение вопросов, представляющих основной интерес в области статистики численностей, было оставлено до настоящей главы.

В последующих главах мы имели дело с нормальным распределением; оно служит основой для многих статистических методов, направленных на разработку данных, полученных в результате измерения величины объектов. Большинство методов настоящей главы, как и главы 1, относится к выборкам из биномиально или мультиномиально распределенных совокупностей, а также из совокупностей, имеющих распределение Пуассона. Более детально эти последние распределения будут рассмотрены в главе 16.

Все методы, рассмотренные в главе 1, за исключением параграфа 14, относились к единичным выборкам из совокупности, в которой объекты разделялись только на два класса, чему соответствует только одна степень свободы. Теперь мы перейдем к рассмотрению случаев, когда имеется не одна, а большее число выборок, а также большее, чем два, число классов.

В последующих двух параграфах этой главы, как это было и в главе 1, проверка гипотез будет производиться на основе *критерия согласия*; в этом случае речь идет о согласованности данных выборки с некоторыми теоретическими отношениями, подобными отношениям $3:1$ или $9:3:3:1$, или классов биномиального распределения. В дальнейшем будут рассмотрены гипотезы других видов — гипотезы о независимости и однородности.

2. Опыт с несколькими выборками. Довольно часто представляется более удобным и для получения более полных сведений выгодным собрать данные в виде нескольких небольших выборок, чем пользоваться одной большой выборкой. Этим путем можно получить существенно более обширную информацию о предмете исследования. В качестве иллюстрации в таблице 74 приведены данные некоторых опытов о наследовании содержания хлорофилла у кукурузы [14]. Эти данные состоят из 11 групп потомства самоопыленных гетерозиготных зеленых растений, разбитых по признаку доминирующей зеленой и рецессивной желтой окраски. Гипотетическое отношение этих признаков $3:1$.

Значения хи-квадрат вычислены для каждого отдельного потомства и даны в последней колонке таблицы. Рассмотрим в качестве примера первую группу потомства, состоящую из 122 растений, из которых 98 зеленых и 24 желтых. При гипотетическом отношении $3:1$ следует ожидать три чет-

верти от 122, т. е. 91,5, зеленых растений и одну четверть, т. е. 30,5, желтых; в этом случае хи-квадрат (если это вам необходимо, посмотрите параграф 10 главы 1) будет

$$\chi^2 = \frac{(98 - 91,5)^2}{91,5} + \frac{(24 - 30,5)^2}{30,5} = 1,85.$$

Это не будет слишком большим значением χ^2 для случайной выборки из однородного материала (см. табл. 6).

Так как большинство значений хи-квадрат очень малы и все эти значения не являются существенными, то у нас мало данных, говорящих за то, что эти 11 выборок потомства не могли появиться из совокупности, в которой отношение признаков 3 : 1. В действительности можно ожидать даже несколько большее варьирование этого отношения; вспомните, что в выборках из биномиального распределения можно встретить 1 выборку из 20, у которой хи-квадрат будет больше 3,841.

ТАБЛИЦА 74

Число зеленых и желтых проростков у 11 групп потомства самоопыленных гетерозиготных зеленых растений, а также ожидаемое число их при гипотетическом отношении 3 : 1 и соответствующие значения хи-квадрат

Число растений	Зеленые	Желтые	Отношение з/ж	Количество ожидаемых зеленых растений	Количество ожидаемых желтых растений	Хи-квадрат
122	98	24	4,1	91,50	30,50	1,85
149	110	39	2,8	111,75	37,25	0,11
86	68	18	3,8	64,50	21,50	0,76
55	42	13	3,2	41,25	13,75	0,05
71	54	17	3,2	53,25	17,75	0,04
179	144	38	3,7	134,25	44,75	1,36
150	120	30	4,0	112,50	37,50	2,00
36	27	9	3,0	27,00	9,00	0,00
91	70	21	3,3	68,25	22,75	0,17
53	39	14	2,8	39,75	13,25	0,05
111	85	26	3,3	83,25	27,75	0,15
В целом 1103	854	249		827,25	275,75	3,46

Одной из замечательных особенностей величины хи-квадрат является то, что сумма n выборочных значений ее сама распределена по закону распределения хи-квадрат с n степенями свободы. Даже в том случае, когда отдельные хи-квадрат будут являться малосущественными, их сумма может оказаться существенной в высокой степени. Таким образом, информация, заключенная в малых выборках, будучи объединена, дает вполне определенное освещение фактов, относящихся к совокупности. Например, возьмем $\chi^2 = 2,706$ из первой строки таблицы 6 при $P = 0,10$. Допустим, что имеется 5 одинаковых выборок, каждая из которых случайно дала одно и то же значение хи-квадрат, указанное выше. Сумма этих значений будет $5 \times 2,706 = 13,53$ при 5 степенях свободы, что превосходит уровень для $P = 0,025$.

Для потомства кукурузы сумма 11 значений хи-квадрат составляет только 6,54, что меньше, чем обычно ожидаемое значение для 11 степеней свободы. Такое объединение информации малых выборок еще больше подчеркивает согласованность группировки потомства с отношением 3 : 1.

Здесь имеются две другие величины хи-квадрат, которые могут дать нам сведения относительно таких совокупностей, как в нашем эксперименте. Значение хи-квадрат, вычисляемое по данным в целом, приведено в последней строке таблицы. Здесь 11 групп потомства обработаны как одна

большая выборка. Используя соответствующую формулу (параграф 16 главы 1), находим:

$$\chi^2 = \frac{(a-rb)^2}{r(a+b)} = \frac{(854-3 \times 249)^2}{3 \times 1103} = 3,46.$$

При одной степени свободы это значение χ^2 почти достигает 5 %-ного уровня. Причиной этого является то обстоятельство, что в 9 случаях из 11 отношение z/k больше 3. Столь устойчивая тенденция может быть вполне достаточной для того, чтобы привести к отказу от гипотезы об отношении 3 : 1. (См. пример 1 этой главы.)

Наконец, здесь имеется и хи-квадрат, измеряющее *неодинаковость* отклонений выборочных отношений от гипотетического. Фишер и Матер [11] назвали это свойство *неоднородностью*. Вычисление этого хи-квадрат очень просто:

	Степень свободы	Хи-квадрат
Сумма 11 хи-квадрат	11	6,54
Объединенное хи-квадрат	1	3,46
Неоднородность (различия между ними)	10	3,08

Несколько неожиданным в данном опыте с кукурузой является слишком малая неоднородность хи-квадрат при 10 степенях свободы. При данной нулевой гипотезе значения, превосходящие полученное здесь, ожидаются в 98 % случайных выборок. Колебания фактического отношения около гипотетического в данном случае много меньше, чем обычно бывают в выборочных наблюдениях. Такого рода чрезмерная устойчивость этого отношения может навести экспериментатора на мысль произвести проверку условий, в связи с которыми могло возникнуть столь малое случайное варьирование. Сравните эти результаты с результатами примера 2.

Теперь мы имеем три величины хи-квадрат, проверяющих одну и ту же нулевую гипотезу о том, что несколько выборок взяты все из одной совокупности, имеющей внутри себя определенное отношение признаков. Но им соответствуют различные альтернативные гипотезы. Для суммы хи-квадрат такой альтернативой является то, что ряд отношений отклоняется от гипотетического отношения, без учета того, каковы эти отклонения в сторону преувеличения или преуменьшения, так как при возведении отклонений в квадрат их знаки не играют никакой роли. Альтернативой для обобщенного хи-квадрат является наличие тенденции преимущественной роли отклонений, имеющих определенный знак. В случае хи-квадрат, характеризующего неоднородность, уже различаются знаки отклонений и определяется их неодинаковость.

Дальнейшее рассмотрение вопроса о хи-квадрат, характеризующего неоднородность, можно найти в параграфах 6 и 9 этой главы.

Пример 1. Из совокупности, в которой ожидается отношение признаков 3 : 1, взяты выборки, давшие следующие отношения: 98 : 41, 71 : 31, 127 : 52, 61 : 25, 86 : 36. Вычислите выборочные отношения и значения хи-квадрат (работа значительно облегчается, если взять формулу параграфа 16 главы 1). Сумма пяти значений хи-квадрат равна 6,73; $P=0,25$. Объединенное хи-квадрат равно 6,66; $P < 0,01$, что указывает на существенность устойчивого превышения числа рецессивных растений. Хи-квадрат для неоднородности, проверяющее наблюдающуюся согласованность выборочных отношений между собой, равно только 0,07.

Пример 2. Выборки с отношениями признаков 47 : 33, 40 : 26, 30 : 42, 24 : 34 были взяты из совокупности, в которой признаки находятся в отношении 1 : 1. Обратите внимание на различия между выборочными отношениями. Это приводит к большой неоднородности хи-квадрат, составляющего 9,01; $P=0,03$. Вывод из этого таков: в данном случае имеется не одна, а две совокупности, из которых взяты эти выборки.

3. Опыт с числом классов, по которым распределены объекты, большим двух. Представление об этом виде эксперимента было дано в параграфе 14 главы 1. Освобождаясь от ограничения иметь только единичную

степень свободы, этим путем мы можем проверить все возможные теоретические отношения генетики, причем это можно сделать как по отдельным выборкам, так и по их различным комбинациям. Например, в таблице 75 даны две группы растений поколения F_2 , полученных в результате скрещивания кукурузы антоциановой окраски и тонкополосатой. Группировка растений ожидалась в отношении 9 : 3 : 3 : 1. Следовательно, из 168 растений ожидалось зеленых $9/16$ частей, антоциановой окраски $3/16$ и т. д.

Формула хи-квадрат в параграфе 14 главы 1 дает сумму четырех составляющих частей:

$$(117 - 94,5)^2/94,5 = 5,36$$

$$(26 - 31,5)^2/31,5 = 0,96$$

$$(18 - 31,5)^2/31,5 = 5,79$$

$$(7 - 10,5)^2/10,5 = 1,17$$

$$\chi^2 = 13,28$$

Это значение χ^2 основывается на трех степенях свободы, на единицу меньше, чем число частей, на которое распадается выборка. Оно превосходит 1 %-ный уровень — 11,34.

ТАБЛИЦА 75

Растения F_2 , полученные от скрещивания кукурузы антоциановой окраски с тонкополосатой. Данные Линдстрёма [14]

Гипотетические отношения 9 : 3 : 3 : 1. В скобках даны гипотетические численности

Номера потомства	Число растений	Зеленые	С антоцианом	Тонкополосатые	Комбинированные	Хи-квадрат
1, 4, 6	168	117 (94,5)	26 (31,5)	18 (31,5)	7 (10,5)	13,28
2, 3	135	82 (76,0)	12 (25,3)	33 (25,3)	8 (8,4)	9,82
Сумма	303	199	38	51	15	

Проверяемая здесь гипотеза состоит в том, что совокупность разлагается по данному признаку в отношении 9 : 3 : 3 : 1. Следовательно, имеется немного шансов в пользу того, что растения этой группы образуют выборку из такой совокупности, т. е. эта выборка дает отношения, существенно отличающиеся от гипотетических.

Вторая группа потомства дает $\chi^2 = 9,82$ при 3 степенях свободы, что соответствует 2 %-ному уровню. Отклонение от гипотетического отношения здесь также существенно, но в меньшей степени, чем отклонение первой выборки.

Сумма этих двух значений хи-квадрат $13,28 + 9,82 = 23,10$ при 6 степенях свободы еще более подчеркивает наличие отклонений от гипотетических численностей, но не дает сведений относительно знака этих отклонений.

Объединенное хи-квадрат, определяемое по суммам последней строки таблицы, равно 12,37 при 3 степенях свободы и тоже существенно. Эта величина характеризует общую тенденцию отклонений обеих выборок отклоняться от гипотетических отношений в одном и том же направлении, хотя эта тенденция меняется на обратную в двух классах тонкополосатых растений.

Хи-квадрат для неоднородности вычисляется следующим образом:

	Число степеней свободы	Хи-квадрат
Сумма двух хи-квадрат	6	23,10
Объединенное хи-квадрат	3	12,37
Неоднородность	3	10,73

Отклонения с учетом их знаков существенны, но не в такой степени, как отклонения двух других видов.

Линдстром замечает: «Это распределение растений F_2 указывает на независимость наследования факторов антоциановой и тонкополосатой окраски, хотя эти данные несколько необычны в связи с тем, что имеются определенные трудности при классификации различных типов полосатости, а также в связи с тем, что имеет место различная выживаемость растений этих типов».

На этом мы заканчиваем рассмотрение вопросов о применении хи-квадрат к оценке согласованности наблюдаемых численностей с теоретическими. В главе 1 и в особенной в настоящей главе мы противопоставляли наблюдаемые данные некоторым теоретическим распределениям и отношениям. В дальнейшем, через два параграфа, мы перейдем к применениям хи-квадрат к оценке зависимости и однородности в таблицах сопряженности признаков при отсутствии предварительных сведений относительно соотношений этих признаков в совокупности.

Пример 3. Бейтсон и Паннетт [2] изучали душистый горошек, имевший цветы синей окраски (C), доминирующей над красной (k), и удлинненную форму зерен пыльцы ($У$), доминирующую над округлой (o). В своих двух опытах они вместо ожидаемого отношения $9 : 3 : 3 : 1$ получили следующее распределение растений по классам:

	CY	Co	kY	ko
Номер 61, 1910 г.	85	33	41	1
F 32, 1910 г.	72	35	28	0
В целом	157	68	69	1

В числе нескольких гипотез относительно данного явления они «...включили в рассмотрение частичное сцепление между C и $У$ », один из видов которого приводит к отношению $129 : 63 : 61 : 1$. Анализ хи-квадрат при такой нулевой гипотезе имеет следующий вид:

	Число степеней свободы	Хи-квадрат
Сумма двух хи-квадрат . . .	6	3,249
Объединенное хи-квадрат . .	3	0,961
Неоднородность	3	2,288

Так как эта гипотеза была взята в порядке подбора к имеющимся данным, то критерии гипотез здесь не вполне правомерны. Этот анализ указывает просто на то, что неоднородность хи-квадрат составляет основную часть суммы.

4. Объединение вероятностей. Данные о влиянии изучаемых факторов могут быть получены из опытов различного типа. Такие данные не могут непосредственно быть объединены и приведены к значениям хи-квадрат, которые можно было бы сложить. Но если каждый эксперимент дает определенные вероятности, то эти последние могут быть соответствующим образом объединены в целях обобщения данных, полученных из разных источников. Фишер (глава 10, параграф 21.1, а также см. в списке литературы [21]) показал, что величина $-2 \log P$ распределена по закону распределения хи-квадрат с 2 степенями свободы. Это приводит к методу, который мы применяем к условным данным.

В трех опытах с мелкими животными изучалось различие между двумя видами рациона. В первом опыте был произведен простой подсчет животных, давших или не давших превышение над определенным стандартным уровнем. В результате получено хи-квадрат, равное 2,7, с одной степенью свободы и $P=0,10$. Во втором опыте производилось сравнение двух групп животных, получавших различные рационы, что дало $t=1,65$ при 18 степенях свободы и $P=0,12$. В третьем опыте сравнивались отдельные особи и получено $t=2,00$ при 9 степенях свободы и $P=0,08$. Каждый из этих экспериментов указывал на наличие определенного влияния вида рациона, но ни один из них не был существенным при 5%-ном уровне. Однако можно думать, что объединение всех этих результатов может быть более убедительным. Таблица 76 показывает, как производится объединение вероятностей.

На основе полученных в ней результатов экспериментатор приобретает уверенность для отказа от гипотезы о том, что различие рационов не дает эффекта.

ТАБЛИЦА 76

Объединение вероятностей, полученных в трех опытах

Опыт	P	$\log_{10} P$
1	0,10	9,0000—10
2	0,12	9,0792—10
3	0,08	8,9031—10
Сумма		26,9823—30 = -3,0177

$$\log_e P = 2,3026 \times (-3,0177) = -6,95$$

$$\chi^2 = -2 \log_e P = 13,90; \text{ число степеней свободы} = 3 \times 2 = 6; P = 0,035$$

Если эти вероятности будут около 0,3, то объединение их не приводит к результатам, отличным от результатов отдельных опытов. Допустим, что вы имеете три опыта, для каждого из которых $P=0,3$; в этом случае вы можете убедиться в том, что объединенная вероятность все еще остается на уровне 0,3.

5. Поправка на непрерывность. В главах 4 и 9 распределение хи-квадрат, которое теоретически непрерывно, применялось к оценке дискретных данных. Скачкообразное изменение наблюдаемых данных на целое число единиц предопределяет то, что вычисленное хи-квадрат отличается от соответствующего точного значения, относящегося к одной из вероятностей, указанных в таблице 6. В результате получается некоторая неточность, смещение вероятности, взятой из этой таблицы, причем имеется тенденция преуменьшить ее, вследствие чего получается повышенное число случаев опровержения нулевых гипотез. В случаях, когда имеется единичная степень свободы, это смещение корректируется весьма просто. Соответствующая формула исправляется путем вычитания 0,5 из абсолютных значений отклонений. После этого все последующие вычисления и использование таблицы χ^2 производятся как обычно. При больших выборках эта небольшая поправка оказывает слабое влияние на вычисленное значение хи-квадрат, но введение ее столь просто, что она может быть принята в качестве обязательной во всех случаях.

Рассмотрим, например, одну из выборок цифр, полученных в параграфе 5 главы 1. Допустим, что в выборке из 10 цифр оказалось 8 нечетных. Как вы помните, фактическое отношение нечетных цифр к общему их числу равно 0,506. В этом случае мы имеем:

$$f_1 = 8, \quad f_2 = 2, \quad F_1 = 5,06, \quad F_2 = 4,94.$$

Подстановка в исправленную формулу хи-квадрат дает:

$$\begin{aligned} \chi^2 &= \sum (|f - F| - 0,5)^2 / F = \\ &= \frac{(|8 - 5,06| - 0,5)^2}{5,06} + \frac{(|2 - 4,94| - 0,5)^2}{4,94} = \\ &= \frac{(2,94 - 0,5)^2}{5,06} + \frac{(2,94 - 0,5)^2}{4,94} = \\ &= \frac{2,44^2}{5,06} + \frac{2,44^2}{4,94} = 2,38. \end{aligned}$$

Все изменение по сравнению с ранее применявшейся формулой состоит только в том, что абсолютные значения отклонений уменьшились с 2,94 до 2,44; знаки отклонений не принимаются во внимание, так как они не играют никакой роли при возведении в квадрат.

Преобразованная формула параграфа 11 главы 1 также легко поддается исправлению:

$$\chi^2 = (|f - F| - 0,5)^2 \times (1/F_1 + 1/F_2) = (|8 - 5,06| - 0,5)^2 (1/5,06 + 1/4,94) = 2,44^2 \times 0,40006 = 2,38,$$

т. е. то же значение, что и ранее.

Очевидно, что исправленное значение хи-квадрат всегда меньше неисправленного; последнее в нашем примере было равно 3,46. Это обстоятельство приводит к тому, что при исправленных значениях хи-квадрат из таблицы χ^2 берутся большие вероятности, т. е. эти вероятности меньше отклоняются от нормы; следовательно, нулевые гипотезы в этом случае имеют меньшую возможность быть отвергнутыми. Теоретическое обоснование этой поправки на непрерывность смотри в списке литературы [6, 10, 21, 22].

В тех случаях, когда $|f - F|$ оказывается меньше 0,5, данная поправка не вводится. В таких случаях хи-квадрат столь мало, что поправка не влияет на выводы.

В условиях, когда берется комбинация нескольких хи-квадрат, поправки на непрерывность должны быть опущены. Примерами этого могут служить выборки хи-квадрат (параграфы 11 и 12 главы 1) и анализ хи-квадрат, произведенный в параграфе 2 этой главы. Если в значения хи-квадрат будут введены поправки на непрерывность, то тем самым будет нарушена слагаемость частей, лежащая в основе анализа, описанного в параграфе 2.

Рассмотренная здесь поправка на непрерывность не применяется к данным, имеющим число степеней свободы больше единицы.

Формула параграфа 16 главы 1 корректируется путем вычитания из $(a - rb)$ величины $(r + 1)/2$. Применим эту поправку к данным указанного параграфа:

$$\chi^2 = \frac{(|a - rb| - (r + 1)/2)^2}{r(a + b)} = \frac{(|310 - 270| - 4/2)^2}{3 \times 400} = 1,20$$

вместо неисправленного значения 1,33. Заметим, что если совокупность делится на две равные части, то

$$r = 1 \quad \text{и} \quad \chi^2 = \frac{(|a - b| - 1)^2}{a + b}.$$

Пример 4. В примере 27 главы 1 $a = 2$, $b = 19$, $r = 1$. Исправленное хи-квадрат будет

$$(|2 - 19| - 1)^2 / 21 = 12,2$$

вместо неисправленного 13,76. Выводы не меняются.

Пример 5. В примере, который иллюстрировался в параграфе 2 этой главы, разделенные первой выборки на части было таким: 98 зеленых и 24 желтых растения, что дало неисправленное значение хи-квадрат, равное 1,85. Вычислите исправленное хи-квадрат по каждой из подходящих для этого формул. *Ответ:* 1,57.

6. Критерий независимости. Четырехклеточная таблица сопряженности признаков. После группировки выборки на основе одного качественного признака объекты, входящие в каждую из этих частей, в свою очередь, могут быть подразделены в соответствии с другим признаком. В качестве иллюстрации можно взять данные об обследовании посевов кукурузы округа Бун, упомянутого в параграфе 3 главы 1. Одним из качественных признаков было наличие яиц от второго потомства кукурузного мотылька, наблюдение над которым было проведено в промежутке между 14 августа и 6 сентября. В таблице 77 указано, что 125 полей из 176, взятых в выборку, были заражены этим вредителем. Вторым исследуемым качественным признаком было применение или неприменение минерального удобрения или навоза. Таблица 2×2 дает разложение выборки в соответствии с этими двумя признаками.

В связи с тем, что данная выборка участков является случайной, представляется возможным на основе методов главы 1 сделать целый ряд выводов. Для показа этих возможностей можно привести три таких примера.

а) В 31/176 доле случаев, т. е. на 17,6 % полей, были найдены оба признака — яйца второго поколения кукурузного мотылька на удобренном участке; 95 %-ный доверительный интервал для этого процента будет от 11,6 до 24,6 %. Если вся совокупность участков известна, то как оценка значения, так и оценка интервала может быть распространена на все число зараженных вредителем и удобренных полей округа Бун.

ТАБЛИЦА 77

Наличие или отсутствие яиц второго поколения кукурузного мотылька при выборочном обследовании посевов кукурузы округа Бун в зависимости от применения удобрений

Яйца второго поколения кукурузного мотылька	Удобрение		Итого
	есть	нет	
Наличие	31	94	125
Отсутствие	9	42	51
Итого	40	136	176

б) В данной выборке участков 125/176 часть, или 71,0 %, оказалась зараженной яйцами второго поколения мотылька; соответствующий 95 %-ный доверительный интервал будет от 63,0 до 78,5 %.

в) Из 40 удобренных полей, вошедших в выборку, 31 поле оказалось зараженным и 9 свободными от яиц. Теперь можно произвести проверку любой интересующей нас гипотезы; например, гипотезу о том, что половина удобренных полей округа Бун заражена яйцами вредителя. Проверка такой гипотезы сводится к ответу на вопрос: «Имеется ли существенное различие между долями зараженных и незараженных полей из числа тех, которые были удобрены?» Применяя третью формулу параграфа этой главы, находим $a=31$, $b=9$, $r=1$ и

$$\chi^2 = (31 - 9 - 1)^2 / 40 = 11,0.$$

Существенность различия очевидна.

При переходе к новым факторам четырехклеточной таблицы данный вопрос расширяется и становится таким: оказывает ли удобрение какое-либо влияние на наличие яиц второго поколения вредителя или оба эти признака независимы один от другого? Другая постановка этого вопроса касается двух отношений удобренных полей к неудобренным; эти отношения будут:

$$31/94 = 0,330 \text{ среди зараженных полей и}$$

$$9/42 = 0,214 \text{ среди незараженных полей.}$$

Являются ли эти отношения существенно различными? Третья форма этого же вопроса относится к проценту зараженных полей:

$$31/125, \text{ или } 24,8\%, \text{ среди удобренных полей и}$$

$$9/51, \text{ или } 17,6\%, \text{ среди неудобренных полей.}$$

Различаются ли эти проценты существенно?

Эти три вопроса логически являются одним и тем же вопросом: все они требуют ответа на основе нулевой гипотезы о независимости признаков, которая проверяется при помощи хи-квадрат. Если эта нулевая гипотеза верна, то приведенные выше пары могут рассматриваться как выборочные значения соответствующих отношений совокупности. Обоснование этого будет представлено в следующем параграфе, а пока остановимся на формуле, удобной для расчета хи-квадрат.

Четырехклеточная таблица 2×2 символически может быть представлена в таком виде:

$$\begin{array}{ccc} a & b & a+b \\ c & d & c+d \\ \hline a+c & b+d & a+b+c+d=n \end{array}$$

Одна из наиболее простых расчетных формул хи-квадрат для проверки гипотезы о независимости имеет вид:

$$\chi^2 = \frac{n(ad-bc-n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Дробь $n/2$ является поправкой на непрерывность. В выборках, состоящих из нескольких сотен наблюдений, и при тех условиях, когда эта поправка неприменима, эта дробь может быть опущена.

По данным выборочного обследования полей округа Бун имеем: $a=31$, $b=94$, $c=9$, $d=42$ и $n=176$. Делая подстановку этих данных, получаем:

$$\chi^2 = \frac{176 \times (31 \times 42 - 94 \times 9 - 176/2)^2}{125 \times 51 \times 40 \times 136} = \frac{176 \times 368 \times 368}{125 \times 51 \times 40 \times 136}$$

В этих расчетах необходимо иметь только 3 или 4 значащие цифры. Чтобы избежать громоздкости расчетов, удобно разделить числитель и знаменатель на одну и ту же подходящую степень 10, в данном случае на 10^6 . Это производится переносом запятых, отделяющих десятичные знаки, слева направо в целом на 6 знаков в каждом члене дроби:

$$\chi^2 = \frac{1,76 \times 3,68 \times 3,68}{1,25 \times 5,1 \times 4,0 \times 1,36} = \frac{23,8}{34,7} = 0,69$$

Как будет показано далее, критерий независимости в случае таблицы 2×2 основывается только на одной степени свободы; в данном случае P около 0,42, и поэтому нулевая гипотеза не отвергается. Каждая пара отношений, о которых говорилось выше, вполне могут быть в совокупности одинаковыми. Вывод таков, что удобрение поля не влияет на отложение яиц второго поколения вредителя; эти два процесса являются независимыми друг от друга.

Теоретически этот вывод имеет ту ценность, что на отложение яиц второго поколения мотылька не влияет внесение на поля удобрений. Практическое значение этого вывода состоит только в определенности его формы.

Если бы этот вывод был другим, то мы должны были бы пересмотреть второй пункт из тех трех заключений, о которых говорилось в начале настоящего параграфа. Раз доля зараженных полей зависит от удобрения, то общий процент зараженных полей уже не имеет почти никакого практического значения. Вместо него мы будем интересоваться двумя различными процентами: одним для удобренных и другим для неудобренных полей. В условиях разнородности (или зависимости) общий процент зараженных полей если и может иметь, то только весьма небольшое практическое значение.

Пример 6. Д-р К. Г. Ричардсон на сельскохозяйственной опытной станции штата Айова нашел после опрыскивания посевов двумя растворами натриевой соли олеиновой кислоты следующие количества живых и погибших экземпляров тли (*Aphis rumicis* L.):

Концентрация (в %)	Живые	Погибшие	Итого	Процент погибших
1,40	3	62	62	95,4
0,65	43	55	68	80,9
Итого . . .	46	117	133	

Вычислите исправленное хи-квадрат 5,31. Определите 95%-ный доверительный интервал для процента погибших насекомых в совокупности при высшей (87—99%) и при низ-

шей (72—92%) концентрации. Имеется ли здесь какое-либо основание для того, чтобы взять процент погибших насекомых во всей выборке в целом? Заметим, что итоговые данные 65 и 68 здесь выбраны до некоторой степени произвольно; они не дают никаких сведений об относительных размерах соответствующих двух совокупностей.

Пример 7. Ниже приводятся данные о прорастании семян мятлика, взятых из двух мешков:

Номер мешка	Число проросших семян	Число непроросших семян	Итого	Процент проросших семян
1	340	60	400	85
2	356	44	400	89
Итого . . .	696	104	800	87

Вычислите исправленное $\chi^2=2,49$. Имеющееся различие между процентами проросших семян может быть приписано выборочному варьированию, и поэтому 87% может рассматриваться в качестве несмещенной оценки процента проросших семян в обоих мешках вместе. Установите с вероятностью 0,99 доверительные границы для этого процента в совокупности. *Ответ:* 84—90%.

Пример 8. В 1953 г. Дэйвис [8] сделал попытку определить причину необычного развития пыльной головни (*Ustilago zeae*) на растении кукурузы. Так как все растения были поражены, то возможность заражения представлялась одинаковой. Поверх початков 174 растений непосредственно перед появлением столбиков были надеты мешочки, благодаря чему было предупреждено опыление и задержано развитие. Среди этих растений 38 были поражены пыльной головней. Среди же 262 растений, не покрытых мешочками, на 15 сентября оказалось 31 растение, пораженное болезнью. Хи-квадрат здесь равен 7,13, что указывает на существенное различие в проценте пораженных пыльной головней растений; это следует приписать влиянию изоляции.

Пример 9. При изучении действия различных способов опрыскивания для предупреждения повреждений яблок яблонной молью Хансберри и Ричардсон [12] произвели подсчет поврежденных плодов на каждом из 48 деревьев. Два дерева, опрысканные мышьяковокислым свинцом, дали:

- A : 2130 плодов, из которых 1299, или 61%, были повреждены;
- B : 2190 плодов, из которых 1183, или 54%, были повреждены.

Хи-квадрат=21,16 со всей определенностью свидетельствует о том, что размеры повреждений плодов на этих двух деревьях различны. Но они все же во всех отношениях были обработаны одинаково. Данное положение является характерным для опытов по опрыскиванию: во каком-то неизвестным причинам степень повреждения плодов при одинаковой обработке деревьев оказывается существенно различной. Следовательно, сравнение результатов опрыскивания на отдельных деревьях является нежелательным, так как различие в проценте поврежденных плодов может быть в одинаковой мере приписано как способам обработки, так и неизвестным причинам варьирования. Определение методами статистики однородности или разнородности подопытного материала при соблюдении одинаковых внешних условий получило название проверки техники эксперимента. Хансберри и Ричардсон объединили вместе изучение техники эксперимента и свой опыт по опрыскиванию.

7. Четырехклеточная таблица; частоты, основанные на гипотезе о независимости. Если гипотеза о независимости верна, то численности 4 клеток таблицы должны быть пропорциональны. Обозначим такие гипотетические частоты через F :

$$\frac{F_1}{F_3} = \frac{F_2}{F_4} = \frac{a+b}{c+d} = \frac{a+c}{n}$$

Краевые итоги здесь считаются неизменными, т. е. $F_1+F_2=a+b$ и т. д. В этом случае, если признаки независимы, должны наблюдаться равенства $F_1/F_2 = F_3/F_4$ и $F_1/F_3 = F_2/F_4$.

Эти гипотетические частоты легко определить на основе краевых итогов. Так,

$$F_1 = \frac{(a+b)(a+c)}{n} \quad \text{и т. д.}$$

Применим эти положения к выборочному обследованию полей округа Бун:

$$\begin{array}{ccc} F_1 & F_2 & 125 \\ \frac{F_3}{40} & \frac{F_4}{136} & \frac{51}{176} \end{array}$$

$F_1 = (125 \times 40) / 176 = 28,4$ и т. д. В этом случае в непосредственном вычислении нуждается только одна численность; остальные же находятся простым вычитанием из соответствующих краевых итогов:

$$F_2 = 125 - F_1 = 125 - 28,4 = 96,6 \text{ и т. д.}$$

В результате гипотетические частоты F в нашем случае будут

$$\begin{array}{cc} 28,4 & 96,6 \\ 11,6 & 39,4 \end{array}$$

При вычислении хи-квадрат эти гипотетические численности сравниваются с выборочными численностями f :

$$\begin{array}{cc} 31 & 94 \\ 9 & 42 \end{array}$$

Четыре отклонения $f - F$ будут такими:

$$\begin{array}{cc} 2,6 & -2,6 \\ -2,6 & 2,6 \end{array}$$

Сумма этих отклонений по каждому столбцу и по каждой строке равна нулю; вместе с тем все эти отклонения, если не обращать внимания на их знаки, одинаковы.

Так как абсолютные значения этих 4 разностей $|f - F|$ одинаковы, то обычная формула для хи-квадрат без поправки на непрерывность

$$\chi^2 = \sum \frac{(f - F)^2}{F}$$

превращается в такую:

$$\chi^2 = (f - F)^2 \sum \frac{1}{F}.$$

Подстановка данных обследования полей округа Бун дает

$$\chi^2 = 2,6^2 \left(\frac{1}{28,4} + \frac{1}{96,6} + \frac{1}{11,6} + \frac{1}{39,4} \right) = 4,06.$$

Поправка на непрерывность производится путем вычитания 0,5 из $|f - F|$. Формула скорректированного хи-квадрат будет:

$$\chi^2 = (|f - F| - 0,5)^2 \cdot \sum 1/F = (2,6 - 0,5)^2 \times 0,1572 = 0,69.$$

Это приводит нас к тому же ответу, который был получен в предыдущем параграфе более простым способом. Все эти формулы эквивалентны друг другу.

Теперь ясна и причина, почему в данном случае имеется только одна степень свободы. В связи с тем что теоретические численности основаны на краевых итогах, только одно из F свободно, остальные же целиком определяются через это F и краевые итоги.

При принятии гипотезы о независимости четырехклеточная таблица содержит в себе уже мало каких-либо дополнительных сведений, представляющих интерес. Данные выборки из 176 участков в этом случае можно рассматривать с двух точек зрения. Во-первых, установлено, что яйца второго поколения вредителя обнаружены на 71 % полей при 95 %-ном доверительном интервале от 63,0 до 78,5 %. Во-вторых, установлено, что около 23 % полей были удобрены при 95 %-ном доверительном интервале от 16 до 31 %. Все другие вопросы, так же как указанные в п. а и в в парагра-

фе 6, теперь уже не имеют смысла. Как и в случае других применений хи-квадрат при анализе таблиц сопряженности признаков, описанные выше методы, даже при введении поправки на непрерывность, являются приближенными. По мере уменьшения размера выборки это приближение становится все меньше и меньше. Если численность в некоторых клетках меньше 5, то это может сказаться на правильности выводов. Для ознакомления с такими крайними случаями читатель отсылается к [10], где дано описание точного метода, и к [7], где дано дальнейшее обсуждение этого вопроса.

Пример 10. Ниже приводятся результаты клинических наблюдений над 171 пациентом при терапевтическом заражении с помощью комара анофеллис штаммом *Plasmodium vivax*; применялись спороносители не старше 50-дневного возраста [3]. Наличие или отсутствие заражения малярией в январе — марте и в апреле—декабре:

Месяцы	Заражение		Итого
	было	отсутствовало	
Январь — март	13	6	19
Апрель — декабрь	142	10	152
Итого	155	16	171

Вычислите исправленное $\chi^2=9,67$. При опровержении нулевой гипотезы здесь можно сделать только два заключения по поводу совокупности; они будут основываться на двух парах оценок (значения или интервала) процента: 1) наличия (или отсутствия) заражения в январе — марте и 2) наличия (или отсутствия) заражения в апреле — декабре. Эти оценки не могут истолковываться расширенно, так как в данном случае нет никаких сведений относительно размера совокупности, из которой взята случайная выборка.

Пример 11. Рассматривая данные обследования полей округа Бун, обратите внимание на наличие между гипотетическими частотами такого соотношения: $F_1/F_2=F_3/F_4$ или $F_1F_4=F_2F_3$. Проверьте это расчетом.

Пример 12. В течение осени 1943 г. примерно одна из 1000 семей штата Айова была опрошена по поводу употребления консервированных фруктов или овощей в предшествующем сезоне. Из опрошенных 392 семей, живущих в сельской местности, 378 не употребляли консервов, в то время как среди 300 городских (т. е. поселений с 2500 и более жителей) семей 274 употребляли их. Проверьте гипотезу о том, что процент семей, употребляющих консервы, был в 1943 г. один и тот же в городских и сельских поселениях штата Айова. *Ответ:* $\chi^2=7,19$.

Два замечания: 1) При предположении, что каждая семья штата Айова имела одну и ту же вероятность 0,001 быть опрошенной, представляется возможным произвести оценку и обобщение данных каждой из 4 клеток таблицы и 4 крайних итогов. При интерпретации последних следует считаться с доказанной здесь неоднородностью данных. 2) Допущение о полной рендомизации в данном случае не вполне соответствует действительности; проведение такой рендомизации обходится слишком дорого. Разработанные практические способы выборочных наблюдений дают возможность обойти это затруднение. (См. главу 17.)

8. Критерий независимости в таблице $R \times C$. В тех случаях, когда один или оба качественных признака подразделяются на число градаций, большее двух, в результате чего образуется таблица с R строками и C столбцами, оценка независимости может быть произведена при помощи хи-квадрат с $(R-1)(C-1)$ степенями свободы. Как и в параграфе 7 этой главы, гипотетические частоты F могут быть рассчитаны по крайним итогам. В результате при наличии фактических численностей f мы приходим к знакомой нам форме χ^2 :

$$\chi^2 = \sum (f - F)^2 / F.$$

Странд и Джессен [19] изучали распределение аренды в соответствии с уровнем почвенного плодородия в округе Одюсон (штат Айова). Результаты случайного отбора ферм представлены в таблице 78. Перед тем как обратиться к выводам на основе крайних итогов, следует решить вопрос: не зависит ли

аренда от плодородия или не является ли распределение аренды одним и тем же при всех трех уровнях плодородия?

ТАБЛИЦА 78

Число ферм округа Одобон штата Айова в трех группах по уровню плодородия почвы. Распределение ферм по признаку аренды

	Почва	Собственная	Арендованная	Смешанная	Итого
I	<i>f</i>	36	67	49	152
	<i>F</i>	36,75	62,92	52,33	
	<i>f</i> - <i>F</i>	-0,75	4,08	-3,33	
II	<i>f</i>	31	60	49	140
	<i>F</i>	33,85	57,95	48,20	
	<i>f</i> - <i>F</i>	-2,85	2,05	0,80	
III	<i>f</i>	58	87	80	225
	<i>F</i>	54,40	93,13	77,47	
	<i>f</i> - <i>F</i>	3,60	-6,13	2,53	
	Итого	125	214	178	517

$$\chi^2 = \sum \frac{(f-F)^2}{F} = \frac{(0,75)^2}{36,75} + \dots + \frac{(2,53)^2}{77,47} = 1,54; \text{ число степеней свободы} = (R-1)(C-1) = 4$$

Как и ранее, гипотетические численности для любой клетки вычисляются по крайним итогам соответствующих строк и столбцов.

$$F = \frac{(\text{итог строки}) \times (\text{итог столбца})}{n} = \frac{\text{итог строки}}{n} \times (\text{итог столбца}).$$

Например: для первой строки,

$$\frac{\text{итог строки}}{n} = \frac{152}{517} = 0,29400$$

$$F_1 = 0,29400 \times 125 = 36,75$$

$$F_2 = 0,29400 \times 214 = 62,92$$

$$F_3 = 0,29400 \times 178 = 52,33$$

Эти вычисления становятся особенно простыми, если применяется счетная машина. Для контроля следует принять во внимание, что: 1) сумма *F* для любой строки или любого столбца равна соответствующему итогу и, следовательно, 2) сумма отклонений по столбцам и строкам равна нулю.

Эти положения сразу и полностью определяют число степеней свободы. Можно свободно выбрать *R*-1 теоретических частот в любом столбце, но оставшаяся клетка будет в этом случае заполнена определенной численностью, полученной в результате вычитания из итога столбца суммы *R*-1 значений *F*. Подобно этому, в каждом из *R*-1 ряду будет иметься *C*-1 степеней свободы. Следовательно, число степеней свободы = (*R*-1) (*C*-1).

Расчет хи-квадрат производится, как показано в таблице. Так как *P* > 0,8, то в данном случае нулевая гипотеза о независимости не будет отвергнута.

В связи с тем, что в наших данных о плодородии почвы и об аренде не выявилась сколько-нибудь существенная неоднородность, имеется возможность проверить любую гипотезу, которую может высказать экспериментатор по поводу крайних итогов. Например, если ему известно, что распределение ферм по признаку аренды в данном штате, взятом в целом, подчиняется отношениям 25 : 45 : 30, то он может произвести проверку (как указано в параграфах 14 главы 1 и 3 этой главы) того, является ли его выборка случайно

взятой из совокупности, относящейся ко всему штату, или округ Одиубон имеет распределение ферм по этому признаку иное, чем во всем штате. Или же он может задаться вопросом, не являются ли фермы обследуемого округа, имеющие различный уровень плодородия почвы, одинаковыми по численности; это будет уже особая гипотеза об отношении 1 : 1 : 1.

В тех случаях, когда одна или несколько клеток таблицы заняты необычно большими значениями $(j-F)^2/F$, то это может быть поводом для специального изучения. Следует ли эти большие компоненты приписать капризам выборочного метода или они могут получить некоторые, имеющие практический смысл, истолкования?

Если экспериментатор не находит нужным рассматривать компоненты отдельных клеток в своей таблице $R \times C$, то он может сэкономить половину времени, применяя сокращенный способ П. Г. Лесли [13]. Этот способ особенно выгоден, когда производится обработка целого ряда таких таблиц.

Пример 13. В трех опытах по искусственному осеменению коров получены следующие данные об успешном и неудачном оплодотворении:

Опыт	Число оплодотворенных коров	Число неоплодотворенных коров
1	194	158
2	80	86
3	165	151

Вычислите хи-квадрат. *Ответ:* 2,21; число степеней свободы=2.

Пример 14. Обращаясь к данным таблицы 74, допустим, что они не свидетельствуют в пользу того, что потомство разделяется в отношении 3 : 1; или допустим, что нет никакого теоретико-генетического отношения, которое было бы применимо в этих условиях. В таком случае исследователь может применить критерий независимости, как это сделано в настоящем параграфе. Вычислите хи-квадрат. *Ответ:* 3,31; число степеней свободы=10. Вы видите, что эта величина не та, что вычисленная в параграфе 2 для хи-квадрат или однородности. Это различие будет обсуждено в следующем параграфе.

9. Специальный метод вычисления хи-квадрат в таблице $R \times 2$. Во многих случаях таблицы, характеризующие явление на основе численностей, имеют особую форму, представленную в первых трех столбцах таблицы 79. Эти данные отобраны из опыта, в котором отдельным группам крыс были введены одинаковые дозы бактерий Danysz. Размеры отдельных групп определялись числом животных, подготовленных к инъекции на отдельные даты.

Обращают на себя внимание вероятности выживания p_i , приведенные в третьем столбце таблицы. Для таблиц такого вида Снедекором и Ирвингом [18] предложен специальный метод вычисления хи-квадрат, представляющий известные удобства (этот метод иногда называется методом Брандта и Снедекора). Конечно, окончательный результат в этом случае получается тот же самый, что и при обработке по методу предыдущего параграфа. В столбцах втором и третьем можно было бы поставить численности смертельных случаев, но работа упрощается, когда берутся меньшие вероятности. В последнем столбце приведена сумма произведений, которая получается, как обычно, путем накопления на счетчике вычислительной машины без выписки отдельных произведений.

Результаты обработки показывают, что вероятность гибели не зависит от группы животных, и поэтому можно считать ее характеристикой степени заразительности инъецированных бактерий.

В данном опыте контроль над техникой его проведения встречал определенные препятствия: неподопытные крысы в отношении их реакции на инъекцию представляли, вероятно, довольно варьирующий материал; инъецируемый организм, несомненно, в какой-то мере различался по своей вирулентности, и количество вводимого заразного начала не было, по всей вероят-

Процент крыс, выживших после инъекции бактерии *Danysz*

Численность группы	Число выживших X_i	Вероятность ¹ выживания p_i	Произведения $p_i X_i$
40	9	0,2250	2,0250
12	2	0,1667	0,3334
22	3	0,1364	0,4092
11	1	0,0909	0,0909
37	2	0,0541	0,1082
20	3	0,1500	0,4500
$n=142$	$\Sigma X_i=20$	$\bar{p}=\Sigma X_i/n=20/142=0,1408$ $q=1-\bar{p}=0,8592$	$\Sigma p_i X_i=3,4167$

$$\chi^2 = \frac{\Sigma p_i X_i - \bar{p} \Sigma X_i}{pq} = \frac{3,4167 - 2,8160}{0,1408 \times 0,8592} = 4,97; \text{ число степеней свободы} = 5; P = 0,43$$

¹ Если вы предпочитаете пользоваться процентным выражением вероятностей, то ознакомьтесь с примером 18.

ности, абсолютно одинаковым, — и тем не менее вероятность гибели животного по всем группам оказалась не более варьирующей, чем это можно ожидать в группах животных, случайно отобранных из одной совокупности, имеющей некоторую определенную вероятность гибели.

Данный метод [11] может представляться особенно удобным, когда при разработке генетического материала есть основание предполагать кроссинговер, или сцепление, и когда теоретически этот материал подразделяется в отношении $r : 1$. В целях удобства сравнения этот метод применим к данным таблицы 74, которые перенесены теперь в таблицу 80. Новым здесь является то, что дано в нижней части этой таблицы, где производится вычисление хи-квадрат для двух допущений:

а) что в совокупности имеется отношение 3 : 1; в этом случае значение хи-квадрат совпадает с показателем разнородности хи-квадрат, вычисленным в параграфе 2 этой главы как разность между суммой хи-квадрат и обобщенным хи-квадрат; б) что нет никакого теоретического отношения между частями совокупности. Что дает это сравнение?

Во-первых, так как $\chi^2=3,08$ при числе степеней свободы=10 слишком мало, что указывает на отсутствие разнородности между выборочными отношениями, то основной интерес сосредоточивается на объединенной величине хи-квадрат, которую удобнее всего вычислить по формуле (параграф 16 главы 1):

$$\chi^2 = \frac{(a-rb)^2}{r(a+b)} = \frac{(854-3 \times 249)^2}{3 \times 1103} = 3,46.$$

Теперь мы имеем как обобщенное хи-квадрат, так и характеристику неоднородности группы значений хи-квадрат; обе эти величины имеют определенное содержание; если это нужно, то в дополнение к ним может быть установлено итоговое хи-квадрат. В данном случае вычисления заметно менее трудоемки, чем расчеты таблицы 74.

Во-вторых, мы можем, как об этом говорилось ранее, использовать эти данные для решения вопроса: имеет ли здесь место кроссинговер или его нет. В то время как обобщенное хи-квадрат ниже 5%-ного уровня, наблюдается явно выраженная тенденция — недостаток рецессивных желтых растений. Если исследователь приписывает это кроссинговеру (или какому-либо другому генетическому явлению), то он отвергнет гипотезу: 3 : 1. В этом

случае он обратится к $\chi^2=3,31$, основанному только на предположении о независимости, проведя небольшое количество дополнительных вычислений.

Подразделение растений кукурузы на зеленые и желтые в 11 потомствах.
Данные таблицы 74

Альтернативный метод расчета

Число наблюдений	Число желтых растений X_i	Вероятность желтых растений p_i	Произведения $p_i X_i$
122	24	0,1967	4,7208
149	39	0,2617	10,2063
86	18	0,2093	3,7674
55	13	0,2364	3,0732
71	17	0,2394	4,0698
179	38	0,2123	8,0674
150	30	0,2000	6,0000
36	9	0,2500	2,2500
91	21	0,2308	4,8468
53	14	0,2642	3,6988
111	26	0,2342	6,0892
<hr/>			
1103	$\Sigma X_i = 249$	$\bar{p} = 0,22575$ $\bar{q} = 0,77425$	$\Sigma p_i X_i = 56,7897$ $\bar{p} \Sigma X_i = 56,2148$
			$\Sigma p_i X_i - \bar{p} \Sigma X_i = 0,5779$

Допущение отношения 3 : 1; $p = 0,75$; $q = 0,25$;

$$\chi^2 = \frac{\Sigma p_i X_i - \bar{p} \Sigma X_i}{pq} = \frac{0,5779}{0,75 \times 0,25} = 3,08.$$

При отсутствии какого-либо допущения об отношении

$$\chi^2 = \frac{\Sigma p_i X_i - \bar{p} \Sigma X_i}{pq} = \frac{0,5779}{0,2258 \times 0,7742} = 3,31.$$

Теперь мы имеем две величины хи-квадрат, характеризующие неодинаковость размера и знака отклонений от теоретических соотношений; эти величины в нашем случае не слишком различаются друг от друга. В отношении расчета они отличаются одна от другой только своими знаменателями. Если возникает необходимость в определении, как в параграфе 2, хи-квадрат, характеризующего неоднородность, то следует производить деление на гипотетические вероятности. Если же желательно проверить независимость признаков, то следует использовать вероятности, определенные на основе краевых итогов таблицы по примеру параграфа 8.

Пример 15. В другой выборке зараженных крыс данные были такими: 25, 21; 50, 48; 20, 15 и 20,17; здесь первое число в каждой паре чисел означает размер частных выборок, а второе — число погибших животных. Вычислите $\chi^2=6,69$; $P=8,6\%$.

Пример 16. Объедините значения хи-квадрат для этих двух выборок крыс, что должно дать $\chi^2=11,66$; число степеней свободы=8 и $P=17\%$.

Пример 17. Если у вас есть основание принять гипотезу о том, что каждая из двух этих выборок крыс была взята из однородной совокупности, то это даст возможность проверить гипотезу о том, что вероятность гибели животных в этих двух совокупностях одинакова, т. е. что две выборочные вероятности 85,9% и 87,8% различаются только в такой степени, которую можно ожидать в условиях случайных выборок из общей совокупности. *Ответ:* корректирование $\chi^2=0,070$ (параграф 6 этой главы).

Пример 18. Примените метод таблицы 80 к данным таблицы 74, используя для этого процентное выражение вероятностей вместо обычного. Следует обратить внимание на те небольшие изменения в формуле, когда вместо вероятностей берутся проценты:

Численность групп	Число желтых растений X_i	Процент желтых растений p_i	
139	41	29,4964 ¹	
102	31	30,3922	
179	52	29,0503	
86	25	29,0698	
122	36	29,5082	
628	$\Sigma X_i = 185$	$\bar{p} = 29,4586$ $\bar{q} = 70,5414$	$\Sigma p_i X_i = 5451,17$ $\bar{p} \Sigma X_i = 5449,84$ $\Sigma p_i X_i - \bar{p} \Sigma X_i = 1,33$

$$\text{Для проверки разнородности: } \chi^2 = \frac{100 (\Sigma p_i X_i - \bar{p} \Sigma X_i)}{25 \times 75} = 0,071$$

$$\text{Для проверки независимости: } \chi^2 = \frac{100 (\Sigma p_i X_i - \bar{p} \Sigma X_i)}{pq} = 0,064$$

¹ Для того чтобы получить результат, совпадающий с примером этой главы, я ввел два добавочных десятичных знака.

Поскольку мы из предыдущих вычислений знаем, что обобщенное хи-квадрат существенно на 1%-ном уровне, то получает положительное свое решение χ^2 -тест о наличии некоторого смещения отношения, возникшего от таких причин, как кроссинговер. Если это так, то гипотеза об отношении 3 : 1 отбрасывается и неоднородность определяется величиной $\chi^2 = 0,064$. Легко видеть, что эта неоднородность слишком мала в обоих случаях как при отношении 3 : 1, так и при независимости.

Пример 19. Бернетт [4] изучал влияние пяти способов хранения на жизнеспособность семян кукурузы. На влажном чердаке над кухней 111 зерен из 120 дали проростки, в закрытом подвале — 55 из 60, в открытом подвале — 55 из 60, на открытом воздухе — 41 из 48 и на сухом чердаке — 50 из 60. Вычислите $\chi^2 = 5,09$, число степеней свободы = 4, $P = 28\%$. Вычисления будут несколько проще, если взять число непроросших семян. Эти результаты содержат в себе мало данных, свидетельствующих против гипотезы, что в данном случае имеется однородная совокупность процента проросших семян, независимо от места хранения.

Пример 20. В ряде экспериментов, данные которых опубликованы Линдстромом [15], наблюдались следующие количества початков кукурузы, в которых обнаружено наличие рецессивного гена, определяющего признак сахаристости: 18 из 33, имевших 8 рядов зерен; 37 из 63, имевших 10 рядов; 27 из 70, имевших 12 рядов, и ни одного из 4, имевших 14 рядов. Легко видеть, что ожидаемое число сахаристых початков в последней группе меньше 5, т. е. меньше того минимума теоретической численности, который допустим при определении хи-квадрат. Поэтому следует объединить 12- и 14-рядковые початки в одну группу, в которой будет 27 сахаристых початков из 74. Если это сделать, то $\chi^2 = 7,4$, число степеней свободы = 2 и $P = 3\%$. Это свидетельствует, хотя и не слишком определенно, против однородности материала по признаку сахаристости. Представляется, что ген сахаристости связан с наличием в початке небольшого числа рядов зерен.

Пример 21. Взяв данные предыдущего примера, вычислите хи-квадрат при помощи формулы параграфа 8. Так как из общего числа 170 початков сахаристых оказалось 48,24%, то среди 33 восьмьрядковых початков следует ожидать 48,24%, или 15,92 початка, с этим геном. Отклонение фактической численности от ожидаемой составляет 18 — 15,92 = 2,08 початка. Подобно этому могут быть определены и остальные пять отклонений (всего три пары отклонений, в каждой из которых будут отклонения абсолютного размера, но противоположных знаков). В результате, как и прежде, получим хи-квадрат, равное 7,4.

10. Три ряда градаций качественных признаков; таблица сопряженности признаков с тремя входами. Не так уж редко встречаются наблюдения над тремя и большим числом качественных признаков; например, берут животных разной породы, разного пола при разных вариантах их кормления. При трех признаках соответствующую таблицу можно мыслить, как образованную путем наложения друг на друга в высоту L таблиц вида $R \times C$, что даст прямоугольную призму с тремя измерениями $R \times C \times L$.

В биологии такие таблицы с тремя входами чаще всего по каждому признаку имеют только две градации: успех или неудача мероприятия, сохранение жизни или гибель организма, мужской или женский пол и т. д. Обычно основной интерес при изучении таких таблиц представляет процент успеха мероприятия и т. д., а не отношение случаев успеха к случаям неудачи и т. д. Таблицы этого вида приводятся к таблицам с двумя входами, в которых переменной является указанный выше процент. Методы обработки этого вида данных будут приведены в примерах 6, 27 и 29 главы 12.

В тех случаях, когда признаки имеют три или более градации, для проверки гипотез о независимости применяются методы, описанные в [17]. Исчисление гипотетических частот приводит к тем же значениям хи-квадрат и в то же время дает возможность определить местоположение больших отклонений.

В таблице $2 \times 2 \times 2$

a	b		a'	b'
c	d		c'	d'

возможна проверка гипотезы о независимости, т. е. о равенстве отношений

$$\frac{a/b}{c/d} = \frac{a'/b'}{c'/d'} \quad \text{или} \quad \frac{ad}{bc} = \frac{a'd'}{b'c'}$$

В таблице 2×2 проверка равенства процентов $a/(a+b) = c/(c+d)$ равносильна проверке равенства отношений $a/b = c/d$, но теперь в таблице $2 \times 2 \times 2$ соответствующее соотношение уже не сохраняется. Поэтому только в редких случаях при наличии таблицы с 3 входами интересуются указанной выше гипотезой о независимости. Если же она все-таки представляет интерес, то применяется метод, описанный в [1]. Некоторые сведения об этом методе были даны в параграфе 8 главы 9 четвертого издания этой книги.

11. Эксперимент и проверка техники его проведения. При применении современных экспериментальных методов исследователь среди целого ряда других моментов должен определить, сумеет ли он снова воспроизвести свои результаты, т. е. имеет ли он достаточный контроль над условиями, в которых проведен данный опыт. В параграфе 9 мы уже ссылались на эту проблему. Если результаты опыта не могут быть проверены при контролируемых условиях, которые считаются идентичными условиям данного опыта, тогда нет никакого смысла в установлении того влияния, которое произойдет при изменении этих условий; никто не может сказать, будут ли различия (или сходства) таких результатов происходить от изменения контролируемых условий или от каких-либо неизвестных причин, не попавших под контроль.

Деккер и Андре [9] столкнулись с этой проблемой в начале своего исследования вопроса о влиянии короткого и быстро устанавливающегося холода на взрослые особи клопа-черепашки. Подопытные насекомые были собраны прямо с поля, и поэтому однородность данного материала оставалась неизвестной. В каждую из 50 пробирок было помещено по 10 насекомых, которых в течение 15 минут содержали при температуре -8°C . В связи с тем, что метод хи-квадрат не может дать точных результатов, если ожидаемые численности меньше 5, при подсчете числа погибших насекомых произведено объединение наблюдений в 5 группы по 10 случайно отобранных пробирок в каждой, т. е. образованы группы по 100 насекомых в каждой. Смертность в этих группах соответственно была 14, 14, 23, 17 и 20 насекомых. По этим данным определено $\chi^2 = 4,22$, число степеней свободы = 4 и $P = 39\%$. Это примерно соответствует тому варьированию, которое следует ожидать в случайных выборках из однородного материала. Средняя вероятность гибели составляет 17,6%. Эти результаты находятся в соответствии с гипотезой о том, что каждое насекомое имеет одну и ту же вероятность гибели при снижении температуры.

Во второй выборке из 500 насекомых, отобранных и испытанных при тех же условиях, исключая только то, что теперь температура была снижена

до -9° С, смертность в группах по 100 насекомых составляла 38, 30, 30, 40 и 27. Хи-квадрат 5,79 характеризует технику проведения опыта и указывает на то, что здесь имеет место только выборочное варьирование около средней смертности 33%. Хотя это в данном случае и не представляет практического значения, все же интересно сложить две величины хи-квадрат и получить $\chi^2 = 4,22 + 5,79 = 10,01$; число степеней свободы = 8 и $P = 27\%$.

Явно выраженная выравниваемость результатов этих опытов приводит к уверенному выводу о несколько неожиданном факте существенного различия между количеством погибших насекомых при переходе от -8° до -9° С. Общее число погибших насекомых в обеих выборках по 500 составляет 88 и 165. Значение $\chi^2 = 31,4$, число степеней свободы = 1 и P меньше 0,0002 (параграф 8 главы 8) дает полную уверенность в том, что повышение процента гибели насекомых при понижении температуры с -8° С до -9° С является характерной особенностью совокупности, а не просто результатом варьирования при случайном отборе.

Легкость применения такого рода проверки техники эксперимента делает ее почти обязательной, исключая, разве, случаи работы с крайне стандартизированным материалом. Для того чтобы осуществить эту проверку, нужно только получать экспериментальные данные в виде нескольких небольших групп вместо одной большой их массы. Полученная таким образом дополнительная информация может в корне изменить выводы.

Пример 22. Р. Г. Уокер получил следующие неопубликованные данные о количестве бактерий в суспензии почвы и частиц идиго, составленной по методу Торнтона и Грея [20]. Число частиц идиго в поле зрения определило количество их в почвенном образце. Техника эксперимента считается хорошей, если число бактерий отклоняется от пропорциональности числу частиц идиго только в такой степени, в какой это допустимо при случайном отборе наблюдений из совокупности, в которой отношение между ними постоянно. Подсчет был произведен при скольжении по 16 полям зрения микроскопа, по четыре от каждой капли суспензии. Запись данных производилась по общему количеству в скользком поле микроскопа, по четыре таких поля на каждую почвенную суспензию. Для расчетов мы возьмем только два вида данных: 1) общее количество как бактерий, так и частиц идиго; 2) число бактерий. Для четырех пластинок с почвенной суспензией 1: 204,78; 206,75; 246,76; 278,95. Хи-квадрат = 5,22, число степеней свободы = 3, $P = 16\%$. Для суспензии 2: 260,60; 196,50; 198,45; 186,50. Хи-квадрат = 1,30, число степеней свободы = 3, $P = 73\%$. Для суспензии 3: 177,22; 177,23; 150,16; 177,20. Хи-квадрат = 0,53, число степеней свободы = 3, $P = 91\%$. Для суспензии 4: 289,46; 356,63; 281,45; 250,42. Хи-квадрат = 0,48, число степеней свободы = 3, $P = 92\%$. В данном случае согласованность данных большая, чем обычно можно ожидать при случайных выборках.

Пример 23. Для того чтобы определить, различаются ли существенно четыре почвенных суспензии по признаку числа бактерий на единицу объема, были взяты из предыдущего примера итоговые данные о числе бактерий: для суспензии № 1: 988,324; для № 2: 840,205; для № 3: 681,84 и для № 4 — 1176,196. Хи-квадрат = 130, число степеней свободы = 3, P меньше 1%. Нет никакого сомнения в различии по этому признаку четырех почвенных суспензий.

Пример 24. Метод вычисления хи-квадрат таблицы 79 применим в одинаковой мере как в случае, когда численности наблюдений в группах одинаковы, так и в случае, когда они разные. Если численности равны, то математик легко может привести данную ранее формулу к виду, соответствующему этому особому случаю: $\chi^2 = k \sum x^2 / \bar{x}(k - \bar{x})$, где k размер группы, $\sum x^2$ — сумма квадратов отклонений от средней, определяемая обычным порядком при помощи формулы $\sum x^2 = \sum X^2 - (\sum X)^2 / n$. Тот, кому часто приходится вычислять хи-квадрат, при использовании этой специальной формулы для групп одинаковой численности может сэкономить некоторое количество труда и времени. Эту формулу можно применить к данным Девкера и Андреса. Она впервые была применена «Мететес» [16].

12. Опыт, в котором качественный признак выражен количественно. Данные о численностях, о которых говорилось выше, относились к некоторому качественному признаку, подлежащему наблюдению. Но имеются и другого вида данные, когда различие признака может быть выражено несколькими единицами, благодаря чему наблюдения становятся количественными уже на первой стадии работы. Оба эти наблюдения применяются, например, во многих исследованиях, касающихся изучения насекомых. Так, в исследованиях интенсивности повреждений плодов яблонной молью можно не просто провести качественное различие между поврежденными и неповрежденными плодами, но подсчитать численности плодов, имеющих 0, 1, 2, и т. д.

личинок насекомого. Последнего вида данные часто приводят к распределению, известному под названием распределения Пуассона, подробное описание которого будет дано в главе 16. Соответствующий пример мы находим в опыте, предназначенном для исследования влияния различных видов обработки почвы на личинок капустной пяденицы [5]. Некоторая часть данных этого опыта в подытоженном виде приведена в таблице 81. Первоначальные записи результатов подсчета насекомых на отдельных растениях, например делянки № 2 варианта четвертого, были такими: 0, 1, 0, 0, 1, 2, 1, 0, 0, 0, 0, 4, 0 и т. д. На четырех делянках, составляющих один из вариантов опыта, не было обнаружено личинок, кроме только единичных экземпляров. В связи с этим для наших целей мы отобрали только такие варианты, которые дают по крайней мере пять личинок, ожидаемых для каждой делянки. На делянках с меньшим, чем это, числом личинок однородность данных оказалась даже бо́льшая, чем на тех, которые были отобраны нами.

Для варианта № 1 ожидаемое для каждой делянки число личинок равно просто средней из четырех поделяночных наблюдений, т. е. 6 личинкам. Поэтому отклонения от ожидаемой численности в этом случае являются не чем иным, как отклонениями от средней наших количественных наблюдений: 11—6, 4—6 и т. д. Для вычисления хи-квадрат каждое из этих отклонений должно быть возведено в квадрат и после этого разделено на ожидаемую численность. Но так как последняя одинакова для всех четырех делянок, то деление может быть произведено после того, как проведено суммирование квадратов. Так, для варианта № 1 имеем:

$$\chi^2 = \frac{5^2 + (-2)^2 + (-2)^2 + (-1)^2}{6} = \frac{34}{6} = 5,67.$$

В таблице 81 применен обычный метод обработки на счетной машине. Легко видеть, что данная формула является простым приспособлением общей формулы хи-квадрат к этому частному случаю.

ТАБЛИЦА 81

Число личинок пяденицы, обнаруженных на 50 растениях капусты, имеющих на делянке

Каждые четыре делянки обработаны одинаково. Пять вариантов обработки

Варианты	Число личинок на каждой из 4 делянок X	Сумма по варианту ΣX	Средняя \bar{x}	Хи-квадрат $\Sigma x^2/\bar{x}$
1	11, 4, 4, 5	24	6,00	5,67
2	6, 4, 3, 6	19	4,75	1,42
3	8, 6, 7, 11	29	7,25	2,69
4	14, 27, 8, 18	67	16,75	11,39
5	7, 4, 9, 14	34	8,50	6,24
Итого		173		28,41

Вычисления для варианта I

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 11^2 + 4^2 + 4^2 + 5^2 - 24^2/4 = 34$$

$$\chi^2 = \Sigma x^2/\bar{x} = 34/6,00 = 5,67; \text{ число степеней свободы} = 3; P = 14\%$$

Вычисления по всем пяти вариантам

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 24^2 + 19^2 + 29^2 + 67^2 + 34^2 - 173^2/5 = 1437,2$$

$$\bar{x} = 173/5 = 34,6$$

$$\chi^2 = \Sigma x^2/\bar{x} = 1437,2/34,6 = 41,5; \text{ число степеней свободы} = 4; P \text{ меньше } 0,0002$$

Значение хи-квадрат, вычисленное в таблице для варианта № 1, может появиться при однородном материале чаще чем 14 раз на 100 случаев. Однородность плотности расселения личинок находится вне всякого сомнения. Сумма значений хи-квадрат для пяти вариантов равна 28,41, число степеней свободы = 15 и $P = 2\%$ (см. табл. 81).

Это указывает на существенность отклонений от однородности расселения личинок на делянках одного и того же варианта. При просмотре первоначального материала видно, что уменьшение однородности относится главным образом к делянкам с более высокой плотностью расселения. Однако это отклонение от однородности представляется небольшим по сравнению с варьированием показателя по вариантам опыта. Последнее определено путем вычислений в конце таблицы, где обработке подвергнуты итоги по вариантам. Хи-квадрат 41,5, число степеней свободы = 4 приводит к заключению о несомненном влиянии вариантов опыта, несмотря на то, что сюда включается некоторая неопределенная часть варьирования, связанная с неоднородностью делянок одного и того же варианта.

Пример 25. Дейвис [8] подсчитал количество галлов пузырчатой головни на растениях кукурузы восьми делянок. Растения четырех делянок были заражены возбудителем этой болезни. Численности галлов, подсчитанных на этих делянках, были: 56, 60, 41 и 75. Хи-квадрат = 10,1, число степеней свободы = 3, $P = 2\%$. На четырех делянках, растения которых не были заражены, численности галлов были 20, 12, 26, 22. Хи-квадрат = 5,2, число степеней свободы = 3, $P = 17\%$. Для обеих групп: $\chi^2 = 15,3$, число степеней свободы 6, $P = 2\%$. Эти данные указывают на существенную неоднородность одинаково обработанных делянок. Тем не менее, так как эта неоднородность не очень велика, то можно произвести оценку различия между итогами обеих групп делянок. Для сумм 232 и 80 имеем $\chi^2 = 7,4$, число степеней свободы = 1, P меньше 1%. Здесь нет почти никакого сомнения в отношении существенности различия между этими двумя суммами.

Пример 26. Если в примере 22 оставить без внимания количество частиц индиго в суспензии № 1, то для одних только бактерий имеем $\chi^2 = 3,3$; число степеней свободы = 3 и $P = 36\%$. Таким образом, ни для бактерий, ни для отношения численностей бактерий к итогам нет существенного отклонения от случайного варьирования при выборках из однородной совокупности. Отношения, как указывает на это большее значение хи-квадрат, более вариабильны, и, возможно, поэтому данный критерий более строг: первый критерий основывался на варьировании как численности бактерий, так и численности частиц индиго.

Пример 27. Дейвис [8] подсчитал численности початков на двух делянках кукурузы, растения которых были заражены пузырчатой головней, а также на двух контрольных делянках. Численности соответственно были 436 и 432. Очевидное отсутствие влияния зараженности на численность початков подтверждается некорректированным значением $\chi^2 = 0,02$, число степеней свободы = 1. Однако число пораженных головней початков было 69 и 39. На то, что разность соответствующих процентов существенна, указывает скорректированное значение $\chi^2 = 8,59$; число степеней свободы = 1.

ЛИТЕРАТУРА

1. Bartlett M. S., Supplement to the Journal of the Royal Statistical Society, 2 248, 1935.
2. Bateson W., Punnett R. C., Journal of Genetics, 1, 297, 1911.
3. Boyd Mark F., Kitchen S. F., Muench Hugo, The American Journal of Tropical Medicine, 16, 589, 1936.
4. Burnett L. C., Thesis submitted for the degree of Master of Science, Iowa State College, 1906.
5. Caffrey D. J., Smith C. E., Bureau of Entomology and Plant Quarantine, U.S.D.A. (Baton Rouge), 1934.
6. Cochran W. G., Iowa State College Journal of Science, 16, 421, 1942.
7. Cochran W. G., Annals of Mathematical Statistics, 23, 315, 1952. Biometrics, 10, 447, 1954.
8. Davis Glen N., Iowa Agricultural Experiment Station Research Bulletin 199, 1936.
9. Decker George C., Andre Floyd, Iowa State College Journal of Science, 10, 403, 1936.
10. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd Edinburgh, 1934, и след.
11. Fisher R. A., Mather K., Annals of Eugenics, 7, 265, 1936.
12. Hansberry T. Roy, Richardson C. H., Iowa State College Journal of Science, 10, 27, 1935.
13. Leslie P. H., Biometrics, 7, 283, 1951.

14. Lindstrom E. W., Cornell University Agricultural Experiment Station Memoir 13, 1918.
15. Lindstrom E. W., Iowa Agricultural Experiment Station Bulletin 142, 1931.
16. «Mathetes». Annals of Applied Biology, **11**, 220, 1924.
17. Mood A. M., Introduction to the Theory of Statistics. McGraw-Hill Book Company, Inc., New York, 1950.
18. Snedecor George W., Irwin M. R., Iowa State College Journal of Science, **8**, 75, 1933.
19. Strand Norman V., Jessen Raymond J., Iowa Agricultural Experiment Station Research Bulletin 315, 1943.
20. Thornton H. P., Gray P. H. H., Proceedings of the Royal Society of London, Series B, **115**, 522, 1934.
21. Wallis W. Allen, Econometrica, **10**, 229, 1942.
22. Yates F., Supplement to the Journal of the Royal Statistical Society, **1**, 217, 1934.

ДВЕ ИЛИ БОЛЬШЕЕ ЧИСЛО ВЫБОРОК КОЛИЧЕСТВЕННОГО ПРИЗНАКА. ДИСПЕРСИОННЫЙ АНАЛИЗ

1. Распространение теории двух выборок на большее число их. Статистические и экспериментальные методы, основанные на двух выборках, были рассмотрены в главе 4, но задачи исследователя редко ограничиваются изучением только двух вариантов опыта. Увеличение числа вариантов требует соответствующего расширения методов обработки. Такого рода обобщение методов, относящихся к качественным признакам, было сделано в предыдущей главе. Теперь мы переходим к такому же обобщению методов в отношении данных о количественном признаке и распространим методы главы 4 на случай одновременного изучения трех или более вариантов.

В начале 20-х годов настоящего столетия был достигнут большой прогресс в отношении разработки статистических методов для множественной группировки [5]. В начале этого десятилетия Фишер [6] разрешил ряд проблем, относящихся к распределениям, которые легли в основу его *дисперсионного анализа*. Предложенный им способ привел к чрезвычайному развитию теории планирования опытов и соответствующей системы статистического анализа. Некоторые из этих форм эксперимента и его анализа будут предметом обсуждения в настоящей главе.

2. Несколько выборок из общей совокупности. Дисперсионный анализ. В начальных параграфах главы 4 мы рассматривали две выборки равного размера; теперь коснемся вопроса о трех или большем числе таких выборок. Например, такими выборками могут быть несколько групп животных, каждая из которых получает различный рацион, или несколько групп школьников 6-го класса, по отношению к которым применяются различные методы обучения. Для того чтобы ознакомиться с статистическим анализом, соответствующим такого рода экспериментальным данным, лучше всего обратиться к выборкам главы 3. Там мы отобрали случайным порядком некоторое число выборок из нормального распределения привеса свиней, имеющего среднюю 30 и дисперсию 100. Если мы возьмем вместе несколько таких выборок, положим 4, то получим условный эксперимент с 40 свиньями, состоящими из 4 группы по 10 животных в каждой. В фактическом опыте такие группы должны относиться к различным вариантам, но наши данные только имитируют опыт и образуют *условный эксперимент*, или *однородный опыт*, в котором все варианты одинаковы.

Результаты одного опыта только что описанного типа даны в таблице 82. Для облегчения вычислений численность групп снижена до 5 свиней в каждой и их привесы несколько изменены, чтобы избежать дробных величин. Здесь имеется a выборок или групп (варианты), каждая из которых состоит из n наблюдений (повторения) над весом животных. Таким образом, этот эксперимент включает в себе an объектов.

Данные этой таблицы приводят к трем оценкам дисперсии $\sigma^2=100$, характеризующей совокупность, из которой произведены выборки этого эксперимента. Первая берется из правого столбца таблицы, где объединены все 20 привесов. Этой оценкой является средний квадрат $1918/19=100,9$.

Вторая оценка дисперсии совокупности получается из сумм квадратов внутри четырех групп. В соответствии с методом главы 4 эти суммы квадратов суммируются, или, иначе говоря, обобщаются:

$$472 + 396 + 616 + 464 = 1648.$$

Каким будет в данном случае число степеней свободы? Из пяти наблюдений каждой группы следует исключить по одному, так как средняя группа использована в этих расчетах в качестве начала отсчета отклонений. В целом, поскольку для каждой из 4 группы имеется 4 степени свободы, сумме квадратов 1648 будет соответствовать 16 степеней свободы. В результате этого средний квадрат $1648/16=103$ будет являться второй выборочной оценкой σ^2 .

ТАБЛИЦА 82

*Привесы (в фунтах) 4 групп, по 5 свиней в каждой.
Выборки взяты из таблицы 15*

	a=4 группам, n=5 свиньям в группе				По опыту в целом
	1	2	3	4	
	40	29	41	17	
	24	27	31	24	
	46	20	47	28	
	20	39	37	33	
	35	45	39	24	
ΣX	165	160	135	120	580
\bar{x}	33	32	27	24	29
ΣX^2	5917	5516	4261	3044	18738
$(\Sigma X)^2/n$	5445	5420	3645	2880	16820
Σx^2	472	396	616	464	1918

Групповые средние приводят к третьей оценке дисперсии совокупности. Так как эти 4 средние составляют случайную выборку из одной и той же нормальной совокупности, их средний квадрат будет несмещенной оценкой $\sigma^2_{\bar{x}}=100/5=20$. Средний квадрат этих средних равен

$$\frac{(33-29)^2 + (32-29)^2 + (27-29)^2 + (24-29)^2}{3} = 18$$

вместо ожидаемых 20. Так как 18 является оценкой $\sigma^2/5$, то $5 \times 18 = 90$ будет третьей оценкой σ^2 ; она основана на групповых средних при 3 степенях свободы.

Последующий анализ представлен в таблице 83. Одна дополнительная деталь: средний квадрат групповых средних в таблице умножен на соответствующее число степеней свободы, что дает сумму квадратов 270, применение которой будет ясным из последующего.

ТАБЛИЦА 83

Дисперсионный анализ для привеса свиней по данным таблицы 82

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Объекты отдельных групп	16	1648	103
Групповые средние	3	270	90
Итого	19	1918	100,9

Эта таблица выявляет два замечательных факта: не только общая сумма квадратов, но и соответствующее число степеней свободы подразделяются на две части, которые согласуются со структурой данного эксперимента. Так как эти части, так же как и общая сумма квадратов, вычислены независимо одна от другой, то все это подчеркивает их свойство слагаемости. Мы напоминаем здесь подобные теоремы слагаемости параграфа 11 главы 6. Такое подразделение степеней свободы и соответствующих сумм квадратов называется *дисперсионным анализом*. При использовании случайного отбора в чистом виде каждый из трех средних квадратов является оценкой дисперсии совокупности σ^2 .

Пример 1. Проведите дисперсионный анализ по таким данным, подобранным в целях упрощения вычислений.

Помер выборки			
1	2	3	4
11	13	21	19
4	9	18	4
6	14	15	19

Дисперсионный анализ

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Объекты	8	172	21,5
Выборки	3	186	62,0
Итого	11	358	32,5

Пример 2. Вычислите сумму квадратов для выборок 186 путем сложения четырех поправок на среднюю каждой выборки 147, 432 и т. д. и вычитанием из полученной суммы поправки на среднюю всего эксперимента 1728.

Пример 3. Вычислите сумму квадратов для выборок следующим образом:

$$\frac{21^2 + 36^2 + 54^2 + 33^2}{3} - \frac{144^2}{12}$$

Числа, которые стоят в числителе и возводятся в квадрат, являются итогами выборок и всего опыта, а то время как в знаменателях стоят n и an .

Какой из двух методов вычисления этой суммы квадратов представляется более простым?

3. Обычный способ вычислений. Одной из причин широкого распространения дисперсионного анализа является изящество его вычислительных операций. После того как выяснен смысл таблицы 82, ее вычисления могут быть значительно сокращены. При наличии соответствующего контроля за вычислениями многие предпочитают опустить при обработке данных таблицы 82 последние три строки нижней части этой таблицы, вычисляя непосредственно только две суммы квадратов для итогов и групповых средних и находя после этого сумму квадратов для объектов простым вычитанием из первой суммы квадратов второй. Это и будет обычно применяемый способ вычислений (те, кто предпочитают иметь дело с алгебраической символикой, найдут ее в параграфе 10 этой главы):

1. Сумма всех наблюдений: $\Sigma X = 40 + 24 + \dots + 21 = 580$.
2. Поправка на среднюю: $C = (\Sigma X)^2/an = 580^2/20 = 16\ 820$.

$$3. \text{ Общая сумма квадратов: } \Sigma X^2 - C = 40^2 + 24^2 + \dots + 21^2 - C = \\ = 18\,738 - 16\,820 = 1\,918.$$

4. Сумма квадратов для групповых средних:

$$\frac{\Sigma (\Sigma X)^2}{n} - C = \frac{165^2 + 160^2 + \dots + 120^2}{5} - C = \frac{85\,450}{5} - C = 17\,090 - 16\,820 = 270.$$

Результаты двух последних пунктов с соответствующими им степенями свободы вносятся в таблицу 84, которая окончательно заполняется путем вычисления показателей для объектов.

ТАБЛИЦА 84

Дисперсионный анализ привеса свиней. Обычная форма

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	19	1 918	
Групповые средние	3	270	90
Объекты	16	1 648	103

Пояснения: 1) число степеней свободы и сумма квадратов последней строки определяются вычитанием на основе теоремы слагаемости, на которой базируется этот анализ; 2) здесь нет возможности сделать проверку вычислений, как это было в таблице 82, поэтому для начинающих полезно произвести несколько раз непосредственное вычисление суммы квадратов для объектов внутри группы; 3) в отношении техники расчета п. 4 примите во внимание примеры 3 и 2 этой главы.

Если все это вначале покажется несколько сложным, то следует надеяться, что небольшая практика вскоре приведет к почти автоматическому проведению всех этих расчетов.

При отсутствии счетной машины следует поступать так: 1) по каждой группе вычисляются отклонения $x = X - \bar{x}$; сумма их квадратов даст Σx^2 ; сумма всех таких частных сумм Σx^2 является суммой квадратов для объектов 1648; 2) определяются отклонения каждой групповой средней от общей средней всего опыта, квадраты этих отклонений суммируются и умножаются на 5; в результате получается сумма квадратов для групповых средних; 3) если в этом возникает необходимость, то общая сумма квадратов может быть найдена путем суммирования первых двух сумм квадратов, но обычно нет никакой надобности вычислять эту общую сумму квадратов; 4) если средние не «слишком гладкие», то для вычисления сумм квадратов следует использовать метод произвольного начала параграфа 11 главы 2.

Пример 4. Часть более обширного опыта [25] включала 3 группы по три свиньи, получавших обычный рацион с добавлением антибиотиков и 3 доз витамина В₁₂. Средние привесы за день этих свиней (при 75 фунтах живого веса) были:

Доза В ₁₂ (в мг на 1 фунт корма)	Средний привес за 1 день		
5	1,52	1,56	1,54
10	1,63	1,57	1,54
20	1,44	1,52	1,63

Дисперсионный анализ будет таким:

Источник варьирования	Число степеней свободы	Число квадратов	Средний квадрат
Общее	8	0,0274	0,0034
Групповые средние	2	0,0042	0,0021
Объекты	6	0,0232	0,0039

У к а з а н и е. Если вы от каждого привеса отнимете по 1,00, то это упростит расчеты и даст экономию времени (параграф 4 главы 5).

Пример 5. Процент чистой шерсти в 7 тюках был определен путем взятия 3 случайных образцов из каждого тюка. Проценты чистой шерсти по каждому образцу были следующие:

Номер тюка						
1	2	3	4	5	6	7
41,8	33,0	38,5	43,7	34,2	32,6	36,2
38,9	37,5	35,9	38,9	38,6	38,4	33,4
36,1	33,1	33,9	36,3	40,2	34,8	37,9

Вычислите средние квадраты для тюков (11,11) и образцов (8,22).

Пример 6. Средний квадрат M , который является оценкой некоторого σ^2 — параметра нормальной совокупности, имеет точно установленное выборочное распределение, о котором говорилось в параграфе 5 главы 3. Вычислите s_M для объектов в таблице 84. *Ответ:* 34. (См. пример 5 главы 3; $n-1=16$; $n+1=18$).

Пример 7. Постройте 95%-ный доверительный интервал для σ^2 , основываясь на оценке M для объектов в таблице 84 (см. параграф 14 главы 2). *Ответ:* $57 \leq \sigma_M^2 \leq 239$. Не слишком ли широк этот интервал?

4. Случайные выборки из двух или большего числа совокупностей. Пока обсуждался вопрос относительно дисперсионного анализа случайных выборок, взятых из одной совокупности. Но в большинстве приложений дисперсионного анализа изучаемые варианты опыта, к которым относятся группы, так или иначе влияют на средние. В результате этого группы становятся выборками из различных совокупностей. Считается, что эти совокупности имеют различные средние, определяемые вариантами опыта, и в то же время общую дисперсию, не зависящую от вариантов. Теоретически они являются нормальными совокупностями с одной и той же дисперсией, но каждая из них имеет свое частное значение средней μ . При дисперсионном анализе данных такого рода экспериментов средний квадрат для объектов оценивает, как и ранее, σ^2 , но средний квадрат групповых средних является преувеличенным в связи с различиями между μ .

Таблица 85 представляет данные подобного рода эксперимента. В процессе выпечки тесто может поглощать различное количество жира. Лоу [19] поставил перед собой задачу изучить вопрос: является ли количество поглощенного тестом жира его характерной особенностью. Шесть выпечек, составивших 24 булки, были проведены с каждым из четырех видов жира. Числа, приведенные в таблице, дают количества (в г) поглощенного жира, кодированные путем вычитания по 100 г из каждого первоначального числа. Средние же указаны в их первоначальном виде (параграф 4 главы 5). Таблица 86 дает дисперсионный анализ.

При изложенных выше предположениях дисперсия выборок из нормальных совокупностей с общей σ^2 , равная 100,9, является оценкой этой σ^2 . Но средний квадрат для видов жира 545,5 явным образом содержит в себе

Количество жира (в г), поглощенного при 6 выпечках теста, для каждого вида жира (от данных каждого образца отнято по 100 г)

Вид жира	Поглощенных граммов—100						Сумма	Средний
	1	2	3	4	5	6		
1	64	72	68	77	56	95	432	172
2	78	91	97	82	85	77	510	185
3	75	93	78	71	63	76	456	176
4	55	66	49	64	70	68	372	162

Вычисления

- $\Sigma X = 64 + 72 + \dots + 70 + 68 = 1770$.
- $C = 1770^2 / 24 = 130537,5$.
- Общее: $\Sigma x^2 = 64^2 + 72^2 + \dots + 70^2 + 68^2 - C = 134192 - C = 3654,5$.
- Для средних: $\frac{432^2 + \dots + 372^2}{6} - C = 132174 - C = 1636,5$.

ТАБЛИЦА 86

Дисперсионный анализ данных о выпечке хлеба

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	23	3654,5	
Виды жира	3	1636,5	545,5
Образцы	20	2018,0	100,9

некоторый добавочный компонент, обусловленный различиями между этими видами жира. Оценка этого компонента и критерий существенности его будет предметом последующего обсуждения.

Для проверки константности дисперсий по группам можно рассмотреть ряд размахов варьирования этих выборок. Для вида жира 1 размах равен $95 - 56 = 39$, для других видов жира он будет 20, 30 и 21. Если вы вспомните о варьировании выборок, взятых из данных таблицы 15 и имеющих одно и то же значение σ^2 , то у вас не будет никакого сомнения в том, что приведенные выше размахи однородны. Более точное доказательство этого будет дано в параграфе 20 этой главы.

Пример 8. В таблице 81 были даны количества личинок мушкетера, найденных на 50 растениях капусты каждой делянки: в опыте изучались 5 вариантов, каждый на четырех полностью рандомизированных делянках. Эти численности были таковы:

Вариант	1	11	4	4	5
»	2	6	4	3	6
»	3	8	6	4	11
»	4	14	27	8	18
»	5	7	4	9	14

В отношении данных, подобных этим, полученным в результате подсчета, а не измерения, нельзя ожидать нормального распределения, и влияние вариантов не может считаться складывающимся со случайным варьированием, как это предполагается в настоящей главе (см. параграф 11 главы 11). Но все же в качестве иллюстрации мы обработаем эти данные по методу дисперсионного анализа.

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	19	670,55	
Варианты	4	359,30	89,82
Долянки одного и того же варианта	15	311,25	20,75

Пример 9. Скопструируйте опыт, подобный опыту с выпечкой хлеба, взяв для этого данные таблицы 82, изменив средние и оставляя дисперсию такой, какой она была в этой таблице. Например, увеличьте среднюю группы 1 так, чтобы ее значение в совокупности возросло с 30 до 45 фунтов. Пять привесов этой группы в таком случае будут 40+15+55 и т. д., т. е. возрастут на 15. Оставляя группу 2 без изменений, уменьшите среднюю группы 3 на 6 фунтов, а группы 4 — на 9 фунтов. В результате получаются следующие экспериментальные данные:

55	29	5	8
39	27	25	12
61	20	11	19
35	29	31	24
50	45	33	12

Дисперсионный анализ для этого нового опыта будет:

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	19	4798	
Группы	3	3150	1050
Объекты	16	1648	103

Так как $\sum x^2$ (и размах варьирования) внутри групп остаются неизменными, то и средний квадрат объектов сохраняет свое значение 103. Но средний квадрат групп в связи с изменением групповых средних возрос с 90 до 1050. Новый средний квадрат уже не является оценкой $\sigma^2=100$; в него входит добавочный компонент, обусловленный различием средних μ . Этот результат будет более подробно рассмотрен далее в параграфе 11.

Пример 10. Сумма квадратов для групп в предыдущем примере неожиданно возросла с 270 до 3150; это объясняется тем, что большие первоначальные средние были увеличены, в то время как малые уменьшены. Скопструируйте на основе таблицы 82 другие данные путем вычитания из средней первой группы 9, из средней второй группы 6 и прибавления к средней четвертой группы 15, оставив третью группу без изменения. Проведите дисперсионный анализ. *Ответ:* сумма квадратов для групп=690.

Пример 11. Не представляет никакого труда показать, каким образом сумма квадратов примера 9 изменилась от 270 до 3150:

Первоначальное	Изменение	Новое \bar{x}	Отклонения от 29	(Отклонения) ²
33	15	33+15	4+15	16+120+225
32	0	32	3	9
27	-6	27-6	-2-6	4+24+36
24	-9	27-9	-5-9	25+90+81
29	0	29	0	54+234+342

Новая сумма квадратов отклонений средних равна $54+234+342=630$. Так как $n=5$, то сумма квадратов в пересчете на единичное наблюдение будет $630 \times 5=3150$.

5. Критерий равенства μ . Отношение дисперсий F . Данные о выпечке хлеба приводят нас к знакомому уже вопросу: обусловливается ли большое различие между средними квадратами для видов жира и образцами хлеба обычным варьированием случайных выборок или же мы должны сделать заключение, что средние для отдельных видов жира различаются друг от друга по другим причинам, а не в связи с выборочным варьированием? Лоу считал возможным, что свойства отдельных видов жира таковы, что они поглощаются тестом в различных количествах. Если это так, то образцы хлеба образуют выборки не из одной общей совокупности, а из различных совокупностей.

Соответствующая такой постановке вопроса нулевая гипотеза такова: $H_0: \mu_1 = \mu_2 = \dots = \mu_n$; она определяет собой совокупность, подобную той, которую мы имели для привеса свиней в таблице 82. Для проверки этой гипотезы H_0 мы должны познакомиться с выборочным распределением, относящимся к случаю нескольких выборок, взятых из общей совокупности. Для этого вернемся назад и произведем выборки из таблицы 15. Теперь первоначальные выборки, состоящие из $n=10$ членов, взятые вместе, будут образовывать систему опытов при $a=10$; образец одного из таких опытов приведен в таблице 87. Дисперсионный анализ для каждого такого эксперимента проводится по указанному там же образцу. После этого вычисляется новый критерий, который является отношением:

$$\frac{\text{средний квадрат выборочных средних}}{\text{средний квадрат объектов}}$$

Это отношение имеет распределение, закон которого открыл Р. А. Фишер [6]. Я обозначил этот критерий буквой F в честь его имени [26]. Фишер дал таблицу этого распределения в преобразованной форме $z = \log_e \sqrt{F}$. Фишер и Нейте называют F *отношением дисперсий*; Махаланобис [20], который его впервые вычислил, обозначает буквой x .

В таблице 87 значение $F=1,51$; это значение получено в условиях, когда мы знаем, что нарушение равенства между двумя средними квадратами объясняется только выборочным варьированием; каждый из этих средних квадратов в данном случае оценивает $\sigma^2=100$.

ТАБЛИЦА 87

Десять выборок, каждая из 10 наблюдений, взятые случайным порядком из нормальной совокупности таблицы 15. Здесь имитируется условный опыт с 10 группами свиней

1	2	3	4	5	6	7	8	9	10
35	34	17	39	18	7	33	42	7	39
42	38	29	34	7	22	29	35	48	34
42	26	30	22	31	40	25	31	43	33
30	17	36	27	41	29	27	33	53	33
15	42	41	42	21	31	21	53	7	33
31	28	30	33	17	41	46	21	33	39
29	35	3	24	21	30	19	41	17	36
29	33	23	36	40	12	43	29	57	32
17	16	38	29	14	44	22	34	42	32
21	40	30	25	14	30	21	49	42	30

Источник варьирования	Число степеней свободы	Средний квадрат
Выборочные средние	9	172,3
Объекты	90	114,0

$$F = 172,3/114,0 = 1,51$$

Таблица 88 дает выборочное распределение 100 значений F , каждое из которых получено на основе случайно отобранного эксперимента, подобного представленному в таблице 87. Прежде всего обращает на себя внимание скошенность распределения, выражающаяся в большой концентрации малых значений и в длинном «хвосте» больших значений F . Следующий приметный факт состоит в том, что 65 значений F меньше 1. Если вы вспомните, что оба числа данного отношения являются оценкой σ^2 , то вы, может быть, удивитесь тому, что 1 не является медианой распределения. Однако среднее значение F , вычисленное по методу параграфа 3 главы 8, равно 0,96, т. е. близко к ожидаемой единице. Наконец, можно видеть, что 5% значений F превосходят 2,25 и 1% выше 2,75; эти два уровня приблизительно соответствуют уровням теоретического распределения, которое будет рассмотрено в дальнейшем.

ТАБЛИЦА 88

Распределение F в 100 выборках из таблицы 15. Число степеней свободы 9 и 90

Интервалы классов	Численности	Интервалы классов	Численности
0—0,24	7	1,50—1,74	5
0,25—0,49	16	1,75—1,99	2
0,50—0,74	16	2,00—2,24	4
0,75—0,99	26	2,25—2,49	2
1,00—1,24	11	2,50—2,74	2
1,25—1,49	8	2,75—2,99	1

Таблица 89 содержит в себе 5%- и 1%-ные уровни F для некоторых комбинаций степеней свободы. Сверху таблицы даны f_1 —числа степеней свободы, соответствующие количеству выборок: $f_1 = a - 1$. Слева даны f_2 — числа степеней свободы для объектов: $a (n - 1)$.

Для нахождения 5%- и 1%-ных уровней по данным таблицы 87 в таблице 89 следует взять столбец, озаглавленный $f_1 = 9$, и пойти по нему вниз до строк $f_2 = 80$ и 100. Искомые уровни будут 1,98 и 2,62; они вычислены как средние из табличных значений для $f_2 = 80$ и 100. Сравните с этими уровнями уровни 2,25 и 2,75, найденные экспериментально в таблице 88 для выборки, состоящей из 100 опытов; это неплохое совпадение. Для более полной проверки выборочного распределения я вернулся к первоначальным расчетам F и нашел 8% значений, превосходящих 5%-ный уровень, и 2%, превосходящих 1%-ный уровень. Это дает некоторое представление о варьировании F , наблюдаемое при выборочном методе.

В опыте с выпечкой хлеба гипотеза о том, что образцы хлеба составляют случайные выборки из общей совокупности, также может быть проверена при помощи таблицы 89. Из дисперсионного анализа таблицы 86 следует:

$$F = 545,5/100,9 = 5,41.$$

Для $f_1 = 3$ и $f_2 = 20$ значение 1%-ного уровня по новой таблице равно 4,94. Таким образом, в распределении, определяемом данной гипотезой, имеется меньше одного шанса из 100 получить выборку, дающую значение F , большее, чем наблюдаемое. Очевидно, что данные выборки принадлежат совокупностям с различными μ . Вывод из этого: различные виды ячменя имеют различную способность впитываться тестом.

Пример 12. Каждый из четырех видов кормов скармливался группе из 5 молодых мышей [24]. Привесы были следующие:

Вид корма 1	55	49	42	21	52
» 2	61	112	30	89	63
» 3	42	97	81	95	92
» 4	169	137	169	85	154

f_2	1	2	3	4	5	6	7	8	9	10	11
1	161	200	216	225	230	234	237	239	241	242	243
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40
3	10,43	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38	2,34
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28
22	4,30	3,44	3,05	2,82	2,66	2,55	2,47	2,40	2,35	2,30	2,26
23	4,28	3,42	3,03	2,80	2,64	2,53	2,45	2,38	2,32	2,28	2,24
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22
25	4,24	3,38	2,99	2,76	2,60	2,49	2,41	2,34	2,28	2,24	2,20
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,30	2,25	2,20	2,16
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19	2,15
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18	2,14
30	4,17	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,12
32	4,15	3,30	2,90	2,67	2,51	2,40	2,32	2,25	2,19	2,14	2,10
34	4,13	3,28	2,88	2,65	2,49	2,38	2,30	2,23	2,17	2,12	2,08
36	4,11	3,26	2,86	2,63	2,48	2,36	2,28	2,21	2,15	2,10	2,06
38	4,10	3,25	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09	2,05
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,07	2,04
42	4,07	3,22	2,83	2,59	2,44	2,32	2,24	2,17	2,11	2,06	2,02
44	4,06	3,21	2,82	2,58	2,43	2,31	2,23	2,16	2,10	2,05	2,01
46	4,05	3,20	2,81	2,57	2,42	2,30	2,22	2,14	2,09	2,04	2,00
48	4,04	3,19	2,80	2,56	2,41	2,30	2,21	2,14	2,08	2,03	1,99
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98
55	4,02	3,17	2,78	2,54	2,38	2,27	2,18	2,11	2,05	2,00	1,97
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99	1,95
65	3,99	3,14	2,75	2,51	2,36	2,24	2,15	2,08	2,02	1,98	1,94
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93
80	3,96	3,11	2,72	2,48	2,33	2,21	2,12	2,05	1,99	1,95	1,91
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88
125	3,92	3,07	2,68	2,44	2,29	2,17	2,08	2,01	1,95	1,90	1,86
150	3,91	3,06	2,67	2,43	2,27	2,16	2,07	2,00	1,94	1,89	1,85
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,83
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81
1000	3,85	3,00	2,61	2,38	2,22	2,10	2,02	1,95	1,89	1,84	1,80
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83	1,79

Функция F (e в степени $2z$) вычислена частично по таблице VI Фишера [7].
Остальные данные получены путем интерполяции, в большинстве случаев графической.

в определении F

большого среднего квадрата)

12	14	16	20	24	30	40	50	75	100	200	500	∞
244	245	246	248	249	250	251	252	253	253	254	254	254
19,41	19,42	19,43	19,44	19,45	19,46	19,47	19,47	19,48	19,49	19,49	19,50	19,50
8,74	8,71	8,69	8,66	8,64	8,62	8,60	8,58	8,57	8,56	8,54	8,54	8,53
5,91	5,87	5,84	5,80	5,77	5,74	5,71	5,70	5,68	5,66	5,65	5,64	5,63
4,68	4,64	4,60	4,56	4,53	4,50	4,46	4,44	4,42	4,40	4,38	4,37	4,36
4,00	3,96	3,92	3,87	3,84	3,81	3,77	3,75	3,72	3,71	3,69	3,68	3,67
3,57	3,52	3,49	3,44	3,41	3,38	3,34	3,32	3,29	3,28	3,25	3,24	3,23
3,28	3,23	3,20	3,15	3,12	3,08	3,05	3,03	3,00	2,98	2,96	2,94	2,93
3,07	3,02	2,98	2,93	2,90	2,86	2,82	2,80	2,77	2,76	2,73	2,72	2,71
2,91	2,86	2,82	2,77	2,74	2,70	2,67	2,64	2,61	2,59	2,56	2,55	2,54
2,79	2,74	2,70	2,65	2,61	2,57	2,53	2,50	2,47	2,45	2,42	2,41	2,40
2,69	2,64	2,60	2,54	2,50	2,46	2,42	2,40	2,36	2,35	2,32	2,31	2,30
2,60	2,55	2,51	2,46	2,42	2,38	2,34	2,32	2,28	2,26	2,24	2,22	2,21
2,53	2,48	2,44	2,39	2,35	2,31	2,27	2,24	2,21	2,19	2,16	2,14	2,13
2,48	2,43	2,39	2,33	2,29	2,25	2,21	2,18	2,15	2,12	2,10	2,08	2,07
2,42	2,37	2,33	2,28	2,24	2,20	2,16	2,13	2,09	2,07	2,04	2,02	2,01
2,38	2,33	2,29	2,23	2,19	2,15	2,11	2,08	2,04	2,02	1,99	1,97	1,96
2,34	2,29	2,25	2,19	2,15	2,11	2,07	2,04	2,00	1,98	1,95	1,93	1,92
2,31	2,26	2,21	2,15	2,11	2,07	2,02	2,00	1,96	1,94	1,91	1,90	1,88
2,28	2,23	2,18	2,12	2,08	2,04	1,99	1,96	1,92	1,90	1,87	1,85	1,84
2,25	2,20	2,15	2,09	2,05	2,00	1,96	1,93	1,89	1,87	1,84	1,82	1,81
2,23	2,18	2,13	2,07	2,03	1,98	1,93	1,91	1,87	1,84	1,81	1,80	1,78
2,20	2,14	2,10	2,04	2,00	1,96	1,91	1,88	1,84	1,82	1,79	1,77	1,76
2,18	2,13	2,09	2,02	1,98	1,94	1,89	1,86	1,82	1,80	1,76	1,74	1,73
2,16	2,11	2,06	2,00	1,96	1,92	1,87	1,84	1,80	1,77	1,74	1,72	1,71
2,15	2,10	2,05	1,99	1,95	1,90	1,85	1,82	1,78	1,76	1,72	1,70	1,69
2,13	2,08	2,03	1,97	1,93	1,88	1,84	1,80	1,76	1,74	1,71	1,68	1,67
2,12	2,06	2,02	1,96	1,91	1,87	1,81	1,78	1,75	1,72	1,69	1,67	1,65
2,10	2,05	2,00	1,94	1,90	1,85	1,80	1,77	1,73	1,71	1,68	1,65	1,64
2,09	2,04	1,99	1,93	1,89	1,84	1,79	1,76	1,72	1,69	1,66	1,64	1,62
2,07	2,02	1,97	1,91	1,86	1,82	1,76	1,74	1,69	1,67	1,64	1,61	1,59
2,05	2,00	1,95	1,89	1,84	1,80	1,74	1,71	1,67	1,64	1,61	1,59	1,57
2,03	1,98	1,93	1,87	1,82	1,78	1,72	1,69	1,65	1,62	1,59	1,56	1,55
2,02	1,96	1,92	1,85	1,80	1,76	1,71	1,67	1,63	1,60	1,57	1,54	1,53
2,00	1,95	1,90	1,84	1,79	1,74	1,69	1,66	1,61	1,59	1,55	1,53	1,51
1,99	1,94	1,89	1,82	1,78	1,73	1,68	1,64	1,60	1,57	1,54	1,51	1,49
1,98	1,92	1,88	1,81	1,76	1,72	1,66	1,63	1,58	1,56	1,52	1,50	1,48
1,97	1,91	1,87	1,80	1,75	1,71	1,65	1,62	1,57	1,54	1,51	1,48	1,46
1,96	1,90	1,86	1,79	1,74	1,70	1,64	1,61	1,56	1,53	1,50	1,47	1,45
1,95	1,90	1,85	1,78	1,74	1,69	1,63	1,60	1,55	1,52	1,48	1,46	1,44
1,93	1,88	1,83	1,76	1,72	1,67	1,61	1,58	1,52	1,50	1,46	1,43	1,41
1,92	1,86	1,81	1,75	1,70	1,65	1,59	1,56	1,50	1,48	1,44	1,41	1,39
1,91	1,85	1,80	1,73	1,68	1,63	1,57	1,54	1,49	1,46	1,42	1,39	1,37
1,89	1,84	1,79	1,72	1,67	1,62	1,56	1,53	1,47	1,45	1,40	1,37	1,35
1,88	1,82	1,77	1,70	1,65	1,60	1,54	1,51	1,45	1,42	1,38	1,35	1,32
1,87	1,79	1,75	1,68	1,63	1,57	1,51	1,48	1,42	1,39	1,34	1,30	1,28
1,85	1,77	1,72	1,65	1,60	1,55	1,49	1,45	1,39	1,36	1,31	1,27	1,25
1,82	1,76	1,71	1,64	1,59	1,54	1,47	1,44	1,37	1,34	1,29	1,25	1,22
1,80	1,74	1,69	1,62	1,57	1,52	1,45	1,42	1,35	1,32	1,26	1,22	1,19
1,78	1,72	1,67	1,60	1,54	1,49	1,42	1,38	1,32	1,28	1,22	1,16	1,13
1,76	1,70	1,65	1,58	1,53	1,47	1,41	1,36	1,30	1,26	1,19	1,13	1,08
1,75	1,69	1,64	1,57	1,52	1,46	1,40	1,35	1,28	1,24	1,17	1,11	1,00

f_1 степени свободы (для

f_2	1	2	3	4	5	6	7	8	9	10	11
1	4,052	4,999	5,403	5,625	5,764	5,859	5,928	5,981	6,022	6,056	6,082
2	98,49	99,00	99,17	99,25	99,30	99,33	99,34	99,36	99,38	99,40	99,41
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78
11	9,65	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52
18	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60	3,51	3,44
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37	3,30
21	8,02	5,78	4,87	4,37	4,04	3,81	3,65	3,51	3,40	3,31	3,24
22	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,25	3,17	3,09
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,21	3,13	3,05
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,17	3,09	3,02
27	7,68	5,49	4,60	4,11	3,79	3,56	3,39	3,26	3,14	3,06	2,98
28	7,64	5,45	4,57	4,07	3,76	3,53	3,36	3,23	3,11	3,03	2,95
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,08	3,00	2,92
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,06	2,98	2,90
32	7,50	5,34	4,46	3,97	3,66	3,42	3,25	3,12	3,01	2,94	2,86
34	7,44	5,29	4,42	3,93	3,61	3,38	3,21	3,08	2,97	2,89	2,82
36	7,39	5,25	4,38	3,89	3,58	3,35	3,18	3,04	2,94	2,86	2,78
38	7,35	5,21	4,34	3,86	3,54	3,32	3,15	3,02	2,91	2,82	2,75
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,88	2,80	2,73
42	7,27	5,15	4,29	3,80	3,49	3,26	3,10	2,96	2,86	2,77	2,70
44	7,24	5,12	4,26	3,78	3,46	3,24	3,07	2,94	2,84	2,75	2,68
46	7,21	5,10	4,24	3,76	3,44	3,22	3,05	2,92	2,82	2,73	2,66
48	7,19	5,08	4,22	3,74	3,42	3,20	3,04	2,90	2,80	2,71	2,64
50	7,17	5,06	4,20	3,72	3,41	3,18	3,02	2,88	2,78	2,70	2,62
55	7,12	5,01	4,16	3,68	3,37	3,15	2,98	2,85	2,75	2,66	2,59
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56
65	7,04	4,95	4,10	3,62	3,31	3,09	2,93	2,79	2,70	2,61	2,54
70	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,77	2,67	2,59	2,51
80	6,96	4,88	4,04	3,56	3,25	3,04	2,87	2,74	2,64	2,55	2,48
100	6,90	4,82	3,98	3,51	3,20	2,99	2,82	2,69	2,59	2,51	2,43
125	6,84	4,78	3,94	3,47	3,17	2,95	2,79	2,65	2,56	2,47	2,40
150	6,81	4,75	3,91	3,44	3,14	2,92	2,76	2,62	2,53	2,44	2,37
200	6,76	4,71	3,88	3,41	3,11	2,90	2,73	2,60	2,50	2,41	2,34
400	6,70	4,66	3,83	3,36	3,06	2,85	2,69	2,55	2,46	2,37	2,29
1000	6,66	4,62	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,26
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24

Функция F (α в степени $2z$) вычислена частично по таблице VI Фишера [7].
Остальные данные получены путем интерполяции, в большинстве случаев графической.

распределения F

большого среднего квадрата)

12	14	16	20	24	30	40	50	75	100	200	500	∞
6,106	6,142	6,169	6,208	6,234	6,258	6,286	6,302	6,323	6,334	6,352	6,361	6,366
99,42	99,43	99,44	99,45	99,46	99,47	99,48	99,48	99,49	99,49	99,49	99,50	99,50
27,05	26,92	26,83	26,69	26,60	26,50	26,41	26,35	26,27	26,23	26,18	26,14	26,12
14,37	14,24	14,15	14,02	13,93	13,83	13,74	13,69	13,61	13,57	13,52	13,48	13,46
9,89	9,77	9,68	9,55	9,47	9,38	9,29	9,24	9,17	9,13	9,07	9,04	9,02
7,72	7,60	7,52	7,39	7,31	7,23	7,14	7,09	7,02	6,99	6,94	6,90	6,88
6,47	6,35	6,27	6,15	6,07	5,98	5,90	5,85	5,78	5,75	5,70	5,67	5,65
5,67	5,56	5,48	5,36	5,28	5,20	5,11	5,06	5,00	4,96	4,91	4,88	4,86
5,11	5,00	4,92	4,80	4,73	4,64	4,56	4,51	4,45	4,41	4,36	4,33	4,31
4,71	4,60	4,52	4,41	4,33	4,25	4,17	4,12	4,05	4,01	3,96	3,93	3,91
4,40	4,29	4,21	4,10	4,02	3,94	3,86	3,80	3,74	3,70	3,66	3,62	3,60
4,16	4,05	3,98	3,86	3,78	3,70	3,61	3,56	3,49	3,46	3,41	3,38	3,36
3,96	3,85	3,78	3,67	3,59	3,51	3,42	3,37	3,30	3,27	3,21	3,18	3,16
3,80	3,70	3,62	3,51	3,43	3,34	3,26	3,21	3,14	3,11	3,06	3,02	3,00
3,67	3,56	3,48	3,36	3,29	3,20	3,12	3,07	3,00	2,97	2,92	2,89	2,87
3,55	3,45	3,37	3,25	3,18	3,10	3,01	2,96	2,89	2,86	2,80	2,77	2,75
3,45	3,35	3,27	3,16	3,08	3,00	2,92	2,86	2,79	2,76	2,70	2,67	2,65
3,37	3,27	3,19	3,07	3,00	2,91	2,83	2,78	2,71	2,68	2,62	2,59	2,57
3,30	3,19	3,12	3,00	2,92	2,84	2,76	2,70	2,63	2,60	2,54	2,51	2,49
3,23	3,13	3,05	2,94	2,86	2,77	2,69	2,63	2,56	2,53	2,47	2,44	2,42
3,17	3,07	2,99	2,88	2,80	2,72	2,63	2,58	2,51	2,47	2,42	2,38	2,36
3,12	3,02	2,94	2,83	2,75	2,67	2,58	2,53	2,46	2,42	2,37	2,33	2,31
3,07	2,97	2,89	2,78	2,70	2,62	2,53	2,48	2,41	2,37	2,32	2,28	2,26
3,03	2,93	2,85	2,74	2,66	2,58	2,49	2,44	2,36	2,33	2,27	2,23	2,21
2,99	2,89	2,81	2,70	2,62	2,54	2,45	2,40	2,32	2,29	2,23	2,19	2,17
2,96	2,86	2,77	2,66	2,58	2,50	2,41	2,36	2,28	2,25	2,19	2,15	2,13
2,93	2,83	2,74	2,63	2,55	2,47	2,38	2,33	2,25	2,21	2,16	2,12	2,10
2,90	2,80	2,71	2,60	2,52	2,44	2,35	2,30	2,22	2,18	2,13	2,09	2,06
2,87	2,77	2,68	2,57	2,49	2,41	2,32	2,27	2,19	2,15	2,10	2,06	2,03
2,84	2,74	2,66	2,55	2,47	2,38	2,29	2,24	2,16	2,13	2,07	2,03	2,01
2,80	2,70	2,62	2,51	2,42	2,34	2,25	2,20	2,12	2,08	2,02	1,98	1,96
2,76	2,66	2,58	2,47	2,38	2,30	2,21	2,15	2,08	2,04	1,98	1,94	1,91
2,72	2,62	2,54	2,43	2,35	2,26	2,17	2,12	2,04	2,00	1,94	1,90	1,87
2,69	2,59	2,51	2,40	2,32	2,22	2,14	2,08	2,00	1,97	1,90	1,86	1,84
2,66	2,56	2,49	2,37	2,29	2,20	2,11	2,05	1,97	1,94	1,88	1,84	1,81
2,64	2,54	2,46	2,35	2,26	2,17	2,08	2,02	1,94	1,91	1,85	1,80	1,78
2,62	2,52	2,44	2,32	2,24	2,15	2,06	2,00	1,92	1,88	1,82	1,78	1,75
2,60	2,50	2,42	2,30	2,22	2,13	2,04	1,98	1,90	1,86	1,80	1,76	1,72
2,58	2,48	2,40	2,28	2,20	2,11	2,02	1,96	1,88	1,84	1,78	1,73	1,70
2,56	2,46	2,39	2,26	2,18	2,10	2,00	1,94	1,86	1,82	1,76	1,71	1,68
2,53	2,43	2,35	2,23	2,15	2,06	1,96	1,90	1,82	1,78	1,71	1,66	1,64
2,50	2,40	2,32	2,20	2,12	2,03	1,93	1,87	1,79	1,74	1,68	1,63	1,60
2,47	2,37	2,30	2,18	2,09	2,00	1,90	1,84	1,76	1,71	1,64	1,60	1,56
2,45	2,35	2,28	2,15	2,07	1,98	1,88	1,82	1,74	1,69	1,62	1,56	1,53
2,41	2,32	2,24	2,11	2,03	1,94	1,84	1,78	1,70	1,65	1,57	1,52	1,49
2,36	2,26	2,19	2,06	1,98	1,89	1,79	1,73	1,64	1,59	1,51	1,46	1,43
2,33	2,23	2,15	2,03	1,94	1,85	1,75	1,68	1,59	1,54	1,46	1,40	1,37
2,30	2,20	2,12	2,00	1,91	1,83	1,72	1,66	1,56	1,51	1,43	1,37	1,33
2,28	2,17	2,09	1,97	1,88	1,79	1,69	1,62	1,53	1,48	1,39	1,33	1,28
2,23	2,12	2,04	1,92	1,84	1,74	1,64	1,57	1,47	1,42	1,32	1,24	1,19
2,20	2,09	2,01	1,89	1,81	1,71	1,61	1,54	1,44	1,38	1,28	1,19	1,11
2,18	2,07	1,99	1,87	1,79	1,69	1,59	1,52	1,41	1,36	1,25	1,15	1,06

Проведите дисперсионный анализ и сделайте проверку гипотезы о равенстве всех μ . *Ответ:* средние квадраты: 1) группа 8745; 2) пыллет 722; $F=12,4$. Так как выборочное F далеко выходит за пределы табличного 1%-ного уровня, то почти нет никакого сомнения, что совокупности, относящиеся к различным видам корма, имеют различные μ .

Пример 13. Взяв данные по шерсти из примера 5, проверьте гипотезу, что все токи взяты из совокупностей, имеющих общую среднюю. *Ответ:* $F=1,35$, $F_{0,05}=2,85$. Следовательно, здесь нет веских данных против гипотезы о том, что все токи могут иметь один и тот же процент чистой шерсти.

Пример 14. Выборочные средние в опыте с витамином B_{12} примера 4 различаются меньше, чем это можно ожидать, исходя из среднего квадрата для объектов; такое положение, как показывает таблица 88, встречается во многих случаях. Хотя в этом случае нет смысла вычислять F , все же определите $F=0,54$. Конечно, здесь ничто не указывает на различие между μ .

Пример 15. По данным примера 8 проверьте гипотезу о том, что варианты опыта не оказывают никакого влияния на количество личинок пяденицы. *Ответ:* $F=3,33$. Какое вы сделаете заключение?

Настоящим сообщением с основы дисперсионного анализа читатель поощряется для ознакомления с начальными параграфами главы II.

6. Оценка сравнений между всеми средними. В опыте по выпечке хлеба были получены вполне убедительные свидетельства в пользу того, что между μ существуют определенные различия. Но критерий F сам по себе не указывает, каково количество таких различий. Между a групповыми средними можно определить всего $a(a-1)/2$ разностей; в нашем случае между видами жара имеется $4 \times 3/2=6$ разностей. Будет ли каждое μ отличаться от всех остальных или же некоторые из μ будут различаться несущественно? Если установлен только сам факт наличия существенных различий, то вслед за этим необходимо установить, где они помещаются, т. е. к каким μ они относятся.

Имеется несколько методов для оценки существенности разности между средними [4, 17, 22]. Я здесь принимаю метод, предложенный Дж. Н. Тьюки (с некоторыми видоизменениями), частично по причине удобства его применения и частично потому, что результаты эксперимента представляют собой как бы смесь существенных и несущественных различий.

Оценка различий между средними производится путем определения разности D , которая существенна на 5%-ном уровне; далее с ней сравниваются все $a(a-1)/2$ выборочные разности, полученные в эксперименте. Величина D получается путем умножения $s_{\bar{x}}$ на множитель Q , который берется из таблицы 90. Данные этой таблицы связаны с числом вариантов опыта, указанным в верхнем заголовке ее, и числом степеней свободы f для внутригруппового варьирования (ошибка), указанным слева.

В опыте по выпечке хлеба $s_{\bar{x}}=1/\sqrt{100,9/6}=4,1$, где $s^2=100,9$ является оценкой σ^2 . При четырех вариантах и $f=20$ величина Q по этой таблице равна 3,96. Таким образом:

$$D = Qs_{\bar{x}} = 3,96 \times 4,1 = 16,2.$$

Обращаясь к таблице 85, определяем фактические разности между средними и сравниваем их с D :

Вид жара	\bar{x}	$\bar{x}-162$	$\bar{x}-172$	$\bar{x}-176$
2	185	23	13	9
3	176	14	4	
4	172	10		
4	162			

Здесь выборочные средние расположены по порядку от высшей к низшей, и каждая из них вычитается от всех вышестоящих. Из шести разностей только одна превосходит $D=16,2$. Вывод такой, что жар 2 вытвывается тестом в больших количествах, чем жар 4, и что больше нет никаких других разностей, существенных на 5%-ном уровне.

Верхние 5%-ные уровни Q студентизированного размаха каррирования¹

Число степеней свободы f	Число вариантов o																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	18,0	26,7	32,8	37,2	40,5	43,4	45,4	47,3	49,1	50,6	51,9	53,2	54,3	55,4	56,3	57,2	58,0	58,8	59,9	
2	6,09	8,28	9,80	10,89	11,73	12,43	13,03	13,54	13,99	14,39	14,75	15,08	15,38	15,65	15,91	16,14	16,36	16,57	16,77	
3	4,50	5,88	6,83	7,51	8,04	8,47	8,85	9,18	9,46	9,72	9,95	10,16	10,35	10,52	10,69	10,84	10,98	11,12	11,24	
4	3,93	5,00	5,76	6,31	6,73	7,06	7,35	7,60	7,83	8,03	8,21	8,37	8,52	8,67	8,80	8,92	9,03	9,14	9,24	
5	3,61	4,54	5,18	5,63	5,99	6,28	6,52	6,74	6,93	7,10	7,25	7,39	7,52	7,64	7,75	7,86	7,95	8,04	8,13	
6	3,46	4,34	4,90	5,31	5,63	5,89	6,12	6,32	6,49	6,65	6,79	6,92	7,04	7,14	7,24	7,34	7,43	7,51	7,59	
7	3,34	4,16	4,68	5,06	5,35	5,59	5,80	5,99	6,15	6,29	6,42	6,54	6,65	6,75	6,84	6,93	7,01	7,08	7,16	
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,39	6,48	6,57	6,65	6,73	6,80	6,87	
9	3,20	3,95	4,42	4,76	5,02	5,24	5,43	5,60	5,74	5,87	5,98	6,09	6,19	6,28	6,36	6,44	6,51	6,58	6,65	
10	3,15	3,88	4,33	4,66	4,91	5,12	5,30	5,46	5,60	5,72	5,83	5,93	6,03	6,12	6,20	6,27	6,34	6,41	6,47	
11	3,11	3,82	4,26	4,58	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,91	5,98	6,06	6,14	6,20	6,27	6,33	
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,40	5,51	5,61	5,71	5,80	5,88	5,95	6,02	6,09	6,15	6,21	
13	3,06	3,73	4,15	4,46	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,71	5,79	5,86	5,93	6,00	6,06	6,11	
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,56	5,64	5,72	5,79	5,86	5,92	5,98	6,03	
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,57	5,65	5,72	5,79	5,85	5,91	5,96	
16	3,00	3,65	4,05	4,34	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,52	5,59	5,66	5,73	5,79	5,84	5,90	
17	2,98	3,62	4,02	4,31	4,52	4,70	4,86	4,99	5,11	5,21	5,31	5,39	5,47	5,55	5,61	5,68	5,74	5,79	5,84	
18	2,97	3,61	4,00	4,28	4,49	4,67	4,83	4,96	5,07	5,17	5,27	5,35	5,43	5,50	5,57	5,63	5,69	5,74	5,79	
19	2,96	3,59	3,98	4,26	4,47	4,64	4,79	4,92	5,04	5,14	5,23	5,32	5,39	5,46	5,53	5,59	5,65	5,70	5,75	
20	2,95	3,58	3,96	4,24	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,36	5,43	5,50	5,56	5,61	5,66	5,71	
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,25	5,32	5,38	5,44	5,50	5,55	5,59	
30	2,89	3,48	3,84	4,11	4,30	4,46	4,60	4,72	4,83	4,92	5,00	5,08	5,15	5,21	5,27	5,33	5,38	5,43	5,48	
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,74	4,82	4,90	4,98	5,05	5,11	5,17	5,22	5,27	5,32	5,36	
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,94	5,00	5,06	5,11	5,15	5,20	5,24	
120	2,80	3,36	3,69	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,84	4,90	4,95	5,00	5,04	5,09	5,13	
∞	2,77	3,32	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,74	4,80	4,84	4,89	4,93	4,97	5,01	

¹ Перепечатано из «Биометрики», 39: 192 (1952 г.) с разрешения автора Джойс М. Мей и редакции.

