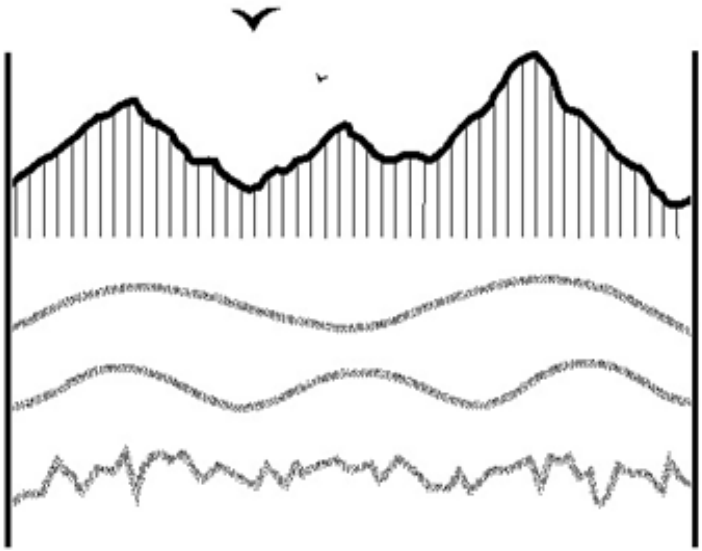
А. В. КОРОСОВ

Специальные методы биометрии



Петрозаводск 2007

А. В. Коросов

Специальные методы биометрии

Учебное пособие

Петрозаводск
Издательство ПетрГУ
2007

ББК 28.08:22.172
К686
УДК 578.087.1

Рецензенты:

член-корр. РАН, доктор биологических наук *Э. В. Ивантер;*
доцент, кандидат физико-математических наук *В. Б. Ефлов*

Коросов А. В.
К686 Специальные методы биометрии: Учеб. пособие / А. В. Коросов. — Петрозаводск: Изд-во ПетрГУ, 2007. — 363 с.
ISBN 978-5-8021-0615-0

ISBN 978-5-8021-0615-0

Книга рассчитана на читателей, знакомых с основами и практикой статистического анализа данных, и предназначается, в первую очередь, студентам старших курсов эколого-биологических специальностей. Она будет полезна специалистам разного профиля в качестве введения в практику разнообразных сложных методов обработки числовых данных, шкалирования и моделирования.

Работа выполнена при поддержке РФФИ (грант 05-04-97506-р_север_а)

ББК 28.08:22.172
УДК 578.087.1

ISBN 978-5-8021-0615-0

Введение

Термин «биометрия» традиционно ассоциируется с применением алгоритмов вариационной статистики для решения эколого-биологических задач. При этом из поля зрения выпадает широкий спектр собственно методов измерения живого (bios – жизнь, metron – мера) и в целом опыт метрологии, науки об измерениях. От этого страдает качество биологических исследований, использующих математические методы. Дело в том, что метрология теоретически обосновывает и показывает пути к тому, чтобы между значениями биологических показателей (полученных при тех или иных измерениях) и состоянием измеряемого природного объекта устанавливалось соответствие, понятное исследователю, чтобы *число* было адекватно *реальности*. Метрология предлагает приемы для корректного формирования измерительных шкал, в единицах которых оценивается состояние природы. Существуют разнообразные способы, улучшающие свойства измерений. В их число входит создание неких расчетных признаков (индексов), в обобщенной форме выражающих основные свойства объектов исследования на базе ряда исходных характеристик, замеренных в природе. Помимо этого, метрология ставит проблему создания собственно биологических шкал. В настоящее время у биологических объектов определяют физико-химические свойства (штука, масса, размеры, температура, плотность, концентрация и пр.). Даже такие явно биологические показатели, как толерантность и адаптивная ценность, имеют химические и физические единицы измерения (доза фактора и доля выживших особей). Введению в метрологию посвящена глава 2.

Лишь те из биологических и близких к ней дисциплин, которые изучают не физически заданный объект, а «эфемерные субстанции» (чувства, мышление, память, знания), оказались лидерами введения нефизических шкал – это психология, социология, педагогика и др. В их недрах оформились методы построения специфических численных характеристик, имеющих биологический смысл; они успешно работают в своих областях. В пособии сделана попытка популяризовать некоторые из этих приемов (многомерное шкалирование, глава 8), которые могут применяться в эколого-биологических исследованиях.

Есть в биометрии области, где формированию парабиологических шкал постоянно уделяется большое внимание, – это методы изучения биологического разнообразия, вычисление разнообразных «расстояний» между объектами. Описание видового состава (видового богатство) ценозов, выравненность, их сходство и отличие от сообществ на других территориях представляют большой интерес для науки и общества. В последнее время тема биоразнообразия распространилась и на внутривидовую изменчивость, оказались востребованными методы описания и сравнения фенотипической (фенетической) структуры популяций. В то же время приемов корректной и статистически выверенной обработки подобных материалов не так и много. Этот момент учтен при рассмотрении алгоритмов исследования биоразнообразия (глава 5).

Разработке биологических шкал мешает, на наш взгляд, настороженное отношение биологов к качественным признакам и балльным оценкам, как к менее точным, нежели собственно числовые показатели. Однако современные методы приведения номинальных и порядковых шкал к шкалам отношения (глава 2) во многом снимают эту проблему. Кроме того, к балльным оценкам применимы алгоритмы непараметрической статистики, которыми зачастую пренебрегают, несмотря на их эффективность (глава 4).

Большое внимание мы уделили методам изучения последовательностей – временным и пространственным рядам данных. На наш взгляд, подача этой очень перспективной темы чрезмерно переусложнена в немногочисленных источниках. В то же время подобные алгоритмы имеют особую ценность в связи с развитием ГИС-технологий, открывающих возможность исследования космоснимков и аэрофотоснимков (глава 10).

Важно отметить, что настоящее издание продолжает серию биометрических пособий, начатую Э. В. Ивантером (см. библиографию). Мы пытались постоянно следовать их главному принципу – обеспечивать предметность изложения. Именно поэтому каждая из статистических процедур, отраженная в пособии, иллюстрируется конкретным примером решения той или иной эколого-биологический задачи. Ввиду существенной сложности некоторых методов (особенно многомерных), разделы обычно начинаются с небольшого логико-теоретического введения. Текст доступен читателю, знакомому с основами математической статистики.

Глава 1
СБОР ДАННЫХ

Любое биометрическое исследование начинается с формирования выборки – множества значений случайной величины, оценок изучаемого свойства. В этом процессе участвует несколько агентов, которые и порождают собственно *изменчивость*, отличие отдельных вариант (значений, чисел) друг от друга. Изучая случайную и сопряженную изменчивость их свойств, можно выяснить природу факторов, ее порождающих. Для понимания структурно-логической сущности выборки и числа требуется рассмотреть *объект, признак, фактор, метод*.

1.1. Фрейм значения

Единицей выборки служит отдельная варианта, единичное значение, число (рис. 1.1.1). *Значение* есть качественное или количественное выражение *свойства* некоего *объекта*, полученного при данном уровне *фактора* внешней среды вполне определенным *методом*. С помощью этого фрейма покажем основные направления *тиражирования* значений, то есть составления выборок вариант.

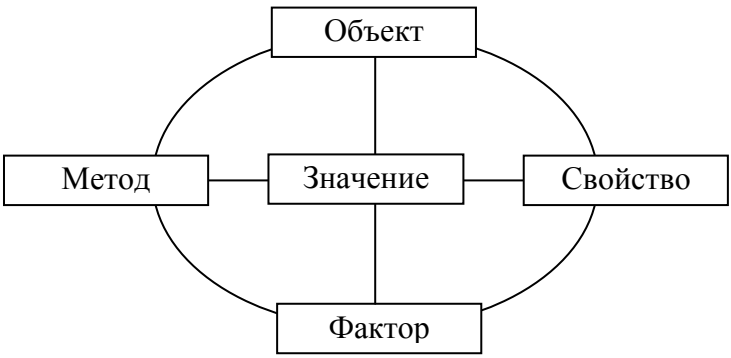


Рис. 1.1.1. Элементарный фрейм значения

Объект

Объект исследования – это биологический феномен – предмет (организм, популяция, экосистема) или процесс (размножение, динамика численности, сукцессия), на который направлено внимание исследователя. Объект измерения – это конкретный представи-

тель объекта исследования, свойства которого оцениваются непосредственно с помощью органов чувств или измерительного инструмента. В результате наблюдения появляется *варианта* – некий носитель частных характеристик объекта исследования. Варианта как отображение объекта измерения может «нести» одно значение (объект охарактеризован одним признаком) или несколько значений (оценены несколько качеств). Расширяя спектр зарегистрированных свойств, мы получаем возможность усложнить методы статистической обработки и от одномерных методов (описательная статистика) переходить к поиску зависимостей между двумя характеристиками (дисперсионный, регрессионный, корреляционный анализ) и многомерному анализу (кластерный, дискриминантный, компонентный анализы).

Важнейший прием формирования выборок – это отбор и измерение более или менее однородных представителей объекта исследования. Отличие между такими вариантами имеет внутренний, эндогенный, источник – индивидуальные отличия *по статусу* и по *состоянию*. Например, животные одного возраста индивидуально различны по полу, фенотипу, генотипу; множество их однотипных характеристик образует выборку организмов, отличающихся по статусу. В то же время каждая особь в разные годы, сезоны, время суток имеет разные морфофизиологические характеристики; множество подобных замеров составит рад вариант, характеризующих отличия по состоянию. Еще один путь получения выборок состоит в наблюдении многообразия реакций объекта исследования на разные внешние условия существования (см. раздел **Фактор**).

Зная природу объекта, можно правильно выбрать метод обработки данных. Статистические алгоритмы ориентированы на изучение случайных величин разного типа. Желательно, чтобы они подчинялись нормальному закону распределения (непрерывные признаки) или биномиальному закону (дискретные признаки) (подробнее законы распределения рассмотрены в главе 3). Зачастую отклонение поведения случайных величин от этих законов связано с непродуманным способом получения выборок, с методическими ошибками и неточностями измерений, хотя для описания многих биологических свойств нормальный закон не подходит. В общем случае для приведения распределения к более «чистому» виду нужно пытаться выявить, учесть, изучить «вредные» факторы, влияю-

щие на изменчивость вариант, и самые сильные из них ликвидировать. Эту унификацию можно проводить путем разделения одной исходной выборки на несколько более однородных выборок (с вариантами отчетливо разного статуса). Так, в популяционной морфологии считается необходимым разделение животных по видам. Но не менее важно отдельно характеризовать самок и самцов, разновозрастных животных и даже представителей разных генераций. Ликвидируя сильные причины варьирования, мы формируем выборки, лучше соответствующие объекту исследования. Теоретически мыслима ситуация, когда исключены все факторы варьирования и многократные повторные измерения объекта исследования дают одно и то же единственное значение. Если в физике такая ситуация возможна («чистый эксперимент»), то в биологии с ее необозримым числом внешних и внутренних факторов практически не удается получить абсолютную повторяемость значений. Обычно множество малосущественных (случайных) причин изменчивости обеспечивает нормальное распределение признака.

Фактор

Фактор – это условия проведения наблюдений, среда существования объекта, возможная причина, определяющая текущее состояние объекта. В кибернетической модели «влияние – зависимость» фактор есть активное начало, вход (переменная x), действующий на признак, выход (y) (рис. 1.1.2):

$$y = f(x).$$

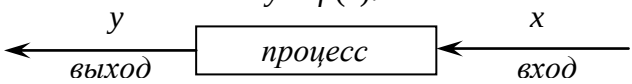


Рис. 1.1.2. Блок-схема зависимости

С методической точки зрения различают факторы контролируемые и неконтролируемые. В первом случае степень проявления фактора точно устанавливается, получение выборок организуется при разных заданных дозах фактора. Таковы условия проведения лабораторных экспериментов, когда имеется возможность сразу получать выборки, не загрязненные эффектами действия посторонних агентов. Во втором случае (натурные наблюдения) факторы неподвластны исследователю. Некоторые из них удается регистрировать,

другие – нет. Эта ситуация наиболее обычна для экологии, и важно понимать, как здесь может помочь статистика. Существуют биометрические методы, которые (при достаточно большом числе наблюдений объекта в разных условиях) позволяют из общей изменчивости объектов исследования выделять эффекты воздействия разных факторов по отдельности. Современный путь биометрии – измерение большого количества свойств объектов и уровней факторов среды с последующим изучением зависимостей между ними методами многомерной статистики, дисперсионного и регрессионного анализов. Отсюда следует общая рекомендация при составлении выборки – учитывать по возможности все условия ее получения.

По существу, цель любого биометрического исследования состоит в том, чтобы доказать достоверность действия какого-либо фактора, определить, влияет ли изменение дозы (силы) данного агента на изменение значений данного признака. Сравнение двух выборок уже есть задача сравнения двух доз некоего фактора, представленных, соответственно, двумя группами вариант. Если несколько выборок вариант выражают несколько доз, или градаций, фактора, то можно оценить и интенсивность этого влияния (дисперсионный анализ), а также его характер (регрессионный анализ).

Вопрос о влиянии может быть поставлен не только в отношении контролируемого (регистрируемого) фактора, но также для таких (существенных) факторов, которые сказались на изменчивости значений сразу нескольких признаков, хотя и *не измерялись непосредственно*. Метод главных компонент, многомерное шкалирование, гармонический и спектральный анализ позволяют извлекать из многомерных данных и временных рядов некие обобщенные характеристики их изменчивости. Расчетные «факторы» и «гармоники» играют роль пока не наблюдаемых гипотетических характеристик внешней или внутренней среды объекта, поиск которых в природе тем самым обретает бóльшую определенность.

Другая группа неучтенных (незарегистрированных) факторов – случайные. Эти факторы действуют, вызывая ненаправленное варьирование. Общей задачей биометрического исследования остается отделение доминирующих факторов от случайных, точнее – оценка статистической значимости любого фактора изменчивости и их классификация на доминирующие (существенные) и случайные.

Свойство

Свойство (качество, признак, показатель, величина, характеристика, переменная) – это любая информация о наблюдаемом объекте, выраженная качественно или количественно определенная. В рамках вариационной статистики любые признаки выступают в роли случайной величины. *Случайная величина – численная характеристика, принимающая те или иные заранее точно не известные значения.* Точный прогноз случайной величины получить нельзя, но математическая статистика способна дать *вероятностное* описание, когда за множеством частных случаев просматривается их единство, и с помощью специальных расчетов позволяет получить *интервальные* предсказания (диапазон вероятностного ожидания случайной величины). Максимально эффективно это можно сделать, если не упускать из вида требования к выбору (конструированию) признака. «Список» потенциальных свойств любого объекта бесконечен, поэтому выбор того или иного признака должен хорошо соответствовать цели исследования (см. подробнее главу 2).

Приступая к составлению выборки, регистрируемые признаки следует соотнести с теми статистическими методами, которые планируются для их обработки. Грубые методы, ориентированные на чувственные оценки, позволяют получить только приблизительные значения (качественные, балльные показатели); точные инструментальные методы дают числа. Количественные признаки доступны для обработки точными параметрическими методами, тогда как балльные оценки можно статистически исследовать только с помощью менее точных непараметрических методов.

Метод

Процедура получения чисел (вариант) включает субъект, методику, инструмент и процесс их измерения. Использование разных методов измерения одного и того же объекта порождает выборку различающихся вариант. Отличия повторных промеров в какой-то степени характеризуют разнокачественность применяемых методик, инструментов или уровня подготовки исполнителей. Разные методы обладают разной способностью сообщать вариантам случайные ошибки (неточность оценок) и систематические ошибки (смещение оценок). По этой причине те выборки, варианты которых получены разными методами, обладают заведомо большей изменчивостью,

чем методически однородные выборки. Рекомендация очевидна – для формирования сравнимых выборок следует использовать единую методику, одинаковый откалиброванный инструмент, «одни руки».

Различают точность инструмента измерения и точность метода измерения. В первом случае говорят о технической характеристике прибора. Под точностью метода подразумевают точность (погрешность) измерительной процедуры, т. е. возможность воспроизведения тех же результатов при повторном измерении одного и того же объекта. Помимо характеристики прибора здесь фигурируют навыки исследователя, точность инструкции, особенности условий проведения измерений (влажность, радиация и др.). Точность инструмента и трудоемкость измерения должна быть явно соотнесена с погрешностью самой процедуры измерения. Для биологической практики обычна ситуация, когда погрешность метода выше, чем погрешность инструмента (как правило, предназначенного для замера физических или химических величин). Нет смысла проводить измерения очень точными приборами, если сама процедура измерения предполагает широкое варьирование. В частности, измерять длину тела живой взрослой гадюки довольно сложно, погрешность таких замеров составляет около 2 см. В таком случае удобнее (проще и быстрее) воспользоваться рейкой с сантиметровой шкалой, нежели делать промеры линейкой с ценой деления 1 мм. Для более точной оценки выполняются измерения «усыпленных» животных. Выбирая метод регистрации вариант, следует предварительно оценить его погрешность (причем разными исполнителями).

Погрешность

Отличие отдельного измерения (x_i) от истинных (X) значений величины называется *погрешностью*: $\Delta = x_i - X$. Качество множества (n) измерений одного и того же объекта оценивается с помощью *средней квадратической погрешности*: $S = \sqrt{\frac{\sum (x_i - X)^2}{n-1}}$.

Поскольку истинное значение признака практически никогда неизвестно, вместо него используют среднюю арифметическую величину $M = \Sigma x_i / n$, рассчитанную по n повторных измерений одного

и того же объекта. Тогда $S = \sqrt{\frac{\sum (x_i - M)^2}{n-1}}$. Зачастую вместо абсолютной величины погрешности удобнее пользоваться относительной погрешностью (коэффициентом вариации): $CV = \frac{S}{M} \cdot 100\%$.

Определение точности наблюдений

Для случаев, когда измеряется множество разных объектов одного вида, судить о точности полученных результатов (о возможности по части характеризовать целое) позволяет *статистическая ошибка* средней арифметической: $m = \frac{S}{\sqrt{n}}$. Относительная ошибка измерений вычисляется как отношение ошибки к средней арифметической изучаемого показателя: $\varepsilon = \frac{m}{M} \cdot 100\%$. Чем меньше значение показателя ε, тем более репрезентативна средняя арифметическая. При $\varepsilon < 3\%$ точность считается хорошей, при $3\% < \varepsilon < 5\%$ – удовлетворительной. При больших значениях наблюдения следует уточнить (повторить опыт, собрать дополнительный материал). Например, показатель точности измерения массы тела бурозубок ($M = 9.3 \pm 0.11$ г) составил $\varepsilon = (0.11/ 9.3) \cdot 100 = 1.2\%$, что говорит о достаточной надежности выборочной оценки.

Оптимальный объем выборки

При планировании биологических наблюдений объемы репрезентативных выборок рассчитывают заранее, ориентируясь на известные свойства изучаемых показателей и требования к материалу. Теоретической основой служит известное соотношение между изменчивостью и ошибкой репрезентативности: $m = \frac{S}{\sqrt{n}}$, откуда $n = \left(\frac{S}{m}\right)^2$. Однако требования к выборке проще предъявлять в относительных понятия, поэтому используя известные соотношения между средней арифметической, стандартным отклонением, ошибкой

средней, точностью наблюдений (ε, %), коэффициентом вариации (CV, %) и распределением Стьюдента (t, табл. 6*П* для $\alpha = 0.05$), объем выборки (n) при заданном уровне значимости (α) вычисляют по формуле: $n = \left(\frac{t \cdot CV}{\varepsilon}\right)^2$. Например, необходимый объем условной выборки, обеспечивающий хорошую точность $\varepsilon = 3\%$, для уровня значимости $\alpha = 0.05$ ($t \approx 2$) и для коэффициента вариации $CV = 12\%$ (такова относительная изменчивость многих размерно-весовых признаков животных) равен: $n = \left(\frac{2 \cdot 12}{3}\right)^2 = 64$ экз. Для обеспечения доверительной вероятности 99% ($\alpha = 0.01$; $t = 2.62$) необходимый объем выборки составляет: $n = \left(\frac{2.62 \cdot 12}{3}\right)^2 \approx 110$ экз.

Если исследуется фенотипическое (видовое) разнообразие, может возникнуть задача определения минимального объема выборки, в которой будет присутствовать хотя бы один экземпляр с определенным фенотипом. С позиций теории вероятности задача ставится так: определить объем выборки, в которой с вероятностью P можно ожидать присутствие особи с признаком, частота которого в генеральной совокупности составляет π. Предлагается следующая формула (Животовский, 1991): $N = \frac{\ln(1-P)}{\ln(1-\pi)}$.

В первом приближении значение π можно определить приблизительно по имеющимся данным. Задаваемый уровень вероятности P довольно сильно влияет на величину необходимого объема выборки. Для большей надежности следует брать $P = 0.99$, но тогда возрастет объем работ; не столь высокие требования ($P = 0.95$) могут и не позволить найти искомый фенотип. При уровне вероятности $P = 0.95$ и предположительной частоте фенотипа в популяции $\pi = 0.05$ потребуется отловить $N = \frac{\ln(1-0.95)}{\ln(1-0.05)} = 58.4 \approx 59$ экз., чтобы обнаружить хотя бы одну особь с этим дискретным признаком. Для практического использования нетрудно рассчитать небольшую таблицу, содержащую оценки необходимых объемов выборок для разной вероятности отлова и различных частот фенотипов в популяции.

1.2. Методы отбора проб

Сбор данных ведет к формированию *статистической совокупности*, которая далее анализируется с помощью статистических методов. Полученные при наблюдении материалы практически всегда представляют собой лишь часть от той общей информации об объекте исследования, которая «заключена» в природе. О целом (популяции, совокупности) приходится судить по части (группе особей, вариант, значений). Приступая к сбору эмпирических материалов, важно хорошо представлять себе соотношение между целым и частью, между природным явлением и попавшими в руки исследователя его фрагментарными описаниями.

Выражаясь на языке статистики, по выборке из генеральной совокупности приходится судить обо всей генеральной совокупности. *Генеральная совокупность* – все варианты одного типа. Это математическое понятие относится к группам значений, имеющих бесконечный объем. В предметной биологии его можно интерпретировать как *мыслимое* множество вариант, сформированных при одинаковых (внешних и внутренних) условиях. Так, генеральной совокупностью будет выступать чистая линия рачков-дафний, выращенных при температуре помещения 20.2 ºС. Может статься, что кроме группы из 100 особей в мире не существует других дафний, выращенных при таких условиях. Все равно в определении генеральной совокупности важно не реальное ее существование, но мыслимое однообразие условий, порождающих выборки, уверенность, что при воссоздании условий восстановятся и причины формирования выборок именно с такими свойствами. Поэтому эти 100 экз. будут выборкой из генеральной совокупности дафний, выращенных при 20.2 ºС, хотя в природе таких же особей может больше и не быть.

Важнейшее свойство генеральной совокупности состоит в том, что на всех ее вариантах (значениях) сказываются одни и те же систематические и случайные факторы, их набор уникален для данной генеральной совокупности. Для другой генеральной совокупности он будет другим, или же другой будет сила действия тех же факторов. Мысленно меняя условия, можно сформировать бесконечное множество бесконечных по объему генеральных совокупностей, отличающихся нюансами условий своего формирования. Только математике доступно исследование свойств бесконечного числа значе-

ний случайной величины; на основании открытых законов их поведения предложено множество моделей для описания и сравнения случайных величин, наблюдаемых в действительности (п. 3.2).

Вследствие бесконечности объема генеральной совокупности ее нельзя познать до конца, в действительности мы всегда имеем дело с выборками. *Выборка* из генеральной совокупности – это множество вариант одного типа, ограниченное способом отбора (методами получения вариант), это часть целого. Отличие выборок от генеральной совокупности состоит в разных объемах, в реальности первых и умозрительности вторых, а также в том, что в отдельной выборке *в полной мере* не могут проявиться все факторы, действующие в генеральной совокупности. Если систематический фактор действует на каждую варианту строго одинаковым образом, то случайные факторы сказываются на значениях вариант по-разному: на одну варианту сильно («большая прибавка значения»), на другую – слабо («малая прибавка»), на одну сильно повлияет много случайных факторов, на другую – мало. В результате такого влияния все варианты, оставаясь в целом единообразными (влияние доминирующих причин), все же будут отличаться друг от друга (влияние случайных причин). Нельзя составить набор из вариант, на которые воздействовали бы все комбинации случайных факторов. И нельзя составить две выборки, в которых эти комбинации случайных факторов были бы в точности одинаковыми. Разные выборки вариант из одной и той же генеральной совокупности всегда будут отличаться друг от друга. Вследствие этого любые обобщенные характеристики частных выборок всегда будут численно отличаться от параметров других выборок и от параметров породившей их генеральной совокупности – в этом есть *ошибки репрезентативности*.

Условие равной вероятности отдельных наблюдений

Несмотря на то, что выборка не может включить в себя все проявления случайных факторов, наблюдения следует организовать так, чтобы обеспечить *равную вероятность обнаружения* разных вариант (значений). Время, место, условия и орудия наблюдений (отлова, сбора) материала должны быть так подобраны, чтобы соблюдались равные пропорции в исследовании пространственно-временного континуума в поисках предметов измерения. Конечно, всегда существует неравенство вероятности *появления* разных вари-

ант, как следствие биолого-экологических законов. Например, на юге Карелии белых грибов гораздо меньше, чем подберезовиков. Однако одинаковая вероятность *обнаружения* объектов как раз и является необходимым приемом для того, чтобы выборка отражала истинное представительство разных видов значений, помогала обнаружить закономерное неравенство их частот и понять биологические причины этого явления. По этой причине, изучая распространение грибов, миколог бродит всюду, а не только по сосновым борам. Иными словами, при сборе данных должны учитываться *временные и пространственные пропорции* природных явлений, естественный набор природных объектов должен быть пропорционально представлен в выборках. Иногда это требование неточно называют правилом случайного отбора проб.

Пропорциональный отбор проб в пространстве

Методы отбора проб на территории хорошо разработаны в геоботанике. Подмечено, что «однородных участков в природе не бывает» и что «делянка должна охватить всю возможную пестроту участка под опытом». Выделяют 6 основных методик сбора.

1. *Сплошное* (тотальное) исследование – сбор всех объектов, выборка равна генеральной совокупности (рис. 1.2.1, *А*). Это была бы лучшая из характеристик природных объектов, но ее невозможно реализовать. Приходится пользоваться выборочным методом.

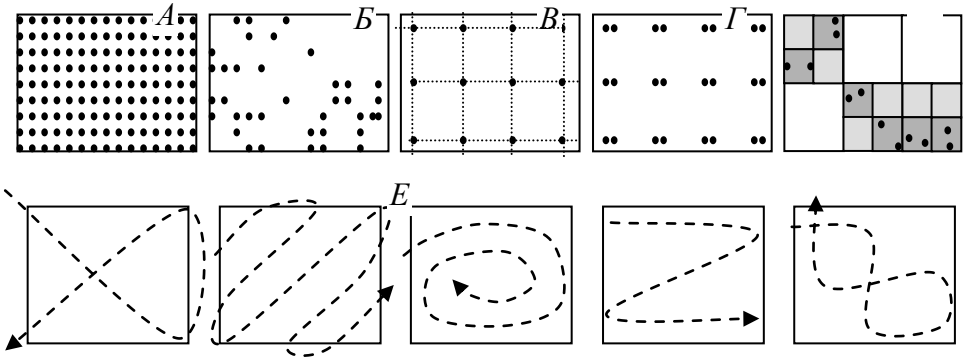


Рис. 1.2.1. Виды методов отбора проб (*А – Д*) и маршрутов (*Е*)

2. *Случайный* (рандомизированный) отбор (рис. 1.2.1, *Б*). Место отбора проб назначается по координатам, определенным с ис-

пользованием функции среды Excel =СЛЧИС() или с помощью таблицы случайных чисел. Вначале осям X и У территории придают значения от 0 до 1 и в качестве начала осей координат (0,0) назначается единственная точка на местности. Далее берут по два случайных числа и на карте ставят точку с этими координатами; процесс повторяют *n* раз. *Случайная прогулка* отличается тем, что за начало осей координат (0, 0) каждый раз принимается последняя из назначенных точек; процесс отбора проб – суть блуждание по местности. Достоинство метода – возможность быстрого формирования выборки с равными шансами обнаружить разные варианты, если территория однородна. Недостаток состоит в том, что территории обычно не однородны и случайный отбор не может обеспечить пропорционального представительства разных выделов территории, а то и вовсе не отображает их.

3. *Шахматный* (равномерный, регулярный, систематический, механический отбор) (рис. 1.2.1, *В*). На территорию накладывается регулярная квадратная сетка, пробы берутся в местах пересечения линий. Это лучший метод отбора проб, но, как и сплошной отбор, слишком громоздкий. Однако в природных условиях трудно отбирать пробы точно по квадратной сетке, а отклонения от методики отнимают преимущества метода. Увеличение шага сетки ведет к тому, что мелкие контуры могут быть представлены слишком малым числом вариант.

4. *Парно-шахматный* метод. Этот метод решает проблему снижения ошибки измерения шахматного метода при равных затратах труда (рис. 1.2.1, *Г*). Если для каждого назначенного места собирать пробы в двух ближайших точках, а затем усреднять, то полученная оценка будет лучше представлять качество среды в данной точке (или времени): 16 равноудаленных проб и 8 попарно связанных проб дадут одинаковую информацию о качестве среды, но во втором случае ошибка измерений будет меньше.

5. *Стратификационный* (расслоенный, многоступенчатый) метод – места отбора проб назначаются в два или несколько этапов (рис. 1.2.1, *Д*).

5.1. *Типический* метод состоит в том, что сначала территория делится на множество выделов, имеющих естественные границы, а затем в пределах этих типичных районов места отбора проб назначаются шахматным или случайным методом. При этом необходимо

помнить, что оптимальное квантирование информации предполагает *примерное равенство* выделяемых типологических подразделений территории.

5.2. *Пропорциональный* (ограниченно рандомизированный) метод есть соединение шахматного со случайным. Территория разбивается на серию квадратных выделов, среди них случайным образом выбираются нескольких выделов, для которых алгоритм повторяется 2–5 раз. В заключении имеем случайный набор равномерно распределенных точек отбора проб.

5.3. *Гнездовой* (серийный) метод состоит в том, что сначала случайным образом выбирается серия опорных точек, вокруг которых затем назначаются несколько мест отбора проб.

6. *Маршрут* (трансекта, траверс) – методы сбора данных при экспедиционных обследованиях обширных территорий. Практика показывает, что на маршруте очень трудно реализовать случайный отбор, поэтому пользуются систематическим, отбирая пробы через равные расстояния или равные промежутки времени, или равномерно покрывая территорию траекторией маршрута. Предложено множество схем (рис. 1.2.1, *E*).

Пропорциональный отбор проб во времени

Общие проблемы порождают сходство приемов исследования пространственного распределения и временного следования природных явлений.

1. *Регулярные* наблюдения (через равные промежутки времени) рекомендуются при составлении временных рядов. Достоинство метода – репрезентативность данных и простота последующей обработки. Недостаток состоит в том, что в интервалах между моментами наблюдений могут происходить важные события, которые не будут зафиксированы. Если замеры выполнять чаще, метод станет слишком громоздким, трудновыполнимым, а для отрезков слабого изменения функции будут давать множество продублированных значений.

2. *Стратификационный* метод – помимо регулярных наблюдений наиболее плотно изучаются самые *информативные интервалы*, когда переменная быстро изменяет свои значения (рис. 1.2.2). Если же требуется получить ряда значений через равные временные шаги, то для интервалов с незначительным изменением функции

можно провести интерполяцию: построить уравнение динамики показателя по известным наблюдениям и затем вычислить его значения в нужные моменты времени.

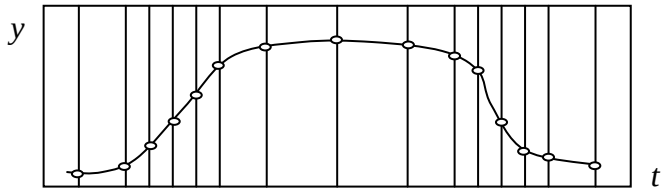


Рис. 1.2.2. Увеличения частоты замеров (*y*) в наиболее информативные интервалы времени (*t*)

3. Опыты с *повторностями*. Противоречие между точностью и стоимостью (временно́й, финансовой, ресурсной) наблюдений заставляет постоянно решать разные варианты общей задачи: как уменьшить ошибки измерений без роста издержек? Например, как лучше провести замеры функции – в четырех разных точках или в двух точках, но зато два раза? Ошибка измерений будет меньше во втором случае (рис. 1.2.3, А), значит, линия регрессии будет лучше обоснована. Возражения против этого метода, состоят в том, что через две точки можно провести и кривую линию (пунктир), а через 4 – только прямую. Но они не выдерживают критики, поскольку через 4 точки тоже можно провести кривые, хотя и другого вида. Видимо, лучше доказать наличие хотя бы линейной регрессии, чем при том же объеме материала остаться без доказательства криволинейной регрессии. Практика постановки дублированных опытов и наблюдений широко распространена в биохимии и токсикологии.

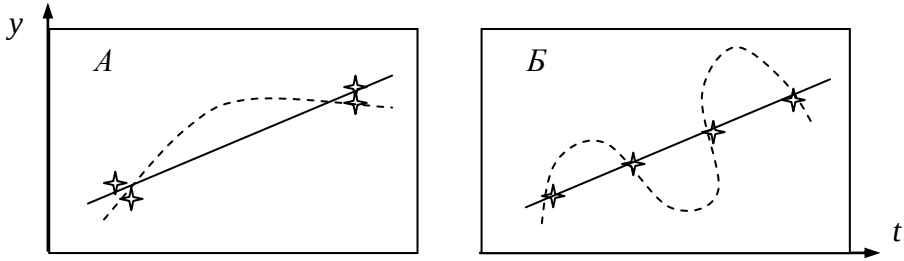


Рис. 1.2.3. Варианты оценки регрессии по четырем точкам

Глава 2
СВОЙСТВА И ШКАЛЫ

Наука начинается с тех пор, как начинают измерять.
Д. И. Менделеев

Математические методы применимы к биологии лишь в той мере, в какой она способна количественно охарактеризовать объекты своего исследования. Биометрия служит для обработки значений, полученных в результате измерения свойств биологических объектов. Количественные оценки (числа, подсчеты, замеры, промеры) свойств объектов позволяют достаточно строго высказываться о биологических явлениях и, главное, переводить их на математический язык – создавать модели, выражающие законы и закономерности их существования. Точный закон формулируется только как соотношение между величинами изучаемых свойств. Но что такое число, в чем его смысл, насколько полноценно оно отражает биологическую специфику, имеются ли ограничения при описании числом биологических проявлений, жизни вообще? Вот круг вопросов, на которые следует искать ответы еще до рассмотрения собственно методов биометрии.

2.1. Свойство, число, измерение

Ключевую роль в обсуждаемой проблеме играет понятие *свойство*. Свойства отражают процессы взаимодействия объектов между собой и их воздействия на наши органы чувств. Фактически существуют только *отношения* между (например, двумя) объектами. И если в поле зрения исследователя нет второго объекта («акцептора» свойств первого), это *отношение воспринимается как свойство*. «Небо синее» не само по себе, а потому, что зрительная система человека воспринимает излучение и специфически его интерпретирует. Конечно, оценка «синее» покоится на объективном (не зависящем от человека) отношении солнца со всеми объектами Земли: электромагнитные излучения с определенной длиной волны поступают «от неба» к поверхности Земли. Эта простая модель хорошо иллюстрирует любое биологическое наблюдение: есть (мини-

мум) два объекта («небо» и Земля), есть реальное отношение между ними (посредством излучения) и наблюдатель, который воспринимает небольшую часть энергии этого отношения и которое он воспринимает как *свойство* «неба». Эта модель позволяет дать практически важные определения. *Отношение* – сопричастное бытие вещей. *Свойство* – свернутое (или одноместное) отношение. *Наблюдение свойств* – восприятие человеком объективных отношений между объектами исследования.

Из предложенной модели наблюдения следует: любой объект природы обладает потенциально бесконечным числом свойств, поскольку он, в принципе, может вступить в бесчисленное множество отношений с другими объектами природы. Разные стороны (органы, качества) объектов участвуют во многих отношениях с внешним миром. Взять, к примеру, массу тела животного, стандартный зоологический показатель. Внешне это свойство выражает отношение тела животного только к весам. Объективный характер свойства «масса тела» состоит в отношении данного вида к доступным кормам, к условиям укрытия в данных местообитаниях, к опасным для него хищникам (в том числе к человеку), в массе тела «скрыты» отношения между брачными партнерами, между матерью и плодом (или зрелыми половыми продуктами) и т. п. Смысл измерения свойств живого и состоит в том, чтобы количественно (аналитически) выразить объективные отношения, царящие в природе.

При выполнении биологических исследований важно вспомнить принцип неопределенности Гейзенберга и определиться с тем, какую именно часть «энергии отношения» между объектами изымает данное наблюдение. Дискретность жизни часто заставляет действовать по схеме: «для изучения живого его нужно убить». В стремлении получить объемные, репрезентативные выборки (требование биометрии) следует иметь в виду тот невосполнимый урон (изъятие «кванта жизни»), который может нанести «измерение».

Число

Касаясь понятия числа, «легче всего дышится... когда решаешься вовсе оставить в стороне эти трудные вещи» (Клейн, 1987; с. 26). Числа – это *символы* (слова, понятия) для выражения количественных свойств объекта исследования. Углубляясь в тему, важно различать э*мпирические* и *логические* истоки термина.

Эмпирические корни количественных (счетных, числовых) понятий следует искать в быту древних людей. Не имея ни малейшего представления о теоретической арифметике, они вводили в обиход числовые оси и арифметические действия, справедливость которых доказывалась повседневным опытом в операциях с личными вещами, едой, при обмене и торговле. Например, примитивная система счета австралийского племени Камиларон (живущего по законам каменного века) основана на трех простых числах: мал (1), булан (2) и гулиба (3) и нескольких составных: булан-булан (4), булан-гулиба (5), гулиба-гулиба (6). Понятие о счете было тесно связано с представлением о расположении объектов в пространстве и о последовательности явлений во времени. Например, у первобытных племен Америки было около 307 систем счисления (с разными названиями чисел), основанными на количестве пальцев на руке (пятичная система), на двух руках (десятичная), на руках и ногах (двадцатичная), на смешанных основаниях. Наблюдения за числом лунных циклов в течение года породили двенадцатичную систему счисления. Видимо, попытка объединить пятичную (пространственную) и двенадцатичную (временну́ю) системы привела к созданию шестидесятичной системы, сохранившейся с шумерских времен по сей день для обозначения пространственных координат и времени.

Математическое (логическое) содержание числа сформировалось в античные времена. «Нечто радикально новое свершается тогда – и это есть рождение математики, – когда не просто принимают в качестве данных случайно встретившиеся в действительности числа, а располагают *в ряд* |, ||, |||, ... *все возможные числа*. Это осуществляется посредством *производящего процесса*, в котором постоянно повторяется одна и та же операция – переход от числа *n* к ближайшему числу *n*'. Как знаковая операция она выполняется посредством нового штриха. ...Мы уверены в самой *возможности* продолжать процесс дальше после каждого достигнутого пункта. *Действительное проецируется здесь на фон возможного – открытого в бесконечность многообразия, свободно сознаваемого умом при помощи надежно установленного способа*. ...Производство чисел в процессе постоянно повторяемого перехода от числа *n* к ближайшему числу *n*' находит себе методическое выражение *в определении и в умозаключении через совершенную индукцию*...» (Вейль, 1989; с. 61, 62). Полноты ради следует привести и формулировку

закона совершенной индукции: «если некоторое предположение справедливо для небольших чисел и если сверх того оно остается справедливым для числа n+1 всякий раз, как оно справедливо для числа n, то оно справедливо вообще для всякого числа» (Клейн, 1987; с. 27). Эти цитаты из произведений известных математиков мы привели для того, чтобы подчеркнуть логический (умственный) характер математического понятия числа, которое не нуждается в подтверждении практикой.

Введенное логическое понятие числа и ряда чисел позволяет логическим же образом исследовать их свойства (т. е. всевозможные отношения между числами) и выразить их в виде законов. Поскольку законы выполняются для всех без исключения чисел (включая неизвестные, неназванные), числа можно обозначать буквами. Приведем для примера пять законов сложения натуральных (целых положительных) чисел:

1. $a + b$ всегда представляют собой число, т. е. действие сложения выполнимо,
2. сумма $a + b$ всегда определена однозначно,
3. ассоциативный (сочетательный) закон: $(a + b) + c = a + (b + c)$,
4. коммутативный (переместительный) закон: $a + b = b + a$,
5. закон монотонности: если $b > c$, то $a + b > a + c$.

Исходная числовая ось натуральных чисел постепенно развивалась за счет пополнения новыми математическими понятиями. Введение таких несуществующих в природе чисел, как отрицательные, нуль, иррациональные ($\pi, \sqrt{2}$), мнимые ($\sqrt{-1}$), все больше увеличивало отрыв математического понятия числа от эмпирического. Все известное разнообразие чисел составляет современную *числовую ось – абсолютную шкалу*, которую в той или иной мере используют для создания других, частных числовых шкал.

Здесь не место рассматривать (серьезнейшие) философские проблемы соотношения реальности и математики. Но нам важно заметить, что объекты природы (реальность) и числа (математика) представляют собой два *разных* мира, между которыми нет простой *непосредственной* связи. В то же время какая-то глубинная («вселенская») связь между ними есть, поскольку законы математики ус-

пешно работают в приложении к материальным объектам. Но это «перекрывание» лишь частичное.

Природа «богаче» математики, которая не в силах описать ее во всех проявлениях: для любой математической модели (формулы) найдутся реальные ситуации, когда эта модель не справится с их описанием. Причина состоит в (необходимом) формально-логическом подходе к количественному описанию действительности, когда рассматривается одно или несколько свойств объекта, *важных с точки зрения исследователя*. Модель целесообразна и поэтому всегда включает в себя небольшое (ограниченное целью) число переменных и параметров. Природа же бесконечна в выборе возможных партнеров по взаимодействиям и зачастую демонстрирует «*контринтуитивное*» поведение. Например, если смешать 1 л воды и 1 л спирта, то в силу специфического взаимного упорядочивания молекул жидкостей получится не 2, а около 1.9 л смеси, сумма из двух (разнополых) кроликов через 2 месяца будет равна восьми. В этих и во многих других случаях даже простая арифметическая модель (сложение) не работает.

С другой стороны, мир математики «богаче» реальности, поскольку в нем непротиворечиво сосуществуют реалистичные и физически *невозможные* конструкции (объекты). Например, физические законы, которые организуют наш реальный четырехмерный мир (3 координатных оси пространства плюс время), без нарушений выполнялись бы и в мире с большей размерностью (5, 10... осей). Иными словами, не существует формально-математических «запрещений» для многомерных миров, структура которых, тем не менее, воплощена лишь на бумаге (в компьютере), но не во Вселенной.

В то же время имеются обширные области, где можно установить четкие отношения между математикой и реальностью, между *свойствами чисел* и *свойствами природных объектов*. При этом важно всегда помнить, что природа не наделена числовой системой. По этому поводу К. Поппер пошутил: думать о том, что природа устроена математически из-за того, что ее можно описать математическими формулами, все равно, если думать, что природа устроена по-английски, поскольку ее можно описать на английском языке. Даже действительные числа не находятся сами по себе в каком-либо *естественном* соответствии с материальным миром. Это соответствие приходится специально исследовать и *устанавливать*.

Измерение

Исследованием и установлением отношений между миром чисел и природой занимается *метрология*, наука об измерениях. *Измерение* – способ выражения свойств объектов с помощью чисел соответствующих шкал. *Шкала* – последовательность чисел, служащих для количественной оценки какой-либо величины.

В центре внимания метрологии находятся две системы: *эмпирическая система с отношениями* (ЭСО) (множество реальных природных объектов и отношений между ними) и *числовая система с отношениями* (ЧСО) (множество чисел и отношений между ними).

Первая задача метрологии состоит в том, чтобы *определить шкалу*: установить такое соответствие между системами ЭСО и ЧСО, что если эмпирические объекты (*a, b*…) вступают в некоторые эмпирические отношения друг с другом, то их числовые образы (*M*(*a*), *M*(*b*)…) должны вступать в соответствующие отношения. Ориентируясь на реальность, для конструируемой шкалы устанавливаются некоторые из разнообразных отношений, существующих между числами абсолютной (числовой) шкалы. Практика показывает, что отношения между «арифметическими» числами гораздо разнообразнее, чем наблюдаемые отношения между реальными объектами и *многие свойства чисел отсутствуют у природных объектов*. В стремлении приблизить шкалы к объектам измерения и исследования разработаны 5 видов эмпирических шкал с модификациями. Значения, принадлежащие разным эмпирическим шкалам, неравноценны; одни из них обладают бóльшим объемом свойств, присущих математическим числам, другие обладают ограниченным их списком. Чем богаче структура эмпирического свойства, тем шире спектр адекватных числовых свойств, которые можно привлечь для его описания. Примером *слабой шкалы* служит балльная оценка знаний (это порядковая шкала): для нее не выполняются некоторые арифметические законы. Так, нельзя сказать, что объем знаний у студентов, получивших оценки 5 и 4, различается настолько же, что и оцененных на 4 и 3, хотя, казалось бы, 5 – 4 = 4 – 3 = 1. Аналогично объем знаний «хорошиста» вовсе не в два раза больше, чем у «двоечника». Здесь реальное «количество знаний», «заключенное в студентах», не соответствует идеальному количеству, заключенному в числах 5, 4 , 3 и 2, то есть баллы 2, 3, 4 и 5 – это не числа!

Почему так важно иметь точное представление о комплиментарности ЭСО и ЧСО, о соотношении свойств чисел и объектов еще до акта измерения? Основная *цель измерения* состоит вовсе не в том, чтобы приписать эмпирическому свойству какие-то числа. Устанавливая соответствие между объектами и числами, мы тем самым приписываем объектам свойства чисел, а *числам приписываем свойства объектов*! Так мы получаем возможность *из чисел извлекать информацию о свойствах объекта*. Получить числа – не самоцель науки, главное здесь – изучить (как можно точнее) законы отношений между реальными объектами (закономерности изменения их свойств). Числа должны быть *адекватны* свойствам тел; это позволяет (корректно выполняя их обработку) добывать новое знание, устанавливать закономерности, формулировать законы. С этих позиций измерить свойство означает поставить в соответствие каждому объекту из ЭСО определенное число из ЧСО и расширить наше понимание свойств эмпирической системы.

Помимо определения шкал в функции метрологии входит составление предписаний, как *правильно выполнять* измерения величин. Здесь необходимо ввести ряд определений. *Величина* – все, что способно увеличиваться или уменьшаться. *Физическая величина* – это свойство, качественно общее множеству объектов, но количественно индивидуальное для каждого из них. *Измерить физическую величину* значит получить ее *значение* в результате сравнения этой величины с единицей физической величины с помощью средства измерения, хранящего эту единицу. Средства измерения градуируются, а результаты измерения выражаются в установленных единицах с помощью уравнения $Q = n\,[Q]$, где $[Q]$ – единица, n – число единиц. *Единица физической величины* – значение величины, которое по определению считается равным 1.

На протяжении истории человечества единицы одних и тех же свойств постепенно менялись, унифицировались и уточнялись. В качестве эталонов для первых единицы длины выбирались части тела человека (дюйм – ширина сгиба большого пальца; фут – длина ступни; локоть, сажень), которые существенно различались в разных уголках Земли. В XVII в. использовалось около 100 разных футов. Постепенно система мер становилась строже, во-первых, из-за введения сопряженных единиц (1 фут = 12 дюймов), во-вторых, благодаря выбору более объективных единиц (считать дюймом длину

трех ячменных зерен). В Европе это привело к созданию метрической системы (8.05.1790), утвердившей единые эталоны многих величин. Например, метром была принята единица, равная одной сорокамиллионной «парижского меридиана», расстояния от Барселоны до Дюнкерка; килограммом стала масса воды объемом 1 $дм^3$ при температуре 4 °С. В дальнейшем эталоны заменялись более строгими, а состав единиц измерения пересматривался. Сейчас метр – это длина, равная 1656763,83 длины волны в вакууме некоторого типа излучения атома криптона 86, а секунда – 1/315569259747 часть 1900 года (между двумя весенними равноденствиями).

Последняя Международная система единиц СИ (**S**ysteme **I**nternational d'Unites, **SI**) определяет состав четырех групп единиц: основных, производных, допустимых внесистемных и недопустимых. К основным относятся метр, килограмм, секунда, ампер (сила тока), кельвин (термодинамическая температура), моль (количество вещества), кандела (сила света). Производные единицы получены как отношения между основными – это герц, джоуль, вольт, ом, люмен, грей и др. В число внесистемных, но допустимых для использования единиц, включены тонна, минута, градус, литр, градус Цельсия и др. К недопустимым единицам, подлежащим изъятию из обращения, отнесены ангстрем, центнер, дина, лошадиная сила, калория, рад, рентген, кюри, мм ртутного столба и некоторые другие.

Измерения обладают разными качествами, в число которых включаюn следующие. *Точность* результатов измерений, характеризуемая величиной погрешности (стандартной ошибки) (п. 1.2). *Сходимость*, или близость результатов измерений в одинаковых условиях. *Воспроизводимость* – близость результатов измерений в различных местах при разных условиях. *Быстрота* получения промеров. *Единство* измерений, определяемое равенством единиц измерения. Каждое из этих качеств измерения определяется как характером организации измерительной процедуры, так и настройкой (исправностью) инструмента измерения, который хранит единицы измерения. В этом отношении ответственность исследователя состоит как в соблюдении правил измерительной процедуры, так и в проверке исправности (калибровке) измерительной аппаратуры, даже если речь идет о простой линейке.

2.2. Шкалы

Шкала – это не просто ряд чисел, служащих для отображения свойств эмпирических объектов, это – утверждение специфических отношений между ними (*гомоморфизм*), т. е. таких, что некоторые свойства эмпирических объектов однозначно соответствуют некоторым свойствам чисел (это значит, числа способны отображать учтенные свойства объектов). В своем предельном выражении подобные отношения носят название *изоморфизм*, или взаимнооднозначное отображение двух совокупностей. Основой для формирования шкал различных признаков служит множество «математических» чисел, свойства которых лишь в некоторой степени присущи числам частных шкал. Иными словами, «объемы» гомоморфизма для разных шкал отличаются.

Проблема шкал биологических признаков мало привлекает внимание исследователей и довольно слабо отражена в литературе. Так, если в общем понятно, что длину тела рептилий измерять удобнее в сантиметрах, чем в попугаях, то остается вопрос, в каких единицах измерять приспособленность *организмов*, толерантность *особей*, конкурентоспособность *видов*, устойчивость *экосистем*? Традиционно подобные измерения пытаются свести к физическим и химическим шкалам (например, толерантность измеряют в концентрациях загрязнителя). Однако не поэтому ли биология остается пока слабо развитой дисциплиной, что переход на собственные меры затянулся? Разве достаточно одних деклараций о специфическом биологическом типе организации материи, и не пора ли измерять ее в биологических единицах? Опыты такого рода предпринимаются редко (пример – введение «собственного» биологического времени для разных видов) и зачастую встречаются «в штыки».

В практике эколого-биологических исследований используются все типы шкал: номинативная, порядковая, интервалов, разностей, отношений и абсолютная шкала.

Абсолютная шкала

Шкала имеет и абсолютный нуль, и абсолютную единицу – это шкала чисел, числовая ось. Она получена логическим путем и не нуждается для своего определения в реальности, в предметах. Различают несколько типов чисел: натуральные (целые положитель-

ные), рациональные (натуральные плюс нуль, отрицательные, дроби, логарифмы), иррациональные (не имеющие конечного определения – e, π, $\sqrt{2}$ и др.), мнимые (не имеющие формы $\sqrt{-1}$), комплексные и др.

Только в одном случае абсолютная шкала используется непосредственно – при счете предметов. В других случаях она служит «вспомогательным» средством для операций над числами, полученными в других шкалах. К ней применимы (из нее исходят) все арифметические операции, ее члены могут использоваться в качестве показателей степени чисел (в природе логарифмы «не живут»).

Числа абсолютной оси обладают всеми свойствами чисел, многие из которых еще не открыты, не известны, не изобретены. Все другие шкалы лишь в той или иной мере приближаются к абсолютной; наиболее близка шкала отношений.

Абсолютная шкала не имеет именованных единиц измерения, она безразмерна. К такой же форме выражения признаков прибегают для характеристики относительных свойств объектов, исходно оцененных в иных шкалах, – это индексы, доли, коэффициенты. Например, нормированное отклонение $t = (x - M)/S$ переводит любой непрерывный признак в форму безразмерного показателя. Однако простое избавление от единиц измерения не превращает эмпирический признак в абсолютный, его свойства не становятся в большей мере адекватными абсолютной числовой оси. Для различения этих шкал в последнем случае говорят о *частной абсолютной шкале* (см. ниже).

