П. Ф. Рокицкий

ОСНОВЫ ВАРИАЦИОННОЙ СТАТИСТИКИ ДЛЯ БИОЛОГОВ

МИНСК 1961

Проф. П. Ф. РОКИЦКИЙ

ОСНОВЫ ВАРИАЦИОННОЙ СТАТИСТИКИ ДЛЯ БИОЛОГОВ

ИЗДАТЕЛЬСТВО
БЕЛГОСУНИВЕРСИТЕТА имени В. И. ЛЕНИНА
МИНСК — 1961

Книга представляет собой учебное пособие по курсу вариационной статистики для студентов биологических факультетов университетов и других вузов биологического профиля, а также может быть использована научными и практическими работниками—биологами различных специальностей.

В работе подробно и последовательно изложены необходимые для биологических исследований статистические методы: группировка материала, составление вариационных рядов, вычисление важнейших статистических показателей, характеризующих совокупности, измерение корреляции и регрессии. Уделено внимание понятиям вероятности и достоверности и их значению для анализа биологических данных.

Изложенный в книге материал иллюстрируется большим числом конкретных примеров из различных областей биологии. Каждая глава содержит проверочные вопросы и задачи на материале ботаники, зоологии, животноводства, растениеводства, физиологии, генетики. В приложении дается перечень статистических показателей и их формул и 9 статистических таблиц, необходимых для оценки ряда показателей и проверки результатов.

ВВЕДЕНИЕ

Общеизвестно, что в физике и химии, а также в различных областях техники и народного хозяйства широко применяются математические приемы и методы, с помощью которых можно точно характеризовать те или иные явления и выражать с помощью математических формул разнообразные связи и зависимости между ними. В настоящее время биология претерпевает очень резкий переход от чисто описательной науки, которой она была в прошлом, к точной и экспериментальной, пользующейся количественными методами исследования и основывающейся не только на наблюдении, но и на эксперименте. Значение количественных методов особенно подчеркивается тем фактом, что в биологии приходится иметь дело, как правило, не с одним, а со многими объектами. В озерах и реках водится множество рыб, относящихся к одному и тому же и к разным видам, миллионы экземпляров различных видов рачков, моллюсков, коловраток, водорослей, инфузорий и других животных и растений. В совхозах и колхозах разводят сотни и тысячи голов крупного рогатого скота, тысячи и десятки тысяч голов овец, тысячи свиней, кур и других животных, выращивают миллионы различных полевых, огородных, садовых и иных растений.

Все эти организмы характеризуются самыми различными показателями: коровы—удоями за лактацию или за 300 дней лактации, процентом жира молока, живым весом; рыбы—весом, длиной тела; овцы—настригами, длиной и тониной шерсти; куры-несушки—количест-

вом снесенных за год яиц; колосья пшеницы—количеством зерен в колосе, весом отдельных зерен и т. д.

Для того, чтобы суметь разобраться в разнообразных количественных данных о животных и растениях как в условиях производства, так и в научных опытах с ними, нужны определенные приемы и методы, заимствованные из математики.

Сравнение между собою сортов растений или пород животных, изучение влияния различных внешних факторов на физиологические и биологические процессы, на хозяйственные и биологические свойства организмов, изучение видов, подвидов, экотипов, географических рас в природе—все это может быть сделано только путем применения точных количественных методов наблюдения или эксперимента и последующего математического анализа результатов. Только в этом случае можно сделать правильные выводы. Особенно это относится к быстро развивающимся новым отраслям биологии: экологии, генетике, теоретической систематике, селекции, физиологии и фармакологии.

Роль количественного подхода к явлениям природы была замечательно выражена еще в известном завещании Галилея: измерять все, что измеряется, и делать пригодным к измерению то, что пока не поддается измерению.

Дело не только в применении математических приемов для количественного изучения, но и в правильной оценке количественных закономерностей в явлениях природы, ибо на основе количественного анализа можно выявить их качественное своеобразие и отсюда правильно подойти к проведению необходимых наблюдений и к постановке опытов, оценить полученные цифровые данные и предостеречь против возможных исходных ошибок в наблюдениях или опытах.

Применение математических методов к живым существам составляет особую науку—биометрию.

Термин «биометрия» (или «биометрика») был предложен еще в конце XIX века. Биометрия рассматривалась как наука о применении математических методов для изучения разнообразия живых существ. В настоящее время она дает в распоряжение биолога большое количество приемов, подчас очень сложных, с помощью которых можно значительно лучше выяснить особенности

того или иного материала (животных, растений или микроорганизмов). Однако мы ограничимся рассмотрением небольшой группы элементарных математических приемов, которые необходимы для работы биолога, зоолога, ботаника, растениевода, животновода, входящих в область лишь одной математической науки,—вариационной статистики. Современная же биометрия (или биоматематика, как иногда ее называют) использует приемы и некоторых других областей математики. Но приемы вариационной статистики имеют наибольшее значение, ибо они позволяют проводить количественный анализ массовых явлений, то есть таких явлений, которые охватывают значительное количество отдельных изменяющихся величин (слово «вариация» в переводе и означает «изменение», «колебание»). Как раз вариационная статистика и занимается математическим изучением закономерностей, проявляющихся в массовых явлениях. Вот почему наш курс называется не биометрией, а вариационной статистикой.

В излагаемых ниже главах будут даны лишь элементарные основы вариационной статистики в размерах, отведенных на этот курс учебными планами биологических факультетов университетов.

Более полные сведения о вариационной статистике дают специальные руководства, которые указаны в заключительной части книги.

Глава 1

ГРУППИРОВКА ДАННЫХ, СОВОКУПНОСТЬ И ВАРИАЦИОННЫЙ РЯД

Характеристика совокупности. Всякое множество отдельных изменяющихся объектов составляет так называемую совокупность. Совокупностями являются популяции рыжих полевок того или иного района, стадо коров, потомство быка, заготовляемые в определенном районе беличьи шкурки, растения на опытных делянках, группа цыплят, на которых ставится опыт по применению антибиотиков, мальки окуня в озере и т. д. Понятие совокупности приложимо не только к животным и растениям. Такими же совокупностями являются, например, дети, родившиеся в стране в течение какого-то года или месяца, молекулы газа в том или другом объеме. В состав совокупности входят различные члены или единицы: для популяции животных—каждое отдельное животное, для стада коров единицей является каждая корова, для совокупности шкурок—каждая шкурка, для потомства быка— все телки и бычки, от него полученные, для совокупности зерен гречихи—каждое отдельное зерно.

Обычно число единиц совокупности называют объемом совокупности и обозначают латинской буквой n. Единица совокупности может характеризоваться определенными признаками, например, коровы—удоями за лактацию, весом, мастью, молекулы газа—скоростями их движения и т. д. Каждый изучаемый признак принимает разные значения у различных единиц совокупности, он меняется в своем значении от одной единицы совокупности к другой. Это изменение называется вариацией (т. е. изменчивостью) или дисперсией (т. е. рассеяни-

ем). Мы говорим—«признак варьирует». Это означает, что он принимает разные значения у различных членов совокупности, например, коров данной породы, мышей опытной группы, поросят одного помета и т. д. Значение или меру признака для той или иной единицы совокупности называют вариантой и обозначают определенной буквой. Раньше обозначали варианты буквой v, теперь чаще обозначают буквой x. В таком случае ряд вариант в совокупности следует обозначать как $x_1, x_2, x_3, \ldots, x_n$. Самую же варьирующую величину называют случайной переменной. Варианты являются ее числовыми значениями.

Чаще всего в состав совокупностей входят отдельные особи. Так, например, при характеристике стада коров по весу во взрослом состоянии (на 1 января определенного года) за единицу совокупности следует взять каждую корову. Однако возможны случаи, когда единицей совокупности может быть не каждое животное в отдельности, а только какая-то его характеристика. Так, изучая вариацию коров стада по молочной продуктивности, можно взять за единицу каждую лактацию. Тогда при общем количестве коров в стаде, например 100 голов, количество изучаемых за несколько лет лактаций может быть 500 или 600. Отдельными вариантами будут величины удоев за каждую лактацию. Можно изучать вариацию того или иного признака во времени даже на одном животном. Как известно, жирность молока изменяется не только по дням лактации, но и по отдельным дойкам того же дня. Варьирующие данные о проценте жира в молоке определенной коровы, полученные путем измерения жирности за ряд доек и дней лактации, также составляют совокупность, которую можно изучить вариационно-статистическими методами.

Такой же совокупностью, очевидно, является ряд показателей состава крови у одной морской свинки при его изучении в течение какого-то времени.

В общем виде можно сказать, что сумма наблюдений или измерений есть тоже совокупность. Каждое отдельное наблюдение, при котором устанавливается значение случайной переменной, тогда будет единицей совокупности.

Следует иметь в виду, что совокупность может состоять из других, более частных совокупностей. Совокуп-

ность, представляющая собой всех животных данной породы, распадается на частные совокупности—стада отдельных хозяйств, колхозов или совхозов. В пределах стада одного хозяйства можно выделить еще более частные совокупности, например потомство определенных быков.

При постановке опытов по изучению влияния каких-либо антибиотиков на рост крыс внутри совокупности, охватывающей всех опытных и контрольных животных, можно отдельно рассматривать каждую группу, подвергавшуюся воздействию определенных факторов, как самостоятельную, более частную совокупность. Во всех случаях мы будем встречаться с постоянной вариацией как внутри отдельных частных совокупностей, так и между ними.

Задачей изучения всякой совокупности является получение статистических (или, как иногда говорят, биометрических) характеристик или показателей, которые позволяют судить о данной совокупности в целом, о вариации внутри совокупности и об отличии ее от других, сходных с ней или близких к ней совокупностей.

Именно тогда совокупность становится статистической, когда в ее описание вносится количественный метод. Применение количественного метода изучения совокупности и позволяет получать для нее ряд статистических показателей (в специальной литературе статистические показатели нередко называют сокращенно «статистиками»). С их помощью мы получаем основную информацию о совокупности.

Варьирующие признаки и их учет. При изучении единиц совокупности по тем или другим признакам необходимо записать полученные данные.

В настоящее время предпочитают производить такого рода записи на карточках, так как их можно затем группировать любым образом. При большом количестве карточек обработка записей может производиться счетной машиной. В этом случае карточки должны быть перфорированными, т. е. в определенных местах на них должны быть пробиты дырочки или сделаны вырезы в соответствии с записанными цифрами. Машина сама производит необходимые подсчеты по этим дырочкам или вырезам. Наконец, в особо сложных случаях все полученные при опытах или наблюдениях данные переводятся

на условный код, который записывается в соответствующих частях электронно-счетных машин. Такие машины в дальнейшем могут обработать полученные данные, при этом с большой скоростью.

Способы обработки данных сильно зависят от того, каков характер изменчивости (вариации) изучаемых признаков. Различия между вариантами могут выражаться в каких-то качествах. Такую изменчивость называют качественной, или альтернативной. Так, если совокупность животных характеризуют по масти, тогда каждая варианта должна получить качественную характеристику в соответствии с заранее принятыми обозначениями: черная, рыжая, черно-пестрая, черно-рыжая и т. д. В этом простейшем случае подсчет числа особей в каждой из выделенных групп дает представление о совокупности в целом. Подсчет, произведенный в абсолютных числах, можно выразить в процентах и представить в виде диаграммы (столбиками или секторами круга). Однако, как мы увидим в дальнейшем, и при альтернативной изменчивости возможно получение ряда статистических показателей.

В других случаях различия между вариантами являются количественными. Количественная вариация может быть двух типов: прерывистая (дискретная) и непрерывная. В первом случае различия между вариантами, отдельными значениями случайной переменной, выражаются целыми числами, между которыми нет и не может быть переходов. Примером может служить количество детенышей в помете (поросят у свиноматок или щенков у серебристо-черных лисиц), число сосков у свиноматок, число лучей в плавниках рыб, количество лепестков в цветке, число позвонков у птиц и т. д. Для изучения подобного варьирования надо сосчитать у каждой единицы совокупности число изучаемых элементов и записать его на соответствующую карточку. При непрерывной изменчивости значения вариант не обязательно выражаются только целыми числами. Все зависит от того, какая степень точности принимается для характеристики данного количественного признака. Так, например, при изучении веса крупного рогатого скота можно ограничиться значениями вариант, выраженными в килограммах, отбросив граммы, но совершенно недостаточно округлять веса рыб до килограммов, а необ-

ходимо выражать их в граммах, так как грамм здесь имеет большое значение. В опытах же по изучению влияния гормонов на рост гребня у цыплят вес гребня придется измерять в миллиграммах. Молочную продуктивность за лактацию обычно выражают в килограммах, но общая картина удоев не изменится, если округлять количество молока за период лактации до десятков килограммов. Оценка же жирности молока в целых процентах явно недостаточна, ее надо давать в десятых или даже в сотых долях процента. Однако во всех этих и им подобных случаях существует непрерывная вариация, выражающаяся в том, что между вариантами возможны все переходы. При изучении непрерывной изменчивости надо все единицы совокупности характеризовать количественно с той степенью точности, которая заранее намечена и больше всего подходит в данном конкретном случае, и полученные данные (варианты) внести в карточки.

Группировка данных при качественной и дискретной изменчивостях. Чтобы проанализировать ту или иную совокупность, необходимо сгруппировать полученные отдельные варианты и затем представить эту группировку в виде таблицы или ряда. Только при упорядочении полученных данных можно их обработать математически и вывести вариационно-статистические показатели, которые будут исчерпывающе характеризовать изучаемую совокупность. Проблема группировки занимает большое место в статистике вообще (особенно в экономической), так как ошибочная группировка данных может привести к неправильным выводам о существе изучаемого явления.

Наиболее проста группировка при качественной, альтернативной, изменчивости. Так, если при просмотре совокупности из 150 коров рогатых было обнаружено 120, а комолых—30, то полученные данные можно свести в таблицу (табл. 1).

При количественной изменчивости надо предварительно наметить для таблицы классы, охватывающие все изученные количественные данные от минимальных их значений до максимальных. Это легко сделать при прерывистой (дискретной) количественной изменчивости.

Допустим, что была изучена плодовитость 80 самок серебристо-черных лисиц, т. е. число родившихся у каж-

дой самки щенков. Варианты x_1, x_2, x_3, ..., x_{80} этой совокупности будут выражены цифрами, представленными в табл. 2.

Таблица 1

Группировка коров по рогатости и комолости

Типы коров	Количество	Процент
Рогатые	120	80
Комолые	30	20
Всего	150	100

Таблица 2

Количество щенков у 80 самок серебристо-черных лисиц

4	5	3	4	6	7	8	3	1	4
6	4	4	3	2	5	3	4	5	4
5	3	4	5	4	4	4	6	5	7
6	4	5	4	4	4	4	2	3	4
5	5	4	5	4	4	6	4	4	4
4	8	7	5	4	9	4	3	4	4
5	4	6	4	4	3	4	4	4	2
4	4	5	4	6	4	3	3	4	2

Легко видеть, что минимальное число щенков 1, максимальное—9. Отсюда естественно установить 9 классов: с 1 щенком, с 2, с 3 и т. д. и распределить все варианты по этим 9 классам. Наиболее простым способом разнесения вариант по классам является следующий. Составляется таблица с намеченными 9 классами и в соответствующие горизонтальные строчки разносятся все варианты, начиная от первой. Обозначаются они так: первые четыре данного класса—точками, а последующие—черточками, соединяющими четыре точки. Число 10 будет в таком случае фигурой $\boxed{\times}$. В табл. 3 произведена разноска первых 20 вариант, записанных в двух верхних строчках табл. 2.

Предоставляем каждому самому довести разноску данных табл. 2 до конца. В окончательном виде таблица будет иметь вид табл. 4.

Таблица 3

Разноска 20 вариант по классам

Классы (число щенков в помете каждой самки)	Частоты (количество вариант в каждом классе)
1	. 1
2	. 1
3	:: 4
4	⌐ 7
5	:. 3
6	.. 2
7	. 1
8	. 1

Таблица 4

Распределение 80 самок серебристо-черных лисиц по классам

Классы	Частоты
1	1
2	4
3	10
4	39
5	13
6	7
7	3
8	2
9	1
	$n = 80$

Вариационный ряд и его графическое изображение. Таким образом, после распределения всех вариант по

классам получился ряд, в котором показано, как часто встречается каждый тип или класс вариант и как варьирует число щенков в помете у отдельных самок, начиная от минимальной величины (1 щенок) и кончая максимальной (9 щенков). Поэтому подобные ряды были названы вариационными. По вариационному ряду можно судить не только о границах колеблемости изучаемого количественного признака, но и о характере вариации. В данном примере в вариационном ряду наиболее частым является класс «4 щенка», следующими за ним по частоте являются классы «3 щенка» и «5 щенков». Наиболее же редкими являются крайние классы «1 щенок» и «9 щенков». Класс, обладающий наибольшей частотой, получил название модального, значения же крайних классов называются лимитами, или пределами.

Всякий вариационный ряд можно изобразить графически. Графическое изображение вариационного ряда получило название кривой распределения, или вариационной кривой. При прерывистой изменчивости это будет, конечно, не кривая, а многоугольник, или так называемый полигон, который показан на рис. 1. При построении кривых распределения, или полигонов, нужно всегда доводить их справа и слева до нулевых классов, т. е. тех соседних классов, в которых уже не было ни одной варианты. В нашем примере ими являются классы «0 щенков» и «10 щенков».

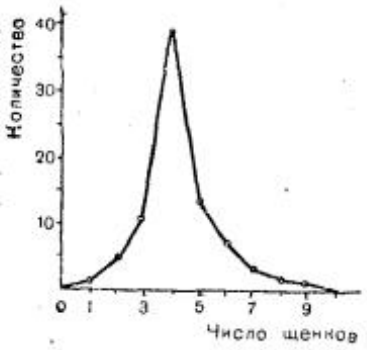


Рис. 1. Кривая распределения 80 самок серебристо-черных лисиц по числу щенков в помете

В приведенном примере модальным был только один класс «4 щенка». Поэтому полигон распределения самок лисиц по числу щенков в помете имел только одну вершину. Однако возможны случаи, когда в вариационном ряду обнаруживается несколько модальных классов, и тогда полигон является многовершинным. Наиболее простой причиной многовершинности, особенно при очень растянутых рядах, является недостаточное количество вариант в изученной группе. При малом числе особей в

некоторых классах вариационного ряда может вообще не быть ни одной варианты. Вариационный ряд окажется прерывистым, а вариационная кривая—разорванной на части. Однако, если при большом числе особей в изучаемой совокупности сохранится дву- или многовершинность, причину этого надо искать в том, что биологический материал в действительности является смешением двух качественно различных совокупностей, которые или находились в резко отличных условиях внешней среды или принадлежат к разным типам, морфам. Так как многие виды в природе являются полиморфными или диморфными, то соединение в одном ряду особей разных морф может дать внешнюю картину дву- или многовершинности. Например, платиновые лисицы отличаются по числу щенков от серебристо-черных, поэтому было бы неправильно помещать в один вариационный ряд по этому признаку и платиновых, и серебристо-черных лисиц.

Группировка данных при количественной непрерывной изменчивости. В этом случае группировка данных является наиболее трудной. Допустим, что в результате измерения веса 25 кроликов различных пород были получены варианты, представленные в табл. 5, при этом они расположены в так называемом ранжированном виде, т. е. от меньших величин к большим.

Таблица 5

Веса 25 кролнков в кг (для большей наглядности взяты кролики различных пород)

3,2	4,5	5,2	5,6	6,0
3,8	4,7	5,2	5,7	6,3
4,1	4,9	5,3	5,8	6,4
4,3	5,0	5,3	5,8	6,7
4,3	5,1	5,4	5,9	7,3

Здесь нет тех естественных классов, с которыми мы встречались при анализе прерывистой, дискретной вариации. Их необходимо наметить произвольно. Разница между наибольшим и наименьшим значениями вариант в нашем примере равна 4,1 (7,3—3,2). Но чтобы иметь примерно 8—9 классов, размеры их должны быть 0,5 кг. В таком случае можно наметить следующие классы:

3,0—3,4 кг; 3,5—3,9 кг; 4,0—4,4 кг; 4,5—4,9 кг; 5,0—5,4 кг; 5,5—5,9 кг; 6,0—6,4 кг и т. д.

На правильное построение шкалы для классов надо обращать очень большое внимание. Во-первых, необходимо, чтобы величина классового промежутка была всегда одной и той же. Было бы неправильно, если бы в начале ряда был взят классовый промежуток 0,5 кг, как в нашем примере, а в конце ряда—1,0 кг (величину классового промежутка обычно обозначают буквой i). Во-вторых, границы классов должны быть намечены таким образом, чтобы одна и та же цифра не повторялась в двух классах. Если первый класс заканчивался величиной 3,4, то второй класс должен начинаться со следующей по порядку цифры—3,5. Если бы классы были намечены следующим образом: 4,0—4,5; 4,5—5,0; 5,0—5,5 и т. д., то всегда было бы сомнение, к какому классу отнести особь со значением 4,5 или 5,0. Если же один класс будет охватывать значения вариант от 4,0 до 4,4 включительно, а другой—от 4,5 до 4,9 включительно, разноска вариант по намеченным классам не вызовет затруднений. Ее можно проводить тем же методом, который использован при составлении табл. 3 и 4.

В результате данные по весам кроликов будут представлены в табл. 6.

Таблица 6

Распределение 25 кроликов по весу

Классы, кг	Частоты	Классы, кг	Частоты
3,0 — 3,4	1	3,0 — 3,9	2
3,5 — 3,9	1	4,0 — 4,9	6
4,0 — 4,4	3	5,0 — 5,9	12
4,5 — 4,9	3	6,0 — 6,9	4
5,0 — 5,4	7	7,0 — 7,9	1
5,5 — 5,9	5		
6,0 — 6,4	3		
6,5 — 6,9	1		
7,0 — 7,4	1		
$i = 0{,}5$ кг	$n = 25$	$i = 1{,}0$ кг	$n = 25$

В левой части табл. 6 показан вариационный ряд распределения 25 кроликов при разбивке на классы с i=0,5 кг. Ряд получился несколько растянутым—9 классов. Его можно сделать более сжатым, приняв i=1,0 кг, как это сделано в правой части таблицы.

Возникает вопрос, какое же число классов выгоднее. Это зависит от размеров совокупности, т. е. от n. На практике можно руководствоваться примерно следующими правилами:

Количество вариант	Число классов
25 — 40	5 — 6
40 — 60	6 — 8
60 — 100	7 — 10
100 — 200	8 — 12
200 — 500 и более	10 — 15

При выборе числа классов надо одновременно иметь в виду размеры классового промежутка. Они должны быть или целыми числами или округленными дробями. Лучше, чтобы i было равно 0,5; 1; 5; 10, а не 0,45; 1,1; 6; 11, если даже число классов при этом будет несколько меньшим или большим указанного выше.

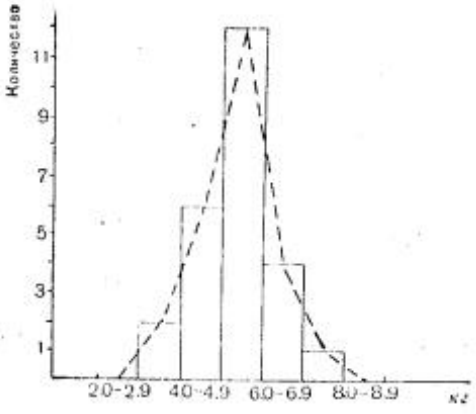


Рис. 2. Гистограмма распределения 25 кроликов по весу

Вариационный ряд при непрерывной изменчивости также может быть изображен на графике. В этом случае на горизонтальной линии (оси абсцисс) надо нанести классы, а на вертикальной (оси ординат) — численности классов в виде столбиков, расположенных основаниями на обозначениях каждого

класса и имеющих высоту, пропорциональную частотам, как это показано на рис. 2. Такой ступенчатый график носит название гистограммы. Из гистограммы легко получить полигон, или кривую распределения, соединив линиями середины верхних сторон всех столбиков, как это показано на этом же рисунке. Нужно помнить, что началом и концом полигонов должны быть середины соседних нулевых классов (2,0—2,9 и 8,0—8,9). Однако правильнее пользоваться при непрерывной изменчивости только гистограммами.

Характер распределения вариант в вариационном ряду. Изучая распределение вариант в вариационных рядах, представленных в табл. 4 и 6 и выраженных в виде графиков на рис. 1 и 2, легко заметить некоторые общие закономерности, а именно: 1) большинство вариант располагается в средней части вариационного ряда или около середины вариационной кривой, здесь наблюдается максимум вариант, как бы их сгущение; 2) распределение вариант в обе стороны от этого максимума более или менее симметрично; 3) частота вариант постепенно убывает к краям вариационного ряда.

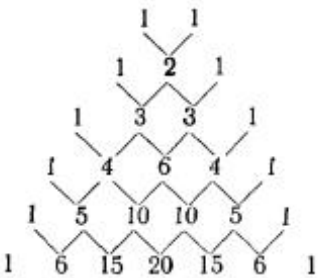
Эти закономерности в той или иной степени присущи любому вариационному ряду. В середине XIX в. указал на них бельгийский статистик Кетле, который впервые построил вариационную кривую, изучив распределение по росту 26 000 солдат американской армии. Кетле пришел к выводу, что распределение особей в вариационном ряду следует коэффициентам разложения двучлена, возведенного в известную степень. Вспомним, какими будут коэффициенты при отдельных членах разложения бинома Ньютона $(a+b)$ при возведении его в разные степени:

$$(a+b)^1 = a + b$$
$$(a+b)^2 = a^2 + 2\,a\,b + b^2$$
$$(a+b)^3 = a^3 + 3\,a^2\,b + 3\,a\,b^2 + b^3$$
$$(a+b)^4 = a^4 + 4\,a^3\,b + 6\,a^2\,b^2 + 4\,a\,b^3 + b^4 \text{ и т. д.}$$

Эти коэффициенты легко получить с помощью треугольника Паскаля, в котором цифры каждого последующего ряда получаются путем сложения двух цифр ряда, расположенного над ним (см. на стр. 18).

Распределение вариант в виде вариационного ряда, частоты в котором следуют коэффициентам разложения

бинома, может быть наглядно показано с помощью аппарата Гальтона (рис. 3). Этот аппарат представляет собой коробку, в верхней части которой расположен ящичек с выходным отверстием посредине. В средней части



коробки воткнуты булавки, причем булавки каждого последующего ряда расположены против середин промежутков предыдущего ряда. В нижней части коробка разделена перегородками на ряд отделений. Коробка ставится наклонно, примерно под углом 30° к поверхности пола или стола. В верхний ящичек насыпается дробь. Отдельные дробинки, падая через отверстие ящичка, встречают на своем пути булавки, при столкновении с ними отклоняются вправо или влево и, наконец, падают в отделение нижней части аппарата. Оказывается, что накопление дробинок в этих отделениях образует фигуру, аналогичную гистограмме вариационного ряда, с характерной концентрацией большинства вариант в средней части и рассеянием их вправо и влево. Подобно тому как положение вариант в вариационном ряду является результатом суммирования очень многих случайных факторов, создававших отклонения вариант от некоторого среднего положения, так и расположение дробинок в отделениях аппарата является результатом встреч дробинок со многими булавками, при которых дробинки могли многократно отклоняться в сторону от прямого пути. Чаще всего происходило взаимное погашение этих отклонений: дробинка, первый раз отклонившись вправо, второй раз отклонялась влево и т. д. и в конечном счете попадала в одно из средних отделений. В других, более редких случаях отклонения в одном и том же направлении вправо или влево совпадали, и дробинка попадала в одно из крайних правых или левых отделений. В силу разнонаправленности воздействий на дробинки (и на от-

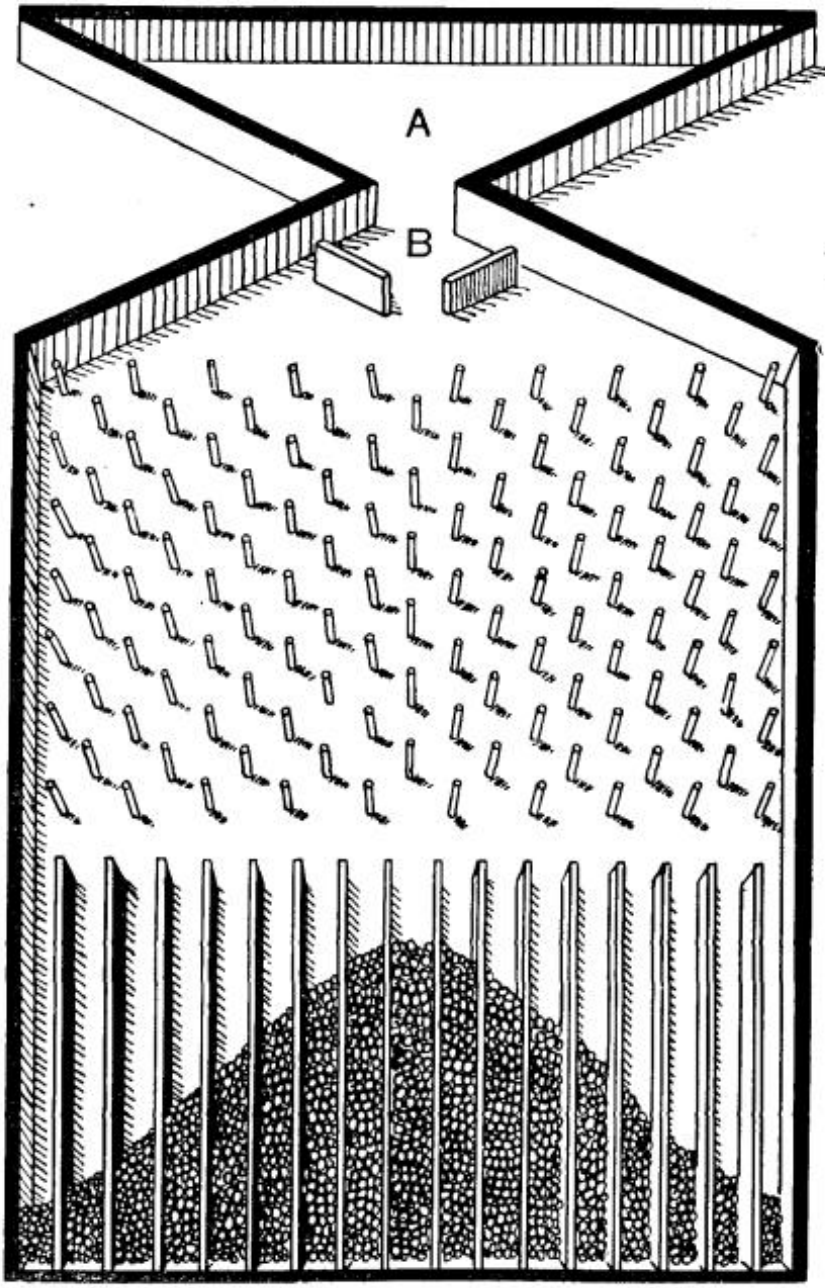


Рис. 3. Аппарат Гальтона

дельные варианты вариационного ряда) отклонения их в обе стороны одинаковы, но максимальные отклонения являются самыми редкими.

В дальнейшем мы увидим, что закономерности вариационного ряда основываются на закономерностях случайной вариации, изучаемых теорией вероятностей.

ВОПРОСЫ

1. Что такое совокупность? Примеры различных совокупностей.
2. Что такое варианта? Случайная переменная?
3. Какими могут быть различия между отдельными значениями (вариантами) случайной переменной?
4. Принципы группировки данных при альтернативной изменчивости? При количественной дискретной? При количественной непрерывной изменчивости?
5. На сколько классов надо разбивать фактические данные при количественной изменчивости? Целесообразно ли намечать 10—15 классов, когда $n<100$?
6. Что такое вариационный ряд? Особенности распределения вариант в вариационном ряду.
7. В чем разница между гистограммой и полигоном распределения? Можно ли превратить гистограмму в полигон?
8. Каковы возможные причины многовершинности вариационных кривых?
9. Что иллюстрирует аппарат Гальтона?

ЗАДАЧИ

1. Представьте в виде вариационного ряда и графически данные о длине листьев садовой земляники (в см):

8,2	9,7	6,6	7,4	8,0	6,4	6,6	6,8	8,4	7,1
9,0	6,0	7,6	8,1	11,8	5,8	9,3	7,3	8,2	7,2
7,2	6,4	7,7	9,0	8,1	7,1	7,1	8,8	7,5	9,2
7,5	6,8	7,0	6,4	7,4	8,2	6,3	7,0	8,1	10,0
7,0	7,1	8,7	6,3	8,6	7,7	7,3	8,0	8,4	9,3
7,3	6,0	7,7	6,1	9,6	7,4	7,2	7,2	8,7	7,5
9,1	6,4	8,3	6,5	8,2	7,2	6,9	6,9	8,2	9,0
7,4	8,0	8,4	7,0	7,1	7,4	6,6	6,4	8,3	7,9
8,3	7,2	7,2	6,6	6,6	7,7	8,7	5,6	7,5	5,7
6,9	7,4	7,2	6,2	6,9	6,8	9,2	9,2	7,1	6,5
5,2	8,0	7,1	8,4	8,1	6,8	6,1	6,8	7,9	8,0
5,6	7,8	7,2	8,8	6,6	6,6	5,6	8,1	9,0	8,4
7,1	7,4	8,7	8,9	7,8	7,3	8,6	8,7	8,2	8,9
6,4	8,6	7,8	5,7	8,5	10,4	8,6	7,7	8,1	8,2
8,5	7,8	7,9	7,5	6,7	7,0	7,9	7,5	8,7	6,8
8,1	7,8	7,8	8,2	7,2	7,9	9,5	7,6	7,0	7,0
7,7	8,1	7,3	7,0	7,4	7,6	8,4	7,3	5,9	9,4
7,8	7,0	7,6	6,6	7,5	9,3	8,1	7,4	8,6	8,2
8,0	7,0	7,0	10,2	6,3	9,6	8,4	8,4	8,0	7,4
8,0	6,2	6,8	10,3	8,5	7,0	7,8	8,1	7,0	7,2

2. В 400 квадратах гемацитометра было подсчитано число дрожжевых клеток. Представьте эти данные в виде вариационного ряда, а также графически:

2 2 4 44	524 77	475 2 8	67344
3 3 2 42	542 86	366 10 8	35644
7 9 5 27	442 44	435 6 5	41426
4 1 4 73	235 82	953 9 5	52434
4 1 5 93	446 65	465 5 4	35964
4 4 5 104	438 32	141 5 6	42333
3 7 4 51	857 95	895 6 6	43744
7 5 6 36	745 86	334 3 7	44453
8 10 6 33	652 53	1137 4 7	35534
1 3 7 25	553 34	656 1 6	44464
4 2 5 48	634 65	266 1 2	22522
5 9 3 56	465 71	365 4 2	89453
2 211 46	646 25	357 2 6	55127
512 5 82	421 64	512 9 1	34736
5 6 5 44	527 62	735 4 4	54754
8 4 6 65	335 74	555 6 10	23835
6 6 4 26	675 45	867 6 4	26114
7 2 5 74	645 15	1087 5 4	64475
4 3 1 62	533 37	437 8 4	73144
7 6 7 24	513 124	228 7 6	76354

3. Было подсчитано число лучей в хвостовых плавниках камбалы:

53	51	52	55	56	49	51	52	54	56
54	53	52	53	51	55	53	55	53	54
51	51	56	54	54	53	54	53	55	53
52	55	53	53	56	53	52	56	52	52
56	55	50	54	49	54	54	55	54	55
52	51	55	52	55	54	51	54	53	54
54	56	54	55	53	53	56	55	54	53
55	52	53	52	51	55	53	54	51	50
53	54	55	52	55	52	53	50	53	52
58	57	57	58	56	57	56	58	57	57

Составьте вариационный ряд и начертите полигон распределения.

4. У 100 свиней были получены следующие привесы за 20 дней (в кг):

1,2	2,8	4,4	4,8	5,2	5,6	6,1	6,4	6,8	6,8
7,2	7,2	7,2	7,6	7,6	7,6	8,0	8,1	8,4	8,4
8,5	8,8	8,8	9,2	9,2	9,6	9,6	9,7	10,0	10,1
10,2	10,4	10,4	10,4	10,5	10,7	10,8	10,8	11,1	11,2
11,2	11,4	11,6	11,6	11,7	12,0	12,0	12,0	12,1	12,1
12,1	12,1	12,1	12,1	12,2	12,3	12,4	12,4	12,5	12,6
12,7	13,1	13,2	13,3	13,3	13,5	13,6	13,7	13,8	13,8
14,0	14,2	14,3	14,4	14,6	14,8	15,1	15,2	15,5	15,6
15,6	15,8	15,9	16,0	16,3	16,4	16,7	16,8	16,8	17,1
17,2	17,3	17,6	18,0	18,4	18,8	19,2	19,6	21,2	22,8

Составьте вариационный ряд и начертите гистограмму.

5. По данным Г. В. Гладкого, длина тела у 77 экземпляров плотвы оз. Швакшта (в мм) была следующей:

143	157	148	153	150	142	164	139	139	140
143	120	144	130	138	124	127	137	139	129
128	119	120	138	130	114	126	138	117	132
130	145	140	153	137	142	145	137	141	125
143	138	140	135	135	139	125	137	131	120
127	118	120	124	134	111	132	133	100	132
143	134	138	130	135	133	134	151	107	110
94	95	142	148	136	165	172			

Составьте вариационный ряд и начертите гистограмму.

6. По данным Г. В. Гладкого, обхват тела у густеры оз. Швакшта (в мм) выражался следующими числами ($n = 80$):

80	75	75	85	78	85	80	77	83	85
88	94	95	86	80	75	78	90	95	90
80	75	83	70	78	83	75	78	86	81
62	77	75	73	80	80	74	73	82	72
80	90	80	78	60	65	75	72	64	67
74	80	68	75	76	65	70	78	75	83
85	70	88	73	56	75	70	73	68	66
65	68	75	78	63	68	68	70	60	56

Составьте вариационный ряд и начертите гистограмму.

7. Изучен живой вес 63 телят холмогорских помесей при рождении (в кг):

27	32	32	31	32	28	37	35	26	28
32	39	34	30	37	26	27	40	35	37
28	43	26	35	45	26	35	32	32	35
35	28	32	36	32	36	37	33	28	31
36	33	33	28	23	26	34	32	36	27
32	39	30	30	36	38	24	32	30	31
28	36	36							

Составьте вариационный ряд и изобразите его на графике.

8. Составьте вариационный ряд и изобразите его графически для следующих данных об удоях коров за 300 дней лактации (в кг):

3586	2761	2825	3807	3858	3904	3530	1951	2362	2729
3453	2635	3752	2666	3331	923	2948	3428	2574	2581
3165	2361	4055	2440	2763	2838	2893	2461	791	4011
2148	2144	2856	2293	3246	2965	3920	3205	2949	2559
2358	2766	2849	3420	2833	3528	3250	1474	2632	2108
2580	3468	903	3027	3177	3666	3242	2715	2730	2748
3115	2330	3339	2033	1850	2093	3642	3736	3847	4080

3847	2934	3676	4155	3306	3734	2199	2468	2448	3293
3465	2540	4288	3685	4708	3758	2735	3363	3306	3511
4052	3380	3154	4571	1426	2981	3224	1480	1586	1953
2340	2520	2855	2600	3711	3073	3708	4167	4526	1600
1360	2192	2610	3390	3350	3009	3940	3510	3658	2326
3445	3170	2271	2007	2107	4901	3002	2934	3007	1687
3458	1915	3090	1917	3382	4773	2331	1420	3656	1966
3651	4174	1274	2247	3859	1548	2620	3564	4507	2562
4659	4985	2132	3047	4582	2815	2973	4305	2340	3043
3021	4194	2654	3001	5290	2665	3230	5235	3936	4980
3148	3015	1785	2088	2026	2390	2064	4207	2540	4853
1450	2118	2936	4510	4216	3315	2821	3431	3354	4106
1501	2454	3287	4580	1965	1563	3559	3401	2728	3491

Глава 2

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ ДЛЯ ХАРАКТЕРИСТИКИ СОВОКУПНОСТИ

Относительные числа. Характерным свойством всякой совокупности является наличие постоянных различий между ее членами. Совокупность представляет собой всегда какое-то множество варьирующих единиц. Возникает вопрос, каким образом, с помощью каких показателей можно охарактеризовать совокупность как целое. Если различия между вариантами являются качественными, или альтернативными, то наиболее простыми показателями являются относительные числа, показывающие, какую долю, выраженную в процентах от объема всей совокупности, занимают варианты с теми или другими качествами. Примером подобных относительных чисел является процент рогатых и комолых коров в стаде, приведенный в табл. 1. Относительные числа показывают структуру совокупности, т. е. соотношение между всей совокупностью и ее частью (или частями).

Лимиты. При количественной изменчивости способы характеристики совокупности могут быть очень разнообразными. Одним из них является указание на границы или лимиты изменчивости. Лимиты позволяют в известной степени судить о характере совокупности. Если известно, что лимиты молочной продуктивности одного стада 2000 и 4000 кг, а другого 2100 и 6800 кг, то на первый взгляд можно сделать вывод о более высоком качестве второго стада по сравнению с первым. Однако лимиты не дают указаний на то, как распределяются по изученному признаку отдельные члены совокупности и какое значение признака является наиболее характерным для

совокупности. Возможно, что большинство коров и в первом и во втором стадах имеет удои около 300 кг, крайние же варианты, по которым определены лимиты, были в этих стадах единичными животными. Первоначальное, основанное только на лимитах мнение о более высокой продуктивности второго стада может быть ошибочным. Вот почему нужны такие показатели для характеристики, которые отражали бы свойства всех ее членов.

Мода и медиана. При изучении распределения самок лисиц по числу щенков в помете обнаружилось, что 39 самок из общего числа 80 имели по 4 щенка, т. е. класс «4 щенка» обладал наибольшей частотой. Такой класс был назван модальным. Значение модального класса называют модой. Мода обозначается символом Мо. Величина моды является как бы типичной для совокупности 80 лисиц, так как действительно почти половина всех самок имела в помете именно 4 щенка. К числу средних величин относится также медиана (обозначается Ме). Медиана—это значение варианты, находящейся точно в середине ряда. Для того, чтобы найти такую варианту, надо сначала расположить все варианты по порядку от минимальных их значений до максимальных. Такое расположение вариант называют ранжировкой. В табл. 5 веса 25 кроликов представлены в ранжированном виде. 13-я по счету варианта разделяет ряд из 25 вариант точно пополам. Она имеет значение 5,3 кг. Это число и является медианой данного ряда. Чтобы определить Ме при четном числе вариант, надо взять значения двух соседних срединных вариант, например при $n=80$ значения вариант с порядковыми номерами 40 и 41, и разделить их сумму на 2. В примере, представленном в табл. 4, обе эти варианты будут иметь значения «4 щенка». Ме данного ряда равна 4,0. Медиана и мода дают известное представление о совокупности в целом. Они характеризуют как бы тип, типичное в данной совокупности (конечно, речь идет только о данном изучаемом признаке).

Средняя арифметическая. Всякая средняя представляет собой абстрактную характеристику совокупности. Нахождение средней — это в сущности замена индивидуальных варьирующих значений признаков отдельных членов совокупности некоторой уравненной величиной при сохранении основных свойств всех членов совокупности. Этому условию в наибольшей степени удовлетво-

ряет так называемая средняя арифметическая, обозначаемая $\bar{x}$ (ранее ее обозначали M).

Представим себе, что ряд членов совокупности, т. е. ряд значений случайной переменной $x_1, x_2, x_3, \ldots, x_n$, заменим таким же рядом из одинаковых величин $\bar{x}$, т. е.

$$\bar{x}, \bar{x}, \bar{x}, \ldots, \bar{x} \ (n \text{ раз}).$$

Тогда сумма всех вариант совокупности $x_1+x_2+x_3+\ldots+x_n$ будет равна $\bar{x}+\bar{x}+\bar{x}, \ldots, \bar{x}$ (n раз), т. е. $n\bar{x}$. Сумму всех вариант совокупности можно сокращенно обозначить Σx. (Греческая буква Σ —большая сигма — обозначает суммирование; конкретные суммы часто обозначают также латинской буквой S.)

Тогда $\Sigma x = n\bar{x}$, откуда $\bar{x} = \frac{\Sigma x}{n}$. (1)

Иногда пишут также $\bar{x} = \frac{1}{n}\Sigma x$. (1$a$)

Мы получили наиболее простую формулу средней арифметической. Для того, чтобы вычислить среднюю арифметическую, достаточно сложить значения всех вариант (на счетах или арифмометре) и сумму разделить на общее число вариант. В простейших случаях так и делают. Приведенные в табл. 5 веса 25 кроликов в сумме составляют 131,8 кг. Тогда $\bar{x} = \frac{131{,}8}{25} = 5{,}27$ кг. Очевидно, в таких случаях можно пользоваться данными, полученными непосредственно при анализе членов совокупности, не прибегая к группировке вариант. Однако при большом количестве вариант оказывается более выгодным пользоваться сгруппированными данными и прибегнуть к более сложному методу вычисления $\bar{x}$, так как тогда можно будет легче и проще определить некоторые другие показатели, о которых будет говориться ниже.

Если все варианты разнесены по классам, каждый из которых характеризуется определенным значением вариант v, а частота каждого класса f, то $\bar{x}$ можно вычислить по формуле

$$\bar{x} = \frac{\Sigma fv}{n}. \quad (2)$$

Для вариационного ряда, представленного в табл. 4, придется добавить еще одну колонку, в которой надо будет записать произведения чисел первого (v) и второ-

го (f) столбцов, т. е. 1.1; 2.4; 3.10 и т. д. Сумма произведений равна 348. В результате получим $\bar{x} = \frac{\Sigma fv}{n} = \frac{348}{80} =$ = 4,35 щенка.

Вариационный ряд (табл. 6) не может быть обработан таким методом без существенного дополнения. В графе «классы» даны две цифры, в пределах которых располагаются все варианты данного класса. Их надо заменить так называемым центральным значением класса, которое будет соответствовать среднему значению v данного класса. Для этого надо сумму двух чисел, указанных в графе «классы», разделить на два. В данном примере для первого по счету класса центральным значением будет величина $3{,}45\left(\frac{3{,}0+3{,}9}{2}\right)$, для следующего $-4{,}45\left(\frac{4{,}0+4{,}9}{2}\right)$ и т. д. Табл. 6 (правая часть) после преобразования примет тогда следующий вид (табл. 7).

Таблица 7

Распределение 25 кроликов по весу

Классы, кг	Центральные значения классов, кг v	Частоты f	Произведения fv
3,0 — 3,9	3,45	2	6,90
4,0 — 4,9	4,45	6	26,70
5,0 — 5,9	5,45	12	65,40
6,0 — 6,9	6,45	4	25,80
7,0 — 7,9	7,45	1	7,45
		$n = 25$	$\Sigma fv = 132{,}25$

Отсюда $\bar{x} = \frac{\Sigma fv}{n} = \frac{132{,}25}{25} = 5{,}29$ кг.

При обоих способах вычисления средней арифметической используются непосредственно данные о вариантах совокупности. В том случае, если анализируется сложная совокупность, состоящая из нескольких частных, для каждой из которых уже известна средняя арифметическая, можно вычислить так называемую взвешен-

ную среднюю арифметическую для сложной совокупности по формуле:

$$\bar{x} = \frac{\bar{x}_1 p_1 + \bar{x}_2 p_2 + \bar{x}_3 p_3 + \ldots + \bar{x}_n p_n}{p_1 + p_2 + p_3 + \ldots + p_n}, \qquad (3)$$

где $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, ..., $\bar{x}_n$ — средние арифметические отдельных частных совокупностей, а p_1, p_2, p_3, ..., p_n — число членов в каждой частной совокупности (их называют также весами частных совокупностей).

Свойства средней арифметической. Прямой способ определения средней арифметической по указанным формулам в ряде случаев оказывается довольно трудоемкой и кропотливой вычислительной операцией. Кроме того, при его применении нет возможности вычислить некоторые другие биометрические показатели. Поэтому на практике часто пользуются окольными методами вычисления средней арифметической. Они основаны на определенных математических свойствах средних арифметических, которые можно изложить в простой форме без специального доказательства следующим образом.