Если в опыте только два варианта, то $Q = t\sqrt{2}$, и поэтому $D = ts_x \sqrt{2}$. Это — получившая широкое распространение *наименьшая существенная разность*, или НСР. Она часто неправильно применяется ко всем разностям. когда в опыте имеется 3 или более средних, в результате чего получается преувеличенное число существенных разностей. Критерий НСР правилен только при $a=2$.

Тьюки определил D как доверительный интервал для ряда выборочных средних. Таким образом, например, при $\mu_2 - \mu_1$ можно с уверенностью сказать, что она лежит в интервале $23 \pm 16,2$ или между 6,8 и 39,2 г.

Так как интервал не покрывает собой нуль, то эти μ , очевидно, различны. Но, например, разность $\mu_2 - \mu_1$ имеет доверительный интервал $13 \pm 16,2$. т. е. от $-2,2$ до 30,2 г, включающий в себя нуль, и поэтому $\mu_2 - \mu_1$ может равняться нулю. Особенность этого доверительного интервала состоит в том, что здесь констатируется факт покрытия нуля, в то же время расстояние верхней и нижней границы его в этом случае не играет никакой роли.

При применении критерия Тьюки возможная ошибка в наших выводах приписывается эксперименту в целом. Риск этой ошибки в таблице 90 определен так, что в 5 опытах из 100 можно прийти к неправильным заключениям. Число же ошибочных частных заключений внутри какого-либо опыта остается неопределенным.

Хартли показал [13], что метод последовательной оценки, подобный методу Кьюлса, до некоторой степени является более мощным, т. е. чувствительным, чем только что описанный метод. В этом случае применяется не одно табличное значение Q , а несколько значений, по одному для каждого интервала между средними по вариантам. В опыте по выпечке хлеба соседние средние в ранжированном их ряде оцениваются при помощи $Q=2,95$ для $a=2$; для двух следующих, более широких интервалов применяется $Q=3,58$ при $a=3$, а $Q=3,96$ применяется только для наибольшего интервала, когда $a=4$. Соответствующие D будут $2,95 \times 4,1=12,1$; $3,58 \times 4,1=14,7$ и $3,96 \times 4,1=16,2$. Эти D располагаются в таблице разностей по диагоналям, проходящим, так сказать, с северо-востока на юго-запад:

Вид жира	\bar{x}	$\bar{x}-162$	$\bar{x}-172$	$\bar{x}-176$
2	185	23 (16,2)	13 (14,7)	9 (12,1)
3	176	14 (14,7)	4 (12,1)	
1	172	10 (12,1)		
4	162			

Каждая разность теперь сравнивается со своим D ; разность будет считаться существенной, если она больше соответствующего D . Общее правило при рассмотрении такой таблицы состоит в том, что, как только будет установлена разность, меньшая соответствующего D , дальнейшее сопоставление разностей, лежащих вправо от нее по ее строке и ниже нее по столбцу, в котором она стоит, с соответствующими D становится бесполезным.

В опыте с выпечкой хлеба мы приходим при обоих способах оценки существенности к одному и тому же выводу. Однако последовательная оценка чаще обнаруживает большее число существенных разностей.

Но не будут ли эти *оценки* разностей в совокупности иллюзорными? Нет, они являются несмещенными оценками, и, будут ли они существенными или несущественными, они все же содержат определенную информацию относительно реальных разностей.

Пример 16. В отношении искусственно построенного примера 9 этой главы мы знаем, что все четыре μ различны. Однако выборочное варьирование затушевывает некоторые

из этих разностей. Вычислите $D=18,4$ и покажите, что только два из всех доверительных интервалов не покрывают нуль. Заметьте, что все x имеют тот же порядок, что и μ ; этого может и не быть.

Пример 17. Вычислите, как это делалось в опыте с выпечкой хлеба, наименьшую существенную разность при $a=2$, $s^2=100,9$, $n=6$, $f=20$, $t_{0,95}=2,086$. *Ответ:* $D=HSP=12,1$. Если этот критерий ошибочно применить к 6 разностям между средними в опыте по выпечке хлеба, то 3 из них должны были бы считаться существенными.

Пример 18. По данным опыта с цыплятами в примере 12 покажите, что четвертая группа существенно отличается от каждой из остальных трех групп.

При первом чтении лица, интересующиеся только основными методами, могут перейти прямо к параграфам 14 и 16 этой главы и после этого к главе 11.

7. Упрощенный метод вычислений при помощи размахов варьирования. Один из упрощенных способов оценки всех сравнений между средними основан на размахах варьирования выборок [18]. В опыте по выпечке хлеба таблицы 85 имеется 4 размаха от отдельных групп: 39, 20, 30, 21 и их сумма 110. Эта сумма размахов умножается на коэффициент, взятый из таблицы 91. В столбце $a=4$ и строке $n=6$ этот коэффициент 0,95. После этого $D'=(\text{коэффициент}) \times (\text{сумма размахов})=0,95 \times 110=10,4$. Величина D' используется так же, как D в критерии Тьюки предыдущего параграфа. Сравнивая с ним 6 разностей между вариантами, находим, что 3 из них теперь следует считать существенными.

Пример 19. Используя упрощенный способ, проверьте все разности в опыте с цыплятами примера 12. *Ответ:* $D'=245$. Получаются те же выводы, что и в примере 18.

ТАБЛИЦА 91

Значения коэффициента для принятия гипотезы при 5%-ном риске¹

Размер выборки n	Количество выборок a									
	2	3	4	5	6	7	8	9	10	
2	3,43	2,37	1,78	1,40	1,16	1,00	0,87	0,78	0,70	
3	1,91	1,44	1,13	0,94	0,80	0,70	0,62	0,56	0,51	
4	1,63	1,25	1,01	0,84	0,72	0,63	0,57	0,51	0,47	
5	1,53	1,19	0,96	0,81	0,70	0,61	0,55	0,50	0,45	
6	1,50	1,18	0,95	0,80	0,69	0,61	0,55	0,49	0,45	
7	1,49	1,17	0,95	0,80	0,69	0,61	0,55	0,50	0,45	
8	1,49	1,17	0,96	0,81	0,70	0,62	0,55	0,50	0,46	
9	1,50	1,18	0,97	0,82	0,71	0,62	0,56	0,51	0,47	
10	1,52	1,20	0,98	0,83	0,72	0,63	0,57	0,52	0,47	

¹ Извлечение из таблицы Куртца, Линка, Тьюки и Уоллеса [18].

8. Критерии для заранее намеченных сравнений между средними. Во многих опытах само их построение предопределяет те сравнения, которые должны быть сделаны; решение этого вопроса не откладывается до тех пор, пока не будут видны результаты опыта, а намечается заранее (обычно число таких намечаемых сравнений равно числу степеней свободы, относящемуся к вариантам). Это предварительное планирование состоит в определении разностей, подлежащих оценке при некотором уровне вероятности. Такой эксперимент является более *целенаправленным*. Планирование сравнений перед проведением опыта должно производиться всегда, как только для этого представляется возможность.

В опыт с витамином B_{12} , приведенный в примере 4, была включена контрольная группа животных, не получавшая витамина B_{12} . Некоторые дополнительные результаты анализа этого опыта приведены в таблице 92. Часто эксперименты подобного рода планируются из расчета иметь два сравнения: одно между двумя вариантами и другое между контролем и вариантом, взятыми вместе.

Оценка первого сравнения между $\bar{x}_{10}=1,58$ и $\bar{x}_5=1,54$ производится путем вычисления суммы квадратов, для чего удобно использовать соответствующие суммы наблюдений

$$\frac{4,62^2 + 4,74^2}{3} - \frac{(4,62 + 4,74)^2}{6} = 0,0024.$$

Эта сумма квадратов с единичной степенью свободы сравнима с ошибкой. Так как она меньше, чем средний квадрат 0,0131, то нет никаких оснований считать различие между вариантами в совокупности существенным.

ТАБЛИЦА 92

Средние суточные привесы 3 групп свиней, получивших антибиотики и 3 дозы витамина В₁₂

В ₁₂ (в мг)	Средние привесы за сутки (в г)			Суммы	Средние
	0	5	10		
0	1,04	1,00	0,59	2,73	0,91
5	1,52	1,56	1,54	4,62	1,54
10	1,63	1,57	1,54	4,74	1,58
Источник варьирования			Число степеней свободы	Сумма квадратов	Средний квадрат
Дозы			2	0,8474	0,4237
Объекты			6	0,0784	0,0131

$$F = 0,4237/0,0131 = 32; \quad F_{0,91} = 10,92$$

Второе сравнение является разностью между средней из двух вариантов $(4,62 + 4,74)/6 = 1,56$ и средней для контроля 0,91. Так как эти средние базируются на разном числе наблюдений, то соответствующая сумма квадратов разности их вычисляется следующим образом (параграф 16 этой главы):

$$\frac{(4,62 + 4,74)^2}{6} + \frac{2,73^2}{3} - \frac{(4,62 + 4,74 + 2,73)^2}{9} = 0,8450.$$

Теперь $F = 0,8450/0,0131 = 64$. Таким образом, большая часть суммы квадратов для трех средних (0,8474) сосредотачивается во втором сравнении, а на первое остается только незначительная ее часть.

Следует отметить тот факт, что две суммы квадратов, каждая из которых соответствует одной степени свободы, складываясь вместе, дают общую сумму квадратов для доз витамина В₁₂:

$$0,0024 + 0,8450 = 0,8474.$$

Эта слагаемость является следствием *ортогональности* сравнений, что более подробно будет рассмотрено в главе 12. Если ваш эксперимент допускает ряд ортогональных сравнений, запланированных перед постановкой опыта, то при обработке данных опыта метод настоящего параграфа будет более эффективным, чем простое сравнение между всеми средними. (См. пример 23 этой главы и главу 12).

9. Дисперсионный анализ при двух группах. В главе 4 был описан частный случай анализа, рассматриваемого в общем виде в настоящей главе, так как там было дано только предварительное представление о дисперсионном анализе на примере двух групп цыплят. Как можно заметить, объединение сумм квадратов отдельных выборок является моментом, общим для обеих этих глав. По данным опыта с цыплятами таблицы 27, такая объединенная сумма квадратов внутри групп цыплят равна 16220 при $f=20$; она

внесена теперь в таблицу 93. Чтобы сделать дисперсионный анализ полным, вычислим сумму квадратов для двух выборочных средних 97 и 56 мг. Для

ТАБЛИЦА 93

Дисперсионный анализ для опыта с цыплятами,
данные таблицы 27

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Внутри групп цыплят	20	16220	811
Групповые средние	1	9245,5	9245,5

$$F = 9245,5/811 = 11,4; \quad \sqrt{F} = 3,38 = t$$

большей ясности, может быть, лучше сначала вычислить две суммы: $97 \times 11 = 1067$ и $56 \times 11 = 616$. После этого сумма квадратов определяется, как обычно:

$$\frac{1067^2 + 616^2}{11} - \frac{1683^2}{22} = 9245,5.$$

Это число также вносится в таблицу 93, и анализ после этого становится полным. F больше своего 1%-ного уровня 8,10; соответственно этому t таблицы 28 также выходит за пределы 1%-ного уровня.

Здесь представляют интерес два положения. 1) Если групповым средним соответствует единичная степень свободы, то $\sqrt{F} = t$. Вероятность для F превзойти 11,4 равна вероятности t превзойти 3,38, и поэтому здесь представляется свободный выбор между тем или другим критерием. 2) При наличии только двух выборок сумма квадратов для их средних проще всего вычисляется так:

$$\frac{(\sum X_1 - \sum X_2)^2}{2n} = \frac{(1067 - 616)^2}{2 \times 11} = 9245,5,$$

это дает то же число, что и ранее. В этом последнем случае просто берется разность между двумя суммами, возводится в квадрат, после чего производится деление на общее число наблюдений. Равным образом, тот же результат может быть получен и таким путем:

$$\frac{n(\bar{x}_1 - \bar{x}_2)^2}{2} = \frac{11(97 - 56)^2}{2} = 9245,5.$$

Я думаю, вы согласитесь с тем, что критерий F более удобен для применения, чем критерий t .

Пример 20. Хансберри и Ричардсон [11] дают проценты поврежденных яблок в двух группах деревьев, по 12 штук каждая. Деревья группы A , опрыснутые мышьяковокислым свинцом, имели 19, 26, 22, 13, 26, 25, 38, 40, 36, 12, 16 и 8% поврежденных яблок. Группа B , опрыснутая мышьяковокислым кальцием и буферными веществами, имела 36, 42, 20, 43, 47, 49, 59, 37, 28, 49, 31 и 39% повреждения плодов. Вычислите средний квадрат внутри выборок 111,41 при 22 степенях свободы и средний квадрат, между выборочными средними 1650,04 при 1 степени свободы. После этого

$$F = 1650,04/111,41 = 14,8.$$

Теперь оцените существенность различия между выборочными средними по способу таблицы 28. Значение $t = 3,85 = \sqrt{14,8}$.

Пример 21. Докажите, что три выражения для суммы квадратов, приведенные в предыдущем параграфе, идентичны друг другу.

Пример 22. Пользуясь таблицей Q , определите 95%-ные доверительные пределы для разности между средними весами гребней цыплят. *Ответ:* $41 \pm 25,3$; от 15,7 до 66,3 мг.

Пример 23. Показать, что критерий планируемого сравнения $x_{10} - \bar{x}_5$ предыдущего параграфа эквивалентен критерию t настоящего параграфа. А также, что он эквивалентен критерию НСР.

10. Модель I. Фиксированные эффекты вариантов. Пришло время дать более строгое формальное изложение вопроса о модели дисперсионного анализа, которая применялась в параграфе 4 этой главы и последующих параграфах. В этих случаях предполагалось, что влияние вариантов опыта приводит к фиксированным значениям средних в совокупности. Выборочные средние являются оценками этих фиксированных μ .

Представляется удобным определить средние совокупностей через общую среднюю всех μ и отклонения от этой последней. Эти отклонения, характеризующие варианты опыта, обозначим через α_i , где i принимает значения 1, 2, ..., a . Так как α_i являются отклонениями от μ , то отсюда следует, что $\sum \alpha_i = 0$. Соответственно каждому α_i имеется n наблюдений или повторений, обозначаемых X_{ij} , где $j=1, 2, \dots, n$. В целом эксперимент символически изображается, как в таблице 94.

Такие обозначения имеют широкое применение в статистической литературе. Первый подстрочный указатель определяет варианты (или группы, или выборки), в то время как второй относится к отдельным наблюдениям, входящим в данный вариант. Подстрочная точка указывает на суммирование: так, $X_{i.}$ означает сумму $j=1 \dots n$ наблюдений варианта за номером i . Две точки, поставленные рядом, означают суммирование по всему эксперименту в целом.

Модель I дисперсионного анализа может быть выражена словами так:

Любое наблюдаемое значение является суммой трех частей: 1) общей средней, 2) отклонения варианта и 3) случайного элемента, взятого из нормально распределенной совокупности, имеющей среднюю, равную нулю, и стандартное отклонение σ .

В алгебраической компактной форме это будет:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, a; \quad j = 1, \dots, n; \quad \varepsilon_{ij} \text{ из } N(0, \sigma).$$

Следует подчеркнуть слагаемость этих элементов. Если, например, один вариант относится к другому, как произведение или отношение, то схема дисперсионного анализа неприменима. Как быть в таком случае, будет обсуждено далее.

Данные примера 9 этой главы были построены в полном соответствии с этой моделью. Общая средняя была $\mu=30$, значения α_i были $\alpha_1=15$, $\alpha_2=0$, $\alpha_3=-6$ и $\alpha_4=-9$. Значения ε_{ij} были взяты из таблицы нормального распределения весов свиней, что более ясно можно видеть из таблицы 82;

ТАБЛИЦА 94

Символическое изображение опыта с a фиксированными вариантами, в каждом по n наблюдений. Общая средняя $=\mu$

Константы вариантов $\mu + \alpha_i$	Отдельные наблюдения				Итого	Средняя
	1	2	... j ...	n		
$\mu + \alpha_1$	X_{11}	X_{12}	... X_{1j} ...	X_{1n}	$X_{1.}$	\bar{x}_1
$\mu + \alpha_2$	X_{21}	X_{22}	... X_{2j} ...	X_{2n}	$X_{2.}$	\bar{x}_2
...
$\mu + \alpha_i$	X_{i1}	X_{i2}	... X_{ij} ...	X_{in}	$X_{i.}$	\bar{x}_i
...
$\mu + \alpha_a$	X_{a1}	X_{a2}	... X_{aj} ...	X_{an}	$X_{a.}$	\bar{x}_a
Итого . . .	Столбцы здесь не имеют определенного смысла, так как X_{ij} по рядам рандомизированы				$X_{..}$	$\bar{x}_{..}$
Средняя						$\bar{x}_{..}$

$\varepsilon_{11}=40-30=10$, $\varepsilon_{12}=24-30=-6$, $\varepsilon_{21}=29-30=-1$, $\varepsilon_{45}=21-30=-9$. Это все ε , взятые из $N(0,10)$.

Правомерность дисперсионного анализа всецело определяется этим свойством слагаемости у эффектов вариантов в совокупности. Доверительные интервалы и критерий существенности основаны на случайном отборе величин ε из одной и той же нормально распределенной совокупности со средней, равной нулю, однако было показано [23], что если имеет место умеренная скошенность распределения, то критерий F будет испытывать только небольшое смещение.

11. Компоненты. Модель I. В примере 9 этой главы было установлено, что изменения групповых средних приводят к заметным изменениям в соответствующем среднем квадрате; он больше уже не является оценкой $\sigma^2=100$. Сюда кое-что прибавилось. Теперь мы установим, что эта прибавка является оценкой величины $n\sum\alpha_i^2/(a-1)$, что для привеса свиней составляет $5(15^2+0^2+6^2+9^2)/3=5 \times 114=570$. Вы узнаете в этой величине средний квадрат α_i , увеличенный до уровня первичных наблюдений путем умножения на $n=5$. Пусть $\kappa^2=\sum\alpha_i^2/(a-1)$. Это κ^2 вычисляется точно так же, как s^2 . Но между ними есть различие, заключающееся в том, что s^2 основывается на наблюдаемых значениях случайной переменной, в то время как κ^2 образуется из фиксированных α_i . Таким образом, $n\kappa^2$ является добавочным компонентом. В таблице 95 представлены точные соотношения, соответствующие этому случаю.

Возможно, что несколько удивительной покажется слишком завышенная оценка параметров; действительно, $n\kappa^2=570$ оценивается значением 947, а $\kappa^2=114$ значением 189,4. Объяснение тех особенностей системы выборок, которые привели к этому результату, даны в примерах 10 и 11 этой главы.

ТАБЛИЦА 95

Дисперсионный анализ данных о привесе свиней из примера 9

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Варианты	3	1050	$\sigma^2 + n\kappa^2$
Объекты	16	103	

1050 — 103 = 947 дает оценку $n\kappa^2=570$;
 947/5 = 189,4 дает оценку $\kappa^2=114$

Выделение в последнем столбце этой таблицы оцениваемых параметров проливает свет на содержание критерия существенности F . Теперь видно, что

$$F \text{ оценивает отношение } \frac{\sigma^2 + n\kappa^2}{\sigma^2}.$$

Нулевая гипотеза $H_0: \kappa^2=0$ идентична гипотезе $\alpha_i=0$. Таким образом, если H_0 верна, то F оценивает 1. Если же H_0 ошибочна, то F оценивает величину, превосходящую единицу. Выборочные же значения F , как показывает таблица 88, при $\kappa^2=0$ могут варьировать в довольно широких размерах, начиная с нуля. Редко может быть так, чтобы малое значение F (т. е. $F < 1$) указывало на существование чего-либо иного, кроме естественного выборочного варьирования. Только большие значения F позволяют предполагать влияние вариантов опыта.

Рассмотренная здесь связь *компонентного анализа* с критерием F будет иметь все более возрастающее значение, когда мы перейдем к ознакомлению с более сложными экспериментами. Другие случаи применения этого анализа будут рассмотрены по мере необходимости.

Пример 24. Оцените $\chi^2=114$ в примере 10. Ответ: 25,4.

Пример 25. Оцените χ^2 для неизвестной системы α_i в опыте по выпечке хлеба по таблице 86. Ответ: 74,1.

Пример 26. В связи с наличием константного члена $n\chi^2$ средний квадрат для вариантов в случаях, подобных представленному в таблице 95, не подчиняется закону распределения σ^2 . Поэтому формулы параграфов 14 главы 2 и 5 главы 3 неприменимы ([1], стр. 342).

12. Модель II. Случайные эффекты. В предыдущем мы обсудили вопрос об эффектах вариантов опыта, которые связаны с некоторыми фиксированными параметрами. Теперь мы перейдем к вопросу о случайных эффектах, возникающих выборочным порядком из нормальных совокупностей. Влияние на результаты опыта особенностей помета поросят таково, каким оно было бы при случайном отборе из совокупности пометов. В данном случае интерес сосредоточивается на оценке компонента дисперсии, обусловленного различием между пометами, и на сравнении его с компонентом дисперсии, относимым к отдельным особям. Еще пример: повторные почвенные образцы могут быть взяты из различных случайно расположенных точек поля. В то время как основным объектом изучения в этом случае является среднее содержание в почве некоторого элемента, все же определенное размещение проб приводит к оценке компонента, относящегося к влиянию данного размещения постольку, поскольку это находит выражение в выборочном варьировании внутри этого размещения. С точки зрения расчетов здесь нет никакого различия между данной моделью и моделью I. Различия между ними возникают при их интерпретации (и их теории).

Символически модель II может быть представлена так:

$$X_{ij} = \mu + A_i + \varepsilon_{ij}; \quad i = 1, \dots, a; \quad j = 1, \dots, n;$$

$$A_i = N(0, \sigma_A); \quad \varepsilon_{ij} = N(0, \sigma).$$

В таблице 96 дан пример опыта со случайными эффектами вариантов. Эти данные взяты из более обширного и более сложного опыта [28]. По четырем пробам с каждого из четырех случайно отобранных листьев одного растения турнепса было определено содержание кальция.

ТАБЛИЦА 96

Содержание кальция в зеленом растении турнепса
(в % от сухого веса)

Лист	Содержание кальция (в %)				Сумма	Средняя
1	3,28	3,09	3,03	3,03	12,43	3,11
2	3,52	3,48	3,38	3,38	13,76	3,44
3	2,88	2,80	2,81	2,76	11,25	2,81
4	3,34	3,38	3,23	3,26	13,21	3,30

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Листья	3	0,2961	$\sigma^2 + 4\sigma_A^2$
Отдельные определения	12	0,0066	σ^2

$s^2=0,0066$ дает оценку σ^2 : $s_A^2=(0,2961-0,0066)/4=0,0724$ дает оценку σ_A^2 .

В процессе данного анализа компонент дисперсии листьев определяется величиной σ_A^2 , а не χ^2 ; это обусловлено тем, что в данном случае мы имеем дело с переменной величиной, дисперсия которой подлежит оценке, а не с рядом константных величин α_i . Заметим, что в качестве выборочных оценок σ^2 и σ_A^2 берутся, как обычно, s^2 и s_A^2 .

Обычно проверяемой нулевой гипотезой является $\sigma_A^2 = 0$. В данном случае:

$$F = \frac{0,2961}{0,0066} = 45 \text{ дает оценку отношению } \frac{\sigma^2 + 4\sigma_A^2}{\sigma^2}.$$

Если результаты окажутся менее заметными, то оценка указанного отношения должна предшествовать разделению компонентов, так как если вы придете к заключению, что σ_A^2 может быть нулем, то нет никакого смысла вычислять ее выборочную оценку s_A^2 .

В нашем случае очевидно, что варьирование при переходе от одного листа к другому значительно больше, чем варьирование между отдельными пробами одного и того же листа. Ясно, что уменьшение проб на каждом листе при увеличении численности этих последних приведет к увеличению точности и экономии средств, так как отбор листьев дешев, а проведение анализа образца — дело относительно дорогое. Для того чтобы сделать это положение более наглядным, рассмотрим средний квадрат для средних по листьям при $an=16$ образцам:

$$s_x^2 = \frac{0,2961}{16} = \frac{0,0066 + 4 \times 0,0724}{16} = \frac{0,0066}{16} + \frac{0,0724}{4}.$$

Теперь допустим, что этот опыт видоизменяется так, что берутся другие n и a , но компоненты дисперсии сохраняются теми же самыми, причем целью этой перепланировки является увеличение информации (т. е. уменьшение s_x^2) при сохранении той же стоимости опыта. В этом новом эксперименте будет:

$$(s_x^2)' = \frac{0,0066 + n' \times 0,0724}{a'n'} = \frac{0,0066}{a'n'} + \frac{0,0724}{a'}.$$

Так как здесь больший числитель равен 0,0724, то совершенно очевидно, что a' должно быть увеличено, а n' уменьшено в пределах, которые допускает условие неизменной стоимости всего эксперимента. Испробуем $n'=1$ и $a'=15$; так как анализ образца стоит, вероятно, раз в 10 дороже сбора листьев, то в данном случае общие издержки на проведение эксперимента изменятся очень мало. При этом, новом плане опыта мы ожидаем:

$$(s_x^2)' = \frac{0,0066}{15} + \frac{0,0724}{15} = 0,0053.$$

При таком уменьшении среднего квадрата «эффективность» предполагаемого опыта по сравнению с первоначальным будет

$$s_x^2 / (s_x^2)' = 0,0185 / 0,0053 = 349\%.$$

Это происходит потому, что дорогой анализ образцов, дающий небольшую изменчивость данных, распространяется на большее число листьев, варьирование которых велико. В параграфе 11 главы 17 будет дано более детальное рассмотрение этого вопроса.

Средний квадрат компонента модели II определяется на основе средних квадратов M_A и M , которые и используются при его вычислении. Рассмотрим компонент s_A^2 в таблице 96. Он вычислен таким путем:

$$s_A^2 = \frac{M_A - M}{n} = \frac{0,2961 - 0,0066}{4}.$$

Отсюда следует, что средний квадрат для этого компонента будет (см. параграф 5 главы 3 и примеры 20 и 21 главы 3):

$$\frac{1}{n^2} (s_{M_A}^2 + s_M^2) = \frac{1}{n^2} \left(\frac{2M_A^2}{f_A + 2} + \frac{2M^2}{f + 2} \right) = \frac{2}{16} \left(\frac{0,2961^2}{5} + \frac{0,0066^2}{14} \right) = 0,002192.$$

Отсюда следует, что выборочное стандартное отклонение s_A^2 равно $\sqrt{0,002192} = 0,047$, что составляет больше половины $s_A^2 = 0,0724$. Это дает

некоторое представление о выборочном варьировании компонента s_A^2 . Я не могу дать точной рекомендации относительно доверительного интервала для этого компонента, так как выборочное распределение его неизвестно. О некоторых приближенных решениях см. [1], стр. 231.

Пример 27. Следующие данные взяты из отчетов об одной линии свиней польско-китайской породы, выведенной на сельскохозяйственной опытной станции штата А.ова. От каждого из четырех пометов, полученных от одного производителя, было взято по 2 борозка; эти животные в период от момента отъема их до достижения ими веса примерно 225 фунтов получали обычный рацион. Ниже приводятся средние суточные привесы:

Пометы	1	2	3	4
Привесы	1,18 1,11	1,36 1,65	1,37 1,40	1,07 0,90

Допуская, что варьирование по пометам имеет нормальное распределение, показать, что σ_A^2 существенно отличается от нуля ($F=7,41$) и что ее оценка равна 0,0474.

Пример 28. Вычислите стандартное отклонение s_A^2 для предыдущего примера. *Ответ:* 0,035. Если s_A^2 распределено нормально, то это выборочное стандартное отклонение определяет доверительный интервал, который покрывает нуль. И все же s_A^2 существенно.

13. Искусственное построение модели II. Не представляет трудностей на основе выборок и известных совокупностей построить эксперимент по модели II. Одну совокупность можно взять для получения варьирующих эффектов вариантов с дисперсией σ_A^2 , а вторую — для отдельных объектов с дисперсией σ^2 ; из каждой совокупности можно взять выборки и сочетать их в некоторой установленной пропорции. В таблице 97 даны результаты такого построения условного эксперимента. Компоненты для пометов (случайных вариантов) были взяты из привесов свиней в таблице 24, и поэтому $\sigma_A^2 = 25$; эта величина одна и та же для каждого помета. Привесы отдельных свиней взяты из таблицы 15 при $\sigma^2 = 100$ по два на каждый помет. Сумма этих данных представляет условно-наблюденные привесы, которые приведены в четвертой колонке таблицы.

Забудем на момент, каким путем образованы эти привесы. Вообразим себе, что они являются фактическими наблюдениями над 20 свиньями, взятыми попарно из случайно отобранных 10 пометов. Дисперсия их разлагается на части обычным порядком, в результате чего происходит выделение компонентов этой дисперсии.

Теперь посмотрите на последнюю строку таблицы и кое-что припомните из предыдущего. По 20 наблюдениям мы определяем оценки для $\sigma^2 = 100$ и $\sigma_A^2 = 25$, т. е. для двух компонентов, которые были положены в основу наших условных привесов свиней; сюда, конечно, включено и некоторое выборочное варьирование.

Этот пример был выбран нами потому, что он дает довольно точные оценки. По таблице 98, где даны результаты обработки 25 подобных условных опытов, можно составить представление об обычном варьировании такого рода данных. Здесь бросается в глаза очень большое варьирование оценок σ_A^2 ; некоторые из них даже отрицательны! Это последнее обстоятельство просто указывает на то, что средний квадрат для групп меньше среднего квадрата для отдельных животных; средние по группам, если они взяты из одной нормальной совокупности, как правило, должны варьировать меньше, чем единичные случайные наблюдения. Ясно, что нельзя ожидать сколь угодно точных оценок σ^2 и σ_A^2 при столь малых выборках, как наши.

Я огорчен тем, что должен указать вам на невозможность измерения этого компонента для пометов; он остается скрытым. Что может быть здесь измеримым, так это только привес в расчете на одну свинью.

Привесы 20 свиней из 10 пометов, по 2 свиньи из каждого

Каждый привес является суммой 3 компонентов. Компонент для пометов является выборкой из таблицы 24 при $\sigma_A^2 = 25$; компонент для отдельных животных является выборкой из таблицы 15 при $\sigma^2 = 100$; каждое наблюдение уменьшено на 30. Третьим компонентом является $\mu = 30$

Номер помета	Компонент помета A_i	Компонент отдельной свиньи ε_{ij}	Выборка привесов свиней $X_{ij} = \mu + A_i + \varepsilon_{ij}$	Выборка привеса помета
(1)	(2)	(3)	(4) = (36) + (2) + (3)	(5)
1	-1	7	36	
		9	38	74
2	2	-4	28	
		-23	9	37
3	-1	0	29	
		19	48	77
4	0	2	32	
		2	32	64
5	-4	3	29	
		12	38	67
6	-10	9	29	
		3	23	52
7	10	5	45	
		-4	36	81
8	2	-19	13	
		-10	22	35
9	4	-4	30	
		18	52	82
10	-2	15	43	
		-6	22	65

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Средние по пометам	9	144,6	$\sigma^2 + 2\sigma_A^2$
Отдельные свиньи	10	96,5	σ^2

$s^2 = 96,5$ дает оценку 100; $s_A^2 = (144,6 - 96,5)/2 = 24,0$ дает оценку 25.

ТАБЛИЦА 98

Оценки $\sigma_A^2 = 25$ и $\sigma^2 = 100$, установленные по 25 выборкам, полученным подобно тому, как это сделано в таблице 97

Номер выборки	Оценка $\sigma_A^2 = 25$	Оценка $\sigma^2 = 100$	Номер выборки	Оценка $\sigma_A^2 = 25$	Оценка $\sigma^2 = 100$
1	60	127	14	56	112
2	56	104	15	-33	159
3	28	97	16	67	54
4	6	91	17	-18	90
5	48	60	18	33	65
6	-5	91	19	-21	127
7	7	53	20	-48	126
8	-1	87	21	4	43
9	0	66	22	3	145
10	-78	210	23	49	142
11	14	148	24	75	23
12	7	162	25	77	106
13	68	76	В среднем	17,0	102,6

14. **Выборки из выборок. Модель II.** Каждая выборка может быть составлена из субвыборок, а эти, в свою очередь, из других субвыборок и т. д. Соотношение выборки и субвыборок дает представление о *гнездовом отборе*, или о *последовательной классификации*. Отдельные серии выборок могут все относиться к модели II или иметь смешанный характер.

В таблице 99 дан пример выборки, относящейся к модели II и разделенной на субвыборки. Эти данные являются частью опыта с турнепсом, о котором говорилось ранее [28]. Случайным порядком были отобраны 4 растения; после этого с каждого из растений также случайно было взято 3 листа. С каждого листа было взято по 2 образца в 100 мг, по которым и было определено микрохимическими методами содержание кальция. Наша задача состоит в том, чтобы выделить суммы квадратов, относящиеся к источникам варьирования: к растениям, к листьям на том же самом растении и к образцам одного и того же листа.

Вычисления сумм квадратов для образцов, листьев и растений производятся знакомым уже нам способом. Новым моментом является здесь то, что теперь у нас два «итога» с соответствующими им и вычитаемыми из них «вари-

ТАБЛИЦА 99

Содержание кальция (в процентах к сухому веществу) в $b=3$ листьям с каждого $a=4$ растениям турнепса; $n=2$ образца с каждого листа. Дисперсионный анализ

Растение, i $i=1, \dots, a$	Лист, j $j=1, \dots, b$	Образцы X_{ijk}		X_{ij}	$X_{i..}$	$X_{...}$
1	1	3,28	3,09	6,37	19,05	72,29
	2	3,52	3,48	7,00		
	3	2,88	2,80	5,68		
2	1	2,46	2,44	4,90	13,07	
	2	1,87	1,92	3,79		
	3	2,19	2,19	4,38		
3	1	2,77	2,66	5,43	17,71	
	2	3,74	3,44	7,18		
	3	2,55	2,55	5,10		
4	1	3,78	3,87	7,65	22,46	
	2	4,07	4,12	8,19		
	3	3,31	3,31	6,62		

Общий размер выборки $= abn = 4 \times 3 \times 2 = 24$ образца.

$$C = X_{...}^2 / abn = 72,29^2 / 24 = 217,7435.$$

$$\text{Образцы: } \sum X_{ijk}^2 - C = 3,28^2 + \dots + 3,31^2 - C = 10,2704.$$

$$\text{Листья: } \sum X_{ij.}^2 / n - C = (6,37^2 + \dots + 6,62^2) / 2 - C = 10,1905.$$

$$\text{Растения: } \sum X_{i..}^2 / bn - C = (19,05^2 + \dots + 22,46^2) / 6 - C = 7,5603.$$

$$\text{Листья на одном и том растении} = \text{листья} - \text{растения} = 10,1905 - 7,5603 = 2,6302.$$

$$\text{Образцы на одном и том же листе} = \text{образцы} - \text{листья} = 10,2704 - 10,1905 = 0,0799.$$

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Растения	3	7,5603	2,5201
Листья на одном растении	8	2,6302	0,3288
Образцы на одном листе	12	0,0799	0,0067
Итого	23	10,2704	

антами»: 1) сумма квадратов для листьев является «итогом», из которого вычитается сумма квадратов для растений, что дает сумму квадратов, обычно приписываемую «объектам», а у нас *листьям на одном и том же растении*, или, как кратко говорят, *листьям на растениях*; 2) сумма квадратов образцов также является «итогом», из которого вычитается сумма квадратов «вариантов», у нас *листьев*, что приводит к образцам на одном и том же листе, или просто к образцам на листьях. Этот процесс может быть распространен и на последующие субвыборки.

Эта модель может быть выражена так:

$$X_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}; \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad k = 1, \dots, n;$$

$$A_i = N(0, \sigma_A); \quad B_{ij} = N(0, \sigma_B); \quad \varepsilon_{ijk} = N(0, \sigma),$$

где A — относится к растениям и B к листьям. Латинские буквы здесь использованы для обозначения растений и листьев потому, что они являются случайными переменными, а не константами.

В развернутом дисперсионном анализе таблицы 100 даны все компоненты дисперсии. Каждый компонент субвыборки входит совместно с компонентом выборки в дисперсию этой последней. Соответствующие оценки вычисляются, как указано в этой таблице.

Нулевые гипотезы, которые в этом случае могут быть проверены, такие:

$$1) \sigma_A^2 = 0; \quad F = \frac{2,5201}{0,3288} = 7,66 \text{ дает оценку } \frac{\sigma^2 + n\sigma_B^2 + nb\sigma_A^2}{\sigma^2 + n\sigma_B^2}, \quad j = 3 \text{ и } 8;$$

$$2) \sigma_B^2 = 0; \quad F = \frac{0,3288}{0,0067} = 49 \text{ дает оценку } \frac{\sigma^2 + n\sigma_B^2}{\sigma^2}, \quad j = 8 \text{ и } 12.$$

В первом случае при степенях свободы $j_1 = 3$ и $j_2 = 8$ значение находится почти на своем 1%-ном уровне — 7,59; во втором же случае при степенях свободы 8 и 12 F лежит далеко за пределами своего 1%-ного уровня — 4,50. Вполне очевидно, что в совокупности, из которой взята выборка, процент кальция варьирует существенным образом как при переходе от листа к листу, так и при переходе от растения к растению.

ТАБЛИЦА 100

Полный дисперсионный анализ данных по турнепсу

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Растения	3	2,5201	$\sigma^2 + n\sigma_B^2 + nb\sigma_A^2$
Листья	8	0,3288	$\sigma^2 + n\sigma_B^2$
Образцы	12	0,0067	σ^2

$$n = 2, \quad b = 3, \quad a = 4, \quad s^2 = 0,0067 \text{ дает оценку } \sigma^2;$$

$$s_B^2 = (0,3288 - 0,0067)/2 = 0,1610 \text{ дает оценку } \sigma_B^2;$$

$$s_A^2 = (2,5201 - 0,3288)/6 = 0,3652 \text{ дает оценку } \sigma_A^2.$$

Средний квадрат для средних по растениям определяется:

$$s_{\bar{X}}^2 = \frac{2,5201}{24} = 0,105 = \frac{0,0067 + n \times 0,1610 + nb \times 0,3652}{nab} =$$

$$= \frac{0,0067}{nab} + \frac{0,1610}{ab} + \frac{0,3652}{a}.$$

Как и в случае параграфа 12 этой главы, большую информацию в пересчете на один доллар можно получить путем уменьшения числа дорогостоящих образцов на листе, потому что они имеют наименьший компонент дисперсии, а за счет этого увеличить число листьев или растений. Отбор растений, вероятно,

обходится дороже, чем отбор листьев на растении, но и компонент растений больше. Каким образом можно сбалансировать эти элементы, будет показано в параграфе 12 главы 17.

Использование компонентов дисперсии становится все более и более широким. Не только при планировании опытов (как показано выше), но и для оценки программы работ они, как об этом было замечено, почти необходимы. В генетике они также находят эффективное применение для определения распределений наследственных признаков.

Пример 29. Вычислите стандартное отклонение для s_B^2 по таблице 100. *Отсет:* 0,074.

Пример 30. Покажите, что стандартное отклонение для $s_A^2 = 0,27$.

15. Выборки из выборок. Смешанная модель. В повторных субвыборках, относящихся к основным выборкам, могут существовать константы, связанные с этими последними. Например, при определении качественной ценности производителей в животноводстве каждый из них покрывает случайную группу маток; критерием в этом случае служит ряд характеристик потомства. Этому будет соответствовать модель:

$$X_{ijk} = \mu + \alpha_i + B_{ij} + \varepsilon_{ijk}, \quad i = 1, \dots, a; \quad j = 1, \dots, b.$$

Величины α_i константны ($\sum \alpha_i = 0$) и связаны с производителями, но B и ε являются случайными переменными, соответствующими маткам и потомству. В таблице 101 приведен соответствующий пример. Здесь приведены более обширные, чем ранее, данные из отчетов о разведении свиней голско-китайской породы на сельскохозяйственной станции штата Айова. Вычисления проводятся, как и прежде. В этой модели имеется только одно изменение, состоящее в том, что средний квадрат для констант α_i обозначен через χ^2 .

Здесь мы имеем несколько необычную ситуацию: средний квадрат для выборок (производители) меньше, чем средний квадрат для субвыборок (матки). Этого в совокупности не может быть даже в случае, если $\chi^2 = 0$. Данное положение возникло в связи с тем, что выборочное варьирование или уменьшило действительный средний квадрат для производителей, или увеличило его для маток, или произвело и то и другое вместе. Но все это, очевидно, не является свидетельством против гипотезы: $H_0: \chi^2 = 0$.

Если бы средний квадрат для производителей был больше среднего квадрата для маток, то нулевая гипотеза $\chi^2 = 0$ (т. е. $\alpha_i = 0$) могла бы быть проверена при помощи критерия:

$$F = \frac{\text{средний квадрат для производителей}}{\text{средний квадрат для маток}}.$$

ТАБЛИЦА 101

Средний суточный привес двух поросят из каждого помета

Производитель	Матка	Привес поросят		Суммы	
1	1	2,77	2,38	5,15	
	2	2,58	2,94	5,52	10,67
2	1	2,28	2,22	4,50	
	2	3,01	2,61	5,62	10,12
3	1	2,36	2,71	5,07	
	2	2,72	2,74	5,46	10,53

4	1 2	2,87 2,31	2,46 2,24	5,33 4,55 9,88
5	1 2	2,74 2,50	2,56 2,48	5,30 4,98 10,28 51,48

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Производители	4	0,0249	$\sigma^2 + n\sigma_B^2 + nb\chi^2$
Матки при том же производителе	5	0,1127	$\sigma^2 + n\sigma_B^2$
Пары у одной матки	10	0,0387	σ^2

$n=2$, $b=2$, $s^2=0,0387$ дает оценку σ^2 ; $s_B^2=(0,1127-0,0387)/2=0,0370$ дает оценку σ_B^2 ; 0 дает оценку χ^2 .

Для проверки гипотезы $\sigma_B^2=0$; $F=0,1127/0,0387=2,91$; $F_{0,05}=3,33$.

16. Выборки разного размера. Иногда не представляется возможным получить выборки одинакового размера, а в некоторых случаях это и нежелательно. В качестве примера этого последнего случая допустим, что производится исследование содержания жира в молоке коров в различных хозяйствах, снабжающих сливочным маслом некоторый район страны.

Допустим, что требуется провести это исследование с помощью выборочного метода в такой его форме, чтобы полученное этим путем среднее количество масла на одну корову, будучи умножено на численность продуктивного скота в районе, давало примерную величину общего количества продукции. В этом случае предпочтительней выборку из отдельных хозяйств брать так, чтобы число коров было бы пропорциональным распределению их по стадам района. В противном случае равное участие некоторых малопродуктивных стад приведет к искажению средней продуктивности. Это указывает как на важность специального отбора выборок, точно отражающих изучаемую совокупность, так и на возможность проведения полностью случайного отбора по отношению к основным единицам наблюдения. Эти два процесса отнюдь не исключают друг друга, но оба входят как составные части хорошо спланированного выборочного наблюдения (глава 17). Кстати, из этого следует, что такое представление о репрезентативности требует, чтобы исследователь ограничивал свои заключения совокупностью, из которой взята выборка, а не распространял бы их на все совокупности вообще.

Другая возможность получения выборок неодинакового размера представлена в таблице 102. Во многих случаях экспериментальный материал имеет именно такую естественную группировку. Здесь приведены данные об опоросе восьми свиней; различия между средними весами поросят при рождении могут быть частично приписаны индивидуальным особенностям маток и частично размеру помета.

Что касается метода расчетов, то для распространения его на случай неравных групп требуется только одно изменение. Это относится к сумме квадратов между средними по пометам. Так как n_i меняется от помета к помету, то для перехода к расчету на единичное наблюдение каждое X_i^2 должно

делиться на соответствующее n_i . Число степеней свободы, получаемое как остаток в таблице дисперсионного анализа, может рассматриваться как сумма 8 значений $n_i - 1$, относящихся к отдельным пометам:

$$\Sigma (n_i - 1) = 9 + 7 + \dots + 3 = 48.$$

Этот путь расчета имеет большой смысл, так как он обращает внимание на тот факт, что соответствующая сумма квадратов 17, 17 является суммой 8 сумм квадратов для отдельных пометов, каждая из которых вычисляется самостоятельно и независимо (табл. 82).

Настоящая схема соответствует модели II:

$$X_{ij} = \mu + A_i + \varepsilon_{ij}; \quad i = 1 \dots a; \quad j = 1 \dots n_i; \quad \varepsilon_{ij} = N(0, \sigma).$$

Новым условием здесь является то, что значения j варьируют от выборки к выборке. Это обстоятельство создает определенные затруднения для выделения компонентов. При равных n_i ($= n$) компонент для вариантов умножается на n . Теперь же необходимо это умножение производить на некоторое усредненное значение из n_i . Теоретически доказано, что соответствующая средняя из n_i определяется так:

$$n_0 = \frac{1}{a-1} \left(n - \frac{\Sigma n_i^2}{n} \right).$$

Для наших данных $\Sigma n_i^2 = 10^2 + 8^2 + \dots + 4^2 = 432$, откуда

$$n_0 = \frac{1}{8-1} \left(56 - \frac{432}{56} \right) = 6,90.$$

Полный дисперсионный анализ приведен в таблице 103. Здесь сделано допущение о наличии общей σ^2 для всех совокупностей, из которых взяты выборки. Случай, когда дисперсии разнородны, см. в параграфе 20 этой главы.

Проверка гипотезы $H_0: \sigma_A^2 = 0$ производится, как ранее:

$$F = \frac{1,07}{0,36} = 2,97 \text{ дает оценку } \frac{\sigma^2 + n_0 \sigma_A^2}{\sigma^2}.$$

Здесь возникает некоторое сомнение в том, что именно особенности маток являются источником увеличения дисперсии поросят внутри помета. Эта дополнительная дисперсия может иметь два источника: 1) матки могут различаться в отношении их способности питать плод, в результате чего будет

ТАБЛИЦА 102

Вес при рождении (в фунтах) 8 пометов поросят польско-китайской породы

Помет i	Вес при рождении X_{ij}										Суммы X_i	Численности n_i	Сумма квадратов ΣX_{ij}^2
1	2,0	2,8	3,3	3,2	4,4	3,6	1,9	3,3	2,8	1,1	28,4	10	
2	3,5	2,8	3,2	3,5	2,3	2,4	2,0	1,6			21,3	8	
3	3,3	3,6	2,6	3,1	3,2	3,3	2,9	3,4	3,2	3,2	31,8	10	
4	3,2	3,3	3,2	2,9	3,3	2,5	2,6	2,8			23,8	8	
5	2,6	2,6	2,9	2,0	2,0	2,1					14,2	6	
6	3,1	2,9	3,1	2,5							11,6	4	
7	2,6	2,2	2,2	2,5	1,2	1,2					11,9	6	
8	2,5	2,4	3,0	1,5							9,4	4	
$a=8$	$\Sigma X_{ij} = \Sigma X_i = 152,4$											$n. = 56$	439,40

$$C = X^2/n. = 152,4^2/56 = 414,75$$

$$\text{Общее: } \Sigma X_{ij}^2 - C = 2,0^2 + 2,8^2 + \dots + 1,5^2 - C = 439,40 - 414,75 = 24,65$$

$$\text{Пометы: } \Sigma \frac{X_i^2}{n_i} - C = \frac{23,4^2}{10} + \dots + \frac{9,4^2}{4} - C = 422,23 - 414,75 = 7,48$$

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	55	24,65	
Пометов	7	7,48	1,07
Поросята в пометах	48	17,17	0,36

$$F = 1,07/0,36 = 2,97; F_{0,01} = 3,04$$

наблюдаться различие в весе порослят при рождении даже в случае, если все остальные факторы одинаковы, и 2) на вес при рождении может влиять размер помета. Разделение этих двух источников варьирования будет предметом обсуждения в следующем параграфе.

ТАБЛИЦА 103

Дисперсионный анализ по данным о пометах таблицы 102

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Пометы	7	1,07	$\sigma^2 + n_0 \sigma_A^2$
Поросята	48	0,36	σ^2

$$n_0 = \frac{1}{a-1} \left(n - \frac{\sum n_i^2}{n} \right) = \frac{1}{7} (56 - \frac{432}{56}) = 6,90; s^2 = 0,36 \text{ дает оценку } \sigma^2;$$

$$s_A^2 = (1,07 - 0,36)/6,90 = 0,103 \text{ дает оценку } \sigma_A^2$$

Если компоненты группы фиксированы (α_i заменяет A_i), то вычисления остаются теми же, но для сохранения принятой системы записей средний квадрат для вариантов должен иметь вид $\sigma^2 + n_0 \alpha^2$.

Последнее в этом параграфе замечание о выборках неодинакового размера: как было установлено в параграфе 5 главы 4, если это осуществимо, то выгоднее иметь дело с группами одинакового размера. При предположении о наличии общей для всех групп σ^2 наилучшее использование ресурсов получается при равных n_i .

Пример 31. Для приобретения практики расчетов ниже приводится несложный пример:

Группа 1: 18, 11, 16, 12, 14, 17, 18, 20.

Группа 2: 19.

Группа 3: 16, 18, 26.

Группа 4: 19, 15, 15, 10, 24, 17, 13, 17, 13, 16.

Общая сумма квадратов, равная 367,65 при 22 степенях свободы, подразделяется на 50,15 с $f=3$ для групповых средних и на 257,50 для объектов. $F=1,23$.

Пример 32. Применяя дисперсионный анализ к двум группам привесов у крыс из параграфа 5 главы 4, покажите, что $\sqrt{F}=t=1,99$. Какие части таблицы 29 можно использовать при данных расчетах?

Пример 33. Дженкинс [16] и Снедекор [26] произвели сравнение урожаев нескольких сортов кукурузы, причем каждый сорт был представлен несколькими линиями. Ниже приводятся урожаи (в бушелях на акр) шести сортов с их линиями:

1. Фор Каунти: 7,3; 4,5; 7,4; 7,4; 5,0; 5,9; 6,4; 6,3; 5,0; 6,1; 7,9; 5,7.

2. Сильвер Кинг: 7,7; 5,4; 5,2; 4,0.

3. Айодент: 6,9; 6,3; 7,6; 8,1; 9,4; 12,0; 15,9; 7,4; 9,0; 5,2; 9,2; 8,6.

4. Ланкастер: 9,6; 7,8; 9,6; 7,7; 8,2; 7,3; 11,3; 9,5; 8,8; 8,4; 6,8.

5. Остерленд: 4,8; 9,2; 8,5; 8,8; 7,9; 5,9; 9,2.

6. Кларк: 4,3; 8,4; 6,6; 4,0; 5,8; 7,6; 3,7.

Вычислите $F=17,55/3,38=5,19$ при $f=5$ и 47. Оцените компонент дисперсии, относящейся к сортам приведенного выше примера. Ответ: $n_0=8,63$, $\alpha_A^2=1,64$.

Пример 34. Для 224 мышей, зараженных 3 штаммами тифа, получены следующие данные о числе дней, прошедших с момента заражения до гибели животных. При проведении дисперсионного анализа метод параграфа 2 главы 8, пожалуй, лучше применять без кодирования чисел дней до гибели X , а использовать их в представленном в таблице виде. Итоговый столбец здесь введен только для контроля. Вычислите $F=179, 9/5,78=31,4$, $f=2$ и 224.

Дней до гибели	Число мышей, зараженных указанным штаммом тифа			Итого
	9D	11C	DSC1	
2	6	1	3	10
3	4	3	5	12
4	9	3	5	17
5	8	6	8	22
6	3	6	19	28
7	1	14	23	38
8		11	22	33
9		4	14	18
10		6	14	20
11		2	7	9
12		3	8	11
13		1	4	5
14			1	1
Итого	31	60	133	224
$\sum X$	125	442	1 037	1 604
$\sum X^2$	561	3 302	8 961	13 124

17. Выборки из выборок при разном их размере. Как выборки, так и субвыборки могут иметь различный размер. Это обстоятельство значительно усложняет дело, но в большинстве практических случаев можно применять упрощенные методы анализа.

Для изучения этого вопроса вернемся к предыдущему параграфу и распределим пометы по четырем группам в соответствии с их размером: 10, 8, 6 и 4 поросенка на помёт. Сумма весов при рождении используется для вычисления «помета» (внутри группы по размеру помета), как это показано в таблице 104. Соответствующий средний квадрат вместе с другими данными указывает на то, что как матки, так и размер помета вносят свою долю в общее варьирование веса поросят при рождении. Вопрос о существенности этих влияний отложим до проведения полного компонентного анализа.

Некоторые затруднения возникают при вычислении коэффициентов, стоящих перед компонентами дисперсии. Вычисление их приведено в таблице 105, а сам компонентный анализ — в таблице 106. Может вызвать удивление то обстоятельство, что здесь имеется два различных усредненных размера помета и ни один из них не вычисляется по схеме в таблице 103. Но средний размер группы $(nb)_0$ вычисляется по ранее проведенной формуле для n_0 .

Относящаяся к данному случаю модель такова:

$$X_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}; \quad i = 1, \dots, a; \quad j = 1, \dots, b_i; \quad k = 1, \dots, n_{ij}.$$

Следствием различных путей усреднения является то, что нет точного критерия для гипотезы $\sigma_A^2 = 0$, потому что ожидаемое отношение, взятое из таблицы 106,

$$\frac{\sigma^2 + 6,76\sigma_B^2 + 13,52\sigma_A^2}{\sigma^2 + 7,00\sigma_B^2}$$

не превращается в единицу при $\sigma_A^2 = 0$. Андерсон и Банкрофт [1] предложили один приближенный критерий, но я в экспериментах, не слишком детальных, допускаю следующий, более простой, способ.

Веса при рождении поросят польско-китайской породы (таблица 102), распределенные по размерам пометов

Размер i	Помет ij	Веса при рождении X_{ijk}	Сумма		Численность		
			в помете X_{ij}	в размере $X_{i..}$	в помете n_{ij}	в размере $n_{i.}$	пометов b_i
1	1	2,0...1,1	28,4	60,2	10	20	2
	2	3,3...3,2	31,8		10		
2	1	3,5...1,6	21,3	45,1	8	16	2
	2	3,2...2,8	23,8		8		
3	1	2,6...2,1	14,2	26,1	6	12	2
	2	2,6...1,2	11,9		6		
4	1	3,1...2,5	11,6	21,0	4	8	2
	2	2,5...1,5	9,4		4		

$a=4$ $\Sigma X_{ijk} = \Sigma X_{ij.} = \Sigma X_{i..} = 152,4$ $n_{..} = 56; \Sigma b_i = 8$

Размер: $\Sigma \frac{X_{i..}^2}{n_{i.}} = \frac{60,2^2}{20} + \dots + \frac{21,0^2}{8} - C = 5,47$

Пометы одного размера: пометы — размер = 7,48 — 5,47 = 2,01

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Размер	3	5,47	1,82
Пометы в размере	4	2,01	0,50
Поросята в помете (таблица 102)	48	17,17	0,36

Вычисление средних численностей в случае последовательной классификации при выборках разного размера

n_{ij} (1)	$n_{i.}$ (2)	n_{ij}^2 (3)	$\Sigma_j n_{ij}^2$ (4)	$\Sigma_j n_{ij}^2 / n_{i.}$ (5)	$n_{i.}^2$ (6)	$b_i - 1$ (7)
10	20	100	200	10	400	1
10		100				
8	16	64	128	8	256	1
8		64				
6	12	36	72	6	144	1
6		36				
4	8	16	32	4	64	1
4		16				
56		432		28	864	4

$n_{0B} = \frac{\Sigma (1) - \Sigma (5)}{\Sigma (7)} = \frac{56 - 28}{4} = 7,00;$

$n_{0A} = \frac{\Sigma (5) - \Sigma (3) / \Sigma (1)}{a - 1} = \frac{28 - 432 / 56}{4 - 1} = 6,76;$

$(nb)_{00} = \frac{\Sigma (1) - \Sigma (6) / \Sigma (1)}{a - 1} = \frac{56 - 864 / 56}{4 - 1} = 13,52.$

Так как выборочные стандартные ошибки этих компонентов велики, то обычно с достаточной точностью можно произвести вычисление обоих усредненных размеров пометов при помощи формулы таблицы 103. Результат этих расчетов 6,90 является числом промежуточным между точными значениями 7,00 и 6,76. Та же самая формула, примененная к размеру групп, дает, так же как и в таблице 105, $(nb)_0 = 13,52$. Этот приближенный способ приводит к средним квадратам:

$$\begin{aligned} \text{Размер помета} & s^2 + 6,90 s_B^2 + 13,52 s_A^2 \\ \text{Пометы} & s^2 + 6,90 s_B^2 \\ \text{Поросята} & s^2 \end{aligned}$$

«Размер помета» теперь сравним с «пометами»: $F = 1,82/0,50 = 3,64$; $F_{0,05} = 6,59$. Здесь мы имеем точный критерий, основанный на приближенных значениях компонентов. Отсюда с двумя десятичными знаками находим $s_B^2 = 0,02$ и $s_A^2 = 0,10$, т. е. те же значения, что и ранее.

ТАБЛИЦА 106

Компонентный анализ дисперсии весов при рождении у поросят

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Размер пометов	3	1,82	$\sigma^2 + n_{0A} \sigma_B^2 + (nb)_0 \sigma_A^2$
Пометы	4	0,50	$\sigma^2 + n_{0B} \sigma_B^2$
Поросята	48	0,36	σ^2

Для пометов: $F = 0,50/0,36 = 1,39$; $F_{0,05} = 2,56$; $s^2 = 0,36$; $s_B^2 = 0,62$; $s_A^2 = 0,10$.

Попытаемся разобраться в той путанице, которая возникла в связи с множественностью статистических показателей, и установить, что же получается в результате исследования. В предыдущем параграфе было доказано, что компонент помета добавляется к варьированию поросят. Теперь обратим внимание на две части s_B^2 и s_A^2 , которые хотя и имеют определенные значения, но все же ни одна из них не оказалась существенной. Первая относится к способности матки оказывать влияние на начальный вес поросят при данном размере помета. Для выявления этой способности следует сравнить X_{ij} таблицы 104: 28,4 фунта с 31,8, 21,3 с 23,8 и т. д. Второй компонент относится к различиям, обусловленным размером помета. Следует ожидать, что поскольку должны быть определенные границы для общего утробного питания будущего помета, то поросята в небольших пометах должны быть в целом крупнее. В нашей выборке этот компонент хотя и оказался небольшим, все же обнаружился. Но рассмотрим соответствующие средние:

$$\begin{aligned} \text{для 10 поросят: } & \bar{x}_1 = 60,2/20 = 3,01 \\ \text{» 8 »} & \bar{x}_2 = 45,1/16 = 2,82 \\ \text{» 6 »} & \bar{x}_3 = 26,1/12 = 2,18 \\ \text{» 4 «} & \bar{x}_4 = 21,0/8 = 2,62 \end{aligned}$$

Если бы этот компонент был существенным, то интерпретация данного факта была бы затруднительной. Дальнейшее рассмотрение вопроса будет проведено в параграфе 20 этой главы.

По мере углубления дробной классификации происходит все большее и большее усложнение точных формул для расчета усредненных размеров выборок [9]. В этих случаях лучше применить указанную приближенную формулу, которая дает точность, достаточную для большинства задач, в особенности, если принять во внимание наличие у этих компонентов довольно крупных стандартных ошибок.

Пример 35. Кокран [3] публикует данные о выборочном изучении урожая пшеницы в 6 районах Великобритании. В каждом районе были взяты одна или несколько ферм на ферме отобрано одно или несколько полей, из которых в случайном порядке отобрано 2 и 4 пробных площадки. Дисперсионный анализ средних по площадкам был таким:

Источник варьирования	Число степеней свободы	Средний квадрат
Районы	5	10,284 = M_1
Фермы в районах	19	6,591 = M_2
Поля на фермах	11	3,026 = M_3
Площадки на полях	39	0,825 = M_4

Распределение пробных площадок было следующим:

№	Районы	Фермы		Поля	
	площадок на район	площадок на ферму	число ферм	площадок на поле	число полей
1	8	4	2	2	4
2	6	4	1	2	2
		2	1	2	1
3	12	12	1	4	3
4	22	4	2	2	11
		2	7		
5	4	4	1	2	2
6	26	4	3	2	13
		2	7		
$P=6$	78	78	25	78	36

Кокран поясняет, что это выборочное наблюдение проведено по модели смешанного характера, так как районы фиксированы и фермы не были случайно отобранными. Однако в качестве числовой иллюстрации допустим, что эти данные относятся к модели П. Андерсон и Банкрофт ([1], стр. 328) произвели компонентный анализ этих данных по точным формулам:

Источник варьирования	Средний квадрат	Параметры
Районы	10,284	$\sigma^2 + 2,34 \sigma_C^2 + 4,90 \sigma_B^2 + 11,96 \sigma_A^2$
Фермы	6,591	$\sigma^2 + 2,00 \sigma_C^2 + 2,58 \sigma_B^2$
Поля	3,026	$\sigma^2 + 2,36 \sigma_C^2$
Площадки	0,825	σ^2

$$s^2 = 0,825; s_C^2 = 0,933; s_B^2 = 1,512; s_A^2 = -0,011$$

Применяя приближенную формулу, я получил:

Источник варьирования	Средний квадрат	Параметры
Районы	10,284	$\sigma^2 + 2,163 \sigma_C^2 + 3,064 \sigma_B^2 + 11,959 \sigma_A^2$
Фермы	6,591	$\sigma^2 + 2,163 \sigma_C^2 + 3,064 \sigma_B^2$
Поля	3,026	$\sigma^2 + 2,163 \sigma_C^2$
Площадки	0,825	σ^2

$$s^2 = 0,825; s_C^2 = 1,018; s_B^2 = 1,164; s_A^2 = 0,309$$

При сравнении этих двух систем оценок учтите большие величины стандартных отклонений, которые вы вычислите для некоторых компонент. Например, стандартное отклонение для s_B^2 находится где-то вблизи 0,9. В случае ответственных и тонких исследований точные формулы могут быть необходимыми, но обычно достаточны и значительно легче приближенные решения.

18. Размер выборки. Эта задача уже рассматривалась в параграфах 15 главы 2 и 7 главы 4. При наличии приведенных после этого таблиц теперь можно дать более действенные решения задачи. Это дает возможность экспериментатору при планировании своего опыта лучше обеспечить выполнение своих намерений [29]. Конечно, не всегда можно получить дополнительную уверенность за счет увеличения повторности наблюдений.

Как и ранее, мы допускаем возможность основываться на некоторой заранее заданной оценке σ ; назовем ее s_0 при f_0 степенях свободы. Теперь представим себе опыт, состоящий из a вариантов по n наблюдений в каждом. Это означает новое число степеней свободы $f = a(n-1)$. Нам требуется установить, какая разность δ может быть обнаружена при этих условиях. Здесь мы встречаемся с новым моментом, в отношении которого вы в предыдущем изложении отсылались к настоящему параграфу, а именно: к выбору вероятности того, что наш прогноз будет правильным. Эту вероятность, если не будет оговорено противного, будем брать равной 0,75. Это означает, что мы имеем 1 шанс из 4 испытать неудачу при попытке обнаружить такую разность δ , которая существует в действительности. Соответствующая формула будет:

$$\delta = \frac{Q_{\alpha, f} \cdot s_0 \sqrt{F_{f, f_0}}}{\sqrt{n}}.$$

Таблица 90 дает $Q_{\alpha, f}$ для a вариантов в строке, соответствующей f степеням свободы. Значение F_{f, f_0} для $P = (1 - 0,75) = 0,25$ берется из таблицы 107 при $f_1 = f$ и $f_2 = f_0$. Таким образом, в каждом подстрочном указателе первая буква относится к столбцу, а вторая к строке соответствующей таблицы.

Для иллюстрации вернемся к примерам 12 и 18 этой главы. В последнем было найдено, что $Q_{s_x} = 48,7$ и что только 4-й вариант отличается от других. Величина 48,7 составляет 57% от средней опыта 84,8; этот эксперимент имеет малую точность. Насколько повысится точность опыта, если в каждую группу включить по 20 случайно отобранных цыплят?

Из данных первоначального опыта находим $s_0 = \sqrt{722} = 26,9$ при $f_0 = 16$. Для предполагаемого же опыта будет $a = 4$, $n = 20$, $f = 76$, отсюда $Q_{4, 76} = 3,72$ и $F_{76, 16} = 1,36$. (Указания по поводу интерполяции: Q и f находятся во взаимном отношении.) Из этих данных находим:

$$\delta = \frac{3,73 \times 26,9 \times \sqrt{1,36}}{\sqrt{20}} = 26.$$

Следовательно, при 20 цыплятах в группе предполагаемый эксперимент должен ($P = 0,75$) обнаруживать любую разность, большую $100 \times 26/84,8 = 31\%$ от средней. Этим путем можно испробовать и другие размеры групп, пока не будет получен опыт с достаточной чувствительностью.

Можно поставить вопрос и о прямом определении n таким образом, чтобы достигнуть определенной точности опыта, положим, $\delta = 20\%$ от 84,8, т. е.

17. Решая то же уравнение относительно n , имеем³

$$n = \frac{Q_{\alpha, f}^2 \cdot s_0^2 \cdot F_{f, f_0}}{\delta^2}.$$

Так как Q и F зависят от n , то для решения этого уравнения приходится применять метод последовательного приближения. Это облегчается тем, что Q и F в некоторых частях соответствующих таблиц изменяются очень медленно. Возьмите явно завышенное значение n и определите первое

25, 10, 2,5 и 0,5%-ные уровни в распределении F^1

11 степени свободы (для большего среднего квадрата)

f_2	P	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	0,250	5,83	7,50	8,20	8,58	8,82	8,98	9,10	9,19	9,26	9,32	9,41	9,49	9,58	9,63	9,67	9,71	9,76	9,80	9,85
	0,100	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,20	60,70	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
	0,025	648	800	864	900	922	937	948	957	963	969	977	985	993	997	1,001	1,003	1,004	1,004	1,018
	0,005	16,211	20,000	21,615	22,500	23,056	23,437	23,715	23,925	24,091	24,224	24,426	24,630	24,836	24,940	25,044	25,148	25,253	25,359	25,465
2	0,250	2,57	3,00	3,15	3,23	3,28	3,31	3,34	3,35	3,37	3,38	3,39	3,41	3,43	3,43	3,44	3,45	3,46	3,47	3,48
	0,100	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
	0,025	38,51	39,00	39,16	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,42	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
	0,005	198	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	199	200
3	0,250	2,02	2,28	2,36	2,39	2,41	2,42	2,43	2,44	2,44	2,44	2,45	2,46	2,46	2,46	2,46	2,47	2,47	2,47	2,47
	0,100	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
	0,025	17,44	16,04	15,44	15,10	14,88	14,74	14,62	14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
	0,005	55,55	49,80	47,47	46,20	45,39	44,84	44,43	44,13	43,88	43,69	43,39	43,08	42,78	42,62	42,47	42,34	42,15	41,99	41,83
4	0,250	1,81	2,00	2,05	2,06	2,07	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08
	0,100	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
	0,005	31,33	26,28	24,26	23,16	22,46	21,98	21,62	21,35	21,14	20,97	20,70	20,44	20,17	20,03	19,89	19,75	19,61	19,47	19,32
5	0,250	1,69	1,85	1,88	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,88	1,88	1,88	1,88	1,87	1,87	1,87
	0,100	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
	0,005	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62	13,38	13,15	12,90	12,78	12,66	12,53	12,40	12,27	12,14
6	0,250	1,62	1,76	1,78	1,79	1,79	1,78	1,78	1,77	1,77	1,77	1,77	1,76	1,76	1,75	1,75	1,75	1,74	1,74	1,74
	0,100	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
	0,005	18,64	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25	10,03	9,81	9,59	9,47	9,36	9,24	9,12	9,00	8,88
7	0,250	1,57	1,70	1,72	1,72	1,71	1,71	1,70	1,70	1,69	1,69	1,68	1,68	1,67	1,67	1,66	1,66	1,65	1,65	1,65
	0,100	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
	0,005	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38	8,18	7,97	7,75	7,64	7,53	7,42	7,31	7,19	7,08
8	0,250	1,54	1,66	1,67	1,66	1,66	1,65	1,64	1,64	1,64	1,63	1,62	1,62	1,61	1,60	1,60	1,59	1,59	1,58	1,58
	0,100	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
	0,005	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21	7,01	6,81	6,61	6,50	6,40	6,29	6,18	6,06	5,95

f₁ степени свободы (для большего среднего квадрата)

f ₂	P	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
9	0,250	1,51	1,62	1,63	1,63	1,62	1,61	1,60	1,60	1,59	1,59	1,58	1,57	1,56	1,56	1,55	1,54	1,54	1,53	1,53
	0,100	3,36	3,61	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
	0,005	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42	6,23	6,03	5,83	5,73	5,62	5,52	5,41	5,30	5,19
10	0,250	1,49	1,60	1,60	1,60	1,59	1,58	1,57	1,56	1,56	1,55	1,54	1,53	1,52	1,52	1,51	1,51	1,50	1,49	1,48
	0,100	3,28	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
	0,005	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85	5,66	5,47	5,27	5,17	5,07	4,97	4,86	4,75	4,64
11	0,250	1,47	1,58	1,58	1,57	1,56	1,55	1,54	1,53	1,53	1,52	1,51	1,50	1,49	1,49	1,48	1,47	1,47	1,46	1,45
	0,100	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
	0,005	12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42	5,24	5,05	4,86	4,76	4,65	4,55	4,44	4,34	4,23
12	0,250	1,46	1,56	1,56	1,55	1,54	1,53	1,52	1,51	1,51	1,50	1,49	1,48	1,47	1,46	1,45	1,45	1,44	1,43	1,42
	0,100	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
	0,005	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09	4,91	4,72	4,53	4,43	4,33	4,23	4,12	4,01	3,90
13	0,250	1,45	1,55	1,55	1,53	1,52	1,51	1,50	1,49	1,49	1,48	1,47	1,46	1,45	1,44	1,43	1,42	1,42	1,41	1,40
	0,100	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
	0,005	11,37	8,19	6,93	6,23	5,79	5,48	5,25	5,08	4,94	4,82	4,64	4,46	4,27	4,17	4,07	3,97	3,87	3,76	3,65
14	0,250	1,44	1,53	1,53	1,52	1,51	1,50	1,49	1,48	1,47	1,46	1,45	1,44	1,43	1,42	1,41	1,41	1,40	1,39	1,38
	0,100	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
	0,005	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,72	4,60	4,43	4,25	4,06	3,96	3,86	3,76	3,66	3,55	3,44
15	0,250	1,43	1,52	1,52	1,51	1,49	1,48	1,47	1,46	1,46	1,45	1,44	1,43	1,41	1,41	1,40	1,39	1,38	1,37	1,36
	0,100	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
	0,025	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,76	2,70	2,64	2,58	2,52	2,46	2,40
	0,005	10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42	4,25	4,07	3,88	3,79	3,69	3,58	3,48	3,37	3,26
16	0,250	1,42	1,51	1,51	1,50	1,48	1,47	1,46	1,45	1,44	1,44	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,35	1,34
	0,100	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
	0,005	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,38	4,27	4,10	3,92	3,73	3,64	3,54	3,44	3,33	3,22	3,11
17	0,250	1,42	1,51	1,50	1,49	1,47	1,46	1,45	1,44	1,43	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,35	1,34	1,33
	0,100	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
	0,005	10,38	7,35	6,16	5,50	5,07	4,78	4,56	4,39	4,25	4,14	3,97	3,79	3,61	3,51	3,41	3,31	3,21	3,10	2,98

F₁ степени свободы (для большого среднего квадрата)

f ₂	P	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
18	0,250	1,41	1,50	1,49	1,48	1,46	1,45	1,44	1,43	1,42	1,42	1,40	1,39	1,38	1,37	1,36	1,35	1,34	1,33	1,32
	0,100	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
	0,005	10,22	7,21	6,03	5,37	4,96	4,66	4,44	4,28	4,14	4,03	3,86	3,68	3,50	3,40	3,30	3,20	3,10	2,99	2,87
19	0,250	1,41	1,49	1,49	1,47	1,46	1,44	1,43	1,42	1,41	1,41	1,40	1,38	1,37	1,36	1,35	1,34	1,33	1,32	1,30
	0,100	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
	0,025	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
	0,005	10,07	7,09	5,92	5,27	4,85	4,56	4,34	4,18	4,04	3,93	3,76	3,59	3,40	3,31	3,21	3,11	3,00	2,89	2,78
20	0,250	1,40	1,49	1,48	1,47	1,45	1,44	1,43	1,42	1,41	1,40	1,39	1,37	1,36	1,35	1,34	1,33	1,32	1,31	1,29
	0,100	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
	0,005	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85	3,68	3,50	3,32	3,22	3,12	3,02	2,92	2,81	2,69
21	0,250	1,40	1,48	1,48	1,46	1,44	1,43	1,42	1,41	1,40	1,39	1,38	1,37	1,35	1,34	1,33	1,32	1,31	1,30	1,28
	0,100	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,88	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
	0,025	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
	0,005	9,83	6,89	5,73	5,09	4,68	4,39	4,18	4,01	3,88	3,77	3,60	3,43	3,24	3,15	3,05	2,95	2,84	2,73	2,61
22	0,250	1,40	1,48	1,47	1,45	1,44	1,42	1,41	1,40	1,39	1,39	1,37	1,36	1,34	1,33	1,32	1,31	1,30	1,29	1,28
	0,100	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
	0,025	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
	0,005	9,73	6,81	5,65	5,02	4,61	4,32	4,11	3,94	3,81	3,70	3,54	3,36	3,18	3,08	2,98	2,88	2,77	2,66	2,55
23	0,250	1,39	1,47	1,47	1,45	1,43	1,42	1,41	1,40	1,39	1,38	1,37	1,35	1,34	1,33	1,32	1,31	1,30	1,28	1,27
	0,100	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
	0,025	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
	0,005	9,63	6,73	5,58	4,95	4,54	4,26	4,05	3,88	3,75	3,64	3,47	3,30	3,12	3,02	2,92	2,82	2,71	2,60	2,48
24	0,250	1,39	1,47	1,46	1,44	1,43	1,41	1,40	1,39	1,38	1,38	1,36	1,35	1,33	1,32	1,31	1,30	1,29	1,28	1,26
	0,100	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
	0,025	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
	0,005	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59	3,42	3,25	3,06	2,97	2,87	2,77	2,66	2,55	2,43
25	0,250	1,39	1,47	1,46	1,44	1,42	1,41	1,40	1,39	1,38	1,37	1,36	1,34	1,33	1,32	1,31	1,29	1,28	1,27	1,25
	0,100	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
	0,025	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
	0,005	9,48	6,60	5,46	4,84	4,43	4,15	3,94	3,78	3,64	3,54	3,37	3,20	3,04	2,92	2,82	2,72	2,61	2,50	2,38
26	0,250	1,38	1,46	1,45	1,44	1,42	1,41	1,40	1,39	1,38	1,37	1,35	1,34	1,32	1,31	1,30	1,29	1,28	1,26	1,25
	0,100	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
	0,025	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
	0,005	9,41	6,54	5,41	4,79	4,38	4,10	3,89	3,73	3,60	3,49	3,33	3,15	2,97	2,87	2,77	2,67	2,56	2,45	2,33

1/1 степени свободы (для большого среднего квадрата)

f_2	P	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40	60	120	∞
27	0,250	1,38	1,46	1,45	1,43	1,42	1,40	1,39	1,38	1,37	1,36	1,35	1,33	1,32	1,31	1,30	1,28	1,27	1,26	1,24
	0,100	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
	0,025	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
	0,005	9,34	6,49	5,36	4,74	4,34	4,06	3,85	3,69	3,56	3,45	3,28	3,11	2,93	2,83	2,73	2,63	2,52	2,41	2,29
28	0,250	1,38	1,43	1,45	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,34	1,33	1,31	1,30	1,29	1,28	1,27	1,25	1,24
	0,100	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
	0,025	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
	0,005	9,28	6,44	5,32	4,70	4,30	4,02	3,81	3,65	3,52	3,41	3,25	3,07	2,89	2,79	2,69	2,59	2,48	2,37	2,25
29	0,250	1,38	1,45	1,45	1,43	1,41	1,40	1,38	1,37	1,35	1,35	1,34	1,32	1,31	1,30	1,29	1,27	1,26	1,25	1,23
	0,100	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
	0,025	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
	0,005	9,23	6,40	5,28	4,66	4,26	3,98	3,77	3,61	3,48	3,38	3,21	3,04	2,86	2,76	2,66	2,56	2,45	2,33	2,21
30	0,250	1,38	1,45	1,44	1,42	1,41	1,39	1,38	1,37	1,36	1,35	1,34	1,32	1,30	1,29	1,28	1,27	1,26	1,24	1,23
	0,100	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
	0,005	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34	3,18	3,01	2,82	2,73	2,63	2,52	2,42	2,30	2,18
40	0,250	1,36	1,44	1,42	1,40	1,39	1,37	1,36	1,35	1,34	1,33	1,31	1,30	1,28	1,26	1,25	1,24	1,22	1,21	1,19
	0,100	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
	0,005	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12	2,95	2,78	2,60	2,50	2,40	2,30	2,18	2,06	1,93
60	0,250	1,35	1,42	1,41	1,38	1,37	1,35	1,33	1,32	1,31	1,30	1,29	1,27	1,25	1,24	1,22	1,21	1,19	1,17	1,15
	0,100	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
	0,005	8,49	5,80	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90	2,74	2,57	2,39	2,29	2,19	2,08	1,96	1,83	1,69
120	0,250	1,34	1,40	1,39	1,37	1,35	1,33	1,31	1,30	1,29	1,28	1,26	1,24	1,22	1,21	1,19	1,18	1,16	1,13	1,10
	0,100	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,54	1,48	1,45	1,41	1,37	1,32	1,26	1,19
	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
	0,005	8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71	2,54	2,37	2,19	2,09	1,98	1,87	1,75	1,61	1,43
∞	0,250	1,32	1,39	1,37	1,35	1,33	1,31	1,29	1,28	1,27	1,25	1,24	1,22	1,19	1,18	1,16	1,14	1,12	1,08	1,00
	0,100	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00
	0,025	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00
	0,005	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,62	2,52	2,36	2,19	2,00	1,90	1,79	1,67	1,53	1,36	1,00

¹ Перепечатано с разрешения авторов и редакции из «Таблиц процентных уровней обращенного бета (P) распределения», составленных Мэксис Меррингтон и Кэтрин М. Томпсон. *Biometrika* 33, 73 (1943).

приближение. Например, $n=121$ представляется значением, превышающим приемлемый размер группы. Для этого n находим $j=120 \times 4=480$. Интерполяция дает $Q_{4,480}=3,65$ и $F_{480,16}=1,34$. Подставляя эти данные в формулу, находим:

$$1\text{-е приближение для } n = \frac{3,65^2 \times 722 \times 1,34}{17^2} = 45.$$

Теперь это приближение проверяется по формуле для δ . При $j=44 \times 4=176$, $Q_{4,176}=3,67$ и $F_{176,16}=1,34$ получаем:

$$\delta = \frac{3,67 \times 2,69 \times \sqrt{1,35}}{\sqrt{45}} = 17,1.$$

Так как это значение несколько больше заданного $\delta=17$, то увеличим данное 1-е приближение на единицу. Вы найдете, что при $n=46$ получается заданное $\delta=17$. Если экспериментатор увеличит размер группы до 46 цыплят, то он получит достаточно хороший шанс ($P=0,75$) обнаружить любую разность порядка 17.

Рекомендуется брать f_0 столь большим, чтобы подлежащая информация о варьировании была бы обеспечена. Для этого можно иногда объединять экспериментальные ошибки нескольких однотипных опытов. При некоторых лабораторных работах изменчивость подопытного материала столь хорошо известна, что можно считать f_0 равным ∞ .

Если же, однако, нет никаких достаточно определенных сведений относительно s_0 и f_0 , то экспериментатор может составить себе о них представление, исходя из общего знания особенностей подопытного материала. Он, например, имеет представление о размахе варьирования веса или урожая, которые можно встретить в условиях опыта. В этом случае он может воспользоваться таблицей 10 для перевода размаха в стандартное отклонение. Знание соответствующих коэффициентов вариации также может оказать известную помощь. Исходя из такого рода сведений, экспериментатор может дать оценку верхнего и нижнего пределов σ , в отношении которых он ожидает, что истинное значение σ с равным шансом расположено между ними. Тем самым он делает попытку фиксировать 50%-ный доверительный интервал для σ . Пусть g и h соответственно верхний и нижний пределы. Тогда s_0 определяется как $(g+h)/2$. Для f_0 следует взять числа из такой таблицы [14]:

Отношение g/h	3,6	2,2	1,84	1,67	1,58	1,51	1,46	1,42	1,39	1,36
Число степеней свободы f_0	1	2	3	4	5	6	7	8	9	10

Приведенная выше обширная таблица 107 требует некоторых дальнейших пояснений. Кроме применения ее для расчетов размера выборки при 25 и 10%-ных уровнях, она дает и широкие возможности для оценки однородности дисперсий двух выборок. Сравнение ее с таблицей 31 показывает, что последняя содержит в себе данные, относящиеся к 2,5%-ному уровню новой таблицы, т. е. дает двусторонний критерий, который всегда необходим при делении большего среднего квадрата на меньший. В новой таблице при оценке однородности дисперсий, например, на 1%-ном уровне следует пользоваться 0,5%-ным уровнем. Эта таблица при желании может быть использована для оценки существенности и при других, кроме 5% и 1%-ных, уровнях. Если следствия из такой оценки не очень ответственные, то может быть допустимой и оценка на 10%-ном уровне. В некоторых случаях бывает необходимо знать, какое количество значений F находится вне интервала 5—1%, что также можно определить по этой таблице.

Для определения промежуточных вероятностей, не вошедших в таблицу, обычно достаточна линейная интерполяция. Более точные результаты можно

получить, связывая табличные значения F с логарифмами вероятностей (или, что проще, процентов) и применяя после этого графический способ интерполяции.

Часто возникает потребность в интерполяции между табличными значениями степеней свободы, особенно при применении таблицы 107. Обычно достаточно линейная интерполяция. Если нужна более высокая точность, следует связать табличные значения F с обратными значениями степеней свободы. Получаемый таким образом график имеет почти линейную форму, что позволяет легко определить F , соответствующее любому числу степеней свободы.

Пример 36. Было отобрано 4 отчета, относящихся к периоду с 1941 по 1947 г., о содержании альбумина в плазменном протеине у нормальных лиц на основе показаний электрофора [27].

Число лиц	Средний процент	Число степеней свободы	s	s^2	Σx^2
12	62,3	11	3,60	12,96	142,56
15	60,3	14	2,80	7,84	109,6
7	59,5	6	5,78	33,41	200,46
16	61,5	15	4,30	18,49	277,35
50	Взвешенная $\bar{x} = 61,05$	46			730,13

При помощи метода, изложенного в параграфе 20 этой главы, была проверена и принята гипотеза об однородности дисперсий. Это позволило произвести оценку общей σ^2 , в качестве которой взято значение $730,13/46 = 15,87$, откуда $s_0 = 3,98$ при $f_0 = 46$. Взвешенный средний процент равен 61,05 ($C = 6,5\%$). Исследователь планирует применить данную технику для изучения 4 вариантов опыта. Он желает, чтобы при 5%-ном уровне обнаруживались различия, не меньше 3% содержания альбумина, при шансе успеха этого обнаружения 3 : 1. Сколько он должен взять лиц для каждого отдельного варианта? *Ответ:* 29. Замечание: 3% альбумина составляет около 5% от средней.

Пример 37. Некоторый исследователь не может дать действительную оценку σ , но, исходя из своих знаний о размахе варьирования подошного материала, он с шансом на успех 1 : 4 считает, что σ находится в пределах между $g = 18$ и $h = 13$. Определите оценку $s_0 = 15,5$ и $f_0 = 9$. Экспериментатор считает, что $\mu = 120$. Планируя 4 варианта опыта, он желает обнаружить различия, не меньше 12. Какой в этом случае требуется размер групп? *Ответ:* 35.

Пример 38. По данным предыдущего примера определите размер группы при условии, чтобы исследователь имел 9 : 1 шансов на успех обнаружения заданной разности. *Ответ:* 49.

Пример 39. Взяв данные третьего от конца абзаца в параграфе 15 главы 2, проверьте размер выборки, необходимой для отношения шансов 3 : 1.

Пример 40. По данным последнего абзаца параграфа 7 главы 4 произведите проверку размера выборки — 25.

19. Внутрикласовая корреляция. Ранее было установлено, что средний квадрат «вариантов» может быть подразделен на две части, одна из которых отражает естественное варьирование объектов при одинаковых условиях, а вторая возникает в зависимости от генетических условий или особенностей окружающей обстановки, присущих отдельным выборкам. Первая является оценкой σ^2 , которую можно считать одинаковой для всех совокупностей, откуда взяты выборки, в то время как вторая является оценкой σ_A^2 , добавочной частью дисперсии, относящейся к различиям между средними совокупностей. Сумма $\sigma^2 + \sigma_A^2$ является дисперсией объектов, на которые мы случайно наталкиваемся в том целом, что образуется из объединения всех совокупностей. Отношение этих двух дисперсий

$$\frac{\sigma_A^2}{\sigma^2 + \sigma_A^2}$$

известно под названием *внутриклассовой корреляции* Q_1 . В числителе этого отношения стоит дисперсия, которая является общей для всех объектов

внутри выборок или группы, потому что все они в одинаковой мере зависят от варьирования, присущего группе как единому целому. В выборках таблицы 97, взятых для иллюстрации, эта величина $\sigma_A^2 = 25$. В знаменателе же отношения стоит некоторая усредненная дисперсия, которая приписывается объектам, когда они выбираются случайно из объединенной совокупности, без учета границ между составляющими ее частными совокупностями.

Представление о том, почему здесь применяется название «корреляция», можно получить, возвращаясь к параграфу 4 главы 7, где было установлено, что r также является отношением двух дисперсий. Числитель в обеих этих корреляциях относится к дисперсии, которая является общей для объектов: это — *ковариация* в одном случае и σ_A^2 — в другом случае. Знаменатель характеризует усредненное варьирование, которое испытывают объекты, когда они не классифицируются по X_1 и X_2 или по субвыборкам.

Если все выборки взяты из одной и той же совокупности, как это имело место по отношению к данным о привесе свиней в таблицах 82 и 88, то $\sigma_A^2 = 0$, и, следовательно, $q_I = 0$; вариация в целом как объектов, так и группы здесь возникает из случайного варьирования выборок, взятых из одной совокупности. Конечно, в частной выборке при $q_I = 0$ значение внутриклассовой r_I не обязательно равно нулю. Так, в таблице 87 $s^2 = 114,0$, $s_A^2 = (172,3 - 114,0) / 10 = 5,83$ и

$$r_I = \frac{5,83}{114,0 + 5,83} = 0,049.$$

Гипотеза о том, что $q_I = 0$, тождественна с гипотезой, по которой $\sigma_A^2 = 0$, и поэтому несущественное значение

$$F = 172,3 / 114,0 = 1,51$$

является правомерным критерием обеих этих гипотез.

Большая величина внутриклассовой корреляции указывает на относительно малое варьирование объектов внутри субвыборок. Действительно, в предельном случае, когда $\sigma^2 = 0$,

$$q_I = \frac{\sigma_A^2}{\sigma + \sigma_A^2} = 1.$$

Так как этот случай возможен только тогда, когда все объекты в каждой группе имеют точно одну и ту же величину, то он представляет собой только теоретический интерес. Однако приводимые в таблице 108 данные о двойнях одного пола иллюстрируют поразительную близость к полной корреляции. Общее число складок на пальцах почти одно и то же у членов каждой пары, но заметно различается между парами. Так как эти величины, вероятно, не зависят от возраста, то полученная здесь высокая корреляция для двоен $r_I = 0,966$ измеряет наследственную схожесть их по данному признаку.

Довольно полное представление о варьировании r_I в малых выборках можно получить, вычисляя его значения по дисперсиям случайных выборок в таблице 98. Например, в выборке 1:

$$r_I = 60 / (60 + 127) = 0,321.$$

ТАБЛИЦА 108

Количество складок на пальцах обеих рук у 12 двоен женского пола
Данные Неймана, Фрмана и Холзигера [21]

Пары	Количество складок внутри двоен	Пары	Количество складок внутри двоен	Пары	Количество складок внутри двоен
1	71, 71	5	76, 70	9	114, 113
2	79, 82	6	83, 82	10	94, 91
3	105, 99	7	114, 113	11	75, 83
4	115, 114	8	57, 44	12	76, 72

Дисперсионный анализ

Источник варьирования	Число степеней свободы	Средний квадрат
Двойни попарно	11	817,31
Внутри двоек	12	14,29

$$s^2 = 14,29; s_A^2 = 401,51; r_I = 0,966$$

Среди этих 25 выборок выборка за номером 20 имеет наименьшую корреляцию — 0,615, а за номером 24 — наибольшую 0,765. Это и будет обычное выборочное варьирование корреляции, при ее значении в совокупности:

$$r_I = 25/(25 + 100) = 0,2.$$

При $n=2$ внутриклассовая корреляция может быть усреднена и ее доверительные пределы могут быть установлены так, как показано в параграфе 6 главы 7 ([7], глава 7). Значение z имеет отрицательное смещение, которое может быть приблизительно исправлено путем прибавления $1/(2a-1)$, где a — число парных наблюдений. Так, для выборки за номером 1 из таблицы 98 имеем:

$$\begin{aligned} r_I &= 0,321 \\ z &= 0,333 \\ 1/(2a-1) &= 0,053 \end{aligned}$$

$$\text{Несмещенное } z = 0,386$$

При непосредственном вычислении внутриклассовой корреляции по таблице дисперсионного анализа удобно придать формуле такой вид:

$$r_I = \frac{M_A - M}{M_A + (n-1)M},$$

где M_A и M обозначают средние квадраты для «вариантов» (группы) и объектов. По данным для двоек, $M_A=817,31$, $M=14,29$ и $n=2$, откуда

$$r_I = \frac{817,31 - 14,29}{817,31 + (2-1) \times 14,29} = 0,97 \text{ (как и ранее).}$$

В случаях, когда субвыборки имеют различный размер, применяют усредненное n_0 . Например, для данных о поросятах в таблице 102 имеем:

$$r_I = \frac{1,07 - 0,36}{1,07 + (6,90 - 1) \times 0,36} = 0,222.$$

Критерий F для этой таблицы указывает на то, что эта невысокая корреляция все же существенна.

Отрицательное r_I встречается тогда, когда выборочные средние варьируют меньше, чем это ожидается, исходя из варьирования отдельных объектов, то есть когда M_A меньше M . Если это случится, то s_A^2 приобретает отрицательное значение. Гипотеза $\sigma_A^2=0$ (или $\chi^2=0$) в этом случае не отвергается, а альтернативное заключение будет состоять в том, что нет неравенства $\alpha_i \neq 0$; здесь мы имеем *некоторый контроль над случайным отбором выборок*. Это означает, что или наблюдения внутри групп имели неестественный слишком большой разброс, или средние групп были сжаты в более узком промежутке, чем это обычно бывает при выборочных наблюдениях. Первый случай может встретиться при сбалансированном подборе животных в группы, когда отдельные особи выбираются так, чтобы по всем группам выдерживать одинаковый размах варьирования (или стандартное отклонение). Второй случай возникает от поправок, производимых время от времени и направленных

ных на выравнивание средних. Если бы потребовалось графически изобразить данные о двойках, то было бы неизвестно, показатели какого индивидуума в каждой паре следует взять в качестве X_1 и какого индивидуума в качестве X_2 . Решение вопроса будет состоять в том, чтобы отмечать каждую пару наблюдений двумя точками: например, вторая пара в таблице 108 определяет точки (79,82) и (82,79), которые расположены симметрично относительно прямой, проходящей через начало координат под углом 45° (угол равен 45° , так как шкалы для X_1 и X_2 одинаковы). Если будут тройки вместо двоек, то каждый ряд наблюдений, a , b и c будет давать 6 точек (a , b), (b , a), (a , c), (c , a), (b , c) и (c , b). Число точек на каждую выборку по мере увеличения n возрастает очень быстро: для $n=4$ их 12, для $n=5$ их 20 и т. д. При большом n работа по вычислению r_I непосредственно по всем таким парам была просто невыполнимой, пока Гаррис [12] не открыл упрощенный способ, подобный только что изложенному дисперсионному анализу.

Пример 41. В выборках шерсти примера 5 этой главы $M_A=11,11$ и $M=8,22$. Допуская теперь, что этот пример удовлетворяет модели II, вычислите $r_I=0,105$. Значение $F=1,35$ указывает, что r_I не является существенным. В совокупности, из которой взята выборка, образцы, относящиеся к одному и тому же типу сходны между собой не больше, чем образцы, взятые из различных типов.

Пример 42. По данным о турнеесе в таблице 99 можно определить две внутрикласовые корреляции. Определите внутрикласовую корреляцию между наблюдениями для «листьев одного растения» 0,960 и между образцами листа 0,690. В соответствии с критерием F каждая из них существенна.

20. Критерии однородности дисперсий. Время от времени мы поднимали вопрос, могут ли два или несколько средних квадратов считаться существенно различающимися. По отношению к двум выборкам ответ на этот вопрос был дан в параграфе 8 главы 4. Для случая же большего числа выборок Бартлеттом [2] был предложен специальный критерий однородности.

В случае одинакового размера выборок данный критерий состоит из сравнения логарифма обобщенного среднего квадрата, умноженного на a , с суммой логарифмов отдельных средних квадратов. В таблице 109 этот метод применен к четырем выборочным s^2 , взятым из таблицы 82; в этом случае известно, что выборки взяты из одной общей совокупности. Множитель

ТАБЛИЦА 109

Вычисление критерия однородности дисперсий Бартлетта для выборок одинакового размера

Данные из таблицы 82: $a=4$, $n=5$

Выборки	Сумма квадратов	Средний квадрат s^2	$\log s^2$
1	472	118	2,07188
2	396	99	1,99584
3	616	154	2,18752
4	164	41	1,61278

$$\Sigma s^2 = 412 \quad \Sigma \log s^2 = 7,86782$$

$$\text{Среднее } \bar{s}^2 = \Sigma s^2/a = 412/4 = 403 \quad \log \bar{s}^2 = 2,01284$$

$$a \log \bar{s}^2 = 4 \times 2,01284 = 8,05136$$

$$\Sigma \log s^2 = 7,86782$$

$$\text{Разность} = 0,18354$$

$$\chi^2 = 2,3026 (n-1) (a \log \bar{s}^2 - \Sigma \log s^2) = 2,3026 \times (5-1) \times 0,18354 = 1,69$$

$$f = a - 1 = 3$$

$$\text{Поправочный делитель } C = 1 + \frac{a-1}{3a(n-1)} = 1 + \frac{4-1}{3 \times 4 \times (5-1)} = 1,1042$$

$$\text{Исправленное } \chi^2 = \chi^2/C = 1,69/1,1042 = 1,53$$

2,3026 является константным ($\log_e 10$) и введен для того, чтобы можно было бы пользоваться обычными логарифмами. Непосредственно вычисленное хи-квадрат несколько смещено в сторону преувеличения. В нашем примере даже это смещенное значение хи-квадрат указывает на то, что варьирование s^2 здесь меньше, чем ожидаемое в среднем при случайном выборочном наблюдении (табл. 6). Необходимость введения поправки в χ^2 возникает только тогда, когда наблюдаемое значение χ^2 лежит вблизи от критического табличного значения, да и то только в случае, если требуется определить точно значение P .

При выборках разного размера вычисление хи-квадрат следует тому же принципу, хотя и становится более сложным. В таблице 110 критерий однородности применен к данным о весе поросят при рождении из таблицы 102. Каждая Σx^2 вычисляется по соответствующей сумме ΣX^2 путем вычитания обычной поправки на среднюю $(\Sigma X)^2/n$. Эта таблица не требует каких-либо дальнейших пояснений.

В отношении данных о весе поросят при рождении здесь обнаруживается некоторая аномалия: не представляется возможным дать подходящее физиологическое объяснение имеющемуся различию между дисперсиями. Действительно, в пометах с 10 поросятами наблюдаются наибольшее и наименьшее значения дисперсии, в то время как остальные различия дисперсий кажутся возникшими случайно. Все это совместно с необъяснимым изменением веса поросят в связи с изменением размера помета (параграф 17 этой главы) приводит к тому, что данная выборка содержит в себе мало полезных сведений.

Кроме этой странности биологического характера, имеется здесь и чисто статистическое обстоятельство, меняющее некоторые из допущений, сделанных в параграфе 16 этой главы. Оценка среднего квадрата внутри группы 0,36,

ТАБЛИЦА 110

Вычисление критерия однородности дисперсий Барлетта. Выборки разного размера
Вес поросят при рождении по данным таблицы 102

выборки	Σx^2	Число степеней свободы $n-1$	Обратные величины $1/(n-1)$	Средний квадрат s^2	$\log s^2$	$(n-1) \log s^2$
1	8,18	9	0,11111	0,9089	-0,04148	-0,3733
2	3,48	7	0,14286	0,4972	-0,30347	-2,1243
3	0,68	9	0,11111	0,0756	-1,12148	-10,0933
4	0,72	7	0,14286	0,4029	-0,98758	-6,9131
5	0,73	5	0,20000	0,1460	-0,83565	-4,1782
6	0,24	3	0,33333	0,0800	-1,09391	-3,2907
7	1,97	5	0,20000	0,3940	-0,40450	-2,0225
8	1,17	3	0,33333	0,3900	-0,40894	-1,2268

$$a=8 \quad 17,17 = \Sigma x^2 \quad 48 = \Sigma (n-1) \quad 1,5756 = \Sigma \{1/(n-1)\} \quad \Sigma (n-1) \log s^2 = -30,2222$$

$$\bar{s}^2 = \Sigma x^2 / \Sigma (n-1) = 17,17 / 48 = 0,3577;$$

$$(\log \bar{s}^2) \times \Sigma (n-1) = -0,44948 \times 48 = -21,4310$$

$$\chi^2 = 2,3026 [(\log \bar{s}^2) \times \Sigma (n-1) - \Sigma (n-1) \log s^2] = 2,3026 [-21,4310 - (-30,2222)] = 20,24.$$

$$\text{Исправочный делитель: } C = 1 + \frac{1}{3(a-1)} \left(\Sigma \frac{1}{n-1} - \frac{1}{\Sigma (n-1)} \right) = 1 + \frac{1}{3 \times 7} \left(1,5746 - \frac{1}{48} \right) = 1,074.$$

$$\text{Исправленное } \chi^2 = 20,24 / 1,074 = 18,85^{**}; \quad f = a - 1 = 7.$$

2,3026 является константным ($\log_e 10$) и введен для того, чтобы можно было бы пользоваться обычными логарифмами. Непосредственно вычисленное хи-квадрат несколько смещено в сторону преувеличения. В нашем примере даже это смещенное значение хи-квадрат указывает на то, что варьирование s^2 здесь меньше, чем ожидаемое в среднем при случайном выборочном наблюдении (табл. 6). Необходимость введения поправки в χ^2 возникает только тогда, когда наблюдаемое значение χ^2 лежит вблизи от критического табличного значения, да и то только в случае, если требуется определить точно значение P .

При выборках разного размера вычисление хи-квадрат следует тому же принципу, хотя и становится более сложным. В таблице 110 критерий однородности применен к данным о весе поросят при рождении из таблицы 102. Каждая Σx^2 вычисляется по соответствующей сумме ΣX^2 путем вычитания обычной поправки на среднюю $(\Sigma X)^2/n$. Эта таблица не требует каких-либо дальнейших пояснений.

В отношении данных о весе поросят при рождении здесь обнаруживается некоторая аномалия: не представляется возможным дать подходящее физиологическое объяснение имеющемуся различию между дисперсиями. Действительно, в пометах с 10 поросятами наблюдаются наибольшее и наименьшее значения дисперсии, в то время как остальные различия дисперсий кажутся возникшими случайно. Все это совместно с необъяснимым изменением веса поросят в связи с изменением размера помета (параграф 17 этой главы) приводит к тому, что данная выборка содержит в себе мало полезных сведений.

Кроме этой странности биологического характера, имеется здесь и чисто статистическое обстоятельство, меняющее некоторые из допущений, сделанных в параграфе 16 этой главы. Оценка среднего квадрата внутри группы 0,36,

ТАБЛИЦА 110

Вычисление критерия однородности дисперсий Бартлетта. Выборки разного размера
Вес поросят при рождении по данным таблицы 102

Выборки	Σx^2	Число степеней свободы $n-1$	Обратные величины $1/(n-1)$	Средний квадрат s^2	$\log s^2$	$(n-1) \log s^2$
1	8,18	9	0,11111	0,9089	-0,04148	-0,3733
2	3,48	7	0,14286	0,4972	-0,30347	-2,1243
3	0,68	9	0,11111	0,0756	-1,12148	-10,0933
4	0,72	7	0,14286	0,1029	-0,98758	-6,9131
5	0,73	5	0,20000	0,1460	-0,83565	-4,1782
6	0,24	3	0,33333	0,0800	-1,09391	-3,2907
7	1,97	5	0,20000	0,3940	-0,40450	-2,0225
8	1,17	3	0,33333	0,3900	-0,40894	-1,2268
$a=8$	$17,17 = \Sigma x^2$	$48 = \Sigma (n-1)$	$1,5756 = \Sigma [1/(n-1)]$			$\Sigma (n-1) \log s^2 = -30,2222$

$$\bar{s}^2 = \Sigma x^2 / \Sigma (n-1) = 17,17 / 48 = 0,3577;$$

$$(\log \bar{s}^2) \times \Sigma (n-1) = -0,44648 \times 48 = -21,4310$$

$$\chi^2 = 2,3026 [(\log \bar{s}^2) \times \Sigma (n-1) - \Sigma (n-1) \log s^2] =$$

$$= 2,3026 [-21,4310 - (-30,2222)] = 20,24.$$

$$\text{Поправочный делитель: } C = 1 + \frac{1}{3(a-1)} \left(\Sigma \frac{1}{n-1} - \frac{1}{\Sigma (n-1)} \right) =$$

$$= 1 + \frac{1}{3 \times 7} \left(1,5746 - \frac{1}{48} \right) = 1,074.$$

$$\text{Исправленное } \chi^2 = 20,24 / 1,074 = 18,85^{**}; f = a - 1 = 7.$$

как теперь установлено, является средней из средних квадратов выборок, взятых из совокупностей с различными σ^2 . Если это так, то отношение

$$\frac{s^2 + n_0 s_A^2}{s^2}$$

не подчиняется закону распределения F , и этот критерий в таблице 102 теряет свою силу.

Таким образом, и с этой стороны данные о пометах пороят не представляются ценными: здесь имеется непонятная тенденция как в изменении весов при рождении, так и в соотношении средних квадратов, в связи с чем этот материал не содержит в себе сколько-нибудь определенных сведений. Конечно, возможны опыты, представляющие более значительный интерес. Первый вопрос, на который следует в таких сомнительных случаях ответить, это: «Какой смысл имеет наблюдаемое различие дисперсий?» Раскрытие этого явления может быть более важным для практики делом, чем простое обнаружение фактов, относящихся к средним. Если же этого сделать не удастся, то естественно прийти к заключению, что различие между средними квадратами представляет собой явление только случайного характера, и поэтому применение обычного критерия для оценки гипотезы $\sigma_A = 0$ становится законным.

Если очевидно (в частности, на основе критерия Бартлетта), что σ^2 различны, то иногда можно произвести проверку гипотезы о том, что средние равны между собой даже при наличии неравенства дисперсий [15, 30, 40]. Соответствующий метод состоит в вычислении взвешенных средних квадратов, после чего производится обычная оценка их отношения при помощи распределения F (эта оценка не является абсолютно точной, но я думаю, что вы не будете слишком обеспокоены этим обстоятельством). Все вычисления, необходимые в этом случае, приведены в таблице 111.

ТАБЛИЦА 111

Вычисление критерия для гипотезы $H_0: \mu_i = \mu$ при неравенстве дисперсий
Данные из таблицы 102

Размер выборки n_i (1)	Средняя \bar{x}_i (2)	Средний квадрат s_i^2 (3)	Вес $w_i = n_i / s_i^2$ (4)	Отклонения $\bar{x}_i - \bar{x}_w$ (5)	$(\bar{x}_i - \bar{x}_w)^2 = (5)^2$ (6)	$\frac{w_i}{\sum w_i} = \frac{(4)}{\sum (4)}$ (7)	$\left(1 - \frac{w_i}{\sum w_i}\right)^2 = -[1 - (7)]^2$ (8)	$\frac{\left(1 - \frac{w_i}{\sum w_i}\right)^2}{\frac{w_i - 1}{(1) - 1}} = \frac{(8)}{(1) - 1}$ (9)
10	2,84	0,9089	11,00	-0,05	0,0025	0,03110	0,9388	0,4043
8	2,66	0,4972	16,09	-0,23	0,0529	0,04549	0,9111	0,1302
10	3,18	0,0756	132,28	0,29	0,0841	0,37398	0,3919	0,0435
8	2,98	0,1029	77,75	0,09	0,0081	0,21981	0,6087	0,0870
6	2,37	0,1460	41,10	-0,52	0,2704	0,11620	0,7811	0,1562
4	2,90	0,0800	50,00	0,01	0,0001	0,14136	0,7373	0,2458
6	1,98	0,3940	15,23	-0,91	0,8281	0,04306	0,9157	0,1831
4	2,35	0,3900	10,26	-0,54	0,2916	0,02901	0,9428	0,3143

$$a = 8 \quad \bar{x}_w = 2,89 \quad \Sigma (4) = 353,71 \quad \Sigma (9) = 1,2644$$

$$\text{Взвешенная средняя } \bar{x}_w = \frac{\Sigma (2) (4)}{\Sigma (4)} = \frac{2,84 \times 11,00 + \dots + 2,35 \times 10,26}{353,71} = 2,89 \text{ фунта.}$$

$$\text{Взвешенная сумма квадратов: } \Sigma (4) (6) = 11,00 \times 0,0025 + \dots + 10,26 \times 0,2916 = 39,36;$$

$$F' = \frac{\Sigma (4) (6) / (a - 1)}{1 + \frac{2(a - 2)}{a^2 - 1} \cdot \Sigma (9)} = \frac{39,36 / (8 - 1)}{1 + \frac{2 \times (8 - 2)}{8^2 - 1} \times 1,2644} = 4,53;$$

$$f_1 = a - 1 = 7; \quad f_2 = \frac{1}{3 \Sigma (9) / (a^2 - 1)} = \frac{1}{3 \times 1,2644 / 63} = 16,6.$$

Величина F' приблизительно распределена как F при f_1 и f_2 степенях свободы. Из таблицы 107 следует, что P близко к 0,005: H_0 отвергается с весьма малым риском впасть при этом в ошибку. Здесь опровержение гипотезы даже более решительное, чем это было при сомнительном критерии таблицы 102. Иногда возможно и обратное этому. Нельзя производить выбор критерия в зависимости от размера P ; его следует выбирать в соответствии с гипотезой относительно σ^2 .

Л И Т Е Р А Т У Р А

1. Anderson R. L., Bancroft T. A., Statistical Theory in Research. McGraw-Hill Book Company, Inc., New York, 1952.
2. Bartlett M. S., Supplement to the Journal of the Royal Statistical Society, 4, 137, 1937.
3. Cochran W. G., Journal of the American Statistical Association, 34, 492, 1939.
4. Duncan David B., Biometrics, 11, 1, 1955.
5. Engledow F. L., Yule G. Udny, The Principles and Practice of Field Trials. Empire Cotton Growing Corporation, London, 1926.
6. Fisher R. A., International Mathematical Conference, Toronto, 1924.
7. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925--1950.
8. Fisher R. A., Yates F., Statistical Tables. Oliver and Boyd, Edinburgh, 1938--1953.
9. Ganguli M., Sankhya, 5, 449, 1941.
10. Gurland John, Queries in Biometrics, 11, December, 1955.
11. Hansberry T. Roy, Richardson Charles H., Iowa State College Journal of Science, 10, 27, 1935.
12. Harris J. A., Biometrika, 9, 446, 1913.
13. Hartley H. O., Communications on Pure and Applied Mathematics, 8, 47, 1955.
14. Harris Marilyn, Horvitz D. G., Mood A. M., Journal of the American Statistical Association, 43, 391, 1948.
15. James G. S., Biometrika, 38, 324, 1951.
16. Jenkins M. T., Journal of Agricultural Research, 39, 677, 1929.
17. Keuls M., Euphytica, 1, 112, 1952.
18. Kurtz T. E., Link R. F., Tukey J. W., Wallace D. L., Journal of the American Statistical Association, 1956.
19. Lowe Belle, Данные Айовской с.-х. опытной станции. 1935.
20. Mahalanobis P. C., Indian Journal of Agricultural Science, 2, 694, 1932.
21. Newman Horatio H., Freeman Frank N., Holzinger Karl J., Twins. The University of Chicago Press, 1937.
22. Newman D., Biometrika, 31, 20, 1939.
23. Pearson E. S., Biometrika, 23, 114, 1931.
24. Query in Biometrics, 5, 250, 1949.
25. Richardson D., Catron D. V., Underkoffler L. A., Maddock H. M., Friedland W. C., The Journal of Nutrition, 44, 371, 1951.
26. Snedecor George W., Analysis of Variance and Covariance. Collegiate Press, Inc., Ames, Iowa, 1934.
27. Snedecor George W., Annals of the New York Academy of Sciences, 52, 792, 1950.
28. Studies of Sampling Techniques and Chemical Analyses of Vegetables. Southern Cooperative Series Bulletin No. 10, 1951.
29. Tukey J. W., The Problem of Multiple Comparisons. Mimeographed for limited circulation, 1953.
30. Welch B. L., Biometrika, 38, 330, 1951.

ОПЫТЫ ПРИ ДВОЙНОЙ ГРУППИРОВКЕ. ДИСПЕРСИОННЫЙ АНАЛИЗ

1. **Использование предварительных сведений об опытном материале.** Исследователь часто имеет возможность предвидеть в общих чертах поведение того экспериментального материала, с которым он собирается работать. Ему, например, известно, что молодые самцы крыс будут расти быстрее, чем самки при тех же условиях. Растения в одной части теплицы могут образовывать больше цветков, чем растения в другой части ее, если даже те и другие находятся на одном и том же стеллаже. Такого рода сведения о материале могут быть использованы для повышения ценности опыта. В предыдущей главе это обстоятельство не было использовано при планировании опыта. В этом случае объекты распределялись в группы по жребию; ошибка опыта обуславливалась не только непредугадываемым варьированием, возникающим при случайном отборе, но также и некоторым варьированием материала, которое можно было бы предвидеть. Построение опыта с двойной группировкой материала, к описанию которого мы теперь переходим, дает экспериментатору возможность найти полезное применение предварительных знаний особенностей опытного материала. В главе 2 этот вопрос был уже рассмотрен для случая двух вариантов. Теперь эти методы будут распространены на случаи сравнений трех и более вариантов.

2. **Опыт с группировкой по двум признакам.** Из опытов по кормлению свиней известно, что один помет может реагировать на один и тот же рацион кормления совсем по-иному, чем другой помет. Ричардсон и др. [20] запланировали исключение из ошибки опыта этого заранее известного источника варьирования, для чего они испытывали все варианты на поросятах одного помета, делая пометы повторениями опыта. Каждое животное здесь классифицируется по двум признакам: помет, к которому он относится, и вариант, который на нем испытывается.

В таблице 112 приведены данные из более обширного опыта, о котором уже упоминалось ранее (пример 4 главы 10, в котором принадлежность к помету не учитывалась). Здесь опущены нулевая доза витамина B_{12} и варианты с антибиотиками.

Вычисления, при одном исключении, проводятся по знакомой уже схеме. В дополнение к суммам квадратов «общая» и «вариантов» теперь появилась сумма квадратов для «пометов». Остаток — «различие», как будет показано далее, является оценкой экспериментальной ошибки. Этот остаток совместно с соответствующими степенями свободы возникает здесь как следствие теоремы сложения, относящейся к ортогональным зависимостям. Отметим также, что число степеней свободы «различий» является произведением чисел степеней свободы главных эффектов, т. е. «вариантов» и «пометов».

Характерной особенностью данного дисперсионного анализа является разложение общей суммы квадратов и общего числа степеней свободы на части, каждая из которых относится к некоторому элементу опыта.

Выделение из ошибки влияния различий между пометами дает экспериментатору возможность использовать свои предварительные знания об опытных животных для уменьшения s^2 и тем самым для повышения точности опыта.

Прежде чем перейти к сравнению полученных средних квадратов, полезно ознакомиться с теми результатами данной обработки материала, которые можно ожидать при проведении случайного отбора из общей совокупности. Приводимая ниже таблица 113 отличается от таблицы опыта с витамином B_{12} только тем, что привесы свиней все взяты случайно из нормального распределения таблицы 15. Преимущество этого примера в том, что мы точно знаем дисперсию $\sigma^2 = 100$ совокупности, из которой взята выборка. Можно видеть, что, когда все наблюдения взяты из общей совокупности, все 4 средних квадрата являются оценками общей σ^2 ; в данном примере эти оценки очень близки к σ^2 .

ТАБЛИЦА 112

Средний суточный привес (в фунтах) в 75 фунтам живого веса свиней по каждому из 3 пометов

Без применения антибиотиков

Доза B_{12} (в мг/фунт)	Пометы			Сумма	Средний
	1	2	3		
5	1,26	1,21	1,19	3,66	1,22
10	1,29	1,23	1,23	3,75	1,25
15	1,38	1,27	1,22	3,87	1,29
Сумма	3,93	3,71	3,64	11,28	1,25

Поправочный член: $11,28^2 \cdot 9 = 14,1376$.

Общее: $1,26^2 + 1,29^2 + \dots + 1,22^2 - C = 14,1634 - 14,1376 = 0,0258$.

Пометы: $\frac{3,93^2 + 3,71^2 + 3,64^2}{3} - C = 14,1529 - 14,1376 = 0,0153$.

Варианты: $\frac{3,66^2 + 3,75^2 + 3,87^2}{3} - C = 14,1450 - 14,1376 = 0,0074$.

Различия = Остаток = $0,0258 - (0,0153 + 0,0074) = 0,0031$.

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Пометы	2	0,0153	0,0076
Варианты	2	0,0074	0,0037
Различия (ошибка)	4	0,0031	0,0008
Общее	8	0,0258	

В случае же фактического опыта, когда сказываются влияния вариантов и пометов, соответствующие средние квадраты возрастут, но средний квадрат «различий» будет по-прежнему оценивать σ^2 ; обычно его называют *ошибкой опыта*. Обоснование этого названия будет дано позднее.

Вернемся к опыту с витамином B_{12} ; оценка существенности главных эффектов здесь производится при помощи критерия F . Для «пометов» $F = 0,0076/0,0008 = 9,5$, $f = 2$ и 4, $P = 0,04$; для «вариантов» $F = 4,6$; $P > 0,05$. Ясно, что стоило затратить труд на элиминирование различий по пометам;

А словный опыт с 5 вариантами, распределенными внутри 4 пометов

Данные взяты из таблицы 15. $\sigma^2 = 100$.

Варианты	Пометы				Сумма	Средняя
	1	2	3	4		
A	15	31	20	30	96	24
B	22	14	45	26	104	26
C	33	37	30	44	144	36
D	18	31	49	34	132	33
E	37	30	36	21	124	31
Сумма	125	140	180	155	600	

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Варианты	4	392	98
Пометы	3	330	110
Различия	12	1212	101
Общее	19	1934	101,8

эффективность этого построения опыта по сравнению с опытом при полной рендомизации главы 10 будет определена в параграфе 6 этой главы. В данном случае варианты не оказали влияния; известно, что 5 мг витамина повышают привес свиней, но влияние дополнительных доз его остается под вопросом. Дополнительные сведения об этом опыте см. в примерах 5 и 6 этой главы.

Пример 1. У трех видов citrusовых деревьев было определено при трех условиях затенения отношение листовой поверхности к сухому весу листьев [17].

Степень затенения	Апельсин Шамути	Грейфрут Марш	Мандарин Клементин
На солнце	112	90	123
Наполовину затенение	86	73	89
В тени	80	62	81

Проведите дисперсионный анализ. *Ответ:* средние квадраты для «затенения» и ошибки равны 942,1 и 21,8. Затенение оказывает явное влияние на уменьшение изучаемой переменной величины (см. пример 25 этой главы).

Пример 2. Опыт с вирусом мозаики в параграфе 9 главы 2 является опытом с двойной группировкой при 8-кратной повторности и с двумя вариантами. Проведите дисперсионный анализ и получите следующие данные:

варианты	1	64	64
повторения	7	575	82,2
ошибка	7	65	9,29

Компоненты вариантов явно существенны ($F=6,89$; $f=1$ и 7; $P=0,04$). Значимость компонента повторений уже обезждалась, и ее можно было предвидеть. Отметить, что при $f_1=1$; $\sqrt{F}=\sqrt{6,89}=2,63 \approx 1$.

Пример 3. Вычислите средний квадрат для вариантов во каждой из частных формул для двух выборок, приведенных в параграфе 9 главы 10.

Пример 4. Вычислите средний квадрат для различий по среднему квадрату разностей, $s^2=18,57$ из таблицы 12.

3. Сравнения средних. Здесь дается небольшое добавление к тому, что говорилось по этому поводу в параграфах 6 и 8 главы 10. Оба вида сравнений, описанных в этих параграфах, можно применить в приводимом в таблице 114 опыте на овощной опытной станции Южной Каролины [19], который является частью коллективного опыта по изучению способов протравливания семян. Намечаемое сравнение между обработанными и необработанными семенами, т. е. сравнение действия химикатов и контроля, проводится путем вычислений по методу выборок неодинакового размера (параграф 16 главы 10).