Шкала наименований

В мышлении и разговорном языке человек пользуется почти исключительно *шкалами наименований* (номинальными шкалами). Любое понятие (мысленный образ) обозначает группы явлений или вещей, сходных друг с другом и отличающихся от других подобных групп. Если мы способны отличить черное от белого, красного, синего, воздух от воды, земли и огня, значит, мы пользуемся номинальными шкалами. Рассматривая только эмпирические свойства, нетрудно найти сходство нескольких объектов по степени его выраженности, т. е. определить *классы эквивалентности*. Для обозначения одинаковых классов объектов (*a, b, c*…) должны выбираться

одинаковые символы, или равные значения некой величины *M*, соответствующие одинаковой выраженности свойств наблюдаемых объектов. Выражаясь более формализовано, шкала наименований основана на *отношении эквивалентности* между значениями. Оно представляет собой *первое правило для определения количественного понятия*: если между эмпирическими объектами *a* и *b* имеется отношение эквивалентности E_M (=), то эти объекты буду иметь равные значения величины *M*: если E_M (*a, b*), то *M*(*a*) = *M*(*b*).

Каждому классу ставят в соответствие обозначение *M*(*a*), отличное от других классов *M*(*b*), *M*(*c*)... Результаты измерений записываются с помощью слов, символов чисел, обозначающих данный класс объектов. Так, цвета радуги обозначаются словами (красный, синий...). В других случаях пользуются буквами (горизонты почвы: A, B…), буквенными сочетаниями (репродуктивные группы животных: juv, sad, ad, sen), числами (номера пробных площадок). Эквивалентность объектов внутри класса носит условный характер, то есть в группу, обозначенную общим символом, могут попасть довольно сильно различающиеся предметы (например, объекты *синего* цвета могут иметь разные оттенки, интенсивность, яркость, тон).

Отношение эквивалентности включает следующие утверждения: либо *M*(*a*) = *M*(*b*), либо *M*(*a*) ≠ *M*(*b*). Если *M*(*a*) = *M*(*b*), то *M*(*b*) = *M*(*a*). Если *M*(*a*) = *M*(*b*) и *M*(*b*) = *M*(*c*), то *M*(*a*) = *M*(*c*).

По существу, эти формулы выражают три закона логики (тождества, непротиворечия, исключения третьего) и соответствуют формализации «здравого смысла». *Формализация* – это логическая операция исключения из рассмотрения всех свойств вещи, кроме одного, важного с точки зрения исследователя. Простой пример – обозначение самок животных с помощью символа ♀, «значение» которого одинаково для новорожденных, неполовозрелых, взрослых и старых особей, которые по другим качествам, кроме половой принадлежности, сильно отличаются друг от друга.

При обработке экспериментальных данных, измеренных в шкале наименований, можно выполнять лишь *операции проверки совпадения* или *несовпадения* значений. Они позволяют идентифицировать объекты, заключить, равны ли (или эквивалентны) два элемента друг другу, но и только. *Запрещены операции сравнения* «выраженности» свойств объектов, с использованием символов, обозначающих классы. Так, номера объектов лишь внешне выглядят

как числа, но на самом деле не обладают большинством их свойств, т. е. практически числами не являются. Нельзя сказать, например, что среди участников соревнований спортсмен под номером 2 хуже носителя номера 4 (тем более, в два раза).

Принято считать, что шкала наименований – это самая «слабая» шкала, поскольку она рассматривает лишь общие качественные характеристики. Однако она широко используется в биологии, психологии, социологии политике, поскольку позволяет выразить *сложные* свойства, с трудом поддающиеся формализации. Как, например, выразить числом «силу материнской любви»? А словами можно. В биологии шкалы наименований широко используются и чрезвычайно важны для классификационных построений. В XVII веке центральной задачей биологии считалась именно классификация объектов природы (по К. Линнею: «назвать и расположить в системе»), которая до сих пор не утратила своей актуальности, поскольку на ней основана эволюционная теория. В медицине любое лечение начинается с установления диагноза – отнесения заболевания к определенному классу. Более того, самые изощренные методы математической обработки чисел возвращаются к номинальным шкалам как к главному результату расчетов. Кластерный анализ и диагностические процедуры (глава 5), компонентный анализ и многомерное шкалирование (глава 8) служат для выявления кластеров – групп сходных объектов, которые по сути есть номинальные классы. Мысль человека стремится к привычным номинальным образам, подчеркивающим качественное (значит, сильное) различие сравниваемых объектов, главные черты рассматриваемого явления.

Оцифровка номинальных шкал

Несмотря на качественную заданность категорий номинальной шкалы, на ее основе допустимы операции количественного обобщения, поставляющие новую информацию об объектах исследования. Во-первых, можно *подсчитывать количество* совпадений (*частоты*, или объемы классов эквивалентности), получая *распределения частот встречаемости* носителей данного качества (см. рис. 3.2.2, 3.2.11). Здесь осуществляется переход от качественного признака к количественному, что дает возможность выполнять численные сравнения и получать строгие доказательные выводы. Например, распределения можно сравнивать с помощью критерия

хи-квадрат (χ^2). Во-вторых, анализ формы распределения дает возможность определить *моду* (*Mo*), самое распространенное значение из числа полученных вариант.

Другой путь трансформации порядковых шкал связан с разработкой объективных методов измерения. Так, каждый из цветов радуги можно однозначно и точно охарактеризовать длиной волны соответствующего видимого излучения (например, $\lambda \approx 0.3$ мкм соответствует синему свету, $\lambda \approx 0.7$ – красному). Измерить можно все, главное – захотеть и разработать соответствующий метод.

Существует третий путь превращения классификаций в более «сильные» шкалы. Основанием введения новой шкалы является решаемая проблема. Именно *в фокусе определенной цели* различные объекты (явления, вещи, понятия, показатели), служащие для решения проблемы, получают свою *оценку относительной важности* – характеристику своей значимости в деле улучшения состояния системы. С определенной точки зрения номинальные категории удается ранжировать, выстроить в определенном порядке, соответствующем их желательности, важности, ценности относительно целевого критерия. Это означает построение следующей по строгости *порядковой шкалы*. Более того, предлагаются разнообразные эффективные процедуры трансформации номинальных (и порядковых) шкал – в *шкалу отношений*. В основу этих алгоритмов положен *метод парных сравнений*, использующий различные *функции предпочтений*:

$$w_{ij} = \begin{cases} a, \text{ если } i\text{-й объект предпочтительнее } j\text{-го объекта} \\ b, \text{ если } i\text{-й объект эквивалентен } j\text{-му объекту} \\ c, \text{ если } j\text{-й объект предпочтительнее } i\text{-го объекта} \end{cases}.$$

В качестве показателей результатов сравнений (значения *a, b* и *c*) предлагаются разные числа, например 1, 0.5, 0 или -1, 0, 1 или 2, 1, 0 (Михеев, 2004). Результатом сличения друг с другом *n* элементов становится квадратная *матрица парных сравнений* размерностью *n*n*. Простейшая обработка этой матрицы состоит в поиске усредненных по строкам оценок, которые численно характеризуют значимость каждой номинальной категории.

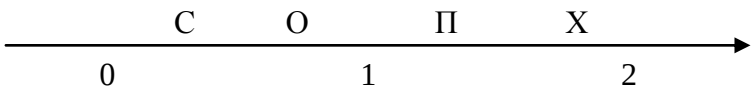
Положим, что для выбора маршрута экологической тропы (скорости движения, времени маршрута, транспортного средства)

нам нужно учесть пожелания экскурсантов наблюдать ту или иную группу птиц. Грубо выделив четыре номинальных значения (певчие, околоводные, синантропные и хищные), попробуем оценить их относительную значимость, сравнивая категорию строки с категорией столбца. Для функции предпочтений выберем значения $a = 2$, $b = 1$, $c = 0$. Наблюдать певчих интереснее, чем водных (2) и синантропных (2), но менее интересно, чем хищных (0). Продолжая сопоставления, строим таблицу 2.1.1.

Таблица 2.1.1. Шкала предпочтений (w_i)

Насколько строка важнее столбца?					Среднее
w_i	Певчие	Околоводные	Синантропные	Хищные	
Певчие	1	2	2	0	1.25
Околоводные	0	1	2	0	0.75
Синантропные	0	0	1	0	0.25
Хищные	2	2	2	1	1.75

После усреднения определились ранги приоритетных групп, их относительное расположение на оси сравнений.



Матрицу парных различий можно обработать и достаточно сложным *методом многомерного шкалирования* (п. 8.3), который выполняет детальную расшифровку информации, скрытой в исходных данных. Промежуточным по сложности, но очень эффективным методом выступает оценка *относительной важности* разнокачественных объектов.

Шкала относительной важности

Метод предложен для придания количественной определенности нашим суждениям об относительном вкладе в достижение намеченной цели того или иного элемента изучаемой системы (свойства, фактора, причины, критерия нашей деятельности). В от-

личие от уже рассмотренных показателей (значения *a*, *b* и *c*) важность объекта *i*, относительно объекта *j*, выражается баллом от 0 до 9, а важность объекта *j*, относительно объекта *i*, – дробью: если $a = 5$, то $c = 1 / a = 1 / 5 = 0.2$; $b = 1$ (табл. 2.1.2).

Выработка количественной меры идет в три этапа. Сначала каждому элементу назначается серия оценок его важности относительно всех других элементов, принятых к рассмотрению; значения заносятся в таблицу (табл. 2.1.3). Затем полученные оценки для каждого элемента усредняются по формуле средней геометрической: $W = \sqrt[n]{\prod \frac{w_j}{w_i}}$ (значок $\prod$ выражает произведение нескольких членов). В завершение они суммируются (ΣW) и для каждого элемента рассчитывается доля средней от суммы $p_j = \frac{W_j}{\sum W}$.

Таблица 2.1.2. Шкала относительной важности (w_i / w_j)
(Саати, Кернс, 1991)

Интенсивность важности	Определение	Объяснения
1	Равная важность	Равный вклад двух видов деятельности в цель
3	Умеренное превосходство	Опыт и суждения дают превосходство одного над другим
5	Существенное превосходство	
7	Значительное превосходство	Практически значимое превосходство одного над другим
9	Очень сильное превосходство	
2, 4, 6, 8	Промежуточные решения	Применяется как компромисс
Обратные величины оценок	Если «превосходство» А над Б выражено числом 3, то «превосходство» Б над А получает значение 1/3	

Таблица 2.1.3. Оценки относительной важности для двух элементов

	Элемент 1	Элемент 2	W_j	p_j
Элемент 1	$\frac{w_1}{w_1} = 1$	$\frac{w_1}{w_2}$	$\sqrt[n]{\prod \frac{w_1}{w_i}}$	$\frac{W_1}{\sum W}$
Элемент 2	$\frac{w_2}{w_1}$	$\frac{w_2}{w_2} = 1$	$\sqrt[n]{\prod \frac{w_2}{w_i}}$	$\frac{W_2}{\sum W}$
Сумма			$\sum W$	1.00

Рассмотрим оценку предпочтений отлова того или иного вида озерной рыбы в Карелии при ловле удочкой на червя. Состав улова обычно ограничен шестью видами. Зададимся вопросом («фокус проблемы»): какой рыбки хотелось бы наловить (при прочих равных условиях)? Составим таблицу относительной ценности улова (табл. 2.1.4), сравнивая по парам виды рыб с точек зрения, например, интереса к процессу лова, вкусовых качеств и возможности приготовить разные блюда. Категории в строке сопоставляются с категориями в столбце. Поскольку отлов леща эквивалентен отлову леща, в первой ячейке ставим 1.

Таблица 2.1.4. Относительная важность отлова рыбки

Насколько строка предпочтительнее столбца?									
w_i / w_j	Лещ	Окунь	Щука	Плотва	Ерш	Уклея	П	W_j	p_j
Лещ	1	0.3333	0.5	5	5	5	20.83	1.66	0.21
Окунь	3	1	3	6	4	3	648.00	2.94	0.37
Щука	2	0.33	1	5	5	3	50.00	1.92	0.24
Плотва	0.20	0.17	0.20	1	2	0.3333	0.00	0.41	0.05
Ерш	0.20	0.25	0.20	0.50	1	0.3333	0.00	0.34	0.04
Уклея	0.20	0.33	0.33	3	3	1	0.20	0.76	0.10
								8.03	1.00

Следующий вопрос: насколько лещ предпочтительнее окуня? Поймать леща (подлещика) на червя удается много реже, но ведет он себя вяло; вкусен, но костляв; годится в основном на жарку.

Умеренный проигрыш леща оцениваем $w_{\text{лещ}} / w_{\text{окунь}} = 1 / 3 = 0.3333$. Соответственно, в графу предпочтения окуня над лещом ставим $w_{\text{окунь}} / w_{\text{лещ}} = 1 / 0.3333 = 3$. На червя может клюнуть только маленькая щука (щуренок), ее нежное молодое мясо годится для самых разных кулинарных изысков и вкуснее, чем у подлещика; за ней умеренный приоритет: $w_{\text{лещ}} / w_{\text{щука}} = 1/2 = 0.5$, $w_{\text{щука}} / w_{\text{лещ}} = 2$. По сравнению с плотвой (а также ершом и уклеей) лещ намного выигрывает: $w_{\text{лещ}} / w_{\text{плотва}} = 5$, $w_{\text{плотва}} / w_{\text{лещ}} = 0.2$. Аналогичным образом выполняем парное сравнение всех объектов лова.

Заполнив матрицу предпочтений, находим произведения n членов строки (П). Для ряда Лещ имеем $1 \cdot 0.3333 \cdot 0.5 \cdot 5 \cdot 5 \cdot 5 = 20.83$. Извлекаем из произведения корень n степени (у нас $n = 6$): $W_1 = \sqrt[6]{20.83} = 1.66$ (формула Excel: =20.83^(1/6)).

Суммируем все значения ($\Sigma W = 8.03$) и находим относительные величины $p_1 = 1.66 / 8.03 = 0.21$. Проведя сортировку видов рыб по значениям относительной важности, получаем ряд приоритетов: окунь (0.37), щука (0.24), лещ (0.21), уклея (0.1), плотва (0.05), ерш (0.04).

Не все читатели согласятся с назначенными в таблице 2.1.4 приоритетами. Для многих романтический факт поимки на удочку леща будет важнее его утилитарных (кулинарных) качеств. Иными словами, перед классификацией объектов следует определиться со значимостью критериев. Полная процедура оценки относительной важности (названная *методом анализа иерархий*) одним из этапов как раз и предполагает оценку относительной важности критериев (Саати, Кернс, 1989), но более подробное рассмотрение процедуры выходит за рамки нашей книги.

Порядковая шкала

Когда объекты номинальной шкалы удается сравнить по изучаемому свойству и упорядочить (рассортировать), говорят о порядковой шкале. В известной (искусственной) классификации цветковых растений К. Линнея таким признаком выступило число тычинок. Современная же (естественная, генеалогическая) классификация являет пример типичной номинальной шкалы.

Данной шкале присуще отношение порядка, которое представляет собой *второе правило определения количественного понятия*. Если между эмпирическими объектами *a* и *b* имеется *отношение порядка* L_M (свойства предмета *a* выражены слабее, чем у предмета *b*), то значения величины *M* будут меньше для *a*, чем для *b*: если L_M (*a*, *b*), то $M(a) < M(b)$. Иными словами, порядковая шкала кроме аксиом эквивалентности удовлетворяет *аксиомам упорядоченности*: если $M(a) < M(b)$, то $M(b) > M(a)$. Если $M(a) > M(b)$ и $M(b) > M(c)$, то $M(a) > M(c)$.

Порядковые шкалы вводятся для тех свойств объектов, которые не имеют количественно выраженного стандарта (единицы измерения) и основаны зачастую на чувственном восприятии человека. Для многих явлений созданы *балльные* порядковые шкалы.

Общеприняты *шкалы балльной оценки знаний*, имеющие от 2 до 100 делений (значений): 2 (зачет), 4 (экзамен), 5 (школа) (Россия), 10 (школы Европы), 100 (колледж США) и др. *Шкала силы ветра* Ф. Бофорта (1806 г., английский гидрограф) имеет 12 градаций: 0 – штиль, 4 – умеренный ветер, 6 – сильный ветер, 10 – буря, 12 – ураган. *Шкала твердости* Ф. Мооса (1811 г., немецкий минералог) дает 10 значений: 1 – тальк, 2 – гипс, 3 – кальций, 4 – флюорит, 5 – апатит, 6 – ортоклаз, 7 – кварц, 8 – топаз, 9 – корунд, 10 – алмаз. В биологии балльные градации распространены достаточно широко, в их число входят баллы проективного покрытия почвы (6), классы бонитета (5 градаций), тип трофности водоемов (3) и др.

При назначении числа градаций балльных шкал (процесс *квантификации*) свою роль играют субъективные и объективные факторы. Исследования по психологии показывают, что человек одновременно оперирует (держит перед внутренним взором, в кратковременной памяти) в среднем 5 (2–10) объектов или понятий. Поскольку большинство балльных оценок дают эксперты, ориентирующиеся на свои чувственные восприятия, обычно шкала имеет менее 10 градаций. Исследования в области теории информации дают правила для распознавания *сигналов* с использованием понятий мощность и помехи (Экоинформатика, 1992). Показано, что *уровень квантирования* (число квантов разной информации, число градаций, число групп на балльной шкале) пропорционален двоич-

ному логарифму отношения мощности сигнала к мощности шумов: $L \cong a \log_2 \left(\frac{сигнал}{шум} \right)$.

Так, при отношении *сигнал / шум*, равном 10 (ошибка порядка 25%), удается выделить лишь 4 градации, для 100 (ошибка 12%) – 8 баллов, для 1000 – 32 интервала. При среднем уровне величины сигнала человек способен воспринимать информацию с ошибкой около 12%, т. е. хорошо различать около 8 уровней величины признака. Говорят, что такие шкалы имеют логарифмическое (показательное, степенное) основание. Это подтверждает сопоставление балльных оценок (B) с прямыми измерениями одной и той же величины (X): между ними обнаруживается не прямая пропорция, а степенная (показательная, аллометрическая) связь вида $B = bX^a$. Например, шкала балльной оценки проективного покрытия почвы травянистыми растениями имеет лишь 5 зрительно отличающихся состояний: отсутствие покрытия (0%, 0 баллов), единичные объекты (1–5%, 1), слабое покрытие (5–30%, 2), сильное покрытие (30–70%, 3), сплошное (70–100%, 4). Эти соотношения можно описать степенным уравнением кривой $B = 0.4\ X^{0.5}$ (рис. 2.2.1). Зависимость баллов (B) от истинной характеристики (X) не позволяет непосредственно пересчитать балльные оценки в точный показатель (B – в X), приемы преобразования баллов в более сильные шкалы рассмотрены ниже.

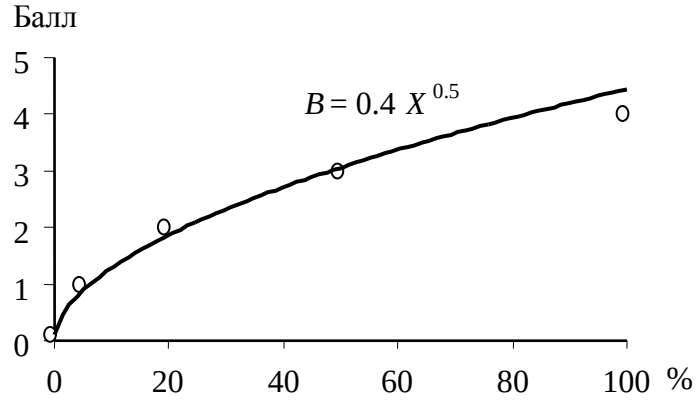


Рис. 2.2.1 Соотношение балльной и инструментальной оценок

Для величин, выраженных шкалой порядка, допустимо сравнение, выяснение, какое из двух значений больше, какое из двух наблюдений предпочтительнее. Однако ориентируясь на эти значения, ничего нельзя сказать о дистанции (разности) между сравниваемыми классами.

Используя значения баллов, можно утверждать, например, что студент, получивший за экзамен 4, существенно лучше знает материал, чем получивший 2. Но при этом нельзя сказать, что этих знаний у него в два раза больше. Здесь неизмеряемое реальное «количество знаний» студентов не соответствует идеальному количеству, заключенному в балльных числах 2 и 4. Аналогично нельзя утверждать, что алмаз (твердость 10 баллов) в 2 раза тверже апатита (5 баллов).

Раз неизвестна дистанция между значениями данной шкалы, то порядковые экспериментальные данные (даже изображенные числами) нельзя рассматривать как полноценные числа. Для них не выполняются арифметические законы, поэтому результаты арифметических действий над ними невозможно интерпретировать. В частности, средняя арифметическая нескольких баллов не имеет смыслового содержания. Используя диаграмму 2.2.1, в этом нетрудно убедиться. Например, на одном участке местности фактическое проективное покрытие может варьировать в пределах 30–40%, а на другом – 60–70%. Однако по разработанной шкале они получат *равные* средние арифметические баллы, хотя фактически будут отличаться в два раза. Для травяного покрова можно уточнить методы измерения, но что делать с повсеместно принятым усреднением экзаменационных баллов учащихся?

Значения шкалы порядка могут обрабатываться методами, приемлемыми для номинальных шкал: сравнивать значения и относить к определенному классу (классификация), подсчитывать частоты (построение распределений), определять моду. Кроме них становятся доступными другие обобщающие показатели. С этой целью значения ряда ранжируют. *Ранг* выражает номер объекта в упорядоченном ряду. Сходные объекты получают средний ранг, или *мидранг*. Для ранжированных рядов (упорядоченных и пронумерованных вариант) можно определить *медиану* (аналог средней) и ее ошибку (п. 4.1), вычислить *коэффициенты ранговой корреляции* между двумя выборками, сравнить две выборки с помощью *непара-*

метрических критериев (п. 4.2) и сравнить несколько выборок – с помощью *непараметрического дисперсионного анализа* (Ивантер, Коросов, 2003).

Следует отметить, что в эколого-биологических исследования ранговыми показателями часто пренебрегают, но стремятся использовать аппарат средней арифметической и других параметрических статистик даже в тех случаях, когда ими пользоваться нельзя или неудобно. Скорее всего, это связано с тем, что руководства по биометрии содержат мало методов статистической оценки и сравнения ранговых показателей. В 4-й главе рассмотрены характерные задачи для таких методов.

Как и в случае с качественными признаками, порядковые характеристики стремятся выразить в более сильной *шкале отношения*, чтобы расширить набор приемов количественной обработки данных.

Шкала желательности Е. Харрингтона

Данная метрика разработана для решения задач планирования экспериментов, когда важно по серии разнокачественных характеристик получить единообразную «функцию отклика системы». Она служит для перевода признаков любой природы (как количественных, так и качественных) в безразмерную шкалу относительного непрерывного показателя, принимающего значения от 0 до 1. Для перевода отдельного признака в шкалу желательности используются 6 стандартных отметок (табл. 2.2.1). Каждой отметке шкалы (d) ставят в соответствие определенные уровни выраженности свойств объектов измерения (X). Характеристика выраженности свойств в ключевых точках (0.37, 0.63) должна быть как можно более точной. После построения шкалы все выполняемые эмпирические наблюдения (измерения) оцениваются как значения функции желательности.

В основе шкалы лежит экспоненциальная функция $d = e^{-e^{-X}}$. Ключевые значения функции подобраны так, чтобы им соответствовали простые выражения $0.63 \approx 1 - (1 / e)$, $0.37 \approx 1 / e$ (рис. 2.2.2). Величина $d = 0.37$ задает границу допустимых значений.

Таблица 2.2.1. Стандартные отметки на шкале желательности *d* (Адлер и др., 1976)

Желательность	Диапазон	Отметки
Очень хорошо	1.00–0.80	0.8
Хорошо	0.80–0.63	0.63
Удовлетворительно	0.63–0.37	0.37
Плохо	0.37–0.20	0.20
Очень плохо	0.20–0.00	

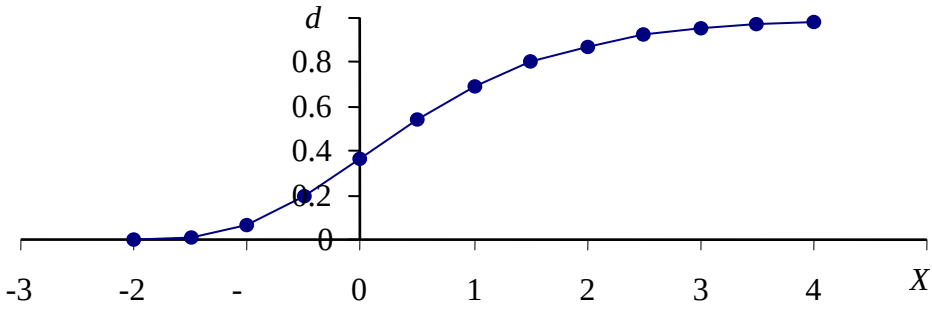


Рис. 2.2.2. Соотношение признака *X* и функции желательности *d*

Как и показательная, функция желательности хорошо описывает соотношения между балльными оценками и реальными значениями изучаемого признака. В частности, по графику видно, что вблизи от значений 0 и 1 «чувствительность» показателя *d* к изменению преобразуемого признака значительно ниже, чем в средней зоне. Это вполне согласуется с опытом: для органов чувств легче отличить сильное проявление свойств от слабого, чем небольшие отличия пары маленьких или пары больших значений признака.

Для функции желательности характерны важные качества: непрерывность, монотонность, гладкость. Они позволяют выполнять операцию объединения нескольких разных признаков, переопределенных в шкалу d_i, в форму обобщенной функции желательности, имеющей вид средней геометрической из *n* частных функций: $D = \sqrt[n]{d_1 \cdot d_2 \cdot ... \cdot d_n}$. Рассмотренный показатель служит достаточно простым способом интеграции нескольких изучаемых характери-

стик с тем, чтобы в количественной форме выразить приоритеты научной и практической деятельности.

Построим шкалу для количественной оценки качества природы при выборе места воскресного пикника. Сколько людей, столько и мнений, но все же почти всеми высоко ценится чистая вода, возможность позагорать, отсутствие других отдыхающих и следов их пребывания. Выбрав эти четыре критерия, дадим характеристику всем интервалам шкалы (табл. 2.2.2).

Таблица 2.2.2. Шкала желательности для свойств места отдыха

Желательность	Свойство			
	Вода	Открытость	Соседи	Следы
1.00–0.80	Прозрачная, чистая	Широкий (10 м) песчаный пляж	Никого, тишина	Кроме кострища все чисто
0.80–0.63	Желтоватая, без запаха	Узкий пляж, бараний лоб	1 чел./ час	Тропки, кострище, дрова, окурки
0.63–0.37	Буроватая, прозрачная	Обширные луга на берегу озера	Изредка появляются люди	Бутылки и пакеты легко убрать
0.37–0.20	Бурая с болотным запахом	Поле с перелесками, лесной луг	В час видишь 5 и более человек	Много бутылок, банок, пакетов
0.20–0.00	Мутная, непрозрачная, пахучая	Глухой лес	Люди постоянно в поле зрения	Многолетние запасы мусора, ямы и др.

Используя эту таблицу, нетрудно дать интегральную характеристику любому месту предполагаемого отдыха. Рассмотрим че-

тыре точки, которые знакомы многим жителям Петрозаводска: 1) берег Онежского озера в районе Октябрьского проспекта, 2) берега островов Кижского архипелага, 3) восточный берег Ладоги, 4) берега Падозера (дачные поселки Лучевое) (табл. 2.2.3).

Таблица 2.2.3. Оценка качества четырех мест отдыха

Место	Вода	Открытость	Соседи	Следы	*D*
Петрозаводск	0.3	0.6	0.1	0.2	0.24
Острова Онеги	0.8	0.3	0.7	0.8	0.61
Ладога	0.4	1.0	0.5	0.3	0.49
Падозеро	0.7	0.1	0.1	0.7	0.26

Оценки конкретным местам возможного отдыха назначаются по шкале желательности. Например, летом на берегу Ладожского озера в районе Устьобжанки вода светлая, желтоватая, но мутная, с органическими частицами ($d_{\text{вода}} = 0.4$). Обширные песчаные пляжи имею ширину 15–50 м ($d_{\text{открытость}} = 1.0$); людей немного, но при обзоре в 1–2 км они постоянно виды ($d_{\text{соседи}} = 0.5$); из-за частого посещения отдыхающими и прибоя по всему берегу раскиданы бутылки, тара, обрывки, тряпки, веревки, сети, но в небольшом количестве ($d_{\text{следы}} = 0.3$). Обобщаем оценки: $D_{\text{Ладога}} = \sqrt[4]{0.4 \cdot 1.0 \cdot 0.5 \cdot 0.3} = 0.49$.

Интегральные значения желательности на первое место ставят острова Кижского архипелага (0.61). Расширив список регистрируемых показателей, можно существенно уточнить правила составления рейтинга. Со своей стороны, увеличивая число анализируемых точек, мы получаем возможность составлять карты распространения предпочтительных мест отдыха, точно спланировать свое место под воскресным солнцем.

Шкала интервалов

Многие свойства биологических объектов имеют довольно простую структуру, благодаря чему их удается описать более полноценными числами. Переход к количественной шкале связан с применением еще трех правил, устанавливающих единицы измерения. *Третье правило определения количественного понятия* состоит в том, что некоему легко узнаваемому состоянию приписывается

значение нуль, $M_0 = 0$. *Четвертое правило* состоит в том, что другому легко опознаваемому состоянию приписывается значение единицы, $M_1 = 1$. *Пятое правило* определяет точную форму шкалы. Если различие (отношение D) между объектами a и b такое же (отношение E), что и между объектами c и d, то разности между соответствующими значениями величины M также равны: если ED (a, b, c, d), то $M(a) - M(b) = M(c) - M(d)$.

Иными словами, используя характеристику M, упорядочивание объектов удается выполнить настолько точно, что становятся известны расстояния между любыми двумя объектами, которые поэтому могут быть измерены в одинаковых единицах, т. е. одинаковыми по длине участками шкалы. Примером служит температурная шкала, для которой отношения интервалов из разных областей шкал остаются равными: десять градусов, заключенные между 10 и 20°, и десять градусов, заключенные между 100 и 110°, одинаковы.

Для интервальных характеристик могут использоваться различные точки начала и конца шкалы и методы расчета единицы измерения, эти отметки не являются единственно возможными. Например, на шкале Цельсия температура замерзания воды равна 0 º, закипания – 100º (диапазон делится на 100 частей); на шкале Фаренгейта температура замерзания воды принята за 32º, закипания – за 212º, а диапазон между ними поделен на 180 частей. Обе шкалы равноценны и удовлетворяют пяти правилам количественного показателя. Важно помнить, что в интервальной шкале *свойствами чисел обладают* не сами значения температуры, но *интервалы* (разности между отдельными значениями), то есть только интервалы правильно отражают физический смысл процесса. Например, теплота, обеспечивающая нагрев тела с 10 до 20°, равна теплоте, обеспечивающей его же нагрев со 100 до 110°. В то же время сами значения температуры не в полной мере обладают свойствами чисел. Например, нельзя сказать, что температура 18 °С в два раза выше, чем 9 °С, поскольку объект при $t = 18$ °С вовсе не в два раза больше содержит теплоты, чем объекте при $t = 9$ °С! Причина кроется в том, что для шкал интервалов выбор нулевого значения произволен, фактически он не соответствует полному отсутствию данного свойства. Как известно, в промежутке от 0 ºС до абсолютного нуля 0 ºК заключено целых 273º!

Еще ярче ограниченность интервальной оценки проявляется при сравнении разных шкал. Казалось бы, 9 °C (по шкале Цельсия) в два раза меньше, чем 18 °C. Но по шкале Фаренгейта соответствующие числа 37 и 42 °F не поддерживают этой пропорции (42 / 37 = 1.14). Установить реальные тепловые соотношения было бы возможно, если воспользоваться «сильной» шкалой отношений, которую имеет физическое свойство – теплота, Q. В интервальной шкале между свойствами объектов и свойствами чисел нет полного тождества; тождество есть только между свойствами чисел и свойствами *интервалов* данной шкалы.

Для шкалы интервалов возможно вычисление разнобразных статистических характеристик (параметров) – средней арифметической, дисперсии, коэффициента корреляции и др., модно использовать параметрические методы статистического оценивания. При этом следует помнить, что средняя арифметическая (и другие центральные моменты) имеет узкий смысл – только относительно начала отсчета, хотя дисперсия (учитывающая интервалы, разности между значениями) имеет абсолютный физический смысл.

Шкала отношений

В отличие от шкалы интервалов вводится абсолютный единственный нуль, соответствующий состоянию объектов при полном отсутствии изучаемого свойства; шкала обладает единственным нулем, где свойство исчерпывается. Примером могут служить разные системы измерения расстояний, например с единицами «сантиметр» и «дюйм». Для них значение нуль (0) имеет общий смысл: при отсутствии «длины» 0 дюймов равен 0 см.

При этом возможно использование разных единиц измерения. Например, за единицу длины в континентальной Европе приняты метр, а в Британии и Америке – фут, ярд, дюйм. Процесс унификации шкал физических величин идет постоянно.

Изменения, выполненные в данной шкале, обладают свойствами настоящих чисел, над которыми можно выполнять любые арифметические действия. Средняя арифметическая величина наряду с дисперсией имеет абсолютный физический смысл.

Шкала разностей

Вариант шкалы интервалов, для которой *принят* единственный нуль, т. е. приписано свойство шкалы отношений. Это возможно благодаря особой организации расположения шкалы на инструментах измерения – по кругу (с помощью циферблата). Таковы компас (направления, °), время суток (с), фаза колебаний (°, рад.). Главное свойство подобной шкалы состоит в том, что она повторяет свои значения через некоторый период (цикл) и поэтому ее называют еще и *периодической*, или *циклической* шкалой. Общепринятый нуль позволяет рассматривать измерения в этой шкале как настоящие числа и выполнять над ними все возможные арифметические операции.

Частная абсолютная шкала

Многие свойства объектов исследования, исходно измеренные в разных шкалах, стремятся выразить в относительных единицах, лишенных размерности, – с помощью индексов, долей, отношений. На первый взгляд, таким путем получают величины, подобные числам абсолютной шкалы, которые не имеют единиц измерения. Тем не менее их свойства не могут «улучшиться» от простой операции деления, степень близости индексов к абсолютной числовой оси «наследуется» от исходных шкал. В то же время индексы приобретают новые *негативные* свойства, которые приходится дополнительно изучать и учитывать. Используя тот или иной индекс, всегда необходимо точно определять, в чем смысл их нулевого и единичного значения, каков характер распределения индекса, каковы его статистические свойства. Обычно статистические характеристики индексов оказываются хуже, чем у исходных признаков, и ничего, кроме дополнительной сложности для понимания сути явления, они не дают. К большинству индексов, предложенных биологами с позиций «здравого смысла», следует относиться с настороженностью. Даже такой простой показатель, как относительная масса органа ($M_{орг.} / M_{тела}$), во-первых, включает в себя две причины изменения и изменчивости: бóльшие значения индекса могут получиться как при большой массе органа, так и при низкой массе тела – понять истинные факторы варьирования индекса оказывается сложно. Во-вторых, сама операция деления искажает исходно нормальные распределения показателей $M_{орг.}$ и $M_{тела}$: распределение индекса приобретает правостороннюю асимметрию и приближается к лог-

нормальному. Квадраты, логарифмы, степени, используемые в индексотворчестве, привносят свои специфические искажения, запутывающие суть дела (см. п. 3.3, 5.3). Прежде чем предлагать свой индекс, мы рекомендуем ознакомиться с уже имеющимся набором показателей с известными статистическим свойствами, либо обратиться к математическим процедурам, формирующим универсальные линейные индексы или уравнение регрессии (главы 5, 6, 8).

Путаница возникает и при исследовании индексов с помощью корреляционного и регрессионного анализов; обычная ошибка состоит в том, что зависимость между индексами интерпретируется как зависимость между исходными признаками. Например, уровень *корреляции между массой* разных внутренних органов в группе животных будет определяться степенью изменчивости их размеров (чем крупнее особь, тем крупнее все ее органы). В свою очередь, величина корреляции между индексами органов для той же группы будет связана со степенью физиологической напряженности, с уровнем развития стресса, когда рост одних органов (например, генеративных) подавлен, а другие гипертрофированы (органы барьерные и внутренней секреции). По этой причине к*орреляция между индексами* будет определяться уже тем, в какой мере наблюдения охватывают все фазы пика популяционного цикла.

Несмотря на большие сложности в построении адекватных биологических характеристик, создание индексов представляет собой, видимо, единственный путь получения *биологических шкал* (главы 1, 5, 8).

Один из простых и распространенных методов создания частной абсолютной шкалы – оценки *расстояний между особями по нескольким признакам* с последующей классификацией объектов по группам (кластеризация). По сути дела, на основе многих количественных характеристик особей (биообъектов) создается новая номинальная шкала, которую можно сопоставить с аналогичными номинальными шкалами для факторов среды и прийти к содержательным выводам. В качестве меры различия между объектами часто используется *евклидово расстояние*: $d_i = \sqrt{\sum^{k}(x_{i1} - x_{i2})^2}$, где $(x_{i1} - x_{i2})$ – – разность между свойствами двух объектов (всего дано k свойств). Необходимый пример рассмотрен в разделе 5.4.

К сожалению, эта мера не свободна от недостатков и, в частности, зависит от единиц измерения признаков. Например, длина тела мелких позвоночных обычно выражается в граммах (35.5 г), а масса внутренних органов – в миллиграммах (2300 мг); это значит, что вклад в значение d массы тела будет в 100 раз меньше, чем вклад внутренних органов. Поскольку такая игра цифр не имеет никакого отношения к сути биологических различий между особями, эти величины следует унифицировать, стандартизировать.

Известны методы центрирования, двойного центрирования (п. 8.2) и *нормирования*, когда каждое реальное значение признака x_i заменяют *нормированным отклонением*: то есть вычитают из него среднюю данного признака (M) и делят на свое стандартное отклонение (S): $z_i = \frac{x_i - M}{S}$. В силу свойств нормального распределения нормированные отклонения обычно принимают значения в диапазоне ± 3. Рассчитав нормированные значения для каждой варианты каждого признака, их можно использовать для вычисления обобщенного расстояния между объектами: $d_i = \sqrt{\sum^{k}(z_{i1} - z_{i2})^2}$.

Нормирование лишь отчасти снимает проблему избавления от единиц измерения. Не все признаки распределены нормально и диапазон изменчивости может быть весьма велик (± 10). Нормированные значения контекстуальны, т. е. определяются тем, каковы были оценки средних и дисперсий; отличия в их репрезентативности могут привести к тому, что *одинаковые* объекты в контексте разных выборок получат *разные* количественные характеристики, что негативно скажется на результатах сравнений.

Предлагаются и другие оценки различий (Котов, 1985). Хорошие статистические свойства имеет логарифмическая мера расстояния l, которая порождает одни и те же отношения равенства и порядка, независимо от единиц измерения признаков: $l_i = \left(\sum^{k} \left| \ln \frac{x_{i1}^{j}}{x_{i2}^{j}} \right|^r \right)^{\frac{1}{r}}$, где i – номер признака, 1 и 2 – номера сравни-

ваемых объектов, j, r – произвольные числа. В случае, эквивалентном метрике Евклида ($j = 1$, $r = 0.5$), мера примет такой вид:

$$l_i = \sqrt{\sum^{k} \left| \ln \frac{x_{i1}}{x_{i2}} \right|^2} .$$

Особенность нового показателя состоит в том, что теперь отыскивается не разность значений признаков $x_{i1} - x_{i2}$, но логарифм отношения значений признаков $\ln \frac{x_{i1}}{x_{i2}}$, то есть *разность степенных показателей* значений признаков, которая не зависит от единиц измерения. Таким образом, безо всякого предварительного нормирования приходят к адекватной оси биологических расстояний между объектами! К сожалению, мера не подходит для случаев, когда значения признаков равны нулю.

Глава 3
РАСПРЕДЕЛЕНИЕ ПОКАЗАТЕЛЕЙ

В практике биологических исследований приходится иметь дело с количественными оценками свойств, которые являются следствиями не только закономерных проявлений живого, но несут в себе и элементы случайного. Стохастическое «поведение» выражается в определенном группировании вариант на числовой оси: одни значения встречаются чаще, другие реже. Распределение – это соотношение между значениями признака и частотой их встречаемости. Распределение представляет собой не иллюстрацию, а природную закономерность, результат действия систематических и случайных факторов на каждую варианту, представителя объекта исследования.

3.1. Модель варианты

Случайное

Источники случайного выступают в роли объектов исследования не только для математики, физики, биологии, но и философии. Марк Аврелий писал, что «либо мир является огромным хаосом, либо в нем царствует порядок и закономерность; какая из этих двух взаимоисключающих возможностей реализуется, мыслящий человек должен решить сам». Однако дело состоит не столько в выборе позиции мыслителя, сколько в объективных сложностях процесса пополнения наших знаний.

В первую очередь, случайными принято считать те события, причины которых скрыты от нашего текущего знания. Любое описание не учитывает некоторые важные факторы, влияющие на явление. Биологические системы разных уровней – от макромолекул до биосферы – настолько сложны и недостаточно исследованы, что внутренне закономерные (причинно детерминированные) явления внешне выглядят как непредсказуемые, случайные. Неполнота наших знаний будет ощущаться всегда, поскольку каждая вещь имеет бесконечное число свойств, познать которые до конца невозможно.

Не так давно обнаружились и другие источники неточности даже текущего знания. Случайное порождается самим научным наблюдением, поскольку играет роль достаточно сильного вмешатель-

ства в наблюдаемую систему, которое изменяет или даже разрушает ее. «Принцип неопределенности Гейзенберга», родившийся в недрах квантовой физики, гласит: «Мы в принципе не можем знать настоящее во всех его детерминированных подробностях». Например, электрон может «сообщить» наблюдателю лишь какую-либо одну характеристику своего текущего состояния, исчерпав весь запас свободной энергии (в виде дискретного кванта), которой попросту больше нет, чтобы проявить другие свои свойства. Можно вспомнить, что основные биологические знания получены на трупах, то есть на объектах, тоже утративших свой «квант жизни».

Третье начало случайного, как выяснила школа И. Пригожина, состоит в том, что мир и устойчив, и изменчив одновременно. Порядок способен перерастать в хаос, а хаос – структурироваться, упорядочиваться. Было показано, что многие равновесные динамические системы в условиях притока энергии совершают быстрый переход в одно из немногих новых известных состояний, но в какое именно, предсказать нельзя. С ростом потока энергии число потенциальных состояний множится, упорядоченность поведения системы утрачивается. Примеры таких систем известны в математике (системы из трех дифференциальных уравнений), в физике, химии, биологии, социологии и др. науках. При специфическом взаимодействии элементов строго детерминированных системы у них настолько усиливаются прежде несущественные свойства, что их проявления выводят ситуацию из равновесия (из-под контроля); поведение предсказуемых по отдельности динамических систем при их интеграции становится хаотическим, случайным.

Случайная величина

С бытовой точки зрения *случай* – это такой ход событий, который невозможно однозначно предсказать на базе имеющихся сведений. Более формализованное определение с*лучайности* – это возможность появления разных исходов испытаний (наблюдений). Если «испытания природы» выразить какими-либо величинами, мы получаем ключевое определение: с*лучайная величина* – величина, принимающее те или иные заранее точно не известные значения. *Варьирование, изменчивость* – появление различных значений случайной величины при наблюдении. *Варианта* – отдельное значение варьирующего признака (случайной величины). *Статистическая*

совокупность – множество вариант определенного типа, множество значений случайной величины.

В качестве примера можно положить результаты измерения размеров гадюки, который для южной Карелии составил от 48 до 68 см. Значение 53 см – варианта. Множество вариант образует распределение, графически отображаемое в форме полигона частот (рис. 3.1.1) или столбчатой диаграммы – гистограммы.

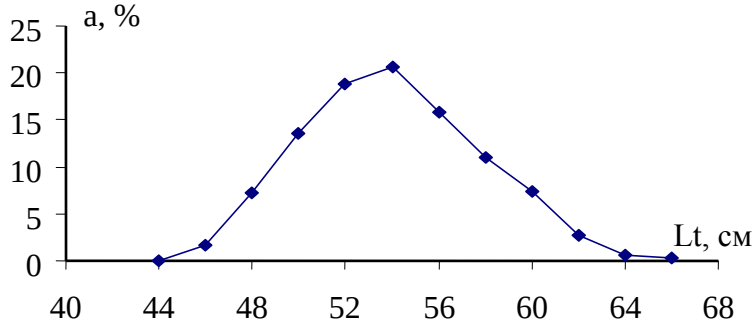


Рис. 3.1.1. Встречаемость взрослых самок гадюки с разной длиной тела (без хвоста)

«Жизнь варианты»

Разобраться в причинах формирования оригинальной формы распределения разноразмерных особей помогает прием разделения всех факторов, обеспечивающих процесс формирования каждой варианты, на две группы – случайные и доминирующие (систематические). На каждую варианту выборки доминирующие факторы действуют одинаково, случайные же могут подействовать, могут и не действовать. Проиллюстрируем эту мысль диаграммой «траектории жизни» отдельной варианты (рис. 3.1.2).

Доминирующий фактор определяют ту точку, от которой «берет старт» каждая варианта. В случае с гадюками – это средняя видовая норма самок гадюки $Lt \approx 54$ см (x_2). Данный доминирующий фактор вносит одинаковый вклад в формирование величины любой варианты. Несколько случайных факторов, нам не известных и нами не контролируемых, обуславливают конкретное для каждой варианты отклонение от уровня, предписанного доминирующим фактором. В число случайных причин входят генетические особенности (для варианты x_1 – отклонение вправо), условия развития (вправо), усло-

вия жизни (влево), здоровье (влево). В данном примере в число случайных (неучтенных) причин входит и возраст особей. Конкретная варианта (x_i) занимает свое место на числовой шкале.

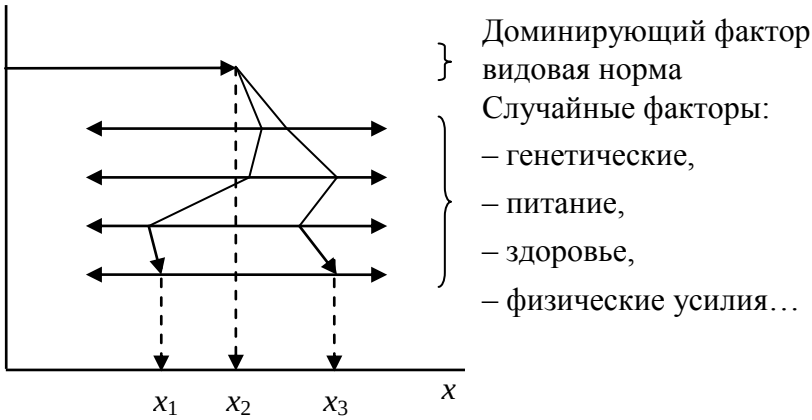


Рис. 3.1.2. Траектория формирования значения случайной величины

Как правило, число доминирующих (изучаемых, известных) факторов мало или он всего один, случайных факторов множество и они не известны.

Модель варианты

Значение каждой варианты (x_i) складывается из доли, определенной действием доминирующего фактора ($x_{домин.}$), и доли, определенной действием случайных факторов ($x_{случ.}$), – это можно записать как *модель варианты*:

$$x_i = x_{домин.} \pm x_{случ.},$$

где x_i – значение варианты, i – индекс варианты (i = 1, 2, … , n), $x_{домин} = \sum x_{j\ домин.}$ – общий вклад j доминирующих факторов, $x_{случ.} = \sum x_{k\ случ.}$ – суммарный вклад k случайных факторов.

Важно отметить, что вклад случайных факторов в формирование значения варианты может быть как положительным, так и отрицательным, часть факторов благоприятствует росту значения от уровня, обеспеченного доминирующим фактором, часть действует в противоположном направлении. Выборочные значения оказываются распыленными в некоторой области. Например, гадюку с длиной тела 58 см можно воспринимать как имеющую видовую норму

54 см, невысокие потенции прироста (+2 см), испытавшую в ранние годы дефицит благоприятных физических условий среды (–3 см), но хорошо питавшуюся (+5 см): 58 = 54 + (2 – 3 + 5).

Аддитивный (суммативный) подход к представлению отдельного значения очень важен для статистических моделей, особенно многомерных (см. главу 8).

3.2. Типы распределения признаков

Для описания «статистического поведения» случайных величин было предложено довольно большое количество моделей (типов распределений), связывающих значения признака (x) и частоту их встречаемости (A): $A = f(x)$. С целью придать этим моделям универсальный характер в теоретических формулах используют относительные величины: теперь формула выражает вероятность встречи (P) нормированных отклонений вариант (t): $P = f(t)$.

Можно выделить несколько целей практического применения законов распределения. В первую очередь, подбирая («примеривая») к эмпирическим данным ту или иную модель, мы фактически переносим принципы организации этих распределений на биологический материал и тем самым выясняем механизмы биологического явления. Например, распределение самок животных по плодовитости позволяет вывести вероятность рождения или гибели отдельного эмбриона (зависящего от условий жизни, стрессов матери, интоксикации и пр.), т. е. по морфологическому феномену дать физиологическое заключение (см. *Распределение биномиальное, альтернативное*). Кроме того, соответствие эмпирических данных тому или иному типу распределения позволяет воспользоваться статистическими методами, разработанными специально для этой модели, то есть получить наиболее полую информацию, сделать более точный прогноз, правильнее оценить выборочные параметры. Большое число статистических методов разработано для переменных, имеющих нормальное распределение. Установив «нормальность» эмпирической совокупности, всем этим арсеналом можно пользоваться. Если же распределение имеет иной вид, то применять следует иные статистики и критерии. Короче говоря, приступая к обработке данных, необходимо сначала исследовать частоты встречаемости вариант.

Нормальное распределение

Наиболее характерный тип распределения *непрерывных случайных величин* имеет колоколообразную симметричную форму (кривая Гаусса): крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем они чаще встречаются (рис. 3.2.1). Среднее квадратичное отклонение примерно 4 раза укладывается в размахе изменчивости признака $S \approx Lim/4$ и по величине в 3–5 раз меньше средней арифметической $S < M/3$. Модель нормального распределения имеет вид: $p_i = (1/\sqrt{2\pi}) \cdot e^{-(x_i - M)^2 / 2 \cdot S^2}$ или $p_i = (1/\sqrt{2\pi}) \cdot e^{-t^2/2}$, $t = (x_i - M) / S$. Формулы расчета основных параметров (средней, дисперсии) обычно подходят и для других типов распределений: $M = \sum x_i / n$ = 15.4 г, Ошибка средней равна: $m_M = \frac{S}{\sqrt{n}}$ = 3.8 / 136 = 0.326; $M \pm m_M$ = 15.4 ± 0.36 г. Стандартное отклонение с ошибкой равно: $S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}}$ = 3.8 г., $m_S = \frac{S}{\sqrt{2 \cdot n}}$ = 0.162; $S \pm m_S$ = 3.80 ± 0.162 г.

Для построения нормального распределения по эмпирическим данным применяют метод *интервального оценивания*. Диапазон значений изучаемого признака разбивают на интервалы (с шагом *dx*), подсчитывают число вариант, попавших в них (эмпирическая частота *a*), далее по формуле нормального распределения *вычисляют* вероятность попадания очередной варианты в тот или иной интервал значений. Зная вероятность, можно рассчитать ожидаемую (теоретическую) частоту появления значений в том или ином интервале ($A = p \cdot n$). Работа идет в несколько этапов (рис. 3.2.1):

1. Определение параметров выборки *n, M, S, k, dx* и поправки $C = n \cdot dx / S$ = 136 · 2.6 / 3.8 ≈ 91.
2. Вычисление границ и центров интервалов, составление вариационного ряда (см. подробнее: Ивантер, Коросов, 2003).
3. Вычисление нормированных отклонений для центральных зна-

чений: $t_i = (x_i - M) / S$ … $t_3 = (9.9 - 15.37) / 3.8 \approx -1.44$…

4. Вычисление частостей (вероятностей): $p_i = (1/\sqrt{2\pi}) \cdot e^{-t^2/2}$; … $p_3 = (1/2.507) \cdot e^{-(-1.44)^2/2} = 0.3989 \cdot 0.3546 = 0.142$… (табл. 2С, стр. 350). Рассчитать вероятность для данного значения *t* можно с помощью функции Excel =НОРМРАСП(t,0,1,ЛОЖЬ).
5. Вычисление теоретических частот, используя поправку *C*: $A_i = C \cdot p_i$ … $A_3 = 91 \cdot 0.142 = 12.9$ …
6. Построение гистограммы.