1. Если каждую из вариант совокупности, для которой вычисляется средняя арифметическая, увеличить или уменьшить на одну и ту же величину, то и средняя арифметическая соответственно увеличится или уменьшится на ту же величину.

В алгебраическом выражении это означает, что если совокупность x_1, x_2, x_3, ..., x_n, имеющая среднюю арифметическую $\bar{x}$, будет заменена совокупностью (x_1-a), (x_2-a), (x_3-a), ..., (x_n-a), а средняя арифметическая для новой совокупности будет равна $\bar{x}-a$.

2. Алгебраическая сумма отклонений отдельных вариант от средней арифметической равняется нулю, т. е. $(x_1-\bar{x})+(x_2-\bar{x})+(x_3-\bar{x})+\ldots+(x_n-\bar{x})=0$. Это положение очень важно для понимания сущности средней арифметической, как своего рода равнодействующего показателя для всех варьирующих величин совокупности. В то же время оно дает возможность проверить правильность вычисления средней арифметической.

3. Сумма квадратов отклонений от средней арифметической меньше суммы квадратов отклонений от любой другой величины A, не равной $\bar{x}$, т. е. $\Sigma(x-\bar{x})^2 < \Sigma(x-A)^2$, если A не равно $\bar{x}$.

Два последние положения позволяют применить непрямой способ вычисления средней арифметической и других биометрических показателей с помощью так называемой условной средней A (иногда его называют также *окольным* способом).

Средняя арифметическая, как и некоторые другие средние, известна издавна. Она имеет очень большое значение в науке и технике. В то же время нужно предостеречь от возможных ошибок в ее понимании. Средняя арифметическая характеризует всю совокупность в целом, а не отдельные члены совокупности, ибо она представляет собой обобщающую абстрактную характеристику совокупности, являющуюся как бы равнодействующей всех определяющих условий, участвовавших в образовании входящих в данную совокупность индивидуальных величин, отдельных значений данной случайной переменной. Если для группы лисиц выведено среднее число щенков в помете 4,3, то эта величина относится только ко всей группе, каждая же отдельная лисица характеризуется своим числом щенков в помете от 1 до 9. Далее, средняя имеет смысл только по отношению к качественно однородной совокупности. Так, нельзя вычислять средний вес или размеры животных без учета состава по возрасту. Надо взять каждую возрастную группу отдельно и для них вычислить $\bar{x}$. Поскольку средняя относится к данной совокупности, перенесение ее на явления, выходящие за ее рамки, рискованно без специального анализа вопроса о правомерности такого перенесения.

В дальнейшем мы увидим, что особое место в вариационной статистике занимает вопрос о том, каким образом на основе данных о той или иной частной совокупности можно делать выводы о других совокупностях подобного же рода. Наконец, средняя относится лишь к отдельным изучаемым признакам и не может быть автоматически перенесена на их сумму.

Непрямой способ вычисления $\bar{x}$. В качестве условной средней A можно взять любую величину, однако выгоднее всего для большей простоты вычислений выбрать такое значение A, которое было бы близко к средней, о чем можно судить по расположению частот в вариационном ряду. Практически это значит, что условной средней A можно принять значение того класса, в котором распо-

лагается наибольшее количество вариант или который находится примерно в середине ряда. Кроме того, A должно быть целым числом. Это упростит все расчеты. В дальнейшем вместо вычисления отклонений всех вариант совокупности от средней арифметической $\bar{x}$ берут их отклонения от принятой условной средней A. Часть из этих отклонений будет иметь знак плюс, другая же часть—минус. Если сумма положительных и отрицательных отклонений от A окажется равной нулю, то условная средняя A полностью совпадет с истинной средней арифметической $\bar{x}$, как это вытекает из второго свойства средней арифметической. Если сумма всех отклонений окажется величиной положительной, значит принятая условная средняя меньше истинной. Если же сумма всех отклонений будет величиной отрицательной, принятая условная средняя больше истинной. В обоих случаях для того, чтобы перейти от условной средней A к средней арифметической $\bar{x}$, надо внести в принятую величину A поправку b.

Таблица 8

Вариационный ряд распределения коров по глубине груди

Классы, см	Центральные значения классов, см	Частоты f	Отклонения каждого класса от условной средней a	fa	fa^2*
66,6—67,5	67	2	—7	—14	98
67,6—68,5	68	1	—6	— 6	36
68,6—69,5	69	3	—5	—15	75
69,6—70,5	70	12	—4	—48	192
70,6—71,5	71	11	—3	—33	99
71,6—72,5	72	20	—2	—40	80
72,6—73,5	73	21	—1	—21	21
73,6—74,5	74	24	0	0	0
74,6—75,5	75	29	1	29	29
75,6—76,5	76	24	2	48	96
76,6—77,5	77	14	3	42	126
77,6—78,5	78	3	4	12	48
78,6—79,5	79	3	5	15	75
79,6—80,5	80	1	6	6	36
		$n = 168$		$\Sigma fa = -177$ $+152 = -25$	$\Sigma fa^2 =$ $= 1011$

* Графа fa^2 в этой и последующих таблицах понадобится для вычисления других показателей.

Она равна сумме всех положительных и отрицательных отклонений вариант совокупности от A, деленной на общее число вариант, т. е. $b=\frac{\Sigma fa}{n}$, где a—отклонение значения вариант каждого класса от A.

Таким образом $\bar{x}=A+b=A+\frac{\Sigma fa}{n}$. (4)

Ход вычислений можно проиллюстрировать на приведенных в табл. 8 данных о промерах глубины груди у 168 коров симментальской породы.

В качестве условной средней A принята величина 74 см. Тогда отклонения от A вариант, находящихся в классах с центральными значениями 67—73 см, будут отрицательными, а отклонения вариант из классов 75—80 см—положительными.

$\Sigma fa=-25$. В таком случае

$$b=\frac{\Sigma fa}{n}=\frac{-25}{168}=-0{,}15 \text{ см}.$$

Отсюда $\bar{x}=A+b=74+(-0{,}15)=73{,}85$ см.

В приведенном примере величина классового промежутка равна 1, поэтому ее можно не учитывать. Однако в тех случаях, когда $i\neq 1$, в формулу надо включать величину i, а именно

$$\bar{x}=A+b.i. \quad (4а).$$

Таблица 9

Распределение 25 кроликов по весу

Классы, см	Центральные значения классов, см	Частоты f	Отклонения a	fa	fa^2
3,0—3,4	3,2	1	—4	—4	16
3,5—3,9	3,7	1	—3	—3	9
4,0—4,4	4,2	3	—2	—6	12
4,5—4,9	4,7	3	—1	—3	3
5,0—5,4	5,2	7	0	0	0
5,5—5,9	5,7	5	1	5	5
6,0—6,4	6,2	3	2	6	12
6,5—6,9	6,7	1	3	3	9
7,0—7,4	7,2	1	4	4	16
$i=0{,}5$ кг		$n=25$		$\Sigma fa=-16+18=+2$	$\Sigma fa^2=82$

Вычисление следует производить в тех же условных отклонениях a, равных 1, 2, 3 или —1, —2, —3 и т. д., как это было сделано в табл. 8, но в дальнейшем полученное значение b надо умножить на величину i, как это показано в табл. 9.

По данным табл. 9 можно вычислить $\bar{x}$ с помощью A и b, а именно:

$$b = \frac{\Sigma fa}{n} = \frac{2}{25} = 0{,}08,$$

$$\bar{x} = A + b.i = 5{,}2 + 0{,}08.0{,}5 = 5{,}24 \text{ кг}.$$

Выше, в табл. 7, было получено несколько иное значение $\bar{x}$=5,29 кг. Причиной различия является то, что в табл. 7 и 9 одни и те же данные сведены в разное число классов (5 и 8). При применении прямого способа вычисления средней арифметической было получено значение $\bar{x}$=5,27 кг. Таким образом, при разных способах обработки материала могут быть получены несколько различные значения $\bar{x}$, однако различия выражаются лишь в сотых долях килограмма.

Следует иметь в виду, что в некоторых случаях целесообразно принимать A=0, тогда отклонения a превратятся в варианты x, и b будет равно $\frac{\Sigma x}{n} = \bar{x}$.

Средняя геометрическая. Средняя арифметическая является наиболее часто применяемым статистическим показателем, в том числе в биологии. Однако в некоторых случаях (например, при изучении темпов роста организмов или роста целых популяций) приходится пользоваться другой средней величиной—средней геометрической.

Формула для ее вычисления следующая:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \ldots x_n}. \qquad (5)$$

Очевидно, что при ее вычислении надо исключать варианты, выражающиеся нулем или отрицательным числом.

На практике вычисление средней геометрической производится с помощью логарифмов по следующей рабочей формуле:

$$\log G = \frac{1}{n}(\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n), \qquad (5a)$$

т. е. логарифм средней геометрической равен арифметической средней суммы логарифмов отдельных значений x. По значению $\log G$ затем определяется величина G.

Основным критерием для применения средней геометрической является то, что возрастание данного признака происходит не путем арифметического прибавления к первоначальному значению какой-то величины, а умножением пропорционально какой-то степени.

При таком характере возрастания значений x арифметическая средняя дает очень неточные результаты, и лучше пользоваться геометрической средней. Это значит, что надо заменить арифметические значения признака их логарифмами и оперировать в дальнейшем уже с ними. Так, например, если в пробах планктона были получены показатели от 428 до 43 300, то их следует перевести в логарифмы ($\log 428=2{,}63$; $\log 43\,300=4{,}64$ и т. д.), сложить все логарифмы и разделить на n. Полученное значение $\log G$ потенцируется, и таким образом получается значение средней геометрической, выраженное в конкретных значениях изучаемого признака.

Измерение вариации. Средняя арифметическая указывает на то, какое значение признака наиболее характерно для данной совокупности. Но она сама по себе еще недостаточна для характеристики совокупности, так как главной особенностью совокупности является наличие вариации между ее членами. Два стада коров могут иметь очень близкие средние удои, но в одном величины удоев могут сильно колебаться, в другом же коровы могут представлять собой довольно однородную группу с небольшим размахом колебаний по удоям. Учет степени колеблемости того или другого признака в совокупности имеет очень большое значение для биолога, так как всякая вариация в популяции животных или растений в конечном счете отражает различия между организмами—в их наследственной природе и в тех условиях, при которых они выращивались. Приемы работы с животными должны меняться в зависимости от характера их вариации. Сравнение лимитов может в известной степени указывать на разницу в степени изменчивости внутри вариационных рядов, но оно недостаточно. Во-первых, крайние величины в рядах не очень устойчивы, и при изменении количества изучаемых особей они лег-

ко сдвигаются. Во-вторых, при одних и тех же пределах вариации распределение отдельных вариант в рядах может быть различным. Иллюстрацией сказанного является распределение частот по классам в четырех вариационных рядах, представленных в табл. 10.

Таблица 10

Распределение частот по классам в четырех вариационных рядах

Классы	Частоты			
	ряд 1	ряд 2	ряд 3	ряд 4
1			1	
2			1	1
3			3	2
4	1	3	9	4
5	6	6	20	9
6	14	10	40	15
7	6	6	20	9
8	1	3	9	4
9			3	2
10			1	1
11			1	
	$n = 28$	$n = 28$	$n = 108$	$n = 47$

Ряды 1 и 2 имеют одинаковые значения крайних классов, но распределение частот в них различно. Ряды 3 и 4 очень близки друг к другу по характеру распределения частот, однако ряд 3 более растянут и охватывает больше классов.

Вот почему для характеристики колеблемости отдельных значений случайной переменной x, иначе говоря, вариации между членами совокупности, нужен такой показатель, который обобщал бы колеблемость всех вариант. Для этого надо сравнивать варианты или друг с другом или с какой-то одной постоянной величиной. В качестве последней лучше всего взять среднюю арифметическую. Мы уже видели, что каждое значение $x_1, x_2, x_3, \ldots, x_n$ в какой-то степени отличается от $\bar{x}$, т. е. отклоняется от него в плюс или минус сторону. Казалось бы, наиболее простым способом характеристики вариации в совокупности было бы сложить все значения $(x-\bar{x})$, т. е. получить сумму $(x-\bar{x})$ и разделить ее

на n. Но, согласно второму свойству средней арифметической, $\Sigma (x-\bar{x})=0$. На ранних этапах развития вариационной статистики пробовали складывать значения всех вариант без алгебраического знака и характеризовать вариацию с помощью среднего отклонения. Однако этот метод оказался очень упрощенным, так как среднее отклонение не улавливает истинной закономерности рассеяния вариант в совокупности, в вариационном ряду. Характер вариации или рассеяния вариант требует, чтобы использованная для характеристики вариации величина имела размерность второй степени. Были предложены в качестве мерила вариации—квадрат отклонений, иначе называемый вариансой или дисперсией, и среднее квадратическое отклонение, которое иногда называют также стандартным отклонением. Вариансу обозначают σ^2 (греческая буква сигма) или s^2 (латинская буква эс), а среднее квадратическое отклонение—σ или s. В специальной литературе греческие и латинские обозначения относят к различным типам совокупностей, в частности, в применении к конкретным выборкам часто пишут s^2 и s. Но так как в советской литературе является привычным обозначение среднего квадратического отклонения через σ, мы решили сохранить именно это обозначение.

Чтобы лучше представить, почему правильнее иметь дело с квадратическим отклонением, а не с простым отклонением, разберем простой пример о попадании в мишень пуль двух стрелков. Предположим, что средние отклонения от центра мишени пуль одного стрелка 2 см, а второго—4 см. С первого взгляда можно сделать вывод, что меткость первого стрелка только в 2 раза больше, нежели второго. Но в действительности степень попадания зависит не от расстояния от центра мишени, а от площади рассеяния пуль, т. е. изменяется пропорционально квадрату расстояния от центра. Квадрат отклонения пуль второго стрелка $4^2=16$, а первого $2^2=4$. Попасть в цель в пределах окружности радиусом в 4 см в четыре раза легче, чем в окружность радиусом в 2 см. В принципе близкие соображения лежали в основе положений об установлении в качестве мерила изменчивости вариансы.

Формула для варианты в общем виде следующая:

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}. \qquad (6)$$

Словами ее можно формулировать так: варианса — это сумма квадратов отклонений отдельных значений данной переменной от средней, деленная на число вариант. Иногда в литературе ее называют сокращенно средний квадрат, понимая под этим, что имеется в виду средний квадрат отклонений.

Отсюда формула для среднего квадратического отклонения, которое часто называют просто сигмой:

$$\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}}. \qquad (7)$$

Среднее квадратическое отклонение является числом именованным и будет выражаться в тех же измерениях, как и средняя арифметическая. При этом надо помнить, что она имеет 2 знака: плюс и минус, но их можно не писать. В случае, если общее число вариант мало (меньше 25—30), лучше применять формулы

$$\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n-1} \qquad (6а)$$

и

$$\sigma = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}} \qquad (7а)$$

Иногда также пишут

$$\sigma^2 = \frac{1}{n-1} \Sigma (x-\bar{x})^2. \qquad (6б)$$

Степени свободы. Величина $n-1$ получила особое название—число степеней свободы (точнее число степеней свободы вариации). Она обозначается в англо-американской литературе буквами *d.f.*, а в немецкой *FG*. Так как во многих разделах статистики приходится пользоваться числом степеней свободы, то следует объяснить его значение. Выше уже указывалось, что если известен ряд от x_1 до x_n, состоящий из n членов или наблюдений, то для него может быть вычислена характеристика в виде средней арифметической. Возникает вопрос, как может быть определено каждое отдельное

значение ряда. Очевидно, его всегда можно узнать, если известны средняя арифметическая и остальные наблюдения, т. е. $n-1$. Иначе говоря, определение одного значения в данной совокупности зависит от остальных значений. Так, например, если известно, что 2 кролика в сумме весят 6 кг, а один из них весит 2,5 кг, то вес второго уже точно определен весом первого, т. е. имеется лишь 1 степень свободы ($2-1=1$). Если 3 кролика весят 3 кг, то вес одного всегда точно определяется весом двух других, между которыми уже возможна вариация, т. е. в этом случае имеются 2 степени свободы ($3-1=2$) и т. д. В общем виде при численности членов совокупности n число степеней свободы $d.f.=n-1$. Вот почему точнее вычислять σ^2 и σ, пользуясь знаменателем $n-1$. При большом n разница между n и $n-1$ настолько невелика, что она мало отразится на значении вариансы (и сигмы). Но при малом n разница будет значительна. Так, если $n=6$, а сумма квадратов равна 60, то средний квадрат отклонений от средней арифметической будет равен не $\frac{60}{6}=10{,}0$, а $\frac{60}{5}=12{,}5$. Поэтому надо разделить сумму квадратов на число степеней свободы, т. е. на $n-1=5$.

В некоторых случаях, как это будет видно в дальнейшем, число степеней вычисляется более сложно.

Различные способы вычисления вариансы и среднего квадратического отклонения. Если отдельные варианты данной совокупности по тем или иным причинам не сгруппированы в вариационный ряд, например при малой численности опытных животных, то можно вычислить σ^2 и σ путем прямого применения указанных выше формул (6) и (7) или (6а) и (7а). Тогда целесообразно составить подсобную табличку, в которую должны быть записаны значения всех вариант, как это показано в табл. 11.

В этом примере в знаменателе для вычисления вариансы и сигмы лучше взять $n-1$, а не n.
Тогда

$$\sigma^2=\frac{0{,}28}{7}=0{,}04, \text{ а } \sigma=\sqrt{0{,}04}=0{,}2\ \%\ \text{жира}.$$

Если взять знаменатель n, то величина вариансы

(и сигмы) окажется несколько заниженной:

$$\sigma^2 = \frac{0{,}28}{8} = 0{,}0350 \text{ и } \sigma = \sqrt{0{,}0350} = 0{,}18 \text{ \% жира.}$$

Таблица 11

Вариация процента жира в молоке 8 опытных коров

Процент жира x	$x - \bar{x}$	$(x - \bar{x})^2$
4,1	0,3	0,09
3,8	0	0
3,5	0,3	0,09
4,0	0,2	0,04
3,9	0,1	0,01
3,8	0	0
3,7	0,1	0,01
3,6	0,2	0,04
$\Sigma x = 30{,}4$		$\Sigma (x - \bar{x})^2 = 0{,}28$

Исходным для вычисления вариансы и среднего квадратического отклонения является сумма квадратов отклонений от $\bar{x}$, или просто «сумма квадратов». Сумма квадратов и средний квадрат—это две важнейшие величины, с которыми приходится встречаться в очень многих вычислениях. Поэтому необходимо хорошо уяснить их смысл. В дальнейшем все формулы будут построены на сумме квадратов.

Средняя арифметическая $\bar{x}$ часто выражается числом с десятичной дробью, имеющей несколько десятичных знаков. Отклонения от $\bar{x}$ отдельных вариант, т. е. $(x - \bar{x})$, будут также дробными величинами, возведение которых в квадрат затруднит вычисления. Для облегчения вычислительной работы можно прибегнуть к непрямому способу вычисления вариансы и среднего квадратического отклонения с помощью так называемой условной средней, уже использовавшейся для вычисления средней арифметической.

Среди свойств средней арифметической было одно, имеющее прямое отношение к непрямому способу вычисления вариансы, а именно, что сумма квадратов отклонений от средней арифметической меньше суммы квадратов отклонений от любой другой величины A, не равной $\bar{x}$.

Значит, сумма квадратов отклонений от условной величины A всегда больше суммы квадратов отклонений от $\bar{x}$, при этом на определенную величину, а именно на величину $(A-\bar{x})^2$, умноженную на n, т. е.

$$\Sigma (x-A)^2 = \Sigma (x-\bar{x})^2 + n (A-\bar{x})^2 .$$

Отсюда формула для среднего квадрата отклонений от средней арифметической:

$$\sigma^2 = \frac{\Sigma (x-\bar{x})^2}{n} = \frac{\Sigma (x-A)^2}{n} (-A-\bar{x})^2 , \qquad (8)$$

а

$$\sigma = \sqrt{\frac{\Sigma (x-A)^2}{n} - (A-\bar{x})^2} . \qquad (9)$$

При малом n лучше:

$$\sigma^2 = \frac{\Sigma (x-A)^2 - n (A-\bar{x})^2}{n-1} \qquad (8а)$$

Эта формула в случае, если варианты уже сгруппированы в вариационный ряд с частотами для каждого класса f, может быть записана и так:

$$\sigma^2 = \frac{\Sigma f (x-\bar{x})^2}{n} = \frac{\Sigma f (x-A)^2}{n} - (A-\bar{x})^2 . \qquad (8б)$$

Применять эту формулу очень легко на примере, приведенном в табл. 9. Если вспомнить, что $x-A$ мы обозначили через a, а $A-\bar{x}$ через b *, то формула с учетом величины i для среднего квадратического отклонения будет следующей:

$$\sigma = i \cdot \sqrt{\frac{\Sigma f a^2}{n} - b^2} . \qquad (9а)$$

А так как $$b = \frac{\Sigma fa}{n} ,$$

то $$\sigma = i \cdot \sqrt{\frac{\Sigma fa^2}{n} - \left(\frac{\Sigma fa}{n}\right)^2} . \qquad (9б)$$

В табл. 8 и 9 были введены для вычисления сигмы графы fa^2. Можно сейчас воспользоваться этими

* При условии, что $i = 1$. Если же $i \neq 1$, то $x - A = ai$, а $A-\bar{x} = bi$.

данными. Подставив в формулу (9*а*) все необходимые значения из табл. 8, мы получим:

$$\sigma = \sqrt{\frac{1011}{168} - (-0{,}15)^2} = 2{,}45 \text{ см}.$$

Таким образом, среднее квадратическое отклонение по промерам глубины груди для изученной группы симменталов равно 2,45 см. Варианса в данном случае определяется как подкоренное значение, т. е. σ^2=5,9954, или, округляя до второго знака, 6,00.

При вычислении сигмы по данным табл. 9 необходимо учесть величину классового промежутка, равную 0,5.

$$\sigma = 0{,}5 \cdot \sqrt{\frac{82}{25} - 0{,}08^2} = 0{,}90 \text{ кг}.$$

Таблица 12

Вычисление $\bar{x}$ и σ для данных о весе при рождении 20 морских свинок

Веса при рождении (в г) x	Отклонения от A $x - A$	Квадраты отклонений $(x - A)^2$
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
Всего:	+5	247

Варианту вычисляем обратным путем, т. е. путем возведения в квадрат вычисленной сигмы. Тогда

$$\sigma^2 = 0{,}90^2 = 0{,}81.$$

Варианса для данных, несгруппированных в вариационный ряд. Способ вычисления варианты и среднего квадратического отклонения с помощью условной средней может быть применен и тогда, когда данные не сведены в вариационный ряд.

В табл. 12 приведены 20 значений живого веса при рождении морских свинок (из пометов с 2 детенышами) и значения $\Sigma(x-A)$ и $\Sigma(x-A)^2$.

Тогда

$$\Sigma(x-A) = 5,$$

$$b = \frac{\Sigma(x-A)}{n} = \frac{5}{20} = 0{,}25,$$

$$\bar{x} = A + b = 30{,}0 + 0{,}25 = 30{,}25 \text{ г}.$$

Варианса (с учетом малого значения n):

$$\sigma^2 = \frac{\Sigma(x-A)^2 - n(A-\bar{x})^2}{n-1} =$$

$$= \frac{247 - 20\ (0{,}25)^2}{19} = \frac{245{,}75}{19} = 12{,}93,$$

а

$$\sigma = \sqrt{12{,}93} = 3{,}60 \text{ г}.$$

Некоторые преобразования формулы варианты. Указанная выше формула для варианты, в которой использована условная средняя, может приобрести и иной вид, если в качестве A будет принят 0. Тогда

$$\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \bar{x}^2.$$

Иначе говоря, можно получить средний квадрат отклонений от средней арифметической, вычислив средний квадрат отклонений от нуля, т. е. средний квадрат вариант, разделив на число степеней свободы и вычтя из этой величины квадрат средней арифметической. Такое вычисление варианты можно проделать и на основе данных табл. 11. Для этого понадобится лишь графа x^2. Ее

легко получить с помощью таблицы квадратов или арифмометра.

Учитывая, что $n=8$, для вычисления σ^2 (и соответственно σ) знаменателем надо взять число степеней свободы $n-1=7$.

Дальнейшее преобразование формулы для вариансы позволяет еще более упростить схему вычислений, сделав ее, в частности, очень удобной для машинных вычислений, а именно:

$$\sigma^2 = \frac{\Sigma x^2}{n} - \bar{x}^2,$$

$$\sigma^2 = \frac{\Sigma x^2 - n \cdot \bar{x} \cdot \bar{x}}{n};$$

так как $$n \cdot \bar{x} = \Sigma x,$$

то $$\sigma^2 = \frac{\Sigma x^2 - \Sigma x \cdot \bar{x}}{n} \qquad (10)$$

или точнее $$\sigma^2 = \frac{\Sigma x^2 - \Sigma x \cdot \bar{x}}{n-1}. \qquad (10a)$$

В этом случае для определения вариансы нужны только: сумма всех вариант (Σx), сумма квадратов всех вариант (Σx^2), средняя арифметическая ($\bar{x}$) и число вариант или наблюдений (n).

Но так как $\bar{x} = \frac{\Sigma x}{n}$, то, подставив значение $\bar{x}$ в предыдущую формулу, можно получить и такую формулу:

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}. \qquad (10б)$$

В зависимости от того, какие данные следует обработать и какие технические возможности имеются для проведения вычислений, можно применять любую из этих формул. Результаты будут одинаковыми. Так, например, данные табл. 12 по весу при рождении морских свинок могут быть обработаны с помощью счетной машины на основе следующих исходных данных:

$$n = 20;\ \Sigma x = 605;\ \Sigma x^2 = 18\,547.$$

Тогда $$\bar{x} = \frac{\Sigma x}{n} = \frac{605}{20} = 30{,}25 \text{ г.},$$

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{18547 - \frac{605^2}{20}}{19} =$$

$$= \frac{245{,}75}{19} = 12{,}93.$$

Легко видеть, что в числителе для σ^2 получено по этой формуле то же число 245,75, как и при применении формул (8) и (8а) С помощью различных формул вычисляется в конечном счете одна и та же величина в числителе для σ^2, а именно $\Sigma(x-\bar{x})^2$, т е сумма квадратов отклонении от средней арифметической Так как сумма квадратов отклонений представляет собои важнейшую величину в целом ряде разделов вариационной статистики, небесполезно дать сводку различных ее значений, вытекающих из указанных выше формул

$$\Sigma (x-\bar{x})^2 = \Sigma f \cdot (x - \bar{x})^2$$

$$\Sigma f \cdot (x - A)^2 - n \ (A - \bar{x})^2$$

$$\Sigma x^2 - n \cdot \bar{x}^2$$

$$\Sigma x^2 - \Sigma x \cdot \bar{x}$$

$$\Sigma x^2 - \frac{(\Sigma x)^2}{n}.$$

Сумму квадратов отклонений, указанную в левои части ее, можно обозначить каким либо условным символом, например SQ, как это применяется в немецкой литературе, однако в советской и англо американской литературе предпочитают писать словами «сумма квадратов», а *варианс*у *обозначают также словами «средний* ьквадрат»

Взвешенная варианса. Выше была приведена формула (3) для вычисления средневзвешенной средней арифметической, т е средней арифметической при объединении нескольких рядов (совокупностей), каждый из которых имеет разные n и $\bar{x}$ При объединении нескольких совокупностей может быть вычислена и общая варианса

по следующей формуле

$$\sigma_g^2 = \frac{\sigma_1^2 (n_1-1) + \sigma_2^2 (n_2-1) + \sigma_3^2 (n_3-1) + \ldots + \sigma_k^2 (n_k-1)}{n-k}, \quad (11)$$

где n_1, n_2, n_3, ..., n_k — численности отдельных совокупностей или рядов (всего их k), σ_1^2, σ_2^2, σ_3^2, ..., σ_k^2 — варианты отдельных рядов.

Сумма $n_1 + n_2 + n_3 + \ldots + n_k$ равна n.

Так, например, если для трех групп:

$$n_1 = 6 \qquad \sigma_1^2 = 4$$
$$n_2 = 10 \qquad \sigma_2^2 = 3$$
$$n_3 = 18 \qquad \sigma_3^2 = 2,$$

$$\text{то } \sigma_g^2 = \frac{4.5 + 3.9 + 2.17}{34-3} = 2{,}61,$$

$$\text{а} \quad \sigma_g = \sqrt{2{,}61} = 1{,}6.$$

Существенно, что знаменатель представляет собой сумму числа степеней свободы трех отдельных групп n_1, n_2 и n_3, которая в данном случае равняется $n-3$.

Закон сложения вариации. Всякая изучаемая совокупность может состоять из нескольких, более частных совокупностей или групп. Соответственно этому при изучении изменчивости мы встречаемся с вариансами отдельных групп, вариансами средних показателей этих групп и, наконец, вариансой объединенной совокупности. Существует общий закон сложения вариации ряда групп в общей совокупности, который может быть записан в виде следующей формулы:

$$\sigma_0^2 = \sigma_I^2 + \overline{\sigma^2}, \quad (12)$$

где σ_0^2 — общая варианса, σ_I^2 — варианса групповых средних, $\overline{\sigma^2}$ — варианса внутри групп, т. е. средняя характеристика индивидуальной внутригрупповой изменчивости, иначе называемая остаточной вариацией.

Та же формула в более общем виде:

$$\sigma_0^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots + \overline{\sigma^2}. \quad (12a)$$

При этом надо помнить, что общая варианса представляет собой не просто сумму частных вариант, а

средневзвешенную из частных вариансс с весами, равными численностям отдельных групп. Сумма же квадратов отклонений для всей группы просто равна сумме квадратов отклонений отдельных анализируемых групп, включая и индивидуальную вариацию внутри групп.

Эти общие положения лежат в основе особого метода, т. н. дисперсионного (или вариансного) анализа. Дисперсионный анализ имеет в современной вариационной статистике громадное значение, так как с его помощью могут быть тщательно проанализированы результаты наблюдений и опытов по изучению влияния различных факторов на те или иные признаки или биологические свойства организмов. Поэтому он подлежит отдельному рассмотрению.

Для иллюстрации приведем простейший пример на биологическом материале.

В течение 20 дней изучали прирост в весе 20 свиней, распределенных в 4 группы по 5 свиней в каждой. В результате были получены данные, сведенные в табл. 13.

Таблица 13

Привесы 20 свиней (в фунтах)

Обозначения сводных величин и показателей	Группы				В целом
	1	2	3	4	
	40	29	11	17	
	24	27	31	21	
	20	20	17	28	
	46	39	37	33	
	35	45	39	21	
Σx	165	160	135	120	580
$\bar{x}$	33	32	27	24	29
Σx^2	5917	5516	4261	3044	18738
$\frac{(\Sigma x)^2}{n}$	5445	5120	3645	2880	16820
Сумма квадратов $[\Sigma (x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}]$	472	396	616	164	1918
		Σ — 1648			

Исходя из общего положения о сложении вариации,

можно записать данные по анализу вариансы привесов свиней в табл. 14.

Таблица 14

Анализ вариансы данных привесов свиней, приведенных в табл. 13

Источники вариации	Степени свободы	Сумма квадратов	Средний квадрат
Общая	19	1918	100,9
Между особями различных групп	16	1648	103,0
Между средними групп	3	270	90,0

Сумма квадратов для общей изменчивости имеется в табл. 13 (последняя цифра в правой колонке). Сумму же квадратов отклонений между особями различных групп можно взять как сумму четырех сумм квадратов, т. е. 472+396+616+164=1648. Сумма квадратов отклонений между средними групп получается путем простого вычитания. Вычитанием же получится и число степеней свободы для последней строчки. Поскольку групп 4, число степеней свободы должно быть 3. Можно проверить и правильность среднего квадрата, т. е вариансы для суммарных данных как средней взвешенной, а именно:

$$\sigma^2 = \frac{103 \cdot 16 + 90 \cdot 3}{19} = 100{,}9.$$

Обычно при дисперсионном анализе применяется несколько иной способ расчетов, хотя в принципе он не отличается от изложенного в примере со свиньями.

Коэффициент изменчивости. Среднее квадратическое отклонение (σ) выражается в тех же единицах, что и $\bar{x}$, например, при измерении веса кроликов в кг и г, при измерении глубины груди крупного рогатого скота в см. Поэтому сравнивать изменчивость разных групп животных можно только в отношении одного и того же признака. Если же одна сигма выражена в см, а другая в кг, судить о том, в каком случае вариация больше, а в каком меньше, нет возможности. Для сравнения изменчивости различных признаков (а также изменчивости групп животных разных видов) применяют так называе-

мый коэффициент изменчивости или коэффициент вариации (обозначается *с. v.* или просто *C*).

Коэффициент изменчивости представляет собой отношение σ к $\bar{x}$, выраженное в процентах, иначе говоря, он показывает, какой процент от $\bar{x}$ составляет σ:

$$c.\ v.\ = \frac{\sigma \cdot 100}{\bar{x}}. \qquad (13)$$

Можно применить эту формулу к вариационным рядам распределения кроликов и скота по различным признакам, а именно: кроликов—по живому весу, крупного рогатого скота—по промерам груди, коров—по жирности молока.

Для первого из них $\bar{x} = 5{,}24$ кг; $\sigma = 0{,}90$ кг;

$$c.\ v.\ = \frac{0{,}90 \cdot 100}{5{,}24} = 17{,}1\ \%.$$

Для второго $\bar{x} = 73{,}85$ см; $\sigma = 2{,}45$ см;

$$c.\ v.\ = \frac{2{,}45 \cdot 100}{73{,}85} = 3{,}3\ \%.$$

Для третьего $\bar{x} = 3{,}8\ \%$ жира; $\sigma = 0{,}18\ \%$ жира;

$$c.\ v.\ = \frac{0{,}18 \cdot 100}{3{,}8} = 4{,}7\ \%.$$

Наименьший коэффициент изменчивости характеризует данные по промерам груди. Очень высокий коэффициент изменчивости получен для данных по весу 25 кроликов, что неудивительно, так как в этой группе были взяты кролики разных пород. Чем более однороден изучаемый материал (по происхождению, условиям выращивания и т. д.), тем меньшими окажутся коэффициенты изменчивости. Однако даже при достаточной однородности материала степень изменчивости различных признаков может быть различной, что зависит от особенностей самих признаков. Известно, например, что жирность молока—признак значительно менее изменчивый, нежели удой за лактацию. Если в стаде коров показатели по удою за лактацию: $\bar{x}$=3000 кг, а σ=400 кг, по жирности же молока $\bar{x}$=3,8% и σ=0,24%, то соответствую-

щие коэффициенты изменчивости будут следующими:

$$c.\ v.\ \text{по удою равен}\ 13{,}3\%\left(\frac{400\cdot 100}{3000}\right),$$

$$c.\ v.\ \text{по жирности молока равен}\ 6{,}3\%\left(\frac{0{,}24\cdot 100}{3{,}8}\right).$$

Таким образом, коэффициент изменчивости дает возможность сравнивать изменчивость признаков, выражающихся в различных единицах измерения, и устанавливать различие в степени изменчивости. Для биолога, животновода, растениевода очень важно знать, насколько изучаемый ими материал выравнен или, наоборот, разнороден, в какой степени устойчивы взятые для сравнения признаки.

В частности, это важно при планировании опытов, установлении величины необходимых опытных групп, а также при оценке результатов опытов. Так, если было ранее установлено, что изменчивость изучаемых признаков колеблется в пределах 10—15%, а в опыте были получены данные, выходящие за эти пределы, то искать допущенную ошибку нужно или в самой постановке опытов, или в вычислениях, или, наконец, предположить, что какое-то непредвиденное обстоятельство повлияло на степень точности опытов.

Известно, что средняя арифметическая и среднее квадратическое отклонение некоторых признаков изменяются более или менее параллельно в зависимости от возраста, сезона года и других причин.

Величина σ также может иногда увеличиваться и в связи с увеличением самой $\bar{x}$. В таких случаях удобно пользоваться коэффициентом изменчивости, так как он оказывается более устойчивой величиной. Как указано, $c.\ v.$ определяется на основе уже известных x и σ. Но, зная $c.\ v.$ и $\bar{x}$, можно определить σ. Если известно, например, что $c.\ v.$ для веса крыс возраста 56—84 дней был равен 13%, а $\bar{x}$=200 г, то $\sigma = \frac{(c.\ v.)\cdot \bar{x}}{100} =$

$$= \frac{13\cdot 200}{100} = 26\ \text{г}.$$

Однако один коэффициент изменчивости явно недостаточен для характеристики совокупности. Он является

лишь дополнительным показателем, полезным при наличии $\bar{x}$ и σ или σ^2. Показателями же, действительно характеризующими всякую совокупность значений случайной переменной x, являются $\bar{x}$, σ^2 и σ. Они дают возможность, не имея самой совокупности, как бы построить ее, так как $\bar{x}$ указывает на наиболее типичное значение x, около которого сосредоточивается большинство вариант, а σ и σ^2 измеряют изменчивость. Поэтому их можно назвать на математическом языке параметрами совокупности.

Средняя арифметическая и варианса при альтернативной изменчивости. Выше уже указывалось, что при альтернативной, или качественной, изменчивости группировка данных сводится к подсчету количества особей, относящихся к каждой качественной группе и к выражению этого количества в процентах к общему объему совокупности.

В общем виде данные при альтернативной изменчивости могут быть представлены в виде двух классов: класса «0», охватывающего варианты с отсутствием данного признака, и «1»—с присутствием его. Сокращенный вариационный ряд, состоящий только из двух классов, также можно обработать, подобно ряду количественной изменчивости (табл. 15).

Таблица 15

Общая схема обработки ряда при качественной изменчивости

Классы	Частоты f	Отклонения от условной средней a	fa	fa^2
0	p_0	0	0	0
1	p_1	1	p_1	p_1
	$p_0+p_1=n$		$\Sigma = p_1$	$\Sigma = p_1$

Применив обычные формулы

$$x = A + b = A + \frac{\Sigma fa}{n}$$

и $$\sigma = \sqrt{\frac{\Sigma fa^2}{n} - b^2},$$

получим

$$\bar{x} = 0 + \frac{p_1}{n} = \frac{p_1}{n}. \qquad (14)$$

Так как это фактически доля определенного качественного класса в общей совокупности, можно писать вместо $\bar{x}$ букву p, т. е.

$$p = \frac{p_1}{n}. \qquad (14a)$$

Среднее же квадратическое отклонение

$$\sigma = \sqrt{\frac{p_1}{n} - \left(\frac{p_1}{n}\right)^2}.$$

Но так как $n = p_0 + p_1$, то подкоренную величину можно преобразовать следующим образом:

$$\sigma = \sqrt{\frac{np_1}{n^2} - \frac{p_1^2}{n^2}} = \sqrt{\frac{(p_0 + p_1)\,p_1 - p_1^2}{n^2}} = \sqrt{\frac{p_0 \cdot p_1}{n^2}}. \qquad (15)$$

Отношение $\frac{p_1}{n} = p$, а отношение $\frac{p_0}{n} = 1 - p = q$.

Тогда

$$\sigma = \sqrt{p\,(1-p)}, \qquad (15a)$$

а поскольку $1 - p = q$, то $\sigma = \sqrt{p\,q}$. $\qquad (15б)$

Применим указанные формулы к данным о 284 коровах, которые были подвергнуты туберкулинизации. Отрицательную реакцию дала 201 корова, положительную — 83. Эти данные можно внести в табл. 16.

Таблица 16

Распределение коров по реакции на туберкулез

Классы	Частоты f	a	fa	fa^2
0	201	0	0	0
1	83	1	83	83
	$n = 284$		$\Sigma = 83$	$\Sigma = 83$

В таком случае

$$p = \frac{83}{284} = 0{,}29 \text{ (или } 29\%)$$

и $\sigma = \sqrt{0{,}29 - 0{,}0841} = \sqrt{0{,}2059} = 0{,}45$ (или 45%).

Соответственно $\sigma^2 = 0{,}21$ (или 21%).

Применив формулу (15б), получим ту же величину

$$\sigma = \sqrt{0{,}29 \cdot 0{,}71} = 0{,}45 \text{ (или } 45\%).$$

С помощью указанных формул значения $\bar{x}$ и σ выражаются в долях единицы или в процентах. Но бывает необходимость выразить их в абсолютных величинах. Тогда

а) $\bar{x} = n\frac{p_1}{n} = np = p_1 = 83$,

б) $\sigma = \sqrt{npq} = \sqrt{284 \cdot 0{,}29 \cdot 0{,}71} = 7{,}3$.

ВОПРОСЫ

1 Как характеризовать структуру совокупности при качественных различиях между вариантами?

2 Какой класс является модальным?

3 Что такое медиана?

4 Основная формула средней арифметической

5 Могут ли совпасть значения $\bar{x}$, Мо и Ме?

6 Взвешенная средняя арифметическая для сложной совокупности

7 Свойства средней арифметической

8 В чем заключается непрямой способ вычисления $\bar{x}$?

9 В каких случаях целесообразно пользоваться средней геометрической? Формула средней геометрической и ее преобразование с помощью логарифмов

10 Среднее квадратическое отклонение как мерило изменчивости совокупности

11 Что такое варианса?

12 Методы вычисления суммы квадратов отклонений с помощью различных формул. Вычисление вариансы для данных, несгруппированных в вариационный ряд

13 Степень свободы и значение этого показателя при вычислении σ^2 и σ. При каких значениях n более точным является использование числа степеней свободы, а не количества вариант (наблюдений)?

14 Можно ли приравнять условную среднюю A к нулю? Каков вид тогда примут формулы для вычисления $\bar{x}$ и σ^2?

15 Как вычисляется взвешенная варианса? Определение числа степеней свободы для объединенной совокупности

16. В чем заключается закон сложения вариации?

17. Какая разница между σ и $c.v.$? В каких случаях важно использование $c.v.$?

18. Почему $\bar{x}$ и σ являются основными характеристиками вариационного ряда?

19. Методы вычисления $\bar{x}$ и σ при альтернативной изменчивости.

ЗАДАЧИ*)

9. Вычислите $\bar{x}$, σ^2 и σ, пользуясь методом условной средней, для вариационного ряда задачи 1.

10. То же, для задачи 2.

11. То же, для задачи 3.

12. То же, для задачи 4.

13. То же, для задачи 5.

14. То же, для задачи 6.

15. То же, для задачи 7.

16. Было сделано 5 определений содержания кальция в крови (в усл. единицах): 11,27; 11,36; 11,09; 11,16; 11,47. Вычислите $\bar{x}$, σ^2 и σ.

17. Живой вес при рождении 11 поросят (в кг) был следующий: 1,2; 1,1; 1,3; 0,9, 1,4; 1,0; 1,5; 1,3; 1,2; 1,4; 1,0 Вычислите $\bar{x}$, σ^2 и σ. Какую формулу для вычисления среднего квадрата удобнее применить?

18. Вес цыплят белых леггорнов (в г) за 2 месяца был следующим: 1-я неделя—62,7; 2-я—121,4; 3-я—193,0; 4-я—380,0; 5-я—481,0; 6-я—504,0; 7-я—719,0 и 8-я неделя—759,0. Определите, насколько увеличивался вес по неделям, и после этого вычислите средний привес по формуле средней геометрической.

19. У 1060 студентов исследовали биение пульса. Колебания были от 43 до 108 ударов в минуту. Данные были сгруппированы в следующий вариационный ряд ($i = 4$):

Классы	Частоты
43—46	1
47—50	2
51—54	6
55—58	22
59—62	52
63—66	79
67—70	118
71—74	165
75—78	186
79—82	165
83—86	103
87—90	82
91—94	45
95—98	19
99—102	11
103—106	3
107—110	1
	1060

*) Для удобства пользования задачами им дана сплошная нумерация независимо от глав, где они помещены.

Вычислите $\bar{x}$ и σ методом условной средней. Постройте гистограмму. Уменьшите число классов вдвое, приняв $i = 8$, и вычислите $\bar{x}$ и σ. Насколько изменились результаты при увеличении размеров классов?

20. Были получены следующие средние арифметические для пяти групп телок:

$\bar{x}_1 = 262$ кг $(n = 10)$; $\bar{x}_2 = 238$ кг $(n = 3)$; $\bar{x}_3 = 260{,}5$ кг $(n = 7)$; $\bar{x}_4 = 275$ кг $(n = 15)$ и $\bar{x}_5 = 255{,}4$ кг $(n = 5)$.

Вычислите взвешенную среднюю арифметическую. Насколько взвешенная $\bar{x}$ будет отличаться от $\bar{x}$, полученной без учета весов отдельных групп телят?

21. У 48 коров холмогорских помесей соски были коническими, у 12—цилиндрическими. Определите $\bar{x}$ и σ для числа цилиндрических сосков (в долях единицы и в абсолютных величинах).

22. Были установлены следующие показатели высоты в холке (в см):

	$\bar{x}$	σ
для телят	60	3
для молодых коров	100	5

Отличаются ли они по изменчивости?

23. Применили три разных метода определения хлорофилла на выборках из 12 листьев растений, при этом получили следующие статистические показатели (в мг):

$\bar{x}_1 = 61{,}4$ $\quad \sigma_1 = 5{,}22$
$\bar{x}_2 = 337$ $\quad \sigma_2 = 31{,}2$
$\bar{x}_3 = 13{,}71$ $\quad \sigma_3 = 1{,}2$

Сравните коэффициенты изменчивости при разных методах и сделайте выводы.

24. При изучении роста лабораторных крыс *с. v.* веса самцов 56—84-дневного возраста был примерно 13%, а $\bar{x} = 200$ г. Чему равны среднее квадратическое отклонение и варианса веса крыс?

25. Было установлено, что в полевых опытах с пшеницей коэффициент изменчивости урожая с га около 5%. Будет ли неожиданным, если при среднем урожае 25 ц с га среднее квадратическое отклонение окажется около 0,5 ц.

26. Было установлено, что в группе свиней средняя скорость роста составляла 1,4 фунта в день. Определите σ, если известно, что *с. v.* $\cong 10\%$.

Глава 3

ЗАКОНОМЕРНОСТИ СЛУЧАЙНОЙ ВАРИАЦИИ

Вероятность и ее исчисление. Основной особенностью вариационной статистики является то, что она имеет дело не с единичными явлениями или объектами, а с их совокупностями. Отдельные члены совокупности, как правило, в той или другой степени отличаются друг от друга, варьируют. Каждый из них представляет собой как бы отдельный случай, который осуществляется под влиянием многих определяющих условий. Однако этих причин может быть так много и они группируются в столь сложные сочетания, что обнаружение их для каждого отдельного случая становится невозможным. Приходится говорить только об известной возможности или вероятности значения, которое приобретает тот или иной член изучаемой совокупности. Для зоотехника, работающего с крупным рогатым скотом холмогорской породы, ясно, что возможность или вероятность найти корову холмогорку с жирностью молока 3,5% очень велика, но встретить корову с жирностью 4,5% —маловероятно. Растениевод хорошо знает, что случаи появления двух зародышей в семени ржи редки, поэтому вероятность найти такое семя в группе семян ржи мала. Известно, что у каждого вида млекопитающих рождается примерно одинаковое число самок и самцов. Поэтому вероятность рождения в семье мальчика или девочки достаточно велика и примерно одинакова. В этих примерах оценка вероятности как возможности кажется очень ясной и понятной. Но так бывает далеко не всегда. Возьмем такой пример. Ветеринарный врач применил для лечения заболевших какой-то болезнью лисиц в совхозе новое лекарство. Выяснилось, что из лисиц, получивших лекарст-

во, погибло только 4%, а не получивших —отход был равен 13%. От чего же зависит разница в проценте отхода, от того ли, что в опытной группе применили новое лекарство, или, может быть, лекарство никакой роли не сыграло. Когда нет точной уверенности в правильности того или другого суждения, часто употребляют слово «вероятно». Прибавкой к нему слов «очень» или «мало» выражают степень уверенности. Сторонник применения лекарства может сказать: «Очень вероятно, что именно благодаря применению лекарства отход в опытной группе был меньше». Но настроенный скептически противник лекарства может утверждать обратное: «Если бы и не давали лекарства, все равно в одной группе лисиц отход был бы больше, а в другой меньше в силу других причин, создавших разницу между группами».