ТАБЛИЦА 114
Число взойдящих растений из 100 высеванных семян сои. Сорт Канро

Варианты	Повторения					Сумма	Средний
	1	2	3	4	5		
Контроль	92	90	88	87	89	446	89,2
Арасан	98	94	93	89	95	469	93,8
Свергол	96	90	91	92	90	459	91,8
Симесан	97	95	91	90	94	467	93,4
Фермейт	91	93	95	95	97	471	94,2
Сумма	474	462	458	453	465	2312	

Поправка: $2312^2/25=213813,76$.

Варианты: $\frac{446^2 + \dots + 471^2}{5} - C = 83,84$.

Химикаты с контролем: $\frac{446^2}{5} + \frac{(469 + \dots + 471)^2}{20} - C = 67,24$.

Между химикатами: $83,84 - 67,24 = 16,60$.

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Повторения	4	49,84	
Варианты:			
Химикаты с контролем . . .	1	67,24	67,24
Между химикатами	3	16,60	5,53
Ошибка	16	86,56	5,41

Для сравнения химикатов с контролем: $F=67,24/5,41=12,43$; $F_{0,01}=8,53$.

Особенностью этого сравнения является слагаемость соответствующей ему суммы квадратов с суммой квадратов «между химикатами»: $67,24 + 16,60 = 83,84$. Такие сравнения называются *ортгоналными*. Вы, может быть, помните, что в параграфе 8 главы 10 мы уже встречались с этим свойством.

Следует заметить, что намеченные здесь сравнения оцениваются без учета существенности результатов опыта в целом. Вообще в опыте, в котором варианты в целом не дают существенного значения F , могут быть отдельные существенные сравнения.

В данном опыте не выявлены различия во влиянии четырех химикатов, и поэтому нет надобности в разыскании существенных различий среди 6 сравнений, которые можно было бы произвести при четырех вариантах. Но

все же для иллюстрации вычислим 5%-ную существенную разность D :

$$s_{\bar{x}} = \sqrt{5,41/5} = 1,04 \text{ растения}; \quad Q_{16,4} = 4,05;$$

$$D = Qs_{\bar{x}} = 4,21 \text{ растения.}$$

Как и предполагалось, в опыте нет разностей между средними по химикатам столь большой величины. Этот опыт убеждает в том, что химическая обработка семян приводит в условиях опыта к увеличению всхожести семян сорта Канро, но не доказывает, что между химикатами имеются различия по их эффективности.

Пример 5. По данным таблицы 112 проверьте гипотезу о том, что наименьшая доза витамина B_{12} столь же эффективна, как и низшая. *Ответ:* средний квадрат = 0,00735; $F = 0,04$. Это является дополнительной информацией, не использованной при первоначальной обработке данных.

Пример 6. Дополните предыдущее проверкой гипотезы, что в совокупности средняя промежуточной дозы та же самая, что и объединенная средняя двух других доз. Наиболее простой способ состоит в сопоставлении суммы 3,66 + 3,87 с $2 \times 3,75$. *Ответ:* средний квадрат = 0,00005. На то, что эти два сравнения ортогональны, указывает факт получения суммы квадратов для «вариантов» из таблицы 112 при сложении средних квадратов $0,00735 + 0,00005 = 0,0074$.

Пример 7. Оцените существенность всех разностей между вариантами затенения в примере 1. *Ответ:* $D = 13,5$.

При первом чтении параграфы 4 и 5 этой главы можно опустить.

4. Символика для таблицы с двумя входами. В математической статистике становится общим правилом применение подстрочных точек для обозначения суммирования. У нас это обозначение будет применяться в ограниченных размерах. В помощь тем, кто предпочитает буквенную символику, ниже дается таблица 115.

Конечно, не является случайностью то, что эта книга написана так, чтобы читатель меньше всего зависел от буквенной символики. Результаты большинства опытов выражаются в цифрах, и эти цифры являются тем, в чем экспериментатор усматривает нужные ему сведения. Я здесь стараюсь сохранить связь читателя с цифровым материалом, как можно меньше отвлекая его внимание символическими выражениями.

5. Структура опыта по двойной группировке. Модель, которая здесь применяется (модель I), такова:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}; \quad i = 1, \dots, a; \quad j = 1, \dots, b; \quad \varepsilon_{ij} = N(0, \sigma); \quad \sum \alpha_i = \sum \beta_j = 0.$$

Здесь α_i обозначает эффекты вариантов; β_j — эффекты повторений (блоков или в прошлом — пометов) и ε_{ij} — случайная переменная, представляющая выборочное варьирование.

Наилучшим образом можно понять эту модель путем построения на ее основе искусственного эксперимента. Пусть $\mu = 30$ фунтам.

$$\alpha_1 = 10; \quad \alpha_2 = 3; \quad \alpha_3 = 0; \quad \alpha_4 = -13; \quad \sum \alpha_i = 0;$$

$$\beta_1 = 1; \quad \beta_2 = -4; \quad \beta_3 = 3; \quad \sum \beta_j = 0.$$

ε_{ij} берутся случайно из таблицы 24 при уменьшении каждого показателя на $\mu = 30$ фунтам.

В каждую клетку таблицы 116 сначала вносится $\mu = 30$. После этого присоединяется компонент варианта, различный по своей величине при переходе от одного ряда к другому. Далее вносится компонент блока, одинаковый для каждого отдельного столбца. Сумма этих трех частей в каждой клетке определяется фиксированными значениями μ , α_i и β_j . Выборочное варьирование дает четвертый член — отклонение, взятое по жребию из таблицы 24. Этот элемент имеет нормальное распределение со средней, равной нулю, и дисперсией, равной 25. В соответствии с этой моделью X_{ij} является суммой четырех перечисленных выше членов.

Символическое изображение таблицы с двумя входами при a вариантах и b блоках.
Схема вычислений и дисперсионный анализ

Варианты $i=1, \dots, a$	Блоки $j=1, \dots, b$ $1, \dots, j, \dots, b$	Сумма	Среднее
1	$X_{11} \dots X_{1j} \dots X_{1b}$	$X_{1.}$	$\bar{x}_{1.}$
2	$X_{21} \dots X_{2j} \dots X_{2b}$	$X_{2.}$	$\bar{x}_{2.}$
⋮	⋮	⋮	⋮
i	$X_{i1} \dots X_{ij} \dots X_{ib}$	$X_{i.}$	$\bar{x}_{i.}$
⋮	⋮	⋮	⋮
a	$X_{a1} \dots X_{aj} \dots X_{ab}$	$X_{a.}$	$\bar{x}_{a.}$
Сумма	$X_{.1} \dots X_{.j} \dots X_{.b}$	$X_{..}$	$\bar{x}_{..}$
Средняя	$\bar{x}_{.1} \dots \bar{x}_{.j} \dots \bar{x}_{.b}$		

Поправка: $C = (\sum X_{ij})^2 / ab = X_{..}^2 / ab$.

Общая: $\sum X_{ij}^2 - C$.

Вариантов: $A = \frac{X_{1.}^2 + \dots + X_{a.}^2}{b} - C$.

Блоков: $B = \frac{X_{.1}^2 + \dots + X_{.b}^2}{a} - C$.

Различия: $D = \text{общая} - (\text{варианты} + \text{блоки})$.

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Варианты	$a - 1$	A	$A / (a - 1)$
Блоки	$b - 1$	B	$B / (b - 1)$
Различия	$(a - 1)(b - 1)$	D	$D / (a - 1)(b - 1)$
Общее	$ab - 1$	$A + B + D$	

Теперь становятся очевидными некоторые особенности этой модели. 1) Главные эффекты вариантов не влияют на β_j , так как сумма этих последних в каждом ряду равна нулю; если оставить в стороне ε_{ij} , то сумма наблюдений, например для варианта 1, будет $41 + 36 + 43 = 120$, а средняя $40 = \mu + \alpha_1$. Но наличие двух слишком больших отклонений -11 и -7 для величин ε_{1j} снизило наблюдаемую среднюю до 35. 2) В блоке 3 случилось так, что не только $\sum \alpha_i = 0$, но и $\sum \varepsilon_{i3} = 0$. Следовательно, эта средняя имеет значение фиксированной суммы $\mu + \beta_3 = 33$. 3) Так как по всему опыту $\sum \varepsilon_{ij} = -12$, то общая средняя меньше $\mu = 30$ на среднее значение ошибки $\varepsilon_{ij} = -12/12 = -1$. 4) Главное, что надо помнить, это то, что α_i балансируются в каждом блоке и β_j в каждой строке, вследствие чего один главный эффект не влияет на другой (хотя оба они находятся под влиянием ε_{ij}). 5) Слово бросается в глаза слабость сумм квадратов и степеней свободы. Данная модель не может применяться, если показатели вариантов меняются, например пропорционально уровням блоков.

В приведенном выше дисперсионном анализе представляют интерес два момента: 1) Средний квадрат «ошибки» 22 является выборочной оценкой $\sigma^2 = 25$. 2) Средние квадраты «блоков» и «вариантов» увеличены за счет β_j и

Опыт, сконструированный в соответствии с моделью I. $\mu=30$

Вариант	Блок			$X_{i.}$	$\bar{x}_{i.}$			
	$\beta_1=1$	$\beta_2=-1$	$\beta_3=3$					
$\alpha_1=10$	30 10 1 -11	30 10 -4 -7	30 10 3 3	$X_{11}=30$	$X_{12}=29$	$X_{13}=46$	105	35
$\alpha_2=3$	30 3 1 1	30 3 -4 5	30 3 3 -3	$X_{21}=35$	$X_{22}=34$	$X_{23}=33$	102	34
$\alpha_3=0$	30 0 1 0	30 0 -4 4	30 0 3 -1	$X_{31}=31$	$X_{32}=30$	$X_{33}=32$	93	31
$\alpha_4=-13$	30 -13 1 2	30 -13 -4 -2	30 -13 3 1	$X_{41}=16$	$X_{42}=11$	$X_{43}=21$	48	16
$X_{.j}$	112	104	132	348				
$\bar{x}_{.j}$	28	26	33					29
Источник варьирования		Число степеней свободы	Сумма квадратов	Средний квадрат				
Блоки		2	104	52				
Варианты		3	702	234				
Ошибка		6	132	22				

z_j ; каково это увеличение, показывает компонентный анализ в таблице 117, где параметрические величины и средние квадраты для соответствующих компонентов вычислены так, чтобы они были сравнимы с выборочной ошибкой.

В резком контрасте с анализом главы 10 здесь находится то, что ни один из этих компонентов не входит в главный эффект другого фактора, т. е. α_A не входит в средний квадрат блоков, а α_B в средний квадрат вариантов. Это остается верным также для модели II и для смешанной модели; действительно, следует только заменить стоящие у нас буквы «канна» на соответствующие «сигма».

Резюме и р е з у л т а т ы: Мы сконструировали опыт по модели I с двумя рядами фиксированных компонентов и случайной ошибкой. Анализ этих данных позволяет произвести оценку как фиксированных, так и случайных частей и сравнить их с известными в нашем случае параметрами. В целом

Компонентный анализ искусственного опыта

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Блоки	2	52	$\sigma^2 + a\chi_B^2$
Варианты	3	234	$\sigma^2 + b\chi_A^2$
Ошибка	6	22	σ^2

$$\chi_B^2 = \frac{\Sigma \beta_j^2}{b-1} = \frac{1^2 + (-4)^2 + 3^2}{2} = 13.$$

$$\chi_A^2 = \frac{\Sigma \alpha_i^2}{a-1} = \frac{10^2 + 3^2 + 0^2 + (-13)^2}{3} = 92^2/3.$$

$s_B^2 = (52 - 22)/4 = 8$ является оценкой 13; $s_A^2 = (234 - 22)/3 = 71$ является оценкой $92^2/3$.

Средний квадрат ошибки = 22 оценивает 25.
Средний квадрат блоков = 52 оценивает $25 + 4 \times 13 = 77$.
Средний квадрат вариантов = 234 оценивает $25 + 3 \times 93 = 304$.

этот процесс дает представление о подобного вида приближенных оценках, получаемых из фактического опыта, параметры которого поддаются оценке, но (в связи с ошибками ϵ_{ij}) никогда точно не известны.

Аналогично этому можно подобрать такую же линейную модель, но уже по экспериментальным данным. Теперь параметры только оцениваются; сами же они остаются неизвестными. Процесс анализа в дальнейшем будет углублен, а в настоящее время основное значение его сводится к показу только того, как образуются «различия».

ТАБЛИЦА 118

Высота (в см) растений сои в течение 5 недель
Четыре блока или повторений

Блок	Недели					Сумма Σx_i	Средняя \bar{x}_i	Отклонение $\bar{x}_i - \bar{x}_{..}$
	1	2	3	4	5			
1	4	18	26	38	44	130	26	2
2	3	19	25	35	43	125	25	1
3	6	18	24	28	39	115	23	-1
4	7	13	21	31	38	110	22	-2
Сумма $\Sigma_{.j}$	20	68	96	132	164	480		
Средняя $\bar{x}_{.j}$	5	17	24	33	41	$\bar{x}_{..} = 24$		
Отклонение $\bar{x}_{.j} - \bar{x}_{..}$	-19	-7	0	9	17			

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	19	3 250	
Недели	4	3 120	780
Блоки	3	50	16,67
Остаток (различия)	12	80	6,67

Данные таблицы 118 взяты из вегетационного опыта по выращиванию сои. Для облегчения вычислений произведены небольшие изменения дан-

ных. В окаймляющих графах таблицы приведены суммы, средние и отклонения этих средних от общей средней опыта $\bar{x}_{..} = 24$ сантиметрам.

В таблице 119 произведен подбор линейной модели:

$$\hat{X}_{ij} = \bar{x}_{..} + (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}).$$

Здесь в каждой клетке берется общая средняя 24, к ней добавляется отклонение средней строки и отклонение средней столбца от этой общей средней. Сумма \hat{X}_{ij} подписывается непосредственно под наблюдаемой высотой растения X_{ij} .

ТАБЛИЦА 119

Линейная модель, подобранная по выборочным данным. Высота растений сои

Блок	Неделя						
	1	2	3	4	5		
1	X_{1j}	4	18	26	38	44	
	$\bar{x}_{1.}$	24	24	24	24	24	
	$\bar{x}_{1.} - \bar{x}_{..}$	2	2	2	2	2	
	$\bar{x}_{.j} - \bar{x}_{..}$	-19	7 ¹	-7	19	0	26
	\hat{X}_{1j}	-19	7	0	26	9	35
d_{1j}	-3	-1	0	3	1		
2	X_{2j}	3	19	25	35	43	
	$\bar{x}_{2.}$	24	24	24	24	24	
	$\bar{x}_{2.} - \bar{x}_{..}$	1	1	1	1	1	
	$\bar{x}_{.j} - \bar{x}_{..}$	-19	6	-7	18	0	25
	\hat{X}_{2j}	-19	6	-7	18	9	34
d_{2j}	-3	1	0	1	1		
3	X_{3j}	6	18	24	28	39	
	$\bar{x}_{3.}$	24	24	24	24	24	
	$\bar{x}_{3.} - \bar{x}_{..}$	-1	-1	-1	-1	-1	
	$\bar{x}_{.j} - \bar{x}_{..}$	-19	4	-7	16	0	23
	\hat{X}_{3j}	-19	4	-7	16	9	32
d_{3j}	2	2	1	-4	-1		
4	X_{4j}	7	13	21	31	38	
	$\bar{x}_{4.}$	24	24	24	24	24	
	$\bar{x}_{4.} - \bar{x}_{..}$	-2	-2	-2	-2	-2	
	$\bar{x}_{.j} - \bar{x}_{..}$	-19	3	-7	15	0	22
	\hat{X}_{4j}	-19	3	-7	15	9	31
d_{4j}	4	-2	-1	0	-1		

¹ $\hat{X}_{11} = \bar{x}_{..} + (\bar{x}_{1.} - \bar{x}_{..}) + (\bar{x}_{.1} - \bar{x}_{..}) = 24 + 2 - 19 = 7$.

Суммы \hat{X}_{ij} являются высотами, которые растения при данной линейной модели имели, если бы не было никакого дополнительного случайного варьирования. Если вы проведете дисперсионный анализ в отношении \hat{X}_{ij} , то найдете, что «различия» равны нулю. Разность между \hat{X}_{ij} в любой паре столбцов будет одной и той же для каждого ряда. Например, разность для столбцов 1 и 3:

$$\hat{X}_{11} - \hat{X}_{13} = 7 - 26 = -19 = \hat{X}_{21} - \hat{X}_{23} = \hat{X}_{31} - \hat{X}_{33} = \hat{X}_{41} - \hat{X}_{43}.$$

То же самое относится и к парам рядов. Этот факт имеет полезное практическое применение. Каждый может легко определить размер расхождений в таблице с двумя входами путем просмотра поперек таблицы пары рядов и вдоль таблицы пары столбцов, наблюдая нарушения равенства разностей между парами рядов или столбцов.

В последней строке каждого блока в таблице указаны разности:

$$d_{ij} = X_{ij} - \bar{X}_{ij}$$

Эти расхождения между наблюдаемыми и вычисленными высотами растений обладают таким замечательным свойством:

$$\sum d_{ij}^2 = \text{сумме квадратов различий.}$$

Для проверки вычислим: $(-3)^2 + (-1)^2 + \dots + (-1)^2 = 80$, что совпадает с «различиями» в таблице 118. Это делает наглядным тот факт, что средний квадрат этих различий оценивает σ^2 в такой таблице, как таблица 116. Заметим, что величины d_{ij} в выборке аналогичны величинам ϵ_{ij} в совокупности.

6. Рандомизированные блоки в полевом опыте. Описанное в предыдущем построение опыта находит широкое применение в полевом экспериментировании. Накопленный опыт подсказывает экспериментатору-агроному,

Блок I	Г	29,3	Блок II	Б	33,0	Блок V	Г	28,8
	Б	33,3		А	34,0		В	35,8
	В	30,8		В	34,3		Б	34,5
	А	32,3		Г	26,0		А	38,5
Блок III	Г	29,8	Блок IV	Б	36,8			
	А	34,3		А	35,0			
	Б	36,3		Г	28,0			
	В	35,3		В	32,3			

Рис. 23. План полевого опыта с четырьмя сортами А, Б, В и Г гапшинольцевской пшеницы. Урожай в фунтах на одну деланку.

что различия плодородия почвы увеличиваются с расстоянием. В связи с этим он испытывает все свои варианты в блоке, деланки которого находятся друг с другом в возможно близком соседстве, и повторяет эти блоки несколько раз. В этом случае различия между блоками не входят в ошибку опыта.

Каждый блок подразделяется на деланки по числу вариантов. Теперь перейдем к сущности этого построения опыта. Так как соседние деланки имеют большую вероятность попасть на одинаково плодородную почву, причем это сходство условий уменьшается с расстоянием, то урожай соседних деланок не будет независимыми. Для преодоления этого затруднения необходима рандомизация. Расположение вариантов по длине преследует также и другие цели, среди которых выделяется необходимость выравнивания ошибки по всем разностям вариантов. В каждом блоке местоположение вариантов должно быть установлено в случайном порядке.

Для получения оценки экспериментальной ошибки необходимо иметь два или более блока (повторений). Точность опыта увеличивается по мере возрастания числа повторений.

На рисунке 23 изображен план участка, на котором проводилось испытание четырех сортов пшеницы в пяти блоках [13]. Узкая форма деланок, вытянутых по длине блоков, содействует повышению точности опыта. Причина этому та, что при переходе от одного конца блока к другому может происходить изменение в плодородии почвы; длинными узкими деланками имеется возможность в одинаковой мере перекрыть его и более или менее рав-

номерно распределить эти различия между собой. Однако многие экспериментаторы не придерживаются этой практики. Если делянки расположены попеременно блоками, то две делянки одного блока могут иногда различаться больше, чем соответствующие делянки в соседних блоках. В результате этого может получиться так, что опыт с рендомизированными блоками окажется менее эффективным, чем простое рендомизированное размещение вариантов, подобное описанному в главе 10. Цель же данного построения опыта состоит в том, чтобы делянки внутри блока были бы по возможности одинаковы. Максимальные различия должны наблюдаться при переходе от блока к блоку, они будут элиминированы из ошибки. (О применении квадратных делянок см. в следующем параграфе.)

Данные этого опыта вместе с соответствующим дисперсионным анализом приведены в таблице 120. Различия между сортами вполне убедительны. Применение метода параграфа 6 главы 10 дает $D=2,7$ фунта, откуда следует, что четвертый сорт низкоурожайный, остальные же три не отличаются друг от друга по урожайности.

ТАБЛИЦА 120

Урожай (фунты на делянку) четырех сортов пшеницы, выращенной в пяти рендомизированных повторениях

Сорта	Блоки					Сумма	Средняя
	1	2	3	4	5		
А	32,3	34,0	34,3	35,0	36,5	172,1	34,4
Б	33,3	33,0	36,3	36,8	34,5	173,9	34,8
В	30,8	34,3	35,3	32,3	35,8	168,5	33,7
Г	29,3	26,0	29,8	28,0	28,8	141,9	28,4
Сумма	125,7	127,3	135,7	132,1	135,6	656,4	32,8
Источник варьирования		Число степеней свободы		Сумма квадратов		Средний квадрат	
Блоки		4		21,46		5,36	
Сорта		3		134,45		44,82	
Ошибка		12		26,26		2,19	
Для сортов				$F = 44,82/2,19 = 20$, $F_{0,01} = 5,95$			

Отметим, что коэффициент вариации $C=2,19/32,8$ составляет только 4,5%, что указывает на довольно высокую точность технической стороны этого полевого опыта. Данный факт нашел свое отражение и в малой величине существенной разности 2,7 фунта, что составляет только 8% от средней.

Эффективность опыта с рендомизированными блоками по сравнению с опытом при полностью случайном расположении делянок определяется обратным отношением их оценок ошибки опыта; чем меньше ошибка, тем более эффективен опыт.

Предполагая, что эти два построения опыта наложены на одни и те же делянки, можно определить относительную эффективность метода блоков (8.14):

$$\frac{(b-1)M_B + b(a-1)M_E}{(ab-1)M_E}$$

Здесь средние квадраты M_B и M_E берутся из данных фактического опыта с рендомизированными блоками. Взяв для иллюстрации данные таблицы 120,

имеем $M_B=5,36$, $M_E=2,19$. $b=5$, $a=4$. отсюда:

$$\text{Относительная эффективность} = \frac{s_R^2}{s_B^2} = \frac{(5-1) \times 5,36 - 5 \times (4-1) \times 2,19}{(20-1) \times 2,19} = 1,30.$$

Значит $s_R^2/s_B^2=1,30$, где символы R и B относятся к случайному и блоковому расположениям. Для получения одной и той же стандартной ошибки средней необходимо n и b выбрать так, чтобы

$$\frac{s_R^2}{n} = \frac{s_B^2}{b}.$$

Решая это уравнение по n , мы находим повторность полностью рандомизированного опыта, необходимую для достижения уровня эффективности опыта с b блоками

$$n = \frac{bs_R^2}{s_B^2} = 1,30b.$$

По данным опыта с пшеницей имеем $n=1,30 \times 5=6,5$ — повторность опыта при полной рандомизации, эквивалентная 5-кратной повторности при схеме рандомизированных блоков. Это означает, что при полной рандомизации потребовалось бы на 30% больше деценок, или на 30% больше денежных расходов на опыт.

Эффективность метода блоков здесь несколько завышена в связи с потерей степеней свободы, так как их число 16 при полном рандомизированном расположении вариантов и 12 в опыте с блоками. Это учтено в формуле, которую дал Фишер [10]. Относительное количество информации $= \frac{(j_B+1)(j_R+3)s_R^2}{(j_B+3)(j_R+1)s_B^2} = \frac{(12+1)(16+3)}{(12+3)(16+1)} < 1,30 = 126\%$.

Формула содержит множитель для отношения s_R^2/s_B^2 , определяемый по степеням свободы. Этот множитель и приводит к окончательной формуле относительной эффективности.

Пример 8. Ниже приводятся урожай (бушелл на акр) 7 линий соев, взятые из более обширного опыта в Гудзоне, Айова [9].

Линии	Блок			
	1	2	3	4
A	21,3	19,7	28,7	27,3
B	27,7	15,9	28,1	29,0
C	22,0	22,1	22,0	26,9
D	28,3	20,7	26,0	34,1
E	25,1	20,1	24,9	29,8
F	28,1	19,4	31,5	28,7
G	32,7	26,7	34,2	35,6

Проведите дисперсионный анализ:

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Повторение	3	347,78	
Варианты	6	215,47	35,91
Ошибки	18	111,75	6,21

Пример 9. Показать, что в опыте с соей линия G существенно отличается от A, B, C и E. *Ответ:* $D=5,6$.

Пример 10. Определите 95%-ые пределы для разности δ между линиями G и F. *Ответ:* $-0,2 \leq \delta \leq 11,0$. Если бы это потребовалось, я без всяких колебаний принял бы эту разность в качестве характеристики совокупности.

Пример 11. Вычислите в опыте с соей коэффициент вариации. *Ответ:* 9,5%.

7. Латинский квадрат. В опыте с рендомизированными блоками делаянки блока должны быть по возможности сходными друг с другом; каждая делаянка должна включать в себя все варьирование условий внутри блока. Однако часто делаянки не могут быть расположены в соответствии с этим условием. Некоторые варианты требуют, чтобы делаянки были почти квадратной формы, в связи с чем изменения плодородия внутри блоков становятся такими, что в значительной мере уменьшается эффективность опыта. По этой и по другим причинам может быть выгодным расположить делаянки в столбцы попереки блоков и расположить варианты так, чтобы каждый из них встречался один раз в каждом столбце и один раз в каждом блоке. В этом случае столбцов и блоков требуется столько, сколько в опыте вариантов. Такая схема опыта называется латинским квадратом. Например, для 4 вариантов *A*, *B*, *C* и *D* он может быть таким:

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>B</i>	<i>C</i>	<i>D</i>	<i>A</i>
<i>C</i>	<i>D</i>	<i>A</i>	<i>B</i>
<i>D</i>	<i>A</i>	<i>B</i>	<i>C</i>

Здесь при оценке каждой средней по вариантам опыта все блоки и столбцы представлены в одинаковой степени. В процессе дисперсионного анализа различия между столбцами, подобно различиям между рядами (блоками), исключаются из ошибки. Применение латинского квадрата в сельскохозяйственных опытах приводит обычно к повышению их эффективности, которое означает, что для достижения той же точности требуется меньшая повторность опыта или достигается более высокая точность опыта при данном числе повторений.

Модель опыта по схеме латинского квадрата (модель I) такова:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}; \quad i, j, k = 1, \dots, a; \quad \epsilon_{ijk} = N(0, \sigma),$$

где α , β и γ относятся к эффектам вариантов, рядов и столбцов при обычных ограничениях.

В таблице 121 приведено расположение на поле и урожай опыта по изучению ширины междурядий при посеве проса [15]. В дополнение к суммам по вариантам здесь вычисляются суммы по рядам и столбцам. В соответствии с этим в таблицу дисперсионного анализа вводится новый эффект («столбцы»). Оценка ошибки определяется путем вычитания из «общего» этих трех главных эффектов.

Компонентный анализ здесь также включает в себя только одно добавление к тому, что было при анализе при рендомизированных блоках. В таблице 122 я сделал допущение о соответствии этого опыта смешанной модели и считал, что данные по столбцам и рядам являются случайными переменными, а эффекты вариантов фиксированы, что нашло отражение только в обозначениях σ^2 и κ^2 .

Многочисленные возможности, кроме условий полевого опыта, эффективного использования схемы латинского квадрата возникают, когда опыт проводится на фоне двух переменных, независимость которых заранее известна экспериментатору. Например, в опытах по кормлению животных влияние таких факторов, как помет и условия содержания или способ разведения и условия содержания скота, может быть этим путем исключено из «вариантов» и «ошибки». Затруднения, возникающие в связи с невозможностью изучения некоторой последовательности вариантов на одном и том же объекте, можно отчасти избежать, пользуясь схемой латинского квадрата в качестве способа, устанавливающего полную случайность в этой последовательности вариантов. Эта схема может обеспечить, например, сбалансирование вариантов и объектов в течение определенной последовательности периодов времени (пример 14). Как и в случае полевого опыта, схема латинского квадрата

Урожай (в г) деленок проса, расположенных латинским квадратом.
 Расстояния между рядами: А—2, В—4, С—6, D—8 и Е—10 дюймов

Ряды	Столбцы					Сумма
	1	2	3	4	5	
1	В: 257	Е: 230	А: 279	С: 287	Д: 202	1255
2	Д: 245	А: 283	Е: 245	В: 280	С: 260	1313
3	Е: 482	В: 252	С: 280	Д: 246	А: 250	1210
4	А: 203	С: 204	Д: 227	Е: 493	В: 259	1086
5	С: 231	Д: 271	В: 266	А: 334	Е: 338	1440
Сумма	1118	1240	1297	1340	1309	6304
Суммирование по расстояниям						
А: 2 дюйма В: 4 дюйма С: 6 дюймов D: 8 дюймов Е: 10 дюймов						
Сумма	1349	1314	1262	1194	1188	6304
Средняя	269,8	262,8	252,4	238,2	237,6	252,2

Поправка: $6304^2/25 = 1589617$.

Общее: $257^2 + \dots + 338^2 \dots 1589617 = 36571$.

Рядов: $\frac{1255^2 + \dots + 1440^2}{5} - 1589617 = 13601$.

Столбцов: $\frac{1118^2 + \dots + 1309^2}{5} - 1589617 = 6146$.

Расстояний: $\frac{1349^2 + \dots + 1188^2}{5} - 1589617 = 4156$.

Остаток: 12668.

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	24	36571	3400
Ряды	4	13601	1536
Столбцы	4	6146	1039
Расстояния	4	4156	1039
Ошибка	12	12668	1056

ТАБЛИЦА 122

Компонентный анализ при латинском квадрате

Источник варьирования	Число степеней свободы	Средний квадрат	Компоненты
Ряды В	$a - 1$	M_B	$\sigma^2 + a\sigma_B^2$
Столбцы С	$a - 1$	M_C	$\sigma^2 + a\sigma_C^2$
Варианты А	$a - 1$	M_A	$\sigma^2 + az_A^2$
Ошибка	$(a - 1)(a - 2)$	M_E	σ^2

позволяет перекрывать в двух направлениях различия в лабораторных условиях или на стеллажах теплицы.

Упрощенное правило рендомизации в латинском квадрате следующее: выписав какое-нибудь систематическое размещение букв, удовлетворяющее условиям латинского квадрата, производим перераспределение по жребью рядов и столбцов, после чего опять по жребью производится отождествление вариантов с буквами. Более обоснованный способ см. у Фишера и Нейтса [12].

Знакомясь со средними квадратами таблицы 121, можно прийти к заключению, что расстояния между рядами растений не оказывают никакого влияния на урожай; варьирование урожая такое, которое можно ожидать в выборках из одной нормальной совокупности урожаев. Но данная схема опыта содержит в себе некоторый момент, который пока не был выявлен в процессе анализа, а именно регрессию урожая на величину ширины междурядья. Результаты опыта таблицы 121 указывают на то, что урожай снижается по мере увеличения ширины междурядий. В связи с этим очевидно, что сумма квадратов для вариантов 4156 содержит в себе значительную часть, относящуюся к линейной регрессии, и что соответствующее разложение этой суммы квадратов может дать дополнительные сведения об изучаемом явлении. Для этого, согласно главе 6, нам необходимо иметь $\sum x^2$, $\sum xy$ и $\sum y^2$. Последняя сумма уже вычислена — 4156. Две другие суммы удобнее всего вычислить по средним, после чего для перехода к основным единицам (в которых выражены суммы квадратов и средние квадраты) следует их умножить на $a=5$:

$$\sum x^2/5 = 2^2 + \dots + 10^2 = 30^2/5 = 40;$$

$$\sum xy/5 = 2 \times 269,8 + \dots + 10 \times 237,6 = \frac{30 \times 1260,8}{5} = -178.$$

Следовательно, искомые суммы квадратов и сумма произведений будут 200 и —890. Теперь сумма $\sum y^2$ подразделяется на две части: 1) на часть, относящуюся к регрессии $(\sum xy)^2/\sum x^2 = (-890)^2/200 = 3960$, и 2) на часть, характеризующую отклонения от регрессии $4156 - 3960 = 196$. Результаты такого пересмотренного дисперсионного анализа (опуская данные, относящиеся к рядам и столбцам) приведены в таблице 123.

ТАБЛИЦА 123

Анализ регрессии среднего урожая на величину междурядья
Опыт с просом

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Регрессия	1	3960	3960
Отклонения	3	196	65,7
Ошибка (табл. 121)	12		1056

Для регрессии $F = 3960/1056 = 3,75$; $F_{0,05} = 4,75$

Теперь ясно, что большая часть варьирования урожая связана с регрессией, отклонения же от регрессии ничтожны. Иными словами, главная часть суммы квадратов вариантов связана с единичной степенью свободы, относящейся к регрессии. Однако даже при этих условиях регрессия не является существенной, т. е. в совокупности нет никакой связи между урожаем и величиной междурядья. Тем не менее представляет интерес тот факт, что в этом опыте имеется значительное увеличение урожая на одно растение, при переходе к широким междурядьям, обусловленное тем, что устранение конкуренции между растениями полностью компенсирует уменьшение числа их на этих делянках. Отметим и величину коэффициента вариа-

ции $C = \sqrt{1056}/252,2 = 13\%$, которая может объяснить несущественность довольно хорошо выраженной выборочной регрессии.

В связи с тем что число повторений в латинском квадрате равно числу вариантов, экспериментатор, если он хочет применить этот метод, обычно принужден ограничить себя изучением 8 или 10 вариантов. При 4 и меньшем числе вариантов число степеней свободы для ошибки становится меньше допустимого, например при квадрате 4×4 оно равно только $(a-1)(a-2) = 3 \times 2 = 6$. Это затруднение можно преодолеть путем повторения таких квадратов (пример 15 этой главы).

Относительная эффективность опыта по схеме латинского квадрата по сравнению с опытом при полной рандомизации определяется формулой:

$$\frac{M_B + M_C + (a-1)M_E}{(a+1)M_E}$$

Подстановка сюда данных опыта с просом дает:

$$\text{Относительная эффективность} = \frac{s_R^2}{s_L^2} = \frac{3400 + 1536 + (5-1) \times 1056}{(5+1) \times 1056} = 145\%$$

что на 45% превышает эффективность полной рандомизации.

Представляет некоторый интерес и определение относительной эффективности по сравнению с опытом при рандомизированных блоках, в котором не используются ряды или столбцы. В опыте с просом в связи с тем, что средний квадрат столбцов довольно мал (он может быть приписан просто случайному выборочному варьированию), его можно опустить и рассматривать ряды в качестве блоков. Относительная эффективность латинского квадрата в этом случае будет:

$$\frac{M_C + (a-1)M_E}{aM_E} = \frac{1536 + (5-1) \times 1056}{5 \times 1056} = 109\%$$

Кемпторн [14] предупреждает, что такое сравнение может быть не совсем правильным. В опыте с рандомизированными блоками форма делянок, по всей вероятности, будет иной и более подходящей, что приведет к повышению эффективности такого опыта. Например, в данном опыте с просом выбор надлежащей формы делянок может вполне компенсировать всю выгоду от учета столбцов.

В опытах, в которых нет заранее намеченных сравнений, возможно сопоставление всех средних между собой так, как указано в параграфе 6 главы 10.

Пример 12. Ниже приводится латинский квадрат, составленный из чисел, облегчающих вычисления. Варианты обозначены *A*, *B* и *C*.

Ряды	Столбцы		
	1	2	3
1	<i>B</i> : 23	<i>A</i> : 17	<i>C</i> : 29
2	<i>A</i> : 16	<i>C</i> : 25	<i>B</i> : 16
3	<i>C</i> : 24	<i>B</i> : 18	<i>A</i> : 12

Средние квадраты таковы: для рядов 21, для столбцов 3, для вариантов 93 и остаток 3.

Пример 13. Подберите линейную модель для латинского квадрата по данным примера 12. Проверьте это при помощи равенства $\sum d_{ijk}^2 = 2 \times 3 = 6$.

Пример 14. В опытах по изучению молочной продуктивности коров приходится сталкиваться с большим варьированием индивидуальных особенностей животных, что требует большого числа животных для установления даже умеренных различий между вариантами. Возможности применения нескольких вариантов последовательно в отношении одной и той же коровы ограничены в связи с уменьшением удоев коров с течением

времени, формой лактационной кривой, влиянием прогулок и возможной корреляцией между ошибками ϵ_{ijk} . Была сделана попытка преодолеть эти затруднения применением нескольких пар ортогональных латинских квадратов [7], столбцы которых представлены отдельными коровами, а ряды — последовательными периодами в течение одной лактации; вариантами были: *A* — грубые корма; *B* — ограниченное количество зерна; *C* — полное количество зерна.

В этом примере представлен один из таких латинских квадратов, в котором не исключено влияние прогулок. В некоторых последних своих исследованиях Паттерсон [18] и Лукас [16] изучали нелинейное влияние прогулок. Если существует такое влияние, то ошибка прямого эффекта будет смещенной в сторону уменьшения. Можно вычислить поправки на нелинейность, но Лукас замечает, что они не имеют практического значения в работах с молочными коровами.

Периоды	Коровы		
	1	2	3
I	A: 608	B: 885	C: 940
II	B: 715	C: 1087	A: 766
III	C: 844	A: 711	B: 832

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Периоды	2	5 900	2 950
Коровы	2	47 214	23 607
Варианты	2	103 436	51 718
Ошибка	2	4 843	2 422

Пример 15. Один квадрат 3×3 с двумя степенями свободы для характеристики ошибки опыта явно недостаточен. Кокран и другие опубликовали результаты анализа данных об общей усвояемости потребленных кормов (игнорируя влияние прогулок), полученные в опыте с 5 такими латинскими квадратами:

Источник варьирования	Число степеней свободы	Средний квадрат
Квадраты, или группы	4	39 484
Ряды (периоды) внутри квадратов	10	4 053
Столбцы (коровы) внутри квадратов	10	10 534
Рацнаны	2	266 934
Обобщенная ошибка	10	1 440
Остаток	8	1 503

Указанный здесь остаток является взаимодействием между «рационами» и «группами», что будет объяснено в главе 12.

8. Размер опыта. Методы параграфа 18 главы 10 непосредственно применимы к опытам, проведенным по схеме рендомизированных блоков и латинских квадратов. В качестве иллюстрации можно взять данные примера 9 этой главы. Разность, которую можно здесь обнаружить, составляла $D/x_{..} = 5,6/26,3 = 21\%$. Допустим, что требуется произвести испытание первых 6 линий сои в новом опыте с рендомизированными блоками, который должен ($P=0,75$) обнаруживать ($P=0,05$) разности, составляющие 10% от средней; 10% от 25,3 дает 2,5 бушеля на акр. Таким образом, $\delta=2,5$, а также для вариантов $a=6$, $s_0 = \sqrt{6,21} = 2,49$ бушеля на акр и $f_0=18$. Возьмем сначала завышенное число блоков $b=31$. Тогда $f=(6-1)(31-1)=150$; $Q_{6,150}=4,09$

и $F_{150,18} = 1,33$. Итак, предварительная оценка будет:

$$b = \frac{4,09^2 \times 6,21 \times 1,33}{2,5^2} = 23.$$

Это число повторений при $f = (6-1)(23-1) = 110$, $Q_{6,110} = 4,11$ и $F_{110,18} = 1,33$ позволяет обнаруживать разности порядка:

$$\delta = \frac{4,11 \times 2,49 \times \sqrt{1,33}}{1,23} = 2,5,$$

что полностью соответствует заданию.

Вряд ли можно представить себе осуществление опыта с 23 повторениями в одно время и в одном месте. Действительно, если, например, изучаются перспективные сорта, то они, конечно, будут испытываться в небольших опытах, распределенных в нескольких местах и в течение нескольких лет.

Допустим, что планируется испытание 6 сортов методом рендомизированных блоков в 4 пунктах в течение нескольких лет. В каждом месте и в каждый следующий год решено закладывать 3 повторения. В этом случае представляют интерес два вопроса. Во-первых, какая процентная разность (коэффициент вариации) будет с заданным шансом (3:4) обнаружена ($P=0,05$) в каждом пункте в течение первого года? Допустим, что общая σ_0 и что поэтому можно использовать показатели $s_0 = 2,49/25,3 = 9,8\%$ при $f_0 = 18$. Для предполагаемого опыта имеем $f = 10$, так что $Q_{6,10} = 4,91$ и $F_{10,18} = 1,42$. Отсюда находим δ в процентах от средней

$$\delta = \frac{4,91 \times 9,8 \times \sqrt{1,42}}{\sqrt{3}} = 33\%.$$

Следующий вопрос: «Какая процентная разность будет при данном шансе обнаружена на 4 пунктах совместно?» Если мы в данном случае сделаем довольно оптимистическое допущение, что $s_0 = 9,8\%$ является приемлемой оценкой для процентной ошибки по всем пунктам, то $f = 4 \times 10 = 40$, $Q_{6,40} = 4,23$, $F_{40,18} = 1,35$ и

$$\delta = \frac{4,23 \times 9,8 \times \sqrt{1,35}}{\sqrt{12}} = 14\%.$$

В целом эти предварительные расчеты довольно грубы, но и они имеют ценность при планировании программы будущих опытов. (Здесь предполагается, что взаимодействие сорт — пункт равно нулю. О дисперсионном анализе таких, как наша, серий опытов см. в главе 12.)

Пример 16. Планируется испытание 6 сортов по схеме латинского квадрата. На основе прежних опытов экспериментатор готов принять в качестве 50%-ного верхнего и верхнего пределов для σ соответственно 6 и 8 бушелей на акр. Тогда $s_0 = 7$ бушелей на акр и (по параграфу 18 главы 10) $f_0 = 11$. Для предполагаемого эксперимента $f = 20$, $Q_{6,20} = 4,45$ и $F_{20,11} = 1,49$. Вычислите $\delta = 16$ бушелям на акр.

9. Выпавшие данные. Различные случайности часто приводят к потере данных в одной или нескольких клетках таблицы с двумя входами. Например, урожай делянки может погибнуть, животное в момент проведения опыта пасть или может быть обнаружена ошибочность записей. Такое выпадение данных не позволяет применять теорему о слагаемости сумм квадратов, и, следовательно, мы не имеем возможности использовать описанные выше методы расчетов для построения дисперсионного анализа. Но последующее развитие теории этих методов дало возможность применения их в таких особых случаях и получения правильного анализа опытных результатов с выпавшими данными (параграф 17 главы 12). К счастью лиц, применяющих статистические методы обработки опытных данных, оказывается, что выпавшие данные могут быть теоретически вычислены приближенно и вставлены в пустые клетки таблицы. При одном выпавшем наблюдении применение обычных методов (с соответствующим изменением степеней свободы) дает приближенный

дисперсионный анализ, который при введении поправки на смещение суммы квадратов для вариантов становится уже точным.

Единичное выпавшее наблюдение в опыте с рендомизированными блоками определяется по формуле [1, 26]:

$$X = \frac{aT + bB - S}{(a-1)(b-1)},$$

где:

a — число вариантов;

b — число блоков;

T — сумма данных того варианта, где находится выпавшее наблюдение;

B — сумма данных того блока, где находится выпавшее наблюдение;

S — сумма всех наблюдений.

Например, допустим, что в таблице 120 в блоке 1 и в разновидности Г вышло наблюдение 29,3 фунта. Исключая этот урожай из сумм блока, разновидности и из общей суммы, получим величины, которые были бы при отсутствии урожая этой делянки:

$$B = 125,7 - 29,3 = 96,4; \quad T = 141,9 - 29,3 = 112,6; \quad S = 656,4 - 29,3 = 627,1;$$

$$a = 4; \quad b = 5.$$

Отсюда:

$$X = \frac{4 \times 112,6 + 5 \times 96,4 - 627,1}{(4-1)(5-1)} = 25,4 \text{ фунта.}$$

Это значение и вставляется в таблицу в качестве урожая выпавшей делянки. Дисперсионный анализ в этом случае проводится как обычно, но с двумя изменениями: каждое из числа степеней свободы в графах «общее» и «ошибка» уменьшается на единицу, а сумма квадратов для «вариантов» уменьшается на

$$\text{поправку} = \frac{[B - (a-1)X]^2}{a(a-1)} = \frac{[96,4 - (4-1) \times 25,4]^2}{4(4-1)} = 34,00.$$

В этом случае дисперсионный анализ данных по урожайности пшеницы при подстановке на место выпавшего наблюдения 25,4 фунта будет:

блоки	4		35,39	
разновидности	3	171,36 - 34,00 =	137,36	45,79
ошибка	12 - 1 = 11		17,33	1,58

Сравнение этих результатов с данными таблицы 120 приводит к двум выводам. С одной стороны, несмотря на потерю урожайных данных одной делянки, мы имеем правильный анализ остальных данных, но, с другой стороны, утрата некоторой доли информации является невозвратимой. Потеря данных по одной делянке из 20 означает утрату одной двадцатой части средств, вложенных в данный опыт. Наибольшее, что может дать наш анализ, это полное оправдание полученными из опыта сведениями затрат остальных средств.

Для латинского квадрата соответствующая формула будет

$$X = \frac{a(R + C + T) - 2S}{(a-1)(a-2)}.$$

Поправка:

$$\frac{[S - R - C - (a-1)T]^2}{[(a-1)(a-2)]^2},$$

где a — число вариантов, рядов и столбцов.

Допустим для иллюстрации, что в примере 14 вышло первое наблюдение — 608 фунтов. В этом случае:

$$R = 885 + 940 = 1825$$

$$C = 715 + 844 = 1559$$

$$T = 711 + 766 = 1477$$

$$S = 715 + \dots + 832 = 6780$$

$$X = \frac{3(1825 + 1559 + 1477) - 2 \times 6780}{(3-1)(3-2)} = 512 \text{ фунтам.}$$

$$\text{Поправка} = \frac{[6780 - 1825 - 1559 - (3-1) \times 1477]^2}{(3-1)(3-2)^2} = 48\,814.$$

После определения выпавшего урожая — 512 фунтов — в таблице дисперсионного анализа следует взять:

сумма квадратов для вариантов	129 655
поправка	48 814

Исправленная сумма квадратов 80 814

Дисперсионный анализ в окончательном виде будет:

ряды	2	9 847	
столбцы	2	68 185	
варианты	2	80 814	40 407
ошибка	1		2 773

Сравнение этого с первоначальным анализом дает представление о тех количественных изменениях, которые возникли от отсутствия данных о молочной продуктивности одной коровы. Однако вряд ли стоит затрачивать время на выводы, основанные на данных только одного единственного квадрата 3×3 . Пример 15 показывает, как могут анализироваться данные нескольких таких квадратов.

В случае выпадения двух или большего числа данных необходимы более сложные методы. В опыте с рандомизированными блоками можно непосредственно применить метод наименьших квадратов, описанный в параграфе 17 главы 12. Но в случае небольшого числа выпавших данных можно применить и метод последовательного приближения, который сводится к применению специальных способов вычисления исправленных значений сумм квадратов для вариантов.

Для облегчения иллюстрации возьмем простой пример таблицы 124.

ТАБЛИЦА 124

*Опыт с рандомизированными блоками.
Случай выпадения двух данных*

Варианты	Блоки			Сумма
	1	2	3	
A	6	5	4	15
B	15	X_{22}	8	23
C	X_3	15	12	27
Сумма	21	20	24	65

Начнем с внесения в таблицу для одного из выпавших наблюдений какого-либо более или менее подходящего значения; так, возьмем $X_{22} = 10,5$. Здесь общая средняя \bar{x} равна примерно 9, но как блок, так и вариант, к которым относится X_{22} , имеют средние выше этой общей средней, поэтому значение $X_{22} = 10,5$ представляется более подходящим. Теперь по известной формуле определяется значение второго выпавшего наблюдения X_{31} :

$$X_{31} = \frac{3 \times 27 + 3 \times 21 - 75,5}{2} = 17,1.$$

$$T = 711 + 766 = 1477$$

$$S = 715 + \dots + 832 = 6780$$

$$X = \frac{3(1825 + 1559 + 1477) - 2 \cdot 6780}{(3-1)(3-2)} = 512 \text{ фунтам.}$$

$$\text{Поправка} = \frac{[6780 - 1825 - 1559 - (3-1) \times 1477]^2}{(3-1)(3-2)^2} = 48\,811.$$

После определения выпавшего урожая — 512 фунтов — в таблице дисперсионного анализа следует взять:

сумма квадратов для вариантов	129 655
поправка	48 841

Исправленная сумма квадратов 80 814

Дисперсионный анализ в окончательном виде будет:

ряды	2	9 847	
столбцы	2	68 185	
варианты	2	80 814	40 407
ошибка	1		2 773

Сравнение этого с первоначальным анализом дает представление о тех количественных изменениях, которые возникли от отсутствия данных о молочной продуктивности одной коровы. Однако вряд ли стоит затрачивать время на выводы, основанные на данных только одного единственного квадрата 3×3 . Пример 15 показывает, как могут анализироваться данные нескольких таких квадратов.

В случае выпадения двух или большего числа данных необходимы более сложные методы. В опыте с рандомизированными блоками можно непосредственно применить метод наименьших квадратов, описанный в параграфе 17 главы 12. Но в случае небольшого числа выпавших данных можно применить и метод последовательного приближения, который сводится к применению специальных способов вычисления исправленных значений сумм квадратов для вариантов.

Для облегчения иллюстрации возьмем простой пример таблицы 124.

ТАБЛИЦА 124

*Опыт с рандомизированными блоками.
Случай выпадения двух данных*

Варианты	Блоки			Сумма
	1	2	3	
A	6	5	4	15
B	15	X_{22}	8	23
C	X_{31}	15	12	27
Сумма	21	20	24	65

Начнем с внесения в таблицу для одного из выпавших наблюдений какого-либо более или менее подходящего значения; так, возьмем $X_{22} = 10,5$. Здесь общая средняя \bar{x} равна примерно 9, но как блок, так и вариант, к которым относится X_{22} , имеют средние выше этой общей средней, поэтому значение $X_{22} = 10,5$ представляется более подходящим. Теперь по известной формуле определится значение второго выпавшего наблюдения X_{31} :

$$X_{31} = \frac{3 \times 27 + 3 \times 21 - 75,5}{(3-1)(3-1)} = 17,4.$$

Подстановка в таблицу $X_{31}=17,1$ позволяет найти более подходящее значение уже для выпавшего наблюдения X_{22} , применяя ту же формулу:

$$X_{22} = \frac{3 \times 23 + 3 \times 20 - 82,1}{4} = 11,7.$$

С этим уточненным значением X_{22} обращаемся к пересчету X_{31} :

$$X_{31} = \frac{3 \times 27 + 3 \times 21 - 76,7}{4} = 16,8.$$

Наконец, при внесении в таблицу этого нового значения X_{31} вычисляем $X_{22}=11,8$. После этого требуется еще один этап расчетов для того, чтобы убедиться, что $X_{22}=11,8$ не меняется от этого пересчета X_{31} .

Вычисленные таким образом два наблюдения вносятся в таблицу и проводятся обычные расчеты для дисперсионного анализа. Это делается для получения только одного результата анализа: суммы квадратов для ошибки $4-2=2$ степеням свободы (вычитается по одной степени свободы для каждого выпавшего наблюдения). Сумма квадратов в нашем случае равна 6,40.

После этого возвращаемся к первоначальной таблице и, игнорируя пустующие клетки ее, вычисляем общую сумму квадратов:

$$\text{Общее: } \Sigma X_{ij}^2 - C = 735 - 65^2/7 = 131,43.$$

Точно так же вычисляется по первоначальным данным сумма квадратов для блоков:

$$\text{Блоки: } \frac{24^2 + 20^2}{2} + \frac{24^2}{3} - C = 8,93.$$

Несмещенная оценка суммы квадратов для вариантов получается теперь следующим образом:

		Сумма квадратов
Общее, первоначальные данные . . .		131,43
Блоки, первоначальные данные . . .	8,93	
Ошибка, восполненные данные . . .	6,40	15,33
Сумма квадратов для вариантов . . .		116,10

Окончательный дисперсионный анализ будет таким:

Варианты	2	116,10	58,05
Ошибка	2	6,40	3,20

Полный дисперсионный анализ восполненной (за счет оценок выпавших данных) таблицы будет иметь неправильную, смещенную сумму квадратов для вариантов 144. Разность $144-116,1=27,9$ и является искомым смещением.

Кемпторн [14] дал метод несмещенной оценки двух или большего числа выпавших данных для случая латинского квадрата. Мне никогда не встречался латинский квадрат с двумя или большим числом выпавших наблюдений, и поэтому я для иллюстрации построил (по способу параграфа этой главы 7) искусственную модель квадрата 4×4 . Я взял:

$$\mu = 12$$

$$\alpha_i = -5, -1, 2, 4; \quad \alpha_A^2 = \Sigma \alpha_i^2 / (4-1) = 15,3$$

$$\beta_j = -2, -1, 0, 3$$

$$\gamma_k = -4, 0, 2, 2$$

$$\varepsilon_{ijkl} = N(0, 1) \text{ (из таблицы 15; каждое отклонение разделено на 10).}$$

После исключения по жребию двух деянок я имел данные таблицы 125 исключенными значениями были $X_{111}=1$ и $X_{324}=18$.

Латинский квадрат 4×4 с двумя выпавшими делянками

Ряд	Столбцы				Сумма
	1	2	3	4	
1	$A X_{111}$	$B \ 8$	$C \ 17$	$D \ 15$	40
2	$B \ 6$	$C \ 14$	$D \ 16$	$A \ 9$	45
3	$C \ 10$	$D X_{324}$	$A \ 7$	$B \ 12$	29
4	$D \ 15$	$A \ 9$	$B \ 16$	$C \ 19$	59
Сумма	31	31	56	55	173

На первой стадии работы подбираются оценки для выпавших данных. Теперь метод последовательного приближения, описанный выше, осуществляется при помощи формулы для выпавших данных в латинском квадрате. В результате получается $X_{111}=2,2$ и $X_{324}=12,3$.

Далее проводится дисперсионный анализ пополненного квадрата (т. е. с включением двух вычисленных X). Отсюда берется только одна необходимая нам часть — сумма квадратов ошибки $SSE_S=5,33$.

Затем обрабатывается таблица 125, как опыт с рандомизированными блоками, т. е. учитывая ряды и столбцы, но игнорируя буквы, что дает два новых значения для X , вычисленных по формуле выпавших данных в рандомизированных блоках: $X_{11}=11,7$ и $\bar{X}_{21}=6,1$.

Далее, проводится дисперсионный анализ данных опыта с пополненными этими значениями рандомизированными блоками. Все, что здесь нам необходимо, это сумма квадратов ошибки $SSE_B=119,49$.

Наконец, в таблице 126 проводится дисперсионный анализ для латинского квадрата, в котором в качестве суммы квадратов «вариантов» берется разность: $SSE_B-SSE_S=119,49-5,33=114,16$.

ТАБЛИЦА 126

Дисперсионный анализ латинского квадрата с двумя выпавшими делянками

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат	Оцениваемые параметры
Варианты	3	114,16	38,05	$\sigma^2 + 4\chi^2*$
Ошибка	4	5,33	1,33	σ^2

$$5,33 = SSE_S; 114,16 = SSE_B - SSE_S = 119,49 - 5,33;$$

$$1,33 \text{ оценивает } \sigma^2 = 1; (38,05 - 1,33)/4 = 9,18 \text{ оценивает } \chi^2 = 15,3$$

* Здесь не приняты во внимание изменения, возникающие при выпадении наблюдений.

Как вы можете видеть, успех, достигнутый нами в оценке заранее известных нам параметров, посредственный. При применении этого метода в случаях, когда выпадает более двух наблюдений, необходимо производить вычитание из степеней свободы «ошибки» по одной степени на каждое восстанавливаемое наблюдение.

Здесь мы делаем последнее направляющее указание читателю, интересующемуся кратким курсом. Он вполне подготовлен к чтению начальных параграфов большинства последующих глав.

10. Несогласованность с моделью. Преобразования. В предыдущем мы использовали модель опытов, которая характеризуется слагаемостью

эффектов при нормальном и независимом распределении ошибок, имеющем постоянную дисперсию $\varepsilon = N(0, \sigma)$. Однако не столь редко можно встретить выборки из совокупностей, к которым эта модель непосредственно не подходит. Один путь преодоления такого положения состоит в изменении самой модели; для этого были предложены (параграфы 7—10 главы 5, параграф 12 главы 7) разнообразные непараметрические методы. Другой путь состоит в изменении шкалы количественного признака при помощи *преобразования* ее. Для выборок из некоторых, имеющих определенный характер совокупностей эта задача имеет вполне точное решение [11], но чаще всего выборка носит смешанный характер. Поэтому обычно находят приближенные решения, которые практически представляются достаточными [3]. Ниже будут описаны три из таких решений.

Аномальность, слагаемость и неоднородность варьирования обычно встречаются вместе. Было бы идеально, если бы преобразование шкалы могло бы быть средством от всех этих нарушений, но это встречается не часто. Наиболее существенным требованием является слагаемость эффектов, а следующим — однородность дисперсий.

Теоретически определенного вида совокупности должна соответствовать определенная модель. Когда выборки берутся из биномиального или пуассоновского распределений, то известно, что дисперсия находится в связи со средней (глава 16). Для таких случаев можно точно определить подходящее преобразование. Но чаще всего приходится руководствоваться только тем, что можно усмотреть в самой выборке. Иногда ключ к решению вопроса может дать сравнение размахов варьирования по вариантам опыта. См. также [23].

Условие слагаемости по выборочным данным можно проверить (параграф 14 этой главы) по методу Тьюки. Этот критерий применяется в тех случаях, когда имеются сомнения в необходимости преобразования или в ее успехе.

11. Преобразование данных подсчета численности путем извлечения квадратного корня. Результаты подсчета таких переменных, как количество червячков на делянке или насекомых, попавших в ловушку, иногда приближаются к распределению, имеющему форму распределения Пуассона с дисперсией, пропорциональной средней и не подчиняющейся правилу слагаемости эффектов. В этих случаях часто подходящим является преобразование \sqrt{X} (или $\sqrt{X+1}$, если некоторые подсчеты дают небольшие значения). В таблице 127 приведены в качестве примера данные о количестве растений мака

ТАБЛИЦА 127

Число растений мака в посевах овса. Делянки $3\frac{3}{4}$ кв. фута

Блоки	Варианты				
	A	B	C	D	E
1	438	538	77	17	18
2	442	422	61	31	26
3	319	377	157	87	77
4	380	315	52	16	20
Средняя . . .	395	413	87	38	35
Размах . . .	123	223	105	71	59

в посевах овса. Большое различие размахов варьирования наводит на мысль о неоднородности дисперсий. Если в этом случае вычислить средний квадрат ошибки, то, вообще говоря, он будет завышенным для разностей между C, D и E и заниженным для разностей между A и B.

В таблице 128 выписаны квадратные корни из этих чисел и по ним произведен соответствующий анализ. Теперь размахи варьирования примерно одинаковы. Условие слагаемости будет исследовано в параграфе 14.

Что здесь имеются различия между вариантами, видно как из той, так и из другой таблицы; нет необходимости устанавливать это при помощи критерия *F*. Для всех сравнений доверительный интервал средней разности (параграф 6 главы 10) будет здесь таким:

$$D = 4,51 \sqrt{4,06/4} = 4,54.$$

Это указывает на то, что между тремя эффективными вариантами нет существенных различий: нет этих различий и между двумя неэффективными вариантами (*A* был контролем), но варианты *C*, *D* и *E* существенно отличаются от *A* и *B*.

После того как оценка и сравнения проведены, производится обратный переход к первоначальным единицам измерения, что дает информацию

ТАБЛИЦА 128

Квадратные корни из численностей растений мака по данным таблицы 127

Блок	A	B	C	D	E
1	20,9	23,2	8,8	4,1	4,2
2	21,0	20,5	7,8	5,6	5,1
3	17,9	19,4	12,5	9,3	8,8
4	19,5	17,7	7,2	4,0	4,5
	19,8	20,2	9,4	5,8	5,6

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Блоки	3	22,65	
Варианты	4	865,44	216,36
Ошибка	12	48,69	4,06

в форме исходных данных. Исправленные оценки средних в нашем случае будут:

$$\bar{x}_A = 19,8^2 = 392 \text{ растениям, } \bar{x}_B = 408, \bar{x}_C = 83, \bar{x}_D = 34 \text{ и } \bar{x}_E = 31 \text{ растению.}$$

Сравнение с первоначальными оценками дает представление о влиянии неслагаемости и неоднородности дисперсий.

Пример 17. На делянках опыта по изучению некоторых способов фумигации почвы в предшествующий год, поставленного по схеме латинского квадрата, были обнаружены такие количества проволочного червя:

Ряды	Столбцы				
	1	2	3	4	5
1	P 3	O 2	N 5	K 1	M 4
2	M 6	K 0	N 6	P 4	P 4
3	O 4	M 9	K 1	P 6	N 5
4	N 17	P 8	M 8	O 9	K 0
5	K 4	N 4	P 2	M 4	O 8

Числа, соответствующие процентам: угол = арксинус V процент, по данным Блесс (4)

0	1	2	3	4	5	6	7	8	9	
0,0	0	0,57	0,81	0,99	1,15	1,28	1,40	1,52	1,62	1,72
0,1	1,81	1,90	1,99	2,07	2,14	2,22	2,29	2,36	2,43	2,50
0,2	2,56	2,63	2,69	2,75	2,81	2,87	2,92	2,98	3,03	3,09
0,3	3,14	3,19	3,24	3,29	3,34	3,39	3,44	3,49	3,53	3,58
0,4	3,63	3,67	3,72	3,76	3,80	3,85	3,89	3,93	3,97	4,01
0,5	4,05	4,09	4,13	4,17	4,21	4,25	4,29	4,33	4,37	4,40
0,6	4,44	4,48	4,52	4,55	4,59	4,62	4,66	4,69	4,73	4,76
0,7	4,80	4,83	4,87	4,90	4,93	4,97	5,00	5,03	5,07	5,10
0,8	5,13	5,16	5,20	5,23	5,26	5,29	5,32	5,35	5,38	5,41
0,9	5,44	5,47	5,50	5,53	5,56	5,59	5,62	5,65	5,68	5,71
1	5,74	6,02	6,29	6,55	6,80	7,04	7,27	7,49	7,71	7,92
2	8,13	8,33	8,53	8,72	8,91	9,10	9,28	9,46	9,63	9,81
3	9,98	10,14	10,31	10,47	10,63	10,78	10,94	11,09	11,24	11,39
4	11,54	11,68	11,83	11,97	12,11	12,25	12,39	12,52	12,66	12,79
5	12,92	13,05	13,18	13,31	13,44	13,56	13,69	13,81	13,94	14,06
6	14,18	14,30	14,42	14,54	14,65	14,77	14,89	15,00	15,12	15,23
7	15,34	15,45	15,56	15,68	15,79	15,89	16,00	16,11	16,22	16,32
8	16,43	16,54	16,64	16,74	16,85	16,95	17,05	17,16	17,26	17,36
9	17,46	17,56	17,66	17,76	17,85	17,95	18,05	18,15	18,24	18,34
10	18,44	18,53	18,63	18,72	18,81	18,91	19,00	19,09	19,19	19,28
11	19,37	19,46	19,55	19,64	19,73	19,82	19,91	20,00	20,09	20,18
12	20,27	20,36	20,44	20,53	20,62	20,70	20,79	20,88	20,96	21,05
13	21,13	21,22	21,30	21,39	21,47	21,56	21,64	21,72	21,81	21,89
14	21,97	22,06	22,14	22,22	22,30	22,38	22,46	22,55	22,63	22,71
15	22,79	22,87	22,95	23,03	23,11	23,19	23,26	23,34	23,42	23,50
16	23,58	23,66	23,73	23,81	23,89	23,97	24,04	24,12	24,20	24,27
17	24,35	24,43	24,50	24,58	24,65	24,73	24,80	24,88	24,95	25,03
18	25,10	25,18	25,25	25,33	25,40	25,48	25,55	25,62	25,70	25,77
19	25,84	25,92	25,99	26,06	26,13	26,21	26,28	26,35	26,42	26,49
20	26,56	26,64	26,71	26,78	26,85	26,92	26,99	27,06	27,13	27,20
21	27,28	27,35	27,42	27,49	27,56	27,63	27,69	27,76	27,83	27,90
22	27,97	28,04	28,11	28,18	28,25	28,32	28,38	28,45	28,52	28,59
23	28,66	28,73	28,79	28,86	28,93	29,00	29,06	29,13	29,20	29,27
24	29,33	29,40	29,47	29,53	29,60	29,67	29,73	29,80	29,87	29,93
25	30,00	30,07	30,13	30,20	30,26	30,33	30,40	30,46	30,53	30,59
26	30,66	30,72	30,79	30,85	30,92	30,98	31,05	31,11	31,18	31,24
27	31,11	31,37	31,44	31,50	31,56	31,63	31,69	31,76	31,82	31,88
28	31,95	32,01	32,08	32,14	32,20	32,27	32,33	32,39	32,46	32,52
29	32,58	32,65	32,71	32,77	32,83	32,90	32,96	33,02	33,09	33,15
30	33,21	33,27	33,34	33,40	33,46	33,52	33,58	33,65	33,71	33,77
31	33,83	33,89	33,96	34,02	34,08	34,14	34,20	34,27	34,33	34,39
32	34,45	34,51	34,57	34,63	34,70	34,76	34,82	34,88	34,94	35,00
33	35,06	35,12	35,18	35,24	35,30	35,37	35,43	35,49	35,55	35,61
34	35,67	35,73	35,79	35,85	35,91	35,97	36,03	36,09	36,15	36,21
35	36,27	36,33	36,39	36,45	36,51	36,57	36,63	36,69	36,75	36,81
36	36,87	36,93	36,99	37,05	37,11	37,17	37,23	37,29	37,35	37,41
37	37,47	37,52	37,58	37,64	37,70	37,76	37,82	37,88	37,94	38,00
38	38,06	38,12	38,17	38,23	38,29	38,35	38,41	38,47	38,53	38,59
39	38,65	38,70	38,76	38,82	38,88	38,94	39,00	39,06	39,11	39,17
40	39,23	39,29	39,35	39,41	39,47	39,52	39,58	39,64	39,70	39,76
41	39,82	39,87	39,93	39,99	40,05	40,11	40,16	40,22	40,28	40,34
42	40,40	40,46	40,51	40,57	40,63	40,69	40,74	40,80	40,86	40,92
43	40,98	41,03	41,09	41,15	41,21	41,27	41,32	41,38	41,44	41,50
44	41,55	41,61	41,67	41,73	41,78	41,84	41,90	41,96	42,02	42,07
45	42,13	42,19	42,25	42,30	42,36	42,42	42,48	42,53	42,59	42,65
46	42,71	42,76	42,82	42,88	42,94	42,99	43,05	43,11	43,17	43,22
47	43,28	43,34	43,39	43,45	43,51	43,57	43,62	43,68	43,74	43,80

%	0	1	2	3	4	5	6	7	8	9
48	43,85	43,91	43,97	44,03	44,08	44,14	44,20	44,25	44,31	44,37
49	44,43	44,48	44,54	44,60	44,66	44,71	44,77	44,83	44,89	44,94
50	45,00	45,05	45,11	45,17	45,23	45,29	45,34	45,40	45,46	45,52
51	45,57	45,63	45,69	45,75	45,80	45,86	45,92	45,97	46,03	46,09
52	46,15	46,20	46,26	46,32	46,38	46,43	46,49	46,55	46,61	46,66
53	46,72	46,78	46,83	46,89	46,95	47,01	47,06	47,12	47,18	47,24
54	47,29	47,35	47,41	47,47	47,52	47,58	47,64	47,70	47,75	47,81
55	47,87	47,93	47,98	48,04	48,10	48,16	48,22	48,27	48,33	48,39
56	48,45	48,50	48,56	48,62	48,68	48,73	48,79	48,85	48,91	48,97
57	49,02	49,08	49,14	49,20	49,26	49,31	49,37	49,43	49,49	49,54
58	49,60	49,66	49,72	49,78	49,84	49,89	49,95	50,01	50,07	50,13
59	50,18	50,24	50,30	50,36	50,42	50,48	50,53	50,59	50,65	50,71
60	50,77	50,83	50,89	50,94	51,00	51,06	51,12	51,18	51,24	51,30
61	51,35	51,41	51,47	51,53	51,59	51,65	51,71	51,77	51,83	51,88
62	51,94	52,00	52,06	52,12	52,18	52,24	52,30	52,36	52,42	52,48
63	52,53	52,59	52,65	52,71	52,77	52,83	52,89	52,95	53,01	53,07
64	53,13	53,19	53,25	53,31	53,37	53,43	53,49	53,55	53,61	53,67
65	53,73	53,79	53,85	53,91	53,97	54,03	54,09	54,15	54,21	54,27
66	54,33	54,39	54,45	54,51	54,57	54,63	54,70	54,76	54,82	54,88
67	54,94	55,00	55,06	55,12	55,18	55,24	55,30	55,37	55,43	55,49
68	55,55	55,61	55,67	55,73	55,80	55,86	55,92	55,98	56,04	56,11
69	56,17	56,23	56,29	56,35	56,42	56,48	56,54	56,60	56,66	56,73
70	56,79	56,85	56,91	56,98	57,04	57,10	57,17	57,23	57,29	57,35
71	57,42	57,48	57,54	57,61	57,67	57,73	57,80	57,86	57,92	57,99
72	58,05	58,12	58,18	58,24	58,31	58,37	58,44	58,50	58,56	58,63
73	58,69	58,76	58,82	58,89	58,95	59,02	59,08	59,15	59,21	59,28
74	59,34	59,41	59,47	59,54	59,60	59,67	59,74	59,80	59,87	59,93
75	60,00	60,07	60,13	60,20	60,27	60,33	60,40	60,47	60,53	60,60
76	60,67	60,73	60,80	60,87	60,94	61,00	61,07	61,14	61,21	61,27
77	61,34	61,41	61,48	61,55	61,62	61,68	61,75	61,82	61,89	61,96
78	62,03	62,10	62,17	62,24	62,31	62,37	62,44	62,51	62,58	62,65
79	62,72	62,80	62,87	62,94	63,01	63,08	63,15	63,22	63,29	63,36
80	63,44	63,51	63,58	63,65	63,72	63,79	63,87	63,94	64,01	64,08
81	64,16	64,23	64,30	64,38	64,45	64,52	64,60	64,67	64,75	64,82
82	64,90	64,97	65,05	65,12	65,20	65,27	65,35	65,42	65,50	65,57
83	65,65	65,73	65,80	65,88	65,96	66,03	66,11	66,19	66,27	66,34
84	66,42	66,50	66,58	66,66	66,74	66,81	66,89	66,97	67,05	67,13
85	67,21	67,29	67,37	67,45	67,54	67,62	67,70	67,78	67,86	67,94
86	68,03	68,11	68,19	68,28	68,36	68,44	68,53	68,61	68,70	68,78
87	68,87	68,95	69,04	69,12	69,21	69,30	69,38	69,47	69,56	69,64
88	69,73	69,82	69,91	70,00	70,09	70,18	70,27	70,36	70,45	70,54
89	70,63	70,72	70,81	70,91	71,00	71,09	71,19	71,28	71,37	71,47
90	71,56	71,66	71,76	71,85	71,95	72,05	72,15	72,24	72,34	72,44
91	72,54	72,64	72,74	72,84	72,95	73,05	73,15	73,26	73,36	73,46
92	73,57	73,68	73,78	73,89	74,00	74,11	74,21	74,32	74,44	74,55
93	74,66	74,77	74,88	75,00	75,11	75,23	75,35	75,46	75,58	75,70
94	75,82	75,94	76,06	76,19	76,31	76,44	76,56	76,69	76,82	76,95
95	77,08	77,21	77,34	77,48	77,61	77,75	77,89	78,03	78,17	78,32
96	78,46	78,61	78,76	78,91	79,06	79,22	79,37	79,53	79,69	79,86
97	80,02	80,19	80,37	80,54	80,72	80,90	81,09	81,28	81,47	81,67
98	81,87	82,08	82,29	82,51	82,73	82,96	83,20	83,45	83,71	83,98
99,0	84,26	84,29	84,32	84,35	84,38	84,41	84,44	84,47	84,50	84,53
99,1	84,56	84,59	84,62	84,65	84,68	84,71	84,74	84,77	84,80	84,84
99,2	84,87	84,90	84,93	84,97	85,00	85,03	85,07	85,10	85,13	85,17
99,3	85,20	85,24	85,27	85,31	85,34	85,38	85,41	85,45	85,48	85,52
99,4	85,56	85,60	85,63	85,67	85,71	85,75	85,79	85,83	85,87	85,91
99,5	85,95	85,99	86,03	86,07	86,11	86,15	86,20	86,24	86,28	86,33
99,6	86,37	86,42	86,47	86,51	86,56	86,61	86,66	86,71	86,76	86,81
99,7	86,86	86,91	86,97	87,02	87,08	87,13	87,19	87,25	87,31	87,37
99,8	87,44	87,50	87,57	87,64	87,71	87,78	87,86	87,93	88,01	88,10
99,9	88,19	88,28	88,38	88,48	88,60	88,72	88,85	89,01	89,19	89,43
100,0	90,00	—	—	—	—	—	—	—	—	—

Так как здесь числа небольшие, то преобразование следует провести по формуле $1/\sqrt{3-1}$. Первое число 3 приводит к $\sqrt{3-1}=2$ и т. д.

Проведите дисперсионный анализ. *Ответ:* средний квадрат для «вариаптов» 1,4457, для «блоков» 0,3259.

Пример 18. Вариапт *O* был «контролем», но он не выделяется среди других. Вычислите *D* = 1,06 и покажите, что *K* приводит к существенно меньшему количеству проволочного червя, чем *M*, *N* и *O*.

Пример 19. Определите средние числа проволочника на делянку по отдельным вариаптам. *Ответ:* *K*—0,99; *M*—6,08; *N*—6,40; *O*—5,53; *P*—4,38.

12. Преобразование долей через арксинусы. В тех случаях, когда переменная представляет собой долю определенного вида объектов, распределение ее приближается к бинomialной форме. Соответствующее преобразование производится через угол, синус которого является квадратным корнем из доли или процента, т. е. при помощи таблицы 129. Эта таблица придает больший вес малым процентам, которые имеют и малую дисперсию. Применение этой таблицы показано в таблице 130, где приведены данные о доле растений, пораженных раком стеблей [9].

ТАБЛИЦА 129

Доли растений сор, пораженных раком стеблей. Гудсон (Майо), 1919 г.

Блоки	Сорта					
	A	B	C	D	E	F
1	19,3	10,1	25,2	14,0	3,3	3,1
2	29,2	34,7	36,5	30,2	35,8	9,6
3	1,0	14,0	23,4	7,2	1,1	1,0
4	6,4	5,6	12,9	8,9	2,0	1,0

Арксинус $\sqrt{\text{доля}}$

1	26,1	18,5	30,1	22,0	10,5	10,1
2	32,7	36,1	37,2	33,3	36,8	18,0
3	5,7	22,0	28,9	15,6	6,0	5,7
4	14,6	13,7	21,0	17,4	8,1	5,7
Средняя % растений	19,8	22,6	29,3	22,1	15,4	9,9
	11,5	14,8	24,0	14,0	7,1	3,0

Источники варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Блоки	3	1392,85	
Сорта	5	885,72	177,14
Ошибка	15	364,15	24,28

Если большинство процентов превосходит 50, то, применяя обработку другого из двух противоположных случаев, например у нас доли непораженных растений, можно упростить работу и достигнуть экономии времени.

Применение критерия $D=11,3$ указывает на то, что сорт *F* более устойчив против рака, чем сорта *B*, *C* и *D*, и что *E* более устойчив, чем *C*. Результаты обратного перехода к процентам приведены в графе, стоящей над дисперсионным анализом.

Пример 20. Дж. К. Гейнс на Техасской сельскохозяйственной опытной станции изучал воздействие четырех видов дуста на коробочного долгоносика. Вместе с необработанной делянкой *E* в каждом повторении было 5 вариантов: опыт был заложен по схеме латинского квадрата 5×5. На каждой делянке выделено 200 квадратов для наблюдения проколов бутонов хлопчатника долгоносиком. Процент квадратов, на которых такие проколы были обнаружены, по отдельным делянкам был:

<i>C</i> 27	<i>E</i> 42	<i>A</i> 18	<i>D</i> 34	<i>B</i> 17
<i>A</i> 18	<i>D</i> 27	<i>E</i> 42	<i>B</i> 14	<i>C</i> 12
<i>D</i> 23	<i>B</i> 14	<i>C</i> 17	<i>E</i> 25	<i>A</i> 14
<i>E</i> 24	<i>A</i> 10	<i>B</i> 8	<i>C</i> 12	<i>D</i> 26
<i>B</i> 9	<i>C</i> 11	<i>D</i> 15	<i>A</i> 11	<i>E</i> 22

Преобразуйте проценты в углы и после этого постройте следующий дисперсионный анализ:

Ряды	4	311,3	
Столбцы	4	6,0	
Варианты	4	637,5	159,38
Ошибка	12	108,8	9,07

Пример 21. Показать, что $D=6,1$ и что дусты *A*, *B* и *C* дают существенно меньшие проценты, чем *D* или контроль *E*.

13. Логарифмическое преобразование. Логарифмы применяются, когда известно, что эффекты перемножаются, а не складываются, и также в случаях, когда стандартное отклонение изменяется прямо пропорционально средней. Пропорциональность влияний—широко распространенное явление в экономике и, по мнению К. Б. Вильямса [24], имеет преобладающее значение в таких явлениях, как количество насекомых, попавших в ловушку. Применение логарифмов может привести к исправлению более серьезных случаев нарушения слагаемости, когда преобразование через квадратные корни не имеет успеха (см. статью Бартлетта [3]).

Приведенные в таблице 131 данные об улове планктона [25] прекрасно подходят для логарифмического преобразования. Фактические размахи варьирования и средние здесь близки к полной пропорциональности; отно-

ТАБЛИЦА 131

Определение численностей четырех видов планктона, попавшего при шести уловах двумя сетями

Пробы	Установленные численности				Логарифмы			
	I	II	III	IV	I	II	III	IV
1	895	1520	43 300	11 000	2,95	3,18	4,64	4,04
2	540	1610	32 800	8 600	2,73	3,21	4,52	3,93
3	1020	1900	28 800	8 260	3,01	3,28	4,46	3,92
4	470	1350	34 600	9 830	2,67	3,13	4,54	3,99
5	428	980	27 800	7 600	2,63	2,99	4,44	3,88
6	620	1710	32 800	9 650	2,79	3,23	4,52	3,98
7	760	1930	28 100	8 900	2,88	3,29	4,45	3,95
8	537	1960	18 900	6 060	2,73	3,29	4,28	3,78
9	845	1840	31 400	10 200	2,93	3,26	4,50	4,01
10	1050	2410	39 500	15 500	3,02	3,38	4,60	4,19
11	387	1520	29 000	9 250	2,59	3,18	4,46	3,97
12	497	1685	22 300	7 900	2,70	3,23	4,35	3,90
Средняя	671	1701	30 775	9 396	2,802	3,221	4,480	3,962
Размах	663	1480	24 400	9 440	0,43	0,39	0,36	0,41

Дисперсионный анализ логарифмов

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Вид планктона	3	20,2070	6,7357
Почва	11	0,3387	0,0308
Повышки	33	0,2300	0,0070

ления средней к размаху соответственно 1,01, 1,15, 1,26 и 1,00. Иначе говоря, здесь коэффициент вариации имеет примерно одно и то же значение. После преобразования размахи стали почти одинаковыми и некоррелированными со средней.

При переходе к первоначальным численностям уловов правильно рассчитанными средними для четырех видов планктона будут: антилогарифм $1,802 = 634$, 1663, 30200 и 9162. Это будут *геометрические средние*.

После вычитания $D = 0,092$ все средние логарифмов (и, следовательно, средние геометрические из первоначальных численностей) представляются как существенно различные.

Стандартное отклонение логарифмов $\sqrt{0,0070} = 0,084$, что дает для первоначальных численностей антилогарифм 1,21. Цитируем Уинсора и Кларка (стр. 5): «Отклонение 0,084 у логарифма улова означает, что улов умножается (или делится) на 1,21. Следовательно, мы можем сказать, что однократное стандартное отклонение у логарифма соответствует процентному стандартному отклонению или коэффициенту вариации в 21% улова».

Пример 22. Следующие данные взяты из опыта [24], который имел несколько более сложную схему. Данные представляют собой средние геометрические количеств насекомых (макролепидоптера), попавших в ловушку в течение трех следующих одна за другой ночей. Ловушку каждую ночь помещали в разных местах. Место наблюдения — Ротамедская опытная станция.

Ловушка	Периоды по три ночи, Август 1959 г.				
	16—18	19—21	22—24	25—27	28—30
1	19,1	23,4	29,5	23,4	16,6
2	50,1	166,0	223,9	58,9	64,6
3	123,0	407,4	398,1	229,1	251,2

Здесь не возникает никакого сомнения относительно существования различий, но все же для получения точных оценок требуется преобразование данных. Вильямс применил здесь логарифмическое преобразование, хотя в данном случае можно было бы ожидать распределение численностей в форме распределения Пуассона. Проведите преобразование в логарифмы и проведите по ним дисперсионный анализ. *Ответ:* средний квадрат для ловушек 1,4455, для ошибки 0,0172.

Показать, что все различия существенны и что геометрические средние по ловушкам будут 21,9, 93,3 и 257,0 насекомых. Этот пример более подробно обсуждается в следующем параграфе.

14. Критерий слагаемости Тьюки. Этот критерий применяется для различных целей: 1) для определения, оправдало ли себя некоторое преобразование данных; 2) для выяснения вопроса относительно уклоняющихся наблюдений; 3) для решения вопроса, является ли преобразование данных необходимым; 4) для подсказа подходящего вида преобразования. Для более полного выяснения вопроса и для знакомства с другими иллюстрациями читатель отсылается к статьям Тьюки [21, 23].

Неслагаемость дисперсий встречается тогда, когда основные эффекты образуются при помощи таких процессов, как перемножение, т. е. путем процентного изменения, или же, когда случайным образом некоторое наблюдение появляется из посторонней совокупности, а не из той, которая исследуется; наличие неоднородности дисперсий также вызывает необходимость взвешивания средних; наконец, если имеет место дополнительный эффект, присоединяющийся к главным эффектам и ошибке, т. е. если имеется взаимодействие между ними (глава 12).

Применение этого метода проверки гипотезы о слагаемости дисперсий при схеме рандомизированных блоков показано в таблице 132 путем обработки логарифмов от численности пойманных экземпляров макролепидонтера. Первоначальные численности приведены в примере 22.

Новые особенности вычислений отражены в пунктах 1 и 2, поставленных внизу расчетов. Во-первых, величины p_j , внесенные в правый столбец, полу-

ТАБЛИЦА 132

Логарифмы (X_{ij}) численностей пойманных экземпляров макролепидонтера. Пример 22.

Расчеты критерия слагаемости

Период	Ловушки			Сумма X_{ij}	Средняя \bar{X}_j	Отклонение $d_j = \frac{X_{1j} - \bar{X}_j}{\bar{X}_j}$	Произведение $p_j = \sum X_{ij} d_j$
	1	2	3				
1	1,28	1,70	2,09	5,07	1,69	-0,22	0,4344
2	1,37	2,22	2,61	6,20	2,07	0,16	0,6795
3	1,47	2,35	2,60	6,42	2,14	0,23	0,6266
4	1,37	1,77	2,36	5,50	1,83	-0,07	0,5230
5	1,22	1,81	2,40	5,43	1,81	-0,10	0,6313
Сумма X_{ij}	6,71	9,85	12,06	28,62			$\sum p_j = 2,8948$
Средняя \bar{X}_j	1,34	1,97	2,41	$\bar{X}_{..} = 1,91$			
Отклонение d_j	-0,57	0,07	0,50				

1) Произведения $p_1 = \sum X_{1j} d_j = 1,28 \cdot (-0,57) + 1,70 \cdot 0,07 + 2,09 \cdot 0,50 = 0,4344$

$p_5 = \sum X_{5j} d_j = 1,22 \cdot (-0,57) + 1,81 \cdot 0,07 + 2,40 \cdot 0,50 = 0,6313$

2) $P = \sum d_j p_j = (-0,22) \cdot 0,4344 + \dots + (-0,10) \cdot 0,6313 = 0,05753$

3) $\sum d_j^2 = (-0,22)^2 + \dots + (-0,10)^2 = 0,1418$

4) $\sum d_j^2 = (-0,57)^2 + 0,07^2 + 0,50^2 = 0,5798$

5) Сумма квадратов от неслагаемости = $\frac{P^2}{\sum d_j^2 \cdot \sum d_j^2} = \frac{0,05753^2}{0,1418 \cdot 0,5798} = 0,0403$

Дисперсионный анализ логарифмов численностей

Период	4	0,4251	
Ловушка	2	2,8910	1,4455
Ошибка	8	0,4381	0,0172

* Сумма d должна равняться нулю.

чаются как суммы произведений логарифмов данного ряда на отклонение, приведенные в последней строке таблицы. Во-вторых, P является суммой произведений данных в двух крайних столбцах справа.

Пункт 5 дает сумму квадратов (с одной степенью свободы), возникшую в связи с неслагаемостью дисперсий. Она составляет часть суммы квадратов «ошибки». Проверка гипотезы о слагаемости дисперсий производится так:

ошибка	8	0,4381	
неслагаемость	1	0,0403	0,0403
для проверки		7	0,0978
			0,0140

Здесь $F = 0,0403/0,0140 = 2,88$ при $f = 1$ и 7 . Так как $F_{0,05} = 5,59$, то в отношении слагаемости не возникает сомнения. По крайней мере в этом отношении логарифмическое преобразование можно считать оправдавшим себя.

Если F окажется большим, то это будет указывать на неслагаемость эффектов и требует решения вопроса о том, как поступить далее. В помощь решению этого вопроса можно обратиться к графику (рис. 24), в котором в качестве ординат берутся p_i , а в качестве абсцисс средние \bar{x}_i . Среднее значение p_i у нас $\bar{p}_i = 2,8948/5 = 0,58$, выражается горизонтальной прямой. Подобным же образом указывается и 95%-ный доверительный интервал:

$$\begin{aligned} \bar{p}_i \pm 2\sqrt{\Sigma d_i^2} \times (\text{средний квадрат для проверки слагаемости}) &= \\ &= 0,58 \pm 2\sqrt{0,58 \times 0,0140} = \\ &= 0,58 \pm 0,18, \text{ т. е. от } 0,40 \text{ до } 0,76. \end{aligned}$$

Тот факт, что все точки у нас лежат внутри доверительных пределов, находится в соответствии с несущественностью одной степени свободы, относящейся к неслагаемости. Если же была бы фактическая неслагаемость, то следует ожидать, что одна или несколько из отмеченных точек выйдут из этих доверительных пределов.

Так, если взять первоначальные численности насекомых (пример 22) и произвести проверку слагаемости, то одна степень свободы для слагаемости окажется весьма и весьма существенной, а точки графика расположатся в направлении вверх и вправо (положительная регрессия), так что некоторые

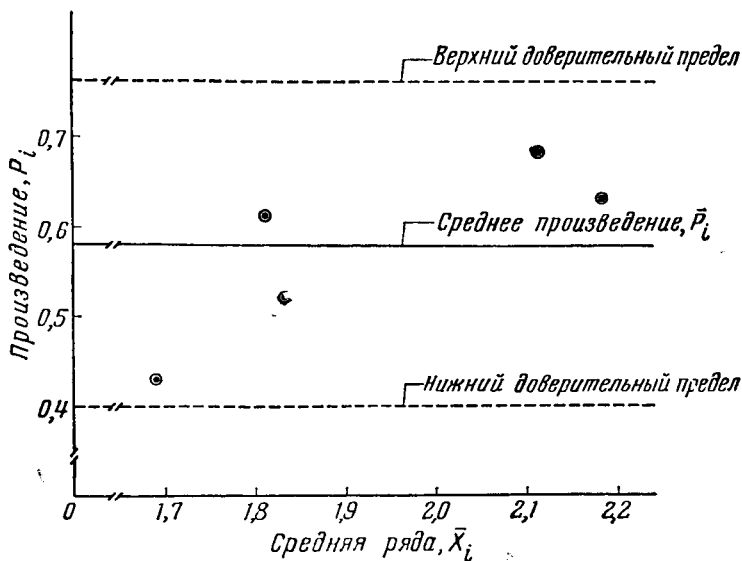


Рис. 24. Произведение p_i , как функция средней ряда \bar{x}_i . Средняя p_i указана с ее доверительными пределами.

из них окажутся выше и ниже соответствующих границ доверительного интервала. Это является образцом того, как при помощи графика можно установить необходимость преобразования данных при помощи квадратных корней или логарифмов. Логарифмическое преобразование в данном случае является более действенным, так как более радикально корректирует влияние неслагаемости.

Иногда может быть так, что по точкам графика нельзя выявить никакой тенденции, но в то же время одна или две точки выйдут из доверительных границ. Это заставляет думать, что данные точки относятся к уклоняющимся наблюдениям, т. е. наблюдениям из некоторой нереальной совокупности. В этом случае следует внимательно рассмотреть (пример 23), не было ли

ошибок в записях или не был ли взят ненормальный экспериментальный материал, и, если это возможно, исправить ошибки или исключить соответствующие данные. Если не будет найдено надлежащего объяснения этому факту, то экспериментатор сталкивается с довольно трудной задачей. Такие отклоняющиеся от нормы наблюдения могут оказаться в стороне в связи с тем, что они относятся к восторженной совокупности, что снижает качество информации, которую дает опыт, но они могут быть восприняты и как резко уклоняющиеся члены изучаемой совокупности.

Дальнейшие сведения о проверке свойства неслагаемости можно найти в [3].

Критерий слагаемости для латинского квадрата [22] требует модификации соответствующей линейной модели и проведения этого процесса дальше, чем это было необходимо при схеме рандомизированных блоков. Данные таблицы 133 взяты из опыта с обезьянами [5]. Наблюдения велись над частотой реагирования на слуховые и (или) зрительные стимулы, подчиненные 5 различным условиям. Для того чтобы избежать влияния корреляции между ошибками (e_{ijh}), периоды наложения этих стимулов были случайным порядком распределены между 5 неделями, так что каждая пара обезьян получала один определенный стимул в каждую неделю.

При рассмотрении данных было установлено, что стандартное отклонение частоты реакции было почти пропорциональным средней, и поэтому данные были преобразованы в логарифмы. Каждый показатель, внесенный в таблицу, является логарифмом геометрической средней из частот каждого члена пары. Возникает вопрос: была ли этим путем достигнута слагаемость эффектов?

ТАБЛИЦА 133

Логарифмы геометрических средних из данных о частоте реакции у отдельных пар обезьян на 5 видов стимулирования. Критерий слагаемости в латинском квадрате

Пары обезьян	Недели					Сумма $X_{i..}$	Средняя $\bar{x}_{i..}$	$\frac{\bar{x}_{i..}}{\bar{x}_{...}}$
	1	2	3	4	5			
1	B 4,99	D 2,25	C 2,48	A 2,48	E 2,51	11,11	2,222	0,007
	2,20	2,27	2,22	2,08	2,52			
	-0,03	-0,02	-0,04	0,10	-0,01			
2	D 2,00	B 1,85	A 1,79	E 2,14	C 2,31	10,09	2,018	-0,197
	1,95	1,93	1,85	2,15	2,21			
	0,05	-0,08	-0,06	-0,01	0,10			
3	C 2,47	A 2,10	E 2,34	B 2,20	D 2,30	11,21	2,242	0,027
	2,43	2,08	2,35	2,18	2,47			
	0,04	0,02	-0,01	0,02	-0,07			
4	E 2,41	C 2,47	B 2,44	D 2,53	A 2,44	12,29	2,458	0,243
	2,45	2,46	2,37	2,53	2,48			
	-0,04	0,01	0,07	0,00	-0,04			
5	A 1,85	E 2,32	D 2,21	C 2,05	B 2,25	10,68	2,136	-0,079
	1,87	2,25	2,17	2,16	2,23			
	-0,02	0,07	0,04	-0,11	0,02			
Сумма $X_{.j.}$	10,42	10,99	10,96	11,10	11,91	$X_{...} =$ $=55,38$		
Средняя $\bar{x}_{.j.}$	2,084	2,198	2,192	2,220	2,382		$\bar{x}_{...} =$ $=2,245$	
$\frac{\bar{x}_{.j.}}{\bar{x}_{...}}$	-0,131	-0,017	-0,023	0,005	0,167			

Варианты	Сумма $X_{..k}$	Средняя $\bar{x}_{..k}$	Отклонение $\bar{x}_{.jk} - \bar{x}_{..k}$
A	10,36	2,072	-0,143
B	10,73	2,146	-0,069
C	11,18	2,236	0,021
D	11,39	2,278	0,063
E	11,72	2,344	0,129

Сначала произведено заполнение окаймляющих таблицу граф средними и отклонениями от общей средней; эти показатели для вариантов вынесены в нижний конец таблицы.

Далее применена линейная модель

$$\hat{X}_{ijk} = \bar{x}_{...} + (\bar{x}_{i..} - \bar{x}_{...}) + (\bar{x}_{.j.} - \bar{x}_{...}) + (\bar{x}_{..k} - \bar{x}_{...}).$$

Таким образом, регрессия оценки складывается из: 1) общей средней опыта $\bar{x}_{...}$; 2) отклонения данного ряда $(\bar{x}_{i..} - \bar{x}_{...})$; 3) отклонения данного столбца $(\bar{x}_{.j.} - \bar{x}_{...})$ и 4) отклонения варианта $(\bar{x}_{..k} - \bar{x}_{...})$. Например, для первого наблюдения:

$$2,215 + 0,007 - 0,131 - 0,069 = 2,02.$$

Эти значения, вычисленные по регрессии, строго подчиняются линейной модели, что является их характерной чертой; если их подвергнуть дисперсионному анализу, то сумма квадратов ошибки должна равняться нулю.

Третьим этапом является вычисление разностей по третьей строке каждой клетки. Эти разности обладают следующими свойствами.

1) Сумма их по каждому ряду, столбцу или варианту равна нулю. Если сюда вкравлись ошибки округления, то должны быть произведены небольшие изменения для доведения этих сумм до нуля.

2) Сумма их квадратов является суммой квадратов «ошибки» в дисперсионном анализе, т. е. $SSE = 0,0706$. Если провести этот дисперсионный анализ непосредственно, избегая ошибок округления, то будет получен более точный результат 0,0725 (пример 28).

На четвертом этапе вычисляются 25 разностей $\hat{X}_{ijk} - \bar{x}_{...}$, каждая из них возводится в квадрат, умножается на 1000, и результат вносится в таблицу 134. Например:

$$1000 (\hat{X}_{112} - \bar{x}_{...})^2 = 1000 (2,02 - 2,215)^2 = 38.$$

Для кодирования здесь можно применить любой удобный множитель или делитель; кодированные квадраты входят в приводимой ниже формуле как в числитель, так и в знаменатель, и поэтому кодирование не меняет результаты.

ТАБЛИЦА 134

Квадраты отклонений, определенных по регрессии значений от общей средней опыта, умноженные на 1000

B 38	D 3	C 0	A 18	E 93
D 70	B 81	A 133	E 4	C 0
C 7	A 18	E 18	B 1	D 65
E 55	C 60	B 24	D 99	A 70
A 119	E 1	D 2	C 3	B 0

Различия = 22 216

На пятом этапе по этой новой таблице определяется сумма квадратов «различий» ($SSD = 22216$) методом дисперсионного анализа или путем подбора линейной модели с последующим суммированием квадратов отклонений.

Затем определяется сумма произведений данных таблицы 134 на отклонения таблицы 133:

$$P = (-0,03) \times 38 + (-0,02) + 3 + \dots + 0,02 \times 0 = -20,68.$$

Далее вычисляется сумма квадратов (с одной степенью свободы) для неслагаемости:

$$SSN = P^2/SSD = (-20,68)^2/22216 = 0,0192.$$

Наконец, проводится дисперсионный анализ:

ошибка SSE	12	0,0725	
неслагаемость SSN	1	0,0192	0,0192
для проверки	11	0,0533	0,0048

Так как здесь $F = 4,00$ меньше 5%-ного уровня 4,84, то условие слагаемости может быть принято. Выводы из этого опыта находятся вне критической области (примеры 28—30 этой главы), и поэтому даже некоторое небольшое нарушение слагаемости не может их изменить.

Пример 23. В параграфе 9 главы 2 и в примере 2 этой главы мы изучали данные о количестве мест поражения на двух половинах листьев, зараженных двумя препаратами вируса табачной мозаики. Там был поставлен вопрос о допущении нормальности распределения. Хотя наличие некоторого отклонения от нормальности не имеет столь большого значения, но обычно с ним связана неслагаемость эффектов. Для проверки данных о вирусе относительно условия слагаемости отведем под варианты 2 столбца, а под повторения 8 рядов. Вы придете к таким результатам:

различия	7	65	
неслагаемость	1	38	38
для проверки	6	27	4,5

На графике эти 7 точек не дают никакой определенной тенденции, но имеется одна из них, которая оказалась выше верхней доверительной границы. Эта точка соответствует паре наблюдений 31,18. График дан в указанной ранее статье Тьюки.

Пример 24. Примените преобразование $\sqrt{X+1}$ к данным о заражении вирусом. Несмотря на то, что F оказалось несущественным, одна из точек на новом графике лежит выше верхней доверительной границы. Первоначальное наблюдение 31 представляется исключительно большим. Если вы примените логарифмическое преобразование к этим данным о вирусе табачной мозаики, то вам удастся полностью исключить неслагаемость.

Пример 25. В примере 1 этой главы оценка ошибки не производилась, и за основу была взята модель рендомизированных блоков. Для полного выяснения вопроса примените здесь критерий слагаемости. *Ответ:* для неслагаемости $F = 5,66$.

Пример 26. Используя отклонения $\bar{x}_i - \bar{x}_..$ и $\bar{x}_j - \bar{x}_..$ совместно с общей средней $\bar{x}_..$ вычислите по уравнению регрессии оценки для 15 клеток таблицы 132 и после этого отклонения $X_{ij} - \hat{X}_{ij}$. Сумма квадратов этих последних должна быть $0,1381 = SSE$ (исключая ошибки округления).

Пример 27. После заполнения таблицы 125, имеющей два вычавших наблюдения, определите по уравнению регрессии 16 оценок для всех наблюдений и соответствующие отклонения. Отклонения для клеток 111 и 324 должны быть равны нулю (если не считать погрешности округления), а сумма квадратов этих отклонений по остальным 14 клеткам должна быть равна $5,33 = SSE$. Два отклонения, равные нулю, соответствуют двум степеням свободы, исключаемым из «ошибки».

Пример 28. Проведите дисперсионный анализ по преобразованным в логарифмы данным о реакции обезьян на стимулы. Вы получите:

пары обезьян	4	0,5244	0,1311
недели	4	0,2294	0,0574
стимулы	4	0,2313	0,0578
ошибка	12	0,0725	0,00894

Пример 29. Опишите все различия между средними таблицы 136, применив метод последовательного сравнения, описанный в параграфе 6 главы 10. *Ответ:* E и D больше A и B ; C больше A .

Пример 30. Обратите логарифмы средних по вариантам таблицы 136 в первоначальные численности. *Ответ:* $x_{12} = 220,8$ и т. д.

Пример 31. Вычислите сумму квадратов, относящуюся к регрессии логарифма частоты реакции на порядковый номер недели. Для этого последнего удобно ввести код x : $-2, -1, 0, 1$ и 2 . После этого, взяв средние по неделям в качестве Y , найдем $\sum xy = 0,618$ и $(\sum xy)^2 / \sum x^2 = 0,03819$. При переходе к основным единицам измерения эта сумма квадратов регрессии будет $5 \times 0,03819 = 0,1910$. Теперь строка дисперсионного анализа для «неделей» в примере 28 разделяется на две части:

линейная регрессия . . .	1	0,1910	0,1910
отклонения от регрессии	3	0,0384	0,0128

При сравнении среднего квадрата с ошибкой видно, что отклонения от регрессии не являются существенными и что большая часть суммы квадратов по строке «недели» обусловлена этой регрессией.

Л И Т Е Р А Т У Р А

1. Allen F. E., Wishart J., *Journal of Agricultural Science*, 20, 399, 1930.
2. Bartlett M. S., *Supplement to the Journal of the Royal Statistical Society*, 3, 68, 1936.
3. Bartlett M. S., *Biometrics*, 3, 39, 1947.
4. Блисс К. Защита растений, № 12, Ленинград, 1937.
5. Butler R. A., *Journal of Experimental Psychology*, 48, 19, 1954.
6. Cochran W. G., *The Empire Journal of Experimental Agriculture*, 6, 157, 1938.
7. Cochran W. G., Autrey K. M., Cannon C. Y., *Journal of Dairy Science*, 24, 937, 1941.
8. Cochran William G., Cox Gertrude M., *Experimental Designs*. John Wiley and Sons, Inc., New York, 1950.
9. Crall James M., Данные Айовской с.-х. опытной станции, 1949.
10. Fisher R. A., *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935—1951.
11. Fisher R. A., *Biometrics*, 10, 130, 1954.
12. Fisher R. A., Yates F., *Statistical Tables*. Oliver and Boyd, Edinburgh, 1938—1953.
13. Forster H. C., Vasey A. J., *Journal of the Department of Agriculture of Victoria, Australia*, 30, 35, 1932.
14. Kempthorne Oscar, *Design and Analysis of Experiments*. John Wiley and Sons, Inc., New York, 1952.
15. Li H. W., Meng C. J., Liu T. N., *Journal of the American Society of Agronomy*, 28, 1, 1936.
16. Lucas H. L. Цит. по [18].
17. Monselise S. P., *Palestine Journal of Botany*, 8, 1, 1951.
18. Patterson H. D., *Journal of Agricultural Science*, 40, 375, 1950.
19. Porter R. H., *Cooperative Soybean Seed Treatment Trials*, Iowa State College Seed Laboratory, 1936.
20. Richardson D. et al. См. ссылку 25, гл. 10.
21. Tukey J. W., *Biometrics*, 5, 232, 1949.
22. Tukey J. W., *Queries in Biometrics*, 10, 562, 1954.
23. Tukey J. W., *Queries in Biometrics*, 11, 111, 1955.
24. Williams C. B., *Bulletin of Entomological Research*, 42, 513, 1951.
25. Winsor C. P., Clarke G. L., *Sears Foundation, Journal of Marine Research*, 3, 1, 1940.
26. Yates F., *The Empire Journal of Experimental Agriculture*, 1, 129, 1933.

СРАВНЕНИЯ. ФАКТОРИАЛЬНЫЕ РАСПОЛОЖЕНИЯ ВАРИАНТОВ

1. Сравнения. Сравнениями называются разности между двумя средними и группами средних. С ними мы уже встречались много раз, причем в одних случаях это были заранее намечаемые сравнения, в других случаях подвергались рассмотрению все возможные разности. В настоящей главе наше внимание будет сосредоточено на тех видах сравнений, которые планируются заранее. В первых трех параграфах будут рассмотрены некоторые свойства таких сравнений и будут даны наиболее удобные методы соответствующих вычислений.

Рассмотрим некоторые заранее намеченные сравнения, которые были проведены нами ранее. В опыте с витамином В₁₂ таблицы 92 мы сначала сравнивали средние группы по вариантам:

$$\bar{x}_{10} - \bar{x}_3 = 1,58 - 1,54 = 0,04 \text{ фунта за день.}$$

Это же было сделано и с суммами:

$$S_{10} - S_3 = 4,74 - 4,62 = 0,12.$$

На суммах не сказываются погрешности округлений, которые часто включаются в средние; вместе с тем использование сумм приводит к более удобным формулам.

Второе сравнение производилось между контролем и вариантами, взятыми вместе:

$$\frac{S_{10} + S_3}{2} - S_0 = \frac{4,74 + 4,62}{2} - 2,73 = 1,95.$$

Это сравнение обычно производится при удвоении обеих сравниваемых количеств:

$$S_{10} + S_3 - 2S_0 = 4,74 + 4,62 - 2 \times 2,73 = 3,90.$$

В опыте с семенами сои в таблице 114 было произведено сравнение контроля со средней для четырех химикатов. Теперь это сравнение будет таким:

$$S_1 + S_2 + S_3 + S_4 - 4S_0 = 459 + 459 + 467 + 471 - 4 \times 446 = 82.$$

Для сохранения результатов прямого расчета вы должны от этой учетверенной разности перейти к сравнению средней из сумм 4 вариантов с суммой контроля; кроме того, так как в опыте 5 повторений, то число 82 является $4 \times 5 = 20$ -кратным результатом сравнения средней из вариантов с контролем, т. е. $(93,8 + 94,8 + 93,4 + 94,2) / 4 = 89,2 = 4,1$ растения.

Прежде чем идти далее, полезно познакомиться с некоторыми правилами, которым подчиняются сравнения и суммы квадратов, основанные на них.

2. Правила для проведения сравнений при одинаковой повторности. В этом параграфе и в большинстве последующих параграфов считается, что все суммы основаны на одном и том же числе повторных наблюдений n . Нам придется иметь дело с линейными связями между суммами

$$c = \lambda_1 S_1 + \dots + \lambda_a S_a,$$

где a число вариантов, входящих в сравнение, и λ_i являются коэффициентами, с которыми суммы входят в сравнение. Например, в опыте с витамином B_{12} сравнение $S_{10} - S_5$ имеет $\lambda_1 = 1$ и $\lambda_2 = -1$, сравнение $S_{10} + S_5 - 2S_0$ имеет $\lambda_1 = 1$, $\lambda_2 = 1$ и $\lambda_3 = -2$.

П р а в и л о 1. Величина $c = \sum \lambda_i S_i$ является сравнением, если $\sum \lambda_i = 0$.

Это означает, что нет никакого смысла в сопоставлении, например, $S_1 + S_2$ с S_3 ; интерес будет представлять сравнение $S_1 + S_2$ с $2S_3$ или с $S_3 + S_4$. В терминах коэффициентов λ_i эти два сравнения могут быть представлены так:

$$\begin{array}{cccc} +1 & +1 & -2 & \Sigma \lambda_i = 0 \\ +1 & +1 & -1 & -1 & \Sigma \lambda_i = 0 \end{array}$$

П р а в и л о 2. Если c является некоторым сравнением между S , то величина

$$\frac{c^2}{n \Sigma \lambda^2}$$

является частью суммы квадратов для вариантов, которой соответствует единичная степень свободы.

Сопоставим это правило с тем, которое было применено в параграфе 8 главы 10. Для суммы квадратов сравнения между двумя дозами витамина мы теперь имеем:

$$\frac{(4,74 - 4,62)^2}{3 \times 2} = 0,0024.$$

Сравнение между вариантами, взятыми вместе, и контролем дает

$$\frac{(4,74 + 4,62 - 2 \times 2,73)^2}{3 \times 6} = 0,8450.$$

В тот и другой знаменатель здесь входит 3 — число повторений. Второй множитель знаменателя является суммой $\Sigma \lambda^2$ $1^2 + (-1)^2 = 2$ в первом сравнении и $1^2 + 1^2 + (-2)^2 = 6$ во втором.

П р а в и л о 3. Два сравнения

$$\begin{aligned} c_1 &= \lambda_{11} S_1 + \lambda_{12} S_2 + \dots + \lambda_{1a} S_a \\ c_2 &= \lambda_{21} S_1 + \lambda_{22} S_2 + \dots + \lambda_{2a} S_a \end{aligned}$$

ортогональны, если

$$\lambda_{11} \lambda_{21} + \lambda_{12} \lambda_{22} + \dots + \lambda_{1a} \lambda_{2a} = 0.$$

При применении этого правила для суммы S_i , которая не входит в данное сравнение, следует за коэффициент брать нуль.

Что два сравнения в опыте с витамином B_{12} , приведенных ранее, ортогональны, можно установить при помощи коэффициентов:

$$\begin{array}{ccc} S_{10} & S_5 & S_0 \\ c_1 & +1 & -1 & 0 \\ c_2 & +1 & +1 & -2 \end{array}$$

где $(+1) \times (+1) + (-1) \times (+1) + (0) \times (-2) = 0$.

П р а в и л о 4. При a вариантах, если составленные из них $(a-1)$ сравнения взаимно ортогональны, существует равенство:

$$\frac{c_1^2}{n \Sigma \lambda_1^2} + \frac{c_2^2}{n \Sigma \lambda_2^2} + \dots + \frac{c_{a-1}^2}{n \Sigma \lambda_{a-1}^2} = \text{сумме квадратов для вариантов.}$$

Это правило можно иллюстрировать на примере сумм квадратов для двух ортогональных сравнений средних в опыте с витамином В₁₂:

$$0,0024 + 0,8450 = 0,8774,$$

что представляет собой сумму квадратов для вариантов в таблице 92.

Удобным способом применения этих правил является проведение вычислений по схеме, показанной на примере опыта с витамином В₁₂.

Дозы В ₁₂	0	5	10			
Суммы	2,73	4,62	4,74	Сравне- ние	Дели- тель	Сумма квadra- тов
В ₁₂ с контролем	-2	+1	+1	3,90	18	0,8450
5 мг с 10 мг	0	-1	+1	0,12	6	0,0024
Сумма						0,8774

П р а в и л о 5. Если имеется k сравнений c_i одного и того же вида (например, линейные сравнения), то сумма $\sum c_i$ равна сравнению c_S соответствующих сумм вариантов. Вместе с этим сумма квадратов c_S равна

$$\frac{(\sum c_i)^2}{kn \sum \lambda^2},$$

а сумма квадратов для взаимодействия между этими k сравнениями будет

$$\frac{\sum c_i^2}{n \sum \lambda^2} - \frac{(\sum c_i)^2}{kn \sum \lambda^2}.$$

Применение этого и следующего за ним правила будет показано в параграфе 8 этой главы.

П р а в и л о 6. После того как произведено одно сравнение c_i , можно провести вторичное сравнение c_j значений c_i . Сумма квадратов этого второго сравнения будет:

$$\frac{c_j}{n (\sum \lambda_i^2) (\sum \lambda_j^2)}.$$

Пример 1. По данным таблицы 114 проверить методами настоящего параграфа сумму квадратов для сравнения контроля со всеми химикатами — 67,24.

Пример 2. В следующей таблице приведены данные о привесе (в г) 6 групп самцов крыс в опыте с рендомизированными блоками.

Высокая доза протеина			Низкая доза протеина		
мясо	верно	сало	мясо	верно	сало
73	98	94	90	107	49
102	74	79	76	95	82
118	56	96	90	97	73
104	111	98	64	80	86
81	95	102	86	98	81
107	88	102	51	74	97
100	82	108	72	74	106
87	77	91	90	67	70
117	83	120	95	89	61
111	92	105	78	58	82
Источник варьирования			Число степеней свободы	Сумма квадратов	Средний квадрат
Варианты			5	4613	923
Ошиб.			54	11586	214,6

Здесь могут быть произведены следующие сравнения:

Сумма весов	Много протеина			Мало протеина		
	мясо	верно	сало	мясо	верно	сало
	1000	859	995	792	839	787
1. Высокая доза протеина с низкой дозой	+1	+1	+1	-1	-1	-1
2. Мясо с салом. Высокая доза протеина	+1	0	-1	0	0	0
3. Мясо с салом. Низкая доза протеина	0	0	0	+1	0	-1
4. Корм животного происхождения с растительным. Высокая доза протеина . .	+1	-2	+1	0	0	0
5. Корм животного происхождения с растительным. Низкая доза протеина . .	0	0	0	+1	-2	+1

Проверьте взаимную ортогональность этих сравнений. Вычислите все 5 сравнений. Ответ: 436, 5, 5, 277, -99. Вычислите сумму квадратов для каждого сравнения, после чего проверьте равенство суммы этих пяти сумм квадратов с суммой квадратов для вариантов 4613. Системы других сравнений, представляющих больший практический интерес, будут указаны в примерах 4 и 8 этой главы.

3. Правила для построения сравнений при различной повторности.

П р а в и л о 1. Величина $c = \sum \lambda_i S_i$ является сравнением, если $\sum n_i \lambda_i = 0$.

П р а в и л о 2. Если c это некоторое сравнение сумм S , то

$$c^2 / \sum n_i \lambda_i^2$$

является суммой квадратов этого сравнения.

П р а в и л о 3. Два сравнения c_1 и c_2 ортогональны, если

$$\sum n_i \lambda_{1i} \lambda_{2i} = 0.$$

П р а в и л о 4 остается без изменения.

4. Факториальное расположение вариантов. Опыт 2×2 или 2^2 .

Имея два или более ряда вариантов, можно провести их изучение во всех комбинациях между собой. Такие ряды называются *факторами*, а осуществление всех их комбинаций известно под названием *факториального расположения* их. Широким распространением пользуется термин *факториальный опыт*.

Факториальный опыт строится таким образом, что контролируются все, а не один фактор. В контролируемом опыте изучается действие одного фактора при некоторых определенных условиях. Если есть основание ожидать, что действие его меняется при изменении этих условий, то данный фактор может быть испытан при нескольких таких условиях. Ряд таких условий становится вторым фактором. Так, ряд сортов (первый фактор) может быть испытан в нескольких местах (второй фактор), а это двухфакторное сочетание может быть повторено в течение ряда лет (третий фактор). Такое факториальное построение опыта дает сведения не только относительно поведения каждого сорта в каждом месте и в каждом году, но также и о различиях в поведении при переходе от одного пункта к другому и от одного года к другому. Один какой-нибудь сорт может занять преимущественное положение в одном пункте и уступить его другим сортам в другом месте. Некоторые сорта могут быть одинаково продуктивными при всем разнообразии сезонных условий, в то время как другие будут резко реагировать на условия погоды.

Каждый фактор может изучаться при двух или большем числе градаций: например, в опыте с удобрением такой фактор, как азот, может изучаться при градациях: нуль (контроль), 20 и 40 фунтов на акр. Пользуясь этой условной терминологией, отдельные сорта или годы можно также считать градационными факторами. Два фактора, каждый с двумя градациями, образуют факториальный опыт 2×2 или 2^2 . Когда изучаются f факторов, каждый из которых имеет k градаций, то это может быть изображено, как k^f ; число факторов здесь стоит в качестве показателя степени. При разном числе градаций построение опыта изображается в виде произведения; например, опыт 3×4 включает в себя 2 фактора, один с 3, а другой с 4 градациями.

В таблице 135 приведены несколько измененные результаты факториального опыта 2^2 , проведенного по полностью рандомизированной схеме; на этот опыт мы ссылались уже несколько раз [20]. При отсутствии антибиотиков витамин практически не дает никакого эффекта, по-видимому, в связи с деятельностью флоры кишечника, усваивающей B_{12} ; но при использовании антибиотиков, воздействующих на эту флору, эффект витамина становится заметным. Данную таблицу можно рассматривать и под другим углом зрения: антибиотики, взятые в отдельности, уменьшают привес, вероятно, вследствие подавления деятельности флоры, синтезирующей B_{12} , но при добавлении B_{12} антибиотики, уменьшая в первую очередь активность вредной флоры, дают уже прямой эффект.

Здесь мы встречаемся с особым видом эффекта, на обнаружение которого направлено факториальное построение опыта. Этот эффект называется *взаим-*

ТАБЛИЦА 135

Факториальный опыт с антибиотиками и витамином B_{12}

Средний привес свиней (в фунтах) за день

Градация антибиотиков (в мг)	Градация B_{12} (в мг)			Итого по антибиотикам
	0	5		
0		1,30	1,26	7,23
		1,19	1,21	
		1,08	1,19	
	Сумма	3,57	3,66	
	Средняя	1,19	1,22	
40		1,05	1,52	7,73
		1,00	1,56	
		1,05	1,55	
	Сумма	3,10	4,63	
	Средняя	1,03	1,54	
Итого по B_{12}		6,67	8,29	14,96

Дисперсионный анализ¹

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Варианты	3	0,4124	
Ошибка	8	0,0293	0,00366

¹ Применены методы главы 10, полностью рандомизированный опыт с 4 группами, по 3 наблюдения в каждой.

модействием. Его наличие нарушает слагаемость основных эффектов; смысл взаимодействия таков: то, что добавляется одним фактором при первой гра-

дации другого фактора, отличается от того, что добавляется этим же первым фактором при другой градации второго фактора.

Взаимодействие измеряется разностью между двумя разностями, т. е. отклонением этих разностей от полного равенства. Для свиней при двух дозах V_{12} имеем:

разность для 40 мг антибиотиков	1,54—1,03=0,51
разность при отсутствии антибиотиков	1,22—1,19=0,03

взаимодействие, равное разности 0,48

Подобно этому, при двух градациях антибиотиков:

разность для 5 мкг V_{12}	1,54—1,22=0,32
разность при отсутствии V_{12}	1,03—1,19=—0,16

взаимодействие, равное разности 0,48

Как будет видно из дальнейшего, взаимодействие в этом опыте не может быть приписано просто выборочному варьированию.

Понятие о взаимодействии различно у разных лиц. Химик отождествляет его с реакцией. Физиолог, работающий с гормонами, думает о взаимном стимулировании между основными действующими началами. Агроном может интересоваться эффектом одного вида питательного вещества для растения в условиях надлежащего подбора второго вида. Для статистика взаимодействие является мерой нарушения слагаемости у эффектов факторов.

Арифметически взаимодействие в таблице 2² является сравнением таких сумм:

высшая градация+низшая градация обоих факторов	1,54+1,19=2,73
(высшая—низшая) + (низшая—высшая) градации	1,03+1,22=2,25

взаимодействие=разность 0,48

Сумма квадратов для этого сравнения вычисляется, как обычно, на основе соответствующих сумм:

$$\frac{(4,63+3,57-3,66-3,10)^2}{3 \times 4} = \frac{1,44^2}{12} = 0,1728,$$

где $n=3$ и $\Sigma \lambda^2=4$. Разность, стоящая в числителе, является n -кратным взаимодействием: $3 \times 0,48=1,44$.

Теперь средний квадрат 0,1728 с одной степенью свободы является величиной, сопоставимой со средним квадратом ошибки из таблицы 135.

$$F = 0,1728/0,00366 = 47, \quad f = 1 \text{ и } 8.$$

Отсюда ясно, что эти два вещества имеют совместный эффект, отличающийся от суммы их эффектов, когда они применяются раздельно.

При наличии взаимодействия обычно мало что можно сказать по поводу главных эффектов каждого отдельного фактора. Например, нет никакого общего эффекта от витамина V_{12} . Вместо этого имеется два отдельных эффекта — один без антибиотиков: 1,22—1,19=0,03 и другой при антибиотиках: 1,54—1,03=0,51. Сумма, или средняя из этих эффектов, не относится ни к каким определенным условиям.

Каждый из этих отдельных эффектов может быть сопоставлен с экспериментальной ошибкой. Так, для эффекта витамина при антибиотиках имеем:

$$\frac{(4,63-3,10)^2}{3 \times 2} = 0,3902.$$

В данном случае F будет $0,3902/0,09366 = 416,6$, $f = 1$ и 8. Это значение выходит за пределы 5%-ного уровня $F_{0,05} = 14,69$. Следует обратить внимание на то, что имеющимся в данном случае четырем таким частным эффектам соответствует две степени свободы (так как одна из трех степеней свободы отошла к взаимодействию). Отсюда следует, что соответствующие четыре сравнения не будут независимыми или ортогональными и что указанная оценка каждого из них в отдельности не будет точной.