E13 =(1/((2*ПИ())^0.5))*EXP(-(F13^2)/2)

	A	B	C	D	E	F
1			n =	136		
2			M =	15.37		
3			S =	3.805		
4			Lim =	18		
5			k =	7		
6			dx =	2.6		
7			C = n*dx/S =	91		
8	x max	x	a	A	P	t
9	3.4					
10	6	4.7	0	0.7	0.008	-2.81
11	8.6	7.3	4	3.8	0.042	-2.12
12	11.2	9.9	11	12.9	0.142	-1.44
13	13.8	12.5	29	27.3	0.300	-0.76
14	16.4	15.1	46	36.2	0.398	-0.07
15	19	17.7	26	30.1	0.331	0.61
16	21.6	20.3	10	15.7	0.173	1.29
17	24.2	22.9	8.0	5.1	0.056	1.98
18	26.8	25.5	2	1.1	0.012	2.66
19	29.4	28.1	0	0.1	0.001	3.34
20	Σ		126	126.0		

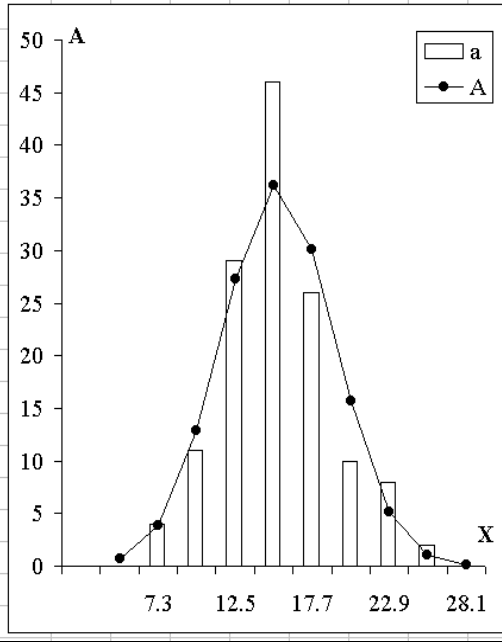


Рис. 3.2.1. Распределение нормальное. Расчет теоретических частот (*A*) с использованием эмпирических частот (*a*) для значений массы тела прибылых рыжих полевок в августе (x max – верхняя граница интервала, x – центр интервала, t – нормированное отклонение, P – плотность вероятности, ординаты нормальной кривой)

Вследствие того, что в примере применялся метод интервального оценивания (оценивалась не вероятность определенного значения непрерывной величины, но вероятность появления этой величины в определенном интервале), вместо *непрерывного* нормального распределения получили *дискретное*. Используя выборки, всегда ограниченные по объему, нельзя построить *гладкий* график распределения, между наблюдаемыми значениями всегда будут пробелы. Обобщая группы значений в интервалах, можно построить, хоть и грубоватое, но близкое к нормальному распределение. Мысленно можно выполнить эксперимент: если увеличивать объем выборки, сужая ширину интервалов и увеличивая их число, то «в пределе» ($n \to \infty$) мы получим гладкую гауссову кривую.

Распределение альтернативное

Распределение дискретной случайной величины, имеющей лишь два противоположных значения (например, орел и решка, белое или черное, 0 и 1, больные и здоровые организмы, сработавшие и пустые ловушки на одной учетной линии, два варианта аллельных признаков, вакцинированные и невакцинированные пациенты среди заболевших). Число классов распределения всегда равно двум, $k = 2$. Распределения дискретных величин получают при отборе проб. *Пробой* называют фиксированное множество вариант, полученных при одном наблюдении; объем пробы равен m. При альтернативном распределении в одной пробе содержится одна варианта, одно из двух возможных значений. Вероятности каждого из них могут быть равны ($p = q$), либо не равны ($p < q$; $p > q$). Вычисления параметров просты и не требуют построения вариационного ряда.

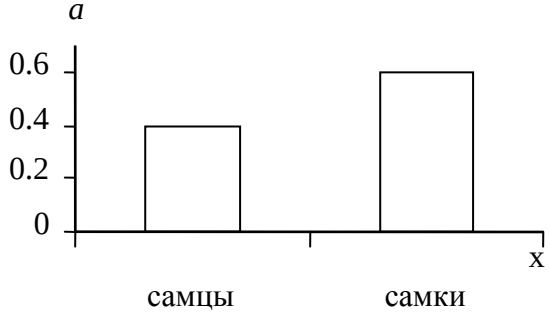


Рис. 3.2.2. Распределение альтернативное. Выборочные доли самок и самцов рыжей полевки

Роль средней арифметической $M = \frac{\sum x}{n}$ играет *доля* (*p*) вариант определенного вида (*A*), представленных общим числом n_A в пределах выборки объемом *n*: $p = \frac{n_A}{n}$, если исходы испытаний выражать числами 0 или 1. Мы имеем право придавать испытаниям вид чисел 0 и 1, руководствуясь идеей отбора вариант группами, *пробами*, когда объем пробы *m* = 1. Так, если для группы особей, состоящей из самцов и самок, подсчитывать число самок в пробе из одной варианты, получаем набор единиц (самка есть) и нулей (самки нет, это самец), т. е. выборку частот встречаемости самок в пробах. В этом случае есть все основания применять формулу средней арифметической и рассматривать долю вариант с качеством 1 для всего ряда значений именно как среднее число встреч самок в пробе.

Один из результатов отловов полевок из природных популяций показал, что в исследуемой группе (200 особей) было 120 самок и 80 самцов. В данном случае мы имеем дело с альтернативным признаком (самка – самец). Из 200 проб 120 проб содержат самок (значение 1), 80 – не содержат (значение 0); так получаем выборку *n* = 200:

00000000000000000000000000000000000000000000000000000000000000000
01111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111.

Доля вариант со значением 1 составляет: $p = \frac{n_A}{n} = \frac{120}{200} = 0.6$, что совпадает со средней арифметической для всего ряда: $M = \frac{\sum x}{n} = \frac{120}{200} = \mathbf{0.6}$. Среднюю (долю самок) можно вычислить и по-другому: $M = m \cdot p = 1 \cdot 0.6 = p = 0.6$.

Стандартное отклонение (для объема пробы *m* = 1) равно: $S = \sqrt{m \cdot p \cdot q} = \sqrt{p \cdot q} = \sqrt{0.6 \cdot 0.4} = 0.24$. Ошибка средней (ошибка доли самок): $m = \frac{S}{\sqrt{n}} = \frac{0.24}{\sqrt{200}} = 0.017$.

Доверительный интервал для альтернативных признаков (их долей, процентов и частот) строится с помощью φ-преобразования Фишера, что дает более точные границы, особенно если доли сильно

отличаются. Сначала вместо значения доли (процента) одного признака объектов берут значение φ (фи), найденное по формуле, $\varphi = 2 \cdot \arcsin\sqrt{p}$ или по табл. 3С (стр. 348). Затем вычисляют ошибку: $m_\varphi = 1/\sqrt{n}$, обе доверительные границы $\varphi_{лев.} = \varphi - Tm_\varphi$, $\varphi_{прав.} = \varphi + Tm_\varphi$, после чего с помощью табл. 3С переводят найденные значения обратно в проценты.

Найдем доверительные границы для доли самок полевок $p = 0.6$ при уровне значимости $\alpha = 0.05$. Проведя расчеты, получаем: $\varphi(60\%) = 1.772$, $m_\varphi = 1/\sqrt{200} = 0.0707$, $\varphi_{лев.} = 1.772 - 1.96 \cdot 0.0707 = 1.6334$, $\varphi_{прав.} = 1.772 + 1.96 \cdot 0.0707 = 1.9106$, $p_{лев.}(1.6334) = 53.1\%$, $p_{прав.}(1.9106) = 66.4\%$.

Доля самок в генеральной совокупности (популяции полевок) составляет минимум 53.1%, максимум 66.4%.

Распределение биномиальное

Биномиальный закон описывает поведение дискретных признаков (выраженных целыми числами) и подходит для описания более или менее симметричного распределения биологического признака, у которого дисперсия меньше средней: $S < M / 3$. Можно привести следующие примеры таких признаков: число больных корнеплодов в пробе, число поврежденных участков на листьях, число волосков на единице площади шкурки, количество лучей в плавниках рыб, число хвостовых щитков у рептилий, плодовитость (размер выводка) самок животных. Для формирования дискретных распределений используется большая *группа* объектов, либо *территория* (маршрут), либо *процесс*. Совокупности разбиваются на *пробы* – порции (группы), участки (площадки), отрезки (этапы). Затем идет подсчет числа проб, содержащих то или иное число известных объектов (явлений) (объемы всех проб равны). Значение отдельной варианты есть число объектов определенного качества в пробе заданного объема, например, число гнилых клубней в пробе из 10 штук.

В основе биномиального распределения лежит альтернативное (качественно отличное) проявление изучаемого признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет. Подсчет количества определенных объектов в

пробе из нескольких объектов превращает качественный признак в количественный.

Выражаясь более формальным языком, рассматривается ситуация, когда из большой совокупности отбирают n проб по m вариант в каждой пробе, т. е. всего в выборке содержится $N = n \cdot m$ вариант. Имеется два типа вариант – обладающие и не обладающие неким свойством. В выборке объемом N изучаемым качеством обладают n_p вариант, а n_q им не обладают ($N = n_p + n_q$). В одной пробе может оказаться x вариант одного качества (от $x = 0$ до $x = m$). При испытании получаем распределение частоты (a) встречаемости отдельных проб с разным числом вариант данного типа (x) (рис. 3.2.3). Доля вариант одного качества равна $p = n_p / N$, другого – $q = n_q / N$. Вероятность того, что в отдельной пробе находится x вариант данного качества выражается формулой бинома Ньютона: $P_x = (p + q)^m$.

Главной «организующей силой» распределения является способ, с помощью которого получают значения случайной величины, – это отбор проб. Для биномиальной модели предполагается, что очередное испытание (отбор проб из m вариант) проходят после возврата предыдущей пробы в исходную совокупность и последующего тщательного перемешивания. Для практики биологических исследований более характерен сбор безвозвратно изымаемых проб из небольших совокупностей. Такая методика изменяет соотношения вариант разного качества, значит, нарушает условие равновероятной возможности обнаружения вариант разного типа. Когда же изучаемые совокупности велики, а требованиях к точности оценок не очень строги, биномиальный закон вполне можно использовать для аппроксимации эмпирических распределений. Сужая круг адресатов, можно сказать, что для целей популяционной экологии этот метод подходит практически всегда, а для популяционной генетики – только при соблюдении указанных условий.

По многим чертам (форма, соотношения параметров) биномиальное распределение (при условии $p \approx q$) сходно с нормальным. Если есть возможность брать большое число вариант в пробах (m), то количество классов распределения будет очень большим (а ширина класса маленькой). Форма такого распределения будет близка к гладкой нормальной кривой. Теоретически при $m \to \infty$ биномиальное распределение превращается в нормальное.

D12 =D11*(E$6/E$7)*((E$1-(A12-1))/(A12))

	A	B	C	D	E
1				m =	10
2				n =	76
3				N = m*n =	760
4				np = Σa*x =	377
5				nq = N-np =	383
6				p = np /N	0.496
7				q = 1-p	0.504
8	x	a	A	P	a*x
9	0	0	0.1	0.00	0
10	1	1	0.8	0.01	1
11	2	1	3.5	0.05	2
12	3	8	9.2	0.12	24
13	4	16	16	0.21	64
14	5	23	19	0.25	115
15	6	21	15	0.20	126
16	7	3	8.6	0.11	21
17	8	3	3.2	0.04	24
18	9	0	0.7	0.01	0
19	10	0	0.1	0.00	0
20	Σ	76	76	0.99	377

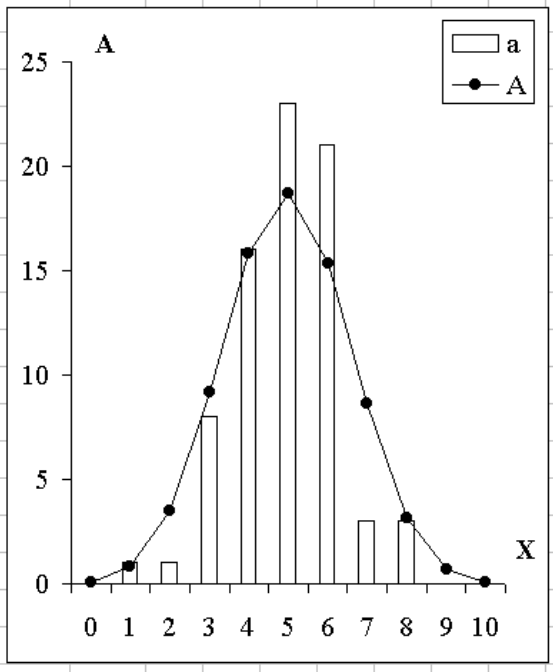


Рис. 3.2.3. Распределение биномиальное. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения плодовитости черно-бурых лисиц (x – число детенышей у самки)

Для практических целей расчеты теоретических частот распределения удобно вести следующим методом. Вначале рассчитывается частость нулевого класса ($x = 0$): $P_0 = q^m$. Затем по рекуррентной формуле определяются частоты в других классах ($x = 1, 2 \ldots m$): $P_x = P_{x-1} \cdot \frac{m - x_{i-1}}{x_i} \cdot \frac{p}{q}$ (x_{i-1} – значения признака в классе, предшествующем текущему i-му). Теоретические частоты равны: $A_x = n \cdot P_x$.

В примере рассматривается распределение плодовитости (x) самок черно-бурой лисицы, $n = 76$. В качестве объема *пробы* можно взять число «потенциальных зародышей» у одной самки $m = 10$. Общее число потенциальных детенышей у всех самок составит $N = 76 \cdot 10 = 760$. Реально появилось на свет $n_p = \Sigma(a \cdot m) = 377$ экз. Вероятность родиться (доля «реализовавшихся» детенышей из числа возможных) составила: $p = n_p / N = 377/760 = 0.496$. Доля неродившихся: $q = 1 - p = 0.504$.

Теоретическая частость для нулевого класса (нет новорожденных, $x = 0$) равна:

$$P_0 = q^m = 0.504^{10} = 0.001056$$

(на листе Excel приведено округленное значение $P_0 = 0.0$; рис. 3.2.3). Рассчитаем частость других классов:

$$P_1 = P_0 \cdot \frac{10-0}{1} \cdot \frac{0.496}{0.504} = 0.001056 \cdot 10 \cdot 0.984 = 0.0104,$$

$$P_2 = P_1 \cdot \frac{10-1}{2} \cdot \frac{0.496}{0.504} = 0.0104 \cdot 4.5 \cdot 0.984 = 0.0461 \text{ и т. д.}$$

Соответствующие этим классам теоретические частоты составят: $A_0 = 76 \cdot 0.001056 = 0.08$, $A_1 = 76 \cdot 0.0104 = 0.79$, $A_2 = 76 \cdot 0.0461 = 3.5$ … Рассчитаем параметры по формулам нормального распределения:

$$M = \frac{\sum x}{n} = 4.96 \text{ экз./самку},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = 1.33 \text{ экз./самку},$$

$$m_M = \frac{s}{\sqrt{n}} = \frac{1.33}{\sqrt{63}} = 0.1676 \text{ экз./самку},$$

$$m_S = \frac{s}{\sqrt{2 \cdot n}} = \frac{1.33}{\sqrt{2 \cdot 63}} = 0.1185 \text{ экз./самку}.$$

Найденные выше вероятности p и q позволяют по-иному рассчитать параметры биномиального распределения (но результаты совпали):

$$M = m \cdot p = 10 \cdot 0.496 = 4.96 \text{ экз./самку},$$

$$S = \sqrt{m \cdot p \cdot q} = \sqrt{10 \cdot 0.496 \cdot 0.504} = 1.58 \text{ экз./самку}.$$

Расхождения в оценках S (1.33 и 1.58) вызваны тем, что первая величина относится к исходной выборке (частотам a), а второе значение соответствует стандартному отклонению для построенного теоретического симметричного распределения (частоты A).

Доверительный интервал для параметров биномиального распределения строится так же, как и для нормального распределения: $M \pm tm_M$, $S \pm tm_S$. Так, при уровне значимости $\alpha = 0.05$ находим доверительный интервал, например, для стандартного отклонения: $S \pm tm_s = 1.33 \pm 1.96 \cdot 0.118 = 1.33 \pm 0.231$ экз./ самку. Значение генерального стандартного отклонения находится в диапазоне от 1.09 до 1.56 экз./ самку.

Распределение Пуассона

Описывает редкие события, происходящие лишь несколько раз на протяжении опыта из десятков и сотен наблюдений. Иными словами, вероятность элементарных альтернативных событий резко неодинакова, одно из них наблюдается много реже, чем другое ($p << q$). В числе примеров – частота нарушений хромосомного аппарата на каждую тысячу митозов, встречаемость семян сорняка в большой серии навесок семян культурного растения, число повторных попаданий животных в ловушки, встречаемость животных на отрезках длинных маршрутов (или на пробных площадках обширной территории), отловы животных в отдельные промежутки времени при длительных наблюдениях. Эта модель служит для описания стохастического поведения дискретных признаков, имеющих резко выраженную правостороннюю асимметрию: очень много проб не содержат ни одного объекта исследуемого качества (много значений $x = 0$). Характерной особенностью является примерное равенство оценок средней и дисперсии $M \approx S^2$; которое используется в качестве критерия данного типа распределения.

Как и в случае с биномиальным распределением, одна варианта может иметь лишь два качества (0 и 1, орел и решка). Значение случайной величины определяется путем подсчета числа вариант данного вида в пробе (в группе, в навеске, на участке, на этапе). Поэтому и теоретической платформой этого закона служит биномиальное распределение, задающее исходную формулу расчета частот в определенном классе: $p_x = (p + q)^m$. Для расчета теоретической вероятности частоты данного класса требуется выполнить разложение

или воспользоваться формулами комбинаторики:

$$p_x = \frac{m!}{x!(m-x)!} \cdot p^x \cdot q^{m-x}.$$

Если же принять, что вероятность *найти* объекты искомого качества в пробе мала ($p \to 0$), а вероятность их *не найти*, соответственно, высока ($q \to \infty$), эти формулы существенно упрощаются. Теперь вероятность отсутствия объекта в пробе (т. е. вероятность попадания значения в нулевой класс, $x = 0$) можно выразить формулой $p_0 = e^{-M}$. Аналогично, вероятность попадания очередной варианты в следующий i-й класс в i раз меньше, чем вероятность попадания в предыдущий (i-1-й): $\frac{p_i}{p_{i-1}} = \frac{M}{i}$.

Например, вероятность появления очередного значения в первом классе (проба содержит лишь один искомый объект, $x = 1$) равна $\frac{p_1}{p_0} = \frac{M}{1}$, откуда $p_1 = p_0 \cdot M = e^{-M} \cdot M$. Для значения $x = 2$ имеем $\frac{p_2}{p_1} = \frac{M}{2}$, откуда $p_2 = \frac{p_1 \cdot M}{2} = \frac{p_0 \cdot M \cdot M}{1 \cdot 2} = \frac{e^{-M} \cdot M \cdot M}{1 \cdot 2} = \frac{e^{-M} \cdot M^2}{1 \cdot 2}$.

Для общего случая (x) получаем формулу закона Пуассона:

$$p_x = \frac{M^x}{x!} e^{-M}.$$

Форма распределения определяется единственным параметром M (часто параметр обозначают буквой «лямбда» $\lambda = M = S^2$). В среде Excel для расчета вероятностей распределения Пуассона для данного класса x при данной средней M служит функция: =ПУАССОН(x,M,ЛОЖЬ).

Теоретические частоты вычисляются как обычно: $A_x = n \cdot p_x$. Если объем выборки n подставить в приведенные выше выражения, сразу получаем формулы расчета теоретических частот вариант в классах. Частота нулевого класса равна: $A_0 = n \cdot e^{-M}$. Рекуррентная формула определяет частоты в других классах (x = 1, 2… m): $A_x = A_{x-1} \cdot \frac{M^x}{x_i}$ (x_i – значения признака в данном классе) (рис. 3.2.4).

В качестве примера рассмотрим результаты мечения с повторным отловом буревестников. В течение одного года (1946) пометили кольцами и выпустили на волю 32 птицы. В последующие пять лет часть из них отлавливали повторно: 7 экз. по одному разу, 7 – по два, 2 – по три, 1 экз. – четыре раза, 15 экз. окольцованных птиц повторно не попадались. Число классов составляет $k = 4$, интервал $dx = 1$. Асимметрия в частотах встречаемости птиц позволяет предполагать распределение Пуассона.

Расчеты показали, что средняя арифметическая (M) примерно равна дисперсии (S^2):

$$M = \frac{\sum x}{n} = \frac{31}{32} = m \cdot p = 4 \cdot 0.242 = 0.968 \text{ экз.},$$

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n-1)}} = \sqrt{\frac{69 - \frac{(32)^2}{32}}{(32-1)}} = 1.121 \text{ экз.}, S^2 = 1.257.$$

C7 =C6*(D$2^A7)/A7

	A	B	C	D
1		n = Σa =		32
2	M = Σ(a*x) / n =			0.97
3	x	a	A	a*x
4	0	15	12.1	0
5	1	7	11.8	7
6	2	7	5.5	14
7	3	2	1.7	6
8	4	1	0.4	4
9	5	0	0.1	0
10	6	0	0.0	0
11	7	0	0.0	0
12	Σ	32	31.5	31
13				
14				

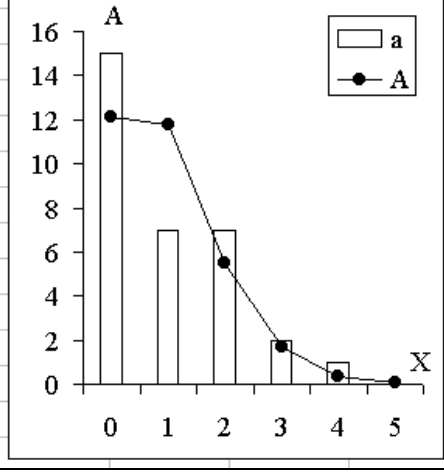


Рис. 3.2.4. Распределение Пуассона. Расчет теоретических частот (A) для эмпирического (a) распределения частоты повторных встреч окольцованных буревестников (x – число повторных отловов)

Проверка по критерию Фишера не выявила достоверных отличий между ними: $F = 1.257 / 0.968 = 1.157 < F_{(0.05,31,31)} = 1.8$ (табл. 5С, стр. 352). Это свидетельствует о соответствии наблюдаемого распределения закону Пуассона.

Доверительный интервал для параметров распределения Пуассона определить несколько сложнее, чем для других типов. В связи с его асимметричностью для расчета левой и правой доверительных границ и средней арифметической (и равной ей дисперсии) применяют формулы с использованием статистики хи-квадрат χ^2:

левая граница $x_{лев.} = 0.5 \cdot \chi^2_{(P, df1)}$,

правая граница $x_{прав.} = 0.5 \cdot \chi^2_{(\alpha, df2)}$,

где P – доверительная вероятность (обычно $P = 0.95$), α – уровень значимости (обычно $\alpha = 0.05$), df – число степеней свободы: $df_1 = 2 \cdot M$, $df_2 = 2 \cdot (M+1)$, M – средняя арифметическая выборки, χ^2 – значение из табл. 6С (стр. 353).

Доверительные границы для среднего числа повторных отловов составят: левая граница $x_{лев.} = 0.5 \cdot \chi^2_{(0.95,2)} = 0.5 \cdot 0.1 = 0.05$, правая граница $x_{прав.} = 0.5 \cdot \chi^2_{(0.05,4)} = 0.5 \cdot 9.49 = 5$. Иными словами, число повторных отловов птиц может варьировать от 0 до 5 раз. Точнее определить среднюю повторную встречаемость нельзя вследствие слишком маленькой исходной выборки.

Распределение гипергеометрическое

Это – симметричное распределение дискретных признаков. При больших объемах выборок оно совпадает с биномиальным распределением. Для относительно небольших выборок моделирует реалистичную ситуацию, когда пробы после испытания не возвращаются обратно в исходную совокупность и варианты после изъятия не перемешиваются (биномиальная модель предполагает возврат и перемешивание). Средняя и дисперсия оцениваются по формулам: $M = mp$, $S^2 = \dfrac{M \cdot q \cdot (N - m)}{(N - 1)}$.

Рассматривается ситуация, когда из большой совокупности отбирают n проб по m вариант в каждой пробе, т. е. всего в выборке содержится $N = n \cdot m$ вариант. Имеется два типа вариант – обладающие и не обладающие неким свойством (например, больной или здоровый корнеплод). В выборке объемом N изучаемым качеством

обладают n_p вариант, а n_q им не обладают ($N = n_p + n_q$). В одной пробе может оказаться x вариант одного качества (от $x = 0$ до $x = m$). При испытании получаем распределение частоты встречаемости отдельных проб (a) с разным числом вариант данного типа (x) (рис. 3.2.5). Вероятность того, что в отдельной пробе находится x вариант данного качества, выражается формулой гипергеометрического распределения: $P_x = \dfrac{C_{n_p}^{x} \cdot C_{n_q}^{m-x}}{C_N^m}$. Расчеты проще вести по другим формулам. Вначале находят частость нулевого класса ($x = 0$):

$$P_0 = \frac{C_{n_q}^{m}}{C_N^m} = \frac{n_q \cdot (n_q - 1) \cdot (n_q - 2)...(n_q - m - 1)}{N \cdot (N - 1) \cdot (N - 2)...(N - m - 1)}.$$

Затем по рекуррентной формуле определяются частоты в других классах ($x = 1$, $x = 2$…):

$$P_x = \frac{P_{x-1} \cdot (m - x_{i-1}) \cdot (n_p - x_{i-1})}{(x_{i-1} + 1) \cdot (N - n_p - x_{i-1} + 1)},$$ где x_{i-1} – значения признака в классе, предшествующем текущему i-му.

Теоретические частоты вычисляются по формуле: $A = n \cdot P_x$.

В примере анализируется распределение в $n = 112$ пробах из $m = 9$ клубней число гнилых (x). Среди всех $N = 112 \cdot 9 = 1008$ изученных клубней гнилых было $n_q = \Sigma(a \cdot m) = 639$, а здоровых – $n_з = N - n_q = 1008 - 639 = 369$ шт. Теоретическая частость для нулевого класса (нет гнилых, $x = 0$) равна:

$$P_0 = \frac{369 \cdot 368 \cdot 367 \cdot 366 \cdot 365 \cdot 364 \cdot 363 \cdot 362 \cdot 361}{1008 \cdot 1007 \cdot 1006 \cdot 1005 \cdot 1004 \cdot 1003 \cdot 1002 \cdot 1001 \cdot 1000} = 0.000111,$$

но в форматированных ячейках Excel она отображается как 0.0 или 1E-04 (рис. 3.2.5). Рассчитаем частость первого и второго классов:

$$P_1 = \frac{P_0 \cdot (9 - 0) \cdot (639 - 0)}{(0 + 1) \cdot (1008 - 639 - 0 + 1)} = 0.000111 \cdot 15.931 = 0.00177,$$

$$P_2 = \frac{P_1 \cdot (9 - 1) \cdot (639 - 1)}{(1 + 1) \cdot (1008 - 639 - 1 + 1)} = 0.\ 00177 \cdot 7.0497 = 0.01245.$$

Соответствующие этим классам теоретические частоты составят: $A_0 = 112 \cdot 0.000111 = 0$, $A_1 = 112 \cdot 0.00177 = 0.2$, $A_2 = 112 \cdot 0.01245 = 1.4$...

D10 =(D9*(F$1-A9)*(F$4-A9))/((A9+1)*(F$3-F$4-F$1+A9+1))

	A	B	C	D	E	F	G
1					m =	9	
2					n =	112	
3					N = m*n =	1008	
4					nq = Σa*x =	639	
5					np = N-nq =	369	
6			P0 = Π(np -x) / Π(N-x) =			1E-04	
7	x	a	A	P	a*x	np -x	N-x
8	0	0	0.0	0.0	0	369	1008
9	1	0	0.2	0.0	0	368	1007
10	2	7	1.4	0.0	14	367	1006
11	3	4	5.7	0.1	12	366	1005
12	4	18	15.0	0.1	72	365	1004
13	5	27	26.0	0.2	135	364	1003
14	6	14	30.1	0.3	84	363	1002
15	7	18	22.2	0.2	126	362	1001
16	8	20	9.5	0.1	160	361	1000
17	9	4	1.8	0.0	36	-	-
18	Σ	112	112.0	1	639		
19	Π					1E+23	1E+27
20							

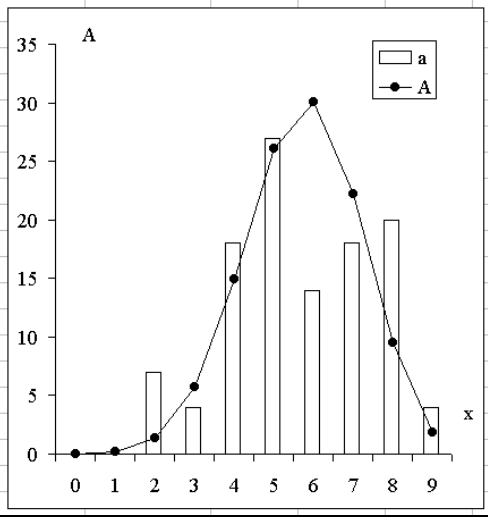


Рис. 3.2.5. Распределение гипергеометрическое. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения числа гнилых клубней в пробе (x – число гнилых клубней в пробе, П – знак произведения между всеми значениями блоков F8:F16 и G8:G16)

Для расчета теоретических частот для данного класса *x* (при общих значениях *m*, n_q, *N*) можно воспользоваться функцией Excel: =ГИПЕРГЕОМЕТ(x,m,n*q*,N), например, для той же ячейки D10 (рис. 3.2.5) функция имела бы такой вид:
=ГИПЕРГЕОМЕТ(A10,F$1,F$4,F$3).

Распределение Парето

Служит для описания распределений с резко выраженной правосторонней асимметрией и резким перепадом частоты встречаемости значений (рис. 3.2.6). Оно определяется одним параметром: $\alpha = \dfrac{M}{M - x_1}$, где *M* – заранее вычисленная средняя, x_1 – центр первого классового интервала. Частоты можно рассчитать в среде пакета Excel по формуле: $A = N \cdot x_1^{\alpha} \cdot (x_{i-1}^{-\alpha} - x_i^{-\alpha})$, где x_{i-1}, x_i – центры двух соседних интервалов, *N* – объем выборки.

D9 =D$4*D$2^D$3*(B8^-D$3-B9^-D$3)

	A	B	C	D	E	F	G	H	I
1			M=	1.195					
2			x1=	0.5					
3			*a*=	1.72					
4			N=	339					
5	Lim	x	a	A					
6	0	0.5							
7	1	1.5	310	287.7					
8	2	2.5	13	30.0					
9	3	3.5	7	9.4					
10	4	4.5	2	4.2					
11	5	5.5	3	2.3					
12	6	6.5	3	1.4					
13	7	7.5	1	0.9					
14	8	8.5	0	0.6					
15									
16	Σ		339	336.4					
17									

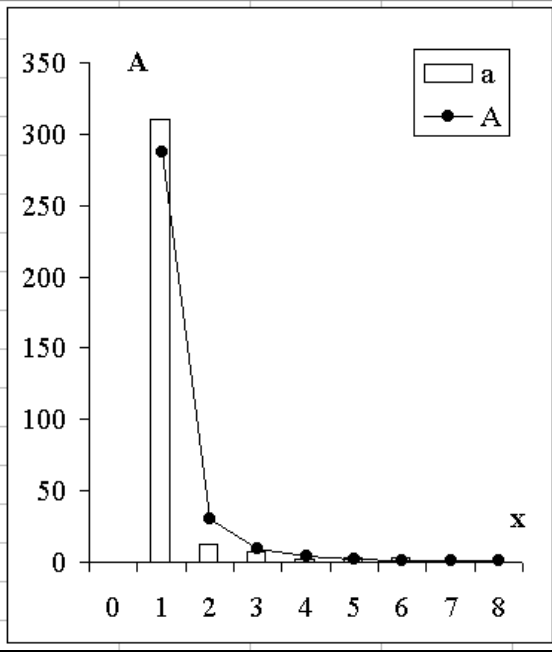


Рис. 3.2.6. Распределение Парето. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения числа побегов на одной особи гелениума (Lim – границы классовых интервалов, x – их центры, *a* – параметр α)

Распределение Рэлея

Используется для описания распределений с умеренной асимметрией (рис. 3.2.7). Средняя в два раза больше стандартного отклонения: $S = M\sqrt{(4/\pi)-1} = 0.522 \cdot M$. Основной параметр – это мода $a = M_o = \sqrt{2M^2/\pi}$. Дисперсия вычисляется по формуле: $S^2 = a^2(2-\pi/2) = 310.2(2-3.14/2) = 133.4$. Частоты вычисляются по формуле: $A = N \cdot dx \cdot \frac{x}{a^2} \cdot e^{-x^2/2a^2}$, где x – центр классовых интервалов, dx – ширина интервалов.

D10 =(B10/D$2)*EXP(-(B10^2)/(2*D$2))*D$3*D$4

	A	B	C	D
1			M=	22.08
2			*a*^2=	310.2
3			N=	205
4			dx=	7
5	Lim	x	a	A
6	1.5			
7	8.5	5	21	22.2
8	15.5	12	48	44.0
9	22.5	19	50	49.1
10	29.5	26	29	40.5
11	36.5	33	34	26.4
12	43.5	40	13	14.0
13	50.5	47	7	6.2
14	57.5	54	2	2.3
15	64.5	61	1	0.7
16	Σ		205	205.4

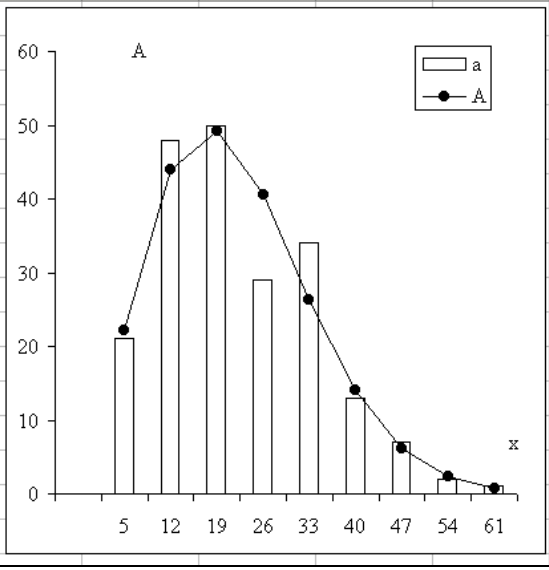


Рис. 3.2.7. Распределение Рэлея. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения числа побегов на растениях пиона (Lim – границы классовых интервалов, x – их центры, *dx* – их ширина, *a*^2 – квадрат параметра a^2)

Распределение Максвелла

Используется для описания распределений с умеренной правосторонней асимметрией (рис. 3.2.8). Параметр распределения равен $a = M / 2\sqrt{2/\pi} = 0.62666M$. Частоты вычисляются по формуле:

$$A = (N \cdot dx / a) \cdot \sqrt{2/\pi} \cdot t^2 \cdot e^{-\frac{t^2}{2}} = \frac{N \cdot dx}{a^2} \cdot \frac{x^2}{a^2} \cdot \sqrt{\frac{2}{\pi}} \cdot e^{-\frac{x^2}{2a^2}},$$

где x – центр классовых интервалов, dx – ширина классовых интервалов, t – нормированное отклонение x / a.

	A	B	C	D	E	F	G	H	I	J
1		M=	16.36							
2		*a* =	10.25							
3		N=	381							
4		dx=	10							
5	x	a	A							
6	5	37	62.6							
7	15	238	217.6							
8	25	78	90.2							
9	35	9	10.2							
10	45	8	=(C$3*C$4/C$2)*(A10^2/C$2^2)*КОРЕНЬ(2/ПИ())*EXP(-(A10^2)/(2*C$2^2))							
11	55	3	0.0							
12	65	3	0.0							
13	75	1	0.0							
14	85	2	0.0							
15	95	0	0.0							
16	105	1	0.0							
17	115	1	0.0							
18	Σ	381	381.0							
19										

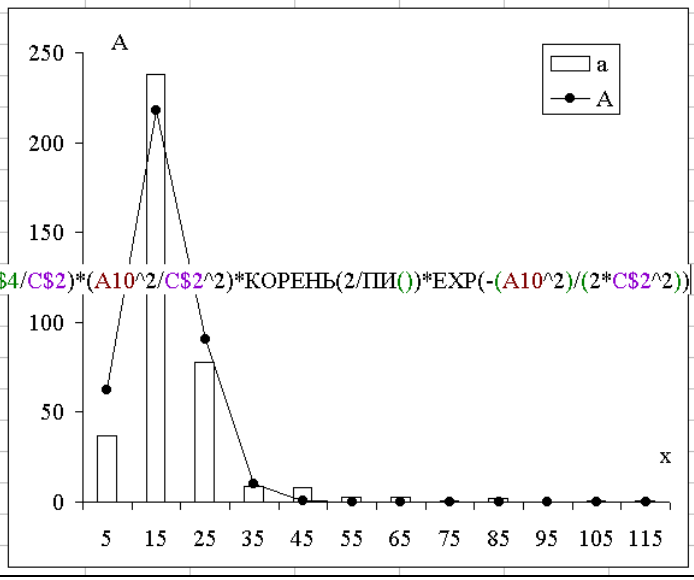


Рис. 3.2.8. Распределение Максвелла. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения продолжительности цветения древесных растений в Ленинградской области (Lim – границы классовых интервалов, x – их центры, *a* – параметр *a*)

Логнормальное распределение

Достаточно часто встречается гладкое асимметричное распределение непрерывных случайных величин, имеющих длинный «хвост» больших значений. Его название отображает тот факт, что логарифмы от исходных значений имеют форму нормального распределения. Эффект связан с тем, что большие значения уменьшаются после логарифмирования гораздо более существенно, чем невысокие величины. Например, lg2 = 0.3 в семь раз меньше исходной величины (2/0.3 ≈ 7), lg10 = 1 – в 10 раз меньше, lg100 = 2 – в пятьдесят раз меньше (100/2 = 50). После логарифмирования асимметричный хвост «подтягивается», сообщая распределению форму кривой Гаусса. С технической точки зрения параметры логнормального распределения можно отыскать после логарифмирования исходных значений как параметры нормального распределения, а затем выполнить обратные преобразования. Например, после логарифмиро-

вания (по основанию e, $\ln x$) и усреднения преобразованных значений массы селезенки получаем среднюю $M_{lnx} = 4.66$, в исходных единицах имеем $M_x = e^{4.66} = 106$ мг.

Интересна природа такой неслучайной асимметрии случайного распределения. Наблюдения показывают, что этот тип поведения характерен для показателей, выражающих пропорции, доли, соотношения частей некой развивающейся системы (выборки объектов). Наиболее наглядна «теория дробления» (Дэвис, 1977), описывающая происхождение логнормального распределения размеров песчинок на пляже. В процессе выветривания (разрушение горных пород в силу физико-метеорологических и биотических причин) каждый обломок породы имеет определенную вероятность быть разрушенным в следующем «акте» дробления. Но имеется и некоторая вероятность не быть раздробленным. Небольшому числу частиц «удается» «накапливать» вероятность не быть раздробленными и поэтому они некоторое время сохраняются относительно крупными по сравнению с основной массой частиц. Выполнив промеры на некотором этапе процесса дробления, мы обнаруживает повышенную долю «недоразрушенных» частиц. Хвост распределения соответствует незавершенному процессу дробления (завершение есть превращение песка в пыль и глину). Логнормальное распределение характерно для концентраций веществ в природных водоемах (содержит хвосты «недорастворенных» порций), для биохимических веществ в тканях организмов. Если логнормальное распределение отображает незавершенные процессы на определенном шаге своего развития, значит, оно сохраняет на себе следы истории развития явления, «память» о прошлых состояниях системы. В рассмотренных примерах правый хвост распределения представляет собой арьергард выборки, который через некоторое время обретет меньшие значения (рис. 3.2.9, *А*). Процесс становления промера как бы ориентирован по времени справа налево.

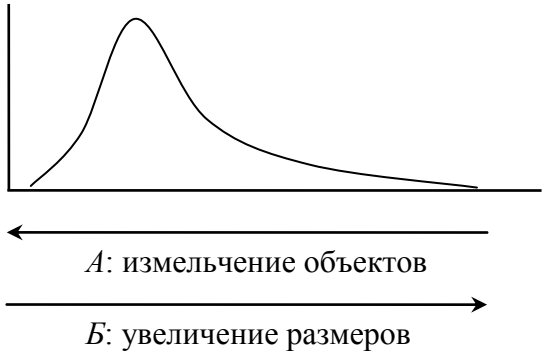


Рис. 3.2.9. Направления процессов, создающие логнормальное распределение

Следующая группа примеров имеет противоположное направление: события «начинаются слева», но все варианты постепенно перейдут в правую область, основное распределение «перетечет на хвост» (рис. 3.2.9, *Б*). Так, из числа биоэкологических характеристик к нашему случаю подходит диаметр бактериальных колоний (большие колонии стремятся расшириться в большей степени, чем малые, но все станут большими), плодовитость грызунов на северной периферии ареала (шансы повысить свою долю летом больше у высокоплодовитых животных, но шансы выжить – меньше), размеры селезенки, надпочечников, гонад неполовозрелых млекопитающих (в любой группе есть особи с гипертрофированными органами, отражающие начало полового созревания).

Распределение показательное (экспоненциальное)

Используется для описания распределений с резко выраженной правосторонней асимметрией. Средняя арифметическая примерно равна стандартному отклонению $M \approx S$. Единственный параметр равен обратной величине от средней $\lambda = 1/M$. Частоты можно найти по формуле (рис. 3.2.10):

$$A = N(e^{-\lambda \cdot g_{i-1}} - e^{-\lambda \cdot g_i}),$$

где g_i, g_{i-1} – верхние границы двух смежных классовых интервалов.

D10 =D$3*(EXP(-D$2*A9)-EXP(-D$2*A10))

	A	B	C	D	E	F	G	H	I	J
1			M=	44.828						
2			λ=	0.0223						
3			N=	2619						
4			dx=	30						
5	Lim	x	a	A						
6	0									
7	30	15	1249	1277.8						
8	60	45	639	654.4						
9	90	75	419	335.1						
10	120	105	185	171.6						
11	150	135	89	87.9						
12	180	165	20	45.0						
13	210	195	10	23.0						
14	240	225	8	11.8						
15	270	255	0	6.0						
16	Σ		2619	2612.7						

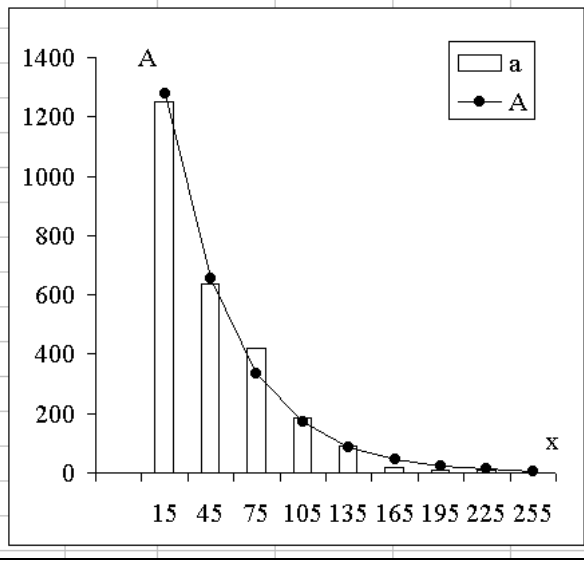


Рис. 3.2.10. Распределение экспоненциальное. Расчет теоретических частот (*A*) для эмпирического (*a*) распределения числа побегов у однолетних растений гелениума (Lim – границы классовых интервалов, x – их центры, верхняя граница первого интервала равна $g_1 = 30$, второго $g_2 = 60$)

Полиномиальное распределение

Наблюдается для качественных признаков, имеющих не два альтернативных свойства, но несколько возможных проявлений качества ($k > 2$). Примеры полиморфизма популяций (в широком смысле слова) – из этой области. В их числе варианты окраски покровов и волос, типы рисунков в определенных областях тела, способы жилкования листьев растений или крыльев насекомых, варианты расположения и формы щитков рептилий и другие проявления множественности фенотипов особей. Формализуя описание, укажем, что в одной пробе содержится одна варианта ($m = 1$), но типов вариант (морф, фенотипов) больше, чем два ($k > 2$).

Примером полиномиального (мультиномиального) распределения может служить встречаемость 4 фенов головы живородящей

ящерицы – 4 варианта контакта лобно-носового, предлобных и лобного щитков (рис. 3.2.11).

Лучше всего выборка может быть представлена вариационным рядом – частотами (p_j) встречаемости в популяции особей с данным (j-м) проявлением качественного признака и общим числом морф (k).

Для более емкого представления ряда используется величина «среднее число фенотипов», учитывающая характер распределения частот между разными морфами: $\mu = \Sigma(p_j)^2$,

статистическая ошибка показателя равна: $m_\mu = \sqrt{\frac{\mu \cdot (k - \mu)}{n}}$.

Среднее число фенотипов (μ) равно числу фенотипов (k) только тогда, когда частоты всех фенотипов одинаковы ($p_1 = p_2 = \ldots = p_j \ldots = p_k$), и меньше во всех других случаях.

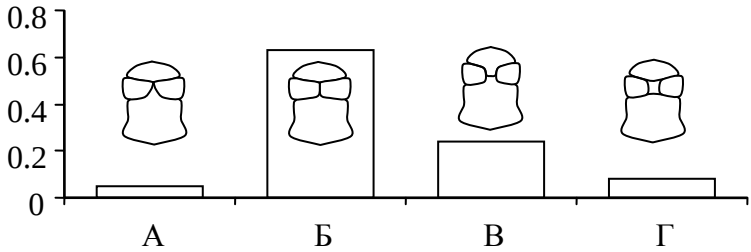


Рис. 3.2.11. Полиномиальное распределение (4 фена головы живородящей ящерицы). По оси ординат – частости фенов среди 64 сеголетков, отловленных под Петрозаводском

Для полиномиального распределения предлагается еще одна характеристика – «доля редких морф»: $h = 1 - \mu \cdot k$, статистическая ошибка показателя равна: $m_h = \sqrt{\frac{h \cdot (1 - h)}{n}}$. Доля редких фенотипов равна нулю при равенстве частот всех морф и отличается от нуля при других вариантах распределения.

Равномерное распределение

Частный случай распределения альтернативного и полиномиального. Равномерное распределение характеризуется одинаковой частотой встречаемости всех значений дискретного призна-

ка ($p = q$ для двух классов или $p_1 = p_2 = \ldots = p_j \ldots = p_k$ для нескольких классов). Такой тип распределения можно использовать для формулирования гипотез при анализе частот фенов в популяциях, при подсчете тест-организмов, выживших в токсикометрическом эксперименте, и т. п. В частности, можно предположить, что ветви дерева могут равномерно располагаться по сторонам света (рис. 3.2.12). Теоретические частоты рассчитываются по общей формуле: $A_x = n \cdot p_x$.

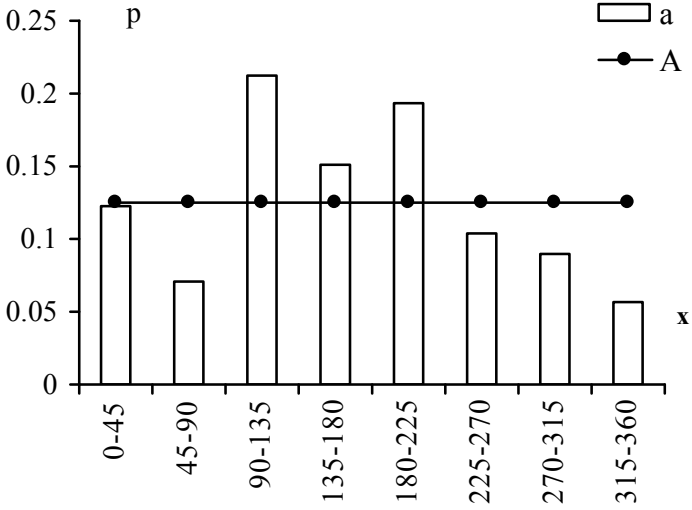


Рис. 3.2.12. Предположительно равномерное распределение числа ветвей ели по секторам азимута (x, º)

Случайное число

Случайными числами пользуются, когда необходима жеребьевка. Например, следует выполнить случайный выбор из нескольких вариантов или нужно формировать ограниченную выборку из более обширной совокупности (для формирования равномерных дисперсионных комплексов) и пр.

Под *случайным числом* обычно понимают безразмерную непрерывную случайную величину, принимающую дробные значения в диапазоне от 0 до 1. Более того, обычно считается, что эта величина имеет равномерное распределение, т. е. вероятность появления любого (i-го) числа равна вероятности появления любого другого (j-го) числа из этого диапазона: $p_i = p_j = \ldots$ Отдельное значение слу-

чайного числа можно получить с помощью функции Excel =СЛЧИС(), которая реализует конгруэнтный метод:

$$z_{i+1}=(a\cdot z_i + b)(\text{mod } c),$$

где z – случайное число в диапазоне 0–1, a, b, c – любые константы, mod c – остаток от деления $(az + b)/ c$ (Нивергельт и др., 1977).

При желании получать свои собственные случайные числа можно применять эту формулу или ее упрощенный аналог: $z_{i+1}=az_i$ (mod c); остаток от деления возвращает функция Excel =ОСТАТ(делимое, делитель).

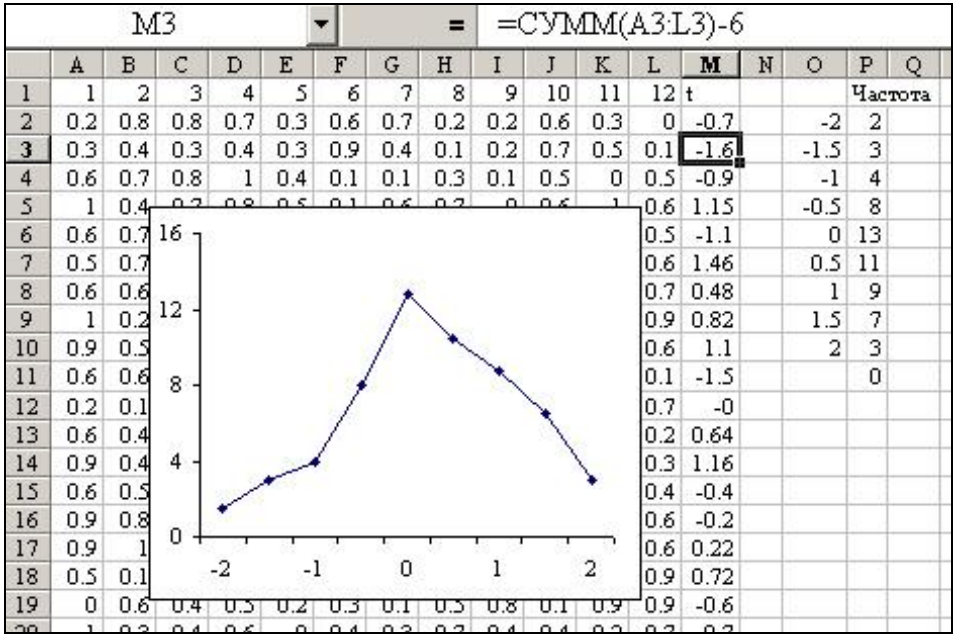
Основная проблема получения «искусственных» случайных чисел состоит в том, что через некоторое количество расчетов они начинают повторяться, т. е. утрачивают требуемое свойство случайности. «Хорошие» случайные числа имеют очень длинный период повторения, их можно получить при таких константах, как $a = 125$, $c = 8192$.

Иногда требуется датчик случайных чисел, имеющих нормальное распределение. Его можно получить на основе равномерно распределенного случайного числа и с использованием следствия центральной предельной теоремы (сумма случайных величин любой природы приближается к величине с нормальным распределением). Величина t аппроксимирует нормальное распределение:

$$t = (1/n)(\Sigma z_i - 0.5)(12n)^{0.5},$$

где z_i – значения случайных величин, распределенных равномерно на интервале [0...1], i=1, 2,... n; t – значение случайной величины, распределенной нормально со средней $M = 0$ и стандартным отклонением $S = 1$.