Эти примеры показывают, что в практике биолога встречаются вопросы, дать ответы на которые можно только зная некоторые, хотя бы самые элементарные положения вариационной статистики и лежащей в ее основе теории вероятности.

Что же такое вероятность? Это возможность осуществления определенного события в некотором количестве случаев из общего числа возможных, или иначе говоря, степень уверенности в том, что событие наступит.

Исходным в понятии вероятности является понятие равновозможности, на основе которого можно отделить необходимые явления от случайных. Так, если при осуществлении события A возможно только событие B, такую связь явлений можно назвать необходимой. Если же при A равновозможны и B и C, мы имеем дело с проявлением возможности в виде случайного. Случайное— это такое же объективное явление, как и необходимое, и оно также обусловлено различными причинами, как и необходимое, только характер причинности здесь иной, а именно, возможен не один, а два или более результатов. Эти возможности и являются вероятностями. Процесс осуществления явления на основе известной его возможности или вероятности называется вероятностным или стохастическим. Теория вероятности и изучает математические законы таких процессов.

Вероятность можно выразить математически по следующей формуле: $p = \frac{F}{S}$, где F—число благоприятных

случаев, а S—число всех возможных или правильнее равновозможных случаев.

Так, если на каждой из сторон кубика написаны цифры 1, 2, 3, 4, 5, 6, то вероятность того, что наверху будет цифра 4, равна $\frac{1}{6}$, ибо всех возможных положений кубика может быть шесть и лишь один случай благоприятный. Значительно сложнее рассчитать вероятность того, что при выбрасывании двух кубиков сумма цифр наверху равна 6. Для этого следует рассчитать все возможные случаи сочетания цифр в двух кубиках:

1 + 1	2 + 1	3 + 1	4 + 1	5 + 1	6 + 1
1 + 2	2 + 2	3 + 2	4 + 2	5 + 2	6 + 2
1 + 3	2 + 3	3 + 3	4 + 3	5 + 3	6 + 3
1 + 4	2 + 4	3 + 4	4 + 4	5 + 4	6 + 4
1 + 5	2 + 5	3 + 5	4 + 5	5 + 5	6 + 5
1 + 6	2 + 6	3 + 6	4 + 6	5 + 6	6 + 5

Таким образом, возможно 36 случаев (S=36). Благоприятных же случаев, когда цифры двух кубиков дают в сумме число 6, будет только 5 (они подчеркнуты). Значит вероятность выбрасывания 2 кубиков с суммой цифр наверху, равной 6, может быть выражена по формуле $p = \frac{F}{S}$:

$$p = \frac{5}{36}.$$

Эти примеры являются теоретическими. На подобных примерах или моделях решаются многие задачи по теории вероятностей. Но легко дать немало и биологических примеров осуществления событий на основе той *или иной* вероятности. Выше уже указывалось, что особи того или другого пола у очень многих видов животных рождаются в примерно равном количестве. Это значит, что на каждые 100 потомков в среднем должно родиться 50 самок и 50 самцов, отсюда вероятность рождения от коровы телочки или бычка равна $p = \frac{50}{100} = 0{,}5$ (или 50%).

Другой биологический пример. Чтобы оценить вероятность рождения комолого теленка, надо знать количество *рождавшихся* ранее в данном стаде или породе комолых и рогатых животных. Так, если в данной породе было обнаружено за несколько последних лет 110 комолых телят среди общего количества 55 000 родивших-

ся телят, то вероятность рождения от коровы данной породы комолого теленка равна $p=\frac{110}{55000}=0{,}002$. Это значит, что в среднем на каждые 1000 случаев будет только 2 случая рождения комолых телят. На этом же примере легко понять и другую вероятность, как бы обратную величине p, что родится не комолый, а рогатый теленок. Это значение вероятности обозначается другой буквой q. Оно выражается в данном случае величиной, равной 0,998. Алгебраически сумма величин p и q равна 1, т. е. сумма вероятностей противоположных событий равна единице.

Таким образом, по относительной частоте определенной категории животных можно судить о ее вероятности.

Приведенные примеры показывают, что вероятности p и соответствующие им q могут иметь самые разные значения—от величин, близких к нулю или равных ему, до величин, близких к единице или равных ей. В нашем примере вероятность рождения комолого теленка очень мала. Однако существует немало событий, обладающих еще меньшей вероятностью. Наконец, если $p=0$, то на появление данного события вообще нельзя рассчитывать.

Но могут быть события, вероятность которых, хотя и очень близка к нулю, но все же нулю не равняется. Практически можно утверждать, что вероятность обнаружения в стаде холмогорской породы коровы с жирностью молока 6,5% равна нулю, но возможен все же какой-то исключительный случай, когда в силу своеобразных физиологических условий холмогорская корова может дать молоко исключительно высокой жирности. Очевидно, вероятность подобного события может быть выражена очень малой дробью.

Однако и события, обладающие очень малой вероятностью, осуществляются вполне закономерно, хотя они могут казаться невозможными. Маловероятные события при многократном повторении явления приобретают вполне устойчивую и определенную вероятность их осуществления, хотя бы оно было в одном случае из многих миллионов. С точки зрения вероятности возникновение жизни на земле является необычайно редким событием. Но как бы ни невероятным казалось возникновение жизни или тех этапов, из которых складывалось возникновение жизни на земле, времени для него было достаточ-

но, поэтому оно наверняка могло произойти хотя бы один раз, что было уже достаточным для дальнейшего развития жизни.

Оценка того, насколько должна быть мала вероятность, для того чтобы с ней можно было не считаться, в значительной степени зависит от степени важности события, о котором идет речь. Так, если вероятность воздействия нового удобрения не на повышение, а на понижение урожая равна 0,05, это не должно помешать применению удобрения, так как все же оно окажется полезным в 0,95 случаев. Совсем другое дело, если оказывается, что новое лекарственное средство может с вероятностью, равной 0,05, принести не пользу, а вред организму больного. В этом случае его применение не может быть допущено.

Эти примеры мы приводим для того, чтобы подготовить изучающего вариационную статистику к принципу, широко применяемому сейчас в опытах и наблюдениях, когда заранее намечают приемлемую величину (или уровень) вероятности и ее считают достаточной для доказательства получения того или иного эффекта.

По мере приближения величины p к единице событие становится все более достоверным.

Если $p=1$, то событие бесспорно наступит. Оно вполне достоверно.

Теоремы сложения и умножения вероятностей. Для понимания закономерностей случайной вариации важны две теоремы. Первая из них—теорема сложения вероятностей—относится к таким независимым друг от друга событиям, которые несовместимы друг с другом; вторая—умножения вероятностей—также к независимым событиям, но совместимым друг с другом или следующим друг за другом. Эти теоремы можно проиллюстрировать на следующих элементарных примерах. На клумбе растут 20 красных, 30 синих и 40 белых астр. Какова вероятность сорвать в темноте цветную астру? Она равна сумме вероятностей сорвать красную астру или синюю астру, т. е.

$$p = \frac{20}{90} + \frac{30}{90} = \frac{50}{90}.$$

Второй пример несколько сложнее. Какова вероятность, что при выбрасывании двух кубиков, на гранях

которых написаны цифры от 1 до 6, наверху будет сумма не менее 10? Эта вероятность составляется из суммы трех вероятностей: получить сумму цифр 10, сумму 11 и сумму 12. Первая вероятность, как легко рассчитать из данных выше сочетаний двух цифр, $p_{10}=\frac{3}{36}$, вторая — $p_{11}=\frac{2}{36}$ и третья — $p_{12}=\frac{1}{36}$. Сумма их составит $^{6}/_{36}$, или $^{1}/_{6}$.

Для умножения вероятностей необходимо, чтобы второе событие E_2 осуществлялось только при осуществлении события E_1, при этом осуществление E_1 не влияет на вероятность осуществления E_2, т. е. E_1 и E_2 независимы.

Какова вероятность наличия цифры 4 наверху двух выброшенных одновременно кубиков? При выбрасывании одного кубика вероятность появления цифры 4 равна $^{1}/_{6}$. При выбрасывании второго кубика вероятность та же: $^{1}/_{6}$. Общая вероятность $p=\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$.

Второй пример. Какова вероятность прохождения по лабиринту с шестью развилками и шестью тупиками. Очевидно, что на каждом развилке вероятности попасть или в тупик или к следующему развилку одинаковы — по $^{1}/_{2}$. Тогда при наличии шести развилков общая вероятность будет равна

$$\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{2^6}=\frac{1}{64}.$$

Выборочная, генеральная и стохастическая совокупности. В приведенных выше примерах исчислялись так называемые эмпирические вероятности. Они приложимы только к тем конкретным совокупностям, на которых они вычислены. Вероятность появления комолых телят относится к определенной изученной группе скота. Для породы, в которой очень много комолых животных, вероятность появления комолых телят окажется во много раз выше вычисленной. Для практики же очень важно судить не только об отдельных конкретных случаях, но и о всех возможных случаях этого рода. Математическая теория, имея дело с отдельными, частными наблюдениями, выработала методы, позволяющие по по-

лученным результатам наблюдений судить о тех результатах, которые имели бы место, если бы была изучена не только данная совокупность осуществившихся случаев, но и теоретически мыслимая совокупность всех возможных случаев этого рода. Иначе говоря, по эмпирическим опытным вероятностям, основанным на учете конкретных относительных частостей тех или других явлений, судить о теоретических или так называемых априорных вероятностях, т. е. таких, которые можно брать заранее, до проведения опыта.

В предыдущих главах было приведено несколько вариационных рядов. Каждый из них являлся результатом изучения некоторого, сравнительно небольшого числа животных. Так, суждение о плодовитости серебристо-черных лисиц, ее средней величине и изменчивости было сделано по 80 лисицам. Но можно было бы изучить не эту маленькую группу лисиц, а всех лисиц, разводимых в СССР. Такая совокупность всех конкретных объектов, которую можно было бы изучить, называется генеральной. Иногда ее называют также популяцией. Изученная же небольшая группа представляет собой как бы выборку из генеральной совокупности, поэтому ее называют выборочной. Наконец, можно себе представить и теоретически мыслимую совокупность, т. е. совокупность всех возможных наблюдений, в том числе и таких, которые практически не были осуществлены. Такую совокупность называют *стохастической.*

Теория вероятностей как раз дает возможность построить абстрактные совокупности, представляющие собой отображение реальных совокупностей. В таких абстрактных стохастических совокупностях, доступных точному математическому анализу, вероятности становятся теоретическими. Очевидно в жизни мы встречаемся, как правило, с выборочными совокупностями, но по ним мы стремимся судить о генеральной или стохастической совокупности. Так, для изучения окуня данного озера нет надобности изучать всю его популяцию, т. е. генеральную совокупность, а достаточно взять выборочную совокупность в количестве 100, 200 или 1000 особей. По капле крови больного можно делать выводы о состоянии всей крови, данные об изменчивости нескольких десятков леммингов позволяют судить о всей популяции леммингов и т. д.

Если бы все особи популяции были сходны, то уже по одной особи можно было бы получить полную информацию о всей генеральной совокупности, всей популяции. Но в действительности существует очень большая изменчивость как среди самих особей популяции, так и в отношении условий внешней среды, в которой они живут и развиваются. Поэтому и проводимые многократно выборки из генеральной совокупности никогда не будут одинаковыми. Нужно учесть, что далеко не всякая выборочная совокупность содержит информацию о генеральной совокупности. Вот почему перед биологами, применяющими методы вариационной статистики, стоит очень важная задача—так составлять выборочные совокупности, чтобы они отображали генеральную совокупность, т. е. популяцию.

Эта задача стоит и при постановке опытов на ограниченной группе животных и растений. Так, когда мы ставим опыт по влиянию рационов кормления коров на жирность молока, мы выделяем несколько групп коров и каждую из них ставим на определенный рацион. Таким образом, создается ограниченное количество групп животных, также составляющих выборочные совокупности, т. е. выборки из генеральной совокупности всех коров данной породы. Однако можно было бы представить и рассчитать все возможные теоретически варианты опыта. Составленная из них теоретическая совокупность является тогда стохастической.

Естественно, может возникнуть вопрос, каковы же закономерности вариации внутри каждой совокупности и каково взаимоотношение между разными типами совокупностей. Это дает возможность подойти и к другому важному вопросу, можно ли по статистическим показателям, полученным на основании изучения одной совокупности, например выборочной, судить о статистических показателях других видов совокупности, например генеральной. Иначе говоря, это вопрос о том, насколько достоверны статистические показатели, полученные по выборочной совокупности, для того чтобы можно было судить по ним о генеральной совокупности.

Распределение вероятностей—основа вариации. Обратимся опять к вариационному ряду. Выше было разобрано несколько эмпирических вариационных рядов и показано, что для всех их характерно определенное рас-

пределение вариант, а именно: чем ближе значения вариант к средней арифметической, тем выше их частота, а чем дальше, тем реже они встречаются. В конечном счете это распределение вариант основано на теоретической закономерности уменьшения вероятности встречаемости той или иной варианты по мере ее удаления от средней. Для иллюстрации того, что вариационный ряд действительно основан на вероятности, покажем, как распределяются вероятности появления курочек среди 10 цыплят. Начнем со случая, когда среди них нет ни одной курочки (0), далее 1 курочка из 10 цыплят, 2 курочки, 3 курочки и т. д. и, наконец, когда все цыплята — курочки (табл. 17).

Таблица 17

Распределение вероятностей появления разного количества курочек среди 10 цыплят

Количество курочек	0	1	2	3	4	5	6	7	8	9	10
Количество случаев	1	10	45	120	210	252	210	120	45	10	1
Вероятности	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001

Если графически выразить данные табл. 17, то будет получена изображенная на рис. 4 вариационная кривая распределения числа случаев появления разного количества курочек среди 10 цыплят (на общее число 1024 случая). Она является так называемой биномиальной кривой распределения, соответствующей разложению бинома Ньютона, о чем уже упоминалось в главе 1. Биномиальность кривой распределения можно уяснить на следующем примере. Представим себе, что мы подбрасываем одновременно две монеты. Будем считать выпадение герба (Г) благоприятным случаем, а выпадение решетки (Р) неблагоприятным. Возможны 4 случая выпадения герба и решетки.

В первом случае обе монеты выпадут гербами вверх (ГГ). Во втором, на первой монете вверху герб, на второй—решетка (ГР). В третьем случае на первой монете вверху решетка, а на второй герб (РГ). Второй и третий случаи совпадают по результату. Каждый из них является комбинацией одного благоприятного (Г) и одного неблагоприятного (Р) случаев. Наконец, в четвертом случае обе монеты выпадут решетками вверх (РР). Какова

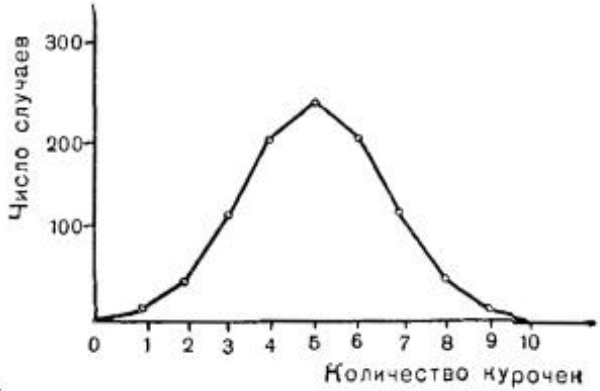


Рис. 4. Кривая распределения числа случаев появления разного количества курочек среди 10 цыплят (на общее число 1024 случая).

же вероятность каждого результата? Вероятность выпадения герба обозначим буквой p, а вероятность выпадения решетки—q. В данном случае $p=q=\frac{1}{2}$. Тогда вероятность выпадения двух монет одновременно гербами вверх равна произведению вероятностей, т. е. $p.p=p^2$.

Вероятность выпадения одной монеты гербом вверх и другой решеткой вверх равно $p.q$. Так как таких случаев с разным порядком наступления благоприятного и неблагоприятного результата два, то их вероятности суммируются: $pq+pq=2pq$. Наконец, вероятность сочетания двух неблагоприятных случаев, т. е. выпадение решетки, равна $q.q=q^2$. Таким образом, для простейшего примера из 2 событий мы имеем следующее их распределение:

$$(p+q)^2 = p^2 + 2\,pq + q^2.$$

Такое же рассуждение может быть применено к случаю сочетания 3, 4 и т. д. событий. Во всех случаях получение вероятности различных сочетаний независимых со-

бытий основывается на том, что вероятности нескольких комбинаций выражаются членами разложения бинома $(p+q)^k$, где k—число независимых случайных событий, p и q—соответствующие вероятности благоприятных и неблагоприятных событий. Чтобы получить не отдельные вероятности, а вероятные численности разных результатов при данном общем числе n, надо умножить их на это общее число случаев, т. е. испытаний. В приведенном выше примере число сочетаний разного количества *курочек и петушков* $k=10$, т. е. мы имеем дело с биномом

$$(p+q)^{10}.$$

Его разложение в виде конкретного количества случаев каждого сочетания дано во второй строчке табл. 17 ($n=1024$), а вероятности отдельных случаев—в третьей строчке. Сумма вероятностей должна быть равна 1.

Таким образом, вариационный ряд с характерным для него расположением большинства вариант вблизи его центральной части и рассеиванием к краям ряда является в то же время и распределением вероятностей. Это значит, что в вариационном ряду случайная переменная x принимает разные значения:

$$x_1, x_2, x_3, \ldots, x_n$$

в зависимости от большого количества самых разнообразных причин, независимых по отношению друг к другу. Поэтому вариацию величины x можно рассматривать как случайную. Отдельным значениям x можно придать соответствующие вероятности $p_1, p_2, p_3, \ldots, p_n$. Совокупность значений x и соответствующих им вероятностей и называется распределением.

Биномиальное распределение. Если вероятности появления отдельных значений x выражаются величинами, соответствующими коэффициентам разложения бинома Ньютона, как это было в приведенном выше примере, распределение называется биномиальным.

Для биномиального распределения характерна дискретность, прерывистость. Переход от одного класса к другому здесь совершается не постепенно, а прерывисто, так как вероятности отдельных классов определяются коэффициентами разложения бинома, имеющего определенную степень. Характер биномиального распределения зависит от соотношения вероятностей p и q, при этом принимается, что разница между p и q сравнительно не-

велика. При очень малом значении p распределение будет значительно отличаться от биномиального. Такое распределение будет рассмотрено ниже.

Как и для большинства других распределений, параметрами для биномиального распределения являются средняя арифметическая и среднее квадратическое отклонение.

Теоретически их значения определяются значениями вероятностей p и q, а также значением k, т. е. числа независимых событий, распределение которых изучается.

Отсюда средняя арифметическая при биномиальном распределении

$$\bar{x} = kp \qquad (16)$$

и среднее квадратическое отклонение

$$\sigma = \sqrt{kpq}\,. \qquad (17)$$

При биномиальном распределении можно связать определенные $\bar{x}$ и σ, вычисленные на основе данного конкретного материала, с вероятностями p и q.

Возьмем следующий пример, схематизированный для упрощения подсчетов. В 100 группах цыплят, подвергавшихся экспериментальному воздействию, определяли количество погибших цыплят (табл. 18).

Таблица 18

Распределение погибших цыплят в 100 группах

Количество погибших цыплят x	Частота f	fx	fx^2
0	6	0	0
1	24	24	24
2	36	72	144
3	24	72	216
4	6	24	96
	$n = 96$	$\Sigma = 192$	$\Sigma = 480$

По данным табл. 18 можно вычислить обычными методами, указанными во второй главе, $\bar{x}$ и σ:

$$\bar{x} = \frac{192}{96} = 2 \text{ цыпленка на группу,}$$

$$\sigma = \sqrt{\frac{480 - \frac{192^2}{96}}{96}} = 1 \text{ цыпленок на группу.}$$

С другой стороны, исходя из математических взаимоотношений, при биномиальном распределении можно определить $\bar{x}$ и σ через величины k, p и q, а именно:

$$\bar{x} = kp = 4 \cdot 0{,}5 = 2,$$

$$\sigma = \sqrt{kpq} = \sqrt{4 \cdot 0{,}5 \cdot 0{,}5} = 1.$$

В данном случае принято, что $p=q=\frac{1}{2}$, как точно установленные теоретические вероятности. Но возможны и другие значения вероятностей p и q. Было получено следующее распределение самок в 103 пометах с 4 мышками каждая:

Количество самок	0	1	2	3	4
Число пометов	8	32	34	24	5

Тогда $\bar{x} = \frac{8 \cdot 0 + 1 \cdot 32 + 2 \cdot 34 + 3 \cdot 24 + 4 \cdot 5}{103} = 1{,}864.$

Но так как $\bar{x}=kp$, а $k=4$, то $p=\frac{1{,}864}{4} = 0{,}466$.

Исходя же из формулы $\sigma^2=kpq$, можно получить

$$\sigma^2 = 4 \cdot 0{,}466 \cdot 0{,}534 = 1{,}0.$$

Так как данный ряд является рядом разложения бинома $(0{,}534+0{,}466)^4$ при $n=103$, то путем разложения бинома легко вычислить, сколько особей следует ожидать в каждом классе. Получатся следующие цифры для ожидаемых частот каждого класса:

Количество самок	0	1	2	3	4
Ожидаемое число пометов	8	29	38	23	5

Уже на глаз видно большое совпадение фактически полученных величин с ожидаемыми.

В главе 7 будут изложены методы установления степени соответствия фактических опытных данных с ожидаемыми с помощью метода $\varkappa$-квадрат.

Нормальное распределение и его характеристика с помощью нормированного отклонения. Если при биномиальном распределении значение показателя бинома $(p+q)^k$ является конечным, то при приближении k к бесконечности распределение становится непрерывным.

Графическое его изображение будет выражаться плавной симметричной кривой, носящей название нормальной вариационной кривой. Само же распределение получило название нормального. Нормальное распределение занимает важнейшее место в вариационной статистике, поэтому очень важно разобрать закономерности случайной вариации при нормальном распределении.

Для изучения закономерностей случайной вариации в настоящее время широко пользуются понятием так называемого нормированного отклонения, которое обозначается буквой t. Нормированное отклонение представляет собой отклонение той или другой варианты (или группы вариант) от средней арифметической, выраженное в сигмах, т. е.

$$t = \frac{x_i - \bar{x}}{\sigma}. \qquad (18)$$

Отсюда $$x_i - \bar{x} = t\sigma.$$

В дальнейшем будет показано, что t имеет несколько более широкий смысл и что оно может выражаться не только в сигмах.

Изменение величины t характеризует различные типы распределения. При нормальном распределении значения t колеблются в пределах примерно ± 3, т. е. отклонения от средней охватывают несколько больше 6 сигм, 3 сигмы вправо от средней и 3 сигмы влево, как это видно на представленной на рис. 5 нормальной вариационной кривой.

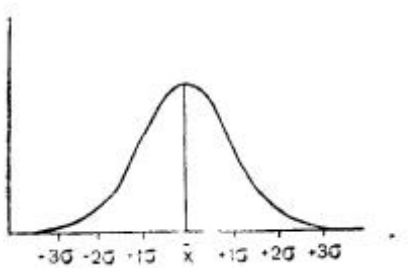


Рис. 5. Нормальная вариационная кривая.

Зная вариационную кривую распределения вариант по тому или иному признаку и предполагая, что распределение является нормальным, можно установить, какой процент изученных особей (или вариант) укладывается в пределах $\pm 1\sigma$, в пределах $\pm 2\sigma$, в пределах $\pm 3\sigma$. Так, в пределах $\pm 1\sigma$ располагается 68,3% всех вариант данного ряда, в пределах $\pm 2\sigma$—95,5% и в пределах $\pm 3\sigma$—99,7% всех вариант.

Таким образом, различные значения t ограничивают определенные части вариационного ряда, что очень хорошо видно на графике нормальной вариационной кривой. Но в то же время распределение t указывает на закономерность уменьшения количества вариант по мере отдаления от средней арифметической. Эта закономерность основана на закономерности распределения вероятностей.

Вероятность любого отклонения от средней есть функция нормированного отклонения. Эта функция выражается довольно сложной формулой, которую мы здесь приводить не будем, но на ее основе была составлена готовая таблица так называемого нормального интеграла вероятностей. Так как к ней придется не раз обращаться в связи с разбором материала самых различных частей нашего курса, она дана в приложении (табл. I), а не в тексте.[1]

В табл. I первая колонка слева дает значения t с одним десятичным знаком, второй десятичный знак t представлен 10 столбцами, на которых вверху стоят цифры от 0 до 9. Тогда t=0,11 соответствует значение вероятности 0876 (в таблице 2-я строчка, вторая цифра), значению t=1,00 — 6827 (11-я строчка, первая цифра) и т. д.

В целях упрощения для вероятностей даны лишь десятичные знаки, поэтому слева надо к ним присоединять ноль и запятую, т. е. число 0876 надо записать как 0,0876, число 6827 как 0,6827 и т. д.

С помощью таблицы интеграла вероятностей можно определить вероятность нахождения вариант в данных пределах величины $\pm t$, т. е. в зависимости от того, на сколько сигм или долей сигмы они отклоняются от средней арифметической.

Так, вероятность того, что взятая наугад особь из части вариационного ряда, ограниченной справа и слева от средней одной сигмой, т. е. $\pm 1t$ ($t=\pm 1\sigma$), равна 0,6827; двумя сигмами, т. е. $\pm 2t$, равна 0,9545 и, наконец, тремя сигмами, т. е. в пределах $\pm 3t$, равна 0,9973.

Доверительные вероятности. Существенно важны две последние вероятности, которые постоянно упоминаются в биологических, зоотехнических и агрономических рабо-

[1] Эта и другие таблицы, помещенные в приложении, будут обозначаться римскими цифрами.

тах с использованием методов биометрии. Округленно их обычно выражают величинами 0,95 и 0,99. Из табл. I можно установить, что с вероятностью 0,9500 любая случайно взятая особь будет отклоняться от $\bar{x}$ не более чем на $1{,}96\sigma$, или иначе с вероятностью 0,05 она будет за пределами $1{,}96\sigma$. С вероятностью же, равной 0,9973, она будет отклоняться от $\bar{x}$ не более чем на 3σ. Соответственно этому вероятность отклонения от $\bar{x}$ больше чем на 3σ ($t > \pm 3$) очень мала—всего 0,0027. Это очень важное правило часто называют правилом трех сигм. Три сигмы как бы ограничивают пределы случайного рассеяния внутри вариационного ряда. То, что находится в пределах 3σ, относится к данному ряду, за пределами 3σ вероятнее всего уже не относится. В сущности для достижения вероятности 0,9900 достаточно взять границы даже не $\pm 3\sigma$, а только $\pm 2{,}58\sigma$.

Вероятность, выражающаяся величиной 0,99, достаточно велика, и в тех случаях, когда достигнута такая вероятность, можно с очень большой степенью уверенности делать вывод, в частности, по поводу отнесения особи к той или иной группе, результатов опыта и т. д. Но нередко можно остановиться и на более низком уровне вероятности, например 0,95. В этом случае отклонения от ожидаемого будут уже в 5% случаев (вероятность 0,05). Вероятности 0,95 и 0,99, или 95% и 99%, получили название доверительных вероятностей, т. е. таких, значениям которых можно достаточно доверять или которыми можно уверенно пользоваться.

Понятие доверительной вероятности, в настоящее время широко используемое в статистике, было введено английским биологом и статистиком Р. Фишером.

Вероятности, принятые как доверительные, в свою очередь определяют доверительные границы и доверительный интервал между ними, на которых можно основывать оценку той или иной величины и те границы, в которых она может находиться при разных вероятностях.

Для различных вероятностей доверительные интервалы будут следующими:

Вероятности	Интервалы
0,95	$-1{,}960\sigma \ldots +1{,}960\sigma$
0,99	$-2{,}576\sigma \ldots +2{,}576\sigma$
0,999	$-3{,}291\sigma \ldots +3{,}291\sigma$

Вероятности можно обозначать как в долях единицы, так и в процентах, поэтому в последующем мы будем употреблять параллельно оба обозначения.

Уровни значимости. Определенным значениям вероятностей соответствуют так называемые уровни значимости. Вероятности 0,95 (95%) соответствует уровень значимости 0,05 (5%). По отношению к закономерностям нормального распределения это означает, что выход за пределы принятых границ возможен в порядке случайности с вероятностью в 0,05, т. е. в 5% случаев рискуют ошибиться в своих выводах.

При вероятности 0,99 уровень значимости 0,01 (1%). Случайное отклонение возможно лишь с вероятностью 0,01, т. е. риск ошибиться в оценках составляет только 1% (1 случай на 100).

Таким образом, уровень значимости обозначает вероятность получения случайного отклонения от установленных с определенной вероятностью результатов. С помощью уровня значимости можно оценить, в каком проценте случаев (или с какой вероятностью, если обозначать уровень значимости не в процентах, а в долях единицы) все же возможна ошибка в результатах,в тех выводах, которые делаются на основе опыта, в оценке достоверности того или иного показателя или величины, в оценке различий между какими-то величинами, полученными в опытах или при проведении наблюдений. Так как при научном исследовании надо не только получить те или другие результаты, но и сделать выводы, очень важно, чтобы получаемые выводы имели достаточно высокую достоверность (употребляют также термины—значимость, существенность). 5%-ный уровень значимости (0,05) указывает, что возможна в силу случайности ошибка в 5% случаев. В некоторых случаях можно удовлетвориться и таким результатом. Но если нужна большая доказательность результатов, то уровень значимости должен быть повышен до 1% (0,01). Чем цифра меньше, тем уровень значимости, а следовательно, и достоверность результатов, выше. При уровне значимости 0,01 (1%) вывод необоснован только в одном случае из 100. Такую значимость считают уже высокой и широко ею пользуются. Но бывают случаи, когда уровень значимости может быть еще выше—0,001. Тогда вывод не обоснован только в одном случае из 1000.

Уравнение нормальной кривой распределения. Кривая нормального распределения может быть охарактеризована математически с помощью определенного уравнения. Это уравнение носит следующую форму:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

В него входят все уже известные величины, а именно μ (греческая буква ми), т. е. средняя арифметическая генеральной совокупности (или вообще средняя нормального распределения при $k \to \infty$), σ и σ^2—среднее квадратическое отклонение и варианса, характеризующие степень колеблемости вокруг средней. Эти две величины являются параметрами нормального распределения. π—(читается пи)— число, равное 3,14159, а e—основание натуральных логарифмов, равное 2, 71828. В показателе степени величины e находится возведенное в квадрат нормированное отклонение $t = \frac{x-\mu}{\sigma}$. При нормальном распределении большая часть площади кривой укладывается в пределах $\pm 3t$ (или в пределах $\pm 3\sigma$, так как t выражено в сигмах). Точки перегиба нормальной кривой приходятся на $+1\sigma$ и -1σ.

Если принять $\sigma=1$ и заменить значение $\frac{x-\mu}{\sigma}$ величиной t, то уравнение кривой нормального распределения примет следующую, более простую форму:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Высчитанные для разных значений t величины y и дадут ординаты нормальной кривой.

Если взять значения t только с одним десятичным знаком, то в сокращенном виде значения ординат будут следующими (см. на стр. 72).

Таким образом, получается, что в случае нормального распределения, при $t=3{,}0$, кривая практически сливается с осью абсцисс. Продолжение кривой за пределами

$\pm 3t$ можно заметить только при очень большом числе изучаемых особей.

t	Ординаты	t	Ординаты
0,0	0,3989	1,6	0,1109
0,1	0,3970	1,7	0,0940
0,2	0,3910	1,8	0,0790
0,3	0,3814	1,9	0,0656
0,4	0,3683	2,0	0,0540
0,5	0,3521	2,1	0,0440
0,6	0,3332	2,2	0,0355
0,7	0,3123	2,3	0,0283
0,8	0,2897	2,4	0,0224
0,9	0,2661	2,5	0,0175
1,0	0,2420	2,6	0,0136
1,1	0,2179	2,7	0,0104
1,2	0,1942	2,8	0,0079
1,3	0,1714	2,9	0,0060
1,4	0,1497	3,0	0,0044
1,5	0,1295	3,9	0,0002
		4,0	0,0001

Распределение при малых значениях n. Закономерности нормального распределения вариант можно наблюдать лишь при значительном количестве наблюдений. Они основаны на так называемом законе больших чисел, формулировка которого будет дана ниже. На практике же, и в частности в биологии, нередко приходится встречаться с очень ограниченным числом вариант или наблюдений. Особенно это относится к выборочным совокупностям. Возникает вопрос о том, каковы в этом случае закономерности распределения и насколько они будут отличаться от закономерностей нормального распределения. Ответ на него практически дал английский математик Госсет, который писал под псевдонимом «Стюдент». Поэтому изученное им распределение вероятностей получило название t-распределения по Стюденту.

Теоретическое обоснование закона распределения, открытого Стюдентом, было дано Фишером. Наряду с некоторыми математическими особенностями распределения Стюдента, о которых сейчас мы говорить не будем, существенно то, что оно может быть использовано и при очень малых количествах вариант. В этих случаях оно несколько отличается от нормального распределения,

приближаясь к нему по мере увеличения n. Практически при $n=20$ оно уже мало отличается от нормального. t-распределение по Стюденту также симметрично.

Таблица распределения вероятностей по Стюденту в *пределах* $\pm t$ *для малого числа наблюдений* n *дана* в приложении (табл. II).

При малом n вероятности нахождения вариант в пределах тех значений t, о которых говорилось выше, значительно снижаются, иначе говоря, для достижения тех же *вероятностей нужно взять значительно большие интервалы* $\bar{x} \pm t$. Так, при $n=5$ (число степени свободы 4) вероятность 0,95 достигается лишь при $t=\pm 2{,}8$, а вероятность 0,99—при $t=\pm 4{,}6$. На рис. 6 представлены для сравнения 2 кривые: верхняя—для нормального распределения при $n=\infty$ и нижняя — для t-распределения по Стюденту при $n=5$. Кривые несколько отличаются друг от друга. Нижняя — с более крутой вершиной и с краями, более растянутыми вправо и влево, нежели верхняя нормальная кривая. По 2,5% вариант справа и слева отсекаются: в верхней кривой при $t=1{,}96$, в нижней—при $t=2{,}78$. В обоих случаях вероятность 0,95, а уровень значимости 0,05. Так как в практической работе надо исходить из определенных уровней значимости, то были составлены рабочие таблицы, с помощью которых можно определить без таблиц I и II минимальное зна-

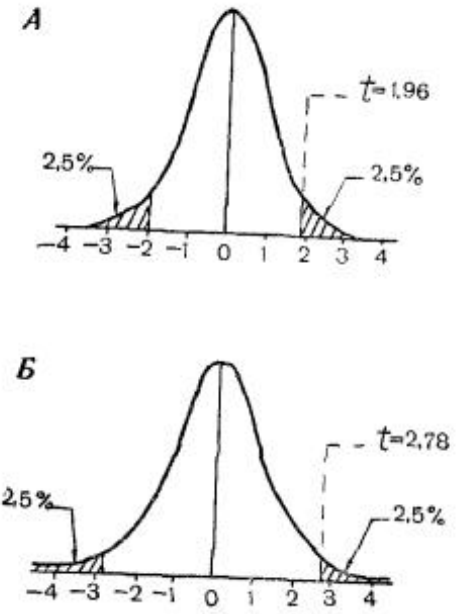


Рис. 6. Разные значения t, отсекающие 2,5% вариант справа и 2,5% слева: а) на кривой нормального распределения при $n=\infty$, $t=1{,}96$; б) на кривой t-распределения по Стюденту при $n=5$, $t=2{,}78$

чение t, обязательно требующееся для данной вероятности (табл. III).

Табл. III построена на основе того, что заранее приняты необходимые доверительные вероятности и соответствующие им уровни значимости. Для упрощения в ней даны только 4 уровня значимости (0,1, 0,05, 0,02 и 0,01), в полных таблицах обычно приводят и иные уровни значимости. Если, например, выборка включает только 10 наблюдений (число степеней свободы 9), а требуется по условиям опыта уровень значимости 0,01 (и доверительная вероятность 0,99), то величина t должна быть не менее 3,25. Уровню значимости 0,05 (и доверительной вероятности 0,95) удовлетворяет при $d.f.=9$ величина $t=2{,}62$. В то же время при большом n ($n>30$) этим уровням значимости будут удовлетворять значения $t=2{,}576$ и $t=1{,}96$, как это было показано выше при разборе нормального распределения. Нижняя строчка, где $n=\infty$, связывает значения по Стюденту со значениями, приведенными в табл. I для нормального интеграла вероятностей. Для более точных расчетов вероятности надо пользоваться таблицами Стюдента при малом n ($n<30$) и таблицами нормального интеграла вероятности при бо́льших n ($n>30$). Погрешности в последнем случае даже при небольших n будут незначительными.

Кроме биномиального и нормального распределений, существует еще несколько других видов распределения, обычно носящих названия по именам открывших их ученых (распределение Пуассона, Неймана и др.).

Следует упомянуть о распределении Пуассона, так как с ним приходится встречаться в физике и в биологии. Распределение Пуассона является частным случаем биномиального распределения. Оно осуществляется в соответствии с разложением бинома $(p+q)^k$. Но при пуассоновском распределении p очень мало, т. е. наблюдаемые события осуществляются очень редко, а k очень велико и стремится к бесконечности. Поэтому физики применяют закономерности пуассоновского распределения к таким явлениям, как испускание радиоактивными веществами α-частиц, где число α-частиц очень мало по сравнению с общим числом атомов. В биологии пуассоновскому распределению удовлетворяют редко наблюдаемые явления, например появление полиэмбрионии в се-

менах растений, частота островков Лангерганса в тканях поджелудочной железы и др. Многие расчеты в современной радиобиологии основываются на анализе пуассоновского распределения, в частности, при применении так называемой теории мишени, так как и здесь приходится встречаться с очень редкими событиями. Если, например, происходит облучение группы клеток или бактерий γ-лучами, то число облучаемых объектов, т. е. наблюдаемых событий ($n=k$), очевидно, очень велико, наблюдаемые же изменения (смерть отдельных бактерий, цитологические изменения в клетках) являются редкими событиями i, вероятность которых ($p=\frac{i}{n}$) является очень малым числом.

Само распределение отдельных наблюдений при этом является обычно асимметричным. Асимметрия тем больше, чем меньше p. При увеличении p, а отсюда и $\bar{x}$, оно приближается к нормальному. Пуассоновское распределение характеризуется в сущности только одним параметром—средней арифметической $\bar{x}$, так как σ^2 в этом случае обычно равна $\bar{x}$ или близка ей по значению. Именно по этому равенству $\bar{x}$ и σ^2 легче всего определить, что данное распределение является пуассоновским.

Средняя арифметическая для пуассоновского распределения (обычно она обозначается не $\bar{x}$, а греческой буквой лямбда—λ) равна, как и при биномиальном распределении, kp, где p—вероятность обнаружения данного признака, а k—количество фактически проведенных наблюдений (формула 16). σ^2 будет равна или очень близка к λ.

ВОПРОСЫ

1. Что такое вероятность? По какой формуле вычисляется вероятность?
2. Какие процессы называются вероятностными или стохастическими?
3. Дайте примеры некоторых биологических явлений, осуществление которых может быть оценено известной вероятностью.
4. Можно ли не считаться с возможностью событий, обладающих малой вероятностью?
5. Какое значение имеет p для очень достоверных событий?
6. Какая связь существует между частотой определенной категории животных и вероятностью?
7. Чему равна сумма $p+q$?

8. Дайте определения теорем сложения и умножения вероятностей. Проиллюстрируйте их на примерах.

9. Что такое выборочная совокупность, генеральная, стохастическая? Можно ли называть генеральную совокупность популяцией?

10. С какими популяциями чаще всего приходится иметь дело биологу?

11. Если бы все особи популяции были бы одинаковы, по какому количеству особей можно было бы получить информацию о популяции?

12. Какая связь существует между вариацией в пределах вариационного ряда и распределением вероятностей?

13. Что такое биномиальная кривая распределения? Какая общая формула является основой для биномиального распределения?

14. Что такое k в биноме $(p+q)^k$?

15 Какими признаками и параметрами характеризуется биномиальное распределение? Является ли оно дискретным или непрерывным?

16. Как можно связать значения $\bar{x}$ и σ^2 при биномиальном распределении со значениями p, q и k?

17. Что такое нормальное распределение и как оно связано с биномиальным?

18. Почему нормальное распределение является непрерывным?

19. Что такое нормированное отклонение? Сколько t охватывает вариационный ряд при нормальном распределении?

20. Что показывает таблица нормального интеграла вероятностей?

21 Какой процент особей укладывается в пределах $\pm 1\,\sigma$, $\pm 2\sigma$, $\pm 3\,\sigma$?

22. Какова вероятность, что взятая наугад варианта будет отклоняться от средней не более чем на $1{,}9\,\sigma$, на $2{,}5\,\sigma$?

23. Какие вероятности считаются доверительными?

24. Дайте определение терминов «доверительные границы» и «доверительный интервал».

25. Каков доверительный интервал при нормальном распределении с вероятностью 0,95, 0,99?

26. Что такое уровень значимости? Какая связь между уровнем значимости и вероятностью? Можно ли выражать уровень значимости в процентах? На что указывает процентная величина уровня значимости?

27. Каков характер распределения при малых значениях n?

28. Чем отличается t-распределение по Стьюденту от распределения t в нормальном ряду?

29. Для достижения одной и той же вероятности, в каком случае значения t должны быть бо́льшими, при малом n или при большом?

30. Чем отличается распределение Пуассона от биномиального? Дайте примеры использования распределения Пуассона в биологии.

31. Можно ли заметить распределение Пуассона по значениям $\bar{x}$ и σ^2?

Какими параметрами характеризуется распределение Пуассона?

ЗАДАЧИ

27 При каком значении $\pm t$ 50% вариант находится в пределах данного значения t, а другие 50%—за его пределами?

28 Какому уровню значимости соответствует $t = 2{,}575$ (при $n = \infty$, при $n = 30$, при $n = 15$)?

29 Какая часть вариант нормального распределения находится в интервалах $\bar{x} \pm 1{,}645\,\sigma$, $\bar{x} \pm 2\,2\sigma$, $\bar{x} \pm 2\,86\,\sigma$?

30 На 1000 мальчиках 13 летнего возраста было установлено, что 390 из них отклоняются от средней арифметической по росту (высоте тела) не более чем на 1,4 дюйма ($\bar{x} = 57{,}3$ дюйма) Можно ли по этим данным определить примерную величину σ, если предусматривается нормальное распределение?

31 Какое значение t нужно взять, чтобы оно ограничивало 95% площади вариационной кривой при разных значениях n, а именно если $n = 4$, $n = 12$, $n = 20$?

32 Если совокупность очень большая, при каком значении t возможны случайные отклонения за его пределами в сторону плюс. в 2,5% случаев, в 5,0% случаев?

33 105 опоросов по 8 поросят в каждом распределялись по числу самцов следующим образом

Число самцов	1	2	3	4	5	6	7	8
Количество опоросов	5	9	22	25	26	14	4	1

Приняв, что в данном случае имеется биномиальное распределение, вычислите $\bar{x}$ и с помощью его определите p и q Попробуйте вычислить отдельные значения количества опоросов, развернув формулу $(p + q)^8$, при $n = 106$

34 На 10 000 семей с 4 детьми было все девочки — в 641 семье, 3 девочки и 1 мальчик — в 2625 семьях, 2 девочки и 2 мальчика— в 3748 семьях, 1 девочка и 3 мальчика — в 2420 семьях, все мальчики — в 566 семьях Исходя из предположения о биномиальности распределения, вычислите вероятность рождения мальчиков и девочек.

35 В 221 выборках по 6 поросят учитывали количество самцов 0,1,2 3 и т д Распределение оказалось следующим.

Количество самцов	0	1	2	3	4	5	6
Число выборок	3	16	53	78	53	10	8

Какая частота классов ожидается, если распределение должно соответствовать коэффициентам разложения бинома $(p + q)^6$, где $p = q = 1/2$

36 Среди 402 опоросов свиней дюрок-джерзейской породы, в каждом из которых было 8 поросят, пометы распределялись следующим образом

Количество самцов в помете	0	1	2	3	4	5	6	7	8
Количество опоросов	1	8	37	81	162	77	30	5	1

Определите $\bar{x}$ и σ обычным методом и сравните их с получаемыми по формулам $\bar{x} = kp$ и $\sigma = \sqrt{kpq}$.

37. В 100 пробах, в каждой из которых находилось по 1200 зерен ржи, проверяли наличие двойных зародышей Оказалось, что в некоторых пробах находили от 1 до 6 таких зародышей. Распределение найденных зерен с 2 зародышами по пробам было следующим:

Количество зерен с двумя зародышами	0	1	2	3	4	5	6
Число проб	6	24	32	18	9	6	5

Вычислите обычным путем (с помощью условнои средней) среднюю арифметическую количества зерен с 2 зародышами на пробу, а также вариансу данного ряда и обратите внимание на почти полное их равенство. К какому типу распределения следует отнести этот ряд? Какова вероятность нахождения зерен с 2 зародышами в общей популяции зерен ржи?

38. В табуне лошадей гнедых было 250, а вороных — 150. Какова вероятность того, что одна из пойманных на удачу лошадей будет гнедой, вороной? Чему равна сумма этих двух вероятностей?

Глава 4

ОЦЕНКА ДОСТОВЕРНОСТИ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ

Вариация в различных совокупностях. Средняя ошибка. После рассмотрения общих закономерностей случайной вариации надо вернуться к вопросу о различиях между тремя типами совокупностей—выборочной, генеральной и стохастической. В общем виде можно утверждать, что эти совокупности характеризуются одинаковыми закономерностями случайной вариации. Для них могут быть вычислены соответствующие статистические показатели: средняя арифметическая и варианса (или среднее квадратическое отклонение). Было установлено, что средняя арифметическая генеральной совокупности всегда равна средней арифметической стохастической совокупности. Иначе говоря, оценка генеральной совокупности, под которой обычно понимают всю популяцию, в то же время является и оценкой теоретически мыслимой совокупности. Средняя арифметическая выборочной совокупности $\bar{x}$ характеризует среднюю арифметическую генеральной совокупности μ (греческая буква ми) лишь приближенно, отличаясь от нее на некоторую величину.

Сказанное станет ясным из разбора данных одного из вариационных рядов (табл. 8). Для 168 коров симментальской породы была получена средняя арифметическая $\bar{x}=73{,}85$ см. 168 коров представляют собой выборку из генеральной совокупности, охватывающей популяцию всех коров симментальской породы. Если бы мы взяли ряд выборок из популяции симментальской породы, то обнаружилось бы, что $\bar{x}$ этих выборок будут различными. Одни из $\bar{x}$ будут несколько больше, чем

73,85 см, другие несколько меньше. Значения $\bar{x}$ для отдельных выборок будут обладать вариацией, которая в свою очередь может быть изображена вариационной кривой и измерена своей сигмой. Эта сигма получила название средней ошибки или средней квадратической ошибки. Иногда ее называют также стандартной ошибкой. Она является мерилом достоверности показателей выборочной совокупности. В то же время она указывает на возможные границы, в пределах которых находится средняя арифметическая генеральной совокупности—μ.

Средняя ошибка для $\bar{x}$ может быть вычислена по формуле

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \qquad (19)$$

В прежних руководствах по статистике среднюю ошибку обозначали латинской буквой m или греческой Δ. Ее можно было бы также изобразить как $\sigma_{\bar{x}}$. Из формулы видно, что средний квадрат отклонений выборочных средних от μ пропорционален среднему квадрату отклонений от средней арифметической выборочной совокупности и обратно пропорционален объему выборочной совокупности.