Несмотря на то, что в этом случае смысл главных эффектов становится сомнительным, все же средние квадраты их определяются даже при наличии взаимодействия. Это вошло в традицию практики, часто подвергающуюся критике, но почти неизбежную (см. ниже). Читатель должен разобраться во всем этом вопросе.

ТАБЛИЦА 136

Выявленная концентрация рибофлавина (мг/г) в листьях коларда при двух размерах пробы (при обработке и без обработки перекисью перманганата)

Повторения	Перманганат		Без перманганата		Сравнения	Делитель	Сумма квадратов
	проба 0,25 г	проба 1,00 г	проба 0,25 г	проба 1,00 г			
1	27,2	24,6	39,5	38,6			
2	23,2	24,2	43,1	39,5			
3	24,8	22,2	45,2	33,0			
Сумма	75,2	71,0	127,8	111,1			
Перманганат . . .	-+1	-+1	+1	+1	92,7	12	716,11
Размер пробы . . .	+1	-+1	-+1	-+1	20,9	12	36,40
Взаимодействие	-+1	+1	+1	-+1	12,5	12	13,02

Итого 765,53

Источники варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Повторения	2	3,76	
Варианты	(3)	(765,53)	
Перманганат	1	716,11	716,11**
Размер выборки	1	36,40	36,40
Взаимодействие	1	13,02	13,02
Ошибка	6	49,08	8,18

**—означает $P \leq 0,01$. Другие эффекты не будут существенными на 5%-ном уровне.

Определение возможного взаимодействия часто производится в виде некоторой страховки. Оно может привести к выводу, что взаимодействие отсутствует. Но даже в этом случае факториальное построение опыта является по крайней мере столь же эффективным, как два опыта, в которых, положим, витамин будет изучаться при каждой градации антибиотиков. Соединение таких опытов вместе уменьшает риск искажений в результате резко выраженного варьирования окружающих условий.

Случай, когда не проявляется никакого взаимодействия, иллюстрируется в таблице 136 опытом [28] по изучению влияния обработки перекисью перманганата на выявленную концентрацию рибофлавина в листьях коларда (вид капусты). Здесь была осуществлена схема рандомизированных блоков,

повторенных в 3 последующие дня. Изучались два фактора: перманганат и размер пробы. Первоначальный анализ проводится методами главы 11.

Вслед за выписанными в этой таблице экспериментальными данными произведено вычисление трех ортогональных сравнений между четырьмя вариантами. Сложение трех соответствующих сумм квадратов дает сумму квадратов для «вариантов» в первоначальном дисперсионном анализе. Замечание: каждый коэффициент для взаимодействия является произведением двух вышестоящих коэффициентов.

Здесь нет достаточно выраженного проявления взаимодействия, и поэтому можно приступить прямо к исследованию главных эффектов факторов. Отметим, что каждый из них является сравнением, получаемым путем сложения двух частных эффектов. Например, главный эффект перманганата может вычисляться так: $(127,8 - 75,2) + (111,1 - 71,0) = 92,7$. Таким образом, при отсутствии взаимодействия сравнение для «перманганата» строится на основе всех данных опыта: 6 одних наблюдений сравниваются с 6 другими. Сравнение «размер пробы» подобным же путем вычисляется по всем 12 наблюдениям. «Отсутствие взаимодействия» означает, что к главным эффектам применимо свойство слагаемости, т. е. что они независимы.

Вывод из данного опыта был таким: «При флуорометрическом определении рибофлавина в стандартной пробе сухого вещества колларда является существенным этапом очищение гидрогеном перекиси перманганата. Без этого среднее значение равно 39/8 мкг, в то время при применении очищения получается более приемлемая средняя 24,4».

Пример 3. В опыте с сахарной свеклой, проведенном по схеме рендомизированных блоков, число сохранившихся растений по делянкам опыта было таким:

Варианты	Блоки			
	1	2	3	4
Без удобрения	183	176	291	254
Суперфосфат P	353	300	301	271
Калий K	224	258	244	217
P + K	329	283	308	326

Численности здесь столь велики, что преобразование этих данных не является необходимым. Вычислите средний квадрат ошибки. *Ответ:* 1494. Кроме того, вычислите и следующие средние квадраты: P=2457, K=203 и PK (взаимодействие) — 28. Остается непонятным, почему суперфосфат столь сильно влияет на число растений. Возможно, что повышенная энергия прорастания или более раннее появление всходов позволяют избежать повреждение растений некоторыми почвенными организмами в наиболее опасную фазу их развития (сравните с параграфом 6 главы 13).

Пример 4. Обработайте данные полностью рендомизированного опыта примера 2 этой главы, взяв только два варианта: мясной и растительный корм при высшей и низшей дозе протеина. Вы здесь найдете, что средний квадрат ошибки 223,6 при среднем квадрате взаимодействия 884, т. е. он выходит за 5%-ный уровень. По-видимому, крысы могут использовать повышенную норму мяса для увеличения своего веса, но не могут усвоить добавочное зерно.

Пример 5. В таком опыте, как опыт, приведенный в таблице 135, обычно, несмотря на наличие взаимодействия, все же проводится полный дисперсионный анализ. В данном случае он будет таким:

	Число степеней свободы	Средний квадрат
Антибиотики	1	0,0208
Витамин	1	0,2187
Взаимодействие	1	0,1728
Ошибка	8	0,00366

Произвести проверку средних квадратов для главных эффектов факторов.

Пример 6. Хоблин и Пальмер [14], работая с черенками сливы, брали для испытания две длины (6 и 12 см) их и два диаметра (3—6 и 9—12 мм); 240 саженцев каждого вида были высажены в момент их срезы, а 240 сохранялись для посадки следующей весной.

	Высаженные сразу			Высаженные весной		
	прижились	не прижились	всего	прижились	не прижились	всего
Длинные	156	84	240	84	156	240
Короткие	107	133	240	31	209	240

Как указывалось в параграфе 10 главы 9, в данном случае интерес представляет процент прижившихся саженцев:

65,0	35,0
44,6	12,9

Преобразуйте эти проценты в углы (табл. 129) и проведите дисперсионный анализ, но помните, что углы являются средними, а не суммами и что поэтому квадраты должны умножаться, а не делиться на 240. *Ответ:* средние квадраты: для времени посадки 88059, длины саженцев 43902, для взаимодействия 690.

Теоретически установлено, что преобразование в арксинусы, примененное к биномиальному распределению, имеет определенную дисперсию ошибки: всегда 824 при бесконечном числе степеней свободы (17, стр. 156). Например, для длины саженцев $F=43902/824$, для взаимодействия $F=690/824$.

Пример 7. В предыдущем примере проще не применять умножения на 240, а вместо этого взять за ошибку $824/240$.

5. Двухфакторный опыт. Оставляя теперь частный случай, когда фактор имеет две градации, рассмотрим более общее построение опыта с a градациями первого фактора и b градациями второго фактора, т. е. схему $a \times b$. Эта схема относится как к ранее рассмотренным случаям, так и к большому числу других опытов.

Модель I (с константами) для полностью рандомизированной схемы в данном случае будет такой:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n; \sum \alpha_i = \sum \beta_j = 0; \varepsilon_{ijk} = N(0, \sigma)$$

Взаимодействия $(\alpha\beta)_{ij}$ подчиняются строгим ограничениям: они должны в сумме как по отдельным рядам, так и по отдельным столбцам таблицы $a \times b$ составлять нуль, т. е. $\sum (\alpha\beta)_{i.} = \sum (\alpha\beta)_{.j} = 0$. Эта модель имеет столько взаимодействий, сколько клеток в таблице.

Присутствие взаимодействий нарушает слагаемость главных эффектов α_i и β_j . Это будет выяснено в последующем.

Для полного понимания этой модели лучше всего обратиться к построению на ее основе условного опыта. Представляется удобным использовать готовую конструкцию таблицы 116, считая блоки за градации второго фактора. Для этого сложим три первых числа каждой клетки этой таблицы и внесем эту сумму в соответствующую клетку новой таблицы 137; например, для первой клетки $\mu + \alpha_1 + \beta_1 = 30 + 10 + 1 = 41$. Далее выбираем константы взаимодействия для каждой клетки. Это могут быть любые, какие вы пожелаете, числа, но только их суммы по каждому ряду и по каждому столбцу должны быть равны нулю. Вслед за этим в каждую клетку добавляется два случайных значения ε_{ijk} . В качестве первого из них (слева) я взял ε_{11} в готовом виде из таблицы 116; второе было взято из таблицы 24 также в случайном порядке, как и ранее, в первом случае. Две суммы, указанные в каждой клетке, являются двумя случайными наблюдениями для одной из $a \times b$ комбинаций вариантов. Суммы из двух этих наблюдений, являющиеся суммами по вариантам, для удобства также поставлены в каждой отдельной клетке; после этого они просуммированы для получения краевых итогов.

Теперь мы имеем опыт 4×3 с двумя случайными повторными наблюдениями, которые, соединяясь, дают 12 сумм по вариантам.

На первом этапе анализа применяется метод главы 10. Общая сумма квадратов подразделяется на две части:

варианты	$ab - 1 = 11$	2468	
объекты (ошибка)	12	305	25

Следующий этап состоит в разложении суммы квадратов для вариантов на части, относящиеся к двум главным эффектам A и B и к взаимодействию AB . Последнее вычисляется вычитанием из «вариантов» сумм квадратов для A и B .

Теперь вам станет ясным не всегда правильно понимаемое соотношение между главными эффектами и взаимодействиями. Если вы сложите

$$X_{111} + X_{121} + X_{131} = 26 + 30 + 49 = 105$$

и после этого сравните результат с соответствующей суммой в таблице 116, то вам бросится в глаза тот факт, что взаимодействия не меняют главный эффект, так как $\sum(\alpha\beta)_i = 0$. Представляется естественным думать, что принцип слагаемости всех этих эффектов приемлем. Однако против этого имеется два возражения. Во-первых, биологические взаимодействия сами по себе не приспосабливаются к этой модели; посмотрите на взаимодействие таблицы 135, которое главным образом связано с четвертой клеткой этой таблицы, и вы найдете еще один аргумент против слагаемости эффектов. Математическая модель, приведенная выше, преследует определенную цель: произвести разложение дисперсии, определить компоненты и дать оценку их. Но она не дает описание тех видов взаимодействия, которые обычно представляют интерес для специалиста в данной области. Второе обстоятельство, говорящее против принципа слагаемости, менее явное. Для его объяснения допустим, что первый фактор таблицы 137 относится к 4 дозам удобрения, а второй к 3 сортам. При отсутствии взаимодействия эффекты доз удобрения будут одинаковы по всем сортам; если удобрение a_1 увеличивает урожай сорта b_1 на 10 бушелей на акр, то оно дает то же самое увеличение урожая (если не считать ошибок опыта) и для сортов b_2 и b_3 . Этот главный эффект одинаковым образом накладывается на каждый отдельный эффект сортов. Напротив, при наличии взаимодействия влияние a_1 на урожай сортов будет различно; если взять взаимодействия таблицы 137, то видно, что, в то время как увеличение урожая b_1 составит 10—4 бушелей, увеличение урожая b_3 будет 10+3 бушелей. При соблюдении требования $\sum(\alpha\beta)_i = 0$ главные эффекты в этом случае остаются неизменными, но это не распространяется на частные эффекты.

Если выявлено взаимодействие факторов, то обычно нет оснований для вычисления средних квадратов главных эффектов факторов, хотя это по традиции и производится. (Я это сделал для построения в дальнейшем таблицы 139.) Вместо этого следует выписать в таблицу $a \times b$ суммы (или средние) по вариантам и изучить содержание такой таблицы. Для того чтобы сделать это ясным, я вычислил по таблице 137 средние и внес их в таблицу 138. В качестве примера такого изучения содержания этой таблицы рассмотрим средние двух первых рядов. Если бы здесь не было взаимодействия или ошибки опыта, то все разности $a_1 - a_2$ должны были бы равны $\alpha_1 - \alpha_2 = 10 - 3 = 7$ (табл. 116). В противоположность этому мы в данном случае имеем: $32 - 42 = -10$; $34 - 28 = 6$; $48 - 32 = 16$. Если бы это был фактический опыт, то нарушение полного равенства между этими разностями (оставляя в стороне ошибки опыта) могло бы послужить для экспериментатора поводом к интерпретации его как взаимодействия.

Более наглядный способ изучения взаимодействий состоит в подборе линейной модели (табл. 119) для приведенной выше таблицы 4×3 . В этом случае отклонения дают представление о взаимодействии плюс ошибка опыта. В таблице 139 даны компоненты модели 1, т. е. дисперсии константных эффектов. В условиях нашего искусственно созданного опыта, константы которого известны, могут быть вычислены, как это показано в таблице, ожидаемые значения (параметры) этих компонентов и сравнены с фактическими,

ТАБЛИЦА 1

Двухфакторный опыт, построенный по модели 1. Полностью рандомизированная схема

Вариант a_i Фактор А	Вариант b_j , фактор В			X_{ij}
	b_1	b_2	b_3	
a_1	$\begin{array}{r} 41 \\ -4 \\ \hline 37 \\ -11 \quad 1 \\ \hline 26 \quad 38 \\ \hline 64 \end{array}$	$\begin{array}{r} 35 \\ 1 \\ \hline 37 \\ -7 \quad 2 \\ \hline 30 \quad 39 \\ \hline 69 \end{array}$	$\begin{array}{r} 43 \\ 3 \\ \hline 46 \\ 3 \quad 1 \\ \hline 49 \quad 47 \\ \hline 96 \end{array}$	229
a_2	$\begin{array}{r} 34 \\ 8 \\ \hline 42 \\ 1 \quad -2 \\ \hline 43 \quad 40 \\ \hline 83 \end{array}$	$\begin{array}{r} 29 \\ -5 \\ \hline 24 \\ 5 \quad 4 \\ \hline 29 \quad 28 \\ \hline 57 \end{array}$	$\begin{array}{r} 36 \\ -3 \\ \hline 33 \\ -3 \quad 2 \\ \hline 30 \quad 35 \\ \hline 65 \end{array}$	205
a_3	$\begin{array}{r} 31 \\ -3 \\ \hline 28 \\ 0 \quad 5 \\ \hline 28 \quad 33 \\ \hline 61 \end{array}$	$\begin{array}{r} 26 \\ -1 \\ \hline 25 \\ 4 \quad -13 \\ \hline 29 \quad 12 \\ \hline 41 \end{array}$	$\begin{array}{r} 33 \\ 4 \\ \hline 37 \\ -1 \quad 0 \\ \hline 36 \quad 37 \\ \hline 73 \end{array}$	175
a_4	$\begin{array}{r} 18 \\ -1 \\ \hline 17 \\ -2 \quad 0 \\ \hline 15 \quad 17 \\ \hline 32 \end{array}$	$\begin{array}{r} 13 \\ 5 \\ \hline 18 \\ -2 \quad -1 \\ \hline 15 \quad 17 \\ \hline 33 \end{array}$	$\begin{array}{r} 20 \\ -4 \\ \hline 16 \\ 1 \quad -4 \\ \hline 17 \quad 12 \\ \hline 29 \end{array}$	94
$X_{.j}$	240	200	233	703

Поправка $(\sum X_{ijk})^2/nab = 703^2/24 = 20592$.

Общее: $\sum X_{ijk}^2 - C = 26^2 + 38^2 + \dots + 12^2 - C = 2773$.

Вариантов: $\frac{\sum X_{ij}^2}{n} - C = \frac{64^2 + 69^2 + \dots + 29^2}{2} - C = 2468$.

Ошибка: $2773 - 2438 = 305$.

А: $\frac{\sum X_{i..}^2}{nb} - C = \frac{229^2 + \dots + 94^2}{2 \times 3} - C = 1729$.

В: $\frac{\sum X_{.j}^2}{na} - C = \frac{240^2 + 200^2 + 233^2}{2 \times 4} - C = 254$.

AB: Варианты $- A - B = 24 \cdot 8 = 1729 + 254 = 485$.

Источник варьиро- вания	Число степеней свободы	Сумма квадратов	Средний квадрат
Эффекты вариантов:			
А	3	1729	(576)
В	2	254	(127)
AB	6	485	81*
Ошибка	12	305	25
Общее	23	2773	

Средние вариантов, выражающие взаимодействие.
Данные таблицы 137

Фактор А	Фактор В		
	b ₁	b ₂	b ₃
a ₁	32	34	48
a ₂	42	28	32
a ₃	30	20	36
a ₄	16	16	14

выбранными нами заранее. Выборочные оценки в данном случае оказались необычно близкими к фактическим, в особенности это относится к $s^2=25,4$ (что при округлении дает 25), которая является оценкой $\sigma^2=25$. Оценка существенности здесь проста: средний квадрат ошибки будет служить знаменателем для всех F .

В тех случаях, когда опыт проводится по схеме рандомизированных блоков, в эту модель входят члены, соответствующие блокам:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varrho_k + \epsilon_{ijk}.$$

Предварительный анализ в этом случае проводится методами, указанными в главе 11. После этого он усложняется тем, что из суммы квадратов для вариантов выделяются главные эффекты факторов и взаимодействия, как это было сделано в таблице 137. При компонентном анализе эффекты блоков не представляют интереса и поэтому не определяются.

ТАБЛИЦА 139

Компонентный анализ двухфакторного опыта; модель I. Полностью рандомизированная схема. Данные из таблицы 137

Источник варьирования	Средний квадрат	Оцениваемый параметр
Эффекты вариантов: А	576	$\sigma^2 + nb\kappa_A^2$
В	127	$\sigma^2 + na\kappa_B^2$
АВ	81	$\sigma^2 + n\kappa_{AB}^2$
Ошибка	25	σ^2

$$n=2; a=4; b=3$$

$$\kappa_{AB}^2 = \frac{\sum(\alpha\beta)_{ij}^2}{(a-1)(b-1)} = \frac{(-4)^2 + 1^2 + \dots + (-4)^2}{3 \times 2} = 32$$

$$\kappa_A^2 = 92^2/3, \kappa_B^2 = 13 \text{ (из таблицы 117)}$$

$$s_{AB}^2 = (81 - 25)/2 = 28 \text{ оценивает } 32; s_B^2 = (127 - 25)/8 = 12,8 \text{ оценивает } 13.$$

$$s_A^2 = (576 - 25)/6 = 92 \text{ оценивает } 92^2/3.$$

Компонентный анализ для случая модели II и смешанной модели представлен в таблице 140. Из этого анализа для модели II видно, что при оценке существенности главных эффектов факторов в знаменатель критерия F входит средний квадрат взаимодействия AB . То же самое относится и к фиксированному эффекту смешанной модели, но эффект фактора, имеющий случайный характер, оценивается, как и ранее, на основе среднего квадрата ошибки в качестве знаменателя в критерии F . В настоящей главе мы впервые встречаемся с случаем, когда в качестве знаменателя F берется другой средний квадрат, а не средний квадрат ошибки. Как и в главе 10, критерием для выбора знаменателя F является не обязательное сравнение с ошибкой, а то

положение, что знаменатель должен содержать в себе все компоненты, входящие в один (или нескольких), подлежащего оценке.

ТАБЛИЦА 1

Компонентный анализ двухфакторного опыта при модели II и смешанной модели

	Оцениваемые параметры	
	модель II	смешанная модель, A фиксировано
Эффекты вариантов: A	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$
B	$\sigma^2 + n\sigma_{AB}^2 + na\sigma_B^2$	$\sigma^2 + na\sigma_B^2$
AB	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$
Ошибка	σ^2	σ^2

¹ Если фиксировано B, а A переменное, то σ_{AB}^2 будет компонентом B, но не A.

В параграфе 10 этой главы даны общие правила для построения компонентного анализа для различного вида моделей. Но уже в настоящий момент следует отметить три факта: 1) главный эффект фактора, который является случайной переменной, не включает в себя компонент взаимодействия, потому что сумма взаимодействий в этом направлении равна нулю; 2) коэффициенты и подстрочные указатели, прикрепленные к σ^2 (исключая ошибку опыта), всегда составляют пару (т. е. AB, или b и A, или a и B), причем число этих пар всегда точно определенное, в данном случае 3; 3) иронические буквы обозначают фактор, строчные — число вариантов или повторений; если ироническая буква поставлена подстрочным указателем, то соответствующая строчная буква исключается из коэффициента при σ^2 .

Примечание. Здесь имеется опасность спутать обозначения, так как я не использовал символы *a* и *b*, для обозначения вариантов и в то же время символы *a* и *b* (без подстрочных указателей) для обозначения числа градаций. Но эти обозначения столь удобны и легко воспринимаемы, что, я думаю, можно пойти на этот риск.

Пример 8. Теперь можно взять более подходящий ряд сравнений для опыта с крысами, приведенного в примере 2. Проверьте суммы квадратов, указанные в последней колонке приводимой ниже таблицы, и сравните их сумму с общей суммой квадратов для вариантов в первоначальном примере:

Сравнения	Много протеина			Мало протеина			Сумма квадратов
	мясо 1000	зерно 839	сало 995	мясо 792	зерно 839	сало 787	
Протеин	+1	+1	+1	-1	-1	-1	3168,0
Мясо с салом	+1	0	+1	+1	0	+1	2,5
Взаимодействие с протеином	+1	0	+1	-1	0	+1	0,0
Корм животного происхождения с растительным	+1	+2	+1	+1	+2	+1	234,0
Взаимодействие с протеином	+1	+2	+1	-1	-2	+1	1178,4

Установленное здесь взаимодействие протеина со сравнением животного корма с растительным заслуживает внимания. Умножая приведенные в заголовке суммы вариантов на коэффициент последней строки, можно построить таблицу 2х2:

Корм	Животный	Растительный
Много протеина	1995	1718
Мало протеина	1579	1678

Отсюда видно, что высокая доза протеина животного происхождения лучше усваивается крысами, чем растительный протеин, вероятно, по причине лучшего сбалансирования аминокислот.

6. Опыт с рандомизированными блоками при учете деязинок выборочным способом. С формальной точки зрения опыты с рандомизированными бло-

ками являются факториальными опытами: варианты относятся к одному фактору, а блоки — к другому. В тех случаях, когда вместо полного учета делянок применяется выборочный учет, дисперсионный анализ арифметически остается тем же самым, что и в предыдущем параграфе. Однако соответствующая модель такого опыта не будет содержать в себе взаимодействия, а в ней появятся два компонента для случайных ошибок выборочного наблюдения:

$$X_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ij} + \delta_{ijk},$$

где δ_{ijk} относится к выборочным единицам внутри делянки. Обе величины ε_{ij} и δ_{ijk} появляются из $N(0, \sigma)$, но σ у них обычно различны.

В качестве примера возьмем данные из коллективного опыта по удобрению посевов овса, выращенного на кэррингтоновском суглинке. Урожай учитывался с четырех квадратов $3,6 \times 3,6$ дюйма в декаграммах на квадрат; эти квадраты были взяты в случайном порядке при уборке делянок 21×52 дюйма; результаты приведены в таблице 141.

ТАБЛИЦА 141

Урожай (декаграммы на квадрат) овса на кэррингтоновском суглинке. Выборки по четыре квадрата $3,6 \times 3,6$ дюйма с делянки

Блоки	Контроль	N	P	NP	Сумма
1	19	18	16	13	306
	16	16	22	27	
	18	17	18	28	
	18	14	18	28	
	71	65	74	96	
2	15	22	19	15	296
	16	24	20	14	
	16	22	20	18	
	17	20	24	14	
	64	88	83	61	
4	21	20	24	25	328
	19	21	15	20	
	18	20	19	24	
	22	20	20	20	
	80	81	78	89	
5	18	20	20	20	300
	20	22	20	18	
	20	18	17	12	
	18	19	20	18	
	76	79	77	68	
Сумма	291	313	312	314	1230

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемый параметр
Блоки	3		
Варианты	3	7,61	$\sigma_S^2 + 4\sigma_E^2 + 16\kappa_T^2$
Ошибка опыта	9	31,77	$\sigma_S^2 + 4\sigma_E^2$
Ошибка выборочного учета	48	7,46	σ_S^2

$$s_E^2 = (31,77 - 7,46) / 4 = 6,08$$

Как и в случае таблицы 137, сначала производится вычисление на основе урожаев по квадратам и их поделочным суммам, что дает:

общее — деланки = ошибка выборки.

Второй этап вычислений проводится как обычно для рандомизированных блоков, но с изменением делителя. Например, $\sum X_i^2$ для деланок перед вычитанием поправки на среднюю должна быть разделена на 4.

Экспериментальная ошибка здесь оказалась много больше, чем должна была быть при выборочном варьировании пробных квадратов. Для выяснения этого рассмотрите состав поделочных сумм так, как это делали вы при изучении взаимодействия. Это повышенное варьирование поделочных данных дает

$$e = \frac{\sqrt{31,77/4}}{x} = 14,7\% \text{ на деланку.}$$

(делитель 4 соответствует пересчету экспериментальной ошибки на деланку). Компонентный анализ позволяет установить дополнительную к варьированию квадратов часть экспериментальной ошибки, которая будет $(31,77 - 7,46) / 4 = 6,08$.

Если вы произведете вычисление компонента для «вариантов», то обнаружите, что он отрицателен. Но в данном случае нет оснований ожидать, чтобы урожаи вариантов были более выравниваемы, чем это возможно в случайной выборке. Более вероятно, что случайности выборочного метода дали преувеличенную «ошибку опыта» и (или) преуменьшенный средний квадрат для «вариантов». В той мере, в какой об этом позволяет судить данный опыт, следует сделать вывод, что компонент для «вариантов» равен нулю.

Для предвычисления ошибки взятия проб при тех или иных условиях следует определить добавочную величину, обусловленную ростом выборки; этот вид задачи будет обсужден в параграфе 11 главы 17. Если вы, например, определяете ошибку опыта при 8 пробных квадратах на деланку, то найдете $7,46 + 8 \times 6,08 = 56,10$ вместо 31,77. Отсюда $s_{\bar{x}}$ теперь будет равно $56,10 / 32 = 1,753$ вместо фактического $31,77 / 16 = 1,986$, что является slightком небольшим уменьшением ошибки и не компенсирует удвоение числа проб. Любое более или менее существенное уменьшение стандартной ошибки средней может быть достигнуто только за счет увеличения повторности опыта.

В предыдущем параграфе были даны некоторые правила относительно коэффициентов и подстрочных показателей при σ^2 . При применении их в данном случае следует под символом ошибки опыта E понимать тождественную ей величину T/B .

Пример 9. Хомейером и Блэком [15] был изучен вопрос о двух размерах пробных площадок при выборочном учете деланок с оном. Данные, полученные ими на одном участке, приведены в следующей таблице.

r-k	Блок			Итого по вариантам
	1	2	3	
0-0	190	140	190	1337
	149	157	120	
	205	170	156	
	514	467	376	
0-10	190	80	143	1445
	215	170	92	
	253	192	170	
	628	442	375	
20-0	208	221	232	1935
	235	205	141	
	240	242	211	
	683	668	584	

P-K	Блок			Итого по вариантам
	1	2	3	
20-10	185 204 207 <u>596</u>	265 243 132 <u>640</u>	122 225 189 <u>536</u>	1772
Итого блока	2421	2197	1871	6489

Варианты	3	25 892	8 634
Блоки	2	12 749	6 374
Ошибка	6	5 595	932
Ошибка проб	24	42 987	1 791

Здесь даны урожай (в граммах) трех квадратов 3×3 дюйма, взятых с делянок в одну сороковую акра. Удобрение внесено в дозе 300 фунтов на акр. Проведите дисперсионный анализ.

Ошибка взятия проб здесь оказалась выше ошибки опыта. Это превышение ошибки выбора проб говорит о том, что нужно или увеличение размера проб, или большее их число, или и то и другое вместе.

Пример 10. Разложите сумму квадратов для вариантов на части:

Фосфор	1	23 767
Калий	1	84
Взаимодействие	1	2 040
В целом	3	<u>25 891</u>

Для фосфора $F = 23\,767/932 = 25,5$; $f = 1$ и 6.

7. Регрессия в однофакторном опыте. Варианты опыта часто имеют более двух градаций, в связи с чем возникают вопросы о регрессии. Если градации X находятся на одинаковом расстоянии, то они могут быть кодированы путем образования последовательного ряда целых чисел, например, $-1, 0, +1$; это значительно облегчает вычисления. В качестве примера возьмем опыт с рендомизированными блоками по изучению влияния ширины междурядий, превышающих 3 фута, на урожай сена коровьего гороха [22]. В таблице 142 приведены данные для одного из сортов.

Предварительный анализ дает сумму квадратов для «междурядий» 204,17. Она разлагается на две части, одна из которых соответствует линейной регрессии, а вторая — отклонениям от регрессии.

ТАБЛИЦА 142

Урожай сена коровьего гороха (в фунтах на делянку в $1/100$ моргена)

Междурядья (дюймы)	Блоки				Сумма	Средняя
	1	2	3	4		
4	65	61	60	63	249	62
8	60	58	56	60	234	58
12	53	53	48	55	209	52
Сумма	178	172	164	178	692	

Блоки	3	44,0	
Междурядья	(2)	(204,17)	
Линейная регрессия	1	200,00	200,00**
Отклонение	1	4,17	4,17
Ошибка	6	8,50	1,42

Полагая $x = -1, 0, +1$, получаем $\Sigma x^2 = 2$ и $\Sigma xy = (-1) \times 249 + 0 \times 234 + 1 \times 209 = -40$. Отсюда коэффициент регрессии

$$b = \frac{\Sigma xy}{n\Sigma x^2} = \frac{-40}{4 \times 2} = -5 \text{ фунтам на кодированное } X.$$

Множитель $n=4$ в знаменателе введен для перехода от сумм к средним. (Примечание: при первоначальной единице измерения междурядий в дюймах коэффициент регрессии будет: $-5/4 = -1,25$ фунта на дюйм).

Сумма квадратов для регрессии равна:

$$\frac{(\Sigma xy)^2}{n\Sigma x^2} = \frac{(-40)^2}{4 \times 2} = 200.$$

Это значение вносится в таблицу дисперсионного анализа как сумма квадратов для линейной регрессии. Остаток суммы квадратов «междурядий» $204,17 - 200,00 = 4,17$ вносится в эту таблицу как сумма квадратов «отклонений от регрессии».

Общий вывод такой, что увеличение ширины междурядья приводит к уменьшению урожая, который уменьшается на 1,25 фунта на каждый дюйм увеличивающегося междурядья.

Я здесь считал, что две части суммы квадратов для «междурядий» обладают свойством слагаемости. Таблица 143 показывает, что эти части соответствуют двум ортогональным сравнениям. Первое из них является общим уменьшением урожая $209 - 249 = -40$ при переходе к широким междурядьям. Второе представляет собой удвоенное отклонение урожая при промежуточном междурядье от среднего урожая двух других, т. е. крайних междурядий:

$$2 \left(\frac{249 + 209}{2} - 234 \right) = 2 \times (229 - 234) = -10.$$

Если вы построите график, взяв за X ширину междурядья и за Y сумму урожая, то, соединив крайние точки прямой линии, увидите, что середина этой прямой находится выше оси X на 229 единиц и на 5 единиц ниже урожая промежуточного междурядья. Коэффициенты, которые применяются при этих расчетах, выбраны так, чтобы с меньшим трудом получить точные суммы квадратов.

ТАБЛИЦА 143

Ортогональные сравнения при определении регрессии урожая на ширину междурядья

Междурядья (дюймы)	4	8	12	Сравнения	Делитель	Сумма квадратов
Суммы	249	234	209			
Линейная регрессия	-1	0	+1	-40	8	200,00
Отклонения	+1	-2	+1	-10	24	4,17
Итого						204,17

Отметим, что этот опыт не дает каких-либо указаний относительно того междурядья, при котором получается наивысший урожай. Ясно только, что оно меньше 4 дюймов. Экспериментатор в данном случае имел намерение взять такие междурядья, среди которых можно ожидать оптимум, но условия погоды сделали этот опыт нетипичным. Если бы ожидания экспериментатора полностью оправдались, то линия регрессии была бы строго горизонтальной и вся сумма квадратов «междурядий» отошла бы к «отклонениям».

Два ряда ортогональных коэффициентов, использованных выше, находятся во второй секции таблицы 144. Эти коэффициенты относятся к двум сравнениям, из которых одно может быть выделено в качестве характеристики

привизны регрессии, проходящей через три точки. Такая регрессия называется регрессией *второго порядка*, или *параболической*; полиномиальная кривая второго порядка, или парабола второго порядка, может точно проходить через любые три точки. В первой колонке таблицы 144 для каждого ряда коэффициентов дана $\Sigma \lambda^2$; эта величина, умноженная на n , берется в качестве делителя в выражении

$$\frac{c^2}{n \Sigma \lambda^2}.$$

ТАБЛИЦА 144

Коэффициенты и делители для рядов ортогональных сравнений при определении регрессии с равными промежутками между X

Порядок сравнения	Сравнение	Число градаций							Делитель $\Sigma \lambda^2$
		1	2	3	4	5	6	7	
1	Линейное	-1	+1						2
2	Линейное	-1	0	+1					2
	Квадратичное	+1	-2	+1					6
3	Линейное	-3	-1	+1	+3				20
	Квадратичное	+1	-1	-1	+1				4
	Кубичное	-1	+3	-3	+1				20
4	Линейное	-2	-1	0	+1	+2			10
	Квадратичное	+2	-1	-2	-1	+2			14
	Кубичное	-1	+2	0	-2	+1			10
	Четвертой степени	+1	-4	+6	-4	+1			70
5	Линейное	-5	-3	-1	+1	+3	+5		70
	Квадратичное	+5	-1	-4	-4	-1	+5		84
	Кубичное	-5	+7	+4	-4	-7	+5		180
	Четвертой степени	+1	-3	+2	+2	-3	+1		28
	Пятой степени	-1	+5	-10	+10	-5	+1		252
6	Линейное	-3	-2	-1	0	+1	+2	+3	28
	Квадратичное	+5	0	-3	-4	-3	0	+5	84
	Кубичное	-1	+1	+1	0	-1	-1	+1	6
	Четвертой степени	+3	-7	+1	+6	+1	-7	+3	154
	Пятой степени	-1	+4	-5	0	+5	-4	+1	84
	Шестой степени	+1	-6	+15	-20	+15	-6	+1	924

В качестве другой иллюстрации применения этой таблицы возьмем данные об урожайности при 5 расстояниях между растениями в опыте с просом; эти данные, взятые из таблицы 121, перенесены в таблицу 145 вместе с коэффициентами таблицы 144 для 5 градаций фактора. Суммы $\Sigma \lambda^2$ умножены на $n = 5$ и использованы в качестве делителей. Первая сумма квадратов для линейного сравнения 3960 совпадает с соответствующей суммой таблицы 123. Сумма трех других $29 + 144 + 23 = 196$ согласуется с суммой квадратов для «отклонений». Но теперь эти «отклонения» разложены на три части. Первая новая часть является дополнительным извлечением из суммы квадратов, оставшейся после подбора прямой линии; эта часть приписывается параболе. Это означает, что число $3960 + 29 = 3989$ является извлечением из суммы квадратов и относится к параболе, которая может быть непосредственно подобрана по первоначальным средним вариантов. Подобные положения относятся и к другим нашим сравнениям. Кривая четвертого порядка будет точно проходить через все 5 точек, так что после этого в сумме квадратов для вариантов уже не останется никакого остатка.

Сравнения регрессии средних из опыта с просом по данным таблицы 121

Расстояния	2	4	6	8	10	Сравне- ние	Делите- ли	Сумма квадратов
Сумма	1349	1314	1262	1191	1188			
Линейное	-2	-1	0	+1	+2	-445	50	3960
Квадратичное	+2	-1	-2	-1	+2	45	70	29
Кубичное	-1	+2	0	-2	+1	85	50	144
Четвертой степени	+1	-4	+6	-4	+1	89	350	23
Итого								4 156

Конечно, подбор такой сложной кривой регрессии по данным опыта с просом представляет собой просто некоторое упражнение, так как основная масса варьирования урожая вариантов объясняется только линейной регрессией, причем она, как мы видели, не оказалась существенной. Для некоторых читателей может показаться неожиданным наличие существенных сравнений внутри несущественно различающегося ряда данных по вариантам опыта. Можно часто встретиться, например, с таким дисперсионным анализом:

варианты	10	(400)	(40)
линейная регрессия	1	350	350
отклонения	9	50	
ошибка	30	1200	40

По всем вариантам критерий $F=1$, но при оценке линейной регрессии $F = 350/40 = 8,75$, чему соответствует $P < 0,01$. Большая часть суммы квадратов для вариантов относится к одному сравнению «линейная регрессия». Это обстоятельство подчеркивает еще раз различие между заранее планируемыми (как у нас в опыте с просом) и непланируемыми сравнениями; только при последних сравнениях должен иметь значение факт существенности или несущественности общей оценки всех вариантов в целом.

8. Регрессия в двухфакторных опытах. Регрессия может иметь место не только при изучении одного фактора, но и нескольких. В качестве примера регрессии одного фактора в присутствии другого фактора я беру дополнительные данные к опыту о междурядьях коровьего гороха, приведенному в предыдущем параграфе. В таблице 146 данные по сорту II повторены в том виде, какими они были взяты ранее, и к ним добавлены данные для двух других сортов. Теперь здесь два фактора, каждый с тремя градациями; значит, мы имеем факторный опыт 3^2 с 9 вариантами.

В этом опыте обращает на себя внимание то обстоятельство, что у сортов I и III имеет место тенденция к повышению урожайности при расширении междурядья, в то время как у сорта II наблюдается обратная тенденция. Это является причиной большого значения среднего квадрата для взаимодействия и настораживает против общих заключений, сделанных на основе главных эффектов.

Для изучения указанных тенденций изменения урожая при изменении междурядья в таблице 147 для каждого сорта произведено вычисление линейного и квадратичного сравнений. Отрицательное значение линейного сравнения у сорта II резко выделяется среди положительных значений этого сравнения у сортов I и III. Средние же квадраты у каждого сорта ясно указывают на наличие линейной тенденции в изменении урожая и на отсутствие криволинейности в этом изменении. Можно видеть, что сумма всех шести средних квадратов, относящихся к линейным и квадратичным сравнениям, равна суммам квадратов для «междурядья» и «взаимодействия», взятым вместе с 6 степенями свободы (табл. 146). При наличии взаимодействия средний

квадрат для «сортов» не представляет, собственно говоря, никакого интереса.

Если произвести сравнение между сортами I и III, то и между ними по признаку крутизны линейного подъема будет наблюдаться существенное различие.

В случае, если эти взаимодействия между линейными сравнениями незначительны, то по тем же схемам, что и ранее, могут быть проведены вычисления нижней части таблицы. При этом следует основываться на правиле 5 параграфа 2. Так, для сравнения по междуурядьям в целом, т. е. по всем трем сортам, следует брать $kn=12$, что является числом делянок, относящихся к каждому междуурядью. Взаимодействия же, характеризующие различия между тремя внутрисортными сравнениями, будут иметь $n=4$, т. е. равное числу повторений.

ТАБЛИЦА 146

Урожай сена трех сортов коровьего гороха (фунтов на делянку в $1/100$ моргена)

Сорта	Междуурядья (дюймы)	Блоки				Сумма
		1	2	3	4	
I	4	56	45	43	46	190
	8	60	50	45	48	203
	12	66	57	50	50	223
II	4	65	61	60	63	249
	8	60	58	56	60	234
	12	53	53	48	55	209
III	4	60	61	50	53	224
	8	62	68	67	60	257
	12	73	77	77	65	292
Сумма		555	530	496	500	2081

Сорта	Междуурядья			Сумма
	4	8	12	
I	190	203	223	616
II	249	234	209	692
III	224	257	292	773
Сумма	663	694	724	2081

Блоки	3	255,64	
Сорта V	2	1027,39	513,70
Междуурядья S	2	155,06	77,53
Взаимодействие VS	4	765,44	191,36**
Ошибка	24	424,11	17,67

Теперь мы можем произвести оценку существенности выявленных в данном опыте фактов. Сорта имеют прямолинейное изменение урожая, которое оказалось неодинаковым. Так как кривизна регрессий несущественна, то остается невыясненным, каково оптимальное междуурядье у данных сортов. По-видимому, сорта I и III имеют столь мощное вегетативное развитие, что для получения максимального урожая требуют более широкого междуурядья, чем 12 дюймов. Другое следствие из эксперимента таково: междуурядья для сортов I и III должны быть иными, чем для сорта II.

Вы можете спросить, почему суммы 9 вариантов не были расположены в ряд с соответствующими коэффициентами, как это было сделано в таблице 136 по отношению к опыту 2². Причина этого лежит в том, что между сортами

нельзя заранее установить строго определенный порядок сравнения, т. е. сорта нельзя заранее отождествить с числами $-1, 0, 1$. Конечно, чисто формально легко могут быть определены некоторые ортогональные сравнения, однако они не будут иметь никакого реального смысла.

Теперь мы перейдем к опыту 3×4 , в котором можно рассматривать вопрос о регрессии в отношении обоих факторов. Для этого возьмем данные отдела кормов сельскохозяйственной опытной станции штата Айова [25]. Предметом изучения было уменьшение содержания аскорбиновой кислоты в фасоли, происходящее при воздействии трех температур за четыре периода хранения, продолжительность каждого из которых на две недели больше предыдущего. Уборка фасоли производилась при одинаковых условиях до восьми часов одного и того же утра. Далее они были подготовлены и быстро заморожены до начала второй половины этого же дня. Для каждого из 12 вариантов было предназначено по три пробных образца; все пакеты с этими образцами были в случайном порядке помещены на хранение в ящик, в результате чего получена полностью рандомизированная схема опыта. В таблице 147 приведены суммы по результатам трех определений содержания аскорбиновой кислоты.

ТАБЛИЦА 147

Линейные и квадратичные сравнения для каждого сорта в опыте с коровьим горохом

Линейное Квадратичное	4 дюйма	8 дюймов	12 дюймов	Сравнения с	
	-1 +1	0 -2	+1 +1	линейное	квадратичное
Сорт I	190	203	223	33	7
II	249	234	209	-40	-10
III	224	257	292	68	2
Сумма	663	694	724	61	-1

Сорт I: линейное	$\frac{33^2}{4 \times 2} = 136,12^{**}$;	квадратичное	$\frac{7^2}{4 \times 6} = 2,04$.
Сорт II:	$\frac{(-40)^2}{4 \times 2} = 200,00^{**}$;		$\frac{(-10)^2}{4 \times 6} = 4,17$.
Сорт III:	$\frac{68^2}{4 \times 2} = 578,00^{**}$;		$\frac{2^2}{4 \times 6} = 0,17$.
Итого	914,12		6,38.

Проверка: $914,12 + 6,38 = 155,06 + 765,44 (=S + SV)$; $f = 6$.

Междурядья: линейное	$\frac{(\Sigma c_i)^2}{kn \Sigma \lambda^2} = \frac{61^2}{3 \times 4 \times 2} = 155,04$;	квадратичное	$\frac{(-1)^2}{3 \times 4 \times 6} = 0,01$.
Взаимодействие: линейное	$\frac{\Sigma c_i^2}{n \Sigma \lambda^2} - S_L = \frac{33^2 + (-40)^2 + 68^2}{4 \times 2} - 155,05 = 759,08$;		
квадратичное	$\frac{\Sigma c_i^2}{n \Sigma \lambda^2} - S_Q = \frac{7^2 + (-10)^2 + 2^2}{4 \times 6} - 0,01 = 6,37$.		

Междурядья	(2)	(155,06)	
Линейное	1		155,04
Квадратичное	1		0,01
Взаимодействие	(4)	(765,44)	
Линейное	2	759,08	379,54**
Квадратичное	2	6,37	3,18
Ошибка	24		17,67

Из этих данных ясно видно, что содержание аскорбиновой кислоты уменьшается с повышением температуры хранения и с увеличением времени хра-

нения (последнее не относится к нулевой температуре). Точно так же видно, что скорость уменьшения содержания кислоты с температурой не идет по прямой линии и что она различна в отдельные периоды хранения. В данном случае имеется возможность установить строгую последовательность наблюдений в отношении градаций факторов, и поэтому они позволяют применить к ним рассматриваемые здесь критерии.

ТАБЛИЦА 148

Схема результатов трех определений содержания аскорбиновой кислоты (мг/100 г) для каждого из 12 вариантов опыта 3×4 с фасолью

Температура F°	Неделя хранения				Сумма
	2	4	6	8	
0	45	57	46	46	184
10	45	43	41	37	166
20	34	28	21	16	99
Сумма	124	118	108	99	449

Температура T	2	334,39
Двухнедельный период P	3	40,53
Взаимодействие TP	6	34,05
Ошибка ¹	24	0,706

¹ Ошибка (образцы одного варианта) вычислена по первоначальным, не приведенным здесь данным.

Анализ можно начать как с температуры, так и с периодов; я выбираю температуру. В этом случае для каждого из периодов при помощи коэффициентов $-1, 0, +1$ и $+1, -2 +1$ вычисляются линейные и квадратичные сравнения:

Неделя хранения	2	4	6	8	Сумма
Линейное T_L	-11	-19	-25	-30	-85
Квадратичное T_Q	-11	-11	-15	-12	-49

По мере увеличения времени хранения наклоны прямых линий регрессии становятся более крутыми. Этот вопрос будет рассмотрен в дальнейшем; в настоящее же время вычислим следующие суммы квадратов:

$$T_L = \frac{(-85)^2}{12 \times 2} = 301,04^{**};$$

$$T_Q = \frac{(-49)^2}{12 \times 6} = 33,35^{**}.$$

Сумма этих величин $301,04 + 33,35 = 334,39$ является суммой квадратов для T в приведенной выше таблице дисперсионного анализа. Существенность этих эффектов (отмеченная звездочками) установлена путем сравнения их со средним квадратом ошибки 0,706. Очевидно, что регрессии криволинейны, так как параболическое сравнение оказалось существенным: это значит, что по мере увеличения температуры скорость уменьшения содержания аскорбиновой кислоты быстро возрастает. (Отметим, что каждой из сумм для определенной температуры соответствует число повторений, образованное из четырех периодов и трех образцов, т. е. 12).

Одинаковы ли регрессии для всех четырех периодов? Для того чтобы ответить на этот вопрос, вычислим взаимодействия этих регрессий с периодами (см. правило 5 параграфа 2):

$$T_L P = \frac{(-11)^2 + \dots + (-30)^2}{3 \times 2} - T_L = 33,46^{**};$$

$$T_Q P = \frac{(-11)^2 + \dots + (-12)^2}{3 \times 6} - T_Q = 0,59.$$

Сумма этих величин равна сумме квадратов для TP в таблице дисперсионного анализа. По мере удлинения периода хранения линейные регрессии существенно уменьшаются (т. е. принимают все большие и большие отрицательные значения), но кривизна регрессий может считаться одинаковой для всех периодов.

Обращаясь теперь к суммам для четырех периодов, определим три сравнения:

Суммы	124	118	108	99	Сравнения	Сумма квадратов
Линейное	-3	-1	+1	+3	-85	40,14**
Квадратичное	+1	-1	-1	+1	-3	0,25
Кубичное	-1	+3	-3	+1	5	0,14
Сумма = Сумма квадратов для периодов						40,53

Это указывает на то, что регрессия на период может считаться линейной.

Теперь мы перейдем к исследованию нового, не встречавшегося до настоящего времени вопроса, а именно рассмотрим регрессии T_L и T_Q по периодам. Будем вычислять для каждого периода T_L , т. е. наклон прямой действия температуры на содержание витамина. Это ответит на вопрос: как величина T_L меняется по периодам?

Эти вычисления проводятся так: к вычисленным ранее T_L применяются ортогональные коэффициенты для четырех градаций. Например:

$$T_L P_L = (-3) \times (-11) + (-1) \times (-19) + 1 \times (-25) + 3 \times (-30) = -63.$$

Для этого сравнения сумма квадратов будет:

$$T_L P_L = \frac{(-63)^2}{3 \times 2 \times 20} = 33,08^{**}.$$

Новым моментом здесь является применение правила 6 для двух последовательных группировок. Множитель 2 является $\Sigma \lambda_1^2$ в первой группировке (по температурам), в то время как 20 является $\Sigma \lambda_2^2$ во второй группировке (по периодам).

Подобно предыдущему, находим:

$$T_L P_Q = \frac{3^2}{3 \times 2 \times 4} = 0,38;$$

$$T_L P_C = \frac{(-1)^2}{3 \times 2 \times 20} = 0,01.$$

Сумма трех сумм квадратов этих регрессий составляет 33,47, что равно вычисленному ранее $T_L P$. Теперь очевидно, что линейная регрессия на температуру сама уменьшается линейно по мере увеличения периода хранения, что было установлено ранее при рассмотрении значений T_L .

Применяя этот же метод и для T_Q , получим:

$$r_{QP_L} = \frac{(-7)}{3 \times 6 \times 20} = 0,14;$$

$$T_Q P_Q = \frac{3^2}{3 \times 6 \times 4} = 0,12;$$

$$r_{QP_C} = \frac{11^2}{3 \times 6 \times 20} = 0,34.$$

Сумма всех этих величин $T_Q P = 0,60$. Ясно, что величины T_Q не меняются по периодам. Все полученные результаты собраны в таблице 149. Содержание взаимодействия сводится к одной только регрессии на регрессию ($T_L P_L$).

Обычно нет необходимости выписывать все несущественные сравнения, их просто объединяют в одну группу в качестве остатка.

Резюмируя результаты опыта, можно сказать, что регрессия количества аскорбиновой кислоты на температуру имеет форму параболы, причем линейная часть наклона сверху вниз становится все более и более крутой по мере увеличения периода хранения. Регрессия на период хранения линейна, причем наклон сверху вниз становится более значительным по мере увеличения температуры. На основе этих выводов была построена геометрическая поверхность, позволяющая определить содержание аскорбиновой кислоты для некоторого заданного периода хранения и для некоторой температуры (и то и другое в пределах данных опыта). Все это изложено в цитированной здесь литературе.

ТАБЛИЦА 149

Дисперсионный анализ для данных о содержании аскорбиновой кислоты в фасоле

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Температура	(2)	(334,39)	
T_L	1		301,04**
T_Q	1		33,35**
Периоды	(3)	(40,53)	
P_L	1		40,14**
P_Q	1		0,25
P_C	1		0,14
Взаимодействие	1	(34,05)	
$T_L P_L$			33,08**
$T_L P_Q$			0,38
$T_L P_C$			0,01
$T_Q P_L$			0,14
$T_Q P_Q$			0,12
$T_Q P_C$			0,34
Ошибка	24		0,706

Пример 11. Так как в таблице 146 установлено взаимодействие, то произведите исследование двух сравнений для «междурядий» по каждому сорту в отдельности.

Пример 12. В опыте, подобном приведенному в таблице 148, представляется удобным расположить все суммы по вариантам в один ряд, после чего под ними выписать коэффициенты для определения всех сравнений:

	0°				10°				20°			
	2	4	6	8	2	4	6	8	2	4	6	8
T_L	-1	-1	-1	-1	0	0	0	0	+1	+1	+1	+1
T_Q	+1	+1	+1	+1	-2	-2	-2	-2	+1	+1	+1	+1
S_L	-3	-1	+1	+3	-3	-1	+1	+3	-3	-1	+1	+3
S_Q	+1	-1	-1	+1	+1	-1	-1	+1	+1	-1	-1	+1
S_C	-1	+3	-3	+1	-1	+3	-3	+1	-1	+3	-3	+1
$T_L S_L$	+3	+1	-1	-3	0	0	0	0	-3	-1	+1	+3

и т. д.

Заполните эту таблицу, вычислите суммы квадратов ($n=3$) и произведите сопоставление их общей суммы с суммой квадратов для «вариантов» из таблицы 148 ($334,39+40,53+34,05=408,97$).

9. Трехфакторные опыты 2³. Экспериментатор часто нуждается в получении сведений относительно эффектов трех или большего числа факторов при их совместном действии. В простейшем случае это будет опыт с тремя факторами при двух градациях каждый, т. е. опыт $2 \times 2 \times 2$, или 2³. Здесь в общей экспериментальной схеме испытывается 8 комбинаций этих градаций.

В таблице 150 приведены данные, взятые из одного неопубликованного опыта с рендомизированными блоками [16] по изучению влияния добавок соевой муки в рацион с кукурузой для свиней. Так как в кукурузе имеется недостаток лизина, то в корм добавлялось 0,6% этой аминокислоты. Соевая же мука содержит лизин. Из последней колонки таблицы видно, что добавление протеина в форме соевой муки приводит к увеличению скорости роста при отсутствии добавочного лизина, но в протеине нет никакой потребности, когда эта аминокислота добавляется непосредственно. Наличие этого взаимодействия указывает на то, что главные эффекты в этом случае представляют малый интерес, но все же для целей иллюстрации метода обработки анализ этих данных будет проведен до конца.

ТАБЛИЦА 150

Средние суточные привесы свиней в опыте по схеме 2³ с рендомизированными блоками

Лизин (в %)	Протеин (в %)	Пол	Повторения (блоки)								Сумма вариантов	Сумма для 2 полов
			1	2	3	4	5	6	7	8		
0	12	М	1,11	0,97	1,09	0,99	0,85	1,21	1,29	0,96	8,47	17,08
		Ж	1,03	0,97	0,99	0,99	0,99	1,21	1,19	1,24	8,61	
0,6	14	М	1,52	1,45	1,27	1,22	1,67	1,24	1,34	1,32	11,03	
		Ж	1,48	1,22	1,53	1,19	1,16	1,57	1,13	1,43	10,71	
0,6	12	М	1,22	1,13	1,34	1,41	1,34	1,19	1,25	1,32	10,20	19,34
		Ж	0,87	1,00	1,16	1,29	1,00	1,14	1,36	1,32	9,14	
0,6	14	М	1,38	1,08	1,40	1,21	1,46	1,39	1,17	1,21	10,30	
		Ж	1,09	1,09	1,47	1,43	1,24	1,17	1,01	1,13	9,63	
Сумма повторения			9,70	8,91	10,25	9,73	9,71	10,12	9,74	9,93		78,09

Повторения	7	0,1411	
Варианты	7	0,7986	0,1141**
Ошибка	49	1,0994	0,0224

Разложение суммы квадратов вариантов на 7 сравнений, каждому из которых соответствует единичная степень свободы, удобно произвести по схеме таблицы 151. Сюда из предпоследнего столбца таблицы 150 перенесены суммы по вариантам. Далее выписываются коэффициенты главных эффектов, после чего путем перемножения их получают коэффициенты взаимодействия. Здесь теперь имеется три двухфакторных взаимодействия и одно трехфакторное. Для этого последнего взаимодействия коэффициенты получаются путем умножения коэффициентов любого из главных эффектов на коэффициенты того взаимодействия, которое относится к двум другим факторам; например, коэффициенты L на коэффициенты SP . То, что эти 7 сравнений ортогональны, можно проверить при помощи правила 3 параграфа 2.

Чтобы определить эти сравнения, достаточно присоединить к суммам по вариантам соответствующие коэффициенты. Например, для сравнения

Семь сравнений в факториальном опыте 2³ со свиньями

Эффекты	Лизин = 0				Лизин = 0,6%				Сравнение	Сумма квадратов
	P = 12%		P = 14%		P = 12%		P = 14%			
	М	Ж	М	Ж	М	Ж	М	Ж		
	8,47	8,61	11,03	10,71	10,20	9,14	10,30	9,63		
Пол S	-1	+1	-1	+1	-1	+1	-1	+1	-1,91	0,0570
Протеин P	-1	-1	+1	+1	-1	-1	+1	+1	5,25	0,4307
SP	+1	-1	-1	+1	+1	-1	-1	+1	-0,07	0,0001
Лизин L	-1	-1	-1	-1	+1	+1	+1	+1	0,45	0,0032
SL	+1	-1	+1	-1	-1	+1	-1	+1	-1,55	0,0375
PL	+1	+1	-1	-1	-1	-1	+1	+1	-4,07	0,2588*
SPL	-1	+1	+1	-1	+1	-1	-1	+1	0,85	0,0113
Итого										0,7986

между полами = Ж - М: $c = +8,61 + 10,71 + 9,14 + 9,63 - 8,47 - 11,03 - 10,20 - 10,30 = -1,91$.

Суммы же квадратов получаются на основе правила 2 параграфа 2, где при $n = 8$ и $\Sigma\lambda^2 = 8$ для каждого случая следует брать делитель 64. Правило 4 параграфа 2 дает возможность произвести проверку этих вычислений. Итоговая сумма из этих 7 сумм квадратов должна равняться сумме квадратов для вариантов в таблице дисперсионного анализа.

В данном случае нет достоверного влияния различия полов как в отношении главного эффекта, так и в отношении взаимодействий с другими факторами. Это может служить основанием для объединения сумм по полам и изучения по ним вопроса о существенности взаимодействия протеин - лизин. Эти суммы из последней колонки таблицы 150 можно перенести в таблицу 2×2 :

	L_0	$L_{0,6}$
P_{12}	17,08	19,34
P_{14}	21,74	19,93

Отсюда следует, что протеин кукурузы при добавлении 0,6% лизина был эффективным, но при добавлении протеина в форме соевой муки, богатой лизином, добавление этой аминокислоты уже не дает эффекта.

Напомним, что сравнение взаимодействия является разностью между двумя такими разностями:

$$\begin{aligned} \text{эффект лизина при } P_{14}: & 19,93 - 21,74 = -1,81 \\ \text{эффект лизина при } P_{12}: & 19,34 - 17,08 = 2,26 \\ \hline \text{взаимодействие} = \text{разность} & = -4,07, \end{aligned}$$

т. е. то же значение, как и в таблице 151. Для того чтобы по этой более компактной таблице вычислить соответствующую сумму квадратов, следует объединить вместе повторения и «пол», т. е. взять $n = 8 \times 2 = 16$. Это увеличение повторности будет компенсироваться уменьшением числа коэффициентов, которых теперь вместо 8 будет 4: +1, -1, -1, +1, так что $\Sigma\lambda^2 = 4$. В результате получится та же, что и ранее, величина:

$$\frac{c^2}{n\Sigma\lambda^2} = \frac{(-4,07)^2}{16 \times 4} = 0,2588.$$

Общий эффект протеина в данном случае при наличии взаимодействия не представляет никакого практического интереса. Конечно, нельзя рекомендовать добавление соевого протеина как с лизином, так и без лизина. Эта

аминокислота имеет значение только при отсутствии в корме соевой муки, а отнюдь не при ее наличии. Экспериментатору после этого остается только решить, какой из источников аминокислоты является главным.

Если бы здесь оказалось существенным трехфакторное взаимодействие, то его содержание могло бы быть выяснено путем сравнения двух двухфакторных взаимодействий:

	Хряки		Матки	
	L_0	$L_{0,6}$	L_0	$L_{0,6}$
P_{12}	8,47	10,20	8,61	9,14
P_{14}	11,03	10,30	10,71	9,63
$P_{14} - P_{12}$	2,56	0,10	2,10	0,49
P_L	- 2,46		- 1,61	

Взаимодействие в группе маток менее выражено, чем взаимодействие в группе хряков. Разность

$$- 1,61 - (- 2,46) = 0,85$$

и будет трехфакторным взаимодействием, указанным в таблице 151.

Если вычислять сумму квадратов данного сравнения в форме разности $- 1,61 + 2,46$, то следует брать $n = 32$ и $\Sigma \lambda^2 = 2$. Это подсказывается требованием полного соответствия всех коэффициентов, входящих в приведенные выше две таблицы, коэффициентам последней колонки таблицы 151.

Моя практика говорит, что опыты редко бывают достаточно чувствительными для того, чтобы в них можно было обнаружить трехфакторные взаимодействия, даже если они реально существуют.

10. Трехфакторные опыты $2 \times 3 \times 4$. Настоящий параграф будет посвящен описанию некоторых более общих методов, применяемых к обработке факториальных опытов, включая сюда и случай двухфакторного опыта. Здесь будут рассмотрены I, II и смешанная модели с соответствующими правилами оценки существенности компонентов. Мы начнем с числового примера и уже после этого перейдем к общему случаю.

Пример я снова возьму из опыта, из которого уже брал пример для предыдущего параграфа. В опыте был четвертый фактор — дозы метионина, из которых я возьму здесь три. Вместе с тем теперь будет включено вместо двух четыре дозы лизина. Возьмем также только два повторения и только данные о хряках. Все это дает факториальный опыт по схеме $2 \times 3 \times 4$ с рендомизированными блоками. Данные приведены в таблице 152. Вычисления не требуют объяснений.

Принятая здесь модель такова:

$$X_{ijkm} = \mu + R_i + L_j + M_k + P_m + (LM)_{jk} + (LP)_{jm} + (MP)_{km} + (LPM)_{jkm} + \varepsilon_{ijkm}.$$

Это модель опыта I, в которой значения факторов считаются фиксированными. В таблице 153 дан соответствующий этому случаю компонентный анализ. Правила для выписывания компонентов будут указаны в последующих абзацах настоящего параграфа, а в примерах 13, 14 и 15 этой главы будет дана соответствующая интерпретация результатов.

Теперь перейдем к общему случаю. Чтобы построить для этого случая схему компонентного анализа и испытания гипотез, удобнее начать с модели II при полной рендомизации. При эти условиях можно написать

$$X_{ijkm} = \mu + A_i + B_j + C_k + (AB)_{ij} + (AC)_{ik} + (BC)_{jk} + (ABC)_{ijk} + \varepsilon_{ijkm};$$

$$i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; m = 1, \dots, n; \varepsilon_{ijkm} = N(0, \sigma).$$

В связи с тем, что теперь все эффекты факторов являются случайными переменными, будем считать, что они распределены нормально с соответствующими дисперсиями $\sigma_A^2, \dots, \sigma_{ABC}^2$. Компоненты для каждого из этих эффектов представлены в таблице 154.

Здесь полезно руководствоваться двумя правилами [23]. О втором из этих правил уже упоминалось в параграфе 5 этой главы. Первое правило можно применять только тогда, когда источники варьирования расположены в указанном порядке: сначала главные эффекты, далее взаимодействия.

ТАБЛИЦА 152

Трехфакторный опыт ($2 \times 3 \times 4$) с рандомизированными блоками. Средние суточные привесы свиней, получающих различные добавки (в %) лизина, метионина и протеина

Лизин L	Метионин M	Протеин P	Повторения (блоки) R		Итого по вариантам	
			1	2		
0,05	0	12	1,11	0,97	2,08	
		14	1,52	1,45	2,97	
	0,025	12	1,09	0,99	2,08	
		14	1,27	1,22	2,49	
	0,050	12	0,85	1,21	2,06	
		14	1,67	1,24	2,91	
	0,10	0	12	1,30	1,00	2,30
			14	1,55	1,53	3,08
0,025		12	1,03	1,21	2,24	
		14	1,24	1,34	2,58	
0,050		12	1,12	0,96	2,08	
		14	1,76	1,27	3,03	
0,15		0	12	1,22	1,13	2,35
			14	1,38	1,08	2,46
	0,025	12	1,34	1,41	2,75	
		14	1,40	1,21	2,61	
	0,050	12	1,34	1,19	2,53	
		14	1,46	1,39	2,85	
	Итого	0	12	1,19	1,03	2,22
			14	0,80	1,29	2,09
0,025		12	1,36	1,16	2,52	
		14	1,42	1,39	2,81	
0,050		12	1,46	1,03	2,49	
		14	1,62	1,27	2,89	
Итого			31,50	28,97	60,47	

Вычисления: 1) $C = 60,47^2/48 = 76,1796$.

2) Общее: $1,11^2 + 0,97^2 + \dots + 1,62^2 + 1,27^2 - C = 2,0409$.

3) Варианты: $(2,08^2 + 2,97^2 + \dots + 2,89^2)/2 - C = 1,2756$.

4) Повторения: $(31,50^2 + 28,97^2)/24 - C = 0,1334$.

5) Ошибка: $2,0409 - (1,2756 + 0,1334) = 0,6319$.

Метионин	Сводная таблица А				
	Лизин				Итого
	0	0,05	0,10	0,15	
0	5,05	5,38	4,81	4,31	19,55
0,025	4,57	4,82	5,36	5,33	20,08
0,050	4,97	5,11	5,38	5,38	20,84
Итого	14,59	15,31	15,55	15,02	60,47

Вычисления (продолжение):

6) В таблицу А вносятся суммы двух доз протеина: $5,05 = 2,08 + 2,96$ и т. д.

7) Общее в таблице А: $(5,05^2 + \dots + 5,38^2)/4 - C = 0,3496$.

8) Лизин L: $(14,59^2 + \dots + 15,02^2)/12 - C = 0,0427$.

9) Метионин M: $(19,55^2 + 20,08^2 + 20,84^2)/16 - C = 0,0526$.

10) LM: $0,3496 - (0,0427 + 0,0526) = 0,2543$.

Метионин	Сводная таблица В		
	Протеин		Итого
	12	14	
0	8,95	10,60	19,55
0,025	9,59	10,49	20,08
0,050	9,16	11,68	20,84
Итого . .	27,70	32,77	60,47

Вычисление (продолжение):

- 11) В таблицу В вносятся суммы 4 доз лизина: $8,95 = 2,08 + 2,30 + 2,35 + 2,22$ и т. д.
 12) Общее в таблице В: $(8,95^2 + \dots + 11,68^2)/8 - C = 0,6702$.
 13) Протеин Р: $(27,70^2 + 32,77^2)/24 - C = 0,5355$.
 14) $MP: 0,6702 - (0,5355 + 0,0526) = 0,0821$.

Сводная таблица С

Протеин	Лизин				Итого
	0	0,05	0,10	0,15	
12	6,22	6,62	7,63	7,23	27,70
14	8,37	8,69	7,92	7,79	32,77
Итого	14,59	15,31	15,55	15,02	60,47

Вычисления (продолжение):

- 15) В таблицу С вносятся суммы 3 доз метионина: $6,22 = 2,08 + 2,08 + 2,06$ и т. д.
 16) Общее для таблицы С: $(3,22^2 + \dots + 7,79^2)/6 - C = 0,8181$.
 17) $LP: 0,8181 - (0,5355 + 0,0427) = 0,2399$.
 18) $LMP: 1,2756 - (0,0427 + 0,0526 + 0,5355 + 0,2543 + 0,0821 + 0,2399) = 0,0685$.

П р а в и л о 1. Компоненты любого среднего квадрата (кроме последнего из них) определяются прописной буквой (или буквами), относящимися к соответствующим источникам варьирования. Последний компонент имеет эту, относящуюся к нему букву (или буквы) в подстрочном указателе. Каждый параметр содержит в себе все компоненты, имеющие его подстрочную букву (или буквы), определяющую соответствующий источник варьирования.

П р а в и л о 2. К каждой σ^2 (исключая σ^2 для ошибки) прикреплены определенные коэффициенты и подстрочные указатели, общее число которых константно. Прописная буква обозначает фактор; соответствующая строчная буква — число градаций, или повторений. Если прописная буква поставлена в подстрочный указатель, то в коэффициенте опускается соответствующая строчная буква.

Оценка существенности взаимодействий производится непосредственно, но при оценке главных эффектов встречаются некоторые затруднения. Если производится проверка гипотезы $\sigma_A^2 = 0$, то нельзя указать подходящий для этого знаменатель в критерии F . Однако задача о приближенных критериях все же решена [21, 8]. То, что параметр для A содержит в себе компоненты как AB , так и AC , наводит на мысль сложить соответствующие им средние квадраты. Эта сумма, если ее выразить через компоненты, дает средний квадрат:

$$MS_2 = 2\sigma^2 + 2n\sigma_{ABC}^2 + n\sigma_{AB}^2 + n\sigma_{AC}^2.$$

В противовес MS_2 сложим компоненты для A и ABC :

$$MS_1 = 2\sigma^2 + 2n\sigma_{ABC}^2 + n\sigma_{AB}^2 + n\sigma_{AC}^2 + nb\sigma_A^2.$$

Дисперсионный анализ трехфакторного опыта со свиньями.
Компонентный анализ для модели I. Рендомизированные блоки

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат	Компоненты, оцениваемые средним квадратом
Повторения	1	0,1334		
Лизин L ($l=4$)	3	0,0427	0,0142	$\sigma^2 + rmpk^2_L$
Метионин M ($m=3$)	2	0,0526	0,0263	$\sigma^2 + rlpk^2_M$
Протеин P ($p=2$)	1	0,5355	0,5355**	$\sigma^2 + rlmk^2_P$
LM	6	0,2543	0,0424	$\sigma^2 + rpk^2_{LM}$
LP	3	0,2399	0,0800	$\sigma^2 + rmk^2_{LP}$
MP	2	0,0821	0,0410	$\sigma^2 + rlpk^2_{MP}$
LMP	6	0,0685	0,0114	$\sigma^2 + rpk^2_{LMP}$
Ошибка ($r=2$)	23	0,6319	0,0275	σ^2

Отношение $F' = MS_1/MS_2$ имеет форму, пригодную для проверки гипотезы $\sigma^2_A=0$, но оно не подчиняется строго закону распределения F.

ТАБЛИЦА 154

Компонентный анализ трехфакторного опыта, Модель II.
Полностью рендомизированное расположение

Источник варьирования	Параметры, оцениваемые средним квадратом
A	$\sigma^2 + n\sigma^2_{ABC} + n\sigma^2_{AB} + nb\sigma^2_{AC} + nb\sigma^2_A$
B	$\sigma^2 + n\sigma^2_{ABC} + n\sigma^2_{AB} + na\sigma^2_{BC} + na\sigma^2_B$
C	$\sigma^2 + n\sigma^2_{ABC} + nb\sigma^2_{AC} + na\sigma^2_{BC} + nab\sigma^2_C$
AB	$\sigma^2 + n\sigma^2_{ABC} + n\sigma^2_{AB}$
AC	$\sigma^2 + n\sigma^2_{ABC} + nb\sigma^2_{AC}$
BC	$\sigma^2 + n\sigma^2_{ABC} + na\sigma^2_{BC}$
ABC	$\sigma^2 + n\sigma^2_{ABC}$
Повторность	σ^2

Однако вероятность превзойти этой величине некоторое заданное значение может быть приближенно определена по таблице F, если для этого исчислить соответствующие числа степеней свободы:

$$n_1 = \frac{MS^2}{\frac{MS^2_A}{f_A} + \frac{MS^2_{ABC}}{f_{ABC}}};$$

$$n_2 = \frac{4f_2}{\frac{MS^2_{AB}}{f_{AB}} + \frac{MS^2_{AC}}{f_{AC}}},$$

где MS означает средний квадрат и f — соответствующее число степеней свободы.

Хотя опыт с кормлением свиней (табл. 153) не относится к типу II, все же его данные могут быть использованы для иллюстрации указанного выше критерия существенности. Положим, что надлежит произвести оценку действия протеина в предположении, что все эффекты имеют распределение случайного

характера. Средние квадраты, входящие в приведенные ранее формулы, будут иметь теперь подстрочный указатель P :

$$MS_1 = MS_P + MS_{LMP} = 0,5355 + 0,0114 = 0,5469;$$

$$MS_2 = MS_{LP} + MS_{MP} = 0,0800 + 0,0410 = 0,1210;$$

$$F' = 0,5469/0,1210 = 4,52;$$

$$n_1 = \frac{0,5469^2}{\frac{0,5355^2}{1} + \frac{0,0114^2}{6}} = 1,04;$$

$$n_2 = \frac{0,1210^2}{\frac{0,0800^2}{3} + \frac{0,0410^2}{2}} = 4,9.$$

Обычно в таких случаях вычисления n_1 и n_2 допустимо производить с округлением до целых единиц. В данном случае нет никаких сомнений, что F не достигает 5%-ного уровня. В противоположность этому, правильный критерий (соответствующий модели I) показывает, что эффект протеина весьма существен.

При смешанных моделях следует начать с выписки компонентов для модели II, после чего применить следующее простое правило.

П р а в и л о 3. Просматриваются подстрочные указатели у всех компонентов данного эффекта, при этом не обращается внимание на подстрочные буквы, одинаковые с обозначением этого эффекта. Если какая-нибудь из других подстрочных букв соответствует фиксированному фактору, то этот компонент отбрасывается. Это правило не распространяется ни на первый (σ^2), ни на последний компоненты каждой строки.

Для иллюстрации допустим, что A в таблице 154 имеет заранее фиксированные градации. Рассматриваем первый компонент в первой строке σ_{ABC}^2 . Игнорируя A , видим, что как B , так и C соответствуют случайно меняющимся эффектам, поэтому данный компонент сохраняется. Подобным же образом устанавливается необходимость сохранения σ_{AB}^2 и σ_{AC}^2 . Последний компонент σ_A^2 всегда сохраняется, так как на него приведенное выше правило не распространяется.

Во второй строке, игнорируя в подстрочном указателе σ_{ABC}^2 букву B , встречаем A , относящееся к фиксированному фактору, и поэтому данный компонент отбрасываем. Уничтожается также и σ_{AB}^2 , но σ_{BC}^2 удерживается.

В AB имеется только один компонент σ_{ABC}^2 , к которому применимо это правило. Игнорируя в подстрочном указателе символ рассматриваемого эффекта AB , имеем C — случайно меняющийся фактор, и поэтому данный компонент необходимо сохранить. Но по строке BC такой же компонент вычеркивается, потому что A — фактор с фиксированными градациями. Окончательный компонентный анализ для этого случая представлен в таблице 155.

Оценка существенности здесь не встречается затруднений, кроме оценки одного главного эффекта A . В этом последнем случае следует применить указанный приближенный метод оценки.

Если эффекты двух факторов A и B фиксированы и только у фактора C имеет место случайное варьирование эффекта, то применение правила к таблице 155 (или к таблице 154) приводит к вычеркиванию таких добавочных эффектов:

$$\text{в } A: \sigma_{ABC}^2, \sigma_{AB}^2$$

$$\text{в } C: \sigma_{BC}^2$$

$$\text{в } AC: \sigma_{ABC}^2$$

Если, наконец, фиксированы все факторы A , B и C , то дальнейшее применение данного правила приводит к компонентному анализу таблицы 153.

Компонентный анализ трехфакторного опыта. Смешанная модель:
A — фиксировано, *B* и *C* — случайны. Полностью рендомизированное
 расположение

<i>A</i>	$\sigma^2 + n\sigma_{ABC}^2 + nc\sigma_{AB}^2 + nb\sigma_{AC}^2 + nbck_A^2$
<i>B</i>	$\sigma^2 + na\sigma_{BC}^2 + nac\sigma_B^2$
<i>C</i>	$\sigma^2 + na\sigma_{BC}^2 + nab\sigma_C^2$
<i>AB</i>	$\sigma^2 + n\sigma_{ABC}^2 + nc\sigma_{AB}^2$
<i>AC</i>	$\sigma^2 + n\sigma_{ABC}^2 + nb\sigma_{AC}^2$
<i>BC</i>	$\sigma^2 + na\sigma_{BC}^2$
<i>ABC</i>	$\sigma^2 + n\sigma_{ABC}^2$
<i>R</i>	σ^2

Пример 13. В таблице 153 суммы квадратов для *LM*, *MP* и *LMP* все столь малы, что ни одна из единичных степеней свободы, выделенная из них, не может достигнуть уровня существенности. Но *LM* и *LP* подлежат в этом отношении дальнейшему изучению.

При рассмотрении *LM* в сводной табличке *A* таблицы 152 создается впечатление о наличии некоторого взаимодействия, хотя общий для всех 6 степеней свободы критерий не обнаруживает его. Но приглядитесь к линейным эффектам. Сначала вычислите M_L (−1, 0, +1) для каждой дозы лизина (здесь и в следующем примере следует различать: *L* — подстрочное означает «линейное», а *L* строчное — «лизин»):

$$-0,08, -0,27, 0,57 \text{ и } 1,07.$$

Далее возьмите линейный эффект лизина (−3, −1, +1, +3), определенный по этим M_L ; в результате получите 4,29. Наконец, применение правила 6 параграфа 2 этой главы дает сумму квадратов:

$$L_L M_L = \frac{4,29^2}{4 \times 2 \times 20} = 0,1150,$$

что все же ниже 5%-ного уровня существенности. Ни одно из остальных пяти сравнений этого взаимодействия также не будет существенным. В более обширном опыте, частью которого являются представленные здесь данные, $L_L M_L$ было существенным. Какую интерпретацию вы дадите этому факту?

Пример 14. В сводной табличке *C* взаимодействие *LP* характеризуется разностями между данными для 14% и 12% протенна:

$$2,15; 2,07; 0,29; 0,56.$$

Линейные эффекты лизина (−3, −1, 1, 3) при двух дозах протенна составляют 4,04 и −2,51. Беря разность между ними в качестве линейного эффекта протенна (−1, +1), мы получим 4,04 − (−2,51) = 6,55. По правилу 6 параграфа 2:

$$L_L P_L = \frac{6,55^2}{6 \times 2 \times 20} = 0,1788.$$

$F = 0,1788 / 0,0275 = 6,50$; $P = 0,025$. Это согласуется с весьма существенным эффектом, который был установлен в таблице 151, где была дана и соответствующая интерпретация данного факта.

Вычитание из суммы квадратов *LP* таблицы 153 определенного сейчас среднего квадрата $L_L P_L$, т. е. 0,2339 − 0,1788 = 0,0611, убеждает в том, что ни одно из двух оставшихся сравнений не будет существенным.

Пример 15. Исследователь часто интересуется больше оценкой разностей, чем оценкой существенности взаимодействий. Поэтому он, например, может пожелать вместо взаимодействия *LP* оценить эффект протенна без лизина. Сводная таблица *C* даст соответствующую среднюю разность (8,37 − 6,22)/6 = 0,36 фунта привеса в день. (В данном случае взяты все дозы метионина, так как в опыте не выявлены ни главный эффект этого фактора, ни взаимодействие его с протенином.) Доверительное отклонение будет $2,069 \times \sqrt{2 \times 0,0275 / 6} = 0,20$ фунта в день. Таким образом, 95%-ный интервал будет от 0,16 до 0,56 фунта в день.

Пример 16. Некоторые опыты представляют собой комбинацию полностью рендомизированного расположения с построением опыта при двойной группировке, как это имеет место в рендомизированных блоках. В приведенной ниже таблице помещены данные, взятые из отчета, на который мы уже ссылались ранее (28). В случайном порядке были взяты три растения, и с каждого из них также случайно отобрано по четыре листа. Из водного раствора золы от каждого листа были взяты по две повторные пробы в 2 мл

для анализа на кальций. Второй ряд анализов был проведен на следующий день. В данном случае модель эксперимента такова:

$$X_{ijkm} = \mu + P_i + L_{ij} + A_k + (PA)_{ik} + (LA)_{ijh} + \varepsilon_{ijkm};$$

$$i = 1, \dots, 3; j = 1, \dots, 4; k = 1, 2; m = 1, 2; \varepsilon = N(0, \sigma).$$

Все эффекты можно считать случайно изменяющимися; так что $\sigma_P^2, \dots, \sigma_{LA}^2$ берутся в качестве дисперсий из нормальных совокупностей.

Отличительной особенностью этого построения опыта является то, что лист имеет предназначенный номер j на каждом растении i , что соответствует схеме, рассмотренной в главе 10.

Данные этого опыта и дисперсионный анализ для них приведены в объединенной таблице. Символ $L(P)$ означает «листья одного растения»; введение его необходимо для

Трехфакторный опыт с турнепсом ($2 \times 3 \times 4$), полностью рандомизированное размещение. Содержание кальция (в % к сухому веществу)

Растение	Лист	Сжигание 1-е			Сжигание 2-е			Сумма для листа	Сумма для растения
		R_1	R_2	сумма	R_1	R_2	сумма		
1	1	3,28	3,09	6,37	3,03	3,03	6,06	12,43	
	2	3,52	3,48	7,00	3,38	3,38	6,76	13,76	
	3	2,88	2,80	5,68	2,81	2,76	5,57	11,25	
	4	3,34	3,38	6,72	3,23	3,26	6,49	13,21	
			13,02	12,75	25,77	12,45	12,43	24,88	
2	1	2,63	2,66	5,29	2,54	2,59	5,13	10,42	
	2	3,74	3,44	7,18	3,58	3,64	7,22	14,40	
	3	2,55	2,55	5,10	2,49	2,49	4,98	10,08	
	4	2,77	2,66	5,43	2,86	2,80	5,66	11,09	
			11,69	11,31	23,00	11,47	11,52	22,99	
3	1	3,78	3,87	7,65	3,88	3,82	7,70	15,35	
	2	4,31	4,40	8,71	3,99	4,05	8,04	16,75	
	3	3,31	3,31	6,62	3,19	3,13	6,32	12,94	
	4	4,07	4,12	8,19	4,00	4,08	8,08	16,27	
			15,47	15,70	31,17	15,06	15,08	30,14	
Сумма				79,94			78,01		157,95

Дисперсионный анализ данных о содержании кальция. Модель II

Источник варьирования	Число степеней свободы	Средний квадрат	Параметры, оцениваемые средним квадратом
Растения $P(p=3)$. . .	2	3,8548*	$\sigma^2 + 2\sigma_{AL(P)}^2 + 8\sigma_{AP}^2 + 4\sigma_{L(P)}^2 + 16\sigma_P^2$
Листья на растении $L(P) (l=4)$	9	0,6646**	$\sigma^2 + 2\sigma_{AL(P)}^2 + 4\sigma_{L(P)}^2$
Сжигания $A (a=2)$. . .	1	0,0776	$\sigma^2 + 2\sigma_{AL(P)}^2 + 8\sigma_{AP}^2 + 24\sigma_A^2$
AP	2	0,0191	$\sigma^2 + 2\sigma_{AL(P)}^2 + 8\sigma_{AP}^2$
$AL(P)$	9	0,0112*	$\sigma^2 + 2\sigma_{AL(P)}^2$
Повторения $R(n=2)$. .	24	0,00412	σ^2

применения правил компонентного анализа. В соответствии с особенностями построения опыта правило 1 параграфа 10 прилагается к компонентам P по-разному. Правило 2 этого параграфа приводит к коэффициенту для P , равному $na=4$. Буква P поставлена в скобки, чтобы показать отличие варьирования внутри растения от символа взаимодействия. Применяя приведенные ранее правила, проверьте результаты анализа.

Пример 17. Проверьте указанные ранее критерии существенности для опыта с кальцием. Для растений $F' = 5,65$; $f' = 2,0$ и $9,5$.

Пример 18. Какие компоненты в опыте с кальцием будут вычеркнуты, если считать сжигание фиксированным фактором. *Ответ:* из $L(P)$ вычеркивается $\sigma_{AL(P)}^2$, из P вычеркиваются $\sigma_{AL(P)}^2$ и σ_{AP}^2 . Каково влияние это окажет на критерии существенности?

Пример 19. Какие компоненты дополнительно к этому исчезнут, если вместе с A будет фиксированным и P ? *Ответ:* из A вычеркивается $\sigma_{AL(P)}^2$ и σ_{AP}^2 . Каково будет влияние этого на критерии существенности?

11. Способ расщепленных делянок. Довольно часто в отношении какого-либо одного фактора и его взаимодействия с другим фактором требуется получить достаточно точную информацию, а в отношении второго фактора нет необходимости добиваться столь большой точности. Например, при изучении трех источников получения животным некоторого витамина можно производить нужные для этого сравнения путем испытания их, положим, на трех самцах одного и того же помета, повторяя эксперимент на 20 пометах. Это будет схема рендомизированных блоков, дающая высокую точность в связи с тем, что на ошибку опыта приходится 38 степеней свободы. На этот опыт может быть наложен некоторый другой эксперимент, в котором единичей наблюдения является помет в целом. Положим, на отдельных пометах должно быть испытано четыре вида содержания животных, что даст 5 повторений при 12 степенях свободы для ошибки. Главные варианты (способы содержания животных) не будут между собой сравнимы с такой точностью, с какой сравнимы субварианты (источники витамина), так как, во-первых, здесь будет меньшая повторность, и, во-вторых, здесь для определения эффектов содержания в ошибку будут включаться различия между пометами. Тем не менее в этом случае при небольших дополнительных затратах может быть получена определенная информация относительно способов содержания животных, а вместе с этим будет с достаточной точностью определено и взаимодействие между содержанием животных и источниками снабжения их витамином.

В опытах по сортоиспытанию или с удобрениями, проводимыми на малых делянках, можно объединять целые группы таких делянок и на укрупненных делянках можно производить испытание различных видов обработки почвы с применением крупных машин (в частности, полив посевов является одним из таких мероприятий), которые требуют довольно больших площадей для отдельных вариантов). Набор таких вариантов обработки почвы обычно повторяется в опыте только очень небольшое число раз, но сорта или удобрения при этом способе повторяются по всем делянкам обработки и имеют, таким образом, большое число повторений. Опыты такого вида называются опытами с *расщепленными делянками*; здесь делянки, на которых изучаются варианты обработки почвы, называемые *главными делянками*, расщепляются на малые *субделянки*, предназначенные для сортов. Эта схема опыта имеет широкую известность.

Существенной особенностью опыта с расщепленными делянками является то, что варианты, изучаемые на субделянках, не рендомизируются по всему крупному блоку опыта, а распределяются по жребью только внутри каждой главной делянки. Таким образом, рендомизация субвариантов производится заново по каждой главной делянке, а варианты главных делянок рендомизируются по большим блокам. Вследствие этого ошибка опыта для субвариантов шире (обычно значительно меньшая), чем ошибка вариантов, изучаемых на главных делянках.

Рисунок 25 показывает расположение на поле опыта с расщепленными делянками по изучению трех сортов люцерны, в котором субвариантами являлись четыре даты последнего укоса культуры [31]. Первые два укоса были одновременными по всем делянкам опыта, причем второй укос был проведен

27 июля 1943 г. Третий укос был таким: на делянках *A* не проводился, на делянках *B* проведен 1 сентября, на делянках *B* — 20 сентября и на делянках *Г* — 7 октября. Урожай 1944 г. представлен в таблице 156. Конечно, опыт этого вида не должен ограничиваться одним годом; статистические методы обработки данных по многолетним культурам рассмотрены ниже.

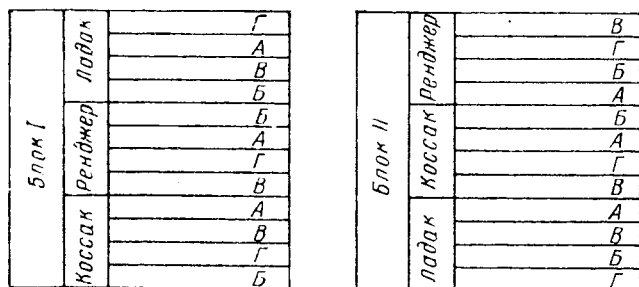


Рис. 25. Первые два блока опыта с люцерной на расщепленных делянках, иллюстрирующие случайное расположение главных делянок и субделянок.

зируются, как опыт с четырьмя датами укосов, рендомизированными в каждой из 18 главных делянок (табл. 157).

Существенные различия между датами укосов не были неожиданными, как нет неожиданности в уменьшении урожая при *B* и *B*. Последний срок уборки был или слишком ранним для возобновления роста и восстановления запасов питательных веществ в корнях, или столь поздним, что ни вегета-

ТАБЛИЦА 156

Урожай трех сортов люцерны (в т на 1 акр) в 1944 г. в зависимости от четырех дат проведения последнего укоса в 1943 г.

Сорта	Даты	Блоки					
		1	2	3	4	5	6
Ладак	<i>A</i>	2,17	1,88	1,62	2,34	1,58	1,66
	<i>B</i>	1,58	1,26	1,22	1,59	1,25	0,94
	<i>B</i>	2,29	1,60	1,67	1,91	1,39	1,12
	<i>Г</i>	2,23	2,01	1,82	2,10	1,66	1,10
	.	.	8,27	6,75	6,33	7,94	5,88
Коссак	<i>A</i>	2,33	2,01	1,70	1,78	1,42	1,35
	<i>B</i>	1,38	1,30	1,85	1,09	1,13	1,06
	<i>B</i>	1,86	1,70	1,81	1,54	1,67	0,88
	<i>Г</i>	2,27	1,81	2,01	1,40	1,31	1,06
	.	.	7,84	6,82	7,37	5,81	5,53
Ренджер	<i>A</i>	1,75	1,95	2,13	1,78	1,31	1,30
	<i>B</i>	1,52	1,47	1,80	1,37	1,01	1,31
	<i>B</i>	1,55	1,61	1,82	1,56	1,23	1,13
	<i>Г</i>	1,56	1,72	1,99	1,55	1,51	1,33
	.	.	6,38	6,75	7,74	6,26	5,06
Итого		22,49	20,32	21,44	20,01	16,47	14,24

Сорта	Даты уборки				Итого
	А	Б	В	Г	
Ладак	11,25	7,84	9,98	10,92	39,99
Коссак	10,59	7,81	9,46	9,86	37,72
Ренджер	10,22	8,48	8,90	9,66	37,26
Итого	32,06	24,13	28,34	30,44	114,97
Средняя (в т на 1 акр)	1,78	1,34	1,57	1,69	

ция, ни использование запасов не были возможными. В этом опыте два неожиданных момента: во-первых, урожай *В* оказался выше *Б* ввиду того, что конец сентября считается в штате Айова неблагоприятным для уборки люцерны, и, во-вторых, отсутствие взаимодействия между датами укосов и сортами, так как сорт Ладак обычно после укоса медленно возобновляет свой рост и поэтому должен иначе реагировать на сроки укоса, чем другие сорта. Вероятно, это было обусловлено особенностями сезона.

Модель для опыта с расщепленными делянками в рендомизированных блоках будет такой:

$$X_{ijk} = \mu + B_i + M_j + \varepsilon_{ij} + T_k + (MT)_{jk} + \delta_{ijk}$$

$$i = 1, \dots, b; \quad j = 1, \dots, m; \quad k = 1, \dots, t; \quad \varepsilon_{ij} = N(0, \sigma_M); \quad \delta_{ijk} = N(0, \sigma_I).$$

Здесь *M* означает вариант, испытываемый на главной делянке, и *T* — вариант субделянки.

ТАБЛИЦА 157

Дисперсионный анализ опыта с люцерной на расщепленных делянках

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Главные делянки:			
сорта	2	0,1781	0,0890
блоки	5	4,1499	0,8300
Ошибка главных делянок	10	1,3622	0,1362
Субделянки			
даты уборки	3	1,9625	0,6542**
даты × сорта	6	0,2105	0,0351
Ошибка субделянок	45	1,2586	0,0280

1. Поправка: $C = 114,9^2/72 = 183,5847$.
2. Общее: $2,17^2 + \dots + 1,33^2 - C = 9,1268$.
3. Главные делянки $(8,27^2 + \dots + 5,07^2)/4 - C = 5,6902$.
4. Сорта: $(39,99^2 + \dots + 37,26^2)/24 - C = 0,1781$.
5. Блоки: $(22,49^2 + \dots + 14,24^2)/12 - C = 4,1499$.
6. Ошибка главных делянок: $5,6902 - (0,1781 + 4,1499) = 1,3622$.
7. Подклассы в таблице сорта × даты: $(11,25^2 + \dots + 9,66^2)/6 - C = 2,3511$.
8. Даты $(32,06^2 + \dots + 30,44^2)/18 - C = 1,9625$.
9. Сорта × даты: $2,3511 - (0,1781 + 1,9625) = 0,2105$.
10. Ошибка субделянок: $9,1218 - (5,6902 + 1,9625 + 0,2105) = 1,2586$.

В таблице 158 в общем виде представлен компонентный анализ для опыта с расщепленными делянками в рендомизированных блоках. Эта модель типа II легко получается из полной структуры трехфакторного опыта таблицы 154 путем объединения в две ошибки (главных делянок и субделянок) всех «взаимодействий» с блоками.

Если *T* фиксировано, то оценка существенности производится непосредственно при помощи *F*; если же *T* случайно меняющееся, то эта оценка вари-

Компонентный анализ опытов с расщепленными делянками: m — вариантов главных делянок; t — вариантов субделянок b — блоков. Случайные блоки

Источник варьирования	M и T случайны	M — случайно, T — фиксировано	M — фиксировано, T — случайно	M и T — фиксированы
Варианты главных делянок M	$E_a + b\sigma_{MT}^2 + bt\sigma_M^2$	$E_a + bt\sigma_M^2$	$E_a + b\sigma_{MT}^2 + bt\kappa_M^2$	$E_a + bt\kappa_M^2$
Ошибка (a) . . .	E_a	E_a	E_a	E_a
Варианты субделянок T . . .	$E_b + b\sigma_{MT}^2 + bm\sigma_T^2$	$E_b + b\sigma_{MT}^2 + bm\kappa_T^2$	$E_b + bm\sigma_T^2$	$E_b + bm\kappa_T^2$
MT	$E_b + b\sigma_{MT}^2$	$E_b + b\sigma_{MT}^2$	$E_b + b\sigma_{MT}^2$	$E_b + b\kappa_{MT}^2$
Ошибка (b)	E_b	E_b	E_b	E_b

Андерсон и Банкрофт ([2], стр. 347) считают, что могут быть случаи, когда возникает необходимость разложения ошибки (b) на две части BT и BM . Но обычно эти две величины ожидаются тождественными и берутся в качестве оценки одного и того же значения σ^2 .

Если взаимодействие отсутствует, то сравнения каждого из двух факторов производятся, как обычно, но только на основе степеней свободы для ошибки (a) или ошибки (b), в зависимости от того, какой берется фактор. Если критерий F указывает на существенность действия того или другого фактора, то оценка разностей может быть произведена в соответствии с параграфом 6 главы 10.

В случаях же, когда имеет место взаимодействие между факторами главных делянок и субделянок, становится необходимым рассмотрение уже таблицы с двумя входами, подобной той, которая приведена в конце таблицы 156. При этом может оказаться необходимым, например, для разных сортов дать совершенно различные рекомендации и оставить без внимания главные эффекты факторов.

Вопрос о выпадении данных в опыте с расщепленными делянками изучен Андерсоном [1]. При одной выпавшей субделянке применяется формула:

$$X_{ijh} = \frac{bX_{ij.} + tX_{.jh} - X_{.j.}}{(b-1)(t-1)},$$

где b — число блоков (повторений) и t — число субвариантов. Другие переменные символы имеют обычный смысл, а именно: суммы остальных урожаев той главной делянки ($X_{ij.}$), того субварианта ($X_{.jh}$) и того варианта ($X_{.j.}$), изучаемого на главных делянках, где находится выпавшее наблюдение.

Пример. Допустим, что в таблице 156 выпало наблюдение $X_{111} = 2,17$; здесь $b = 6$ и $t = 4$; $X_{11.} = 8,27 - 2,17 = 6,10$ является суммой остальных урожаев главной делянки для сорта Ладак в первом блоке; $X_{.11} = 1,88 + 1,62 + \dots + 1,66 = 9,08$ является суммой остальных урожаев сорта Ладак при сроке уборки A ; $X_{.j.} = 39,99 - 2,17 = 37,82$ является суммой остальных урожаев сорта Ладак. Подстановка этих значений дает:

$$X_{111} = \frac{6 \times 6,10 + 4 \times 9,08 - 37,82}{(6-1) \times (4-1)} = 2,34 \text{ т на 1 акр.}$$

Все средние квадраты, кроме ошибки (b), будут несколько преувеличены. Это преувеличение, вероятно, незначительно, но все же Андерсон дает метод соответствующей корректировки.

Экспериментаторы часто применяют расщепление и субделянок и даже дальнейшее дробление этих субсубделянок. Статистические методы обработки таких опытов не усложняются в этом случае очень сильно. В таблице 159 приведена структура модели II для опыта с двойным расщеплением деля-

нок. Как обычно, здесь строчные буквы *b*, *m* и т. д. означают числа градаций у факторов, обозначенных соответствующими прописными буквами. Смешанные модели и модель I получаются путем применения правил, перечисленных в предыдущем параграфе. При модели I каждый эффект будет состоять только из двух компонентов: из первого и последнего в каждой строке таблицы. Вычисления ведутся, начиная с блоков и последовательно переходя сначала к двухфакторной, а потом к трехфакторной схеме. Каждый вид ошибки находится путем вычитания предшествующих.

ТАБЛИЦА 159

Структура модели II для опыта с двойным расщеплением делянок

Блоки <i>B</i>	
Главные делянки <i>M</i>	$E_a + b\sigma^2_{MNT} + bn\sigma^2_{MT} + bt\sigma^2_{MN} + bnt\sigma^2_M$
Ошибка (<i>a</i>) E_a	E_a
Субделянки <i>N</i>	$E_b + b\sigma^2_{MNT} + bm\sigma^2_{NT} + bt\sigma^2_{MN} + bmt\sigma^2_N$
<i>MN</i>	$E_b + b\sigma^2_{MNT} + bt\sigma^2_{MN}$
Ошибка (<i>b</i>) E_b	E_b
Субсубделянки <i>T</i>	$E_c + b\sigma^2_{MNT} + bm\sigma^2_{NT} + bn\sigma^2_{MT} + bmn\sigma^2_T$
<i>MT</i>	$E_c + b\sigma^2_{MNT} + bn\sigma^2_{MT}$
<i>NT</i>	$E_c + b\sigma^2_{MNT} + bm\sigma^2_{NT}$
<i>MNT</i>	$E_c + b\sigma^2_{MNT}$
Ошибка (<i>c</i>) E_c	E_c

Пример 20. При изучении трех густот стояния и трех доз удобрения с поливом и без полива в опыте с кукурузой [4] применен способ двойного расщепления. Схема опыта — рандомизированные блоки в 4-кратной повторности. На главных делянках изучались варианты полива. Каждая из этих делянок разделена на 3 части, на которых испытывались густоты стояния 10 000, 13 000 и 16 000 растений на 1 акр. Наконец, каждая из субделянок была разделена на 3 части соответственно трем изучаемым дозам азота: 60, 120 и 180 фунтов на 1 акр. Урожай даны в бушелях на 1 акр. Проведите дисперсионный анализ.

Полив	Густота стояния	Удобрение	Блоки			
			1	2	3	4
Без полива	1	1	90	83	85	86
		2	95	80	88	78
		3	107	95	88	89
	2	1	92	98	112	79
		2	89	98	104	86
		3	92	106	91	87
	3	1	81	74	82	85
		2	92	81	78	89
		3	93	74	94	83
Полив	1	1	80	102	60	73
		2	87	109	104	114
		3	100	105	114	114
	2	1	121	99	90	109
		2	110	94	118	131
		3	119	123	113	126
	3	1	78	136	119	116
		2	98	133	122	136
		3	122	132	136	133

Источник варьирования	Число степеней свободы	Средний квадрат
Главные делянки:		
Блоки	3	
Полив J	1	8277,56
Ошибка (a)	3	470,59
Субделянки:		
Густота стояния S	2	879,18
JS	2	1373,51*
Ошибка (b)	12	232,33
Субсубделянки:		
Удобрение F	2	988,72
JF	2	476,72**
SF	4	76,22
JSF	4	58,68
Ошибка (c)	36	86,36

Пример 21. В предыдущем примере обращает внимание наличие двух существенных взаимодействий JS и JF ; в то же время JSF меньше соответствующей ошибки. Это означает, что взаимодействие JS остается одним и тем же при всех градациях F , или, по-другому, взаимодействие JF также остается одним и тем же при всех градациях S . Следовательно, каждая из двух соответствующих таблиц с двумя входами дает сведения об этих взаимодействиях:

	F_1	F_2	F_3	S_1	S_2	S_3
Без полива	1047	1058	1099	1064	1134	1006
Полив	1183	1356	1437	1162	1353	1461

Ясно, что ни удобрение, ни густота стояния на делянках без полива не влияют на урожай; при поливе же влияние обоих факторов ярко выражено. Отсюда возникает необходимость самостоятельного изучения опыта с расчепленными делянками только при наличии полива. Проверьте следующие средние квадраты:

Густота стояния:	
линейность эффекта	1 3725**
отклонения	1 96
Ошибка (a)	6 316
Удобрение:	
линейность эффекта	1 2688**
отклонения	1 118
SF	4 92
Ошибка (b)	18 137

Пример 22. Заметим, что в данном опыте степень возрастания градаций густоты стояния и удобрения были хорошо подобраны для варианта без полива, но при поливе они оказались слишком малыми для того, чтобы установить оптимальные значения этих факторов. Отсюда следует, что полив не должен быть фактором, изучаемым в подобного рода опытах. Но в целях сравнения денежных затрат и обобщения результатов за ряд лет вполне уместно параллельное проведение двух таких опытов (с поливом и без полива) при случайном их размещении для исключения почвенных различий.

12. Серия опытов. Серия опытов может быть проведена или в ряде пунктов, или за ряд лет, или и то и другое вместе. В этих случаях в анализ включаются пункты или года в качестве дополнительных факторов. Каждый фактор, вообще говоря, может быть заранее планируемым или случайно

выбираемым. Но я никогда не встречал опыта, в котором год проведения фактически выбирался бы по жребию; случайный же выбор пункта если и встречается, то редко и только в опытах типа обследования, однако *эффекты* годов и пунктов часто носят характер случайной переменной.

Совершенно недостаточно одного указания на то, что анализ каждого отдельного ряда опытов представляет собой как для статистики, так и для экспериментаторов самостоятельную задачу. Величина и схема опыта могут меняться от пункта к пункту или от года к году. Наличие взаимодействий вариантов с пунктами или вариантов с годами приводит к усложнению анализа. Ошибка опыта может различаться по пунктам и годам. В связи с этим экспериментатор, не чувствуя себя хорошо осведомленным в области статистической теории, поступит вполне благоразумно, если перед проведением серии опыта свяжется с компетентным математиком-статистиком. Мои наблюдения показывают, что много времени и средств тратится зря на неправильно проведенные серии экспериментов.

В литературе [7, 33, 10, 17, 9] имеются прекрасные иллюстрации тех затруднений, которые возникают при непосредственном подходе даже к менее сложным случаям. Предлагаемые в настоящей книге методы достаточны для того, чтобы читатель мог обратиться к указанной выше литературе.

Читателю полезно познакомиться с трудностями, которые возникают при изучении многолетних культур в течение нескольких лет. Урожай одной делянки в последовательные годы, по всей вероятности, коррелированы; ошибка опыта в одном году не будет независимой от ошибки другого года. По этим причинам первоначальный анализ проводится на основе суммарных урожаев делянок за весь период проведения эксперимента.

Для иллюстрации возьмем данные таблицы 160 опыта Хейбера [12] по изучению влияния различных вариантов срезки спаржи. Плантация заложена в 1927 г., а срезка началась с 1929 г. Спаржу на одной делянке убирали до 1 июня каждого года, на трех делянках соответственно 15 июня, 1 июля и 15 июля. Урожай определялись весами сборов на 1 июня, независимо от последующей срезки на каждой делянке. В этом случае веса измеряют силу отрастания, и, следовательно, в данном эксперименте изучается относительная эффективность времени уборки урожая.

Во избежание влияния корреляции урожай по годам сложены вместе, и обработка проведена по этим суммам. Конечно, здесь нет никакого сомне-

ТАБЛИЦА 160

Веса (в унциях) урожаев спаржи на 1 июня на делянках с различными вариантами срезки

Блок	Год	Прекращение срезки				Итого
		1 июня	15 июня	1 июля	15 июля	
1	1930	230	212	183	148	773
	1931	324	415	320	246	1305
	1932	512	584	456	304	1856
	1933	399	386	255	144	1184
			1465	1597	1214	842
2	1930	216	190	186	126	718
	1931	317	296	295	201	1109
	1932	448	471	387	289	1595
	1933	361	280	187	83	911
			1342	1237	1055	699

Блок	Год	Преграшение срезки				Итого
		1 июня	15 июня	1 июля	15 июля	
3	1930	219	151	177	107	654
	1931	357	278	298	192	1125
	1932	496	399	427	271	1593
	1933	344	254	239	90	927
			1416	1082	1141	660
4	1930	200	150	209	168	727
	1931	362	336	328	226	1252
	1932	540	485	462	312	1799
	1933	381	279	244	168	1072
			1483	1250	1243	874
Итого		5706	5166	4653	3075	18 600

Блоки	3	30170	
Сроки уборки урожая .	(3)	(241377)	
Линейный компонент	1		220815**
Квадратичный	1		16835*
Остаток	1		3727
Ошибка	9		2429

ния, что срезка побегов влияет на мощность растения при условии, что последняя измеряется урожайностью на 1 июня.

На основе итогов по вариантам можно определить линейный, квадратичный и кубические эффекты; соответствующие суммы квадратов у нас внесены в таблицу дисперсионного анализа. Существенность квадратичного компонента указывает на то, что чем более поздние сроки уборки спаржи, урожайность все более и более быстро падает. Если построить по средним величинам (или соответствующим суммам) график, то это положение становится совсем ясным. Так как здесь «остаток» только немного больше ошибки, то отклонения от параболы можно приписать выборочному варьированию.

Данный опыт содержит в себе много больше сведений, чем это выявлено в приведенном выше дисперсионном анализе. Корреляцию ошибок в большей мере можно избежать, обрабатывая разности между годами. Например, линейный компонент на годы для одного года должен быть независимым от такого же компонента второго года. Результаты этого опыта были подробно рассмотрены Снедекором и Хейбером [13,27], а также в 4-м издании настоящей книги. Методы статистической обработки этих данных были указаны в предыдущих параграфах, и поэтому здесь нет необходимости их повторять.

Пример 23. Приводимые ниже данные получены в серии опытов, заложенных в 5 пунктах [19]. На каждой из 5 принимающих участие в опыте станций высевались семена мукдесной сои; четыре группы по 100 семян были протравлены четырьмя химикатами, а пятая оставлена без обработки; опыт заложен по схеме пяти репдоминированных блоков. Общее число проросших семян (из 500) указано для каждого из 5 пунктов в приводимой ниже таблице. Вместе с этим приведены и результаты дисперсионного анализа по каждому пункту.

Число проросших семян (из 500) в 5 пунктах. Коллективный опыт по протравливанию семян мукденской сои. 1943 г.

Пункты	Не обрабо- таны	Обработаны				Итого
		арасаном	спергоном	семесаном	фермейтом	
Мичиган	360	356	362	350	373	1801
Миннесота	302	354	349	332	332	1669
Висконсин	408	407	391	391	409	2006
Виргиния	244	267	293	235	278	1317
Род Айленд	373	387	406	394	375	1935
Итого	1687	1771	1801	1702	1767	8728

Средние квадраты из первоначальных дисперсионных анализов

Источник варьирова- ния	Число степеней свободы	Пункты				
		Мичиган	Миннесота	Висконсин	Виргиния	Род Айленд
Варианты	4	14,44	82,84*	17,44	114,26*	37,50
Блоки . . .	4	185,14	54,64	5,64	70,76	4,80
Ошибка . .	16	42,29	26,67	30,64	26,34	13,05

Проверьте гипотезу однородности дисперсий ошибки. *Ответ:* исправленное $\chi^2 = 5,22$, $f = 4$.

Пример 24. Для объединенных данных о сое провести следующий дисперсионный анализ:

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Варианты	4	380,29	95,07
Пункты	4	11852,61	2963,15
Взаимодействие	16	685,63	52,85
Блоки в пунктах	20	1283,92	—
Ошибка опыта	80	2223,68	27,80

Данные по блокам и ошибке опыта взяты в суммарном виде из данных анализа по отдельным пунктам.

Пример 25. Выпишите компонентный анализ, считая варианты фиксированными, а пункты и блоки факторами со случайным изменением градаций:

$$\begin{aligned} \text{Варианты } T & \quad \sigma^2 + 5\sigma_{TL}^2 + 25\sigma_T^2 \\ \text{Пункты } L & \quad \sigma^2 + 5\sigma_{B(L)}^2 + 25\sigma_L^2 \\ TL & \quad \sigma^2 + 5\sigma_{TL}^2 \\ \text{Блоки в пунктах } B(L) & \quad \sigma^2 + 5\sigma_{B(L)}^2 \\ \text{Объединенная ошибка} & \quad \sigma^2 \end{aligned}$$

Для вариантов: $F = 95,07/42,85 = 2,22$; $F_{0,05} = 3,01$.

Пример 26. Выделите сумму квадратов для планируемого сравнения всех вариантов взятых вместе с контролем. *Ответ:* 171,70; $F = 4,01$; $F_{0,05} = 4,49$.

13. Пропорциональные численности субклассов. По тем или иным причинам числа наблюдений для вариантов в таблице, подобной нашей 135, могут быть неодинаковыми. Когда по случайной ли причине или в силу ограниче-

ний, налагаемых техникой эксперимента, встречается некоторое выпадение наблюдений, теорема слагаемости сумм квадратов теряет свою силу, и необходимо определенное видоизменение дисперсионного анализа. Соответствующие методы и будут служить темой нижеследующих параграфов. Однако иногда численности субклассов, будучи неодинаковыми, находятся друг с другом в определенной пропорции. Так было в некоторых из предшествующих таблиц; в таких случаях общая схема дисперсионного анализа не нарушается.

В таблице 161 приведены данные о проценте (за вычетом 70%) убойного веса в живом весе свиней разных пород и полов в пропорции примерно такой, какой эти животные были выращены на убой мясной лабораторией колледжа штата Айова [5]. Для иллюстрации мы взяли только небольшую выборку из имеющихся данных. Все представленные здесь свиньи имели живой вес между 200 и 219 фунтами; при таких больших знаменателях, использованных при вычислении указанных процентов, нет необходимости в преобразовании этих процентов, и можно обрабатывать непосредственные данные. Я здесь подчеркиваю тот факт, что численности субклассов для пород репрезентируют соотношение их в совокупности. Представление об этой совокупности было бы искажено, если бы были взяты одинаковые численности. Здесь введено только одно видоизменение первоначальных данных, направленное на то, чтобы сделать численности пропорциональными в двух направлениях:

$$12:6 = 30:15 = 4:2 \text{ и т. д. и } 12:30:4:6:10 = 6:15:2:3:5.$$

Хотя наличие пропорциональности между численностями субклассов не нарушает самой основы дисперсионного анализа, однако вычисления должны быть несколько видоизменены, так как здесь мы имеем дело с выборками неодинакового размера (параграф 16 главы 10). Соответствующий метод вычислений указан в таблице. Так как эти данные взяты только для иллюстрации и представляют лишь небольшую часть всего первоначального материала, наши выводы будут только примерные.

При неодинаковых численностях в субклассах компонентный анализ усложняется, как это было в параграфах 16 и 17 главы 10. В этом случае для каждого компонента, входящего в тот или иной эффект, должен быть вычислен специальный, отличный от других, коэффициент [24].

Уилк и Кемпторн [30] вывели общие формулы, приложимые ко всем моделям трехфакторного опыта; число членов субклассов может быть в этом случае как равным, так и пропорциональным. Я возьму здесь только части проблемы, относящиеся к двухфакторному опыту. Пусть имеется пропорция для фактора $A: u_1:u_2:\dots:u_a$ и для фактора $B: v_1:v_2:\dots:v_b$. Далее пусть:

$$U = \Sigma u, \quad V = \Sigma v, \quad U^* = \frac{\Sigma u^2}{(\Sigma u)^2}, \quad V^* = \frac{\Sigma v^2}{(\Sigma v)^2}.$$

Тогда компоненты в совокупности будут:

$$A: \sigma^2 + \frac{nUV(1-U^*)}{a-1} \left\{ \frac{1}{U^*} - \frac{1}{U} \right\} \sigma_{AB}^2 + \sigma_A^2;$$

$$B: \sigma^2 + \frac{nUV(1-V^*)}{b-1} \left\{ \left(U^* - \frac{1}{A} \right) \sigma_{AB}^2 + \sigma_B^2 \right\}$$

$$AB: \sigma^2 + \frac{nUV(1-U^*)(1-V^*)}{(a-1)(b-1)} \sigma_{AB}^2.$$

Здесь, как обычно, n — число повторений в полностью рандомизированной схеме опыта; оно также является числом блоков при схеме рандомизированных блоков. Величины же A и B имеют следующие значения:

в модели I: $A = a, B = b$;

в модели II: $\frac{1}{A} = 0, \frac{1}{B} = 0$;

Процент убойного веса (минус 70%) 93 свиней, сгруппированных по породам
и по полу. Живой вес 200—219 фунтов

Номер	Порода									
	1		2		3		4		5	
	хряк	матка	хряк	матка	хряк	матка	хряк	матка	хряк	матка
1	13,3	18,2	10,9	14,3	13,6	12,9	11,6	13,8	10,3	12,8
2	12,6	11,3	3,3	15,3	13,1	14,4	13,2	14,4	10,3	8,4
3	11,5	14,2	10,5	11,8	4,1		12,6	4,9	10,1	10,6
4	15,4	15,9	11,6	11,0	10,8		15,2		6,9	13,9
5	12,7	12,9	15,4	10,9			14,7		13,2	10,0
6	15,7	15,1	14,4	10,5			12,4		11,0	
7	13,2		11,6	12,9					12,2	
8	15,0		14,4	12,5					13,3	
9	14,3		7,5	13,0					12,9	
10	16,5		10,8	7,6					9,9	
11	15,0		10,5	12,9						
12	13,7		14,5	12,4						
13			10,9	12,8						
14			13,0	10,9						
15			15,9	13,9						
16			12,8							
17			14,0							
18			11,1							
19			12,1							
20			14,7							
21			12,7							
22			13,1							
23			10,4							
24			11,9							
25			10,7							
26			14,4							
27			11,3							
28			13,0							
29			12,7							
30			12,6							
ΣX	168,9	87,6	362,7	182,7	41,6	27,3	79,7	33,1	110,1	55,7

Общее: $N=93$; $\Sigma X=1149,4$; $\Sigma X^2=14785,62$.

Суммы по породам: 1) 256,5; 2) 545,4; 3) 68,9; 4) 112,8; 5) 165,8.

Сумма по полам: хряки—763,0; матки—386,4.

1. Поправки: $C=(\Sigma X)^2/n=1149,4^2/93=14205,60$.

2. Общее: $\Sigma X^2-C=14785,62-14205,60=580,02$.

3. Субклассы: $\frac{168,9^2}{12} + \frac{87,6^2}{6} + \dots + \frac{55,7^2}{5} - C = 122,83$.

4. Внутри субклассов: $580,02 - 122,83 = 457,19$.

5. Пол: $\frac{763,0^2}{62} + \frac{386,4^2}{31} - C = 0,52$.

6. Порода: $\frac{256,5^2}{18} + \dots + \frac{165,8^2}{15} - C = 97,38$.

7. Взаимодействие: $122,83 - (97,38 + 0,52) = 24,93$.

Пол	1	0,52	0,52
Порода	4	97,38	24,34**
Взаимодействие пол—порода	4	24,93	6,23
Внутри субклассов	83	457,19	5,51

если A фиксировано, а B случайно: $A = a, \frac{1}{B} = 0$;

если A случайно, а B фиксировано: $\frac{1}{A} = 0, B = b$.

Так, для наших данных о свиньях, если через A обозначить пол, а через B — породу, имеем:

$$n = 1; a = 2; b = 5; u_1 = 2; u_2 = 1; v_1 = 6; v_2 = 15; v_3 = 2; v_4 = 3;$$

$$v_5 = 5; U = 2 + 1 = 3; V = 6 + \dots + 5 = 31;$$

$$U^* = \frac{2^2 + 1^2}{3^2} = 0,556; V^* = \frac{6^2 + \dots + 5^2}{31^2} = 0,311.$$

Наконец, так как пол и порода содержат фиксированные параметры, то $A = a = 2$ и $B = b = 5$. Подстановка дает компоненты:

$$A: \sigma^2 + 4,58\kappa_{AB}^2 + 41,3\kappa_A^2;$$

$$B: \sigma^2 + 0,90\kappa_{AB}^2 + 16,0\kappa_B^2;$$

$$AB: \sigma^2 + 7,11\kappa_{AB}^2.$$

По данным о свиньях, оценки этих компонентов таковы:

$$s^2 = 5,51;$$

$$s_{AB}^2 = \frac{6,23 - 5,51}{7,11} = 0,101;$$

$$s_B^2 = \frac{24,34 - 5,51 - 0,90 \times 0,101}{16,0} = 1,17.$$

Для s_A^2 наша выборка не дает ничего другого, кроме нуля.

Определение существенности этих компонентов не представляет трудностей, если использовать приближение Саттертуэйта — Кокрана, параграф 10 этой главы. Для приведения к сравнимому виду среднего квадрата породы начнем со среднего квадрата взаимодействия порода — пол:

$$6,23 = s^2 + 7,11s_{AB}^2.$$

Умножаем на отношение коэффициентов при s_{AB}^2 в B и AB , т. е. на $0,90/7,11$:

$$\frac{0,90}{7,11} \times 6,23 = \frac{0,90}{7,11} s^2 + 0,90s_{AB}^2.$$

или

$$0,79 = 0,13s^2 + 0,90s_{AB}^2.$$

Теперь прибавляем:

$$5,51 = s^2$$

Средний квадрат, $6,30 = 1,13s^2 + 0,90s_{AB}^2$
являющийся знаменателем в F

Для числителя же F начнем с величины

$$\frac{0,90}{7,11} \times 5,51 = \frac{0,90}{7,11} s^2,$$

или

$$0,70 = 0,13s^2.$$

Средний квадрат породы $24,34 = s^2 + 0,90s_{AB}^2 + 16,0s_B^2$
Сумма $25,04 = 1,13s^2 + 0,90s_{AB}^2 + 16,0s_B^2$

Отсюда:

$$F' = \frac{1,13s^2 + 0,90s_{AB}^2 + 16s_B^2}{1,13s^2 + 0,90s_{AB}^2} = \frac{25,04}{6,30} = 3,97.$$

Так как оба члена F' являются комбинированными суммами средних квадратов, то соответствующие степени свободы должны определяться следующим образом:

$$n'_1 = \frac{25,04^2}{\frac{0,10^2}{83} + \frac{24,34^2}{4}} = 4,2;$$

$$n'_2 = \frac{6,30^2}{\frac{0,79^2}{4} + \frac{5,51^2}{83}} = 76.$$

Для $f = 4$ и 76 критерий $F_{0,01} = 3,58$, с чем и следует сравнивать выборочное значение $F' = 3,97$. Несомненно в опыте установлены различия в проценте убойного веса по породам. Соответствующие средние проценты: 84,2, 82,1, 81,5, 82,5 и 81,4%.

В тех случаях, когда численности групп носят случайный характер, т. е. не репрезентируют реальных различий в размере групп основной совокупности, вполне допустимо произвести анализ на основе таблицы с двумя входами, составленной из невзвешенных средних. Этот метод изложен в параграфе 16 этой главы и применен к данным о свиньях в примере 31 этой главы.

14. Непропорциональные численности субклассов. Таблица 2×2 . Как равенство, так и пропорциональность численностей субклассов далеко не всегда достижимы. Например, нельзя регулировать пропорцию полов у цыплят, вылупившихся из различно обработанных групп яиц. Последствия этого могут быть довольно серьезными, если изучаемое явление, например скорость роста, идет по-разному у разных полов. Группа цыплят, в которой доминируют петушки, будет, вероятно, в целом расти быстрее, чем группа, в которой большинство курочек, хотя бы в этой второй группе вариант опыта был более благоприятным для скорости роста. Примерно такое положение можно заметить и в нашей таблице 162 [26]. Взвешенная средняя из двух групп цыплят, получавших гормон B , оказалась меньше, а невзвешенная средняя (без учета численностей субклассов) больше, чем соответствующие средние у цыплят, не получавших гормон B . Как мы увидим, ни та, ни другая пара наблюдений не дает несмещенную оценку эффекта B .