Чем больше число слагаемых, тем лучше величина z_i подчиняется нормальному закону. Для приблизительных расчетов рекомендуется упрощенная формула всего для двенадцати значений z_i ($n = 12$): $t = \Sigma z_i - 6$, i=1, 2, 3,... 12. Формула не годится для ответственных случаев, поскольку полученная величина очень приблизительно соответствует нормальному распределению (рис. 3.2.13).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	1	2	3	4	5	6	7	8	9	10	11	12	t			Частота	
2	0.2	0.8	0.8	0.7	0.3	0.6	0.7	0.2	0.2	0.6	0.3	0	-0.7		-2	2	
3	0.3	0.4	0.3	0.4	0.3	0.9	0.4	0.1	0.2	0.7	0.5	0.1	-1.6		-1.5	3	
4	0.6	0.7	0.8	1	0.4	0.1	0.1	0.3	0.1	0.5	0	0.5	-0.9		-1	4	
5	1	0.4										0.6	1.15		-0.5	8	
6	0.6	0.7										0.5	-1.1		0	13	
7	0.5	0.7										0.6	1.46		0.5	11	
8	0.6	0.6										0.7	0.48		1	9	
9	1	0.2										0.9	0.82		1.5	7	
10	0.9	0.5										0.6	1.1		2	3	
11	0.6	0.6										0.1	-1.5			0	
12	0.2	0.1										0.7	-0				
13	0.6	0.4										0.2	0.64				
14	0.9	0.4										0.3	1.16				
15	0.6	0.5										0.4	-0.4				
16	0.9	0.8										0.6	-0.2				
17	0.9	1										0.6	0.22				
18	0.5	0.1										0.9	0.72				
19	0	0.6	0.4	0.5	0.2	0.3	0.1	0.5	0.8	0.1	0.9	0.9	-0.6				

Рис. 3.2.13. Опыт создания случайной величины, распределенной нормально (N = 60), основанной на сумме 12 равномерно распределенных случайных чисел (в каждую ячейку блока A1:L60 введена одинаковая формула =СЛЧИС())

Тест на «нормальность»

Рассмотрим пример проверки нормального характера распределения молодых рыжих полевок по длине тела (l); 60 промеров введены на электронный лист StatGraphics. Вызвать программу для оценки соответствия данного распределения нормальному закону позволяет команда Describe \ Numeric Data \ Distribution Fitting. Кнопкой Tabular options открываем одноименное окно и ставим галочку в позиции Test for normality, вызывая окно Test for Normality for l. Кнопкой Graphic options открываем одноименное окно и ставим галочку в позиции Frequency Histogram: на фоне гистограммы строится нормальная кривая (рис. 3.2.14).

```
Tests for Normality for l

Computed Chi-Square goodness-of-fit statistic = 17.3333
P-Value = 0.432013

Shapiro-Wilks W statistic = 0.983416
P-Value = 0.810423

Z score for skewness = 0.265203
P-Value = 0.790849

Z score for kurtosis = 0.896635
P-Value = 0.369912
```

Все четыре рассмотренных критерия отличия от нормальности (χ^2, Вилкса, асимметрии и эксцесса) имеют значения ниже критических уровней (которые в таблице не приводятся), поэтому вероятность соответствия эмпирического распределения нормальному закону (P-Value) в каждом случае превышает критический уровень значимости $\alpha = 0.1$. Вот почему, как написано в последнем абзаце обзора (The StatAdvisor), мы не можем отбросить гипотезу о соответствии распределения длины тела нормальному закону с доверительной вероятностью 90%.

```
   The lowest P-value amongst the tests performed equals 0.369912.
Because the P-value for this test is greater than or equal to 0.10, we
can not reject the idea that l comes from a normal distribution with
90% or higher confidence.
```

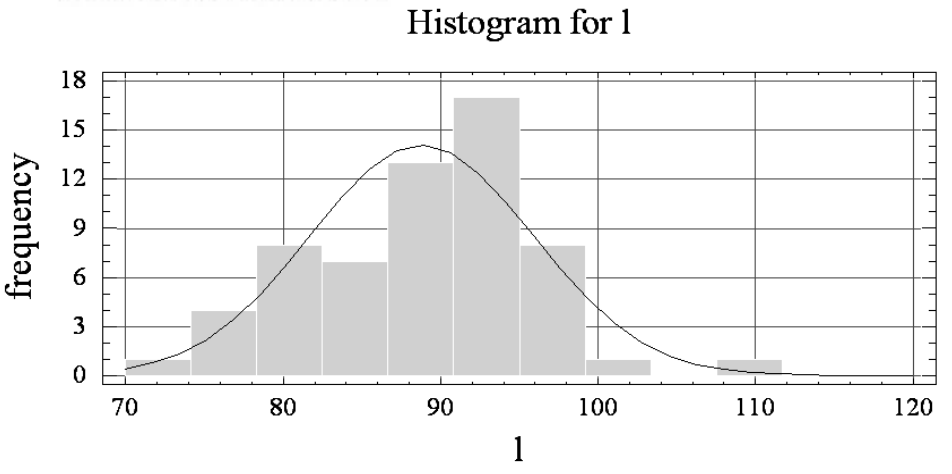


Рис. 3.2.14. Гистограмма частот и нормальная кривая

3.3. Оценка флуктуирующей асимметрии*

Флуктуирующая асимметрия представляет собой незначительные, ненаправленные отклонения от строгой симметрии билатеральных признаков организмов, которые не имеют очевидного адаптивного значения. Считается, что это явление возникает в результате случайностей развития (онтогенетический шум), связанных со спонтанным тепловым движением молекул, которое сравнивают с эффектами зашумления сигналов в теории информации. Основную роль здесь отводят воздействиям среды: условиям экологического пессимума, влиянию различных стрессирующих агентов, включая антропогенные. Изменение уровня флуктуирующей асимметрии может иметь и генетическую подоплеку: отдаленная гибридизация, селекция, отбор (Захаров, 1987).

Формы билатеральной асимметрии

Билатеральная асимметрия обнаруживается как отличия $(L_{ij} - R_{ij})$ в величине *j*-го признака (j = 1, 2… m) на левой (L_{ij}) и на правой (R_{ij}) сторонах тела *i*-й особи (i = 1, 2… n). Выделяют три формы: флуктуирующая, направленная и антисимметрия.

Флуктуирующая асимметрия (ФА) проявляется как нормальное распределение показателя отличия сторон $(L_{ij} - R_{ij})$ с нулевым средним значением $M_{(L_{ij}-R_{ij})} = 0$ (рис. 3.3.1, А). Этот вид изменчивости присущ разнообразным чертам строения организмов и отмечается при других видах асимметрии. Явление флуктуирующей асимметрии отмечено для многих билатеральных характеристик организмов, включая размеры и строение частей скелета и черепа млекопитающих и птиц, метрические и меристические признаки ящериц и рыб, характеристики крыльев насекомых, антенн диптер, губных щупальцев и сифонных сосочков пресноводных моллюсков, признаки вегетативных и генеративных органов сосудистых растений.

Направленная асимметрия характеризуется смещением нормального распределения показателя отличия сторон $(L_{ij} - R_{ij})$

* Раздел написан в соавторстве с А. А. Зориной (Зорина, Коросов, 2007).

(рис. 3.3.1, Б) и не нулевой средней $M_{(L_{ij}-R_{ij})} \neq 0$. В качестве образца направленной асимметрии приводят сердце млекопитающих, лево- и правостороннюю асимметрию в строении тела камболообразных, закрученность раковины у брюхоногих моллюсков и т. д.

Антисимметрии присуще бимодальное распределение разности $\left(L_{ij} - R_{ij}\right)$ или распределение с эксцессом меньше нормального (рис. 3.3.1, В). Экстремальные формы антисимметрии можно наблюдать на примере сигнальных клешней крабов *Uca sp.*, у которых одна клешня намного больше, чем другая, но «правши» и « левши» попадаются примерно с одинаковой частотой у всех видов.

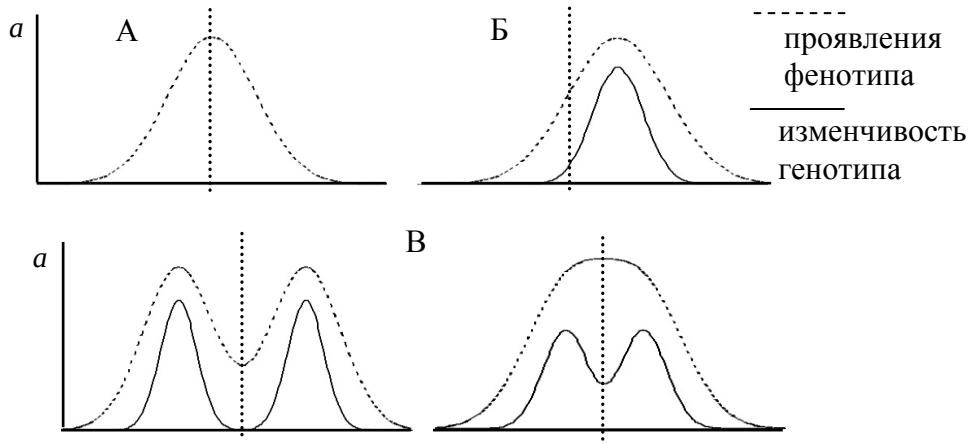


Рис. 3.3.1. Формы билатеральной асимметрии: А – флуктуирующая асимметрия, Б – направленная асимметрия, В – антисимметрия.

Направленная асимметрия и антисимметрия имеют отчетливое адаптивное значение и наследственно детерминированы.

Объединение нескольких форм асимметрий приводит к формированию сложных видов распределения величины $(L_{ij} - R_{ij})$. Для флуктуирующей асимметрии характерно нормальное распределение различий сторон, средняя и эксцесс равны нулю.

Показатели флуктуирующей асимметрии

Оценка флуктуирующей асимметрии, отражающая стабильность развития отдельных организмов, используется для характеристики состояния природных популяций в целом. Считается, что рост

показателей асимметрии этого вида индицирует отклонение параметров внешней среды от оптимальных значений (Захаров, 1987, 2001; Методические рекомендации…, 2003; Гелашвили и др., 2004).

Для практического применения метода предлагаются разнообразные показатели (нам известно около 30 формул), работоспособность и статистические свойства которых пока не получили оценки математиков. Представленные в литературе индексы характеризуют как изменчивость, так и величину (уровень) асимметрии объектов и образуют некую иерархию (табл. 3.3.1).

Таблица 3.3.1. Некоторые оценки флуктуирующей асимметрии

Оценка для особи	Оценка для группы по одному признаку	Интегральный индекс
1 $fa_{ij} = \frac{\lvert L_{ij} - R_{ij} \rvert}{(L_{ij} + R_{ij})}$	$fa_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\lvert L_{ij} - R_{ij} \rvert}{(L_{ij} + R_{ij})}$	$FA = \frac{1}{n}\sum_{i=1}^{n} fa_i$ $fa_i = \frac{1}{m}\sum_{j=1}^{m} \frac{\lvert L_{ij} - R_{ij} \rvert}{(L_{ij} + R_{ij})}$
2 $fa_{ij} = t_{R_{ij}} - t_{L_{ij}}$ $t_{R_{ij}} = (R_{ij} - M_{R_j}) / S_{R_j}$ $t_{L_{ij}} = (L_{ij} - M_{L_j}) / S_{L_j}$	$fa_j = S^2_{fa_{ij}}$	$FA = S^2_{fa_i}$ $fa_i = \frac{1}{m}\sum_{j=1}^{m} fa_{ij}$
3 $fa_{ij} = \frac{2 \cdot L_{ij} \cdot R_{ij}}{L_{ij}^2 + R_{ij}^2}$	$fa_j = 1 - \frac{1}{n}\sum_{i=1}^{n} fa_{ij}$	$FA = 1 - \frac{1}{n}\sum_{i=1}^{n} fa_i$ $fa_i = \frac{2 \cdot \sum_{j=1}^{m} L_{ij} \cdot R_{ij}}{\sum_{j=1}^{m} (L_{ij}^2 + R_{ij}^2)}$
4 $fa_{ij} = b_{ij}$ $b_{ij} = 1$ (L = R); $b_{ij} = 0$ (L ≠ R)	$fa_j = \frac{\sum_{i=1}^{n} b_{ij}}{n}$	$FA = \frac{1}{n \cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m} b_{ij}$

Оценка флуктуирующей асимметрии *у отдельной особи* по *одному* меристическому и пластическому признаку (fa_{ij}) рассчитывается как различие числа структур или промеров на левой и правой сторонах (показатели 1, 2, 3), для качественного признака указывается, симметрично или асимметрично его проявление (показатель 4) (табл. 3.3.1).

Оценка флуктуирующей асимметрии *для группы особей по одному признаку* (fa_j) представляет собой усредненные оценки асимметрии по выборке особей, отрицательный знак для них ликвидирован (берутся абсолютные или квадратичные отклонения).

Интегральный индекс (*FA*), или оценку флуктуирующей асимметрии для выборки особей *по комплексу признаков*, получают в два этапа. Сначала усредняются оценки асимметрии всех m признаков у одной особи (fa_i), затем эти величины усредняются по всем n особям (двойное усреднение).

Групповые и интегральные показатели асимметрии используются при сравнении выборок друг с другом или с условной нормой (оценкой, рассчитанной при нормальном фоновом состоянии особей). Достоверность отличия выборок свидетельствует об существенных отличиях в стабильности развития особей, сформировавших выборки.

Статистические свойства индексов fa_i

Многие индексы флуктуирующей асимметрии преобразуют исходные отличия между левосторонними и правосторонними промерами ($L_{ij} - R_{ij}$) для того, чтобы избавиться от знака разницы и нивелировать масштаб измеряемых признаков. С этой целью учитывают абсолютное значение различия $|L_{ij} - R_{ij}|$, или берут его квадрат $(L_{ij} - R_{ij})^2$, или произведение $L_{ij} \cdot R_{ij}$ и относят эту величину либо к среднему промеру $avg|L_{ij} - R_{ij}|$, либо к сумме промеров ($L_{ij} + R_{ij}$), либо к сумме квадратов $\left(L_{ij}^2 + R_{ij}^2\right)$ и т. п.

По мнению авторов подобных преобразований, это сообщает показателям характер безразмерной величины и позволяет объединять ряд групповых показателей в интегральный индекс путем усреднения.

Такая полуинтуитивная практика «индексотворчества» приводит к одному и тому же результату (рис. 3.3.2) – преобразованные

индексы до такой степени искажают исходно нормальное распределение разности сторон ($L_{ij} - R_{ij}$), что делают невозможным использование параметрических методов статистики для сравнения этих показателей, в том числе по критерию Стьюдента, хотя именно он рекомендуется для сравнения выборок с разных территорий (Методические рекомендации…, 2003). Для этих индексов пока не разработаны методики расчетов статистических ошибок, поэтому сравнение выборок показателей приходится проводить с помощью достаточно грубых непараметрических критериев (Гелашвили и др, 2004), которые оценивают совпадение центров распределений (критерий Уилкоксона–Манна–Уитни) или совпадение частот вариант (критерий Колмогорова–Смирнова).

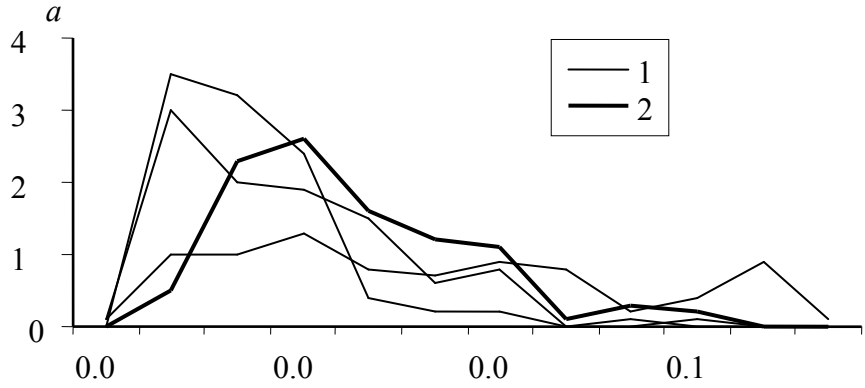


Рис. 3.3.2. Распределения значений $fa_{ij} = \dfrac{\left|L_{ij} - R_{ij}\right|}{\left(L_{ij} + R_{ij}\right)}$ для трех билатеральных признаков листа березы повислой (1) и усредненного интегрального индекса $fa_i = \dfrac{1}{m}\sum_{j=1}^{m}\dfrac{\left|L_{ij} - R_{ij}\right|}{\left(L_{ij} + R_{ij}\right)}$ (2)

На наш взгляд, до появления специальных математических публикаций по этому поводу наиболее оправдано применение традиционных способов нормирования характеристик флуктуирующей асимметрии, например, расчет *нормированного отклонения* (см. п. 2.2); для сравнения таких оценок применим параметрический критерий *F* Фишера (см. ниже).

В качестве критериев величины флуктуирующей асимметрии в литературе предлагаются некие уровни первого показателя *FA*, характеризующие степень приближения к условной норме (Захаров, 2001; Методические рекомендации…, 2003). Диапазон значений, соответствующий фоновому состоянию популяции, принимается за первый балл (условная норма), пятый балл указывает на критическое состояние популяции; еще три уровня соответствуют промежуточным состояниям популяции. Оценить масштаб флуктуирующей асимметрии при разных формах повреждающих факторов позволяют статистические критерии.

Сравнение выборок

Используя разные индексы *FA* промеров листовой пластинки, оценим реакцию березы повислой (Выб1) и березы пушистой (Выб2) на одни и те же условия произрастания. Листья в количестве $n_1 = n_2 = 100$ шт. собирали с брахибластов с нижней части кроны (западная экспозиция) деревьев в возрасте 50–70 лет, произрастающих на заброшенном лугу о. Большой Климецкий (Карелия).

Относительный показатель асимметрии

Для практики фонового мониторинга и оценки последствий антропогенного воздействия рекомендуется первый интегральный индекс (FA_1) из таблицы 3.3.1 (Захаров, 2001; Методические рекомендации…, 2003). Вначале показатель вычисляется для отдельных признаков (рис. 3.3.3). Например, относительные разности промеров для признаков первой особи составили (ячейки I3:K3):

$$\frac{|L_{11} - R_{11}|}{(L_{11} + R_{11})} = \frac{|20.6 - 20.8|}{(20.6 + 20.8)} = 0.005, \quad \frac{|L_{12} - R_{12}|}{(L_{12} + R_{12})} = \frac{|34.5 - 32.7|}{(34.5 + 32.7)} = 0.027.$$

Далее значения усредняются по всем индексам для первой ($i = 1$) особи (ячейка L3):

$$fa1_i = \frac{1}{m}\sum_{j=1}^{m}\frac{|L_{ij} - R_{ij}|}{(L_{ij} + R_{ij})} \quad = \quad fa1_1 = \frac{1}{3}\sum_{j=1}^{3}\frac{|L_{1j} - R_{1j}|}{(L_{1j} + R_{1j})} = 0.046,$$

а затем по всем особям (ячейка L1): $FA_1 = \frac{1}{n}\sum_{i=1}^{n} fa1_i = \frac{1}{100}\sum_{i=1}^{100} fa1_i = 0.050.$

I3 = =(ABS(B3-C3))/(B3+C3)

	A	B	C	D	E	F	G	H	I	J	K	L
1	**m=**	**3**	**n=**	**100**							**FA_1**	**0.0501**
2	**i**	**1L**	**1R**	**2L**	**2R**	**3L**	**3R**		**fa_{i1}**	**fa_{i2}**	**fa_{i3}**	**fa1**
3	**1**	20.6	20.8	34.5	32.7	7.7	6.2		0.005	0.027	0.108	0.047
4	**2**	18.8	17.3	27.9	25.6	5.3	5.9		0.042	0.043	0.054	0.046
5	**3**	14.0	15.0	21.8	22.2	3.0	3.7		0.034	0.009	0.104	0.049
6	**4**	16.7	17.3	27.3	25.2	5.3	5.6		0.018	0.04	0.028	0.028

Рис. 3.3.3. Расчет интегрального индекса для березы повислой (Выб1) по трем признакам: 1) ширина листовой пластинки, 2) длина 2-й жилки второго порядка, 3) расстояние между основаниями 2-й и 3-й жилок второго порядка

Оценки асимметрии объектов-листьев разных видов образовали два множества значений *fa*1 для первой и второй выборок (рис. 3.3.4, блок B3:C103). В силу отмеченных особенностей распределения данных индексов, выполним сравнение выборок разных видов по критерию Уилкоксона–Манна–Уитни (см. п. 4.2).

E4 f_x =РАНГ(B4;$B$4:$C$103;1)

	A	B	C	D	E	F	G	H
1	**FA_1=**	0,0503	0,0498	**U=**	5186	4814	**U_{max}=**	5186
2		**n_1=n_2=**	100	**суммR=**	9864	10236		
3	**fa1**	***Выб1***	***Выб2***		***$R_{выб1}$***	***$R_{выб2}$***	**Тэ=**	0,64
4		0,05	0,04		104	97	**Тт=**	1,98
5		0,05	0,05		106	109		
6		0,05	0,06		111	129		

Рис. 3.3.4. Сравнение разных видов березы по величине *fa*1 с помощью критерия Уилкоксона–Манна–Уитни

С помощью стандартных функций Excel находим ранг каждого значения (блок E3:F103), суммируем их (E2:F2), вычисляем значения *U*, статистику *T* (E1:H1, H3):

$$T_э = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n+1)/12)}} = \frac{5186 - 0.5 \cdot 100 \cdot 100}{\sqrt{(100 \cdot 100 \cdot (100+1)/12)}} = 0.64.$$

Полученное значение $T_э$ = 0.64 меньше табличного (табл. 4С, стр. 351) $t_{(0.1,\infty)}$ = 1.65, следовательно, отличия особей разных видов берез по уровню флуктуирующей асимметрии незначимы.

Нормированный показатель асимметрии

Первый индекс использует нормирование показателя асимметрии признака на величину самого признака. Как известно, такие величины приобретают асимметричные распределения (Шварц и др., 1968). Здесь более уместным было бы использовать базовый способ унификации статистических данных – нормированное отклонение $(x - M) / S$, который реализован при создании второго индекса таблицы 3.3.1 (рис. 3.3.5).

Сначала (рис. 3.3.5, А) для всех эмпирических рядов (n = 100) обоих выборок (3 признака * 2 промера * 2 выборки = 12 рядов, блок B5:N105) вычисляются средние арифметические и стандартные отклонения (M, S, блок A2:N3). Затем отыскиваются нормированные отклонения промеров (рис. 3.3.5. Б; t, блок P5:AB105). Для первого промера слева первого листа имеем (20.6 – 15.1) / 2.1 = 2.6. Далее вычисляются разности между нормированными промерами правой и левой половинок объекта (листа) (рис. 3.3.5, В; блок AD5:AI105). Для первой особи разность между первыми промерами слева и справа составит 2.6 – 2.3 = 0.3. В заключение различия нормированных промеров для всех признаков усредняются по каждой выборке, формируя два ряда значений fa для каждой выборки (блок AJ4:AK105).

Процедура сохраняет нормальное распределение промеров (рис. 3.3.6), однако различие между сторонами при этом не утрачивается. Средние значения индекса отличия сторон для любых выборок равны $fa_1 = fa_2 = 0$, но дисперсии (или стандартные отклонения) представляют собой интегральные показатели асимметрии ($FA = S^2_{fa_i}$), которые можно сравнивать с помощью точного параметрического критерия F Фишера.

А

B2 =СРЗНАЧ(B6:B105)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	m=	3	n=	100										
2	M	15,1	15,2	24,7	24,3	5,4	5,5		16,0	16,5	23,0	23,0	5,5	5,4
3	S	2,1	2,4	3,6	3,4	1,2	1,2		2,2	2,1	2,7	2,5	0,9	0,8
4		*Выб1*							*Выб2*					
5	i	1L	1R	2L	2R	3L	3R		1L	1R	2L	2R	3L	3R
6	1	20,6	20,8	34,5	32,7	7,7	6,2		21,1	20,7	31,3	28,8	6,7	6,4
7	2	18,8	17,3	27,9	25,6	5,3	5,9		18,1	19,7	27,1	27,6	6,2	6,3
8	3	14,0	15,0	21,8	22,2	3,0	3,7		18,1	16,2	26,5	24,4	8,0	5,1

Б

P6 =(B6-B$2)/B$3

	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1														
2							$t(L)=(L-M_L)/S_L$							
3							$t(R)=(L-M_R)/S_R$							
4		*Выб1*							*Выб2*					
5		t(1L)	t(1R)	t(2L)	t(2R)	t(3L)	t(3R)		t(1L)	t(1R)	t(2L)	t(2R)	t(3L)	t(3R)
6		2,6	2,3	2,7	2,5	1,9	0,6		2,3	2,0	3,0	2,3	1,2	1,3
7		1,7	0,9	0,9	0,4	-0,1	0,3		1,0	1,5	1,5	1,8	0,7	1,1
8		-0,5	-0,1	-0,8	-0,6	-1,9	-1,4		0,9	-0,1	1,3	0,6	2,6	-0,3
9		0,7	0,9	0,7	0,3	-0,1	0,1		1,3	-0,1	0,5	-0,1	0,8	-0,3

В

AJ6 = =СРЗНАЧ(AD6:AF6)

	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1							FA_1	FA_2	Fэмп	1.80
2							0.202	0.362	$F_{(0.01)}$ =	1.60
3		t(jL)-t(jR)							$F_{(0.05)}$ =	1.39
4		*Выб1*			*Выб2*		*Выб1*	*Выб2*		
5	1	2	3	1	2	3	fa1	fa2		
6	0.3	0.2	1.3	0.3	0.7	0.0	0.6	0.3		
7	0.9	0.5	-0.4	-0.6	-0.3	-0.4	0.3	-0.4		
8	-0.4	-0.2	-0.5	1.0	0.7	3.0	-0.4	1.6		

Рис. 3.3.5. Сравнение двух видов берез по величине второго интегрального индекса *FA* с помощью критерия Фишера: А – исходные промеры; Б – нормированные отклонения промеров; В – расчет и сравнение показателей флуктуирующей асимметрии

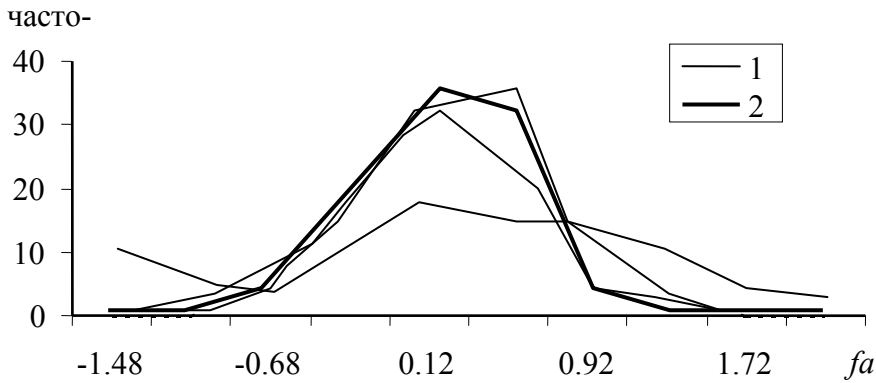


Рис. 3.3.6. Распределение нормированных разностей $(t_{L_{ij}} - t_{R_{ij}})$ трех промеров листа березы повислой (1) и усредненного индекса fa_1 (2)

В примере FA_1 = AJ2 =ДИСП(AJ6:AJ105) = 0.20, FA_2 = 0.36. Критерий Фишера (отношение большей дисперсии к меньшей) равен F = 0.36 / 0.20 = 1.80. Эмпирическое значение критерия (1.80) больше табличного ($F_{(0.01,\ 99,\ 99)}$ = 1.60; =FРАСПОБР(0.01,99,99)), следовательно, дисперсии отличаются значимо. Флуктуирующая асимметрия листьев березы пушистой выше, чем у березы повислой.

Этот вывод противоположен полученному ранее с помощью первого показателя *FA*. Причина, на наш взгляд, состоит в том, что первый метод искажает распределение исходных признаков и вынуждает использовать приблизительные порядковые статистики. Иными словами, вследствие принятой процедуры происходит утрата части информации о специфике сравниваемых выборок. С этих позиций второй метод предпочтительней. Однако его громоздкость заставляет искать иные пути для корректной и лаконичной характеристики флуктуирующей асимметрии.

Глава 4
НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

Планируя использование какого-либо статического метода, предварительно необходимо убедиться, что исходные данные соответствуют модели, лежащей в основе выбранного вида анализа. В частности, в основе многих статистических критериев лежит идея о нормальном распределении изучаемых признаков. По этой причине практически любое биометрическое исследование должно начинаться с установления характера распределения изучаемых признаков. *Распределение* – это соотношение между значениями случайной величины и частотой их встречаемости. Распределение представляет собой не прием обработки данных, это природное явление, закономерность, причинно обусловленное явление бóльшей повторяемости одних значений по сравнению с другими.

Если выборки невелики, то проверить предположение о действии нормального закона нельзя, значит, не появляется уверенности в адекватности применения обычных показателей статистического описания (средней арифметической, стандартного отклонения, коэффициента корреляции) и в правомочности использования обычных параметрических критериев (Стъюдента, Фишера, Пирсона и др.). Дело в том, что параметрические методы оценивания и проверки гипотез резко теряют свою эффективность даже при небольших отклонениях формы распределения от нормального, в частности, они очень чувствительны к сильно уклоняющимся вариантам выборки. Известные рекомендации исключать «чужеродные» значения выполнимы, опять-таки, когда выборки достаточно объемны и соответствуют нормальному распределению. Отбраковка же крайних вариант из небольшой совокупности не оправдана даже по содержательным соображениям – вдруг они являются не результатом методической ошибки, а характеризуют именно биологическую закономерность? Здесь уместно было бы применить методы, мало чувствительные к резким уклонениям (находящимся на периферии рассеяния вариант), но выражающие свойства выборок, ориентируясь на центры распределений. Во всех этих случаях прибегают к помощи *порядковых* статистических показателей.

Вместо средней арифметической, стандартного отклонения, доверительной зоны, критерия различия средних Стьюдента, коэффициента корреляции Пирсона следует определять медиану, медианное стандартное отклонение, медианный критерий различия, медианный коэффициент корреляции (Животовский, 1991).

В этом разделе будут рассмотрены *порядковые* статистики (статистические характеристики), применимые для количественных признаков с неизвестным законом распределения. Для анализа качественных признаков предназначены *ранговые* статистики, использующие не сами значения вариант (x), но их ранги (r_x).

4.1. Порядковые статистики

Расположив наблюдаемые значения в порядке возрастания и перенумеровав, получаем ранжированный ряд. *Ранг* – это номер объекта (значения) в упорядоченном ряду. Если измерения признака выполнялись достаточно грубо или регистрируется счетный показатель, то выборка может содержать повторяющиеся значения; группы (подмножества) одинаковых объектов образуют классы. Сходные объекты получают среднеарифметический ранг, или *мидранг*.

Оценка медианы

В качестве характеристики выборки используется медиана, близкая по смыслу к средней арифметической.

Для рядов с нечетным объемом *медиана* есть значение признака X из середины ранжированного ряда:

$$Me = \mathbf{med}\{x_1, x_2, ..., x_n\} = x_{(n+1)/2},$$

где n – объем выборки, $(n+1)/2$ – средний номер нечетного ряда.

Если n четное, выборочная медиана определяется как среднее арифметическое из двух смежных срединных значений ряда: $Me = \mathrm{med}\{x_1, x_2, ..., x_n\} = [x_{(n/2)} + x_{(1+n)/2)}]/2$.

Например, для выборки из $n = 34$ неполовозрелых прибылых самцов рыжей полевки (Кижский архипелаг, Онежское озеро) масса селезенки варьирует в пределах 32–880 мг (рис. 4.1.1).

Две срединных варианты с номерами $n/2 = 34/2 = 17$ и $1 + n/2 = 1 + 34/2 = 18$ имеют значения $x_{17} = 94$ и $x_{18} = 100$, медиана

равна $Me = [94 + 100]/2 = 97$. Это значение возвращает и функция Excel =МЕДИАНА(A2:A35). Если из выборки исключить чрезмерно большое последнее значение (связанное, предположим, с неточным определением возраста особи), получим немного меньшую оценку медианы (для n = 33) – это будет варианта с номером (n+1)/2 = = (33+1)/2 = 17: $Me = x_{17} = 94$.

G6 = =100*F6/F$24

	A	B	C	D	E	F	G	H	I	J
1	x	ранг	d=20					x-Me	сорт(x-Me)	ранг
2	32	1	Карман					62	0	1
3	35	2		*Карман*	*Частота*	Кумулята	%	59	2	2
4	40	3	30	30	0	0	0	54	4	3
5	40	4	50	50	7	7	21	54	6	4
6	45	5	70	70	2	9	27	49	8	5
7	45	6	90	90	6	15	45	49	9	6
8	50	7	110	110	6	21	64	44	10	7
9	55	8	130	130	0	21	64	39	16	8
10	60	9	150	150	3	24	73	34	18	9
11	72	10	170	170	3	27	82	22	19	10
12	74	11	190	190	2	29	88	20	20	11
13	75	12	210	210	1	30	91	19	22	12
14	76	13	230	230	0	30	91	18	34	13
15	86	14	250	250	1	31	94	8	39	14
16	90	15	270	270	0	31	94	4	39	15
17	92	16	290	290	0	31	94	2	44	16
18	94	17	310	310	0	31	94	0	46	17
19	100	18	330	330	1	32	97	6	49	18
20	103	19	350	350	0	32	97	9	49	19
21	104	20	370	370	0	32	97	10	53	20
22	110	21	390	390	0	32	97	16	54	21
23	133	22	410	410	0	32	97	39	54	22
24	140	23	430	430	1	33	100	46	59	23
25	147	24		Еще	1			53	62	24
26	168	25						74	74	25
27	168	26						74	74	26
28	168	27						74	74	27
29	175	28						81	81	28
30	176	29		Me=	94			82	82	29
31	200	30						106	106	30
32	242	31						148	148	31
33	330	32						236	236	32
34	429	33						335	335	33

Рис. 4.1.1. Ранжированный (A1:B34) и вариационный (D3:E24) ряды массы селезенки прибылых самцов рыжих полевок

Исключение одного (очень большого) значения повлекло за собой уменьшение медианы примерно на 3%. Если же сравнить рассчитанные для этих же рядов средние арифметические величины, получим различие почти в 16%: $M_{n=32} = 119.8$, $M_{n=33} = 142.2$. В отличие от медианы среднее арифметическое значение оперирует со всеми вариантами ряда и сильно реагирует на экстремальные значения, вот почему в нашем примере средняя ($M = 119.8$) существенно превышает медиану ($Me = 97$). Медиана гораздо устойчивее средней арифметической, она в меньшей степени зависит от очень больших и очень маленьких значений в выборке. Дело в том, что медиану можно рассматривать как указатель того места в упорядоченной выборке, слева от которого находятся 50% вариант. Добавление или исключение крайних, самых невероятных, вариант лишь немного смещает этот указатель, который будет показывать на близкие по величине срединные значения ряда. Из этого свойства вытекает графическая интерпретация медианы. Построим вариационный ряд (подсчитаем частоту встречаемости вариант в нескольких интервалах, на которые разбивается диапазон значений изучаемого признака (см.: Ивантер, Коросов, 2003) (рис. 4.1.1, графы Карман, Частоты); в Excel это делается с помощью программы Сервис \ Анализ данных \ Гистограмма. Теперь найдем кумуляту – ряд накопленных частот, выразим их в процентах от объема выборки n и построим график (рис. 4.1.2, графы Кумулята, %). Медиана есть то значение на оси абсцисс, которому соответствует 50% накопленных частот; нормаль, построенная из нее, делит фигуру распределения пополам.

В идеальном случае нормального распределения медиана, средняя арифметическая и мода имеют одно и то же значение.

Стандартная ошибка медианы

Медиана, найденная по выборке, является лишь оценкой некой генеральной медианы изучаемой величины (признака). Для определения области вероятных значений этого параметра, а также для сравнения медиан необходимо рассчитать ошибку репрезентативности медианы. Статистическая ошибка вычисляется по формуле, использующей разность между значениями двух условных вариант, имеющих ранги N_1 и N_2:

$$m = 0.289(x_{N_2} - x_{N_1}),\ N_1 = (n - \sqrt{3n})/2,\ N_2 = (n + \sqrt{3n})/2.$$

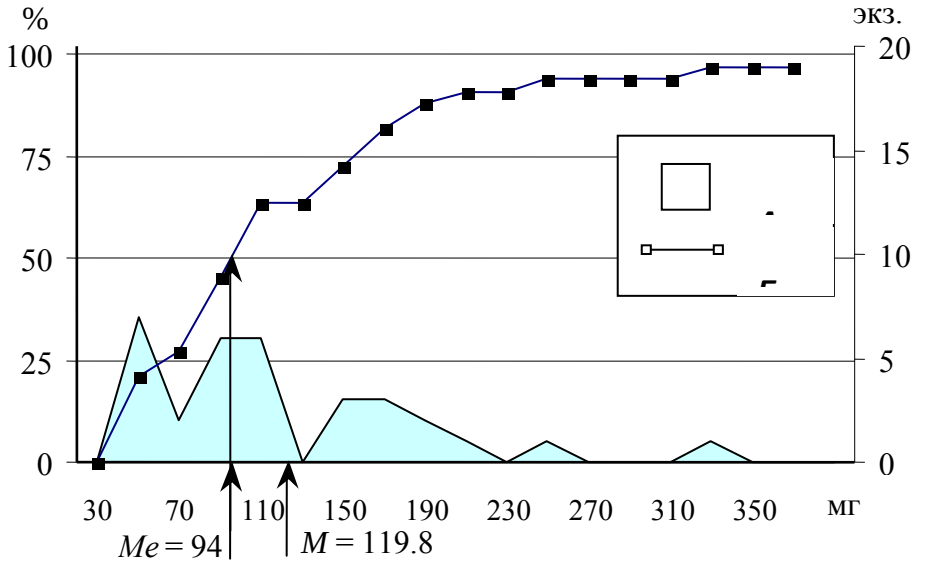


Рис. 4.1.2. Распределение массы селезенки самцов рыжей полевки (А, экз.), кумулята частот этого распределения (Б, % от *n*) и положение на оси абсцисс параметров распределения – медианы (*Me*) и средней арифметической (*M*)

В нашем случае $N_1 = (n - \sqrt{3n}) / 2 = (33 - \sqrt{3 \cdot 33}) / 2 = 11.52$. Ранг 11.52 без дробной части указывает на 11-ю варианту, то есть в качестве первой опорной варианты x_{N_1} нужно брать $x_{11} = 74$. Однако большая дробная часть 0.52 требует учета и соседней варианты $x_{12} = 75$. Их следует учитывать совместно в пропорции этой дроби: $x_{N_1} = 0.48 \cdot x_{11} + 0.52 \cdot x_{12} = 0.48 \cdot 74 + 0.52 \cdot 75 = 74.5$. Ранг второй опорной варианты равен $N_2 = 21.47$. Дробная часть также велика, поэтому для расчетов x_{N_2} вновь используем две варианты $x_{21} = 110$ и $x_{22} = 133$: $x_{N_2} = 0.53 \cdot x_{21} + 0.47 \cdot x_{22} = 0.53 \cdot 110 + 0.47 \cdot 133 = 120.8$.

Ошибка равна $m = 0.289 \cdot (120.8 - 75.5) = 0.289 \cdot 46.3 = 13.4$ (мг). Расчетные ранги можно округлять $N_1 = 11.52 \approx 12$, $N_2 = 21.47 \approx 21$ мг (им соответствуют величины $x_{12} = 75$ и $x_{21} = 110$ мг). Теперь ошибка менее точна: $m = 0.289 \cdot (110 - 75) = 0.289 \cdot 35 = 10.1$ (мг).

Доверительный интервал для генерального значения медианы μ

Приблизительные значения доверительного интервала оцениваются по стандартной формуле: $Me \pm t \cdot m$, где t – табличное значение критерия Стьюдента (табл. 4С, стр. 350) для $df = n - 1$. У нас $df = 33 - 1 = 32$, $t_{(0.05,\,32)} = 2.04$, $m = 13.4$ мг, $Me = 94$ мг. Генеральная медиана находится в диапазоне $94 \pm 2.04 \cdot 13.4 = 94 \pm 27.3$, то есть от 66.6 до 121.4 мг.

Более точные границы доверительного интервала дают варианты с расчетными рангами h и g: $x_h < \mu < x_g$; $h = (n - t_н\sqrt{n} - 1)/2$, $g = (n - h + 1)$, $t_н$ – нормально распределенная величина (для принятого уровня значимости: для $\alpha = 0.05$ $t_н = 1.96$).

В примере ранги равны: $h = (33 - 1.96\sqrt{33} - 1)/2 = 10.3 \approx 10$ и $g = (n - h + 1) = (33 - 10 + 1) = 24$, точные границы доверительного интервала составляют: $x_{10} = 72$, $x_{24} = 147$. Генеральная медиана находится в диапазоне от 72 до 147 мг.

Оценка стандартного отклонения

Медианное стандартное отклонение определяется по формуле $S_{Me} = 1.481 \cdot$ **med**$[(x_1 - Me), (x_2 - Me), \ldots, (x_N - Me)]$. Вычисления организуются на листе Excel (рис. 4.1.1). Вначале находим все отклонения вариант от медианы (колонка H), затем копируем этот столбец (H2:H34) в буфер обмена и вставляем с помощью Специальной вставки как Значения в блок (I2:I34), сортируем в порядке возрастания и отыскиваем срединное значений (его ранг равен 17): $x_{17} = 46$ мг. Можно применить и функцию =МЕДИАНА(H2:H34). Медианное стандартное отклонение составит $S_{Me} = 1.481 \cdot 46 = 68.13$ мг. Стандартное отклонение, рассчитанное обычным способом, равно $S = 86.6$ мг. Большое отличие величин (на 18.5 мг, 21%) оказалось вызвано всего лишь присутствием варианты $x_{34} = 429$ мг. Исключив ее из расчетов получаем значение $S = 67.5$ мг. Это лишний раз показывает, что пользуясь медианной мерой изменчивости, можно получать адекватные оценки рассеяния, мало зависящие от особенных «выскакивающих» вариант выборки.

4.2. Непараметрические критерии различия

Для ориентировочной оценки расхождений между двумя небольшими выборками, имеющих неопределенный тип распределения количественных и качественных признаков, служат *порядковые* (или непараметрические) *критерии*, ориентированные на исследование соотношений *рангов* исходных значений вариант. Конструкции таких критериев отличаются простотой. Процедура состоит из трех этапов – упорядочивание и ранжирование вариант, подсчет сумм рангов в соответствии с правилами данного критерия, сравнение полученной величины с табличным значением критерия. Нулевая гипотеза гласит: «характер распределения вариант сходен». Исследуется вопрос, насколько равномерно варианты разных выборок «перемешаны» между собой. Если они более или менее регулярно чередуются в общем упорядоченном ряду, значит, распределены сходным образом и отличий между совокупностями (и их медианами) нет. Если же выборки пересекаются не полно (смешиваются только краями распределений, либо одна выборка поглощает другую), то становится ясно, что эти выборки взяты из разных генеральных совокупностей – со смещенными центрами или разными дисперсиями. Среди множества известных методов можно выделить критерий Уилкоксона–Манна–Уитни, ориентированный на сравнение центральных тенденций (медиан), то есть отвечающий на вопрос, существенно ли выборки смещены друг относительно друга.

Критерий U Уилкоксона–Манна–Уитни

Техника метода состоит в том, что все варианты сравниваемых совокупностей ранжируют в одном *общем ряду*: каждому значению присваивают ранг, порядковый номер. При этом одинаковым (повторяющимся) значениям вариант должен соответствовать один и тот же средний ранг. Затем ранги вариант суммируют отдельно по каждой выборке: $R_1 = \Sigma r_i$, $R_2 = \Sigma r_j$, i = 1, 2, …, n_1, i = 1, 2, …, n_2 и вычисляют критерий: $T = \dfrac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n+1)/12)}}$, где $U = \max(U_1, U_2)$ – максимальное значение из двух величин:

$U_1 = n_1 \cdot n_2 + 0.5 \cdot n_1(n_1 + 1) - R_1$, $U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2(n_2 + 1) - R_2$,

n – объем выборки с максимальной суммой рангов U.

Если выборка достаточно велика ($n > 20$), величина статистики T сравнивается с табличным значением критерия Стьюдента (табл. 4С, стр. 351) для $df = \infty$ и $\alpha = 0.1$ (т. е. только *для верхней* 95% области распределения одностороннего критерий Стьюдента). При $n < 20$ нужно пользоваться таблицами Уилкоксона–Манна–Уитни (табл. 7С, стр. 354). Метод хорошо работает при $n > 10$.

Сравнивали 5- и 35-дневных щенков песцов по активности фермента каталазы в сердце (E): 5-дневные: 41, 44, 31, 38, 43, 29, 71, 45; $n_1 = 8$; 35-дневные: 52, 51, 62, 52, 52, 50, 54, 62, 31; $n_2 = 9$. Поскольку биохимические показатели, как правило, не подчиняются нормальному закону, воспользуемся непараметрическим критерием. Упорядочим по возрастанию все варианты выборок вместе, но так, чтобы значения каждой выборки располагались в двух отдельных рядах (E_5, E_{35}) (табл. 4.2.1). Такое расположение упрощает назначение рангов (ряды r_5, r_{35}) и их суммирование (R).

Таблица 4.2.1. Совместное ранжирование двух выборок

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	*R*
E_5	29		31	38	41	43	44	45									71	
E_{35}		31							50	51	52	52	52	54	62	62		
r_5	1		2.5	4	5	6	7	8									17	50.5
r_{35}		2.5							9	10	12	12	12	14	15.5	15.5		102.5

Далее вычисляем вспомогательные величины и значение критерия T: $U_1 = 9 \cdot 8 + 0.5 \cdot 9 \cdot (9+1) - 50.5 = 66.5$,

$$U_2 = 9 \cdot 8 + 0.5 \cdot 8 \cdot (8+1) - 102.5 = 5.5,$$

$$U = \max(U_1, U_2) = 66.5,\ T = \frac{66.5 - 0.5 \cdot 9 \cdot 8}{\sqrt{(9 \cdot 8 \cdot 10/12)}} = 3.81.$$

Полученное значение (3.81) больше табличного ($t_{(0.1,\,\infty)} = 1.65$; табл. 4С), отличия между выборками достоверны, т. е. активность каталазы с возрастом меняется. Раз выборки малы, воспользуемся точными таблицами (табл. 8С). Получаем $T_{(0.05,\,n1,\,n2)} = T_{(0.05,\,8,\,9)} = 51$. Расчетное значение (66.5) больше табличного (51), следовательно, различия между выборками достоверны.

Непараметрический однофакторный дисперсионный анализ

Когда необходимо сравнить несколько небольших выборок, применяют схему непараметрического дисперсионного анализа. Нулевая гипотеза состоит в том, что выборочные распределения одинаковы, то есть выборки взяты из одной генеральной совокупности.

Изучали плодовитость дафний ($n = 14$) в 4 градациях фактора: чистая вода (градация $A1$; $x = 6, 5, 5, 7$), слабая (5 мг/л, $A2$; 8, 7, 6, 6), средняя (15 мг/л, $A3$; 8, 8, 7) и сильная концентрация токсиканта (30 мг/л, $A4$; 8, 7, 9).

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

После упорядочивания, варианты ранжируются (п. 2.2).

Градация	1	1	1	2	2	1	2	3	4	2	3	3	4	4
Значение	5	5	6	6	6	7	7	7	7	8	8	8	8	9
Ранг	1.5	1.5	4	4	4	7.5	7.5	7.5	7.5	11.5	11.5	11.5	11.5	14

Разносим ранги по градациям, подсчитаем суммы и отношения.

Градация, i	1				2				3				4			
Значение	5	5	6	7	6	6	7	8	7	8	8		7	8	9	
Ранг, R	1.5	1.5	4	7.5	4	4	7.5	11.5	7.5	11.5	11.5		7.5	11.5	14	
Сумма, R_i				14.5				27				30.5				33
n_i				4				4				3				3
R_i^2 / n_i				52.56				182.3				310.1				363

Критерий равен:

$$H = \frac{12}{n \cdot (n-1)} \cdot \left(\frac{R_1^{\,2}}{n_1} + \ldots + \frac{R_j^{\,2}}{n_j} + \ldots + \frac{R_k^{\,2}}{n_k} \right) - 3 \cdot (n+1) =$$

$$= \frac{12}{14 \cdot 13} \cdot (52.56 + 182.3 + 310.1 + 363) - 3 \cdot 15 = 14.86$$

При $n > 5$ статистика H имеет распределение хи-квадрат ($df = k - 1$). Полученное значение критерия (14.86) больше табличного ($\chi^2_{(0.05,3)} = 7.81$), отличие выборочных распределений достоверно. Химическая добавка действительно увеличивает плодовитость дафний.

Глава 5
ИЗУЧЕНИЕ БИОРАЗНООБРАЗИЯ

Понятие биоразнообразия традиционно связывают с характеристиками видового разнообразия – числом сходных в экологическом отношении видов, населяющих данную территорию, и их численностью. Здесь возникают методические задачи определения видового богатства и выравненности территориальных группировок животных и растений (таксоценозов), решение которых подробно рассмотрено ниже. Вместе с тем этот богатый арсенал методов позволяет эффективно исследовать и внутривидовое биоразнообразие, рассматривая вместо видов дискретные хорошо различимые морфы. В общем случае исследования полиморфизма популяций такая подмена оправдана, однако для изучения генетической структуры популяций требуются специальные методы, которые в этом пособии не рассмотрены.

5.1. Видовое богатство: α-разнообразие

Состав видовых списков населения (фауны, флоры) данной территории обозначается как *альфа*-разнообразие. *Видовое богатство* – это число видов, обитающих на данной территории. Основные сложности получения таких характеристик связаны с обычной для биоэкологии невозможностью проведения тотальных учетов, поэтому число видов в выборках всегда меньше, чем реально живет в изучаемом районе. По мере расширения исследований видовой список пополняется, хотя и с постепенно снижающейся скоростью. Это явление наблюдается при увеличении продолжительности исследований и росте числа изученных проб (рис. 5.1.1), при увеличении исследуемых территорий (рис. 5.1.2), на сериях разноразмерных изолированных ареалов (озера, острова, города) (рис. 5.1.3). Динамику наполнения списка видов удобно аппроксимировать параболическими или логарифмическими уравнениями (например, при увеличении изучаемой площади имеем $s = a\log(S) + b$, где S – площадь обследованной территории, s – число найденных видов). На графике насыщения можно выделить две области – начальный быстрый рост

числа выявленных видов и последующее гораздо более плавное повышение.

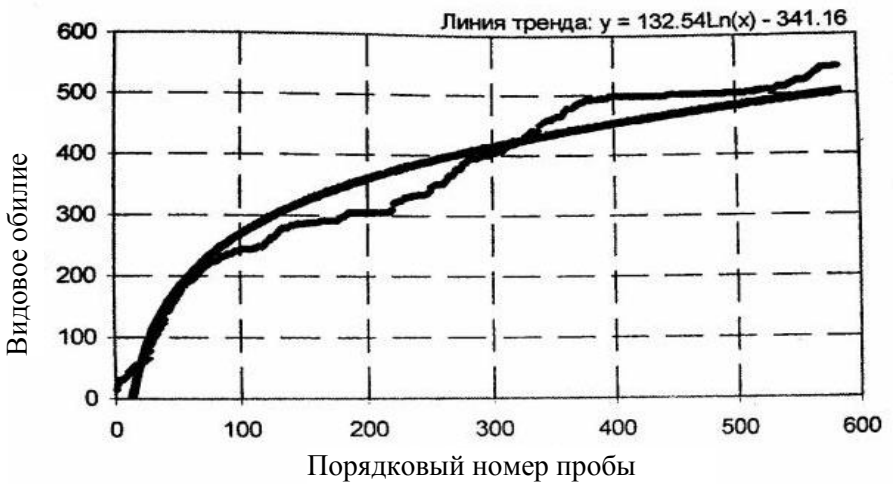


Рис. 5.1.1. Рост видового списка зообентоса малых рек Самарской области (Шитиков и др., 2003)

Форма представленной кривой определяется двумя родами факторов – статистическими и биологическими. Если бы разные виды имели равные численности, а их представители равномерно и пропорционально покрывали поверхность земли, то насыщение списка, например, при маршрутном учете произошло бы очень быстро и прямо пропорционально длине пройденного маршрута; этой схеме в определенной мере соответствует начальная крутая часть кривой насыщения, имеющая почти линейную форму. В реальности особи разных видов распределены в пространстве не равномерно и к тому же имеют разную численность. В силу этого скорость наполнения видового списка замедля-

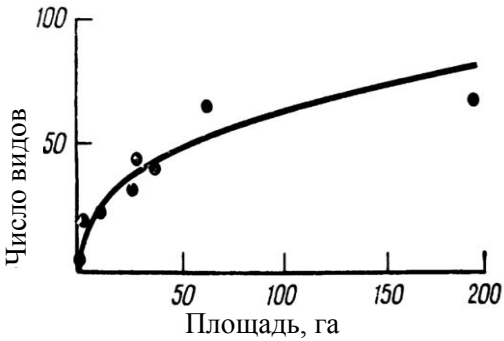


Рис. 5.1.2. Зависимость числа видов гнездящихся птиц от площади участков листопадного леса

ется и чем дальше, тем больше; график насыщения искривляется и даже при длительных наблюдениях (при больших объемах материала) может не давать полной картины биологического разнообразия.

Опыт показывает, что нельзя давать оценку видовому богатству и выравненности населения данной территории, используя единичную пробу, одну коллекцию. Проводя исследования по оценке биоразнообразия, необходимо строить кривую насыщения видового списка по мере увеличения объема работы (продолжительности наблюдений, исследованной площади, величины сборов, числа маршрутных учетов и пр.). Для расчета показателей биоразнообразия использовать только такие описания, число видов которых уже находится на слабо наклоненной части кривой насыщения; например, на рис. 5.1.3 эта точка соответствует 10 000 км^2 обследованной территории.

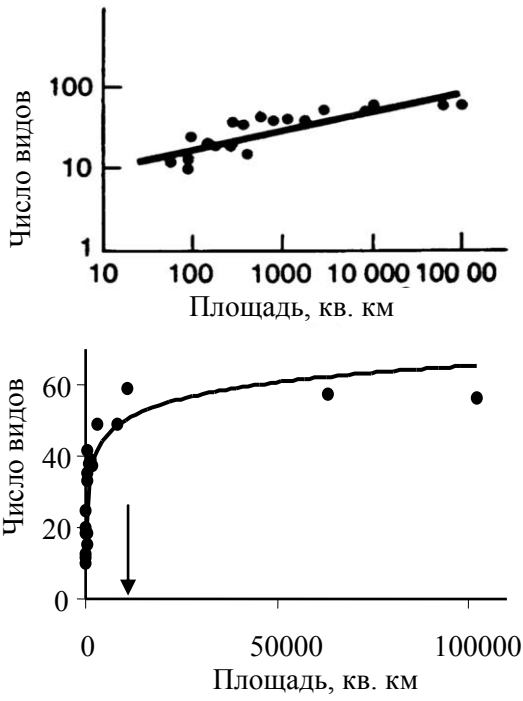


Рис. 5.1.3. Зависимость числа видов гнездящихся птиц от площади островов в логарифмическом и натуральном масштабе (Уилкокс, 1983)

Сравнение кривых насыщения видовых списков

Иногда возникает необходимость сопоставить между собой по видовому богатству территории, ограниченные естественными границами, но имеющие разную площадь (острова, города). При прочих равных условиях насыщенность видами должна быть выше у наиболее крупного выдела в силу отмеченного выше явления роста списка видов при увеличении обследуемой территории. Но кроме этой «механической» причины возможны и другие (биологические, антропогенные), которые оказывают влияние на видовое разнообра-

зие. Поэтому прямое сопоставление видовых списков для разноразмерных территорий не даст возможности их содержательно интерпретировать – сравнивать следует сравнимые объекты (площади).