Входящие в формулу показатели σ и n получаются, как известно, для любого вариационного ряда. В примере с коровами (табл. 8) они равны: σ=2,45 см; n=168. Отсюда средняя ошибка для средней арифметической глубины груди изученных 168 симментальских коров

$$s_{\bar{x}} = \frac{2{,}45}{\sqrt{168}} = 0{,}17 \text{ см}.$$

Таким образом, по $\bar{x}$ можно с некоторой вероятностью судить о μ. Оказалось, что вероятность появления данной величины средней арифметической для выборки из генеральной совокупности является функцией того же нормированного отклонения, с помощью которого была дана выше характеристика нормального распределения. Поэтому и здесь с помощью нормированного отклонения t, только выраженного не в σ ряда, а в $s_{\bar{x}}$ $\left(t = \frac{\bar{x} - \mu}{s_{\bar{x}}}\right)$, можно установить возможные границы, в пределах кото-

рых находится средняя арифметическая генеральной совокупности, а именно:

$$\bar{x} - ts_{\bar{x}} \leqslant \mu \leqslant \bar{x} + ts_{\bar{x}}, \qquad (20)$$

или иначе $\bar{x} - t\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + t\frac{\sigma}{\sqrt{n}}$.

Это означает, что истинное значение μ, т. е. средняя арифметическая для генеральной совокупности, может находиться:

при $t=1$ с вероятностью 0,68 в границах $73{,}85 \pm 0{,}17$ см
при $t=2$ „ 0,95 „ $73{,}85 \pm 2 \cdot 0{,}17$ см
при $t=3$ „ 0,997 „ $73{,}85 \pm 3 \cdot 0{,}17$ см

Если бы было взято несколько выборок из всей совокупности симментальских коров, то можно заранее предвидеть, что больше половины из них будут иметь средние в пределах 73,85+0,17 и 73,85—0,17 см, т. е. между 74,02 и 73,68 см, и громадное большинство, т. е. 997 случаев из 1000, в пределах $73{,}85 \pm 3.0{,}17$ см, т. е. между 74,36 (73,85+0,51) и 73,34 см (73,85—0,51).

Средняя ошибка—ошибка выборочности. Термин «ошибка» часто вводит в заблуждение начинающих, которые предполагают, что она является результатом недостаточной аккуратности в работе. Это не так. Статистическая ошибка, в данном случае средняя ошибка, ничего общего не имеет с ошибкой точности. Само собою разумеется, что все измерения (вес и промеры рыб, удои коров и жирность их молока, настриги шерсти овец и ее длина) надо делать точно и добросовестно. Но статистические показатели для выборочной совокупности всегда имеют так называемые ошибки выборочности, которые представляют собой величины расхождения между значениями изучаемых признаков в отобранной группе, выборке, и в генеральной совокупности.

Так как $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, то очевидно, что размер определяемой средней ошибки зависит от сигмы выборочной популяции и от ее объема. Чем лучше взята выборка, чем больше ее размеры, тем меньше и средняя ошибка, тем меньше расхождение между значениями признаков в выборочной и генеральной совокупностях.

Закон больших чисел. В связи между статистическими показателями выборочных, генеральных и стохастических совокупностей выражается так называемый закон больших чисел. Именно так он был назван Пуассоном. В наиболее общем виде этот закон заключается в том, что чем больше число n некоторых случайных величин, тем их средняя арифметическая ближе к средней арифметической генеральной или стохастической совокупности, т. е. средней их математических ожиданий, тем меньше разница между μ и $\bar{x}$. По мере увеличения n вероятность осуществления приближения $\bar{x}$ к μ становится все больше, стремясь при $n = \infty$ к единице, т. е. к *полной достоверности*.

В этом заключается теорема одного из основоположников вариационной статистики русского математика Чебышева.

А так как всякое явление, как правило, складывается из массы единичных, случайных явлений, то закон больших чисел выступает как реальный закон объективной действительности. Именно он лежит в основе нормального распределения вариант в вариационном ряду, т. е. распределения значений случайной переменной $\bar{x}$ вокруг $\bar{x}$, в основе распределения $\bar{x}$ отдельных выборок из одной генеральной совокупности. Но так как средняя ошибка представляет собой то же среднее квадратическое отклонение, только для вариационного ряда, составленного из многих $\bar{x}$ отдельных выборок одной и той же совокупности, то и к ряду значений статистических ошибок относятся изложенные в предыдущей главе закономерности случайной вариации. Ошибки также следуют закону нормального распределения, что было открыто еще в начале XIX в. математиками Гауссом и Лапласом.

Установление доверительных границ для μ с помощью t. С помощью t можно установить доверительные границы для μ, т. е. возможные границы, в пределах которых находится средняя арифметическая генеральной совокупности μ при той или иной доверительной вероятности (0,95; 0,99 и т. д., что соответствует определенным уровням значимости: 0,05, 0,01 и т. д.). Чтобы указать, какой уровень значимости или вероятности принимается в данном случае, при букве t записывают показатель уровня значимости, например, t_{05} или t_{01}.

Кроме того, необходимо обращать внимание на n. При большом n значение t можно взять из таблицы нормального интеграла вероятностей (табл. I), при малом n—из таблицы Стюдента (табл. II).

В примере с глубиной груди симментальского скота $n=168$. Величина t_{05} (т. е. вероятность 0,95, а уровень значимости 0,05) при $n=168$ по табл. I будет 1,96. Так как $s_{\bar{x}}=0{,}17$ см, то это значит, что доверительный интервал для колебаний μ при уровне значимости 0,05 будет от 73,52 (=73,85—1,96.0,17) до 74,18 (=73,85+1,96.0,17).

Можно записать и так: 73,85±0,33 (при 95% вероятности или уровне значимости 0,05).

Для иллюстрации определения доверительного интервала средней арифметической при малом n возьмем такой пример. Определяли концентрацию витамина C в томатном соке (в миллиграммах на 100 г сока). При этом $\bar{x}=20$ (мг/100 г), $s_{\bar{x}}=0{,}965$ (мг/100 г), $n=17$.

Надо определить интервал с доверительной вероятностью 0,95 (уровнем 0,05). Так как n меньше 20, надо воспользоваться табл. II для t-распределения по Стюденту.

Так как в табл. II нет графы $n=17$, надо взять цифры вероятностей, средние между $n=16$ и $n=18$. Для вероятности 0,95 значение t будет между 2,1 и 2,2, примерно 2,12. Тогда $t_{05}.s_{\bar{x}}=2{,}12.0{,}965=2{,}05$ (мг/100 г), а доверительные границы будут 17,95 (=20—2,05) и 22,05 (=20+2,05) (мг/100 г).

Еще проще воспользоваться для установления t табл. III. При $n=17$ количество степеней свободы равно 16 (n—1). В пересечении графы для уровня значимости 0,05 и строчки $d.f.=16$ находим $t=2{,}12$.

Оценка достоверности средней арифметической выборочной совокупности. Средняя ошибка позволяет определить степень достоверности, или значимости, самой величины средней арифметической данной выборочной совокупности. Что надо понимать под достоверностью средней арифметической? Так как средняя арифметическая является результатом сложения ряда значений случайной переменной, т. е. x_1, x_2, x_3, ..., x_n, то при наличии некоторых x с отрицательным знаком возможен и такой теоретически мыслимый случай, когда $\bar{x}$ будет равно 0. Именно с нулем и надо сравнивать $\bar{x}$, т. е. надо

доказать, что средняя арифметическая достоверно отличается от нуля. Мерилом достоверности явится, как и ранее, нормированное отклонение, только знаменателем надо взять не σ ряда, а σ выборочных средних, т. е. $s_{\bar{x}}$.

В таком случае t будет равняться:

$$t = \frac{\bar{x} - 0}{s_{\bar{x}}}. \qquad (21)$$

При большом n значения t можно брать из таблицы нормального интеграла вероятности (табл. I), при малом—из таблицы распределения по Стюденту (табл. II).

Если же обратиться к табл. III, то достаточно будет, чтобы полученное фактически значение превышало табличное при данном значении *d.f.* и принятом уровне значимости.

В качестве примера можно взять данные табл. 11 по вариации процента жира в молоке 8 коров. Здесь $\bar{x}=3{,}8$;

$$\sigma = 0{,}2;\ s_{\bar{x}} = \frac{0{,}2}{2{,}83} = 0{,}07.$$

Отсюда $t = \frac{3{,}8}{0{,}07} = 54{,}3.$

В данном случае и без помощи таблиц можно утверждать, что полученное значение $\bar{x}$ высоко достоверно. Вообще следует указать, что $\bar{x}$, вычисленные для каких-либо конкретных биологических показателей, даже на *сравнительно малых по размерам* выборочных совокупностях чаще всего являются достаточно достоверными, если только ряд не является слишком растянутым. Однако иначе может получиться, если приходится оперировать с экспериментальными данными, в которых фигурируют какие-либо условные или относительные величины, часть которых может иметь и отрицательный знак. Тогда установление достоверности $\bar{x}$ совершенно необходимо, и оно может подчас привести к неожиданным результатам. Фишер приводит в качестве примера данные по анализу дополнительных часов сна, полученных в опытах по применению 2 снотворных лекарств, где $\bar{x}=+1{,}58$ и $s_{\bar{x}}=0{,}3890$. Тогда $t=4{,}06$ при $n=9$. По таблице Стюдента находим, что достоверность $\bar{x}$ равна 0,996, т. е. уровень значимости 0,004. Это значит, что шансов на то, что полученное значение $\bar{x}$ случайно, всего только 4

на 1000. Полученную величину $\bar{x}$ можно считать вполне достоверной.

Средние ошибки для σ и с. v. В некоторых случаях могут понадобиться средние ошибки для других биометрических показателей. Они вычисляются по следующим формулам:

s для σ $$(s_\sigma) = \frac{\sigma}{\sqrt{2n}}, \qquad (22)$$

s для с. v. $$(s_{\text{с. }v.}) = \frac{c.\ v.}{\sqrt{2n}}. \qquad (22a)$$

Эти формулы можно применять голько при большом числе наблюдений. При малом n применяются другие, более сложные методы.

Средняя ошибка и здесь дает возможность по такому же принципу, как для $\bar{x}$ определить доверительные границы для σ и $c.v$. Допустим, что σ=3,5, n=200. Тогда $s_\sigma = \frac{3,5}{400}$=0,175. При уровне значимости 0,01 t=2,58.

Доверительные границы для σ будут 3,05 (3,5—2,58.0,175) и 3,95 (3,5+2,58.0175).

Это значит, что среднее квадратическое отклонение при уровне значимости 0,01 находится между 3,05 и 3,95.

Таким образом, с помощью средней ошибки можно судить о степени достоверности полученных на основании выборочной совокупности статистических показателей, можно установить, приняв определенную вероятность, возможные границы для колебаний средней арифметической и среднего квадратического отклонения. Это дает возможность предсказать по выборочной совокупности свойства генеральной совокупности. В дополнение к сказанному можно еще добавить, что варианса выборочной совокупности (σ^2) очень близка к варианса генеральной совокупности (σ_δ^2).

Нулевая гипотеза. Метод средней ошибки позволяет сравнивать между собой любые две группы животных или растений, например две выборочные совокупности, взятые из природной, неизученной популяции; выборку из какой-то уже известной группы и группу, из которой она взята; опытную и контрольную группы при постановке опытов, и установить, насколько достоверны различия между их статистическими показателями (средними

арифметическими, вариансами и др.). Общие принципы сравнения основываются на анализе так называемой нулевой гипотезы. Согласно этой гипотезе, первоначально принимается, что между данными показателями (или группами, на основе которых они получены) достоверного различия нет, т. е. что обе группы вместе составляют один и тот же однородный материал, одну совокупность. Статистический анализ должен привести или к отклонению нулевой гипотезы, если доказана достоверность полученных различий, или к ее сохранению, если достоверность различий не доказана, т. е. различия признаны случайными. Но так как все статистические показатели и различия между ними характеризуются определенными уровнями значимости, то отбрасывание нулевой гипотезы должно быть связано с принятием определенного уровня значимости. Так, если признан необходимым уровень значимости 0,01 и если вероятность достоверности данного статистического показателя или разницы между показателями не удовлетворяет этому условию, т. е. она ниже 0,99 (например, 0,97, 0,91, 0,88), то тогда нет оснований для отбрасывания нулевой гипотезы. Ее надо по-прежнему считать правильной, по крайней мере, до тех пор, пока новые данные не дадут возможности ее опровергнуть, доказав, что существующие различия не являются чисто случайными.

Конечно, и в том случае, когда нулевая гипотеза считается опровергнутой, какой-то шанс, что она в действительности верна, остается. При уровне значимости 0,01 этот шанс составляет 1 на 100, т. е. в 1% случаев отбрасывание нулевой гипотезы было ошибкой. Если достигнут уровень значимости не 0,01, а 0,001, то уверенность в том, что нулевая гипотеза действительно отвергнута, резко возрастает (лишь 1 шанс на 1000 случаев, что она все же верна). Сравнительно редко в научных работах по биологии можно удовлетвориться уровнем значимости 0,05, ибо тогда уверенность в правильности вывода составляет лишь 95 случаев из 100, а в 5 возможен неправильный вывод.

Проверкой нулевой гипотезы является также оценка достоверности средней арифметической выборочной совокупности. Первоначально принималось, что $\bar{x}=0$. Надо было доказать, что $\bar{x}$ достоверно отличается от

нуля. Если же это не удавалось сделать, то оставалась правильной нулевая гипотеза.

Оценка достоверности разницы между средними арифметическими двух совокупностей. Нулевая гипотеза в данном случае будет сводиться к признанию того, что две совокупности являются в сущности одной совокупностью и разница между их средними арифметическими чисто случайна, т. е. лежит в пределах ошибки выборочности.

Чтобы иметь право отвергнуть нулевую гипотезу, надо доказать, что разница между средними арифметическими достоверна, т. е. удовлетворяет требуемому уровню значимости.

Для установления достоверности разницы между средними арифметическими надо воспользоваться нормированным отклонением. Нормированное отклонение примет следующую форму:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}. \qquad (23)$$

Числителем является разница между средними арифметическими двух групп. Ее можно обозначить сокращенно буквой d. В знаменателе же—средняя ошибка этой разницы, т. е. s_d. Тогда

$$t = \frac{d}{s_d}. \qquad (23a)$$

В том случае, когда обе сравниваемые группы обладают достаточно большой численностью, большей чем 30 особей каждая,[1] легко применить эту общую формулу. Зная $\bar{x}_1$ и $\bar{x}_2$, а также $s_{\bar{x}_1}$ и $s_{\bar{x}_2}$, можно определить t, вычислив d и s_d.

Средняя ошибка разницы определяется по формуле:

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}. \qquad (24)$$

Допустим, что мы хотим сравнить по удою 2 группы коров. В одной группе n_1=50. В другой n_2=40. Средний

[1] В некоторых руководствах по вариационной статистике при больших значениях n пишут вместо буквы t букву u или T и относят обозначение t только к малым выборкам. Но чтобы не усложнять формул, мы для всех случаев нормированных отклонений будем употреблять одно обозначение t.

удой коров первой группы: $\bar{x}_1 \pm s_{\bar{x}} = 2100 \pm 120$ кг; второй группы: $\bar{x}_2 + s_{\bar{x}_2} = 2635 \pm 140$ кг. Разница между средними удоями двух групп

$$d = \bar{x}_2 - \bar{x}_1 = 2635 - 2100 = 535 \text{ кг.}$$

Ошибка разницы

$$s_d = \sqrt{s_{x_1}^2 + s_{x_2}^2} = \sqrt{140^2 + 120^2} = 184 \text{ кг.}$$

Таким образом, $d \pm s_d = 535 \pm 184$ кг,

т. е. $$t = \frac{535}{184} = 2{,}91.$$

По таблице нормального интеграла вероятности (табл. I) находим, что в этом случае вероятность достоверности очень велика—0,9963. При отсутствии таблиц можно исходить из правила трех сигм: если разница превышает свою ошибку почти в три раза, она достоверна с вероятностью не менее 0,99. Уровень же значимости в таком случае 0,01, т. е. только 0,01 шансов (1 из 100 случаев) за то, что эта разница все же случайна. При сравнении двух групп с малыми n формула t для установления достоверности или недостоверности различия между двумя группами будет более сложной, так как величина s_d определяется по формуле

$$s_d = \sqrt{\frac{\Sigma_1 (x_1 - \bar{x}_1)^2 + \Sigma_2 (x_2 - x_2)^2}{(n_1 - 1) + (n_2 - 1)} \left(\frac{n_1 + n_2}{n_1 \; n_2}\right)}; \quad (25)$$

Смысл этой формулы заключается в том, что нельзя пользоваться просто готовыми средними ошибками, вычисленными заранее для двух сравниваемых групп, как это было при применении формулы (24), а нужно сначала сложить суммы квадратов отклонений из обеих групп, т. е. получить объединенную сумму квадратов отклонений, затем определить вариансу объединенных рядов (путем деления объединенной суммы квадратов на число степеней свободы обеих групп) и, наконец, после умножения на $\frac{n_1 + n_2}{n_1 \; n_2}$ и извлечения квадратного корня получить ошибку разницы.

Для иллюстрации сказанного возьмем следующий пример. На двух группах крыс был поставлен опыт по сравнению влияния двух рационов на рост. Крысы первой группы (n=12) получали рацион с высоким содержанием белка, крысы второй (n=7)—с низким. Привесы (в г) за 56 дней опыта для каждой крысы составляли: первой группы: 134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97, 123; второй группы: 70, 118, 101, 85, 107, 132, 94.

После обработки данных с помощью формул для среднего квадрата (например, 10б) без составления вариационных рядов можно составить таблицу 19.

Таблица 19

Сводные данные по сравнению двух групп крыс, получавших разные рационы

Рационы	Количество крыс	Число степеней свободы $d\ f.$	Средний привес, г $\bar{x}$	Сумма квадратов отклонений $\Sigma(x-\bar{x})^2$
Высокобелковый	12	11	120	5023
Низкобелковый	7	6	101	2552

$$d\ f = 17;\ d = \bar{x}_1 - \bar{x}_2 = 19;\ \Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2 = 7584$$

Подставив все значения в формулу, получим:

$$s_d = \sqrt{\frac{7584}{17} \cdot \frac{(12+7)}{12\,7}} = \sqrt{100{,}9262} = 10{,}04.$$

Отсюда $t = \frac{19}{10{,}04} = 1{,}89.$

По табл. III находим, что (при $d.f.$=17 и уровне значимости 0,05) t должно быть не менее 2,11, полученное значение t ниже табличного. Для уточнения вероятности достоверности воспользуемся табл. II. Из нее видно, что t=1,89 соответствует вероятности только 0,92, т. е. уровень значимости равен 0,08. Таким образом, можно считать, что разные рационы не привели к разделению популяции крыс по привесам на две достоверно отличающиеся друг от друга популяции, иначе говоря, нулевая

гипотеза не может быть отвергнута. Конечно, опытные группы были слишком малы. Возможно, что при их увеличении была бы получена более достоверная разница между группами крыс, находившимися на разных рационах кормления.

Достоверность разницы между попарными данными. В некоторых случаях можно значительно упростить все расчеты по проверке достоверности разницы, оперируя непосредственно значениями разниц между вариантами обеих групп. Для этого надо, чтобы последние были сгруппированы попарно. Такой случай как раз имеет место, если опытная и контрольная группы (или две опытные группы) составлены из отдельных партнеров однояйцевых двоен того вида, у которого бывают однояйцевые двойни. Один член каждой пары двоен помещается в одну опытную группу и подвергается воздействию фактора *A*, а другой—в другую группу и подвергается воздействию фактора *B*. Подобную же парность данных можно получить, если, например, для изучения влияния микроэлементов на число крольчат в помете экспериментировать с одними и теми же крольчихами, которые в период одних окролов рассматриваются как контрольные, а в период других—как опытные. При оценке быков-производителей по потомству сравнивают попарно удои коров-дочерей с удоями их матерей. Такой по-

Таблица 20

Попарное сравнение веса самок и самцов мышей, г

Номер помета	Вес ♀	Вес ♂	*d*	Номер помета	Вес ♀	Вес ♂	*d*
1	26	16,5	9,5	14	22,5	20,5	2
2	20	17	3	15	23,5	19,5	4
3	18	16	2	16	23,5	22,5	1
4	28,5	21	7,5	17	25	20	5
5	23,5	23	0,5	18	24,5	20,5	4
6	20	19,5	0,5	19	23,5	18	5,5
7	22,5	18	4,5	20	20,5	24,5	−4
8	24	18,5	5,5	21	20	22	−2
9	24	20	4	22	20,5	20	0,5
10	25	28	−3	23	25	20	5
11	22	27,5	−5,5	24	23,5	23	0,5
12	24	20,5	3,5	25	22	24	−2
13	22,5	23	−0,5				

парный метод имеет ряд преимуществ перед методом создания опытной и контрольной групп из случайно взятых особей или методом аналогов.

В качестве примера возьмем данные о весах самок и самцов мышей (табл. 20) в возрасте 125 дней в 25 пометах (в каждом помете были 1 самка и 1 самец).

Данные последнего столбца, т. е. 25 разниц d, можно обработать как вариационный ряд.

Предоставляем каждому самому обработать его, применив для вычисления σ^2 одну из формул (6а, 10 или 10 б).

Приведем лишь готовые данные:

$$d = 2{,}04,$$
$$\sigma^2 = 13{,}177,$$
$$\sigma = 3{,}63,$$
$$s_d = 0{,}73,$$
$$t = \frac{\bar{d}}{s_{\bar{d}}} = 2{,}81.$$

Число степеней свободы $d.f. = n - 1 = 24$.

Из табл. III видно, что при $d.f. = 24$ для уровня значимости 0,01, т. е. 99% вероятности того, что разница достоверна, t должно быть 2,80. Полученное значение $t = 2{,}81$ как раз лежит в границах требуемой достоверности, иными словами между средним весом ♂♂ и ♀♀ мышей разница достоверна, или как еще говорят в статистике, значима, существенна. Этим самым нулевая гипотеза должна быть отвергнута.

Сравнение средних квадратических отклонений и варианс. Если сравниваемые группы численно достаточно велики, сравнение их изменчивости может быть проведено по тому же принципу, как и сравнение $\bar{x}$, с помощью показателя t. В данном случае

$$t = \frac{\sigma_1 - \sigma_2}{s_{\sigma_1 - \sigma_2}}. \qquad (26)$$

В знаменателе—ошибка разницы между средними квадратическими отклонениями. Она вычисляется по формуле

$$s_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}. \qquad (27)$$

При $t \geqslant 3$ разницу между сигмами можно считать, как обычно, достоверной, существенной.

Однако в силу ряда теоретических соображений, изложение которых выходит за рамки элементарного курса, значительно более точным методом для сравнения изменчивости и установления достоверности различий между степенями изменчивости сравниваемых групп является так называемый критерий F, представляющий собой отношение варианс:

$$F = \frac{\sigma_1^2}{\sigma_2^2}. \qquad (28)$$

Теоретически значения F могут колебаться от 0 до ∞. Если обе вариансы σ_1^2 и σ_2^2 равны, то тогда $F=1$. Очевидно, что нулевой гипотезой является признание равенства варианс. Если они не равны, то нужно доказать, что это неравенство неслучайно, достоверно. Значения F, являющиеся границами для признания достоверности разницы между вариансами, приводятся в специальных таблицах, где учитываются разные объемы сравниваемых групп (вернее для разных чисел степеней свободы этих групп) и принимаются различные уровни значимости. В несколько сокращенном виде они представлены в табл. IV (для уровня значимости 0,05) и в табл. V (для уровня значимости 0,01).

Обычно отношение варианс берут таким образом, чтобы в числителе была большая варианса.

Если полученная величина F больше табличного значения при принятом уровне значимости, различие между вариансами признается достоверным; если она меньше, то расхождение между вариансами может считаться несущественным, случайным, т. е. нулевая гипотеза остается неопровергнутой.

Практическое значение F очень велико в целом ряде специальных глав вариационной статистики, особенно в так называемом дисперсионном (или вариансном) анализе. Если различия между вариансами групп в опытах, где анализируется влияние различных факторов (удобрение, корма, лекарства, химические вещества, наследственные свойства производителей и т. д.) на растения или животных, могут быть признаны достоверными, это позволяет устанавливать влияние тех или иных факторов на изучаемые признаки или биологиче-

ские свойства (урожайность, молочность, устойчивость к заболеваниям и т. д.).

Разбор методов дисперсионного анализа не входит в задачу нашего элементарного курса,[1] поэтому важно только получить общее представление о критерии F как отношении варианс.

В качестве примера вычисления F возьмем данные опытов по влиянию шести различных рационов кормления на яйценоскость кур.

Между группами с разным кормлением варианса оказалась равной $\sigma_1^2 = 1074{,}5$, при этом $d.f._1 = 5$.

Варианса же внутри групп, получавших одинаковые рационы кормления, равна $\sigma_2^2 = 312{,}4$, при этом $d.f._2 = 114$.

Таким образом, видно, что между группами кур с разным кормлением разнообразие по яйценоскости больше, нежели внутри групп. Чтобы доказать, достоверно ли это различие в изменчивости, обратимся к критерию F:

$$F = \frac{1074{,}5}{312{,}4} = 3{,}44.$$

По табл. V находим цифру в пересечении строчки, где $d.f._2 = 120$ (так как нет 114), и столбца, где $d.f._1 = 5$ (вертикальные столбцы указывают число степеней свободы для большей вариансы). Она равна 3,17. Полученное значение F превышает табличное. Значит различия по яйценоскости между группами кур с разными рационами кормления достоверны с вероятностью 0,99 (только в 1 случае из 100 эта разница может быть следствием случайности).

Можно привести и более простой пример использования критерия F. Нужно сравнить изменчивость по высоте в холке групп черно-пестрого и красно-пестрого скота. Для первого $n_1 = 100$ и $\sigma_1^2 = 16{,}32$, для второго $n_2 = 42$ и $\sigma_2^2 = 14{,}44$.

Тогда $$F = \frac{16{,}32}{14{,}44} = 1{,}13.$$

[1] См. книгу Н. А. Плохинского. Дисперсионный анализ. Изд. Сибирского отделен. АН СССР. Новосибирск, 1960.

В таблицах IV и V в вертикальных столбцах нет цифры 100 Тогда надо взять $df_1 = \infty$. По горизонтали же можно взять $df_2 = 40$ Обратимся сначала к табл V При уровне значимости 0,01 F должно быть больше 1,80 Этому уровню значимости полученное значение F явно не удовлетворяет В таком случае, может быть, различие между вариансами σ_1^2 и σ_2^2 удовлетворяет уровню значимости 0,05 По табл IV F для $df_1 = \infty$ и $df_2 = 40$ равно 1,51 Фактическая величина F ниже и этой величины Отсюда можно сделать вывод, что хотя черно-пестрый и красно-пестрый скот отличаются по масти, но их варианс по высоте в холке достоверно не отличаются друг от друга Вероятность различия между вариансами, как случайного, более 0,05 Нулевая гипотеза о равенстве варианс сохраняет свое значение и остается неопровергнутой Можно считать, что группы черно-пестрого и красно-пестрого скота по высоте в холке составляют одну популяцию

Вычисление ошибки при альтернативной изменчивости. Учет достоверности получаемых показателей необходим не только в случае количественной, но и качественной, альтернативной, изменчивости Как уже указывалось выше, при альтернативной изменчивости имеется налицо одна из двух возможностей данный признак присутствует или его нет Достаточно сосчитать число тех и других случаев, и мы получим картину альтернативной изменчивости, выраженную или в абсолютных числах, или в процентах, или в долях единицы

Выше были даны формулы для p и σ_p (14а, 15, 15а, 15б) Средняя ошибка для оценки достоверности вычисляется по обычной формуле (19) Применительно к альтернативной изменчивости она примет следующий вид

$$s_p = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}, \qquad (29)$$

или

$$s_p = \sqrt{\frac{pq}{n}}. \qquad (29a)$$

В данном случае p и q—это выраженные в долях единицы частоты присутствия или отсутствия качественного признака, а n—общее число вариант или наблюдений Однако в некоторых случаях желательно определить

ошибки не для частот, выраженных в долях единицы, а для конкретных чисел особей с тем или другим признаком. Тогда может быть применена та же формула ошибки, только в числителе подкоренного количества требуется записать $p_1\ (n-p_1)$, где под p_1 следует понимать конкретное число особей с данным качественным признаком:

$$s_{p_1} = \sqrt{\frac{p_1(n-p_1)}{n}}. \qquad (30)$$

Допустим, что стадо ярославского скота состояло из 200 голов, в том числе 120 голов было черно-пестрых, а 80 нечерно-пестрых (рыжих и рыже-пестрых).

Тогда

$$p = \frac{120}{200} = 0{,}6\ ,$$

$$q = 0{,}4\ ,$$

$$\sigma_p = \sqrt{0{,}6 \cdot 0{,}4} = 0{,}49,$$

$$s_p = \sqrt{\frac{0{,}6 \cdot 0{,}4}{200}} = 0{,}034.$$

Тогда доля черно-пестрых $p \pm s_p = 0{,}6 \pm 0{,}034$.

Значение p во много раз превышает свою ошибку. То же относится и к q.

Применив формулу ошибки для конкретных чисел черно-пестрых и нечерно-пестрых животных, получим

$$s_p = \sqrt{\frac{120 \cdot 80}{200}} = 6{,}9$$

(для 120 черно-пестрых),

$s_p = 6{,}9$ (для 80 нечерно-пестрых).

Таким образом, распределение на черно-пестрых и нечерно-пестрых можно записать так:

черно-пестрых $120 \pm 6{,}9$ голов,

нечерно-пестрых $80 \pm 6{,}9$ голов.

Методы определения разницы между средними при альтернативной изменчивости проиллюстрируем на 2 группах данных по реакции коров на туберкулез. В одной группе, состоявшей из 284 коров, реагировавших

(P_{1x}) было 83, в другой, состоявшей из 50 коров, реагировавших (P_{1y}) было 6. В таком случае $\bar{p}_x = \frac{83}{284} = 0{,}29$ и $\bar{p}_y = \frac{6}{50} = 0{,}12$.

$$d = \bar{p}_x - \bar{p}_y = 0{,}17.$$

Можно было бы для определения s_d использовать обычную формулу $s_d = \sqrt{s_1^2 + s_2^2}$. Она вполне применима для сравнения двух групп в пределах одной совокупности, например к только что приведенной группе черно-пестрых и нечерно-пестрых животных. Но в нашем примере мы имеем дело с двумя группами данных. Полученные на них показатели надо сравнивать не друг с другом, а с показателями теоретически мыслимой единой совокупности, из которой взяты 2 выборки, одна—284 коровы, а другая—50 коров.

Для такой совокупности p можно вычислить по формуле:

$$p = \frac{n_1 p_x + n_2 p_y}{n_1 + n_2} = \frac{p_{1x} + p_{1y}}{n_1 + n_2}$$

В данном случае $p = \frac{83 + 6}{284 + 50} = 0{,}27$.

Поэтому ошибки для средних двух групп следует вычислять, исходя не из частных p_1 и p_2, а из p генеральной совокупности, т. е. надо взять в качестве ошибок s_1^0 и s^0

$$s_1^0 = \sqrt{\frac{p(1-p)}{n_1}} \text{ и } s_2^0 = \sqrt{\frac{p(1-p)}{n_2}}.$$

Тогда

$$s_d = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}} = \sqrt{p(1-p)\cdot\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

или, если помнить, что $1-p=q$, то

$$s_d = \sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}} = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Подставив в формулы конкретные значения из нашего примера, получим:

$$s_d = \sqrt{0{,}27 \cdot 0.73 \cdot \left(\frac{1}{284} + \frac{1}{50}\right)} = 0{,}068.$$

Отсюда

$$t = \frac{d}{s_d} = \frac{0{,}17}{0{,}068} = 2{,}52.$$

По таблице интеграла вероятности такое значение t дает вероятность 0,9883 и уровень значимости 0,012. Отсюда можно сделать вывод о существенном различии в реакции на туберкулинизацию двух групп коров, т. е. что нулевая гипотеза должна быть отброшена.

Ввиду некоторой трудности изложенного материала и важности этих методов для биологических исследований приведем еще один пример, на этот раз из генетики. При воздействии рентгеновых лучей на дрозофил был получен несколько более высокий процент мутаций тогда, когда перед облучением их кормили кормом, содержавшим железо. Данные были следующими. При облучении дозой в 300 p дрозофил, питавшихся обычным кормом, на 805 культур F_2 было получено 80 сцепленных с полом мутаций. При облучении той же дозой дрозофил, кормившихся кормом, содержавшим железо, было получено на 2756 культур F_2 357 сцепленных с полом мутаций. Таким образом,

$$n_1 = 805;\ p_1 = \frac{80}{805} = 0{,}0994 \text{ (или } 9{,}94\%);$$

$$n_2 = 2756;\ p_2 = \frac{357}{2756} = 0{,}1295 \text{ (или } 12{,}95\%);$$

$$d = p_2 - p_1 = 0{,}0301 \text{ (или } 3{,}01\%).$$

Возникает вопрос, является ли это повышение частоты мутаций результатом чисто случайных причин (в частности, может быть сыграла роль значительная разница в численности опытной и контрольной групп) или же оно вызвано наличием солей железа в корме.

Примем, что доказательство должно удовлетворять уровню значимости 0,01.

Идя тем же путем, как в предыдущем примере, получим

$$p = \frac{437}{3512} = 0{,}123$$

$$\text{и } q = 1 - 0{,}123 = 0{,}877,$$

$$s^2_d = 0{,}123 \cdot 0{,}877 \cdot \left(\frac{1}{805} + \frac{1}{2756}\right) = 0{,}0001726.$$

Откуда $s_d = \sqrt{0{,}0001726} = 0{,}013$

и $t = \frac{d}{s_d} = \frac{0\,0301}{0{,}013} = 2{,}31.$

Для уровня значимости 0,01 (т е вероятности достоверности 0,99) t должно быть равно 2,58 Значение же t=2,31 дает вероятность только 0,9791 и уровень значимости 0,021

Таким образом, полученные результаты не удовлетворяют поставленному условию — уровню значимости 0,01, поэтому нулевую гипотезу приходится считать неопровергнутой, т е разницу надо считать случайной В этом опыте был получен все же довольно высокий уровень значимости, выше, чем 0,05, хотя и ниже требуемого, поэтому при менее строгом подходе исследователи могли бы им удовлетвориться Однако более правильно считать, что влияние железа на частоту вызванных облучением мутаций остается все же недоказанным и что необходимы новые опыты, в которых опытная и контрольная группы были бы приблизительно равными и большими по численности, чем в проведенных опытах

ВОПРОСЫ

1 Отличаются ли друг от друга по закономерностям случайной вариации выборочная, генеральная и стохастическая совокупности?

2 В какой степени средняя арифметическая выборочной совокупности характеризует среднюю арифметическую генеральной совокупности?

3 Как колеблются $\bar{x}$ отдельных выборок вокруг средней арифметической генеральной совокупности?

4 Что такое средняя ошибка? Какова ее формула?

5 В каких пределах по отношению к $\bar{x}$ выборочной совокупности может находиться средняя арифметическая генеральной совокупности? С какой вероятностью?

6 В чем заключается ошибка выборочности?

7 Объясните в чем заключается закон больших чисел?

8 Закон больших чисел как основа распределения вариант в вариационном ряду, $\bar{x}$ отдельных выборок и ошибок

9 Какова зависимость между значением средней ошибки и объемом совокупности?

10 Изменяются ли доверительные границы и доверительный интервал для μ при разных величинах n? Когда надо пользоваться t распределением Стюдента?

11 Как оценивается достоверность среднеи арифметической выборочной совокупности?

12 Каковы формулы средних ошибок для σ и cv?

13 Какова связь между вариансами выборочной и генеральной совокупностей?

14 Объясните сущность нулевой гипотезы и дайте примеры

15 Как оценивается достоверность разницы между средними арифметическими? Одинаковы ли способы оценки при малых и больших n?

16 Как получить объединенную сумму квадратов отклонений для двух рядов?

17 Как формулируется нулевая гипотеза при сравнении двух средних арифметических?

18 В чем преимущество попарного сравнения данных? Дайте примеры из биологии

19 Можно ли установить достоверность разницы между средними квадратическими отклонениями с помощью t?

20 Что такое критерий F? Симметрично или асимметрично распределение F?

21 В чем заключается нулевая гипотеза при сравнении вариант?

22 Можно ли считать достоверным различие между вариансами если фактическое значение F больше табличного? если оно меньше табличного? если оно равно табличному?

23 Как вычисляется ошибка при альтернативной изменчивости? Покажите ее применение к исчислению доли данного качественного признака в совокупности, к конкретным числам особей с качественным признаком

24 Методы определения достоверности разницы между средними арифметическими при альтернативной изменчивости? Различие между ними при сравнении двух групп в пределах одной совокупности и при сравнении долей качественного признака двух совокупностей

ЗАДАЧИ

39 Средний процент жира в молоке за лактацию коров холмогорских помесей был следующим 3,4, 3 6, 3,2, 3,1, 2,9, 3,7, 3,2, 3,6, 4,0, 3,4, 4 1, 3,8, 3,4, 4,0, 3,3, 3,7, 3,5, 3,6, 3,4, 3 8

Определите $\bar{x}$, σ и $s_{\bar{x}}$. Установите доверительные границы для $\bar{x}$ при вероятности 0,99, при вероятности 0,95

40 На 400 растениях гибридной ржи первые цветки появлялись в среднем на 70,5 дня после посева Среднее квадратическое отклонение было 6,9 дня Определите среднюю ошибку для $\bar{x}$ и доверительные границы при вероятности 0,95

41 При изучении длины листьев садовой земляники были получены $\bar{x} = 7{,}86$ см, $\sigma = 1{,}32$ см Так как $n = 502$, то $s_{\bar{x}} = \pm 0{,}06$ см

Определите доверительные интервалы для средней арифметической генеральной совокупности с уровнями значимости 0,01, 0,02 и 0,05 Можно ли пользоваться в данном случае таблицей нормального интеграла вероятности?

42 Было измерено 9 листочков земляники Получены значения $\bar{x} = 5{,}0$ см, $\sigma = 1{,}5$ см, $s_{\bar{x}} = \pm 0{,}5$ см

Каковы доверительные интервалы для $\bar{x}$ при уровнях значимости 0,05, 0,01, 0,001?

43 При изучении содержания жира в молоке (в %) было показано, что:

у 50 коров черно-пестрой породы $\bar{x} = 3{,}58$; $\sigma = 0{,}12$

у 12 коров джерзейской породы $\bar{x} = 6{,}04$; $\sigma = 0{,}43$

Надо установить, достоверна ли разница по проценту жира в молоке между джерзейской и черно пестрой породой скота

44 Для 7 коров известны следующие данные об их убойном весе (в кг) в теплом состоянии (1) и после охлаждения (2):

1	2
322,6	318,9
250,6	247,0
287,3	279,7
408,1	403,0
338,0	334,0
213,5	209,3
323,3	319,2

Определите достоверность разницы между средним убойным весом в теплом состоянии и средним убойным весом после охлаждения двумя способами: путем сравнения $\bar{x}$ обоих рядов и путем обработки разниц между двумя убойными весами каждой коровы, как вариационного ряда

45 На 10 парах крыс определяли биологическую ценность белков земляного ореха сырого *(P)* и жареного *(R)* Пары данных (в условных единицах) были следующими: 61,55, 60,54, 56,47, 63,59, 56,51; 63,61; 59,57; 56,54; 44,63, 61,58 Достоверна ли разница? Какой метод можно применить для установления ошибки разницы? Насколько изменятся результаты, если исключить резко отличающуюся от остальных пару данных 44,63? Достаточны ли полученные данные для того, чтобы можно было сделать какой-либо вывод?

46 Для определения *pH* применили 2 типа электродов При первом показания *pH* были 5,78, 5,74; 5,84, 5,80; при втором —5,82; 5,87; 5,96, 5,89. Следует ли отбросить нулевую гипотезу?

47. Пробы по 15 зерен кукурузы разных стадий зрелости проверяли на устойчивость к раздавливанию Первая проба дала следующие цифры (в единицах давления): 42, 50, 36, 34, 45, 56, 42, 53, 25, 65, 33, 40, 39, 43, 42; вторая — 43, 44, 51, 40, 29, 49, 39, 59, 43, 48, 67, 44, 46, 54, 64. Прозерьте, достоверно ли различие между двумя $\bar{x}$?

48. На двух группах бычков сравнивали влияние на их суточный привес подкормок: льняного жмыха и сои. Привесы (в фунтах) бычков, получавших льняной жмых, были: 1,95, 2,17, 2,06, 2,11, 2,24; 2,52, 2,04; 1,95; получавших сою: 1,82, 1,85, 1,87; 1,74, 2,04, 1,78, 1,76; 1,86. Вычислите разницу между средними и установите ее достоверность. Какой таблицей надо пользоваться для установления достоверности?

49. Было изучено общее содержание азота в плазме крови крыс-альбиносов в возрасте 37 и 180 дней. Результаты были выражены в граммах на 100 куб см плазмы. В возрасте 37 дней 9 крыс имели (в г): 0,98; 0,83, 0,99, 0,86, 0,90; 0.81; 0,94; 0,92 и 0,87 В возрасте 180 дней 8 крыс имели (в г): 1,20, 1,18; 1,33, 1,21; 1,20, 1,07;

1,13 и 1,12. Установите доверительные интервалы для разницы с вероятностью 0,95

50 Для изучения влияния рационов с добавкой 10 мкг витамина B_{12} на рост свиней было составлено попарно 16 групп, в каждой из которых было по 6 голов Средние суточные привесы в фунтах (на 100 фунтов живого веса) представлены в следующей таблице:

Рационы	Пары групп							
	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
	0,04	0,16	0,23	0,15	0,16	0,23	0,18	0,21

Какова достоверность разницы?

51. Был проведен опыт по подкормке 32 свиноматок препаратом афаромом, содержавшим железо и медь, в целях уменьшения процента мертворожденных поросят От каждой матки получали: 1 опорос, когда маткам добавляли в корм афаром, и 1 опорос контрольный, когда добавки препарата не было Маток покрывали всегда одними и теми же хряками Были получены следующие результаты:

Номера маток	Мертворожденных, %		Номера маток	Мертворожденных, %	
	при афароме	без афарома		при афароме	без афарома
1	0	8,3	12	11,1	0
2	0	12,5	13	11,1	0
3	0	9,1	14	0	25,0
4	18,2	22,2	15	0	9,1
5	0	10,0	16	0	14,3
6	25,0	33,3	17	0	35,7
7	10,0	0	18	0	63,6
8	11,1	0	19	0	9,1
9	0	16,7	20	0	10,0
10	0	28,6	21	22,2	40,0
11	0	25,0	22—32	0	0

Установите, достоверна ли разница в проценте мертворожденных между опытной и контрольной группами? Можно ли применить метод обработки значений d как вариационного ряда?

52 Имеются следующие данные об удоях 12 коров-матерей и их дочерей (по полновозрастным лактациям).

Удой матери (1)	3770	3817	2450	3463	3500	5544	3112	3150	3118	3018	4291	3463
Удой дочери (2)	2991	4593	3529	4274	3103	3947	3491	3559	2916	4580	4510	4144

Достоверна ли разница между удоями матерей и дочерей? Какой метод сравнения можно применить?

53 В опытах с рисом сравнивали 2 метода обработки почвы Вариансы контрольной (1) и опытной (2) групп были следующими: $\sigma_1^2 = 16{,}65$ (колич. делянок 7), $\sigma_2^2 = 19{,}67$ (число делянок 4). Достоверны ли различия между вариансами?

54 При скрещивании дрозофил, гетерозиготных по черной окраске тела (ebony) и киноварной окраске глаз (cinnabar), было получено: нормальных по окраске тела и глаз 482, черных с нормальными глазами—156, с киноварными глазами, но с нормальной окраской тела—144 и особей, имевших черное тело и киноварные глаза,—72 Рассчитайте, сколько особей каждой категории мушек ожидалось при расщеплении по формуле 9 : 3 : 3 : 1 и какова ошибка для каждого ожидаемого числа? Сравните полученные значения с ожидаемыми и установите, в каких случаях разница между ними достоверна?

55 По окраске зерен у гороха было получено расщепление 6022 желтых и 2001 зеленых Проверьте, достоверна ли разница между фактически полученным числом желтых зерен и ожидаемым числом при моногибридном расщеплении с 25% рецессивных форм.

56 Из 30 больных определенной болезнью умерло 4 человека. По массовым же статистическим данным частота смертей от этого заболевания была равна 0,133. Достоверна ли разница между фактически наблюдавшейся частотой смертных случаев и ожидаемой? Какими таблицами для определения достоверности t надо пользоваться?

57. При изучении суточных привесов 30 баранчиков выяснилось, что они происходят от 4 разных производителей. Данные о привесах потомков этих производителей были следующими:

Производитель	*A*:	124,	151,	196,	141,	174,	201,	147,	157
„	*B*:	183,	150,	198,	191,	154,	173,	157,	159
„	*C*:	234,	167,	189,	165,	175,	190,	176	
„	*Д*:	173,	184,	277,	214,	182,	191,	204	

Определите $\bar{x}$ и σ^2 для привесов каждой группы баранчиков, общую вариансу и вариансу между группами. Установите с помощью критерия F достоверность разницы между этими вариансами

58 К 20 больным тифом было применено новое лечебное средство *C*, к 20 другим — средство *E*. В первой группе ни одного смертного случая не было, во второй — умерло 10 Достоверна ли разница между группами, т. е можно ли говорить о достоверном эффекте средства *C*?

59 Для популяции мужчин возраста от 25 до 30 лет σ длины тела 4,5 см Для выборочной группы 400 спортсменов $\sigma = 3{,}5$ Случайно ли отклонение в величине σ по длине тела у спортсменов от σ популяции?

60. Для определения содержания хлора в химическом соединении были применены методы A и B. Результаты были следующими (в %):

при применении метода A — 27,5; 27,0; 27,3; 27,6; 27,8;

при применении метода B — 27,9; 26,5; 27,2; 26,3; 27,0; 27,4; 27,3; 26,8.

Примените критерий F для установления разницы между вариансами данных, полученных этими методами (для упрощения вычислений можно от всех отдельных значений x отнять 27,0%).

61. В опыте по откорму 15 баранов получали ежедневно в качестве подкормки по 5 г фосфорной муки. 15 других баранов в среднем того же возраста, веса, происхождения и родившихся в тот же период были контрольными. Суточный привес (в г) был следующим:

в опытной группе:	234	277	214	201	174	167	184	157
	196	173	190	191	141	150	191	
в контрольной группе:	183	154	175	159	157	189	198	165
	176	124	173	182	204	151	147	

Каким методом можно установить, достоверна ли разница между опытной и контрольной группами по суточному привесу? Определите эту разницу и выясните, достоверна ли она. Выясните также, отличались ли опытная и контрольная группы по варианс и достоверно ли это различие.

62. На 18 000 больных было зарегистрировано 72 больных диабетом. Определите процент диабетиков в популяции и его ошибку. Установите доверительные границы для процента диабетиков при вероятности 0,95 (95%).

63. За длительный период наблюдали случаи гастрической геморрагии, при этом оказалось, что за первые 4 года их было 40, 10 из них было со смертным исходом, а за последующие 6 лет — 60 случаев, в том числе 5 смертных. Достоверна ли разница в проценте смертных случаев за первые 4 года и последующие 6 лет?

Глава 5

ИЗМЕРЕНИЕ СВЯЗИ. КОРРЕЛЯЦИЯ

Понятие о корреляции. Изложенные в предыдущих главах методы анализа дают возможность изучать изменчивость животных по каждому отдельному признаку—весу, промерам, плодовитости и др. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или даже нескольких признаков, изменяются ли два признака самостоятельно, независимо друг от друга или может быть изменчивость одного признака в какой-то степени связана с изменчивостью другого.

Связь в изменчивости разнородных признаков называется корреляцией.

Связь или зависимость корреляционную надо отличать от связи функциональной. При функциональной связи каждому значению одного признака (аргумента) соответствует, как правило, одно определенное значение другого признака (функции) или же, если несколько значений, то во всяком случае также вполне определенных. При статистических же зависимостях, которые и называются корреляционными, одной и той же величине первого изменяющегося признака соответствуют различные, не вполне определенные значения второго также изменчивого признака. Обоим признакам свойственна случайная вариация. Функциональная связь имеет место по отношению к каждому отдельному наблюдению. Корреляционная же связь проявляется лишь в среднем для всей совокупности наблюдений. В отношении же отдельных наблюдений она очень неполна и неточна.

Так, например, известно, что существует корреляция между весом животного и его высотой. Это означает, что

более высокие животные обычно тяжелее более низких. Но полного соответствия между изменениями этих признаков нет. В некоторых случаях более низкое животное окажется более тяжелым и наоборот.