Другой характерной особенностью таблицы с двумя входами при непропорциональности численностей субклассов является полная неприменимость теоремы о слагаемости суммы квадратов. Это не очень бросается в глаза в таблице 162, но становится очевидным при анализе данных о крысах в таблице 164, где две суммы квадратов, относящиеся к полу и поколению, будучи сложены, оказываются больше общей суммы квадратов. Ясно, что такая таблица не способна дать оценку взаимодействия факторов. В этом случае все оценки и критерии существенности будут в связи с непропорциональностью численностей субклассов неправильными, вследствие чего становится необходимым применение соответствующих усложненных статистических методов [32]. Случаи выпадения данных в рендомизированных блоках и в латинском квадрате являются частными случаями более общей задачи, к рассмотрению которой мы теперь и переходим.

При решении задач, рассматриваемых в настоящем и следующих параграфах, необходим предварительный дисперсионный анализ без учета непропорциональности численностей субклассов. Этот анализ проводится по образцу анализа с пропорциональными численностями, описание которого дано в предшествующем параграфе. Теперь мы рассмотрим, какими должны быть несмещенные оценки и критерии существенности для определения главных эффектов (т. е. средних по строкам и столбцам) и взаимодействий. Но следует заметить, что средние квадраты ни между субклассами, ни внутри субклассов в данном случае не испытывают никакого смещения. Например, суммы квадратов между субклассами для таблицы 161:

$$\frac{240^2}{3} + \frac{1440^2}{12} + \frac{1200^2}{12} + \frac{672^2}{6} - \frac{3552^2}{33} = 4940$$

Вес гребешка (в мг) у отдельных групп цыплят, получавших отдельно и в комбинации два вида гормона

	Без гормона			Гормон В		
	численность	ΣX	\bar{x}	численность	ΣX	\bar{x}
Без гормона	3	240	80	12	1200	100
Гормон А	12	1440	120	6	672	112
	15	1680		18	1872	

Взвешенная средняя $1680/15 = 112$
 Невзвешенная средняя $(80+120)/2 = 100$

Предварительный дисперсионный анализ без учета непропорциональности численностей субклассов

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Гормон А	1	3 724	
Гормон В	1	524	
Внутри групп	29	23 519	811

и внутри этих субклассов [23519] являются в соответствии с параграфом 16 главы 10 вполне правильными оценками.

Вычисления таблицы 163 требуют небольших пояснений. Веса W определяются по численностям цыплят n , что приводит к взвешенной сумме квадратов [1216] средних разностей. Поправочный член дает средний квадрат взаимодействия, который несколько больше, чем выборочное его значение. Следовательно, это взаимодействие в совокупности здесь считается несуществующим.

ТАБЛИЦА 163

Исключение взаимодействия в таблице 2×2 при непропорциональности численностей субклассов.

Продолжение расчетов таблицы 162. Данные о влиянии гормонов на цыплят

	Без гормона		Гормон В		$\frac{n_1 n_2}{n_1 + n_2} = W$	$\bar{x}_2 - \bar{x}_1 = D$	WD	WD ²
	n_1	\bar{x}_1	n_2	\bar{x}_2				
Без гормона	3	80	12	100	2,4	20	48	960
Гормон А	12	120	6	112	4,0	-8	-32	256
					6,4		16	1216

Сумма квадратов для взаимодействия $= \Sigma WD^2 - (\Sigma WD)^2 / \Sigma W = 1216 - 16^2 / 6,4 = 1216 - 40 = 1176$;

$F = 1176/811 = 1,45$, $f = 1$ и 29, несущественно

Погрешка на непропорциональность $= (\Sigma WL)^2 / \Sigma W = 524 - 40 = 484$ (см. анализ таблицы 162)

Полный дисперсионный анализ

Источник варьирования	Число степеней свободы	Средний квадрат
Гормон <i>A</i>	1	(3 724 — 484) = 3 240
Гормон <i>B</i>	1	40
Взаимодействие	1	1 176
Внутри субклассов	29	811

Для *A*: $F = 3\,240/811 = 4,00$; для *B*: $F = 40/811 = 0,05$

Это приводит к тому, что поправочный член $(\Sigma WD)^2/\Sigma W = 40$ является несмещенной оценкой среднего квадрата между средними по столбцам (гормон *B*), поэтому приближенное значение этого среднего квадрата в таблице 162 должно быть уменьшено на поправку на непропорциональность $= 524 - 40 = 484$. Введение той же самой поправки и для гормона *A* дает полный анализ, приведенный в конце таблицы 163.

Теперь установлена надлежащая несмещенная оценка средней разности, относящейся к гормону *B*, действующему при наличии и при отсутствии гормона *A*; эта взвешенная средняя разность равна:

$$\Sigma WD/\Sigma W = 16/6,4 = 2,5 \text{ мг.}$$

Отметим, что здесь нет различия как между этими взвешенными, так и невзвешенными средними таблицы 162. Подобная оценка разности, связанной с гормоном *A*, может быть рассчитана по колонкам таблицы. В этом случае веса оказываются теми же самыми, как и по рядам, и поэтому

$$\Sigma WD/\Sigma W = (2,4 \times 40 + 4,0 \times 12)/6,4 = 22,5 \text{ мг.}$$

Существенность этих взвешенных разностей устанавливается в таблице при помощи критерия *F*.

Если взаимодействие окажется явным, то обычно нет оснований для проведения этого анализа, так как все равно главные эффекты будут иметь небольшой смысл или даже не будут иметь никакого смысла. В этом случае установление существования взаимодействия и его интерпретация обычно являются окончанием самого исследования. Если же требуется дальнейшее сравнение групп, то эффекты гормона *A* и гормона *B*, взятых по отдельности, могут быть исследованы в соответствии с параграфом 5 главы 4 или параграфом 16 главы 10.

Пример 27. При изучении механизма спаривания у двух видов дрозофилы [18] исследователи оставили на несколько дней мужских особей *D. persimilis* с женскими особями того же самого вида при благоприятных для спаривания условиях, после этого был дан доступ этим мужским особям к женским как того вида, так и к женским особям *D. pseudoobscura*. Затем при неблагоприятных условиях были оставлены мужские особи *D. persimilis* с женскими *D. pseudoobscura* с последующим переходом, как и ранее, к свободному выбору. Число оплодотворенных женских особей было таким:

	Благоприятные условия		Неблагоприятные условия	
	оплодотворение	нет	оплодотворение	нет
<i>D. persimilis</i>	18	14	41	6
<i>D. pseudoobscura</i>	5	32	20	32

Вспоминая параграф 10 главы 9, преобразуйте проценты оплодотворенных в углы (табл. 129). Численности и углы будут следующие:

32	48,59	47	69,06
37	21,57	52	38,33

Дисперсионный анализ (даны средние квадраты) будет таким:

условия	1	13 910
виды	1	35 701
взаимодействие	1	139
ошибка (см. пример б)		821

Как вы интерпретируете эти результаты?

15. Непропорциональность численностей субклассов. Таблица $R \times 2$. Этот вид таблицы встречается довольно часто, обычно когда берутся два пола. Данные таблицы 164 прежде всего останавливают наше внимание некоторой особенностью их, вытекающей из непропорциональности численностей субклассов [6]. Эта особенность состоит в том, что в предварительном дисперсионном анализе две суммы квадратов для полов и поколений при сложении дают величину большую, чем общая сумма квадратов для субклассов, что подчеркивает неортогональность, возникающую в связи с изменчивостью численностей по клеткам таблицы.

Здесь взаимодействие хотя и велико, все же находится ниже уровня существенности. Вместе с этим здесь нет оснований ожидать, что различия полов в отношении скорости роста будут меняться по поколениям, и поэтому можно допустить, что взаимодействие отсутствует.

В остальном вычисления ведутся так же, как и в таблице 163.

ТАБЛИЦА 164

Исключения взаимодействия в таблице $R \times 2$ с непропорциональными численностями субклассов. Численности и средние привесы (в г) 149 крыс, изученных в течение 1928—1929 гг. Четыре последовательных поколения.

Привесы взяты за 6-недельный период, начиная с 28-дневного возраста.

Все данные уменьшены на 100 г

Поколения	Самцы		Самки		$\frac{n_1 n_2}{n_1 + n_2} = W$	$\bar{x}_1 - \bar{x}_2 = D$	WD
	n_1	\bar{x}_1	n_2	\bar{x}_2			
1	21	76,952	27	9,518	11,8125	67,434	796,564
2	15	61,467	25	14,080	9,3750	47,387	444,253
3	12	55,667	23	8,522	7,8857	47,145	371,771
4	7	71,000	19	6,790	5,1154	64,210	328,460
					34,1886		1941,048

Предварительный дисперсионный анализ по исходным данным

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Средние субклассов	7	119 141	
Пол	1	114 287	
Поколения	3	5 756	
Внутри субклассов	141		409

Сумма квадратов для взаимодействия: $\Sigma WD^2 - (\Sigma WD)^2 / \Sigma W = 113 385 - 110 202 = 3183$. Поправки на непропорциональность = $114 287 - 110 202 = 4085$

Полный дисперсионный анализ

Пол	1		110 202	110 202
Поколения	3	(5756 — 4085) =	1 671	557
Взаимодействие	3		3 183	1 061
Внутри субклассов	141			409

Этот метод применим и в тех случаях, когда полностью выпадают данные всего субкласса. В этих случаях при заполнении пустующих мест для n и \bar{x} нулями все вычисления сохраняются; из степеней свободы взаимодействия вычитается 1 за каждую незаполненную клетку таблицы.

Несмещенная оценка полового различия в отношении среднего прироста в весе у нас будет:

$$\Sigma WD / \Sigma W = 1941,048 / 34,1886 = 56,77 \text{ г.}$$

Оценка существенности этой разности производится при помощи критерия $F = 110\ 201/409$.

Несмещенные оценки средних разностей по поколениям будут определены специальным методом подыскания констант, о чем будет речь в параграфе 17 этой главы. Очень хорошее приближение может быть получено и при помощи метода, указанного в таблице 163. В качестве примера в таблице 165 произведено вычисление взвешенной средней разности между поколениями 1 и 2 и произведена оценка существенности ее.

ТАБЛИЦА 165

Метод приближенного вычисления взвешенной средней разности между поколениями 1 и 2. Точный метод дан в параграфе 17 главы 12

	Поколения 1		Поколения 2		D	W	WD
	n_1	\bar{x}_1	n_2	\bar{x}_2			
Самцы	21	76,952	15	61,487	8,75	15,485	135,49
Самки	27	9,518	25	14,080	12,98	-4,562	-59,21
					21,73		76,28

Взвешенная средняя разность $\Sigma WD / \Sigma W = 3,51$ г.

Средний квадрат средней разности $= 409 / 21,73 = 18,82$; $s_d = 4,34$; $t = 3,51 / 4,34 = 0,81$; $f = 141$.

Пример 28. Беккер и Холл [3] определили численности ооцистов, продуцированных крысами 5 рас в период иммунизации их *Eimeria miyairii*. Единицей измерения было 10^6 ооцистов.

Пол		Расы				
		Ламберт	Lo	Hi	W. E. L	Уистар (A)
Самцы	$\frac{n}{x}$	8	14	20	8	9
		36,1	94,9	194,4	64,1	175,7
Самки	$\frac{n}{x}$	7	14	21	10	8
		31,9	68,6	187,3	89,2	148,4

Проверить полный дисперсионный анализ, указанный в первоисточнике

Пол	1	2594,6	2594,6
Раса	4	417565,6	104391,4
Взаимодействие	4	8805,3	2201,3
Ошибка	109	332962,9	3054,7

Вы не получите точное совпадение своих результатов с этими данными потому, что приведенные здесь средние даны только с тремя значащими цифрами. Ваши результаты дадут примерно три первые цифры у средних квадратов, что вполне достаточно для оценки существенности.

Пример 29. Многие исследователи изучали влияние способа искусственного осеменения молочных коров (в шейку матки и в матку). По данным трех отчетов я составил следующую таблицу сопряженности трех признаков (параграф 10 главы 9):

Исследователь	Осеменение в шейку матки		Осеменение в матку	
	оплодотворены	нет	оплодотворены	нет
1	64	39	97	53
2	123	70	121	72
3	292	208	324	176

Вычислите проценты случаев оплодотворения, преобразуйте их в углы и получите такой дисперсионный анализ (даны средние квадраты)

способ осеменения	1	2357
исследователь	2	174,5
взаимодействие	2	688
ошибка	∞	821

Обычно в пользу способа осеменения в шейку матки приводится тот аргумент, что при этом способе меньше затрачивается времени и меньше шанс внесения инфекции. Какое решение примете вы относительно способа осеменения?

16. Непрорпциональность численностей субклассов в таблице $R \times C$.

Приближенные методы. Общий метод подбора констант, излагаемый в следующем параграфе, довольно трудоемок. Приближенный метод невзвешенных средних, к описанию которого мы переходим, имеет целый ряд преимуществ, включая сюда и знаковую форму расчетов. Если наличие взаимодействия очевидно, то в этом случае нет никакой необходимости производить какие-либо дальнейшие вычисления. Если же взаимодействием можно с полным основанием пренебречь, то наличие или отсутствие главных эффектов может быть выражено в такой явной форме, что отпадает всякая необходимость в точном методе следующего параграфа. Приближенный метод сводится к использованию заранее планируемых сравнений (параграф 2 этой главы) или всех возможных сравнений по примеру параграфа 6 главы 10. В тех случаях, когда численности субклассов только слегка отклоняются от полного равенства, в особенности когда они достаточно велики (скажем, 10 и более), то приближение предлагаемого метода к методу подбора констант будет весьма близким.

Данный метод мы иллюстрируем примером, в котором диспропорция численностей субклассов довольно значительна. Эти же данные будут использованы и в следующем параграфе, что позволит произвести сравнение обоих методов.

Три расы мышей были заражены тремя штаммами тифозной бациллы [11]. Для каждой мыши был определен срок от момента заражения до гибели (в днях). Первоначальные данные, которые здесь не приводятся, нужны только для определения ошибки, т. е. варьирования данных о мышах внутри каждого варианта. Эта величина была вычислена, в результате чего получено 5,015 при $f = 774$. В таблице 166 приведены средние по субклассам и соответствующие численности. Последние при построении дисперсионного анализа не принимались в расчет, а использовались только средние, по которым велась простая обработка таблицы с двумя входами, подобная обработке таблицы 112.

Средние квадраты средних для приведения их к сравнимому с ошибкой виду должны умножаться на численность мышей в субклассе. Но так как эти последние различны, то берется их средняя гармоническая. Она является обратной величиной средней из обратных значений численности по субклассам; ее вычисление дано в конце таблицы. Произведение на средний квадрат взаимодействия записано в последнем столбце (в качестве альтернативы можно 5,015 разделить на n_0 и тем самым привести ошибку к сравнимому виду со средним квадратом средней; в результате получится 0,0841).

Численности и средние числа дней до гибели по субклассам трех рас мышей, зараженных тремя штаммами тифозной бациллы

Штамм	Расы мышей			Сумма
	RI	Z	Ba	
9D	34 4,0000	31 4,0323	33 3,7576	11,7899
11C	66 6,4545	78 6,7821	113 4,3097	17,5463
DSC1	107 6,6262	133 7,8045	188 4,1277	18,5584
Сумма . . .	17,0807	18,6189	12,1950	47,8946

Дисперсионный анализ невзвешенных средних

На единичное наблюдение

Штамм	2	8,8859		
Расы	2	7,5003		
Взаимодействие	4	3,2014	0,8004**	47,71**
Внутри групп	774		0,0841	5,015

$$\frac{1}{n_0} = \frac{1}{9} \left(\frac{1}{34} + \dots + \frac{1}{188} \right) = 0,0167754; n_0 = 59,6111$$

Как это следует из внимательного рассмотрения данных, взаимодействие факторов здесь имеет место. Расы RI и Z были стойкими против 11C и DSC1, но раса Ba оказалась восприимчивой ко всем штаммам тифозной бациллы. Поэтому представляется невозможным какое-либо сравнение главных эффектов. Правда, некоторые сравнения между 4 крупными средними и 5 малыми средними могли бы быть сделаны, но это будут в некоторой степени искаженные сравнения, потому что они выбраны на основе предварительной оценки и после рассмотрения конкретных результатов. Однако я все же взял 4 наибольших средних, построил таблицу 2×2, обработал по способу таблицы 162 и нашел, что взаимодействие не достигает 5%-ного уровня существенности.

Помимо метода невзвешенных средних, имеется другой метод, особенно полезный, когда есть основание считать, что численности выборки репрезентируют численности в совокупности, из которой взята выборка. Пользуясь крайевыми итогами, можно вычислить численности внутренних клеток, которые теперь будут пропорциональными, благодаря чему становится применимым метод *пропорциональных численностей субклассов*, о котором говорилось в параграфе 13 этой главы. Этот случай нам встретился, когда в параграфе 8 главы 9 рассматривались данные обследования форм землевладения на фермах [29]. Было установлено, что численности субклассов должны располагаться по образцу крайевых итогов независимо от почвы и владения. В данном случае представляется оправданным взять эти пропорциональные численности вместо фактических непропорциональных.

Таблица 167 содержит средние данные о площадях посевов кукурузы, вычисленные по данным этого обследования. Умножение средней каждой клетки на пропорциональные численности субклассов дает для каждой клетки новые значения ΣX. Эти-то новые величины ΣX совместно с пропорциональными численностями и кладутся в основу дисперсионного анализа.

*Площади посевов кукурузы, сгруппированные по виду землевладения
и по плодородию почвы*

Классы почвы		Собственность		Аренда		Смешанное		Σn	ΣX
		фак- тиче- ски	про- пор- цио- наль- но	фак- тиче- ски	про- пор- цио- наль- но	фак- тиче- ски	про- пор- цио- наль- но		
I	$\frac{n}{x}$	36	36,75	67	62,92	49	52,33	152	7323
	ΣX	32,7	1202	55,2	3473	50,6	2684		
II	$\frac{n}{x}$	31	33,85	60	57,95	49	48,20	140	6584
	ΣX	36,0	1219	53,4	3095	47,1	2270		
III	$\frac{n}{x}$	58	54,40	87	93,13	80	77,47	225	9102
	ΣX	30,1	1637	49,8	4358	40,1	3107		
Σn		125		214		178		517	
ΣX			4058		10 926		8025		23 009

Дисперсионный анализ на основе пропорциональных численностей

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Почвы	2	6 635	3 318*
Виды землевладения	2	27 337	13 684**
Взаимодействие	4	883	221
Ошибка (по исходным данным)	507		830

Преимуществом этого анализа является то, что он построен на принципе ортогональности и что он дает компоненты, которые могут быть оценены по методу параграфа 13 этой главы.

Здесь не проявляется никакого взаимодействия факторов в совокупности, но средние по видам землевладения, несомненно, различны. Если представляется необходимым применить к почвенным различиям более строгий критерий, то надо обратиться к методу следующего параграфа.

Пример 30. Примените к данным обследования таблицы 167 метод невзвешенных средних. Вы в этом случае получите:

почвы	2	2 436*
виды землевладения	2	14 490**
взаимодействие	4	255
ошибка	507	830

Выводы не изменились, хотя теперь эффект почвенных различий менее выражен (см. пример 32).

Пример 31. В таблице 161 размеры выборок по породам предположительно пропорциональны распространению пород в районе, но соотношение полов определяется условиями рынка и не нашло своего отражения в таблице. Если представленные там данные рассматривать как характеристику местной совокупности свиней, сгруппированной по породам и полу, то могут быть достаточные основания для анализа невзвешенных средних для каждой клетки таблицы 2×5:

пол	1	1,3
порода	4	15,5*
взаимодействие	4	10,1
ошибка	83	5,51

17. Подбор констант для двухфакторной таблицы. Этот метод является самым общим при анализе двухфакторного опыта; большинство методов, рассмотренных ранее, включая сюда и случай выпадения наблюдений, представляют собой частные случаи этого метода. Подбор констант приводит к несмещенным оценкам и позволяет оценить взаимодействие, если последнее имеет место. Когда же взаимодействие отсутствует, то этот метод даст несмещенные оценки и критерии существенности для главных эффектов.

Для сравнения этого метода с методом невзвешенных средних произведем подбор констант к ранее рассмотренным данным о мышах, которые с дополнением их фактическими значениями ΣX представлены в таблице 166. Искомые константы являются значениями α и β в уравнении:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

ТАБЛИЦА 168

Метод подбора констант. Данные о числе дней с момента заражения до гибели для трех рас мышей, зараженных тремя штаммами тифозной бактерии

Виды тифа		Расы мышей			Σn_i	ΣX	\bar{x}_w	$b+x$
		<i>h1</i>	<i>Z</i>	<i>Ва</i>				
9D	n $n/\Sigma n_i$	34 0,34694	31 0,31633	33 0,33373	98	385	3,9286	3,9490
11C	n $n/\Sigma n_i$	66 0,25681	78 0,30350	113 0,43969	257	1442	5,6109	5,8600
DSC1	n $n/\Sigma n_i$	107 0,25000	133 0,31075	188 0,43925	428	2523	5,8949	6,1375
Σn_i ΣX		207 1271	242 1692	334 1387	783	4350		
a		0,4506	1,2240	-1,6746				$\bar{x}=5,3155$

Предварительный дисперсионный анализ

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Субклассы	8	1785,58	1536,66
Штамм	2	309,47	
Расы мышей	2	1227,19	
Внутри группы	774		5,015

$a_1: (34 \times 0,34694 + \dots + 107 \times 0,25000 - 207) a_1 + (34 \times 0,31633 + \dots + 107 \times 0,31075) a_2 + (34 \times 0,33673 + \dots + 107 \times 0,43925) a_3 = 34 \times 3,9286 + \dots + 107 \times 5,8949 - 1271$ и так далее вплоть до последнего a ;

$a_3: (33 \times 0,34694 + \dots + 188 \times 0,25000) a_1 + (33 \times 0,31633 + \dots + 188 \times 0,31075) a_2 + (33 \times 0,33673 + \dots + 188 \times 0,43925 - 334) a_3 = 33 \times 3,9286 + \dots + 188 \times 5,8949 - 1387$.

где α и β являются градиентами двух факторов *A* и *B*. Их оценки a_i и b_j , причем $\Sigma a_i = \Sigma b_j = 0$.

В каждую клетку таблицы вносится отношение n к Σn_i ; сумма этих отношений по каждой строке должна равняться единице. Для каждой строки определяется взвешенная средняя $\bar{x}_j = \Sigma X / \Sigma n_i$. Если бы наша таблица была вместо квадратной (3×3) прямоугольной с *R* рядами с *C* столбцами, то ее

следовало бы расположить так, чтобы R было большим, а C — меньшим (меньшее число столбцов упрощает вычисления).

Вычисления проводятся по строгой системе. В соответствии с каждым из a столбцов составляется a уравнений по правилу, указанному в конце таблицы 168. Для определения уравнения для a_1 все время используются численности столбца 1 (34, 66 и 107), умножаемые последовательно на отношения $(n/\Sigma n_i)$ первого, второго и третьего столбцов и на столбец средних; последнее произведение является правой частью соответствующего уравнения. Для определения уравнения для a_2 тем же порядком используются численности второго столбца совместно со всеми столбцами отношений и средних. Изменение порядка расчета состоит еще только в том, что при составлении каждого уравнения производится вычитание различных сумм Σn_j .

Эти коэффициенты при a_i легко вычисляются на счетной машине, после этого составляется следующая система линейных уравнений:

$$\begin{aligned} -151,505a_1 + 64,036a_2 + 87,468a_3 &= -136,35; \\ 64,036a_1 - 167,191a_2 + 103,155a_3 &= -348,54; \\ 87,468a_1 + 103,155a_2 - 190,624a_3 &= 484,92. \end{aligned}$$

В качестве контроля за вычислениями может служить то обстоятельство, что коэффициенты располагаются симметрично относительно главной диагонали и что сумма коэффициентов каждого столбца и свободных членов в правой части уравнений равна нулю (в пределах ошибки округления).

Так как $\Sigma a = 0$, то первый этап решения системы уравнений сводится к подстановке в каждое уравнение $a_3 = -a_1 - a_2$. Это производится весьма просто: 87,468 вычитается из каждого предыдущего коэффициента первого уравнения, а 103,155 из каждого коэффициента второго уравнения. В результате получается:

$$\begin{aligned} -238,973a_1 - 23,432a_2 &= -136,35; \\ -39,119a_1 - 270,346a_2 &= -348,54. \end{aligned}$$

Третье уравнение будет просто-напросто являться суммой этих двух уравнений и может быть отброшено. Решение двух уравнений дает:

$$a_1 = 0,4506,$$

$$a_2 = 1,2240,$$

откуда $a_3 = -1,6746.$

Проверка производится подстановкой этих значений в одно или несколько исходных уравнений. Полученные значения a_i рекомендуется ввести в последнюю строку таблицы.

Далее для каждого ряда вычисляются величины $b + \bar{x}$, для чего используются \bar{x}_j и a_i совместно с отношениями $n/\Sigma n_i$ данного ряда. Например, для 1-го ряда:

$$b_1 + \bar{x} = 3,9286 - 0,34694 \times 0,4506 - \dots - 0,33673 \times (-1,6746) = 3,9490.$$

Подобно этому $b_2 + \bar{x} = 5,8600$

$$b_3 + \bar{x} = 6,1375$$

Складываем: $3x = 15,9465$ (так как $\Sigma b = 0$).

Отсюда $\bar{x} = 5,3155,$

и, наконец, $b_1 = -1,3665, b_2 = 0,5445, b_3 = 0,8220.$

Теперь для вычисления «части суммы квадратов, обусловленной подобранными константами» следует взять эти константы и исходные суммы

ΣX , приведенные в окаймляющих графах таблицы; расчеты ведутся так: $0,4506 \times 1271 + 1,2240 \times 1692 + \dots + 6,1375 \times 2523 - 4350^2/783 = 1609,78$.

Вычитание этого результата из суммы квадратов для субклассов приводит к сумме квадратов взаимодействия:

$$1785,58 - 1609,78 = 175,80.$$

Так как для этого взаимодействия у нас имеется 4 степени, то имеем: средний квадрат для взаимодействия = $175,80/4 = 43,95$;

$$F = 43,95/5,015 = 8,76; f = 4 \text{ и } 774; F_{0,01} = 3,35.$$

Как и прежде, здесь определенно установлено наличие взаимодействия, и поэтому отпадает какая-либо необходимость в дальнейшем анализе; данные должны рассматриваться в их развернутом виде, как это было сделано в предыдущем параграфе. Но для иллюстрации способа дальнейших расчетов мы продолжим обработку данных так, как будто взаимодействия здесь не обнаружено.

В таблице 169 произведено вычисление оценок для средних по субклассам и средних по вариантам $\bar{x} + a_i + b_j$. Если взаимодействие в совокупности

ТАБЛИЦА 169

Оценки, полученные при подборе констант. Взаимодействие считается отсутствующим. Данные о мышах из таблицы 168

Штамм	Константы	R1	Z	Va	Средняя
		$a_1 = 0,4506$	$a_2 = 1,2240$	$a_3 = -1,6746$	
9D	$b_1 + \bar{x} = 3,9490$	4,3996	5,1730	2,2744	3,9490
11C	$b_2 + \bar{x} = 5,8600$	6,3106	7,0840	4,1854	5,8600
DSC1	$b_3 + \bar{x} = 6,1375$	6,5881	7,3615	4,4629	6,1375
Средняя		5,7661	6,5395	3,6409	

Полный дисперсионный анализ

Штамм	2	309,47 - (-73,12) ¹	= 382,59	
Расы мышей	2	1227,19 - (-73,12)	= 1300,31	
Взаимодействия	4		175,80	43,95
Внутри групп	774			5,015

¹ Поправка = (объединенная сумма квадратов для штаммов тифозной бациллы и рас мышей) - (сумма квадратов, относящаяся к константам) = $1536,66 - 1609,78 = -73,12$.

равно нулю, то эти показатели будут несмещенными оценками соответствующих параметров совокупности. В данном случае ясно, что при подборе констант не может быть и речи о взаимодействии между средними по субклассам (взаимодействия отсутствуют в приведенной выше модели). Следовательно, наши средние не могут репрезентировать средние совокупности, из которой взята выборка и в которой имеет место взаимодействие факторов.

Метод подбора констант не требует обязательного наличия данных в каждой клетке таблицы. Если некоторые клетки окажутся пустыми, то подстановка вместо n и $n/\Sigma n_i$ нулей не меняет дальнейших вычислений; только в дисперсионном анализе свободы для субклассов и взаимодействия следует уменьшить на столько единиц, сколько было выпавших данных.

15. Homeyer Paul G., Black C. A., Soil Science Society of America Proceedings, 11, 341, 1946.
16. Iowa Agricultural Experiment Station, Animal Husbandry Swine Nutrition Experiment No. 577, 1952.
17. Kempthorne Oscar, The Design and Analysis of Experiments. John Wiley and Sons, Inc., New York, 1952.
18. Mayr Ernst, Dobzhansky Th., Proceedings of the National Academy of Science, 31, 75, 1945.
19. Porter R. H., Cooperative Soybean Seed Treatment Trials, Iowa State College Seed Laboratory, 1943.
20. Richardson D., et al. См. ссылку 26, гл. 10.
21. Satterthwaite F. E., Biometrics (Bulletin), 2, 110, 1946.
22. Saunders A. R., Union of South Africa Department of Agriculture and Forestry Science Bulletin No. 200, 1939.
23. Schultz E. F., Jr., Biometrics, 11, 123, 1955.
24. Smith H. Fairfield, Biometrics, 7, 70, 1951.
25. Snedecor George W., Proceedings of the International Statistical Conferences, 3, 440, 1947.
26. Snedecor George W., Breneman W. R., Iowa State College Journal of Science, 19, 333, 1945.
27. Snedecor George W., Haber E. S., Biometrics (Bulletin), 2, 61, 1946.
28. Southern Cooperative Series Bulletin No. 10, page 114, 1951.
29. Strand Norman, Jessen Raymond, Iowa Agricultural Experiment Station Research Bulletin 315, 1943.
30. Wilk M. B., Kempthorne O., Iowa State College Statistical Laboratory, 1954.
31. Wilsie C. P., Iowa State College Agricultural Experiment Station, 1944.
32. Yates F., Journal of Agricultural Science, 23, 108, 1933.
33. Yates F., Cochran W. G., Journal of Agricultural Science, 28, 556, 1938.
34. Yates F., Imperial Bureau of Soil Science, Technical Communication No. 35, 1937.

Глава 13 КОВАРИАЦИЯ

1. Введение. Неконтролируемые условия, сопровождающие опыт, оказывают свое влияние как на экспериментальную ошибку, так и на оценки эффектов изучаемых в опыте вариантов. В тех случаях, когда предоставляется

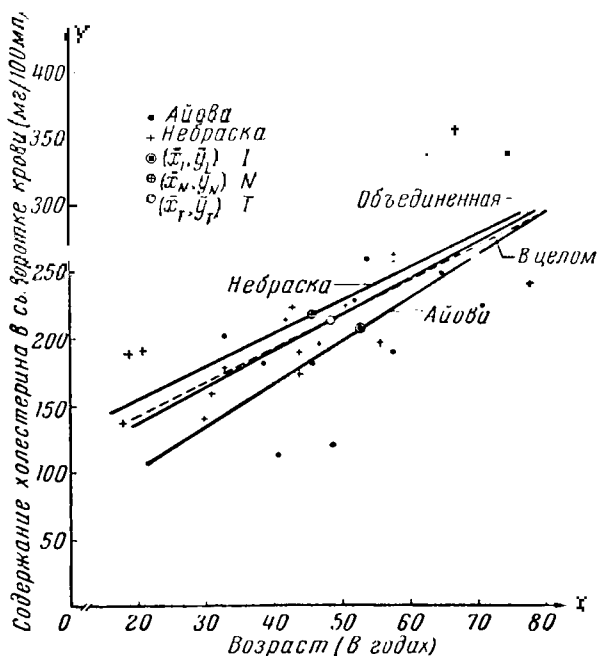


Рис. 26. График для 11 наблюдений по штату Айова и 19 наблюдений по штату Небраска. X — возраст; Y — содержание холестерина.

эти варианты воздействуют также и на X, для того, чтобы выделить непосредственное влияние вариантов опыта на Y.

2. Ковариация в полностью рандомизированном опыте с двумя вариантами. При обследовании населения Среднего Запада на предмет изучения состояния здоровья в зависимости от питания было определено, в частности, содержание в сыворотке крови холестерина у 56 случайно отобранных женщин штата Айова и у 130 женщин штата Небраска. В таблице 170 приведены выборочные данные из результатов этого обследования. Рисунок 26 дает графическое изображение двух рядов данных с явно выраженной линейной регрессией на возраст и несомненной однородностью дисперсий. Первое, т. е. наличие

возможность произвести измерение, хотя бы приближенное, таких сопутствующих условий, часто можно эти данные использовать для получения большего количества сведений, получаемых из опыта. Соответствующие этому статистические методы известны под названием *ковариации*. Они являются распространением методов регрессии главы 6 на общие схемы экспериментальных исследований. В этом случае сопутствующая величина, которая была измерена, будет являться независимой переменной X, а в качестве зависимой переменной Y будет участвовать наблюдаемый в опыте результат. Предметом изучения здесь является влияние X на Y для того, чтобы можно было сравнить воздействие вариантов опыта на Y независимо от X (или с поправкой на X), или, если

линейной регрессии, является предпосылкой, на которой строится настоящая глава; второе же, т. е. однородность дисперсий, должна быть проверена при помощи известных критериев.

ТАБЛИЦА 170

Возраст и содержание холестерина (в мг/100 мл) в сыворотке крови у женщин штата Айова и Небраска

Айова, n=11		Небраска, n=19			
возраст X	холестерин Y	возраст X	холестерин Y	возраст X	холестерин Y
46	181	18	137	30	140
52	228	44	173	47	196
39	182	33	177	58	262
65	249	78	241	70	261
54	259	51	225	67	356
33	201	43	223	31	159
49	121	44	190	21	191
76	339	58	257	56	197
71	224	63	337		
41	112	19	189		
58	189	42	214		
Сумма 584	2285			873	4 125
$\bar{x}_I = 53,1$	$\bar{y}_I = 207,7$			$\bar{x}_N = 45,9$	$\bar{y}_N = 217,1$
Айова					
$\Sigma X^2 = 32\ 834$		$\Sigma XY = 127\ 235$		$\Sigma Y^2 = 515\ 355$	
$C = 31\ 005$		121 313		474 657	
$\Sigma x^2 = 1\ 829$		$\Sigma xy = 5\ 922$		$\Sigma y^2 = 40\ 698$	
Небраска					
$\Sigma X^2 = 45\ 677$		$\Sigma XY = 203\ 559$		$\Sigma Y^2 = 957\ 785$	
$C = 40\ 112$		189 533		895 559	
$\Sigma x^2 = 5\ 565$		$\Sigma xy = 14\ 026$		$\Sigma y^2 = 62\ 226$	
В целом: n = 30					
$\Sigma X = 1\ 457$	$\bar{x}_T = 48,6$	$\Sigma X^2 = 78\ 511$	$\Sigma XY = 330\ 794$	$\Sigma Y^2 = 1\ 473\ 140$	
$\Sigma Y = 6\ 410$	$\bar{y}_T = 213,7$	$C = 70\ 762$	311 312	1 369 603	
		$\Sigma x^2 = 7\ 749$	$\Sigma xy = 19\ 482$	$\Sigma y^2 = 103\ 537$	

Вычисления проводятся по образцу главы 6, что дает скорректированные суммы квадратов и произведений для каждой выборки в отдельности и для итога (т. е. объединения двух выборок вместе). Для последнего имеем $\Sigma X^2 = 32\ 834 + 45\ 677 = 78\ 511$, т. е. сумму значений ΣX^2 для отдельных выборок, и т. д. Скорректированные суммы квадратов и произведений вносятся в строки 1, 2 и 7 таблицы 171.

Прежде чем идти дальше, полезно ознакомиться с дисперсионными анализами для X и Y в отдельности.

Для X: Общее	29	7 749	
Внутри выборок	28	$1\ 829 + 5\ 565 = 7\ 394$	264
Между выборками	1	355	355
Для Y: Общее	29	103 537	
Внутри выборок	28	$40\ 698 + 62\ 226 = 102\ 924$	3676
Между выборками	1	613	613

Ковариационный анализ. Данные о холестерине

Стро- ка	Штат	f	Σx^2	Σxy	Σy^2	Коеф. регрес- сии	Отклонения от регрессии		
							f	$\Sigma y^2 -$ $-(\Sigma xy)^2/\Sigma x^2$	средний квадрат
1	Айова	10	1 829	5 922	40 698	3,24 2,52	9	21 524	2 392
2	Небраска	18	5 565	14 026	62 226		17	26 875	1 581
3	Внутри выборки						26	48 399	1 862
4	Кoeffициент рег-						1	708	708
5	рессии	28	7 394	19 948	102 924	2,70	27	49 107	1 819
6	Общее								
7	Приведенные						1	5 450	5 450
	средние								
	Итого	29	7 749	19 482	103 537		28	54 557	

Средние двух выборок переменной X — возраста различаются друг от друга ненамного больше, чем это можно ожидать в выборках из общей совокупности со средней μ_x . Если бы это было не так, т. е. если бы между средними X по штатам было существенное различие, то следовало бы считать, что как X , так и Y находятся под влиянием вариантов (у нас — условий в каждом из этих штатов); в таком случае требовалось бы объяснить это влияние.

Подобно этому, средние Y на рисунке находятся примерно на одном и том же уровне над осью X . В данном случае указаны *приведенные средние*, подлежащие дальнейшему изучению; в нашем примере они, как будет это установлено, различаются между собой в большей степени, чем неприведенные средние.

Кoeffициенты регрессии, вычисляемые по обычной формуле $b = \Sigma xy / \Sigma x^2$, и соответствующие средние \bar{x} и \bar{y} позволяют провести на графике линии регрессии. Суммы квадратов отклонений для каждой из этих регрессий $\Sigma d_{y..x}^2$ внесены в таблицу 171. В порядке компенсации использования данных для выявления коoeffициента регрессии соответствующие числа степеней свободы уменьшаются на единицу.

Конечной целью исследования является решение вопроса, можно ли считать линейные регрессии холестерина на возраст одинаковыми для штатов Айова и Небраска. Эти регрессии могут различаться друг от друга в нескольких отношениях: по наклону прямой, по приподнятости ее над осью X и по системе отклонений, характеризуемой $\sigma_{y..x}^2$. Основным интерес обычно представляет уровень над осью X ; будет ли среднее содержание холестерина одинаковым, если в том и другом штате взят один и тот же средний возраст? Для того чтобы сделать этот вопрос определенным, мы должны ввести следующие допущения.

1. Наши две выборки взяты из совокупностей, имеющих общую σ^2 .

2. Наклоны двух линий регрессии одинаковы, т. е. в совокупности эти две прямые расположены параллельно друг другу.

При просмотре данных таблицы 171 мы видим, что нет никаких данных, свидетельствующих против первого допущения: средние квадраты, стоящие в правой графе по строкам 1 и 2, почти одинаковы. Когда в этом возникнет необходимость, то вопрос об общей σ^2 может быть решен на основе методов параграфа 8 главы 4 и параграфа 20 главы 10. Если будет в этом случае установлена разнородность дисперсий, то тем самым уже будет получена из выборки определенная, имеющая значение информация. Другими словами, это приводит к требованию раздельного рассмотрения этих отличающихся друг от друга дисперсий, или указывает на то, что, возможно, будет эффективным то или иное преобразование переменной. Во всех остальных случаях, когда нет доказательств в пользу противного, допустимо предположение

об однородности дисперсий. Следовательно, в нашем случае можно произвести по третьей строке таблицы объединение соответствующих степеней свободы и двух значений $\Sigma d_{y..x}^2$. В результате получится средний квадрат 1862 при $f=26$, который является оценкой σ^2 , полученной при объединении варьирования внутри штатов.

Для решения вопроса о втором допущении следует посмотреть на значения коэффициентов регрессии, а также для наглядности на рисунок 26. При таких малых выборках, как наши, наблюдающееся различие в наклонах прямых часто может быть полностью приписано выборочному варьированию. Доказательство этого можно получить следующим путем: производится объединение сумм квадратов и произведений этих двух выборок с целью получить общую регрессию, представленную у нас в строке 5 таблицы 171. Линия этой регрессии проходит через точку общих средних $\bar{x}_T=48,6$ лет и $\bar{y}_T=213,7$ мг/100 мл и имеет наклон, определяемый коэффициентом регрессии 2,70. Этот наклон является взвешенной средней из наклонов двух выборочных регрессий. При допущении параллельности линий регрессии в совокупности эта обобщенная регрессия и отклонения от нее дают возможность определить на основе всех выборочных данных оценки для β и σ^2 .

Характерной особенностью только что описанного способа является то, что он может быть интерпретирован как движение двух исходных выборочных линий регрессии параллельно самим себе до тех пор, пока их средние (\bar{x}, \bar{y}) не сольются. Все соответствующие этим линиям выборочные точки также движутся вместе с ними, в результате чего отклонения x и y остаются неизменными. В этой новой позиции $\Sigma d_{y..x}^2$ для выборочной регрессии останется такой же, как прежде, а именно, она будет наименьшей из возможных сумм квадратов отклонений по отдельным регрессиям. Но если только эти линии регрессии с самого начала не были параллельными, $\Sigma d_{y..x}^2$ для общей регрессии будет больше, чем сумма для двух выборок, взятых вместе, так как первая минимальна. Разность $49\ 107 - 48\ 399 = 708$ строки 4 измеряет различие между двумя выборочными коэффициентами регрессии. Этот средний квадрат можно сравнивать со средним квадратом «внутри выборок»—1862. Следовательно, второе допущение может быть принято, так как средний квадрат «коэффициентов регрессий» оказался даже меньше, чем выборочное варьирование «внутри штатов».

Если бы несущественность не была столь очевидна, то понадобилась бы проверка гипотезы $\beta_1 = \beta_2$ при помощи критерия:

$$F = \frac{\text{средний квадрат коэффициентов регрессии}}{\text{средний квадрат внутри выборок}}; f = 1 \text{ и } 26.$$

Если будет установлена существенность различий между коэффициентами регрессии, то тем самым мы придем к концу исследования. В таком случае рассмотрению подлежит вопрос «Почему это произошло?» Ставить вопрос относительно различных уровней, на которых находятся линии регрессии в совокупности, не имеет смысла, пока не установлено, что линии параллельны.

Так как в нашем случае допущения 1 и 2 представляются приемлемыми, то на очередь встает вопрос о проверке совпадения регрессий в совокупности или, поскольку они параллельны, вопрос о том, находятся ли они на одном и том же уровне. Это может быть установлено на основе сравнения строки 5 с 7 таблицы дисперсионного анализа. Разность между $d_{y..x}^2$ здесь равна $54\ 557 - 49\ 107 = 5450$, что соответствует выборочной разности между уровнями. $F = 5450/1819 = 3,00$; $f = 1$ и 27, что меньше 5%-ного уровня достоверности. Вывод из всего этого тот, что данные две совокупности могут иметь одну и ту же линию регрессии; по крайней мере имеющиеся различия не столь велики, чтобы они были бы обнаружены при настоящем выборочном наблюдении.

В исследовании этого вопроса на более обширном материале были получены еще менее выраженные различия, чем установленные в данных двух

частных выборках. В результате этого исследователи пришли к убеждению о возможности объединения данных для дальнейшей изучения интересующего их вопроса по всему обобщенному ими материалу.

При указанных выше двух допущениях принятая здесь модель величины Y такова:

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij},$$

$i=1$ и 2 ; $\sum \alpha_i = 0$ и $\varepsilon_{ij} = N(0, \sigma)$. Величина β характеризует наклон общей линии регрессии в совокупности, и x_{ij} является отклонением любого X от общей средней $\bar{x}_T = 48,6$ лет. Разность $\alpha_1 - \alpha_2$ представляет различие между уровнями линий регрессии для штатов Айова и Небраска. Проверяемая выше гипотеза сводится к равенству $\alpha_i = 0$.

Разность $\alpha_1 - \alpha_2$ оценивается при помощи разности двух приведенных средних (параграф 5 главы 6).

$$\text{Для Небраски: } \bar{y}_N - b(\bar{x}_N - \bar{x}_T) = 217,1 - 2,70(45,9 - 48,6) = 224,4$$

$$\text{Для Айовы: } \bar{y}_I - b(\bar{x}_I - \bar{x}_T) = 207,7 - 2,70(53,1 - 48,6) = 195,5$$

$$\text{Разность: } \frac{28,9}{28,9}$$

Как было ранее установлено, разность 28,9 мг/100 мл незначительна.

Можно видеть, что эта разность между приведенными средними больше, чем разность между первоначальными средними Y , которая равна $217,1 - 207,7 = 9,4$. В других случаях возможно и обратное положение. Но ошибка для оценки существенности различий в ковариационном анализе значительно меньше, чем ошибка при анализе величины Y , взятой сама по себе.

Пример 1. Ниже приводятся две выборки, подобранные так, чтобы облегчить расчеты.

Вариант 1	X	29	20	14	21	6
	Y	22	22	20	24	12
Вариант 2	X	15	9	1	6	19
	Y	30	32	26	25	37

Разместите эти данные на одном графике. Легко видеть, что эти две регрессии не слишком отклоняются от параллельности.

Пример 2: Проведите по данным предыдущего примера ковариационный анализ:

Строка	Выборка	f	$\sum x^2$	$\sum xy$	$\sum y^2$	b	f	$\sum d^2$	Средний квадрат
1	1	4	294	134	88	0,456	3	26,93	8,98
2	2	4	204	117	94	0,574	3	26,90	8,97
3	Вместе						6	53,83	8,97
4	Коэффициент регрессии						1	1,66	1,66
5	Общее	8	498	251	182	0,504	7	55,49	7,93
6	Приведенные средние						1	372,56	372,56
7	В целом	9	658	51	432		8	428,05	

Для различий между приведенными средними $F=47,0$

Пример 3. Вычислите две приведенные средние и нанесите их на график совместно с двумя частными линиями регрессии и общей регрессией. *Ответ:* 17,98 и 32,02.

Пример 4. Напишите уравнение общей регрессии по данным таблицы 171. *Ответ:* $\hat{Y} = 2,70X + 82,5$ мг/100 мл.

Пример 5. Найдите доказательство того, что после передвижения какой-либо точки к новому началу координат в (\bar{x}_T, \bar{y}_T) отклонение ее от общей регрессии иное, чем отклонение от передвинутой выборочной регрессии (если только выборочные регрессии не будут строго параллельны).

Пример 6. Нанесите на график выборочных данных о холестерине точки, соответствующие приведенным средним: для штата Небраска (48,6; 224,4), для штата Айова (48,6; 195,5).

Пример 7. Вычтите из строки 7 таблицы 171 строку 5 и определите тем самым суммы квадратов и произведений «между средним». Установите наклон прямой, проходящей через эти две средние. *Ответ:* $-1,31$. Проведите эту прямую на графике.

Пример 8. Пусть в модели ковариации $\mu = 20$, $\alpha_1 = 4$, $\alpha_2 = -4$, $\beta = 1$, $\bar{x}_T = 10$. Вычислите $Y_{1j} - e_{1j}$, если $X_{1j} = 12$. *Ответ:* 26. Постройте график для этих двух регрессий в совокупности.

3. Ковариация в полностью рендомизированном опыте с числом вариантов, большим двух. Методы предыдущего параграфа легко распространить на опыт при совместном изучении трех или большего числа вариантов. Данные таблицы 172 дают опыт этого типа [7]. В данном случае будем считать, что произведено случайное распределение свиной по группам (см. параграф 7 этой главы). Введение поправок в суммы квадратов и произведений, показанных в конце этой таблицы (как это делалось и в таблице 170), приводит к данным четырех первых строк и девятой строки таблицы 173.

ТАБЛИЦА 172

Начальный вес X и средний суточный привес Y в 4 группах свиней по 10 голов в каждой. Полная рендомизация

Группа	6		7		8		9	
	X	Y	X	Y	X	Y	X	Y
	79	1,96	61	1,40	62	1,22	71	1,15
	65	1,77	59	1,79	73	1,39	60	1,28
	57	1,62	59	1,61	58	1,28	54	1,40
	51	1,76	53	1,47	43	1,28	50	1,37
	57	1,88	56	1,69	50	1,45	60	1,19
	66	1,50	50	1,48	44	1,22	61	1,18
	44	1,60	45	1,40	48	1,31	44	1,20
	41	1,49	39	1,42	51	1,57	53	0,96
	44	1,77	38	1,29	40	1,21	41	1,13
	36	1,27	45	1,26	38	1,06	38	1,12
Суммы	540	16,62	505	14,81	507	12,99	532	11,98

ΣX^2	30770	26143	26771	29248
ΣY^2	28,0068	22,1917	17,0569	14,4992
ΣXY	913,24	756,65	664,20	638,50

Итоги по опыту:
 $\Sigma X^2 = 112932$

$\Sigma X = 2084$
 $\Sigma XY = 2972,59$

$\Sigma Y = 56,40$
 $\Sigma Y^2 = 81,7546$

Сначала рассмотрим дисперсионный анализ для X и Y в отдельности:

в целом	39	$\frac{X}{4356}$		$\frac{Y}{2,2306}$	
внутри групп	36	$\frac{4262}{118}$		$\frac{0,9726}{0,0270}$	
между средними	$\frac{3}{94}$		$\frac{31}{1,2580}$		$0,4193^{**}$

Между средними различие оказалось меньшим, чем то, которое можно обычно ожидать в случайной выборке. В параграфе 7 этой главы этому будет

дано объяснение. В противоположность этому средними по привесам имеются несомненные различия. Теперь мы перейдем к изучению вопроса о том, в какой мере эти различия находятся в связи с различиями в первоначальных весах.

ТАБЛИЦА 173

Ковариационный анализ. Четыре группы по 10 свиней в каждой

Стро-ка	Группа	f	Σx^2	Σxy	Σy^2	b	Отклонения от регрессии		
							f	$\Sigma d_{y \cdot x}^2$	средний квадрат
1	6	9	1 610	15,76	0,3844	0,40	8	0,2301	0,0209 0,0251 0,02123 0,4066**
2	7	9	640	8,75	0,2581	0,14	8	0,1385	
3	8	9	1 066	5,61	0,1829	0,005	8	0,1534	
4	9	9	946	1,16	0,1472	0,001	8	0,1458	
5	Внутри групп						32	0,6678	
6	Коэффициент регрессии						3	0,0752	
7	Общее	36	4 262	31,28	0,9726	0,0073	35	0,7430	
8	Приведенные средние						3	1,2199	
9	В целом	39	4 356	34,15	2,2306		38	1,9629	

Переходя к ковариационному анализу, можно видеть, что здесь представляется возможность произвести оценку существенности различий между приведенными средними, по строкам 1—4 установить однородность дисперсий и приемлемость предположения о параллельности линий регрессий в совокупности (строки 5 и 6). Критерий, построенный по данным строк 7 и 8, указывает на значительные различия между приведенными средними 1,65; 1,49; 1,31 и 1,19. Можно заметить, что приведенные средние только немногим отличаются от первоначальных средних и что критерий F для первоначальных и приведенных средних имеют почти одинаковое значение. В данном опыте ковариация оказалась только умеренно эффективной; какова эта эффективность, будет установлено в дальнейшем.

Для планируемого сравнения между двумя приведенными средними удобнее всего вычислить величину:

$$\bar{y}_{1A} - \bar{y}_{2A} = \bar{y}_1 - \bar{y}_2 - b(\bar{x}_1 - \bar{x}_2) = D.$$

Например, рационы кормления групп 6 и 7 отличались только добавлением для первой группы дрожжей. Для этих групп имеем:

$$D = 1,662 - 1,481 - 0,0073(54,0 - 50,5) = 0,155 \text{ фунта в день.}$$

Средний квадрат для D будет:

$$s_D^2 = s_{y \cdot x}^2 \left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\Sigma x_E^2} \right],$$

где n_1 и n_2 численности животных в каждой из двух групп и Σx_E^2 является суммой квадратов X для «ошибки», т. е. по строке 7; $s_{y \cdot x}^2$ — средний квадрат «ошибки». Для вычисленной выше разности находим:

$$s_D^2 = 0,02123 \left[\frac{2}{10} + \frac{(54,0 - 50,5)^2}{4262} \right] = 0,00431,$$

откуда $s_D = 0,0657$ фунта в день. Доверительные 95 %-ные пределы в данном случае будут $0,155 \pm 2,030 \times 0,0657 = 0,155 \pm 0,133$ фунта в день. Вместе с этим можно, если это желательно, применить и критерий t . Очевидно, что добавка дрожжей приводит к увеличению скорости роста, но, вообще говоря, это различие небольшое.

Если в опыте приходится производить большое число попарных сравнений средних, то вычисление всех s_D становится довольно трудоемкой работой.

В этом случае можно воспользоваться для всех сравнений усредненным значением s_D [6,3]. Потеря точности при этом небольшая, в особенности если число степеней свободы достаточное (Кокран и Кокс [1] считают 20 или более) и если варьирование X небольшое (средний квадрат X для вариантов должен быть несущественным). К этому усреднению мы придем путем вычисления взамен $s_{y \cdot x}^2$ среднего квадрата:

$$\overline{s_{y \cdot x}^2} = s_{y \cdot x}^2 \left[1 + \frac{\Sigma x_T^2}{(a-1) \Sigma x_E^2} \right],$$

где Σx_T^2 является суммой квадратов вариантов, исчисленной по начальным весам таблицы 173, т. е. $4356 - 4262 = 94$. Подстановка в эту формулу других требуемых данных из той же таблицы дает:

$$\overline{s_{y \cdot x}^2} = 0,0212 \left[1 + \frac{94}{3 \times 4262} \right] = 0,0214.$$

Теперь усредненный средний квадрат D при численностях наблюдений n_1 и n_2 будет:

$$\overline{s_D^2} = \overline{s_{y \cdot x}^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

В нашем опыте по кормлению свиней между 4 средними заранее было намечено два сравнения. Однако в целях иллюстрации метода допустим, что требуется произвести все сравнения между этими средними. При этом предположении применим критерий последовательного сравнения:

$$\overline{s_{yA}^2} = 0,0214/10 = 0,00214; \quad \overline{s_{yA}} = 0,046 \text{ фунта в день.}$$

Из таблицы 90:

$$Q = 2,87; 3,46; 3,81.$$

Тогда:

$$Q \overline{s_{yA}} = 0,13; 0,16; 0,18.$$

Разности	Группы	\overline{y}_A	$\overline{y}_A - 1,19$	$\overline{y}_A - 1,31$	$\overline{y}_A - 1,49$
	6	1,65	0,46*	0,35*	0,16*
	7	1,49	0,30*	0,18*	
	8	1,31	0,12		
	9	1,19			

После полного проведения ковариационного анализа экспериментатор, наверное, спросит: «А стоило ли на это затрачивать время?» Эффективность ковариации относительно простого анализа средних суточных привесов животных определяется отношением соответствующих средних квадратов для ошибки:

$$\frac{s_y^2}{s_{y \cdot x}^2} = \frac{s_y^2}{s_{y \cdot x}^2 \left[1 + \frac{\Sigma x_T^2}{(a-1) \Sigma x_E^2} \right]},$$

где s_y^2 основывается на общей ошибке для Y . В опыте со свиньями $s_y^2 = 0,9726/36 = 0,0270$; $s_{y \cdot x}^2 = 0,0214$. Следовательно, *относительная эффективность* будет $0,0270/0,0214 = 126\%$. Без ковариации для достижения той же точности понадобилось бы иметь в опыте число животных на 26% больше, чем было фактически. В большинстве опытов значение ковариации не ограничивается только уменьшением ошибки. Много ценного в этом случае дает более детальное изучение результатов опыта самих по себе. Действительно, основная масса информации может быть получена из рассмотрения соответствующих графиков; ковариация к этому присоединяет точное числовое уравнение.

Пример 9. Приводимые ниже данные извлечены из более широкого исследования о корреляции между гибридами поколения F_1 и чистыми линиями кукурузы [4,8].

Гибриды F_1	Число дней от 30 июня до появления первых женских соцветий (X) и урожай зерна кукурузы чистых линий на делянку (Y)											
Сильвер Кинг X	24	31	26	30								
	Y	7,7	5,4	5,2	4,0							
Ланкастер X	33	33	32	36	33	38	30	38	31	32	32	
	Y	9,6	7,8	9,6	7,7	8,2	7,3	11,3	9,5	8,8	8,4	6,8
Остерленд X	31	33	33	33	27	32	36					
	Y	4,8	9,2	8,5	8,8	9,2	7,9	5,9				

Проведите дисперсионный анализ для времени появления первых женских соцветий и для урожая. Для времени появления соцветий $F=6,05$, что связано с выбором скрещиваний F_1 ; линия Сильвер Кинг является раннеспелой белозерной. Остерленд ранней желтозерной и Ланкастер поздней с желтым зерном. Таким образом, X определено связано с «вариантами», т. е. скрещиваниями. При этих условиях нет никакого смысла приводить урожай к общему X , кроме того случая, когда изучается вопрос: в какой мере скрещивание определяет различия урожаев? Сводится ли влияние скрещивания к одной только раннеспелости или здесь были некоторые другие генетические и сопутствующие характеристики?

Пример 10. Проведите расчет следующего ковариационного анализа:

Гибриды F_1	f	Σx^2	Σxy	Σy^2	Отклонение от регрессии		
					f	$\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}$	средний квадрат
1 Сильвер Кинг	3	32,75	-11,42	7,17	2	3,19	1,60
2 Ланкастер	10	72,73	-12,38	16,71	9	14,60	1,60
3 Остерланд	6	44,86	-8,46	18,02	5	16,42	3,28
4 Внутри групп					16	34,21	2,14
5 Коэффициент регрессии					2	0,77	0,38
6 Общее	19	150,34	-32,26	41,90	18	34,98	1,94
7 Приведенные средние					2	32,97	16,48**
8 В целом	21	246,00	18,90	69,40	20	67,95	

Отсюда следует, что варьирование в отношении раннеспелости между чистыми линиями не является причиной различий в урожайности растений поколения F_1 .

Пример 11. Сумма квадратов отклонений между приведенными средними по строке 7 состоит из двух частей. Одна часть относится к отклонениям средних F_1 от соответствующей им регрессии, что характеризуется суммами квадратов и произведений, получаемыми путем вычитания из строки 8 и строки 6:

Регрессия средних: 2 95,66 51,16 27,50 1 0,14 0,14.

Вторая же часть $32,97 - 0,14 = 32,83$ с одной степенью свободы относится к разности между общей регрессией и регрессией средних. Для этой разности $F = 32,83 / 1,94 = 16,9$ при $f = 18$, что далеко выходит за пределы уровня существенности 0,005. Это указывает на то, что: 1) отклонения средних от регрессии незначительны: средние расположены почти на одной прямой; 2) эта регрессия средних отличается от общей регрессии; последняя направлена книзу, а первая вверх. Раннеспелость в отношении чистых линий благоприятствует урожаю, т. е. более раннее начало развития приводит к более продолжительному периоду вегетации. Но среди гибридов F_1 раннеспелости сопутствует падение урожайности.

Пример 12. Построить график регрессий для чистых линий в трех скрещиваниях нанеся как отдельные точки, так и средние. Проведите линии общей регрессии и регрессии средних.

4. Ковариация в полностью рендомизированных опытах: второй способ. Практический опыт и графические методы изучения материала часто позво-

ляют вполне уверенно ожидать наличие линейных параллельных друг другу регрессий при однородной дисперсии. В таких случаях оценка существенности различий между приведенными средними может быть проведена с меньшей затратой труда, чем это требуется при полном ковариационном анализе, так как в этом случае отпадает необходимость в оценке различий между коэффициентами регрессии. Так, если взять опыт со свиньями (табл. 172), то для анализа потребуются только суммы квадратов и произведений в целом и такие же суммы для вариантов. Исправленные суммы квадратов и произведений по строке «в целом» вычисляются, как по строке 9 таблицы 173, и вносятся в таблицу 174. Соответствующие вычисления по строке «между средними» проводятся обычным способом:

$$\text{для } X: (540^2 + 505^2 + 507^2 + 532^2)/10 - 2084^2/40 = 94 \text{ и т. д.}$$

Разности третьей строки этой таблицы те же самые, что и разности, вычисленные непосредственно по строке 7 таблицы 173. Критерий существенности остается одинаковым для обеих таблиц.

ТАБЛИЦА 174

Ковариационный анализ для полностью рендомизированного опыта.
Опыт по кормлению свиней

Источник варьирования	f	Σx ²	Σxy	Σy ²	Отклонение от регрессии		
					f	сумма квадратов	средний квадрат
В целом	39	4 356	34,15	2,2306	38	1,9629	
Средние	3	94	2,87	1,2580			
Общее	36	4 262	31,28	0,9726	35	0,7430	0,0212
Приведенные средние . .					3	1,2199	0,4066**

Особенность, которую следует отметить, в данном случае состоит в том, что при разложении в регрессиях «в целом» и «общее» суммы квадратов Σy² не выделяется регрессия средних. Варьирование между фактическими приведенными средними для оценки этой регрессии не может служить подходящей базой потому, что эти средние коррелированы между собой посредством общей регрессии.

5. Ковариация в рендомизированных блоках. Теория, изложенная в предыдущих параграфах, приложима и к опытам с рендомизированными блоками. В этом случае модель опыта становится такой:

$$Y_{ij} = \mu + \alpha_i + \rho_j + \beta x_{ij} + \varepsilon_{ij},$$

где ρ_j — параметры блоков.

Здесь не совсем просто произвести оценку линейности регрессии или однородности дисперсий; и то, и другое приходится предполагать. В связи с этим мы должны ограничиться методами, соответствующими тем, о которых говорилось в параграфе 4 главы 13.

Решение данной задачи представлено на примере таблицы 175 [9]. Здесь наблюдается некоторое варьирование густоты стояния, которая может влиять на средние по сортам и на оценку ошибки. В этих условиях желательно получить эти средние приведенными к одной густоте стояния и использовать такие *приведенные* средние для соответствующих сравнений. Более того, при элиминировании из ошибки сопутствующего влияния густоты стояния, взятой в качестве переменной величины X, здесь происходит повышение эффективности опыта.

Анализ X и Y в отдельности дает:

X: сорта	5	9,17	Y: сорта	5	1898,0
ошибка	15	7,59	ошибка	15	583,5
		F = 1,21			F = 3,25*

Густота стояния (X) и урожай (Y) четырех сортов кукурузы
(в фунтах полевого веса початков). Ковариация в рендомизированных блоках

Сорта	Блок								Итого	
	1-й		2-й		3-й		4-й			
	X	Y	X	Y	X	Y	X	Y	X	Y
A	28	202	22	165	27	191	19	134	96	692
B	23	145	26	201	28	203	24	180	101	729
C	27	188	24	185	27	185	28	220	106	778
D	24	201	28	231	30	238	30	261	112	931
E	30	202	26	178	26	198	29	226	111	804
F	30	228	25	221	27	207	24	106	106	860
Итого:	162	1 166	151	1 181	165	1 222	154	1 225	632	4 794

Источник варьирования	f	Σx^2	Σxy	Σy^2	Отклонения от регрессии		
					f	$\Sigma d^2_{y \cdot x}$	средний квадрат
В целом	23	181,33	1 485,00	18 678,50			
Блоки	3	21,67	8,50	436,17			
Сорта	5	45,83	559,25	9 490,00			
Ошибка	15	113,83	917,25	8 752,33	14	1 361,07	97,22
Сорт плюс ошибка . .	20	159,66	1 476,50	18 242,33	19	4 537,99	
Приведенные средние					5	3 226,92	645,38**

Первый анализ убеждает нас в том, что густота стояния оказывает на сорта только случайное влияние; это означает, что зерно отдельных сортов в одинаковой степени жизнеспособно. Если бы в данном случае было бы обнаружено, что варианты (сорта) так же влияют на X , как это установлено в отношении Y , то интерпретацию результатов ковариационного анализа пришлось бы изменить. В отношении анализа урожая вопросы, требующие ответа, будут такими: отличаются ли приведенные урожаи от непосредственно исчисленных урожаев; находится ли ошибка опыта под влиянием различных густот стояния; различаются ли приведенные урожаи друга от друга существенно?

При построении таблицы 175 не введено ничего нового в отношении расчетов сумм квадратов и произведений. Новым моментом является только суммирование «вариантов» и «ошибки» для изучения частного итога, помещенного в предпоследней строке таблицы. Этот частный итог занимает место итога «в целом» таблицы 174. В остальном оценка существенности в обеих таблицах остается одной и той же. Здесь заметно резкое уменьшение среднего квадрата «ошибки». При оценке различий по сортам критерий F благодаря этому возрастает с 3,25 до $645,38/97,22 = 6,64$.

Приведенные средние в данном случае будут:

$$A = 191,8; B = 191,0; C = 193,1; D = 219,3; E = 189,6; F = 213,6.$$

Применяя последовательный метод оценки всех разностей, находим, что D и F между собой не различаются существенно, но что они существенно отличаются от всех других сортов, которые, в свою очередь, не имеют между собой существенных различий.

Эффективность ковариации по отношению к анализу только самого урожая составляет здесь 555% (см. примеры). Это означает, что без кова-

риации для обнаружения таких же, как в данном случае, различий понадобилось бы 22-кратная повторность опыта. Приведение к одной густоте стояния не оказало большого влияния на оценки средних урожаев. Замечательным в этом опыте является большое снижение ошибки опыта; большая часть различий Y обусловлена различиями в густоте стояния, а не в урожайности на одно растение.

Пример 13. Вычислите в опыте с кукурузой приведенные средние и проведите до конца оценку всех разностей. *Ответ:* $s_{\bar{y}.x} = 5,12$ фунта.

Пример 14. Определите эффективность ковариации по отношению к обычному опыту с рандомизированными блоками.

6. Сравнения при ковариации: факториальные опыты. Об единичном сравнении двух приведенных средних было сказано в параграфе 3 этой главы. В этом случае применялся критерий t . Но можно применить и критерий F , поступая следующим образом:

Группа 6:	$\Sigma X = 540$	$\Sigma Y = 16,62$
Группа 7:	505	14,81
Разность:	35	1,81

На основании правила параграфа 2 главы 12 имеем суммы квадратов для X и Y :

$$\frac{35^2}{10 \times 2} = 61; \quad \frac{1,81^2}{10 \times 2} = 0,1638.$$

Небольшое видоизменение этого правила путем подстановки вместо квадрата произведения дает сумму произведений:

$$\frac{35 \times 1,81}{10 \times 2} = 3,17.$$

Теперь по методам параграфов 4 и 5 этой главы имеем:

	f	Σx^2	Σxy	Σy^2	f	$\Sigma y^2 \cdot x$	Средний квадрат
Варианты T	1	61	3,17	0,1638	35	0,7430	0,02123
Ошибка E	36	4262	31,28	0,9726			
T + E	37	4323	34,45	1,1364	36	0,8619	
Для разностей по вариантам	1				1	0,1189	0,1189

«Варианты» и представляет то сравнение, которое подлежит сделать. Ошибка переносится из таблицы 173. Этот способ сопоставления «варианты + + ошибка» с «ошибкой» является обычным при оценке различий по вариантам.

Из приведенной таблички следует $F = 0,1189 / 0,02123 = 5,60$. В подтверждение правильности этого критерия напомним, что для единичной степени свободы $\sqrt{F} = \sqrt{5,60} = 2,37$, а значение t для проведенного ранее сравнения $t = 0,155 / 0,0657 = 2,36$ (небольшое различие связано с округлением данных). Критерий F является иногда более удобным для проведения некоторого ряда ортогональных сравнений, встречающихся в факториальных опытах.

В целях показа сравнений в факториальных опытах ниже приводится таблица 176, являющаяся извлечением из одного опыта, проведенного с сахарной свеклой в северной части штата Айова. Здесь не требуется никакого предварительного анализа для определения явного влияния суперфосфата как на число растений, так и на урожай. Поскольку это касается свекловода, то, если такой результат опыта типичен, он без колебания будет применять суперфосфат в качестве удобрения.

Число корней на делянку и урожай (в т на 1 акр) в опыте с рендомизированными блоками

Удобрение	Число растений и урожай	Блок						Суммы по вариантам
		1-й	2-й	3-й	4-й	5-й	6-й	
Без удобрений	Число растений	183	176	291	254	225	249	1 378
	Урожай	2,45	2,25	4,38	4,35	3,42	3,27	20,12
Суперфосфат P	Число растений	356	300	301	271	288	258	1 774
	Урожай	6,71	5,44	4,92	5,23	6,74	4,74	33,78
Хлористый калий K	Число растений	224	258	244	217	192	236	1 371
	Урожай	3,22	4,14	2,32	4,42	3,28	4,00	21,38
P+K	Число растений	329	283	308	326	318	318	1 882
	Урожай	6,34	5,44	5,22	8,00	6,96	6,96	38,92
Суммы по блокам	Число растений	1 092	1 017	1 144	1 068	1 023	1 061	6 405
	Урожай	18,72	17,27	16,84	22,00	20,40	18,97	114,20

Источник варьирования	f	Σx ²	Σxy	Σy ²	Отклонения от регрессии		
					f	Σd ² _{y,x}	средний квадрат
Блоки	5	2 756	-36,33	4,6727			
Варианты, T	3	35 253	1224,63	42,8939			
Ошибка, E	15	16 873	381,27	11,7711	14	3,1558	0,2254
T+E	18	52 126	1605,90	54,6650	17	5,1906	
Для приведенных средних					3	2,0348	0,6783

$$F = 0,6783/0,2254 = 3,01; F_{0,05} = 3,34$$

Но этот опыт и сам по себе представляет интерес. Остается неясным, в какой мере удобрение оказывает влияние на густоту стояния; возможно, что оно создает определенные условия питания растений или ускоряет их созревание и тем самым позволяет благополучно пройти через некоторый критический период. Все это в опыте было в известной мере неопределенным.

Так как влияние вариантов на X очевидно, то здесь нет оснований для оценки Y без учета влияния X; обе величины X и Y находятся под влиянием удобрения. Выдвигается для изучения вопрос: оказывает ли удобрение на Y влияние сверх того, которое производится через X. Все это еще усложняется конкуренцией между растениями.

Ковариационный анализ показывает, что большая часть действия суперфосфата сказалась на густоте стояния растений. После исключения влияния удобрения через густоту стояния общий критерий, примененный к приведенным средним, оказывается несущественным. Но отдельные корни на делянках с суперфосфатом будут в целом более крупными. Будет ли это различие существенным?

Критерий t для сравнения, относящегося к фосфору, строится на основе сумм по вариантам:

$$\text{Фосфор: } \bar{x}_1 = (1774 + 1882)/12 = 304,67;$$

$$\bar{y}_1 = (33,78 + 38,92)/12 = 6,058.$$

$$\text{Без фосфора: } \bar{x}_2 = (1378 + 1371)/12 = 229,08;$$

$$\bar{y}_2 = (20,12 + 21,38)/12 = 3,458;$$

$$D_{y.x} = \bar{y}_1 - \bar{y}_2 - b(\bar{x}_1 - \bar{x}_2) = 6,058 - 3,458 - 0,0226(304,67 - 229,08) = 0,892 \text{ тонн на 1 акр.}$$

$$s_D^2 = s_{y.x}^2 \left[\frac{2}{r} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\Sigma x_E^2} \right] = 0,2254 \left[\frac{2}{12} + \frac{75,59^2}{16,873} \right] = 0,1139;$$

$$s_D = 0,338; \quad t = 0,892/0,338 = 2,64; \quad P = 0,02.$$

Результаты этого сравнения указывают на то, что корни сахарной свеклы на делянках с суперфосфатом в целом больше, *несмотря на увеличившуюся конкуренцию между растениями.*

Вместо этого критерия t для оценки различий между приведенными средними можно применить другой критерий F , расчет которого дан в таблице 177. Для сравнения по фосфору вычисления проводятся следующим образом:

$$X: \frac{(1774 + 1882 - 1378 - 1371)^2}{6 \times 4} = \frac{(3656 - 2749)^2}{12 \times 2} = 34\,277.$$

$$Y: \frac{(72,70 - 41,50)^2}{12 \times 2} = 40,5600.$$

$$XY: \frac{(3652 - 2749)(72,70 - 41,50)}{12 \times 2} = 1179,10.$$

Показатели для «ошибки» переносятся из таблицы 176. Процесс оценки действия суперфосфата дан в таблице 177; там же приведены и значения F для оценки действия хлористого калия и взаимодействия этих двух удобрений. Идентичность оценок при помощи критериев F и t очевидна из сравнения \sqrt{F} со значением t , полученным ранее.

ТАБЛИЦА 177

Суммы квадратов и произведений для P , K и PK . Критерий F и приближенный критерий F

Сравнения	$b^2 = \frac{0,0095106}{\Sigma x^2}$	$-2b = \frac{-0,04512}{\Sigma xy}$	Σy^2	Средний квадрат	Приближенный критерий* F
P	34 277	1 179,10	40,5600	4,7759	21,19
K	425	26,93	1,7067	0,7067	3,14
PK	551	18,59	0,6273	0,0685	0,30
Сумма	35 253	1 224,62	42,8940	$\Sigma d_{y.x}^2$	Ср. кв.
Ошибка	16 873	381,27	11,7711	3,1558	0,2254
$P+E$	51 150	1 560,37	52,3311	4,7324	
Для оценки приведенной средней разности по фосфору				1,5766	1,5766

$$F = 1,5766/0,2254 = 6,99, \quad t = \sqrt{F} = 2,64.$$

$$\text{Для } K: \quad F = 3,06. \quad \text{Для } PK: \quad F = 0,29.$$

* Этот критерий не совсем подходит для данного опыта, где варьирование \bar{x} существенно. См. текст.

Заметим, что сами приведенные средние в расчетах критерия F не участвуют. Они должны вычисляться самостоятельно, по мере выявления величин, подлежащих оценке. Это делает критерий F более трудоемким, чем критерий t , если только не отсутствует сама необходимость производить оценку средних.

В тех опытах, в которых X имеет средний квадрат, достаточно малый по сравнению с ошибкой, вполне пригоден приближенный критерий F . Хотя он не подходит к условиям опыта с сахарной свеклой в связи с большим варьированием X по вариантам, однако в целях иллюстрации в таблице 177 произведен расчет этого критерия. Для каждого из 3 эффектов изучаемых вариантов производится переход к сумме квадратов, обусловленной ошибкой регрессии. Так как в этом случае самостоятельные регрессии этих эффектов не используются, а вместо них применяется общая регрессия, то способ вычисления $\sum d_{y,x}^2$ сводится к определению по каждой строке суммы квадратов $y - bx$. Взяв (по строке «ошибка») $b = 381,27/16\ 873 = 0,022596$, ставим в таблице 177 над суммами $\sum x^2$ квадрат $b^2 = 0,0005106$ и над $\sum xy$ произведение $-2b = -0,045192$. После этого по каждой строке вычисляется $\sum y^2 - 2b\sum xy + b^2\sum x^2$. Так, для фосфора P имеем:

$$40,5600 - 0,045192 \times 1179,10 + 0,0005106 \times 34277 = 4,7729;$$

$$F = 4,7729/0,2254 = 21,19.$$

При такой приближенной оценке критерий F всегда несколько преувеличен, но это не является серьезной помехой, если варьирование X небольшое. Но, как вы можете видеть, в опыте с сахарной свеклой, где варьирование X значительно, этот критерий неприменим.

Пример 15. При изучении веса отдельных корней сахарной свеклы может быть применен обычный метод использования отношения Y/X . Этот метод даст правильный результат, если шкала новой переменной будет линейной. Как было показано в параграфе 14 главы 6, если регрессия Y на X линейна, то отношение Y/X не будет линейным, за исключением случая, когда регрессия проходит через начало координат.

Одна из возможностей выяснить этот вопрос состоит в применении критерия слагаемости к таким отношениям (параграф 14 главы 11). Я сделал подобного рода проверку и не получил никаких данных, противоречащих этой гипотезе. Другой путь состоит в подборе независимых линейных моделей для X и для Y таблицы 176. Получаются 24 пары отклонений $X_{ij} - \hat{X}_{ij}$ и $Y_{ij} - \hat{Y}_{ij}$. Нанесение этих парных отклонений на графике дает некоторое представление о наклоне регрессии и о распределении этих отклонений. В данном случае наклон линии регрессии, определенный по строке «ошибка», равен $b = 0,0226$.

Пример 16. В рассматриваемом опыте имеются некоторые данные в пользу того, что регрессия урожая на число корней криволинейна и поэтому отношение Y/X может быть линейным. Сделав это допущение, вычислите 24 значения Y/X (умножая в порядке кодирования каждое из них на 10000) и после этого проведите дисперсионный анализ по среднему весу корней. *Ответ:* $F = 14,30$. Если указанное допущение правильно, то мы получаем доказательство в пользу того, что суперфосфат используется сахарной свеклой, и что его применение приводит к существенному увеличению корней, несмотря на увеличивающуюся конкуренцию между растениями. См. также параграф 5 главы 15.

7. Один вопрос, относящийся к построению опытов. В опытах, в которых представляется возможным заранее использовать некоторые предварительные данные, исследователь может сделать выбор среди нескольких схем проведения опыта. Он может, например, случайным порядком распределить объекты по вариантам и после этого применить метод ковариации в том виде, как это показано в параграфах 2 и 3 этой главы. Или же он может использовать переменную, определяемую предварительными данными, для образования блоков или групп таким образом, чтобы каждый из них содержал в себе один объект, имеющий высокое значение этой вспомогательной переменной, другой объект с низким ее значением и прочие объекты с промежуточными значениями ее. Построение этого опыта предназначено для анализа по методу рандомизированных блоков, изложенному в параграфе 11. Третий, часто применяемый план опыта состоит в *сбалансировании* групп таким образом, чтобы по каждой группе получить одну и ту же (или примерно одну и ту же) среднюю величину для вспомогательной переменной. В этом случае можно ожидать, что средние результаты получат лучшую оценку, но что оценка ошибки (если ее вычислять по методам глав 4 и 10) будет смещена в сторону преувеличения.

Метод рандомизированных блоков имеет то преимущество, что он дает приблизительно сбалансированность: каждый вариант опыта получает

по одному объекту от каждого уровня условий или блока. В порядке компенсации недостаточно точного сбалансирования этот метод позволяет применить ковариационный анализ параграфа 5 этой главы. Если между вспомогательной переменной и основными результатами существует высокая корреляция, то этим путем может быть получено надлежащее увеличение эффективности такого опыта.

Лукас [5] приводит некоторые соображения в пользу того, что и при сбалансировании можно получить несмещенную (или примерно несмещенную) оценку ошибки опыта на основе применения ковариационного анализа, подобного анализу в параграфах 2 и 3 этой главы. Аргументом общего характера в данном случае является то, что получение точных данных при приближенной их оценке более важно, чем получение неточных данных при точной оценке их существенности. Сбалансированные группы при применении ковариационного анализа оказались более эффективными, чем случайное размещение при той же ковариации; это увеличение эффективности зависит от размера групп и от корреляции между вспомогательной переменной и основной переменной. При небольших группах (5 или менее объектов) относительная эффективность случайного размещения неожиданно оказалась очень небольшой. Это объясняется двумя обстоятельствами: 1) широкий размах варьирования X внутри каждого варианта приводит к улучшению оценки регрессии Y на X и 2) небольшие разности $\bar{x}_1 - \bar{x}_2$, возникающие в связи со сбалансированием X , уменьшают s_D^2 благодаря почти полному или даже полному элиминированию неточности, обусловленной ошибкой b .

Допущения, связанные с ковариацией в группах, будут ли они сбалансированы или не сбалансированы, являются до некоторой степени более жесткими, чем допущения при применении ковариации в рендомизированных блоках, и в особенности по сравнению со случаями, когда ковариация не применяется. Ковариация, о которой говорится здесь, предполагает линейность регрессии и однородность дисперсий. В дополнение к этому ковариация в указанных группах предполагает параллелизм регрессий, соответствующих этим группам. Если такого рода допущения не могут быть приняты, то более правильным будет переход от плана опыта с такого рода сбалансированным построением к рендомизированным блокам, несмотря на то, что это приведет к уменьшению числа степеней свободы для ошибки.

8. Ковариация в латинском квадрате. Описанные в предыдущем методы ковариационного анализа полностью применимы к опытам, в которых при схеме латинского квадрата производится наблюдение над двумя переменными. Для иллюстрации возьмем данные таблицы 178, извлеченные из опыта с соей в Гудзоне, штат Айова, в 1949 г. [2]. Фактическое построение опыта здесь изменено так, чтобы получить подходящий для наших целей пример, но выводы от этого не меняются (см. пример 17 этой главы).

Можно предполагать, что эти четыре линии сои в одинаковой мере подвержены заражению. Это позволяет считать, что X не влияет на варианты опыта — линии. Дисперсионный анализ X не дает никаких доказательств в пользу противного.

В самих расчетах в этом случае нет ничего нового. Ковариационный анализ таблицы 179 указывает на различия линий по их урожайности.

Приведенные средние будут такими: $A=23,77$; $B=27,53$; $C=25,20$; $D=24,88$ бушеля на акр. Здесь наблюдается только одно изменение в порядке расположений линий по сравнению с первоначальными средними.

Применение метода последовательной оценки разностей указывает на наличие только одной существенной разности между приведенными средними линий B и A .

Эффективность применения ковариации по отношению к прямому анализу урожая в данном случае составляет 460%.

Пример 17. Опыт с соей таблицы 178 фактически был проведен по схеме рендомизированных блоков, причем блоками были ряды в нашем примере. Фактически столбцов

Процент растений, пораженных раком стеблей, X и урожай Y (в бушелях на акр) четырех линий сои A , B , C и D . Схема латинского квадрата

Ряды	Столбцы				Итого
	1	2	3	4	
1-я линия					
X	C 4,3	A 19,3	B 10,1	D 14,0	47,7
Y	26,7	21,3	28,3	25,1	101,4
2-я линия					
X	A 29,2	B 34,7	D 30,2	C 48,2	142,3
Y	19,7	20,7	20,1	14,7	75,2
3-я линия					
X	B 14,0	D 7,2	C 6,3	A 1,0	28,5
Y	26,0	24,9	29,0	28,7	108,6
4-я линия					
X	D 8,9	C 6,7	A 6,4	B 5,6	27,6
Y	29,8	29,0	27,3	34,1	120,2
Итого:					
X	56,4	67,9	53,0	68,8	246,1
Y	102,2	95,9	104,7	102,6	405,4
	ΣX	ΣY	\bar{x}	y	
A	55,9	97,0	13,98	24,25	
B	64,4	109,1	16,40	27,28	
C	65,5	99,4	16,38	24,85	
D	60,3	99,9	15,08	24,98	

ТАБЛИЦА 179

Ковариационный анализ данных опыта с соей. Латинский квадрат

Источник варьирования	f	Σx^2	Σxy	Σy^2	Отклонения		
					f	Σd_{y-x}^2	средний квадрат
Ряды	3	2239,32	-747,97	272,93			
Столбцы	3	48,12	-14,64	10,80			
Варианты T	3	14,30	10,19	21,22			
Ошибка E	6	378,88	-131,04	55,23	5	9,91	1,98
$T+E$	9	393,18	-120,85	76,45	8	39,30	
Приведенные средние					3	29,39	9,80*

не было. В соответствии с этим измените обработку таблицы 179. Для оценки существенности вариантов вы получите $F=4,79^*$.

ЛИТЕРАТУРА

1. Cochran William G., Cox Gertrude M., Experimental Designs. John Wiley and Sons, Inc., New York, 1950.
2. Crall James M., Данные Айовской с.-х. опытной станции, 1949.
3. Finney D. J., Biometrics (Bulletin), 2, 53, 1946.
4. Jenkins M. T., Journal of Agricultural Research, 39, 677, 1929.
5. Lucas H. L., Institute of Statistics Mimeograph Series No. 18. University of North Carolina, 1951.
6. Nair K. R., Sankhya, 6, 167, 1942.
7. Snedecor George W., Culbertson C. C., Proceedings, American Society of Animal Production, p. 25, 1932.
8. Snedecor George W., Cox Gertrude M., Journal of Agricultural Research, 54, 449, 1937.
9. Sprague G. J., Данные Айовской с.-х. опытной станции, 1952.
10. Swanson Pearl P., et al. См. ссылку 12, гл. 14.

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ И КОВАРИАЦИЯ

1. Введение. Регрессия Y на одну независимую переменную (глава 6) часто бывает недостаточна для исследователя, и только две или три переменные X могут дать надлежащую информацию относительно изменения Y . В этом случае можно использовать многое из того, что говорилось по поводу одной независимой переменной X . Добавочное здесь состоит в том, что возникает необходимость в оценке той относительной роли, которую выполняют различные независимые переменные X при получении нами информации об Y .

Соединяя вместе нити предыдущих методов, множественная регрессия образует ткань весьма сложного рисунка. Начиная с единичной выборки (глава 2), это расширение идет как в направлении численности групп, так и числа переменных. Сначала мы рассмотрим вопрос о единичной выборке с тремя переменными, далее будет обращено внимание на группу выборок тоже с тремя переменными. Четыре и большее число переменных не требуют введения каких-либо новых принципов и только увеличивают размеры и сложность расчетов.

2. Две независимые переменные в единичной выборке. Увеличение числа независимых переменных с одной до двух неожиданным образом приводит к некоторому числу новых концепций в дополнение к тем, которые даны в главе 6. Вероятно, наиболее удивительной окажется необходимость пользоваться трехмерным пространством, в котором располагается график этих переменных. Здесь вместо геометрии на плоскости необходима геометрия в пространстве. Если Y частично зависит от X_1 и частично от X_2 , то для графического изображения требуется три взаимно-перпендикулярных оси координат. X_1 откладывается вдоль одной из них, X_2 располагается параллельно второй, а Y находится над плоскостью X_1X_2 на прямой, параллельной третьей оси. Точки такого пространства, фиксированные тремя значениями (X_1 , X_2 , Y), определяют *плоскость регрессии*.

В настоящей главе будет рассматриваться такая *плоскость* связи; ее уравнение:

$$\hat{Y} = \bar{y} + b_{Y1.2}(X_1 - \bar{x}_1) + b_{Y2.1}(X_2 - \bar{x}_2),$$

где $b_{Y1.2}$ и $b_{Y2.1}$ являются *коэффициентами частной регрессии*. Первый символ читается так: «регрессия Y на X_1 , независимая от X_2 ». Такое уравнение регрессии или плоскость определяется по наблюдаемым данным при помощи метода наименьших квадратов, особенность которого состоит в том, что получаемая при этом сумма квадратов расстояний фактических точек от плоскости будет минимальной.

Как и в главах 6 и 7, здесь применимы две модели. Одна из них относится к выборкам из нормальной совокупности *нескольких переменных*, например трех переменных X_1 , X_2 и Y , каждая из которых является случайной переменной. Во второй модели значения величин X *определенно фикси-*

рованы, т. е. выбраны исследователем, и только Y является случайной, нормально распределенной переменной.

Пример таблицы 180 взят нами из исследования о потребности растений в фосфоре на различных почвах штата Айова [5]. Содержание неорганического и органического фосфора в почвах было определено химическими мето-

ТАБЛИЦА 180

Неорганический фосфор X_1 , органический фосфор X_2 и оценка потребности растений в доступном фосфоре Y для 18 почв штата Айова при $20^\circ C$ (частей на миллион)

Почвенные пробы	X_1	X_2	Y	\hat{Y}	$Y - \hat{Y}$
1	0,4	53	64	61,6 ¹	2,4 ¹
2	0,4	23	60	59,0	1,0
3	3,1	19	71	63,4	7,6
4	0,6	34	61	60,3	0,7
5	4,7	24	54	66,7	-12,7
6	1,7	65	77	64,9	12,1
7	9,4	44	81	76,9	4,1
8	10,1	31	93	77,0	16,0
9	11,6	29	93	79,6	13,4
10	12,6	58	51	83,8	-32,8
11	10,9	37	76	79,0	-3,0
12	23,1	46	96	101,6	-5,6
13	23,1	50	77	101,9	-24,9
14	21,6	44	93	98,7	-5,7
15	23,1	56	95	102,4	-7,4
16	1,9	36	54	62,8	-8,8
17	26,8	58	168	109,2	58,8
18	29,9	51	99	114,2	-15,2
Сумма	215,0	758	1463	1463,0	0,0
Средняя	11,94	42,11	81,28		

$$\Sigma X_1^2 = 4321,02$$

$$C = 2568,06$$

$$\Sigma X_1 X_2 = 10139,50$$

$$C = 9053,89$$

$$\Sigma X_1 Y = 20706,20$$

$$C = 17474,72$$

$$\Sigma x_1^2 = 1752,96$$

$$\Sigma x_1 x_2 = 1085,61$$

$$\Sigma x_1 y = 3231,48$$

$$\Sigma X_2^2 = 35076,00$$

$$C = 31920,22$$

$$\Sigma X_2 Y = 63825,00$$

$$C = 61608,56$$

$$\Sigma Y^2 = 131299,00$$

$$C = 118909,39$$

$$\Sigma x_2^2 = 3155,78$$

$$\Sigma x_2 y = 2216,44$$

$$\Sigma y^2 = 12389,61$$

¹ Число десятичных знаков, сохраненных в предыдущих вычислениях, таково, что результаты этих двух столбцов определены с точностью $\pm 0,1$.

дами. Потребность растений в фосфоре устанавливалась путем выращивания на этих почвах кукурузы. Целью исследования являлось установление источника питательного вещества, используемого растением.

Знакомый способ расчетов в этой таблице приводит к скорректированным суммам квадратов и произведений. Согласно теории метода наименьших квадратов, два коэффициента частной регрессии определяются из такой системы двух нормальных уравнений:

$$\Sigma x_1^2 b_{Y1.2} + \Sigma x_1 x_2 \cdot b_{Y2.1} = \Sigma x_1 y;$$

$$\Sigma x_1 x_2 \cdot b_{Y1.2} + \Sigma x_2^2 \cdot b_{Y2.1} = \Sigma x_2 y.$$

Решение этих уравнений с помощью обычных алгебраических методов приводит к формулам:

$$b_{Y1.2} = \frac{(\Sigma x_1^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{D};$$

$$b_{Y2.1} = \frac{(\Sigma x_2^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{D},$$

где

$$D = (\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2.$$

Подставляя наши данные, находим:

$$D = 1752,96 \times 3155,78 - 1085,61^2 = 4\,353\,400;$$

$$b_{Y1.2} = \frac{3155,78 \times 3231,48 - 1085,61 \times 2216,44}{4\,353\,400} = 1,7898;$$

$$b_{Y2.1} = \frac{1752,96 \times 2216,44 - 1085,61 \times 3231,48}{4\,353\,400} = 0,0866.$$

Теперь уравнение множественной регрессии становится таким:

$$\begin{aligned} \hat{Y} &= 81,28 + 1,7898(X_1 - 11,94) + 0,0866(X_2 - 42,11) = \\ &= 56,26 + 1,7898X_1 + 0,0866X_2. \end{aligned}$$

Оно имеет следующий смысл: усвоение фосфора растением увеличивается в среднем на 1,7898 части на миллион на каждую добавочную одну миллионную часть неорганического фосфора и только на 0,0866 части на добавочную миллионную часть органического фосфора. Это указывает на то, что при температуре 20° неорганический фосфор является главным источником удовлетворения потребностей растения в фосфоре. Этот последний вывод выходит из рамок самого уравнения регрессии, которое дает только количественное выражение для тесноты связи между изменениями Y и соответствующими изменениями двух переменных X .

Моделью, более соответствующей данным условиям, является модель с фиксированными X . Хотя здесь и имеют место некоторые случайные моменты, определяющие значения X , тем не менее сами типы почв были специально отобраны, и в эти почвы были внесены определенные, специально установленные дополнительные количества питательных веществ. При предположении фиксированных значений X модель будет:

$$Y = \alpha + \beta_{Y1.2}(X_1 - \bar{x}_1) + \beta_{Y2.1}(X_2 - \bar{x}_2) + \varepsilon,$$

где α и β являются параметрами, подлежащими оценке, и ε из $N(0, \sigma)$. Здесь каждое Y состоит из частей, определяемых параметрами (совместно с фиксированными X) и случайным компонентом ε . Первая часть оценивается величиной \hat{Y} , т. е. точкой на выборочной плоскости регрессии, соответствующей X_1 и X_2 . Вопрос о случайной части будет рассмотрен в дальнейшем.

Два параметра β требуются для того, чтобы определить *направленность* плоскости регрессии; параметр же α определяет *высоту* этой плоскости на вертикальной оси координат. В опыте с фосфором число 81,28 является оценкой α , число 1,7898 — оценкой $\beta_{Y1.2}$ и число 0,0866 — оценкой $\beta_{Y2.1}$.

Для каждой почвы \hat{Y} является оценкой значения Y , определяемого регрессией в совокупности для X_1 и X_2 , соответствующих данной почве. Например, для почвы 1: $Y = 56,26 + 1,7898 \times 0,4 + 0,0866 \times 53 = 61,6$ части на миллион. Наблюдаемое значение $Y = 64$ частям на миллион отклоняется от этой оценки на $64 - 61,6 = 2,4$ части на миллион; это значит, что наблюдаемое значение Y лежит на 2,4 единицы выше плоскости регрессии. Это и является в данной модели выражением выборочного случайного варьирования. Все 18 значений \hat{Y} вычислены нами и выписаны в таблице 180. В последнем столбце даны отклонения $Y - \hat{Y}$; они дают меру того, что знание X недостаточно для точного предсказания значений Y .

Теперь исследователь сталкивается с вопросом об изучении этих отклонений от регрессии. В некоторой своей части они могут быть связаны с другими переменными, не включенными в данное исследование. Те или иные объяснения могут быть найдены для некоторых, в особенности больших, отклонений. Такого рода рассмотрение отдельных отклонений может представлять большую ценность для понимания результатов опыта или оно может послужить поводом для исключения одного или нескольких наблюдений и для пересчета после этого уравнения регрессии. Но исключение уклоняющихся наблюдений всегда является мероприятием сомнительного характера и не должно производиться без достаточного основания.

Часто основной задачей исследователя является вопрос об установлении относительной силы связи между Y и каждой из двух независимых переменных. Для решения этого вопроса нет единственного и безусловного пути, но некоторые приближения к нему все же возможны. Если выборочные стандартные отклонения для X_1 и X_2 одинаковы, то простым и обычно само собой напрашивающимся решением вопроса будет сравнение двух коэффициентов регрессии. Но чтобы исключить неравенство варьирования переменных X , следует определить и сравнить *стандартные коэффициенты частной регрессии*:

$$b'_{Y1.2} = b_{Y1.2} \sqrt{\frac{\Sigma x_1^2 / \Sigma y^2}{\Sigma x_2^2 / \Sigma y^2}} \quad \text{и} \quad b'_{Y2.1} = b_{Y2.1} \sqrt{\frac{\Sigma x_2^2 / \Sigma y^2}{\Sigma x_1^2 / \Sigma y^2}}$$

В опыте с фосфором это будет:

$$b'_{Y1.2} = 1,7898 \sqrt{1752,96/12389,61} = 0,67;$$

$$b'_{Y2.1} = 0,0866 \sqrt{3155,78/12389,61} = 0,04.$$

Здесь нет никакого сомнения в том, что неорганический фосфор дает большую возможность определить потребность растения в фосфоре. В последующих параграфах вы встретитесь с методами, которые приводят к более точным заключениям.

В связи с использованием модели регрессии с фиксированными X возникает вопрос о возможных ошибках при измерении этих X . Если такие ошибки имеют место, то как оценки параметров, так и критерии существенности испытывают смещение, т. е. становятся в известной мере искаженными. Величина этого смещения зависит от размера ошибки по отношению к размаху варьирования X . В дополнение к нашим ссылкам на [1] и [14] в главе 6 смотрите ссылку на [11].

Пример 1. Ниже приводится ряд из 20 наблюдений над тремя величинами, построенный так, чтобы облегчить вычисления:

X_1	X_2	Y	X_1	X_2	Y
29	2	22	21	2	24
1	4	26	12	1	7
5	3	23	24	3	23
31	1	8	16	3	28
25	3	25	6	4	25
16	1	12	20	2	22
26	1	13	35	1	25
15	4	30	9	4	32
6	2	12	19	4	37
10	3	26	14	2	20

Определите уравнение регрессии:

$$\hat{Y} = 0,35 X_1 + 7,16 X_2 - 1,9.$$

3. Интервальные оценки и критерии существенности. В соответствии с данной моделью значения ϵ получаются из $N(0, \sigma)$. Величина σ оценивается

на основе отклонений от выборочной регрессии $Y - \hat{Y} = d_{Y.12}$. Вычисляя по данным последнего столбца таблицы 180 способом нарастающего итога сумму квадратов, имеем $\Sigma d_{Y.12}^2 = 6414,46$. Применяя более простой и более точный метод вычислений, указанный ниже, получим более точный результат $\Sigma d_{Y.12}^2 = 6414,00$. Отсюда $s_{Y.12}^2 = 6414,00/15 = 427,6$; число степеней свободы здесь равно n минус число переменных. Вообще же одна степень свободы из общего их числа предназначается для \bar{y} и по одной для каждого из коэффициентов b .

Остаточная сумма квадратов $\Sigma y^2 - \Sigma d_{Y.12}^2 = 12389,6 - 6414,0 = 5975,6$ относится к регрессии. Она символически может быть обозначена $\Sigma \hat{y}_{12}^2$. Соотношение между этими суммами квадратов дано в таблице 181.

Нулевая гипотеза $\beta_{Y1.2} = \beta_{Y2.1} = 0$ проверяется при помощи критерия F . В данном случае производится оценка существенности регрессии в целом.

ТАБЛИЦА 181

Дисперсионный анализ данных о доступном для растений фосфоре

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	17	$\Sigma y^2 = 12\ 389,6$	
Регрессия	2	$\Sigma \hat{y}_{12}^2 = 5\ 975,6$	2 987,8
Отклонения	15	$\Sigma d_{Y.12}^2 = 6\ 414,0$	427,6

$$F = 2987,8/427,6 = 6,99; P < 0,01.$$

На практике сумма квадратов Y , относящаяся к регрессии, вычисляется следующим образом:

$$\Sigma \hat{y}_{12}^2 = b_{Y1.2} \Sigma x_1 y + b_{Y2.1} \Sigma x_2 y = 1,7898 \times 3231,48 + 0,0866 \times 2216,44 = 5975,6.$$

Значение же $\Sigma d_{Y.12}^2$ будет теперь разностью:

$$\Sigma y^2 - \Sigma \hat{y}_{12}^2 = 12389,6 - 5975,6 = 6414,0.$$

Этот способ расчетов в меньшей мере зависит от ошибок округлений, чем тот, который был применен ранее.

Доверительный интервал для α будет:

$$\bar{y} \pm t_{0,05} \cdot s_{Y.12} / \sqrt{n}; f = n - m,$$

где m — число переменных. Подставляя наши значения, получаем:

$$81,28 \pm 2,131 \sqrt{427,6/18}, \text{ или от } 70,9 \text{ до } 91,7 \text{ части на миллион.}$$

Для построения доверительных интервалов и оценки существенности коэффициентов регрессии необходимо определение двух *весовых коэффициентов*:

$$c_{11} = \Sigma x_2^2 / D = 3155,78 / 4353400 = 0,0007249;$$

$$c_{22} = \Sigma x_1^2 / D = 1752,96 / 4353400 = 0,0004027.$$

После этого вычисляются:

$$s_{b_{Y1.2}}^2 = s_{Y.12}^2 \cdot c_{11} = 427,6 \times 0,0007249 = 0,3100; s_{b_{Y1.2}} = 0,557;$$

$$s_{b_{Y2.1}}^2 = s_{Y.12}^2 \cdot c_{22} = 427,6 \times 0,0004027 = 0,1722; s_{b_{Y2.1}} = 0,415.$$

Отношение $(b_{Y1.2} - \beta_{Y1.2}) / s_{b_{Y1.2}}$ распределено, как t . Поэтому гипотеза $\beta_{Y1.2} = 0$ проверяется обычным способом:

$$t_1 = b_{Y1.2} / s_{b_{Y1.2}} = 1,7898 / 0,557 = 3,21^{**};$$

$$t_2 = b_{Y2.1} / s_{b_{Y2.1}} = 0,0866 / 0,415 = 0,21.$$

Отсюда следует, что в той совокупности, из которой взята выборка, доступный для растения фосфор лучше всего определяется через содержание в почве неорганического фосфора. Опыт показывает, «...что органический фосфор почвы *сам по себе* непригоден для растений. Вероятно, органический фосфор становится в достаточной мере доступным для растения только при его минерализации и что скорость минерализации его в опытах при 20°С слишком мала для того, чтобы иметь какое-либо количественное значение». Сравните это с данными параграфа 10 этой главы.

В порядке дальнейшего исследования может возникнуть потребность в определении роли только одного неорганического фосфора, оставляя в стороне органический фосфор. В этом случае может быть вычислен коэффициент регрессии Y на X_1 : $\Sigma x_1 y / \Sigma x_1^2 = 3231,48 / 1752,96 = 1,843$ вместо $b_{Y1.2} = 1,7898$. Здесь явственно выделяется тот факт, что как оценки параметров, так и значения критериев зависят от всех тех независимых переменных, которые включены в регрессию.

Во всякой регрессии эти оценки и критерии коррелированы между собой, поэтому в регрессии происходят более или менее сложные изменения, когда включаются или исключаются некоторые независимые переменные. В этом смысле положения, высказываемые относительно роли какой-либо независимой переменной в изменении зависимой переменной, не являются строго определенными; они зависят от других переменных, использованных для построения регрессии.

Вообще говоря, имеется другой, более общий путь решения вопроса об относительной роли независимой переменной в предсказании значений зависимой переменной. Он показан в таблице 182 и начинается с определе-

ТАБЛИЦА 182

Критерии для отдельных X после того, как влияние другого X исключено

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
X_1 и X_2	2	$\Sigma \hat{y}_{12}^2 = 5\,975,6$	
Только X_1	1	$\Sigma \hat{y}_1^2 = 5\,957,0$	
X_2 после X_1	1	18,6	18,6
X_1 и X_2	2	$\Sigma \hat{y}_{12}^2 = 5\,975,6$	
Только X_2	1	$\Sigma \hat{y}_2^2 = 1\,556,7$	
X_1 после X_2	1	4 418,9	4418,9**
Ошибка	15	6 414,0	427,6

ния суммы квадратов, относящейся к регрессии Y на X_1 и X_2 . После этого из нее вычитается сумма квадратов, относящаяся к регрессии Y на X_1 , вычисленная по формуле 4 параграфа 15 главы 6. Остаток является суммой квадратов включения X_2 после того, как влияние X_1 учтено. В опыте с фосфором этот остаток ничтожен. Далее из $\Sigma \hat{y}_{12}^2$ вычитается сумма квадратов, относящаяся к регрессии Y на X_2 . Эта в нашем случае вполне существенная остаточная сумма квадратов относится к X_1 , включенному в регрессию после того, как принято во внимание влияние X_2 . Нетождественность результатов особенно видна при сравнении сумм квадратов, относящихся к X_2 , взятому отдельно, и к X_2 , включенному после X_1 . Этот факт знакомит нас с тем положением, что заключения по поводу роли отдельных факторов зависят в известной мере от порядка, в котором применяются соответствующие

щие критерии. Эти критерии уже не будут ортогональными, а являются взаимозависимыми.

Обобщая только что описанный метод, можно сказать: когда имеется несколько независимых переменных, то некоторая *часть их* может быть отброшена, если их общая доля в сумме квадратов мала. Иллюстрация этого правила будет дана в параграфе 9 этой главы.

При элиминировании только одного X критерий F таблицы 182 становится идентичным с критерием t , как это подтверждается сравнением $\sqrt{\bar{F}} = \sqrt{4418,9/427,6} = 3,21$ с соответствующим t_1 , вычисленным ранее. Это проливает свет на смысл проверки нулевой гипотезы $\beta_{Y1.2} = 0$ при помощи t_1 ; данная гипотеза то же самое, что и гипотеза о том, что после учета влияния X_2 на Y не остается никакой остаточной суммы квадратов \bar{Y} (если оставить в стороне ошибку опыта). Эта тождественность критериев позволяет применять различные методы оценки; может оказаться, что более удобно применить способ таблицы 182, чем вычислять весовые коэффициенты, требуемые для критерия t . При трех и более переменных X мы пока, до знакомства с параграфами 10 и 11 этой главы, не даем такого же способа решения задачи.

Для определения доверительного интервала значения Y , относящегося к регрессии в совокупности, необходим третий весовой коэффициент $c_{12} = -\Sigma x_1 x_2 / D = -1085,61 / 4353400 = -0,0002494$.

Доверительный интервал для Y , соответствующего (X_1, X_2) , будет:

$$\hat{Y} \pm t_{0,05} \cdot s_{Y.12} \sqrt{(1/n + c_{11}x_1^2 + c_{22}x_2^2 + 2c_{12}x_1x_2)}.$$

Пример: Для значения Y в точке $X_1 = 4,7$ и $X_2 = 24$, соответствующей в таблице 180 почвенному образцу 5, имеем $x_1 = 4,7 - 11,9 = -7,2$ и $x_2 = 24 - 42 = -18$; отсюда доверительный интервал:

$$66,7 \pm 2,131 \times \sqrt{427,6} \times \sqrt{[1/18 + 0,0007249 \times (-7,2)^2 + 0,0004027 \times (-18)^2 + (-0,0002494) \times (-7,2) \times (-18)]} = 66,7 \pm 17,6 \text{ части на миллион.}$$

Некоторые сведения относительно линейности регрессии можно получить путем построения точечной диаграммы сначала для отклонений $d_{Y.12}$ (приведенных в последней колонке таблицы 180) в соответствии с X_1 , а потом для тех же отклонений в соответствии с X_2 . Если регрессия фактически не является линейной, то на одной или на обоих этих графиках вместо горизонтальной прямой линии будет обнаружена кривизна линии, около которой расположены точки, соответствующие наблюдениям. (По поводу криволинейной множественной регрессии см. пример 13 этой главы.)

Пример 2. Опуская X_2 , проведите дисперсионный анализ для регрессии Y на X_1 . Вы придете к таким средним квадратам:

регрессия	1	5927	
отклонения	16	402	$F = 14,82$.

Это отличается от того, что было получено в таблице 181. По сравнению с X_1 сведения, которые дает нам X_2 , столь незначительны, что это приводит даже к уменьшению F при включении в регрессию X_2 . Объясните механизм этого уменьшения.

Пример 3. Опуская X_1 , проведите дисперсионный анализ для регрессии Y на X_2 :

регрессия	1	1557	
отклонения	16	677	$F = 2,30$.

Отметьте размер суммы квадратов, относящейся к X_2 в отсутствии X_1 . Это кажущееся противоречие обусловлено корреляцией между X_1 и X_2 . В этом смысле распределения двух переменных X_1 и X_2 не являются независимыми. Вычислите $r_{12} = 0,46$.

Пример 4. Постройте 95%-ные доверительные интервалы для $\beta_{Y1.2}$ и $\beta_{Y2.1}$. *Ответ:* от 0,60 до 2,98 и от $-0,79$ до 0,97.

Пример 5. Вычислите по таблице 180 корреляцию между Y и \hat{Y} . Вы получите примерно $R = 0,69$. Этот показатель называется *коэффициентом множественной корреляции*. Как вы видите, он дает меру подбора к фактическим точкам плоскости регрессии.

Пример 6. Вычислите R по формуле $R^2 = \hat{\Sigma y}_{12}^2 / \Sigma y^2 = 0,4823$. Это показывает, что R^2 является долей общей суммы квадратов Y , относящейся к регрессии. В этом смысле R^2 является мерой того, сколь удовлетворительна подобранная регрессия.

Пример 7. Вычислите $1 - R^2 = \Sigma d_{Y.12}^2 / \Sigma y^2 = 0,5177$. Величина $1 - R^2$ составляет долю Σy^2 , измеряющую отклонения от регрессии, т. е. дает меру недостаточности этой регрессии для объяснения варьирования Y .

ТАБЛИЦА 183

Начальный возраст (X_1), начальный вес (X_2) и привес (Y) для 40 свиней.
Четыре варианта с группами одинакового размера

Варианты	Начальный возраст X_1 (в днях)	Вес X_2 (в фунтах)	Привес Y (фунтов в день)	Суммы			Средние		
				ΣX_1	ΣX_2	ΣY	\bar{x}_1	\bar{x}_2	\bar{y}
1	78	61	1,40	799	544	14,64	79,9	54,4	1,46
	90	59	1,79						
	94	76	1,72						
	71	50	1,47						
	99	61	1,26						
	80	54	1,28						
	83	57	1,34						
	75	45	1,55						
	62	41	1,57						
	67	40	1,26						
2	78	74	1,61	784	550	13,19	78,4	55,0	1,32
	99	75	1,31						
	80	64	1,12						
	75	48	1,35						
	94	62	1,29						
	91	42	1,24						
	75	52	1,29						
	63	43	1,43						
	62	50	1,29						
	67	40	1,26						
3	78	80	1,67	749	520	14,45	74,9	52,0	1,44
	83	61	1,41						
	79	62	1,73						
	70	47	1,23						
	85	59	1,49						
	83	42	1,22						
	71	47	1,39						
	66	42	1,39						
	67	40	1,56						
	67	40	1,36						
4	77	62	1,40	750	495	13,25	75,0	49,5	1,35
	71	55	1,47						
	78	62	1,37						
	70	43	1,15						
	95	57	1,22						
	96	51	1,48						
	71	41	1,31						
	63	40	1,27						
	62	45	1,22						
	67	39	1,36						
Итого				3082	2109	55,53	77,0	52,7	1,39

4. Множественная ковариация в полностью рендомизированном опыте.
На случай множественной ковариации при наличии двух или нескольких

независимых переменных можно распространить методы параграфов 2, 3 и 4 главы 13, причем при добавлении новых переменных не происходит никаких изменений ни в модели, ни в теории вопроса. Этот метод показан в таблице 183 на примере среднего суточного привеса свиней. Вероятно, на привесе свиней в какой-то мере сказались возраст и исходный вес каждой свиньи до начала опыта. Если это так, то при помощи ковариации можно улучшить как оценки влияния вариантов, так и оценку ошибки опыта. Более того, так как из дисперсионного анализа X_1 и X_2 видно сбалансирование их влияний, то можно ожидать, что ковариация даст вполне надежную оценку ошибки опыта (параграф 7 главы 13).

Вычисление общих сумм квадратов и произведений проводится по образцу таблицы 180, т. е. не обращая внимания на подразделения свиней на группы. Значения этих сумм с поправками на средние внесены в первую строку таблицы 184.

ТАБЛИЦА 184

Расчет сумм квадратов и произведений для множественной ковариации в полностью рендомизированном опыте. Данные о привесе свиней из таблицы 183

	Σx_1^2	$\Sigma x_1 x_2$	$\Sigma x_1 y$
Общее	4 735,90	3 037,55	6,9235
Варианты	187,70	160,15	1,3005
Внутри групп	4 548,20	2 877,40	5,6230
	Σx_2^2	$\Sigma x_2 y$	Σy^2
Общее	5 065,98	27,5408	1,0228
Варианты	189,08	1,3218	0,1776
Внутри групп	4 876,90	26,2190	0,8452

Вычисление этих сумм для Y и двух X по вариантам производится обычным порядком. Два примера:

$$\Sigma x_1^2 = \frac{799^2 + 784^2 + 749^2 + 750^2}{10} - \frac{3082^2}{40} = 187,70;$$

$$\Sigma x_1 x_2 = \frac{799 \times 544 + \dots + 750 \times 495}{10} - \frac{3082 \times 2109}{40} = 160,15.$$

Теперь переходим к вычислению двух коэффициентов b , $\Sigma \hat{y}_{12}^2$ и $\Sigma d_{Y.12}^2$ сначала «в целом», а после «внутри групп», для чего дважды применяем в соответствующей последовательности формулы параграфов 2 и 3 этой главы. В результате получаем:

$$\text{В целом: } b_{Y1.2}^* = -0,0032903; b_{Y2.1} = 0,0074093; \Sigma \hat{y}_{12}^2 = 0,1803;$$

$$\Sigma d_{Y.12}^2 = 0,8415.$$

$$\text{Внутри групп: } b_{Y1.2} = -0,0034542; b_{Y2.1} = 0,0074142; \Sigma \hat{y}_{12}^2 = 0,1750;$$

$$\Sigma d_{Y.12}^2 = 0,6702.$$

Эти суммы $\Sigma d_{Y.12}^2$ и внесены в таблицу 185, в результате образуется дисперсионный анализ для отклонений от регрессии.

Этот анализ подобен анализу параграфа 4 главы 13. В его основе лежит допущение, что регрессии в совокупности линейны и параллельны при общей σ , характеризующей нормальное и независимое распределение ε . При вычислении b , $\Sigma \hat{y}_{12}^2$ и $\Sigma d_{Y.12}^2$ по каждому из четырех вариантов и для общей

Ковариационный анализ привеса свиней. Отклонения от регрессии

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
В целом	37	0,8415	
Ошибка (внутри групп) . . .	34	0,6702	0,0197
Для оценки приведенных средних по вариантам . . .	3	0,1713	0,0571*

регрессии («ошибка» в таблице 185) следует руководствоваться методом параграфа 3 главы 13. В этом случае анализ будет подобен анализу таблицы 173.

Приведенные средние по вариантам оцениваются в таблице 185 и вычисляются по формуле:

$$\bar{y}_{Ai} = \bar{y}_i - b_{Y1.2} (\bar{x}_{1i} - \bar{x}_{1.}) - b_{Y2.1} (\bar{x}_{2i} - \bar{x}_{2.}),$$

где подстрочный указатель i обозначает группу и b — коэффициент регрессии по строке «ошибка». Так, для четвертой группы:

$$\begin{aligned} \bar{y}_{A4} &= 1,35 - (-0,0034543) (75,0 - 77,0) - 0,0074142 (49,5 - 52,7) = \\ &= 1,35 - 0,0069 + 0,0237 = 1,37 \text{ фунта в день.} \end{aligned}$$

В данном опыте имеются только небольшие изменения при переходе от \bar{y}_i к \bar{y}_{Ai} потому, что групповые средние X_1 и X_2 очень близки к соответствующим средним по всему опыту в целом. Все четыре приведенные средние равны:

$$1,46, \quad 1,31, \quad 1,44 \text{ и } 1,37 \text{ фунта в день.}$$

Для дальнейших оценок и для проверки существенности необходимо вычисление трех весовых коэффициентов по строке «ошибка», которое проводится, как указано в параграфе 3 этой главы:

$$c_{11} = 0,0003508; \quad c_{22} = 0,0003272; \quad c_{12} = -0,0002070.$$

Средний квадрат для приведенной средней Y будет:

$$s_{\bar{y}_{Ai}}^2 = s_{Y.12}^2 [1/n + c_{11}(\bar{x}_{1i} - \bar{x}_{1.})^2 + c_{22}(\bar{x}_{2i} - \bar{x}_{2.})^2 + 2c_{12}(\bar{x}_{1i} - \bar{x}_{1.})(\bar{x}_{2i} - \bar{x}_{2.})].$$

Для четвертой группы:

$$\begin{aligned} s_{\bar{y}_{A4}}^2 &= 0,0197 \times [1/10 + 0,0003508 \times (75,0 - 77,0)^2 + 0,0003272 \times (49,5 - 52,7)^2 + \\ &+ 2 \times (-0,0002070) \times (75,0 - 77,0) \times (49,5 - 52,7)] = 0,00201. \end{aligned}$$

$$s_{\bar{y}_{A4}} = 0,045 \text{ фунта в день.}$$

Это позволяет определить доверительный интервал для приведенной средней в четвертой группе: $1,37 \pm 2,032 \times 0,045$, т. е. от 1,28 до 1,46 фунта в день.

При проведении заранее планируемых сравнений разность между двумя приведенными средними определяется так:

$$D = \bar{y}_i - \bar{y}_j - b_{Y1.2} (\bar{x}_{1i} - \bar{x}_{1j}) - b_{Y2.1} (\bar{x}_{2i} - \bar{x}_{2j}),$$

а ее средний квадрат будет:

$$s_D^2 = s_{Y.12}^2 [2/n + c_{11}(\bar{x}_{1i} - \bar{x}_{1j})^2 + c_{22}(\bar{x}_{2i} - \bar{x}_{2j})^2 + 2c_{12}(\bar{x}_{1i} - \bar{x}_{1j})(\bar{x}_{2i} - \bar{x}_{2j})].$$

Допустим для иллюстрации, что было запланировано сравнение средних привесов в группах 1 и 2.

$$D = 1,46 - 1,32 - (-0,003454) \times (79,9 - 78,4) - 0,007414 \times (54,4 - 55,0) = 0,15 \text{ фунта в день.}$$

(Эту разность можно получить и непосредственно из приведенных средних, если они были вычислены заранее, как это было сделано выше.)

$$s_D^2 = 0,0197 \left[\frac{1}{10} + 0,0003508 \times (79,9 - 78,4)^2 + 0,0003272 \times (54,4 - 55,0)^2 + 2 \times (-0,0002070) \times (79,9 - 78,4) \times (54,4 - 55,0) \right] = 0,00395.$$

Отсюда 95%-ный доверительный интервал для этой разности будет:

$$0,15 \pm 2,032 \sqrt{0,00395}; \text{ от } 0,02 \text{ до } 0,28 \text{ фунта в день.}$$

В этом опыте не планировались сравнения между определенными вариантами. Для оценки всех разностей требуется вычисление усредненной стандартной ошибки:

$$\begin{aligned} \overline{s_{y \cdot 12}^2} &= \frac{s_{\bar{Y} \cdot 12}^2}{n} \left[1 + \frac{c_{11} (\sum x_1^2)_T + c_{22} (\sum x_2^2)_T + 2c_{12} (\sum x_1 x_2)_T}{a-1} \right] = \\ &= \frac{0,0197}{10} \left[1 + \frac{0,0003508 \times 187,70 + 0,0003272 \times 189,08 + 2(-0,0002070) \times 160,15}{4-1} \right] = \\ &= 0,00201; \quad \overline{s_{y \cdot 12}} = 0,0448. \end{aligned}$$

Оценка всех разностей проводится теперь, как обычно:

Группы	\bar{x}_i	$\bar{x}_i - \bar{x}_2$	$\bar{x}_i - \bar{x}_4$	$\bar{x}_i - \bar{x}_3$
1	1,46	0,15 (0,13)	0,09 (0,16)	0,02 (0,17)
3	1,44	0,13 (0,16)	0,07 (0,17)	
4	1,37	0,06 (0,17)		
2	1,31			

Только одна самая большая разность оказалась существенной.