Интереснее всего сопоставлять друг с другом две кривые насыщения видовых списков, составленные для сравниваемых территорий (рис. 5.1.4). Теоретически мыслятся три возможные ситуации: на меньшей территории динамика такая же (А), либо ускоренная (число видов относительно больше, Б), либо замедленная (видов относительно меньше, В).

Второй случай соответствует ситуации, когда на территории действуют факторы, вызывающие агрегирование видов. Это могут быть какие-либо особенно благоприятные условия (в городе, например, для многих видов растений микроклиматические и трофические условия лучше, чем в окрестной природе) или изолирующие механизмы (водные границы островов, препятствующие свободному перемещению как самих животных, так и их хищников).

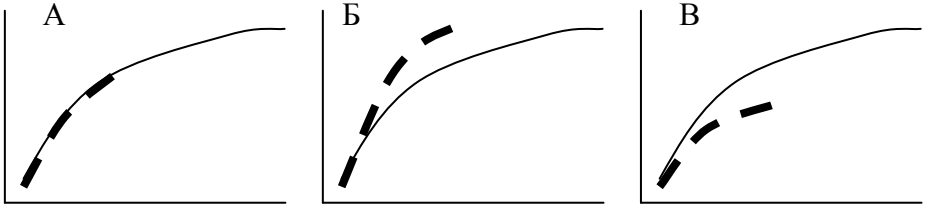


Рис. 5.1.4. Варианты кривых насыщения при увеличении обследованной площади

В контексте нашего изложения следует дать важную методическую рекомендацию. При изучении видового богатства сообществ научный интерес имеет не только представление результата исследования (составленные видовые списки), но и регистрация хода процесса исследования (например, раздельный учет всех маршрутов или отбора проб на площадках), который даст возможность построить кривые насыщения, служащие основой для репрезентативного сравнения и содержательного обсуждения найденных отличий в составе флоры или фауны.

Есть другой выход из этой ситуации. Для большей территории можно построить кривую насыщения видового списка, имитируя отбор проб со все увеличивающихся площадок (вплоть до раз-

мера полного крупного выдела). Тогда видовой список меньшего выдела (площадью S_l) может быть сопоставлен с тем отрезком кривой насыщения крупного контура (площадью S_b), который соответствует равной с ним площади.

Рассмотрим задачу сравнения флористического состава пяти городов Карелии (Антипина, 2002): Петрозаводск (изучена территория размером 20 км^2), Костомукша (2.6), Пудож (4.9), Олонец (6.1), Медвежьегорск (7.5), Сегежа (11.5). В Петрозаводске получены видовые списки растений отдельно для девяти районов: Ключевая (площадь 2.3 км^2), Соломенное (2.4), Сулажгора (1), Октябрьский (2.8), Центр (3), Старая Кукковка (1), Новая Кукковка (1.8), Перевалка (2.5), Древлянка (2).

На первом этапе анализа были составлены несколько условных кривых насыщения видового списка города Петрозаводска, выражающих процесс последовательного исследования района за районом и прибавление новых видов растений к первоначальному списку некоего исходного района. Начнем, например, с Ключевой. Здесь на площади 2.3 км^2 найдено 292 вида растений. В Соломенном (2.4) обнаружено 300 видов, из которых 61 вид не найден на Ключевой. Значит, на общей площади 4.7 км^2 найдено уже 353 вида. В Сулажгоре (1 км^2) найдено 272 вида, в том числе 39 новых; получаем 5.6 км^2 и 392 вида. Прибавляя районы и объединяя списки, получаем рост числа видов на растущей исследованной территории (рис. 5.1.5). Можно заметить, что полученная кривая находится на стадии явного выполаживания, что свидетельствует о тщательном исследовании флористического состава и близком пределе объема видового списка. Если менять порядок прибавления районов, полученные кривые видового насыщения будет разниться. На диаграмме отображены 10 рядов, полученных случайными перестановками исследованных районов г. Петрозаводска. Теперь можно провести регрессионный анализ, построить для всех рядов усредненную кривую логарифмической функции (получилось $s = 120\log(S) + 208$, $p < 0.01$), доверительную зону области возможных значений, а также добавить точки, соответствующие видовым спискам сравниваемых городов. Учитывая, что изучаемые видовые списки не лишены случайной изменчивости и могут иметь сходные по величине доверительные интервалы, можно считать, что флоры всех городов, кроме Сегежи, вполне удовлетворительно соответствуют кривой насыще-

ния видового богатства города Петрозаводска, то есть отличиу флоры этих городов связано лишь с их разной площадью. В г. Сегежа, напротив, видов существенно меньше, чем можно было бы ожидать теоретически. Поскольку этот город расположен много севернее остальных, налицо влияние климатической зональности.

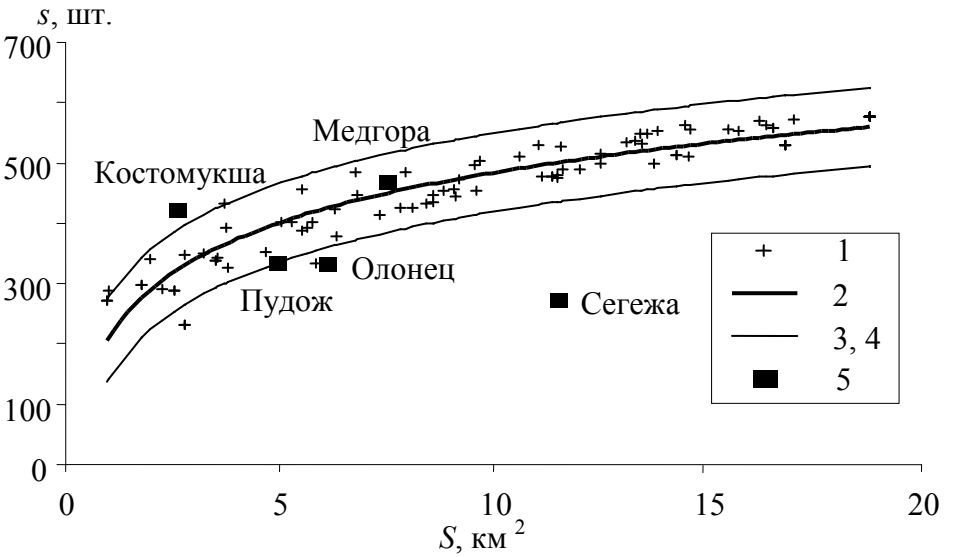


Рис. 5.1.5. Рост числа видов высших растений (*s*) по мере увеличения обследованной площади (*S*): 1 – число видов, зарегистрированных в разных районах г. Петрозаводска, 2, 3, 4 – логарифмическая регрессия и границы доверительной зоны ($P = 0.95$), 5 – положение видовых списков разных городов (занимающих разную площадь)

5.2. Видовое богатство: β-разнообразие

Изменение видового богатства по градиенту местообитаний обозначается как *бета*-разнообразие. Из числа наиболее общих случаев исследований такого рода можно назвать сравнение видовых списков двух или нескольких территорий, рассмотрение динамики изменения разнообразия во времени в связи с естественным развитием биосистем или при антропогенных нагрузках, изменение биоразнообразия вдоль по градиенту экологического фактора, территориальное распространение показателей видового богатства.

Показатели сходства видового богатства

В качестве объекта исследования возьмем биоценотические группировки мелких млекопитающих Прибайкальской равнины (Южное Прибайкалье) (Коросов, Демидович, 1987). Видовой состав изучен в 7 основных биотопах; это объекты, им соответствуют строки таблицы данных (табл. 5.2.1).

Таблица 5.2.1 Встречаемость разных видов мелких млекопитающих в разных биотопах Южного Прибайкалья в годы пика численности

Биотоп	Крот сибирский	Обыкновенная бурозубка	Средняя бурозубка	Малая бурозубка	Равнозубая бурозубка	Крошечная бурозубка	Водяная кутора	Лесная мышовка	Крыса серая	Крыса черная	Мышь домовая	*Восточно-азиатская мышь*	Лесной лемминг	Полевка обыкновенная	Полевка темная	Полевка-экономка	Полевка красная	Полевка красно-серая	Всего видов
	1	1	1	1	1	1	0	1	0	0	0	1	1	0	0	0	1	1	11
П	1	1	1	1	1	0	1	1	0	0	0	1	1	0	1	1	1	1	13
Э	1	1	1	1	1	0	0	1	0	0	0	1	1	0	1	1	1	1	12
С	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1	8
Б	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	16
Л	1	1	1	1	1	0	0	1	1	1	0	1	0	1	1	1	1	1	14
Г	0	1	1	1	0	0	0	0	1	1	1	1	0	1	1	0	1	0	10
Всего																			18

Биотопы представляют собой серию коренных и трансформированных фитоценозов: кедровник (К), пойменные пихтово-еловые смешанные леса (П), мозаичные смешанные хвойно-лиственные леса на границе с коренными лесами (с локальными вырубками, гарями, дорогами, отвалами; экотон, Э), вторичные сосня-

ки (С), вторичные березняки (Б), производные суходольные луга (Л), территория г. Байкальска (Г). Встречаемость 18 видов насекомоядных и грызунов представлена в таблице 5.2.1 (это показатели, им соответствуют колонки).

Сравнение двух территорий по видовым спискам возможно только при условии равной репрезентативности исследований. Если объемы работ (продолжительность исследований, число проб, площади участков) различаются для сравниваемых территорий, мы неизбежно получим *ложные* различия (п. 5.1), связанные не с действительными биотическими особенностями сообществ, но с методом их изучения («d – эффект»). Для случаев выполнения условия равной репрезентативности предложен ряд разнообразных мер сходства и различия между двумя видовыми списками. Отношения между ними соответствуют схеме пересечения двух множеств – **X** и **Y,** отображенных в *четырехпольной таблице* (табл. 5.2.2).

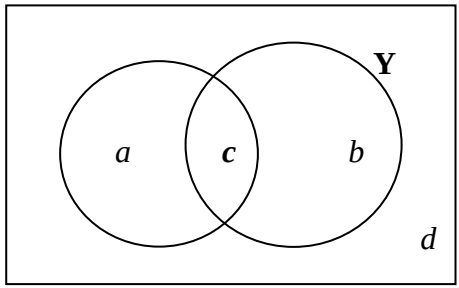


Таблица 5.2.2. Пересечения видовых списков двух территорий

c Виды, общие для обоих списков	*b* Виды из списка **Y**, отсутствующие в **X**	***c + b*** Виды из списка **Y**
a Виды из списка **X**, отсутствующие в **Y**,	*d* Виды, отсутствующие в обоих списках	*a + d* Виды, отсутствующие в списке **Y**
c + a Виды из списка **X**	*d + b* Виды, отсутствующие в списке **X**	***s*** *= a + b + c + d* Общее число видов в регионе

В литературе показано (Шитиков и др., 2003), что большинство *мер сходства* видовых списков могут быть сведены к обобщающей формуле: $C = \dfrac{2c}{(1+u)(a+b+2c)-2uc}$.

При $u = 0$ имеем *коэффициент общности Съёренсена* (отношение удвоенного числа общих видов к сумме видов в двух списках): $C = \frac{2c}{(a+c)+(b+c)}$.

При $u = 1$ после преобразований получаем *коэффициент Жаккара* (отношение числа общих видов к сумме видовых списков без числа общих): $C = \frac{c}{(a+b-c)}$ и т. д. (Миркин и др., 1989; Лебедева и др., 2004).

Большинство известных коэффициентов отличаются друг от друга лишь на величину некого постоянного множителя, то есть информативно равнозначны и не могут «лучше» или «хуже» выполнять роль меры сходства видовых списков.

Из коэффициентов сходства двух списков нетрудно сделать *меру расстояния* между ними: $d = 1 - C$.

В нашем примере данные таблицы 5.2.1 позволили определить число общих видов для всех пар биотопов и рассчитать матрицу расстояний Съёренсена между разными биотопами (табл. 5.2.3).

Таблица 5.2.3. Число общих видов (слева внизу) и различия по Съёренсену (%) между биотопическими группировками (справа вверху)

		К	**П**	**Э**	**С**	**Б**	**Л**	**Г**
	s	**11**	**13**	**12**	**8**	**16**	**14**	**10**
К	**11**		17	13	16	26	28	52
П	**13**	10		4	24	10	19	48
Э	**12**	10	12		20	14	15	45
С	**8**	8	8	8		33	36	44
Б	**16**	10	13	12	8		7	31
Л	**14**	9	11	11	7	14		25
Г	**10**	5	6	6	5	9	9	

Например, в экотоне (Э) отмечено 12 видов мелких млекопитающих, а на лугах (Л) – 14; из них общих было 11. Мера Съёренсе-

на составит: $C_{ЭЛ} = \frac{2c}{(a+c)+(b+c)} = \frac{2 \cdot 11}{12+14} = 0.85$, а расстояние $d_{ЭЛ} = 100 \cdot (1 - 0.85) = 15$ %.

К сожалению, статистическую ошибку показателя определить нельзя. Простой факт наличия вида в списке не позволяет сделать статистический прогноз: для вероятностной оценки нужны данные по его обилию (см. п. 5.3, Индексы видового богатства).

Корреляционная мера сходства списков

При сравнении видовых списков (или спектров фенотипов) двух локальных территорий, как правило, остается неизвестным число видов, не попавших ни в один список, но обитающих в регионе (значение *d* из четырехпольной таблицы). Если же сравниваются два локальных видовых описания в пределах хорошо изученной территории или когда сравниваются описания, выполненные в отдельные годы на фоне многолетних исследований (т. е. информация об отсутствующих видах есть), то возможно применение различных корреляционных мер сходства, или *мер связи*. Один из таких показателей – коэффициент корреляции Пирсона (коэффициент Бравэ):

$$r = \frac{(cd - ba)}{\sqrt{(c+b) \cdot (c+a) \cdot (b+d) \cdot (a+d)}} = \sqrt{\chi^2 / (a+b+c+d)}\,.$$

Если объемы проб велики, возможен расчет ошибки коэффициента корреляции и оценка его значимости по критерию Стьюдента $n = s$: $m_r = \sqrt{(1-r^2)/(n-2)}$, $t = r / m_r$ (табл. 9С, стр. 354).

Сравним видовые списки млекопитающих, обнаруженных в кедровнике (К, 11 видов) и в березняках (Б, 16 видов), представленных в форме четырехпольной таблицы:

c	*b*	***c* + *b***		**10**	1	**11**
a	*d*	*a* + *d*	=	6	8	14
c* + *a	*d* + *b*	***s***		**16**	9	**18**

Помимо 17 видов, обитающих в обоих биотопах (10 общих), на прибайкальской равнине отловлен еще 1 вид. Отсюда коэффициент Бравэ составит: $r_{КБ} = \frac{(10 \cdot 8 - 6 \cdot 1)}{\sqrt{11 \cdot 16 \cdot 9 \cdot 14}} = \frac{74}{149} = 0.497 \approx 0.5$.

Выполнив такие же расчеты относительно прочих биотопов, получаем массив корреляционных мер сходства (табл. 5.2.4). Дополнение коэффициента корреляции до единицы составляет показатель различия: $d = 1 - r$. Эти характеристики используются для составления различных деревьев.

Таблица 5.2.4. Корреляционная мера сходства видовых списков (выделены наибольшие значения, на основании которых будет построено дерево на рис. 5.2.4)

К	0.65	**0.72**	**0.75**	0.5	0.46	0.3
	П	**0.89**	0.64	0.71	0.55	0.34
		Э	**0.69**	**0.67**	0.64	0.38
			С	0.53	0.45	0.48
				Б	**0.76**	0.6
					Л	**0.69**
						Г

Важно отметить, что значения 0 и 1 относятся к дихотомической шкале (Михеев, 2004), которая может представлять любой вид шкал, в том числе относительную, то есть ее значения есть числа и поэтому могут обрабатываться методами статистики. Значит, для оценки сходства двух рядов можно применить коэффициент корреляции Пирсона (табл. 5.2.5).

Таблица 5.2.5 Корреляция между видовыми списками биотопов

К	0.52	0.64	0.71	0.08	0.12	-0.3
	П	0.88	0.55	0.57	0.27	-0.3
		Э	0.63	0.5	0.47	-0.2
			С	0.32	0.21	0.13
				Б	0.66	0.04
					Л	0.33
						Г

Дендрограмма, коррелограмма

Для графического представления разобщенности или, напротив, сходства территорий по видовому составу изучаемых групп разработаны особые дендроидные структуры, или графы, – серия точек (узлов, вершин), соединенных линиями (ребрами, дугами). Вершины символизируют отдельный видовой список, полученный на данной территории, а ребра разной длины или толщины – величину относительных отличий двух списков, связанных этими ребрами. Дендрограммы, по определению, должны объединить все сравниваемые объекты, которые, группируясь, образуют *кластеры* (*группы* сходных объектов). Процедура объединения объектов в группы и построение «деревьев» называется кластеризацией. Существует множество разнообразных способов кластеризации; все они основываются на анализе матрицы сходства видовых списков (какую бы меру сходства ни применяли).

Одним из самых распространенных и «понятных» оказывается «метод ближайшего соседа» (Nearest Neighbor), когда дугами соединяются только те вершины, которые удалены друг от друга на наименьшее расстояние. Вначале процесса анализа матрицы сходства отыскиваются минимальные расстояния (d_{min}) между объектами и в первый кластер добавляются объекты, максимально сходные друг с другом (так может сформироваться несколько кластеров). На следующих шагах в этот кластер добавляется такой объект, который имеет наименьшее отличие от какого-либо из объектов данного кластера по сравнению со всеми прочими объектами. Все сравнения выполняются на основе единственной исходной матрицы расстояний и новые не рассчитываются. Алгоритм кластеризации реализован во многих пакетах статистики, где, к сожалению, отсутствует расчет мер сходства Съёренсена или Жаккара. Поэтому кластеризацию для случаев β-разнообразия приходится выполнять вручную.

Подготовим таблицу кластеризации для нашего примера (табл. 5.2.6), в которой будем отмечать характеристики, необходимые для объединения объектов. В первый столбец (Сосед 1) выпишем индексы всех объектов. Далее в таблице различий (табл. 5.2.3) отыщем первое минимальное расстояние – это отличия между биотопами пихтач и экотон, $d_{ПЭ} = d_{ЭП} = 4\%$. Запишем его в две ячейки графы d, индексы биотопов запишем в графу Сосед 2, отметим номер кластера (1) в поле Кластер 1 (это были шаги 1 и 2). На третьем

шаге отыскиваем следующее минимальное расстояние: на сей раз между березняком и лугом ($d_{БЛ}$ = 7 %), это кластер 2. Четвертый шаг предписывает организовать связь между пихтачом и березняком ($d_{БЛ}$ = 10 %), поэтому записываем напротив П индекс Б вместо Э; в поле Кластер 2 отмечаем объединение обоих кластеров, дополняем поле Шаг (4). Еще через 2 шага все биотопы оказываются связанными воедино, кроме луга, добавляем его и завершаем кластеризацию. В данном случае все объекты сразу слились в один кластер и поле Сосед 3 не понадобилось.

Таблица 5.2.6. Таблица кластеризации видовых списков мелких млекопитающих

Сосед 1	*d*	Сосед 2	*d*	Сосед 3	Кластер 1	Кластер 2	Шаг
К	13	***Э***			1		5
П	10	**Б**			1		1, 4
Э	4	**П**			1		2
С	16	**К**			1		6
Б	7	**Л**			2	1	3
Л	25	**Г**			1		7
Г							

На основе таблицы кластеризации нетрудно построить линейное «дерево» (рис. 5.2.1), который в нашем случае оказался неразветвленным; длина его ребер пропорциональна рассчитанным расстояниям между вершинами (биотопическими группировками).

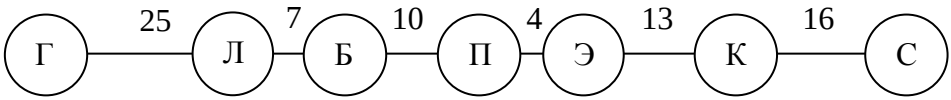


Рис. 5.2.1. Дендрограмма сходства биотопических группировок мелких млекопитающих

Когда дерево построено, его легко можно превратить в более структурированную (но часто менее наглядную) *дендрограмму* (рис. 5.2.2). Здесь все ребра расположены параллельно оси ординат, от-

градуированной в единицах меры сходства, то есть по оси ординат откладывается расстояние между ближайшими соседями (или другие значения выбранной метрики кластеризации). Все объекты расположены вдоль оси абсцисс в том порядке, который продиктован логикой их связей (и субъективными вкусами исследователя); наиболее сходные объекты соседствуют друг с другом. Соединяются объекты друг с другом дугой, треугольником или прямоугольником, высота которых равна расстоянию *d* между объектами при условии, что отдельные ветви древа не должны пересекаться.

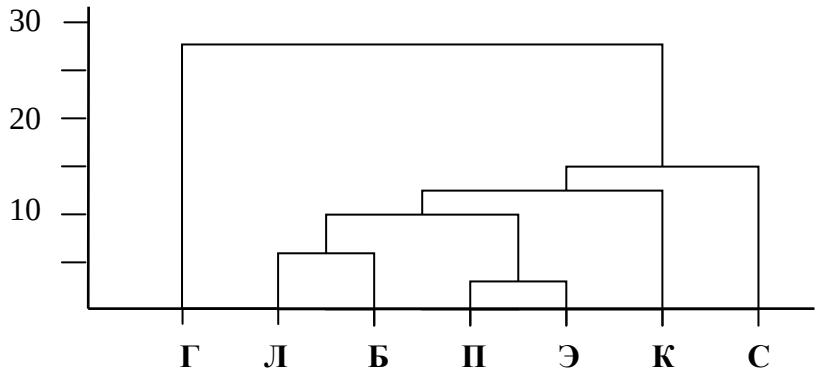


Рис. 5.2.2. Дендрограмма сходства биотопических группировок мелких млекопитающих по видовому составу

Другим приемом структурного изображения отношений между объектами служит *коррелограмма*, в которой объекты располагаются на равных расстояниях друг от друга по периферии окружности. Все вершины (объекты) соединяются линиями разной толщины, подобранной сообразно степени их сходства. Другой вариант – построение нескольких «корреляционных колец», на которых изображают дуги одинаковой толщины, но только для тех пар объектов, степень сходство между которыми превышает заданный порог. При низком уровне сходства все объекты будут связаны со всеми, по мере повышения порогового значения сохраняются только наиболее тесные связи. Анализ серии коррелограмм также позволяет выявить кластеры сходных объектов (Терентьев, 1959).

На основании таблицы корреляционных мер сходства (табл. 5.2.4) построим корреляционное кольцо, используя три поро-

говых величины (0.65, 0.7, 0.8), при которых все биотопы объединяются в один кластер (рис. 5.2.3).

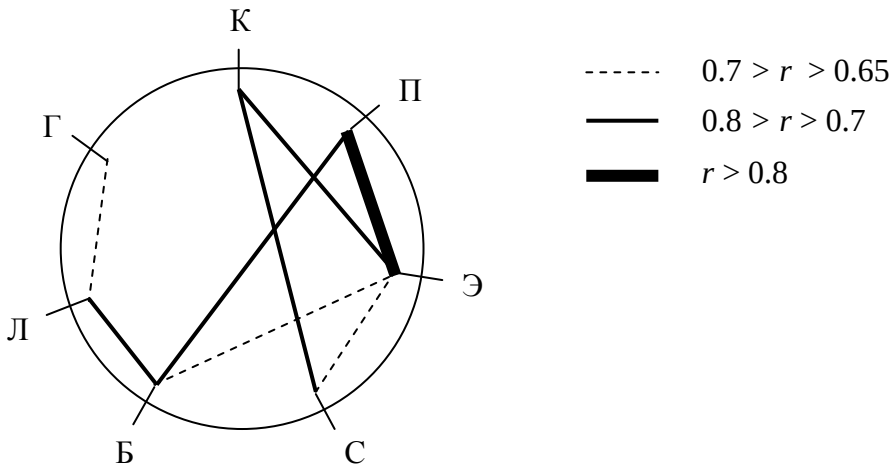


Рис. 5.2.3. Коррелограмма сходства видовых списков мелких млекопитающих

Используя корреляционный коэффициент в форме расстояния, было построено дерево (рис. 5.2.4), который отобразил аналогичную структуру с построением на основе меры Съёренсена (см. рис. 5.2.1).

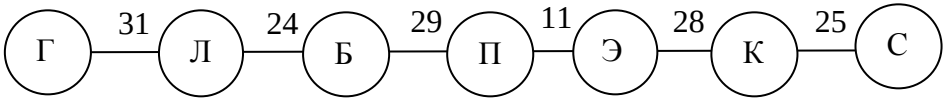


Рис. 5.2.4. Дерево корреляционных расстояний 100·(1–*r*) между биотопическими группировками мелких млекопитающих

Сопоставление встречаемости видов (табл. 5.2.1) с результатами кластерного анализа позволило обнаружить, что все отличия разных местообитаний сопряжены со спецификой распространения трех групп видов, имеющих специфические биологические потребности. Особенность города связана с обитанием здесь синантропных видов (крысы, мыши). Луг и березняк похожи из-за присутствия в уловах серых полевок, характерных только для вторичных стаций. Пихтач, экотон и недалеко отстоящий кедровник характеризуются

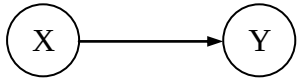
преимущественным обитанием здесь лесных полевок и лесной азиатской мыши. Обобщения многочисленных связей в форме деревьев ярко высветили имеющиеся отличия между биотопическими группировками мелких млекопитающих.

Несимметричные меры сходства

Мера включения (или показатель банальности, *B*) оценивает долю общих видов, включенных в каждый из сравниваемых списков по отдельности, всего получаем два показателя при одном сравнении (Андреев, 1980):

$$B_{XY} = c/(a+c),\ B_{YX} = c/(b+c).$$

Чем большую долю отдельного видового списка занимают общие виды, тем менее своеобразным (более «банальным») он оказывается. Самый «банальный» список (B = 1) не имеет ни одного «своего» вида, самый «экзотичный» (B = 0) – не включает в себя ни одного вида из других коллекций. Два показателя банальности показывают несимметричность отношений между списками, которую удобно отображать графически в форме *орграфа* (ориентированного графа), в котором вершины связаны дугами со стрелками, направленными от менее – к более банальным описаниям:



Графы этого типа передают информацию о структурных отношениях в многокомпонентных системах. Они позволяют на качественном уровне составить первое впечатление о возможных направлениях взаимодействия биосистем, о путях перераспределения видов (фенотипов) между сравниваемыми территориями. Рассмотрим применение меры банальности на примере с мелкими млекопитающими (исходные данные см. в табл. 5.2.1)

Расчет мер включения дает две треугольные матрицы коэффициентов (табл. 5.2.7). Например, из 11 видов, обитающих в кедровнике, 10 видов обнаружены и в экотоне; мера включения общих видов в кедровник составляет $B_{\text{КБ}} = 10/11 = 0.91$. В экотоне же найдено 12 видов, значит, мера включения общих видов в экотон равна $B_{\text{ЭК}} = 10/12 = 0.83$. Кедровник менее своеобразен, чем экотон, что определяет ориентацию стрелки орграфа от экотона – к кедровнику.

Таблица 5.2.7. Таблица отношений включения видовых списков мелких млекопитающих семи биотопов: справа вверху дана мера включения списка строки в список столбца, слева внизу – список столбца – в список строки (жирным выделены числа из примера)

	К	**П**	**Э**	**С**	**Б**	**Л**	**Г**
К		91	**91**	73	91	82	45
П	77		92	62	100	85	46
Э	**83**	100		67	100	92	50
С	100	100	100		100	88	63
Б	63	81	75	50		88	56
Л	64	79	79	50	100		64
Г	50	60	60	50	90	90	

Если из значений верхней матрицы вычесть значения нижней и построить таблицу, содержащую только знак этой разности, то мы получим характеристику парного включения «в чистом виде». Например, виды кедровника включены в список пихтача на 91%, виды пихтача в список кедровника – на 77%, разность 91 – 77 = 14 дает знак «+», видовой состав кедровника «банальнее» пихтача. С сосняком у кедровника отношения иные: 73 – 100 = –27, имеем «–», кедровник «оригинальнее» сосняка.

Таблица 5.2.8. Отношения включения списков микромаммалий:
«+» – виды строки включены в список столбца,
«–» – виды столбца включены в список строки.

К	+	+	–	+	+	–
	П	–	–	+	+	–
		Э	–	+	+	–
			С	+	+	+
				Б	–	–
					Л	–
						Г

Дополнив полученное ранее дерево сходства биотипических группировок мелких млекопитающих (рис. 5.2.1), выясняем основные пути их генезиса с помощью орграфа банальности (рис. 5.2.5).

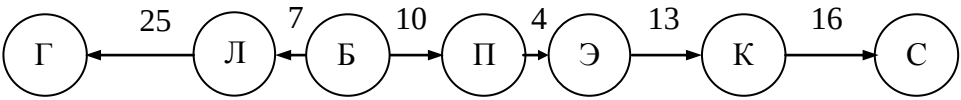


Рис. 5.2.5. Орграф отношения банальности биотопических группировок микромаммалий, построенный на основе дерева сходства

Самый своеобразный таксоценоз обнаружен в березняке. Здесь в различные фазы динамики численности встречаются почти все виды животных, отмеченные в регионе. Эти вторичные леса граничат со всеми другими биотопами и не только обмениваются фаунистическими элементами, но и служат руслами обмена между удаленными биотопами. Через березняки красные полевки проникают на луга и в сосняки, а темные – в экотоны и пихтачи. Здесь устойчиво встречаются красно-серые полевки и многие виды бурозубок. Остальные (вторичные) местообитания оказываются временными приемниками мигрантов, а списки их населения оказываются неустойчивой смесью, меняющейся каждый год.

Типичные таежные местообитания (кедровник, пихтач) населены в основном таежными видами (бурозубки, мыши, лесные полевки), куда проникают жители поймы (полевка-экономка), а иногда и обитатели периферических лугов и березняков (темные полевки). Луговые стации также обладают специфическим компонентом (два вида серых полевок), но во многом зависят от проникновения расселяющихся мигрантов из березняка и тайги. Фауна города (Байкальска) своеобразна из-за обитающих здесь серых и черных крыс, домовых мышей.

Показатели банальности позволяют существенно детализировать представление об источниках сходства и отличий изучаемых коллекций.

Известен ряд других мер сходства-различия: коэффициенты связи Юла, Скотта, Кохена (Песенко, 1982, с. 174–175), Коула, Дайса (Миркин и др., 1989, с. 74–75) и др. Важно отметить, что сравнение только видовых списков есть достаточно примитивная процедура, не учитывающая разную значимость (численность, встречае-

мость, плотность, биомассу, долю) отдельных видов для сообщества, то есть их выравненность. Сопоставление простых списков по информативности и достоверности оценок сходства заметно уступает сравнению распределений видовых значимостей. Не зная оценок встречаемости видов, нельзя определить и теоретическую вероятность обнаружения отдельного вида в списках, то есть нельзя выполнить статистическое оценивание мер сходства. Для точной характеристики отличия между флорой или фауной разных районов следует переходить к оценке видовой значимости (раздел 5.3). Однако для разведочного анализа сравнение территорий по видовому богатству может быть очень полезным.

5.3. Выравненность: α-разнообразие

Показатели видового богатства (видового состава) очень поверхностно описывают таксоценоз как группу сходных видов и совершенно не отражают экологическую роль каждого из сочленов. В определение структуры изучаемой группировки организмов должны быть включены оценки относительной значимости видов, показатели их обилия. *Выравненность* – это характеристика соотношения значимостей (численности, биомассы, продукции, площади и пр.) отдельных видов; в самом общем смысле – распределение особей по видам.

Обычно под значимостью подразумевают *численность* n_i (общее число особей). Показатель подходит для исследования биотопических группировок животных сходных размеров, в силу чего биомасса прямо пропорциональна численности. Для растений это далеко не так – масса совместно обитающих растений, относящихся к одному семейству, может отличаться на несколько порядков. Поэтому при оценке их экологической значимости лучше ориентировать на *биомассу* B_i всех особей данного вида. В гидробиологии применяется синтетическая характеристика – *индекс плотности населения* $\rho_i = \sqrt{n_i B_i}$ (Шитиков и др., 2003; с. 163), вбирающий в себя информацию обеих исходных характеристик. Для древесных пород растений эффективной оценкой значимости служит *сумма площадей поперечных сечений стволов* (Казенс, 1982).

Индекс доминирования

Для описания ценозов зачастую удобнее пользоваться относительными единицами значимости – индексами доминирования. Они рассчитываются на основе измеряемых видовых показателей (численность, биомасса, площадь) и позволяют сопоставлять роль разных видов в едином сообществе, а также сравнивать группировки разного объема с различным видовым богатством. Наиболее понятен, прост и распространен *индекс доминирования*, выраженный как доля особей данного вида во всей выборке: $p_i = n_i / N$.

Например, из 4094 зверьков, отловленных нами в Южном Прибайкалье, 1394 красно-серые полевки составляют долю $p_1 = 0.340$, а пять обыкновенных кутор – $p_{16} = 0.001$. Вместо этого показателя в гидробиологических исследованиях предлагается *индекс доминирования Палия*, учитывающий, кроме численности, представленность вида в разных пробах: $p_i = 100 \cdot \frac{m_i}{M} \cdot \frac{n_i}{N}$, где M – общее число проб, m_i – число проб, в которых обнаружен данный вид. Аналогичная мера предлагается для биомассы: $p_i = 100 \cdot \frac{m_i}{M} \cdot \frac{B_i}{B}$ и для индекса плотности населения: $p_i = 100 \cdot \frac{m_i}{M} \cdot \frac{\sqrt{n_i B_i}}{\sqrt{NB}}$ (Шитиков и др., 2003). Однако большинство методов анализа и показателей выравненности (например, мера информативности Шеннона) основаны на первой формуле индекса доминирования.

Распределение видов и значимостей

В результате натурных наблюдений на одной пробной площадке исследователь получает ряды парных данных – название конкретного вида животных или растений сопровождается оценкой его значимости. Обнаружить закономерную упорядоченность, структуру сообщества позволяют обобщения этих данных, представленные в виде распределений. Для этого используются два основных вида распределений и соответствующие им иллюстрации:

- диаграмма частоты встречаемости видов с данной значимостью (численностью) и
- диаграмма распределения значимостей по видам.

Распределение видов по значимостям создается в несколько этапов. По сути дела, строится вариационный ряд: подсчитывается количество видов, значимости которых попадают в разные предварительно намеченные интервалы значимостей.

Этапы работы таковы. На листе Excel размещаем данные в подписанных столбцах: у нас номера видов (i) заданы в блоке A1:A19, а оценки значимости (число особей, N) – в блоке B1:B19 (рис. 5.3.1).

Число видов (размер коллекции) равно номеру последнего вида $s = 18$.

Определяем максимальное C2 =МАКС(A2:A19), минимальное C3 =МИН(A2:A19) значения и их разность C4 =C2-C3, размах изменчивости значений, *лимит*: $Lim = 1392$.

Задаем число классов вариационного ряда, исходя из содержательных соображений, или по формуле: $k = 1 + 3.32*\lg(18) \approx 5$, C5 =ОКРУГЛ(1+3.32*LOG(18),0).

Найдем ширину интервала: $dx = Lim / k = 1392/5 \approx 278$, C6 =ОКРУГЛ(C4/C5,0).

Установим левую границу первого класса – нуль C7 =0. Находим следующую границу, прибавляя к предыдущей значение ширины интервала: C8 =C7+C$6; далее вводим формулу еще в пять ячеек, применяя *автозаполнение*: C9 =C8+C$6… (блок C8:C13).

Производим разноску вариант в соответствующие классы, подсчитывая их частоты с помощью макроса пакета Excel «Гистограмма» из меню «Сервис \ Анализ данных».

Следует иметь в виду, что макрос «Гистограмма» в результирующей таблице (блок A21:B28) в графе *Карман* указывает не центр, а *правую границу* интервала, ставя напротив число попавших в него значений (графа *Частота*). Например, записи 278 и 14 в ячейках A22 и B22 означают, что интервал от 0 до 278 содержит 14 значений (14 видов с численностью меньше 278 экз.). Оценки значимостей у разных видов могут очень сильно различаться, иногда на 3–4 порядка (у нас – 2 и 1394). При количественной обработке столь несоизмеримых величин преимущество получат наибольшие значения, именно они будут определять уровень обобщенных показателей, оценивающих представленность видов в выборке или степень различия сравниваемых выборок.

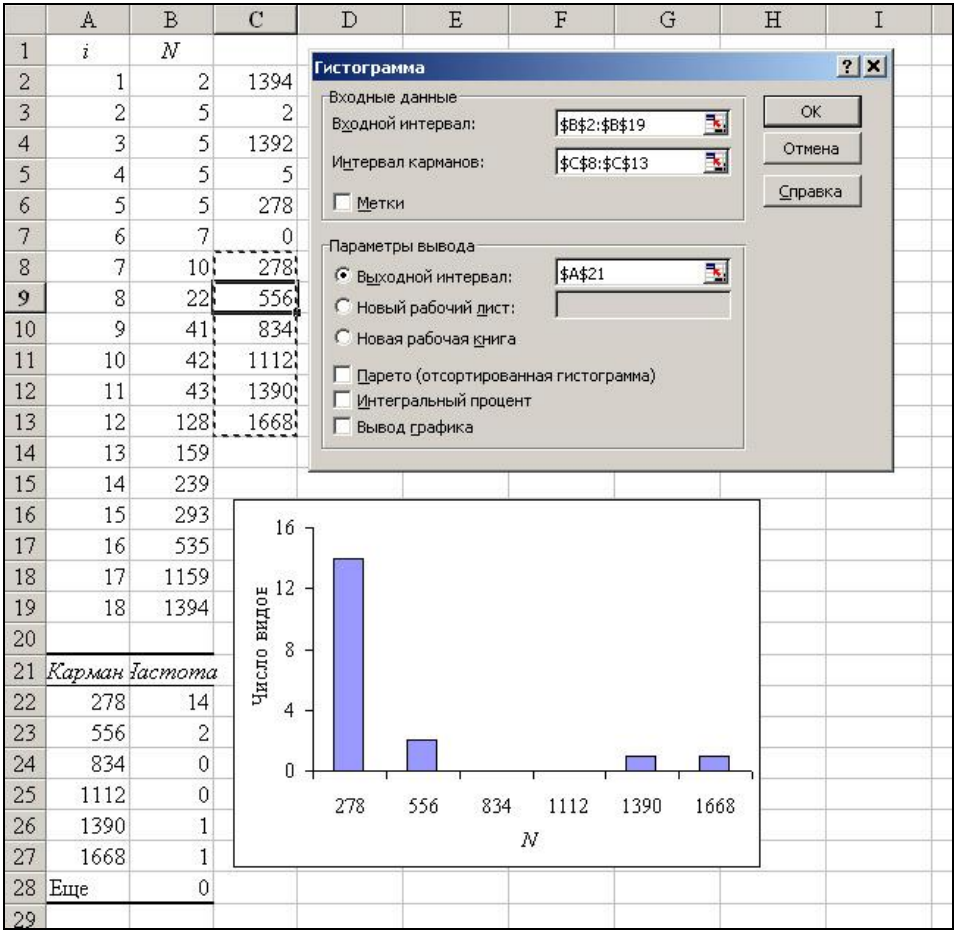
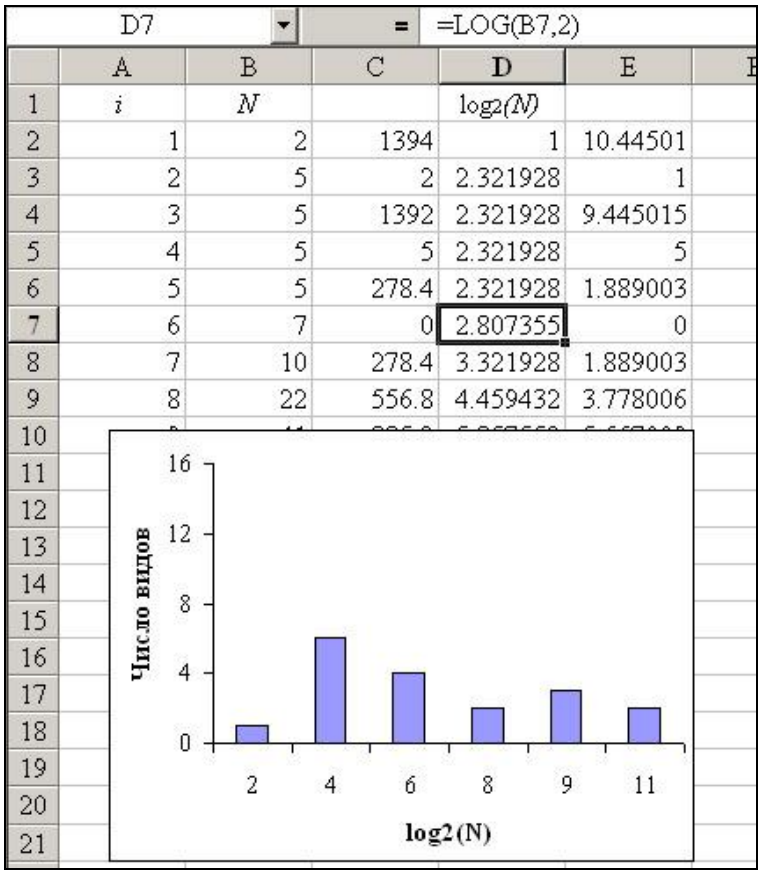


Рис. 5.3.1. Составление вариационного ряда значимостей видов и построение распределения видов по значимостям

С целью сближения оценок разных видов широко используется *логарифмирование значимостей* (расчет значений $\ln(n_i)$, $\ln(B_i)$, $\ln(\sqrt{n_i B_i})$ и пр.). При этом почти безразлично, какое брать основание логарифма (2, *e* или 10), они отличаются друг от друга на постоянный множитель: $\lg(x) = 2.3 \cdot \ln(x)$, $\lg(x) = 3.32 \cdot \log_2(x)$. Логарифмирование (исчисление степени числа) приводит к уменьшению исходных значений, причем существенно сильнее уменьшаются большие значения, нежели небольшие (например, десятичные логариф-

мы чисел 10 и 100 будут равны 1 и 2, то есть 10 уменьшилось в 5 раз, а 100 – в 50). К подобным же результатам, приводит преобразование исходных численностей с помощью извлечения квадратного корня, ($\sqrt{x} = x^{0.5}$, действует мягче логарифмирования), вычисление обратной величины ($1/x = x^{-1}$, действует сильнее логарифмирования). Получив логарифмы исходных значимостей, следует вновь рассчитать классовые интервалы, построить вариационный ряд и гистограмму. В результате этих операций форма распределения видов по численности становится гораздо более симметричной (рис. 5.3.2), а кривые доминирования выпрямляются (см. ниже).



	A	B	C	D	E
1	*i*	*N*		log2(*N*)	
2	1	2	1394	1	10.44501
3	2	5	2	2.321928	1
4	3	5	1392	2.321928	9.445015
5	4	5	5	2.321928	5
6	5	5	278.4	2.321928	1.889003
7	6	7	0	2.807355	0
8	7	10	278.4	3.321928	1.889003
9	8	22	556.8	4.459432	3.778006

Рис. 5.3.2. Распределение видов по логарифмам $\log_2(N)$ значимостей

Анализ распределения видов по значимости дает возможность судить о структуре ценоза, о качестве условий обитания и напряженности конкурентных отношений. Если гистограмма показывает, что количество мало- и среднечисленных видов велико, то, даже не зная никакой экологической теории, можно понять, что условия в этом местообитании благоприятны для самых разных организмов. Если же на графике представлены в основном многочисленные виды, то здравый рассудок подсказывает – условия (или отношения между видами) не способствуют выживанию особенным, редким видам, то есть в целом ситуация в изучаемом районе неблагоприятна. Детальное исследование видов распределений дает основания и для более содержательных заключений.

Распределение значимостей видов получается в том случае, если по оси абсцисс расположить виды, а по оси ординат откладывать их частоты. При этом виды следует расположить в порядке *уменьшения* их значимостей (показателей обилия, биомассы или индексов доминирования). Соединяя точки, строим *кривые значимости видов*, или *кривые разнообразия-доминирования*. Диаграммы этого типа строить проще, поскольку не требуется предварительно подсчитывать частоты. Тем не менее анализ полученных кривых позволяет судить о структуре взаимоотношений между видами сообщества, в том числе о степени перекрывания ниш и связанной с этим конкуренции, а также о качестве среды обитания, насколько она обеспечивает изучаемые виды необходимыми ресурсами. Известна классификация разных типов выравненности таксоценозов, для описания которых используется большое число моделей с различными теоретическими основаниями.

Если большое число видов имеет близкую или одинаковую численность, то условия среды можно считать весьма благоприятными, в которых многие сочлены находят достаточное количество необходимых ресурсов и не вступают в жесткие конкурентные отношения (их ниши почти не перекрываются); таковы условия для сосудистых растений в дождевых лесах. Кривая доминирования в логарифмическом масштабе на большом протяжении плавно спускается к оси абсцисс, образуя два изгиба (см. рис. 5.3.5, 4). Такие сообщества обладают высокой степенью выравненности и могут быть описаны с помощью *модели «разломанного стержня» Макартура*.

В менее благоприятных условиях среды наблюдается перекрывание ниш и ужесточение конкуренции, взаимное подавление выдерживает меньшее число видов. Лимитирующим фактором может служить, например, пространство; хорошими примерами конкурентных отношений служат сообщества гнездящихся птиц. Выравненность сообществ снижается, что наиболее успешно можно описать с помощью *модели логнормального распределения*.

В случае еще более жесткого давления неких лимитирующих факторов кривые значимости видов лучше описываются *логарифмическим рядом значимостей*. Примером могут служить растения наземного яруса в хвойных культурах в условиях низкой освещенности, сообщества многочисленных насекомых.

В суровых условиях среды (север, горы, пустыни) ниши захватываются редкими видами, почти не оставляя ресурсов для прочих видов. Такие сообщества крайне невыровнены, кривая доминирования идет круто вниз и аппроксимируется в помощью *модели геометрического распределения*.

Распределение «разломанного стержня»

Изучая сообщества разнообразных организмов, Р. Макартур предложил несколько теоретических моделей, основанных на идее случайного распределения ниш (Песенко, 1982). Положим, что виды, размножаясь, осваивают объем ресурса экосистемы до того момента, пока не произойдет его полное распределение. Если их ниши не будут перекрываться, то множество объемов потребляемых ресурсов (следовательно, и численности видов) будет представлять собой набор случайных значений, который можно уподобить множеству осколков разбившейся стеклянной палочки. Длина отдельного фрагмента будет пропорциональна численности отдельного вида (при условии равенства числа видов и числа фрагментов).

Не обращаясь пока к аналитическим выкладкам, попробуем воспроизвести ситуацию случайного разбивания стеклянного стержня с помощью простейшей имитационной модели, построенной в среде Excel (рис. 5.3.3). Роль стеклянного стека сыграет отрезок единичной длины, который мы будем разбивать случайным образом на 18 частей (в изучаемом нами сообществе 18 видов мелких млекопитающих). На листе Excel перенумеруем будущие отрезки от $i = 1$ до 18 (колонка А).

D6 = =C6-C5

	A	B	C	D	E	F
1	*i*	сл.число	сорт.	разлом	сорт.	*Mi*
2	1	0.87	0.02	0.02	0.21	859
3	2	0.35	0.03	0.01	0.15	601
4	3	0.58	0.18	0.15	0.11	446
5	4	0.02	0.19	0.01	0.09	374
6	5	0.38	0.24	0.05	0.09	360
7	6	0.03	0.35	0.11	0.09	356
8	7	0.18	0.38	0.03	0.05	214
9	8	0.62	0.47	0.09	0.04	169
10	9	0.63	0.49	0.02	0.03	122
11	10	0.24	0.58	0.09	0.03	105
12	11	0.19	0.62	0.04	0.02	94
13	12	0.49	0.63	0.01	0.02	87
14	13	0.66	0.66	0.03	0.02	74
15	14	0.47	0.87	0.21	0.02	68
16	15	0.98	0.88	0.01	0.01	53
17	16	0.97	0.97	0.09	0.01	46
18	17	0.88	0.98	0.02	0.01	34
19	18		1.00	0.02	0.01	32
20	Сумма				1	4094

Рис. 5.3.3. Имитация случайного разбивания отрезка на 18 частей

Наметим 17 случайных точек разлома с помощью датчика случайных чисел, заполнив формулой =СЛЧИС() ячейки B2:B18. Выделим этот диапазон B2:B18, скопируем в буфер обмена и вставим как Значения (с помощью Специальной вставки контекстного меню: клик правой кнопкой мыши) в следующий блок ячеек C2:C18. Отсортируем значения (Данные \ Сортировка) в порядке возрастания, имитируя последовательность разломов; добавим в ячейку B19 значение 1, как символ края стерженька. Рассчитаем длину получившихся 18 осколков, находя разницу между следующей и предыдущей точками разлома. Поскольку первый отрезок равен разности $0 - 0.02$, можно написать D2 =C2. Длина второго осколка составит D3 =C3-C2: $0.03 - 0.02 = 0.01$. Эту формулу следует скопировать в блок C3:C19.

Выделим столбец рассчитанных значений, скопируем в буфер обмена и вставим с помощью Специальной вставки как Значения в блок E2:E19. Отсортируем значения в порядке убывания; это и будет искомое распределение относительных значимостей p_i для 18 видов, случайным образом поделивших нишу объемом 1. Умножив каждое значение p_i на общую численность сообщества, получим численность данного вида в соответствии с идеей разломанного стержня: $N_i = p_i \cdot N$, например, 0.11· 4094 = 446 экз.

Конечно, случайный набор из 18 случайных осколков не будет точно соответствовать идеальному случайному ряду, но его можно получить с помощью теоретически обоснованной формулы (Лебедева и др., 2004): $p_k = \frac{1}{s} \cdot \sum_{i=k}^{s} \frac{1}{i}$, где s – число видов, i – номер вида по порядку, k – номер вида, принятого в обработку, $\sum_{i=k}^{s} \frac{1}{i}$ – знак суммирования, начиная от вида $i = k$ и заканчивая последним видом.

Расчеты выполняются в среде Excel (рис. 5.3.4). После ввода нумерации видов i (столбец A) рассчитываем отношение $1 / i$ для каждого вида (колонка B). Далее подсчитываем суммы этих отношений, начиная с текущего вида и заканчивая последним. Для этого достаточно в ячейку C2 один раз ввести формулу =СУММ(B2:B19) и скопировать ее в остальные 17 ячеек столбца C. Смещение диапазона суммирования на нижележащие *пустые* ячейки не будет искажать суммы. Теперь умножим значения столбца C на заранее рассчитанную дробь 1 / s = 0.056 (колонка D) и получим искомое теоретическое (гладкое) случайное распределение значимостей видов (D2 = 3.495 · 0.056 = 0.194). Далее умножаем относительные значимости p_i изученных видов на общую численность (N = 4094 экз.) и определяем конкретный вид теоретической кривой разнообразия–доминирования (E2 = 0.194 · 4094 = 795…) (рис. 5.3.5).

В литературе встречается другой вариант формулы рангового распределения (Шитиков и др., 2003): $p_i = \frac{1}{s} \cdot \sum_{i=1}^{k} \frac{1}{s+1-i}$; s – число

видов, i – номер вида по порядку, k – номер вида, $\sum_{i=1}^{k}$ – знак суммирования, начиная от вида $i = 1$ и заканчивая k-м видом.

Эту формулу применяют к ряду видов, упорядоченных *по возрастанию* значимостей (рис. 5.3.4). Формула же, рассмотренная ранее и использованная нами, ориентирована на ряды, упорядоченные *по убыванию* значимостей (рис. 5.3.3). Конечно, можно сортировать и ранжировать виды любым способом, но более привычны диаграммы распределения видов по возрастанию значимостей, поэтому мы подробнее рассмотрели расчеты по первой формуле, хотя обе формулы дают идентичные результаты.