Если функциональную связь выразить математически, в виде определенного уравнения, то изменению аргумента соответствует вполне определенное приращение функции. При корреляции же приходится иметь дело с сопряженной вариацией изучаемых признаков. Это выражается в том, что отклонения от средних значений по обоим признакам идут в какой-то степени сопряженно, параллельно. При этом они могут идти или в одном направлении, т. е. с увеличением одного признака другой также увеличивается, или в разных, т. е. с увеличением одного другой уменьшается. Поэтому различают положительную и отрицательную корреляции.

При положительной корреляции зависимость между признаками прямая: при увеличении одного увеличивается и другой. При отрицательной корреляции зависимость между признаками обратная: увеличение одного признака соответственно связано с уменьшением другого. В случае качественных признаков отрицательная корреляция будет обозначать, что присутствие одного признака преимущественно совпадает с отсутствием другого, а при положительной — присутствие одного преимущественно совпадает с присутствием же другого.

В биологических явлениях очень часто приходится иметь дело с сопряженной вариацией различных признаков, т. е. с их корреляцией. Зачастую познание корреляционных зависимостей имеет большое практическое значение. Так, для животновода очень важно знать, какова связь между общим удоем за лактацию и процентом жира в молоке, иначе говоря, дают ли более высокоудойные коровы молоко с повышенным содержанием жира или, наоборот, с пониженным, и насколько часто встречаются исключения из той или другой зависимости. Оценку качества коровы обычно производят по полновозрастной, т. е. по третьей или четвертой лактации. Но насколько была бы облегчена оценка, если бы была установлена положительная корреляция между удоем за первую и удоем за третью лактацию. Тогда можно было бы предсказывать удой коровы за третью лактацию по ее удою за первую лактацию. Степень уверенности, до-

стоверности подобного предсказания, очевидно, зависит от степени корреляции. Вот почему возникает потребность в количественном измерении корреляции. Для этого служит ряд методов, наиболее распространенным из которых является вычисление так называемого коэффициента корреляции.

Коэффициент корреляции и методы его вычисления. Смысл корреляции заключается в сопряженности вариации признаков (пока мы будем говорить о корреляции только двух признаков и такую корреляцию условимся называть простой). Если мы хотим установить наличие корреляционной зависимости и ее степень, нам надо узнать, насколько параллельно идет вариация по двум признакам. Наиболее простым и в то же время очень грубым способом такого сравнения вариации является построение графиков, на которых была бы выражена кривыми вариация признаков у особей данной совокупности изучаемых животных. По тому, насколько параллельно шли бы кривые изменения признаков, можно было бы судить о корреляции. Однако этот изредка применяемый на практике способ не дает никакого мерила корреляционной зависимости, кроме чисто зрительного впечатления о колебаниях линий на графике. Непосредственное сравнение вариации двух признаков затрудняется и тем обстоятельством, что они, как правило, выражены в разных измерениях. Вот почему при изучении корреляции прибегают к нормированному отклонению t.

Напомним, что нормированное отклонение t представляет собой отклонение тех или других вариант от их средней арифметической, выраженное в долях среднего квадратического отклонения. Выражая отклонения отдельных особей от средних арифметических по обоим признакам одновременно, можно сопоставлять вариацию по обоим признакам. Так, например, если данный экземпляр лисицы отклоняется от средней популяции по длине туловища на $+2\sigma$, а по длине хвоста на $+1{,}8\sigma$, то это дает уже возможность говорить о наличии какой-то положительной связи или положительной корреляции в изменчивости длины туловища и длины хвоста, правда, пока по одной особи. Но ясно, чем теснее связана вариация по этим двум признакам, тем чаще особи, отклоняющиеся от средней по длине туловища на определенное количество сигм или долей сигм, например на $1{,}0\sigma$,

1,2σ, 1,8σ и т. д., будут занимать такие же места и по длине хвоста, т.е отклоняться также на 1,0σ, на 1,2σ, на 1,8σ и т. д. Наоборот, при отсутствии корреляции, совпадение величин t по обоим признакам будет чисто случайным.

Таким образом, зависимость между t обоих рядов, т. е. между величинами $t_x = \frac{x-\bar{x}}{\sigma_x}$ и $t_y = \frac{y-\bar{y}}{\sigma_y}$ (или отношение между ними), может быть мерилом корреляционной связи.

Оказалось, что нормированные отклонения обладают рядом ценных математических свойств. Приведем некоторые из них в готовом виде без математического обоснования.

Первое свойство заключается в том, что среднее произведение двух нормированных отклонений, то есть

$$\frac{\Sigma t_y \cdot t_x}{n},$$

колеблется от 0 до единицы.

Среднее произведение нормированных отклонений $\frac{\Sigma t_y \cdot t_x}{n}$ можно записать и в иной форме, а именно $\overline{t_x \cdot t_y}$. Черта наверху, охватывающая обе буквы, будет обозначать, что взята средняя.

При полном отсутствии связи между изучаемыми признаками $\overline{t_y \cdot t_x} = 0$, а при полной, т. е. уже функциональной связи между признаками, $\overline{t_x \cdot t_y} = 1$.

Второе свойство среднего произведения t заключается в том, что его знак будет разным в зависимости от типа связи: если увеличивающемуся значению одного признака соответствует увеличивающееся значение второго—знак плюс, если с увеличением значения одного признака значение второго уменьшается—знак минус.

Наконец, оказалось, что теми же свойствами характеризуется не только среднее произведение $\overline{t_y \cdot t_x}$, но и их среднее отношение $\left[\frac{t_y}{t_x}\right]$.

Квадратные скобки в данном случае будут обозначать, что отношение также берется в среднем.

Вот почему обе эти величины были приняты как мерило тесноты корреляционной связи двух признаков и

получили название коэффициента корреляции, который обозначается буквой r.

Таким образом,

$$r = \overline{t_y \cdot t_x}$$

или

$$r = \left[\frac{t_y}{t_x}\right].$$

Последнее выражение мы сейчас рассматривать не будем. Укажем только, что оно приводит к так называемому основному корреляционному уравнению или уравнению регрессии $t_y = r \cdot t_x$, преобразование которого дает обычное уравнение прямой $y = a + bx$. С ним мы познакомимся в главе 6, когда будем говорить о регрессии.

Первое же выражение является общим видом формулы, применяемой при вычислениях простого коэффициента корреляции

$$r = \frac{\Sigma t_x t_y}{n}. \qquad (31)$$

Так как $t_x = \frac{(x - \bar{x})}{\sigma_x}$, а $t_y = \frac{(y - \bar{y})}{\sigma_y}$,

то

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y} \qquad (32)$$

или

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{(n-1) \sigma_x \sigma_y}. \qquad (32a)$$

В формуле (32) величина $(x - \bar{x})$ обозначает отклонение каждой изучаемой особи от средней $\bar{x}$ одного признака, распределенного по ряду x, а величина $(y - \bar{y})$ — отклонение той же особи от средней $\bar{y}$ другого признака, выражающегося рядом y. Таким образом, для того чтобы получить числитель дроби, надо учесть отклонения каждой особи от средних по обоим признакам, перемножить их и затем просуммировать. В знаменателе же формулы уже известные величины: n—число особей, σ_x — среднее квадратическое отклонение ряда по признаку x (или просто ряда x) и σ_y—среднее квадратическое отклонение ряда по признаку y (ряду y).

Модификации общей формулы коэффициента корреляции. Практически можно использовать как эту общую формулу, так и ее видоизменения, которые легко получить путем алгебраического преобразования числителя и знаменателя. Так, если произвести перемножение в числителе, то формула преобразуется в следующую:

$$r = \frac{\Sigma xy - n\bar{x}\cdot\bar{y}}{n\,\sigma_x\,\sigma_y}. \qquad (33)$$

В этой формуле используются заранее вычисленные значения $\bar{y}$ и $\bar{x}$.

Так как $$n\overline{yx} = n\frac{\Sigma y}{n}\,\frac{\Sigma x}{n} = \frac{\Sigma y\ \Sigma x}{n},$$

то $$r = \frac{\Sigma xy - \frac{\Sigma x\ \Sigma y}{n}}{n\,\sigma_x\,\sigma_y}. \qquad (34)$$

Удобство этой формулы в том, что вместо средних арифметических берутся сумма вариант ряда x и сумма вариант ряда y.

Если и числитель и знаменатель разделить на n, то мы получим формулу, в которую входят различные средние:

$$r = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x\,\sigma_y}. \qquad (35)$$

Далее, можно проделать некоторые упрощения со знаменателем, чтобы не пользоваться средними квадратическими отклонениями.

Для этого вместо $n\sigma_x\,\sigma_y$ запишем

$$n\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}\cdot\frac{\Sigma(y-\bar{y})^2}{n}}.$$

При перенесении n под корень знаменатель примет вид

$$\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}.$$

Его можно внести в любую из указанных выше формул, например в (32).

Тогда $$r = \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}}. \qquad (36)$$

Коэффициент корреляции в данном случае выражен только с помощью отклонений от средних, так что все вычисления становятся однотипными. Во многих случаях она наиболее удобна для практического использования.

Замена знаменателя $n\,\sigma_x\,\sigma_y$ через $\sqrt{\Sigma (x-\bar{x})^2\ \Sigma (y-\bar{y})^2}$ может быть сделана и в остальных формулах. Формула (33) будет тогда выглядеть следующим образом:

$$r = \frac{\Sigma yx - n\,\bar{y}\,\bar{x}}{\sqrt{\Sigma (x - \bar{x})^2\ \Sigma (y - \bar{y})^2}}. \qquad (33а)$$

Чтобы освободиться от необходимости вычислять непосредственно квадраты отклонений, можно их заменить равными им величинами:

$$\Sigma x^2 - n\bar{x}^2 \quad \text{и} \quad \Sigma y^2 - n\bar{y}^2.$$

Тогда $$r = \frac{\Sigma yx - n\,\bar{y}\,\bar{x}}{\sqrt{(\Sigma x^2 - n\bar{x}^2)\ (\Sigma y^2 - n\bar{y}^2)}}. \qquad (33б)$$

Этой формулой удобнее пользоваться в тех случаях, когда отклонения от средних выражаются слишком дробными числами.

Мы приводим эти модификации основной формулы коэффициента корреляции потому, что в различных руководствах и работах можно встретиться с неодинаковыми методами вычисления r. Кроме того, и конкретные условия полученного фактического материала могут побудить отдать предпочтение одной из перечисленных формул в зависимости от того, какие показатели легче вычислить и с какими легче оперировать, надо ли группировать полученные данные в классы или без этого можно обойтись, велико ли количество наблюдений или мало и т. д.

Примеры использования различных формул для вычисления r. Допустим, что в нашем распоряжении имеют-

ся следующие данные о весах при рождении 10 бычков и среднем суточном привесе их за период от рождения до годичного возраста.

Вес, кг	Прирост, г	Вес, кг	Прирост, г
38,5	694	44,0	743
46,0	901	38,0	896
43,0	736	35,0	863
43,0	1005	40,5	855
40,5	841	54,0	830

Возникает вопрос, какую формулу выгоднее применить для вычисления коэффициента корреляции. Так как и вес, и прирост выражаются довольно большими числами (и первое из них еще с десятичными дробями), формулы, в которые входят отклонения вариант от средней, менее желательны. Если в нашем распоряжении есть арифмометр и таблицы квадратов, то лучше всего воспользоваться формулой (33б).

Нам понадобится составить таблицу, в которой были бы наряду с исходными данными x и y также величины yx, x^2 и y^2 (табл. 21).

Таблица 21

Данные для вычисления коэффициента корреляции между живым весом бычков при рождении x и средним суточным привесом y

x, кг	y, г	x^2	y^2	xy
38,5	694	1482,25	481636	26719,0
46,0	901	2116,00	811801	41446,0
43,0	736	1849,00	541696	31648,0
43,0	1005	1849,00	1010025	43215,0
40,5	841	1640,25	707281	34060,5
44,0	743	1936,00	552049	32692,0
38,0	896	1444,00	802816	34048,0
35,0	863	1225,00	744769	30205,0
40,5	855	1640,25	731025	34627,5
54,0	830	2916,00	688900	44820,0
Σ=422,5	8364	18097,75	7071998	353481,0

Так как $n=10$, то легко вычислить необходимые величины:

$$\bar{x} = 42{,}25\ ,$$
$$\bar{y} = 836{,}4\ ,$$
$$n\bar{x}\bar{y} = 353379\ ,$$
$$n\bar{x}^2 = 17850{,}625\ ,$$
$$n\bar{y}^2 = 6995649{,}6.$$

В табл. же 21 имеются следующие итоговые цифры:

$$\Sigma xy = 353481\ ,$$
$$\Sigma x^2 = 18097{,}75\ ,$$
$$\Sigma y^2 = 7071998.$$

Подставив все эти величины в формулу (33б), получим:

$$r = \frac{353481 - 353379}{\sqrt{(18097{,}75 - 17850{,}625)\ (7071998 - 6995649{,}6)}} = +0{,}023$$

Таблица 22

Рост братьев x и рост сестер y (в дюймах)

Номера пар	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	71	69	2	4	5	25	10
2	68	64	−1	1	0	0	0
3	66	65	−3	9	1	1	−3
4	67	63	−2	4	−1	1	2
5	79	65	1	1	1	1	1
6	71	62	2	4	−2	4	−4
7	70	65	1	1	1	1	1
8	73	64	4	16	0	0	0
9	72	66	3	9	2	4	6
10	65	59	−4	16	−5	25	20
11	66	62	−3	9	−2	4	6
	$\bar{x}=69$	$\bar{y}=64$		$\Sigma(x-\bar{x})^2=74$		$\Sigma(y-\bar{y})^2=66$	$\Sigma(x-\bar{x})(y-\bar{y})$ $46-7=39$

В качестве второго примера приведем данные одного из первых биометриков Пирсона о корреляции в росте

(длине тела) братьев и сестер по 11 парам брат—сестра (табл. 22) и обработаем их таким образом, чтобы можно было применить формулу (36), в которой все исходные величины являются только отклонениями.

Тогда

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}} = \frac{+39}{\sqrt{74 \cdot 66}} = +0{,}56.$$

Таким образом, если по техническим соображениям легче работать с величинами x^2, y^2 и xy, надо применять формулу (33б), если же легче вычислять $(x—\bar{x})$, $(y—\bar{y})$ и их квадраты (а также произведение), то удобнее формула (36). Применение этих формул освобождает от необходимости самостоятельного вычисления сигм.

Во всех этих формулах применяется прямой путь вычисления r на основе использования средних $\bar{x}$ и $\bar{y}$ и отклонений от них.

Непрямой путь вычисления r, формула Бравэ. При значительном числе особей в изучаемой совокупности техника вычисления r может быть упрощена подобно тому, как это было сделано в случае вычисления $\bar{x}$ и σ, путем учета отклонений не от $\bar{x}$ или $\bar{y}$, а от условных средних A_x и A_y.

В этом случае применяется следующая рабочая формула для коэффициента корреляции (формула Бравэ):

$$r = \frac{\Sigma f a_x a_y - n b_x b_y}{n \sigma_x \sigma_y}. \qquad (37)$$

Эта формула отличается от общей формулы (32) только структурой числителя. Она позволяет проделать все вычисления как для получения сигмы каждого ряда, так и для вычисления r, пользуясь одной и той же таблицей, которая называется корреляционной решеткой. Все величины в формуле могут быть взяты в условных значениях, т. е. без умножения на величины классовых промежутков обоих вариационных рядов i_x и i_y. Так как обе эти величины входят во все части формулы и в знаменателе, и в числителе, то они сокращаются. Замена в знаменателе величины n величиной $n—1$ очень мало влияет на величину r, поэтому в ней нет практической необходимости.

Таблица 23

Корреляционная решетка для удоев за лактацию x и среднего процента жира в молоке y 100 коров холмогоро-печорских помесей

y \ x	4000—4499	4500—4999	5000—5499	5500—5999	6000—6499	6500—6999	7000—7499	7500—7999	8000—8499	f	a	fa	fa^2
2,9—3,0	1					1				2	—3	—6	18
3,1—3,2			3	1		1			1	7	—2	—14	28
3,3—3,4	1	3	2	4	1	2	1	1	2	17	—1	—17	17
3,5—3,6	3	2	7	6	1					19	0	0	0
3,7—3,8	7	6	6	5						24	1	24	24
3,9—4,0	7	7	1		1	1				17	2	34	68
4,1—4,2	3	5								8	3	24	72
4,3—4,4	1	2	1				2			6	4	24	96
f	23	25	20	16	4	5	3	1	3	100		+69	323
a	—2	—1	0	1	2	3	4	5	6				
fa	—46	—25	0	16	8	15	12	5	18	+3			
fa^2	92	25	0	16	16	45	48	25	108	375			

$$b_x^* = \frac{+3}{100} = 0{,}03; \quad \sigma_x^* = \sqrt{\frac{375}{100} - 0{,}03^2} = 1{,}93;$$

$$b_y^* = \frac{+69}{100} = 0{,}69; \quad \sigma_y^* = \sqrt{\frac{323}{100} - 0{,}62^2} = 1{,}66.$$

(Звездочками отмечены значения b и σ, не умноженные на i.)

Для составления корреляционной решетки и вычисления коэффициента корреляции используем данные об удоях по максимальной лактации и средней жирности молока за эту же лактацию 100 коров холмогоро-печорских помесей, взятых без всякого выбора из группы учтенных нами в 1954—1957 гг. лучших коров колхозов и совхозов Коми АССР. По каждой корове имелись два показателя: удой за лактацию и средний процент жира в молоке. Так как учитывались коровы с удоями не ниже 4000 л, то нижним лимитом является удой 4000 л, а верхним—8400 л (для упрощения удои округлены). Жирность молока колебалась от 3,0 до 4,4%. Классы для вариационных рядов: по удою молока за лактацию—4000—4499 л, 4500—4999 л, 5000—5499 л, 5500—5999 л и т. д.; по проценту жира—2,9—3,0, 3,1—3,2, 3,3—3,4 и т. д. Определив классы, следует построить корреляционную решетку. На двух сторонах квадрата (вверху по горизонтали и слева по вертикали) надо нанести значения классов обоих рядов, как это сделано в табл. 23. Центральные значения классов в данном случае не требуются.

В макет корреляционной решетки надо разнести значения всех 100 коров. Однако особенностью этой разноски является то, что надо вносить в соответствующие клетки решетки данные об изученных особях одновременно по обоим признакам. Так, например, если первые 3 коровы из списка имели показатели удоя: корова № 1—4100 л и 3,0% жира; корова № 2—5700 л и 3,1% жира и корова № 3—4900 л и 4,4% жира, то корова № 1 должна быть внесена в клетку на пересечении класса в 4000—4499 л по удою и класса 2,9—3,0 по проценту жира, то есть в верхнюю левую клетку; корова № 2 должна быть внесена в клетку на пересечении классов 5500—5999 и 3,1—3,2, то есть в четвертую клетку второго горизонтального ряда; корова № 3—во вторую клетку самого нижнего горизонтального ряда. При разноске можно пользоваться тем же приемом, который был применен при построении вариационного ряда, а именно точками и соединяющими черточками.

В табл. 23 разноска 100 коров уже проведена. Цифры, стоящие в каждой клетке, обозначают, таким образом, число коров, имеющих удой и процент жира согласно классам. Суммы всех особей в горизонтальных строч-

ках пишутся справа (ряд y по проценту жира), суммы всех особей в вертикальных столбцах пишутся внизу (ряд x по удою) Справа внизу в угловой клетке надо записать сумму всех особей (100) Она относится как к ряду x, так и к ряду y

С помощью корреляционной решетки можно получить все величины, необходимые для вычисления коэффициента корреляции Величина n в нашем случае равна 100 Чтобы получить b_x и b_y, а также σ_x и σ_y, достаточно обработать 2 вариационных ряда ряд x (по удою за лактацию) и ряд y (по проценту жира в молоке), которые расположены справа и внизу в этой же таблице Если в ней нет места, можно выписать эти ряды на отдельных листках Однако предпочтительнее первое, как это будет видно в дальнейшем Поэтому в корреляционной решетке, представленной в табл 23, проделана и обработка вариационных рядов x и y

Вновь обращаем внимание на то, что вычисленные для обоих рядов значения b и σ в данном случае неполные, как бы условные

Они не умножены на величины классовых промежутков i, поэтому отмечены звездочками Для ряда x величина классового промежутка равна 500 л, а для ряда y— 0,2% жира Очевидно, что для получения окончательных значений b и σ (и $\bar{x}$, если она нужна) надо ввести поправки на i Для вычисления же коэффициента корреляции умножать на i_x и соответственно на i_y нет необходимости только потому, что они все равно сокращаются, находясь в числителе и в знаменателе формулы

Осталось вычислить последнюю величину для включения в формулу коэффициента корреляции—первый член знаменателя $\Sigma f a_x a_y$.

Для этого можно воспользоваться той же корреляционной решеткой табл 23, но для лучшего уяснения всех действий нужно переписать ее вновь без граф fa и fa^2 (табл 24)

Величина $\Sigma f a_x a_y$ представляет собой сумму произведений отклонений каждого класса от условной средней по ряду x и от условной средней по ряду y (a_x a_y), умноженных на число особей в данном классе f В табл 24 клетки тех классов, которые были приняты за условные средние как по ряду x, так и по ряду y, можно зачеркнуть или выделить жирными линиями, так как для

каждой такой клетки произведение $a_x a_y$ равно нулю. Для всех остальных клеток корреляционной решетки надо получить произведения отклонений каждого класса по ряду x и по ряду y, перемножить их и записать в углу каждой клетки, как это сделано в таблице.

Таблица 24

Корреляционная решетка для удоев за лактацию x и среднего процента жира в молоке y 100 коров холмогоро-печорских помесей[1]

y \ x	4000—4499	4500—4999	5000—5499	5500—5999	6000—6499	6500—6999	7000—7499	7500—7999	8000—8499	f	a
2,9—3,0	6 / 1					9 / 1				2	−3
3,1—3,2			0 / 3	2 / 1	4 / 1	6 / 1			12 / 1	7	−2
3,3—3,4	2 / 1	1 / 3	0 / 2	1 / 4	2 / 1	3 / 2	4 / 1	5 / 1	6 / 2	17	−1
3,5—3,6	0 / 3	0 / 2	0 / 7	0 / 6	0 / 1					19	0
3,7—3,8	2 / 7	1 / 6	0 / 6	1 / 1						24	+1
3,9—4,0	4 / 7	2 / 7	0 / 1		4 / 1	6 / 1				17	+2
4,1—4,2	6 / 3	3 / 5								8	+3
4,3—4,4	8 / 1	4 / 2	0 / 1				16 / 2			6	+4
f	23	25	20	16	4	5	3	1	3	100	
a	−2	−1	0	+1	+2	+3	+4	+5	+6		

[1] Эта таблица дана для вычисления произведений $a_x a_y$; они записаны в верхнем правом углу каждой клетки решетки, где имеются особи.

Разберем теперь, как были вычислены необходимые произведения отклонений В клетке, расположенной в верхнем левом углу решетки, помечена лишь одна особь (класс по удою 4000—4999 л и по проценту жира 2,9—3,0). Чтобы получить для нее произведение $a_x\ a_y$, надо посмотреть на расположенные справа и внизу вариационные ряды Из них видно, что для данной особи отклонение $a_x = -2$, а отклонение $a_y = -3$ Отсюда произведение $a_x\ a_y = (-2)\ (-3) = 6$ Цифра 6 записана в верхнем правом углу этой клетки Можно взять для примера какую-либо другую клетку, например в нижнем ряду третью справа, где помечены две особи Для данной клетки $a_x\ a_y = (+4)\ (+4) = 16$ Для последней правой клетки третьего ряда $a_x\ a_y = (-1).(+6) = -6$ Однако в табл 24 произведения отклонений даны без знака Дело в том, что они будут иметь всегда определенный знак в зависимости от того, в какой части корреляционной решетки расположены соответствующие клетки Решетка разделяется нулевыми рядами (горизонтальным и вертикальным) на четыре части, называемые обычно квадрантами В левом верхнем квадранте и a_x, и a_y отрицательны, поэтому их произведения всегда будут иметь знак плюс В нижнем правом квадранте и a_x, и a_y являются положительными величинами, их произведения также положительны В правом верхнем квадранте все a_x имеют знак плюс, а a_y—знак минус, поэтому их произведения являются величинами отрицательными В левом нижнем квадранте, наоборот, a_y имеет знак плюс, а a_x —знак минус, их произведение будет также величиной отрицательной. Вот почему писать в каждой клетке знак плюс или минус нет необходимости, достаточно запомнить знак для каждого квадранта решетки Не нужно также писать произведения $a_x a_y$ в тех клетках, где не было помечено ни одной особи

После записи значений $a_x\ a_y$ в клетках корреляционной решетки надо помножить каждое произведение $a_x\ a_y$ на число особей в классе и после этого просуммировать все произведения Лучше это сделать в подсобной таблице (табл. 25).

Таблица 25

Произведения $fa_x a_y$ для всех клеток табл. 24 по квадрантам

I квадрант (знак +)

$1 \cdot 6 = 6$
$1 \cdot 2 = 2$
$3 \cdot 1 = 3$
$+11$

2 квадрант (знак —)

$1 \cdot 2 = 2$
$4 \cdot 1 = 4$
$1 \cdot 4 = 4$
$1 \cdot 2 = 2$
$1 \cdot 9 = 9$
$1 \cdot 6 = 6$
$2 \cdot 3 = 6$
$1 \cdot 4 = 4$
$1 \cdot 5 = 5$
$1 \cdot 12 = 12$
$2 \cdot 6 = 12$
-66

3 квадрант (знак —)

$7 \cdot 2 = 14$
$7 \cdot 4 = 28$
$3 \cdot 6 = 18$
$1 \cdot 8 = 8$
$6 \cdot 1 = 6$
$7 \cdot 2 = 14$
$5 \cdot 3 = 15$
$2 \cdot 4 = 8$
-111

4 квадрант (знак +)

$5 \cdot 1 = 5$
$1 \cdot 4 = 4$
$2 \cdot 16 = 32$
$1 \cdot 6 = 6$
$+47$

По данным табл. 25 можно получить значение $\Sigma fa_x a_y$. Оно равно $+11-66-111+47=+58-177=-119$.

Теперь можно проставить все полученные значения в формулу коэффициента корреляции:

$$r = \frac{\Sigma fa_x a_y - n b_x b_y}{n \sigma_x \sigma_y} = \frac{-119 - 100 \cdot 0{,}03 \cdot 0{,}69}{100 \cdot 1{,}66 \cdot 1{,}93} = -0{,}38.$$

Возможные значения коэффициента корреляции. Прежде всего необходимо обратить внимание на знак при коэффициенте корреляции. При положительной корреляции r будет иметь знак плюс, при отрицательной — знак минус. В нашем примере знак минус указывает на отрицательную корреляцию между величиной удоев за лактацию и процентом жира в молоке, т. е. на то, что в изученной группе коров с увеличением удоев жирность молока несколько снижается.

Коэффициенты корреляции могут колебаться от 0 до $+1$ при положительной корреляции и от 0 до -1 при отрицательной корреляции. Если $r=0$, то это означает, что изменчивость обоих признаков происходит независимо. При значениях r, близких к 1, изменчивость обоих признаков взаимосвязана, т. е. с изменением одного признака меняется и другой (в том же направлении—при

положительной корреляции и в противоположном направлении—при отрицательной корреляции).

Предварительные выводы о характере связи можно сделать из анализа расположения вариант в корреляционной решетке, что видно из сравнения следующих трех таблиц, в которых схематически показано возможное распределение вариант по отдельным клеткам решетки при разных типах корреляции. При высокой положительной корреляции варианты расположены в корреляционной решетке преимущественно вблизи диагонали, проходящей от верхнего левого угла к нижнему правому. Это показано в табл. 26.

Таблица 26

Распределение вариант в корреляционной решетке при высокой положительной корреляции[1]

y \ x	1	2	3	4	5	6	7
1	а	в	е				
2	б	д	з	л			
3	г	ж	к	н	р		
4		и	м	п	т	х	
5			о	с	ф	ч	э
6				у	ц	щ	я
7					ш	ю	а

При высокой отрицательной корреляции варианты расположены вокруг другой диагонали—от верхнего правого угла к нижнему левому, как это представлено в табл. 27.

Таблица 27

Распределение вариант в корреляционной решетке при высокой отрицательной корреляции

y \ x	1	2	3	4	5	6	7
1					ш	ю	а
2				у	ц	щ	я
3			о	с	ф	ч	э
4		и	м	п	т	х	
5	г	ж	к	н	р		
6	б	д	з	л			
7	а	в	е				

[1] В табл. 26, 27, 28 классы по рядам x и y отмечены цифрами, количество вариант в классах—буквами.

Наконец, если варианты расположены равномерно по решетке, наблюдается сгущение вариант ближе к средним по ряду *x* и соответственно по ряду *y*, как это требуется по законам случайной вариации. В этом случае можно говорить о независимом варьировании признаков *x* и *y*, т. е. об отсутствии корреляции между ними (табл. 28).

Таблица 28

Распределение вариант по классам корреляционной решетки при отсутствии корреляции

y \ x	1	2	3	4	5	6	7
1				ш			
2			т	у	ф		
3		з	и	ь	л	м	
4	а	б	в	г	д	е	ж
5		и	о	п	р	с	
6			х	ц	ч		
7				щ			

Однако для более точного измерения степени связи в изменчивости двух величин необходимо вычислить коэффициент корреляции. Какие же значения *r* можно считать большими, а какие средними и малыми? С первого взгляда может показаться, что величина *r*, близкая к 0,5, является достаточно высоким коэффициентом корреляции и что при этом совпадение вариации двух признаков должно быть в 50% случаев. На самом деле это не так. Даже при полном отсутствии корреляции будут случаи, когда отклонения от средних по обоим признакам для данной особи окажутся примерно одинаковыми, иначе говоря, когда по обоим признакам особь будет находиться примерно в одном и том же месте вариационного ряда. Так, в табл. 28 группы вариант, отмеченные буквами г, д, п, р, расположены примерно в одних и тех же порядковых классах обоих рядов *x* и *y*, а именно в классах, обозначенных цифрами 4 и 5. Таким образом, и при отсутствии корреляции будет наблюдаться в силу закономерностей случайной вариации как случайное совпадение в вариации двух признаков, так и случайное несовпадение примерно в равном соотношении. При на-

личии же корреляции какая-то доля изменчивости одного признака будет вполне закономерно определяться изменчивостью другого признака. Математическая теория корреляции показывает, что степень «связанности» в вариации двух величин более точно измеряется квадратом коэффициента корреляции r^2. Это значит, что при $r=0{,}5$ 25% изменчивости одного признака объясняется изменчивостью другого признака, по остальной же части изменчивости соотношение между признаками чисто случайное. При $r=0{,}3$ менее 10% изменчивости объясняется таким же образом. При $r=0{,}7$ около 50% изменчивости одного признака определяется изменчивостью другого признака. При таком же коэффициенте корреляции, как 0,9, действительно 81% изменчивости одного признака закономерно связан с изменчивостью другого признака, в остальных же 19% случаев совпадение или несовпадение изменчивости двух признаков является чисто случайным.

Таким образом, хотя коэффициент корреляции и указывает на общность элементов в двух коррелированных рядах признаков, но не вся эта общность объясняется закономерной связью в изменчивости двух признаков.

Из сказанного ясно, что о тесной корреляции можно говорить только в тех случаях, когда r не ниже 0,7. Коэффициенты корреляции порядка 0,5—0,6 следует считать средними, коэффициенты же ниже 0,5 указывают на слабую связь.

Оценка достоверности коэффициента корреляции и сравнение двух коэффициентов корреляции. Критерием для применения того или иного способа оценки является характер распределения показателей выборок и их ошибок, что, в свою очередь, зависит от численности изучаемой группы. Если распределение данного показателя нормально, можно пользоваться таблицами нормального интеграла вероятности при большом n и распределения t по Стюденту—при малом (меньше 30). Однако для вычисления средней ошибки мы пользовались одними и теми же формулами как при больших, так и при малых выборках. Иначе обстоит дело, когда необходимо оценить коэффициент корреляции. При больших выборках (в данном случае $n>100$) и при не очень высоком коэффициенте корреляции среднюю

ошибку для r можно вычислить по формуле

$$s_r = \frac{1 - r^2}{\sqrt{n}}. \qquad (38)$$

Для примера, приведенного в табл. 23 и 24,

$$s_r = \frac{1 - (0{,}38)^2}{\sqrt{100}} = 0{,}0856.$$

Так как $t = \frac{r}{s_r}$ (40), то $t = \frac{0{,}38}{0{,}0856} = 4{,}4$.

Очевидно, что коэффициент корреляции обладает высокой достоверностью. Уровень значимости очень высокий, ниже чем 0,01. Нулевая гипотеза, что $r=0$, опровергается.

Значения коэффициента корреляции, определенные по малым выборкам (с ними в биологических работах особенно часто приходится иметь дело), могут сильно отличаться от его значения для генеральной совокупности. Кроме того, и распределение r при малых выборках может значительно отличаться от нормального, поэтому применение формулы (38) в тех случаях, когда численность изученных групп мала, может привести к неверной оценке достоверности коэффициента корреляции. Оказались необходимы более точные методы оценки. Одним из них является следующая формула для s_r:

$$s_r = \frac{\sqrt{1 - r^2}}{\sqrt{n-2}}. \qquad (39)$$

Отсюда

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n-2}. \qquad (40а)$$

По данным табл. 22 был получен коэффициент корреляции $r=+0{,}56$ при $n=11$.

В таком случае $t = \frac{0{,}56}{\sqrt{1-0{,}56^2}} \cdot \sqrt{11-2} = 2{,}45$.

В табл. II t-распределения по Стюденту нет колонки с $n=11$ и строчки с $t=2{,}45$. Поэтому надо взять цифры вероятности, средние между значениями $n=10$ и $n=12$ и значениями $t=2{,}4$ и $t=2{,}5$. Получаем вероятность

0,965, т е уровень значимости 0,03 Если исходить из жестких требований к уровню значимости, т е принять его за 0,01, надежность полученного коэффициента корреляции недостаточна Однако, так как часто принимается и уровень значимости 0,05, полученный коэффициент корреляции этому условию удовлетворяет Если на этом же примере применить для вычисления ошибки формулу (38), то

$$t = \frac{0{,}56}{0{,}6864} \sqrt{11} = 2{,}79.$$

Значение t значительно больше, чем вычисленное по формуле (39), тем самым сильно завышается и оценка достоверности (получается, что вероятность достоверности равна 0,981 вместо 0,965) Чтобы не усложнять работы вычислением ошибки и последующим обращением к таблице распределения t, можно пользоваться табл VI, с помощью которой легко определить достоверность r при разных n непосредственно по значению коэффициента корреляции

В табл VI даны два уровня значимости 0,05 и 0,01 В более подробных таблицах приводятся и иные уровни значимости Чтобы можно было считать полученный коэффициент корреляции достоверным, он должен превышать табличное значение при данном n Так, например, при уровне значимости 0,05 и $n=11$ коэффициент корреляции должен быть не менее 0,55. При $n=11$ и уровне значимости 0,01 r должен быть не менее 0,68 Полученный выше коэффициент корреляции 0,56 при $n=11$ удовлетворяет уровню значимости 0,05, но не удовлетворяет уровню значимости 0,01, как и было установлено выше с помощью формул

Из табл. VI видно, что значимость одного и того же коэффициента корреляции может быть весьма различной в зависимости от величины выборки, по которой он вычислен. Так, например, если $r=0{,}60$, то при $n=10$ он может быть признан достаточным только при уровне вероятности 0,95, т е при уровне значимости 0,05 Если же принять уровень значимости 0,01, то он не может считаться достоверным Это значит, что при уровне значимости 0,01 нулевая гипотеза не может быть отброшена Если же $n=16$, а $r=0{,}60$, то нулевая гипотеза отвергается при обоих уровнях значимости При $n=8$ $r=0{,}60$

является недостоверным, т. е. нулевая гипотеза остается в силе при обоих уровнях значимости.

При больших n даже значительно меньшие коэффициенты корреляции могут быть достоверными. Так, коэффициент корреляции $r=-0{,}21$ достоверен при $n=150$ с вероятностью 0,99, т. е. при уровне значимости 0,01. При $n=100$ и $r=-0{,}21$ нулевая гипотеза не может быть отвергнута при уровне значимости 0,01. Наконец, при $n=70$ коэффициент корреляции $r=-0{,}21$ недостоверен. Нулевая гипотеза по-прежнему остается в силе даже при уровне значимости 0,05.

В некоторых случаях даже этот улучшенный метод оценки может оказаться недостаточным в связи со значительным отклонением распределения r от нормального (особенно при высоких значениях r). Вот почему Фишером было предложено заменять r другой величиной z. Преимуществом z является то, что распределение величин z значительно ближе к нормальному, чем распределение r. Преобразование r в z производится по определенной формуле, давать которую нет надобности, так как перевод r в z и обратно осуществляется по табл. VII.

Средняя ошибка для z вычисляется по формуле:

$$s_z = \frac{1}{\sqrt{n-3}}. \qquad (41)$$

Оценка достоверности z может производиться, как обычно, с помощью t, при этом:

$$t = \frac{z}{s_z}. \qquad (42)$$

Допустим, что $r=0{,}606$ и $n=10$.

Определяем z по табл VII. Оно будет равно 0,7.

Ошибка для $z = \frac{1}{\sqrt{10-3}} = 0{,}378.$

Тогда $t = \frac{0{,}7}{0{,}378} = 1{,}93.$

По таблице Стюдента (табл. II), при $n=10$ значение $t=1{,}93$ дает вероятность только 0,914, т. е. уровень значимости 0,09. Очевидно, корреляция не доказана. Нулевая гипотеза остается в силе.

Значение метода z заключается еще и в том, что только с помощью z можно определить достоверность или недостоверность разницы между двумя коэффициентами корреляции или между фактически полученным коэффициентом корреляции и теоретически ожидаемым, а также провести объединение данных по нескольким корреляциям, вычисленным на основе малых выборок.

Так, например, при изучении корреляции между высотой в холке и живым весом у крупного рогатого скота были получены следующие коэффициенты корреляции:

На группе черно-пестрого скота $r_1 = +0{,}675$.

На группе рыже-пестрого скота $r_2 = +0{,}761$.

Количество пар наблюдений в каждой группе равно 100.

По табл. VII переводим значения r в значения z. Тогда $z_1 = +0{,}82$ и $z_2 = +1{,}0$.

Ошибка для разницы между z определяется по обычной формуле ошибки разницы.

$$s_d = \sqrt{s_1^2 + s_2^2}.$$

Здесь $s_1 = s_2 = \frac{1}{\sqrt{97}}$.

Тогда

$$s_{dz} = \sqrt{\frac{1}{97} + \frac{1}{97}} = 0{,}1437.$$

Отсюда $$t = \frac{z_1 - z_2}{s_{dz}} = \frac{-0{,}18}{0{,}1437} = 1{,}252.$$

Даже без обращения к таблицам видно, что разница недостоверна. Если мы проверим вероятность по таблице нормального интеграла вероятности (или по таблице Стюдента при $n = \infty$), то увидим, что вероятность достоверности всего только 0,789. Это значит, что в 211 случаев из 1000 (или в 21 из 100) разница между z_1 и z_2 может возникать по чисто случайным причинам. Поэтому естественно, что такое различие между z_1 и z_2 статистически не достоверно. Это значит, что нулевая гипотеза об отсутствии различий между группами черно-пестрого и рыже-пестрого скота в отношении корреляционной зависимости высоты в холке и живого веса не может быть отброшена. Когда устанавливается высокая достоверность того или иного коэффициента корреляции непо-

средственно по нему самому или путем перевода в z, то в понятиях нулевой гипотезы это означает следующее. Нулевая гипотеза предусматривает отсутствие корреляционной связи, т. е. отсутствие сопряженности в вариации двух признаков. Доказывая же достоверность корреляции, мы этим самым выдвигаем аргументы против признания нулевой гипотезы правильной. Как и в других случаях применения нулевой гипотезы, если полученный коэффициент корреляции не удовлетворяет принятому уровню значимости, т. е. его вероятность ниже 0,95 (например, 0,93 или 0,89), то это является основанием признавать нулевую гипотезу по-прежнему правильной. Чтобы ее отбросить, необходимо получить новые данные, расширить опыты и тем самым доказать достоверность r (или z). Однако вполне возможно, что дальнейшее исследование может не привести к обоснованию факта корреляционной зависимости и только подтвердит правильность нулевой гипотезы.

Корреляция и причинность. Если корреляция доказана, то это значит, что существует сопряженность в вариации двух (или нескольких—при более сложных корреляциях) признаков. Но было бы неправильно делать из этого вывод о наличии причинной зависимости между изучаемыми признаками. Так, при отрицательной корреляции между величиной удоя за лактацию и средней жирностью молока было бы неправильно видеть причину снижения жирности молока в самом факте повышения удоев. В действительности здесь дело в сложном характере физиологических процессов, лежащих в основе молоко- и жирообразования, что приводит к сопряженной вариации обоих признаков. Но обнаружение определенного рода зависимости между удоем и процентом жира может быть полезным для понимания соотношения этих процессов. По-видимому, молокообразование и жирообразование—два различные процесса, протекающие с неодинаковой скоростью. В определенных случаях возможно отставание в процессе образования молочного жира по сравнению с образованием молока, что и является причиной создания отрицательной корреляции между количеством молока и его жирностью. Однако соотношение этих процессов у различных коров неодинаково. В корреляционной решетке табл. 24 в нижнем правом квадранте представлены коровы, совмещающие вы-

сокие удои (до 7500 л) с высокой жирностью молока (до 4,4%). Очевидно, у таких коров соотношение процессов молоко- и жирообразования несколько иное, чем у большинства изученных коров, что и приводит к неполной сопряженности вариации по молочной продуктивности и жирности молока.

Корреляция между признаками может создаться в силу действия какой-то дополнительной причины, влиявшей на оба признака, установить которую можно только с помощью специального биологического анализа. Известным примером такого рода является случай установления в одном стаде крупного рогатого скота корреляции между величиной удоя и дополнительными сосками, в то время как в других стадах корреляция не была обнаружена. Причиной корреляции оказалось то, что среди оставляемых на племя телят часть происходила от хороших коров, имевших дополнительные соски. Поэтому при распределении коров в корреляционной решетке *по признакам удоя и многососковости получилось*, что у группы коров более высокие удои совпали с наличием дополнительных сосков, унаследованных ими от коров-родоначальниц.

Наконец, можно обнаружить корреляцию в силу того, что один из взятых признаков является частью другого признака или оба они являются частями какого-то третьего признака. Так было бы, если бы мы стали вычислять корреляцию между высотой лошадей и длиной их ног Естественно, что в высоте лошади важнейшую часть составляют размеры ее конечностей.

Корреляция между тремя и более изменчивыми признаками. Корреляционный метод может быть применен к изучению сопряженной вариации не только двух, но трех и более признаков, что значительно расширяет рамки его применения. Общеизвестно, что в биологических явлениях обычно участвуют многочисленные факторы, связи между которыми очень часто являются статистическими, и поэтому могут быть изучены методами вариационной статистики. Урожай кукурузы на поле может колебаться в зависимости от исходного материала (сорт, линия, характер гибридности), от почвенных условий, внесенных удобрений, влажности и температуры в период роста и развития растений и т. д. Вариация любого признака животных также связана с вариацией

многих факторов (наследственности и внешней среды в широком смысле этого слова). Наконец, изменчивость многих признаков у одних и тех же животных и растений часто происходит сопряженно, связанно.

К числу приемов изучения связи между многими признаками относится установление коэффициентов множественной и частной корреляции. Под множественной корреляцией обычно понимают зависимость изменения величины x от одновременного изменения величин y, z и т. д. Однако ввиду того, что значение множественной корреляции в биологии невелико, в нашем кратком курсе ее можно не рассматривать. Значительно более ценным является метод частной корреляции. Допустим, что три изменчивые величины или три признака x, y, z коррелируют друг с другом и можно было вычислить 3 коэффициента простой корреляции r_{xy}, r_{xz} и r_{yz}. Если есть основание предполагать, что корреляция между любыми двумя признаками возникает за счет связи с третьим, необходимо изучить связь между первыми двумя признаками, выключив влияние на эту связь третьего, как бы элиминировав его. Так, например, при изучении корреляции между разными промерами животных и их живым весом можно предполагать, что зависимость между определенными промерами создается за счет влияния живого веса, т. е. что существует сопряженная вариация между промерами и живым весом. Это обстоятельство маскирует настоящую зависимость между промерами.

Во многих случаях изучение корреляции между признаками животных затрудняется тем, что их вариация находится под влиянием возраста. Почти невозможно изучить группу животных одного возраста, даже если взята казалось бы очень узкая возрастная группа. Поэтому часто бывает необходимо элиминировать из корреляций между теми или иными морфологическими или биологическими показателями влияние возраста.

Для этой цели служит частный коэффициент корреляции (коэффициент частной корреляции), формула которого следующая:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r^2_{xz})\,(1 - r^2_{yz})}} \qquad (43)$$

Путем соответствующей перестановки букв можно записать формулу для $r_{xz \cdot y}$ и $r_{yz \cdot x}$.

Для удобства буквы x, y, z часто обозначают цифрами 1, 2, 3. В указанной выше формуле элиминируется одна из переменных изменчивых величин. Но можно применить тот же метод для элиминации двух величин при четырех переменных и т. д.

Тогда

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} \cdot r_{23.4}}{\sqrt{(1-r^2_{13.4})\ (1-r^2_{23.4})}}. \qquad (43a)$$

Точки в подстрочных показателях указывают не умножение, а лишь отделяют признаки, корреляции с которыми элиминируются.

В качестве примера использования частной корреляции можно взять данные о давлении крови (1), концентрации холестерина в крови (2) и возрасте (3) 142 женщин пожилого возраста. Были получены следующие коэффициенты корреляции:

$$r_{12} = 0{,}2495;\ r_{13} = 0{,}3332;\ r_{23} = 0{,}5029.$$

Так как высокое кровяное давление может быть связано с высоким содержанием холестерина в стенках кровеносных сосудов, целесообразно тщательно проанализировать коэффициент r_{12}. Но очевидно, что и давление крови, и концентрация холестерина увеличиваются с возрастом. Поэтому возникает вопрос, создается ли корреляция между 1 и 2 за счет их общей связи с возрастом или же она реально существует для каждого возраста. Эффект возраста может быть элиминирован по формуле (43). Предоставляем каждому самому произвести необходимые вычисления. В конечном счете

$$r_{12.3} = 0{,}123.$$

По табл VI можно установить, что при $n=150$ (в таблице нет строчки $n=142$) для достоверности коэффициента корреляции даже при уровне значимости 0,05 он должен быть не менее 0,159 Полученный коэффициент корреляции, очевидно, недостоверен. Внутри отдельных возрастных групп между давлением крови и содержанием холестерина корреляции нет. Можно это выра-

зить в несколько более осторожной форме: взятая для изучения группа не дала возможности обнаружить эту связь, если она все же существует. Пока нет оснований отбрасывать нулевую гипотезу.

Коэффициент корреляции при альтернативной изменчивости. При альтернативной (качественной) изменчивости также можно изучать связь между признаками. В этом случае выясняется вопрос о том, встречается ли совпадение присутствия обоих качественных признаков или присутствия одного и, наоборот, отсутствия другого чаще, чем это должно быть по случайным причинам. Корреляционная решетка будет выглядеть при этом значительно проще, как показано на табл. 29.

Таблица 29

Схема корреляционной решетки при альтернативной изменчивости

y \ x	0	1	Σ
0	a	b	$a+b$
I	c	d	$c+d$
Σ	$a+c$	$b+d$	$n=a+b+c+d$

Классами нулевым (0) и первым (1) обозначаются или два качественных признака (например, голубая окраска, черная окраска) или отсутствие и присутствие какого-либо одного признака (например, безрогость, рогатость). В клетках указывается количество особей с тем или иным сочетанием признаков. Суммы горизонтальных строчек пишутся справа и вертикальных столбцов внизу, как и в корреляционной решетке при количественной изменчивости. Коэффициент корреляции вычисляется по следующей формуле:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \qquad (44)$$

Все сказанное о свойствах коэффициента корреляции

относится и к коэффициенту корреляции при альтернативной изменчивости.