В этом опыте эффективность применения ковариации по отношению к ошибке опыта, рассчитанной по непосредственным привесам, составляет

$$s_D^2 / \overline{s_{y \cdot 12}^2} = 0,00235 / 0,00201 = 117\%.$$

Дисперсионный анализ для X_1 , X_2 , и Y таблицы 184 заставляет подозревать, что свиньи не были случайно распределены по группам, хотя можно считать, что это оказало на Y небольшое влияние. Фактически группы были сбалансированы по начальному весу и возрасту. Учитывая положения параграфа 7 главы 13, есть основание считать, что ковариация даст правильные оценки и критерии. По причине этого сбалансирования приведенные средние только немного изменились по отношению к первоначальным групповым средним, а усредненный средний квадрат средних разностей только немного больше $2s_{\bar{Y} \cdot 12}^2 / n$. В этом заключается положительная сторона такого сбалансирования.

Наблюдательный исследователь, пользующийся множественной регрессией, почти наверное выразит недоумение по поводу такого факта: регрессия Y на X_1 , взятая сама по себе, рассматриваемая наряду с другими переменными X , может полностью отличаться от той же самой регрессии Y

на X_1 при другом ряде переменных X ; может оказаться даже так, что изменится знак регрессии на обратный. Рассмотрим, например, регрессию Y на X_1 (возраст) в опыте со свиньями. По строке «в целом» (т. е. принимая во внимание группировку животных) коэффициент регрессии будет:

$b_{Y1} = 6,9235/4735,90 = 0,001462$ фунта в день на день возраста. Сравните это с $b_{Y1.2} = -0,003290$, вычисленным ранее также по строке «в целом». Почему средний привес в день увеличивается с возрастом в первом случае и уменьшается во втором?

Первая регрессия определяет влияние в целом, не учитывая специально начальный вес. В данной выборке имеет место некоторая тенденция к более быстрому росту свиней с большим начальным возрастом. Но среди животных одного и того же начального веса (начальный вес «удерживается неизменным») более взрослые свиньи растут медленнее. По этой причине более взрослые, но медленно растущие животные имеют вес не больший, чем более молодые быстрорастущие особи.

Эти факты можно наблюдать в таблице 186. В правом столбце здесь даны как начальный возраст, так и привесы, увеличивающиеся по мере увеличения начального веса; оба эти показателя находятся между собой в прямой связи, так как они оба связаны с начальным весом.

ТАБЛИЦА 186

Данные о 40 свиньях (табл. 183), сгруппированные по начальному весу

Начальный вес	Число свиней	Начальный возраст и средний суточный привес	Итого	Средняя
39—44	13	62 67 91 63 67 83 66 67	904 17,38	69,5 1,34
		1,57 1,26 1,24 1,43 1,26 1,22 1,39 1,56		
		67 70 71 63 67 1,36 1,15 1,31 1,27 1,36		
45—49	5	75 75 70 71 62	353	70,6
		1,55 1,35 1,23 1,39 1,22	6,74	1,35
50—54	5	71 80 75 62 96	384	76,8
		1,47 1,28 1,29 1,29 1,48	6,81	1,36
55—59	5	90 83 85 71 95	424	84,8
		1,79 1,34 1,49 1,47 1,22	7,31	1,46
60—64	8	78 99 80 94 83 79 77 78	668	83,5
		1,40 1,26 1,12 1,29 1,41 1,73 1,40 1,37	10,98	1,37
74—80	4	78 94 99 78	349	87,2
		1,61 1,72 1,34 1,67	6,31	1,58
Итого	40		3082 55,53	77,05 1,388

Но внутри этих рядов таблицы, где начальный вес меняется в узких пределах, существует тенденция обратного характера. Более взрослые свиньи дают, как правило, меньший привес. График, построенный по таким рядам, дает более наглядное представление об этом, а таблица 187 дает соответствующие точные показатели. В последней строке этой таблицы дана *общая регрессия*, которая оказалась отрицательной. Эта средняя регрессия только немногим отличается от средней регрессии $b_{Y2.1} = -0,003290$, оценивающей то же самое влияние фактора, а именно регрессию среднесуточного привеса на начальный возраст такой совокупности свиней, в которой все животные имеют один и тот же начальный вес.

Тем же путем можно подразделить свиней на группы, каждая из которых будет состоять из животных, имеющих приблизительно одинаковый

Ковариационный анализ по группам свиней одного и того же веса

Группы по весу	Число степеней свободы	Суммы квадратов и произведений			Регрессия Y на X_1
		Σx_1^2	Σx_{1Y}	ΣY^2	
39—44	12	831,2308	-6,1885	0,1917	-0,007445
45—49	4	113,2000	2,0860	0,0729	0,018428
50—54	4	634,8000	2,5720	0,0427	0,004052
55—59	4	324,8000	-0,6480	0,1819	-0,001995
60—64	7	486,0000	-3,6700	0,2140	-0,007551
74—80	3	354,7500	-3,3375	0,1015	-0,009408
Общее	34	2744,7808	-9,1860	0,8047	-0,003347

возраст. Обобщенная регрессия внутри таких групп будет примерно та же самая, что и $b_{Y2.1} = 0,007409$, и будет отличаться от $b_{Y2} = 27,5408/5065,98 = 0,005436$. При построении множественной регрессии мы соединяем в одном уравнении независимое участие X_1 и X_2 в изменении зависимой переменной Y .

5. Множественная ковариация в таблице с двумя входами. В таблице с двумя входами может быть представлена связь Y с двумя или несколькими независимыми переменными. Например, при оценке урожая может иметь значение не только число растений на делянку, но и распространение по площади какого-либо вредителя. В таких случаях, если не считать усложнение расчетов, в методах обработки нет ничего нового, что следовало бы добавить к параграфу 5 главы 13.

В качестве иллюстрации мы взяли данные опыта [1, 14], проведенного в Великобритании в период с 1932 по 1937 г. Конечной целью исследования было определение, в какой мере можно предугадать урожай пшеницы путем выборочного подсчета числа растений. На ранней стадии эксперимента представлялось, что надлежащая информация более всего заключена в двух переменных: высота растения в момент колошения X_1 и число растений в период кущения X_2 . Данные по этим показателям и урожай приведены в таблице 188.

Для изучения этого вопроса необходимо, чтобы соответствующие статистические показатели были освобождены от влияния места и времени наблюдения. При допущении, что X фиксированы, модель, которая в этом случае будет принята, в терминах выборочных показателей может быть записана так:

$$Y_{ij} = \bar{y}_{..} + S_i + P_j + b_{Y1.2} (X_{1ij} - \bar{x}_{1..}) + b_{Y2.1} (X_{2ij} - \bar{x}_{2..}) + e,$$

где S и P константы, относящиеся к времени и месту наблюдения. Для элиминирования этих констант по исходным урожайным данным можно построить модель:

$$\hat{Y}_{ij} = \bar{y}_{..} + S_i + P_j$$

и после этого новую переменную

$$\hat{Y}'_{ij} = Y_{ij} - \hat{Y}_{ij}$$

можно использовать для проведения ковариационного анализа. Этот результат должен быть таким же, как и регрессия, вычисленная по суммам квадратов и произведений по строке «ошибка» в таблице 188:

$$b_{Y1.2} = 1,3148, b_{Y2.1} = -0,6466, \Sigma \hat{y}'_{i2}^2 = 200,59, \Sigma d_{Y.12}^2 = 28,07.$$

Высота стебля при колошении (X_1), число растений в период кущения (X_2) и урожай (Y) пшеницы в Великобритании

(X_1 — в дюймах; X_2 — число растений на 1 фут; Y — в центнерах на 1 акр)

Годы		Пункты						Суммы по годам
		Сил Хейн	Ротам-стед	Ньюпорт	Богхолл	Спрау-стон	Плам-тон	
1933	X_1	25,6	25,4	30,8	33,0	28,5	28,0	171,3
	X_2	14,9	13,3	4,6	14,7	12,8	7,5	67,8
	Y	19,0	22,2	35,3	32,8	25,3	35,8	170,4
1934	X_1	25,4	28,3	35,3	32,4	25,9	24,2	171,5
	X_2	7,2	9,5	6,8	9,7	9,2	7,5	49,9
	Y	32,4	32,2	43,7	35,7	28,3	35,2	207,5
1935	X_1	27,9	34,4	32,5	27,5	23,7	32,9	178,9
	X_2	18,6	22,2	10,0	17,6	14,4	7,9	90,7
	Y	26,2	34,7	40,0	29,6	20,6	47,2	198,3
Суммы по пунктам	X_1	78,9	88,1	98,6	92,9	78,1	85,1	521,7
	X_2	40,7	45,0	21,4	42,0	36,4	22,9	208,4
	Y	77,6	89,1	119,0	98,1	74,2	118,2	576,2
		Скорректированные суммы квадратов и произведений						
		Σx_1^2		$\Sigma x_1 x_2$		$\Sigma x_1 y$		
В целом		230,53		—0,65		341,25		
Пункты		106,34		—47,06		190,83		
Годы		6,26		26,24		8,41		
Ошибка		117,93		20,17		142,01		
		Σx_2^2		$\Sigma x_2 y$		Σy^2		
В целом		385,07		—300,75		982,30		
Пункты		171,46		—257,03		629,22		
Годы		139,41		—22,26		124,42		
Ошибка		74,20		—21,46		228,66		

Эти показатели совместно с некоторыми другими, взятыми из таблицы, приводят к следующим результатам.

а) При освобождении от влияния места и времени наблюдений данные о высоте стебля и числе растений определяют:

$$\Sigma \hat{y}_{12}^2 / \Sigma y^2 = 200,59 / 228,66 = 88\%$$

суммы квадратов, относящейся к урожаю.

б) Точность определения значений Y по этим двум независимым переменным характеризуется следующим анализом $\Sigma \hat{y}^2$:

Регрессия на X_1 и X_2 2 200,59

Регрессия только на X_1 1 171,01

X_2 после X_1 1 29,58*

Регрессия только на X_2	1	6,21	
X_1 после X_2	1	194,38**	
Ошибка	8	28,07	3,51

Хотя каждая из переменных дает существенное уменьшение Σy^2 , все же высота стебля в период колошения является более эффективным фактором.

в) Средний урожай по всему опыту в целом $576,2/18=32,0$ ц на 1 акр имеет доверительный интервал от 31 до 33 ц на 1 акр.

г) Уравнение регрессии по строке «ошибка»:

$$\hat{Y} = 1,393 + 1,3148X_1 - 0,6466X_2$$

Путем подстановки каждой пары значений X вычислены \hat{Y} и отклонения $Y - \hat{Y}$, которые приведены в таблице 189. Этим путем устанавливается влияние места и года проведения опыта. Например: 1) средний урожай в 1933 г. дал отрицательную прибавку $-19,4/6=-3,2$ ц на 1 акр по сравнению со средним урожаем по всему опыту в целом; 2) средняя урожайность в Пламптоне на $16,9/3=5,6$ ц на 1 акр выше общего урожая.

ТАБЛИЦА 189

Фактические и вычисленные урожаи пшеницы

Пункты	1933 г.			1934 г.			1935 г.			Сумма
	Y	\hat{Y}	$Y - \hat{Y}$	Y	\hat{Y}	$Y - \hat{Y}$	Y	\hat{Y}	$Y - \hat{Y}$	
Сил Хейн	19,0	25,4	-6,4	32,4	30,1	2,3	26,2	26,0	0,2	-3,9
Ротамстед	22,2	26,2	-4,0	32,2	32,5	-0,3	34,7	32,3	2,4	-1,9
Н юпорт	35,3	38,9	-3,6	43,7	43,4	0,3	40,0	37,7	2,3	-1,0
Богхолл	32,8	35,3	-2,5	35,7	37,7	-2,0	29,6	26,2	3,4	-1,1
Спраустон	25,3	30,6	-5,3	28,3	29,5	-1,2	20,6	23,2	-2,6	-9,1
Пламптон	35,8	33,4	2,4	35,2	28,4	6,8	47,2	39,5	7,7	16,9
Сумма			-19,4			5,9			13,4	-0,1

Если бы этот опыт был построен по схеме рандомизированных блоков, то для оценки влияния «пунктов» следовало бы применить критерий F , руководствуясь таким порядком:

1. Сложить попарно суммы квадратов и произведений, стоящих в таблице 188 по строкам «пункты» и «ошибка»:

$$\Sigma x_1^2 = 224,27 \quad \Sigma x_1 x_2 = -26,89 \quad \Sigma x_1 y = 332,84$$

$$\Sigma x_2^2 = 245,66 \quad \Sigma x_2 y = -278,49 \quad \Sigma y^2 = 857,88$$

2. Для этой регрессии « $P+E$ » вычисляется $\Sigma d_{Y.12}^2 = 129,20$.

3. По показателям для « $P+E$ » и « E » строится таблица отклонений от регрессии:

$P + E$	13	129,20	
E	8	28,08	3,51

Для оценки «пунктов» 5 101,12 20,22

Оценка разностей между приведенными средними по пунктам основывается на $F = 20,22/3,51 = 5,76$, откуда $P = 0,02$.

Пример 8. По данным об урожае пшеницы в Великобритании вычислите $c_{11} = 0,008893$ и оцените существенность $b_{Y.1.2}$. Ответ: $t = 7,45^{**} = \sqrt{F} = \sqrt{194,38/3,51}$.

Пример 9. Вычислите для регрессии по строке «ошибка» $R^2 = 0,8772$.

Пример 10. Для той же строки «ошибка» вычислите стандартные коэффициенты частной регрессии для сравнительной способности показателей высоты стебля и числа растений давать представление об урожае. *Ответ:* 0,94 и 0,37.

6. Частная корреляция. Кроме случаев, соответствующих предыдущей модели, когда X фиксированы, возможны случаи, когда эти X являются случайными переменными. Здесь рассматривается случайная выборка, взятая из *нормальной совокупности нескольких переменных*, в которой каждая переменная получает свое значение в случайном порядке. Если, например, выборка состоит из 18-летних студентов первого курса, то случайными переменными могут быть: рост, способности, материальное положение, рост родителей и пр. Здесь мы приходим к расширению понятия о выборке, рассмотренного в главе 7, и в связи с этим можно было бы изучить, например, регрессию роста сыновей на рост родителей. В этих условиях применимы все методы предыдущих параграфов. Но предметом изучения в таких случаях может быть и установление корреляций между такими переменными, как рост, вес, основной обмен и пр., без твердого выбора того, какая переменная является зависимой и какая — независимой. В этом случае вместо множественной регрессии более подходящим методом будет метод *частной корреляции*.

При наличии трех переменных существует три *простые*, или *общие*, корреляции между ними: Q_{12} , Q_{13} , Q_{23} . Коэффициент частной корреляции $Q_{12.3}$ характеризует корреляцию между переменными 1 и 2 в той части объектов, в которой эти последние имеют одно и то же значение переменной 3; эта третья переменная *удерживается на постоянном уровне*, в связи с чем только переменные 1 и 2 находятся в корреляционном отношении друг с другом. В этой модели считается, что $Q_{12.3}$ имеет одну и ту же величину при каждом отдельном значении переменной 3.

В выборках из совокупности с тремя переменными оценки $Q_{12.3}$ при различных значениях переменной 3, конечно, будут различаться между собой. Поэтому оценка $Q_{12.3}$, взятая в целом, должна быть некоторой усредненной величиной этих частных оценок. Эта средняя — выборочная частная корреляция определяется по формуле:

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}.$$

В формуле производится исключение влияния переменной 3 из корреляции между переменными 1 и 2.

Оценка существенности частных корреляций производится на основе таблицы 57, но данные из нее берутся в соответствии с $f = n - m$ степенями свободы, где n — число наблюдений каждой переменной и m — число переменных.

В штатах Айова и Небраска была произведена случайная выборка 142 пожилых женщин для изучения вопросов питания населения [12]. Было взято три переменных: возраст, давление крови и содержание в крови холестерина. Получены три общие корреляции: $r_{AB} = 0,3332$; $r_{AC} = 0,5029$ и $r_{BC} = 0,2495$. Так как высота кровяного давления связана с избытком в стенках сосудов холестерина, то представляется интересным установить корреляцию r_{BC} . Но очевидно, что B и C увеличиваются с возрастом. Возникает ли эта корреляция между B и C просто потому, что они имеют общую связь с возрастом, или между ними существует определенная связь в каждом возрасте? Влияние возраста можно исключить путем вычисления корреляции:

$$r_{BC.A} = \frac{0,2495 - 0,3332 \times 0,5029}{\sqrt{(1 - 0,3332^2) \times (1 - 0,5029^2)}} = 0,1233.$$

При $f = 142 - 3 = 139$ корреляция не является существенной. Это позволяет считать, что для каждого определенного возраста давление крови и содержа-

ние холестерина в крови некоррелированы. Во всяком случае данная выборка не столь велика, чтобы могла обнаружить эту корреляцию, даже если она имеет место.

Частная корреляция имеет непосредственную связь с регрессией. Представляется возможным рассматривать как кровяное давление, так и содержание холестерина в крови в качестве переменных, зависящих от возраста. Но (как это установлено выше) сами по себе они не зависят друг от друга. Несколько иное положение можно иллюстрировать примером с наблюдениями над усвоенным организмом протеина (P) и жира (F) у 54 пожилых женщин штата Айова. Общие корреляции в этом случае были:

$$r_{AP} = -0,4865, \quad r_{AF} = -0,5296 \quad \text{и} \quad r_{PF} = 0,5784.$$

Третья корреляция указывает на то, что протеин и жир встречаются вместе во всех диетах, но первые две корреляции устанавливают уменьшение того и другого питательного вещества с повышением возраста; оба они P и F зависят от A . Сколь большей является их зависимость друг от друга при каком-либо одном возрасте?

$$r_{PF.A} = \frac{0,5784 - (-0,4865) \times (-0,5296)}{\sqrt{(1-0,4865^2) \times (1-0,5296^2)}} = 0,4328.$$

Таким образом, часть взаимосвязи P и F зависит от возраста, но часть ее присуща обычной употребляемой пище. Исследуем эту последнюю.

Для этого необходимо вычислить регрессию P на A и после этого определить \hat{P} для каждого A . Полученные таким образом 54 отклонения от регрессии $P - \hat{P}$ будут свободными от влияния возраста. Варьирование их должно быть теперь приписано другим причинам, включая сюда, возможно, и содержание в пище жира. Далее тем же путем должна быть установлена и регрессия F на A . Отклонения $F - \hat{F}$ будут независимыми от возраста. Наконец. Для этих 54 пар отклонений должна быть определена корреляция; эта корреляция должна быть равна $r_{PF.A}$, т. е. корреляции между протеином и жиром. После того, как произведено элиминирование влияния возраста.

Для того чтобы яснее представить себе, в чем состоит независимость $r_{PF.A}$ от возраста, рассмотрим данные о 6 женщинах, имеющих примерно 70-летний возраст. Усвоение протеина и жира их организмами характеризуется данными:

$$P : 56, 47, 33, 39, 42, 38;$$

$$F : 56, 83, 49, 52, 65, 52; \quad r_{PF} = 0,4194.$$

Эта корреляция близка к усредненной корреляции $r_{PF.A} = 0,4328$. Подобные корреляции должны быть определены и для других возрастов.

При некоторых других обстоятельствах может оказаться и так, что эта корреляция будет слабой или даже не будет никакой корреляции между P и F . Если в целях иллюстрации произвести небольшое перераспределение приведенных выше фактических данных, то можно получить:

$$P : 56, 39, 47, 33, 42, 38;$$

$$F : 56, 83, 52, 49, 65, 52; \quad r_{PF} = -0,0621.$$

Но, независимо от такого перераспределения данных, средние этих двух переменных останутся теми же, что и ранее, и могут по-прежнему изменяться с возрастом, как это имеет место со средними P и F в нашем случае:

Возраст	30	35	40	45	50	55	60	65	70	75
Среднее P	67	62	65	57	54	49	55	64	44	47
Среднее F	85	101	130	82	86	81	68	85	59	68

Здесь корреляция между средними для протеина и жира составляет 0,7014. Она обусловлена только изменениями потребления в зависимости от возраста; корреляции же при том или ином определенном возрасте могут иметь любое значение, начиная от -1 и кончая $+1$, и зависят от попарных сочетаний наблюдаемых P и F . Все это показывает, каким образом два влияния — возраст и $r_{PF.A}$ — могут быть независимыми. Построение определенных графиков может помочь вам уяснить эти связи.

При большем числе переменных приходится основываться на других связях между частной корреляцией и множественной регрессией. В параграфе 4 главы 7 было указано, что r является геометрической средней из двух коэффициентов регрессии. Подобно этому

$$r_{12.3} = \sqrt{b_{12.3} \cdot b_{21.3}}$$

Здесь снова подчеркивается факт, что в корреляции нет различия между независимой и зависимой переменными. Она измеряет взаимосвязь переменных и только. Практическое значение приведенной выше формулы состоит в том, что при большом числе переменных для вычисления коэффициентов регрессии имеются удобные и изящные методы [9], в то время как прямые вычисления коэффициентов корреляции высшего порядка становятся крайне сложными.

Пример 11. Брансон и Уилльер [2] определили корреляции между окружностями початка E , окружностью его стержня C и количеством рядков зерен K на основании измерения 900 початков кукурузы:

$$r_{EC} = 0,799; \quad r_{EK} = 0,570; \quad r_{CK} = 0,507.$$

Какой будет корреляция между E и C у початков, имеющих одно и то же число рядков зерен? *Ответ:* $r_{EC.K} = 0,720$.

Пример 12. Имеется ли какая-нибудь корреляция между C и K у початков определенной окружности? *Ответ:* $r_{CK.E} = 0,105$.

Пример 13. В случайной выборке 54 женщин штата Айова [12] усвоение организмом указанных ранее двух питательных веществ определялось при одновременном установлении возраста и содержания холестерина в крови. Если символически обозначать P — протеин, F — жир, A — возраст и C — холестерин, то корреляции будут такими:

	A	P	F
P	$-0,4865$		
F	$-0,5296$	$0,5784$	
C	$0,4737$	$-0,4249$	$-0,3135$

Какой будет корреляция между возрастом и холестерином, независимо от усвоения протеина и жира? *Ответ:*

$$r_{AC.PF} = \frac{r_{AC.F} - r_{AP.F} \cdot r_{CP.F}}{\sqrt{(1 - r_{AP.F}^2)(1 - r_{CP.F}^2)}} = \frac{0,3820 - (-0,2604) \times (-0,3145)}{\sqrt{(1 - 0,2604^2) \times (1 - 0,3145^2)}} = 0,3274.$$

7. Решение нормальных уравнений. Если взять данные таблицы 180 и подставить их в нормальные уравнения этого же параграфа, то получится:

$$1752,96b_{Y1.2} + 1085,61b_{Y2.1} = 3231,48 \quad (1);$$

$$1085,61b_{Y1.2} + 3155,78b_{Y2.1} = 2216,44 \quad (2).$$

Удобный способ решения этих уравнений такой:

$$(1) : 1085,61 : 1,61472b_{Y1.2} + b_{Y2.1} = 2,97665 \quad (3);$$

$$(2) : 3155,78 : 0,34401b_{Y1.2} + b_{Y2.1} = 0,70234 \quad (4);$$

$$(3) - (4) : 1,27071b_{Y1.2} = 2,27431 \quad (5);$$

$$(5) : 1,27071 : b_{Y1.2} = 1,78979 \quad (6);$$

$$\text{Подстановка (6) в (3) : } b_{Y2.1} = 2,97665 - 1,61472 \times 1,78979 = 0,08664 \quad (7).$$

Получены те же результаты, что и при подстановке исходных данных в формулы.

Для проведения таких вычислений в будущем полезно заметить правило расчета $b_{y2.1}$, основанное на числах, набранных жирным шрифтом: (верхнее число) — (левое число) \times (нижнее число). В предыдущей форме расчетов большинство символов может быть опущено; после этого основная их часть будет такой:

1752,96	1085,61	3231,48
1085,61	3155,78	2216,44
1,61472	1	2,97665
0,34401	1	0,70234
1,27071		2,27431
$b_{y1.2}$		1,78979
$b_{y2.1}$		0,08664

Все вычисления могут проводиться на счетной машине без промежуточных записей. Этот способ более удобен и легче запоминается, чем применение двух самостоятельных формул для расчетов b . Он вместе с тем может быть распространен и на случай решения большего числа нормальных уравнений.

Замечание по поводу вычислений: форма расчетов
(верхнее) — (левое) \times (нижнее)

может быть применена и для другого хода вычислений по таким числам:

$$0,70234 - 0,34401 \times 1,78979 = 0,08663,$$

что дает прежний результат, исключая небольшую ошибку округления. Это обстоятельство служит контролем вычислений. Окончательной проверкой является подстановка двух значений b в исходные уравнения:

$$1752,96 \times 1,78979 + 1085,61 \times 0,08664 = 3231,49$$

вместо 3231,48 в уравнение [1]. Подобная же подстановка во второе уравнение дает 2216,43.

Обычно для построения доверительных интервалов и для оценки коэффициентов регрессии возникает потребность в весовых коэффициентах Гаусса. Они находятся при решении двух систем уравнений, подобных системе нормальных уравнений.

Первая система:

$$c_{11}\Sigma x_1^2 + c_{12}\Sigma x_1x_2 = 1;$$

$$c_{11}\Sigma x_1x_2 + c_{12}\Sigma x_2^2 = 0.$$

Вторая система:

$$c_{12}\Sigma x_1^2 + c_{22}\Sigma x_1x_2 = 0;$$

$$c_{12}\Sigma x_1x_2 + c_{22}\Sigma x_2^2 = 1.$$

Обе системы уравнений решаются с помощью одной последовательности арифметических действий. Используя те же данные из опыта с фосфором, можно найти:

		Первая система	Вторая система
1752,96	1085,61	1	0
1085,61	3155,78	0	1
1,61472	1	0,000921141	0
0,34401	1	0	0,000316879
1,27071		0,000921141	-0,000316879
		$c_{11} = 0,000724903$	$c_{12} = -0,000249372$
		$c_{21} = -0,000249374$	$c_{22} = 0,000402666$

Эти результаты совпадают с результатами, полученными в параграфе 3 этой главы.

Замечания по поводу вычислений: расчет по формуле (верхнее) — (левое) \times (нижнее) упрощается, если каждый раз начинать с (верхнее) = 0:

$$c_{21} = -0,34401 \times 0,000724903 = -0,000249374;$$
$$c_{22} = -1,61472 \times (-0,000249372) = 0,000402666.$$

Обе пары значений c следует для проверки подставить в одно или в оба исходных уравнений. Например:

$$1752,96 \times 0,000724903 + 1085,61 + (-0,000249372) = 1,00001.$$

После того как вычислены весовые коэффициенты c , на их основе совместно с суммами Σxy могут быть вычислены коэффициенты регрессии:

$$b_{Y1.2} = c_{11}\Sigma x_1y + c_{12}\Sigma x_2y;$$
$$b_{Y2.1} = c_{21}\Sigma x_1y + c_{22}\Sigma x_2y.$$

В нашем случае подстановка сюда соответствующих данных приводит к коэффициентам, вычисленным в параграфе 2 этой главы:

$$0,000724903 \times 3231,48 + (-0,000249372) \times 2216,44 = 1,7898;$$
$$(-0,000249374) \times 3231,48 + 0,000402666 \times 2216,44 = 0,0866.$$

Каждый из двух описанных здесь методов вычислений имеет свои преимущества. Первый, основанный на применении нормальных уравнений, является более коротким и дает все необходимые данные (включая сюда те сведения о коэффициентах регрессии, которые находятся при помощи b' и таблицы, подобной 182). Но второе решение при помощи весовых коэффициентов является необходимым, если возникает потребность в стандартных ошибках коэффициентов для построения доверительных интервалов соответствующих параметров. Оно также удобно в тех случаях, когда имеется несколько переменных Y , зависящих от одного и того же ряда переменных X (параграф 10 этой главы).

8. Четыре и больше числа переменных в одной группе наблюдений.

Вычисления. Увеличение числа количественных признаков, наблюдаемых у отдельных объектов сверх трех, не приводит к включению каких-либо новых принципов и влечет за собой только усложнение методов множественной регрессии. Поэтому здесь будут даны только основы этого вопроса. Описание техники расчетов и более широкое изложение вопроса можно найти в специальной литературе [3, 4, 6, 13].

Следует предостеречь начинающего исследователя от опрометчивого увлечения регрессии с большим числом переменных. Некоторые исследователи видят какую-то магическую силу в собирании колоссального количества данных; они думают, что некая алхимия множественной регрессии может дать достоверную информацию, извлеченную из сомнительных данных, относящихся к разнородному материалу. Риск впасть в этом случае в ошибку увеличивается с размером и сложностью исследования. Опасность потери некоторых данных или неудачи техники, трудности в подыскании однородного материала, затруднения в истолковании результатов — это только немногое из препятствий, стоящих на пути. Тем не менее не следует воздерживаться от применения этих методов, если вы имеете хорошо сделанные измерения осторожно отобранного материала и если вы можете сформулировать вопросы, ответы на которые находятся в области применения этих методов. По сравнению с трудом, затрачиваемым на получение данных, вычисление показателей регрессии является делом совсем легким.

В тех случаях, когда вы имеете четыре или более переменных, наблюдаемых по нескольким частным выборкам объектов, и эти субвыборки представляют отчасти различные, но в то же время связанные между собой совокупности, применимы методы, аналогичные методам параграфа 4 этой главы.

Следует соблюдать особую осторожность при включении большого числа переменных при малых выборках. Причину этого легко уяснить на основе геометрической интерпретации регрессии; мы начнем с наиболее простого случая двух переменных. На плоскости два наблюдения над двумя переменными определяют линейную регрессию; все дополнительные парные наблюдения дают возможность не только улучшить регрессию, но и употребить их для лучшей оценки ошибки. В трехмерном пространстве требуемое для определения плоскости регрессии число наблюдений над тремя переменными составляет уже три. Это положение может по аналогии быть распространено на случай четырех и большего числа переменных, несмотря на ограниченность измерений нашего трехмерного пространства. Например, регрессия шести переменных будет точно определена по шести рядам соответствующих наблюдений, и при наличии только этих наблюдений ничего не останется для получения сведений об ошибке их. При малых выборках на оценку существенности множественной регрессии остается совсем ничтожное количество данных, и это вызывает определенное скептическое отношение к ней.

Для того, чтобы получить пример с четырьмя переменными, к которому мы переходим, введем в исследование о потребности растений в фосфоре дополнительную переменную X_3 , взятую из тех четырех, которые фактически участвовали в исследовании. Для удобства в таблице 190 вместе с X_3 даны и прежние X_1 и X_2 . Вопрос о температуре почвы 35°C будет рассмотрен в дальнейшем.

Для введения третьей переменной X_3 следует вычислить сумму квадратов ее и суммы произведений этой переменной со всеми остальными переменными: $\Sigma x_3^2 = 35\ 572$; $\Sigma x_1 x_3 = 1200$; $\Sigma x_2 x_3 = 3364$; $\Sigma x_3 y = 7593$.

ТАБЛИЦА 190

Содержание фосфора в различных известковых почвах и определение усвоения фосфора растениями при двух температурах почвы

Номер почвенного образца	Содержание фосфора в почве частей на миллион ¹			Усвоение фосфора почвы растением в частях на миллион	
	X_1	X_2	X_3	температура почвы 20°C Y	температура почвы 35°C Y'
1	0,4	53	153	64	93
2	0,4	23	163	60	73
3	3,1	19	37	71	38
4	0,6	34	157	61	109
5	4,7	24	59	54	54
6	1,7	65	123	77	107
7	9,4	44	46	81	99
8	10,1	31	117	93	94
9	11,6	29	173	93	66
10	12,6	58	112	51	126
11	10,9	37	111	76	75
12	23,1	46	114	96	108
13	23,1	50	134	77	90
14	21,6	44	73	93	72
15	23,1	56	168	95	90
16	1,9	36	143	54	82
17	26,8	58	202	168	128
18	29,9	51	124	99	120

¹ X_1 — неорганический фосфор, определенный по методу Брея и Курца.

X_2 — органический фосфор, растворенный в K_2CO_3 и гидролизанный гипобромитом.

X_3 — органический фосфор, растворенный в K_2CO_3 и не гидролизанный гипобромитом.

Теперь для определения трех коэффициентов регрессии следует решить три нормальных уравнения:

$$\begin{aligned}\Sigma x_1^2 b_{Y1.23} + \Sigma x_1 x_2 b_{Y2.13} + \Sigma x_1 x_3 b_{Y3.12} &= \Sigma x_1 y; \\ \Sigma x_1 x_2 b_{Y1.23} + \Sigma x_2^2 b_{Y2.13} + \Sigma x_2 x_3 b_{Y3.12} &= \Sigma x_2 y; \\ \Sigma x_1 x_3 b_{Y1.23} + \Sigma x_2 x_3 b_{Y2.13} + \Sigma x_3^2 \cdot b_{Y3.12} &= \Sigma x_3 y.\end{aligned}$$

Решение этой системы совместных уравнений дано в таблице 191. Суммы квадратов и произведений взяты из таблицы 180 и дополнены вычисленными в этом параграфе. Понужные символы опущены.

ТАБЛИЦА 191

Решение трех нормальных уравнений. Данные об усвоении растениями фосфора

	X_1	X_2	X_3	Y	Строка
Суммы квад- ратов и X_1 произведе- ний X_2 X_3	1752,96 1085,61 1200,00	1085,61 3155,78 3364,00	1200,00 3364,00 35572,00	3231,48 2216,44 7583,00	(1) (2) (3)
(1): 1200,00	1,460800	0,904675	1	2,692900	(4)
(2): 3364,00	0,322714	0,938103	1	0,658870	(5)
(3): 35572,00	0,033734	0,094569	1	0,213454	(6)
(4) — (6)	1,427066	0,810106		2,479446	(7)
(5) — (6)	0,288980	0,843534		0,445416	(8)
(7): 0,810106	1,761579	1		3,060644	(9)
(8): 0,843534	0,342583	1		0,528036	(10)
(9) — (10)	1,418996			2,532608	(11)
(11): 1,418996	$b_{Y1.23}$			1,784789	(12)
		$b_{Y2.13}$		-0,083403	(13)
			$b_{Y3.12}$	0,161133	(14)

$$b_{Y2.13} = 3,060644 - 1,761579 \times 1,784789 - 0,528036 - 0,342583 \times 1,784789;$$

$$b_{Y3.12} = 2,692900 - 1,460800 \times 1,784789 - 0,904675 \times (-0,083403) =$$

$$= 0,658870 - 0,322714 \times 1,784789 - 0,938103 \times (-0,083403) \text{ и т. д.}$$

З а м е ч а н и я п о п о в о д у в ы ч и с л е н и й. Определение $b_{Y2.13}$ в нижней части таблицы производится так, как это было в предыдущем параграфе, т. е. используя «верхнее» и «левое» — по строке 9 или 10, которые относятся к последнему блоку, содержащему «1». Для определения $b_{Y3.12}$ берется верхний блок с «1», его строка 4. По этой строке имеется два «левое» и два «нижнее» числа, одна их система набрана жирным прифтом, а вторая — курсивом. Схема вычислений в последних двух строках таблицы такая:

$$\text{(верхнее)} - \text{(левое)} \times \text{(нижнее)} - \text{(левое)} \times \text{(нижнее)}.$$

Для проверки вычислений эту схему можно применить по строкам 5 и 6.

Приведенная здесь форма вычислений легко распространяется на число независимых переменных, большее трех. Существуют десятки и, может быть, даже сотни различных способов вычисления коэффициентов множественной регрессии b [4, 13]. Некоторые из них более компактны, чем показанный здесь способ, но преимущество этого последнего состоит в том, что он близок к обычному методу решения системы уравнений и благодаря этому легче укладывается в памяти.

Правильность вычисленных b должна быть проверена путем подстановки их в одно или несколько первоначальных уравнений. Например, подстановка в [1] дает: $1752,96 \times 1,784789 - 1085,61 \times 0,083403 + 1200,00 \times 0,161133 = 3231,48$, что подтверждает точность вычислений.

Смысл этих показателей и выводы будут рассмотрены в следующем параграфе.

9. Четыре или большее число переменных. Выводы. После преодоления всех трудностей вычислений статистик может перейти к рассмотрению полученных этим путем результатов. Следующий его шаг будет зависеть от задачи опыта. «Основной целью настоящего исследования было определение, имеет ли самостоятельное влияние почвенного органического фосфора на количество фосфора, усвояемого растениями». Это значит, что экспериментаторы, поставившие опыт, желали знать, могут ли X_2 и X_3 оказать какую-либо помощь при предсказании значения Y . Мы прежде всего рассмотрим этот вопрос, оставляя суждения по другим вопросам до примеров этого параграфа.

Быстрый, но «черновой» метод состоит в рассмотрении стандартных коэффициентов частной регрессии, $b_i \sqrt{\frac{\sum x_i^2}{\sum y^2}}$

$$b'_{Y2.23} = 1,7848 \sqrt{1752,96/12389,61} = 0,67;$$

$$b'_{Y2.13} = -0,0834 \sqrt{3155,78/12389,61} = -0,04;$$

$$b'_{Y3.12} = 0,1611 \sqrt{35572/12389,61} = 0,27.$$

Это с несомненностью приводит к выводу, что главную часть доступного растению фосфора составляет неорганический фосфор и что часть органического фосфора X_2 неэффективна. Решение же вопроса о второй части органического фосфора требует применения более точных методов.

Для первого из этих методов нам необходимо определить сумму квадратов Y , обусловленную регрессией, и остаточную сумму $\sum y^2$:

$$\sum y_{123}^2 = 1,7848 \times 3231,48 \times (-0,0834) \times 2216,44 + 0,1611 \times 7593,00 = 6806;$$

$$\sum d_{Y.123}^2 = 12390 - 6806 = 5584.$$

Рассматривая эти данные совместно с $\sum \hat{y}_{12}^2$ из таблицы 181, имеем:

Регрессия на X_1, X_2, X_3	3	6 806	
Регрессия на X_1, X_2	2	5 976	
X_3 после X_1 и X_2	1	830	830
Ошибка	14	5 584	399

$F=830/399=2,08$ соответствует вероятности примерно 0,18. Это приводит к прежнему выводу: неорганическая часть фосфора является главным и, возможно, единственным источником доступного для растения при 20° С фосфора.

Теперь можно установить некоторые общие положения относительно множественной регрессии. Во-первых, как уже ранее отмечалось, коэффициент регрессии меняется при переходе к новой группе X . Когда берется только X_2 , то $b_{Y2} = 2216,44/3155,78 = 0,7023$. Добавление X_1 приводит к $b_{Y2.1} = -0,0866$. При трех переменных X $b_{Y2.13} = -0,0834$. Во всякой множественной регрессии коэффициенты регрессии взаимосвязаны, поэтому увеличение или уменьшение числа переменных изменяет все b .

Во-вторых, значение $\sum \hat{y}^2$ при включении новых X никогда не уменьшает-ся, а обычно возрастает. Беря одно только X_1 , имеем $\sum \hat{y}_1^2 = 3231,48^2/1752,96 = 5957$; при X_1 и X_2 будет $\sum \hat{y}_{12}^2 = 5976$, и для всех трех переменных $\sum \hat{y}_{123}^2 = 6806$. Это увеличение может быть небольшим и несущественным, но все же оно представляет собой новые сведения об Y , полученные в связи с добавлением новых X .

В-третьих, высокая корреляция между двумя X может в значительной мере исказиться в ходе вычислений. Если, например, r_{ij} больше 0,95, то даже 6 или 8 значащих цифр может оказаться недостаточным для устранения ошибок округления. Когда коэффициенты корреляции применяются для кодирования X (пример 26 этой главы) и если один из них больше 0,95, то следует

рассмотреть вопрос об исключении одного из двух X , определяющих эту корреляцию.

В-четвертых, если $\Sigma \hat{y}^2$ составляет только незначительную часть Σy^2 , т. е. если R^2 мал (пример 6 этой главы), то это означает, что большая часть варьирования Y остается необъясненной. Это варьирование может быть случайным или оно может быть обусловлено другими независимыми переменными, не принятыми во внимание в данной регрессии. Если эти переменные будут установлены и введены в регрессию, то соотношения между прежними X могут полностью измениться. Не следует слишком спекулировать относительно причин явления и влияний на него, пока не будет получено достаточно высокое значение R^2 , положим около 0,8.

Пример 14. Вычислите регрессию доступного для растения фосфора на указанные ранее три фактора. *Ответ.* $\hat{Y} = 1,7848 X_1 - 0,0834 + 0,1611 X_3 + 43,67$.

Пример 15. Определите доступный для растения фосфор в почвенном образце 17 и сравните его с фактически наблюдаемым значением. *Ответ:* 119 частей на миллион, $Y - \hat{Y} = 49$ частей на миллион.

Пример 16. Экспериментатор мог бы удовлетвориться информацией, которую будет давать ему сохранение в уравнении регрессии только переменной X_3 совместно с X_1 при исключении X_2 . Вычислите новую регрессию. *Ответ.* $\hat{Y} = 1,737 X_1 + 0,155 X_3 + 41,5$.

Пример 17. Вычислите сумму квадратов Y , обусловленную переменной X_2 после включения X_1 и X_3 . *Ответ:* 16.

Пример 18. Какой будет сумма квадратов Y для X_1 , включенного в регрессию после X_2 и X_3 ? *Ответ.* 4394.

Пример 19. Вычислите $R^2 = \Sigma \hat{y}^2 / \Sigma y^2$ для одного только X_1 , для X_1 и X_2 и для X_1 , X_2 и X_3 . *Ответ.* $R_{\hat{Y}.1}^2 = 0,4808$; $R_{\hat{Y}.12}^2 = 0,4823$; $R_{\hat{Y}.123}^2 = 0,5493$. Заметьте, что R^2 при добавлении нового X не уменьшается; как правило, он увеличивается. Свяжите это с соответствующей теоремой относительно $\Sigma \hat{y}^2$.

10. Четыре и большее число переменных. Элементы обратной матрицы; весовые коэффициенты Гаусса. Для построения интервальных оценок и критериев существенности b , \hat{Y} и пр. необходимо вычисление величин c . Это *элементы обратной матрицы*. Обычно они называются *весовыми коэффициентами Гаусса* [8]. Они не только необходимы для многих критериев, но применение их позволяет экономить время в тех случаях, когда несколько величин относятся к одной и той же группе независимых переменных.

Как следует из параграфа 7 главы 4, вычисление этих элементов удобно проводить путем решения стольких систем совместных уравнений, сколько имеется переменных X . При трех X эти системы таковы¹:

$$\begin{aligned} \Sigma x_1^2 c_{11} + \Sigma x_1 x_2 c_{12} + \Sigma x_1 x_3 c_{13} &= 1, 0, 0; \\ \Sigma x_1 x_2 c_{21} + \Sigma x_2^2 c_{22} + \Sigma x_2 x_3 c_{23} &= 0, 1, 0; \\ \Sigma x_1 x_3 c_{31} + \Sigma x_2 x_3 c_{32} + \Sigma x_3^2 c_{33} &= 0, 0, 1. \end{aligned}$$

Левые части этих уравнений являются теми же самыми, какими они были в предыдущем параграфе при замене b на c . Правые части первой системы (1, 0, 0) получаются путем замещения Σxy и т. д.

Если в первую из этих систем уравнений подставить данные о фосфоре и после этого применить методы предыдущего параграфа, то вы будете делить 1 на 1200, что при сохранении 6 значащих цифр даст 0,000833333. В третьей системе вы получите $1/35572 = 0,0000281120$. Арифметические расчеты с 4

¹ Примененное автором символическое изображение трех систем уравнений для определения c_{ij} нельзя признать удачным, так как при изменении правых частей этих уравнений меняются и неизвестные в левой их части. Р. Фишер [7] дает более удачную схему:

$$\begin{aligned} \Sigma x_1^2 b_1 + \Sigma x_1 x_2 b_2 + \Sigma x_1 x_3 b_3 &= 1, 0, 0 & b_1 &= c_{11}, c_{21}, c_{31}; \\ \Sigma x_1 x_2 b_1 + \Sigma x_2^2 b_2 + \Sigma x_2 x_3 b_3 &= 0, 1, 0 & b_2 &= c_{12}, c_{22}, c_{32}; \\ \Sigma x_1 x_3 b_1 + \Sigma x_2 x_3 b_2 + \Sigma x_3^2 b_3 &= 0, 0, 1 & b_3 &= c_{13}, c_{23}, c_{33}. \end{aligned}$$

(Прим. перев.)

и 5 знаками слева от запятой и с 10 знаками справа дело не особенно приятное. Поэтому следует обратиться к тому или иному кодированию данных.

Наилучший способ кодирования, который я знаю, состоит в замене сумм произведений на соответствующие коэффициенты корреляции. Но такое кодирование и обратный от него переход несколько неудобны (пример 26). Предложенный Мэри Клем в нашей вычислительной лаборатории способ, удачный в отношении однообразия действий, состоит в делении каждой левой части уравнений на соответствующую Σx^2 . Практически это производится так: обратное значение Σx^2 ставится на счетную машину и последовательно умножается на коэффициенты при c . Например, по строке (1) таблицы 192 делим: $1/1752,96 = 0,000570464$, что для удобства выписывается в строку (4). Умножая это обратное число на три коэффициента, стоящие в этой строке, вы получите три первых числа по строке (4): 1, 0,6193, 0,6846. Для уменьшения ошибок округления указанное выше обратное число следует брать с двумя знаками больше, чем число знаков, оставляемых в самой таблице. Эти кодированные числа становятся коэффициентами при c . После того, как они внесены в строки (4), (5) и (6), решение уравнений проводится по образцу параграфа 7 этой главы.

Вы можете заметить, что преобразованная таким образом матрица уже не будет симметричной, т. е. коэффициент c_{12} не равен коэффициенту c_{21} и т. д. Но это не влияет на применяемый здесь метод решения уравнений.

Последняя правая колонка таблицы 192 является контрольной колонкой; введение ее было бы полезным и в параграфах 7 и 8 этой главы, но она становится теперь необходимой. Числа этой колонки получаются простым суммированием данных по строкам (4), (5) и (6), после чего с этими суммами поступают так, как с данными других колонок. Например, по строке (7) имеем: $4,8260 = 1,460707 \times 3,3039$. Проверка производится путем сложения всех других данных строки (7): $1,4607 + 0,9046 + 1,0000 + 1,4607 = 4,8260$, т. е. то же число, что и вычисленное ранее. Эти два числа должны быть в пределах ошибок округления равны друг другу.

Пока вычисленная таким образом матрица c несимметрична. Для обратного перехода от кода следует каждую *колонку* разделить на первоначальную сумму Σx^2 этой колонки. Например, решение 1:

$$\begin{aligned} 1,2707 \times 0,000570464 &= 0,0007249; \\ -0,4353 \times 0,000570464 &= -0,0002483; \\ -0,0016 \times 0,000570464 &= -0,0000009. \end{aligned}$$

Вас может удивить применение здесь деления (а не умножения) и проведение этой операции по столбцам (а не по строкам). Это происходит потому, что обратная матрица состоит из элементов, в которых первоначальный числитель становится знаменателем и строки становятся столбцами. Обратный переход от кода приводит к первоначальной симметричности, что само по себе служит прекрасным контролем.

В порядке окончательной проверки вычислений производится подстановка этих элементов матрицы, полученных в результате обратного перехода от кода, в первоначальные уравнения. Например, подставим значения первого решения в первое исходное уравнение и получим:

$$\begin{aligned} (18) \text{ в } (1): 1752,96 \times 0,0007249 - 1085,61 \times 0,0002483 - 1200,00 \times 0,0000009 = \\ = 1,0001 \end{aligned}$$

вместо 1,0000, что объясняется ошибкой округления. Подобно этому, подставим результаты второго решения в (2):

$$-1085,61 \times 0,0002483 + 3155,78 \times 0,0004375 - 3364,00 \times 0,0000330 = 1,0001.$$

Подстановка третьего решения в [3] не дает столь хороших результатов, 1,0012. Здесь ошибки округления вкрадываются в третий десятичный знак.

Действительно, при проведении работы только с четырьмя значащими цифрами, как это было у нас, нельзя надеяться на точность третьего десятичного знака. Для того чтобы показать, какие результаты можно ожидать, я провел расчеты с шестью десятичными знаками, а один из студентов с десятью. Сравнение этих более точных результатов с приведенными ранее произведено в следующей табличке:

	4 знака	6 знаков	10 знаков
c_{11}	0,0007249	0,000724929	0,0007249289466
c_{31}	0,0000009	0,000000969	0,0000009691292
$b_{Y1.23}$	1,7846	1,78479	1,784787539
$b_{Y3.12}$	0,1616	0,161132	0,161132757
$s_{\hat{Y}.123}^2$	6809	6805,9	6806,13

Легко видеть, что окончательные результаты немногим отличаются друг от друга, несмотря на грубое решение уравнения (3).

Дальнейшие вычисления в таблице 192 уже знакомы. Например:

$$b_{Y2.13} = -0,0002483 \times 3231,48 - 0,0004375 \times 2216,44 + (-0,0000330) \times 7593,00 = -0,0833;$$

$$\Sigma \hat{y}_{123}^2 = 1,7846 \times 3231,48 + (-0,0833) \times 2216,44 + 0,1616 \times 7593,00 = 6809;$$

$$s_{b_1} = 20,0 \cdot \sqrt{0,0007249} = 0,538;$$

$$s_{b_2} = 20,0 \cdot \sqrt{0,0004375} = 0,418;$$

$$t_{b_1} = 1,7846/0,538 = 3,32.$$

Указанная в предыдущем параграфе идентичность критериев t и F в нашем случае может быть подтверждена сравнением результатов для X_3 . Оценка $b_{Y3.12}$ дает $t=1,44$, но это есть $\sqrt{2,08}$, где $F=2,08$ является критерием для X_3 , введенного после X_1 и X_2 .

Если для этого имеется какой-либо повод, то теперь может быть произведена проверка гипотезы о том, что два коэффициента регрессии равны между собой. Такое сравнение в опыте с фосфором вряд ли представляет какой-либо интерес, так как здесь регрессия на обе части органического фосфора оказалась несущественной. Но для иллюстрации все же найдем разность:

$$b_{Y3.12} - b_{Y2.13} = 0,1616 - (-0,0833) = 0,2449$$

и оценим ее существенность путем вычисления среднего квадрата

$$s_D^2 = s_{\hat{Y}.123}^2 (c_{22} + c_{33} - 2c_{23}) =$$

$$= 398,6 [0,0004375 + 0,0000313 - 2(-0,0000313)] = 0,2132.$$

Отсюда

$$s_D^2 = 0,462; t = 0,2449/0,462 = 0,53; f = 14.$$

Для определения доверительного интервала Y следует сначала вычислить средний квадрат

$$s_{\hat{Y}}^2 = s_{\hat{Y}.123}^2 (1/n + c_{11}x_1^2 + c_{22}x_2^2 + c_{33}x_3^2 + 2c_{12}x_1x_2 + 2c_{13}x_1x_3 + 2c_{23}x_2x_3),$$

где x является отклонениями от средних. Подставим сюда данные о фосфоре для почвенного образца 17, для которого в примере 15 этой главы было опре-

Вычисление элементов обратной матрицы. Доступный для растения фосфор

	X_1	X_2	X_3	Решения			Контрольная колонка	Строка
				1	2	3		
На таблицы 191	X_1	1752,96	1085,61	1200,00				(1)
	X_2	1085,61	3155,78	3364,00				(2)
	X_3	1200,00	3364,00	35572,00				(3)
(1) \times (1/1752,96). (0,000570464)	1	0,6193	0,6846	1	0	0	3,3039	(4)
(2) \times (1/3155,78). (0,000316879)	0,3440	1	1,0660	0	1	0	3,4100	(5)
(3) \times (1/35572,00). (0,0000281120)	0,0337	0,0946	1	0	0	1	2,1283	(6)
(4) \times (1/0,6846). (1,460707)	1,4607	0,9046	1	1,4607	0	0	4,8260	(7)
(5) \times (1/1,0660). (0,938086)	0,3227	0,9381	1	0	0,9381	0	3,1989	(8)
(6)	0,0337	0,0946	1	0	0	1	2,1283	(9)
(7) - (9)	1,4270	0,8100		1,4607	0	-1	2,6977	(10)
(8) - (9)	0,2890	0,8435		0	0,9381	-1	1,0706	(11)
(10) \times (1/0,8100). (1,234568)	1,7617	1		1,8033	0	-1,2346	3,3305	(12)
(11) \times (1/0,8435). (1,185536)	0,3426	1		0	1,1122	-1,1855	1,2692	(13)
(12) - (13)	1,4191			1,8033	-1,1122	-0,0491	2,0613	(14)
(14) \times (1/1,4191). (0,704672)								
	Обратная матрица			1,2707	-0,7837	-0,0346	1,4525	(15)
				-0,4353	1,3806	-1,1736		(16)
				-0,0016	-0,1041	1,1122		(17)
$\Sigma y^2 = 12,390$								
$\Sigma \hat{y}^2 = \Sigma (b \Sigma xy) = 6809$								
$\Sigma d^2 = 5581$								
	Перехода от кода			0,0007249	-0,0002483	-0,0000010	$\sqrt{c_{11}} = 0,0269$	(18)
				-0,0002483	0,0004375	-0,0000330	$\sqrt{c_{22}} = 0,0209$	(19)
				-0,0000009	-0,0000330	0,0000313	$\sqrt{c_{33}} = 0,0056$	(20)
$s_{Y.123}^2 = \Sigma d^2 / 14 = 398,6$	Σxy (таблица 191)			3231,48	2216,44	7593,00		(21)
$s_{Y.123} = 20,0$	$b = \Sigma (c \Sigma xy)$			1,7846**	-0,0833	0,1616		(22)
	$s_b = s_{Y.123} \sqrt{c}$			0,538	0,418	0,112		(23)
	t			3,32	0,20	1,44		(24)

делено по регрессии значение \hat{Y} :

$$x_1 = 26,8 - 11,9 = 14,9; \quad x_2 = 58 - 42,1 = 15,9; \quad x_3 = 202 - 123 = 79.$$

$$s_{\hat{Y}}^2 = 398,6 \cdot [1/18 + 0,0007249 \times 14,9^2 + 0,0004375 \times 15,9^2 + 0,0000330 \times 79^2 +$$

$$+ 2(-0,0002483) \times 14,9 \times 15,9 + 2(-0,0000010) \times 14,9 \times 79 +$$

$$+ 2(-0,0000330) \times 15,9 \times 79] = 123,6.$$

$$s_{\hat{Y}} = 11,1; \quad t_{0,05} = 2,145; \quad 119 \pm 2,145 \times 11,1; \quad \text{от } 95 \text{ до } 143 \text{ частей}$$

на миллион.

Если возникает потребность в построении доверительного интервала для наблюдаемого значения, то следует к $s_{\hat{Y}}^2$ прибавить $s_{\hat{Y} \cdot 123}^2$:

$$s_{\hat{Y}}^2 = 398,6 + 123,6 = 522,2; \quad s_{\hat{Y}} = 22,9;$$

$$119 \pm 2,145 \times 22,9; \quad \text{от } 70 \text{ до } 168 \text{ частей на миллион.}$$

Верхний предел случайно совпал с наблюдаемым значением Y .

Иногда встречаются случаи, когда одному и тому же ряду наблюдений над независимыми переменными соответствуют несколько отдельных рядов наблюдений над Y . Так в опыте с фосфором каждый почвенный образец изучался как при температуре 20°C , так и при температуре 35°C . Результаты для этой последней температуры даны в последнем столбце таблицы 190. Так как обратная матрица включает в себя только X , то ее очень просто применить и ко второй переменной Y . Остается вычислить только новые суммы произведений:

$$\Sigma x_1 y' = 1720,42; \quad \Sigma x_2 y' = 4337,66; \quad \Sigma x_3 y' = 8324,00.$$

Сочетая эти величины с вычисленными ранее c , получим также коэффициенты регрессии Y' :

$$b_{Y' \cdot 1 \cdot 23} = 0,1618; \quad b_{Y' \cdot 2 \cdot 13} = 1,1958; \quad b_{Y' \cdot 3 \cdot 12} = 0,1159.$$

Отсюда новый ряд результатов:

$$\Sigma (\hat{y}')^2 = 6430, \quad \Sigma d^2 = 12390 - 6430 = 5960, \quad s_{\hat{Y}' \cdot 123}^2 = 425,7.$$

При $s_{Y' \cdot 123} = 20,6$ части на миллион стандартные отклонения трех коэффициентов регрессии будут 0,554; 0,431 и 0,115. Это приводит к трем значениям t : 0,29; 2,77 и 0,10.

«Из результатов этих двух опытов становится очевидным, что органический фосфор почвы *сам по себе* имеет только малое или даже не имеет никакого значения в фосфорном питании растений и что он приобретает некоторую ценность только тогда, когда переходит в неорганическую форму... скорость минерализации органического фосфора при 35°C в несколько раз больше, чем при 20°C ».

Пример 20. При изучении опыления красного клевера пчелами [10] ставилась задача определить влияние различий в длине хоботка насекомого. Измерение хоботка затруднительно, поэтому был произведен предварительный опыт для определения более доступных для измерения признаков насекомого, которые должны иметь высокую корреляцию с длиной хоботка. На 44 пчелах испробовано три таких показателя и получены приводимые ниже результаты:

$n = 44$	Сухой вес X_1 (в мг)	Длина крыла X_2 (в мм)	Ширина крыла X_3 (в мм)	Длина хоботка Y (в мм)
Средняя	13,10	9,61	3,28	6,59

Суммы квадратов и произведений

	X_1	X_2	X_3	Y
X_1	16,6840	1,9279	0,8240	1,5057
X_2		0,9924	0,3351	0,5989
X_3			0,2248	0,1848
Y				0,6831

Кодирование здесь не является необходимым, его можно провести разве только для приобретения навыка. Беря четыре десятичных знака, вычислите регрессию по образцу параграфа 8 этой главы. *Ответ:* 0,0291; 0,6152 и $-0,2017$. *Примечание:* при вычислении b_{Y2} получено два результата: 0,6156 и 0,6149; для b_{Y3} три расчета дали $-0,2013$, $-0,2021$ и $-0,2016$. Я взял средние из этих значений. Эти расхождения характеризуют влияние ошибок округления при проведении вычислений с четырьмя десятичными знаками.

Пример 24. Сделайте грубую оценку относительной роли X при предугадывании Y . *Ответ:* Стандартные частные регрессии 0,14; 0,74 и $-0,12$.

Пример 22. Оцените существенность регрессии в целом. *Ответ:* $F=16,2$; $f=3$ и 40.

Пример 23. Оцените существенность совместного влияния X_1 и X_3 , включенных в регрессию после X_2 . *Ответ:* $F=0,87$.

Пример 24. Вычислите s -матрицу и по ней коэффициенты регрессии. *Ответ:* 0,0291, 0,6148 и $-0,2009$. Это дает представление о влиянии ошибок округления до четырех десятичных знаков.

Пример 25. Какими будут значения t при оценке существенности этих трех коэффициентов регрессии? *Ответ:* 1,18; 4,74 и 0,76.

Пример 26. Произведите кодирование данных опыта с фосфором путем вычисления по суммам квадратов и произведений коэффициентов корреляции.

	X_1	X_2	X_3
Σx_i^2 ; $\Sigma x_1 x_j$	1752,96	1085,61	1200,00
$\sqrt{\Sigma x_1^2}$; $\sqrt{\Sigma x_1^2} \times \sqrt{\Sigma x_j^2}$	41,8684	2352,01	7896,61
r_{1j}	1,0000	0,46156	0,15196
Σx_2^2 ; $\Sigma x_2 x_3$		3155,78	3364,00
$\sqrt{\Sigma x_2^2}$; $\sqrt{\Sigma x_2^2} \times \sqrt{\Sigma x_3^2}$		56,1763	10595,15
r_{2j}		1,0000	0,31750
Σx_3^2			35572,00
$\sqrt{\Sigma x_3^2}$			188,6054

Квадратные корни из трех сумм квадратов здесь перемножены попарно для того, чтобы получить знаменатели соответствующих коэффициентов корреляции. Например:

$$\sqrt{1752,96} \times \sqrt{35572,00} = 41,8684 \times 188,6054 = 7896,61.$$

После этого сумма произведений делится на соответствующее произведение этих квадратных корней, что и дает r . Например:

$$r_{23} = \frac{\Sigma x_2 x_3}{\sqrt{\Sigma x_2^2} \times \sqrt{\Sigma x_3^2}} = \frac{3364,00}{10595,15} = 0,31750.$$

Заметим, что $r_{11}=r_{22}=r_{33}=1$.

Пример 27. Произведите обращение матрицы коэффициентов корреляции:

1,00000	0,46156	0,15196
0,46156	1,00000	0,31750
0,15196	0,31750	1,00000

Ответ.

1,27076	$-0,58411$	$-0,00765$
$-0,58410$	1,38060	$-0,34958$
$-0,00765$	$-0,34958$	1,11215.

Пример 28. Для перехода от кода в этой обратной матрице произведите деление каждого элемента ее на тот же самый делитель, который был использован для кодирования. Например:

$$\begin{aligned} -1,27076/1752,96 &= 0,00072492 \\ -0,58411/2352,01 &= -0,00024834 \\ 1,11215/35572,00 &= 0,00003126. \end{aligned}$$

Эти элементы более точны, чем те, которые вычислены в таблице 192. В худшем случае они ошибочны на одну единицу в последнем знаке. Проверьте свои результаты путем вычисления коэффициентов регрессии и сравнения их с теми, которые были получены в таблице 191.

11. Лишние независимые переменные. После вычисления уравнения регрессии может возникнуть вопрос о бесполезности использования той или иной переменной и об ее исключении. Вместо того чтобы проводить все вычисления заново, можно при помощи приводимых ниже формул (7) перейти к сниженному порядку регрессии, в которой уже не участвует эта переменная.

В опыте с фосфором при 20° С переменная X_2 практически не имеет никакого значения для предсказания Y . В примерах 16 и 17 этой главы эта переменная была выкинута путем вычисления регрессии на две переменные и определения по ним $\Sigma \hat{y}_{13}^2$ и т. д. Но если обратная матрица уже вычислена, то уменьшение суммы Σy_{123}^2 до $\Sigma \hat{y}_{13}^2$ можно вычислить проще:

$$b_{Y2.13}^2/c_{22} = (-0,0833)^2/0,0004375 = 16.$$

Таким образом: $\Sigma \hat{y}_{13}^2 = 6806 - 16 = 6790$.

Более того, легко вычислить и два новых значения оставшихся коэффициентов регрессии:

$$\begin{aligned} b_{Y1.3} &= b_{Y1.32} - c_{12}(b_{Y2.13}/c_{22}) = 4,737; \\ b_{Y3.1} &= b_{Y3.12} - c_{32}(b_{Y2.13}/c_{22}) = 0,155. \end{aligned}$$

Эти результаты согласуются с теми, которые получены в примере 16.

Можно решить задачу и о вычислении вместо прежних новых элементов матрицы сниженного порядка

$$\begin{aligned} (c_{11})_R &= c_{11} - c_{12}^2/c_{22} = 0,0005840; \\ (c_{33})_R &= c_{33} - c_{32}^2/c_{22} = 0,0000288; \\ (c_{13})_R &= c_{13} - c_{12}c_{32}/c_{22} = -0,0000197. \end{aligned}$$

На основе всего этого могут быть вычислены и все другие требуемые показатели. Вы, наверное, видите, что при трех независимых переменных метод параграфа 2 этой главы легче. Описанная сейчас схема выгодна при большом числе переменных X .

Формулы для исключения переменной X_1 или X_3 легко выводятся из приведенных выше; следует только иметь в виду подстрочный указатель исключаемой переменной и того из показателей, который нужно определить. Не представляет затруднений написать аналогичные формулы для исключения переменной из регрессии с четырьмя и большим числом независимых переменных. Однако если необходимо исключить не одну, а большее число переменных, то проведение работы заново может быть более легким, чем последовательное применение этих формул.

Пример 29. Исключите ширину крыла X_3 из регрессии длины хоботка пчелы. Вычислите $\Sigma \hat{y}_{12}^2 = 0,3705$.

Пример 30. На основе предыдущего оцените существенность регрессии Y на X_1 и X_2 в целом. *Ответ:* $F = 24,3$; $j = 2$ и 41. Почему это F больше того, которое было в примере 22.

12. Сводка формул.

A. Метод непосредственного вычисления b :

1. $\Sigma \hat{y}^2 = \Sigma (b_i \Sigma x_i y)$.
2. $\Sigma d^2 = \Sigma y^2 - \Sigma \hat{y}^2$.

$$3. \text{Доля } \Sigma y^2, \text{ обусловленная регрессией} = \Sigma \hat{y}^2 / \Sigma y^2 = R^2.$$

$$4. s_{Y \cdot 12 \dots}^2 = \Sigma d^2 / (n - m), \text{ где } m - \text{число переменных.}$$

$$5. b_{Y1 \cdot 23 \dots} = b_{Y1 \cdot 23 \dots} \cdot \sqrt{\Sigma x_1^2} / \sqrt{\Sigma y^2}.$$

В. После вычисления c -матрицы:

$$1. b_1 = \Sigma (c_{i1} \Sigma x_i y).$$

$$2. \Sigma \hat{y}^2 = \Sigma (b_i \Sigma x_i y).$$

$$3. \Sigma d^2 = \Sigma y^2 - \Sigma \hat{y}^2.$$

$$4. R^2 = \Sigma \hat{y}^2 / \Sigma y^2.$$

$$5. s_{Y \cdot 12 \dots}^2 = \Sigma d^2 / (n - m).$$

$$6. s_{bi} = s_{Y \cdot 12 \dots} \cdot \sqrt{c_{ii}}.$$

$$7. s_{b_1 - b_2}^2 = s_{Y \cdot 12 \dots}^2 (\Sigma c_{ii} - 2 \Sigma c_{ij}); i < j.$$

$$8. s_{\hat{y}}^2 = s_{Y \cdot 12 \dots}^2 (1/n + \Sigma c_{ii} x_i^2 + 2 \Sigma c_{ij} x_i x_j); i < j.$$

$$9. s_{\hat{y}}^2 = s_{Y \cdot 12 \dots}^2 (1 + 1/n + \Sigma c_{ii} x_i^2 + 2 \Sigma c_{ij} x_i x_j); i < j.$$

ЛИТЕРАТУРА

1. Barnard M. M., Journal of Agricultural Science, 26, 456, 1936.
2. Brunson Arthur M., Willier J. G., Journal of the American Society of Agronomy, 21, 912, 1929.
3. Doolittle M. H., United States Coast and Geodetic Survey Report, 115, 1878.
4. Dwyer P. S., Linear Computations. John Wiley and Sons, Inc., New York, 1951.
5. Eid M. T., Black C. A., Kempthorne O., Zoellner J. A., Iowa Agricultural Experiment Station Research Bulletin 406, 1954.
6. Ezekiel Mordecai, Methods of Correlation Analysis. John Wiley and Sons, Jns., New York. Revised Edition, 1941.
7. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925.
8. Gauss Carl Friedrich, Theoria Combinationis Observatorum, Pars Posterior, 1821, Werke, 4, 13, Supplement, 1826, Werke, 4, 71.
9. Goulden Cyril H., Methods of Statistical Analysis, page 147. John Wiley and Sons, Inc., New York, 2nd Edition, 1952.
10. Groust Roy A., Iowa Agricultural Experiment Station Research Bulletin 218, 1937.
11. Lindley D. V., Supplement to the Journal of the Royal Statistical Society, 9, 218, 1947.
12. Swanson Pearl P., Leverton Ruth, Gram Mary R., Roberts Harriett, Pesek Isabel, Journal of Gerontology, 10, 41, 1955.
13. Wallace H. A., Snedecor George W., Correlation and Machine Calculation. Iowa State College Official Publication 30, No. 4, 1925. Revised, 1931.
14. Yates F., Journal of the Ministry of Agriculture, 43, 156, 1936.

КРИВОЛИНЕЙНАЯ РЕГРЕССИЯ

1. Введение. Хотя в большинстве случаев линейная регрессия вполне удовлетворяет потребностям исследования, тем не менее наблюдаются случаи, когда переменные не находятся друг с другом в такой простой связи. Решение задачи более или менее точного описания совместного варьирования двух или нескольких величин приводит к проблемам подбора кривой, характеризующей связи и известной под названием криволинейной регрессии. С этой общей точки зрения определение прямой линии регрессии является наиболее простым и, по-видимому, наиболее широко применимым частным случаем.

Существуют самые различные поводы для подыскания кривой к данным, имеющим нелинейный характер изменения. Иногда это делается с целью получения возможно лучшей оценки зависимой переменной для некоторого частного значения независимой переменной. Сюда входит и задача сглаживания варьирующих данных, а также интерполяция для определения значений Y , соответствующих не вошедшим в серию наблюдаемым значениям X . Иногда целью такого подбора кривой является испытание или раскрытие закона связи между переменными, например установление закономерностей роста. В других случаях форма соотношения между переменными сама по себе не представляет никакого интереса; здесь конечной целью является просто элиминирование тех неточностей, которые вносятся нелинейностью регрессии в коэффициент корреляции или в ошибку эксперимента.

К задаче подбора кривых к наблюдаемым данным можно подойти с различных сторон. Исключая те случаи, когда производится проверка определенного закона, основная трудность задачи состоит в выборе подходящего для данных условий уравнения. Для лиц, имеющих некоторое знакомство с математикой, известную пользу и облегчение работы могут дать графические способы выравнивания [12] и некоторые более теоретические методы [4, 11, 18]. Здесь мы в основном рассмотрим только вопрос об экспоненциальной (степенной) кривой роста и о полиномиальных кривых.

2. Экспоненциальная кривая роста. Характерной особенностью некоторых наиболее простых процессов роста является то, что увеличение соответствующего показателя в каждый данный момент времени пропорционально достигнутому к этому времени его размеру. В этом случае иногда говорят о неизменности темпов роста. В некоторой фазе развития культуры бактерий замечается факт подчиненности численностей организмов этому закону. Прекрасной иллюстрацией для этой закономерности могут служить данные таблицы 193 о сухом весе эмбрионов цыплят между 6-м и 16-м днем развития [16]. Кривая этого веса, изображенная на рисунке 26, с течением времени развития поднимается вверх с увеличивающейся быстротой; соответствующее уравнение регрессии имеет такую общую для данного процесса роста форму:

$$W = A \cdot B^x,$$

где A и B являются константами, подлежащими определению. Логарифмируя это уравнение, находим

$$\log W = \log A + X \log B$$

или

$$Y = a + bX,$$

где $Y = \log W$, $a = \log A$ и $b = \log B$. Это означает, что если вместо W для последовательных значений Y нанесены на график $\log W$, то получится прямая линия. Этот способ использования логарифма вместо самой величины называется способом *спрямления кривой*.

В последнем столбце таблицы 193 приведены $Y = \log W$, которые на рисунке показаны в виде точек, соответствующих значениям X . Уравнение регрессии, вычисленное обычным путем по данным колонок X и Y , в данном случае будет:

$$Y^1 = 0,1959X - 2,689.$$

Эта линия регрессии проходит через точки наблюдений с редко встречающейся точностью; корреляция между Y и X здесь равна 0,9992. Отсюда следует вывод, что эмбрион цыпленка, характеризуемый сухим его весом, увеличивается в полном соответствии с экспоненциальным законом, а логарифм живого веса увеличивается с установленной скоростью 0,1959 в день.

Рис. 27. Подбор кривых для сухого веса эмбрионов цыплят в возрасте 6—16 дней. В натуральном масштабе: $W = 0,002046 \times 1,57^x$. В логарифмическом масштабе: $Y = 0,1959X - 2,689$.

Часто само уравнение регрессии не представляет большого интереса, и задачей исследования является просто констатация того, что данные подчиняются экспоненциальному закону. В этих случаях достаточную информа-

ТАБЛИЦА 193

Сухой вес эмбрионов цыплят между 6-м и 16-м днем развития, а также десятичные логарифмы этих данных

Дней с начала развития X	Сухой вес W (в г)	Десятичный логарифм веса Y
6	0,029	-1,538
7	0,052	-1,284
8	0,079	-1,102
9	0,125	-0,903
10	0,181	-0,742
11	0,261	-0,583
12	0,425	-0,372
13	0,738	-0,132
14	1,130	0,053
15	1,882	0,275
16	2,812	0,449

¹ Здесь следует различать наблюдаемое Y от оценки Y . Символ \hat{Y} будет применяться только тогда, когда речь будет идти об отклонении от регрессии $Y - \hat{Y}$.

цию дает простой график. При этом применение специальной бумаги с *полулогарифмическим* графлением позволяет избежать обращения к таблицам логарифмов. На бумаге с таким графлением горизонтальная шкала построена так, что нанесение данных, подчиняющихся экспоненциальному закону роста, приводит к прямой линии. Такая бумага может быть приобретена в большинстве обычных магазинов.

Иногда значительные удобства представляет такая форма экспоненциального закона роста:

$$W = Ae^{bx},$$

где e — основание натуральных, или неперовских, логарифмов, являющееся некоторым иррациональным числом, приближенное значение которого 2,718. Если решено пользоваться этой формой закона, то удобнее применять натуральные логарифмы:

$$\log_e W = \log_e A + bX.$$

Можно показать, что коэффициент регрессии этого уравнения b является *относительным* темпом роста W , т. е. темпом роста в пересчете на единицу X и на единицу W . Этот относительный темп роста в 2,3026 ($=\log_e 10$) раза больше темпа, определяемого наклоном линии регрессии, выраженной через обычные логарифмы. Так, для эмбрионов цыплят относительный темп роста составляет $2,3026 \times 0,1959 = 0,4511$ гна день и на 1 г достигнутого веса. Отсюда следует, что представление об относительном темпе роста может быть получено как из той, так и другой формы уравнения регрессии. Всегда можно, если это считается более удобным, использовать обычные логарифмы, после чего в результат нужно ввести только поправку при помощи указанного выше множителя. Применение натуральных логарифмов не имеет никаких других преимуществ, кроме того, что работа с таблицами этих логарифмов проводится более быстро, чем с обычными логарифмами.

Если возникает необходимость в переходе от уравнения регрессии $\log W = 0,1959 X - 2,689$ к его экспоненциальной форме¹

$$W = 0,002046 \times 1,57^X$$

или к форме²

$$W = 0,002046 e^{0,4511X},$$

то следует помнить, что сумма квадратов ошибок для этой экспоненциальной формы не будет минимальной. Здесь приводится к минимуму сумма квадратов разностей: $\log W$ минус оценка $\log W$. Для тех случаев, когда необходимо приведение к минимуму суммы квадратов отклонений от самого экспоненциального выражения закона роста, имеются специальные методы (например, в [18, 4]). Все же обычно не возникает необходимости в переходе к экспоненциальной форме уравнения или в проведении несущественных исправлений констант для получения минимальной суммы квадратов отклонений, если этот переход произведен.

Исследователи, работающие в других областях, могут проявить интерес к самым различным видам теоретических уравнений. Иногда берется просто обратная пропорциональность:

$$Y = \frac{1}{X}.$$

Другие возможные случаи дают уравнения $Y = \log X$; $Y = \sqrt{X}$ и $\log Y = a + b \log X$. В этих случаях сначала надо произвести проверку предполагаемого

¹ Для установления этой формы следует иметь в виду, что $0,1959 = \log 1,57$ и $2,689 = \log 488,65$, отсюда $\log W = X \log 1,57 - \log 488,65 = \log 1,57^X - \log 488,65 = \log \frac{1,57^X}{488,65} = \log [(0,002046) \times 1,57^X]$.

² К этому мы приходим из предыдущей формы, так как $\log_e 1,57 = 0,4511$; следовательно, $1,57 = e^{0,4511}$.

закона на графике. В зависимости от того, окажутся ли данные, полученные при помощи этих преобразований одной или обоих переменных, выравненными или нет, можно приступить, если в этом есть необходимость, к расчетам уравнения. Для проверки последнего из приведенных выше уравнений весьма удобна логарифмическая бумага, у которой как по горизонтали, так и по вертикали нанесена логарифмическая шкала. По поводу указаний о применении бумаги со специальными видами графления см. литературу [8].

Пример 1. Дж. У. Гауен и У. К. Прайс подсчитали число мест поражения растений аукуба (кустарник семейства *Sorpaceae*—Прим. ред.) вирусом мозаики, развившихся после облучения X-лучами в течение различных промежутков времени. Следующие данные взяты с любезного разрешения исследователей из не опубликованных пока результатов.

Минут облучения	0	3	7,5	15	30	45	60
Численность (в сотнях) . .	271	226	209	108	59	29	12

Нанесите на график в качестве ординат эти первоначальные данные, а после этого их логарифмы. Определите регрессию $Y=2,432-0,02227X$, где Y —логарифм численности и X — число минут облучения.

Пример 2. Повторите подбор кривой, применяя натуральные логарифмы. Проверьте, что скорость уменьшения мест поражения (в сотнях) в пересчете на минуту и на 100 наблюдений равна $2,3026 \times 0,02227=0,05128$.

Пример 3. Если смысл относительного темпа изменения не вполне ясен, то представление о нем можно получить, применяя приближенный метод его непосредственного расчета. Увеличение веса у эмбриона цыпленка на тринадцатый день развития составляет $1,130-0,738=0,392$ г; это значит, что средний темп роста в этот период равен 0,392 г в день. Но средний вес в этот же период равен $(1,130+0,738) : 2=0,934$ г. Относительный темп, или скорость, роста каждого грамма будет, следовательно, составлять $0,392 : 0,934=0,42$ г в день на 1 г. Это значение отличается от средней, полученной для всего периода между 6-м и 16-м днем развития — 0,4511, что происходит отчасти потому, что как средний вес эмбриона, так и увеличение на тринадцатый день развития подвержены некоторому случайному варьированию, а также частично потому, что точное значение относительного темпа роста основывается на данных о весе и об увеличении в весе в любой сколь угодно малый интервал времени, а не в один день.

Пример 4. Геддес [10] дает оценку хлебопекарного качества муки простого помола (Q) после прогрева ее при $170^\circ F$ в течение различного количества часов (T). Парные наблюдения (T и Q) были такими: 0,25, 93; 0,50, 71; 0,75, 63; 1,0, 54; 1,5, 43; 2,0, 38; 3,0, 29; 4,0, 26; 6,0, 22; 8,0, 20. Нанесите соответствующие точки, откладывая Q по вертикальной, а T по горизонтальной оси. Полагая $\log T=X$ и $\log Q=Y$, нанесите точки Y в соответствии с X . Определите регрессию $Y=1,7116-0,4678X$.

Пример 5. Положив в предыдущем примере $Y=1/Q$, нанесите Y в соответствии с T . Так как этот график оказывается нелинейным, то испробуйте уравнение $Y=\frac{1}{Q-14}$.

Определите регрессию таких спрямленных данных $Y=0,01996 T+0,00546$, из которой следует, что $Q=\frac{1}{0,01996 T+0,00546}+14$. Раннинг [14, VI] дал графический метод определения числа, которое должно вычитаться из W для того, чтобы получилась линейная регрессия. Можете ли вы улучшить результат после введения числа 14? Сколько степеней свободы остается для оценки у среднего квадрата ошибки?

Пример 6. Деккер и Андре [3] изучали гибель клопа-черепашки при температуре $-12,2^\circ C$, удерживавшейся в течение различных промежутков времени:

Число часов	0,25	0,5	1	4	12	24	48	72
Процент гибели . . .	59	63	65	68	70	73	74	75

Пусть Y — процент гибели и X — логарифм числа часов воздействия на насекомых пониженной температуры. Определите для этого случая уравнение $Y=6,051X+64,1$. Нанесите эти данные на логарифмическую бумагу, взяв логарифмическую шкалу по горизонтали и располагая на ней число часов.

Пример 7. В одном опыте с пшеницей в Австралии применялись различные дозы удобрения; получены такие урожайные данные:

Доза X	0	10	20	30	40	60
Урожай y	26,0	34,0	36,3	37,8	38,6	38,9

Подберите к этим данным уравнение Митчерлиха [14, 5] $Y = a - bX$, где $Y = \log(39 - y)$. Урожай 39 здесь берется в качестве максимального урожая, какой бы ни была доза удобрения. Определите уравнение $Y = 0,01434X - 0,2294$.

3. Полином второго порядка. Рассматривая нелинейную регрессию, приходится часто встречаться со случаями, когда теоретическая форма ее

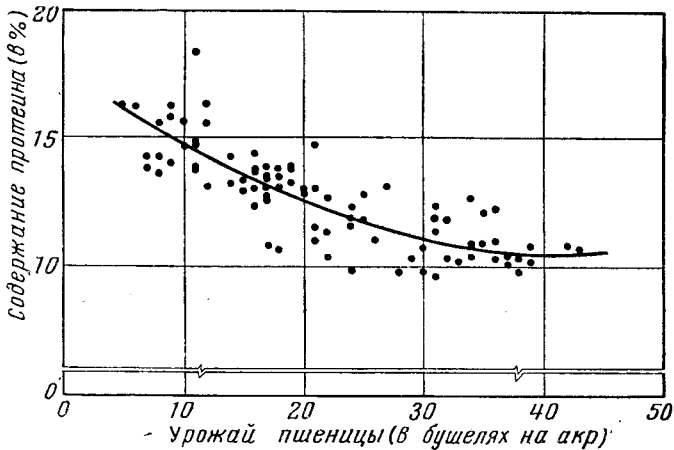


Рис. 28. Регрессия содержания протеина на урожай пшеницы, наблюдения по 91 участку

$$Y = 17,703 - 0,3415X + 0,004075X^2.$$

уравнения остается неизвестной. Во многих исследованиях при этих условиях можно удовлетвориться подбором к наблюдаемым данным полинома второго порядка:

$$Y = a + bX + cX^2.$$

Соответствующий график дает параболу, ось которой направлена по вертикали, но обычно в процессе подбора кривой берется только небольшой отрезок этой параболы. Теперь вместо спрямления данных производится добавление третьей переменной — квадрата X^2 . Тем самым задача сводится к применению методов множественной регрессии. Процесс вычислений точно такой же, как и в главе 14, причем в этом случае в качестве независимых переменных выступают X и X^2 . Следует также отметить, что если того требует исследование, то вместо X^2 могут быть взяты \sqrt{X} , $\log X$ или $1/X$.

В качестве иллюстрации метода и некоторых его приложений ниже (в табл. 194) приводятся данные об урожае пшеницы и содержании протеина [13]; те же данные представлены на рисунке 28. Задача исследователя состояла в определении содержания протеина при различной высоте урожая. Кроме того, мы произведем оценку существенности отклонения данных от линейности.

Во втором столбце таблицы приведены квадраты урожаев столбца 1. Эти квадраты при обработке во всех отношениях подобны третьей переменной в множественной регрессии. Вычисленное обычным способом уравнение регрессии

$$Y = 17,703 - 0,3415X + 0,004075X^2$$

Содержание протеина в процентах (Y) и урожай (X) пшеницы на 91 делянке¹

Урожай (в бушелях на 1 акр) X	Квадрат X ²	Процент протеина Y	Урожай (в бушелях на 1 акр) X	Квадрат X ²	Процент протеина Y
43	1849	10,7	20	400	13,0
42	1764	10,8	19	361	13,9
39	1521	10,8	19	361	13,2
39	1521	10,2	19	361	13,8
38	1444	10,3	18	324	10,6
38	1444	9,8	18	324	13,0
37	1369	10,1	18	324	13,4
37	1369	10,4	18	324	13,7
36	1296	10,3	18	324	13,0
36	1296	11,0	17	289	13,4
36	1296	12,2	17	289	13,5
35	1225	10,9	17	289	10,8
35	1225	12,1	17	289	12,5
34	1156	10,4	17	289	12,7
34	1156	10,8	17	289	13,0
34	1156	10,9	17	289	13,8
34	1156	12,6	16	256	14,3
33	1089	10,2	16	256	13,6
32	1024	11,8	16	256	12,3
32	1024	10,3	16	256	13,0
32	1024	10,4	16	256	13,7
31	961	12,3	15	225	13,3
31	961	9,6	15	225	12,9
31	961	11,9	14	196	14,2
31	961	11,4	14	196	13,2
30	900	9,8	12	144	15,5
30	900	10,7	12	144	13,1
29	841	10,3	12	144	16,3
28	784	9,8	11	121	13,7
27	729	13,1	11	121	18,3
26	676	11,0	11	121	14,7
26	676	11,0	11	121	13,8
25	625	12,8	11	121	14,8
25	625	11,8	10	100	15,6
24	576	9,9	10	100	14,6
24	576	11,6	9	81	14,0
24	576	11,8	9	81	16,2
24	576	12,3	9	81	15,8
22	484	11,3	8	64	15,5
22	484	10,4	8	64	14,2
22	484	12,6	8	64	13,5
21	441	13,0	7	49	13,8
21	441	14,7	7	49	14,2
21	441	11,5	6	36	16,2
21	441	11,0	5	25	16,2
20	400	12,8			

¹ Взято из опубликованного графика, поэтому между корреляцией, к которой приходим мы, и той, что приведена у автора, замечается небольшое различие.

дает кривую, помещенную на рисунке 28. Здесь еле-еле заметна очень небольшая роль члена с урожаем второй степени, что нашло свое выражение в малом значении соответствующего коэффициента; график снижается почти по прямой линии. Однако в правой стороне рисунка член регрессии с X² придает кривой почти горизонтальное направление, т. е. начинает играть заметную роль.

Представляет интерес определить влияние криволинейности регрессии на систему отклонений от регрессии. Если здесь допустить линейную регрессию, то $\sum d_{y.x}^2 = 110,48$; но когда вводится член второй степени, то $\sum d_{y.x}^2 = 97,53$.

Соответствующий дисперсионный анализ и оценка существенности этого члена приведены в таблице 195. Здесь оказывается, что уменьшение суммы квадратов, сопоставляемое с остаточным после введения криволинейной регрессии средним квадратом, является существенным. Следовательно, необходимо отказаться от гипотезы о линейной регрессии; это говорит в пользу криволинейности данной регрессии.

ТАБЛИЦА 195

Оценка существенности отклонения от линейной регрессии

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Отклонения от линейной регрессии	89	110,48	1,11
Отклонения от криволинейной регрессии	88	97,53	
Криволинейность регрессии	1	12,95	12,95**

$$F = 12,95 : 1,11 = 11,7$$

Для некоторого определенного урожая, положим для 20 бушелей на 1 акр, ожидаемое содержание протеина будет:

$$Y = 17,703 - 0,3415 \times 20 + 0,004075 \times 20^2 = 12,51\%.$$

Этим путем можно по уравнению регрессии найти оценки наблюдаемых значений и произвести интерполяцию. Доверительные заключения и проверки гипотез производятся по правилам главы 14.

Может сложиться впечатление о том, что между спрямлением данных в предыдущем параграфе и криволинейностью регрессии, с которой мы познакомились здесь, существует принципиальное различие. Однако это различие в действительности только внешнее. Если представить себе график в трехмерном пространстве, в котором X и X^2 имеют две самостоятельные оси координат, то поверхность регрессии будет плоскостью. В этом случае данные в известном смысле будут, как и прежде, спрямленными. Следовательно, только ограниченность двухмерного графика делает эти методы внешне различными.

Применение регрессии, будет ли она линейной или криволинейной, требует особой осторожности. Имеющиеся данные могут быть недостаточными для того, чтобы быть источником сведений о направлении процесса вне рассматриваемого интервала. Рассматривая, например, рисунок 27 и соблазняясь превосходным подбором кривой к наблюдаемым точкам, можно сделать попытку распространить этот закон роста и на промежуток времени до шестого дня развития или после шестнадцатого дня его. Однако это будет слишком поспешное суждение о скорости роста в двух указанных промежутках времени. Экстраполяция обычно требует широких знаний предмета исследования и глубокой вздумчивости.

Пример 8. Произведите подбор полинома второго порядка к данным примера 1 этой главы. Результат $Y = 262,245 - 9,851X + 0,09681 X^2$ дает не очень хорошее приближение.

Пример 9. Деккер и Андре [3] выдерживали восемь групп по 500 особей клопа черепашки при $-6,6^\circ \text{C}$ в течение различных интервалов времени. В результате гибель насекомых была такой:

Число дней	1	3	3,5	8	13,5	16	21	28
Процент гибели	0,8	3,6	5,8	11,6	22,8	44,6	67,6	92,0

Произведите подбор полинома второго порядка в целях установления, имеет ли здесь место существенное отклонение от линейной регрессии. Значение $F = 6,09$ при $j = 1$ и 5 указывает на то, что отклонение от линейной регрессии, поскольку оно определяется установленной параболой, находится несколько ниже уровня существенности.

Пример 10. Вы можете подумать, что парабола не совсем подходящая кривая для данных последнего примера. Тогда возьмите уравнение $Y = a + bX$, где Y логарифм процента гибели насекомых и X число дней воздействия пониженной температуры.

Как вы произведете сравнение результата при подборе этой кривой с результатом подбора параболы? Для последней средней квадрат отклонения от регрессии второго порядка был равен 32,2 при $f=5$. Чтобы получить сравнимые результаты для логарифмической кривой, следует каждое из \hat{Y} преобразовать в соответствующий антилогарифм P и после этого вычислить отклонения от регрессии. Средний квадрат этих отклонений 1270 значительно больше соответствующего показателя для отклонений от регрессии второго порядка. Почему здесь отсутствует возможность прямого и точного сравнения.

4. Оценка отклонения от линейной регрессии. В предыдущем параграфе эта оценка произведена в таблице 195 после того, как был проведен расчет криволинейной регрессии. Но оценка может быть произведена гораздо проще, если данные собраны в группы и если нас интересует только вопрос об отклонении регрессии от линейности.

Соответствующий метод показан в применении к данным таблицы 196. полученным нами благодаря любезности Б. Дж. Воса и У. Т. Досона. В опыте применены четыре дозы инъецируемого уабанна, каждая из которых удваивает предыдущую. Один из частных вопросов, стоявших перед исследованием, состоял в установлении, не будет ли регрессия Y на X криволинейной. Простой график в данном случае вполне достаточен для того, чтобы убедиться в том, что здесь нет ничего, что свидетельствовало бы о криволинейности. Однако в порядке иллюстрации метода мы все же проведем оценку существенности этой криволинейности.

ТАБЛИЦА 196

Смертельная доза (кодированная путем вычитания 50 единиц) стандартного в США уабанна¹ при медленной внутривенной инъекции в целях прекращения сердечной деятельности кошки

	Инъецируемая доза уабанна — $X \frac{\text{мг/кг/мин}}{1000}$				Итого
	1,04575	2,0915	4,183	8,366	
	5	3	34	51	
	9	6	34	56	
	11	22	38	62	
	13	27	40	63	
	14	27	46	70	
	16	28	58	73	
	17	28	60	76	
	20	37	60	89	
	22	40	65	92	
	28	42			
	31	50			
	31				
ΣY	217	310	435	632	1534
Число кошек	12	11	9	9	41
\bar{y}	18,1	28,2	48,3	70,2	38,9
ΣY^2	4727	10788	22261	45940	83716

¹ Уабанн — ядовитый глюкозид, получаемый из семян *Strophantus gratus* и из древесины *Asokautheria schimperi* (Прим. ред.).

В таблице 197 сначала обычным способом производится разложение общей суммы квадратов для смертельных доз препарата. Из этого анализа следует, что главный источник варьирования обусловлен различными доза-

*Дисперсионный анализ смертельных доз уабаина
при четырех инъекруемых дозах*

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Внутри групп (ошибка)	37	5 651	152,7
Дозы уабаина	3	16 093	5 364**
Итого	40	21 744	

ми уабаина (между групповыми средними имеются существенные различия). Теперь мы ставим вопрос о том, в какой мере существенна тенденция к отклонению этих средних от линейности.

Анализ суммы квадратов групповых средних

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Дозы уабаина (из табл. 197)	3	16 093	
Линейная регрессия $(\Sigma xy)^2/\Sigma x^2$	1	15 700	
Отклонения от линейной регрессии	2	393	196
Ошибка (из табл. 197)	37		153

$$F = 196/153 = 1,3.$$

Для определения линейной регрессии необходимо вычислить значения Σx^2 и Σxy , соответствующие табличной сумме $\Sigma y^2 = 21\,744$. В связи с различной численностью Y в группах вычисления этих сумм несколько отличаются от обычной схемы. Так, сумма

$$\Sigma X = 12 \times 1,04575 + 11 \times 2,0915 + 9 \times 4,183 + 9 \times 8,366 = 148,496$$

является *взвешенной* суммой. Подобно этому

$$\Sigma X^2 = 12 \times 1,04575^2 + \dots + 9 \times 8,366^2 = 848,628.$$

Отсюда находим:

$$\Sigma x^2 = \Sigma X^2 - (\Sigma X)^2/n = 848,628 - 537,831 = 310,797.$$

То же самое для суммы произведений:

$$\Sigma XY = 1,04575 \times 217 + 2,0915 \times 310 + 4,183 \times 435 + 8,366 \times 632 = 7982,21.$$

Вычитая отсюда поправочный член $148,496 \times 1594/41 = 5773,23$, имеем:

$$\Sigma xy = 7982,21 - 5773,23 = 2208,98.$$

После этого в таблицу 198 переносится из таблицы 197 сумма квадратов групповых средних, где она подразделяется на две части, одна из которых относится к линейной регрессии, а вторая — к отклонениям от нее. Первая,

как обычно, определяется по формуле $(\sum xy)^2 / \sum x^2 = 2208,98^2 : 310,797 = 15700$ и имеет одну степень свободы. Остаток 393 с двумя степенями свободы дает нам средний квадрат, соответствующий отклонениям от линейной регрессии. Этот средний квадрат только немного больше, чем средний квадрат ошибки, что указывает на то, что отклонение от линейности обусловлено случайным выборочным варьированием.

Пример 11. В отношении данных примера 6 о гибели клопа-черепашки имеется возможность получить некоторую дополнительную информацию, так как каждый из приведенных ранее восьми процентов гибели насекомых являлся средним из пяти субвыборок по 100 насекомых в каждой. Ниже приводятся проценты гибели по этим субвыборкам.

Субвыборки по 100 насекомых	Число дней воздействия пониженной температуры							
	1	3	3,5	8	13,5	16	21	28
1	0	3	4	9	21	46	68	93
2	0	4	5	21	23	39	70	100
3	1	2	6	8	27	43	66	92
4	1	5	7	10	25	47	61	86
5	2	4	7	10	18	48	73	89
Итого . .	4	18	29	58	114	223	338	460

Дисперсионный анализ для этих данных таков:

внутри сроков 421,2; $f=32$; средний квадрат 13,2

между сроками 39934,4; $f=7$; средний квадрат 5704,9

Хотя здесь наблюдается некоторая неоднородность материала, все же средний квадрат 13,2 может быть взят в качестве ошибки для оценки любого среднего квадрата, взятого из примера 9. Так, $F=32,2/13,2=2,44$; $f=5$ и 32 указывает на то, что отклонение этих данных от регрессии второго порядка не является существенным.

Пример 12. Суонсон и Смит [49] определили содержание общего азота (Y г на 100 куб. см плазмы) в плазме крови крыс девяти возрастов (X дней).

Номер крысы	Возраст (дни)								
	25	37	50	60	80	100	130	180	360
1	0,83	0,98	1,07	1,09	0,97	1,14	1,22	1,20	1,16
2	0,77	0,84	1,01	1,03	1,08	1,04	1,07	1,19	1,29
3	0,88	0,99	1,06	1,06	1,16	1,00	1,09	1,33	1,25
4	0,94	0,87	0,96	1,08	1,11	1,08	1,15	1,21	1,43
5	0,89	0,90	0,88	0,94	1,03	0,89	1,14	1,20	1,20
6	0,83	0,82	1,01	1,01	1,17	1,03	1,19	1,07	1,06
7	0,84	0,95	0,93		0,98	1,08	1,19	1,13	1,29
8	0,75	0,92	1,07		0,99	0,98	1,14	1,12	1,25
9	0,67	0,87	1,03			0,98	1,13		1,23
10	0,70		1,13			1,10	1,06		1,22
11	0,77		1,05				1,11		1,17
12	0,76		0,96				1,04		
13	0,75		1,01				1,29		
14	0,78		1,01				1,09		
15	0,76		0,94				1,14		
16	0,86		1,03				1,12		
17	0,78		1,06				1,29		
18	0,84		1,08				1,10		
19	0,85		1,09				1,07		
20	0,83		0,93				1,10		
21	0,85		0,95						
22	0,85								
23	0,83								
24	0,83								
25	0,70								

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Итого	118	2,8686	
Возраст	8	2,3076	
Животные внутри возрастных групп (ошибки)	110	0,5610	0,0051
Возраст	8	2,3076	
Линейная регрессия	1	1,4851	
Отклонения	7	0,8225	0,1175**

5. Оценка отклонения от линейной регрессии при ковариационном анализе. В тех случаях, когда наряду с другими применениями учения о корреляции и регрессии возникает необходимость проверки того, что при ковариационном анализе регрессия линейна, приходится прибегать к некоторой модификации описанных выше методов. В некоторых случаях, например, рендомизированные блоки могут обладать только случайной регрессией, обусловленной ошибками эксперимента. Если в этих случаях на основе такой регрессии произвести выравнивание, то приведенные средние будут иметь характер правильных оценок, а соответствующий критерий существенности будет обладать вполне законной силой.

В качестве иллюстрации возьмем ранее использованные данные таблицы 176 о числе корней и урожае сахарной свеклы. Положим, что агроном, ознакомившийся с результатами ковариационного анализа, выразил свое удивление по поводу того обстоятельства, что различия между приведенными средними оказались несущественными. В примере 15 главы 13 этот вопрос обсуждался в двух аспектах. Теперь в дополнение к этому мы можем, применяя метод, подобный тому, который дан в параграфе 3 этой главы, произвести изучение вопроса о линейности регрессии. Для этого введем в качестве второй независимой переменной квадрат числа растений X^2 и поставим задачу определить, содержится ли в Σy^2 сколько-нибудь существенная часть, обусловленная криволинейностью регрессии для ошибок опыта.

Эти квадраты числа растений внесем в расчетную таблицу, взяв за единицу 1000, например: $(183)^2 = 33489$ будет внесено в виде $X_2 (= X^2) = 33$. Полученные таким образом три переменные: X_1 — число корней; $X_2 = X_1^2/1000$ и Y — урожай корней в тоннах на акр (по рендомизированным блокам) обрабатываются подобно тому, как это было в параграфе 5 главы 14. Для «ошибки» результаты будут такими:

	X_1	X_2	Y
X_1	16 873,13	8 746,25	381,2750
X_2		4 621,00	202,5650
Y			11,7711

Определение этой множественной регрессии приводит к данным второй строки таблицы 199. Проверяемая в этом случае нулевая гипотеза $\beta_{2.1} = 0$ имеет тот смысл, что криволинейность регрессии отсутствует. Как было установлено и при прежнем обсуждении вопроса, нет ничего, что говорило бы против этой гипотезы.

В других случаях таблицы с двумя входами может оказаться, что регрессия для ошибки будет криволинейной. В этих случаях может возникнуть необходимость в оценке существенности различий между средними, исправленными по криволинейной регрессии. Для этого следует в таблице дисперсионного анализа сложить суммы квадратов и произведений по строкам «варианты» и «ошибки» и произвести расчет Σd_{ij}^2 для множественности

Оценка существенности отклонений от линейной регрессии
для ошибки опыта. Данные об урожае сахарной свеклы
(параграф 6 главы 13)

Источник варьирования	Отклонения от регрессии		
	число степеней свободы	сумма квадратов	средний квадрат
Линейная регрессия (табл. 176)	14	$\Sigma d_{Y..1}^2 = 3,1558$	
Криволинейная регрессия	13	$\Sigma d_{Y..12}^2 = 2,8282$	0,2176
Криволинейность (X_2)	1	0,3276	0,3276

$$F = 0,3276/0,2176 = 1,51; f = 1 \text{ и } 13$$

регрессии. Применение этого процесса к данным об урожайности сахарной свеклы даст таблицу 200.

Вследствие незначительной криволинейности полученное здесь F очень мало отличается от значения F в таблице 176. В тех случаях, когда регрессия ошибки имеет несомненную кривизну, приведенные средние могут иметь как большее, так и меньшее варьирование по сравнению с неприведенными средними.

ТАБЛИЦА 200

Оценка существенности разностей между средними, исправленными
по криволинейной регрессии. Данные о сахарной свекле

Источник варьирования	Отклонения от регрессии		
	число степеней свободы	сумма квадратов	средний квадрат
Варианты + ошибка	16	$\Sigma d_{Y..12}^2 = 4,8248$	
Ошибка (табл. 199)	13	2,8282	0,2176
Для оценки вариантов	3	1,9966	0,6655

$$F = 3,06; f = 3 \text{ и } 13$$

Пример 13. Оценка криволинейности в случае множественной регрессии проводится теми же способами, которые описаны в двух последних параграфах. Эммерт [6] определял количество (частей на миллион) азота N и фосфора P в стеблях картофеля на 15 июня 1935 г.; после этого по каждой из 37 делянок был определен урожай картофеля Y в фунтах. Множественная регрессия Y на N и P дала $\Sigma d_{Y..12}^2 = 2593$. Так как имелись некоторые указания на криволинейность зависимости Y от N и P , то были дополнительно введены переменные N^2 и P^2 . Множественная регрессия Y на N , P , N^2 и P^2 дала $\Sigma d_{Y..1234}^2 = 2414$. Это уменьшение остаточной суммы квадратов указывает на наличие некоторого отклонения от линейности. Применение критерия t показало, что это уменьшение обусловлено только членом, относящимся к переменной P^2 , а член с N^2 никакой дополнительной информации не дал. Поэтому последний член был исключен, после чего регрессия Y на N , P и P^2 привела к $\Sigma d_{Y..123}^2 = 2414$, т. е. к тому же, что и ранее, числу. Результат оценки существенности таков:

Источник варьирования	Число степеней свободы	Σd^2	Средний квадрат
Регрессия Y на N и P	34	2593	
Регрессия Y на N , P и P^2	33	2414	73,15
Криволинейность	1	179,4	179,4

$$F = 179,4/73,15 = 2,45; f = 1 \text{ и } 33.$$

Величина F оказалась ниже 5%-ного уровня, поэтому отклонение от линейности следует приписать случайностям выборки. Исходные данные приводятся ниже.

N	P	P^2	Y	N	P	P^2	Y
139	195	38	33	330	260	68	46
143	146	21	43	380	134	18	60
147	120	14	40	436	130	17	57
154	178	32	49	436	200	40	64
162	266	71	43	470	134	18	55
182	200	40	27	476	160	26	54
186	260	68	55	555	155	24	37
200	240	58	51	556	188	35	42
208	170	29	45	626	125	16	45
213	195	38	41	700	146	21	63
233	146	21	57	750	178	32	52
257	160	26	38	834	188	35	46
264	146	21	48	834	178	32	56
264	146	21	39	908	105	11	65
264	160	26	40	1125	160	26	46
264	250	62	38	1200	122	15	58
265	175	31	52	1250	140	20	80
286	114	13	49	1625	150	22	70
312	155	24	51				

$$Y = 0,0170 N - 0,440 P + 1,10 P^2 + 82,5.$$

6. Определение ортогональных полиномов. Фишером предложены [8] специальные методы подыскания полиномов вида

$$Y = a + bX + cX^2 + dX^3 + \dots$$

В данном случае мы имеем довольно гибкую кривую, приспособленную для сглаживания самых различных данных¹. Прямая линия и парабола являются ее частными случаями. Фишеровские методы основываются на принципах ортогональных полиномов. Один из них описан в [9], где приведены и соответствующие таблицы. Другой метод, излагаемый ниже, является более общим, но приводит иногда к более сложным вычислениям.

Иллюстрацию этого метода мы здесь дадим только для того частного случая, когда X принимает равноотстоящие значения с интервалом, равным единице, и когда каждому X соответствует только одно значение Y . Если X принимает равноотстоящие значения с интервалом, отличающимся от единицы, то путем кодирования при помощи деления каждого значения X на величину этого интервала можно привести интервал к единице. В случаях же, когда X расположены не на одинаковом расстоянии или когда некоторым или всем значениям X соответствуют по нескольку значений Y , приходится применять метод параграфа 3 этой главы, распространяя его на большее, чем два, число переменных X .

Основным преимуществом ортогональных полиномов является то, что подбор кривой проходит через последовательные стадии, на каждой из которых определяется член все более и более высокой степени и производится оценка существенности его. В примере, приводимом ниже, подбор кривой заканчивается членом четвертой степени. Тот, кто пожелал бы иметь формулы для членов более высокой степени, может их вывести с помощью указа-

¹ Метод параболического интерполирования с помощью ортогональных полиномов задолго до Р. Фишера разработан акад. П. Чебышевым. Р. Фишер только конкретизировал этот метод, доведя до рабочих формул. Но даже для наиболее простого, рассматриваемого в этой книге случая последовательная система формул столь сложна, что вряд ли метод Р. Фишера имеет сколько-нибудь заметные преимущества перед обычным способом решения системы нормальных уравнений. Если же применить более компактные схемы решения нормальных уравнений, то потребность в формулах Фишера не так уже велика (*Прим. перев.*)

ний, приведенных в сносках. Общие формулы даны в книге профессора Фишера.

Для иллюстрации возьмем данные таблицы 193 об эмбрионах цыплят. В первую колонку таблицы 201 перенесены данные о сухом весе эмбрионов.

ТАБЛИЦА 201

Суммы и производные от них величины, используемые при подборе полиномов к данным об эмбрионах цыплят

У	2	3	4	5
0,029	0,029	0,029	0,029	0,029
0,052	0,081	0,110	0,139	0,168
0,079	0,160	0,270	0,409	0,577
0,125	0,285	0,555	0,964	1,541
0,181	0,466	1,021	1,985	3,526
0,261	0,727	1,748	3,733	7,259
0,425	1,152	2,900	6,633	13,892
0,738	1,890	4,790	11,423	25,315
1,130	3,020	7,810	19,233	44,548
1,882	4,902	12,712	31,945	76,493
2,812	7,714	20,426	52,371	128,864
$S_1=7,714$	$S_2=20,426$	$S_3=52,371$	$S_4=128,864$	$S_5=302,212$
$a=0,701273;$	$b=0,309485;$	$c=0,183115;$	$d=0,128735;$	$e=0,100637.$
$a'=0,701273;$	$b'=0,391788;$	$c'=0,139048;$	$d'=0,031838;$	$e'=0,003066.$
$A=0,701273;$	$B=0,235073;$	$C=0,0463493;$	$D=0,00619072;$	$E=0,00038325.$

Сумма $S_1=7,714$, деленная на число наблюдений $n=11$, дает среднюю $a=0,701273$, что является применением первой формулы из такого их ряда:

$$a = \frac{S_1}{n} = \bar{y};$$

$$b = \frac{1 \cdot 2}{n(n+1)} S_2;$$

$$c = \frac{1 \cdot 2 \cdot 3}{n(n+1)(n+2)} S_3;$$

$$d = \frac{1 \cdot 2 \cdot 3 \cdot 4}{n(n+1)(n+2)(n+3)} S_4;$$

$$e = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{n(n+1)(n+2)(n+3)(n+4)} S_5.$$

Остальные формулы будут использованы в последующем.

Средняя a переносится в следующую строку, но уже в качестве a' ; в соответствии с первой из ряда приводимых ниже формул эти две величины равны между собой:

$$a' = a$$

$$b' = a - b$$

$$c' = a - 3b + 2c$$

$$d' = a - 6b + 10c - 5d$$

$$e' = a - 10b + 30c - 35d + 14e^1.$$

В третьем ряде уравнений опять $A=a'-a$:

¹ Следующие три уравнения этой серии такие:

$$f' = a - 15b + 70c - 140d + 126e - 42f$$

$$g' = a - 21b + 140c - 420d + 630e - 462f + 132g$$

$$h' = a - 28b + 252c - 1050d + 2310e - 2772f + 1716g - 429h.$$

$$\begin{aligned}
 A &= a' \\
 B &= \frac{6}{n-1} b' \\
 C &= \frac{30}{(n-1)(n-2)} c' \\
 D &= \frac{140}{(n-1)(n-2)(n-3)} d' \\
 E &= \frac{630^*}{(n-1)(n-2)(n-3)(n-4)} e'.
 \end{aligned}$$

Наконец, сумма квадратов отклонений от средней определяется путем вычитания из $\Sigma Y^2 = 13,577730$ поправочного члена $nA^2 = 11 \times 0,701273^2 = 5,409\ 622$, который является первой из величин такого ряда:

$$\begin{aligned}
 nA^2 &= \frac{(\Sigma Y)^2}{n} \\
 &\quad \frac{n(n^2-1)}{12} B^2 \\
 &\quad \frac{n(n^2-1)(n^2-4)}{180} C^2 \\
 &\quad \frac{n(n^2-1)(n^2-4)(n^2-9)}{2800} D^2 \\
 &\quad \frac{n(n^2-1)(n^2-4)(n^2-9)(n^2-16)}{44\ 100^{**}} E^2.
 \end{aligned}$$

Определением остатка $\Sigma y^2 = \Sigma Y^2 - nA^2 = 8,168108$ заканчивается первая стадия подыскания полинома. В этом случае к данным подобрана прямая горизонтальная линия. Сумма квадратов отклонений от этой линии равна Σy^2 .

Вторая стадия начинается с определения сумм колонки 2 таблицы 201. После перенесения первого наблюдения из первой колонки таблицы последующие суммы определяются так:

$$\begin{aligned}
 0,029 + 0,052 &= 0,081; \\
 0,081 + 0,079 &= 0,160; \\
 0,160 + 0,125 &= 0,285 \text{ и т. д.}
 \end{aligned}$$

Этот процесс на суммирующей машине производится весьма быстро. Последняя сумма 7,714 должна быть равна S_1 , что является удобным контролем вычислений. Сумма колонки 2 составляет $S_2 = 20,426$.

Идя по приведенным выше сериям формул, получаем для второй стадии такие результаты.

$$\begin{aligned}
 b &= \frac{1,2}{11,12} \times 20,426 = 0,309485; \\
 b' &= 0,701273 - 0,309485 = 0,391788; \\
 B &= \frac{6}{10} \times 0,391788 = 0,235073; \\
 \Sigma (Y - \hat{Y})^2 &= 8,168108 - \frac{11 \times (11^2 - 1)}{12} \times 0,235073^2 = \\
 &= 8,168108 - 6,078525 = 2,089583.
 \end{aligned}$$

Здесь произведено исключение суммы квадратов 6,078525 с одной степенью свободы, относящейся к линейной регрессии. Оценка ее существенности может быть произведена путем сравнения с остатком 2,089583, имеющим 9 степеней свободы. Следовательно, здесь мы имеем средний квадрат ошибки $2,089583/9 = 0,232$, что дает $F = 6,079/0,232 = 26,2$, имеющую высокую степень достоверности при $f = 1$ и 9.

* Числитель следующих трех членов: 2772, 12 012 и 51 480.

** Знаменатели следующих трех членов: 698 544, 11 099 088 и 176 679 360.

Третья стадия включает в себя подбор члена второй степени. Начиная с $S_3 = 52,371$, имеем:

$$c = \frac{1 \cdot 2 \cdot 3}{11 \cdot 12 \cdot 13} \times 52,371 = 0,183115;$$

$$c' = 0,701273 - 3 \times 0,309485 + 2 \times 0,183115 = 0,139048;$$

$$C = \frac{30}{10 \times 9} \times 0,139048 = 0,0463493;$$

$$\begin{aligned} \Sigma (Y - \hat{Y})^2 &= 2,089583 - \frac{11 \times (11^2 - 1)(11^2 - 4)}{180} \times 0,0463493^2 = \\ &= 2,089583 - 1,843205 = 0,246378. \end{aligned}$$

Средний квадрат для оценки существенности теперь будет $0,246378,8 = 0,0308$; поэтому $F = 1,843/0,0309 = 59,8$, что при $f = 1$ и 8 говорит о высокой существенности этого третьего члена.

Подбор на следующей стадии члена третьей степени дает такой результат:

$$\Sigma (Y - \hat{Y})^2 = 0,246378 - 0,236757 = 0,009621.$$

Отсюда $F = 0,237/0,00137$; $f = 1$ и 7 , т. е. опять происходит весьма существенное уменьшение суммы квадратов.

Можно было бы остановиться на этой стадии. Действительно, остаточная сумма квадратов Σy^2 составляет только

$$0,009621/8,1681 = 0,001178$$

части первоначальной суммы. Однако имеются два повода для того, чтобы продолжить процесс подбора кривой еще на одну стадию дальше, во-первых, в целях более полной иллюстрации метода и, во-вторых, потому, что полным третьей степени дает кривую зигзагообразной формы, что для данных о развитии эмбриона пыленка является не совсем подходящей формой.

Начиная с $S_5 = 302,212$ и применяя последние формулы из каждой приведенной ранее последовательности формул, получим:

$$\Sigma (Y - \hat{Y})^2 = 0,009621 - 0,006049 = 0,003572.$$

Следовательно, $F = 0,006049/0,000595 = 10,2$, что указывает на существенность этого нового члена, хотя и не столь высокую, чем ранее.

В большинстве случаев работа на этом может считаться законченной. Однако некоторые из исследователей могут пожелать иметь графическое изображение кривых, другим могут потребоваться сами уравнения. Те лица, которым нужен только график, а не уравнения, могут последовать той схеме расчетов, которая описана в нескольких приводимых ниже абзацах. Если же необходимо знать уравнения, то следует идти далее, пропустив шестой абзац.

В таблице 202 представлены некоторые полиномы с соответствующими коэффициентами, которые дают возможность построить искомые значения Y путем прибавления серии разностей к *конечному значению* Y . Эти коэффициенты зависят от размера выборки, а полиномы определяются степенью последнего из вводимых в уравнение членов. В качестве первой иллюстрации этого метода мы произведем вычисление точек, определяющих наиболее подходящую прямую линию.

Так как в этом случае последний член уравнения первой степени, то конечное значение Y будет $(a' + 3b') = 0,701273 + 3 \times 0,391788 = 1,8766$.

Первая разность здесь

$$-\frac{6}{n-1} b' = -\frac{6}{11-1} \times 0,391788 = -0,2351.$$

Путем повторного присоединения этой первой разности к конечному значению $1,8766$ можно получить ряд последовательных значений Y , начиная

с наибольшего, соответствующего $X=16$:

$$1,8766 + (-0,2351) = 1,6415;$$

$$1,6415 + (-0,2351) = 1,4064 \text{ и т. д.}$$

X	Полиномиальные значения Y	Первая разность
6	-0,474	
7	-0,239	
8	-0,004	
9	0,231	
10	0,436	
11	0,701	
12	0,936	
13	1,171	
14	1,406	
15	1,642	
16	Конечное значение = 1,8766	-0,2351

ТАБЛИЦА 202

Коэффициенты и полиномы для определения конечных значений и разностей при вычислении членов до седьмой степени

Степень полинома	Коэффициент, зависящий от размера выборки	Полином, кончающийся членом в степени						
		0	1	2	3	4	5	6
	Конечное значение	$a' + 3b' + 5c' + 7d' + 9e' + 11f' + 13g' + 15h'$						
1	$\frac{6}{n-1}$	$b' + 5c' + 14d' + 30e' + 55f' + 91g' + 140h'$						
2	$\frac{60}{(n-1)(n-2)}$	$c' + 7d' + 27e' + 77f' + 182g' + 378h'$						
3	$\frac{-840}{(n-1)(n-2)(n-3)}$	$d' + 9e' + 44f' + 156g' + 450h'$						
4	$\frac{15\,120}{(n-1)\dots(n-4)}$	$e' + 11f' + 65g' + 275h'$						
5	$\frac{-332\,640}{(n-1)\dots(n-5)}$	$f' + 13g' + 90h'$						
6	$\frac{8\,648\,640}{(n-1)\dots(n-6)}$	$g' + 15h'$						
7	$\frac{-259\,459\,200}{(n-1)\dots(n-7)}$	h'						

На счетной машине вычисления следует проводить с четырьмя знаками после запятой, но при заполнении колонки полиномиальных значений последняя цифра опускается. Эти значения нанесены на графике рисунка 29. Конечно, прямая линия производит далеко не удовлетворительное сглаживание данных, но ее преимущество в том, что она проходит через точку (\bar{x}, \bar{y}) , у нас через точку (11, 0,701).

Теперь определим ординаты параболы, наилучшим образом проходящей около фактических точек. Для этого берем из таблицы 202 данные для второй степени:

$$\text{Конечное значение} = a' + 3b' + 5c' =$$

$$= 0,701273 + 3 \times 0,391788 + 5 \times 0,139048 = 2,57188.$$

$$\text{Первая разность} = -\frac{6}{n-1}(b' + 5c') =$$

$$= -\frac{6}{10} \times (0,391788 + 5 \times 0,139048) = -0,65222.$$

$$\text{Вторая разность} = \frac{60}{(n-1)(n-2)}c' =$$

$$= \frac{60}{10 \times 9} \times 0,139048 = 0,092699.$$

Дальнейшие расчеты проводятся по следующей схеме:

X	Y	Первая разность	Вторая разность
6	0,222	0,2748	
7	0,039	0,1821	
8	-0,050	0,0894	
9	-0,047	-0,0033	
10	0,049	-0,0960	
11	0,238	-0,1887	
12	0,519	-0,2814	
13	0,893	-0,3741	
14	1,360	-0,4668	
15	1,920	-0,5595	
16	Конечное значение = 2,57188	-0,65222	0,092699

Здесь вторая разность повторно складывается с нижней первой разностью и каждая сумма ставится на строчку выше. Каждая первая разность, начиная

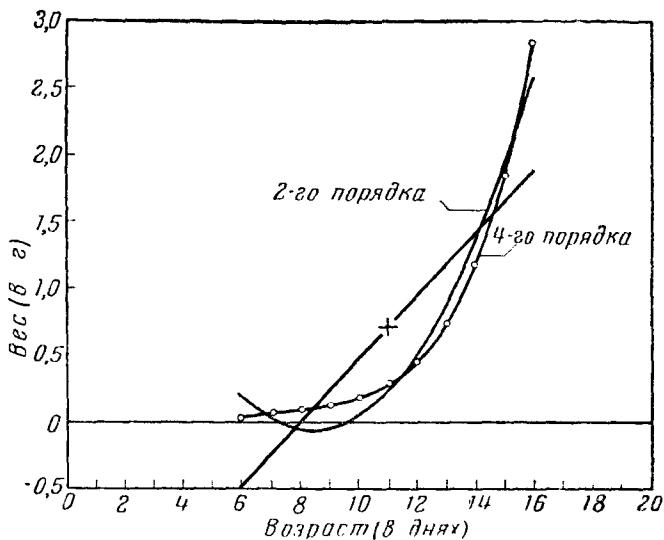


Рис. 29. Графики полиномов первого, второго и четвертого порядка, вычисленных для данных о эмбрионах цыплят из таблицы 201.

с последней, складывается с соответствующим значением Y и дает новое значение, стоящее на строчку выше. Так:

$$-0,65222 + 2,57188 = 1,920;$$

$$-0,5595 + 1,920 = 1,360 \text{ и т. д.}$$

Размещение Y в соответствии со значениями X дает график параболы.