ГРАСПОБР ✕ ✓ = =СУММ(B4:B21)

	A	B	C	D	E	F	G	H
1	*i*	1/*i*	сумма	*Pi*	*Ni*	1/*S* =	0.056	
2	1	1	3.495	0.194	795			
3	2	0.5	2.495	0.139	567			
4	3	0.333	=СУММ(B4:B21)					
5	4	0.25	1.662	0.092	378			
6	5	0.2	1.412	0.078	321			
7	6	0.167	1.212	0.067	276			
8	7	0.143	1.045	0.058	238			
9	8	0.125	0.902	0.05	205			
10	9	0.111	0.777	0.043	177			
11	10	0.1	0.666	0.037	152			
12	11	0.091	0.566	0.031	129			
13	12	0.083	0.475	0.026	108			
14	13	0.077	0.392	0.022	89.1			
15	14	0.071	0.315	0.017	71.6			
16	15	0.067	0.244	0.014	55.4			
17	16	0.063	0.177	0.01	40.2			
18	17	0.059	0.114	0.006	26			
19	18	0.056	0.056	0.003	12.6			
20	Сумма			1	4094			
21								

Рис. 5.3.4. Расчет случайного разбивания отрезка на 18 частей

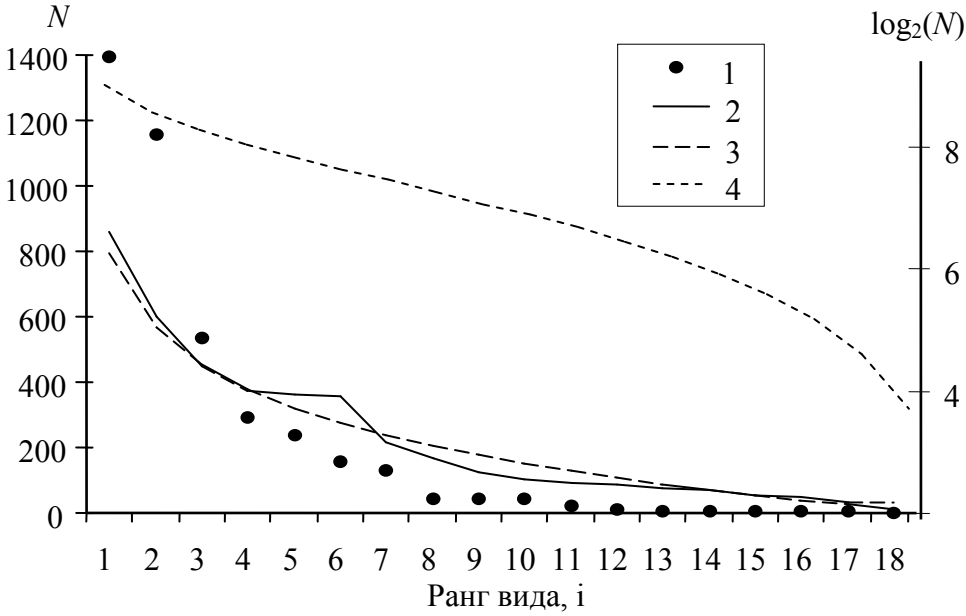


Рис. 5.3.5. Распределение значимостей (числа особей) разных видов мелких млекопитающих: 1 – наблюдения, 2 – имитация случайного разбиения, 3, 4 – модель «разломанного стержня» (3 – натуральные числа; 4 – двоичные логарифмы численности)

Завершая раздел, следует отметить, что построение теоретического распределения рангов с помощью имитационной модели (случайное разламывание стержня длиной 1) в целом хорошо совпало с расчетами по точной формуле (рис. 5.3.5; 2 и 3), что делает идею и формулу более понятными. Отличие имитационной кривой от «истинной» связано с тем, что был рассмотрен один конкретный *выборочный* набор случайных чисел и при других наборах случайных чисел кривая пойдет несколько по-иному. Если же усреднить достаточно большое (бесконечное) множество таких наборов, мы получим гладкий «правильный» ряд.

Результаты расчетов показали сильное отличие исходных данных (оценок численности мелких млекопитающих Южного Прибайкалья) от рассмотренной модели. Существенность этих отличий видна и невооруженным глазом. Но для иллюстрации можно построить график зависимости логарифмов эмпирических и расчетных

значений численности отдельных видов (рис. 5.3.6), построить линию регрессии и оценить достоверность отличия коэффициента регрессии от единицы (что соответствует полному совпадению реальности и теории). Воспользуемся критерием Стьюдента $t = |1 - a| / \sqrt{2m_a^2} = = |1 - 1.77| / \sqrt{2 \cdot 0.1492^2} = 0.77 / 0.21 = 3.7$. Табличное значение составит (табл. 4С, стр. 350): $t_{(0.05,\ 16)} = 2.1$. Отличие коэффициентов ($1.77 \neq 1$) значимо при $\alpha = 0.002$.

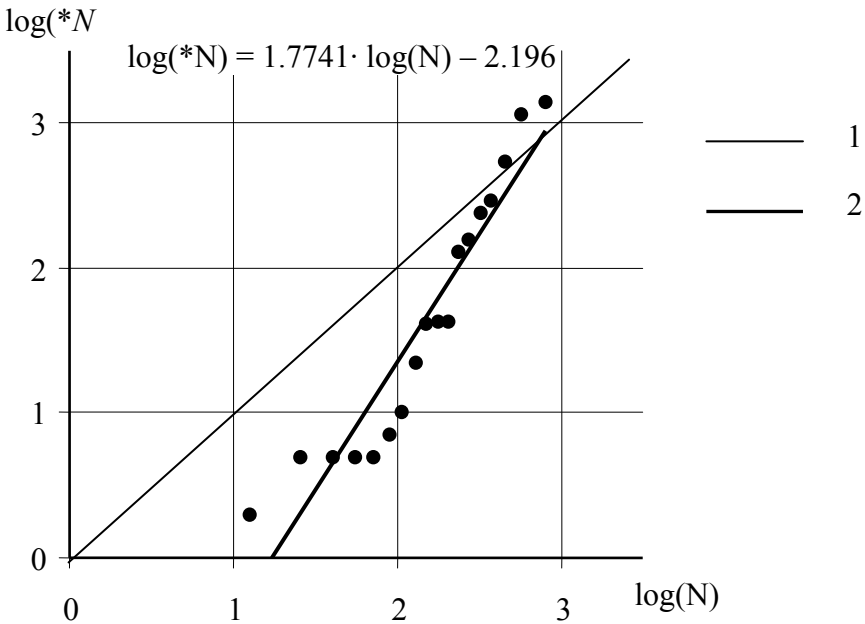


Рис. 5.3.6. Оценка соответствия эмпирических (*N*) и теоретических (**N*) значений численности видов мелких млекопитающих (1 – график, соответствующий модели разломанного стержня; 2 – линия регрессии теоретических значимостей по наблюдаемым)

Получается, что мелкие млекопитающие используют ресурсы биотопов не случайно, а имеет место некое их перераспределение, то есть ниши этих видов явно перекрываются и они в значительной мере конкурируют. Наши материалы подтверждают и численно выражают обычно неопределенные высказывания о существовании конкурентных отношений между представителями этой группы (Европейская рыжая полевка, 1987).

Логарифмическое распределение видов

Идея логарифмического ряда, высказанная Р. Фишером, состоит в описании группы видов, постепенно увеличивающих свою значимость (в примере – численность). Рассматривается число видов (s), обладающих разными значениями численности (n):

$s_1, s_2, \ldots s_n \ldots s_N$ $(n = 1, 2 \ldots N)$ $(s_1 > s_2 > s_3 \ldots > s_N)$.

Начинается этот теоретический ряд с самого большого числа видов, представленных в сообществе одной особью (s_1). Второй член ряда – это число видов, представленных двумя особями, оно меньше первого на константу: $s_2 = s_1 \cdot x/2$. Число видов, представленных тремя особями, еще меньше: $s_3 = s_1 \cdot x^2/3$ и так далее вплоть до очень немногих видов с максимальной численностью N_S: $s_N = s_1 \cdot x^{N-1}/N$. Поскольку s_1 и x на протяжении всего ряда не меняются, их заменяют константой $\alpha = s_1/x$. Теперь по формуле $s_n = \alpha \cdot x^n/n$ можно рассчитать теоретическое число видов S_n с любой численностью n, соответствующее логарифмическому распределению.

Величины α и x являются параметрами логарифмического ряда. Они определяют строгие соотношения между значениями числа видов и их общей численности (Песенко, 1982): $s = \alpha \cdot \ln(1 + N/\alpha) = -a \cdot \ln(1-x) = s_1 + s_2 + \ldots + s_N$, $N = \alpha \cdot x/(1-x)$,

$\alpha = N \cdot (1-x)/x$ (ошибка $m_\alpha^2 = \dfrac{\alpha^2\left[(N+\alpha)^2 \cdot \ln\dfrac{2N+\alpha}{N+\alpha} - N\alpha\right]}{(sN + s\alpha - N\alpha)^2}$),

$$\frac{N}{s} = \frac{x}{-(1-x)\cdot\ln(1-x)}.$$

Две последние формулы позволяют определить параметры и построить теоретическое распределение, соответствующее имеющимся данным.

Для наглядности упростим наш пример и рассчитаем логарифмический ряд только относительно тех видов, которые представлены числом особей менее 100 экз. Таких видов насчитывается $s = 11$ с общим числом особей $N = 187$ экз.

Сначала определим параметр x в среде Excel. На отдельный лист в ячейку B1 вводим исходное приблизительное значение пара-

метра $\alpha \approx 0.9$, в ячейку C1 – формулу отношения $N/s = 187/11 = 17$, в B2 – правую часть последней формулы: $\frac{x}{-(1-x)\cdot\ln(1-x)}$, принимающую вид =B1/(-(1-B1)*LN(1-B1))). Далее вызываем макрос Подбор параметра из меню Сервис и даем задание: установить в ячейке B2 значение 17, изменяя значение ячейки B1 (рис. 5.3.7).

Рис. 5.3.7. Определение параметра *x* с помощью макроса Подбор параметра

Подгонка дала величину *x* = 0.986526436055743. В ячейку B2 вводим формулу для оценки параметра α: B2 =A2*(1-B1)/B1: $\alpha = N\cdot(1-x)/x = 187\cdot(1-0.98652)/0.98652 = 2.55403$ (рис. 5.3.8).

Далее в блок A4: A46 вводим значения количества особей *N* от 1 до 43 (поскольку максимальная численность одного из видов в нашей редуцированной коллекции была 43 экз.). Наконец, в блок B4: B46 вводим формулы расчета теоретического числа видов, представленных данным числом особей. Ссылки на параметры должны быть абсолютными. Создав первую формулу, остальные вводим с помощью «автозаполнение». Так, число видов, имеющих по 2 особи, равно: $s_n = \alpha\cdot x^n/n = 2.55403\cdot 0.9865^2/2 = 1.24$ (в формате Excel: B5 =$B$2*$B$1^A5/A5, рис. 5.3.8). Если расчетные значения числа видов с определенной численностью нанести на диа-

грамму, мы получим диаграмму плавно снижающихся оценок (рис. 5.3.9, 1), на которые совсем не похожи наши эмпирические ряды (рис. 5.3.9, 2).

	A	B	C
1	11	0.987	17
2	187	2.554	
3	*N*	*Sn*	
4	1	2.52	
5	2	=$B$2*$B$1^A5/A5	
6	3	0.817	
7	4	0.605	
8	5	0.477	
9	6	0.392	
10	7	0.332	
11	8	0.286	

Рис. 5.3.8. Формула расчета теоретического числа видов

Для лучшей наглядности многочисленные значения численностей стоит объединить, тогда отличие распределений станет не только наглядным, но и доступным для проверки с помощью критерия хи-квадрат (рис. 5.3.10). На диаграммах хорошо видно, что эмпирический ряд совсем не поход на теоретический, то есть распределение видов мелких млекопитающих не соответствует логарифмическому.

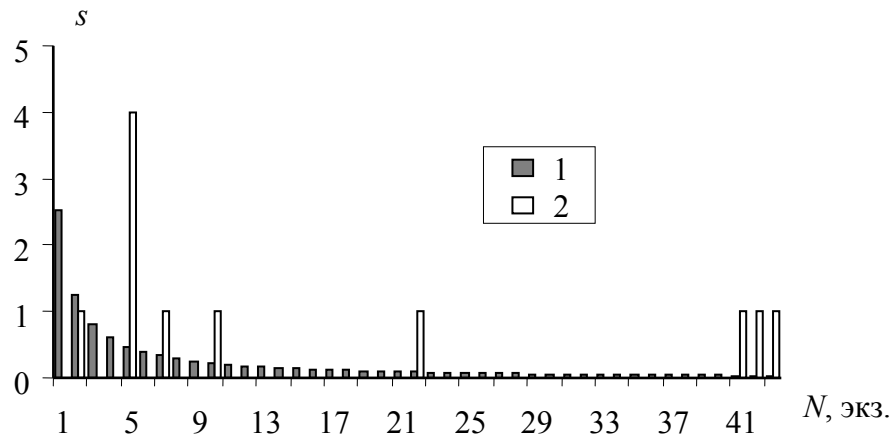


Рис. 5.3.9. Логарифмическое (1) и эмпирическое (2) распределение числа видов

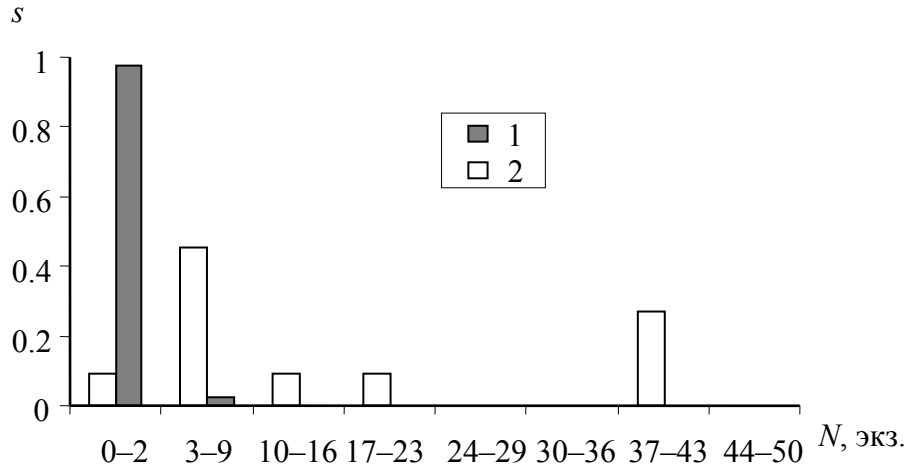


Рис. 5.3.10. Обобщенное теоретическое логарифмическое (1) и эмпирическое (2) распределение числа видов

Логарифмически-нормальное распределение видов

Как можно было заметить, логарифмическое распределение строилось для значений, равноотстоящих друг от друга на оси абсцисс, то есть на равномерной шкале (1, 2, 3 … 43 особи). Получить столь детальные эмпирические данные удается лишь на огромном фактическом материале. Можно, конечно, объединять значения в группы, подобно тому, как сделано в нашей иллюстрации (рис. 5.3.10), но это не снимает известную диспропорцию – видов с небольшой численностью существенно больше, чем многочисленных, значит, и при объединении в распределениях будут зиять пробелы.

Иной путь предложен Ф. Престоном, который разбил ось абсцисс на интервалы, пропорциональные *двоичному логарифму* 2^{N-1}, назвав их *октавами*. В первый интервал попадают виды, представленные одной особью $N = 1$ (2^0), во второй класс – виды с двумя особями $n = 2$ (2^1), в третий – виды с числом особей от 3 до 4 (2^2), в четвертый – от 5 до 8 (2^3), затем – до 16, 32, 64, 128, 256, 518, 1024, 2248 особей и т. д. В этом случае распределение видов по октавам становится значительно более компактным и зачастую приобретает характерную форму нормальной кривой. Логарифмирование численностей приводит к появлению нормального распределения видов – так появилось название этого вида выравненности.

Если выборки не очень велики (несколько тысяч особей), то таких данных недостаточно для обнаружения наиболее малочисленных видов, находящихся в октавах слева от модальной, и полноценно формируется лишь правая ветвь распределения. Условная линия, делящая распределение на неизвестную и реализовавшуюся части, названа *линией вуали* (линией занавеса). С ростом коллекций линия вуали смещается влево. Для описания этого усеченного распределения, то есть для расчета теоретического числа видов в октавах, используется формула расчета ординат нормальной кривой $p = \frac{1}{\sqrt{2\pi}} \cdot e^{-(x_i - M)^2 / 2S^2}$, приспособленная для частного случая: отклонение $(x_i - M)$ заменено на R (отклонение данной октавы от модальной); константа $\frac{1}{\sqrt{2\pi}} = 0.3989$, выражающая максимальную частость в центре нормального распределения, заменена на значение максимальной частоты теоретического ряда (s_0). Теперь формула для расчета частот распределения принимает вид (Песенко, 1982):

$s_R = s_0 \cdot e^{-R^2 / 2\sigma^2}$, где R – номер октавы, начиная с модальной (с наибольшим числом видов), s_R – число видов в октаве с номером R, s_0 – число видов в модальной октаве (при $R = 0$), σ^2 – дисперсия распределения, выраженная в октавах.

Поскольку изначально дисперсия σ^2 неизвестна и по усеченному распределению ее вычислить затруднительно, для расчетов распределения числа видов предложен упрощенный вариант рассмотренной формулы (Шитиков и др., 2003):

$s_R = s_0 \cdot e^{-aR^2}$, где a – параметр, связанный с величиной дисперсии, он примерно равен $a \approx 0.2$.

В начале построения логнормального распределения определяем октавы численностей (табл. 5.3.1, вторая графа): начиная с $N = 1$, каждая следующая граница октавы задается умножением предыдущей на 2: то есть 2, 4, 8 и т. д. Затем подсчитывается количество видов, s_R, численность которых соответствует данным октавам. Например, за время наших наблюдений отловлено 293 экз. обыкновенной бурозубки, значение находится между 256 и 512, значит, этот вид попадает в десятую с начала октаву и т. п.

Таблица 5.3.1. Расчет распределения видов по октавам в соответствии с логнормальным распределением (выделены значения модальной октавы)

Октавы				До настройки		После настройки	
				s_0 = 5		s_0 = 3.45	
численность, N			номер	a = 0.2	Ф = 14.68	a = 0.105	Ф = 11.9
больше	до	s_R	R	* s_R	ф	** s_R	ф
0	1	0	–3	-	-	-	-
1	2	1	–2	-	-	-	-
2	4	0	–1	-	-	-	-
4	**8**	**5**	**0**	**5.00**	**0.00**	**3.45**	**2.39**
8	16	1	1	3.35	5.53	2.80	3.24
16	32	1	2	2.25	1.55	2.27	1.62
32	64	3	3	1.51	2.23	1.84	1.34
64	128	1	4	1.01	0.00	1.50	0.25
128	256	2	5	0.68	1.75	1.21	0.62
256	512	1	6	0.45	0.30	0.98	0.00
512	1024	0	7	0.30	0.09	0.80	0.64
1024	2048	2	8	0.20	3.23	0.65	1.83

Далее в полученном эмпирическом распределении видов по октавам отыскивается *модальная октава*, содержащая наибольшее число видов (в нашем случае s_R = 5), которая далее рассматривается в качестве центра распределения. Уже на этом этапе нетрудно ориентировочно наметить возможный ход нормальной кривой, охватывающей правой ветвью весь эмпирический ряд (рис. 5.3.11).

Модельной октаве присваивается нулевой номер R = 0, октавы, лежащие справа от нее (в таблице – ниже), перенумеровываются в возрастающем порядке, слева – в убывающем; по существу R показывает величину отклонения данной октавы от центра распределения (табл. 5.3.1; графа R). Теперь можно воспользоваться упрощенным уравнением и рассчитать теоретические количества видов, распределенных по октавам в соответствии с логарифмическим законом (табл. 5.3.1; графа $^*s_R$).

Например, в первой октаве (R = 1) число видов должно быть равно: $s_R = 5 \cdot e^{-0.2 \cdot 1^2} = 3.3516$. Значение 5 играет роль частоты центрального класса распределения, а 0.2 – дисперсии распределения.

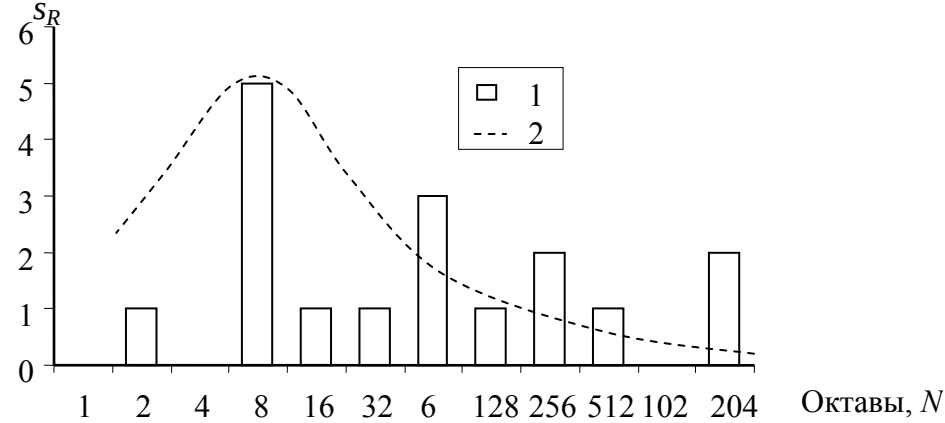


Рис. 5.3.11. Эмпирическое распределение (1) числа видов мелких млекопитающих Южного Прибайкалья по октавам численности и вероятный ход логнормального распределения (2)

Теоретические частоты видов плавно снижаются от модальной октавы к краевой и в целом мало походят на эмпирическое распределение, например, частоты в первой октаве отличаются более чем в три раза (3.35 против 1). Такая ситуация может быть связана как с тем, что изучаемое распределение плохо соответствует логнормальному, так и с тем, что эмпирическое распределение может иметь иные параметры, нежели принятые изначально значения s_0 = 5 и a = 0.2. В частности, величина 5 имеет небольшую репрезентативность, нежели остальные частоты распределения, и вполне может отличаться от истинной в силу случайных причин. Это значит, что при расчете теоретических частот следует ориентироваться не только на частоту модального класса, но и на частоты всех остальных октав. Сделать это можно, если подобрать такие параметры модели (s_0 и a), чтобы различие между эмпирическим и теоретическим распределением стало по возможности минимальным, то есть чтобы сумма квадратов отличий частот свелась к нулю $\Phi = \sum (s_R - {}^*s_R)^2 \to 0$. Рассчитав эти отклонения (например, для

первой октавы имеем $(1 - 3.35)^2 = 5.53$, табл. 5.3.1, графа ф), находим и всю сумму Ф = 14.7. Вызвав макрос Поиск решения, устанавливаем целевую ячейку со значение суммы отличий, равной значению нуль, изменяя ячейки с прежними параметрами 5 и 0.2, ОК.

В результате настройки (табл. 5.3.1, графа ** s_R) отличия распределений уменьшились (Ф = 11.9) за счет уменьшения величины параметров $s_0 = 3.4$ и $a = 0.1$. Однако, как явствует из диаграммы, новое распределение также мало походит на эмпирическое, как и прежнее (рис. 5.3.12).

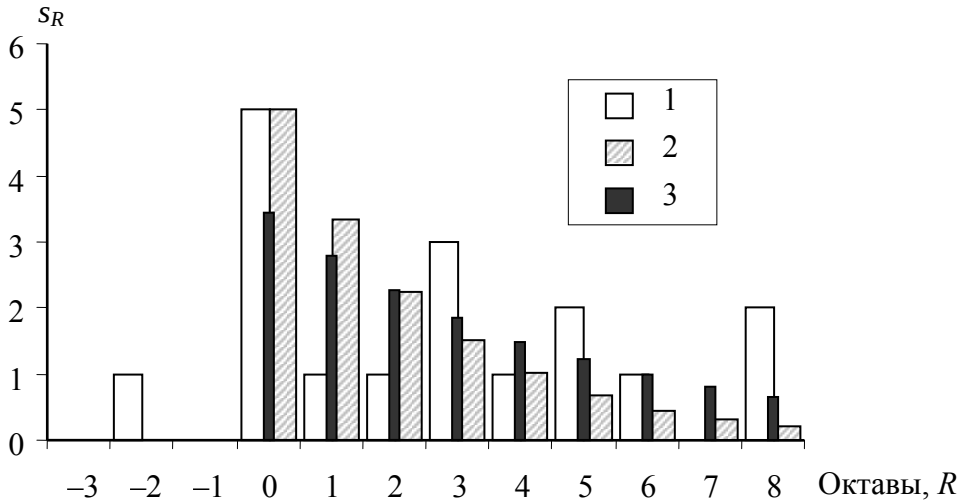


Рис. 5.3.12. Два варианта теоретического логнормального распределения числа видов мелких млекопитающих по октавам (по материалам табл. 5.3.1): 1 – исходные данные (s_R), 2 – вариант с условными параметрами (*s_R), 3 – вариант с настроенными параметрами (**s_R)

Качество модельного описания можно приблизительно оценить по величине коэффициента корреляции между эмпирическими и теоретическими частотами: для октав 0–8 ($n = 9$) он составил $r = 0.57$ и не превысил порог значимости $r = 0.66$ при $df = n - 2 = 7$. Это позволяет сделать вывод, что распределение видов мелких млекопитающих не соответствует логнормальному. В других случаях можно воспользоваться методами сравнения частотных распределений, например критериями χ^2 Пирсона, λ Колмогорова-Смирнова или провести дисперсионный анализ.

Геометрическое распределение значимостей

Для экологических ситуаций, когда видов в сообществе немного, но также имеются ярко выраженный дефицит ресурсов и жесткая конкуренция за них с перекрыванием и перераспределением ниш, И. Мотомура предложил модель геометрического распределения значимостей (численности) видов, ранжированных в порядке убывания этого показателя. Самый многочисленный первый вид ($i = 1$) захватывает самую большую долю (k) доступных ресурсов и приобретает соответствующую численность популяции, оставшаяся часть ресурсов в той же пропорции перераспределяется между остальными: k-я часть остатка потребляется следующим по рангу видом и т. д. Полагая, что наблюдаемая численность видов пропорциональна доле потребляемых ресурсов, величину k мы можем оценить по характеру распределения ранжированных видов:

$n_i = N \cdot k^{i-1}$, где N – общее число особей всех s видов, n_i – численность отдельного вида, k – параметр распределения, доля вида в оставшейся численности, i – ранг (номер) вида.

Путь для определения коэффициента k открывается в том случае, если формулу записать в виде имитационной динамической модели: $n_i = N_{ост.i} \cdot k$, $N_{ост.i+1} = N_{ост.i} - n_i$, где $N_{ост.i}$ – число особей, оставшихся после очередного перераспределения (после изъятия особей предыдущего i–1-го вида). Выражаясь конкретнее, на долю первого вида ($i = 1$) приходится k-я часть из всех особей $n_1 = N \cdot k$ (где $N_{ост.1} = N$). От остатка $N_{ост.2} = N_{ост.1} - n_1 = N - n_1$ особей k-ю часть «забирает» второй по рангу вид $n_2 = N_{ост.2} \cdot k$ ...

Подобрать коэффициент k можно с помощью макроса Excel Поиск решения. Мы рассмотрим два алгоритма – один понятийный, второй упрощенный.

Вначале ранжируем виды по убыванию значимостей (число особей) (рис. 5.3.13, графы *i, n*). Организуем графу $N_{ост.}$(столбец C), в которой будет фигурировать число оставшихся особей. Для первого вида оно составит $N_{ост.1}$ = 4094, введем его в ячейку C5. Для второго вида $N_{ост.2}$,= $N_{ост.1} - n_1$, или C6 =C5-D5.

D7 = =C7*D$3

	A	B	C	D	E	F	G	H	I
1			До настройки			После настройки			
2	Ранг			k =	Φ =		k =	Φ =	Φlog =
3	вида			0.5	481150		0.3599	67878	10.48
4	i	n	$N_{ост.}$	$^*n$	ϕ	$N_{ост.}$	$^*n$	ϕ	ϕlog
5	1	1394	4094	2047	426409	4094	1473	6315.9	0.006
6	2	1159	2047	1024	18360	2621	943	46589	0.088
7	3	535	1024	512	540.56	1677	604	4720.2	0.03
8	4	293	512	256	1378.3	1074	386	8728.1	0.159
9	5	239	256	128	12335	687	247	69.661	0.002
10	6	159	128	64	9030.9	440	158	0.4572	4E-05
11	7	128	64	32	9219	282	101	710.68	0.114
12	8	43	32	16	729.42	180	65	478.19	0.352
13	9	42	16	8	1156.3	115	42	0.2294	3E-04
14	10	41	8	4	1369.1	74	27	208.02	0.391
15	11	22	4	2	400.04	47	17	24.882	0.138
16	12	10	2	1	81.009	30	11	0.7905	0.015
17	13	7	1	0	42.253	19	7	0.0009	4E-05
18	14	5	0	0	22.564	12	4	0.2901	0.027
19	15	5	0	0	23.766	8	3	4.598	0.653
20	16	5	0	0	24.379	5	2	10.062	2.108
21	17	5	0	0	24.689	3	1	14.669	4.391
22	18	2	0	0	3.9378	2	1	1.5652	2.008
23	Сумма	4094		4094			4093		

Рис. 5.3.13. Расчет геометрического распределения числа особей 18 видов мелких млекопитающих (*n)

Путем «автозаполнения» введем эту формулу в ячейки C7:C22 для остальных видов. Далее организуем графу *n (столбец D), в которой будем вычислять теоретическую численность очередного вида. Предварительно зададим параметр распределения величиной k = 0.5 (D3), которая обычно рассматривается как характеристика жестких межвидовых отношений, когда каждый вид «забирает» половину объема доступных ресурсов, оставляя низшим по рангу вторую половину (Шитиков и др., 2003).; на языке Excel эта фор-

мула $n_i = N_{ост.i} \cdot k$ примет вид =C5*D$3 (для первого вида). Отметим, что ссылка на ячейку с параметром k должна быть абсолютной, то есть содержать префикс $. «Автозаполним» формулой остальные ячейки.

Построив диаграмму для колонок n и $*n$, видим, что кривая теоретического распределения проходит вблизи от эмпирической последовательности числа особей разных видов (рис. 5.3.14, 2), но предсказывает избыточно большую численность у доминирующих видов и меньшую – у остальных.

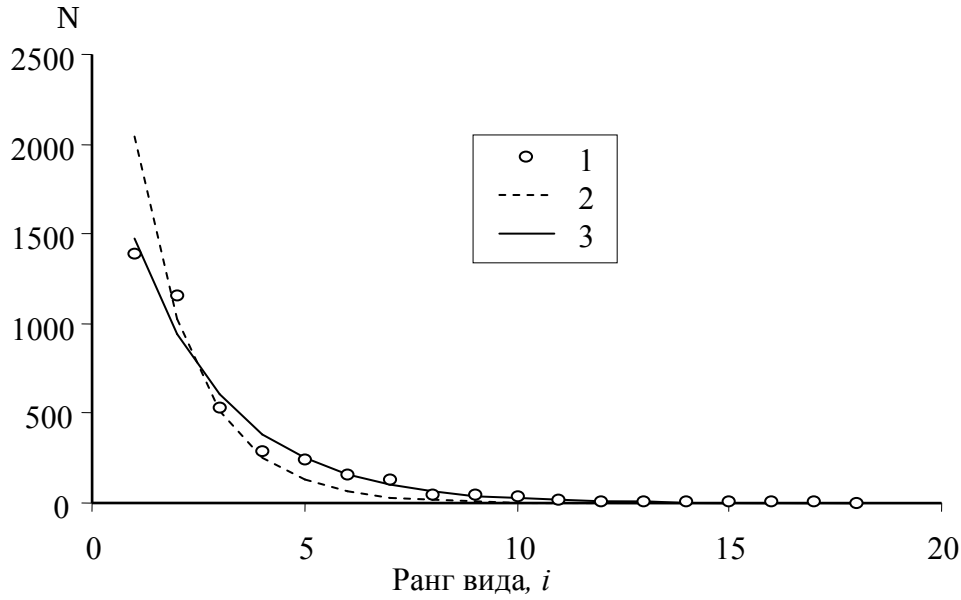


Рис. 5.3.14. Теоретическое геометрическое распределение численности видов мелких млекопитающих (по материалам рис. 5.3.13): 1 – исходные данные (s_R), 2 – вариант с параметром k = 0.5, 3 – вариант с параметром k = 0.3599

Этот недостаток можно исправить, подобрав такое значение параметра k, чтобы отличия между теоретическими и эмпирическими оценками численности видов были минимальны, то есть необходимо обнулить сумму квадратов отличий $\Phi = \sum (n - *n)^2 \to 0$. Введем эти формулы в графу ϕ, рассчитаем сумму, Ф = 481150. Вызываем макрос Поиск решения с целью свести к нулю значение Ф

(ячейка E3), изменяя значение k (ячейка D3). Получаем новое значение параметра $k = 0.3599$ и очень хорошее визуальное совпадение расчетных и реальных численностей видов (рис. 5.3.14, 3). Для построения аналогичного линейного графика взяты двоичные логарифмы ($\log_2 n$), что позволяет сопоставить эти иллюстрации с предыдущим разделом (рис. 5.3.15 и 5.3.11).

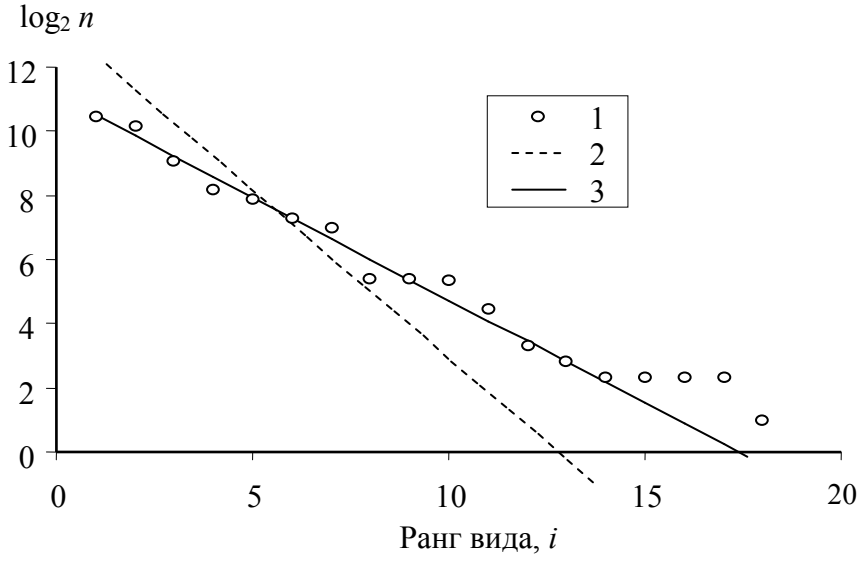


Рис. 5.3.15. Теоретическое геометрическое распределение численности видов мелких млекопитающих (по оси ординат – двоичные логарифмы численности $\log_2 n$): 1 – исходные данные (s_R), 2 – вариант с параметром $k = 0.5$, 3 – вариант с параметром $k = 0.3599$

Подходить к оценке значимости модели можно с позиций критерия χ^2 Пирсона, рассматривая полученные ряды как эмпирическое и теоретическое распределения. В этом случае оказывается, что они существенно не совпадают ($p << 0.001$), например, уже различия в частотах для второго вида дают величину $(1159 - 943)^2 / 943 = 49.4$, превышающую табличное $\chi^2_{(0.05,\ 16)} = 26.3$ (табл. 7С, стр. 357).

Если же рассматривать значения числа отловленных особей как величины, «на которые оказывают большое влияние различные случайные факторы» (Песенко, 1982; с. 66) (то есть как случайные величины в более общем смысле, а не как частоты в классах распре-

делений), то оценить адекватность модельного построения эмпирическим данным можно с помощью дисперсионного анализа линейной модели (Коросов, 2002). Значения численности имеют большой разброс, поэтому предварительно имеет смысл их прологарифмировать (взять $\log_2 n$). Общая сумма квадратов (логарифмов исходных значений численности) равна $C_{общ.} = 151.2$. Остаточная сумма квадратов (представленная в табл. 5.3.13 как $\Phi\log = \Sigma(\log_2 n - \log_2 *n)^2$) составляет $C_{остат.} = 10.48$. Отсюда модельная сумма квадратов равна $C_{мод.} = C_{общ.} - C_{остат.} = 151.2 - 10.\ 5 = 140.7$. При числе степеней свободы $df_{мод.} = 1$ и $df_{осотат.} = s - 2 = 18 - 2 = 16$ получаем дисперсии $S^2_{мод.} = 140.7$, $S^2_{остат.} = 10.48 / 16 = 0.66$. Критерий Фишера равен: $F = S^2_{мод.} / S^2_{остат.} = 0.66 / 140.7 = 214.7$. Полученное значение (214.7) превышает табличное $F_{(0.05,\ 1,\ 16)} = 4.49$ (табл. 5С, стр. 351), следовательно, модель адекватна исходным данным.

Рассчитать теоретические значения численности ранжированных видов можно и с помощью формулы, разработанной Мейем и Мотомурой (Лебедева и др., 2004):

$n_i = N \cdot C \cdot k \cdot (1-k)^{i-1}$, где $C = \left[1-(1-k)^S\right]^{-1}$ – поправочный коэффициент, близкий к единице $C \approx 1$.

Теоретическая численность, например, второго вида ($i = 2$) должна быть равной:

$$n_2 = N \cdot C \cdot k \cdot (1-k)^{i-1} = 4094 \cdot 1.00032 \cdot 0.359 \cdot (1-0.359)^{2-1} = 934.35 \text{ экз.}$$

Для определения истинного значения параметра k приходится также строить имитационную систему и вызывать макрос Поиск решения. Ее отличие от рассмотренной выше (рис. 5.3.13) состоит в том, что столбец $N_{ост.}$ не понадобится. Прочие операции те же.

Существует еще один способ представления геометрического распределения, вытекающий из того положения, что в упорядоченном ряду видовых численностей $n_1, n_2, \ldots n_i, n_{i+1}, \ldots n_S$ для любой смежной пары выполняется отношение $n_{i+1} / n_i = K$ (Песенко, 1982; Шитиков и др., 2003). Тогда запись закона примет вид:

$n_i = N_1 \cdot K^{i-1}$, где N_1 – число особей самого многочисленного вида, n_i – численность отдельного вида, K – доля вида относительно численности предыдущего вида, i – ранг (номер) вида.

Кажется, что здесь искажается идеологическая подоплека: параметр K характеризует не долю вида от оставшейся численности

сообщества (не захват оставшегося ресурса), а просто пропорцию численности данного вида относительно более старшего по рангу. В действительности же этот способ построения геометрического распределения также правомочен, поскольку доля уже захваченного и доля оставшегося ресурса дополняют друг друга до единицы ($N_{\text{оставш.}} + N_{\text{захвачен.}} = N_{\text{общая}}$), следовательно, параметры k и K в сумме образуют единицу: $k + K = 1$. Однако смысловую интерпретацию имеет лишь величина k.

В нашем примере настроенная величина параметра $k = 0.3599$ дает $K = 0.6401$. Расчетная численность первого вида составляет $N_1 = 1473$ (рис. 5.3.13), отсюда численность второго вида равна $n_2 = N_1 \cdot K^{2-1} = 1473 \cdot 0.6401 = 943.2$, что совпадает с расчетами по первой формуле.

Конкуренция мелких млекопитающих

Результаты анализа показывают, что для описания выравненности таксоценоза мелких млекопитающих Южного Прибайкалья, более всего подходит геометрическое распределение. Это относительно неожиданный вывод. Принято считать, что население мелких млекопитающих не представляет собой «сообщество», как функциональное единство видов, состав которого жестко детерминирован отношениями между ними. Обычно говорят о «биотопических группировках» как о наборе видов, наиболее характерных именно для данного местообитания (Ивантер, 1975). При общей слабой изученности этого вопроса, имеются лишь отдельные факты острой конкуренции между некоторыми видами серых полевок (Ивантер, Ивантер, 1986), мышами и полевками (Наумов, 1948), но и о бесконфликтном совместном обитании лесных полевок (Никитина, 1980). В перечне факторов динамики их численности конкурентные отношения упоминаются лишь вскользь (Максимов, 1984).

Складывается впечатление, что для разных видов мелких млекопитающих зоны лесов характерно широкое перекрывание ниш и слабая зависимость друг от друга. Таким отношениям должно соответствовать логнормальное распределение значимостей видов, чего в нашем случае не наблюдалось. Напротив, обобщенные за 8 лет данные хорошо соответствуют именно геометрическому распределению, связанному с конкуренцией за ограниченные ресурсы (в

нашем случае параметр распределения ниже величины $k = 0.5$ и соответствует не слишком жестким отношениям).

Относительно какого же общего ресурса конкурируют грызуны и бурозубки? В Южном Прибайкалье важнейшим кормовым ресурсом для всех видов мелких млекопитающих являются плоды (орешки) кедровой сосны. Исключая кротов, все виды изучаемой группы их поедают. В урожайные годы почти вся паданка (шишки, сбитые ветром) исчезает за 2 дня – ее съедают мелкие зверьки. Не раз отмечено, что в урожайные годы наблюдаются вспышки численности многих видов животных (Соколов, 1979). Хорошее питание сказывается не только на увеличении выживаемости зверьков (включая размножавшихся прибылых особей), но и стимулирует раннее весеннее созревание, позднее осеннее и даже зимнее размножение многих видов – рыжих полевок, мышей. Во многом благодаря кедровым орешкам численность всей группы, например, в Южном Прибайкалье поддерживается на относительно высоком уровне (в среднем около 12 экз./ 100 ловушко-суток). Здесь мы сталкиваемся с противоречием: условия обитания зверьков в регионе никак нельзя считать суровыми (что обычно ассоциируется с геометрическим распределением), и в то же время материал соответствует этой теории. На наш взгляд, все дело в силе указанного фактора, в периодическом режиме его проявления. В редкие годы урожая орехов (раз в 3–4 года) виды-специалисты (обитающие по большей части в местах произрастания кедра) успевают в гораздо большей степени использовать доступный благодатный ресурс (благоприятствует выживаемости, размножению, расселению), нежели редкие неспециалисты (мелкие виды бурозубок, грызуны – обитатели луговых и вторичных стаций). Относительно прочих видов численность первой группы резко увеличивается. Таким образом, геометрическое распределение численности разных видов мелких млекопитающих обусловлено периодически действующим *благоприятным* фактором, поднимающим численность некоторых видов в гораздо большей степени, чем большинства остальных. Оказывается, резкая невыравненность структуры сообществ может быть связана не только с жестким *лимитированием* необходимых средств существования, но и периодическим *редким стимулированием* крайне благоприятным фактором, воспользоваться которым в равной мере разные виды не в состоянии.

Индексы видового богатства

Описанные выше виды распределений есть лишь теоретические конструкции, позволяющие в некотором приближении рассмотреть общие причины сложной организации сообщества и охарактеризовать их с помощью параметров распределений (x, k, α). Зачастую хорошего описания эмпирическим сообществам не дает ни одна из рассмотренных моделей. Тогда общее представление о выравненности сообщества позволяют получить *индексы видового богатства*, которые, в конечном итоге, нужны для сопоставления разных сообществ друг с другом.

Простые индексы видового богатства

Здравый рассудок подсказывает аналогию для таких индексов – среднюю арифметическую, например, среднее число особей, представляющих один вид (N/s), или же среднее число видов, приходящихся на одну особь (s/N). Поскольку кривая доминирования–разнообразия всегда криволинейна, в формулы вводят логарифмы или квадратные корни. Широко известны, например, индексы видового богатства *Менхиника* $s/\sqrt{N}$ и *Маргалефа* $(s-1)/\ln(N)$, которые считаются лучшими из большого ряда подобных величин, предложенных биологами, поскольку менее других зависят от объема пробы. Последний индекс принимает максимальные значения, когда все виды представлены одной особью ($s = N$), а минимальное – при наличии всего одного вида в сообществе. В нашем случае он составил: $(18-1)/\ln(4094) = 2.04$.

Главный недостаток подобных индексов состоит в том, что они плохо отражают тот факт, что число зарегистрированных видов *криволинейно* зависит от величины исследованной площади, длительности наблюдений, интенсивности промысла и пр. (см. п. 5.1). Поэтому значения индексов видового богатства для слабо изученных сообществ будут близки к характеристикам ценозов, где видов действительно мало. Зависимость этих характеристик от объема пробы заставляет обращаться к другим показателям.

Суммативные индексы разнообразия

Общая идея этой группы мер видового разнообразия состоит в том, чтобы каким-либо образом одним числом выразить соотно-

шения значимостей (n_i, p_i и др.) всех s видов коллекции, то есть максимально емко описать кривую доминирования-разнообразия. Основная характеристика биологического разнообразия d должна отвечать нескольким требованиям. Из числа содержательных назовем следующие: 1) d возрастает при росте числа видов (и одинаковом характере кривой выравненности), 2) d возрастает при увеличении выравненности (и одинаковом числе видов); наибольшее значение получаем при полной выравненности (▯▯▯▯), а наименьшее – при полном доминировании одного вида (▯ _ _ _).

Опыт исследования ценозов показал, что кроме одного основного параметра удобно пользоваться несколькими производными величинами, которые дают разностороннюю биологическую интерпретацию наблюдаемым фактам. Если рассчитанный показатель разнообразия разделить на максимально возможную величину (когда значимость всех видов одинакова), получается *мера относительной выравненности* e с диапазоном возможных значений от 0 (доминирование одного вида) до 1 (равное доминирование всех). Поскольку выравненность e всегда меньше 1 (разнообразие d меньше максимального), можно найти величину S_d – *число видов гипотетического полностью выровненного сообщества* ($\ldots = p_i = p_j = \ldots$), имеющего меру разнообразия, равную d. Семейство таких индексов введено М. Хиллом (Песенко, 1982) и выражается общей формулой $S_d = \left(\sum^{S} p_i^{\alpha} \right)^{1/(1-\alpha)}$, где α – коэффициент. При $\alpha = 0$ имеем $S_d = s$ (видовое богатство), при $\alpha = 0.5$ имеем S_μ (мера Животовского), $\alpha = 1$ дает S_H (мера Шеннона), $\alpha = 2$ формирует S_C (индекс полидоминантности) (см. ниже). Так выявляются наиболее значимые (лидирующие) виды, определяющие с точки зрения данной меры основные черты структуры сообщества. Зная число S_d, можно рассчитать и *долю доминирующих видов* S_d / s, а также *долю редких видов*, как дополнение этой величины до единицы $1 - S_d / s$.

Индекс Макинтоша рассчитывает *евклидово расстояние* сообщества от начала координат многомерного пространства, оси которого заданы численностями видов:

$$U = \sqrt{(0-n_1)^2 + (0-n_2)^2 + \ldots + (0-n_i)^2 + \ldots + (0-n_S)^2} = \sqrt{\sum n_i^2} .$$

Индекс будет тем выше, чем больше доля многочисленных видов, то есть чем менее сообщество выровнено. С этой позиции индекс характеризует разнообразие сообщества.

При фиксированной общей численности индекс будет расти с уменьшением числа видов, поскольку вклад каждого вида будет относительно бóльшим. Максимум $U = N$ достигается при $s = 1$ (полное однообразие). С этой точки зрения индекс характеризует выравненность сообщества (характеристика относительной выравненности равна: $e = U / N$). На базе этого показателя выравненности построена *мера разнообразия*: $D = N - U = N - \sqrt{\sum n_i^2}$.

Максимум этой величины равен $D_{\max} = N - \sqrt{N}$. На этом основании Р. Макинтош предлагает независимую от численности N оценку относительного разнообразия: $d = \dfrac{N - \sqrt{\sum n_i^2}}{N - \sqrt{N}}$.

Для нашего примера имеем (расчеты не приводятся): N = 4094 экз., $\sqrt{N}$ = 63.98, U = 1940, D = 4094 – 1940 = 2154, d = 2154 / (4094 –63.98) = 0.536, e = 1940 / 4094 = 0.474.

Индекс Макинтоша критикуется как ориентированный на многочисленные виды, квадрат численности которых в основном и определяет его величину, тогда как малочисленные виды вносят существенно меньший информационный вклад (Шитиков и др., 2003). Так, если из нашего списка удалить два последних вида, относительный показатель разнообразия не изменится, d = 0.534, но стоит удалить два первых, как происходит резкое увеличение индекса: d = 0.844, хотя общая структура кривой доминирования-разнообразия не изменилась.

Индекс Симпсона ориентирован на относительные показатели выравненности видов (p_i) и рассматривается как мера «концентрации»: $C = \sum p_i^2$ или $C = \sum_{i=1}^{S} \left(\dfrac{n_i}{N} \right)^2$.

Для описания эмпирических коллекций предлагается формула, дающая несмещенную оценку: $C' = \sum_{i=1}^{S}\left(\frac{n_i(n_i-1)}{N(N-1)}\right)$, но при достаточно больших выборках (сотни особей) можно пользоваться первой формулой. Статистическая ошибка показателя примерно равна: $m_C \approx \frac{4}{N}\left[\sum p_i^3 - \left(\sum p_i^2\right)^2\right]$.

Исследования свойств этой величины показали, что максимум $C = 1$ она достигает при одновидовом сообществе $s = 1$, а к минимуму $C \to 0$ показатель стремится при $s \to \infty$. Как и в предыдущем случае, индекс Симпсона (за счет возведения значимостей в квадрат) несколько преувеличивает значение многочисленных видов и снижает роль редких.

Индекс имеет отчетливый статистический смысл. Если из нашей выборки видов, входящих в сообщество, извлекать случайным образом (и не возвращать обратно) по одной особи, то величина $p_i = \frac{n_i}{N}$ по существу есть вероятность обнаружения особи *i*-го вида. Величина $p_i = \frac{(n_i-1)}{(N-1)}$ есть вероятность обнаружения особи того же вида после изъятия одной особи. Тогда сумма произведений этих вероятностей $\sum_{i=1}^{S}\left(\frac{n_i(n_i-1)}{N(N-1)}\right)$ есть вероятность того, что две особи, взятые из всей совокупности подряд, будут принадлежать к одному из *s* видов, то есть это *вероятность внутривидовых встреч* в смежных пробах. По этой логике получается, что чем меньше в сообществе видов, тем больше будет вероятность *внутри*видовых встреч, то есть тем меньше разнообразие. Значит, рост индекса *C* означает снижение биоразнообразия (Песенко, 1982). Характеристикой разнообразия тогда должна служить *вероятность межвидовых встреч*, составляющая разность между единицей и вероятностью внутривидовых встреч (1 – *C*). Следовательно, мерой разнообразия должна быть разность между максимально возможным и рассчитанным значением индекса: $d = 1 - \sum p_i^2$, ошибка составляет $m_d = 2m_C$.

В генетике подобную формулу используют для расчета теоретической доли гетерозигот (Животовский, 1991), то есть как оценку вероятности *межаллельных* встреч. Максимальное значение этого индекса составляет $d_{\max} = (s - 1) / s$.

Выполним расчеты для нашего примера (табл. 5.3.2):

$$C = \sum p_i^2 = 0.225 \approx C' = \sum_{i=1}^{S}\left(\frac{n_i(n_i - 1)}{N(N-1)}\right) = \frac{3759314}{16756742} = 0.224,\ d = 0.775.$$

Обратная от коэффициента Симпсона величина рассматривается как *индекс полидоминантности* – число видов некой теоретической выборки, имеющей такое же разнообразие (C), как изучаемая эмпирическая коллекция, но в которой все виды имеют одинаковую значимость: $S_C = \frac{1}{C} = \frac{1}{\sum p_i^2} = \left(\sum p_i^2\right)^{-1}$. Показатель S_C служит оценкой «эффективного» числа видов, образующих сообщество. Максимальное значение $S_{C\max} = s$ имеет коллекция при полной выравненности (условие равной значимости всех видов, $p_i = 1/\mathrm{s}$). В случае наибольшей невыравненности (один вид доминирует, остальные представлены одной особью) $S_{C\min} = N^2 / \left[(N-s)^2 + 2N - s\right]$.

В нашем примере (табл. 5.3.2) показатель имеет значение $S_C = 1/\ 0.224 = 4.45$, то есть при той же численности 4–5 абсолютно выровненных гипотетических видов имеют такое же разнообразие, как 18 наблюдаемых видов. Отсюда доля доминирующих видов равна: $S_C/\mathrm{s} = 4.45/18 = 0.247$, доля редких видов составит: $h_C = 1 - S_C / s = 1 - 4.45/\ 18 = 0.753$. В отличие от меры Макинтоша удаление из списка двух первых видов не вызывает кардинальных изменений индекса $S_C = 4.98$, $h_C = 0.78$.

Индекс Л. А. Животовского (1982) хоть и предлагается для характеристики полиморфизма популяции, но по аналогии вполне пригоден и для описания видового разнообразия: $S_\mu = \left(\sqrt{\sum p_i}\right)^2$, статистическая ошибка: $m_{S\mu} = \sqrt{\mu \cdot (m - \mu)/N}$. Показатель выражает «среднее число фенотипов (видов)» с учетом их встречаемости. Максимальное значение $S_\mu = s$ индекс принимает при полной выравненности сообщества, в общем случае $S_\mu < s$.

Таблица 5.3.2. Расчет разных индексов видового богатства для коллекции мелких млекопитающих с южного берега оз. Байкал

i	n_i	p_i	p_i^2	p_i^3	$n_i \cdot(n_i-1)$	$\sqrt{p_i}$	$p_i \cdot \log_2(p_i)$	$p_i \cdot \ln(p_i)$
1	1394	0.3405	0.116	0.039	2E+06	0.584	–0.529	–0.37
2	1159	0.2831	0.08	0.023	1E+06	0.532	–0.515	–0.36
3	535	0.1307	0.017	0.002	285690	0.361	–0.384	–0.27
4	293	0.0716	0.005	4E–04	85556	0.268	–0.272	–0.19
5	239	0.0584	0.003	2E–04	56882	0.242	–0.239	–0.17
6	159	0.0388	0.002	6E–05	25122	0.197	–0.182	–0.13
7	128	0.0313	1E–03	3E–05	16256	0.177	–0.156	–0.11
8	43	0.0105	1E–04	1E–06	1806	0.102	–0.069	–0.05
9	42	0.0103	1E–04	1E–06	1722	0.101	–0.068	–0.05
10	41	0.01	1E–04	1E–06	1640	0.1	–0.067	–0.05
11	22	0.0054	3E–05	2E–07	462	0.073	–0.041	–0.03
12	10	0.0024	6E–06	1E–08	90	0.049	–0.021	–0.01
13	7	0.0017	3E–06	5E–09	42	0.041	–0.016	–0.01
14	5	0.0012	1E–06	2E–09	20	0.035	–0.012	–0.01
15	5	0.0012	1E–06	2E–09	20	0.035	–0.012	–0.01
16	5	0.0012	1E–06	2E–09	20	0.035	–0.012	–0.01
17	5	0.0012	1E–06	2E–09	20	0.035	–0.012	–0.01
18	2	0.0005	2E–07	1E–10	2	0.022	–0.005	–0
Сумма	4094	1	0.225	0.065	4E+06	2.99	–2.612	–1.81
$N\cdot(N-1)$					2E+07			
C			0.225		0.224	H	2.612	1.810
m_C					0.00001	S_H	6.112	6.112
d			0.775		0.776	e_H	0.626	0.626
m_d					0.00003	h_H	0.66	0.66
S_C			4.454		4.457			
S_μ						8.94		
$m_{S\mu}$						0.141		
h						0.503		
m_h						0.008		

Именно Л. А. Животовский в качестве дополнительной характеристики структуры распределения видов (морф) предложил показатель «доля редких форм»: $h = 1 - S_\mu / s$; ошибка равна $m_h = m_{S\mu} / s$.