В качестве примера приведем следующие данные о связи между окраской шерсти и цветом глаз у кролика (табл. 30).

Таблица 30

Корреляционная решетка для окраски шерсти *y* и цвета глаз *x* у кролика

y \ *x*	Красные глаза	Некрасные глаза	Σ
Белая шерсть	29	11	40
Окрашенная шерсть	1	59	60
Σ	30	70	100

При подстановке всех значений сумм из таблицы в формулу получим:

$$r = \frac{29.59 - 1\ 11}{\sqrt{30.70.40.60}} = +0{,}76.$$

Так как количество наблюдений достаточно велико (100), ошибка для r может быть вычислена по простой формуле:

$$s_r = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - 0{,}76^2}{\sqrt{100}} = 0{,}04.$$

Достоверность r не вызывает никаких сомнений.

Решетку, состоящую из 2×2 полей, можно применять для анализа и более сложных случаев, если только они укладываются в сопоставимые пары, например болезнь—выздоровление, применение лекарства *A*—применение лекарства *B*.

Так, например, при изучении действия препарата на уменьшение заболеваемости можно применить такую же решетку из четырех полей. Тогда по горизонтали надо написать: неподвергавшиеся воздействию, подвергав-

шиеся; по вертикали: невыздоровевшие, выздоровевшие. При наличии связи между применением препарата и выздоровлением численности a и d будут больше, чем численности b и c. Если же связи нет, численности a, b, c и d будут примерно одинаковыми.

Возможны случаи, когда в одном ряду рассматривается вариация по качественному признаку, а в другом — по количественному, т. е. изучается корреляция между количественным и качественным признаками. Тогда корреляционная решетка будет иметь своеобразный характер: по одному ряду, например ряду x, будет несколько классов, а по другому, ряду y, только два класса. Для вычисления коэффициента корреляции должна быть применена обычная формула для количественной изменчивости. Но надо будет учесть, что вместо четырех квадрантов обычной корреляционной решетки будет только два квадранта, так как два другие окажутся нулевыми.

Иногда по условиям анализа нельзя удовлетвориться разбивкой материала по каждому признаку только на две альтернативные группы. Можно привести такой пример. Было изучено 75 пар однополых близнецов, у которых по крайней мере один партнер был туберкулезным. Была поставлена задача выяснить, существует ли корреляция заболеваемости туберкулезом с условиями среды, в которых жили близнецы. Состояние каждой пары близнецов по отношению к заболеванию туберкулезом могло быть следующим: полностью сходные C, сходные, но слабо c, несходные, но в слабой степени d и резко отличавшиеся D. Таким же образом и по отношению к условиям среды, в которых жили партнеры каждой близнецовой пары, были выделены 4 группы: C—полностью сходные условия, c—сходные условия, но сходство довольно слабое, d—несходные условия, однако несходство было не очень резко выражено, и D—условия резко различные. Корреляционная решетка будет состоять не из 4, а из 16 полей. Она представлена в табл. 31.

Ее можно обработать так же, как табл. 24, приняв условно за нулевые классы третью строчку по горизонтали и третий вертикальный столбец.

Ошибка разницы между средними арифметическими при наличии корреляции. После рассмотрения корреляционной связи следует вновь вернуться к некоторым из

Таблица 31

Корреляционная решетка для установления зависимости между туберкулезом и условиями жизни однополых двоен

Состояние по туберкулезу y	Условия среды x				Σ
	C	c	d	D	
C	11	4	3	1	19
c	10	1	3	0	14
d	12	4	7	1	24
D	1	1	12	4	18
Σ	34	10	25	6	75

разобранных в предыдущей главе вопросов. При сравнении совокупностей приходилось вычислять ошибку разницы между средними арифметическими по формуле

$$s_d = \sqrt{s_1^2 + s_2^2}\ .$$

При этом принималось, что между отдельными значениями переменной величины в обоих совокупностях нет корреляции. Однако приходится учитывать возможность некоторой сопряженности в их вариации, например под влиянием какой-то общей причины, фактора, действовавшего на отдельные варианты обеих совокупностей. Такие случаи особенно возможны тогда, когда изучаются две группы животных, живущих одновременно и в сходных условиях внешней среды. Если доказано наличие корреляционной связи между сравниваемыми выборочными совокупностями 1 и 2, ошибка разницы должна вычисляться по формуле:

$$s_d = \sqrt{s_{\bar{x}_1}^2 + s_{\bar{x}_2}^2 - 2 s_{\bar{x}_1} s_{\bar{x}_2} \cdot r_{12}}\ . \qquad (45)$$

Когда проводится парное сравнение каких-либо показателей, необходимо проверить, нет ли между двумя сравниваемыми группами корреляции, чтобы в этом случае внести необходимую поправку в расчеты.

ВОПРОСЫ

1 Что такое корреляция?

2 Какая разница между корреляционной и функциональной зависимостью?

3 Какая разница между положительной и отрицательной корреляциями?

4 Коэффициент корреляции как мерило степени сопряженности в вариации признаков Его определение с помощью двух нормированных отклонений

5 В чем заключаются важнейшие свойства среднего произведения двух нормированных отклонений?

6 Напишите общую формулу для вычисления коэффициента корреляции Какие изменения можно внести в ее числитель и знаменатель?

7 Напишите формулу коэффициента корреляции, в которую входили бы только значения отклонений от средних, только одни средние показатели

8 В чем заключается рабочая формула коэффициента корреляции Бравэ? В каких случаях выгоднее ее применять?

9 Что такое корреляционная решетка? Объясните, как она строится Можно ли судить о характере корреляции по расположению данных в корреляционной решетке?

10 Каковы возможные значения коэффициента корреляции? Какие значения коэффициента корреляции следует считать высокими, средними и почему?

11 Всегда ли при $r=0$ корреляционная связь отсутствует?

12 Чему равен коэффициент корреляции при полной корреляционной связи?

13 Напишите обычную формулу средней ошибки коэффициента корреляции В каких случаях ее можно применять?

14 Какая формула t применяется при малых значениях n?

15 В чем преимущество числа z перед коэффициентом корреляции r? Можно ли переводить r в z и обратно?

16 Напишите формулу средней ошибки и значение t для z

17 Как надо понимать нулевую гипотезу в применении к коэффициенту корреляции, к разнице между двумя коэффициентами корреляции?

18 Является ли наличие корреляции доказательством причинной зависимости между изучаемыми варьирующими признаками?

19 Что такое множественная корреляция?

20 Напишите формулу коэффициента частной корреляции и объясните ее значение

21 Как строится корреляционная решетка при альтернативной изменчивости?

22 Напишите формулу коэффициента корреляции при альтернативной изменчивости

23 Как обрабатывается корреляционная решетка при альтернативной изменчивости, состоящая из многих полей?

24 Изменяется ли формула ошибки разницы средних арифметических при наличии корреляции между ними?

ЗАДАЧИ

64. На 40 пчелах было проведено измерение длины крыла x и длины хоботка y:

x	9,68	9,81	9,59	9,68	9,84	9,59	9,61
y	6,53	6,71	6,70	6,69	6,70	6,62	6,59
x	9,55	9,25	9,08	9,70	9,60	9,50	9,74
y	6,55	6,35	6,25	6,61	6,51	6,55	6,74
x	9,72	9,64	9,73	9,77	9,72	9,54	9,65
y	6,75	6,45	6,75	6,70	6,65	6,68	6,77
x	9,74	9,59	9,71	9,56	9,61	9,61	9,55
y	6,44	6,54	6,64	6,55	6,57	6,61	6,64
x	9,78	9,74	9,48	9,71	9,20	9,53	9,74
y	6,64	6,63	6,62	6,55	6,22	6,43	6,67
x	9,67	9,56	9,49	9,64	9,45		
y	6,68	6,62	6,71	6,70	6,50		

Вычислите r. Для упрощения расчетов можно отнять величины 9 от длины крыла и 6 от длины хоботка.

65. У окуня озера Баторино измерены (Г. В. Гладким) длина головы x и длина грудного плавника y:

x	66	61	67	73	51	59	48	47	58	44	41	54	52	47	51	45
y	38	31	36	43	29	33	28	25	36	26	21	30	28	27	28	26

Определить корреляцию между x и y.

66. Надо было установить, есть ли корреляция между высотой головы x и длиной 3-го членика усика y у *Drosophila funebris*. Для этого с помощью окуляр-микрометра получены следующие данные по x и y (в делениях окуляр-микрометра):

x	15	16	15	15	16	16	17	18	18	17	17	17	15	16		
y	29	31	32	33	32	33	33	36	36	35	35	35	35	33		
x	15	15	15	17	15	13	15	14	17	15	16	15	15	16	15	16	15	16	18
y	31	31	31	35	33	30	32	31	35	33	33	32	30	33	33	33	30	13	34
x	17	14	15	14	15	15	13	15	16	14	15	15	15	14	15	15	15	16	18
y	34	31	33	31	31	33	30	30	33	30	33	31	32	30	31	31	32	33	35
x	15	14	15	15	14	16	17	15	15	15	14	15	14	15	17	15	17	15	14	15	16
y	32	32	32	31	31	33	35	32	31	34	30	33	32	32	35	31	36	33	33	33	33
x	18	17	17	18	17	17	16	17	18	18	16	16	17	17	16	16	17	16			
y	35	36	34	35	33	32	34	34	34	35	35	33	34	33	35	33	33	33			

Вычислите коэффициент корреляции и определите его достоверность.

67. В двух группах свиней изучали корреляцию между привесом и количеством использованного корма. В первой из них ($n = 5$) был получен $r = 0{,}87$, во второй ($n = 12$) — $r = 0{,}56$. Различаются ли эти коэффициенты корреляции? Можно ли проводить сравнение r без перевода их в z?

68. Между живым и убойным весом свиней на материале 533

голов был получен $r = 0{,}986$. Каковы доверительные границы этого коэффициента корреляции при вероятности 0,95?

69. При объединении ряда данных о корреляции между длиной крыла и длиной хоботка у пчел был получен $r = 0{,}721$ ($n = 126$ пчел) Каковы его доверительные границы при вероятности 99%?

70. В двух выборках изучали корреляцию между одними и теми же величинами x и y. Были получены коэффициенты корреляции $r_1 = 0{,}85$ ($n_1 = 20$) и $r_2 = 0{,}70$ ($n_2 = 30$). Достоверно ли различие между ними? Можно ли анализировать различие между r_1 и r_2 непосредственно или надо переводить r в z?

71. Для двух выборок были получены коэффициенты корреляции между x и y: $r_1 = 0{,}75$ ($n_1 = 1000$) и $r_2 = 0{,}80$ ($n_2 = 800$). Можно ли в данном случае определить достоверность разницы, пользуясь только r?

72. По данным А. К. Митропольского для 500 человек в возрасте от 21 до 28 лет были получены следующие коэффициенты корреляции между ростом (1), окружностью груди (2) и весом (3): $r_{12} = 0{,}395$; $r_{13} = 0{,}692$ и $r_{23} = 0{,}646$. Определите частные коэффициенты корреляции: $r_{12 \cdot 3}$ и $r_{23 \cdot 4}$.

73. Обработайте корреляционную решетку табл. 31 из текста главы, вычислите r и определите его достоверность разными методами, в том числе с помощью табл. VII.

74. На однополых однояйцевых двойнях (32 пары) было проведено изучение связи между заболеванием туберкулезом (x), наследственностью (y) и влиянием внешней среды (z). Коэффициенты корреляции были следующими: $r_{xy} = 0{,}47$; $r_{xz} = 0{,}45$; $r_{zy} = 0{,}07$. Значение наследственного предрасположения к туберкулезу могло быть недостаточно выявлено в силу наличия влияния на туберкулез внешних условий Проверьте это путем вычисления коэффициента частной корреляции $r_{xy \cdot z}$. Установите, насколько он достоверен.

75. Коровы холмогорских помесей 2-го поколения по высоте в холке x, глубине груди y и ширине в моклоках z были следующими:

Номера коров	1	2	3	4	5	6	7	8	9
x	125	126	133	130	126	132	130	130	122
y	69	69	70	71	68	73	72	72	66
z	56	52	49	53	42	56	53	53	51

Номера коров	10	11	12	13	14	15	16	17	18	19
x	133	131	131	138	132	127	125	122	123	128
y	76	70	57	73	71	71	68	67	69	70
z	57	50	55	50	54	53	50	50	49	52

Номера коров	20	21	22	23	24	25	26	27	28
x	126	126	124	131	123	131	132	129	133
y	70	65	68	69	70	70	72	67	70
z	52	51	52	54	51	54	52	54	53

Номера коров	29	30	31	32
x	124	124	126	123
y	60	68	64	65
z	46	55	50	47

Вычислите r_{xy}, r_{xz}, r_{yz} и $r_{yz.x}$. Определите также $\overline{x}$, σ и $s_{\overline{x}}$ для всех трех признаков.

76 В опытах по кормлению 35 крыс в течение 28 дней были получены следующие данные (в г). Начальный вес x_1, количество скормленной пищи x_2, конечный вес y.

x_1	25,8		15,8		18,1		13,3		20,1		10,1	17,1
x_2	98		116		104		99		153		98	103
y	14,8		9,7		11,3		26,0		44,7		21,0	25,2
x_1	21,0	23,7	11,2	10,2	16,4	15,9	8,0	26,0	2,4	7,5	15,9	10,7
x_2	112	133	80	87	138	96	102	155	107	142	110	80
y	13,7	38,5	5,8	17,7	40,0	17,1	3,0	37,3	9,7	36,3	21,2	4,5
x_1	6,4	16,9	12,2	13,4	15,0	13,8	17,8	20,4	7,9	16,0	12,8	
x_2	83	105	96	90	24	153	82	88	66	118	135	
y	4,0	20,2	20,5	18,9	26,4	25,4	9,4	21,2	9,2	41,1	31,3	

Вычислите коэффициенты корреляции: 1) между x_1 и x_2, 2) между x_1 и y и 3) между x_2 и y. Определите коэффициент частной корреляции $r_{23.1}$.

77. В 36 анализах крови определяли: x—число эритроцитов (в миллионах), y—содержание гемоглобина (в проц) и z—оседание крови за 24 часа (в мм):

x	y	z	x	y	z
0,80	22	8	4,33	82	34
1,71	45	18	3,80	79	35
2,63	61	24	3,82	87	36
3,19	66	26	3,81	87	37
2,80	72	28	4,20	87	37
3,14	83	29	4,47	90	38
3,21	73	30	3,71	97	40
3,28	82	30	4,22	96	40
3,63	78	30	3,90	92	40
3,30	82	30	4,36	94	44
4,10	81	32	1,30	27	12
3,29	82	32	2,50	50	20
3,46	77	32	2,80	63	26
3,32	80	33	3,10	71	28
3,11	82	33	2,87	70	29
3,28	79	34	3,68	72	30
3,66	84	34	3,59	76	30
3,90	75	34	3,40	71	30

Определите коэффициенты корреляции r_{xy}, r_{xz} и r_{yz} и коэффициенты частной корреляции: $r_{xy\,z}$, $r_{xz\,y}$ и $r_{zy\,x}$.

78 В 1939 г были опубликованы следующие данные о распределении заболевших и незаболевших гриппом среди работников Центрального Универмага в Москве, вдыхавших и невдыхавших противогриппозную сыворотку.

Группы	Незаболевшие	Заболевшие	Итого
Вдыхавшие сыворотку	497	4	501
Не вдыхавшие сыворотку	1675	150	1825
Итого	2172	154	2326

Вычислите коэффициент корреляции между вдыханием противогриппозной сыворотки и незаболеванием гриппом и определите, насколько он достоверен.

Глава 6

ИЗМЕРЕНИЕ СВЯЗИ. РЕГРЕССИЯ

Понятие о регрессии. Коэффициент корреляции указывает лишь на степень связи в изменчивости двух переменных величин или, как иногда говорят, на меру тесноты связи, но не дает возможности судить о форме связи, то есть о том, как меняется одна величина по мере изменения другой. Именно на этот последний вопрос позволяет ответить другой метод определения связи между варьирующими признаками, носящий название метода регрессии. В современной статистике, в том числе биологической, коэффициентами корреляции пользуются сейчас значительно реже, чем прежде. Метод же регрессии приобретает все большее значение. Анализ взаимоотношения двух изменчивых величин с помощью метода регрессии часто может дать очень ценные результаты, особенно в практическом отношении. В некоторых случаях для освещения различных сторон вопроса надо применять и корреляционный, и регрессионный методы анализа.

При простой корреляции изучается зависимость между изменчивостью двух признаков x и y С помощью регрессии ставится дополнительно задача установить, как количественно меняется одна величина при изменении другой на единицу. Так как изменчивых величин две, то регрессия, очевидно, может быть двусторонней: определение изменения y по изменению x и определение изменения x по изменению y. В этом заключается первое отличие метода регрессии от метода корреляции. Второе отличие в том, что степень и характер регрессии можно

установить и при небольшом числе пар значений обоих признаков.

Регрессия может быть выражена несколькими способами: путем построения так называемых эмпирических линий регрессии, путем составления уравнений регрессий и построения теоретических линий регрессии и, наконец, с помощью вычисления коэффициента регрессии. Первые два способа позволяют выразить регрессию графически.

Эмпирические линии регрессии. Для построения эмпирических линий регрессии можно воспользоваться обычной корреляционной решеткой. Но в ней следует заменить границы классов центральными значениями классов. Общая схема решетки с теми данными, которые нужны для построения эмпирических линий регрессии, представлена в табл. 32.

Таблица 32

Схема корреляционной решетки для построения эмпирических линий регрессии

y \ x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Средние по x для классов ряда $y(x^0/y)$
y_1	a_1							x_1^0
y_2	$б_1$	$б_2$	$б_3$					x_2^0
y_3		$в_2$	$в_3$	$в_4$	$в_5$			x_3^0
y_4		$г_2$	$г_3$	$г_4$	$г_5$			x_4^0
y_5			$д_3$	$д_4$	$д_5$	$д_6$		x_5^0
y_6				e_4	e_5	e_6	e_7	x_6^0
Средние по y для классов ряда x (y^0/x)	y_1^0	y_2^0	y_3^0	y_4^0	y_5^0	y_6^0	y_7^0	-

В столбце справа выписаны средние значения признака x для классов ряда y, т. е. регрессия x по y. Шести

значениям y (от y_1 до y_6) соответствуют шесть значений x^0 (от x_1^0 до x_6^0). Важно, что значения y являются в данном случае строго размеренными, т. е. выраженными в классах ряда y, значения же x являются конкретными средними по признаку x тех вариант, которые расположены в каждой горизонтальной строчке. Именно поэтому они обозначены знаками x_1^0, x_2^0 и т. д. Внизу в горизонтальной строчке даны соответствующие значения y для классов ряда x, т. е. регрессия y по x. В этом случае семи значениям x (от x_1 до x_7) соответствуют семь значений y^0 (от y_1^0 до y_7^0). Таким образом, при регрессии y по x точными значениями классов являются значения x_1, x_2, x_3, ..., x_7 значения же y^0 являются средними значениями по данному признаку группы вариант, расположенных в вертикальных столбцах.

Таблица 33

Корреляционная решетка для живого веса x и обхвата груди y группы коров симментальской породы с данными для построения эмпирических линий регрессии[1]

y \ x	225	275	325	375	425	475	525	575	f_y	x^0/y
135	1								1	225
145	2	1	1						4	263
155	1	17	17	17	1				53	325
165		3	40	44	24	8			119	373
175			4	25	35	21	9	1	95	430
185					1	9	15	2	27	508
195								1	1	575
f_x	4	21	62	86	61	38	24	4	300	
y^0/x	145	156	160	166	170	175	182	185		

В качестве конкретного примера в табл. 33 приведена несколько упрощенная решетка живого веса x и обхвата груди y группы коров симментальской породы.

Значения цифр в графах x^0/y и соответственно y^0/x получаются путем обработки данных каждой горизон-

[1] В классах x и y указаны центральные значения классов в кг и соответственно в см.

тальной строчки или вертикального столбца как небольшого вариационного ряда. Так, например во второй строчке табл. 33 (по горизонтали) указаны 4 варианты: 2 из них имеют веса по 225 кг, 1—275 кг и 1—325 кг. Средняя по 4 вариантам 263 кг. В первой строчке только одна варианта, вес которой 225 кг. Поэтому в графе x^0/y записана цифра 225 кг. В третьей строчке все особи, входящие по обхвату груди в один класс «155 см», составляют по весу вариационный ряд, охватывающий классы от 225 кг до 425 кг. В силу полной симметричности ряда среднюю можно определить без подсчета—на глаз. Она равна 325 кг. Для вариант, расположенных в вертикальных столбцах, значения классов надо брать из графы y. Так, например, в первом вертикальном столбце четыре варианты имеют среднюю 145 см. В предпоследнем вертикальном столбце приведены 9 вариант с обхватом груди 175 см и 15—с обхватом груди 185 см. Средняя арифметическая из 24 вариант будет равна

$$\frac{175 \cdot 9 + 185 \cdot 15}{24} = 181 \text{ см.}$$

Корреляция была вычислена на основе изучения 300 особей (эта цифра проставлена в пересечении граф f_y и f_x и представляет собой $\Sigma f_x = \Sigma f_y$. Однако в регрессии x по y число пар значений x и y равно только 7, т. е. $n=7$, а в регрессии y по x $n=8$. Это объясняется тем, что варианты, находящиеся в каждом классе, объединены в единые группы, и в дальнейшем все операции проводятся со средними этих групп. Методом регрессии можно пользоваться и в тех случаях, когда данные сводятся лишь к немногим единичным наблюдениям величин y и соответствующих им значений x.

На основе показателей x^0/y и y^0/x табл. 33 можно построить на одном графике обе линии регрессии, как это сделано на рис. 7. На горизонтальной оси x отмечены центральные значения классов x (от 225 кг до 575 кг). На вертикальной оси y—центральные значения классов y (от 135 см до 195 см). Значения x^0 по классам y нанесены крестиками. Соединяющая их линия представляет собой линию регрессии x по y. Таким же образом построена и линия регрессии y по x (значения y^0 нанесены кружками).

На рис. 7 кружками и крестиками нанесены средние значения x^0 для классов ряда y и средние значения y^0 для классов ряда x. Но в некоторых случаях полезно нанести на поле непосредственно эмпирические точки значений пар $x—y$ и на этом же поле построить эмпирические (или теоретические) линии регрессии. Иногда значений пар $x—y$ так мало, что эмпирическую линию регрессии можно нанести непосредственно по ним (без объединения отдельных вариант по классам), соединяя их линией. Совокупность двух ломаных линий регрессии является графическим изображением связи между x и y, точнее говоря, зависимости x от y и y от x. Чем больше совпадают друг с другом линии регрессии, тем выше связь в изменчивости данных признаков. При полной корреляции линии регрессии должны совпадать друг с другом.

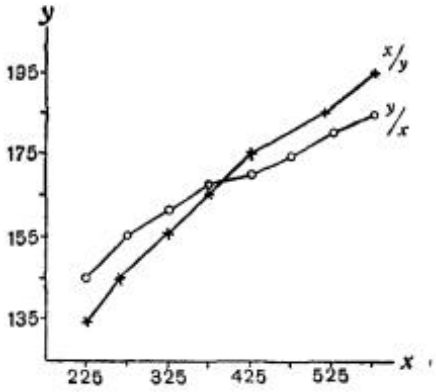


Рис. 7. Эмпирические линии регрессии y по x и x по y (x—живой вес в кг, y—обхват груди в см, коров симментальской породы)

Уравнение регрессии и теоретическая линия регрессии. Эмпирическая линия регрессии обычно представляет собой более или менее ломаную линию. Хотя она достаточно наглядно отображает характер связи между двумя изменчивыми величинами x и y, но не дает еще возможности точно определить любое значение x по заданному значению y или, наоборот, значение y по заданному значению x. Для этой цели могут служить уравнения регрессии.

Уравнение регрессии в общем виде можно записать так:

$$y - \bar{y} = b\ (x - \bar{x}). \qquad (46)$$

Оно выражает определенную зависимость, а именно, что вслед за отклонением x от средней по ряду x происходит и отклонение y от средней по ряду y, причем по-

казатель b является коэффициентом пропорциональности, т. е. величиной, указывающей на количественную связь в изменении y при изменении x.

При переносе $\bar{y}$ в правую часть равенства получим

$$y = \bar{y} + b\ (x - \bar{x}). \qquad (46a)$$

Если $\bar{x}$ приравнять нулю, то $\bar{y}$ будет являться первоначальным значением y, с которого надо начинать при построении линии регрессии при $x=0$. Его можно обозначить через a. Уравнение регрессии примет вид обычного уравнения прямой линии, известного из аналитической геометрии:

$$y = a + b\,x. \qquad (47)$$

Здесь y и x представляют собой коррелирующие в своей изменчивости величины, a—первоначальное значение y при $x=0$, b—коэффициент пропорциональности, который показывает математическую степень зависимости x от y. Это уравнение предусматривает прямолинейную зависимость между x и y, т. е. прямолинейную регрессию. При наличии криволинейной зависимости применяются более сложные уравнения. Для того чтобы определить значения a и b в уравнении $y=a+bx$, надо решить систему двух уравнений:

$$\begin{aligned} &\text{I} \quad na + b\Sigma x = \Sigma y \\ &\text{II} \quad a\Sigma x + b\Sigma x^2 = \Sigma xy. \end{aligned} \qquad (48)$$

Составление этих уравнений, основанных на способе наименьших квадратов, легко видеть на примере, заимствованном из книги акад. В. С. Немчинова «Сельскохозяйственная статистика» (табл. 34). Связь возраста и живого веса поросят установлена по 9 парам наблюдений от возраста «0», т. е. от рождения, до возраста «8 недель». Таким образом, $n=9$. Все остальные данные для включения в уравнения I и II (Σx, Σy, Σx^2 и Σxy) можно взять из табл. 34.

После подстановки соответствующих данных уравнения I и II приобретут следующий вид:

$$\text{I} \quad 9a + 36\ b = 58{,}6$$

$$\text{II} \quad 36a + 204\ b = 319{,}0$$

Таблица 34

Данные о связи между живым весом поросят y и их возрастом x и определение величин, нужных для составления уравнения регрессии y по x

x, недель	y, кг	x^2	xy
0	1,3	0	0
1	2,5	1	2,5
2	3,9	4	7,8
3	5,2	9	15,6
4	5,3	16	21,2
5	7,5	25	37,5
6	9,0	36	54,0
7	10,8	49	75,6
8	13,1	64	104,8
$\Sigma x=36$	$\Sigma y=58,6$	$\Sigma x^2=204$	$\Sigma xy=319,0$

Решать их можно обычными алгебраическими методами. Для этого достаточно все члены уравнения I умножить на 4. Тогда оно приобретет следующий вид: $36a+144b=234,4$.

После этого надо из уравнения II вычесть преобразованное уравнение I, т. е.

$$\begin{array}{r} 36a + 204\ b = 319,0 \\ -(36a + 144\ b = 234,4) \end{array}$$

Получим $36a-36a+204b-144b=319,0-234,4$

$$60b = 84,6.$$

Отсюда $$b = 84,6 : 60 = 1,41.$$

Путем подстановки значения b ($b=1,41$) в уравнение I получим значение a:

$$9a + 36 \cdot 1,41 = 58,6$$

$$9a = 58,6 - 50,8$$

$$9a = 7,8$$

$$a = 0,87.$$

Таким образом, в окончательном виде уравнение регрессии будет следующим: $y=0,87+1,41\ x$. С помощью этого уравнения можно по любому значению x, т. е. для любого возраста поросят от рождения до 8 недель, опре-

делить теоретический средний вес.

Если $x=0$, то $y=0{,}87$ кг; при $x=1$ $y=2{,}28$ кг; при $x=2$ $y=3{,}69$ кг и т. д.

Теоретическая линия регрессии y/x представлена на рис. 8.

На этом же рисунке пунктиром нанесена эмпирическая линия регрессии. Отдельные точки ее отмечены крестиками. Величины a и b определяют местоположение теоретической линии регрессии, а именно: величина a отсекает на оси y от нуля тот отрезок, с конца которого начинается линия регрессии, угол же подъема ее над горизонталью определяется величиной b. С увеличением возраста на одну неделю живой вес увеличивается на $b=1{,}41$ кг.

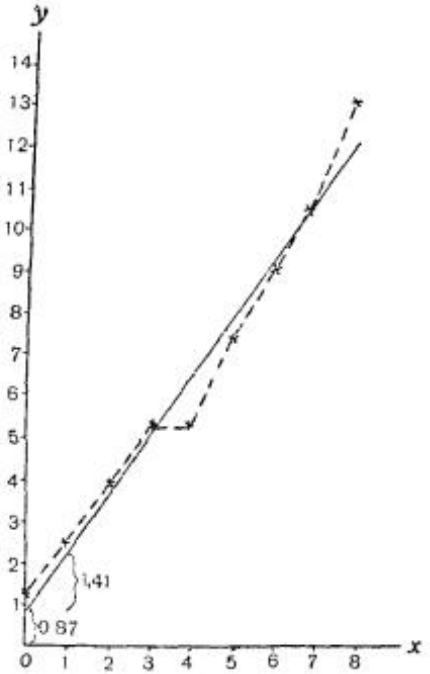


Рис. 8. Эмпирическая и теоретическая линии регрессии y по x (y—живой вес в кг, x—возраст поросят в неделях) На графике нанесены значения $a(a=0{,}87)$ и $b(b=1{,}41)$.

Если бы мы нанесли на рис. 8 также теоретическую линию регрессии x по y, то увидели бы, что две теоретические линии пересекаются в точке, соответствующей среднему значению обоих признаков. При отсутствии корреляции теоретические линии регрессии пересекутся под прямым углом друг к другу, а при полной корреляции они полностью совпадут. Чем меньше угол между линиями регрессии, тем выше корреляция между признаками x и y.

Коэффициент регрессии. Наконец, регрессия может быть выражена с помощью так называемого коэффициента регрессии. Обозначим его буквой R (в литературе

применяют и другие символы). В силу двусторонности регрессии коэффициентов может быть два: $R_{x/y}$ и $R_{y/x}$. Для их вычисления можно применить следующие формулы:

$$R_{x/y} = r\frac{\sigma_x}{\sigma_y} \quad (49) \quad \text{и} \quad R_{y/x} = r\frac{\sigma_y}{\sigma_x}. \qquad (49a)$$

Необходимо помнить, что в данном случае сигмы должны быть выражены в их абсолютных значениях, т е. вычислены с учетом величин классовых промежутков. Эту оговорку приходится делать потому, что при вычислении коэффициента корреляции, как указывалось выше, можно было пользоваться условными сигмами, т е. их значениями, не умноженными на величину классового промежутка.

Таким образом, коэффициент регрессии может быть вычислен, если известны сигмы обоих вариационных рядов по признакам x и y и коэффициент корреляции между ними. Хотя коэффициент регрессии прямо пропорционален коэффициенту корреляции, но он равен ему только в том случае, если отношение $\frac{\sigma_x}{\sigma_y} = 1$, т. е. когда сигмы обоих рядов одинаковы.

В то же время коэффициент регрессии представляет собой не что иное, как коэффициент пропорциональности b в уравнении регрессии $y = a + bx$. В примере зависимости между живым весом поросят и их возрастом (табл. 34) $r = 0{,}988$, $\sigma_y = 3{,}685$, $\sigma_x = 2{,}582$.

Отсюда $$R_{y/x} = 0{,}988 \cdot \frac{3{,}685}{2{,}582} = 1{,}41.$$

Эта же величина 1,41 была получена при решении уравнения регрессии для b. Поэтому коэффициентом регрессии можно называть и величину b.

Значение b может быть вычислено и в том случае, если нет готовых значений r и σ, с помощью средних отклонений и средних квадратов отклонений от средней. Подстановка их вместо σ и r в формулу для $b\left(b = r\frac{\sigma_y}{\sigma_x}\right)$ дает возможность получить следующие рабочие формулы для b:

а) выраженную в отклонениях:

$$b = \frac{\Sigma (x - \bar{x})\ (y - \bar{y})}{\Sigma (x - \bar{x})^2}, \qquad (50)$$

б) выраженную в конкретных значениях x и y:

$$b = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}. \qquad (50a)$$

В обоих формулах имеется в виду значение b для регрессии y по x. Для регрессии x по y в знаменателях надо брать значения не x, а y.

Данные, приведенные в табл. 34, дают возможность вычислить b с помощью формулы (50a):

$$b = \frac{319 - \frac{36 \cdot 58{,}6}{9}}{204 - \frac{36^2}{9}} = 1{,}41.$$

Получена та же величина.

В качестве примера применения формулы (50) используем данные по изучению давления крови у 58 женщин старше 30 лет, сгруппированные по классам (классовый промежуток по возрасту—10 лет) так, как это нужно для построения эмпирической линии регрессии y по x (табл. 35).

Таблица 35

Изменение кровяного давления у 58 женщин

Центральные значения классов по возрасту x	Среднее кровяное давление y	Отклонения		Квадраты отклонений		Произведения отклонений
		$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})$ $(y-y)$
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
$\Sigma=275$	$\Sigma=705$	0	0	$\Sigma=1000$	$\Sigma=1936$	$\Sigma=1330$
$\bar{x}=55$	$\bar{y}=141$					

Коэффициент регрессии

$$b = \frac{1380}{1000} = 1{,}38.$$

Это значит, что с увеличением возраста на 1 год кровяное давление повышалось в среднем на 1,38 единицы. Уравнение регрессии в данном случае будет одним из следующих:

$$y = 141 + 1{,}38\ (x - 55)$$

или

$$y = 65{,}1 + 1{,}38\ x.$$

На основе любого из них может быть построена теоретическая линия регрессии.

Достоверность линии регрессии и коэффициента регрессии. Поскольку в определении линии регрессии участвуют две величины, или, как говорят в статистике, два параметра a и b, следует говорить о них раздельно.

Линия регрессии может быть расположена под бо́льшим или меньшим углом по отношению к оси абсцисс (x). Этот угол определяется величиной b. В геометрическом смысле b есть тангенс угла между линией регрессии и осью абсцисс. При отсутствии регрессии $b = 0$ и линия регрессии должна идти горизонтально, что соответствует нулевой гипотезе. Таким образом, определение достоверности показателя b или коэффициента регрессии R путем установления размеров его ошибки и сопоставления его величины с ошибкой является оценкой достоверности самого наличия регрессии и тем самым сохранения или отбрасывания нулевой гипотезы.

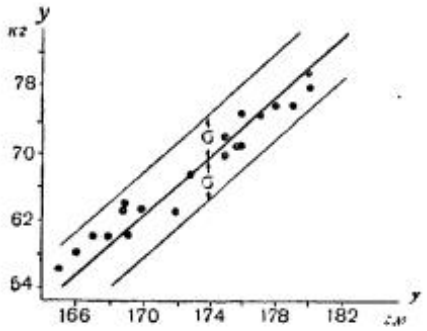


Рис. 9. Теоретическая линия регрессии веса тела человека (y) по его росту (x). Тонкими линиями показаны границы ее колеблемости в пределах $\pm\sigma$. На графике нанесены также эмпирические точки ($n = 20$), по которым построена теоретическая линия регрессии

Ошибка же для величины a является мерой колебле-

мости линии регрессии, так как она указывает возможные верхние и нижние границы линии регрессии при том же угле ее пересечения с осью абсцисс, как это показано на рис. 9.

Основой для определения возможной вариации линии регрессии, как и вычисления ошибки коэффициента регрессии, является сумма квадратов отклонений фактических точек y от вычисленных теоретически значений y для тех же классов ряда x. Так, если использовать приведенные выше данные по кровяному давлению, то можно составить следующую таблицу для расчета этих отклонений (табл. 36).

Таблица 36

Фактические и теоретически вычисленные значения давления крови по классам ряда x

Центральные значения классов по возрасту x	Фактические средние кровяного давления y	Теоретические средние кровяного давления $\hat{y}$	Отклонения от регрессии $y-\hat{y}$	Квадрат отклонений от регрессии $(y-\hat{y})^2$
35	114	113,4	0,6	0,36
45	124	127,2	−3,2	10,24
55	143	141,2	2,0	4,00
65	158	154,8	3,2	10,24
75	166	168,6	−2,6	6,76
			$\Sigma(y-\hat{y})=$ $=\Sigma d_{y\cdot x}=0$	$\Sigma(y-\hat{y})^2=$ $=\Sigma d^2{}_{y\cdot x}=$ $=31{,}60$

Значения для третьего столбца вычисляются по уравнениям регрессии $y=141+1{,}38\ (x-55)$ или $y=65{,}1+1{,}38\ x$ путем подстановки в уравнения значений x для каждого класса по возрасту.

Полученную сумму квадратов отклонений $(y-\hat{y})^2$ (=31,60) надо разделить на число степеней свободы, которое в данном случае равно $n-2$, так как при вычислении отклонений используются две величины, а не одна.

Тогда
$$\sigma^2_{y\cdot x}=\frac{\Sigma(y-\hat{y})^2}{n-2}, \qquad (51)$$

а
$$\sigma_{y\cdot x}=\sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}. \qquad (52)$$

Величину $(y-\hat{y})$ иногда обозначают через $d_{y\cdot x}$.

Подставив соответствующие значения из табл. 36, получим $\sigma^2_{y.x} = 10{,}53$ и $\sigma_{y.x} = 3{,}24$ единицы кровяного давления.

Эта величина $\sigma_{y\cdot x}$ имеет такое же значение, как σ в вариационном ряду. В пределах одной $\sigma_{y\cdot x}$ распределяются отклонения от теоретической линии регрессии вверх и вниз (направление вверх и вниз надо считать по оси y, как это показано на рис. 9) в 68% случаев. С вероятностью 0,997 можно утверждать, что эти отклонения от теоретической линии регрессии расположатся в пределах $\pm 3\sigma_{y\cdot x}$.

Значек $y.x$ показывает, что рассматривается регрессия y по x, т. е. изменение в величине y по точно установленным классам ряда x. При рассмотрении регрессии x по y надо писать $x.y$. $\Sigma (y-\hat{y})^2 = \Sigma d^2_{y\cdot x}$ может быть получена и без составления специальной табл. 36, а из данных предыдущей табл. 35, воспользовавшись следующей формулой:

$$\Sigma (y-\hat{y})^2 = \Sigma (y-\bar{y})^2 - \frac{[\Sigma (x-\bar{x})\ (y-\bar{y})]^2}{\Sigma (x-\bar{x})^2}. \quad (53)$$

Для получения вариансы достаточно разделить эту величину на n—2, а для получения сигмы—извлечь в дальнейшем корень квадратный.

Ошибка коэффициента регрессии и оценка достоверности. Ошибка коэффициента регрессии вычисляется по формуле:

$$s_b = \frac{\sigma_{y.x}}{\sqrt{\Sigma (x-\bar{x})^2}} \quad (54)$$

или, внеся вместо $\sigma_{y\cdot x}$ ее значение в отклонениях,

$$s_b = \sqrt{\frac{\Sigma (y-\bar{y})^2 - \frac{[\Sigma (x-\bar{x})\ (y-\bar{y})]^2}{\Sigma (x-\bar{x})^2}}{(n-2)\qquad \Sigma (x-\bar{x})^2}}, \quad (54a)$$

Возьмем подкоренное значение из табл. 35. Тогда

$$s_b = \frac{3{,}24}{\sqrt{1000}} = 0{,}102.$$

Степень достоверности устанавливается, как обычно, по величине t:

$$t = \frac{b}{s_b}. \qquad (55)$$

При этом надо брать $d.f.=n-2$.

Коэффициент регрессии кровяного давления к возрасту $b=1{,}38$, отсюда

$$t = \frac{1{,}38}{0{,}102} = 13{,}5, \qquad (\text{при } d.f. = 3).$$

По табл. II находим, что полученное значение t превышает требующееся t при уровне значимости 0,01.

В данном случае пришлось пользоваться t-распределением при малых выборках. Но очевидно, что при большом числе наблюдений можно определять вероятность для t по таблице нормального интеграла вероятности.

Если коэффициент регрессии был вычислен с помощью величин r, σ_x и σ_y, то средняя ошибка для него может быть получена по следующей формуле:

$$s_{b\,y\cdot x} = \frac{\sigma_y}{\sigma_x}\sqrt{\frac{1-r^2}{n-2}}. \qquad (56)$$

Соответственно

$$s_{b\,x\cdot y} = \frac{\sigma_x}{\sigma_y}\sqrt{\frac{1-r^2}{n-2}}. \qquad (56a)$$

Сравнение коэффициентов регрессии производится так же, как и сравнение коэффициентов корреляции. Разница между ними делится на ошибку разницы, которая вычисляется путем объединения сумм квадратов обоих выборочных совокупностей по следующей формуле:

$$s_d\,(b_1 - b_2) = \sqrt{\frac{s_1^2}{\Sigma_1 (x_1 - \bar{x}_1)^2} + \frac{s_2^2}{\Sigma_2 (x_2 - \bar{x}_2)^2}}. \qquad (57)$$

При малых величинах совокупностей, на которых

получены коэффициенты регрессии, вносятся некоторые усложнения, подобные тем, с которыми приходилось встречаться при вычислении ошибки разницы между средними арифметическими двух малых выборок. Аналогично им $s_d(b_1 - b_2)$ вычисляется по такой формуле:

$$s_d(b_1 - b_2) = \sqrt{\frac{(n_1 - 2)\, s_1^2 + (n_2 - 2)\, s_2^2}{(n_1 - 2) + (n_2 - 2)} \left(\frac{1}{\Sigma_1 (x_1 - \bar{x}_1)^2} + \frac{1}{\Sigma_2 (x_2 - \bar{x}_2)^2} \right)}. \quad (57a)$$

Достоверность же определяется по значению

$$t = \frac{b_1 - b_2}{s_{d(b_1 - b_2)}} \quad (58)$$

с помощью таблиц II или III.

Связь между регрессией и корреляцией. В начале главы уже указывалось на то, что основное корреляционное уравнение $t_y = r t_x$ может быть преобразовано в обычное уравнение регрессии. Вспомним, что t представляет собой нормированное отклонение:

$$t_y = \frac{y - \bar{y}}{\sigma_y} \text{ и } t_x = \frac{x - \bar{x}}{\sigma_x}.$$

При замене t_y и t_x в формуле $t_y = r t_x$ их полными значениями получим:

$$\frac{y - \bar{y}}{\sigma_y} = r \frac{x - \bar{x}}{\sigma_x}.$$

Если помножить обе половины равенства на σ_y, оно примет следующий вид:

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}).$$

Так как $r \frac{\sigma_y}{\sigma_x} = b$,

то $(y - \bar{y}) = b (x - \bar{x})$.

Мы получили то самое уравнение регрессии, с которого начали рассмотрение вопроса о регрессии. Вот

почему основное корреляционное уравнение $t_y = rt_x$ может быть названо и уравнением регрессии.

Особенностью метода регрессии является то, что зависимость между изменяющимися величинами может рассматриваться как бы в двух разных направлениях, т. е. регрессия может быть двусторонней—x по y и y по x. Отсюда существование двух коэффициентов регрессии. Коэффициент же корреляции является общим мерилом сопряженной вариации двух признаков. Он более искусственен, нежели регрессия. При регрессии один признак выступает в качестве независимой переменной, а другой—в качестве зависимой и наоборот, причем эти зависимости имеют чаще всего совершенно конкретный смысл. Математически коэффициент корреляции представляет собой среднюю геометрическую из двух коэффициентов регрессии. В самом деле

$$b_{x \cdot y} = r \frac{\sigma_x}{\sigma_y} \quad \text{и} \quad b_{y \cdot x} = r \frac{\sigma_y}{\sigma_x}.$$

Отсюда $$b_{x\,y} \cdot b_{y \cdot x} = r^2$$

и $$r = \sqrt{b_{x \cdot y} \cdot b_{y \cdot x}}. \qquad (59)$$

Поэтому как в формуле для коэффициента корреляции, так и в формулах коэффициентов регрессии центральное место занимает сумма произведений отклонений по ряду x и по ряду y, т. е. $\Sigma(x - \bar{x})\ (y - \bar{y})$. Эта сумма является числителем как в общей формуле коэффициента корреляции (32), (36), так и в общей формуле коэффициента регрессии (50) и является в сущности настоящим мерилом сопряженной вариации признаков x и y, или иначе так называемой ковариации (это слово составлено из начальных букв слова корреляция и из слова вариация). Ковариацией называется среднее произведение отклонений двух переменных величин от их средних

$$\text{Cov.} = \frac{\Sigma (x - \bar{x}) \cdot (y - \bar{y})}{d.f.}. \qquad (60)$$

Ковариационный анализ составляет особый раздел современной вариационной статистики. Здесь же ковариация упоминается как связующее звено между корреляционным и регрессионным методами анализа.

Криволинейная зависимость. Зависимость между x и y, выражающаяся прямой линией, является наиболее простой формой связи. Но нередки случаи, когда связи между x и y оказываются более сложными. Так, например, известно, что с повышением возраста коров средние удои за лактацию возрастают. Но эта положительная корреляция наблюдается примерно до 7—8 лактаций, в дальнейшем же, наоборот, с повышением возраста средние удои коров падают. Если выразить эти данные на графике, на котором на оси абсцисс будет нанесен возраст, а на оси ординат—средние удои коров, то будет получена куполообразная кривая, сначала подымающаяся вверх, а затем опускающаяся. Характер кривой, таким образом, отображает реальное биологическое явление, а именно—изменение лактационной способности коров в процессе их индивидуальной жизни и развития.

Математические и статистические приемы анализа должны помочь вскрытию своеобразия тех или иных биологических явлений. Легко убедиться в том, что вычисление обычного коэффициента корреляции по данным об изменении удоев коров в связи с возрастом не поможет установлению своеобразия связи удоев с возрастом. Простой коэффициент корреляции в данном случае будет равен нулю, так как он отражает прямолинейную зависимость между изучаемыми переменными величинами. Вариационная статистика дает методы, с помощью которых можно определить, насколько данная связь отличается от прямолинейной и каков характер этой связи, однако их рассмотрение, ввиду их сложности, выходит за рамки элементарного курса. Укажем лишь, что в общем виде уравнение регрессии, охватывающее как случаи прямолинейной, так и криволинейной зависимости, имеет следующий вид:

$$y = a + bx + cx^2 + dx^3 \ldots \qquad (61)$$

Но есть один случай криволинейной зависимости, который очень важен для биолога. Известно, что рост организмов (или рост популяций организмов) во многих случаях происходит таким образом, что прибавка в весе растущего организма во всякий момент времени пропорциональна уже достигнутому весу. Иллюстрацией могут быть следующие данные, приведенные в табл. 37.

Таблица 37

Изменение сухого веса куриных эмбрионов от 6-дневного до 16-дневного возраста и логарифмы этого веса

Возраст, дней, x	Сухой вес, г, W	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	0,445

Если представить данные второго столбца на графике (рис. 10), то легко видеть, что возрастание веса происходит значительно более быстро, нежели возраста.

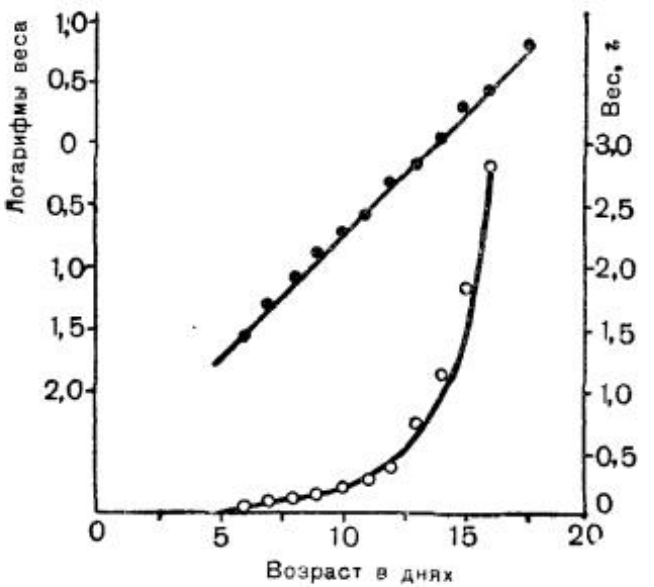


Рис. 10 Изменение сухого веса куриных эмбрионов с возрастом (нижняя кривая линия) и изменение логарифмов веса эмбрионов с возрастом (верхняя прямая линия). На абсциссе логарифмы веса вверх от нуля положительны, вниз—отрицательны

Такая кривая носит название экспоненциальной, и уравнение регрессии для нее выражается следующим образом:

$$W = AB^x, \qquad (62)$$

где A и B—определенные константы.