Переходя к полиному четвертой степени, имеем конечное значение и разности различного порядка:

$$\text{Конечное значение} = a' + 3b' + 5c' + 7d' + 9e' = 2,822337.$$

$$\text{Первая разность} = -\frac{6}{n-1} (b' + 5c' + 14d' + 30e') = -0,974844.$$

$$\text{Вторая разность} = \frac{60}{(n-1)(n-2)} (c' + 7d' + 27e') = 0,296464.$$

$$\text{Третья разность} = \frac{-840}{(n-1)(n-2)(n-3)} (d' + 9e') = -0,069337.$$

$$\text{Четвертая разность} = \frac{15120}{(n-1)(n-2)(n-3)(n-4)} e' = 0,009198.$$

Эти результаты вносятся в последнюю строку таблицы 203. Третьи разности получаются путем последовательного прибавления числа 0,009198. Вторые разности определяются так:

$$-0,069337 + 0,296464 = 0,22713;$$

$$-0,060139 + 0,22713 = 0,16699 \text{ и т. д.}$$

На счетной машине в процессе этих вычислений сохраняются все значащие цифры, но при выписке результатов одна из них (последняя) отбрасывается. Первые разности находятся сходным образом, после чего определяются и сами полиномиальные значения. Эти последние, будучи нанесенными на график, определяют кривую четвертой степени. Она проходит столь близко к фактическим точкам, что отклонения остаются незаметными для глаза.

Теперь у нас в руках вся информация, которая обычно достаточна для большинства случаев практики. Точность подбора кривой здесь устанавливается при помощи критериев существенности последовательных уменьшений суммы квадратов. В то же время график не только позволяет непосредственно наблюдать изменение кривой, но и служит для целей интерполяции с достаточной для обычной практики точностью. Это приводит к тому, что обычно представляется очень мало поводов для установления самого уравнения кривой.

ТАБЛИЦА 203

Вычисления Y-полиномиальных значений четвертой степени по данным о развитии эмбриона цыпленка

X	Y	Первая разность	Вторая разность	Третья разность	Четвертая разность
6	0,026				
7	0,056	-0,0302			
8	0,086	-0,0295	-0,0069		
9	0,119	-0,0337	0,00426	-0,004951	
10	0,171	-0,0521	0,01841	-0,014149	
11	0,265	-0,0939	0,04176	-0,023347	
12	0,434	-0,1682	0,07430	-0,032545	
13	0,718	-0,2843	0,11605	-0,041743	
14	1,169	-0,4512	0,16699	-0,050941	
15	1,847	-0,6784	0,22713	-0,060139	
16	2,8223	-0,97484	0,296464	-0,069337	0,009198

Если все же необходимо иметь уравнение кривой, то описанное выше вычисление полиномиальных значений можно опустить, так как после установления указанным ниже путем уравнения расчет по нему графика производится обычным способом. Искомое уравнение имеет форму:

$$Y = A + BX_1 + CX_2 + \dots,$$

где $A, B, C \dots$ являются величинами, вычисленными в конце таблицы 201, и X_1, X_2 и т. д. являются некоторыми полиномами X . На данной стадии расчетов считается, что X принимает равноотстоящие значения с интервалом в единицу. Если же фактические интервалы хотя и равны, но отличаются от единицы, например составляют 30 дней или 5 лет и т. д., то они путем соответствующего деления легко могут быть кодированы. На практике эти значения просто нумеруются 1, 2, 3 и т. д., причем единица составляет 30 дней, 5 лет и т. д.

В данном случае приходится применять такую серию формул:

$$\begin{aligned} X_1 &= X - \bar{x}; \\ X_2 &= X_1^2 - \frac{n^2 - 1}{12}; \\ X_3 &= X_1^3 - \frac{3n^2 - 7}{20} X_1; \\ X_4 &= X_1^4 - \frac{3n^2 - 13}{14} X_1^2 + \frac{3(n^2 - 1)(n^2 - 9)}{560}; \\ X_5 &= X_1^5 - \frac{5(n^2 - 7)}{18} X_1^3 + \frac{15n^4 - 230n^2 + 407}{1008} X_1. \end{aligned}$$

По данным таблицы 193 имеем $\bar{x} = 11$ дней и $n = 11$. Отсюда

$$\begin{aligned} X_1 &= X - 11; \\ X_2 &= (X - 11)^2 - 10 = X^2 - 22X + 111; \\ X_3 &= (X - 11)^3 - 17,8(X - 11) = X^3 - 33X^2 + 345,2X - 1135,2; \\ X_4 &= (X - 11)^4 - 25(X - 11)^2 + 72 = X^4 - 44X^3 + 701X^2 - 4774X + 11688. \end{aligned}$$

На основе этих имеющихся у нас выражений может быть написано любое уравнение вплоть до уравнения четвертой степени. Например, уравнение наилучшим образом подобранной прямой будет:

$$Y = A + BX_1 = 0,701273 + 0,235073(X - 11) = 0,2351X - 1,885,$$

а уравнение четвертой степени будет:

$$\begin{aligned} Y &= A + BX_1 + CX_2 + DX_3 + EX_4 = \\ &= 0,701273 + 0,235073(X - 11) + 0,046349(X^2 - 22X + 111) + \\ &+ 0,0061907(X^3 - 33X^2 + 345,2X - 1135,2) + 0,000383325 \times \\ &\times (X^4 - 44X^3 + 701X^2 - 4774X + 11688) = \\ &= 0,712 - 0,4772X + 0,11071X^2 - 0,010672X^3 + 0,0003832X^4. \end{aligned}$$

Подстановка в это уравнение определенных значений X дает возможность определить полиномиальные значения Y и нанести их на график.

Читатель может задать вопрос относительно рекомендаций по поводу кривых. Было бы неразумным затрачивать много времени на определение первой попавшейся кривой. Только после того как составлено достаточно ясное представление о форме кривой, следует приступать к этой работе. Часто бывает достаточным нанесение точек на график, для чего применяются небольшие кружки — крупные точки. Эти точки можно соединить прямыми, но следует избегать проведения «глазомерных» кривых. Они слишком субъективны и способны ввести в заблуждение как самого виновника этого дела, так и постороннего наблюдателя. Интерполяция на основе этих прямолинейных связующих звеньев имеет больший шанс оказаться лучшей, чем оценка на основе кривой, проведенной от руки, хотя бы и с большим искусством. Если вы достаточно осведомлены о законе изменения и можете сказать, какая в данном случае должна быть кривая, тогда вы должны обладать и умением придать этому закону соответствующее математическое выражение в виде уравнения, а также произвести подбор уравнения к имеющимся данным. Когда

у вас возникает потребность определения корреляции, если регрессия нелинейна, или когда необходимо произвести оценку существенности криволинейности, становится обязательным подбор наиболее подходящего уравнения криволинейной регрессии. Но чаще всего подбор кривой нужен только для интерполяции. В этих случаях обычно бывает достаточным графическое изображение данных.

Пример 14. Следующие пять точек (0,9), (1,10), (2,9), (3,6) и (4,1) лежат на параболе $Y=9+2X-X^2$. Применяя метод Фишера, показать, что сумма квадратов, остающаяся после подбора кривой второго порядка, равна нулю. Вычислите полиномиальные значения Y . Они должны совпадать с теми парами значений, которые приведены выше. Если вы наложите эти результаты на график, то точки расположатся точно на кривой, соответствующей указанному уравнению. Определите это уравнение, применяя формулы вплоть до второй степени.

Пример 15. Имеется несколько точек (0,0), (1,4), (2,2), (3,0), (4,4), (5,20), лежащих на кубической кривой $Y=9X-6X^2+X^3$. Проведите все вычисления по методу Фишера, включая подбор прямой линии и параболы. После подбора кривой третьей степени не должно быть никакой остаточной суммы квадратов, а полиномиальные значения Y на этой стадии должны оказаться теми же, с которых вы начали работу. Проведите на графике наилучшую прямую и наилучшую параболу, а также напишите их уравнения.

Пример 16. Рид и Холланд [17] приводят следующие данные о средней высоте растений подсолнечника:

Недели	1	2	3	4	5	6	7	8	9	10	11	12
Высота растений . .	18	36	68	98	131	170	206	228	247	250	254	254
Полиномиальные значения	15,0	38,9	68,0	100,6	134,6	168,0	198,9	225,5	245,6	257,5	259,1	248,4

Полиномиальные значения здесь определены по кубическому уравнению. Проверьте и постройте график. Остаточная сумма квадратов после подбора кубического уравнения равна 211,03. Вычислите это уравнение и произведите по нему вторичную проверку приведенных выше полиномиальных значений.

7. Биологическое опробование. Из числа различных приложений метода криволинейного сглаживания в последние два десятилетия предметом особого интереса является *биологическое опробование*. Этим термином обозначается определение потенциальных способностей некоторого стимулятора путем наблюдений над его воздействием на живой организм. Общими случаями таких исследований является измерение силы действия на растения или животных токсических и лекарственных веществ, гормонов и пр.

Новое значительное развитие методов биологического опробования дано в работе Блисса [2], который разработал метод, названный *пробит-анализом* (probit analysis). Берксон [1] применил *логистическую* кривую [15] к задачам современного метода биологического опробования, связав это с названным *логистический анализ*. Книга Финни [7] и статья Берксона дают подробное описание этого метода, таблицы и обильный иллюстрационный материал.

ЛИТЕРАТУРА

1. Berkson Joseph, Journal of the American Statistical Association, 48, 565, 1953.
2. Bliss C. I., Annals of Applied Biology, 22, 134; 22, 307, 1935; 24, 815, 1937.
3. Decker George C., Andre Floyd, Iowa State College Journal of Science, 10, 403, 1936.
4. Deming W. Edwards, Statistical Adjustment of Data. John Wiley and Sons, Inc., New York, 1943.
5. Eid M. T., Black C. A., Kempthorne O., Zoellner J. A., Iowa Agricultural Experiment Station Research Bulletin, 406, 1954.
6. Emmert E. M., Journal of the American Society of Agronomy, 29, 213, 1937.
7. Finney D. J., Probit Analysis. Cambridge University Press, 1952.
8. Fisher R. A., Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 1925, section 27.

9. Fisher R. A., Yates F., Statistical Tables. Oliver and Boyd, Edinburgh, 1938.
10. Geddes W. F., Canadian Journal of Research, 1, 528, 1929.
11. Haskell A. C., How to Make and Use Graphical Charts. Codex Book Co., Inc., New York, 1919.
12. Lipka Joseph, Graphical and Mechanical Computation. John Wiley and Sons, Inc., New York, 1918.
13. Metzger W. H., Journal of the American Society of Agronomy, 27, 653, 1935.
14. Mitscherlich Eilhard Alfred, Landw. Jahrb., 38, 537, 1909.
15. Pearl Raymond, Medical Biometry and Statistics. W. B. Saunders Co., Philadelphia, 2nd Edition, 1930.
16. Penquite Robert, Thesis submitted for the degree Doctor of Philosophy, Iowa State College, 1936.
17. Reed H. S., Holland R. H., Proceedings of the National Academy of Science, 5, 140, 1919.
18. Running T. R., Empirical Formulas. John Wiley and Sons, Inc., New York, 1917.
19. Swanson Pearl P., Smith A. H., The Journal of Biological Chemistry, 97, 745, 1932.

БИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И РАСПРЕДЕЛЕНИЕ ПУАССОНА

1. Введение. В главах 1 и 9 был рассмотрен вопрос о малых выборках из этих распределений. Теперь мы по примеру главы 8, где описывались методы обработки большой выборки из совокупности, образованной данными измерения, перейдем к обсуждению некоторых вопросов, относящихся к большой выборке данных, полученных в результате подсчета численностей. Так как в случае этих двух распределений средние и дисперсии уже не будут независимыми, как это имело место при нормальном распределении, то прежде, чем производить их обработку по методу дисперсионного анализа, необходимо произвести некоторые преобразования этих данных. В главе 11 были описаны три вида таких преобразований.

2. Биномиальное распределение. Если появление некоторого события, подобного появлению при рождении особи мужского пола или выживанию животного в некоторых вредных для него условиях, наблюдается по субвыборкам определенного размера, то эти субвыборки могут быть сгруппированы по признаку числа появления данного события. Полученное в результате этого распределение частот может рассматриваться как выборка из *биномиального распределения*, т. е. распределения, в котором вероятность появления события остается одной и той же для каждого отдельного наблюдения.

Допустим в качестве примера, что имеются группы (субвыборки) по четыре цыпленка, являющихся полными (по отцу и матери) сестрами, и на них испытывается действие некоторого препарата, при котором одинаково вероятно выживание или гибель каждого цыпленка. Событие «выживание» в этих группах может иметь численность от 0 до 4. Допустим, что испытанию было подвергнуто 96 групп и что получены такие результаты:

Число выживших в группе	Число субвыборок с данной численностью выживания
0	1
1	20
2	41
3	25
4	9
Итого	96

Ясно, что отсутствие выживания в группе, как и четыре выживания, маловероятны. Модальная численность групп относится к случаю двух живых цыплят.

Наличие двух признаков, четыре наблюдения в каждой субвыборке и 50%-ная вероятность выживания полностью определяют собой биномиальное распределение.

Относительные частоты появления 0, 1, 2, 3, и 4 выживаний, которые более удобны для расчета, определяются последовательными членами разложения бинома¹:

$$\begin{aligned} \left(\frac{1}{2} + \frac{1}{2}\right)^4 &= \binom{4}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 + 4 \cdot \binom{4}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 + 6 \cdot \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 + 4 \cdot \binom{4}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 + \binom{4}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \\ &= \frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16}. \end{aligned}$$

Это означает, что в одной шестнадцатой части всех субвыборок ожидается отсутствие выживания, в одной четверти — одно выживание и пр. Учитывая это и принимая во внимание, что в выборке фактически было 96 групп, можно определить *гипотетические частоты* по классам: $96 \times \frac{1}{16} = 6$ и т. д., т. е. ряд численностей 6, 24, 36, 24 и 6. Различия между этими численностями и численностями, наблюдаемыми в выборке, могли возникнуть или в связи с капризами выборочного наблюдения, или ошибочности предположения об одинаковой вероятности гибели для всех цыплят. Эти возможности и будут теперь предметом исследования.

Общее положение теории следующее: если появление события имеет постоянную вероятность p , то относительные частоты его появления 0, 1, 2 и т. д. раз в случайной субвыборке из k наблюдений определяются последовательными членами разложения бинома:

$$(q + p)^k,$$

где $q = 1 - p$ означает вероятность непоявления данного события; гипотетические же частоты соответствующих классов выборки получаются путем умножения этих относительных частот на n .

Вы можете воспользоваться удобным способом для проверки этой теории путем взятия выборок из однородной совокупности. Например, можно произвести отбор групп семян бобов одинакового размера, формы и плотности, некоторая часть которых, положим одна треть, имеют определенный цвет (или какую-нибудь другую метку). Отбирая субвыборки, допустим по пяти, при помощи того или иного метода случайного отбора (причем следует за один раз брать по одному семени), отмечаем цвет этого боба, после чего он возвращается обратно в общую массу бобов. Этот процесс производится при перемишивании после каждого наблюдения. Вероятность появления в группе 0, 1, 2, 3, 4 или 5 окрашенных семян определяется в этом случае членами разложения бинома при $p = \frac{1}{3}$, $q = \frac{2}{3}$ и $k = 5$:

$$\begin{aligned} \left(\frac{2}{3} + \frac{1}{3}\right)^5 &= \binom{5}{0} \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^5 + 5 \cdot \binom{5}{1} \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^4 + 10 \cdot \binom{5}{2} \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^3 + 10 \cdot \binom{5}{3} \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^2 + \\ &+ 5 \cdot \binom{5}{4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^1 + \binom{5}{5} \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^0 = \\ &= \frac{32}{243} + \frac{80}{243} + \frac{80}{243} + \frac{40}{243} + \frac{10}{243} + \frac{1}{243}. \end{aligned}$$

Если вы возьмете 243 такие субвыборки, то при $n = 243$ гипотетические частоты будут определяться числителями этих шести дробей.

Вместо бобов можно использовать случайные цифры таблицы 2. Например, цифры 3, 6 и 9 пусть будут «окрашенными бобами», а цифры 1, 2, 4, 5, 7 и 8 — «белыми»; цифра 0 пропускается. При помощи таких условий может быть определена любая взятая вероятность.

¹ Сведения о понятиях: *разложение бинома, вероятность, взятие выборки, пересчитывание и сочетания* можно найти в любом учебнике по математике для высшей школы.

Биномиальные коэффициенты, для образования которых существуют определенные правила, легко удержать в памяти, если их выписать в таком виде:

Размер выборки K	Биномиальные коэффициенты								
1									
2				1	1				
3			1	2	1				
4		1	3	3	1				
5		1	4	6	4	1			
6		1	5	10	10	5	1		
7		1	6	15	20	15	6	1	
8	1	7	21	35	35	21	7	1	
и т. д.	1	8	28	56	70	56	28	8	1
					и т. д.				

Эта схема носит название *треугольник Паскаля*; каждый коэффициент является суммой двух вышестоящих справа и слева коэффициентов. Для приведенных выше расчетов в примерах с цыплятами и бобами можно произвести сверку коэффициентов при $k=4$ и $k=5$. Вы замечаете, что последовательность коэффициентов всегда симметрична. При $p=q=0,5$ и само биномиальное распределение будет симметричным, форма которого по мере увеличения k будет все более и более приближаться к нормальной. Если же вероятности p и q не равны друг другу, то форма биномиального распределения будет скошенной.

В биномиальном распределении, как это было и в нормальном распределении, средняя и стандартное отклонение получаются путем соответствующего усреднения данных. Располагая наши *гипотетические* частоты по обычной вычислительной схеме (параграф 2 главы 8), находим:

Число выживших цыплят X	Гипотетические частоты f	fX	fX^2
0	6	0	0
1	24	24	24
2	36	72	144
3	24	72	216
4	6	24	96
Итого	96	192	480

$$\mu = 192/96 = 2 \text{ цыпленка на группу}$$

$$\sigma = \sqrt{\frac{480 - 192^2/96}{96}} = 1 \text{ цыпленок на группу}$$

Здесь для средней и стандартного отклонения применены символы μ и σ для того, чтобы провести различие между этими параметрами совокупности и выборочными характеристиками \bar{x} и s . Эти последние, вычисленные по *наблюденным* частотам, в данном случае соответственно равны 2,2188 цыпленка на группу и 0,920 цыпленка на группу, но эти показатели найдут у нас очень небольшое применение. Вторая особенность приведенных выше вычислений, на которую вам следует обратить внимание, состоит в том, что делителем при вычислении σ является $n-96$, а не $n-1=95$. Это объясняется тем, что число степеней свободы принимается во внимание при вычислении выборочных характеристик, но не параметров.

Обычно приведенные выше усредненные характеристики μ и σ выражаются в виде долей или процентов, а не в виде численностей на группу. Так, можно сказать, что доля оставшихся в живых цыплят составляет $\mu = 2/4 = 0,5$, а $\sigma = 0,25$, или процент этих цыплят равен 50% при стандартном отклонении 25%. Такая независимость этих характеристик от размера субвыборок позволяет производить непосредственное сравнение их значений для случая, когда в группе четыре наблюдения, со значениями их при любой другой численности группы.

Если основываться на математических соотношениях, свойственных биномиальному распределению, то нет никакой необходимости в проведении указанных выше вычислений. Эти два параметра можно получить более просто по таким формулам:

$$\mu = kp = 4 \times 0,5 = 2 \text{ цыпленка на группу.}$$

$$\sigma = \sqrt{kpq} = \sqrt{4 \times 0,5 \times 0,5} = 1 \text{ цыпленок на группу.}$$

Если же параметры выражаются в долях или процентах, то эти формулы принимают вид:

$$\mu = p = 0,5, \text{ или } 50\%.$$

$$\sigma = \sqrt{pq/k} = \sqrt{0,5 \times 0,5/4} = 0,25, \text{ или } 25\%.$$

Таким образом, p не только вероятность появления события, но оно в то же время является выраженной в долях или процентах средней совокупности.

Заслуживает внимания то обстоятельство, что если бы размер субвыборки был бы не 4, а 16 цыплят, то ожидаемая средняя в процентах осталась бы той же самой $p = 0,5$, или 50%, но ожидаемое значение стандартного отклонения было бы иным:

$$\sigma = \sqrt{pq/k} = \sqrt{0,5 \times 0,5/16} = 0,125, \text{ или } 12,5\%.$$

Таким образом, теоретическое стандартное отклонение при увеличении размера субвыборки уменьшается. Действительно, σ является стандартным отклонением групповой средней и варьирует обратно пропорционально квадратному корню из размера группы — положение, хорошо известное нам для случая обработки данных измерения.

Если же взять биномиально распределенную выборку в целом (объединяя все группы — субвыборки), то стандартная ошибка соответствующей средней будет обратно пропорциональной квадратному корню из общего числа объектов. Это значит, что если имеется n субвыборок, каждая из которых состоит из k объектов, то

$$\sigma_x = \sqrt{\frac{pq}{nk}} = \frac{\sigma}{\sqrt{n}}$$

— опять знакомый нам результат. Для всего распределения цыплят стандартная ошибка в процентах будет

$$\sigma_x = \frac{25}{\sqrt{96}} = 2,55\%.$$

Теперь мы можем обратиться к предложенному ранее вопросу: можно ли различие между предполагаемым значением параметра $\mu = 50\%$ и соответствующей выборочной средней $\bar{x} = 2,2188/4$, т. е. 55,64%, приписать только выборочному варьированию? Если выборка взята действительно из биномиального распределения, то имеется возможность применения критерия хи-квадрат. Но вместо того, чтобы основываться на процентных выражениях, проще решить непосредственно такой вопрос: имеется ли какое-либо существенное различие между гипотетическим числом живых цыплят и наблюдаемым

числом их? Первое число составляет 50% от 384, т. е. 192 цыпленка. Второе же по выборочным данным равно:

$$\Sigma fX = 1 \times 0 + 20 \times 1 + 41 \times 2 + 25 \times 3 + 9 \times 4 = 213.$$

Теперь по обычной схеме находим

$$\chi^2 = \frac{(213 - 192)^2}{192} + \frac{(171 - 192)^2}{192} = 4,59;$$

вероятность появления этой величины равна 0,035. Следовательно, выборочное количество выживших цыплят существенно больше 50%.

Вместо критерия χ^2 обычно применяется критерий t . В рассматриваемом случае эти критерии идентичны друг другу, если в них применяется σ ; применение s может привести иногда к ошибочным заключениям. Ранее мы нашли, что стандартная ошибка средней равна:

$$\sigma_x = 25 / \sqrt{96} = 2,55\%.$$

Разность между выборочной средней и параметром равна:

$$55,47 - 50 = 5,47\%.$$

Поэтому

$$t = 5,47 / 2,55 = 2,15,$$

что при $f = \infty$ близко к уровню 0,035, подобно тому, как хи-квадрат было близким к тому же уровню 0,035. Интересно отметить, что при данных условиях $t = \chi$. Это указывает на законность применения критерия t и к статистическим показателям, определенным по численностям наблюдений.

Пример 1. Вычислите по способу параграфа 2 главы 8 показатели \bar{x} и s по выборочным данным о выживании цыплят.

Пример 2. Показать, что в предложенном ранее отборе семян бобов при $p = 1/3$ и $k = 5$ параметрами будут $\mu = 1,667$ и $\sigma = 1,054$ окрашенных семян на субвыборку, или $\mu = 33,33\%$ и $\sigma = 21,08\%$.

Пример 3. Полезно иметь в виду, что общее число объектов в выборке из биномиального распределения равно nk и что общее число появления события ΣfX . Следовательно, $\bar{x} = 100 \cdot \Sigma fX / nk$ процентов.

Пример 4. Если произведен случайный отбор 243 субвыборок бобов ($k = 5$ и $p = 1/3$) и получено $\bar{x} = 30,37\%$, то показать, что $t^2 = \chi^2 = 4,80$. Какое заключение по поводу этих выборок вы сделаете?

Пример 5. Вы часто встречались с такого рода утверждением: «По мере увеличения размера выборки выборочная средняя все более и более приближается к значению средней в совокупности». Это приводит к требованию о приближении хи-квадрат к нулю. Думаете ли вы, что это так и будет?

Когда размер выборки возрастает, биномиальное распределение приближается к нормальному, причем это происходит более быстро при значениях p , близких к 0,5, и медленнее при p около 0 и 1. Одно из следствий этого состоит в том, что для установления доверительных пределов вместо таблицы 1 можно использовать величину t . Например, если $n = 100$ и в выборке имеется 40 объектов, обладающих изучаемым признаком, то $p = 0,4$, $q = 0,6$ и

$$40 \pm t_{0,05} \sqrt{npq} = 40 \pm 1,96 \sqrt{100 \times 0,4 \times 0,6} = 40 \pm 9,6,$$

откуда определяется доверительный интервал 30,4—49,6. Сравнить это с интервалом 30—50, получаемым на основе указанной таблицы.

Другой пример. Допустим, что 100 объектов, обладающих определенным признаком, были обнаружены в выборке из 1000 объектов. Так как в соответствующей части таблицы 1 приведены не сами численности, а их отношения, то в такой же форме вычислим интервал и для данного случая:

$$0,1 \pm t_{0,05} \sqrt{pq/n} = 0,1 \pm 1,96 \sqrt{0,1 \times 0,9/1000} = 0,1 \pm 0,0186.$$

Таким образом, искомый интервал простирается от 0,0814 до 0,1186, или от 8,14% до 11,86%.

3. Сопоставление выборочного распределения с биномиальным. Экспериментальные данные всегда подвержены выборочному варьированию, с которым приходится считаться. Мы уже познакомились с оценкой существенно-сти отклонения выборочной средней от μ при предположении, что выборка принадлежит к биномиальному распределению. Теперь следует рассмотреть вопрос, можно ли считать данное выборочное распределение, взятое в целом, отличающимся от биномиального. Методы решения этого вопроса подобны описанным в главе 9.

Для иллюстрации мы возьмем данные таблицы 204 [3]. Так как здесь взяты пометы из восьми поросят, то имеется девять классов, в которых 0,1.... 8 боровков. Сначала возьмем за основу гипотезу об одинаковой вероятности появления обоих полов. В этом случае $p=0,5$, $q=0,5$, $k=8$, $n=106$ и гипотетические частоты будут

$$\begin{aligned} & 106 [0,5^8; 8 \times 0,5^7 \times 0,5; 28 \times 0,5^6 \times 0,5^2 \text{ и т. д.}] = \\ & = 106 \times 0,5^8 (1, 8, 28, 56, 70 \text{ и т. д.}) = \\ & = 0,414; 3,312; 11,594 \text{ и т. д.} \end{aligned}$$

ТАБЛИЦА 204

Пометы из восьми поросят, имеющие различное количество боровков. Гипотетические численности и критерий для биномиального распределения

Число боровков в помете X	Число пометов f	Гипотетическая численность	Отклонение фактической от гипотетической численности	Отклонения в квадрате	(Отклонение) ²
					гипотетическая численность
0	0	0,414	-0,414		
1	5	3,312	1,688		
2	9	11,594	-2,594		
3	22	23,188	-1,188		
4	25	28,984	-3,984		
5	26	23,188	2,812		
6	14	11,594	2,406		
7	4	3,312	0,688		
8	1	0,414	0,586		
Итого	106	106,000	0,000		$\chi^2 = 1,948$

Эта система отклонений гипотетических численностей от фактических характеризует отклонение выборочного распределения от данного биномиального распределения. Существенно ли это отклонение? Соответствующим критерием является хи-квадрат. Но так как при применении этого критерия гипотетические численности не должны быть меньше пяти, то для расчета хи-квадрат необходимо объединить вместе три первых и три последних класса. В результате получается значение хи-квадрат, которое при $f=4$ следует считать умеренным. Фактически здесь имеет место даже несколько меньшее отклонение от ожидаемого распределения, чем это может быть в обычных условиях случайного выборочного варьирования.

Почему здесь взято четыре степени свободы? Общая численность 106 пометов, имеющая одну степень свободы, определена на основе выборки и в то же время была использована для подбора биномиального распределения. В результате этого у пяти использованных в расчетах классов остается для оценки хи-квадрат только четыре степени свободы.

Из сравнения фактических и гипотетических численностей следует, что нет никаких данных, свидетельствующих против гипотезы об одинаковой вероятности в 50% для появления обоих полов. Конечно, этот вывод не является доказательством того, что отношение полов точно равно 1:1. Он только означает, что при данном материале не обнаружено существенного отклонения фактического распределения от биномиального распределения, в котором $p=q$.

Пример 6. Парк [4] собрал данные о распределении полов у пометов свиней из отчетов National Duroc Jersey Pig Record. Среди 402 пометов по восемь поросят он установил 1, 8, 37, 81, 162, 77, 30, 5 и 1 пометов, имеющих соответственно 0, 1, 2 ... 8 боровков. Вычислите $\chi^2=38,7$, $f=6$. Можете ли вы дать объяснение существенному отклонению фактического распределения от гипотетического, т. е. сказать, почему вероятность появления женского пола существенно варьирует по пометам?

Пример 7. Оставляя в стороне отличие формы выборочного распределения от биномиальной формы, оцените для предыдущего примера существенность разности между гипотетическим числом боровков (1608) и числом, установленным по наблюдениям (1581). Здесь хи-квадрат равно 0,90. Не проливает ли это некоторый свет на особенности выборочного наблюдения?

Пример 8. Произведите оценку существенности отклонения выборочного распределения цыплят от биномиального распределения при $p=50\%$. Хи-квадрат $=7,07$; $f=4$. Средняя, как вы можете снова убедиться, существенно отличается от 50%.

Пример 9. Произведите оценку существенности отклонения выборочного числа боровков (439) таблицы 204 от числа (424), ожидаемого при гипотезе о 50% вероятности этого пола. Хи-квадрат $=1,06$, $f=1$.

4. Критерий однородности при биномиальной форме распределения.

Часто нет никаких оснований для выдвижения той или иной гипотезы относительно p . В этом случае ни постановка вопроса предыдущего параграфа, ни его решение не имеет никакого смысла. Вместо этого у нас может возникнуть потребность узнать, является ли данная выборка *однородной*, т. е. можно ли считать, что она взята из некоторого биномиального распределения, p которого пока является неопределенным. Так как в этом случае нет никакого теоретического значения вероятности, которое можно было бы поставить в основу расчетов, то в качестве этой вероятности мы просто возьмем выборочный средний процент и определим биномиальное распределение, имеющее это значение p . Фактически здесь мы не применяем критерия для оценки значения p , а рассматриваем вопрос, является ли распределение объектов однородным при данном значении вероятности p .

Хорошей иллюстрацией могут быть данные о выживании цыплят. Непосредственно видно, что в этом случае гипотеза о 50% вероятности выживания не является подходящей (параграф 2 этой главы, пример 8). Можно ли считать, что эти данные являются случайной выборкой из биномиального распределения при $p=55,47\%$, т. е. при вероятности равной выборочной средней? При этих условиях гипотетические частоты определяются последовательными членами разложения бинома: $96 \times (0,4453 + 0,5547)^4$. Эти частоты легко вычисляются по такой схеме:

x	Степени q	Степени p	Биномиальные коэффициенты	Относительные частоты	Гипотетические частоты
0	0,0393		1	0,0393	3,77
1	0,0883	0,5547	4	0,1959	18,81
2	0,1983	0,3077	6	0,3661	35,15
3	0,4453	0,1707	4	0,3040	29,18
4		0,0947	1	0,0947	9,09
Итого				1,0000	96,00

Сначала в предпоследнюю строку и на счетную машину ставится $q=0,4453$. Это число повторно умножается на q , пока не будет получено q^k , и каждая степень q вводится в таблицу. То же самое повторяется и в отношении p , в результате чего получаются последовательные степени p , которые вносятся в таблицу, начиная со второй строки. Относительные частоты пятой колонки являются произведением трех (или двух) колонок, лежащих слева. Например, по третьей строке имеем $0,1983 \times 0,3077 \times 6 = 0,3661$.

Эти относительные частоты в сумме должны давать единицу. Умножение их на $n=96$ приводит к биномиальному распределению частот при $p=55,47\%$.

В таблице 205 дана оценка этого распределения при помощи критерия хи-квадрат. Теперь мы имеем только две степени свободы, так как ожидаемые частоты ограничены не только условием сохранения общей численности выборки $n=96$, но также и тем, что выборочная средняя взята в качестве параметра распределения. В этом случае хотя хи-квадрат превосходит свой 50%-ный уровень, однако имеет небольшое значение. Следовательно, здесь нет ничего, свидетельствующего против допущения о константной вероятности выживания. Соответствующая совокупность может считаться однородной, имеющей постоянную вероятность 55,47%.

ТАБЛИЦА 205

Критерий однородности распределения численностей выживших цыплят

x	f	Гипотетические численности	Отклонение	(Отклонение) ²
				гипотетическая
0	1	3,77 } 22,58 18,81 } 35,15 } 29,18 } 9,09 }	-1,58 5,85 -4,18 -0,09	0,11 0,97 0,60 0,00
1	20			
2	41			
3	25			
4	9			
	96	96,00	0,00	$\chi^2 = 1,68$

При наличии двух выборок после установления их однородности часто возникает задача оценки существенности различия между их средними. Допуская для каждой из них существование постоянной вероятности, можно поставить вопрос: отличаются ли эти вероятности друг от друга существенно? В таблице 206 приведено два распределения численности боровков в пометах, состоящих из пяти и шести поросят.

ТАБЛИЦА 206

Число боровков в пометах свиной дюрок-джерсейской породы

Пометы из пяти поросят			Пометы из шести поросят		
число боровков в помете	число пометов	гипотетические численности	число боровков в помете	число пометов	гипотетические численности
0	2	4,08 } 23,54 19,46 } 37,07 } 35,31 } 16,84 } 20,05 3,21 }	0	3	3,62 } 24,99 21,37 } 52,55 } 69,08 } 50,98 } 20,11 } 23,40 3,29 }
1	20		1	16	
2	41		2	53	
3	35		3	78	
4	14		4	53	
5	4	5	18		
		6	0	18	
Итого	116	$\chi^2 = 0,73$	221		$\chi^2 = 3,92$
	$p = 48,79\%$			$p = 49,62\%$	

Каждая из этих выборок однородна в отношении соответствующей ей вероятности. Имеется ли существенное различие между их средними? Этот вопрос легко разрешить в такой таблице 2×2 :

Пометы	Боровки	Свинки	Итого
С пятью поросятами	283	297	580
С шестью »	658	668	1326
Итого . . .	941	965	1906

$$\chi^2 = \frac{(283 \times 668 + 297 \times 658)^2 1906}{530 \times 1326 \times 941 \times 935} = 0,111,$$

для $f=1$ это значение хи-квадрат оказывается меньше, чем норма при случайном отборе. Столь малая разность между процентами боровков в выборках: $49,62 - 48,79 = 0,83\%$ является, таким образом, несущественной.

Если взять объединенное значение p , полученное путем деления $941/1906 = 49,3704\%$, и вычислить соответствующую стандартную ошибку $\sqrt{pq(1/n_1 + 1/n_2)} = 0,0249$, то оценка существенности может быть проведена при помощи критерия t , как это делалось в предыдущем параграфе. В данном случае χ при одной степени свободы является не чем иным, как t при $f = \infty$.

Но что делать, если выборки окажутся неоднородными? Если для появления события отсутствует постоянная вероятность и если ничего не известно относительно законов, определяющих эту вероятность, то на столь непрочной основе нельзя сделать никаких выводов. В таком случае нет никакой уверенности в том, что при повторении эксперимента будет наблюдаться тот же самый ряд вероятностей. Если остаются неизвестными причины изменения вероятностей и форма этого изменения, то тем самым теряется контроль над экспериментом. Первым шагом в этом случае должно быть изучение техники отбора наблюдений и особенностей имеющегося экспериментального материала. И только после того, как будут вскрыты причины этого варьирования вероятности, можно производить на законном основании необходимые сравнения.

Тем не менее бывают случаи, когда возникает необходимость в оценке существенности разности между средними, несмотря на наличие разнородности выборок. Если такая разность оказывается существенной даже при наличии большого варьирования вероятности, то возможно, что эта разность обусловлена некоторыми действительными фактами. Соответствующий этой оценке метод должен зависеть от наблюдаемого варьирования. По всей вероятности, наилучшим методом для данного случая будет расчет стандартного отклонения для каждой выборки по способу параграфа 2 главы 8 и применение после этого критерия существенности параграфа 8 главы 8. Допущение нормального распределения p в этом случае более уместно, чем допущение константности p .

Для вас может представлять интерес сопоставление методов определения биномиального и нормального распределений. Каждое из них имеет два параметра. Критерий существенности выборочного отклонения от теоретического распределения в обоих случаях один и тот же. Но для всех нормальных распределений имеется один общий ряд относительных частот, приведенный в таблице 72, в то время как для биномиальных распределений эти ряды различны, в зависимости от значений p и k . Отсюда следует, что таблицы для биномиальных распределений, если бы они были составлены, были бы весьма громоздкими; гораздо легче это распределение вычислять каждый раз на основе соответствующих параметров.

Пример 10. Вычислите хи-квадрат для каждой выборки таблицы 206.

Пример 11. Деккер и Андре [1] подвергали взрослые особи клопа чернашки воздействию температуры $-8,5^\circ \text{C}$ в продолжение 15 минут и после этого подсчитывали число погибших насекомых в 100 субвыборках по 10 насекомых в каждой. Количество субвыборок, имевших 0, 1, 2, ... 7 мертвых насекомых, было соответственно равно 4, 21, 22, 28, 14, 8, 2, 1. Субвыборок с восемью или большим числом погибших насекомых

обнаружено не было. Можно ли считать существенным имеющиеся здесь отклонения от константности вероятности гибели насекомых? Хи-квадрат=2,89; $f=8$.

Пример 12. Выборочная вероятность появления самца в примере 6 равна 49,160%. Является ли данная выборка однородной в отношении этого процента? Хи-квадрат=38,15; $f=5$.

Пример 13. В примере 9 по данным таблицы 204 была установлена несущественность отклонения доли особей мужского пола от 0,5. Можно ли считать эту выборку однородной относительно вероятности 0,5177? Хи-квадрат=0,73; $f=3$.

Пример 14. Пользуясь методами главы 8, оценить существенность разности между средними процентами особей мужского пола в выборках таблицы 204 и примера 6. Так как последняя выборка существенно отклоняется от биномиального распределения, то в этом случае нельзя применить расчет хи-квадрат по таблице 2×2 . *Ответ:* $t=0,88$.

Пример 15. Произведите оценку существенности разности между процентами боровков по данным таблицы 206 для пометов с пятью и шестью поросятами, применяя при этой оценке критерий t вместо χ^2 . Вы должны получить $t=\chi=1,0411$.

5. Распределение Пуассона. Этот вид распределения, подобно биномиальному, относится к *дискретной*, или *прерывной*, переменной, возникающей в связи с подсчетом численностей и поэтому обычно принимающей только целочисленные значения. В основе его также лежит предположение о некоторой константной вероятности появления события, но эта вероятность в данном случае обычно столь мала, что часто бывает трудно определенной. Вместе с тем и размер выборки, считающейся также постоянной, обычно остается неизвестным. Однако величина выборки должна быть достаточно большой, и если она не является строго постоянной, то во всяком случае должна находиться внутри узких границ.

В отличие от биномиального и нормального распределений, распределение Пуассона вполне определяется одним параметром — средней. Дисперсия этого распределения равна средней.

Распределение Пуассона может быть выведено в качестве предельной формы биномиального распределения, когда p становится очень малым. Однако это распределение имеет значение и само по себе, независимо от его отношения к биномиальному распределению.

Если μ — ожидаемая средняя частота появления некоторого события в совокупности выборок и если частоты появления его в отдельных выборках подчинены распределению Пуассона, то событие будет встречаться в выборках 0,1,2,... раз с относительными частотами, равными членам такой последовательности:

$$\frac{1}{e^\mu}, \quad \frac{\mu}{e^\mu}, \quad \frac{\mu^2}{2e^\mu}, \quad \frac{\mu^3}{2 \cdot 3 \cdot e^\mu}, \quad \frac{\mu^4}{2 \cdot 3 \cdot 4e^\mu},$$

где: e — основание натуральных логарифмов, являющееся иррациональным числом, приблизительно равным 2,718; логарифм e при основании 10 приблизительно равен 0,434295. Гипотетические же частоты являются произведениями этих относительных частот на численность выборок n .

Процесс подбора распределения Пуассона к выборочным данным показан в таблице 207, где приведены данные о численности семян сорняков [2]. Выборки определялись по весу — $1/4$ унции. Число семян культуры в таких выборках не было постоянным и фактически варьировало в некоторых пределах, но в отношении семян сорняков эти выборки все же сравнимы. Распределение Пуассона, как и биномиальное распределение, строится на основе предположения, что вероятность того, что любое случайно взятое зерно окажется семенем сорняка, остается со всей совокупностью величиной постоянной. Сначала мы определим по имеющимся данным теоретическое распределение, после чего рассмотрим вопрос об оценке существенности отклонения выборки от этого распределения.

Сначала по способу параграфа 2 главы 8 определяется выборочная средняя $\bar{x}=(\Sigma fX)/\Sigma f=296/98=3,02041$ семян сорняков на выборку.

После этого мы устанавливаем распределение Пуассона, взяв эту среднюю в качестве параметра μ . Конечно, всегда можно вычислить это распре-

деление и на основе какого-либо теоретического значения средней, но обычно это не делается, и для построения распределения почти всегда используется выборочная средняя. В данном случае производится оценка только однородности выборки, т. е. решается вопрос: существует ли для всей выборки некоторая константная вероятность встретить семя сорняка?

ТАБЛИЦА 207

Распределение 98 субвыборок семян тимофеевки весом в четверть унции, сгруппированных по числу семян сорняков. Подбор распределения Пуассона и оценка существенности отклонения от этого распределения

Число семян сорняков	Число субвыборок f	Гипотетические численности	Отклонения от теоретического распределения	(Отклонение) ²
				гипотетическая численность
0	3	4,78	-1,78	0,663
1	17	14,44	2,56	0,454
2	26	21,81	4,19	0,805
3	16	21,96	-5,96	1,618
4	18	16,58	1,42	0,122
5	9	10,02	-1,02	0,104
6	3	5,04	0,59	0,041
7	5	2,18		
8	0	0,82		
9	1	0,28		
10		0,08		
Более 10		0,01		
Итого	98	98,00	0,00	$\chi^2 = 3,807$

Приведенные ниже вычисления проводятся при помощи логарифмов. Последовательность системы расчетов вполне ясна из примера. После определения логарифмов для двух первых гипотетических численностей логарифмы последующих численностей находятся путем 1) прибавления логарифма μ и 2) вычитания логарифмов последовательного ряда целых чисел, начиная с 2. Это порядок расчета для вас будет очевидным, если гипотетические численности представить в такой форме:

$$\frac{n}{e^\mu}, \quad \frac{n\mu}{e^\mu}, \quad \left(\frac{n\mu}{e^\mu}\right) \cdot \left(\frac{\mu}{2}\right), \quad \left(\frac{n\mu^2}{2e^\mu}\right) \cdot \left(\frac{\mu}{3}\right).$$

Вычисления проводятся по такой схеме:

Символ	Логарифм	Гипотетическая численность
$n = 98$	1,99123	4,78
e^μ	1,31175	
n/e^μ	$3,02041 \times 0,434295 = 0,67948$	
$\mu = 3,02041$	0,48007	14,44
$\mu n/e^\mu$	1,15955	
μ	0,48007	21,81
2	1,63962	
$\mu^2 n/2e^\mu$	0,30103	
	1,33859	

Символ	Логарифм	Гипотетическая численность
μ	$\frac{0,48007}{1,81866}$	21,96
3	$\frac{0,47712}{1,34154}$	
$\mu^3 n / 2 \cdot 3 e^\mu$		
μ	$\frac{0,48007}{1,82161}$	16,58
4	$\frac{0,60206}{1,21955}$	
$\mu^4 n / 2 \cdot 3 \cdot 4 e^\mu$		
μ	$\frac{0,48007}{1,69962}$	10,02
5	$\frac{0,69897}{1,00065}$	
$\mu^5 n / 2 \cdot 3 \cdot 4 \cdot 5 e^\mu$		
μ	$\frac{0,48007}{1,48072}$	5,04
6	$\frac{0,77815}{0,70257}$	
μ	$\frac{0,48007}{1,18264}$	2,18
7	$\frac{0,84510}{0,33754}$	
μ	$\frac{0,48007}{10,81761 - 10}$	0,82
8	$\frac{0,90309}{9,91452 - 10}$	
μ	$\frac{0,48007}{10,39459 - 10}$	0,28
9	$\frac{0,95424}{9,44035 - 10}$	
μ	$\frac{0,48007}{9,92042 - 10}$	0,08
10	$\frac{1,}{8,92042 - 10}$	
Итого . . .		97,99

Вычисления начинаются с того, что из логарифма 98 вычитается логарифм e^μ , который равен $\mu \log e = 3,02041 \times 0,434295$. Полученная разность является логарифмом искомого частного; антилогарифм будет численностью субвыборки, не имеющих ни одного семени сорняка. Добавление далее логарифма 3,02041 приводит к логарифму следующей гипотетической численности 14,44. В дальнейшем каждый раз производится прибавление логарифма μ и вычитание логарифма последовательно возрастающего целого числа. Этот процесс продолжается до тех пор, пока не будет получена гипотетическая численность, меньшая 5. Для иллюстрации мы провели процесс несколько дальше этого пункта, но для оценки существенности отклонения

фактического распределения от теоретического в таблице 207 нам все равно пришлось прибегнуть к объединению групп.

Критерий хи-квадрат вычисляется как обычно. Но сколько степеней свободы имеет этот критерий? В данном случае в качестве параметра распределения взята выборочная средняя, вместе с этим для вычисления ожидаемых численностей использовано фактическое число субвыборок 98, чем определяется остаток после пятого (или десятого) класса. Следовательно, из фактического числа классов должно быть вычтено две степени свободы, что дает для критерия хи-квадрат $f=5$. Ясно, что данная выборка не отклоняется существенно от распределения Пуассона. Те 98 мешков семян, из которых взяты пробы (субвыборки), образуют совокупность, подчиненную распределению Пуассона. Вероятность того, что любое случайно взятое зерно окажется семенем сорняка, может считаться постоянной для всего этого семенного материала.

Дисперсия теоретического распределения имеет то же самое значение, что и выборочная средняя 3,0204. Следовательно, $s=\sqrt{3,0204}=1,738$ семени на субвыборку. Стандартное отклонение в распределении Пуассона имеет, как правило, совсем другие свойства, чем в нормальном распределении. Например, нельзя сказать, что в распределении Пуассона при его несимметричности 68,27% объектов находится внутри интервала $\mu \pm \sigma$. Однако, если всю выборку из 98 проб семян в четверть унции рассматривать в качестве единицы с 296 семенами сорняков, то можно считать, что она является выборкой из совокупности, средняя которой равна 296, а поэтому и дисперсия ее также равна 296. Распределение Пуассона при столь большой средней весьма близко к нормальному распределению. Поэтому практически столь большая выборка может быть подвергнута обработке так, как будто она была взята из нормальной совокупности, средняя которой равна $\bar{x}=296$ и стандартная ошибка которой $s_x=\sqrt{296}=17,2$ семени сорняков. Численности семян сорной растительности у двух таких выборок можно сравнить, устанавливая существенность их разности методом следующего параграфа.

Пример 16. Леггатт произвел подсчет семян других видов сорняков в своих 98 пробах. Численности двух из этих видов были такими:

Род	Число семян в пробе							Хи-квадрат	P	
	0	1	2	3	4	5	6			7
Potentilla . . .	37	32	16	9	2	0	1	1	2,931	24%
Carex	36	36	23	2	1				4,476	11%

Пример 17. Данные примера 11 о клопе черепашке хорошо подходят к биномиальному распределению. Однако может представлять интерес применение в этом случае и распределения Пуассона. Хи-квадрат=4,634; $f=5$.

Пример 18. Если вы имеете возможность воспользоваться биометрическими таблицами [5], то возьмите там таблицу 39 и сравните приведенное в этой таблице распределение для $\mu=3,0$ с распределением таблицы 207. Произведите при помощи той же самой таблицы грубую проверку вашего решения примера 17.

Пример 19. Когда μ велико, то распределение Пуассона приблизительно нормально. Для проверки этого произведите расчет распределения при $\mu=15$ или воспользуйтесь для этого биометрическими таблицами. Произведите следующую группировку гипотетических частот при $n=100$:

Средины классов	3	6	9	12	15	18	21	24	27
Частоты	1	1	10	24	30	22	9	2	1

Применяя метод параграфа 6 главы 8, найти $g_1=0,094 \pm 0,24$ и $g_2=0,46 \pm 0,48$.

Пример 20. Выборку таблицы 207 можно рассматривать в качестве единичной субвыборки в 24,5 унции весом из распределения Пуассона при $\mu=296$ семенам сорняков.

Дисперсия этой средней будет 296. Сравнивая этот результат с данными предыдущего примера, вы можете установить, почему большие значения средней распределения Пуассона могут считаться распределенными по нормальному закону.

6. Сравнение выборочных средних, полученных из распределений Пуассона. Небольшое размышление делает очевидным то обстоятельство, что средние, полученные из распределений Пуассона, не могут сравниваться так, как это делалось в отношении средних из биномиальных распределений. Действительно, в этом случае наблюдение дает число появлений события, но не дает числа случаев, когда событие не произошло. В параграфе 12 главы 9 был описан метод оценки хи-квадрат, применяемый при малых выборках. Поскольку большие выборочные средние, полученные из распределения Пуассона, распределены почти нормально, можно считать, что к ним применим приближенный метод оценки существенности, описанный в параграфе 8 главы 8. Приводимый ниже пример напомним этот метод.

ТАБЛИЦА 208

Две выборки по подсчету дрожжевых клеток в 400 квадратах гемацитометра и подбор к ним распределений Пуассона

Выборка		Число квадратов, имеющее определенное количество клеток										Среднее	Хи-квадрат
		0	1	2	3	4	5	6	7	8	9		
1	Наблюденное	103	143	98	42	8	4	2				529	3,117
	Гипотетическое	106	141	93	41	14	5						
2	Наблюденное	75	103	121	54	30	13	2	1	0	1	720	7,073
	Гипотетическое	66	119	107	64	29	15						

Стьюдент [6] приводит два распределения дрожжевых клеток в 400 квадратах гемацитометра, воспроизведенные у нас в таблице 208. Значения хи-квадрат указывают на то, что те и другие данные вполне могут быть выборками из распределений Пуассона. Разность между средними составляет $720 - 529 = 191$ дрожжевую клетку на выборку. Для оценки существенности этой разности мы располагаем двумя дисперсиями, которые в то же время являются и средними (720 и 529). Отсюда находим стандартную ошибку разности средних:

$$\sqrt{720 + 529} = 35,3.$$

Следовательно, $t = 191/35,3 = 5,41$, откуда непосредственно видно, что эти две выборки не относятся к одной и той же совокупности.

Пример 21. Вычислите гипотетические численности в таблице 208, а также значения хи-квадрат.

Пример 22. Стьюдент приводит еще одну выборку по подсчету дрожжевых клеток, имеющую такое распределение:

Число дрожжевых клеток	0	1	2	3	4	5
Число квадратов	213	128	37	18	3	1

Вычислите $\chi^2 = 9,683$, $f = 2$. Это распределение существенно отклоняется от распределения Пуассона.

Пример 23. Применяя методы главы 8, оцените существенность разности средних двух выборок в таблице 208. В этом случае вы считаете, что разности средних распределены нормально, но вы не должны пользоваться равенством $\bar{x} = s^2$. Хотя эта формула

ПЛАНИРОВАНИЕ И АНАЛИЗ ВЫБОРОЧНЫХ НАБЛЮДЕНИЙ

УИЛЬЯМ ДЖ. КОБРАН

1. **Совокупности.** Свою статью 1908 г., в которой было сделано открытие критерия t , Стьюдент начинает следующими словами: «Любой эксперимент может рассматриваться в качестве одного из членов *совокупности* экспериментов, которые могли бы быть осуществлены при тех же самых условиях. Всякая серия экспериментов является выборкой, взятой из такой совокупности.

В таком случае любая серия экспериментов представляет для нас ценность только в той мере, в какой она дает нам возможность судить о статистических константах той совокупности, к которой относится данный ряд экспериментов».

Из всего изложенного в предыдущих главах этой книги мы уже знакомы с этой точкой зрения на данные наблюдений. Результаты, получаемые в биологических опытах, подвержены варьированию, поэтому оценки, которые делаются на основе этих данных, также варьируют и, следовательно, в известной мере остаются неопределенными. Однако вы ясно себе представляете, что если будете повторять данный эксперимент много раз, объединяя вместе результаты, то оценка будет вполне определенно приближаться к некоторому неизменному ее значению, которое может быть названо точным, или вполне определенным, результатом опыта. Основная задача статистического анализа единичного эксперимента состоит в том, чтобы вскрыть то, что имеющиеся данные могут сообщить нам относительно этого точного результата. Критерии существенности и доверительные пределы, о которых говорилось на протяжении всей этой книги, и являются основой наших суждений относительно совокупности экспериментов, из которой ваши данные являются просто одной из выборок.

В такого рода задачах выборка вступает как нечто конкретное, а соответствующая совокупность представляется в известной мере гипотетической. Она является совокупностью экспериментов, которые могли бы быть осуществлены при тех же самых условиях, если бы вы обладали необходимыми для этого средствами, временем и интересом.

В настоящей главе мы будем иметь дело с условиями, при которых совокупность вполне конкретна и определена, а наша задача состоит в том, чтобы получить некоторую желаемую информацию относительно такой совокупности. Возьмем такие примеры:

Совокупность

Початки кукурузы в поле.
Семена в большом хранилище.
Вода в резервуаре.

Ученики третьего класса школы.

Требуемая информация

Среднее содержание влаги.
Процент всхожести.
Концентрация определенных бактерий.
Средний вес.

Если совокупность мала, то иногда наиболее удобно получить нужную информацию путем сбора данных по всей совокупности в целом. Однако чаще всего можно сэкономить определенные средства и время, если произвести наблюдение только над выборкой, полученной из этой совокупности. Конечно, выборка становится неизбежной, если само наблюдение приводит к разрушению объекта.

Настоящая глава посвящена рассмотрению некоторых методов получения выборок и вопросам оценки характеристик совокупности, определяемых на основе данных выборки. В течение последних 20 лет выборочный метод получил признание со стороны весьма широкого круга учреждений, включая сюда государственные бюро, исследовательские торговые организации и прессу, собирающую сведения об общественном мнении. С другой стороны, было много сделано и в отношении усовершенствования как теории, так и практики выборочного метода; появилось большое число работ [2, 3, 4, 11, 13], посвященных обзору методов выборочного наблюдения. В настоящей главе мы дадим изложение общих принципов выборочного метода и покажем, как разрешаются некоторые из тех простейших задач, которые имеют широкое распространение в работах биологического характера. Для знакомства с более сложными задачами мы отсылаем к приводимой далее специальной литературе.

2. Простой пример. В первых главах этой книги вам предлагалось производить отбор выборок с целью определения величины варьирования результатов при переходе от одной выборки к другой и для того, чтобы проверить некоторые из наиболее важных положений статистической теории. Этот же прием будет теперь применен и для иллюстрации современных идей, относящихся к производству выборок из определенных совокупностей.

Допустим, что совокупность состоит из $N=6$ членов, обозначенных буквами от a до f . Шесть значений измеряемой величины у этих членов следующие: $a=1$, $b=2$, $c=4$, $d=6$, $e=7$ и $f=16$. Общая сумма совокупности составляет 36. Для того чтобы произвести оценку этой суммы, производится выборка трех из этих членов.

Один из способов, уже знакомый вам, состоит в том, что буквы от a до f выписываются на бобах или кусочках бумаги, перемешиваются в каком-либо сосуде, и после этого производится извлечение трех букв. В работах по выборочному наблюдению этот способ извлечения выборки называется *простым случайным отбором*, или *случайным бесповторным выборочным наблюдением* (потому, что мы не производим возвращение буквы в сосуд после того, как она была извлечена). Обычно такое проведение простого отбора определяет для каждого члена совокупности одинаковую вероятность попасть в выборку. Можно показать, что этот способ также дает одинаковую вероятность совместного образования выборки для каждой из комбинаций трех букв (например, aej или cde).

Теперь спрашивается: сколь удовлетворительную оценку суммы совокупности мы получаем на основе простого случайного отбора? На этот вопрос нельзя сразу дать ответ. Хотя мы и знаем, как была произведена выборка, но пока мы еще не обсудили вопроса, какими результатами выборки оценивается интересующая нас сумма совокупности. Поскольку выборка состоит из 3 членов, а совокупность содержит 6 членов, то наиболее простой операцией представляется умножение суммы выборки на 2; именно эта операция и будет нами применена. Вы должны отсюда сделать вывод, что любой план выборочного наблюдения состоит из двух частей: правила для извлечения выборки и правила образования по результатам выборки искомого оценок.

Выпишем теперь все возможные выборки из 3 членов, определим по каждой из них оценку суммы и посмотрим, сколь близки эти оценки к точному значению ее 36. Здесь имеется 20 возможных выборок, результаты которых представлены в таблице 209, где последовательные колонки дают: состав выборки, выборочную сумму, оценку суммы совокупности и ошибку этой оценки (оценка *минус* точное значение).

Некоторые выборки, например *abf* и *cde*, оказались очень хорошими. в то время как другие, подобные *abc*, дают плохие оценки. Поскольку в каждом отдельном случае мы не знаем, придет ли к нам при отборе выборки удача или нас постигнет неудача, то мы принуждены оценивать ту или иную схему выборочного наблюдения, основываясь на ее усредненных результатах.

ТАБЛИЦА 209

Результаты всех возможных простых случайных выборок по 3 члена

Выборка	Сумма выборки	Оценка суммы совокупности	Ошибка оценки	Выборка	Сумма выборки	Оценка суммы совокупности	Ошибка оценки
<i>abc</i>	7	14	-22	<i>bcd</i>	12	24	-12
<i>abd</i>	9	18	-18	<i>bce</i>	13	26	-10
<i>abe</i>	10	20	-16	<i>bcf</i>	22	44	+8
<i>abf</i>	19	38	+2	<i>bde</i>	15	30	-6
<i>acd</i>	11	22	-14	<i>bdj</i>	24	48	+12
<i>ace</i>	12	24	-12	<i>bef</i>	25	50	+14
<i>acf</i>	21	42	+6	<i>cde</i>	17	34	-2
<i>ade</i>	14	28	-8	<i>cdf</i>	26	52	+16
<i>adf</i>	23	46	+10	<i>cef</i>	27	54	+18
<i>aef</i>	24	48	+12	<i>def</i>	29	58	+22
				В среднем	18	36	0

Средняя из ошибок оценки с учетом их знаков называется *смещением* оценки (или, вообще говоря, плана выборочного наблюдения). Положительное значение смещения означает, что план выборочного наблюдения дает оценки, которые в целом преувеличены, отрицательное смещение указывает, что оценки, наоборот, преуменьшены. Из таблицы 209 видно, что рассматриваемый план дает несмещенные оценки, так как средняя из 20 оценок равна точно 36, и, следовательно, ошибки оценок, слагаясь друг с другом, дают нуль. При простом, случайном отборе такое положение имеет место для любой совокупности и при любом размере выборки. При построении всякого плана выборочного наблюдения желательно иметь такие несмещенные оценки. Но с другой стороны, не следует отказываться от рассмотрения и такого плана, который дает небольшое смещение, если при этом он имеет некоторые другие привлекательные особенности.

В качестве меры точности данного плана выборочного наблюдения мы применяем дисперсию оценок около точного значения величины в совокупности. Эта дисперсия в нашем случае равна:

$$\frac{\sum (\text{ошибка оценки})^2}{20} = \frac{3504}{20} = 175,2.$$

Здесь применен делитель 20 вместо делителя 19, с которым вы были ранее знакомы; таково общее условие, установленное в среде авторов, излагающих выборочный метод. Суммируя все сказанное, можно сказать, что данный план выборочного наблюдения является несмещенным и имеет стандартную ошибку оценки $\sqrt{175,2} = 13,2$. Ошибка составляет около 37% точной суммы совокупности: очевидно, этот план для данной совокупности нельзя считать очень точным.

При простом, случайном отборе получение выборки всецело предоставлено удачам выбора. Здесь никак не используются те или иные наши знания относительно членов совокупности, которые у нас могут быть. Имея такие предварительные знания о предмете, мы должны обладать умением на основе их ввести в процессе простого, случайного отбора соответствующие улучшения. Многие из ранних исследований, относящихся к методам выборочных обследований, были направлены на использование преимуществ,

которые дает надлежащая информация о подвергнутой выборочному наблюдению совокупности.

В порядке иллюстрации этого допустим, что перед планированием выборочного наблюдения мы имеем основание ожидать, что f будет давать более высокие значения, чем любой другой член совокупности. Как можно использовать эти сведения? Ясно, что оценка, полученная из выборки, в значительной мере будет зависеть от того, попадет ли в выборку f или не попадет. Это положение может быть проверено по таблице 209; каждая выборка, содержащая f , дает преувеличенную оценку и каждая выборка, не имеющая f , дает преуменьшенную оценку.

В таком случае представляется, что лучше сделать так, чтобы обеспечить появление f в каждой выборке. Этого мы можем достигнуть, разделяя совокупности на две части или на два слоя. Первая часть, состоящая только из одного f , будет представлена в выборке полностью, а из части второй, состоящей из a, b, c, d и e , мы берем случайную выборку из двух членов, что позволяет сохранить общий размер выборки, равный трем наблюдениям.

В данном случае необходимо некоторое размышление относительно того, какой должна быть выборочная оценка точной суммы совокупности. Если просто удваивать выборочную сумму, как делалось в предыдущем, то это придаст слишком большой вес значению f и, согласно отмеченному выше, всегда будет приводить к преувеличенной оценке точной суммы. Решение задачи состоит в том, чтобы применить этот подход к каждой части совокупности в отдельности. Сумму первой части (16) мы знаем точно, так как всегда будем наблюдать только f . Во второй же части совокупности, из которой выбираются 2 члена из 5, представляется естественным умножение суммы выборки на $\frac{5}{2}$ или 2,5. Следовательно, надлежащей оценкой суммы совокупности будет:

$$16 + 2,5 \times (\text{выборочная средняя слоя II}).$$

ТАБЛИЦА 210

Результаты всех возможных послыхных случайных выборок при неодинаковом доле отбора. Объяснение в тексте

Выборка	Сумма выборки из слоя II (T_2)	Оценка $16+2,5T_2$	Ошибка оценки
<i>abf</i>	3	23,5	-12,5
<i>acf</i>	5	28,5	-7,5
<i>adf</i>	7	33,5	-2,5
<i>aef</i>	8	36,0	0,0
<i>bef</i>	6	31,0	-5,0
<i>bdf</i>	8	36,0	0,0
<i>bef</i>	9	38,5	+2,5
<i>cdf</i>	10	41,0	+5,0
<i>cef</i>	11	43,5	+7,5
<i>def</i>	13	48,5	+12,5
В среднем		36,0	0,0

В таблице 210 даны оценки при всех 10 возможных выборках этого вида. Опять заметим, что оценка здесь несмещенная. Дисперсия ошибок будет:

$$\frac{\sum (\text{ошибка оценки})^2}{10} = \frac{487,50}{10} = 48,75.$$

Стандартная ошибка равна 7,0, или 19% от точной суммы. Это дает заметное улучшение стандартной ошибки, которая при простом, случайном отборе была 13,2.

Такой план выборочного наблюдения получил название *случайного послыоного выборочного наблюдения* при неодинаковых долях отбора. Последняя часть этого термина означает в нашем случае тот факт, что слой I входит в выборку полностью, в то время как из слоя II производится отбор 2 единиц из 5, т. е. этот слой в выборке представлен только на 40%. Послыоный метод отбора состоит в разделении совокупности на частные совокупности, или слои, внутри которых наблюдается меньшее варьирование, чем в первоначальной совокупности; в выборку у нас входили различные, надлежащим образом установленные доли от этих частей совокупности. Более полное обсуждение этого вопроса проведено в параграфах 8 и 9 этой главы.

Пример 1. Допустим в предыдущем примере, что вы ожидаете высокие значения у двух величин e и f . При этом вы решаете, что выборка должна состоять из e и f и одного случайно отобранного члена a , b , c , d . Установить, какой должна быть несмещенная оценка суммы совокупности, и показать, что стандартная ошибка такой оценки равна 7,7. (Эта схема образования выборки оказывается не столь точной, как план с выделением части, состоящей только из одного f , в связи с тем, что фактическое значение e не очень велико.)

Пример 2. Если в предыдущем примере допустить, что f — велико, d и e — среднего размера и a , b , c — малы, то можно произвести расчеты послыоной выборки с тремя частями (слоями) совокупности. Такая выборка будет состоять из f , одного из d и e и одного члена, отобранного из a , b и c . Определите несмещенную оценку суммы совокупности для каждой из 6 возможных в этом случае выборок и покажите, что ее стандартная ошибка равна 3,9.

3. Вероятностное выборочное наблюдение. Приводя предыдущие примеры, мы намеривались дать вам представление о *вероятностном выборочном наблюдении*. Это общее наименование дается плану выборочного наблюдения, в котором:

1) каждый член совокупности имеет известную нам вероятность попасть в выборку;

2) выборка берется при помощи некоторого способа случайного отбора, согласованного с этими вероятностями;

3) эти вероятности отбора учитываются при построении оценок, определяемых по выборке.

Отметим, что вероятности отбора для всех членов совокупности не обязательно должны быть одинаковы; достаточно того, чтобы эти вероятности были известны.

В первом примере предыдущего параграфа каждый член совокупности имел одинаковый шанс попасть в выборку, и поэтому каждый член выборки входил в оценку суммы совокупности с одним и тем же весом. Но во втором примере член совокупности f имел вероятность появления в выборке, равную единице, при вероятности $\frac{2}{5}$ для остальных членов совокупности. Такое неравенство вероятностей при образовании оценки было компенсировано тем, что этим остальным членам выборки был приписан вес, равный $\frac{5}{2}$. Использование неодинаковых вероятностей при некоторых видах совокупностей приводит к существенному выигрышу точности (см. параграф 9 этой главы).

Знание вероятностей выборочного наблюдения имеет значение в нескольких отношениях. Теория вероятностей позволяет определить для различных схем выборочного наблюдения ожидаемые смещения и стандартные ошибки выборочных оценок. Этим путем мы получаем широкую возможность изучения области применения, преимуществ и ограничений каждой такой схемы. Такого рода информация во многом помогает произвести подбор подходящей для некоторого частного случая схемы выборочного наблюдения. Как мы увидим далее, большинство вероятностных схем выборочного наблюдения дает возможность по выборочным результатам произвести расчет стандартной ошибки для той или иной оценки и доверительные границы для точного значения параметра совокупности. Таким образом, при проведении выборки, в основе которой лежит знание вероятностей, мы получаем некоторые представления о том, сколь точны оценки.

Случайный процесс выборочного наблюдения не является единственным путем получения выборки. Один из возможных методов предназначается для изучения совокупности в аспекте «средних», или «типичных», членов ее; в этом случае в выборку берутся только эти члены. Если внутри совокупности наблюдается большая изменчивость и если выборка мала, то этот метод часто дает более точные оценки, чем способ случайного отбора. Другой способ отбора ограничивается тем, что в выборку берутся только такие члены совокупности, которые наиболее доступны для наблюдения. Если тюки товара сложены на складе в компактную массу, то выбор из этой массы внутренних тюков представляет трудности, и поэтому вполне естественно стремление ограничиться отбором только из числа наружных тюков. Во многих биологических исследованиях даже трудно бывает определить способ, каким можно было бы осуществить случайную выборку; например, как определить численность комнатных мух в городе, полевых мышей на поляне или планктона в океане?

Недостатком всех этих неслучайных методов отбора является то, что после получения выборки не представляется возможным определить ее точность. Члены совокупности, собранные в качестве типичных, могут оказаться отнюдь не столь типичными. Наружные тюки товара могут быть не такими, как внутренние. При этих методах отбора становятся неприменимыми формулы стандартной ошибки для оценки и формулы для доверительных границ, относящихся только к методу случайного выборочного наблюдения. Следовательно, всегда желателен применение случайного выборочного наблюдения, кроме тех совершенно очевидных случаев, когда это не представляется возможным или обходится слишком дорого.

4. Определение состава совокупности. Для того чтобы осуществить случайное выборочное наблюдение, мы должны тем или иным путем подразделить совокупность на единицы, называемые *единицами наблюдения*, которые и служат объектами отбора. Единицы наблюдения должны быть легко различимыми, не перекрывать друг друга и в то же время все вместе образовывать всю совокупность. Далее для того, чтобы произвести тот или иной случайный отбор единиц наблюдения, мы должны иметь возможность занумеровать или составить *перечень* всех этих единиц. Как мы увидим, совсем не обязательно всегда фактически составлять их полный список, но возможность перечисления все же должна существовать. Легко, например, составить такой список, если совокупность состоит из 5000 карточек, строго расположенных в ряд, или из 300 початков кукурузы, находящихся в мешке, или из деревьев небольшого сада. Но иногда такое разложение совокупности на отдельные единицы наблюдения, которые могли бы быть перечислены, представляет довольно трудную практическую задачу.

Хотя мы говорим о совокупности, как о чем-то вполне конкретном и определенном, однако иногда может обнаружиться некоторая неясность в отношении границ совокупности, которые не будут точно определены до тех пор, пока не составлен план выборочного наблюдения. Прежде чем мы сможем отграничить совокупность ферм или санаториев, мы должны определить, что следует считать фермой или санаторием. Такое определение может потребовать подробного изучения соответствующих объектов, и окончательное решение, к которому мы придем в результате этого, может быть в известной мере условным.

Всегда следует иметь в виду два требования: во-первых, определение совокупности должно соответствовать целям выборочного наблюдения и, во-вторых, оно должно быть удобным для работы (т. е. лицо, собирающее сведения, должно всегда быть готовым ответить, находится ли тот или иной объект в данной совокупности или вне ее).

Иногда имеющиеся списки ферм, маслобойных заводов, санаториев и др. оказываются недостаточными. Такие списки могут содержать данные о некоторых объектах, которые не относятся к нашей совокупности, и не включать в себя членов, относящихся к ней. Они могут основываться на призна-

ках, отличающихся от тех, которые мы считаем нужным взять для определения нашей совокупности. Перед тем как использовать такого рода списки, необходима самая тщательная проверка их с этой точки зрения. Часто пересмотр списка в целях его пополнения и придания ему удовлетворяющего нас состава вполне оправдывает затрату даже значительных усилий, так как это может оказаться более выгодным, чем составление списка заново. В тех случаях, когда список охватывает только часть совокупности, применим способ, при котором отбор объектов из этой части совокупности производится на основе этого списка и в то же время устанавливается какой-либо специальный метод отбора из не вошедшей в список части совокупности. В этих условиях применим метод послылой выборки: все члены, перечисленные в списке, составляют один слой, не попавшие в список — другой.

Образование перечня, если он не берется в готовом виде, может потребовать изобретательности и довольно большой работы. Приведем такой простой пример. Допустим, что мы желаем взять несколько образцов, характеризующих урожайность некоторой культуры; каждый образец с площадки 2 фута \times 2 фута, и берутся они с делянки 200 футов \times 100 футов. Разделим длину этой делянки на 100 отрезков по 2 фута и ширину на 50 отрезков такого же размера. Таким образом, мы получим координатную систему, разделяющую все поле на 100 \times 50, или 5000 квадратов 2 фута \times 2 фута. Для отбора одного квадрата по методу простой случайной выборки нам следует взять случайным образом одно число из ряда чисел от 1 до 100 и другое число из ряда от 1 до 50. Эти координаты будут определять наиболее удаленный от начала системы координат угол квадрата. Но задача становится более трудной, если делянка имеет размеры 163 фута \times 100 футов, и еще более трудной, если вместо прямоугольной делянки имеется поле неправильной формы. Далее, если мы должны произвести отбор некоторого числа небольших площадок, например 6 дюймов \times 6 дюймов, с поля большой площади, то время, требуемое на такой отбор и установление местоположения выборочных площадок, становится слишком значительным. В частности, именно по этой причине при общепринятом способе отбора почвенных образцов [8] отдается предпочтение методу систематического их расположения (параграф 7 этой главы).

Другой метод проведения выборочного наблюдения (применяемый при ботаническом или химическом анализе) состоит в отборе небольшого образца после того, как делянка уже убрана и получена некоторая масса урожая. Эту массу сначала делят на две части и по результату метания монет (или при помощи случайных чисел) решают, какая из двух частей содержит в себе выборку. Эту часть снова делят надвое и продолжают процесс до тех пор, пока не будет получена выборка требуемого размера. На каждой стадии этого процесса рекомендуется придерживаться хорошего практического правила делать каждые две части по возможности одинаковыми, конечно, при условии, что это производится до того, как брошен жребий. Наиболее легкий и скорый способ состоит просто в том, что из массы берется наугад пригоршня, примерно равная заданному размеру образца; этого часто бывает достаточно, но иногда возникают и искажения.

При обследовании городов в Соединенных Штатах в качестве единицы наблюдения берутся кварталы города; перечень этих кварталов строится на основе плана города. В случае широких сельских обследований производится деление карты страны на территории с границами, легко определяемыми в натуре, и некоторые из этих территорий отбираются для образования выборки. Этому и другим методам, в которых единица наблюдения является некоторая площадь территории, присвоено название *выборочного наблюдения площадей*. Часто основным, но не единственным, преимуществом выборочного наблюдения площадей является то, что оно решает задачу приведения совокупности к перечню единиц наблюдения.

Во многих выборочных исследованиях возможны различные типы и размеры единиц наблюдения, на которые подразделяется совокупность. Например, при взятии образцов почвы при помощи бура размер и форма образца

устанавливаются наблюдателем. То же самое относится и к способу выделения площади для выборочного скашивания какой-нибудь культуры. При зубоврачебном обследовании учащихся пятого класса какого-либо города мы можем за единицу наблюдения взять каждого отдельного ребенка и отобрать нужную выборку из объединенного для всего города списка этих учеников. Однако технически более просто в качестве единицы наблюдения взять школу, произвести выборку из совокупности школ и обследовать каждого ученика пятого класса в выбранных школах. Этот прием, когда единица наблюдения является некоторой естественной группой (у нас школа), состоящей из более мелких единиц, которые нас и интересуют (у нас ученики), получил наименование *гнездового выборочного наблюдения*.

Если вы встретитесь с необходимостью выбора из некоторого числа различных единиц наблюдения, то следует руководствоваться правилом, по которому выбирается единица, дающая наибольшую точность при затрате определенных средств. При фиксированном размере выборки (например, 5% совокупности) обычно большая единица наблюдения дает меньшую точность, чем малая единица, хотя имеются и исключения из этого правила. В противоположность этому 5%-ная выборка при большой единице наблюдения, вообще говоря, стоит дешевле и производится легче, чем при малой единице. Полное и исчерпывающее сопоставление преимуществ двух видов единиц наблюдения может потребовать специального исследования, в процессе которого понадобится вычисление для каждой такой единицы ошибок выбора и стоимости (или затрачиваемого времени).

5. Простое случайное выборочное наблюдение. В этом и последующих параграфах будут рассмотрены некоторые из наиболее известных методов отбора случайных выборок. В нашу задачу будет входить построение такого плана выборочного наблюдения, который давал бы наивысшую точность при затрате определенных средств или, что равносильно предыдущему, давал бы желаемую точность при затрате минимума средств. На знакомство с принципами такого планирования следует затратить некоторое время, так как это позволит нам определить важность любых сведений, которые вы можете получить относительно структуры совокупности и стоимости выборочного наблюдения.

В параграфе 2 этой главы вы уже познакомились с *простым случайным выборочным наблюдением*. Этот метод состоит в том, что члены выборки получены путем независимого отбора при одинаковой вероятности попасть в выборку. В порядке иллюстрации получения случайной выборки при помощи таблицы случайных чисел допустим, что совокупность состоит из $N=372$ членам и что требуемый размер выборки $n=10$. Сначала в таблице 2 наугад берется какое-нибудь трехзначное число, положим 539, стоящее в 11-й строке и 80—82-ом столбцах. Идя по этим столбцам вниз, отмечаем первые десять трехзначных чисел, не превосходящих 372. Они у нас будут: 334, 365, 222, 345, 245, 272, 075, 038, 127 и 112. Выборка будет состоять из тех единиц наблюдения, которые числятся под этими номерами в вашем перечне совокупности. Если какое-либо число встречается дважды или несколько раз, пропустите его после первого появления и продолжайте процесс до тех пор, пока не получите десять *различных* чисел.

Этот способ в тех случаях, когда первой цифрой числа N является 1, 2 или 3, потребует от вас пропусков в таблице большого количества чисел, потому что они будут слишком велики (в приведенном выше примере для того, чтобы набрать в выборку 10 чисел, нам пришлось просмотреть 27 чисел). Это не может служить препятствием, если в нашем распоряжении большое количество случайных чисел. Другой способ состоит в использовании всех трехзначных чисел вплоть до $2 \times 372 = 744$. Начиная с того же самого места, мы найдем, что числами, не превосходящими 744, будут: 539, 334, 615, 736, 365, 222, 345, 660, 431 и 427. Теперь из всех чисел, больших 372, вычтем 372. В результате получим номера членов выборки: 167, 334, 243, 364, 365, 222, 345, 288, 59 и 55. Например, при $N=189$ мы можем отбирать числа вплоть

до $5 \times 189 = 945$, после чего вычитать 189, или 378, или 567 или 756, в зависимости от того, какое из этих вычитаний будет возможным.

Напомним, что простое случайное выборочное наблюдение сводится к отбору членов выборки по жребию. Этот метод часто бывает достаточным, если совокупность варьирует не в слишком широких пределах и если, в частности, ожидаемая доля совокупности, входящая в выборку, составляет от 20 до 80%. Но, с другой стороны, если у вас имеются некоторые сведения об особенностях варьирования совокупности, например, есть основание ожидать более высокого уровня явления в одних частях совокупности, чем в других, то методы, описываемые далее, могут оказаться более точными.

Если Y_i ($i=1, 2, \dots, N$) означает изучаемую переменную, то стандартное отклонение совокупности σ определяется формулой

$$\sigma = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N-1}},$$

где \bar{Y} средняя из совокупности Y_i и \sum обозначает суммирование по всем членам совокупности.

Стандартная ошибка средней для простой случайной выборки объема n будет

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} \sqrt{1-\phi},$$

где $\phi = n/N$ означает *долю выборки*, т. е. долю совокупности, вошедшую в выборку (доля выборки обычно обозначается символом f , но здесь мы применяем ϕ , чтобы избежать смешения с принятым в предыдущем обозначении числа степеней свободы через f).

Член этой формулы σ/\sqrt{n} вам уже знаком: это обычная формула стандартной ошибки выборочной средней. Вторая часть формулы $\sqrt{1-\phi}$ известна под названием *поправки для случая ограниченной совокупности*. Эта поправка вводится потому, что в данном случае выборка берется из конечной совокупности объема N , а не из бесконечной совокупности, как это предполагается в обычной теории выборочного метода. Отметим, что этот член при $n = N$ превращает стандартную ошибку в нуль, что и следует ожидать, ибо в этом случае мы берем в выборку все члены совокупности. Если n/N меньше 10%, т. е. если выборка составляет меньше 10% всей совокупности, то поправка на ограниченность совокупности близка к 1 и может быть опущена.

На этот результат следует обратить внимание. При большой совокупности с фиксированной величиной варьирования (т. е. при данном значении σ) стандартная ошибка средней зависит главным образом от размера выборки и в значительно меньшей степени от доли совокупности, вошедшей в выборку. Для данного значения σ выборка в 100 наблюдений имеет почти одну и ту же точность как при размере совокупности в 200 000 единиц, так и при размере в 20 000 или 2000 единиц. Некоторые исследователи, основываясь на интуиции, думают, что по выборке из 100 наблюдений, взятых из совокупности в 200 000 объектов, нельзя получить точных результатов, так как в этом случае учитывается только очень незначительная часть всей совокупности. Фактически же вопрос о том, будет или не будет данный план выборочного наблюдения точным, зависит прежде всего от величины σ/\sqrt{n} . Это объясняет, почему выборочное наблюдение может привести к значительному сокращению численности необходимых наблюдений и измерений.

Для оценки стандартной ошибки выборочной средней мы имеем величину

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \sqrt{1-\phi},$$

где s стандартное отклонение выборки, вычисленное обычным способом.

Если выборка берется для оценки *суммы* по всем значениям изучаемой переменной, составляющим совокупность, то такой оценкой будет $N\bar{y}$, а ее

стандартной ошибкой

$$s_{N\bar{y}} = \frac{Ns}{\sqrt{n}} \sqrt{1-\varphi}.$$

При применении простого случайного выборочного наблюдения качественных признаков, когда каждый член выборки относится к тому или другому из двух классов, применяется формула

$$s_p = \sqrt{\frac{pq}{n}} \cdot \sqrt{1-\varphi},$$

где p является долей выборки, относящейся к одному из двух классов. Допустим, что из списка 432 семей, имеющих у себя радиоприемник, случайно выбрано 50 семей, из которых 10 семей указали, что они прослушали определенную радиопередачу. В этом случае $p=0,2$, $q=0,8$ и

$$s_p = \sqrt{\frac{0,2 \times 0,8}{50}} \cdot \sqrt{1 - \frac{50}{432}} = 0,053.$$

Если же оставить без внимания ограниченность совокупности, то мы получим $s_p=0,057$.

Отметим, что формула для s_p справедлива только в том случае, если каждая единица как целое относится к тому или другому из двух классов. Если же вы применяете гнездовой метод наблюдения и производите классификацию отдельных элементов внутри каждого гнезда, то необходимо применять уже другую формулу для s_p . Например, при определении процента больных растений на участке при помощи выборки, состоящей из 360 растений, приведенная выше формула сохраняет законную силу только в том случае, если будет произведен независимый и случайный отбор растений. Однако в целях экономии времени мы можем произвести отбор 40 площадок, каждая из которых состоит из 3 соседних рядков по 3 растения в рядке. В этом случае единицей наблюдения будет являться площадка (гнездо из 9 растений). Если распределение болезни растений на поле имеет форму ясно выраженных пятен, то может случиться так, что отдельные площадки будут иметь или все 9 растений, пораженных болезнью, или ни одного больного растения. При этих обстоятельствах выборка из 40 площадок будет не более точной, чем выборка из 40 единичных случайно отобранных растений, и мы введем сами себя в глубокое заблуждение, если будем считать, что перед нами биномиальная выборка из 360 растений.

Правильный способ расчета s_p для этого случая весьма прост. Для каждой площадки (или единицы наблюдения) вычисляется отдельно p , после чего эти значения p подставляются в приведенную ранее формулу для непрерывной переменной. Так, если p_i является процентом больных растений на i -й площадке, то выборочное стандартное отклонение будет...

$$s = \sqrt{\frac{\sum (p_i - p)^2}{n-1}},$$

где n теперь число площадок (гнездовых единиц). После этого

$$s_p = \frac{s}{\sqrt{n}} \sqrt{1-\varphi}.$$

Например, допустим, что число больных растений на 40 площадках оказалось таким, как это указано в таблице 211.

ТАБЛИЦА 211

Число больных растений (из 9) на каждой из 40 площадок

2	5	1	1	1	7	0	0	3	2	3	0	0	0	7	0	4	1	2	6
0	0	1	4	5	0	1	4	2	6	0	2	4	1	7	3	5	0	3	6

Общее число = 99

Стандартное отклонение этой выборки равно 2,331. Так как доли больших растений на 40 площадках находятся путем деления численностей таблицы 211 на 9, то стандартное отклонение этих долей будет:

$$s = \frac{2,331}{9} = 0,259.$$

Следовательно (считая N большим)

$$s_p = \frac{s}{\sqrt{n}} = \frac{0,259}{\sqrt{40}} = 0,041.$$

Для сравнения обработаем эти же данные по формуле биномиального распределения. По общей сумме таблицы 211 находим $p = 99/360 = 0,275$. Формула для биномиального распределения дает

$$s_p = \sqrt{\frac{pq}{360}} = \sqrt{\frac{0,275 \times 0,725}{360}} = 0,024,$$

что приводит к слишком оптимистическому представлению о точности p .

Часто гнезда не все оказываются одинакового размера. Это может случиться, например, тогда, когда единицами наблюдения являются площадки, содержащие в себе различное число растений, подлежащих классификации. Пусть m_i — число элементов, предназначенных для классификации в i -ой единице наблюдения, а a_i — число элементов, попавших в определенный класс, тогда $p_i = a_i/m_i$. Значение p , определенное как доля по всей выборке в целом, будет $\Sigma a_i / \Sigma m_i$, где каждая сумма берется по всем n гнездовым единицам.

В формуле для s , стандартного отклонения p_i , участвуют уже взвешенные средние квадраты m_i :

$$s = \sqrt{\frac{1}{n-1} \cdot \Sigma \left[\left(\frac{m_i}{m} \right)^2 (p_i - p)^2 \right]},$$

где $\bar{m} = \Sigma m_i / n$ является средним размером гнезда в данной выборке. Эта формула является приближенной; точная формула для s в пригодной для практики форме не найдена. Как и ранее:

$$s_p = \frac{s}{\sqrt{n}} \cdot \sqrt{1 - p}.$$

Для облегчения вычислений более удобно такое выражение:

$$s = \frac{1}{m} \sqrt{\frac{1}{n-1} [\Sigma a_i^2 - 2p \Sigma a_i m_i + p^2 \Sigma m_i^2]}.$$

Суммы квадратов Σa_i^2 и Σm_i^2 и сумма произведений $\Sigma a_i m_i$ вычисляются без обычных поправок на среднюю. Независимо от того, будут или не будут применены эти поправки на среднюю, получится одно и то же значение s , но отказ от поправки позволяет сэкономить время.

Пример 3. Из 16 районов округа взята выборка в 4 района и получено стандартное отклонение 45; показать, что стандартная ошибка средней равна 19,5.

Пример 4. В примере, приведенном в параграфе 2 этой главы, мы имели $N=6$ и $n=3$ при 6 членах совокупности: 1, 2, 4, 6, 7 и 16. Формула для точного значения стандартной ошибки при оценке суммы совокупности такова:

$$\sigma_{N\bar{y}} = \frac{N\sigma}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}.$$

Проверить согласованность этой формулы с результатом 13,2, который мы можем получить, выписав все возможные выборки.

Пример 5. Простая случайная выборка в 100 единиц взята для оценки некоторой части совокупности (например, доли мужского пола), истинное значение которой близко к $1/2$. Определить стандартную ошибку выборочной доли p , если размер совокупности равен: (I) 200, (II) 500, (III) 1000, (IV) 10 000 и (V) 100 000. Обратите внимание на то, как мало изменяется стандартная ошибка, когда N превосходит 1000.

Пример 6. Показать, что коэффициент вариации для выборочной средней является в то же время коэффициентом вариации и для оценки суммы совокупности.

Пример 7. Показать, что при простом случайном выборочном наблюдении качественных признаков стандартная ошибка для p при данных N и n максимальна, когда p равно 50%, но коэффициент вариации для p больше тогда, когда p становится очень малым.

6. Размер выборки. При планировании выборки прежде всего может возникнуть вопрос: «Сколь большую выборку я должен взять?» Хотя по причинам, указанным далее, полный ответ на этот вопрос найти не так просто, все же определенный рациональный подход к этому вопросу может быть указан.

Ясно, что нам не следует брать столь малую выборку, при которой оценка будет с практической точки зрения слишком грубой. Равным образом, следует избегать и слишком большую выборку, которая дает большую точность оценки, чем это требуется. Следовательно, на первом этапе работы мы должны сначала решить, какая ошибка оценки может считаться в данном случае допустимой. Это, в свою очередь, требует внимательного обдумывания вопроса о том, как будет использована полученная оценка и какие следствия повлечет за собой тот или иной размер ошибки. Данный вопрос окончательно может быть решен только в известной мере произвольно, но все же многие исследователи после некоторого раздумья чаще всего убеждаются, что при решении этого вопроса у них меньше колебаний, чем они ожидали ранее.

Следующий этап состоит в том, чтобы выразить эту приемлемую ошибку в терминах доверительных пределов. Допустим, что L является дозированной ошибкой выборочной средней и что мы желаем взять 5%-ную вероятность того, что ошибка будет больше L . Другими словами, мы желаем иметь разумную уверенность в том, что ошибка не превосходит L . Вспоминая, что 95%-ные доверительные пределы для выборочной средней определяются формулой

$$\bar{y} \pm \frac{2\sigma}{\sqrt{n}},$$

мы можем положить

$$L = \frac{2\sigma}{\sqrt{n}}.$$

Отсюда получается искомый размер выборки:

$$n = \frac{4\sigma^2}{L^2}.$$

Для того, чтобы эту формулу можно было бы применить для расчетов, мы должны иметь оценку стандартного отклонения совокупности σ . Часто для этого бывает достаточным руководствоваться результатами предыдущих выборочных наблюдений из той же совокупности или из других, сходных с данной совокупностей. Например, в 1938 г. в некоторых районах штата Северная Дакота была произведена пробная выборка для определения урожайности пшеницы [7]. По выборке из 222 участков была определена дисперсия урожайности в пересчете на акр $s^2 = 90,3$ (бушелей в квадрате). Теперь спрашивается, какое количество участков следует учесть, если мы желаем определить фактический средний урожай с точностью ± 1 бушель при 5%-ном риске в отношении того, что ошибка все же будет больше 1 бушеля? Согласно предыдущему, имеем:

$$n = \frac{4\sigma^2}{L^2} = \frac{4 \times 90,3}{1^2} = 361 \text{ участку.}$$

Так как дисперсия урожайности может меняться при переходе от одного года к другому, то при использовании этих расчетов для планирования выборки в каком-либо из последующих лет на эти расчеты следует смотреть только как на ориентировочные.

При отсутствии данных, полученных из ранее проведенной выборки, иногда можно произвести определение σ , если известен размах варьирования

совокупности, используя для этого установленные в параграфе 2 главы 2 отношения между размахом и σ . Если размах известен для совокупности численность которой превышает 500, то в качестве грубой оценки σ можно принять частное (размах)/6.

Когда производится оценка доли какого-либо качественного признака на основе биномиального распределения, то допустимое значение ошибки L при 95%-ной доверительной вероятности будет выражаться равенством

$$L = 2 \sqrt{\frac{pq}{n}}.$$

Поэтому искомый размер выборки для заданного предела ошибки L будет

$$n = \frac{4pq}{L^2}.$$

В этой формуле p , q и L могут быть выражены как в долях, так и в процентах при условии, что все они выражаются одной из этих форм. Этот результат требует предварительного знания p . Если можно ожидать, что величина p находится между 35 и 65%, то можно ограничиться довольно грубой предварительной оценкой ее, так как для этих значений p произведение pq варьирует очень мало. Если же p близко к нулю или 100%, то для точного определения n требуется уже гораздо большее приближение к фактическому значению p .

В формулах настоящего параграфа поправка на ограниченность размера совокупности была опущена. В большинстве случаев практики это вполне допустимо. Если же оказалось, что вычисленное значение n превосходит 10% численности совокупности N , то новое значение n' , учитывающее и эту поправку на ограниченность совокупности, можно получить из соотношения

$$n' = \frac{n}{1 + \varphi}$$

Например, случайное наблюдение группы из 480 проростков показало, что среди них около 15% больны. Допустим, что мы желаем узнать размер выборки, необходимый для определения процента больных растений p в пределах $\pm 5\%$ и при шансе сделать ошибочное заключение 1 из 20. Первоначальная формула дает:

$$n = \frac{4 \times 15 \times 85}{25} = 204 \text{ проросткам.}$$

На основе этих данных мы, пожалуй, решим, что быстрее будет произвести классификацию каждого проростка, чем планировать выборку, которая охватывает столь большую часть всей группы растений. Если мы все же решились сделать выборку, то необходимо определить исправленную величину n' :

$$n' = \frac{n}{1 + \varphi} = \frac{204}{1 + \frac{204}{480}} = 143.$$

Формулы, приведенные в настоящем параграфе, относятся к простому случайному выборочному наблюдению. При некоторых других методах выборочного наблюдения общие принципы определения n остаются теми же, но формула для доверительных пределов и, следовательно, формула, связывающая L с n , уже меняется. Формулы, применимые в случаях более сложных методов выборочного наблюдения, можно найти в книгах, посвященных этому вопросу [2, 4, 11]. На практике формулы настоящего параграфа часто применяются для предварительного определения n даже тогда, когда не предполагается применение простого случайного выборочного наблюдения. Если предполагаемый метод выборочного наблюдения значительно отличается по своей точности от способа простой случайной выборки, то это предварительное значение n в последующем может быть уточнено.

В тех случаях, когда изучается несколько признаков, то сначала устанавливаются значения n для наиболее важных переменных. Если эти значения не слишком отличаются друг от друга, то можно взять наибольшее из этих n . Если n различаются значительно, то один из применяемых методов состоит в том, что также берется наибольшее n , но для некоторых признаков берется из этой первоначальной выборки выборка меньшего размера, например 200 единиц из 1000 в первоначальной выборке. В других случаях большое расхождение значений n указывает на то, что исследование должно состоять из двух или нескольких самостоятельных обследований.

Пример 8. Для оценки процента незаселенных домов должна быть произведена простая случайная выборка. Желательно иметь эту оценку с точностью ± 1 при 95%-ном уровне доверия. По одной предварительной оценке процент незаселенных домов около 6, а по другой около 4. Какие размеры выборки необходимы при этих двух условиях? Какой размер выборки вы рекомендуете?

Пример 9. Требуется определить общее число крыс в заселенной части большого города с ошибкой, не превосходящей 20%, при шансе сделать неверное заключение 1 из 20. Предыдущее обследование дало среднее число крыс на квартал 9 при выборочном стандартном отклонении 19 (распределение оказалось значительно скошенным). Показать, что необходимо взять простую случайную выборку примерно из 450 кварталов.

Пример 10. Уэст [12] приводит следующие данные по 556 фермам округа Сенека в штате Нью-Йорк, ведущим многоотраслевое хозяйство:

	Среднее	Стандартное отклонение на ферму
Число акров под кукурузой	8,8	9,0
Число акров под зерновыми культурами	42,0	39,5
Число акров под сенокосами	27,9	26,9

Показать, что для заданного коэффициента вариации в 5% простая случайная выборка должна состоять из 240 ферм, чтобы с указанной точностью можно было произвести оценку общей площади, занятой на этих 556 фермах под каждую культуру. (Отметим, что здесь необходимо применить поправку на ограниченный размер совокупности.) Наш пример дает иллюстрацию тех результатов, к которым приходят целый ряд исследователей: для получения достаточно точных оценок при столь небольших совокупностях, какими являются округа, в выборку должна быть взята довольно большая часть всей совокупности.

7. Систематическое выборочное наблюдение. Если предстоит взять 10%-ную выборку из комплекта в 730 карточек, то можно случайным порядком выбрать одно из чисел между 1 и 10, положим 3, и после этого брать каждую десятую карточку, т. е. взять карточки за номерами 3, 13, 23 и т. д., кончая номером 723. Выборка этого вида получила название *систематической выборки*, так как выбор первого ее члена полностью определяет всю выборку в целом.

Систематическое выборочное наблюдение имеет два преимущества перед простым случайным выборочным наблюдением. Эта выборка получается гораздо легче, так как необходимо произвести случайный выбор только одного номера, и вместе с этим она более равномерно распределена по всему перечню, составляющему совокупность. По этой последней причине систематическое выборочное наблюдение часто дает более точные результаты, чем простая случайная выборка. Иногда выигрыш в точности бывает значительным. В условиях широкого и постоянного применения выборочных наблюдений систематический отбор становится почти общепринятой техникой таких наблюдений.

Но у этого метода имеются и два потенциальных недостатка. Если совокупность обладает периодической изменчивостью и если интервал между последовательными объектами наблюдений, отбираемыми в систематическую выборку, оказался совпадающим с длиной волны этого периодического изменения (или кратным ей), то мы можем получить выборку, имеющую явное смещение. Укажем на такие возможные и резко выраженные случаи: систематический отбор городских домов может привести к тому, что выборка будет

содержать слишком много или слишком мало угловых домов; систематическая выборка из книги фамилий может содержать слишком много или слишком мало имен, указанных первыми на странице и которые принадлежат преимущественно мужчинам, или главам семейств, или лицам, пользующимся почетом. Систематический отбор растений на поле может привести к выбору растений, находящихся на каждом рядке в одном и том же местоположении. Этот недостаток систематического отбора можно исключить, если за ним наблюдать и если применить в этом случае какой-либо другой метод выборочного наблюдения или произвести выбор нового случайного числа. При отборе растений на поле мы можем произвести выбор отдельного числа для каждого рядка. Следовательно, перед тем как решиться на проведение систематической выборки, нужно иметь некоторые сведения относительно характера изменчивости совокупности.

Другой недостаток этого метода заключается в том, что для результатов систематической выборки нельзя указать подходящего метода определения стандартной ошибки выборочной средней. В работах, посвященных выборочным наблюдениям и дающих различные формулы для $s_{\bar{y}}$, устанавливается, что каждая из этих формул имеет законную силу при определенном типе совокупности, и каждая формула может быть вполне обоснованно применена только в том случае, если мы имеем уверенность в том, что совокупность относится именно к тому типу, для которого применима данная формула. Однако систематический отбор часто является только частью более сложного плана выборочного наблюдения, при осуществлении которого предусматривается получение несмещенных оценок ошибки выборки.

8. Послойное выборочное наблюдение. Проведение послойной, или типической, выборки проводится по трем этапам.

1. Совокупность подразделяется на несколько частей, называемых *слоями*.

2. Из каждой части берется независимая выборка.

3. В качестве оценки средней всей совокупности берется величина

$$\bar{y}_{st} = \frac{\sum N_h \bar{y}_h}{N},$$

где N_h — общее число единиц наблюдения в h -ом слое, \bar{y}_h является средней h -го слоя и $N = \sum N_h$ — объем всей совокупности. Заметим, что для определения \bar{y}_{st} мы должны знать значения N_h (т. е. размеры отдельных слоев).

Имеется целый ряд причин, обусловивших широкое применение при проведении выборочных наблюдений расслоения совокупности. Можно показать, что различия между средними в слоях совокупности не влияют на ошибку выборочной средней \bar{y}_{st} . Другими словами, ошибка выборочной средней \bar{y}_{st} обусловлена только варьированием между единицами наблюдения, принадлежащими к одному и тому же слою. Если мы сможем образовать слои так, чтобы разнородная совокупность разделилась на части, каждая из которых внутри себя более или менее однородна, то следует ожидать определенный выигрыш в точности по сравнению с простым случайным выборочным наблюдением. Например, производя отбор 24 почвенных или растительных образцов с прямоугольного участка, мы могли бы этот участок разделить на 12 компактных делянок и с каждой из них взять по 2 случайных образца. Так как небольшие части поля обычно более однородны, чем более крупные участки, то, по всей вероятности, такого рода расслоение совокупности приведет к некоторому увеличению точности, хотя опыт показывает, что увеличение оказывается более скромным, чем может показаться с первого взгляда. При определении общей площади под пшеницей по выборке ферм мы могли бы эти фермы разбить на группы по их размеру, используя для этой цели соответствующие сведения. При применении послойной выборки этого вида выигрыш в точности часто бывает значительным.

При послойном выборочном наблюдении мы можем выбирать размер выборки для каждого отдельного слоя. Эта свобода выбора дает нам возмож-

ность с наибольшей эффективностью использовать средства для проведения выборок внутри слоев. В ряде приложений это обстоятельство является основным источником увеличения точности послойной выборки. Далее, в тех случаях, когда по отношению к различным частям совокупности выдвигаются разные цели по их описанию и по проведению выборочного наблюдения, расслоение совокупности дает возможность решать эти задачи раздельно. По этой причине при выборочном обследовании жителей города гостиницы и большие доходные дома часто выделяются в виде самостоятельного слоя совокупности.

Теперь мы рассмотрим вопрос об оценке на основе послойной выборки средней и ее стандартной ошибки. Приведенная выше формула, определяющая среднюю совокупности, может быть написана в такой форме:

$$\bar{y}_{st} = \frac{1}{N} \sum N_h \bar{y}_h = \sum W_h \bar{y}_h,$$

где $W_h = N_h/N$ является относительным *весом*, соответствующим данному слою. Отметим, что выборочные средние \bar{y}_h отдельных слоев взвешиваются по численностям N_h этих слоев. Средняя арифметическая из всех выборочных наблюдений здесь неприменима, исключая один специальный, но достаточно важный случай. Этот случай возникает при *пропорциональном распределении* наблюдений по слоям совокупности, когда из каждого слоя берется в выборку одна и та же доля. При этом пропорциональном распределении

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_h}{N_h} = \frac{n}{N}$$

Отсюда следует

$$W_h = \frac{N_h}{N} = \frac{n_h}{n},$$

и, следовательно

$$\bar{y}_{st} = \sum W_h \bar{y}_h = \frac{\sum n_h \bar{y}_h}{n} = \bar{y},$$

так как $\sum n_h \bar{y}_h$ является суммой всех наблюдений в выборке. При пропорциональном распределении мы избавляемся от необходимости взвешивания средних: выборка становится *взвешенной сама собой*.

Стандартная ошибка y_{st} будет:

$$s(\bar{y}_{st}) = \sqrt{\sum W_h^2 \cdot \frac{s_h^2}{n_h}},$$

где s_h^2 — выборочная дисперсия h -го слоя, т. е.

$$s_h^2 = \frac{\sum (Y_{hi} - \bar{y}_h)^2}{n_h - 1},$$

где Y_{hi} — наблюдение за номером i в выборке из h -го слоя. Эта формула стандартной ошибки \bar{y}_{st} предполагает, что внутри каждого слоя производится простой случайный отбор, и не включает в себя поправку на ограниченный размер совокупности. Если доли отбора φ_h в некоторых слоях превосходят 10%, то следует пользоваться более общей формулой:

$$s(\bar{y}_{st}) = \sqrt{\sum W_h^2 \frac{s_h^2}{n_h} \cdot (1 - \varphi_h)}.$$

При пропорциональном распределении, когда все φ_h равны между собой, эта общая формула упрощается:

$$s(\bar{y}_{st}) = \sqrt{\frac{\sum W_h s_h^2}{n}} \sqrt{1 - \varphi}.$$

Далее, если ожидаемые дисперсии во всех слоях одинаковы (что является вполне обоснованным допущением в ряде приложений метода к сельскохо-

зайственным исследованиям), то мы получим еще одно упрощение:

$$s(\bar{y}_{st}) = \frac{s_w}{\sqrt{n}} \cdot \sqrt{1 - \varphi}$$

Это результат совпадает с формулой стандартной ошибки средней при простой случайной выборке, за исключением того, что вместо выборочного стандартного отклонения s теперь стоит обобщенное стандартное отклонение *внутри слоев* s_w . На практике s_w определяется при помощи дисперсионного анализа выборочных данных.

В качестве примера по применению метода пропорционального распределения рассмотрим данные таблицы 212, взятые из одного из ранних исследований Клафема [1] о возможности оценки урожайности при помощи выборки небольших делянок с посевов зерновых культур. Прямоугольные делянки с ячменем были разделены поперек на 3 равные части — слои. Из каждой такой части методом случайного отбора было взято 10 пробных однорядковых площадок длиной в один метр. Задача состоит в определении стандартной ошибки для оценки среднего урожая в пересчете на один метр рядка.

ТАБЛИЦА 212

*Дисперсионный анализ послынной случайной выборки.
Урожай зерна (в г на 1 м)*

Источник варьирования	Число степеней свободы	Сумма квадратов	Средний квадрат
Общее	29	8564	295,3
Между слоями	2	2073	1036,5
Внутри слоев	27	6491	240,4

В этом примере $s_w = \sqrt{240,4} = 15,5$ и $n = 30$. Так как выборка составляет только незначительную часть всей делянки, то величиной n/N можно пренебречь, и поэтому

$$s(\bar{y}_{st}) = \frac{s_w}{\sqrt{n}} = \frac{15,5}{\sqrt{30}} = 2,83 \text{ г.}$$

Насколько эффективным здесь оказалось расслоение совокупности? Из дисперсионного анализа видно, что средний квадрат между слоями оказался превосходящим средний квадрат внутри слоев более чем в четыре раза. Это указывает на наличие существенных различий в уровне урожайности при переходе от одной части делянки к другой. Можно пойти дальше и определить стандартную ошибку средней, если была применена простая случайная выборка без какого-либо расслоения совокупности. Формула стандартной ошибки средней при простой случайной выборке была:

$$s(\bar{y}) = \frac{s}{\sqrt{n}}$$

где s обычное стандартное отклонение выборки. В рассматриваемой здесь выборке s равно $\sqrt{295,3}$ (берется средний квадрат по строке «общее» таблицы 212). Следовательно, в качестве оценки стандартной ошибки средней при условии простой случайной выборки мы имели бы значение:

$$s(\bar{y}) = \frac{\sqrt{295,3}}{\sqrt{30}} = 3,14 \text{ г}$$

вместо 2,83 г при послынном случайном отборе. Расслоение совокупности привело к снижению стандартной ошибки примерно на 10%.

Это сопоставление не совсем правильное, так как можно указать на то обстоятельство, что значение s в данном случае было вычислено по резуль-

татам послышной выборки, а не по простой случайной выборке, как это следовало бы сделать. Обоснованные методы сравнений для всех видов послышных выборок описаны в [2]. Данный же приближенный метод дает достаточно хорошие результаты при пропорциональном распределении и при отборе из каждого слоя не меньше 10 единиц наблюдения.

9. **Подбор размера выборки по отдельным слоям.** Иногда считают, что при послышном выборочном наблюдении мы должны в выборку брать одну и ту же долю каждого слоя, т. е. устанавливать для каждого слоя одно и то же соотношение n_h/N_h , применяя пропорциональное распределение наблюдений по слоям совокупности. Однако более строгий анализ вопроса показывает, что *оптимальным* распределением является такое, когда n_h пропорциональны $N_h\sigma_h/\sqrt{c_h}$, где σ_h является стандартным отклонением единицы наблюдения в h -ом слое и c_h — стоимостью единицы наблюдения в этом h -ом слое. Этот способ распределения наблюдений приводит к наименьшей стандартной ошибке для оценки \bar{y}_{st} при определенной общей стоимости выборки. Данное правило заставляет нас брать большее по сравнению с пропорциональным распределением число наблюдений в том слое, где наблюдается повышенная изменчивость (σ_h велико), и меньшее число наблюдений в слое, где стоимость наблюдения относительно велика (c_h большое). В этом аспекте установленное здесь положение имеет тот же смысл и направленность, которое относится и к другим статистическим правилам, если они продумываются до конца. Когда стандартные отклонения и стоимости проведения наблюдений по всем слоям одинаковы, данное правило сводится к пропорциональному распределению наблюдений по слоям.

Для того чтобы осуществить такое оптимальное распределение наблюдений, необходимо предварительное определение как соотношений между стандартными отклонениями, так и относительных стоимостей наблюдений в различных слоях. Это определение не обязательно должно быть точным; даже примерные оценки этих отношений часто дают вполне удовлетворительное приближение к оптимальному распределению наблюдений. В тех случаях, когда выборочное обследование совокупности производится повторно, то для оценки указанных величин можно воспользоваться результатами предыдущих наблюдений. Даже если выборка из совокупности берется впервые, то и в этих случаях иногда представляется возможность предвидеть, что одни слои совокупности более доступны для наблюдения, чем другие. В таких случаях обычно бывает оправданной и попытка дать примерную оценку различий в стоимости наблюдений. При других условиях мы лишены возможности заранее с достаточной степенью уверенности предсказать, какие из слоев будут более варьирующими и в каких стоимость наблюдений больше, или у нас есть основание думать, что эти различия достаточно малы. В таких случаях следует применять пропорциональное распределение наблюдений.

Существует одно широко распространенное условие, при котором непропорциональное распределение приносит большую выгоду. Оно возникает тогда, когда основная переменная, подлежащая учету, имеет резко выраженное скошенное или асимметричное распределение. Обычно такие совокупности включают в себя небольшое количество единиц наблюдения, обладающих большими значениями данного признака, и большое количество единиц с малыми значениями его. Переменные, выражающие величины различных экономических показателей, часто относятся именно к этому виду величин; например, общий отпуск товаров со складов, число пациентов по клиникам, количество масла, вырабатываемое на маслобойных заводах, и доход фермы (для определенной системы земледелия).

В случае подобного рода совокупностей расслоение их по размеру признака обычно дает большой эффект, и оптимальное распределение наблюдений может быть во много раз лучшим, чем пропорциональное распределение. В качестве иллюстрации возьмем таблицу 213, где приведены данные о количестве учащихся в пересчете на одно учебное заведение для 1019 колледжей и университетов Соединенных Штатов Америки. Эти данные, относящиеся

в основном к 1952/53 учебному году, можно использовать в качестве источника предварительной информации при планировании на будущее выборки, предназначенной для получения быстрой оценки общего численного состава этих учебных заведений. Учебные заведения были разбиты на 4 группы по их размеру.

ТАБЛИЦА 213

Данные об общем численном составе на один колледж или университет, разбитые на 4 группы

Слой: число учащихся на одно учебное заведение	Число учебных заведений	Общий численный состав слоя	Среднее число учащихся на одно заведение	Стандартное отклонение на одно заведение
Менее 1000	661	292 671	443	236
1000—3000	205	345 302	1 683	625
3000—10 000	122	672 728	5 514	2 008
Более 10 000	31	573 693	18 506	10 023
Итого . . .	1 019	1 884 394		

Отметим, что 31 самый крупный университет, составляющий только 3% всех учебных заведений, имеет 30% общего числа учащихся, в то время как в группе мелких учебных заведений, составляющей 65% от общего числа учебных заведений, обучается только 15% учащихся. Заметим также, что внутригрупповое стандартное отклонение по мере увеличения размеров учебных заведений быстро возрастает.

В таблице 214 приведены расчеты, необходимые для подбора оптимальных размеров выборки внутри каждой группы. Для всех слоев мы здесь считаем стоимость проведения наблюдений одинаковой. Расчеты сводятся к тому, что для каждого слоя определяются величины $N_h\sigma_h$, которые далее суммируются по всем слоям. После этого вычисляются относительные размеры выборок $N_h\sigma_h/\sum N_h\sigma_h$. Эти отношения при умножении их на предполагаемый объем всей выборки n дают размеры выборок по отдельным слоям.

ТАБЛИЦА 214

Вычисления для получения оптимальных размеров выборок по отдельным слоям

Слой: число учащихся	Число учебных заведений N_h	$N_h\sigma_h$	Относительные размеры выборок $N_h\sigma_h/\sum N_h\sigma_h$	Фактические размеры выборок	Доля выборки (в %)
Менее 1000	661	155 996	0,1857	65	10
1000—3000	205	128 125	0,1526	53	26
3000—10 000	122	244 976	0,2917	101	83
Более 10 000	31	310 713	0,3700	31	100
Итого . .	1019	839 810	1,0000	250	

В соответствии с большим значением стандартного отклонения в группе крупных университетов наше правило требует, чтобы из этой группы было взято 37% всех выборочных наблюдений. Допустим, что мы задались взять в выборку всего 250 наблюдений. Согласно указанному правилу, в этом случае требуется взять $0,3700 \times 250$, или 92 университета, хотя данная группа содержит всего только 31 университет. При таком сильно скошенном распределении, как данное, нередко случается так, что для оптимального распределения наблюдений требуется 100 или даже более 100% наблюдений из группы наиболее крупных объектов. Когда возникает такого рода положение,

то следует взять все 100% из слоя с наиболее крупными объектами и применить описанный выше способ расчета для распределения остатка выборки между другими слоями. Придерживаясь этого правила, мы включим в выборку 31 наиболее крупный университет, оставив 219 членов выборки для распределения между первыми тремя слоями. Для первого слоя размер выборки будет:

$$219 \left(\frac{0,1857}{0,1857 + 0,1526 + 0,2917} \right) = 65.$$

Распределение наблюдений по слоям, приведенное во втором столбце справа таблицы 214, требует отбора в выборку около 80% (101 из 122) единиц наблюдения из второй по размеру группы учебных заведений и только 10% из группы мелких учебных заведений. Учитывая организационные удобства проведения выборочного наблюдения, мы могли бы в данном случае взять в выборку все 100% объектов этого второго по величине слоя.

Заслуживает внимания следующий вопрос: «Будет ли такое оптимальное распределение иметь большое преимущество перед пропорциональным распределением наблюдений»? Если этого нет, то мало смысла идти на усложнение вычислений и на применение этого оптимального распределения. Конечно, мы не можем дать ответа на этот вопрос в отношении будущей выборки, которая пока еще не произведена, но можно произвести сравнение этих двух распределений в отношении 1952,53 учебного года, данные о котором у нас имеются. Чтобы это сделать, мы воспользуемся данными таблиц 213 и 214 и формулами стандартного отклонения, приведенными в параграфе 8 этой главы, что дает нам возможность рассчитать стандартные ошибки для оценок общего числа учащихся, установленных при помощи этих двух методов. Стандартная ошибка для оптимального распределения оказывается равной 26 000 против 107 000 при пропорциональном распределении наблюдений. Вычисления показывают, что если бы была взята простая случайная выборка без расслоения совокупности, то соответствующая стандартная ошибка была бы 216 000. Здесь ясно видно как значительное снижение стандартной ошибки при расслоении совокупности, так и дополнительное снижение ее при оптимальном распределении наблюдений. При фактическом проведении намечаемого выборочного наблюдения на основе данного расслоения совокупности выигрыш в точности будет, вероятно, несколько меньшим, чем тот, который указан здесь.

10. Послойное выборочное наблюдение качественных признаков. Когда производится выборочное наблюдение над каким-либо качественным признаком, то оценка, соответствующая послойной выборке, определяется по формуле

$$p_{st} = \sum W_h p_h,$$

где p_h доля признака в выборке из слоя за номером h и $W_h = N_h/N$ вес слоя. Для определения стандартной ошибки p_{st} следует в ранее приведенную в параграфе 8 этой главы формулу подставить вместо s_h^2 величины $p_h q_h$.

ТАБЛИЦА 215

Количество огородов у семей штата Айова, подразделенных на три группы

Слой населения	Число семей N_h	Вес W_h	Размер вы- борки n_h	Имеющие огороды	Процент имеющих огороды
Городское	312 393	0,445	300	218	72,7
Сельское неземледель- ческое	161 077	0,230	155	147	94,8
Земледельческое	228 354	0,325	237	229	96,6
	701 824	1,000	692	594	

В качестве примера рассмотрим выборку из 692 семей штата Айова, взятую для изучения среди ряда других вопросов вопроса о том, какое количество огородов было в штате в 1943 г. Эти семьи были разбиты на 3 группы: городское население, сельское население, не занимающееся сельскохозяйственным производством, и население, ведущее сельскохозяйственное производство. Разбивка на группы была сделана потому, что можно предвидеть, что они могут значительно различаться как по численности, так и по размерам огородов. Соответствующие данные приведены в таблице 215.

Общее количество семей по группам было взято из переписи 1940 г. Выборка была распределена примерно пропорционально числу семей в группе: было намечено взять одну семью из 1000.

Взвешенный средний процент семей штата Айова, имеющих огороды, будет

$$\Sigma W_h p_h = 0,445 \times 72,7 + 0,230 \times 94,8 + 0,325 \times 96,6 = 85,6\%.$$

Эта величина практически совпадает со средним процентом всей выборки 594/692, или 85,8%, что объясняется близостью распределения наблюдений к полной пропорциональности.

Дисперсия этой оценки средней определяется по формуле:

$$\Sigma (W_h^2 p_h q_h / n_h) = 0,445^2 \times 72,7 \times 27,3 / 300 + \text{и т. д.} = 1,62.$$

Стандартная ошибка равна 1,27%.

При выборке такого большого размера оценка средней распределена примерно нормально, поэтому доверительные пределы могут быть определены следующим образом:

$$85,6 \pm 2 \times 1,27 \text{ или от } 83,1\% \text{ до } 88,1\%.$$

Для определения оптимальных размеров выборок внутри слоя следует величины n_h взять пропорциональными $N_h \sqrt{p_h q_h / c_h}$. Если по всем слоям стоимость наблюдений остается одной и той же, что имеет место в большинстве обследований, то отсюда следует, что доля, отбираемая в выборку, n_h / N_h должна быть пропорциональной $\sqrt{p_h q_h}$. Теперь вспомним, что величина \sqrt{pq} изменяется относительно мало, когда p находится в пределах от 25% до 75%. Следовательно, пропорциональное распределение при послылойной выборке качественного признака часто будет иметь довольно большую эффективность. Существенное снижение стандартной ошибки по сравнению с пропорциональным распределением наблюдений оптимальное распределение будет давать только тогда, когда p_h будет близким к нулю или 100%, или тогда, когда проведение наблюдений в различных слоях требует различных затрат.

Данный пример не соответствует строго принципам послылойной выборочного наблюдения в том отношении, что для момента обследования ни размеры групп, ни их веса не были известны точно, а были получены из данных переписи, имеющих трехлетнюю давность. Ошибки в весах отдельных слоев приводят к уменьшению выгоды от послылойной выборки и делают формулы стандартного отклонения не вполне правильными. Однако можно думать, что в данном примере эти искажения будут весьма незначительными. Рассмотрение вопроса о методе послылойной выборки при наличии ошибок в определении весов смотри в [2] и [10].

Пример 11. Целью данного примера является сравнение простого случайного выборочного наблюдения с систематическим выборочным наблюдением при небольшом объеме совокупности. Следующие данные являются результатами взвешивания растений кукурузы (за единицу измерения принятого 10 г) в 40 гнездах, последовательно расположенных в один ряд: 104, 38, 105, 86, 63, 32, 47, 0, 80, 42, 37, 48, 85, 66, 110, 0, 73, 65, 101, 47, 0, 36, 16, 33, 22, 32, 33, 0, 35, 82, 37, 45, 30, 76, 45, 70, 70, 63, 83, 34. Чтобы сохранить ваше время, даем стандартное отклонение этой совокупности в готовом виде — 30,1. Вычислите стандартное отклонение средней для простой случайной выборки из 4 гнезд. Систематическая выборка из 4 гнезд может быть взята путем случайного выбора одного числа между 1 и 10 и отбора после этого каждого 10 гнезда. Найдите среднюю \bar{y}_{sy} для каждой из 10 возможных систематических выборок и вычислите стандартное

отклонение этих средних относительно точного значения Y для всей совокупности. Заметим, что соответствующая формула будет такой:

$$\sigma(\bar{y}_{sy}) = \sqrt{\frac{\sum (\bar{y}_{sy} - \bar{Y})^2}{10}}$$

Проверьте, что при систематической выборке стандартное отклонение на 8% ниже. Что вы думаете относительно происхождения этого различия?

Пример 12. В примере послыого выборочного наблюдения, приведенном в параграфе 2 этой главы, можно видеть, что оценка, которую мы взяли для общей суммы совокупности, равна N_{est} . Исходя из общей формулы для дисперсии y_{st} , произведи проверку того, что дисперсия этой оценки суммы всей совокупности равна 48,75, т. е. той величине, которая была определена в параграфе 2 этой главы непосредственным путем (заметим, что слой 1 не участвует в этой дисперсии, так как для этого слоя $n_h = N_h$).