В примере имеем (табл. 5.3.2): $S_\mu = (2.99)^{0.5} = 8.94$, $m_{S\mu} = \sqrt{8.94 \cdot (18 - 8.94)/4094} = 0.14$, $h = 1 - S_\mu / s = 0.503$, $m_h = 0.08$.

Этот показатель по-иному распределяет роли: девять видов выборки считаются доминирующими (имеют довольно солидную представленность – более 40 экз.), а виды из оставшейся части списка с низкой численностью рассматриваются как редкие. Удаление из списка двух первых видов слабо отразилось на показателях: $S_\mu = 9.33$, $h = 0.48$.

Индекс Шеннона в теории информации определяется как *мера неопределенности*, что вполне аналогично понятию структурного разнообразия. Общая формула имеет вид: $H = -\sum p_i \cdot \log p_i$.

Для выяснения истоков этой достаточно сложной формулы рассмотрим понятие *неопределенности результата эксперимента*, который может иметь s разных исходов. Мера должна быть нулевой при отсутствии неопределенности, то есть при наличии полной определенности, когда возможен только единственный исход $s = 1$, и должна увеличиваться при росте возможных исходов (при $s \to \infty$).

Такими свойствами обладает логарифм числа исходов $\log(s)$ ($\log(1) = 0$, $\log(s) > 0$ при $s > 1$). Изначально (в теории сигналов) использовался двоичный логарифм $\log_2(s)$, который в случае двух исходов равен единице $\log_2(2) = 1$ (это двоичная единица информации, бит). Например неопределенность эксперимента, имеющего 18 возможных исходов, равна $\log_2(18) = 4.17$. Переводя разговор в русло исследования многовидовых сообществ, можно сказать, что эта мера характеризует ситуацию с отловом (учетом) очередной особи, которая может принадлежать к одному из 18 видов (как в нашем примере). Если все виды имеют одинаковую значимость (например, численность), то общая неопределенность того, какой же вид будет обнаружен, выражается этой формулой: $H = \log_2(18) = 4.17$. В действительности приходится учитывать, что разные виды имеют не-

одинаковую значимость (численность), значит, и разную вероятность попасть в уловы. Для этого приходится явно вычислять ту долю общей неопределенности ситуации, которая приходится на каждый *i* вид. При равенстве значимостей доля одного вида составляет $\frac{1}{s}$, а доля неопределенности, приходящейся на один вид, составит: $H_i = \frac{1}{s}H = \frac{1}{s}\log_2(s)$. Из правил логарифмирования известно, что $\log(s) = -\log\left(\frac{1}{s}\right)$. Тогда имеем: $H_i = -\frac{1}{s}\log_2\left(\frac{1}{s}\right)$.

На практике значимость видов оценивается с использованием показателей обилия, биомассы и пр., то есть как отношение видовой характеристики к суммарной: $p_i = \frac{n_i}{N}$, при равенстве значимостей имеем: $\frac{1}{s} = p_1 = ... = p_i = ... = p_S$. Тогда неопределенность обнаружения одного вида составит $H_i = -p_i \log_2 p_i$, а неопределенность всей ситуации в целом, точнее, полное разнообразие возможных исходов наблюдений, и составляет искомый индекс разнообразия Шеннона $H_2 = -\sum p_i \log_2 p_i$ (индекс $_2$ указывает на использование двоичного логарифма). Для 18 равнозначных видов имеем: $p_i = \frac{1}{18} = 0.0556$, $H_i = -0.0556 \cdot \log_2(0.0556) = 0.232$, $H = \sum^{18} H_i = 4.17$.

Иными словами, при полностью выровненном сообществе индекс Шеннона равен своему максимальному значению – логарифму числа видов log(*s*). Когда же значимости видов отличаются и выявляется группа немногих доминирующих видов, то ситуация становится *менее неопределенной* – при отловах, скорее всего, будут попадаться особи многочисленных видов. Иными словами, возможное разнообразие исходов ситуации становится меньше, что и будет отображать индекс: если $n_1 > ... > n_i > \ldots > n_S$, то $H < \log_2(s)$. Для нашего примера (табл. 5.3.2) индекс видового богатства Шеннона равен $H_2 = 2.612 < 4.17$.

Биологический смысл полученной величины можно пояснить так. По сути дела, индекс Шеннона – это логарифм, то есть показатель *степени*, в которую нужно возвести *основание* для получения некоего *числа*: $\log_{\text{основания}}$ (*число*) = *степень*. Чтобы получить число, выполняется обратная процедура: *число* = *основание* $^{\text{степень}}$. Какое же число получится, если основание 2 возвести в степень 2.612? Имеем $2^{2.612} = 6.11$ или $S_H = 2^H$. Нами рассчитано число видов (S_H) в гипотетическом *абсолютно выровненном* сообществе ($p_i = p_j \ldots$), имеющем такой же показатель разнообразия (H), что и у изучаемого эмпирического сообщества. Можно сказать и так, что S_H – это группа лидирующих (доминирующих, эффективных) видов, определяющих в главных чертах облик изучаемого ценоза.

Здесь уместно указать на роль разных оснований логарифмов (2, *e*, 10). Двоичные логарифмы ($\log_2 x$) имеют отношение к теории передачи двоичного сигнала (0 и 1) и удобны для анализа дихотомии, но в биологических приложениях отходят в сторону. Десятичные логарифмы ($\log_{10} x = \lg x$) ассоциированы с десятичной шкалой и используются чаще, особенно в иллюстративных целях. Натуральные логарифмы ($\log_e x = \ln x$) связаны с тригонометрическими функциями и используются шире предыдущих, в частности, формулы статистического сравнения индексов Шеннона базируются именно на них. При использовании разных оснований логарифмов (в табл. 5.3.2 рассмотрены расчеты для $\log_2 x$ и $\ln x$) значения индекса Шеннона *будут отличаться* (это важно иметь в виду при сравнении индексов из разных литературных источников). Но основанные на индексе H показатели числа «эффективных» видов S_H, относительной выравненности сообщества e_H и «доля редких видов» h_H примут идентичные значения (табл. 5.3.2).

$$S_H = 2^{H_2} = e^{H_e} = 10^{H_{10}},$$

$$e_H = \frac{H_2}{\log_2(s)} = \frac{H_e}{\ln(s)} = \frac{H_{10}}{\lg(s)},$$

$$h_H = 1 - \frac{S_H}{s}.$$

В примере для двоичных логарифмов получили $H = 2.61$, для натуральных $H = 1.81$. Число эффективных видов будет равно

$S_H = 2^{2.61} = e^{1.81} = 6.1$, относительная выравненность составляет $e_H = \frac{2.61}{\log_2(18)} = \frac{1.81}{\ln(18)} = 0.66$, доля редких видов $h_H = 1 - \frac{6.1}{18} = 0.66$.

Удаление из изучаемой выборки двух наиболее многочисленных видов слабо повлияло на индекс Шеннона, но выравненность немного увеличилась: $H_e = 1.91$, $S_H = 6.74$, $h_H = 0.63$.

С помощью индекса Шеннона можно проводить статистическое сравнение двух сообществ. В этом случае сравнивается сам характер выравненности коллекции, а не доля тех или иных видов в сравниваемых группировках (сравнивается ход кривых доминирования-выравненности, а не два частотных распределения; методы детального сопоставления распределений представлены в следующем разделе).

Для примера сравним логарифмические характеристики ($H_e = -\sum p_i \ln p_i$) выравненности двух кривых доминирования-разнообразия – коллекций животных, полученных при отловах давилками и канавками. Для этого используется критерий Стьюдента:

$$t = (H_1 - H_2) / \sqrt{m_{H_1}^2 + m_{H_2}^2},$$

где $m_H^2 = \left((N \ln^2 N - \sum n_i \ln n_i)/N - H^2 + (s-1)/2N^2\right)/N$ – статистическая ошибка.

Табличное значение $t(\alpha, df)$ (табл. 4С, стр. 351) отыскивается для данного уровня значимости (обычно $\alpha = 0.05$) и числа степеней свободы $df = N_1 \cdot N_2 \cdot (m_{H_1}^2 + m_{H_2}^2)^2 / (N_2 \cdot m_{H_1}^4 + N_1 \cdot m_{H_2}^4)$.

Список видов, отловленных в канавки, был отсортирован по числу особей и пронумерован (i); эти же номера сохранены за видами, отловленными в давилки, также отсортированными по убыванию числа особей. Ориентируясь на номера, можно увидеть, что ранги видов в разных списках не совпадают: виды, часто отлавливаемые в канавки, могут редко попадаться в давилки и наоборот. Например, красная полевка в первом списке имеет номер (ранг) 4, а во втором выходит на первое место. Лидер попадаемости в канавки (красно-серая полевка) занимает вторую позицию в рейтинге по давилкам.

Таблица 5.3.3. Расчет и сравнение индексов видового богатства Шеннона двух выборок мелких млекопитающих, отловленных в канавки и давилки

Отлов в канавки					Отлов в давилки				
i	n_i	n_i^2	p_i	$p_i \cdot \ln(p_i)$	i	n_i	n_i^2	p_i	$p_i \cdot \ln(p_i)$
1	295	1678	0.2373	-0.34	4	859	5803	0.3801	-0.3677
2	231	1257	0.1858	-0.31	1	794	5302	0.3513	-0.3675
3	179	929	0.144	-0.28	6	352	2064	0.1558	-0.2896
4	135	662	0.1086	-0.24	3	113	534.2	0.05	-0.1498
5	132	645	0.1062	-0.24	8	79	345.2	0.035	-0.1172
6	89	399	0.0716	-0.19	5	27	88.99	0.0119	-0.0529
7	40	148	0.0322	-0.11	10	9	19.78	0.004	-0.022
8	37	134	0.0298	-0.1	2	8	16.64	0.0035	-0.02
9	36	129	0.029	-0.1	11	5	8.047	0.0022	-0.0135
10	33	115	0.0265	-0.1	18	5	8.047	0.0022	-0.0135
11	22	68	0.0177	-0.07	7	3	3.296	0.0013	-0.0088
12	5	8.05	0.004	-0.02	9	3	3.296	0.0013	-0.0088
13	3	3.3	0.0024	-0.01	16	2	1.386	0.0009	-0.0062
14	2	1.39	0.0016	-0.01	17	1	0	0.0004	-0.0034
15	2	1.39	0.0016	-0.01	12	0		0	
16	1	0	0.0008	-0.01	13	0		0	
17	1	0	0.0008	-0.01	14	0		0	
18	0		0		15	0		0	
Сумма	1243	6177	1	-2.16		2260	14198	1	-1.441
s =	17		H_K	2.16	s =	15		$H_Д$	1.44
			S_H	8.63				S_H	4.22
			e_H	0.761				e_H	0.532
			h_H	0.492				h_H	0.72

После расчетов необходимых промежуточных сумм вычислим статистические ошибки индексов:

$$m_K^2 = \left((N_K \ln^2 N_K - \sum n_{iK} \ln n_{iK}) / N_K - {H_K}^2 + (s_K - 1)/2{N_K}^2\right) / N_K =$$

= [(1243 · 50.77 – 6177) / 1243 – 4.67 + (17 – 1) / 2· 1549049] / 1243 =

= 0.033, $m_Д^2$ = 0.023 и критерий Стьюдента:

$$t = (H_K - H_Д) / \sqrt{m_K^2 + m_Д^2} = (2.16 - 1.44) / \sqrt{0.033 + 0.023} = 3.02.$$

Для числа степеней свободы:

$$df = 1243 \cdot 2260 \cdot (0.033 + 0.023)^2 / (2260 \cdot 0.033^2 + 1243 \cdot 0.023^2) = 2278$$

табличное значение составит $t_{(0.05,\ 2278)} = 1.96$. Полученная величина (3) больше табличной (1.96), значит, различия между индексами разнообразия достоверны. Коллекции по отловам канавками гораздо более выровнены, чем при использовании давилок (рис. 5.3.16); на это указывают и высокий индекс относительной значимости (0.76 против 0.53) и обширная группа лидирующих доминантов (8.6 против 4.2).

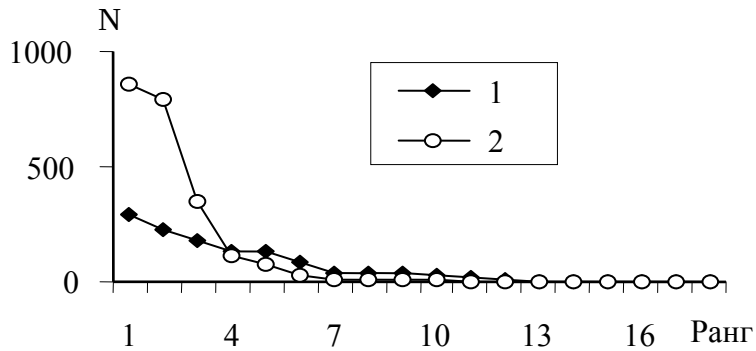


Рис. 5.3.16. Кривые доминирования-разнообразия, построенные по результатам отлова мелких млекопитающих Южного Прибайкалья в канавки (1) и давилки (2)

Причина столь резких отличий сравниваемых выборок, взятых из одной и той же генеральной совокупности (населения мелких млекопитающих данной территории), в первую очередь состоит в том, что давилки конструктивно рассчитаны на довольно крупных грызунов (массой около 15 г), поэтому мелкие виды (бурозубки, мышовки) попадаются в них много реже; канавки же отлавливают всех перемещающихся некрупных животных. Кроме того, давилки

«работают» 2–3 дня и отлавливают зверьков, живущих в непосредственной близости от поставленной линии ловушек, то есть характеризуют численность зверьков на данной территории. Канавки же открыты в течение минимум 10 дней и собирают в себя не только жителей данного местообитания, но и всех мигрантов, проходящих через биотоп; по этой причине канавки регистрируют интенсивность миграционных потоков. Вот основные причины, по которой обычно немногочисленные, очень подвижные мелкие насекомоядные виды в большей мере представлены в отловах канавками, чем давилками (ранги 4–9 в табл. 5.3.3 и на рис. 5.3.16).

Сравнение мер разнообразия

Обстоятельные обзоры достоинств и недостатков рассмотренных нами и многих других мер разнообразия можно найти в литературе (Песенко, 1982; Животовский, 1984; Шитиков и др., 2003). Общий вывод состоит в том, что лучшими статистическими свойствами обладают лишь семейство индексов Симпсона и Животовского. Индекс Макинтоша критикуется в связи с тем, что он сильно преувеличивает роль доминирующих видов, индекс Шеннона явно зависит от добавления новых даже немногочисленных видов. В целом для получения полноценной картины рекомендуется рассчитывать несколько разных индексов. В этой связи имеет смысл сопоставить индексы с точки зрения их содержательной интерпретации, ориентируясь на результаты наших расчетов (табл. 5.3.2, 5.3.4).

Исходные показатели разнообразия (U, C, H) почти не поддаются интерпретации непосредственно, удобнее сравнивать производные величины (S_d, e, h). Так, мера Симпсона дает минимальное число «лидирующих» видов S_d (4.5), мера Животовского – максимальное (8.9). Причина состоит в том, что в структуре индекса Симпсона фигурирует квадрат значимостей видов, резко усиливающий большие значения, поэтому доминирующее положение остается за немногими, наиболее многочисленными видами. В индексы Шеннона и Животовского включены другие преобразования исходных значимостей – логарифм и квадратный корень, которые, напротив, сообщают дополнительный вес немногочисленным видам. Логарифм сильнее сокращает разрыв между числами, чем корень, но в индексе Шеннона он умножается на само значение выравненности, поэтому результат оказывается сглаженным. По этой причине эф-

фективное число лидирующих видов меры Шеннона (6.1) ниже показателя Животовского (8.9).

Таблица 5.3.4. Индексы видового богатства микромаммалий

Меры	Макинтоша	Симпсона	Шеннона	Животовского
d	0.54	0.77	1.8	-
S_d		4.5	6.1	8.9
e	0.47	-	0.63	-
h		0.75	0.66	0.50

Таким образом, каждый индекс «искажает» исходные выравненности по-своему, сообразно с использованными математическими процедурами. Выбор определенной меры для применения в конкретном исследовании, поэтому, должен включать в себя ответ на вопрос: что в рамках цели работы представляет наибольшую важность – акцент на группе лидирующих видов или весомое включение в анализ средне- и малочисленных видов. В первом случае следует пользоваться мерой Симпсона, во втором – показателями Шеннона или Животовского. С этих позиций становится понятным и критикуемое увеличение меры H при включении в коллекцию немногочисленных видов – она и ориентирована на учет видов в первую очередь со средней значимостью. Поэтому отмеченный недостаток является, скорее всего, достоинством индекса Шеннона. Сравнивая показатели, подчеркнем ясное теоретическое содержание мер Симпсона и Шеннона и отсутствие идеологической подоплеки для меры Животовского. Однако расчеты последней характеристики намного проще прочих, да, пожалуй, она в большей мере, чем остальные рассмотренные показатели, соответствует интуитивному образу изучаемого распределения.

5.4. Выравненность: β-разнообразие

Изменение показателей выравненности сообществ в контексте изменения местообитаний (по локализации или градиентам факторов среды) обозначается как β-разнообразие. Типичная задача – сопоставление двух или нескольких коллекций с учетом значимости

входящих в них видов. Следует отметить два главных аспекта этого анализа.

Во-первых, может оцениваться *пересечение коллекций*, когда интересны абсолютные отличия между сообществами и виды сравниваются по численности. Во-вторых, важно оценить *пересечение структуры коллекций*, когда интересны пропорции между видами в сравниваемых группировках; для этого сопоставляются показатели доминирования (доли, значимости) разных видов. На решение первой задачи направлены *индекс Чекановского* и *меры расстояния*, вторая задача решается с помощью коэффициента *общности*, показателя *доли общих форм* и коэффициента *корреляции*.

Сквозным примером для всех наших расчетов послужат результаты отловов мелких млекопитающих в канавки (табл. 5.4.1).

Таблица 5.4.1 Отловы мелких млекопитающих в 1983–1986 гг. в разных биотопах Южного Прибайкалья (экз./ 100 конусо-суток)

Биотоп	Обыкновенная бурозубка	Средняя бурозубка	Малая бурозубка	Равнозубая бурозубка	Лесная мышовка	Азиатская мышь	Лесной лемминг	Полевка темная	Полевка-экономка	Полевка красная	Полевка красно-серая	Всего
Кедровник	6	9	8	4	0.3	8	1	0	0	9	7	52
Пихтач	8	15	4	2	0.3	6	1	0.5	0.5	6	10	53
Экотон	6	3	4	1	0.3	3	2	2	1	5	12	39
Сосняк	8	1	2	0	0	1	1	0	0	5	15	33
Березняк	10	3	4	1	0.3	1	1	1	1	4	12	38
Луг	4	0	4	0.4	0.2	0.2	1	1	4	0.4	3	18
В среднем	7	5	4	1	0.3	3	1	1	1	5	10	38

Показатели пересечения коллекций

В отличие от задачи сравнения двух простых множеств видов, рассмотренной в разделе 5.2, здесь формулируется задача срав-

нения двух *дискриптивных* множеств, в которых каждый элемент (вид) имеет определенный «вес» (значимость, численность).

Решение этой задачи состоит в следующем. Сначала оценивается степень пересечения каждого общего элемента (вида) сравниваемых коллекций, а затем эти оценки объединяются для всех элементов. Мерой пересечения двух множеств особей одного вида служит *меньшая численность*, которая как бы включается в большее значение, она обозначается $\min(N_{1i}, N_{2i})$.

Например, если обыкновенная бурозубка имеет в кедровнике численность 6, а в пихтаче 8 экз./100 к-с, то пересечение составляет $\min(6, 8) = 6$ экз., второе сообщество на 2 экз. богаче первого. Используя формулу Съёренсена (см. раздел 5.2), нетрудно посчитать, что сходство сообществ относительно обыкновенной бурозубки равно: $2 \cdot 6 / (6 + 8) = 12 / 14 = 0.86$. Обобщение для всех видов дает *коэффициент общности* (или *процентное сходство*) *Чекановского*:

$$K = \frac{2 \cdot \sum^{S} \min(N_{1i}, N_{2i})}{N_1 + N_2},$$

где $N_1 + N_2$ – суммарная численность (обилие) всех видов из обеих коллекций, $\min(N_{1i}, N_{2i})$ – меньший показатель численности *i*-го вида, выбранный из двух значений, относящихся к двум сравниваемым коллекциям, $\sum^{S} \min(N_{1i}, N_{2i})$ – сумма минимальных значений численности по всем *s* видам.

Область возможных значений этого коэффициента ограничивается нулем (отсутствие сходства) и единицей (полное сходство).

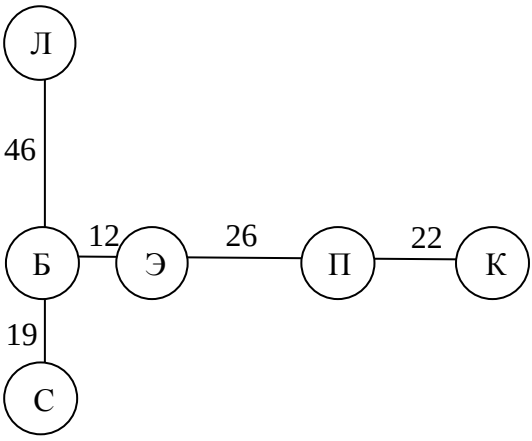
Сопоставляя пары оценок численности разных видов в кедровнике и пихтаче (табл. 5.4.1), получаем ряд минимальных значений: 6, 9, 4, 2, 0, 6, 1, 0, 0, 6, 7, которые дают сумму $\sum^{S} \min(N_{1i}, N_{2i}) = 41$. Суммарная численность всех видов в этих двух биотопах равна $N_1 + N_2 = 52 + 53 = 106$. Тогда индекс Чекановского составит: $2 \cdot 41 / 106 = 0.78$. Сравнения между всеми биотопами порождают таблицу парного сходства сообществ по выравненности (табл. 5.4.2) и графы отношения близости (рис. 5.4.1, 5.4.2).

Таблица 5.4.2. Процентное сходство K и различие d_K биотопических группировок мелких млекопитающих (см. табл. 5.4.1) (петитом и курсивом отмечены минимальные расстояния)

	K				
К	*0.78*	0.66	0.54	0.60	0.37
	П	*0.74*	0.65	0.73	0.40
		Э	*0.77*	**0.88**	0.53
			С	**0.81**	0.41
				Б	0.54
					Л

$d_K = (1 - K) \cdot 100$					
К	**22**	34	46	40	63
	П	**26**	35	27	60
		Э	23	**12**	47
			С	**19**	59
				Б	**46**
					Л

Рис. 5.4.1. Коррелограмма сходства биотопов по численности видов мелких млекопитающих (процентное сходство); линии отмечают сходство на уровне 0.8, пунктир – на уровне 0.7



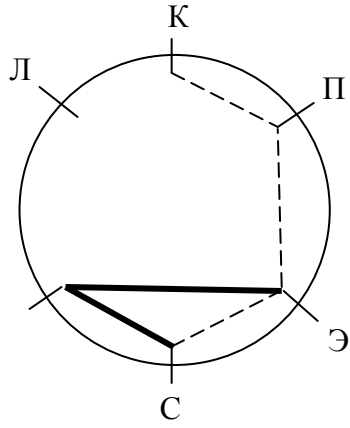


Рис. 5.4.2. Дендроид различия биотопов по численности видов мелких млекопитающих, рассчитанный по процентному сходству

Довольно большие значения попадаемости во всех биотопах вылились в их высокое сходство друг с другом на уровне 0.7 (рис. 5.4.1), однако дерево расстояний (рис. 5.4.2) хорошо подчерки-

вает характерные различия: группа коренных биотопов (К и П) ясно отделилась от вторичных (Э, С, Б) и значительно – от луга (Л).

Сопоставляя дендрограммы различия биотопических группировок мелких млекопитающих, построенные по видовым спискам (рис. 5.2.2) и коллекциям (рис. 5.4.2), можно увидеть как черты сходства (естественные и слабонарушенные биотопы П, Э, К располагаются обособленно от вторичных Б, Л, Г), так и отличия (производные сосняки С отчетливо отошли к группе вторичных биотопов). Использование показателей значимостей видов более рельефно отобразили именно наиболее существенные отношения между сообществами, тогда как на результаты сравнения только по видовому богатству большое влияние оказали редкие виды (в лесу – темная полевка, крысы, на лугах – красная, красно-серая полевки), хотя и представленные в коллекциях, но лишь редкими случайными экземплярами.

От прочих показателей сходства мера Чекановского отличается лучшими статистическими свойствами, простотой интерпретации и техники расчетов (Песенко, 1982).

Оценка расстояний между распределениями

Помимо показателей различия между видовыми списками двух локальных территорий, вычисленных как дополнение меры сходства до единицы ($1 - K$), имеется большой набор мер расстояния, имеющих собственное теоретическое толкование. Преследуя целью найти абсолютное различие между коллекциями, используются характеристики значимости видов, имеющие единицы измерения (численность, биомасса, индекс плотности и др.). Наиболее распространена *мера Минковского*: $d_{12} = \left(\sum^{S} \left| N_{1i} - N_{2i} \right|^{p} \right)^{1/r}$,

где 1 и 2 – номера сравниваемых коллекций, i – номер вида в списке из s видов, N_i – значимость i-го вида, p, r – константы, определяющие некоторые свойства меры.

Для случая $p = r = 1$ получаем простую характеристику – *манхэттенское расстояние* (или *сити-блок метрика*, *хеммингово расстояние*): $d_{12} = \sum^{S} \left| N_{1i} - N_{2i} \right|$. Название отражает специфику из-

мерения расстояния между двумя точками на плоскости, которые можно рассматривать как места расположения двух объектов на двух перекрестках улиц города (Нью-Йорка). Тогда расстояние между ними будет равно длине двух улиц, которые предстоит пройти, чтобы от одного объекта добраться до другого. Другая интерпретация: два объекта на плоскости есть две вершины прямоугольного треугольника, расстояние между ними есть сумма катетов.

Использование коэффициентов $p = r = 2$ дает *евклидово расстояние*: $d_{12} = \sqrt{\sum^{S}\left(N_{1i} - N_{2i}\right)^2}$, которое проще всего интерпретировать как гипотенузу прямоугольного треугольника.

Минимальное значение расстояния ($d = 0$) получается при сравнении идентичных по выравненности коллекций, максимальное значение не определено.

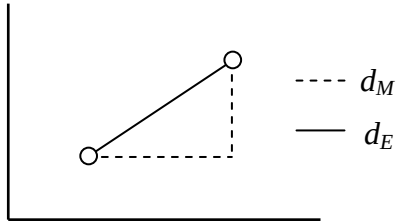


Рис. 5.4.3. Меры расстояний между двумя коллекциями по двум видам: d_M – манхэттенская, d_E – евклидова

Рассмотрим расчеты этих метрик относительно нашего примера (табл. 5.4.3). Ориентируясь на таблицу исходных данных, находим манхеттэново расстояние между кедровником и пихтачем для обыкновенной бурозубки $d_{\text{КП1}} = \left|N_{\text{К1}} - N_{\text{П1}}\right| = |6 - 8| = 2$. Остальные виды дают: |9 – 15| = 6, 4, 2, 0, 2, 0, 1, 1, 3, 3, в сумме $d_{\text{КП}} = 23$ и т. д. Проведя аналогичные расчеты для евклидовой меры, имеем: $d_{\text{КП1}} = \left(6-8\right)^2 = 4$, 36, 16 …, всего 83, $\sqrt{83} \approx 9$ и т. д. В таблице всех парных расстояний нетрудно выбрать наименьшие (метод ближайших соседей, см. раздел 5.2) и построить графы сходства биотопических группировок (рис. 5.4.4). Их анализ почти ничего не добавляет к картине, полученной с помощью меры Чекановского (рис. 5.4.2). Можно отметить лишь, что евклидова мера по сравнению с манхэттенской *увеличивает* большие дистанции: кратность d_E расстояний между парами Л–Б и Б–С, равная 12 / 5 = 2.4, больше, чем кратность d_M расстояний 26 / 13 = 2.

Таблица 5.4.3. Манхэттенское d_M и евклидово d_E расстояния между биотопическими группировками мелких млекопитающих, рассчитанные по табл. 5.4.1 (петитом отмечены минимальные расстояния)

	d_M				
К	**23**	31	39	36	44
	П	**24**	30	25	43
		Э	16	**9**	27
			С	**13**	30
				Б	**26**
					Л

	d_E				
К	**9**	**12**	16	13	17
	П	13	16	14	19
		Э	6	**5**	12
			С	**5**	14
				Б	**12**
					Л

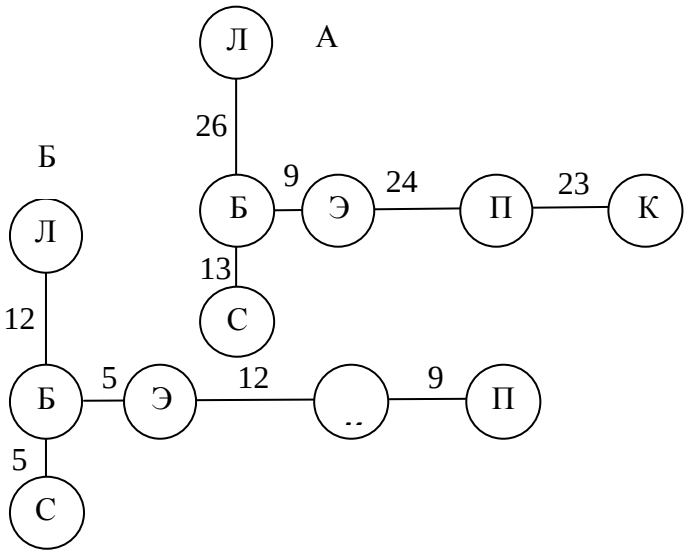


Рис. 5.4.4. Деревья расстояний между группировками мелких млекопитающих: А – манхэттенские, Б – евклидовы

Показатели пересечения структуры коллекций

Локальные флористические или фаунистические описания зачастую сравниваются с целью найти сходство в соотношениях, пропорциях входящих в них видов, поскольку отличия по численности можно получить другими методами, имеющими статистическое обоснование (см. п. 3.2). Задача сравнения структуры коллекций

решается с использованием относительной значимости видов, которая обычно есть просто отношение численности данного вида к общей численности всех видов $p_i = n_i / N$ (см. п. 5.3). Перед расчетами мер сходства следует построить таблицу относительных значений численности (табл. 5.4.4) по данным исходной численности (табл. 5.4.1). Например, доля обыкновенной бурозубки в отловах в кедровнике составляет $p_1 = 6 / 52 = 0.12$.

Таблица 5.4.4. Отловы мелких млекопитающих в разных биотопах (доля от суммы, %)

Биотоп	Обыкновенная бурозубка	Средняя бурозубка	Малая бурозубка	Равнозубая бурозубка	Лесная мышовка	Азиатская мышь	Лесной лемминг	Полевка темная	Полевка-экономка	Полевка красная	Полевка красно-серая	Всего
Кедровник	0.12	0.17	0.15	0.08	0.01	0.15	0.02	0.00	0.00	0.17	0.13	1.00
Пихтач	0.15	0.28	0.07	0.04	0.01	0.11	0.02	0.01	0.01	0.11	0.19	1.00
Экотон	0.15	0.08	0.10	0.02	0.01	0.08	0.05	0.02	0.05	0.13	0.31	1.00
Сосняк	0.25	0.03	0.06	0.00	0.00	0.03	0.03	0.00	0.00	0.15	0.45	1.00
Березняк	0.26	0.08	0.10	0.03	0.01	0.03	0.02	0.03	0.03	0.10	0.31	1.00
Луг	0.22	0.00	0.22	0.03	0.01	0.01	0.05	0.22	0.05	0.03	0.16	1.00
В среднем	0.18	0.13	0.10	0.03	0.01	0.08	0.03	0.02	0.03	0.13	0.26	1.00

Одной из лучших характеристик пересечения структуры коллекций считается *процентное сходство Чекановского*, преобразованное к форме сравнения долей, – сумма минимальных значений относительной значимости по каждому виду:

$K = \sum^{S} \min(p_{1i}, p_{2i})$, где $\min(p_{1i}, p_{2i})$ – минимальное значение p_i для *i*-го вида из двух значений, относящихся к двум коллекциям (1 и 2).

Коэффициент может принимать значения от 0 при отсутствии сходных видов до 1 при полном совпадении распределений видов

обоих сообществ по значимостям. Для выражения отличий выборок можно воспользоваться дополнением до единицы: $100\cdot(1 - K)$.

В нашем примере при сравнении доли обыкновенной бурозубки в кедровнике (0.12) и пихтаче (0.15) находим минимальное значение 0.12, для средней бурозубки имеем 0.17, далее находим 0.07, 0.01 …; суммарное сходство между кедровником и пихтачом составляет $K_{КП} = 0.78$, а отличие – 22% и т. д. (табл. 5.4.5).

Таблица 5.4.5. Показатели пересечения K и различия $100\cdot(1 - K)$ коллекций мелких млекопитающих из разных биотопов, рассчитанные по табл. 5.4.4 (петитом отмечены минимальные расстояния)

	K				
К	0.78	0.69	0.54	0.62	0.50
	П	0.75	0.59	0.70	0.50
		Э	0.74	0.87	0.60
			С	0.80	0.51
				Б	0.50
					Л

	$100\cdot(1 - K)$				
К	**22**	31	46	38	50
	П	**25**	41	30	50
		Э	**26**	**13**	**40**
			С	**20**	49
				Б	50
					Л

Дендрограмма, построенная на этих данных, в целом подтверждает рассмотренные ранее закономерности различий между населением разных биотопов: имеется очевидная специфика коренных стаций (К и П), сильно нарушенных и вторичных лесов (Э, Б, С), остепененных участков (Л).

Для статистического доказательства различий между описаниями двух биотопов предлагается критерий Стьюдента, проверяющий нулевую гипотезу: «выборки принадлежат одной генеральной совокупности и отличие показателя сходства от единицы несущественно». Гипотеза отвергается, если значение критерия превышает табличное $t_{(\alpha,N_{\text{m}})}$ (N_{m} – объем меньшей выборки). Формула критерия имеет вид: $t = (1-K)/m_K$,

где m_K – статистическая ошибка показателя сходства; $m_K = \sqrt{\sum m^2_{\min i}}$, $m^2_{\min i} = p_{\min i}(1-p_{\min i})/N_{\mathbf{m}}$, $p_{\min i} = \min(p_{1i}, p_{2i})$, i – номер (ранг) вида.

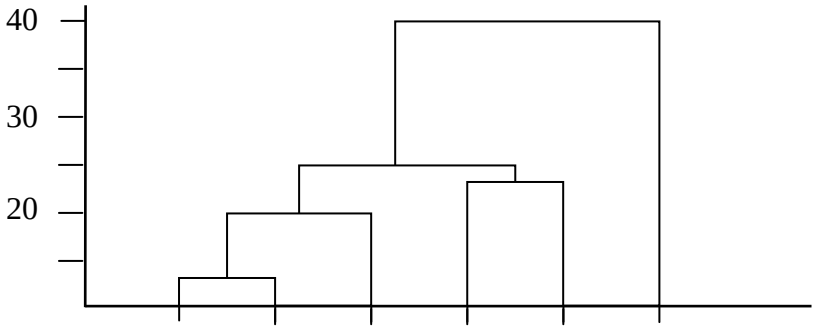


Рис. 5.4.5. Дендрограмма расстояний $100 \cdot (1 - K)$ между биотопическими группировками мелких млекопитающих

Оценим значимость сходства между кедровником и пихтачом. Сначала находим квадраты ошибок каждой из минимальных долей, выбранных для каждого вида. Доля обыкновенной бурозубки в кедровнике меньше, чем в пихтаче: ($p_{\min 1} = 0.12$), общая численность там также меньше ($N_\text{m} = 52$ экз.). Квадрат ошибки этой доли равен: $m^2_{\min 1} = p_{\min 1} \cdot (1 - p_{\min 1}) / N_1 = 0.12 \cdot (1 - 0.12) / 52 = 0.00203$.

Для второго вида, средней бурозубки, получаем: $p_{\min 2} = 0.17$, $m^2_{\min 2} = 0.0027$, для третьего вида $m^2_{\min 3} = 0.0013$ и т. д.; сумма составит $m^2_K = 0.0132$. Отсюда находим ошибку $m_K = 0.1151$ и критерий Стьюдента $t = (1 - K) / m_K = (1 - 0.78) / 0.1151 = 0.22 / 0.1151 = 1.911$. Полученная величина (1.91) меньше табличной $t_{(0.05,52)} = 2.007$ (табл. 4С, стр. 351, или =СТЬЮДРАСПОБР(0.05,52)), следовательно, различия между значением сходства биотопов и единицей не достоверны, нулевая гипотеза сохраняется; нет оснований говорить о том, что выборки из кедровника и пихтача отличаются по структуре доминирования.

Сравнивая кедровник и луг, получаем: $N_\text{m} = 38$, $m_K = 0.0924$, $t = 0.5 / 0.0924 = 5.4$. Эта величина больше табличной $t_{(0.05,38)} = 2.02$, коэффициент сходства выборок достоверно отличается от единицы; спектры видового доминирования в кедровнике и на лугу действительно разные.

В практике генетических и фенетических исследований имеет хождение простой показатель *доля общих форм* (морф, генотипов, фенотипов, видов) (Животовский, 1991): $r = \sum \sqrt{p_{1i} \cdot p_{2i}}$. В качестве меры расстояния рекомендуется выражение $d_r = -\ln r$.

Не вдаваясь в детали расчета данного показателя по нашим данным (которые совершенно аналогичны уже рассмотренным в табл. 5.4.6), отметим лишь, что корреляционное сходство почти эквивалентно процентному сходству долей (рис. 5.4.5 и 5.4.6).

Таблица 5.4.6. Доля общих видов *r* и расстояния – ln *r* ·100 между выборками мелких млекопитающих из разных биотопов, рассчитанные по табл. 5.4.4 (петитом отмечены минимальные расстояния)

r					
К	0.98	0.96	0.91	0.94	0.83
	П	0.973	0.92	0.967	0.85
		Э	0.96	0.99	0.92
			С	0.965	0.86
				Б	0.38
					Л

– ln *r* ·100					
К	**2**	5	9	6	19
	П	**3**	8	3	17
		Э	4	**1**	**8**
			С	**4**	15
				Б	98
					Л

Отличие состоит в том, что зооценоз сосняка вышел из группы вторичных биотопов. Причина, видимо, состоит в том, что в сосняке четыре вида имеют нулевую численность, и в силу конструктивных особенностей показателя (учет произведения долей) для них значения оказались обнуленными и не вошли в обобщающую сумму. *Чувствительность r к нулевым значениям* повлекла за собой существенное уменьшение сходства коллекции сосняка от коллекций из прочих биотопов.

Вторая особенность данного показателя связана с *использованием квадратного корня* для оценки сходства, который существенно *снижает отрыв больших значений от небольших* (см. подробнее раздел 5.3). Поэтому высокие оценки значимости видов сократили свое влияние на общий показатель сходства, а невысокие значимости приобрели относительно больший вес. Например, как видно из сопоставления дендрограмм расстояний (рис. 5.4.5, 5.4.6),

сходство группировок в экотоне и березняке ($r_{ЭП}$), найденное с помощью показателя r, стало меньше (относительно значения $r_{ПК}$), чем оцененное с помощью показателя K. Было (табл. 5.4.5): $K_{ЭБ} = 0.87$ и $K_{ПК} = 0.78$ (отношение равно 0.87 / 0.78 = 1.11), стало (табл. 5.4.5): $r_{ЭБ} = 0.99$ и $r_{КП} = 0.98$ (отношение – 1.01).

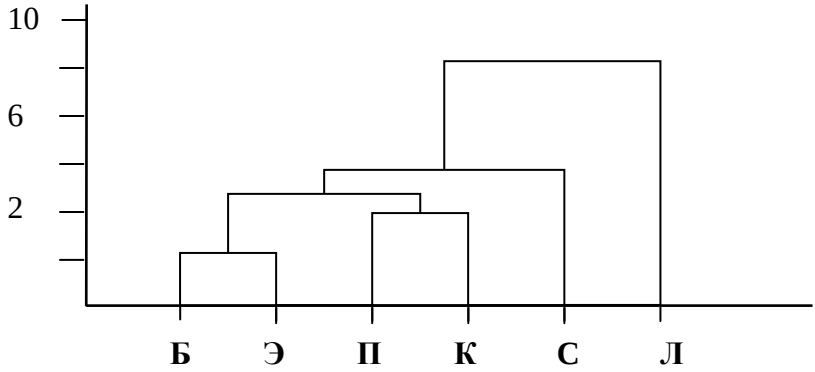


Рис. 5.4.6. Дендрограмма расстояний – ln r ·100 между биотопическими группировками мелких млекопитающих

Статистическая ошибка рассчитывается по формуле:

$$m_r^2 = 0.25 \cdot \left(\frac{1 - p_{01} - r^2}{2N_1} + \frac{1 - p_{02} - r^2}{2N_2} \right),$$

где p_{01} – сумма долей видов первой выборки, не представленных в первой выборки, p_{02} – сумма долей видов второй выборки, не представленных в первой.

В нашем случае (табл. 5.4.4), например в кедровнике по сравнению с пихтачом, не отлавливались два вида полевок (темная и экономка), $p_{01} = 0$, $p_{02} = 0.01 + 0.01 = 0.02$.

$$m_r^2 = 0.25 \cdot \left(\frac{1 - 0 - 0.98^2}{2 \cdot 52} + \frac{1 - 0.02 - 0.98^2}{2 \cdot 53} \right) = 0.000141,\ m_r = 0.012.$$

Проверяя с помощью критерия Стьюдента гипотезу об отличии коэффициента сходства от единицы, получаем: $t = (1 - r) / m_r = 1.68$. Эта величина меньше табличного значения $t_{(0.05,52)} = 2.007$, то есть показатель сходства населения кедровника и

пихтача от единицы значимо не отличается; нет оснований считать, что выборки взяты из разных генеральных совокупностей.

Корреляционная мера сходства коллекций

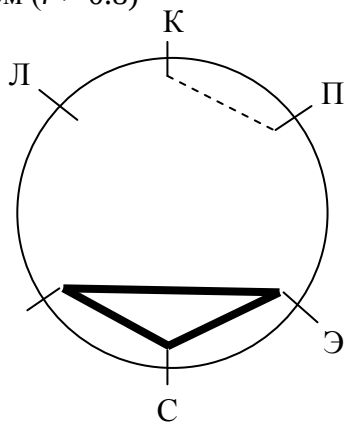
Коэффициент корреляции учитывает сопряженную изменчивость двух случайных величин, то есть изменяется ли один показатель при изменении другого. Иными словами, в центре внимания находится вопрос: насколько синхронно происходит изменение двух показателей, в нашем случае – насколько совпадают перепады численностей (или долей) видов в двух сравниваемых списках. Смысловая формула коэффициента корреляции Пирсона имеет вид: $r = \frac{\sum(N_{i1} - M_{N_1})(N_{i2} - M_{N_2})}{\sqrt{\sum(N_{i1} - M_{N_1}) \cdot \sum(N_{i2} - M_{N_2})}}$, где N_i – численность i-го вида в двух выборках, M_N – средняя арифметическая численность для каждого биотопа. На практике используется иная формула: $r = \frac{\sum N_{i1}N_{i2} - (\sum N_{i1} \cdot \sum N_{i2})/s}{\sqrt{(\sum {N_{i1}}^2 - (\sum N_{i1})^2/s) \cdot (\sum {N_{i2}}^2 - (\sum N_{i2})^2/s)}}$, которая реализована в среде Excel в виде функции =КОРРЕЛ(блок_ячеек1, блок_ячеек2); здесь блок_ячеек1 охватывает значения численностей для одного биотопа, блок_ячеек2 – для другого.

Значения коэффициентов корреляции между показателями обилия мелких млекопитающих в разных биотопах приведены в табл. 5.4.7, на основании которой построена коррелограмма, выявившая по-прежнему три группы (плеяды) биотопов – коренные, вторичные, остепненные.

Обращают на себя внимание присутствующие в матрице отрицательные коэффициенты, которые трудно интерпретировать с точки зрения сходства-различия. Строго говоря, коэффициент корреляции не является мерой сходства, которая (по определению) должна принимать значения от 0 до 1, а область возможных значений r равна двум: $-1 \geq r \geq 1$. С другой стороны, содержание коэффициента корреляции оказывается даже богаче, чем простых мер сходства, поскольку он может ярко выразить степень диспропорции двух распределений видов по значимости. Так, при зеркальной асимметрии корреляция примет значение $r = -1$.

Таблица 5.4.7. Корреляционное сходство *r* между выборками мелких млекопитающих из разных биотопов; разные плеяды отмечены петитом ($r > 0.9$) и курсивом ($r > 0.8$)

	r				
К	*0.80*	0.54	0.42	0.47	-0.03
	П	0.59	0.51	0.58	-0.03
		Э	**0.97**	**0.93**	0.39
			С	**0.95**	0.42
				Б	0.54
					Л



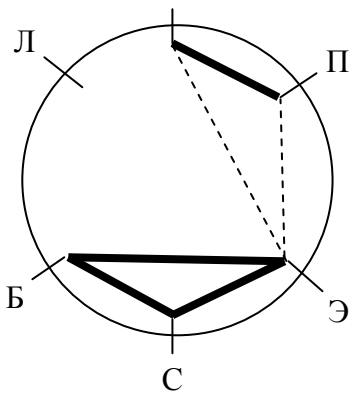
Еще одна позитивная черта этого показателя состоит в том, что на коэффициенте корреляции основан эффективный способ анализа многомерных данных – метод главных компонент, позволяющий интерпретировать корреляционные отношения с точки зрения реакции видов на некие общие факторы. Поиск общих факторов изменчивости можно проводить и на основании матриц сходства-различия (после их определенного преобразования) методом многомерного шкалирования (см. п. 8.3).

Оценка значимости коэффициента корреляции осуществляется стандартным образом по критерию Стьюдента с числом степеней свободы $df = s - 2$ (или по табл. 4С, стр. 354).

В свою очередь, для приведения коэффициента корреляции к шкале расстояний 0–1 предлагается (Песенко, 1982, с. 174) принять в качестве точек отсчета не средние арифметические (M_{N1}, M_{N2}), а начало осей координат (0, 0). В этом случае формула расчета упростится: $r' = \sum N_{i1}N_{i2} / \sqrt{\sum N_{i1}^2 \sum N_{i2}^2}$. Однако теперь мы получаем характеристику пересечения коллекций, принимающую в рассмотрение и пропорции значимостей, и абсолютные их уровни, то есть меру, аналогичную показателю Чекановского. Результаты анализа (табл. 5.4.8) также во многом сходны (см. табл. 5.4.2, рис. 5.4.1).

Таблица 5.4.7. Корреляционное сходство r' между выборками мелких млекопитающих из разных биотопов; разные плеяды отмечены петитом ($r > 0.9$) и курсивом ($r > 0.8$)

	r'				
К	**0.91**	*0.81*	0.65	0.75	0.56
	П	*0.82*	0.70	0.79	0.51
		Э	**0.95**	**0.96**	0.72
			С	**0.96**	0.64
				Б	0.76
					Л



5.5. Диагностические баллы

Рассмотренные выше методы применительно к отдельным видам позволяют решать задачи описания разнообразия фенотипов и на и этой основе сравнивать разные популяции. Одна из важных задач этого плана – экстренная классификация, например определение диагноза больного (часто с использованием качественных признаков). Рассмотрим один из эффективных методов дискриминации, основанный на теории вероятностей и информации, – *метод диагностических таблиц* (Гублер, 1990). Они позволяют буквально за считанные минуты и безо всяких вычислений отнести объект к тому или иному заранее определенному классу. Эта схема особенно важна для работников скорой медицинской помощи, которые должны быстро поставить предварительный диагноз и вынести решение о мерах помощи больному, о необходимости госпитализации и т. п. Этот алгоритм будет крайне полезен и в таких биологических приложениях, как определение вида животного после мимолетной встречи с ним. Не менее актуально и определение пола особей по внешнему виду и пр. Пользоваться диагностическим алгоритмом просто, но его разработка занимает достаточно много времени и требует глубокого анализа исходных данных.

Этапы построения и использования метода мы рассмотрим на примере диагностики пола обыкновенной гадюки в полевых условиях (рис. 5.5.1). Общее требование к количеству используемых признаков состоит в следующем: число признаков должно гарантировать отнесение данного изучаемого объекта к какому-либо классу объектов (с принятым уровнем доверительной вероятности). Опыт показывает, что для однозначного определения пола гадюки (два класса: F – самка, M – самец) без отлова необходимо примерно 6 показателей (x), но для демонстрации метода достаточно рассмотреть два из них: окраска спины (без зигзага) и общие размеры тела.

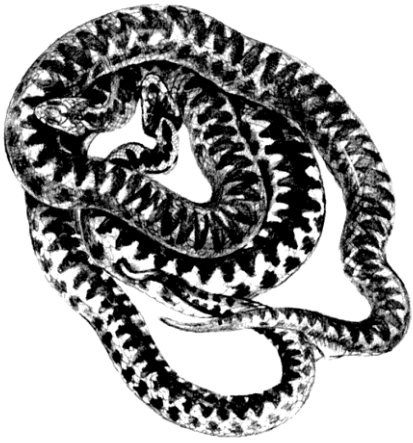


Рис. 5.5.1. Самка и самец гадюки

Процедура составления диагностической таблицы состоит из серии этапов:

- составление схемы диагностического древа,
- сбор данных по каждому признаку (половина данных используется в анализе, вторая половина служит для контроля работы готовой диагностической таблицы),
- проверка достоверности различий между признаками,
- построение и сглаживание распределений показателей,
- определение диагностических баллов,
- оценка информативности показателей,
- составление диагностического дерева (таблицы),
- проверка валидности таблицы по контрольной выборке.

Схема диагностического дерева

В теории информации показано, что дихотомия (раздвоение ветви) в организации испытаний – очень эффективный способ познания качества неизвестного объекта. Классификацию уподобляют движению по дороге с развилками; очередной выбор направления должен (как минимум) в два раза сокращать неопределенность си-

туации, в два раза увеличивать шансы достичь цели. Для того чтобы за минимальное число шагов узнать, к какому из двух классов (A или B) принадлежит данный незнакомый объект, нужно научиться задавать *бинарные* вопросы, когда вероятность получить ответ A равна вероятности получить ответ B; $P(A) = P(B) = 0.5$. Тогда будет получен один бит информации об объекте: $I = H = -\sum(P_i \cdot \log_2 P_i) = -[0.5 \cdot \log_2(0.5) + 0.5 \cdot \log_2(0.5)] = -[0.5 \cdot (-1) + 0.5 \cdot (-1)] = 1$.

Ситуацию можно признать неудовлетворительной, когда при идентификации нескольких объектов (например, в определителе животных) вопросы ориентированы на некий конкретный вид. Так, вопрос «каркает ли встреченная птица?» имеет небольшие шансы получить положительный ответ, поскольку доля ворон в природе составляет едва ли 5% по сравнению с встречаемостью других заметных птиц (хищных, дятлов, голубей, боровой дичи, куликов); информативность такого вопроса невысока: $I = -[0.05 \cdot \log_2(0.05) + 0.95 \cdot \log_2(0.95)] = 0.28$. В поисках истины нам пришлось бы задавать чрезмерно много таких вопросов. Иными словами, объемы групп, сравниваемых в одном вопросе, должны быть равны.

В случае с диагностикой пола животных это условие автоматически выполняется. Число самок и самцов (F, M) в популяции можно считать примерно равным, значит, для каждой очередной отловленной особи *априорная* вероятность «быть самкой» примерно равна *априорной* вероятности «быть самцом». При таком соотношении любой диагностический вопрос относительно свойств новой особи эффективен.

Сбор данных, построение распределений

Характеристики гадюк и условий их обитания, а также половая принадлежность изучались на протяжении ряда лет в полевых условиях (о. Кижи); около 1000 экз. использовано для построения диагностической таблицы, и столько же – для проверки ее работы. Соотношение полов в уловах составило $n_F : n_M = 1.01 : 0.99$.

Определение пола гадюки без ее отлова основано на двух признаках. *Окраска* фона спинной части (за исключением обычно черного зигзага) у самок может быть зеленой, коричневой, серо-коричневой, темной, черной; у самцов обычен серый, голубой, синий, черный цвет фона (табл. 5.5.1). *Размеры тела* встречающихся самок, как правило, больше, чем самцов (табл. 5.5.2).