Однако, логарифмируя это уравнение, его можно привести к форме уравнения прямой:

$$\log W = \log A + (\log B)\, x. \qquad (62a)$$

Здесь log W соответствует y, log A соответствует a и log B соответствует b в обычном уравнении прямой $y=a+bx$.

Третий столбец табл. 37 и дает логарифмы весов, изображение которых на рис. 10 дает прямую линию.

Определить значения для уравнения прямой можно по данным табл. 37, пользуясь теми же приемами, которые были даны выше для составления уравнения прямолинейной регрессии с помощью системы двух уравнений. Она будет иметь следующий вид:

$$y=-2{,}689+0{,}1959\ x.$$

Приемы по изучению роста и применению экспоненциальных кривых изложены в специальной литературе (см. сборник «Рост животных», Биомедгиз, М.—Л., 1935). Практически очень часто ограничиваются лишь графическим изображением кривых роста на особой полулогарифмической бумаге. На такой бумаге на одной из ординат нанесена логарифмическая шкала, а на другой—обыкновенная. Тогда можно не составлять специальных уравнений регрессии.

ВОПРОСЫ

1. Что такое регрессия?
2. В чем преимущество регрессии по сравнению с корреляцией?
3. Какими способами может быть выражена регрессия?
4. Изложите ход работы по построению эмпирической линии регрессии.
5. Под каким углом пересекаются эмпирические линии регрессии при слабой корреляции? при сильной корреляции?
6. Напишите уравнение регрессии в общем виде; в виде уравнения прямой.
7. Напишите систему двух уравнений для определения значений a и b в уравнении $y=a+bx$.
8. Как строится теоретическая линия регрессии если решено уравнение регрессии.

9 Что выражает уравнение регрессии x по y и уравнение регрессии y по x?

10 В каком случае две теоретические линии регрессии пересекаются под прямым углом друг к другу? Когда они совпадают?

11 Чему равен тангенс угла между линией регрессии и осью x?

12 Напишите формулы коэффициента регрессии

13 Может ли коэффициент регрессии быть равным коэффициенту корреляции?

14 Каково взаимоотношение R и b?

15 Напишите формулы для b (в отклонениях и в конкретных значениях x и y)

16 В чем заключается физический смысл ошибки линии регрессии? Как определяются доверительные границы линии регрессии?

17 Каково число степеней свободы при определении ошибки линии регрессии?

18 Напишите несколько формул для ошибки коэффициента регрессии

19 Можно ли вычислить среднюю ошибку для коэффициента регрессии, пользуясь сигмами и коэффициентом корреляции?

20 Как проводится сравнение двух коэффициентов регрессии при больших и малых n?

21 Преобразуйте корреляционное уравнение $t_y = rt_x$ в уравнение регрессии

22 Какова связь коэффициента корреляции с двумя коэффициентами регрессии?

23 Какая величина называется ковариацией?

24 В чем разница между прямолинейной и криволинейной зависимостями? Напишите общую формулу регрессии, охватывающую прямолинейную и криволинейную зависимости

25 Напишите формулу для экспоненциальной кривой регрессии Можно ли преобразовать ее в уравнение прямой?

ЗАДАЧИ[1]

79 У 20 взрослых мужчин были измерены высота (длина тела) в см x и вес y в кг:

Номера мужчин	1	2	3	4	5	6	7	8	9	10
x	165	176	175	163	167	172	175	180	179	173
y	56	75	70	61	61	63	72	80	76	68
Номера мужчин	11	12	13	14	15	16	17	18	19	20
x	166	178	169	169	170	176	180	169	177	176
y	58	76	60	64	63	71	78	63	75	71

[1] Из задач, приведенных в конце главы V, многие могут быть использованы также для построений линий регрессии и вычисления коэффициентов регрессии, например 64, 65, 66, 75, 76, 77 (в последних трех надо брать попарно по 2 признака).

Составьте корреляционную решетку и вычислите r и s_r. Эти же данные используйте для определения регрессии y по x всеми методами.

80. Предполагается, что между количеством остриженной шерсти y и живым весом x овец имеется зависимость. Для 10 овец были получены следующие данные в кг:

Номера овец	1	2	3	4	5	6	7	8	9	10
	50	55	60	50	65	60	50	55	50	65
y	4,0	4,2	4,1	4,2	4,5	4 3	4,1	4,4	4,0	4,2

Постройте линию регрессии y по x (теоретическую и эмпирическую) и установите границы ее колеблемости. Определите коэффициент регрессии.

81. Были получены (Г. В. Гладкий) следующие данные о длине грудного (I) и брюшного (II) плавника у окуня озера Баторино:

I	38	31	36	43	29	33	28	25	36	26	21	30
II	40	34	38	42	26	33	29	26	36	27	22	32
I	27	27	28	26	26	25	24	28	28	27	33	27
II	28	26	32	26	28	27	25	28	30	26	32	27
I	26	23	22	25	24	29	25	25	30	23	24	32
II	29	23	24	30	26	30	27	28	32	23	24	32
I	24	25	30	25	26	30	29	22	29	28	26	28
II	25	27	33	27	27	32	28	24	31	32	27	30
I	25	31	25	32	27	31	28	29	26	32	27	31
II	25	34	26	32	29	30	29	29	26	35	26	33
I	28	28	26	33	30	27	21	28	26	30	23	27
II	29	31	29	33	31	31	23	30	27	29	24	28

Составьте корреляционную решетку и вычислите r и s_r. Постройте эмпирическую и теоретическую линии регрессии II по I и определите коэффициент регрессии

82 Для 10 петушков леггорнов *15-дневного возраста были* получены следующие данные о весе их тела x (в г) и весе гребня y (в мг):

x	83	72	69	90	90	95	95	91	75	70
y	56	42	18	84	56	107	90	68	31	48

Нанесите эти данные на график и составьте уравнение регрессии y по x.

83. Путем еженедельного взятия проб с поля было изучено изменение высоты растений сои y (в см) с возрастом x (в неделях):

x	1	2	3	4	5	6	7
y	5	13	16	23	33	38	40

Выразите эти данные на графике и постройте эмпирическую линию регрессии y по x. Составьте уравнение регрессии и установите до-

верительные интервалы при вероятности 0,95 для линии регрессии

84. Для установления связи между содержанием фосфора в почве x и содержанием фосфора в злаковых растениях y было проведено 9 анализов со следующими результатами:

x	1	4	5	9	13	11	23	23	28
y	64	71	54	81	93	76	77	95	109

Составьте уравнение регрессии и установите, насколько достоверно значение b

85. Для выяснения вопроса о возможности предвидеть размеры удоя коров за 3-ю лактацию по данным об удое за 1-ю лактацию было проведено сравнение удоев за 1-ю x и за 3-ю y лактации по 55 коровам холмогорской помеси (в $л$)

Номера коров	1	2	3	4	5	6	7	8
x	1522	239	1521	2700	1789	2496	1197	1105
y	3693	4453	1446	2134	2940	4353	2066	2152
Номера коров	9	10	11	12	13	14	15	16
x	1701	2218	1790	2964	1287	1756	1406	1810
y	2396	2435	3140	4700	2113	2513	3249	2553
Номера коров	17	18	19	20	21	22	23	24
x	1299	2609	2519	1927	1655	1320	2586	1928
y	2320	4612	3201	3173	3326	1639	4562	3482
Номера коров	25	26	27	28	29	30	31	32
x	3884	2968	2200	1753	1508	1803	1811	2300
y	4257	3465	2448	3435	3747	2112	3061	2985
Номера коров	33	34	35	36	37	38	39	40
x	1697	2870	1284	2049	3730	2300	1804	3064
y	2721	3766	4042	4007	4028	3422	2410	4593

Номера коров	41	42	43	44	45	46	47	48
x	2130	2675	1495	2145	1563	1965	1450	2540
y	3792	3529	2921	4274	3559	4580	2118	4144
Номера коров		49	50	51	52	53	54	55
x		1966	2340	1687	1785	1458	2107	1274
y		2507	4507	2620	2026	2668	3859	2007

Определите корреляцию между удоем за 1-ю лактацию и удоем за 3-ю лактацию. Начертите эмпирические и теоретические линии регрессии и составьте уравнение регрессии.

86. У 10 телят по глубине груди x (в см) и живому весу y (в кг) были получены следующие данные (цифры округлены до десятков):

Номера телят	1	2	3	4	5	6	7	8	9	10
x	90	80	90	95	100	90	90	85	90	90
y	60	40	60	70	85	65	70	50	60	65

Вычислите коэффициент корреляции между глубиной груди и живым весом, постройте эмпирические и теоретические линии регрессии (x по y и y по x) и вычислите коэффициенты регрессии. Определите достоверность r и b.

87. Известны данные для 10 бычков о весе при рождении (в кг) и суточном привесе y (в г):

Номера бычков	1	2	3	4	5	6	7	8	9	10
x	38,5	46,0	43,0	43,0	40,5	44,0	38,0	35,0	40,5	54,0
y	694	901	736	1005	841	743	896	863	855	830

В тексте гл. 5 по этим данным определен коэффициент корреляции. Постройте линию регрессии (эмпирическую и теоретическую) y по x. Вычислите σ теоретической линии регрессии.

88. Были получены следующие данные о потреблении кислорода у пиявками (в мг на кг/час) в зависимости от температуры x (в градусах):

Номера животных	x	y	Номера животных	x	y
1	5,5	16,1	8	16,6	91,0
2	5,6	14,9	9	17,1	94,0
3	6,2	18,8	10	18,8	122,0
4	8,4	32,5	11	19,8	162.0
5	9,0	32,1	12	20,0	167,0
6	10,5	37,1	13	20,7	187,0
7	16,1	88,5	14	26,5	436,0

Определите коэффициент корреляции r_{xy}. Постройте на бумаге график, на котором нанесите точками 14 пар значений x и y. Убедитесь, что они расположены не по прямой, а по кривой линии. После этого замените арифметические значения y их логарифмами и вновь постройте график, где на одной из осей нанесите $\log y$. Вычислите новый коэффициент корреляции $r_{x \cdot \log y}$ и составьте по этим данным уравнение регрессии $\log y$ по x.

89. Под влиянием облучения рентгеновыми лучами наблюдалось следующее замедление развития вируса мозаики Аукуба y (в тыс.) в зависимости от длительности облучения x (в мин):

x	0	3	7,5	15	30	45	60
y	271	226	209	108	59	29	12

Составьте уравнение регрессии, приняв за y логарифм количества вирусов, и за x — минуты облучения. Постройте эмпирическую линию регрессии и теоретическую (ось ординат—логарифмы).

Глава 7

ИЗУЧЕНИЕ СТЕПЕНИ СООТВЕТСТВИЯ ФАКТИЧЕСКИХ ДАННЫХ ТЕОРЕТИЧЕСКИ ОЖИДАЕМЫМ

Фактические данные и научная гипотеза. Количественное изучение биологических явлений обязательно требует создания гипотез, с помощью которых можно объяснить эти явления, вскрыть количественные и качественные закономерности в их проявлении, понять количественные отношения между различными группами изучаемых животных или растений.

Чтобы проверить ту или другую гипотезу, нужно получить с помощью наблюдения или путем проведения специальных опытов ряд фактических данных и сопоставить их с теоретически ожидаемыми, согласно данной гипотезе. Если фактически полученные данные совпадают с теоретически ожидаемыми, то это может быть достаточным основанием для принятия данной гипотезы, для признания ее правильности. Если же фактические данные недостаточно согласуются с теоретическими, не соответствуют им, возникает большое сомнение в правильности предложенной гипотезы. Биолог может быть вынужден отказаться от первоначальной гипотезы и выдвинуть новую, которую также надо проверить. Степень соответствия фактических наблюдений с теоретически ожидаемыми результатами может быть весьма различной. В одних случаях разница между ними очень невелика и может оказаться чисто случайной, в других — она достаточно значительна. Отсюда возникает задача статистической оценки разницы между фактическими данными и теоретически ожидаемыми, установления того, в каких случаях и с какой степенью вероятности

можно считать эту разницу достоверной и, наоборот, когда ее следует считать несущественной, незначимой, находящейся в пределах случайности. В последнем случае сохраняется гипотеза, на основе которой рассчитаны теоретически ожидаемые данные или показатели.

Критерий соответствия хи-квадрат. Степень соответствия фактических данных ожидаемым, иными словами согласия фактических данных с предложенной гипотезой, может быть измерена особым показателем, обозначаемым греческой буквой χ в квадрате (χ^2), отсюда его название «критерий хи-квадрат». В советской литературе его называют по-разному: критерий соответствия, критерий согласия. В дальнейшем мы будем употреблять название «критерий соответствия хи-квадрат» или просто «хи-квадрат». Хи-квадрат был открыт еще в 1900 г. Пирсоном, однако его значение было оценено значительно позже. Очень большое количество самых разнообразных проблем вариационной статистики оказалось в той или иной степени связано с хи-квадратом или основано на его применении. В настоящее время редко можно встретить работу по биологической статистике, где не упоминался бы этот критерий. Нередко даже начинают изложение курса вариационной статистики с разбора критерия хи-квадрат.

В наиболее общем виде формула для критерия соответствия может быть записана следующим образом:

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \qquad (63)$$

где O—фактически наблюдаемое, а E—теоретически ожидаемое число или показатель для данной группы.

Таким образом, хи-квадрат представляет собой сумму частных от деления квадратов отклонений фактически полученных чисел от ожидаемых на число ожидаемых.

Допустим, что при изучении расщепления у томатов по окраске плодов было получено 310 красных плодов и 90 желтых. Ожидалось же при обычном моногибридном скрещивании отношение 3:1, т. е. 300 красных и 100 желтых. Тогда

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

Возникает вопрос, что это за число и как по нему судить, достоверно ли отличается полученное фактически расщепление от теоретически ожидаемого.

Закономерности распределения χ^2. Если бы фактически полученные и теоретически ожидаемые числа полностью совпадали, то χ^2 был бы равен нулю. По мере увеличения разницы между фактическими числами и ожидаемыми величина хи-квадрат будет возрастать. Так как отклонения фактических чисел от ожидаемых возводятся в квадрат, то значения хи-квадрата могут быть только положительными. В этом его отличие от других типов отклонений (например, от t, которое может иметь знаки плюс и минус).

Подобно тому, как это было сделано по отношению к распределению других показателей, было изучено и распределение хи-квадрат. Оказалось, что оно зависит от n, вернее от числа степеней свободы тех данных, по которым производится сравнение фактических и теоретических данных. Каждому же значению χ^2 соответствует и определенная вероятность его появления. Так как значения χ^2 только положительны, то распределение их асимметрично. При изображении этого распределения на графике окажется, что малые значения χ^2 будут обладать наибольшей частотой, с увеличением же значений χ^2 их частота будет падать. В этом отношении есть известная аналогия распределения χ^2 с распределением отклонений от средней арифметической в вариационном ряду. Хи-квадрат может быть вычислен и для них следующим образом:

$$\chi^2 = \Sigma \frac{(x - \bar{x})^2}{\bar{x}}.$$

Однако легко видеть, что сама закономерность распределения χ^2 должна быть иной, так как отклонения возводятся в квадрат.

Значения хи-квадратов могут возрастать от нуля до бесконечности. Соответственно этому вероятности их появления убывают от 1 до 0. Отсюда вытекает возможность рассчитать, какова вероятность появления χ^2 ниже или выше определенной величины. Так как соотношение между хи-квадратом и вероятностью его появления довольно сложное, для практического применения этого критерия пользуются готовыми таблицами. Одна из них

дана в приложении (табл VIII) В этой таблице имеются вертикальные столбцы значений хи-квадрат для различных значений вероятности В левом вертикальном столбце даны степени свободы

Понятия вероятности и значимости в применении к χ^2 На практике не столь важно знать, какое точное значение вероятности соответствует данному значению χ^2, а важно, в какой степени достоверно полученное значение χ^2

Критерии χ^2 используется для проверки определенной гипотезы, которая считается нулевой Нулевая гипотеза обозначает, что нет различия между фактически полученными и исчисленными теоретически данными Значения χ^2, имеющиеся в табл VIII, указывают те границы, до которых полученные значения χ^2 остаются с определеннои вероятностью в рамках случайных отклонений, т е когда нет оснований сомневаться в принятои гипотезе Значения же χ^2, которые будут превышать табличные значения, будут указывать на несостоятельность гипотезы, т е вынуждают отбросить нулевую гипотезу Обычно принято считать допустимой границеи вероятности вероятность 0,05

Значения χ^2, оказывающиеся в пределах значений, указанных в таблице VIII в графах вероятности от 0,99 до 0,10, обеспечивают высокую уверенность в правильности гипотезы Можно удовлетвориться и вероятностью 0,05 Если же значение χ^2 выше табличного, находящегося в графе, где вероятность ниже 0,5, например 0,2 или 0,1, то в этом случае можно считать нулевую гипотезу несостоятельной, т е различие между фактическими и теоретическими результатами является достоверным, значимым

В примере с расщеплением у томатов было получено значение $\chi^2=1{,}33$ Так как групп только 2, то $df=1$ По данным первой строки табл VIII видно, что такое значение χ^2 соответствует вероятности около 0,25 (среднее между 0,30 и 0,20) Значит, совпадение между фактическими результатами и ожидаемыми достаточно велико Принятая гипотеза о том, что имеется расщепление 3 1, подтверждается Но если бы при анализе расщепления и при $df=1$ χ^2 было бы равно, например, 6,4, то вероятность правильности нулевой гипотезы (т е что здесь действительно имеется отношение 3 1) оказалась бы

только около 0,01. Это явилось бы достаточным основанием признать, что наблюдается существенное отклонение от ожидаемого отношения, т. е. что гипотеза о расщеплении в отношении 3:1 должна быть отвергнута. При разборе отдельных конкретных примеров мы будем в дальнейшем еще не раз обращаться к табл. VIII. Сейчас надо лишь отметить, что с помощью критерия χ^2 как бы взвешивается риск ошибиться, сохраняя нулевую гипотезу или, наоборот, ее отбрасывая.

Если отбрасывание первоначальной нулевой гипотезы происходит при $P=0{,}05$, то это означает, что хотя нулевая гипотеза отбрасывается, но еще имеется 5% шансов (5 случаев на 100 или 1 случай на 20), что она правильна. Так что, отбрасывая нулевую гипотезу, исследователь стоит перед возможностью, что он все-таки ошибся. Если отбрасывание нулевой гипотезы производится при $P=0{,}01$, то шанс на ошибку только 1 на 100.

Возьмем теперь противоположный случай. Полученное значение χ^2 несколько превышает табличное при значении $P=0{,}95$, но ниже табличного при $P=0{,}90$. Мы имеем право говорить о значительном совпадении фактических и теоретически ожидаемых данных, т. е. нулевая гипотеза сохраняется. Однако при этом имеется шанс на противоположную ошибку, что все-таки нулевая гипотеза неверна. Этот шанс, правда, очень невелик (5 случаев из 100). Он явно недостаточен, чтобы отбросить первоначальную нулевую гипотезу.

Если вероятность наблюдаемых значений χ^2 находится между 0,5 и 0,6, то считается, что значение χ^2 не выходит из пределов допустимого и достаточных оснований для отбрасывания нулевой гипотезы нет. Но шансы на ошибочность этого мнения уже возрастают.

В биологических исследованиях принято отбрасывать нулевую гипотезу, когда хи-квадрат превышает 3,841. Значения хи-квадрат, превышающие 3,841, составляют как бы область отбрасывания нулевой гипотезы. Они достаточно значимы, достоверны, чтобы отбросить нулевую гипотезу. При этом вероятность того, что нулевая гипотеза все же верна, как раз составляет 0,05.

Так как в понимании вероятности соответствия имеются некоторые тонкости, следует обратить на них внимание и разобрать вопрос подробнее. Когда в гл. 4

рассматривалась оценка разницы между средними арифметическими, то указывалось, что она должна быть достаточно высока, чтобы считать разницу достоверной. При этом в качестве достоверных вероятностей были взяты вероятности 0,99 и 0,95. Уровни же значимости 0,01 и 0,05 являлись величинами, указывающими шансы на признание различия достоверным, в то время как оно только случайно. Имеющиеся в табл. VIII вероятности имеют как бы обратный смысл. Так, в строчке $d.f.=7$ имеется значение $\chi^2=2{,}83$, что соответствует $P=0{,}90$. Допустим, что при анализе был получен $\chi^2=2{,}81$ (при том же числе степеней свободы). Это означает следующее. Если бы было взято большое число выборок из нормальной совокупности, то больше 90% этих выборок имело бы χ^2 той же величины или больше, чем 2,83. Поэтому наблюдаемое в данном примере отклонение фактических частот от теоретически ожидаемых (например, при нормальном распределении вариационного ряда или при любом другом определенном теоретическом отношении между группами) случайно, т. е. эмпирическая выборка имеет тот же характер, что и теоретическая совокупность.

Налицо имеется соответствие, вероятность которого 0,90. Однако вероятности P соответствует как бы дополнительная вероятность Q. В данном случае имеющейся в таблице вероятности 0,90 соответствует дополнительная вероятность 0,10. Это—вероятность противоположного события, а именно: что соответствия нет и изучаемая выборка распределена иначе, чем теоретическая. Но так как эта вероятность мала, нет оснований отбрасывать исходное положение о соответствии, т. е. о том, что получившееся отклонение от ожидаемого, выражающееся $\chi^2=2{,}81$, случайно. Таким образом, только тогда мы можем признавать вероятность достоверности высокой, когда мы не можем считать достаточной для отбрасывания нулевой гипотезы дополнительную к ней вероятность. Так как обычно считается достаточной для признания достоверности разницы вероятности 0,99 и 0,95, то, перенеся это на критерий соответствия, можно считать границами для соответствия 0,05 или 0,01. По отношению к ним дополнительными вероятностями как раз и будут 0,95 и 0,99.

При дополнительной же вероятности, например 0,73,

которая соответствует табличной $P=0{,}27$, нет уверенности в том, что соответствие отсутствует. Поэтому нет оснований для отбрасывания нулевой гипотезы. Рассматривавшиеся выше уровни значимости 0,05 и 0,01 указывали на шансы случайной разницы между изучаемыми статистическими показателями при признании разницы достоверной. Вероятности 0,05 и 0,01 при анализе соответствия указывают на шансы наличия соответствия при признании несоответствия достоверным, т. е. при отбрасывании нулевой гипотезы.

Конечно, надо помнить, что биолог очень редко основывает свои выводы только на проверке гипотезы методом хи-квадрат. Всякий выборочный опыт сам по себе только доставляет известные данные, но не может служить окончательным доказательством гипотезы. В процессе исследования новые доказательства прибавляются к уже существующим. Таким образом, происходит как бы нарастание информации о данном явлении. Если какой-то опыт имеет большую ценность, в результате его может быть создана и новая гипотеза, которая должна быть проверена или путем новых опытов или путем выяснения ее соответствия уже установленным научным положениям. Одним вычислением хи-квадрата и установлением того факта, что он обеспечивает достоверность соответствия на уровне 0,05, ограничиваться в научном исследовании нельзя. При этом вероятность неправильного вывода еще достаточно высока (ошибочный вывод возможен в 5 случаях из 100). Чтобы быть уверенным в выводах, нужно провести такое количество опытов или наблюдений, чтобы возможная ошибочность их была максимально снижена.

Число степеней свободы при пользовании критерием хи-квадрат. Из табл. VIII видно, что распределение хи-квадрат очень сильно зависит от числа степеней свободы. Поэтому надо учитывать именно число степеней свободы, а не просто число наблюдений или групп. Число степеней свободы, как известно, вычислялось как $n-1$. При пользовании критерием хи-квадрат дело обстоит несколько сложнее. Число степеней свободы—это то число классов (или групп), частоты которых могут принимать любые произвольные значения, не связанные с наблюдаемыми частотами. В простейших случаях это будет число классов, уменьшенное на единицу. Так, если при расщепле-

нии возникает 2 класса, то несвязанным с наблюдаемой частотой является лишь 1 класс, второй же уже связан с первым. Тогда $d.f.=1$.

Если при расщеплении изучаются 4 класса (например, в простейшем случае дигибридного наследования), $d.f.=3$. При проверке соответствия частот по классам, распределенным в решетке с числом полей 2×2, 2×3, 4×4 и т. д., обычно пользуются следующей формулой для числа степеней свободы:

$$d.f.=(r-1)(c-1),$$

где r—число горизонтальных рядов, c—число вертикальных столбцов. В таком случае при расположении опытных данных в таблице из 4 полей (2×2) число степеней свободы равно 1, в таблице из 9 полей (3×3) $d.f.=4$, в таблице из 6 полей (2×3) $d.f.=2$ и т. д.

При проверке соответствия полученного распределения вариант в вариационном ряду нормальному, биномиальному и другим видам распределения берется число фактических классов (несколько классов, объединяемых при подсчете в один, считаются за один класс) и из них вычитается 2 или 3, так как фактическое и теоретическое распределения могут совпадать по 2 элементам (например, n и $\bar{x}$) или по 3 (например, n, $\bar{x}$ и σ).

Однако возможны и некоторые другие, более сложные случаи, когда установление числа степеней свободы требует тщательного обдумывания: какие элементы данного изучаемого комплекса могут принимать любые произвольные значения, а какие их определяют, являются как бы фиксированными, выполненными.

Суммирование нескольких χ^2 и критерий разнородности. При проведении нескольких опытов по одному и тому же вопросу можно вычислять частные χ^2 для каждого отдельного опыта, а затем получить значение χ^2 для суммы опытов путем простого суммирования частных χ^2. Число степеней свободы также будет равно сумме чисел степеней свободы складываемых χ^2. Так, например, если в каждом отдельном опыте $d.f.=1$, а опытов было 5, то число степеней свободы для общего χ^2 равно 5. Достоверность полученного значения χ^2 можно проверить по той же табл. VIII.

С другой стороны, можно обработать весь материал в

целом, не считаясь с отдельными опытами, получить соответствующие эмпирические значения, вычислить теоретически ожидаемые величины и получить χ^2. Сравнение χ^2, полученных двумя разными способами объединения опытного материала, позволяет судить о степени его однородности или неоднородности.

Для иллюстрации сказанного возьмем следующий пример. На 11 гетерозиготных растениях кукурузы наблюдали расщепление по окраске проростков на зеленых и желтых. В некоторых случаях оно точно соответствовало отношению 3:1 (например, 27 зеленых и 9 желтых проростков), в других несколько отклонялось в ту или другую сторону (например, 110 зеленых и 39 желтых, 98 зеленых и 24 желтых). Значения хи-квадратов по отдельным растениям колебались от 0,00 до 2,00. При суммировании всех χ^2 было получено значение 6,54 (при $d.f.=11$). При проверке по табл. VIII видно высокое соответствие эмпирических данных теоретически ожидаемым (P около 0,80). Когда весь материал был объединен в одну группу, то расщепление было следующим: 854 зеленых и 249 желтых проростков (при ожидаемых числах 827,25 и 275,25). Вычисленный по этим данным χ^2 равен 3,46, $d.f.=1$. При $d.f.=1$ границей является значение $\chi^2=3{,}84$ ($P=0{,}05$). Таким образом, и по итоговым данным можно говорить о наличии соответствия между эмпирическими и теоретическими числами, т. е. о том, что нулевая гипотеза правильна. Но между значениями χ^2, вычисленными двумя способами, наблюдается разница. Именно она и должна указывать на степень однородности или неоднородности опытных данных. Записать это можно следующим образом:

	Степени свободы	Хи-квадрат
Сумма 11 хи-квадратов	11	6,54
Хи-квадрат по объединенным данным	1	3,46
Разница (разнородность)	10	3,08

Проверка по табл. VIII показывает, что разнородность очень мала. Значение P приблизительно 0,98. Это

значит, что по нулевой гипотезе бо́льшие значения ожидаются в 98%, т е колеблемость исходных данных здесь даже ниже, чем она бывает обычно при случайных выборках

Вычисление теоретически ожидаемых чисел и определение хи-квадратов при анализе расщепления. Формула для определения χ^2 настолько проста, что ее применение чаще всего не вызывает затруднений Более сложным является в некоторых случаях определение теоретически ожидаемых величин Поэтому целесообразно разобрать несколько конкретных примеров, на которых удастся продемонстрировать все приемы определения хи-квадрата

Метод χ^2 очень часто применяется при генетических исследованиях, когда нужно проверить соответствие частот классов, получаемых при расщеплении, свободном комбинировании или сцеплении, с частотами, ожидаемыми при той или иной генетической гипотезе В этом случае для вычисления ожидаемых чисел надо помножить общее число изучаемых фактически особей на соответствующую долю, теоретически ожидаемую при данном типе исследования

Наиболее простым примером расщепления при моногибридном скрещивании были данные о результатах скрещивания томатов, гетерозиготных по окраске плодов Было получено 310 красных и 90 желтых плодов Если ожидать расщепления 3 1, то каждая категория вычисляется как доля от $n \cdot \frac{3}{4} \cdot n$ и $\frac{1}{4} \cdot n$. В данном случае красных плодов должно было быть $\frac{3}{4} \cdot 400 = 300$ и желтых $\frac{1}{4} \cdot 400 = 100$.

Было получено значение $\chi^2 = 1{,}33$ При $df = 1$ фактические результаты хорошо совпадают с теоретически ожидаемыми

Возможно применение и более простой формулы

$$\chi^2 = \frac{(a - rb)^2}{r(a + b)}, \qquad (64)$$

где a и b—фактические числа в каждом классе, а r—теоретическое отношение соответствующих классов в популяции. Для расщепления у томатов вычисление хи квад-

рат по этой формуле дает ту же величину:

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

Второй пример. При скрещивании короткоухих овец (являющихся гетерозиготами, полученными путем скрещивания нормальных длинноухих овец с овцами, лишенными наружного уха) было получено 22 потомка, в том числе 7 овец с нормальными ушами, 9 короткоухих и 6 безухих. Так как гетерозиготы по фактору длины ушей у овец фенотипически отличаются от гомозиготных форм, ожидается в F_2 расщепление 1:2:1. Для получения ожидаемых категорий надо 22 умножить на 1/4 и на 1/2. Получим 5,5; 11,0 и 5,5.

Сопоставление фактических результатов с ожидаемыми может быть произведено с помощью табл. 38.

Таблица 38

Вычисление критерия хи-квадрат для данных о расщеплении по длине ушей у овец

Частоты		$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
O	E			
7	5,5	1,5	2,25	0,410
9	11,0	—2,0	4,00	0,364
6	5,5	0,5	0,25	0,045
$\Sigma=22$	$\Sigma=22$			$\Sigma=0{,}819$

$\chi^2=0{,}82$ при $d.f.=3-1=2$.

По табл VIII находим, что это значение хи-квадрат имеет вероятность, среднюю между 0,75 и 0,50, примерно 0,67.

Таким образом, наблюдается довольно хорошее соответствие между фактическими и теоретически ожидаемыми частотами. Исходную нулевую гипотезу, что в данном случае получено расщепление в отношении 1:2:1, можно считать правильной.

По тому же принципу производится вычисление ожидаемых чисел при более сложных типах расщепления, например 9:3:3:1; 12:3:1 и т. д. Так, например, допустим, что наблюдается расщепление по фенотипу на 4 группы при обычном дигибридном скрещивании: AB—117,

Ab—26, *aB*—18 и *ab*—7. Всего 168. Тогда ожидаемые числа будут следующими:

$$AB - \frac{9}{16} \cdot 168 = 94{,}5; \; aB \text{ и } Ab - \text{по} \frac{3}{16} \cdot 160 = 31{,}5;$$

$$ab - \frac{1}{16} \cdot 168 = 10{,}5.$$

Для получения χ^2 надо вычислить следующие 4 величины:

$$\frac{(117 - 94{,}5)^2}{94{,}5} = 5{,}36$$

$$\frac{(26 - 31{,}5)^2}{31{,}5} = 0{,}96$$

$$\frac{(18 - 31{,}5)^2}{31{,}5} = 5{,}79$$

$$\frac{(7 - 10{,}5)^2}{10{,}5} = 1{,}17$$

Отсюда $\chi^2 = 13{,}28$

Так как групп 4, число степеней свободы 3. Такое высокое значение χ^2 дает основание отвергнуть нулевую гипотезу и считать, что существует достоверное отклонение от ожидаемого отношения.

Вычисление теоретически ожидаемых чисел и определение хи-квадрат при данных, сгруппированных в многопольные таблицы. Критерий хи-квадрат можно применить для анализа многих других опытных данных—физиологических, генетических, медицинских, сельскохозяйственных,—когда анализируется влияние различных факторов на те или иные биологические процессы и явления. Правда, для этих случаев современная вариационная статистика дает в распоряжение биолога также и другие методы, рассмотрение которых выходит за рамки нашего курса. Данные подобных опытов обычно можно сгруппировать в таблицы, состоящие из нескольких полей (2×2, 2×3, 4×4 и т. д.). Исходной нулевой гипотезой, которая должна быть или отвергнута после определения χ^2 или, наоборот, сохранена, является принятие независимости в вариации изучаемых признаков, отсутствие различий во влиянии тех или других факторов и т. д. Выше уже приводился пример с изучением влияния облучения рентгеновыми лучами на частоту

сцепленных с полом мутаций у дрозофилы при подкормке солями железа и без нее. В опытах было получено 2756 культур с применением подкормки, 805 культур без подкормки. Среди первых мутации были получены в 357 культурах, а в 2399 культурах мутаций не было. Среди вторых мутации были в 80 культурах, а в 725 культурах мутации не наблюдались.

Исходная нулевая гипотеза заключается в том, что число мутаций не увеличивается при наличии подкормки, т. е. что частота мутаций, как в группе получавших подкормку, так и в группе не получавших, одинакова. Исходя из такой гипотезы, можно рассчитать ожидаемое число культур с мутациями и без мутаций в каждой группе. Фактически полученные и записанные в скобках теоретически ожидаемые частоты представлены в табл. 39.

Таблица 39

Проверка зависимости частоты вызванных облучением мутаций от подкормки солями железа

Группы	Число культур F_2		Всего
	давшие мутации	не давшие мутации	
С подкормкой	357(338,2)	2399(2417,8)	2756
Без подкормки	80(98,8)	725(706,2)	805
Всего	437	3124	3561

Каждая теоретическая частота вычислена на основе итоговых цифр 3561, 437 и 3124. Так, для верхней левой клетки она будет равна $\frac{2756 \cdot 437}{3561}$, а для верхней правой — $\frac{2756 \cdot 3124}{3561}$. Имеется в виду, что между числом культур, давших мутации и не давших, в группе с подкормкой будет такое же соотношение, как в опыте в целом. Таким же образом вычисляются теоретически ожидаемые частоты и для двух других полей: общая сумма «805» распределяется также пропорционально 437 и 3124.

Дальнейшие расчеты сводятся к получению четырех отклонений фактических чисел от теоретических, возведению каждого из них в квадрат, делению каждого

квадрата отклонений на теоретическое число и, наконец, к суммированию, в результате чего будет получен хи-квадрат. Когда опытный материал может быть расположен в виде таблицы с горизонтальными строчками и вертикальными столбцами, число степеней свободы вычисляется как произведение $(r-1)(c-1)$. В примере с мутациями $d.f.=1$. Для таблиц с 4 полями, частоты в каждом из которых обозначаются как a, b для верхних полей и c и d для нижних, а общая сумма $a+b+c+d=n$, можно применить следующую формулу для критерия соответствия:

$$\chi^2 = \frac{(ad-bc)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}. \qquad (65)$$

Распределение χ^2, как это видно из табл. VIII, является непрерывным, распределение же групп в таблицах, подобных табл. 39, дискретно. Поэтому применение критерия χ^2 к случаям сопоставления фактических и ожидаемых частот при дискретных распределениях сопряжено с некоторой неточностью, особенно если число наблюдений в группах мало. Одним из способов уменьшения неточности является объединение малочисленных групп, примеры чего будут даны ниже. Возможно также внесение поправки в числитель формулы (65), так называемой поправки на непрерывность Иейтса. Формула с поправкой будет следующей:

$$\chi^2 = \frac{\left[(ad-bc)-\frac{1}{2}n\right]^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}. \qquad (65a)$$

Вычисление ожидаемых частот для теоретических вариационных рядов и определение соответствия эмпирических рядов теоретическим. После составления вариационного ряда и вычисления характеризующих его статистических показателей—средней арифметической и среднего квадратического отклонения—возникает необходимость установить, насколько полученное фактически распределение соответствует тому или другому из известных теоретических распределений. В нашем курсе мы ограничились только тремя распределениями: биномиальным, нормальным и пуассоновским. Разберем на простейших примерах применение метода χ^2 для сравнения эмпирических распределений с теоретическими.

В качестве примера биномиального распределения возьмем распределение числа хрячков в пометах свиноматок, в каждом из которых было по 6 поросят. Оно показано в первых 2 графах табл. 40.

Таблица 40

Сравнение эмпирического распределения числа хрячков в пометах свиней с теоретически ожидаемым при биномиальном распределении

Количество хрячков в помете	Фактическое число пометов O	Теоретически ожидаемое E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	3 } 19	3,45 } 24,15	−5,15	26,5225	1,10
1	16	20,70			
2	53	51,75	+1,25	1,5625	0,03
3	78	69,00	+9,00	81,0000	1,17
4	53	51,75	+1,25	1,5625	0,03
5	10 } 18	20,70 } 24,15	−6,15	37,8225	1,57
6	8	3,45			
	$n=221$	$n=220,8$ (~ 221)			$\chi^2=3,90$

Так как в данном случае имеется дискретная, прерывистая, изменчивость, следует ожидать биномиального распределения, в котором ряд получается на основе разложения бинома $(p+q)^k$. Приняв, что $p=q=\frac{1}{2}$, можно вычислить ожидаемые частоты каждого класса, установив величину k. Так как в данном ряду 7 классов, то $k=6$ (напомним, что можно воспользоваться для этой цели треугольником Паскаля). Тогда частоты классов будут выражаться следующими цифрами:

$$221\cdot\frac{1}{64};\ 6.221\cdot\frac{1}{64};\ 15.221\cdot\frac{1}{64};\ 20.221\cdot\frac{1}{64};$$

$$15.221\cdot\frac{1}{64};\ 6.221\cdot\frac{1}{64};\ 221\cdot\frac{1}{64}.$$

После выполнения арифметических вычислений будут получены теоретически ожидаемые частоты. Они записаны в третьей графе табл. 40. Одним из условий правильного применения критерия соответствия является

наличие в каждом из эмпирических или теоретических классов не менее 5 вариант. Поэтому следует объединить две верхних строчки (0 и 1) в один класс и то же проделать с двумя нижними строчками. Все дальнейшие вычисления, приведенные в табл. 40, само собой ясны. По указанному выше правилу число степеней свободы $d.f.=5-2=3$.

Пользуясь табл. VIII, мы находим, что вычисленное значение χ^2 превышает табличное, находящееся в графе $P=0{,}50$, но меньше табличного графы 0,25. Таким образом, обнаруживается хорошее соответствие фактических частот вариационного ряда с ожидаемым при биномиальном распределении.

По тому же принципу проводится сравнение эмпирических частот вариационного ряда с теоретическими и вычисление χ^2 при нормальном и пуассоновском распределениях. Но рассчитать теоретические частоты при этих распределениях значительно труднее.

Напомним, что пуассоновское распределение в принципе является тем же биномиальным, но относится к явлениям, обладающим очень малой вероятностью.

Поэтому оно асимметрично. Как указано в гл. 3, характерным признаком для пуассоновского распределения является то, что средний квадрат отклонений и средняя арифметическая ($\bar{x}$ или λ) количественно почти равны. Именно по этому признаку можно отличить пуассоновское распределение от других распределений.

Теоретические частоты пуассоновского распределения представляют собой следующий ряд:

$$\frac{n}{e^{\lambda}} \text{ (нулевой член)}, \frac{n\lambda}{e^{\lambda}}, \frac{n\lambda^2}{2e^{\lambda}}, \frac{n\lambda^3}{(2)(3)e^{\lambda}}, \frac{n\lambda^4}{(2)(3)(4)e^{\lambda}} \text{ и т. д.}$$

Здесь n—общее количество вариант в вариационном ряду, e—основание натуральных логарифмов (значение его приблизительно равно 2,718, а его логарифм при основании 10 равен 0,43429...), λ—средняя арифметическая вариационного ряда при пуассоновском распределении ($\bar{x}$). Поскольку в значение частот входит величина e, возведенная в степень λ, расчеты надо вести с помощью логарифмов, позднее же по логарифму определить данную частоту.

Для удобства расчетов ряд теоретических частот выгоднее представить в следующем виде:

$$\frac{n}{e^{\lambda}},\ \left(\frac{n}{e^{\lambda}}\right)\cdot\lambda,\ \left(\frac{n\lambda}{e^{\lambda}}\right)\left(\frac{\lambda}{2}\right),\ \left(\frac{n\lambda^2}{2e^{\lambda}}\right)\left(\frac{\lambda}{3}\right),\ \left(\frac{n\lambda^3}{2\cdot 3e^{\lambda}}\right)\left(\frac{\lambda}{4}\right)\text{ и т. д.}$$

Достаточно вычислить первый член и тогда все последующие члены можно получить из предыдущих путем умножения на λ, $\frac{\lambda}{2}$, $\frac{\lambda}{3}$, $\frac{\lambda}{4}$, $\frac{\lambda}{5}$ и т. д. На конкретном примере ход вычислений с применением логарифмов будет выглядеть следующим образом. Допустим, что средняя арифметическая $\lambda=3{,}0204$. Средний квадрат отклонений, т. е. σ^2, также равен 3,0204. Отсюда можно сделать вывод, что распределение пуассоновское. Число вариант $n=98$. Первый член ряда равен $\frac{98}{e^{3{,}0204}}$.

Логарифмируем его: $\log 98 = 1{,}99123$

$$\log\left(e^{3{,}0204}\right) = 3{,}0204\cdot\log e =$$

$$= 3{,}0204\cdot 0{,}43295 = 1{,}31175.$$

При логарифмировании дроби надо от логарифма числителя отнять логарифм знаменателя:

$$1{,}99123 - 1{,}31175 = 0{,}67948.$$

Таков логарифм искомой частоты. По логарифму определяем частоту, которая равна 4,78.

Зная первый член ряда, все последующие можно получить и без помощи логарифмов. Так, для получения второго члена надо умножить число 4,78 на значение λ, т. е. на 3,0204. Получим частоту 14,44. Для получения третьего члена надо помножить частоту второго члена на $^1/_2$ 3,0204. Получим частоту 21,81 и т. д.

Но применение логарифмов для вычисления всех последующих членов пуассоновского ряда избавит от производства кропотливых действий умножения и деления.

Так, для получения второго члена, логарифм которого уже известен и равен 0,67948, к величине 0,67948 надо прибавить log 3,0204, который равен 0,48007. Сумма двух логарифмов равна 1,15255, отсюда частота—14,44. Для получения третьего члена к значению логарифма

1,15955 надо снова прибавить log 3,0204, т. е. 0,48004, и отнять log 2, который равен 0,30103. Получим 1,33859, по которому определяем частоту третьего члена ряда, равную 21,81 и т. д. Все операции с помощью логарифмов можно проводить на одной таблице, прибавляя и вычитая соответствующие логарифмы.

После вычисления теоретических частот для всех классов распределения составляется таблица, аналогичная табл. 40. Сумма величин последней графы $\frac{O-E^2}{E}$ и дает искомое значение χ^2. Если количество вариант в каком-либо из крайних классов меньше 5, следует объединить его с 1—2 соседними. Объединение должно быть проведено одинаково как по фактическому ряду, так и по теоретически ожидаемому.

Число степеней свободы устанавливается тем же путем, как и при биномиальном распределении ($d.f.=k-2$). Так как все дальнейшие операции по вычислению χ^2 просты, мы не даем специального примера на пуассоновское распределение.

Последним из трех рассмотренных распределений является нормальное. При непрерывной количественной изменчивости очень важно знать, в какой степени полученный фактически вариационный ряд следует нормальному распределению. Критерий хи-квадрат позволяет достаточно легко установить степень такого соответствия. И здесь более сложным и трудоемким является установление теоретических численностей каждого класса вариационного ряда при нормальном распределении. Необходимые для этого вычисления можно показать на примере табл. 41.

Статистические показатели для этого ряда следующие: $\bar{x}=134{,}3$ мм, $\sigma=21{,}3$ мм.

Задача вычисления теоретических частот сводится к тому, чтобы отнести к уже имеющимся классам возможное значение частот, если они распределены по законам нормального распределения. Как было показано в гл. 3, нормальное распределение очень хорошо выражается в сигмах. На этом принципе построена табл. I, с помощью которой можно определить, какая часть вариант находится в пределах того или другого значения t, т. е. нормированного отклонения, выраженного в сигмах. Необходимо перевести имеющиеся классы, вы-

раженные в мм, в классы, выраженные в сигмах или долях сигмы, и после этого установить, сколько вариант должно приходиться на каждый данный отрезок нормальной кривой, ограниченный определенными значениями сигмы. Вычисление частот нормальной кривой может быть проделано разными способами. Мы разберем один из них, наиболее простой.

Таблица 41

Фактический вариационный ряд распределения 300 початков **кукурузы по длине (в мм) и теоретически вычисленный ряд в соответствии с нормальным распределением**

Центральные значения классов	Фактические частоты O	Теоретически вычисленные частоты E		$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
80	1	2,17	8,6	— 5,6	31,36	3,646
90	2	6,45				
100	17	15,3		1,7	2,89	1,889
110	39	29,3		9,7	91,09	3,211
120	44	44,8		— 0,8	0,64	0,014
130	66	55,0		11,0	121,00	2,200
140	42	54,2		—12,2	148,84	2,746
150	34	43,0		— 9,0	81,00	1,884
160	29	27,3		1,7	2,89	0,106
170	18	13,9		4,1	16,81	1,209
180	3	5,68	8,0	0,0	0	0,000
190	3	1,86				
200	2	0,49				
	$n=300$	$n=295,4 = \sim 300$				$\chi^2=16,905$

Возьмем в качестве примера класс табл. 41 с центральным значением «100». Границы этого класса 95,0 и 104,9. В значениях сигмы они будут следующими:

$$\frac{95,0-134,3}{21,3}=-1,845\sigma \text{ и } \frac{104,9-134,3}{21,3}=-1,380\sigma.$$

Какая же доля из общего числа вариант при нормальном распределении должна быть в интервале между $-1,845\sigma$ и $-1,380\sigma$?

Ответить на это можно с помощью табл. I, но так как это потребовало бы некоторого дополнительного перерасчета цифр, то удобнее пользоваться табл. IX, в которой

даны готовые частоты для каждого отрезка нормальной кривой в так называемом накопленном виде, т. е. последующие частоты прибавлены к предыдущим. По этой табл. мы находим, что значению $\sigma=1{,}845$ соответствует величина 4673, а значению $\sigma=1{,}380$—величина 4162. Это числа особей при общем числе 10 000. Их можно выразить и как доли—в виде дробей 0,4673 и 0,4162. Тогда доля особей в этом интервале между двумя значениями накопленных частот составит 0,4673—0,4162=0,0511. Ожидаемая частота для данного класса при $n=300$ будет равна 0,0511.300=15,33. Таким же методом можно вычислить теоретические частоты для всех других классов. Они внесены в табл. 41 в готовом виде. Для определения хи-квадрата целесообразно, как это было сделано в других рядах, присоединить классы с малым числом вариант к соседним. В окончательном виде значение $\chi^2=16{,}905$. Число степеней свободы в данном случае 10—3=7, так как теоретический и эмпирические ряды имеют 3 общих элемента: общее количество вариант, среднее квадратическое отклонение и среднюю арифметическую.