Пример 13. Оптимальное распределение наблюдений при послыойной выборке для качественного признака и при одинаковой стоимости наблюдения по всем слоям совокупности получается, если взять n_h пропорциональными $N_h \sqrt{p_h q_h}$. Отсюда следует, что фактические значения n_h будут:

$$n_h = n \frac{N_h \sqrt{p_h q_h}}{\sum N_h \sqrt{p_h q_h}}$$

Допустим, что в обследовании относительно численности огородов в штате Айова полученные в выборке значения p_h можно принять за те, какими они являются в совокупности. Показать, что оптимальное распределение выборки дает для соответствующих слоев совокупности 445, 115 и 132 наблюдения и что в этом случае стандартная ошибка процента семей, имеющих огород, будет 1,17% вместо 1,27% при той выборке, которая была взята фактически.

Пример 14. Для совокупности колледжей и университетов, обсуждавшейся в параграфе 9 этой главы, было установлено, что при послыойной выборке из 250 учебных заведений с пропорциональным ее распределением оценка общего числа учащихся всех 1019 заведений будет иметь стандартную ошибку 107000. Проверить это утверждение по данным таблицы 21}. Заметим, что стандартная ошибка этой оценки общей суммы совокупности при пропорциональном распределении будет:

$$N \sqrt{\frac{\sum W_h \sigma_h^2}{n}} \cdot \sqrt{\left(1 - \frac{n}{N}\right)}$$

11. Выборочное наблюдение, проводимое по двум стадиям. Рассмотрим такую группу разнообразных задач выборочного наблюдения: 1) изучение вопроса о содержании витамина А в масле, выработанном на ряде маслобойных заводов; 2) изучение вопроса о содержании протеина в пшенице, полученной с нескольких полей какого-либо района; 3) изучение вопроса о содержании эритроцитов в крови некоторой группы мужчин в возрасте 20—30 лет; 4) изучение повреждений, нанесенных насекомыми листьям деревьев какого-либо сада, и 5) изучение вопроса о количестве неправильно прорезывавшихся зубов у школьников третьего класса в школах большого города. Что в этих исследованиях имеется общего? Во-первых, в каждом из этих исследований имеется единица наблюдения, напрашивающаяся сама собою, естественным порядком, — маслобойный завод, поле под пшеницей, отдельный мужчина, дерево и школа. Во-вторых, и это является самым главным моментом, в каждом из этих исследований избранная единица наблюдения берется не полностью, а из нее берется *субвыборка*. Действительно, в первых трех исследованиях образование субвыборок является необходимым условием. Никто не позволит нам взять все масло, произведенное маслобойным заводом, для определения в нем витамина А, или взять всю пшеницу, собранную с участка, для определения того, сколько в ней содержится протеина, или взять всю кровь человека для проведения анализа на содержание в ней эритроцитов. При исследовании вопроса о повреждениях листьев, нанесенных насекомыми, может быть было бы возможным, хотя и затруднительным, обследовать все листья какого-нибудь отдельного дерева, но если распределение повреждений по саду носит характер пятен, то мы, пожалуй, решим брать только небольшую выборку листьев с каждого отдельного дерева с тем, чтобы включить в обследование возможно большее число деревьев. При обследовании состояния зубов у детей мы можем взять всех учеников третьего класса какой-ни-

будь отдельной школы, но мы также можем охватить выборочным наблюдением ряд школ, обследуя только выборку детей из третьего класса каждой школы, попавшей в выборку.

Этот вид выборочного наблюдения называется *выборочным наблюдением с двумя стадиями отбора*, или иногда *субвыборочным наблюдением*. На первой стадии отбора берется выборка единиц наблюдения первого порядка — маслобойный завод, участок под пшеницей и т. д. На второй стадии (из каждой единицы первого порядка) берется *субвыборка единиц наблюдения второго порядка*, или *субъединицы*.

Как показано на приведенных выше примерах, метод проведения выборки по двум стадиям иногда является единственным практически осуществимым путем, при котором выборочное наблюдение становится возможным. Даже в тех случаях, когда имеется возможность выбора между субвыборочным наблюдением и полным охватом единиц наблюдения первого порядка, выборочное наблюдение, проведенное по двум стадиям, все же дает ббльший простор, так как в этом случае представляется возможность маневрирования как размером выборки единиц первого порядка, так и размером выборки, которая берется из таких единиц первого порядка. В некоторых исследованиях важным преимуществом метода выборочного наблюдения с двумя стадиями является то обстоятельство, что предоставляется возможность составить перечень всех членов совокупности. Часто относительно легко составить список единиц наблюдения первого порядка, но трудно или слишком дорого образовать список всех субъединиц. Составить перечень деревьев некоторого сада и взять из него выборку обычно довольно просто, но задача проведения случайного отбора листьев с какого-либо дерева может быть весьма затруднительной. При выборочном наблюдении по двум этапам эта задача возникает только в отношении тех деревьев, которые попали в выборку, и не требуется никакого перечня всех листьев сада.

При рассмотрении вопроса о выборочном наблюдении по двум стадиям мы будем предполагать, что все единицы наблюдения первого порядка примерно одного и того же размера. Положим, что взята простая случайная выборка из n_1 единиц первого порядка и из каждой такой единицы отобрано одно и то же количество n_2 субъединиц. Оценка стандартной ошибки выборочной средней \bar{y} в пересчете на *субъединицу* в этом случае определяется по формуле:

$$s_{\bar{y}} = \frac{1}{\sqrt{n_1}} \cdot \sqrt{\frac{\sum (\bar{y}_i - \bar{y})^2}{n_1 - 1}},$$

где \bar{y}_i средняя в пересчете на субъединицу в i -ой единице первого порядка. Эта формула не содержит в себе поправочного члена на ограниченность совокупности, но ею вполне можно пользоваться при условии, что выборка содержит в себе меньше 10% всех единиц первого порядка. Отметим, что в эту формулу не входят отдельные наблюдения по субъединицам, а только средние \bar{y}_i для единиц первого порядка. Поэтому, если, например, субвыборки производятся для химического анализа, то в соответствии с общепринятой практикой можно смешать единицы субвыборки и произвести химические определения по каждой единице первого порядка. По этим данным и производится вычисление $s_{\bar{y}}$.

В параграфе 12 главы 10 вы уже познакомились с техникой выделения «компонентов дисперсии» и с применением этой техники к задаче выборочного наблюдения по двум стадиям. Приведенные там данные относились к содержанию кальция в зеленых растениях турнепса; было произведено 4 определения по каждому из 3 взятых листьев. Здесь лист рассматривается в качестве единицы первого порядка, а отдельные определения как субъединицы. С помощью компонентного анализа дисперсии вы имели возможность установить, в какой мере дисперсия выборочной средней зависит от варьирования определений на одном и том же листе и от варьирования при переходе от одного листа к другому. Вы также имели возможность предсказать, как в дан-

ном опыте будет меняться дисперсия выборочной средней при различном числе взятых в выборку листьев и при различном числе определений на каждом листе.

Так как эта техника анализа находит широкое применение в выборочных наблюдениях с двумя стадиями, то мы здесь повторим некоторые из полученных ранее результатов. Наблюдение над любой субъединицей рассматривается в этом случае как сумма двух независимых частей. Одна часть, связанная с единицей первого порядка, остается неизменной для всех единиц второго порядка, входящих в определенную единицу первого порядка, и варьирует в связи с переходом от одной этой единицы к другой с дисперсией σ_1^2 . Вторая часть, которая служит для измерения различий между единицами второго порядка, варьирует независимо от первой части при переходе от одной субъединицы к другой и характеризуется дисперсией σ_2^2 . Допустим, что выборка состоит из n_1 единиц первого порядка, в каждой из которых взято по n_2 субъединиц. Тогда выборка в целом состоит из n_1 независимых значений первого члена и содержит в себе $n_1 n_2$ независимых значений второго члена. Следовательно, дисперсия выборочной средней в пересчете на субъединицу будет:

$$\sigma_y^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2}.$$

Оценку этих двух компонентов дисперсии σ_1^2 и σ_2^2 можно сделать, как это будет показано, при помощи дисперсионного анализа выборки с двумя стадиями наблюдения. В таблице 216 приводится дисперсионный анализ данных одного из исследований Иммера [6], задачей которого являлась разработка техники выборочного наблюдения по определению процента сахара в полевых опытах с сахарной свеклой. С каждой из 100 делянок дробного учета было взято по 10 корней свеклы; делянки в данном случае являлись единицами первого порядка. Процент сахара был определен по каждому отдельному корню. В порядке воспроизведения условий полевых опытов был вычислен средний квадрат «между делянками», который был получен как средний квадрат между делянками блоков, состоящих из 5 делянок каждый. Этот средний квадрат характеризует дисперсию ошибки опыта, которая была бы в опыте с рендомизированными блоками при 5 вариантах.

ТАБЛИЦА 216

Дисперсионный анализ данных о процентном содержании сахара в корнях сахарной свеклы (в пересчете на один корень)

Источник варьирования	Число степеней свободы	Средний квадрат	Оцениваемые параметры
Между делянками (единицы первого порядка)	80	2,9254	$\sigma_2^2 + 10\sigma_1^2$
Между корнями (субъединицы) внутри делянок	900	2,1374	σ_2^2

Оценкой компонента дисперсии «между делянками» σ_1^2 будет:

$$s_1^2 = \frac{2,9254 - 2,1374}{10} = 0,0788.$$

Здесь делитель 10 является числом корней (субъединиц), взятых с делянки. В качестве же оценки компонента «внутри делянок» σ_2^2 мы имеем:

$$s_2^2 = 2,1374.$$

Следовательно, если какой-либо новый эксперимент будет иметь n_1 повторений при n_2 корней, взятых с каждой делянки в выборку, то ожидаемая дисперсия средних по вариантам будет

$$s_y^2 = \frac{0,0788}{n_1} + \frac{2,1374}{n_1 n_2}.$$

Возьмем некоторое число вопросов, на которые можно ответить, опираясь на эту формулу. Какова будет точность средних по вариантам в опыте с 6 повторениями и 5 корнями, взятыми с каждой делянки? В таком опыте мы ожидаем:

$$s_{\bar{y}} = \sqrt{\frac{0,0188}{6} + \frac{2,1374}{30}} = 0,29\%.$$

Так как в данном случае выборка достаточно велика, то верное значение среднего процента сахара при 95%-ном уровне доверия будет находиться в пределах $\pm 2 \times 0,29 = 0,58\%$.

Какая комбинация n_1 и n_2 требуется для того, чтобы стандартная ошибка средних по вариантам не превосходила 0,2%? Здесь мы имеем

$$\frac{0,0788}{n_1} + \frac{2,1374}{n_1 n_2} = 0,2^2 = 0,04.$$

В связи с тем, что n_1 и n_2 являются числами целыми, данное уравнение не может быть решено точно; поэтому мы будем добиваться только того, чтобы левая часть уравнения не превосходила 0,04. Вы можете легко проверить, что при 4 повторениях ($n_1=4$) для этого нужно брать с делянки 27 корней. при 8 повторениях достаточно 9 корней с делянки, а при 10 повторениях даже 7 корней. Как и следовало ожидать, при увеличении объема выборочного наблюдения единиц первого порядка объем наблюдений внутри этих единиц уменьшается. При этом также уменьшается и общий объем выборки от 108 корней при $n_1=4$ до 70 корней при $n_1=10$.

12. Распределение затрат при выборочном наблюдении с двумя стадиями. Последний из приведенных выше примеров иллюстрирует одну из особенностей выборки с двумя стадиями отбора: одна и та же стандартная ошибка выборочной средней может быть получена при различных комбинациях значений n_1 и n_2 . Какой из этих способов получения выборки лучший? Естественно, что ответ на этот вопрос зависит от относительной стоимости добавления в выборку новых единиц первого порядка (в нашем случае от стоимости добавочных повторений) по сравнению со стоимостью добавочных субъединиц в каждой единице первого порядка (у нас стоимости отбора корней с каждой делянки). Подобно этому, в выборке зеленых растений турнепса (параграф 12 главы 10) наиболее выгодный план выборки зависит от относительной стоимости отбора дополнительных листьев и стоимости проведения дополнительных определений с каждого листа. Очевидно, что если добавление в выборку единиц первого порядка стоит дешево, а увеличение субъединиц обходится дорого, то наиболее экономичным планом выборки будет получение большого числа единиц первого порядка и небольшого числа (возможно, даже только одной) субъединиц из каждой единицы первого порядка. Однако для общего решения этого вопроса необходимо более точное определение стоимости различных планов проведения выборки.

Конечно, в различных практических условиях характер стоимости выборочного наблюдения будет меняться. Во многих исследованиях при этом способе стоимость выборки (если не считать постоянных общих расходов) может быть приближенно выражена уравнением:

$$\text{стоимость} = c_1 n_1 + c_2 n_1 n_2.$$

Множитель c_1 является средней стоимостью единицы первого порядка, складывающейся из таких элементов, которые явным образом зависят от числа единиц первого порядка, а не от размера субвыборки. Так, в выборке корней сахарной свеклы c_1 включает в себя средние затраты на землю и обработку почвы в пересчете на одну делянку. Если в действительном опыте будет произведено объединение всех корней делянки и будет сделано одно общее для всей делянки определение содержания сахара, то c_1 будет также включать в себя и стоимость проведения этого анализа. С другой стороны, множитель c_2 является средними затратами на субъединицу выборки и складывает-

ся из таких затрат, которые прямо пропорциональны общему числу субъединиц. Сюда включается средняя стоимость размещения и отбор корней в выборку в пересчете на один корень. Стоимость измельчения и смешивания корней, предшествующих анализу на сахар, также входит в c_2 при предположении, что эти затраты примерно пропорциональны числу взятых для анализа корней.

Если из какого-нибудь предшествующего исследования представляется возможным дать предварительную оценку этих компонентов стоимости, то можно дать вполне удовлетворительное решение вопроса по отбору наилучшего соотношения между выборкой и субвыборкой. Эта задача может быть поставлена в двух различных формулировках. В некоторых исследованиях мы устанавливаем определенную дисперсию V выборочной средней и ставим задачу достигнуть этого с меньшими по возможности затратами. В других случаях практики мы не должны превысить некоторый уровень общих затрат C и нам желательно получить столь малое значение V , какое можно только достигнуть при данных издержках. Эти две задачи в основном имеют одно и то же решение. В обоих случаях мы должны привести к минимуму произведение

$$VC = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} \right) (c_1 n_1 + c_2 n_1 n_2).$$

После преобразования это выражение принимает вид:

$$VC = (s_1^2 c_1 + s_2^2 c_2) + n_2 s_1^2 c_2 + \frac{s_2^2 c_1}{n_2}.$$

Можно показать, что это выражение принимает наименьшее значение при условии

$$n_2 = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}}.$$

Этот результат дает наилучшее число субъединиц (у нас корней) на единицу первого порядка (у нас делянка). Значение же n_1 находится решением или уравнения стоимости, или уравнения дисперсии относительно n_1 , в зависимости от того, будет ли стоимость или дисперсия наперед заданной величиной.

В опыте с сахарной свеклой мы имели $s_1^2 = 0,0788$ и $s_2^2 = 2,1374$, откуда

$$n_2 = \sqrt{\frac{2,1374}{0,0788}} \cdot \sqrt{\frac{c_1}{c_2}} = 5,2 \sqrt{\frac{c_1}{c_2}}.$$

В данном исследовании стоимость не была указана, хотя, по всей вероятности, c_1 — расходы на делянку были много выше, чем c_2 . Очевидно, что в этом случае можно посоветовать брать возможно большее число корней с делянки. На практике при решении вопроса о числе повторений в опытах с сахарной свеклой следует вместе с определением сахаристости принимать во внимание также и другие объекты наблюдения.

В опыте с растениями турнепса (параграф 12 главы 10) n_1 является числом растений и n_2 числом определений кальция на листе. В обозначениях настоящего параграфа в этом случае имеем:

$$s_1^2 = s_i^2 = 0,0724;$$

$$s_2^2 = s^2 = 0,0066.$$

Следовательно, наиболее экономически выгодным числом определений на лист здесь будет:

$$n_2 = \sqrt{\frac{c_1 s_2^2}{c_2 s_1^2}} = \sqrt{\frac{0,0066}{0,0724}} \cdot \sqrt{\frac{c_1}{c_2}} = 0,30 \sqrt{\frac{c_1}{c_2}}.$$

Практически n_2 должно быть целым числом, и его наименьшее значение равно 1. Данное уравнение показывает, что $n_2 = 1$, т. е. одно определение

на лист, возможно только, если c_1 во много раз больше c_2 . Фактически же, поскольку c_2 включает в себя стоимость химического анализа, эта величина, по всей вероятности, много больше, чем c_1 . Поэтому как относительно большое варьирование между листьями, так и стоимость анализа приводят к тому, что следует остановиться на одной пробе с листа. Этот пример указывает также и на то обстоятельство, что часто выбор n_2 при помощи данного уравнения может быть произведен даже в таких случаях, когда нет достаточно определенных сведений об относительных затратах. Это происходит потому, что данное уравнение часто приводит к одному и тому же значению n_2 для широкого интервала отношений c_1 к c_2 . Значения n_2 , получаемые из уравнения, подвержены случайным ошибкам, о чем смотри в [2].

В параграфе 14 главы 10 вы встретились с примером выборочного трехстадийного наблюдения над зелеными растениями турнепса. Первая стадия относилась к отбору растений, вторая — листьев с растения и третья — проб с листа. В обозначениях настоящего параграфа дисперсия выборочной средней в этом случае будет:

$$s_y^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_1 n_2} + \frac{s_3^2}{n_1 n_2 n_3}.$$

Воспроизводя здесь уравнение, приведенное в параграфе 14 главы 10, мы имеем:

$$s_y^2 = \frac{0,3652}{n_1} + \frac{0,1610}{n_1 n_2} + \frac{0,0067}{n_1 n_2 n_3}.$$

Чтобы определить наиболее выгодные с экономической точки зрения значения n_1 , n_2 и n_3 , возьмем уравнения затрат в таком виде:

$$\text{стоимость} = c_1 n_1 + c_2 n_1 n_2 + c_3 n_1 n_2 n_3$$

и приведем, как и прежде, к минимуму произведение дисперсии на стоимость. В результате получим:

$$n_2 = \sqrt{\frac{c_1 s_3^2}{c_2 s_1^2}}, \quad n_3 = \sqrt{\frac{c_2 s_3^2}{c_3 s_2^2}}$$

в то время, как n_1 определяется решением уравнения относительно стоимости или уравнения относительно дисперсии. Заметим, что формула для n_2 при трехстадийном выборочном наблюдении остается той же, что и при двух стадиях, и что формула для n_3 является простой аналогией формулы, установленной для n_2 . Подставляя сюда числовые значения компонентов дисперсии, находим:

$$n_2 = \sqrt{\frac{0,1610 c_1}{0,3652 c_2}} = 0,66 \sqrt{\frac{c_1}{c_2}}, \quad n_3 = \sqrt{\frac{0,0067 c_2}{0,1610 c_3}} = 0,20 \sqrt{\frac{c_2}{c_3}}.$$

Поскольку значение n_3 будет меньше единицы при любом практически возможном значении отношения c_2/c_3 , то взятие с листа более одной пробы следует считать неэкономичным. Оптимальное число листьев с растения n_2 зависит от отношения c_1/c_2 . Эта величина будет меняться при изменении условий эксперимента. Если в опыте с какой-либо другой целью выращено много растений, благодаря чему всегда имеется широкая возможность отбора их в выборку, то c_1 будет включать в себя только те дополнительные затраты, которые пойдут на образование выборки из большого числа растений вместо малого числа их. В этом случае оптимальное значение n_2 может также свестись к 1. Если же затраты по выращиванию растений включаются в c_1 , то n_2 может быть и больше 1.

Метод субвыборок находит свое широкое применение в исследованиях урожайности, почв и экологических условий в обширных географических зонах, в полевых опытах, когда производится химический анализ урожая, оценка почв или повреждений насекомыми, и в лабораторных исследованиях. Исследователи, которые прежде не обращали внимания на компоненты изменчивости и на факторы затрат, производимых для получения данных, обратятся

к этим вопросам, могут быть вознаграждены определенной экономией в особенности тогда, когда их работы в некоторых стадиях обходятся довольно дорого.

В тех случаях, когда единицы первого порядка довольно значительно отличаются друг от друга по размеру, рекомендуется некоторая модификация описанных здесь методов. Один из достаточно эффективных способов состоит в том, чтобы производить отбор единиц первого порядка с вероятностями, пропорциональными их размеру, а из каждой такой единицы брать одно и то же число единиц субвыборки. Для знакомства с различными другими подходящими для этого случая методами смотри [4, 11, 13].

Пример 15. Ниже приводится дисперсионный анализ данных об урожае пшеницы и протеине в ней, собранных при выборочном обследовании в Канзасе в 1939 г.; обследование проводилось на основе двух стадий отбора.

Источники варьирования	Урожай (в бушелях на 1 акр)		Протеин (в %)	
	число степеней свободы	средний квадрат	число степеней свободы	средний квадрат
Участки	659	434,52	659	21,388
Выборки внутри участков	660	67,54	609	2,870

С каждого из 660 участков было взято по две субвыборки. Вычислите компоненты дисперсии для урожая.

Ответ: $s_1^2 = 183,49$, $s_2^2 = 67,54$.

Пример 16. По приведенным выше данным для урожая определите дисперсию выборочной средней при выборках, состоящих: 1) из 1 субвыборки с каждого из 800 участков, 2) по 2 субвыборки с каждого из 400 участков; 3) по 8 субвыборок с каждого из 100 участков. *Ответ:* 1) 0,313; 2) 0,543; 3) 1,919.

Пример 17. Требуется при 2 субвыборках с участка взять столько участков, чтобы стандартная ошибка среднего урожая не превосходила $\frac{1}{2}$ бушеля и в то же время стандартная ошибка среднего процента протеина была бы не больше, чем $\frac{1}{8}\%$. Спрашивается, сколько нужно взять участков? *Ответ:* около 870.

Пример 18. Допустим, что в среднем на установление местоположения выбранного участка и на то, чтобы дойти до него, требуется один человеко-час. Для каждого участка производится одно определение процента протеина путем объединения субвыборок, взятых с участка. Стоимость этого определения эквивалентна одному человеко-часу. На определение местоположения, уборку и перевозку растительного образца затрачивается 15 минут. На основе этих данных и дисперсионного анализа для процента протеина (пример 15) вычислите произведение дисперсии на стоимость VC для каждого из значений n_2 от 1 до 5. Какое число образцов с участка будет наиболее экономичным? *Ответ:* 2. Как возрастет стоимость, если при заданном V брать с участка 4 образца? *Ответ:* на 12%.

13. Оценки, основанные на отношениях и регрессии. Оценка отношения может считаться особым способом оценки сумм (или средних) совокупностей; эта точка зрения для ряда задач выборочного метода представляет определенные удобства. Допустим, что вы берете выборку для оценки суммы переменной Y по всей совокупности и что полная численность совокупности была известна на основе какого-либо предшествующего учета. Пусть X обозначает значение переменной при этом предварительном учете совокупности. В этом случае вы можете вычислить отношение:

$$R = \frac{\Sigma Y}{\Sigma X},$$

где суммирование производится по всей выборке. Это отношение является оценкой данного уровня переменной по сравнению с уровнем ее при предшествующем учете. При умножении этого отношения на известную сумму всей совокупности, определенную на основе предшествующего учета, вы получаете оценку суммы совокупности Y , произведенную при помощи данного

отношения. Ясно, что если относительное изменение переменной будет по всем единицам наблюдения одинаковым, то отношение R будет точным и оценка суммы всей совокупности будет в достаточной мере удовлетворительной.

Оценка на основе отношения может применяться также тогда, когда X имеет характер сопутствующей переменной, т. е. она будет иного рода, чем Y . Условием для обоснованного применения такой оценки являются требования, чтобы отношение Y/X оставалось относительно константным по всей совокупности и чтобы сумма X по совокупности была бы известна. Рассмотрим пример оценки валового урожая какой-нибудь культуры на основе выборки ферм некоторого района, произведенной сразу же после жатвы. По каждой ферме, взятой в выборку, мы учитываем общий сбор урожая Y и общую площадь под культурой X . В данном случае отношение $R = \Sigma Y / \Sigma X$ является оценкой среднего урожая на 1 акр. Величина R умножается на общую площадь, занятую данной культурой по району; эта площадь должна быть нам известна точно из каких-либо других источников. Такая оценка будет точна в зависимости от того, как средний урожай варьирует при переходе от фермы к ферме.

Стандартная ошибка оценки суммы по всей совокупности \hat{Y}_R , произведенной на основе отношения, при простой случайной выборке объема n определяется по приближенной формуле:

$$s(\hat{Y}_R) = N \sqrt{\frac{\Sigma (Y - RX)^2}{n(n-1)}}.$$

Эта оценка при помощи отношения не всегда будет более точной, чем оценка $N\bar{y}$ (выборочная средняя, увеличенная во столько раз, сколько единиц наблюдения в совокупности). Доказано, что эта оценка через отношение будет более точной только в том случае, если ρ — коэффициент корреляции между Y и X — превосходит $C_X/2C_Y$, где символ C означает коэффициент вариации. Следовательно, оценки через отношения не должны применяться во всех случаях без разбора, хотя при соответствующих условиях они могут привести к большому выигрышу в точности результатов.

Иногда задачей выборочного наблюдения является получение оценки для самого отношения, например отношения сухого веса к общему весу или отношения чистой шерсти к сырой шерсти. Стандартная ошибка для этой оценки будет

$$s(R) = \frac{1}{\bar{X}} \sqrt{\frac{\Sigma (Y - RX)^2}{n(n-1)}}.$$

Эта формула (при других обозначениях) была уже дана в конце параграфа 5, когда рассматривался вопрос об оценках доли при гнездовом выборочном наблюдении.

В главе 6 обсуждался вопрос о линейной регрессии

$$\hat{Y} = a + bx.$$

При наличии вспомогательной переменной X вы можете установить, что при нанесении выборочных значений Y и X на график соответствующие точки расположатся вблизи некоторой прямой, которая не проходит через начало координат. Это означает, что отношение Y/X не будет константным по всей выборке. В этом случае, как указывалось в параграфе 14 главы 6, вместо оценки отношения рекомендуется произвести оценку линейной регрессии. Для суммы Y по всей совокупности оценка, построенная на основе линейной регрессии, будет:

$$N\hat{Y} = N[\bar{y} + b(\bar{X} - \bar{x})],$$

где \bar{X} является средней X по всей совокупности. Член, заключенный в скобки, является выборочной средней \bar{y} , исправленной по регрессии. Чтобы в этом

убедиться, допустим, что вы взяли выборку, из которой получено $\bar{y} = 2,35$; $\bar{x} = 1,70$; $\bar{X} = 1,92$ и $b = +0,4$. Вашей первой оценкой средней совокупности будет $\bar{y} = 2,35$. Но в выборке среднее значение X оказалось заниженным на величину $1,92 - 1,70 = 0,22$. Далее, величина b указывает вам на то, что увеличению X на единицу соответствует увеличение Y в среднем на $0,4$ единицы. Поэтому, корректируя заниженное значение средней X , вы вводите поправку в выборочную среднюю по регрессии на x :

$$2,35 + 0,4 \times 0,22 = 2,44 = \bar{y} + b(\bar{X} - \bar{x}).$$

Для оценки суммы всей совокупности эту величину следует умножить на число единиц наблюдения в совокупности N .

Стандартная ошибка оценки суммы всей совокупности приближенно определяется по формуле:

$$s_{N\hat{y}} = N s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(\bar{X} - \bar{x})^2}{\Sigma x^2}}.$$

Если возникает необходимость введения поправки на ограниченность совокупности, то к формулам стандартной ошибки, приведенным в настоящем параграфе, следует присоединить множитель $\sqrt{1 - \phi}$. При ограниченных совокупностях оценки, основанные на отношениях и регрессиях, бывают несколько смещенными, но это смещение чаще всего практически несущественно.

14. Выборочные методы в биологии. Задачей настоящего параграфа является описание некоторых видов выборочных наблюдений, широко используемых при наиболее важных применениях выборочного метода в биологии. При оценке показателей урожайности и продуктивности животных и при изучении вопросов собственности и экономики в сельском хозяйстве мы имеем дело с совокупностью ферм. Достаточно надежный список ферм в Соединенных Штатах не всегда можно достать. Как об этом упоминалось, общепринятая практика состоит в том, что в качестве единицы наблюдения берется некоторая площадь земли, отмежеванная легко различимыми границами; она содержит в себе в среднем около 6 ферм. Перед производством выборки эти площади отмечаются на картах крупного масштаба. При сельскохозяйственных обследованиях общенационального характера карты, составленные по этому способу, покрывают всю страну в целом и в своем собранном виде носят название «Выборка владений в сельском хозяйстве» (Master Sample of Agriculture). Эти площади в зависимости от целей исследования могут быть подвергнуты различным группировкам. В некоторых исследованиях имеет смысл составить вместе с этим специальный список крупных ферм, которые становятся самостоятельной группой — слоем.

При обследовании не слишком больших районов, таких, как штат, план выборочного наблюдения строится по схеме послышной выборки из этих площадей. Внутри слоя производится случайная выборка площадей, хотя часто для удобства применяют систематическую выборку. На отобранной таким образом площади данные собираются по всем фермам, находящимся на этой площади.

При обследованиях по всей стране план выборочного наблюдения в целях сокращения расходов и более легкого руководства — более сложный, с применением многостадийного выборочного наблюдения.

При обследовании лесов для определения объема древесины по основным видам и возрастным классам деревьев на первом этапе наносятся контуры площадей, занятых лесом, что производится по имеющимся картам или с помощью аэрофотосъемки. После этого лесные районы подразделяются на части в соответствии с видом насаждений, их густотой и размером деревьев. Эта работа может быть выполнена путем тщательного стереоскопического изучения ранее проведенных аэрофотосъемок, но слишком старые съемки могут ввести в заблуждение. Внутри каждой части (слоя) намечаются делянки,

попавшие в выборку; после чего специальные полевые партии производят на месте детальные измерения. При исследовании может возникнуть ряд задач, относящихся к величине затрат, необходимых для образования таких частей (слоев) и для обмера отобранных делянок, к лучшему размеру и форме делянок и к вопросу о преимуществе оптимального распределения учетных делянок перед пропорциональным [5].

Изучение диких животных является весьма затруднительным в связи с тем, что члены совокупности — рыбы, птицы, бобры, полевые мыши и пр. — не остаются все время на одном и том же месте, а иногда их очень трудно обнаружить. Вообще говоря, при этих исследованиях способ послойной выборки применим только в том случае, когда имеются какие-нибудь сведения относительно распределения и обитания изучаемого вида. В этих случаях в каждом слое выделяется площадь, составляющая единицу наблюдения, входящую в случайную или систематическую выборку.

Подсчет численности животных на таких площадях довольно затруднителен. В некоторых случаях может быть произведен в доступной форме сплошной учет. Например, численность оленей может быть установлена путем сплошного обхода полевой партией всей площади; рыба может быть выловлена сетями или учтена после оглушения ее электрическим током. Некоторые крупные животные, такие, как буйволы и антилопы, на подходящей для этого местности могут быть подсчитаны с аэроплана. Один из способов, применяемый для учета дичи, состоит в том, что в определенный час дня и в определенное время года в течение ряда лет по установленному маршруту производится поездка в коляске, во время которой устанавливается численность попавших в поле зрения птиц. Этот прием не может считаться случайной выборкой, а основывается на предположении, что численности попавших в поле зрения птиц за ряд лет примерно пропорциональны общим численностям птиц в эти годы.

Учет тех видов животных, которые не могут быть обнаружены непосредственно, производится при помощи различных ловушек. Иногда путем повторного отлова удается поймать с выбранной площади до 50% всех животных. Замечая скорость уменьшения улова на ловушку в последовательные промежуточные времена, можно получить некоторое представление о количестве животных, избежавших поимки, и на основании этого сделать соответствующую поправку.

В тех случаях, когда может быть отловлена только относительно небольшая часть большой совокупности, можно те особи, которые были пойманы, окольцевать или как-либо отметить и снова отпустить на свободу. В этих случаях определение размера совокупности может быть сделано по доле повторно выловленных в последующем периоде животных. Оба описанные здесь метода зависят от ряда допущений относительно совокупности, в частности от предположения, что все животные имеют одинаковую вероятность быть выловленными. Можно привести доводы, убеждающие в том, что это не так, но в настоящее время мало что можно сказать по поводу погрешностей этих методов.

Наконец, в случаях, когда любой визуальный подсчет крайне затруднен, биологи изучают различные указания на обитание на основе косвенных признаков, таких, как гнезда, логова, следы или экскременты животных. Полное перечисление методов наблюдения над дикими животными дано Скаттергудом [9].

В целях сокращения изложения мы здесь опускаем целый ряд довольно трудных практических проблем применения выборочного метода в биологии. Вместе с тем выборочные методы непрерывно развиваются в соответствии с неуклонным стремлением биологов к усовершенствованию техники наблюдений. Описанные здесь методы несомненно будут изменяться по мере накопления наших знаний и по мере того, как будут открыты новые возможности применения выборочного наблюдения.

1. Clapham A. R., Journal of Agricultural Science, 19, 214, 1929.
2. Cochran W. G., Sampling Techniques. John Wiley and Sons, New York, 1953.
3. Deming W. Edwards, Some Theory of Sampling. John Wiley and Sons, New York, 1950.
4. Hansen M. H., Hurwitz W. N., Madow W. G., Sample Survey Methods and Theory. John Wiley and Sons, New York, 1953.
5. Hasel A. A., Chapter 19 in Statistics in Mathematics and Biology. Iowa State College Press, 1954.
6. Immer F. R., Journal of Agricultural Research, 44, 633, 1932.
7. King A. J., McCarty D. E., McPeak M., U. S. Department of Agriculture Technical Bulletin 814, 1942.
8. Rigney J. A., Reed J. Fielding, Journal of the American Society of Agronomy, 39, 26, 1947.
9. Scattergood L. W., Chapter 20 in Statistics in Mathematics and Biology. Iowa State College Press, 1954.
10. Stephan F. F., Journal of Marketing, 6, 38, 1941.
11. Sukhatme P. V., Sampling Theory of Surveys With Applications. Iowa State College Press, 1954.
12. West Q. M., Mimeographed Report, Cornell University Agricultural Experiment Station, 1951.
13. Yates F., Sampling Methods for Censuses and Surveys. Charles Griffin, London 1954.

УКАЗАТЕЛЬ СИМВОЛОВ

- a число вариантов
 \bar{a} средняя, соответствующая $x = 0$
 a_i средние отклонения
 b выборочный коэффициент регрессии
 b число вариантов
 $b_{Y1.2}$ выборочный частный коэффициент регрессии
 $b'_{Y1.2}$ стандартный частный коэффициент регрессии
 β коэффициент регрессии
 $\beta'_{Y1.2}$ частный коэффициент регрессии
 C коэффициент вариации
 C поправочный член
 C общая стоимость
 C_D коэффициент вариации разностей
 c стоимость
 c элементы обратной матрицы, весовые коэффициенты Гаусса
 D делитель $b_{Y1.2}$
 D разность средних
 D существенная разность
 δ разность совокупности
 $d_{y \cdot x}$ или $d_{Y \cdot 12}$ отклонение от регрессии
 e основание натуральных, или Неперовых, логарифмов
 ε случайная переменная
 F гипотетическая численность
 F отношение дисперсий
 f наблюдаемые численности
 G произвольное начало
 g_1 мера скошенности
 g_2 мера крутизны
 H_A альтернативная гипотеза
 H_0 нулевая гипотеза
 I классовый интервал
 i показатель подразделения
 j показатель подразделения
 k показатель подразделения
 k_2, k_3, k_4 k — характеристики
 χ^2 компонент
 L допустимая ошибка
 λ коэффициент в сравнении
 m число переменных
 μ средняя совокупности
 N размер совокупности или слоя
 $N(\mu, \sigma)$ нормальное распределение с средней μ и стандартным отклонением σ
 n численность группы или выборки
 $n_0, (nb)_0$ усредненная численность выборки
 O начало координат
 O' начало координат в точке (\bar{x}, \bar{y})
 P или p вероятность
 p_{st} выборочная средняя в послышной выборке качественного признака
 Φ доля наблюдений в выборке
 Q студентизированный размах варьирования
 Q_1, Q_2, Q_3 квартили
 q вероятность неоявления события
 R коэффициент множественной корреляции
 r выборочный коэффициент корреляции
 r_I внутриклассовый коэффициент корреляции
 r_S ранговый коэффициент корреляции Спирмана
 $r_{Y1.2}$ или $r_{12.3}$ частный коэффициент корреляции
 q коэффициент корреляции в нормальной совокупности
 Q_I внутриклассовый коэффициент корреляции в совокупности
 S_i суммы
 Σ суммирование
 s выборочное стандартное отклонение
 $s^2, s_x^2, s_{y \cdot x}^2$ и т. д. средние квадраты
 $s_b, s_{b_{Y1.23}}$ выборочные стандартные отклонения коэффициента регрессии
 s_D выборочное стандартное отклонение разности
 s_{g_1}, s_{g_2} см. стр. 194.
 s_{N_y} выборочное стандартное отклонение суммы
 s_p выборочная стандартная ошибка качественного признака в гнездовой выборке
 $s_{\bar{x}}, s_{\bar{y}}$ выборочная стандартная ошибка средней
 s_Y выборочная стандартная ошибка предсказанного Y
 $s_{\hat{Y}}$ выборочная стандартная ошибка \hat{Y}
 $s_{y \cdot x}, s_{y \cdot 12}$ выборочное стандартное отклонение от регрессии
 s_{-yA} выборочная стандартная ошибка при-

\bar{y} — веденной средней	$x = X - \bar{x}$
$s_{y \cdot 12}$ — усредненное стандартное отклонение от регрессии	x' — нормированное отклонение
s (\hat{Y}_R) — выборочная стандартная ошибка для оценки отношения	\bar{x} — выборочная средняя X
s (\bar{y}_{st}) — выборочная стандартная ошибка \bar{y}_{st}	Y — зависимая переменная
σ — стандартное отклонение	Y — переменная, изучаемая выборочным методом
σ^2 — дисперсия	\bar{Y} — средняя совокупности
σ_A^2 — компонент дисперсии	\hat{Y} — значение Y , определенное по регрессии
σ_x — стандартная ошибка средней	\hat{Y}_R — оценка отношения
T — сумма рангов	$y = Y - \bar{y}$
$t = (\bar{x} - \mu) / s$	\bar{y} — выборочная средняя Y
V — средний квадрат выборочной средней	\bar{y}_A — приведенная средняя
W или w — вес	\bar{y}_{st} — послонная выборочная средняя
X — независимая переменная	\hat{y} или \hat{y}_{12} — отклонение \hat{Y} от \bar{y}
	$Y - \hat{Y}$ — отклонение от регрессии
	z — преобразованный r

ПЕРЕЧЕНЬ МАТЕМАТИЧЕСКИХ ТАБЛИЦ

	<i>Стр.</i>
Арксинус $\sqrt{\text{процент}}$	299—300
Биномиальные коэффициенты	441
Графическая таблица для z и r	175
Доверительный интервал для вероятности 95 и 99%	23—24
Коэффициенты корреляции для 5- и 1%-ных уровней	173
Коэффициенты и полиномы для определения конечных значений и разностей	433
Критические коэффициенты для принятия гипотез при 5%-ном уровне	243
Нормальное распределение, ординаты	198
Нормальное распределение, накопленные частоты	200
Ортогональные коэффициенты и делители в регрессии	327
Отношение дисперсий F для 5%-ного уровня	106
Отношение дисперсий F для 5- и 1%-ного уровней	236—239
Отношение дисперсий F для 25-, 10-, 2,5- и 0,5%-ных уровней	263—266
Отношение σ к размаху варьирования, эффективность размаха	54
Размахи варьирования t_w и t_w , аналогичные t	117
Ранговая корреляция при 5- и 1%-ном уровнях	187
Распределение Стьюдента t	61
Случайные цифры	28—32
Стьюдентизированный размах Q при 5%-ном уровне	241
Сумма рангов при 5- и 1%-ном уровнях	123
Сумма рангов T при 5- и 1%-ном уровнях	126
Chi-квадрат	46

УКАЗАТЕЛЬ АВТОРОВ

- Аллен (Allen F. E.) 293
 Андеркофлер (Underkofler L. A.) 230
 Андерсон (Anderson R. L.) 137, 258, 261, 346
 Андре (Andre F.) 222, 420, 423, 447
 Аркин (Arkin H.) 35
 Банкрофт (Bancroft T. A.) 258, 261, 346
 Барлоу (Barlow P.) 195
 Барнард (Barnard M. M.) 389, 407
 Барллетт (Bartlett M. S.) 222, 271, 272, 296, 302, 305
 Батлер (Butler R. A.) 305
 Бейкер (Baker G. A.) 184
 Бейтсон (Bateson W.) 209
 Белл (Bell M. A.) 48
 Беккер (Becker E. R.) 359
 Беренс (Behrens W. U.) 106, 107
 Берксон (Berkson J.) 437
 Бернетт (Burnett L. C.) 221
 Берроуз (Bourroughs W.) 66
 Биггер (Bigger J. H.) 148
 Бидлс (Beadles J. R.) 66
 Бил (Beale H. P.) 64
 Биттенгер (Bittenger M. W.) 347, 366
 Блисс (Bliss C. I.) 299, 437
 Блэк (Black C. A.) 324, 387
 Бойд (Boyd M. F.) 19, 173, 216
 Бойнтон (Boynnton B.) 76, 77
 Брайан (Bryan A. A.) 79
 Браунт (Brandt A. E.) 145, 163, 352
 Брансон (Brunson A. M.) 403
 Браун Б. (Brown B.) 358
 Браун П. (Brown P. E.) 101
 Бренеман (Breneman W. R.) 97, 102, 140
 Бриндли (Brindley T. A.) 21
 Ван-Слайк (Van Slyke D. V.) 195
 Вейси (Vasey A. J.) 284
 Вильямс (Williams C. B.) 302, 303
 Вос Б. (Vos B. J.) 424
 Вос Р. (Vohs R. L.) 72
 Гальтон (Galton F.) 129, 148, 154, 161, 163
 Гангули (Ganguli M.) 260
 Гаррис Дж. (Harris J. A.) 271
 Гаррис М. (Harris M.) 267
 Гаррисон (Harrison C. M.) 113
 Гаусс (Gauss C. F.) 409, 416
 Гауэн (Gowen J. W.) 360, 420
 Геддес (Geddes W. F.) 420
 Гейнс (Gaines J. C.) 301
 Герленд (Gurland J.) 273
 Госсет (Gosset W. S.) 60
 Грам (Gram M.) 401
 Граут (Grout R. A.) 177, 413
 Грей Г. (Gray H.) 63
 Грей П. (Gray P. H. H.) 223
 Гроув (Grove L. C.) 66
 Гулден (Goulden C. H.) 403
 Гурвиц (Hurwitz W. N.) 455
 Давенпорт (Davenport C. B.) 195
 Данган (Dungan G. H.) 148, 158
 Данкан (Duncan D. B.) 240
 Дарвин (Darwin C. R.) 92
 Дейвид Ф. (David F. M.) 85, 177
 Дейвид П. (David P. A.) 177
 Дейвидсон (Davidson O. W.) 112
 Дейвис (Davis G. N.) 214, 225
 Деккер (Decker G. C.) 222, 420, 423, 447, 449
 Деминг (Deming W. E.) 158, 417, 455
 Джеймс (James G. S.) 273
 Дженкинс (Jenkins M. T.) 48, 148, 257
 Джердел (Gerdel R. W.) 171
 Джессеп (Jessen R. J.) 216
 Диксон (Dixon W. J.) 122
 Дип (Dean H. L.) 99
 Доналдсон (Donaldson H. H.) 77
 Дооз (Dawes B.) 153
 Доусон (Dawson W. T.) 424
 Дуайер (Dwyer P. S.) 405, 407
 Дулитл (Doolittle M. H.) 409, 410
 Дьюрост (Durost N.) 28
 Зелнер (Zoellner J. A.) 387
 Иден (Eden T.) 172
 Йейтс (Yates F.) 10, 211, 234, 289, 293, 389, 455
 Иммер (Immer F. R.) 477
 Ирвин (Irwin M. R.) 218
 Калбертсон (Culbertson C. C.) 99, 163, 373
 Каннон (Cannon C. Y.) 291
 Касида (Casida L. E.) 119
 Катрон (Catron D. V.) 230
 Каффри (Caffrey D. J.) 224
 Кемптори (Kempthorne O.) 290, 295, 352, 387
 Кендалл (Kendall M. G.) 137, 187
 Керк (Kirk E.) 195
 Кетчпол (Catchpole H. R.) 102
 Кинг (King A. J.) 465

- Китчен (Kitchen S. F.) 216
 Кларк (Clarke G. S.) 302, 303
 Клафан (Claphan A. R.) 470
 Клем (Clem M.) 410
 Кокран (Cochran W. G.) 19, 211, 216, 261, 291, 354, 375, 454, 455, 474, 480
 Кокс (Cox G. M.) 375
 Коллинс Э. (Collins E. V.) 146
 Коллинс Дж. (Collins G. N.) 91
 Колтон (Colton R. R.) 35
 Крааторн (Crathorne A. R.) 163
 Кремптон (Crempton E. W.) 66, 158
 Кролл (Crall J. M.) 286, 301, 383
 Курц (Kurtz T. E.) 243, 406
 Кьюлс (Keuls M.) 240, 422
- Лав (Love H. H.) 66
 Ламберт (Lambert W. V.) 49, 50
 Лаш (Lush J. L.) 176
 Левертон (Leverton R.) 401
 Леггатт (Leggatt C. W.) 448, 449, 451
 Лесли (Leslie P. H.) 218
 Ли А. (Lee A.) 161
 Ли У. (Li W. H.) 161, 287
 Линк (Link R. F.) 243
 Линдли (Lindley D. V.) 389
 Линдстром (Lindstrom E. W.) 44, 193, 196, 203, 205, 208, 209, 221
 Липка (Lipka J.) 417
 Лорд (Lord E.) 19, 116
 Лоу (Lowe B.) 231, 234
 Лукас (Lucas H. L.) 291, 383
 Лю (Liu T. N.) 287
 Льюис (Lewis W. H.) 195
- Магистрад (Magistrad O. C.) 77
 Маддок (Maddock H. M.) 230
 Мадоу (Madow W. G.) 455
 Мак-Артур (MacArthur J. M.) 43, 48
 Мак-Карги (McCarty D. E.) 465
 Мак-Лафлин (McLaughlin L.) 113
 Мак-Лейн (McLane J. M.) 91
 Мак-Пик (McPeak M.) 465
 Мак-Фи (McPhee H. C.) 444
 Матер (Mather K.) 207, 219
 Махаланобис (Mahalanobis P. C.) 234
 Майр (Mayr E.) 357
 Менг (Meng C. J.) 287
 Меррингтон (Merrington M.) 19, 266
 Мерсер (Mercer W. B.) 70
 Метцгер (Metzger W. H.) 421
 Митчелл (Mitchell H. H.) 66
 Митчерлих (Mitscherlich E. A.) 421
 Монзелзе (Monselze S. P.) 277
 Морган (Morgan E. C.) 113
 Моррис (Morris V. H.) 171
 Муд (Mood A. M.) 43, 73, 83, 122, 127, 137, 222
 Мюнч (Muench H.) 216
- Нейман (Neyman J.) 43, 269
 Нельсон (Nelson P. M.) 113
 Ньюмен (Newman D.) 240
 Нэйр (Nair K. R.) 375
- Остл (Ostle B.) 19
 Отри (Autrey K.) 291
- Паллаи (Pallai K. C. S.) 116
 Палмер (Palmer R. C.) 317
 Паннетт (Punnett R. C.) 209
 Парк (Park O. W.) 99, 109
- Паркс (Parkes A. S.) 445
 Паттерсон (Patterson H. D.) 291
 Пейдж (Page I. H.) 195
 Пенкуайт (Penquite R.) 417
 Перл (Pearl R.) 77
 Песек (Pesek I.) 401
 Пирсон Э. (Pearson E. S.) 53, 102, 247, 451
 Пирсон К. (Pearson K.) 37, 119, 154, 161, 171
 Пирсон П. (Pearson P. B.) 161
 Портер (Porter R. H.) 278, 350
 Прайс (Price W. C.) 420
- Райт (Wright E. B.) 125
 Редди (Reddy C. S.) 57
 Ригни (Rigney J. A.) 460
 Рид Г. (Read H. S.) 437
 Рид Дж. (Reed J. F.) 460
 Рид Л. (Reed L. J.) 184
 Риц (Rietz H. L.) 204
 Ричардсон Ч. (Richardson C. H.) 139, 213, 214, 245
 Ричардсон Д. (Richardson D.) 230, 275, 314
 Робертс (Roberts H.) 401
- Саттертуэйт (Satterthwaite F. E.) 354
 Серфейс (Surface F. M.) 77
 Скаттергуд (Scattergood L. M.) 484
 Смит А. (Smith A. H.) 426
 Смит Ч. (Smith C. E.) 224
 Смит Ф. (Smith F. B.) 101
 Смит Г. (Smith H. F.) 352
 Смит С. (Smith S. N.) 99
 Снелл (Snell M. G.) 163
 Снедекор Дж. У. (Snedecor G. W.) 5, 6, 7, 9, 44, 140, 163, 218, 234, 257, 268, 330, 350
 Снедекор Дж. Дж. (Snedecor J. G.) 44
 Саундерс (Saunders A. R.) 325
 Спир (Speer V.) 72
 Спирман (Spearman C.) 186
 Спрейг (Sprague G. J.) 377
 Старк (Stark A. L.) 148
 Стефан (Stephan F. F.) 474
 Странд (Strand N. V.) 216, 361
 Стюарт (Stewart R. T.) 132
 «Стьюдент» («Student») 60, 452
 Сукхатме (Sukhatme P. V.) 107, 455
 Суонсон (Swanson P. P.) 102, 130, 401, 426
 Сэведж (Savage R.) 118
- Талли (Talley P. J.) 56
 Там (Tam R. K.) 77
 Типпетт (Tippett L. H. C.) 53
 Томпсон Ч. (Thompson C. M.) 46, 266
 Томпсон У. (Thompson W.) 195
 Торнтон (Thornton H. P.) 223
 Тьюки (Tukey J. F.) 240, 243, 262, 297, 303, 305
- Уайт (White C.) 124, 125, 126
 Уиб (Wiebe G. A.) 79, 191
 Уилк (Wilk M. B.) 352
 Уилкоксон (Wilcoxon F.) 123, 126, 127
 Уилкс (Wilks S. S.) 43
 Уилльер (Willier J. G.) 403
 Уилси (Wilsie C. P.) 343
 Уинзор (Winsor C. P.) 302, 303
 Уишарт (Wishart J.) 14, 293
 Уокер (Walker H. M.) 223
 Уоллес Д. (Wallace D. L.) 243

Уоллес Г. (Wallace H. A.) 405, 407
Уоллис (Wallis W. A.) 209, 211
Уоррен (Warren E.) 155
Уэйкли (Wakeley R. E.) 25
Уэлш (Welch B. L.) 273
Уэнц (Wentz J. B.) 132
Уэст (West Q. M.) 467

Финни (Finney D. J.) 437
Фишер Ч. (Fisher C. H.) 180
Фишер Р. (Fisher R. A.) 5, 8, 9, 14, 16,
37, 152, 174, 193, 194, 207, 211, 216,
218, 227, 234, 236, 270, 286, 289, 296,
409, 429
Флинт Л. (Flint L. H.) 91
Флинт У. (Flint W. P.) 148
Форстер (Forster H. C.) 284
Фриленд (Frieland W. C.) 230
Фриман (Freeman F. N.) 269

Хансберри (Hansberry T. R.) 137, 139,
214, 245
Хансен (Hansen M. H.) 455
Хартли (Hartley H. O.) 242, 451
Хаскелл (Haskell A. C.) 417
Хейбер (Haber E. S.) 183, 349, 350
Хейсел (Hasel A. A.) 484
Хеммонд (Hammond W. E.) 99
Хоблин (Hoblyn T. N.) 317

Хойт (Hoyt E. E.) 113
Холберт (Holbert J. R.) 148
Холзингер (Holzinger K. J.) 269
Холл А. (Hall A. D.) 79
Холл Ф. (Holl Phoebe) 359
Холланд (Holland R. H.) 437
Хомейер (Homeyer P. G.) 324
Хорвиц (Horvitz D. G.) 267

Чепмен (Chapman A. B.) 119

Шайв (Shive J. W.) 112
Шеппард (Sheppard W. F.) 193, 197
Шотт (Schott R. G.) 50
Шульц (Schultz E. F.) 337

Эдуардс (Edwards T. I.) 191
Эзекиль (Ezekiel M.) 405
Эйд (Eid M. T.) 387
Эйрс (Ayres J. C.) 63
Экас (Ekas M. P.) 195
Эммерт (Emmert E. M.) 428
Энглдау (Engledow F. L.) 227
Эренкранц (Ehrenkrantz F.) 121

Юден (Youden W. J.) 64
Юл (Yule G. U.) 37, 185, 227

Янг (Young C. M.) 109

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Абсолютное значение (Absolute value) 58
 Асимметрия (Asymmetry) см. Распределение
 Аэрофотоъемка (Aerial photographs) 483
 Бимодальность (Bimodal) 119
 Биномальное распределение (Binomial distribution) 22, 34, 51, 205
 Варьирование (Variation или variability) 20, 58, 67
 Вероятность (Probability) 27, 219, 440, 458
 — объединенная (combination of) 209
 Вес (Weight) 36, 473
 Весовые коэффициенты Гаусса (Gauss multipliers) 409, 416
 Взаимодействие (Interaction) 314
 — второго порядка, или трехфакторное (second order or threefactor) 334
 — неоднородность (heterogeneity) 207
 Взвешенная средняя (Weighted mean) 36
 — сумма (Weighted sum) 425
 — квадратов (Weighted sum of square) 356
 Взвешенный средний квадрат (Weighted mean square) 464
 Виды средних (average) см. Средняя, Мода, Медиана и пр.
 Выборка (Sample) см. также групповые сравнения 20, 454
 — большая (large) 189
 — взвешенная сама по себе (Self-weighted sample) 469
 — репрезентативная (representative) 21, 255
 — случайная (random) 26, 32, 51, 80, 96
 — — из биномиального распределения (from binomial distribution) 21, 439
 — — из непараметрической совокупности (from non-parametric population) 118
 — — из нормальной совокупности (from normal population) 80
 — — из распределения Пуассона (from Poisson population) 224, 448
 — — средних (of means) 82
 Выборочное значение (Sample или estimated)
 — — коэффициента корреляции (correlation coefficient) 166
 — — регрессии (regression coefficient) 128, 390
 — — средней (mean) 53
 Выборочное значение средней послылой выборки (of stratified sample) 468
 — — — — — качественных признаков (of attributes) 473
 — — — — — стандартное отклонение (standart deviation) 58
 — — — — — относительное (relative) 59, 76
 — — — — — от регрессии (from regression) 142, 411
 — — — — — предсказанного Y (of predicted Y) 142, 411
 — — — — — разности (of difference) 74, 395
 — — — — — значение стандартной ошибки (standart error)
 — — — — — \hat{Y} (of \hat{Y}) 141, 411
 — — — — — коэффициента регрессии (of regression coefficient) 129, 390
 — — — — — оценки отношения (of ratio estimate) 482
 — — — — — средней (of mean) 60, 85, 140, 461, 462, 470, 476
 — — — — — приведенной (adjusted) 472, 474, 395, 396
 — — — — — для качественных признаков (of attributes) 462
 — — — — — — — — в гнездовой выборке (in cluster sampling) 463, 464
 — — — — — суммы (of total) 462, 483
 Выборочное наблюдение (Sampling) 20
 — — бесповторное (without replacement) 455
 — — в биологии (in biology) 483
 — — в растениеводстве и животноводстве (crop and livestock estimates) 483
 — — вероятностное (probability) 458
 — — гнездовое (cluster) 460
 — — диких животных (wildlife) 484
 — — качественных признаков (attributes) 26, 205, 439
 — — план (plan) 455
 — — площадей (area) 460
 — — по двум и трем стадиям (in two and three stages) 475, 480
 — — послыльное (stratified) 26, 458, 468
 — — — при качественных признаках (for attributes) 473
 — — — при неодинаковых долях выборки (with unequal fractions) 458
 — — простое (simple) 455, 461
 — — систематическое (systematic) 26, 457
 — — случайное (random) 26, 52, 80, 96, 98, 130, 133, 136, 227, 231, 246, 248, 275, 369, 386, 455, 458

- Выборочное обследование лесов (Forestry surveys) 483
 — распределение (Sampling distribution) 39, 82
 Выпавшее наблюдение или деланка (Missing date or plot) 292
 Вычислительная машина (Calculating machine) 69, 144
- Гипербола (Hyperbola) 142**
Гипотеза (Hypothesis) 37
 — альтернативная (alternative) 43, 72
 — нулевая (null) 42, 43, 65, 71, 72, 73, 140
 — однородности (of homogeneity)
 — — в таблице сопряжения (in contingency table) 209
 — — дисперсии (of variance) 98, 271, 371
 — проверка (test of) 37, 42, 60, 65, 172, 174, 454
Гипотетическая численность (Hypothetical number) 37, 441
Гистограмма (Histogram) 35
Главная деланка (Main plot) 343
Графическое изображение (Graphical representation) 34, 54, 434
Групповые сравнения (Group comparisons) 96, 227
 — — в ковариации (in covariance) 368
 — — встречаемые трудности (difficulties met) 100
 — — неравных групп (unequal groups) 100, 255
 — — равных групп (equal groups) 97, 231
- Дисперсионный анализ (Analysis of variance) 152, 179, 227, 470**
 — — групповых сравнений (group comparisons) 96, 227
 — — Модель I, фиксированные эффекты вариантов (Model I, fixed treatment effects) 247
 — — Модель II, случайные эффекты (Model II, random effects) 248
 — — Модель смешанная (mixed model) 254
 — — отдельных сравнений (individual comparisons) 79, 275
 — — при регрессии (in regression) 152, 280, 325, 368, 390
Дисперсия σ^2 (Variance), см. также Средний квадрат, Дисперсионный анализ 58, 108
 — анализ (analysis of) 152, 227, 477
 — интервальная оценка (interval estimate of) 73
 — компоненты (components of) 249, 477
 — неоднородная (heterogeneous) 108, 272
 — однородность (homogeneity of) 107, 271
 — отношения (ratio) 234
Доверительный (Confidence)
 — интервал или пределы (interval or limits) см. Оценка 21, 22, 35, 62, 87, 142, 174, 443, 454
 — — таблица пределов (table of) 22—23
 — — пояса (belt) 143
Дробный учет (Uniformity trial) 227
- Единица наблюдения (Sampling unit) 459, 476, 479**
- Заключения относительно совокупности (Inferences about population) 20, 21, 26, 37, 46, 65, 139, 231, 454**
Закон сложного процента (Law of compound interest) 417
Звездочка (Asterisk)
 — * существенно: $0,05 \geq P \geq 0,01$ (significant) 132
 — ** более существенно: $P < 0,01$ (highly significant) 132
- Интервал (Interval)**
 — доверительный (confidence) см. Оценка
 — классовой (class) 40
 — между квантилями (inter-quartile) 119
Интерполяция (Interpolation) 135
Исключение лишней переменной (Deletion of a variable) 415
- Качественный признак (Attribute) 26**
 — — выборочное наблюдение (sampling of) 26, 205, 439, 462, 473
 — — выраженный количественно (presence recorded quantitatively) 223
Квантили (Quartiles, Q_1 , Q_2 and Q_3) 118
Класс (Class) 28
 — середина (class mark) 80
 — Классовая численность (Class frequency) 28
Классификация (Classification)
 — по двум признакам (two criteria of) 211, 216, 275, 279
 — по трем признакам (three criteria of) 221
 — последовательная (hierarchical or nested) 252
Классовый интервал (Class interval) 40
Ковариация (Covariance) 168, 368, 427
 — в латинском квадрате (in Latin square) 383
 — в полностью рандомизированном опыте (in completely randomized experiment) 368
 — в рандомизированных блоках (in randomized blocks) 383
 — в таблице с двумя входами (in 2-way table) 398
 — критерий линейности (test of linearity) 427
 — множественная (multiple) 393
Кодирование (Coding) 112, 146, 166, 410
 — правила и предосторожности (rules and precautions) 113
 — при расчете коэффициентов корреляции (using correlation coefficients) 410
Компонент (Component) 248
 — анализ (analysis) 248
 — дисперсии (of variance) 248, 474
Конечное значение (Terminal value) 432
Константы, подбор (Constants, fitting) 363
Корреляция (Correlation)
 — внутриклассовая (intra-class) 161
 — выборочный коэффициент r (sample coefficient) 161, 414
 — и разности (and differences) 168, 180
 — и регрессия (and regression) 167, 180
 — коэффициент ρ (coefficient) 166
 — — — оценка r (estimate of) 161
 — ложная (spurious) 184
 — междуклассовая ρ_I (interclass) 268
 — — — оценка r_I (estimate of) 269
Корреляция между органами животного (organic) 163

- Корреляция между средней и дисперсией (between mean and variance) 297
- множественная (multiple) 393, 415
 - не имеющая смысла (nonsense) 185
 - общая или нулевого порядка (total or zero order) 161
 - отношений (of ratios) 183
 - по методу моментов (product-moment) 161, 187
 - ранговая (rank) 186
 - сумм (of sums) 183
 - таблица (table of) 173
 - части с целым (part-whole) 185
 - частная (partial) 401
- Коэффициент (Coefficient)
- биномиальный (binomial) 441
 - вариации (of variation) 59, 76
 - выборочный (sample)
 - корреляции r (of correlation) 161
 - — регрессии b (of regression) 129
 - — — стандартной частной $b_{Y1.2}$ (standard partial) 398
 - — — частной $b_{Y1.2}$ (partial) 386
 - для сравнений (for comparisons) 311
 - корреляции ρ (of correlation) 166
 - — множественной R (multiple) 393, 415
 - регрессии β (of regression) 133
- Критерий (Test)
- знаков (sign test) 122
 - независимости (independence) 205
 - нормальности (of normality) 195
 - симметрии (of symmetry) 194
 - слагаемости (of additivity) 303
 - согласия (goodness of fit) 205
 - существенности (of significance) 42, 63, 65, 174, 454
 - экспериментальной техники (of experimental technique) 222
- Латинский квадрат (Latin square) 287, 383
- Линейная модель (Linear model) 247, 280, 283, 287
- Линейность регрессии (Linearity of regression) 392
- Логарифм (Logarithm)
- Непера, или натуральный (Naperian or normal) 418
 - обычный (common) 174, 271, 302, 418
- Логарифмическая шкала (Logarithmic scale) 419
- Логарифмическое преобразование (Logarithmic transformation) 302
- Логистический анализ (Logistic analysis) 437
- Масштаб (Scale) 52
- Матрица (Matrix)
- асимметричная (asymmetrical) 410
 - коэффициентов корреляции (of correlation coefficients) 415
 - обратная (inverse) 409
- Медиана (Median) 118
- доверительный интервал (confidence interval on) 120
- Междуквартильный интервал (Inter-quartile interval) 119
- Метод невзвешенных средних (Method unweighted means) 360
- Метод выборок (Sub-sampling) 475
- Многолетняя культура (Perennial crop) 349
- Множественная ковариация (Covariance)
- — в группах (in groups) 393
- Множественная ковариация в таблице с двумя входами (2-way table) 398
- регрессия (regression) 386
- Множественный коэффициент корреляции (Multiple correlation coefficient) 394, 408
- Мода (Mode) 119
- Модель (Model)
- для ковариации (for covariance) 371, 383
 - для латинского квадрата (for Latin square) 287
 - для полностью случайного опыта (for completely random experiment) 246, 248
 - для регрессии (for regression) 133, 137, 155, 388
 - для рандомизированных блоков (for randomized blocks) 279
 - для факториальной схемы опыта (for factorial experiment) 313, 318, 336, 345
- Наблюдения методами вылавливания (Trapping methods) 482
- Наименьшая существенная разность (Least significant difference) 240
- Начало (Origin)
- отклонений (of deviations) 131
 - произвольное (arbitrary) 68
- Независимая переменная X (Independent variable) 129
- Непараметрический (Non-parametric) 111, 118, 122, 124, 127, 186
- Непрерывная переменная (Continuous variable) 51
- Несмещенный (Unbiased) 84
- Нормальное распределение (Normal distribution) см. Распределение
- Нормальные уравнения (Normal equations) 387, 403
- Нормированное отклонение (Standard deviate) 168
- Нулевая гипотеза (Null hypothesis) 71, 88
- Однородность (Homogeneity) 205, 445, 449
- Округление (Rounding) 113
- Опробование биологическое (Assay, biological) 437
- Опыт (Experiment)
- в латинском квадрате (Latin square) 278, 383
 - в рандомизированных блоках (randomized blocks) 275, 284, 380
 - и статистика (and statistics) 109
 - полностью рандомизированный (completely randomized) 96, 227, 314, 368
 - серия (series of) 348
 - с расщепленными делянками (split-plot) 343
 - схема (design) см. Схема эксперимента
 - условный (dummy) 227
 - факториальный (factorial) 313
- Ортогональность (Orthogonality) 278, 379
- Ортогональные (Orthogonal)
- полиномы (polynomials) 429
 - сравнения (comparisons) 311
- Отклонение (Deviation)
- относительное (relative) 59, 76
 - от регрессии (from regression) 131, 390
 - от средней (from mean) 56
 - среднее абсолютное (mean absolute) 58
 - стандартное (standard) см. Стандартное отклонение

Относительная (Relative)
 — скорость роста (growth rate) 419
 — частота (численность) (frequency) 37, 440, 448
 Относительное стандартное отклонение (Relative standart deviation) 76
 Относительные коэффициенты (Rates) 49, 157
 Относительный размер (Relative size) 38
 Отношение полов (Sex ratio) 49
 Отношения и проценты (Rates and percentages) 49
 Отношения и регрессия (Rates and regression) 154, 158, 482
 Оценка (Estimate or estimator)
 — значения (point) 21, 53
 — — α 135
 — — β 135
 — — μ 135
 — — ρ 164
 — — σ 53, 58, 136, 228
 — интервала (interval) 21, 24, 60, 62, 73
 — — для α 141
 — — — β 140
 — — — μ 62, 87, 142
 — — — p 22, 35
 — — — ρ 174
 — — — σ 73
 — — — Y 142
 — — — \hat{Y} 142
 — несмещенная (unbiased) 58, 456
 Ошибка (Error)
 — вида (kind)
 — — первого, Тип I (first, Type I) 43
 — — второго, Тип II (second, Type II) 43
 — выборочного наблюдения (sampling) 323
 — допустимая L (allowable) 465
 — независимой переменной (in independent variable) 389
 — опыта (experimental) 66, 92, 137, 276, 284, 417
 — — снижение (reduction of) 92
 — — оценка (estimate of) 66, 97, 100, 129, 132, 284
 — — оценки (of estimate) 456
 — регрессии (regression) 374, 378
 — стандартная (standart) 60, 85

Парабола (Parabola) 421, 434
 Параметр (Parameter) 52
 Парность (Pairing, pair) 63, 66, 103, 180
 Паскаля треугольник (Pascal's triangle) 441
 Переменная (variable, variate)
 — зависимая Y (dependent) 129
 — независимая X (independent) 129
 — — исключение (deletion of) 415
 — непрерывная (continuous) 42, 52
 — прерывная, или дискретная (discontinuous or discrete) 42, 205, 439, 448
 Перечень (List, listing) 459
 Перфорационные карточки (Punched cards) 203
 План участка (Field plan) 284
 Плоскость регрессии (Regression plane) 386
 Повторение (Replication) 64
 Подбор (Fitting)
 — констант (constants) 363
 — линейной модели (linear model) 283, 305
 Подбор кривой (Curve fitting) 417, 436, 448

Подбор кривой глазомерный (Eye-fitted curve) 436
 Подсчет численностей (Enumeration date) 26, 205, 439
 Полином второго порядка (Polynomial second degree) 421
 Полиномиальное уравнение (Polynomial equation) 435
 Полиномы ортогональные (Polynomial orthogonal) 429
 Полулогарифмическая бумага (Semilogarithmic paper) 419
 Поправка (Correction)
 — для суммы квадратов (for sum of square) 68, 133
 — на группировку (for grouping) 193
 — на непрерывность (for continuity) 210
 — на ограниченность совокупности (finite population) 462
 — на среднюю (for mean) 68
 — Шеппарда (Sheppard's) 193
 Последовательная классификация (Hierarchical classification) 252
 Построение опыта (Design of experiment) 21, 64, 92, 100, 102, 129, 148, 205, 207, 216, 222, 246, 248, 252, 284, 287, 314, 368, 462
 Предугадывание (Prediction) 129, 142
 Преобразование (Transformation) 112, 296
 — в арксинус (arcsin) 298
 — в квадратный корень (square root) 297
 — в логарифм (logarithmic) 302
 Приведенное значение (Adjusted value)
 — — для Y (of Y) 136
 — — средней (of mean) 372, 381, 395
 — — — годной для применения (validity of use) 380
 — — хи-квадрат (chi-square) 210
 Принцип наименьших квадратов (Principle of least squares) 59
 Пробит-анализ (Probit analysis) 437
 Проверка гипотезы (Test of hypothesis)
 См. Гипотеза, проверка
 Промилль (Per mille) 119
 Процентирование (Percentages) 49
 — и хи-квадрат (and chi-square) 47
 Пуассона распределение (Poisson distribution) 205, 297, 448

Размах варьирования (Range) 53
 — — выводы, основанные на (inferences based on) 116
 — — средних (of means) 83
 — — таблица отношения σ / размах (table of σ / range) 54
 Размер опытов или выборки (Size of experiments or sample) 26, 47, 74, 104, 262, 291, 465
 — выборки в отдельных слоях (in individual strata) 471
 Разнородность (Heterogeneity)
 — дисперсий (of variance) 96, 272, 297, 298
 — хи-квадрат (chi-square) 208
 Разности (Differences) 63, 88
 — и корреляция (and correlation) 180
 — между средними (between means) 97
 — среди средних (among means) 240, 278
 Ранги (Ranks) 121
 — сумма (sum of) 123
 Ранговая корреляция (Rank correlation) 186
 Ранжированный ряд (Array) 54
 — — средних (of means) 240

- Распределение, или размещение (Allocation)
- оптимальное (optimum) 473
 - пропорциональное (proportional) 469, 473
 - средств (of resources) 249, 478
- Распределение (Distribution)
- асимметричное (asymmetrical) 41, 84, 118, 194
 - биномиальное (binomial) 23, 34, 51, 205, 298, 439
 - выборочной средней (of sample mean) 82
 - географическое (geographical) 79
 - двухстороннее (two-tail) 62
 - мультиномиальное (multinomial) 27
 - нормальное (normal) 51, 447
 - -- график (graph of) 197
 - -- таблица (table of) 196
 - -- уравнение (equation of) 197
 - нормальное двух переменных (bivariate normal) 203
 - одностороннее (one-tail) 71
 - Пуассона (Poisson) 224, 297, 448
 - среднего квадрата (of mean square) 84
 - студентизированного размаха Q (of Studentized range) 241
 - хи-квадрат (of chi-square) 40, 41, 46
 - частот (frequency) 28, 41, 79
 - -- прерывных (discontinuous) 41
 - F 106, 235
 - r 171
 - s и s^2 84
 - σ /размах 54
 - T 125
 - t 60, 61, 86
 - t_w и t_w^- 117
 - z 175
- Расслоение (Stratification) 26, 457, 468
- Расщепленные делянки (Split plots) 343
- Регрессия (Regression) 129, 155, 325
- в каждой группе (in each group) 370
 - в опытном деле (in experimentation) 94, 129, 148
 - в сравнениях (in comparisons) 325
 - второго порядка (second degree) 421
 - выборочный коэффициент (sample coefficient) 131, 390
 - -- стандартный частный (standard partial) 389
 - -- частный (partial) 386
 - групповых средних (of group means) 382
 - и дисперсионный анализ (and analysis of variance) 152, 289, 328, 368, 390
 - и корреляция (and correlation) 180
 - и отношения (and rates) 155, 158, 481
 - коэффициент (coefficient of) 132, 386
 - криволинейная (curvilinear) 328, 417
 - линейная (linear) 132, 386
 - множественная (multiple) 388
 - Модель I, фиксированное X (Model I, fixed X) 132, 388
 - Модель IA (Model IA) 155
 - Модель II двух переменных (Model II, bivariate) 137, 163, 386
 - отклонения от (deviations from) 133, 137, 390
 - оценка (estimate) 131, 388, 418
 - параболическая (parabolic) 327
 - плоскость (plane) 386
 - -- высота и направление (elevation and orientation) 388
 - поверхность (surface) 386
- Регрессия сниженная (reduced) 415
- уравнение (equation) 131
 - -- криволинейное (curvilinear) 421
 - -- линейное (linear) 386
 - -- множественное (multiple) 388
 - -- полиномиальное (polynomial)
 - -- второго порядка (second degree) 421
 - -- ортогональное (orthogonal) 435
 - усредненная внутри группы (average within groups) 371
- Результат, выход (Outcome) 67, 102, 275
- Рандомизированные блоки (Randomized blocks) 284, 379
- Репрезентативная выборка (Representative sample) 21, 25
- Рост (Growth)
- относительный (relative, rate) 419
 - эмпирическая кривая (experimental curve) 417
- Сбалансированные группы (Balanced lots) 382
- Сглаживание (Smoothing) 417
- Скошенность (Skewness) 84, 118, 194
- Случайные (Random) 26
- выборки (samples) см. Выборка
 - разности (difference) 64, 88, 183
 - числа (numbers) 28
- Смешивание (Confounding) 100
- Смещение (Bias) 58, 456
- поправка на (correction for) 293
- Совокупность (Population) 20, 26, 454
- конечная (finite) 80, 462
 - нескольких переменных (multivariate) 386
 - нормальная (normal) см. Распределение нормальное 51
 - нормальная двух переменных (bivariate normal) 137, 161, 163, 203
 - общая, оценка (total, estimate of) 462
- Сокращенные методы (Short-cuts) 111, 127, 144, 242
- Сопряженность, таблица (Contingency table)
- 2×2 , или четырехклеточная (fourfold) 241
 - -- $R \times C$ 216
 - -- $R \times 2$ 48
 - -- с тремя входами (three-way) 224
- Спрямление (Rectification) 418
- Сравнения (Comparisons) 310
- все, между средними (all, among means) 240, 278
 - группы (group) 96
 - -- неодинакового размера (unequal size) 100
 - -- одинакового размера (equal size) 97
 - отдельные (individual) 63
 - -- в дисперсионном анализе (in analysis of variance) 277
 - -- и групповые (versus group) 103
 - ортогональные ряды (orthogonal sets of) 278, 311
 - предназначенные, планируемые (designed) 243, 278, 310, 328
- Средний квадрат s^2 (Mean square) 58
- -- взвешенный (weighted) 464
 - -- отклонений от регрессии (of deviations from regression) 131
 - -- средней s_x^2 (of mean) 85, 478
- Среднее отклонение от регрессии (deviation from regression) 131

- Средняя (Mean)
- абсолютное отклонение (absolute deviation) 58
 - арифметическая (arithmetic) 52
 - взвешенная (weighted) 469
 - гармоническая (harmonic) 360
 - геометрическая (geometric) 303
 - доверительный интервал (confidence interval on) 62
 - невзвешенная, способ (unweighted, method of) 360
 - поправка на (correction for) 68
 - приведенная (adjusted) 374
 - разность (difference) 63
 - совокупности μ (of population) 52
 - — оценка интервала (interval estimate of) 62
 - сокращенный способ вычисления (short method of calculation) 68, 69, 112, 190
- Стандартная ошибка (Standard error)
- — средней (of mean) 60, 85, 441, 462
 - — суммы совокупности (of population total) 462
- Стандартное отклонение σ (Standard deviation) 52, 58, 462
- — средней (of mean) 52, 85
 - — σ /размах (σ /range) 54
 - — z 174
- Стандартный коэффициент частной регрессии (Standard partial regression coefficient) 389
- Статистический контроль (Statistical control) 148
- Степени свободы d. f. (Degrees of freedom) 44, 58, 131, 235
- — в дисперсионном анализе (in analysis of variance) 229
 - — в корреляции (in correlation) 172
 - — — частной (partial) 401
 - — в регрессии (in regression) 132, 390
 - — хи-квадрат (for chi-square) 45, 206, 216, 217, 444, 451
- Стоимость, затраты средств (Cost) 480
- Субвыборка (Sub-sample) 252
- Субделянка (Sub-plot) 343
- Сумма (Sum)
- взвешенная (weighted) 425
 - квадратов (of squares) 68
 - — в регрессии (in regression) 149, 391
 - — для оценки существенности приведенных средних (for testing significance of adjusted means) 373, 377, 379
 - — и корреляция (and correlation) 180
 - — отклонений от регрессии (of deviations from regression) 132
 - — поправка на (correction for) 68
 - — средних (of means) 228, 230
 - — произведений (of products) 129
 - — поправка на (correction for) 145
- Существенность (Significance of)
- коэффициента корреляции (correlation coefficient) 172
 - коэффициента регрессии (regression coefficient) 381
 - множественной регрессии (multiple regression) 390
 - отклонения от регрессии (deviation from regression) 132, 390
 - разности между двумя (difference between two)
 - — — коэффициентами регрессии (regression coefficient) 371
- Существенность разности между двумя отношениями (ratios) 212
- — — — приведенными средними (adjusted means) 372, 395
 - — — — средними (means) 96
 - — — — частными коэффициентами регрессии (partial regression coefficients) 411
 - — разностей среди (differences among)
 - — — коэффициентов регрессии (regression coefficients) 374
 - — — приведенных средних (adjusted means) 374
 - — — средних строя или в дисперсионном анализе (means in an array or in analysis of variance) 234, 240
 - — хи-квадрат (chi-square) 42
 - — частного коэффициента корреляции (partial correlation coefficient) 172
 - — — регрессии (partial regression coefficient) 411 — F 234
- Теорема сложения (Addition theorem)
- — суммы квадратов и степеней свободы (sum of squares and degrees of freedom) 150, 228
 - — хи-квадрат (chi-square) 206
- Точность (Precision) 276
- Треугольник Паскаля (Triangle Pascal's) 441
- Уровень (Level) 314
- Факториальная схема опыта (Factorial experiment) 314
- Хи-квадрат (Chi-square) 39, 73, 122, 443, 451
- выборочное наблюдение (sampling of) 39
 - и процентирование (and percentages) 47
 - и размер выборки (and sample size) 47
 - малые гипотетические численности (small hypothetical numbers) 215
 - особая формула (special formulas) 43, 45, 48, 216
 - показатель рассеяния (index of dispersion) 37
 - приведенное значение (adjusted value of) 210
 - разнородность (heterogeneity) 207
 - распределение (distribution of) 39, 46
 - существенность (significance of) 42
 - таблица (table of) 46
 - теорема сложения (addition theorem) 207
- Центральная предельная теорема (Central limits theorem) 83
- Цифры (Digits)
- значащие (significant) 114
 - случайные (random) 28
 - — таблица (table of) 28—32
- Частная (Partial)
- корреляция (correlation) 401
 - регрессия (regression) 386
- Частота (численность) (Frequency)
- гипотетическая (hypothetical) 38, 440
 - класса (class) 28
 - наблюдаемая (observed) 38, 440
 - накопленная (cumulative) 87

Частота относительная (relative) 37, 440, 448	Экстраполяция (Extrapolation) 135, 423
Числа, численности (Numbers)	Элементы общие (Elements, common) 169, 180, 185
— гипотетические (hypothetical) 38, 441	Эллипс (Ellipse) 162
— случайные (random) 28	Эффективность (Efficiency of)
— таблица (table of) 28—32	— ковариации (covariance) 396
Численности субклассов (Sub-class numbers)	— латинского квадрата (Latin square) 287
— непропорциональные (disproportionate) 355, 358, 360	— размаха варьирования (range) 58, 116
— одинаковые (equal) 96, 227, 313	— рангового критерия (rank test) 127
— пропорциональные (proportionate) 351, 361	— рандомизированных блоков (randomized blocks) 285
Четырехклеточная таблица (Fourfold table) 211	Эффекты (Effects)
	— отдельные (separate) 315
	— общие, главные (main) 280

КРАТКИЙ КУРС ЭЛЕМЕНТАРНЫХ СТАТИСТИЧЕСКИХ МЕТОДОВ

	Стр.
1. Качественные признаки	20—47
2. Количественные признаки	51—70
3. Распределение выборочных наблюдений	79—95
4. Групповые сравнения	96—104
5. Сокращенные методы	111—116
6. Регрессия	129—145
7. Корреляция	161—173
10. Дисперсионный анализ по группам	227—243
11. Дисперсионный анализ при двух признаках	275—296

СОДЕРЖАНИЕ

Предисловие к русскому изданию	5
Предисловие автора	18
Глава 1. Выборочное наблюдение качественных признаков. Биномиальное распределение	20
1. Выборочное наблюдение	20
2. Две основные задачи выборочного наблюдения	20
3. Выборка данных о фермах. Оценки значения и интервала	21
4. Определения и обсуждение некоторых вопросов. Размер выборки	25
5. Упражнение по проведению выборки	27
6. Распределение частот и его графическое изображение	33
7. Доверительный интервал; проверка теории	35
8. Гипотезы относительно совокупности	37
9. Хи-квадрат — показатель рассеяния	37
10. Формула хи-квадрат	38
11. Экспериментальное изучение выборочного распределения хи-квадрат	39
12. Распределение частот при неравных классовых интервалах	41
13. Проверка нулевой гипотезы или критерий существенности	42
14. Случай, когда число признаков в совокупности больше двух	44
15. Хи-квадрат и размер выборки	47
16. Другая формула хи-квадрат	48
17. Отношения, относительные коэффициенты и процентные отношения	49
Литература	50
Глава 2. Выборочные наблюдения из нормальной совокупности	51
1. Нормальная совокупность	51
2. Оценка μ и σ	52
3. Ранжированный ряд и его графическое изображение	54
4. Символические обозначения	55
5. Отклонения от выборочной средней	56
6. Другая возможность оценки σ . Стандартное отклонение выборки	58
7. Распределение t Стьюдента	60
8. Оценка интервала для μ . Доверительный интервал	62
9. Оценки и критерии существенности для разностей	63
10. Основания и условия парного сравнения	66
11. Вычисление s без счетной машины	68
12. Вычисление s при помощи счетной машины	69

13. Критерии для других видов нулевой гипотезы о значении μ	71
14. Оценка интервала и критерии для σ^2	73
15. Размер выборки	74
16. Относительная изменчивость. Коэффициент вариации	76
Литература	77
 <i>Глава 3. Выборочные наблюдения из нормальной совокупности. Построение выборочных распределений</i>	
1. Введение	79
2. Конечная совокупность, имитирующая нормальную совокупность	80
3. Случайные выборки из нормальной совокупности	80
4. Распределение выборочной средней	82
5. Выборочные распределения s^2 и s	84
6. Оценка стандартной ошибки σ/\sqrt{n}	85
7. Распределение t	86
8. Интервальная оценка μ . Доверительный интервал	87
9. Разности. Проверка нулевых гипотез	88
10. Снижение ошибки выборочного наблюдения	92
Литература	95
 <i>Глава 4. Сравнение двух случайных групп</i>	
1. Введение	96
2. Построение опыта с двумя случайными группами или партиями	96
3. Опыт, в котором производится сравнение двух групп одинакового размера	97
4. Затруднения, встречающиеся в опытах с групповым сравнением	100
5. Группы с разным количеством объектов	100
6. Противоречие между индивидуальными и групповыми сравнениями	102
7. Размер выборки	104
8. Критерий однородности дисперсий	105
9. Способ оценки при $\sigma_1^2 \neq \sigma_2^2$	106
10. Статистика и эксперимент	109
Литература	110
 <i>Глава 5. Сокращенные и приближенные способы. Методы неполной эффективности и непараметрические</i>	
1. Введение	111
2. Линейное преобразование, или кодирование	112
3. Округление и кодирование	113
4. Правила и предосторожности в отношении использования чисел кода	113
5. Значащие цифры	114
6. Выводы, основанные на размахе варьирования в выборках из нормальных совокупностей	116
7. Непараметрические методы: медиана и квартили	118
8. Непараметрические методы: ранжирование двух вариантов	121
9. Непараметрические методы: ранжирование разностей между количественными показателями	123
10. Непараметрические методы: ранжирование непарных количественных показателей	124
11. Непараметрические методы в качестве сокращенных способов обработки выборок из нормальной совокупности	127

Литература	128
Глава 6. Линейная регрессия	129
1. Введение	129
2. Регрессия кровяного давления человека на возраст	129
3. Модель I совокупностей, из которых берутся выборки: фиксированные значения X	133
4. Оценки значений α и β	134
5. \hat{Y} в качестве оценки $\mu = \alpha + \beta x$. Приведенные значения \hat{Y}	134
6. Оценка σ^2	136
7. Модель II совокупности, из которой берется выборка; нормальное распределение двух переменных	137
8. Оценки интервалов и проверка нулевых гипотез	140
9. Сокращенные методы вычисления регрессии	144
10. Использование регрессии при планировании опытов. Статистический контроль	148
11. Разложение суммы квадратов зависимой переменной	150
12. Применение термина «регрессия» Гальтоном	154
13. Регрессия модель IA: σ пропорциональна X и $\alpha = 0$	155
14. О распространенных злоупотреблениях долями и отношениями	158
15. Итоги	159
Литература	160
Глава 7. Корреляция	161
1. Введение	161
2. Выборочный коэффициент корреляции	162
3. Нормальная совокупность двух переменных: модель регрессии II	164
4. Связь между выборочными коэффициентами корреляции и регрессии	167
5. Выборочное варьирование коэффициента корреляции	169
6. Оценка коэффициента корреляции; доверительные утверждения и проверка гипотез	172
7. Корреляция и регрессия	178
8. Корреляция и разложение Σy^2	179
9. Корреляция, общие элементы и регрессия	180
10. Корреляция и разности	180
11. Корреляция между суммами и отношениями. Корреляции, возникающие от общих причин	183
12. Непараметрические методы. Ранговая корреляция	186
Литература	188
Глава 8. Методы обработки больших выборок	189
1. Введение	189
2. Вычисление средней и стандартного отклонения на основе распределения частот	189
3. Вычисление средней и стандартного отклонения на основе распределения частот (продолжение)	191
4. Различные замечания по поводу распределения частот	193
5. Критерий симметричности	194
6. Критерии нормальности при большой выборке	195
7. Построение графика нормального распределения	197
8. Нормальное распределение накопленных частот	199

9. Вычисление по случайной выборке из нормальной совокупности двух переменных	203
Литература	204
Глава 9. Результаты подсчета численностей при числе степеней свободы, превосходящем единицу	205
1. Введение	205
2. Опыт с несколькими выборками	205
3. Опыт с числом классов, по которым распределены объекты, большим двух	207
4. Объединение вероятностей	209
5. Поправка на непрерывность	210
6. Критерий независимости. Четырехклеточная таблица сопряженности признаков	211
7. Четырехклеточная таблица; частоты, основанные на гипотезе о независимости	214
8. Критерий независимости в таблице $R \times C$	216
9. Специальный метод вычисления хи-квадрат в таблице $R \times 2$	218
10. Три ряда градаций качественных признаков, таблица сопряженности признаков с тремя входами	221
11. Эксперимент и проверка техники его проведения	222
12. Опыт, в котором качественный признак выражен количественно	223
Литература	225
Глава 10. Две или большее число выборок количественного признака. Дисперсионный анализ	227
1. Распространение теории двух выборок на большее число их	227
2. Несколько выборок из общей совокупности. Дисперсионный анализ	227
3. Обычный способ вычислений	229
4. Случайные выборки из двух или большего числа совокупностей	231
5. Критерий равенства μ . Отношение дисперсий F	234
6. Оценка сравнений между всеми средними	240
7. Упрощенный метод вычислений при помощи размахов варьирования	243
8. Критерии для заранее намеченных сравнений между средними	243
9. Дисперсионный анализ при двух группах	244
10. Модель I. Фиксированные эффекты вариантов	246
11. Компоненты. Модель I	247
12. Модель II. Случайные эффекты	248
13. Искусственное построение модели II	250
14. Выборки из выборок. Модель II	252
15. Выборки из выборок. Смешанная модель	254
16. Выборки разного размера	255
17. Выборки из выборок при разном их размере	258
18. Размер выборки	262
19. Внутриклассовая корреляция	268
20. Критерии однородности дисперсий	271
Литература	274
Глава 11. Опыты при двойной группировке. Дисперсионный анализ	275
1. Использование предварительных сведений об опытном материале	275
2. Опыт с группировкой по двум признакам	275
3. Сравнения средних	278
4. Символика для таблицы с двумя входами	279

5. Структура опыта при двойной группировке	279
6. Рендомизированные блоки в полевом опыте	284
7. Латинский квадрат	287
8. Размер опыта	291
9. Выпавшие данные	292
10. Несогласованность с моделью. Преобразования	296
11. Преобразование данных подсчета численности путем извлечения квадратного корня	297
12. Преобразование долей через арксинусы	301
13. Логарифмическое преобразование	302
14. Критерий слагаемости Тьюки	303
Литература	309
Глава 12. Сравнения. Факториальные расположения вариантов	310
1. Сравнения	310
2. Правила для проведения сравнений при одинаковой повторности	311
3. Правила для построения сравнений при различной повторности	313
4. Факториальное расположение вариантов. Опыт 2×2 , или 2^2	313
5. Двухфакторный опыт	318
6. Опыт с рендомизированными блоками при учете делянок выборочным способом	322
7. Регрессия в однофакторном опыте	325
8. Регрессия в двухфакторных опытах	328
9. Трехфакторные опыты 2^3	334
10. Трехфакторные опыты $2 \times 3 \times 4$	336
11. Способ расщепленных делянок	343
12. Серия опытов	348
13. Пропорциональные численности субклассов	351
14. Непропорциональные численности субклассов. Таблица 2×2	355
15. Непропорциональность численностей субклассов. Таблица $R \times 2$	358
16. Непропорциональность численностей субклассов в таблице $R \times C$. Приближенные методы	360
17. Подбор констант для двухфакторной таблицы	363
Литература	366
Глава 13. Ковариация	368
1. Введение	368
2. Ковариация в полностью рендомизированном опыте с двумя вариантами	368
3. Ковариация в полностью рендомизированном опыте с числом вариантов, большим двух	373
4. Ковариация в полностью рендомизированных опытах: второй способ	376
5. Ковариация в рендомизированных блоках	377
6. Сравнения при ковариации; факториальные опыты	379
7. Один вопрос, относящийся к построению опытов	382
8. Ковариация в латинском квадрате	383
Литература	385
Глава 14. Множественная регрессия и ковариация	386
1. Введение	386
2. Две независимые переменные в единичной выборке	386
3. Интервальные оценки и критерии существенности	389
4. Множественная ковариация в полностью рендомизированном опыте	393

5. Множественная ковариация в таблице с двумя входами	398
6. Частная корреляция	401
7. Решение нормальных уравнений	403
8. Четыре и большее число переменных в одной группе наблюдений. Вычисления	405
9. Четыре или большее число переменных. Выводы	408
10. Четыре и большее число переменных. Элементы обратной матрицы, весовые коэффициенты Гаусса	409
11. Лишние независимые переменные	415
12. Сводка формул	415
Литература	416
<i>Глава 15. Криволинейная регрессия</i>	<i>417</i>
1. Введение	417
2. Экспоненциальная кривая роста	417
3. Полином второго порядка	421
4. Оценка отклонения от линейной регрессии	424
5. Оценка отклонения от линейной регрессии при ковариационном анализе	427
6. Определение ортогональных полиномов	429
7. Биологическое опробование	437
Литература	437
<i>Глава 16. Биномиальное распределение и распределение Пуассона</i>	<i>439</i>
1. Введение	439
2. Биномиальное распределение	439
3. Сопоставление выборочного распределения с биномиальным	444
4. Критерий однородности при биномиальной форме распределения	445
5. Распределение Пуассона	448
6. Сравнение выборочных средних, полученных из распределений Пуассона	452
Литература	453
<i>Глава 17. Планирование и анализ выборочных наблюдений (Уильям Дж. Кокран)</i>	<i>454</i>
1. Совокупности	454
2. Простой пример	455
3. Вероятностное выборочное наблюдение	458
4. Определение состава совокупности	459
5. Простое случайное выборочное наблюдение	461
6. Размер выборки	465
7. Систематическое выборочное наблюдение	467
8. Послойное выборочное наблюдение	468
9. Подбор размера выборки по отдельным слоям	471
10. Послойное выборочное наблюдение качественных признаков	473
11. Выборочное наблюдение, проводимое по двум стадиям	475
12. Распределение затрат при выборочном наблюдении с двумя стадиями	478
13. Оценки, основанные на отношениях и регрессии	481
14. Выборочные методы в биологии	483
Литература	485
<i>Указатель символов</i>	<i>486</i>
<i>Перечень математических таблиц</i>	<i>487</i>
<i>Указатель авторов</i>	<i>488</i>
<i>Предметный указатель</i>	<i>491</i>
<i>Краткий курс элементарных статистических методов</i>	<i>497</i>

Снедекор Дж. У.
СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРИМЕНЕНИИ К ИССЛЕДОВАНИЯМ
В СЕЛЬСКОМ ХОЗЯЙСТВЕ И БИОЛОГИИ.

Перевод с англ. М., Сельхозиздат, 1961.

503 с.

3332 + 31

Редактор *Е. В. Ластовка.*

Художник *В. В. Еремин.*

Художественный редактор *Н. М. Хохрина.*

Технические редакторы *А. И. Баллод*

и *Л. Н. Прокофьева*

Корректор *Р. Д. Любович*

* * *

Слано в набор 15/IV 1961 г. Подписано к печати 2/X 1961 г. Формат 70×108¹/₁₆.

Печ. л. 31,5 (43,15). Уч.-изд. л. 45,83. Изд.

№ 1691. Заказ № 959. Цена 1 р. 30 к.

* * *

Сельхозиздат, Москва, К-31, ул. Дзержинского, 1/19.

Московская типография № 5 Мосгоссовнархоза.

Москва, К-1, Трехпрудный пер., д. 9.



BIOGRAPHY 17.1 George W. Snedecor (1882 -1974)

George Waddell Snedecor was born in Memphis, Tennessee. He studied mathematics and physics at the Universities of Alabama and Michigan and became a professor of mathematics at Iowa State University. While there, he offered the first statistics course and began a famous collaboration with Henry A. Wallace, then editor of Wallace's Farmer in Des Moines, but later vice-president of the United States. They were interested in agricultural research, organized a seminar to study multiple regression, and did pioneering work in the use of punch cards and punch card machines. Jointly, they published *Correlation and Machine Calculation* in 1925, established a Mathematical Statistical Service in 1927, and the now famous Iowa Statistical Laboratory in 1933. In 1931, Snedecor had invited Ronald A Fisher (Biography 13.1) to Ames and this meeting set in motion developments that produced many of the techniques that are discussed in this chapter. Among the visible signs of this development was the founding, at Iowa State, of the first Department of Statistics in the United States, Snedecor's presidency of the American Statistical Association, and the publication by Snedecor of two famous works: *Calculation and Interpretation of Analysis of Variance and Covariance* (1934) and *Statistical Methods* (1937). The latter work, ultimately co-authored with William G. Cochran, went through seven editions before the death of both authors and sold more than 125,000 copies.

БИОГРАФИЯ 17,1 Джордж Снедекор (1882 -1974) Джордж Уодделл Snedecor родился в Мемфисе, штат Теннесси. Он изучал математику и физику в университетах Алабамы и Мичигане и стал профессором математики в Университете штата Айова. Находясь там, он предложил первый курс статистики и начал знаменитый сотрудничество с Генри А. Уоллес, тогдашний редактор Фермер Уоллеса в Де-Мойн, позже вице-президент Соединенных Штатов. Они были заинтересованы в сельскохозяйственных исследованиях, был организован семинар по изучению множественной регрессии и провели новаторскую работу в области использования перфокарт и перфокартных машин. Совместно они опубликовали корреляции и машины Расчет в 1925 году, установил статистической математической службы в 1927 году, и в настоящее время известный Айова статистическая лаборатория в 1933 году В 1931 году Snedecor пригласил Рональд Фишер (биография 13.1) в Ames и эта встреча в движение события, которые произвели многие из методов, описанных в этой главе. Среди видимых признаков этого развития был основателем, штата Айова, первого отдела статистики в США, президентство Snedecor в

Американской статистической ассоциации, и публикация Snedecor из двух известных работ: Расчет и интерпретация анализа дисперсия и ковариация (1934) и статистические методы (1937). Последняя работа, в конечном счете, в соавторстве с Уильямом Дж Кокран, прошел через семь изданий до смерти обоих авторов и продано более 125 000 экземпляров.

George Waddel SNEDECOR

b. 20 October 1881 - d. 15 February 1974

Summary. George Snedecor is widely known for his text *Statistical Methods*, first published in 1937, now in its eighth edition. He was the founder of the Iowa State University Statistical Laboratory in 1933, the first unit of its kind in the United States.

George W. Snedecor is best remembered as a pioneer in making statistical tools accessible to experimenters in agriculture, biology, and other areas of application. Through his work he laid the foundations for what is today the Department of Statistics at Iowa State University, one of the most prestigious of applied statistics departments.

But perhaps his most remarkable contribution was his book, *Statistical Methods Applied to Experiments in Agriculture and Biology*, published first in 1937, and arguably the most popular and influential statistics text ever written.

1. Origins

George Waddel Snedecor was born in Memphis, Tennessee, the oldest child of James George Snedecor and Emily Alston Estes. George's father was the descendant of Dutch pioneers who had arrived in the United States in the early part of the 17th century. At the time of George's birth, James Snedecor was a practicing lawyer in Memphis. Soon after George was born, the family moved to Florida and later to Alabama, where his father, by then an ordained Presbyterian minister, preached in churches in and around Birmingham.

Most of young Snedecor's school years were spent in Alabama. In 1899, he began two years of college at the Alabama Polytechnic Institute in Auburn, followed by two years of preparatory school teaching. When the family moved to Tuscaloosa in 1903, George transferred to the University of Alabama. In 1905 he graduated with a B.S. in mathematics and physics, with honors. After graduation, George accepted his first academic job, at the Selma Military Academy. As an instructor he taught at the Academy from 1905 through 1907. He then secured a position in Sherman, Texas, teaching mathematics and Greek at Austin College. It was in Sherman where he met and married Gertrude Douglas Crosier in 1908, the daughter of a dormitory matron at Austin College.

In 1910, the Snedecors moved to Ann Arbor, Michigan, while George furthered his education at the University of Michigan. In 1913 he received an A.M. degree with a major in physics.

Finally in 1913, George and Gertrude arrived in Ames, Iowa where they remained until 1958. Both sons, Edward Crosier and James George, were born and raised in Ames. It was there where Snedecor made his mark in the world of statistics.

2. The Early Iowa State College Years

George W. Snedecor joined the faculty at Iowa State College as an Assistant Professor of Mathematics in 1913. He was quickly promoted to Associate Professor, and in 1915 taught a course entitled "Mathematical Theory of Statistics". This was the first statistics course designed to formally introduce statistical methods to researchers in the mathematical and other sciences. The name of the course soon changed to "Statistical Methods of Interpreting Experimental Data", and a second course entitled "Biometric Methods of Interpreting Agricultural Data" was introduced.

By the early 1920s, with several departments across campus offering courses in statistics, researchers in many areas of application were using techniques such as regression analyses on their data. The Graduate College was created in 1920, with the mission of promoting research, particularly in areas of importance to agriculture. Snedecor, with his interest in interdisciplinary research, is likely to have been influential on researchers such as Lindstrom, who by 1925 was using multiple regression models, quite a sophistication for those days (Cox and Homeyer, 1975; Melde, 1990; David, 1998).

In 1927, the Mathematics Statistical Service was established, with Snedecor and A. E. Brandt at the helm. The Service was created to formalize statistical consulting activities, and was the precursor of the Statistical Laboratory, established in 1933. Gertrude Cox (q.v.), who worked part-time in the Mathematics Statistical Service, was the first student to graduate with an M.S. degree in statistics, awarded by the Department of Mathematics in 1931.

By the late 1920s Snedecor was recognized as a preeminent statistician, and in 1931 was promoted to Professor in the Department of Mathematics at Iowa State College. During this time, Snedecor developed an important relationship with R. A. Fisher (q.v.) who was producing path-breaking research at Rothamsted Experimental Station. Their correspondence resulted in two influential six-week visits by Fisher to Iowa State College, in the summers of 1931 and 1936. See also Section 4.

3. The Statistical Laboratory Years, 1933-1947

In recognition of the importance of statistics in other fields of inquiry, Iowa State College established the Statistical Laboratory in 1933, with Snedecor as its Director. For perhaps the first time in the United States, experimental statistics was officially recognized as a science different from mathematics.

Under Snedecor's direction, the Statistical Laboratory became an active center for the research and teaching of statistics. Many visitors were attracted and the permanent staff was substantially increased as a result of a cooperative agreement in 1938 with the U.S. Department of Agriculture. W. G. Cochran was hired at this time and played a key role in strengthening work in survey sampling and in launching a PhD program in statistics (Cox and Homeyer, 1975).

The 1930s were Snedecor's most active research years. He also carried out the administrative duties as the Director of the Statistical Laboratory, in spite of a profound dislike for paperwork. Indeed, it was his reluctance to take on any additional administrative duties that may have led him to resist the creation of a Department of Statistics in the 1930s. As it was, Gertrude Cox went on to establish the Department of Experimental Statistics at North Carolina State College in 1941, the first in the nation. The Iowa State College Department of Statistics was created in 1947.

4. Publications

In 1924 Henry A. Wallace, then editor of *Wallaces' Farmer* and later Secretary of Agriculture and Vice President of the United States, gave a series of ten Saturday afternoon seminars at Iowa State College. He was anxious to pass on to an audience of mainly agricultural and biological research workers the expertise in multiple regression that he had gained primarily in studying the factors influencing corn yields (David, 1998). Wallace summarized these lectures in a bulletin *Correlation and Machine Calculation* that was put in final shape by Snedecor, one of his listeners. This booklet of just 47 pages (Wallace and Snedecor, 1925) reached worldwide circulation and was Snedecor's first publication, at the age of 42.

Soon thereafter Snedecor became aware of R. A. Fisher's fundamental contributions and was among the first in the United States to recognize their great practical importance. This resulted in a major revision of the bulletin (Wallace and Snedecor, 1931), in a small book in which the F-statistic is introduced (Snedecor, 1934, p. 15) and, most important by far, in *Statistical Methods* (Snedecor, 1937).

Statistical Methods was a phenomenal success, and has gone through eight editions. Snedecor was sole author of the first five, the last three being co-authored by W. G. Cochran. The eighth edition was prepared in 1989 by several members of the Iowa State Statistics

Department. Long after the death of the co-authors, the text still had nearly 2000 entries in the Science Citation Index for 1995. Total sales of the editions in English (apart from those published in India) have reached 237,000.

The first edition already makes it plain why this book was to prove so successful. Snedecor had a gift for writing simply and clearly, often using a conversational style. The need for the statistical approach is presented enthusiastically through a series of examples likely to be close to the reader's heart. Snedecor constantly asked questions. There were many tables and charts. The book was studded with helpful and varied exercises. Readers rusty on even elementary mathematics were not ignored. Each topic was introduced unhurriedly and treated in considerable detail, yet a surprising amount of material was covered. Many of the references listed reflected Snedecor's wide experience gained as a consultant to research workers.

In addition, Snedecor filled a real need to satisfy the burgeoning interest of biological research workers in useful statistical methods. The basic texts, Fisher's *Statistical Methods for Research Workers*, first published in 1925, and followed by *The Design of Experiments* in 1935, were impressive and written in a masterfully spare English style. But research workers found them very difficult. It is interesting to note that in the course of a warm correspondence with Fisher, Snedecor writes in 1936 that he is working on an elementary text "designed to lead the beginner to an appreciation of your books." This aim Snedecor accomplished splendidly, even if his presentation was not without occasional shortcomings.

Snedecor's journal articles are mostly in the same spirit of dealing in as simple a way as possible with problems of direct interest to the experimenter. Many of these papers are joint with experimenters or statisticians at Iowa State, on topics including experimental design, analysis of variance and covariance, the chi-square test, disproportionate subclass numbers in multiple classifications, and sampling. Snedecor wrote also on his own on the last two topics, as well as on computing and statistical education and philosophy. The wide knowledge he had gained helping research workers qualified him admirably to be the first editor for *Queries in Biometrics*, from 1945 to 1958. If the flow of queries from readers was slow, he could draw on his vast consulting experience to manufacture a query!

5. The Department of Statistics Years, 1947-1958

In 1947 George Snedecor stepped down as Director of the Statistical Laboratory. He remained as Professor of Statistics until his retirement from Iowa State College in 1958.

By the 1940s, Snedecor was recognized throughout the world as an authority in the development of statistical methods. He was an active member of the American Statistical Association (ASA) and the Biometric Society. He served as Vice President of ASA in 1947, and as President in 1948. In 1950 he was elected member of the International Statistical Institute, the 28th from the United States (Cox and Homeyer, 1975). The Royal Statistical Society made him an Honorary Fellow in 1954.

In 1958, Snedecor retired from Iowa State College after 45 years of service. He was responsible for the establishment and flourishing of perhaps the most influential group of applied statisticians in the United States in the first half of the century and his legacy lasts to this day. The Department of Statistics at Iowa State University is now housed in Snedecor Hall, a fitting tribute to the man.

After retirement in San Diego, California, Snedecor retained a lively interest in Statistics until late in life. He was further honored by the award of the Samuel S. Wilks Memorial Medal in 1970, by election as an Honorary Life Member of the Biometric Society in 1971, by a book of papers in his honor (Bancroft, 1972), and posthumously by the setting up in 1976 of The George W. Snedecor Award for the Best Publication in Biometry, administered by the ASA. He died in 1974, at the age of 92, in Amherst, Massachusetts.

Bibliography

1

Bancroft, T.A. (ed.) (1972). *Statistical Papers in Honor of George W. Snedecor*. The Iowa State University Press, Ames, Iowa.

2

Cox, G.M., and P.G. Homeyer (1975). Professional and personal glimpses of George W. Snedecor. *Biometrics*, 31, 265-301.

3

David, H.A. (1998). Statistics in U.S. universities in 1933 and the establishment of the Statistical Laboratory at Iowa State. *Statistical Science*, 13, 68-74.

4

Snedecor, G.W. (1934). *Calculation and Interpretation of Analysis of Variance and Covariance*. Collegiate Press, Ames, Iowa.

5

Snedecor, G.W. (1937). *Statistical Methods Applied to Experiments in Agriculture and Biology*. Collegiate Press, Ames, Iowa.

6

Wallace, H.A., and G.W. Snedecor (1925). Correlation and machine calculation. *Iowa State College Official Publication*, 23(35).

7

Wallace, H.A., and G.W. Snedecor (1931). Correlation and machine calculation. Revised edition. *Iowa State College Official Publication*, 30(4).

Alicia L. Carriquiry and Herbert A. David

Джордж Waddel SNEDECOR

б. 20 октября 1881 - д. 15 февраля 1974

Аннотация.

Джордж Snedecor широко известен за его текста статистических методов, впервые опубликованный в 1937 году, в настоящее время в его восьмой редакции. Он был основателем Университета штата Айова статистической лаборатории в 1933 году, первый блок, если его вида в Соединенных Штатах. Джордж Снедекор лучше всего помнят как пионер в принятии статистические инструменты доступны для экспериментаторов в области сельского хозяйства, биологии и других областях применения. Благодаря своей работе он заложил основы того, что сегодня Департамент статистики в Университете штата Айова, один из самых престижных прикладной статистики ведомств. Но, возможно, его самый замечательный вклад был его книга, статистические методы прикладной к экспериментам в области сельского хозяйства и биологии, впервые опубликован в 1937 году, и, возможно, текст наиболее популярных и влиятельных Статистика когда-либо написанных.

Происхождение Джордж Waddel Snedecor родился в Мемфисе, штат Теннесси, старший ребенок Джеймса Джорджа Snedecor и Эмили Alston Эстес. Отец Джорджа был потомком голландских пионеров, которые приехали в США в начале 17-го века. Во время рождения Джорджа, Джеймс Snedecor был практикующим юристом в Мемфисе. Вскоре после Георгий родился, семья переехала во Флориду, а затем в Алабаму, где его отец, к тому рукоположенный пресвитерианской церкви, проповедовал в церквях и в окрестностях Бирмингема.

Большинство школьных лет молодой Snedecor провел в Алабаме. В 1899 году он начал два года обучения в колледже в штате Алабама политехнический институт в Auburn, а затем два года подготовки к получению образования в школе. Когда семья переехала в Tuscaloosa в 1903 году, Джордж переданы в университете штата Алабама. В 1905 году он окончил с Б.С. в области математики и физики, диплом с отличием. После окончания школы, Джордж принял его первую научную работу, в Военной академии Сельма. Как преподаватель он преподавал в Академии с 1905 по 1907 г. закрепляют положение в Шерман, штат Техас, преподавания математики и греческого языка в Остин колледжа. Это было в Шерман, где он познакомился и женился на Гертруде Дуглас посох в 1908 году, дочь общезнания матрона в Остине колледжа.

В 1910 году Snedecors переехал в Анн-Арбор, штат Мичиган, в то время как Джордж продолжил свое образование в университете штата Мичиган. В 1913 году он получил утра Степень со специализацией в области физики. Наконец, в 1913 году, Джордж и Гертруда прибыли в Эймс, штат Айова, где они оставались до 1958 года Оба сына, Эдвард Крозье и Джеймс Джордж, не родились и выросли в городе Эймс. Именно там, где Snedecor сделал свой след в мире статистики. 2. В начале штата Айова колледж лет Джордж Снедекор поступил на факультет штата Айова колледже в качестве помощника профессора математики в 1913 году он был быстро назначен адъюнкт-профессором, а в 1915 преподавал курс под названием "Математическая теория статистики". Это было первое статистика курс предназначен для официально представить статистические методы для исследователей в области математических и других наук. Название курса в ближайшее время изменяется на "статистических методов интерпретации экспериментальных данных", и второй курс под названием "Биометрические методы интерпретации данных по сельскому хозяйству" был введен.

К началу 1920-х годов, с несколькими отделами через университетский городок, предлагающие курсы в области статистики, исследователи во многих областях применения используют такие методы, как регрессионный анализ по их данным. Выпускник колледжа был создан в 1920 году, с миссией содействие научным исследованиям, особенно в районах, имеющих важное значение для сельского хозяйства. Snedecor, с его интересом к междисциплинарным исследованиям, скорее всего, были влиятельными на таких исследователей, как Линдстром, который к 1925 году использовал несколько моделей регрессии, довольно сложности для тех дней (Кокс и Хомайера, 1975; Мельде, 1990; David, 1998).

В 1927 году, математика статистическая служба, созданная при Snedecor и А. Е. Брандт у руля. Услуги была создана, чтобы оформить статистические консультационную деятельность, и был предшественником Статистической лаборатории, созданной в 1933 году Гертруда Кокс (QV), который работал неполный рабочий день в Математическом статистической службы, был первым учеником в аспирантуру с MS степень в области статистики, награжден кафедры математики в 1931 году.

К концу 1920-х годов Snedecor был признан как выдающийся статистик, а в 1931 году получил должность профессора на кафедре математики штата Айова колледже. В течение этого времени, Snedecor разработаны важные отношения с Р. А. Фишером (QV), которые производила новаторские исследования в Ротамстедской опытной станции. Их переписка привела к двум влиятельных шести недель визитов Фишер штата Айова колледж, в лето 1931 и 1936

также раздел 4. 3. Статистическая Лаборатория лет, 1933-1947 В знак признания важности статистики в других областях исследования, штата Айова колледж создан статистическая лаборатория в 1933 году, с Snedecor в качестве директора. Ибо, может быть, впервые в Соединенных Штатах, экспериментальные статистика была официально признана как наука, отличной от математики. Под руководством Snedecor, в статистическая лаборатория стала активным центром для научных исследований и преподавания статистики. Многие посетители были привлечены и постоянный персонал был существенно увеличен в результате соглашения о сотрудничестве в 1938 году с Министерством сельского хозяйства США. РГ Кокран был нанят в это время и сыграл ключевую роль в укреплении работу в выборке исследования и запуск программы PhD в области статистики (ЦОГ и Хомайера, 1975). В 1930-е были наиболее активную научно-исследовательскую лет Snedecor-х годов. Он также выполнял административные обязанности в качестве директора Статистической лаборатории, несмотря на глубокую неприязнь к документации. Действительно, это было его нежелание брать на себя каких-либо дополнительных административных обязанностей, которые могли привести его к сопротивлению созданию Департамента статистики в 1930 году. Как это было, Гертруда Кокс пошел на создание кафедры экспериментальной статистики штата Северная

Каролина колледж в 1941 году, первым в стране. Iowa State College Департамент статистики был создан в 1947 году. 4. Публикации В 1924 Генри А. Уоллес, тогдашний редактор Фермер Wallaces а позднее министра сельского хозяйства и вице-президента Соединенных Штатов, дал серию из десяти субботу днем семинары в Айове State College. Он хотел, чтобы перейти к аудитории в основном сельскохозяйственных и биологических научных работников экспертизы в множественной регрессии, что он получил, прежде всего, в изучении факторов, влияющих на кукурузу урожая (David, 1998). Уоллес Обобщив эти лекции в доску корреляции и машины расчетом, что был поставлен в окончательной форме путем Snedecor, один из его слушателей. Это буклет всего 47 страниц (Уоллес и Snedecor, 1925) достиг по всему миру тиражом и был первый публикацию Snedecor в году, в возрасте 42. Вскоре после этого Snedecor стало известно о фундаментальный вклад Р. А. Фишера и был одним из первых в Соединенных Штатах, чтобы признать их большое практическое значение. Это привело к значительному пересмотру бюллетеня (Уоллес и Snedecor, 1931), в небольшой книге, в которой F-статистики вводится (Snedecor, 1934, стр. 15) и, самое главное на сегодняшний день, в статистических методов (Snedecor, 1937)

. Статистические методы был феноменальный успех, и пошел через восемь изданий. Snedecor был единственным автором первой пятерке, три последние являются в соавторстве с WG Кокран. Восьмое издание было подготовлено в 1989 году несколько членов Государственного Департамента статистики Айова. Долгое время после смерти соавторов, текст еще около 2000 записей в Science Citation Index за 1995 Суммарный объем продаж изданий на английском языке (кроме тех, опубликованы в Индии) достигли 237000. Первое издание уже дает понять, почему эта книга была доказать столь успешной. Snedecor был дар для написания просто и ясно, часто используя разговорный стиль. Необходимость статистического подхода представлена с энтузиазмом через ряд примеров, вероятно, будут близко к сердцу читателя. Snedecor постоянно задавали вопросы. Были много таблиц и диаграмм. Книга была усеяна полезные и разнообразные упражнения. Читатели ржавые, даже элементарной математики не были проигнорированы. Каждая тема была введена неторопливо и лечение довольно подробно, но удивительное количество материала было покрыто. Многие из работ, указанных отражение богатый опыт Snedecor набрала в качестве консультанта исследователей. Кроме того, Snedecor заполнено реальную потребность удовлетворить растущего интереса биологических работников научно-исследовательских в полезных статистических методов.

Основные тексты, Фишера Статистические методы для научных работников, впервые опубликованный в 1925 году, и затем к постановке опытов в 1935 году, были впечатляющими и написано в мастерски свободное английском стиле. Но научные работники нашли их очень сложно. Интересно отметить, что в ходе теплой соответствии с Фишером, Snedecor пишет в 1936, что он работает на элементарном текста "", направленные на обеспечение новичка к пониманию ваших книг. " Эта цель Snedecor осуществляется прекрасно, даже если его презентация не обошлось без случайных недостатков. Журнальные статьи Snedecor являются в основном в том же духе борьбы, как простой способ, как можно с проблемами, представляющих непосредственный интерес для экспериментатора. Многие из этих работ являются совместные с экспериментаторами или статистиков штата Айова, по темам, в том числе опытно-конструкторских, анализ дисперсии и ковариации, критерия хи-квадрат, непропорционально большое число подклассов в нескольких классификаций, и отбора проб.

Snedecor написал также сам по себе на последних двух темах, а также о компьютерных и статистической образования и философии. Обширные знания он приобрел, помогая исследователям квалифицировали его превосходно, чтобы быть первым редактором для запросов в области биометрии, с 1945 по 1958 год Если поток запросов от читателей был медленным, он мог опираться на свой огромный опыт

консультирования по производству запрос! 5. Департамента статистики лет, 1947-1958 В 1947 году Джордж Snedecor ушел с поста директора Статистической лаборатории. Он оставался в качестве профессора статистики до своего ухода из штата Айова колледже в 1958. К 1940 году Snedecor был признан во всем мире как авторитет в области разработки статистических методов. Он был активным членом Американской статистической ассоциации (ASA) и биометрического общества. Он служил в качестве вице-президента АСК в 1947 году, и в качестве президента в 1948 году В 1950 году он был избран членом Международного статистического института, 28 из Соединенных Штатов Америки (Кокс и Хомайера, 1975). Королевского статистического общества сделали его почетным членом в 1954.

В 1958 году Snedecor ушел из штата Айова колледж после 45 лет службы. Он был ответственным за создание и расцвет, пожалуй, наиболее влиятельной группы прикладных статистиков в Соединенных Штатах в первой половине века и его наследие длится по сей день. Департамент статистики в Университете штата Айова в настоящее время находится в Snedecor зале, данью человеку. После выхода на пенсию в Сан-Диего, штат Калифорния, Snedecor сохранил живой интерес к статистике, пока в конце жизни. Он был далее удостоен премии Медаль Samuel S. Уилкс Мемориал в 1970 году, в результате выборов в качестве почетного Life Member биометрического общества в 1971 году в книге статей в его честь (Бэнкрофт, 1972), и посмертно создание в 1976 году Джордж Снедекор премии за лучшую публикацию в биометрии, в ведении ASA. Он умер в 1974 году, в возрасте 92, в городе Амхерст, штат Массачусетс.

библиография

1 Бэнкрофт, Т.А. (ред.) (1972). Статистические документы, в честь Джордж Снедекор. Iowa State University Press, Эймс, штат Айова.

2 Кокс, Г.М. и Р.Г. Хомайера (1975). Профессиональные и личные проблески Джордж Снедекор. Биометрия, 31, 265-301.

3 Дэвид, Н.А. (1998). Статистика в американских университетах в 1933 году и создания статистической лаборатории штата Айова. Статистической науки, 13, 68-74.

4 Snedecor, G.W. (1934). Расчет и интерпретация дисперсионного анализа и ковариации. Энциклопедический Press, Эймс, штат Айова.

5 Snedecor, G.W. (1937). Статистические методы, применяемые для экспериментов в области сельского хозяйства и биологии. Энциклопедический Press, Эймс, штат Айова.

6 Уоллес, Н.А. и G.W. Snedecor (1925). Корреляция и расчет машина. Iowa State College официального опубликования, 23 (35).

7 Уоллес, Н.А. и G.W. Snedecor (1931). Корреляция и расчет машина. Дополненное издание. Iowa State College официального опубликования, 30 (4). Алисия Л. Carrigu и Герберт А. Давид