По табл. VIII обнаруживаем, что полученное значение χ^2 выше табличного в графе с $P=0{,}05$, но ниже табличного значения при $P=0{,}01$. Очевидно, мы встречаемся здесь как раз с тем случаем, когда соответствие фактического ряда частот теоретическому нормальному недостаточно, так как границей соответствия обычно считается вероятность 0,05.

Сравнение двух эмпирических распределений между собой. Наряду со сравнением эмпирических распределений с теоретическими иногда нужно сравнить два эмпирические распределения друг с другом. Формула χ^2 в этих случаях несколько сложнее, а именно:

$$\chi^2 = \frac{1}{n_1\ n_2} \Sigma \frac{(f_1 n_2 + f_2 n_1)^2}{f_1 + f_2}, \qquad (66)$$

где f_1 *и* f_2—частота классов первого и второго рядов, а n_1 и n_2—число особей в каждом из них. В качестве примера можно взять данные табл. 42.

$\Sigma \frac{(f_1 n_2 + f_2 n_1)^2}{f_1 + f_2}$ надо помножить на $\frac{1}{n_1\ n_2} = \frac{1}{4104}$.

В результате будет получено следующее значение χ^2:

$$\chi^2 = \frac{523019{,}7}{4104} = 127{,}44.$$

Число степеней 11—1=10, так как единственным связующим элементом двух рядов является то, что их классы сопоставлены попарно.

Таблица 42

Вариационные ряды промеров длины яиц кукушки

x	f_1	f_2	f_1n_2	f_2n_1	$f_1n_2+f_2n_1$	$(f_1n_2+f_2n_1)^2$	$\frac{(f_1n_2+f_2n_1)^2}{f_1+f_2}$
40	1	7	54	532	586	343396	42924,5
41	1	5	54	380	434	188356	31392,7
42	8	14	432	1064	1496	2288016	101728,0
43	3	8	162	608	770	592900	53900,0
44	9	9	486	684	1170	1368900	76050,0
45	13	6	702	456	1158	1340964	70577,1
46	20	3	1080	228	1308	1710864	74385,4
47	6	2	324	152	476	226576	28322,0
48	11	—	594	—	594	352836	32076,0
49	2	—	108	—	108	11664	5832,0
50	2	—	108	—	108	11664	5832,0
	$n_1=76$	$n_2=54$					523019 7

Проверка по табл VIII показывает высокую достоверность различия между рядами ($P<0{,}001$) Нулевая гипотеза, что оба ряда взяты из одной популяции, должна быть отвергнута.

ВОПРОСЫ

1 Зачем нужно измерять соответствие фактических данных ожидаемым?

2 Что такое критерий соответствия хи квадрат? Напишите общую формулу для его вычисления

3 Каковы закономерности распределения хи квадрат? В каком случае χ^2 должен быть равен нулю? Почему распределение χ^2 асимметрично?

4. При каких значениях χ^2 следует отвергнуть нулевую гипотезу?

5 В каких границах вероятности значения χ^2 указывают на соответствие между фактическими и теоретически ожидаемыми данными?

6 Имеется ли шанс на правильность нулевой гипотезы, если она отбрасывается? Каков этот шанс при вероятности 0,01, при вероятности 0,02?

7 Объясните, почему границей для соответствия признается 0 05 Какова в этом случае вероятность того, что несоответствия действительно нет?

8 Что такое область отбрасывания нулевой гипотезы?

9 Проведите параллель между уровнями значимости 0 05 и 0 01 при установлении достоверности разницы между статистическими показателями и вероятностями 0,05 и 0 01 при анализе соответствия Укажите также на различия

10 Можно ли делать выводы о правильности научных гипотез только на основе χ^2?

11 Как устанавливается число степеней свободы при пользовании критерием хи квадрат? Сколько степеней свободы при расщеплении 12 3 1 при расщеплении 9 3 3 1, при распределении данных в таблице с числом полей 4×4 при сравнении эмпирического вариационного ряда, состоящего из 10 классов с нормальным?

12 Как производится суммирование нескольких χ^2?

13 Изложите способ определения степени однородности или разнородности опытного материала

14 Как рассчитать ожидаемые частоты при расщеплении 1 2 1 1 1, 2 1 9 3 3 1?

15 Напишите рабочую формулу χ^2 для случая когда известно теоретическое отношение классов в популяции

16 Как рассчитать ожидаемые частоты в классах таблицы с 4 полями?

17 Напишите формулу критерия соответствия для таблицы с 4 полями

18 В чем заключается поправка на непрерывность в формуле для χ^2?

19 Как вычислить ожидаемые частоты при биномиальном распределении?

20 Как вычислить ожидаемые частоты при пуассоновском распределении? Напишите формулу для первого (нулевого) члена пуассоновского распределения

21 Почему следует объединять эмпирические классы с малым числом вариант?

22 Изложите методы вычисления теоретических частот при нормальном распределении с помощью табл IX

23 Почему при сравнении эмпирического ряда с нормальным число степеней свободы меньше числа классов на 3, а не на 2?

24 Напишите формулу χ^2 при сравнении двух эмпирических вариационных рядов

25 Как устанавливается число степеней свободы при определении соответствия двух эмпирических вариационных рядов?

ЗАДАЧИ

90 При скрещивании ангорских длинношерстных кроликов с кроликами, имевшими нормальную длину шерсти, в первом поколении были получены нормальношерстные кролики От обратного скрещивания их с ангорскими было получено 62 крольченка, в том числе 33 нормальношерстных и 29 длинношерстных Соответствует ли это отношению 1 : 1?

91 По окраске фасолин наблюдали следующее расщепление: сильно окрашенных 92, наполовину окрашенных 182 и имеющих только небольшую окрашенную зону 81. Проверьте соответствие полученных частот с ожидаемыми при расщеплении 1:2:1

92 Среди 162 детей, наследовавших от одного из родителей фактор группы крови M, а от другого фактор N, оказалось: 46 с группой крови M, 68 с группой крови MN и 48 с группой крови N. Рассчитайте ожидаемые частоты при отношении 1:2:1 между группами M, MN и N и определите степень соответствия эмпирических данных теоретически ожидаемым с помощью χ^2.

93 При скрещивании особей, несущих лактоглобулины молока A и B, было получено 14 коров, из которых 2 было с лактоглобулином A, 6 с лактоглобулином B и 6 с обоими лактоглобулинами A и B. Проверьте соответствие полученных данных с ожидаемыми при гипотезе, что расщепление идет по формуле $1A:2AB:1B$

94 У рачков гаммарусов наблюдалось следующее расщепление по окраске глаз:

Номера семей	Количество особей	
	с черными глазами	с красными глазами
1	79	14
2	120	31
3	81	27
4	95	29
5	139	57
6	24	9
7	19	8
8	45	11

Проверьте соответствие полученного расщепления теоретически ожидаемому 3:1 по отдельным семьям и по всему материалу в целом (2 способами)

95. В четырех сериях опытов было получено в F_2 следующее расщепление по окраске у кур

Номера серии	Темноокрашенных	Белых
1	112	43
2	76	22
3	146	12
4	143	44

Проверьте соответствие полученных данных ожидаемым при отношении 3:1 по всему материалу в целом и по каждой серии отдельно. Однороден ли опытный материал?

96. У 10 растений гороха наблюдали следующее количество круглых *A* и морщинистых *a* бобов

Номера растений	1	2	3	4	5	6	7	8	9	10
A	45	27	24	19	32	26	88	22	28	25
a	12	8	7	10	11	6	24	10	6	7

Проверьте с помощью критерия хи-квадрат соответствие полученных данных ожидаемому при расщеплении в отношении 3 : 1 сначала для каждого отдельного растения, а затем для всех 10 растений вместе, получив сводное значение χ^2 Обратите внимание на число степеней свободы сводного χ^2

97 При обратном скрещивании томатов, гетерозиготных по зеленой листве, с томатами, имеющими желтую листву, было получено 671 растение с зеленой листвой и 569 — с желтой Вычислите χ^2 и определите по табл VIII его достоверность Какое биологическое объяснение можно дать факту отставания количества рецессивных форм от ожидаемого?

98 В одной серии опытов по инъекции в яйца кур женских половых гормонов вылупилось на 21 особь только 2 нормальных самца, все остальные были или нормальными самками или проявляли какие-то признаки женского пола Является ли такое отношение полов только крайним отклонением от нормального отношения 1 : 1 или же оно достоверно от него отличается?

99 В стаде крупного рогатого скота за 7 лет было зарегистрировано 6972 бычка и 7126 телочек Проверьте гипотезу, что отношение полов у крупного рогатого скота 1 : 1

100 В F_2 дигибридного скрещивания было получено расщепление по фенотипу 82 *AB*, 12 *Ab*, 33 *aB* и 8 *ab* Проверьте с помощью χ^2 соответствие с ожидаемым отношением 9 : 3 : 3 : 1

101 В опытах с анализом дигибридного распределения в двух группах были получены значения χ^2, равные 13,28 и 9,82 При объединении данных в одну группу $\chi^2 = 12{,}37$ Вычислите степень неоднородности между двумя группами. Какое число степеней свободы для суммы двух χ^2, для χ^2 объединенных данных?

102 Было проверено действие двух концентраций одного и того инсектицида на тлей Результаты оказались следующими:

Концентрация инсектицида, %	Количество тлей	
	выживших	погибших
1	3	62
0,5	13	55

Определите χ^2 и сделайте выводы

103 Во время эпидемии под наблюдением было 32 больных. К 18 из них применили новое лечебное средство В результате 15

человек выздоровело и 3 умерло. Из 14 человек, лечившихся прежними лекарствами, умерло 9 и выздоровело 5. Вычислите χ^2 двумя методами по формуле (65) и по формуле (65а) и сделайте выводы относительно результативности нового лечебного средства

104 Прививки против сыпного тифа 18 483 людям дали следующие результаты:

Группы	Количество		Всего
	заболевших	незаболевших	
Получившие прививку	56	6759	6815
Не получившие прививку	272	11396	11668
Всего	328	18155	18483

Примените критерий χ^2 к анализу роли прививок против сыпного тифа и сделайте выводы.

105. Проверьте соответствие фактического вариационного ряда, составленного по данным задачи 1, теоретическому при нормальном распределении.

106. Проанализируйте вариационный ряд, составленный по данным задачи 2. К какому типу распределения он относится? Проверьте соответствие с предполагаемым теоретически типом распределения.

107. Проверьте соответствие вариационного ряда, составленного по данным задачи 7, с теоретическим при нормальном распределении.

108. Проверьте соответствие вариационного ряда, представленного в табл. 4, теоретическому, предполагая биномиальное распределение.

109. Проверьте соответствие вариационного ряда, представленного в табл. 8, теоретическому при нормальном распределении.

110. Примените χ^2 к данным табл. 16.

111. Соответствует ли нормальному распределению вариационный ряд задачи 19?

ЗАКЛЮЧЕНИЕ

Тех, кто впервые встречается с вариационно-статистическими методами, обычно пугает обилие вычислительной работы. Однако надо помнить, что глубокий анализ биологических вопросов не может быть проведен без применения статистических методов, без вычисления необходимых статистических показателей и установления степени их достоверности. Известно, что между живыми существами постоянно наблюдается изменчивость как по качественным, так и особенно по количественным признакам, которую надо выразить в точных и конкретных показателях. Животные, как правило, неодинаковы по своим наследственным, природным качествам и в то же время находятся под непрерывным воздействием многообразных факторов внешней среды. Реакции различных организмов на условия внешней среды, на кормление, содержание, воспитание также неодинаковы. Ставя любой, самый простейший опыт с животными, необходимо считаться с рядом осложняющих условий, множеством чисто случайных и не поддающихся точному учету факторов, влияющих на опытных и контрольных животных и изменяющих их биологические и хозяйственные признаки. Вот почему статистические понятия и подходы сейчас нераздельно входят в биологическую науку.

Простое вычисление средней арифметической представляет собой анализ статистического процесса, так как главное при этом заключается в установлении свойств не одного какого-либо животного, а всей изучаемой группы, стада, породы. Промеры 50 коров во всех случаях дадут лучшее представление о породе, нежели

промеры одной коровы, но только применение статистических методов позволит ответить на вопрос, достаточно ли пятидесяти коров, чтобы судить о породе, и какова степень достоверности выводов.

Однако как ни важно умение пользоваться теми или иными методами или приемами вычислений, главное заключается в понимании их сути, в понимании значения математического и статистического подхода к биологическим явлениям. При любом анализе—данных ли опыта или результатов наблюдений—громадное значение имеют понятия вероятности, значимости, достоверности. Объективно оценить полученные из опыта или наблюдения данные—это значит суметь оценить их достоверность. Наличие разницы между показателями двух групп животных, опытной и контрольной, само по себе не является доказательством достоверного различия между ними, доказательством влияния того или иного изучаемого фактора, если при этом не установлена достаточная статистическая достоверность этой разницы. Но надо иметь в виду, что недостаточная достоверность выводов еще не является основанием для того, чтобы полностью отвергнуть возможность влияния того или иного фактора. При недостаточной достоверности результатов необходимо вновь повторить опыт или наблюдение, чтобы снизить статистическую ошибку и в конце концов окончательно убедиться в достоверности или, наоборот, в недостоверности выводов. Конечно, всякий опыт должен быть правильно поставлен. Плохой опыт никогда не может дать точных и достоверных результатов, как бы хорошо ни обрабатывали его данные статистически.

Надо постоянно помнить, что математические методы при всем их значении не могут заменить биологических методов. Вариационная статистика лишь помогает биологическому исследованию, делает его более точным. Установить причину тех или иных биологических явлений или связей между ними можно только с помощью биологического исследования.

Учитывая непрерывное расширение применения математических и вариационно-статистических методов в биологии и развитие биостатистики как самостоятельной научной дисциплины автору очень хотелось сделать данное руководство возможно более полноценным и насыщенным современными статистическими методами и

подходами. Но в то же время надо было учитывать его цель, что оно должно служить, прежде всего, учебным пособием по курсу вариационной статистики, для которого на биологических факультетах университетов и в других вузах биологического профиля отводится ограниченное число учебных часов. Поэтому в книге изложены преимущественно элементарные основы вариационной статистики, знание которых необходимо для практической работы биолога, а также для начинающего научного работника в области биологии. Именно такой элементарный курс в настоящее время особенно нужен, так как изданный в 1935 г. прекрасный учебник Сапегина «Вариационная статистика» давно стал библиографической редкостью.

Новые подходы в биологической статистике, особенно связанные со школой Фишера, были в некоторой степени учтены.

В книге использованы новые формулы и данные, а также конкретные примеры из основных советских и иностранных руководств по вариационной статистике. В процессе работы была использована следующая литература:

1. Немчинов В. С. Сельскохозяйственная статистика с основами общей теории. Огиз—Сельхозгиз, М.—Л., 1945.

2. Романовский В. И. Применение математической статистики в опытном деле. Огиз—Гостехиздат, М.—Л., 1947.

3. Финни Д. Применение статистики в опытном деле (сельское хозяйство). Госстатиздат, М., 1957.

4. Фишер Р. А. Статистические методы для исследователей. Госстатиздат, М., 1958.

5. Миллс Ф. Статистические методы. Госстатиздат, М., 1958.

6. Snedocor G. W. Statistical methods. 5-th edition. Iowa State College Press, Ames, Iowa, 1957.

7. Weber E. Grundriss der biologischen Statistik. 2-e Auflage. G. Fischer Verlag, Jena, 1956.

8. Bailey N. T. I. Statistical methods in biology. English Universities Press, London, 1959.

9. Bancroft. Introduction to biostatistics. Cassell a Co, London, 1957.

10. Otto E. Biometrie. Deutscher Bauernverlag, Halle, 1958.

Многие вопросы, сейчас широко освещаемые в руководствах по вариационной статистике и важные для биолога-исследователя, как-то: методика постановки и планирования опытов, дисперсионный, ковариансный и дискриминантный методы, способы взятия проб и выборок и др., естественно, не нашли отражения в данной книге. Поэтому желающим расширить свои знания необходимо ознакомиться со специальными руководствами из числа перечисленных выше. Наиболее систематическое изложение вариационной статистики, исходя из интересов биолога, агронома, зоотехника, имеется в книге акад. Немчинова, а также в иностранных руководствах Снедекора (на английском языке) и Э. Вебер (на немецком языке).

Преподавание курса вариационной статистики обязательно требует упражнений и решения задач. Чтобы облегчить работу преподавателя, были составлены, по собственным данным и по различным литературным источникам, свыше 100 таких задач и упражнений для всех разделов курса. Однако желательно, чтобы в каждом вузе были использованы для обработки вариационно-статистическими методами собственные экспериментальные данные. Для лучшего усвоения теоретического материала в конце каждой главы приведены проверочные вопросы.

В приложениях даны: а) перечень статистических показателей с их символами и формулами; б) девять статистических таблиц (I—IX). Они взяты из различных источников (Немчинов, Снедекор, Бейли, Вебер), но чаще всего несколько упрощены и сокращены по сравнению с исходными таблицами. Совершенно необходимо, чтобы учащиеся привыкли пользоваться статистическими таблицами, так как это не только облегчает статистический анализ конкретного материала, но и значительно углубляет понимание самих статистических закономерностей, приучает учащихся к важнейшим статистическим понятиям, особенно таким, как вероятность, достоверность, значимость, распределение и др.

Все отзывы о книге, критические замечания и указания просьба направлять автору по адресу: г. Минск, Белорусский государственный университет, кафедра зоологии позвоночных животных, проф. П. Ф. Рокицкому.

СТАТИСТИЧЕСКИЕ ПОКАЗАТЕЛИ, СИМВОЛЫ И ФОРМУЛЫ

Ввиду разнообразия обозначений, которые применяются в литературе для одних и тех же показателей, в квадратных скобках приведены разные наименования данного статистического показателя и некоторые из встречающихся символов. В скобках справа формул указаны номера, под которыми они были даны в тексте. Некоторые из более общих формул были приведены в тексте без номеров.

Варианса [средний квадрат, дисперсия; σ^2 или s^2].

Общие формулы $$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}, \qquad (6)$$

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{d.f.};$$

при **малом** n $$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}, \qquad (6а)$$

или $$\sigma^2 = \frac{1}{n - 1} \cdot \Sigma (x - \bar{x})^2. \qquad (6б)$$

Рабочие формулы

а) $$\sigma^2 = \frac{\Sigma (x - A)^2}{n} - (A - \bar{x})^2 \qquad (8)$$

или $$\sigma^2 = \frac{\Sigma f (x - A)^2}{n} - (A - \bar{x})^2; \qquad (8б)$$

б) то же при малом n:

$$\sigma^2 = \frac{\Sigma (x - A)^2 - n (A - \bar{x})^2}{n - 1}; \qquad (8а)$$

в) $$\sigma^2 = \frac{\Sigma x^2 - \Sigma x \cdot \bar{x}}{n}; \qquad (10)$$

г) то же при малом n:

$$\sigma^2 = \frac{\Sigma x^2 - \Sigma x \cdot \bar{x}}{n - 1}; \qquad (10а)$$

д) $$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - k}, \qquad (10б)$$

Варианса взвешенная:

$$\sigma_g^2 = \frac{\sigma_1^2 (n_1 - 1) + \sigma_2^2 (n_2 - 1) + \ldots + \sigma_k (n_k - 1)}{n - k}. \qquad (11)$$

Варианса линии регрессии:

$$\sigma^2_{y\,x} = \frac{\Sigma d^2_{y\,x}}{n - 2}, \qquad (51)$$

где $$\Sigma d^2_{y.x} = \Sigma (y - \bar{y})^2 - \frac{[\Sigma (x - \bar{x})(y - \bar{y})]^2}{\Sigma (x - \bar{x})^2}. \qquad (53)$$

Вариансы разложение:

$$\sigma_0^2 = \sigma_1^2 + \bar{\sigma}^2; \qquad (12)$$

$$\sigma_0^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots + \bar{\sigma}^2. \qquad (12a)$$

Вариансное отношение [критерий значимости F]:

$$F = \frac{\sigma_1^2}{\sigma_2^2}. \qquad (28)$$

Вероятность [p, но вероятность соответствия, а также уровень значимости —P]:

Общая формула $p = \frac{F}{S}$.

Вероятность дополнительная $[q; Q]$ $q = 1 - p$.

Ковариансa

$$Cov. = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{d.f.}. \qquad (60)$$

Корреляционное уравнение

$$t_y = r \cdot t_x.$$

Коэффициент изменчивости [коэффициент вариации; *c. v.*, *C*]:

$$c.v. = \frac{\sigma \cdot 100}{\bar{x}}. \qquad (13)$$

Коэффициент простой корреляции [прямолинейной, обычной, линейной; r]:

Общая формула

$$r = \frac{\Sigma\, t_x\, t_y}{n}. \tag{31}$$

Различные преобразования:

а) $$r = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{n\,\sigma_x\,\sigma_y} \tag{32}$$

или $$r = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{(n-1)\,\sigma_x\,\sigma_y}, \tag{32а}$$

б) $$r = \frac{\Sigma\, xy - n\,\bar{x}\,\bar{y}}{n\,\sigma_x\,\sigma_y}, \tag{33}$$

в) $$r = \frac{\Sigma\, xy - \dfrac{\Sigma x\,\Sigma y}{n}}{n\,\sigma_x\,\sigma_y}, \tag{34}$$

г) $$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x\,\sigma_y}, \tag{35}$$

д) $$r = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{\sqrt{\Sigma\,(x - \bar{x})^2\,\Sigma (y - \bar{y})^2}}, \tag{36}$$

е) $$r = \frac{\Sigma\, xy - n\,\bar{x}\,\bar{y}}{\sqrt{\Sigma\,(x - \bar{x})^2\;\Sigma\,(y - \bar{y})^2}}, \tag{33а}$$

ж) $$r = \frac{\Sigma\, xy - n\,\bar{x}\cdot\bar{y}}{\sqrt{(\Sigma\, x^2 - n\,\bar{x}^2)\;(\Sigma\, y^2 - n\,\bar{y}^2)}}. \tag{33б}$$

Рабочая формула для корреляционной решетки:

$$r = \frac{\Sigma\, fa_x\, a_y - n\, b_x\, b_y}{n\,\sigma_x\,\sigma_y}, \tag{37}$$

при альтернативной изменчивости:

$$r = \frac{a\,d - b\,c}{\sqrt{(a+b)\;(c+d)\;(a+c)\;(b+d)}}. \tag{44}$$

Коэффициент частной корреляции (при 3 признаках):

$$r_{xy\ z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1-r_{xz}^2)\ (1-r_{yz}^2)}}. \qquad (43)$$

Коэффициент частной корреляции (при 4 признаках):

$$r_{12\,34} = \frac{r_{12\,4} - r_{13\cdot 4}\ r_{23\ 4}}{\sqrt{(1-r_{13\ 4}^2)\ (1-r_{23\ 4}^2)}}. \qquad (43a)$$

Коэффициент корреляции в значениях коэффициента регрессии: $r = \sqrt{b_{x\,y} \cdot b_{y\,x}}$. (59)

Коэффициент регрессии

[$R_{x\,y} = R_{\frac{y}{x}}^{x} = b_{x\,y}$, а также $R_{y.x} = R_{\frac{x}{x}}^{y} = b_{y.x}$].

$$R_{x\cdot y} = r\,\frac{\sigma_x}{\sigma_y} \qquad (49)$$

и $$R_{y\cdot x} = r\,\frac{\sigma_y}{\sigma_x}, \qquad (49a)$$

$$b = \frac{\Sigma\,(x-\bar{x})\ (y-\bar{y})}{\Sigma\,(x-\bar{x})^2}, \qquad (50)$$

$$b = \frac{\Sigma\,xy - \frac{\Sigma x\ \Sigma y}{n}}{\Sigma\,x^2 - \frac{(\Sigma x)^2}{n}}. \qquad (50a)$$

Критерий соответствия хи-квадрат $[\chi^2]$.

Общая формула $\chi^2 = \Sigma\frac{(O-E)^2}{E}$. (63)

При определенном теоретическом отношении (r) частот классов в популяции

$$\chi^2 = \frac{(a-rb)^2}{r\,(a+b)}, \qquad (64)$$

Для таблицы из 4 полей

$$\chi^2 = \frac{(ad - bc)^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}, \qquad (65)$$

с поправкой на непрерывность:

$$\chi^2 = \frac{\left[(ad - bc) - \frac{1}{2}n\right]^2 \cdot n}{(a+b)(c+d)(a+c)(b+d)}, \qquad (65a)$$

для сравнения двух эмпирических распределений

$$\chi^2 = \frac{1}{n_1 n_2} \sum \frac{(f_1 n_2 + f_2 n_1)^2}{f_1 + f_2}. \qquad (66)$$

Критерий z [число z; преобразованный коэффициент корреляции]:

$$z = \frac{1}{2}\left[\log_e (1+r) - \log_e (1-r)\right].$$

Нормированное отклонение [критерий значимости t; иногда символ t применяют только для малых n, тогда при больших n обозначают символами u, T, d и др.]:

В общем виде $t = \frac{x_i - \bar{x}}{\sigma}$, (18)

а также $t = \frac{\bar{x} - \mu}{\sigma}$ или $t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$.

Для оценки достоверности $\bar{x}$

$$t = \frac{\bar{x} - 0}{s_{\bar{x}}}, \qquad (21)$$

для оценки достоверности разницы между $\bar{x}_1$ и $\bar{x}_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} \qquad (23)$$

или сокращенно $t = \frac{d}{s_d}$, (23a)

для оценки разницы между сигмами

$$t = \frac{\sigma_1 - \sigma_2}{s_{\sigma_1 - \sigma_2}}, \tag{26}$$

для оценки достоверности коэффициента корреляции

а) при больших n

$$t = \frac{r}{s_r}, \tag{40}$$

б) при малых n

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}, \tag{40a}$$

для оценки достоверности числа z

$$t = \frac{z}{s_z}, \tag{42}$$

для оценки достоверности коэффициента регрессии

$$t = \frac{b}{s_b}, \tag{55}$$

для оценки достоверности разницы между b_1 и b_2

$$t = \frac{b_1 - b_2}{s_{(b_1 - b_2)}}, \tag{58}$$

Средняя арифметическая [$\bar{x}$ или M — для выборочной совокупности; $\bar{x}_0$ или μ—для генеральной совокупности].

Общая формула для несгруппированных данных

$$x = \frac{\Sigma x}{n} \tag{1}$$

или

$$\bar{x} = \frac{1}{n} \Sigma x, \tag{1a}$$

при данных, сгруппированных в классы,

$$\bar{x} = \frac{\Sigma fv}{n}, \tag{2}$$

при использовании условной средней A

$$\bar{x} = A + b \cdot i = A + \frac{\Sigma fa}{n} \cdot i, \qquad (4a,\ 4)$$

взвешенная

$$\bar{x} = \frac{\bar{x}_1 p_1 + \bar{x}_2 p_2 + \ldots + \bar{x}_n p_n}{p_1 + p_2 + \ldots + p_n}, \qquad (3)$$

при альтернативной изменчивости

$$\bar{x} = \frac{p_1}{n}, \qquad (14,\ 14a)$$

при биномиальном распределении (в значениях вероятности)

$$\bar{x} = np; \qquad (16)$$

при пуассоновском распределении [$\bar{x}$ обозначается знаком λ]

$$\lambda = \bar{x} = np = \sigma^2.$$

Средняя арифметическая генеральной совокупности [$\bar{x}_0$; μ] — границы колебаний

$$\bar{x} - ts_{\bar{x}} \leqslant \mu \leqslant \bar{x} + ts_{\bar{x}}. \qquad (20)$$

Средняя геометрическая (G или $\bar{x}_g$)

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \ldots x_n}, \qquad (5)$$

$$\log \bar{x}_g = \frac{1}{n}(\log x_1 + \log x_2 + \ldots + \log x_n) \qquad (5a)$$

Среднее квадратическое отклонение [стандартное отклонение; стандарт; или s]:

Общая формула

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}, \qquad (7)$$

то же при малом n

$$\sigma = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}. \qquad (7a)$$

Рабочие формулы:

а) $$\sigma = \sqrt{\frac{\Sigma (x - A)^2}{n} - (A - \bar{x})^2}, \qquad (9)$$

б) $$\sigma = i \sqrt{\frac{\Sigma f a^2}{n} - b^2}, \qquad (9a)$$

в) $$\sigma = i \sqrt{\frac{\Sigma f a^2}{n} - \left(\frac{\Sigma f a}{n}\right)^2}, \qquad (9б)$$

в частотах или долях альтернативных классов:

$$\sigma_p = \sqrt{\frac{p_0 \cdot p_1}{n^2}}, \qquad (15)$$

$$\sigma_p = \sqrt{p\,(1-p)}, \qquad (15a)$$

$$\sigma_p = \sqrt{p\,q}; \qquad (15б)$$

при биномиальном распределении (в значениях вероятностей) $$\sigma = \sqrt{кр\,q}; \qquad (17)$$

линии регрессии

$$\sigma_{y \cdot x} = \sqrt{\frac{\Sigma d^2_{y\,x}}{n-2}}, \qquad (52)$$

где $$\Sigma d^2_{y,x} = \Sigma (y - \bar{y})^2 - \frac{[\Sigma (x - \bar{x})\ (y - \bar{y})]^2}{\Sigma (x - \bar{x})^2}. \qquad (53)$$

Средняя ошибка [средняя квадратическая ошибка; стандартная ошибка; s; m]

средней арифметической [$s_{\bar{x}}$; $m_{\bar{x}}$]

$$s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}; \qquad (19)$$

среднего квадратического отклонения [s_σ; m_σ]

$$s_\sigma = \frac{\sigma}{\sqrt{2n}}; \qquad (22)$$

коэффициента вариации $[s_{c\cdot v\cdot};\ m_{c\cdot v\cdot}]$

$$s_{c\cdot v\cdot} = \frac{c\,v.}{\sqrt{2n}}; \qquad (22a)$$

разницы между средними арифметическими $\bar{x}_1$ и $\bar{x}_2$ $[s_d;\ m_d;\ m_{diff}]$

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2}}; \qquad (24)$$

разницы между средними арифметическими $\bar{x}_1$ и $\bar{x}_2$ при наличии корреляции

$$s_d = \sqrt{s^2_{\bar{x}_1} + s^2_{\bar{x}_2} - 2\,s_{\bar{x}_1} \cdot s_{\bar{x}_2} \cdot r_{12}}; \qquad (45)$$

разницы между средними арифметическими $\bar{x}_1$ и $\bar{x}_2$ при малых n

$$s_d = \sqrt{s^2 \frac{n_1 + n_2}{n_1 \cdot n_2}}, \qquad (25)$$

где $$s^2 = \frac{\Sigma_1 (x_1 - \bar{x}_1)^2 + \Sigma_2 (x_2 - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} \qquad (25a)$$

разницы между средними квадратическими отклонениями $[s_{d\sigma};\ m_{d\sigma}]$

$$s_{\sigma_1 - \sigma_2} = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}; \qquad (27)$$

доли при альтернативной изменчивости

$$s_p = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{pq}{n}}; \qquad (29)$$

конкретных чисел при альтернативной изменчивости

$$s_{p_1} = \sqrt{\frac{p_1 (n - p_1)}{n}}; \qquad (30)$$

коэффициента корреляции при больших n $[s_r;\ m_r]$

$$s_r = \frac{1 - r^2}{\sqrt{n}}; \qquad (38)$$

коэффициента корреляции при малых n

$$s_r = \frac{\sqrt{1 - r^2}}{\sqrt{n-2}}; \tag{39}$$

для z — числа

$$s_z = \frac{1}{\sqrt{n-3}}; \tag{41}$$

коэффициента регрессии $[s_b;\ m_b]$

в отклонениях

$$s_{by \cdot x} = \frac{\sigma_{y \cdot x}}{\sqrt{\Sigma (x - \bar{x})^2}}, \tag{54}$$

$$s_{by \cdot x} = \sqrt{\frac{\Sigma (y - \bar{y})^2 - \frac{[\Sigma (x - \bar{x})\ (y - \bar{y})]^2}{\Sigma (x - \bar{x})^2}}{(n-2)\ \Sigma (x - \bar{x})^2}}; \tag{54a}$$

в сигмах и r

$$s_{by \cdot x} = \frac{\sigma_y}{\sigma_x} \sqrt{\frac{1 - r^2}{n-2}} \tag{56}$$

и

$$s_{bx \cdot y} = \frac{\sigma_x}{\sigma_y} \sqrt{\frac{1 - r^2}{n-2}}; \tag{56a}$$

разницы между коэффициентами регрессии при больших n

$$s_{d\,(b_1 - b_2)} = \sqrt{\frac{s_1^2}{\Sigma_1 (x_1 - \bar{x}_1)^2} + \frac{s_2^2}{\Sigma_2 (x - \bar{x}_2)^2}}, \tag{57}$$

при малых n $\quad s_{d\,(b_1 - b_2)} =$

$$= \sqrt{\frac{(n_1 - 2)\ s_1^2 + (n_2 - 2)\ s_2^2}{(n_1 - 2) + (n_2 - 2)} \left(\frac{1}{\Sigma_1 (x - \bar{x}_1)^2} + \frac{1}{\Sigma_2 (x_2 - \bar{x}_2)^2} \right)}. \tag{57a}$$

Степени свободы [$d.f.$ или f, иногда v]

В простейших случаях

$$d.f. = n - 1.$$

При сравнении эмпирических распределений с теоретическими

$$d.f. = \kappa - 2 \text{ (биномиальное)}$$
$$d.f. = \kappa - 3 \text{ (нормальное).}$$

В таблицах состава с r-рядами и c-столбцами

$$d.f. = (r-1)\ (c-1).$$

Сумма квадратов [дисперсия]:

Общая формула $\Sigma\,(x - \bar{x})^2$.

Рабочие формулы $\Sigma f\,(x - \bar{x})^2$.

$$\Sigma f\,(x - A)^2 - n\ (A - \bar{x})^2.$$
$$\Sigma\,x^2 - n\,\bar{x}^2$$
$$\Sigma\,x^2 - \Sigma\,x \cdot \bar{x}$$
$$\Sigma\,x^2 - \frac{(\Sigma\,x)^2}{n}.$$

Уравнение кривой нормального распределения

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

или при $\sigma = 1$ и введении величины t

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Уравнение регрессии:

В общем виде

$$y - \bar{y} = b\ (x - \bar{x}). \qquad (46)$$

Преобразованное

$$y = \bar{y} + b\ (x - \bar{x}). \qquad (46a)$$

В виде уравнения прямой

$$y = a + b\,x. \qquad (47)$$

Системы уравнений для его решения

$$1.\ na + b\,\Sigma x = \Sigma y.$$
$$2.\ a\,\Sigma x + b\,\Sigma x^2 = \Sigma xy. \qquad (48)$$

Общее для прямолинейной и криволинейной зависимости

$$y = a + bx + cx^2 + dx^3 + \ldots; \qquad (61)$$

для экспоненциальной кривой

$$W = AB^x; \qquad (62)$$

преобразованное в логарифмическую форму

$$\log W = \log A + (\log B)\,x. \qquad (62)$$

СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

I. Значения нормального интеграла вероятностей в пределах $\pm t$ (из книги Немчинова „Сельскохозяйственная статистика", с изменениями).

II. Площадь кривой вероятностей по Стюденту в пределах $\pm t$ для малого числа наблюдений (той же книги Немчинова).

III. Значения t при различных уровнях значимости P (из книги Фишера „Статистические методы для исследователей; с сокращениями).

IV. Значения F при уровне значимости 0,05 (из книги Bailey— — Statistical methods).

V Значения F при уровне значимости 0,01 (из той же книги Bailey).

VI. Необходимые значения коэффициента корреляции при различных уровнях значимости P и разных n (из книги Snedecor — Statistical methods; с изменениями).

VII. Значения r при разных z (из книги Weber—Grundriss der biologischen Statistik; с изменениями).

VIII. χ-квадрат распределение (из книги Snedecor—Statistical methods; с сокращениями).

IX. Частоты нормального распределения (из книги Snedecor— Statistical methods).

Таблица I

Значения нормального интеграла вероятностей в пределах $\pm t$

t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0000	0080	0160	0239	0319	0399	0478	0558	0638	0717
0,1	0797	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6680	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9606	9616	9625	9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9840	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9909	9912	9915	9917	9920	9922	9924	9926	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9950	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	9973	9981	9986	9990	9993	9995	9997	9998	9999	9999

Таблица II

Площадь кривой вероятностей по Стюденту в пределах $\pm t$ для малого числа наблюдений n

t \ n	2	3	4	5	6	7	8	9	10	12	14	16	18	20	∞
0,1	063	071	073	075	076	076	077	077	077	078	078	078	078	079	080
0,2	126	140	146	149	151	152	153	154	154	155	155	156	156	156	158
0,3	186	208	216	221	224	226	227	228	229	230	231	232	232	233	236
0,4	242	272	284	290	294	297	299	300	302	303	304	305	306	306	311
0,5	295	333	347	357	362	365	368	369	371	373	375	376	377	377	383
0,6	344	391	409	419	425	430	433	435	437	439	441	433	444	444	452
0,7	389	444	466	477	485	490	493	496	498	502	504	505	507	508	516
0,8	430	492	518	531	540	546	550	553	556	558	562	564	565	566	576
0,9	467	537	537	581	591	597	602	606	608	613	616	618	619	621	632
1,0	500	577	609	626	637	644	649	653	657	661	664	667	669	670	683
1,1	530	614	648	667	679	687	692	697	700	705	709	711	713	715	729
1,2	558	647	684	704	716	725	731	736	739	745	748	751	753	755	770
1,3	583	677	716	737	750	759	765	770	774	780	784	788	789	791	806
1,4	605	704	744	766	780	789	796	801	805	811	815	818	821	822	838
1,5	626	728	769	792	806	816	823	828	832	838	842	846	848	850	866
1,6	644	749	792	815	830	839	846	852	856	862	866	870	872	874	890
1,7	661	769	812	836	850	860	867	872	877	883	887	890	893	895	911
1,8	677	786	830	854	868	878	885	890	895	901	905	908	910	912	928
1,9	692	802	846	870	884	894	901	906	910	916	920	923	925	927	943
2,0	705	816	861	884	898	908	914	919	923	929	933	936	938	940	954
2,1	717	829	873	896	910	920	926	931	935	940	944	947	949	951	964
2,2	728	841	885	907	921	930	936	941	945	950	954	956	958	960	972
2,3	739	852	895	917	930	939	945	950	953	958	961	964	966	967	979
2,4	749	862	904	926	938	947	953	957	960	965	968	970	972	973	984
2,5	758	870	912	933	946	953	959	963	966	970	973	975	977	978	988
2,6	766	878	920	940	952	959	965	968	971	975	978	980	981	982	991
2,7	774	886	926	946	957	964	969	973	976	979	982	984	985	986	993
2,8	782	893	932	951	962	969	973	977	979	983	285	987	988	989	995
2,9	789	899	937	956	966	973	977	980	982	986	988	989	990	991	996
3,0	795	905	942	960	970	976	980	983	985	988	990	991	992	993	997
3,1	801	910	947	964	973	979	983	985	987	990	992	993	994	994	998
3,2	807	915	951	967	976	981	985	987	989	992	993	994	995	995	999
3,3	813	919	954	970	979	984	987	989	991	993	994	995	996	996	999
3,4	818	923	958	973	981	986	989	991	992	994	995	996	997	997	999
3,5	823	927	961	975	983	987	990	992	993	995	996	997	997	998	1.
3,6	828	931	963	977	984	989	991	993	994	996	997	997	998	998	
3,7	832	934	966	979	986	990	992	994	995	996	997	998	998	998	
3,8	836	937	968	981	987	991	993	995	996	997	998	998	999	999	
3,9	840	940	970	982	989	992	994	995	996	998	998	999	999	999	
4,0	844	943	972	984	990	993	995	996	997	998	998	999	999	999	
4,1	848	945	974	985	991	994	995	997	997	998	999	999	999	999	
4,2	851	948	975	986	992	994	996	997	998	999	999	999	999	1.	
4,3	855	950	977	987	992	995	996	997	998	999	999	999	999		
4,4	858	952	978	988	993	995	997	998	998	999	999	999	1.		
4,5	861	954	980	989	994	996	997	998	999	999	999	1.			
4,6	864	956	981	990	994	996	998	998	999	999	1.				
4,7	867	958	982	991	995	997	998	998	999	999					
4,8	869	959	983	991	995	997	998	999	999	999					
4,9	872	961	984	992	996	997	998	999	999	1.					

Таблица III

Значения t при разных уровнях значимости (P)

$d.f.$	P			
	0,1	0,05	0,02	0,01
1	6,314	12,706	31,821	63,657
2	2,920	4,303	6,965	9,925
3	2,353	3,182	4,541	5,841
4	2,132	2,776	3,747	4,604
5	2,015	2,571	3,365	4,032
6	1,943	2,447	3,143	3,707
7	1,895	2,365	2,998	3,499
8	1,860	2,306	2,896	3,355
9	1,833	2,262	2,821	3,250
10	1,812	2,228	2,764	3,169
11	1,796	2,201	2,718	3,106
12	1,782	2,179	2,681	3,055
13	1,771	2,160	2,650	3,012
14	1,761	2,145	2,624	2,977
15	1,753	2,131	2,602	2,947
16	1,746	2,120	2,583	2,921
17	1,740	2,110	2,567	2,898
18	1,734	2,101	2,552	2,878
19	1,729	2,093	2,539	2,861
20	1,725	2,086	2,528	2,845
21	1,721	2,080	2,518	2,831
22	1,717	2,074	2,508	2,819
23	1,714	2,069	2,500	2,807
24	1,714	2,064	2,492	2,797
25	1,708	2,060	2,485	2,787
26	1,706	2,056	2,479	2,779
27	1,703	2,052	2,473	2,771
28	1,701	2,048	2,467	2,763
29	1,699	2,045	2,462	2,756
30	1,697	2,042	2,457	2,750
∞	1,645	1,960	2,326	2,576

Таблица IV

Значения F при уровне значимости 0,05 ($d.f._1$ — число степеней свободы для большей вариансы, которая берется числителем)

$d.f._2$ \ $d.f._1$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	250	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,46	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,62	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,94	5,91	5,86	5,80	5,75	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,50	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,81	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,38	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,08	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,86	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,70	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,57	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,47	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,38	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,31	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,25	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,19	2,01

17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,15	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,11	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,07	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,04	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,01	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	1,98	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	1,96	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,94	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,92	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,90	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,88	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,87	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,85	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,84	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,74	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,65	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,55	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,46	1,00

Таблица V

Значения F при уровне значимости 0,01 (df_1 — число степеней свободы для большей вариансы, которая берется числителем)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6261	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,47	99,50
3	34,12	30,82	29,46	28,71	28,42	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,50	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,84	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,38	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,23	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	5,99	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,20	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,65	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,25	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	3,94	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,70	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,51	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,35	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,21	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,10	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,00	2,65

18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	2,92	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,61	3,52	3,43	3,30	3,15	3,00	2,84	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,78	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,72	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,67	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,62	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,58	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,54	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,50	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,47	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,44	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,41	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,39	2,01
40	7,31	5,18	4,31	3,85	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,20	1,80
60	7,08	4,98	4,13	3,63	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,03	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,86	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,70	1,00

Таблица VI

Необходимые значения коэффициента корреляции r при различных уровнях значимости P и разных n

$d.f.$	P		$d.f.$	P	
	0,05	0,01		0,05	0,01
1	0,997	1,000	24	0,388	0,496
2	0,950	0,990	25	0,381	0,487
3	0,878	0,959	26	0,374	0,478
4	0,811	0,917	27	0,367	0,470
5	0,754	0,874	28	0,361	0,463
6	0,707	0,834	29	0,355	0,456
7	0,666	0,798	30	0,349	0,449
8	0,632	0,765	35	0,325	0,418
9	0,602	0,735	40	0,304	0,393
10	0,576	0,708	45	0,288	0,372
11	0,553	0,684	50	0,273	0,354
12	0,532	0,661	60	0,250	0,325
13	0,514	0,641	70	0,232	0,302
14	0,497	0,623	80	0,217	0,283
15	0,482	0,606	90	0,205	0,267
16	0,468	0,590	100	0,195	0,254
17	0,456	0,575	125	0,174	0,228
18	0,444	0,561	150	0,159	0,208
19	0,433	0,549	200	0,138	0,181
20	0,423	0,537	300	0,113	0,148
21	0,413	0,526	400	0,098	0,128
22	0,404	0,515	500	0,088	0,115
23	0,396	0,505	1000	0,062	0,081

Таблица VII

Значения r при разных величинах z (от 0 до 2,99). Для сокращения места ноль перед коэффициентом корреляции пропущен, поэтому 0997 надо читать как 0,0997

z	Сотые доли z									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0000	0100	0200	0300	0400	0500	0599	0699	0798	0898
0,1	0997	1096	1194	1293	1391	1489	1586	1684	1781	1877
0,2	1974	2070	2165	2260	2355	2449	2543	2636	2729	2821
0,3	2913	3004	3095	3185	3275	3364	3452	3540	3627	3714
0,4	3800	3885	3969	4053	4136	4219	4301	4382	4462	4542
0,5	4621	4699	4777	4854	4930	5005	5080	5154	5227	5299
0,6	5370	5441	5511	5580	5649	5717	5784	5850	5915	5980
0,7	6044	6107	6169	6231	6291	6351	6411	6469	6527	6584
0,8	6640	6696	6751	6805	6858	6911	6963	7014	7064	7114
0,9	7163	7211	7259	7306	7352	7398	7443	7487	7531	7574
1,0	7616	7658	7699	7739	7779	7818	7857	7895	7932	7969
1,1	8005	8041	8076	8110	8144	8178	8210	8243	8275	8306
1,2	8337	8367	8397	8426	8455	8483	8511	8538	8565	8591
1,3	8617	8643	8668	8692	8717	8741	8764	8787	8810	8832
1,4	8854	8875	8896	8917	8937	8957	8977	8996	9015	9033
1,5	9051	9069	9087	9104	9121	9138	9154	9170	9186	9201
1,6	9217	9232	9246	9261	9275	9289	9302	9316	9329	9341
1,7	9354	9366	9379	9391	9402	9414	9425	9436	9447	9458
1,8	9468	9478	9488	9498	9508	9517	9527	9536	9545	9554
1,9	9562	9571	9579	9587	9595	9603	9611	9618	9626	9633
2,0	9640	9647	9654	9661	9667	9674	9680	9686	9693	9699
2,1	9704	9710	9716	9721	9727	9732	9737	9743	9748	9753
2,2	9757	9762	9767	9771	9776	9780	9785	9789	9793	9797
2,3	9801	9805	9809	9812	9816	9820	9823	9827	9830	9833
2,4	9837	9840	9843	9846	9849	9852	9855	9858	9861	9863
2,5	9866	9869	9871	9874	9876	9879	9881	9883	9886	9888
2,6	9890	9892	9894	9897	9899	9901	9903	9904	9906	9908
2,7	9910	9912	9914	9915	9917	9919	9920	9922	9923	9925
2,8	9926	9928	9929	9931	9932	9933	9935	9936	9937	9938
2,9	9940	9941	9942	9943	9944	9945	9946	9947	9948	9949

χ^2—распределение

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

Таблица IX

Частоты (накопленные) нормального распределения, $n=10000$.

σ	Сотые доли σ									
	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	1551	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	4656	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	4936
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									

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Рокицкий
Петр Фомич

Основы вариационной статистики для биологов

Издательство
Белгосуниверситета имени В. И. Ленина
Минск—1961

Редактор *Найдович А. Н.*
Художественный редактор *Лысенко П. П.*
Техредактор *Беленькая И. Е.*
Корректоры *Сушко К. В., Сме . И. В.*

Печатается по постановлению РИСо БГУ им. В. И. Ленина

АТ 04716. Сдано в набор 12.X 1960 г. Подписано к печати 14.III-61 г.
Тираж 8000. экз. Бумага 84×108$^{1}/_{32}$. Печ. л. 7. Усл.-печ. л. 11,48.
Уч.-изд. л. 16,5. Заказ 154. Изд. зак. 58. Цена 65 коп.

Типография Издательства Белгосуниверситета имени В. И. Ленина,
г. Минск, ул. Кирова, 24.

Цена 65 коп.

Цена 65 коп.