Таблица 5.5.1. Частота (*a*) встречаемости по-разному окрашенных самок и самцов обыкновенной гадюки

	Окраска спины				
	серая	темная	черная	коричневая	зеленая
a_F	8	37	42	127	74
a_M	139	31	18	10	3

Таблица 5.5.2. Частота (*a*) встречаемости самок и самцов гадюки с разными размерами тела (длина тела + длина хвоста)

	Размеры гадюк (см)			
	маленькие	небольшие	средние	большие
	до 30	до 45	до 60	до 75
a_F	11	22	323	280
a_M	4	28	452	72

Проверка достоверности различий

Минимальное исходное требование к диагностическим признакам двух сравниваемых групп – расхождение в средних тенденциях. Для этого оценивается значимость различий между средними арифметическими величинами (по критерию Стьюдента), или между выборками в целом (например, по непараметрическому критерию Уилкоксона), или же между распределениями признаков (по критерию хи-квадарт). В нашем случае по окраске сравнивали частотные распределения разнополых гадюк с помощью критерия χ^2 Пирсона. Получено достоверное отличие с уровнем значимости $\alpha << 0.001$. Размеры тела у самок и самцов также отличаются как в среднем (58.2 против 53.9, критерий Стьюдента $t = 14.3 << t_{табл.}\ 1.96$), так и по характеру распределения ($\alpha << 0.001$).

Сглаживание распределений

Когда выборки не очень многочисленны, частоты соседних классов в построенных распределениях испытывают достаточно сильное случайное варьирование. В этом случае имеет смысл выровнять частоты отдельных значений относительно друг друга. Для этой цели применяют метод скользящей средней, когда с помощью

весовых коэффициентов вычисляются новые значения частот, например, по следующей формуле (известны и иные фильтры для сглаживания рядов, п. 9.3): $M_i = (y_{i-2} + 2y_{i-1} + 4y_i + 2y_{i+1} + y_{i+2})/10$. Краевые значения вычисляются по модифицированным формулам. Первое сглаженное значение составит $M_1 = (4y_1 + 2y_2 + y_3) / 7$, второе $M_2 = (2y_1 + 4y_2 + 2y_3 + y_4)/9$.

Сглаживание можно проводить для всех видов распределений, *кроме полиномиального*, поскольку в нем частоты соседних значений идеологически не связаны. Выборки гадюк в высокой степени репрезентативны, поэтому в сглаживании частот по размерам нет необходимости, а сглаживание полиномиальных частот по окраске запрещено.

Определение диагностических баллов

Смысл этой операции состоит в том, чтобы простым числом выразить *вклад* любого *значения симптома* (признака) *в определение диагноза* (в установление класса рассматриваемого объекта). Это число названо *диагностическим баллом*. При исследовании отдельного объекта каждая его характеристика служит для определения одного диагностического балла. Баллы по всем показателям складываются, и их сумма показывает, достигнут ли *значимый диагностический порог*, то есть можно ли объект однозначно отнести к определенному классу объектов.

Количественно выразить «диагностическую силу» каждого значения изучаемых признаков позволяет *уравнение Байеса*, модифицированное *для случая равенства априорных вероятностей* двух заданных классов (это наш случай организации процесса классификации). Оно утверждает, что отношение вероятностей состояния A к вероятности состояния B в случае установления симптома x_i равно отношению вероятности симптома x_i при состоянии A к вероятности этого же симптома x_i при состоянии B:

$$\frac{P(A/x_i)}{P(B/x_i)} = \frac{P(x_i/A)}{P(x_i/B)}.$$

Переведем выражение на язык нашего предмета исследований. Допустим, отловлена *серая* гадюка, это симптом x_i = «серая окраска». Нужно установить, каковы *шансы* того, что это *самец* (состояние M), и каковы шансы того, что это *самка* (состояние F). Все-

го было отловлено 8 серых самок, их доля среди изученных (288 экз.) составила $P(x_i / \mathrm{F}) = 8 / 288 = 0.028$; это и есть вероятность быть серой (x_i) и самкой (F). Доля самцов, т. е. вероятность быть серым (x_i) и самцом (M) равна $P(x_i / \mathrm{M}) = 139 / 201 = 0.692$. Можно сказать, что имеется всего 3 шанса из 100, что серая гадюка есть самка и 69 шансов из 100, что это самец.

Теперь формула Байеса может быть выражена предметно: отношение вероятности быть самкой к вероятности быть самцом *для серой особи* равно отношению доли серых самок к доле серых самцов (или отношению шансов быть самкой к шансам быть самцом):

$$\frac{P(F / x_i)}{P(M / x_i)} = \frac{P(x_i / F)}{P(x_i / M)} = \frac{0.028}{0.691} = 0.040.$$

Формула Байеса вычисляет относительную меру информативности данного значения признака для правильной диагностики объекта. Так, серая окраска очень характерна для самцов и отношение шансов «быть самцом» к шансам «быть самкой» *для серой особи* очень велико: $\frac{P(M / x_i)}{P(F / x_i)} = \frac{P(x_i / M)}{P(x_i / F)} = \frac{0.691}{0.028} = 24.9.$

Сравнивая величины 0.040 и 24.9, можно придти к выводу о том, что серая окраска важна для распознавания полов. Аналогично вычисляются диагностические баллы для каждого значения изучаемых признаков (табл. 5.5.3).

Далее нужно определиться с пороговым значением отношения вероятностей (отношения шансов), которое могло бы свидетельствовать о том, что с заданным уровнем вероятности данный объект классифицирован правильно. Если принять, что шансы возможной ошибки не должны превышать 5 против 95 из 100 (что соответствует соглашению о 95%, обычному для биологии уровню значимости $\alpha = 5\%$ и доверительной вероятности $P = 95\%$), получаем пороговые отношения вероятностей: $\frac{0.95}{0.05} = 19$ и $\frac{0.05}{0.95} = 0.053$. Полученное нами значение 24.9 превышает 19, значит, с вероятностью $P > 95\%$ можно утверждать, что любая серая особь есть самец. Аналогично, раз значение 0.04 не достигает порога 0.053, то с вероятностью $P > 95\%$ можно утверждать, что любая серая особь является не самкой.

Таблица 5.5.3. Расчет диагностических баллов для окраски гадюки

x_i	Окраска спины					
	серая	темная	черная	корич-невая	зеленая	Σ
Частота a_F	8	37	42	127	74	288
Частота a_M	139	31	18	10	3	201
$P_F = a_F / \Sigma a_F$	0.028	0.128	0.146	0.441	0.257	1.000
$P_M = a_M / \Sigma a_M$	0.692	0.154	0.090	0.050	0.015	1.000
P_F/P_M	0.040	0.833	1.628	8.864	17.215	
P_M/P_F	24.89	1.200	0.614	0.113	0.058	
lg (P_F/P_M)	**–7**	**0**	**1**	**5**	**6**	
lg (P_M/P_F)	7	0	–1	–5	–6	
J	4.633	0.010	0.060	1.854	1.496	8.052

Таблица 5.5.4. Расчет диагностических баллов для разноразмерных классов гадюки

x_i	Размеры гадюк				
	малень-кие	неболь-шие	сред-ние	большие	Σ
a_F	11	22	323	280	636
a_M	4	28	452	72	556
$P_F = a_F / \Sigma a_F$	0.017	0.035	0.508	0.440	1.000
$P_M = a_M / \Sigma a_M$	0.007	0.050	0.813	0.129	1.000
P_F/P_M	2.404	0.687	0.625	3.400	
P_M/P_F	0.416	1.456	1.601	0.294	
lg (P_F/P_M)	**2**	**–1**	**–1**	**3**	
lg (P_M/P_F)	–2	1	1	–3	
J	0.019	0.013	0.312	0.826	1.17

Для удобства работы дробные значения преобразуют в десятичный логарифм, умножают на константу 5 и округляют. Теперь пороговые диагностические баллы равны: для вероятности 95% – $5 \cdot \lg 19 = 6.39 \approx 6$, для $\alpha = 5\%$ – $5 \cdot \lg 0.053 = -6.39 \approx -6$.

Оценка информативности признаков

Логика подсказывает, что для скорейшего определения классовой принадлежности изучаемого объекта следует в первую очередь принимать во внимание те из них, которые за минимальное число шагов позволят достичь диагностического порога и вынести решение о статусе объекта. Выстраивать признаки на диагностическом древе следует в порядке их информативности. Для ее оценки служит мера Кульбака: $J = \sum^{k} 5 \cdot (P_A - P_B) \cdot \lg\left(\frac{P_A}{P_B}\right)$, где k – число градаций (значений) изучаемого признака (i = 1, 2… k), $P_A = P(A)$, $P_B = P(B)$ – доли определенных значений признака при разных состояниях объекта.

Чем выше окажется величина меры информативности J, тем раньше следует ставить данный признак в диагностической таблице. Расчеты приведены в последней строке таблиц 5.5.3, 5.5.4. Например, для первой градации серой окраски с вероятностями $P_F = 0.028$ и $P_M = 0.692$ имеем:

$$J_1 = 5 \cdot (P_A - P_B) \cdot \lg\left(\frac{P_A}{P_B}\right) = 5 \cdot (0.028 - 0.692) \cdot \lg\left(\frac{0.028}{0.692}\right) = 4.633.$$

Суммирование по всем столбцам дает для признака «окраска спины» значение $J = 8.052$. Информативность признака «размеры тела» существенно ниже: $J = 1.17$, в процессе определения пола он должен оцениваться во вторую очередь, после анализа окраски.

Диагностическая таблица: составление, использование

Теперь остается составить таблицу, в которой каждые две строки содержат значения отдельного диагностического признака и диагностические баллы этих значений. Справа отводится графа для записи результата анализа, то есть для записи полученного балла и суммы накопленных баллов. Внизу подводится окончательный итог.

Порядок работы состоит в том, чтобы, выдвинув некую гипотезу, регистрировать признаки у встретившейся особи, по таблице определять диагностические баллы, соответствующие этой гипотезе, и складывать их. При достижении порога (±6) можно считать, что статус объекта определен.

Допустим, мы рассматриваем гипотезу: не самка ли эта, встретившаяся нам серая (балл –7) небольшая (балл –1) особь: сумма баллов равна –8 (табл. 5.5.5). Поскольку –8 < –6, то с вероятностью $P = 0.95$ гипотезу можно опровергнуть: это не самка, а самец.

Для коричневой (5) большой (3) особи сумма баллов составит 8 (>6), что подтверждает гипотезу – это самка. Однако темные и черные особи любых размеров не набирают пороговой суммы баллов и не могут быть диагностированы по полу. Вывод очевиден: требуется расширить число признаков для уверенной диагностики пола гадюки. Однако визуально их фиксировать трудно и без отлова не обойтись.

Таблица 5.5.5. Таблица для проверки гипотезы «встретилась самка»

Шаг	Признаки					Баллы
1	**серая**	темная	чер-ная	корич-невая	зеленая	–7
	–7	0	1	5	6	(Σ = –7)
2	малень-кие	**небольшие**	сред-ние	боль-шие		–1
	2	–1	–1	3		(Σ = –8)
	Пол особи: М				Итог =	–8

Проверка валидности таблицы по контрольной выборке

Смысл процедуры состоит в том, чтобы по контрольной выборке особей с известным полом и регистрируемыми признаками получить значения диагностических баллов, определить по ним пол и сравнить определение с реальностью. Если данные были предварительно введены в базу данных, возможна автоматическая оценка баллов с помощью логических операторов Excel. Например, следующий оператор (в одной строке) превращает код окраски особи

(серый – **g**rey, темный – **d**ark, черный – **m**elanist, коричневый – **b**rown, зеленый – **s**alad) в диагностический балл:

```
=ЕСЛИ(A1="g",-7,
ЕСЛИ(A1="d",0,ЕСЛИ(A1="m",1,
ЕСЛИ(A1="b",5,ЕСЛИ(A1="s",6,0))))).
```

Аналогично определяется балл для любого другого признака. Их сумма служит для определения статуса объекта.

В нашем случае из 200 проверенных особей пол верно был определен у 84%; у всех темных и черных особей, составляющих оставшиеся 16%, пол остался неопределенными. Ориентируясь на обычный уровень доверительной вероятности $P = 0.95$, следует сделать заключение о плохом качестве предложенной диагностической таблицы и о необходимости ее доработки.

Глава 6
ИЗУЧЕНИЕ ЗАВИСИМОСТЕЙ

Важнейшая проблема обнаружения, описания и доказательства влияния, связи, взаимозависимости в природных системах решается богатым арсеналом разнообразных статистических методов, рассмотренных в многочисленных пособиях. Смысл данной главы состоит в обсуждении некоторых нюансов процедуры регрессионного и корреляционного анализов, которые могут помочь в интерпретации соответствующих коэффициентов. Представляется важным популяризовать интересный метод RMA-регрессии, более соответствующий понятию биологической зависимости, нежели стандартный метод наименьших квадратов.

6.1. МНК-регрессия

Регрессионный анализ изучает эффект влияния одного признака на другой или зависимость признака от фактора. Он дает следующие основные результаты: *уравнение регрессии*, выражающее пропорциональность сопряженного изменения признаков; *оценку значимости параметров* регрессионного уравнения (оценку достоверности их отличия от нуля) с помощью критерия Стьюдента; оценку силы и общей достоверности влияния на признак изучаемого фактора (в форме таблицы дисперсионного анализа с помощью коэффициента детерминации и критерия Фишера).

Эмпирическая регрессия

Основную тенденцию взаимосвязанного изменения двух признаков можно отобразить с помощью простого графического приема обработки эллипса рассеяния. Разобьем ось *x* на несколько интервалов, или выборок (1, 2, 3). Найдем для каждой из них *частное среднее* значение признака *y* (My_i) и проведем через эти средние точки ломаную линию. Это будет линия регрессии. *Регрессия* – изменение среднего уровня одной переменной при изменении значений другой (рис. 6.1.1).

Ход ломаной линии нельзя передать простым уравнением, к тому же на ней сказывается способ интервального разбиения оси абсцисс, а также репрезентативность разных областей распределе-

ния. Лучше использовать единственную прямую линию регрессии, которая подчеркивала бы основные тенденции зависимости признаков и выражалась простым уравнением.

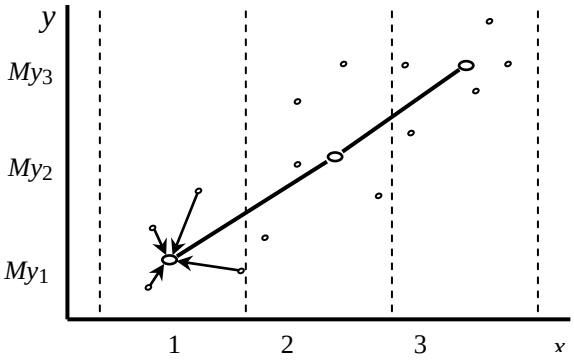


Рис. 6.1.1. Эмпирическая линия регрессии

Уравнение и коэффициент регрессии

Для отображения изменения одной переменной в зависимости от другой в алгебре используется уравнение линии $y = a \cdot x$, коэффициент которого выражает пропорцию между значениями признаков: $a = \frac{y}{x}$. Если линия не проходит через начало координат, уравнение дополняется свободным членом ($y = a \cdot x + b$), равным значению ординаты в точке пересечения с линией: $b = y_0$ при x = 0. В этом случае коэффициент пропорциональности можно найти как отношение разности координат любых двух точек линии с координатами (x_1, y_1) и (x_2, y_2): $a = \frac{y_2 - y_1}{x_2 - x_1}$.

Аналогично можно было бы выразить величину коэффициента регрессии a как пропорцию изменения признака y при изменении признака x (пропорцию одновременного отклонения признаков от своих общих средних величин): $a = \frac{y - M_y}{x - M_x}$ или $y - M_y = a \cdot (x - M_x)$.

Преобразования $y = a \cdot x + M_y - a\, M_x$, $b = M_y - a\, M_x$ приводят к уравнению линии: $y = ax + b$.

Однако применению этой формулы мешает то обстоятельство, что изучаемые признаки, как правило, широко варьируют, и поэтому коэффициенты, вычисленные для каждой пары (x_i, y_i), будут существенно отличаться. Очевидно, что коэффициент регрессии должен быть в каком-то отношении *усредненным*. Такое обобщение можно выполнить, если суммировать все отклонения для всех вариант, но эта сумма будет равна нулю в силу симметрии распределения $a = \frac{\sum(y - M_y)}{\sum(x - M_x)} = \frac{0}{0}$. Если обобщать квадраты отклонений $a = \frac{\sum(y - M_y)^2}{\sum(x - M_x)^2}$, то, поделив их на число степеней свободы ($n - 1$), мы приходим к отношению дисперсий $a = \frac{S_y^2}{S_x^2}$. Эта величина всегда будет положительным числом, независимо от того, существует ли связь между изучаемыми переменными и каков ее характер (положительный или отрицательный). Эта логика может привести к построению специфического уравнения зависимости между признаками (п. 6.3), но в свое время для поиска оценки сопряженной изменчивости признаков был предложен другой прием. Он состоит в следующем: нужно умножить исходную дробь $a = \frac{y - M_y}{x - M_x}$ на отклонение значения x от своей средней ($x - M_x$):

$$a = \frac{(y - M_y)(x - M_x)}{(x - M_x)(x - M_x)}$$

и уже после этого выполнить обобщение для всех вариант:

$$a = \frac{\sum(y - M_y)(x - M_x)/(n-1)}{\sum(x - M_x)(x - M_x)/(n-1)} = \frac{Cov(y, x)}{S_x^2}.$$

В числителе *формулы коэффициента регрессии* стоит *ковариация*, в знаменателе – дисперсия признака x. Рассмотрим смысл выражения в числителе $\Sigma(y - M_y)(x - M_x)$ (Смирнов, 1979). Каждое отклонение варианты от средней (влево со знаком «–» , вправо со знаком «+») можно представить как сумму эффекта действия друго-

го признака (сопряженная изменчивость – x_y, y_x) и эффекта действия случайных причин (случайная изменчивость – $y_{сл.}$, $x_{сл.}$):

$$(y - M_y) \rightarrow (\pm y_x \pm y_{сл.}),$$
$$(x - M_x) \rightarrow (\pm x_y \pm x_{сл.}).$$

Тогда сумма произведения отклонений разложится на компоненты:

$$\Sigma(y - M_y)(x - M_x) \rightarrow \Sigma[(\pm y_x \pm y_{сл.})(\pm x_y \pm x_{сл.})] =$$
$$= \Sigma[(\pm y_x)(\pm x_y)] + \Sigma[(\pm y_x)(\pm \mathbf{x}_{сл.})] + \Sigma[(\pm \mathbf{y}_{сл.})(\pm x_y)] + \Sigma[(\pm \mathbf{y}_{сл.})(\pm \mathbf{x}_{сл.})] =$$
$$= \Sigma[(\pm y_x)(\pm x_y)].$$

Три правых члена содержат случайный элемент какого-либо признака (выделенные петитом) и при суммировании обнуляются. Ненулевым остается лишь первый член, который отражает эффект совместного отклонения признаков в ту или иную сторону от своих средних, то есть только сопряженную изменчивость. Когда признаки действительно взаимосвязаны, то при отклонении *x* в сторону положительных значений *y* отклонится туда же, произведение будет положительным, при отклонении *x* в сторону уменьшения *y* отклоняется тоже в сторону уменьшения, произведение отрицательных отклонений вновь будет положительным. Сумма произведений будет положительной для прямой зависимости (и отрицательной для обратной). *Ковариация* – это численная характеристика сопряженного варьирования двух признаков.

Метод наименьших квадратов

Рабочие формулы для расчета коэффициентов уравнения регрессии были найдены с помощью *метода наименьших квадратов*. Его основная идея состоит в том, что линия регрессии должна пройти на наименьшем удалении от каждой точки, т. е. чтобы сумма *квадратов* расстояний от всех точек до прямой линии была минимальной: $C_{ост.} = \Sigma(y_i - \hat{y}_i)^2 \rightarrow$ *наименьшая*. Эта идея выражается системой дифференциальных уравнений:

$$\frac{dR}{da} = -2\sum\left[\hat{y}_i - a - b(x_i - M_x)\right] = 0,$$
$$\frac{dR}{db} = -2\sum\left[\hat{y}_i - a - b(x_i - M_x)\right](x_i - M_x) = 0,$$

решением которой являются корни, выступающие в роли рабочих формул для определения коэффициентов регрессии:

$a = Cxy / Cx,$

$b = M_y - a \cdot M_x,$

здесь $Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n$, $Cx = \Sigma x^2 - (\Sigma x)^2 / n$, $M_y = \Sigma y / n$, $M_x = \Sigma x / n$.

Идею метода наименьших квадратов помогает понять следующая иллюстрация. Построим два идентичных рисунка двумерного рассеяния, но возможные линии регрессии на каждом из них проведем по-разному. Затем измерим расстояния между линией и отрезками, построенными вдоль оси ординат (регрессия y по x).

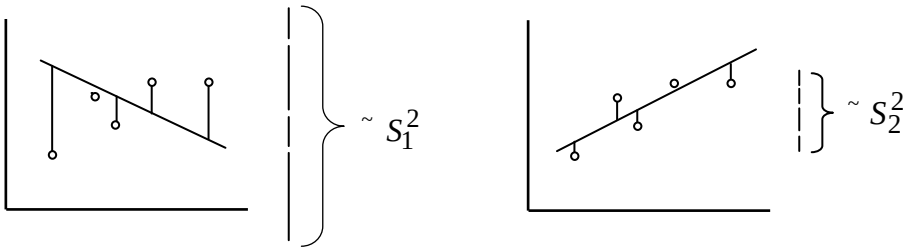


Рис. 6.1.2. Оценки минимальной суммы отклонений линии регрессии от точек

Суммируем длины отрезков. Хорошо видно, что чем дальше отстоит линия от общего направления разброса точек, тем больше сумма отклонений ее от точек (и сумма их квадратов, $S_1^2 > S_2^2$).

В обширных выборках каждому значению признака x_i соответствует множество значений признака y (частная выборка), имеющих в идеале нормальное распределение (рис. 6.1.3). Как известно, лучшей характеристикой выборки значений является средняя арифметическая. Таким образом, лучшей характеристикой для множества частных выборок признака y (полученных для разных значений признака x_i) будет множество частных средних, My_i , которые сформируют прямую линию (рис. 6.1.3, 1).

Метод наименьших квадратов как раз и создает линию регрессии, проходящую через множество частных средних $My_i = \hat{y}_i$, соответствующих отдельным значениям x_i. Такая линия регрессии в статистическом смысле является *лучшей* (эффективной, несмещенной и пр.) линией, описывающей зависимость признаков, она наиболее обоснована со статистической точки зрения.

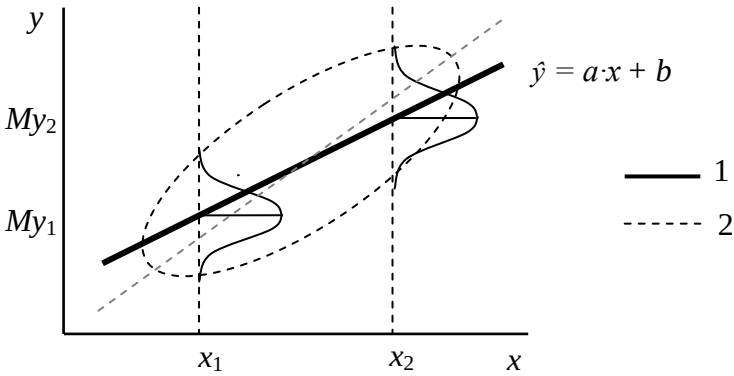


Рис. 6.1.3. Линия регрессии (1) как множество частных средних и ее положение относительно эллипса рассеяния (2). Показаны распределения для двух локальных выборок (со своими средними My_1 и My_2), полученных при фиксированных значениях x_1 и x_2

В то же время всегда следует помнить, что регрессионный анализ односторонне ориентирован на изучение зависимости одного признака от другого, он по-разному выражает влияние x на y. Здесь уместно обратиться к истории.

Термин «регрессия» предложил Ф. Гальтон. Анализируя соотношения между ростом сыновей (y) и ростом отцов (x), он обнаружил, что в соответствии с линейным графиком у низкорослых отцов сыновья должны иметь более высокий рост, чем отцовский. В то же самое время у более высоких отцов сыновья должны быть менее высоки, чем они сами ($x_2 - x_1 > y_2 - y_1$) (рис. 6.1.3, 2). Вместо интуитивно ожидаемой прямой пропорции между ростом отцов и детей (серым пунктиром, это ось эллипса рассеяния) наблюдается определенное *возвращение* к среднему уровню, «регрессия».

Причины такого явления открываются при анализе формулы коэффициента регрессии в терминах действия систематических и случайных факторов: $a = \dfrac{Cov(y,x)}{S_x^2} \to \dfrac{y_x y_y}{(x_y + x_{сл.})^2} \to \dfrac{y_x}{(x_y + x_{сл.})}$.

Коэффициент регрессии, как наиболее обоснованный со статистической точки зрения, используется для предсказания значений одного признака по значениям другого. Биолог хотел бы, чтобы уравнение выражало «чистую» пропорцию между признаками

$a = y_x / x_y$, график которой соответствовал бы оси эллипса рассеяния. Но математик может дать ему лишь наиболее статистически обоснованную пропорцию, выраженную приведенной формулой коэффициента регрессии, которая показывает, на какую величину в *среднем* изменяется один признак при изменении другого на единицу. Случайная изменчивость данных ($x_{сл.}$) не позволяет, к сожалению, точно охарактеризовать истинное соотношение (y_x / x_y), поэтому линия регрессии (линия хода средних) в большей или меньшей степени отклоняется от оси эллипса рассеяния. Чем больше величина случайной составляющей общей изменчивости ($x_{сл.}$ в знаменателе), тем сильнее отклонение; при увеличении знаменателя величина коэффициента регрессии стремится к нулю (рис. 6.1.4).

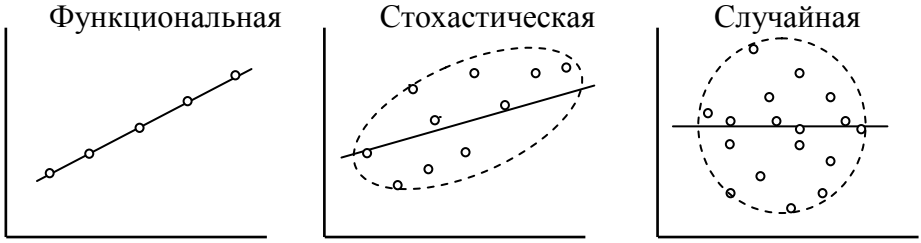


Рис. 6.1.4. Виды зависимостей

Когда зависимость *функциональна*, т. е. действуют только доминирующие факторы взаимного влияния, все варианты выстраиваются в одну линию; коэффициент регрессии велик $a \to y_x/x_y >> 0$ и явно выражает пропорцию именно между признаками.

Если связь *стохастическая* (обычный случай совместного действия доминирующих и случайных факторов), варианты образуют облако эллипса рассеяния; но коэффициент регрессии достоверно превышает нуль ($a \to y_x/(x_y + x_{сл.}) > \pm 0$) и выражает зависимоть среднего уровня одного признака от изменения значений другого.

При варьировании только по *случайным* причинам (признаки не взаимодействуют) область рассеивания вариант принимает округлую форму; a приближается к нулю и значимо от него не отличается, $a \to 0/x_{сл.} \approx 0$ (хотя и никогда не бывает равен ему в точности).

Односторонний подход регрессионного анализа в какой-то мере можно компенсировать изучением взаимозависимости признаков с помощью корреляционного анализа и RMA-регрессии.

Оценка адекватности уравнения регрессии исходным данным

Ответить на вопрос, сказывается ли признак x на изменчивости признака y, то есть является ли он «доминирующим» фактором, можно с привлечением общего принципа статистического оценивания, соотнеся отклонения под действием доминирующего фактора с отклонениями по случайным причинам. На этом основана *модель варианты в регрессионном анализе* (рис. 6.1.5):

$$y_i = M_y \pm y_x \pm y_{сл.},$$

где y_i – значение признака y для i-й варианты (соответствующее значению x_i), M_y – общая средняя арифметическая для всей выборки (общая часть всех вариант), y_x – доля значения y_i, связанная с влиянием признака x, $y_{сл.}$ – доля значения y_i, связанная с действием случайных факторов варьирования.

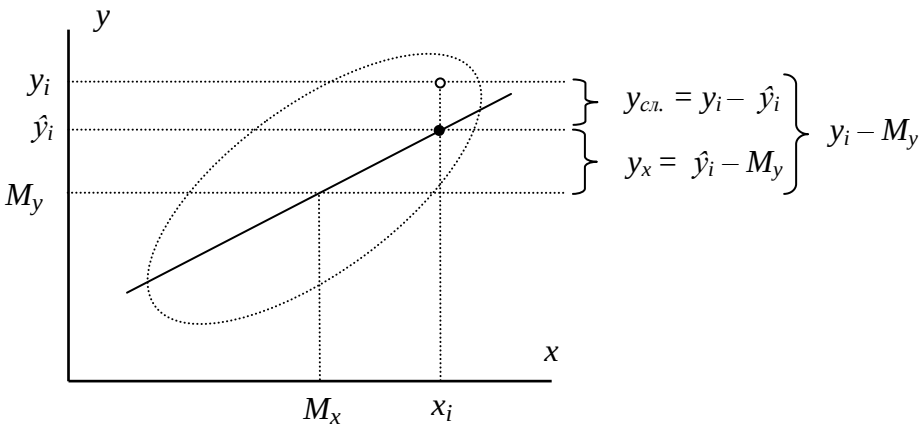


Рис. 6.1.5. Модель варианты в регрессионном анализе

Лучшим показателем взаимосвязи является линия регрессии (множество частных средних $\hat{y}_i$). Характеристикой чисто случайного варьирования выступает отклонение отдельных вариант от линии регрессии. Следовательно, отклонение варианты от общей средней арифметической состоит из отклонения значения регрессии от общей средней и отклонения варианты от линии регрессии:

$$(y_i - M_y) = (y_i - \hat{y}_i) + (\hat{y}_i - M_y),$$

где $y_i - M_y$ – общее отклонение i-й варианты от средней,

$y_x = \hat{y}_i - M_y$ – отклонение линии регрессии (для точки x_i) от средней,

$y_{сл.} = y_i - \hat{y}_i$ – отклонение варианты от линии регрессии по случайным причинам.

Для определения достоверности влияния признака *x* на *y* все рассмотренные отклонения следует объединить по всем вариантам выборки, возведя в квадрат. Получаем оценки *регрессионной* (*модельной*) и *остаточной сумм квадратов* и строим таблицу дисперсионного анализа: изменчивость признака *y* складывается из варьирования, учтенного регрессионной моделью, и из варьирования по случайным причинам (остаточного).

Таблица 6.1.1. Дисперсионный анализ линейной регрессии

Составляющие дисперсии	Суммы квадратов, C	Формулы расчета сумм квадратов	df	S^2	F
Наклон модельной линии	$C_{мод.} = \Sigma(\hat{y}_i - M_y)^2$	$C_{общ.} - C_{остат.}$	1	$S^2_{мод.} = \frac{C_{мод.}}{df_{мод.}}$	$\frac{S^2_{мод.}}{S^2_{остат.}}$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - \hat{y}_i)^2$		$n-2$	$S^2_{остат.} = \frac{C_{остат.}}{df_{остат.}}$	$F_{(0.05,1,n-2)}$
Общая (всего)	$C_{общ.} = \Sigma(y_i - M_y)^2$	$(\Sigma y_i^2 - \Sigma y_i)^2 / n = C_y$			

Общую сумму квадратов ($C_{общ.} = C_y = \Sigma(y_i - M_y)^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$) находят непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и общей средней признака *y*. Остаточную сумму квадратов ($C_{остат.} = \Sigma(y_i - \hat{y}_i)^2$) находят также непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и значением, предварительно рассчитанным по уравнению регрессии $\hat{y}_i = ax_i + b$ (для соответствующих значений x_i). Модельную сумму квадратов ($C_{мод.} = \Sigma(\hat{y}_i - M_y)^2$) рассчитывают как разность между общей и остаточной ($C_{мод.} = C_{общ.} - C_{остат.}$). Можно также рассчитать величину, эквивалентную показателю «силы влияния фактора» – это *коэффициент детерминации*, отношение

регрессионной суммы квадратов к общей сумме квадратов: $R^2 = \frac{C_{мод.}}{C_{общ.}}$. Она принимает значения от 0 до 1.

С помощью критерия Фишера можно проверить нулевую гипотезу Но: предсказания модели в целом *неадекватно* описывают исходные данные, между признаками *зависимости нет*. Конструкция критерия исследует вопрос, превышает ли варьирование, учтенное моделью, случайное (остаточное) варьирование? Критерий Фишера вычисляется как отношение модельной дисперсии к остаточной: $F = S^2_{мод.} / S^2_{остат.} \sim F_{(0.05,1,\,df)}$ (табл. 5С, 6С, стр. 355, 356).

Дисперсии рассчитываются как отношение сумм квадратов к числу степеней свободы. Для остаточной дисперсии берем $df = n - 2$, поскольку в расчетах теоретических значений принимают участие два параметра – a и b. Когда свободный член (b) значимо от нуля не отличается, расчеты теоретических значений проводятся при одном коэффициенте (a) и число степеней свободы будет равно: $df = n - 1$.

Если значение F критерия окажется выше табличного $F_{(0.05,1,\,df)}$, значит, дисперсия реального признака y приближается по величине к дисперсии значений регрессии $\hat{y}_i$, т. е. существенно превышает (случайные) отличия между ними. Значение критерия ниже табличного свидетельствует о существенных отличиях между реальными и модельными данными, о плохом согласовании модели с реальностью, о неадекватности модели.

В качестве примера рассмотрим изучение зависимости между живым весом коров (x) и их приплода (y, кг) (табл. 6.1.2). Сначала определим вспомогательные величины:

$$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n = 103144 - 3150 \cdot 224 / 7 = 2344,$$

$$Cx = \Sigma x^2 - (\Sigma x)^2 / n = 1453158 - 3150^2 / 7 = 35658,$$

$$Cy = C_{общ} = \Sigma y^2 - (\Sigma y)^2 / n = 7330 - 224^2 / 7 = 162.$$

Затем отыщем параметры: $r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$

$M_y = \Sigma y / n = 224 / 7 = 32$, $M_x = \Sigma x / n = 3150 / 7 = 450,$

$a = \frac{Cxy}{Cx} = \frac{2344}{35658} = 0.0657$, $b = M_y - a \cdot M_x = 32 - 0.0657 \cdot 450 = 2.419.$

Получено уравнение линейной регрессии $\hat{y} = 0.0657x + 2.419$, которое позволяет рассчитать теоретические значения $\hat{y}_i$ (табл. 6.1.2) и провести дисперсионный анализ (табл. 6.1.3):
$C_{остат.} = 7.92$, $C_{мод.} = 162 - 7.92 = 154.08$.

Таблица 6.1.2. Расчет линейной регрессии

i	y	x	y^2	x^2	$x \cdot y$	$\hat{y}$	$(y - \hat{y})^2$
1	25	352	625	123904	8800	25.6	0.31
2	26	376	676	141376	9776	27.1	1.29
3	31	402	961	161604	12462	28.8	4.65
4	32	453	1024	205208	14496	32.2	0.04
5	34	484	1156	234256	16456	34.2	0.06
6	38	528	1444	278784	20064	37.1	0.76
7	38	555	1444	308025	21090	38.9	0.81
Σ	224	3150	7330	1453158	103144		7.92

Таблица 6.1.3. Дисперсионный анализ линейной регрессии
$\hat{y} = 0.0657x + 2.419$

Дисперсии	C		df	S^2	F
Наклон модельной линии	$C_{мод.} = \Sigma (\hat{y}_i - M_y)^2 =$	154.08	1	$S^2_{мод.} =$ = 154.08	$F = \frac{154.08}{1.58} =$ = 97.3
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma (y_i - \hat{y}_i)^2 =$	7.92	5	$S^2_{остат.} =$ = 1.58	$F_{(0.05,1,5)} =$ = 6.6
Общая (всего)	$C_{общ.} = \Sigma (y_i - M_y)^2 =$	162	$R^2 = \frac{154.08}{162} = 0.951$		

Расчетное значение F (97.3) превышает табличное (6.6, табл. 5С, стр. 352), следовательно, модель адекватна реальности. Судя по коэффициенту детерминации ($R^2 = 0.95$), влияние веса коров на вес плода велико. Прочие оценки значимости можно найти по известным формулам (Ивантер, Коросов, 2003).

6.2. Корреляционные плеяды

Исследования групп объектов, охарактеризованных многими признаками, показали, что сила связи между разными парами признаков может быть весьма различной: одни коррелируют друг с другом сильнее, другие слабее. Группа признаков, тесно коррелирующих друг с другом, но слабо связанных с остальными признаками, названа *плеядой* (Терентьев, 1959, 1960). Таковы, например, плеяды промеров головы и туловища животного, вегетативной и генеративной частей растений. Если изучается достаточно много показателей (10–30), то плеяд формируется несколько.

Существо явления изучим на примере онтогенеза третьего листа томата. В течение первого месяца жизни один раз в 4 дня промеряли размер 8 частей листа, всего выполнили по 8 промеров каждой части (рис. 6.2.1, А; табл. 6.2.1, А). Затем по этим рядам рассчитали коэффициенты корреляции между всеми признаками. Поскольку практически все части листа росли (а промеры увеличивались), коэффициенты оказались высоки (табл. 6.2.1, Б).

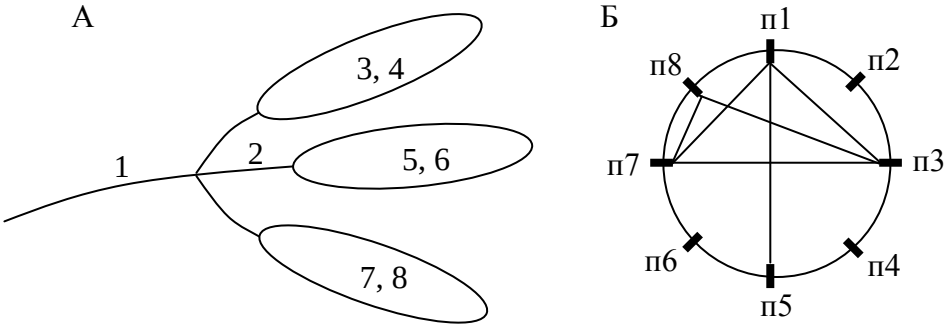


Рис 6.2.1. Промеры листа томата (А: 1 – длина черенка, 2 – расстояние от мутовки до начала второй пластинки, 3, 5, 7 – длина листовых пластинок, 4, 6, 8 – ширина пластинок) и корреляционное кольцо для $r > 0.95$ (Б)

Тем не менее выяснилось, что коэффициенты корреляции не одинаковы и варьируют между разными парами от 0.35 до 0.98. Выделить плеяды тесно связанных признаков можно с помощью «корреляционного кольца»: расположив по кругу индексы всех признаков, соединяем линией те из них, которые имеют коэффициент кор-

реляции выше некоторого заранее принятого порога. Если, например, принять уровень $r > 0.95$, выделится плеяда из пяти наиболее тесно связанных между собой промеров п1, п3, п5, п7, п8 (рис. 6.2.1, Б). Остальные три признака (особенно п2 и п6) хуже коррелируют как с первой плеядой, так и друг с другом. Таким образом, в группе промеров листа томата сформировались одна плеяда из пяти членов и три «свободных» признака.

Таблица 6.2.1. Промеры третьего листа томата (А) и корреляция между ними (Б)

	А. Промеры, мм							
Дата	**п1**	**п2**	**п3**	**п4**	**п5**	**п6**	**п7**	**п8**
15.4	30	7	27	12	36	19	27	12
19.4	39	8	31	14	38	19	30	15
23.4	46	10	32	14	39	19	32	15
27.4	47	12	33	15	39	19	32	15
2.5	49	13	34	15	39	20	34	16
6.5	51	12	34	16	40	20	34	16
10.5	53	10	34	15	41	20	34	16
17.5	53	10	34	15	41	21	34	16

Б. Корреляция между промерами							
	п1	**п2**	**П3**	**п4**	**п5**	**п6**	**п7**
п2	0.73						
п3	0.**97**	0.78					
п4	0.91	0.82	0.91				
п5	0.**97**	0.54	0.91	0.84			
п6	0.70	0.35	0.63	0.56	0.75		
п7	0.**98**	0.79	**0.98**	0.92	0.92	0.71	
п8	0.93	0.71	**0.96**	0.93	0.89	0.62	**0.95**

Причиной этого явления выступает *аллометрия*, неравномерные темпы изменения величины отдельных частей организма, которое проявляется как криволинейная зависимость между показателями.

Коэффициент корреляции, *основанный на линейной модели* метода наименьших квадратов, реагируя на кривизну зависимости, дает *заниженные оценки* в действительности очень тесным биологическим зависимостям.

Когда признаки *изменяются во времени* сходным образом, как, например, промеры п3 и п7 (рис. 6.2.2, А; по осям время и величина промера), график *зависимости между ними* близок к линии (рис. 6.2.2, Б; по осям значения промеров). Линейное соотношение признаков оценивается очень высоким значением коэффициента корреляции $r_{37} = 0.98$. Пластинки параллельно росли и достигли предельного размера.

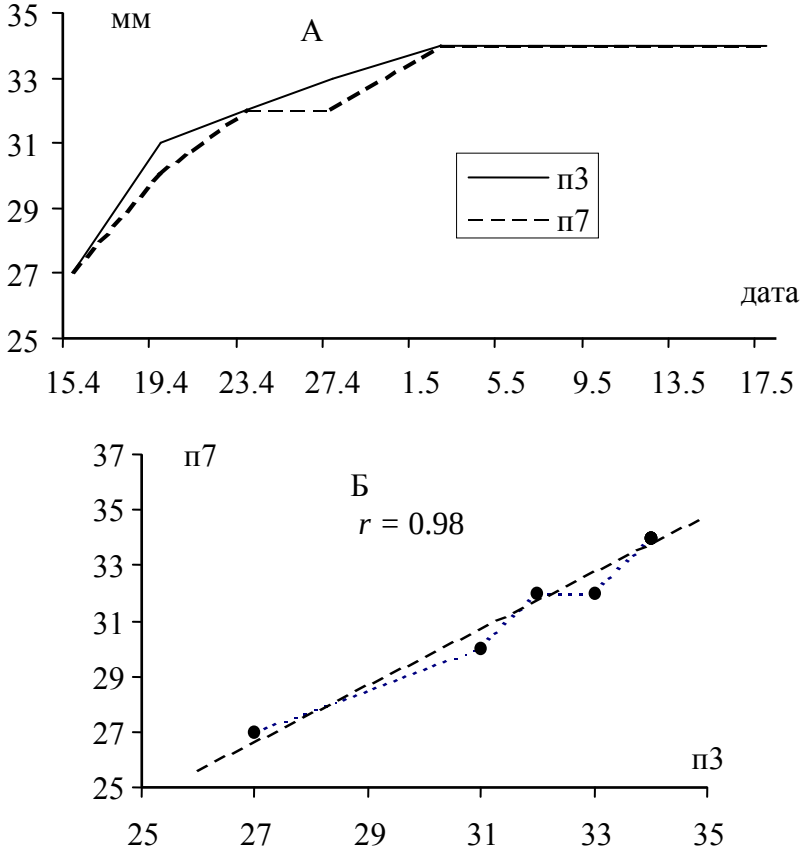


Рис. 6.2.2. Изменение промеров п3 и п7 с течением времени (А) и сопряженное их изменение относительно друг друга (Б)

Если же в процессе роста объекта промеры (п2 и п6) его частей изменяются несинхронно (рис. 6.2.3, А; по осям время и величина промера), то график их взаимозависимости имеет криволинейную форму (рис. 6.2.3, Б; по осям – промеры). Поскольку коэффициент корреляции основан на линейной модели связи признаков, наблюдаемое криволинейное соотношение признаков оценивается им довольно низко $r_{26} = 0.35$. Биологическое явление состоит в том, что листок перестал расти, а черенок стал первым усыхать.

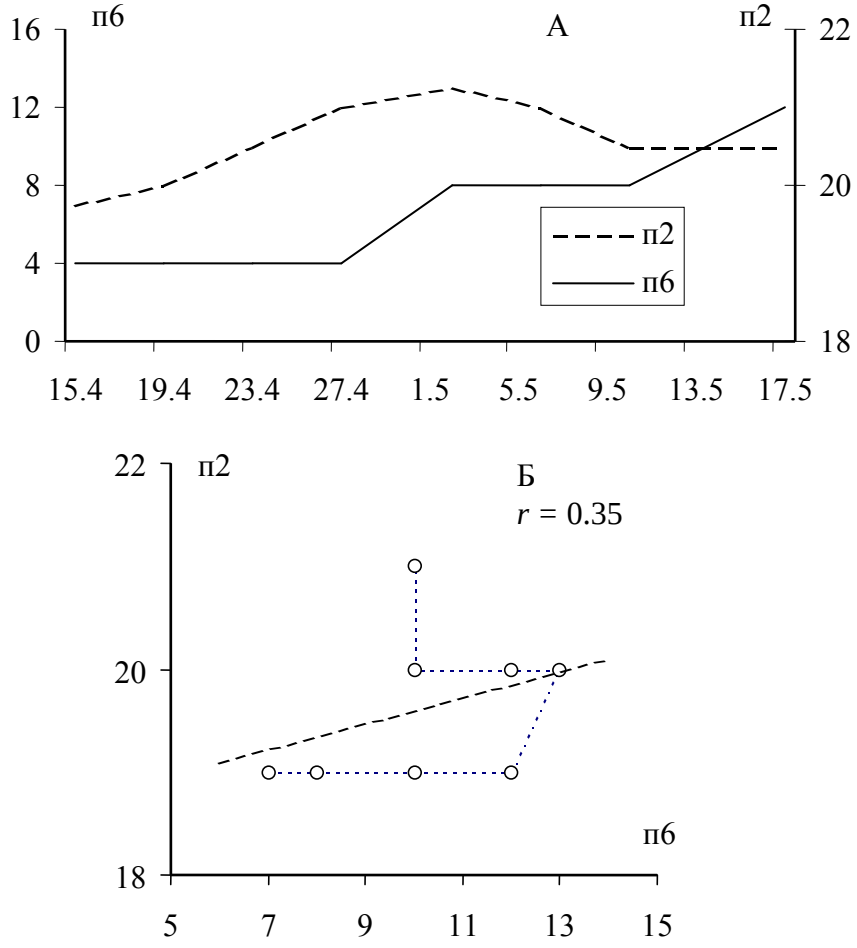


Рис. 6.2.3. Изменение промеров п2 и п6 с течением времени (А) и сопряженное их изменение относительно друг друга (Б)

Иными словами, признаки с синхронной динамикой коррелируют сильнее, чем с асинхронной. Однако эта корреляция является исключительно статистическим феноменом, она указывает лишь на криволинейность взаимосвязи признаков (которую не в силах «правильно» учесть линейный коэффициент корреляции), а не на «силу» их биологической зависимости. Биологическая причина наблюдаемой аллометрии состоит в том, что формирование разных плеяд проходит под управлением разных внешних или внутренних факторов. Разнонаправленная динамика роста отдельных частей организма приводит к тому, что в корреляционной матрице присутствуют коэффициенты разной величины, а вся совокупность признаков распадается на несколько корреляционных плеяд.

Корреляционные плеяды образуются при наблюдении не только динамических рядов, но и выборок статических оценок. Изучали биотопическое распределение одиннадцати видов мелких млекопитающих в шести биотопах Прибайкалья (табл. 6.2.2). В качестве признака здесь выступает показатель «Численность вида в данном биотопе». Корреляции, рассчитанные между столбцами исходной таблицы, показывают корреляционное сходство между разными биотопами по соотношению разных видов. На уровне $r = 0.8$ выделились две плеяды (кедровник и смешанный лес, а также экотон, сосняк и березняк) и один независимый биотоп (луг). Причина этого явления состоит в том, что плеяды признаков (здесь – биотопов) будут отличаться друг от друга по набору именно для них специфичных объектов (видов).

Типичные обитатели хвойных лесов – это средняя, малая и равнозубая бурозубки, лесная мышь и красная полевка. Они пропорционально многочисленны и в кедровнике, и в смешанном лесу (рис. 6.2.4), в результате чего диаграмма их взаимного сопряжения (рис. 6.2.5, А) очень близка к линии, следовательно, высок и коэффициент корреляции $r_{КСм} = 0.81$. В биотопах из другой плеяды, например в березняке, указанные таежные виды имеют низкую численность (особенно средняя бурозубка и лесная мышь) (рис. 6.2.4). На диаграмме соответствия этих оценок (рис. 6.2.5, Б) значения численности данных видов далеко отстоят от корреляционной линии, поэтому коэффициент корреляции оказывается существенно ниже $r_{КБ} = 0.46$. Для этого случая заключение может быть следующим:

сходство двух статических признаков тем выше, чем больше объектов близкого статуса представлено в выборке.

Таблица 6.2.2. Оценки встречаемости (экз./100 конусо-суток) мелких млекопитающих в шести биотопах Прибайкальской равнины (А) и корреляция между ними (Б)

А

Виды	**К**	**См**	**Э**	**С**	**Б**	**Л**
Обыкн. бурозубка	6	8	6	8	9	4
Средняя бурозубка	9	15	3	1	3	0
Малая бурозубка	8	4	4	2	4	4
Равнозубая бурозубка	4	2	1	0	1	0
Кутора	0	0	3	0	0	0
Азиатская лесная мышь	8	6	0	1	1	0
Лесной лемминг	1	1	2	1	1	1
Темная полевка	0	1	2	0	1	5
Полевка-экономка	0	1	2	0	1	1
Красная полевка	9	6	5	5	4	0
Красно-серая полевка	7	10	12	15	12	3

Б

Биотопы	**См**	**Э**	**С**	**Б**	**Л**
Кедровник	**0.81**	0.31	0.4	0.46	–0.08
Смешанный		0.43	0.5	0.57	–0.09
Экотон			**0.95**	**0.92**	0.35
Сосняк				**0.96**	0.33
Березняк					0.46

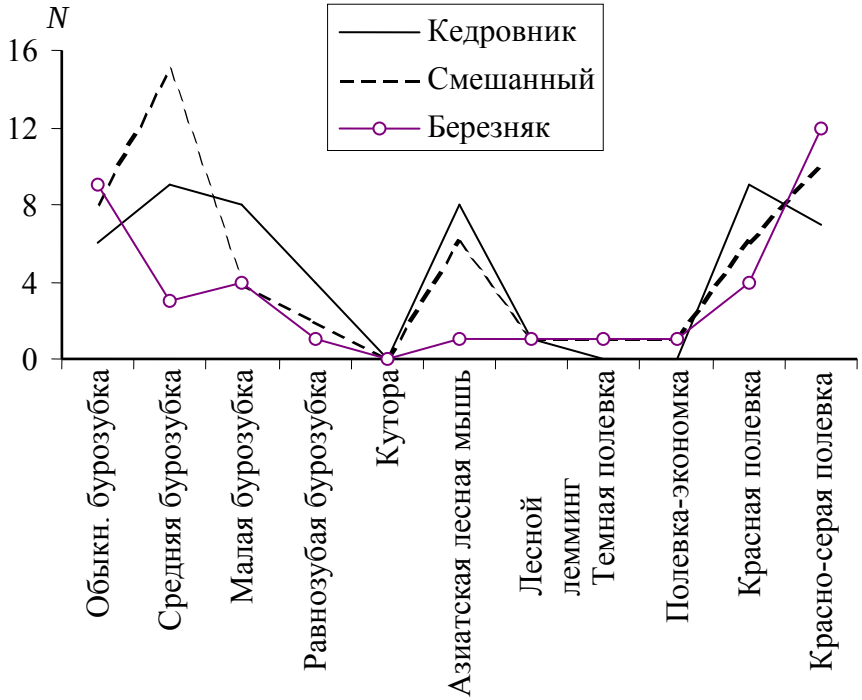


Рис. 6.2.4. Численность (*N*, экз./100 конусо-суток) 11 видов мелких млекопитающих в трех биотопах

Для рассмотренных ситуаций выделяется одно ясное доминирующее направление изменения признаков: в первом случае осуществляется временная динамика, во втором присутствует разнокачественность объектов. Природные выборки, как правило, представляют собой пересечение сразу нескольких тенденций в изменчивости изучаемых параметров: временных, пространственных, статических, что неизмеримо затрудняет интерпретацию плеядной структуры, повышая вероятность ошибочного вывода.

Тем не менее разобраться в тонкостях их взаимосвязей метод корреляционных плеяд вполне позволяет. Основной прием метода корреляционных плеяд – анализ корреляционной матрицы путем последовательного ужесточения критерия вхождения признака в плеяды. Взяв для начала невысокий порог, можно констатировать, что все признаки связаны со всеми. По мере повышения уровня

рассматриваемых корреляций гомогенная масса взаимозависимых признаков постепенно распадается на плеяды, демонстрируя структуру корреляционных связей в изучаемой совокупности. Биолог, отображая изучаемые зависимости в диаграммах, получает возможность интерпретировать общее для нескольких признаков *направление корреляции* как результат влияния на объекты некоего общего фактора, как отражение совместного участия группы органов в осуществлении одной функции.

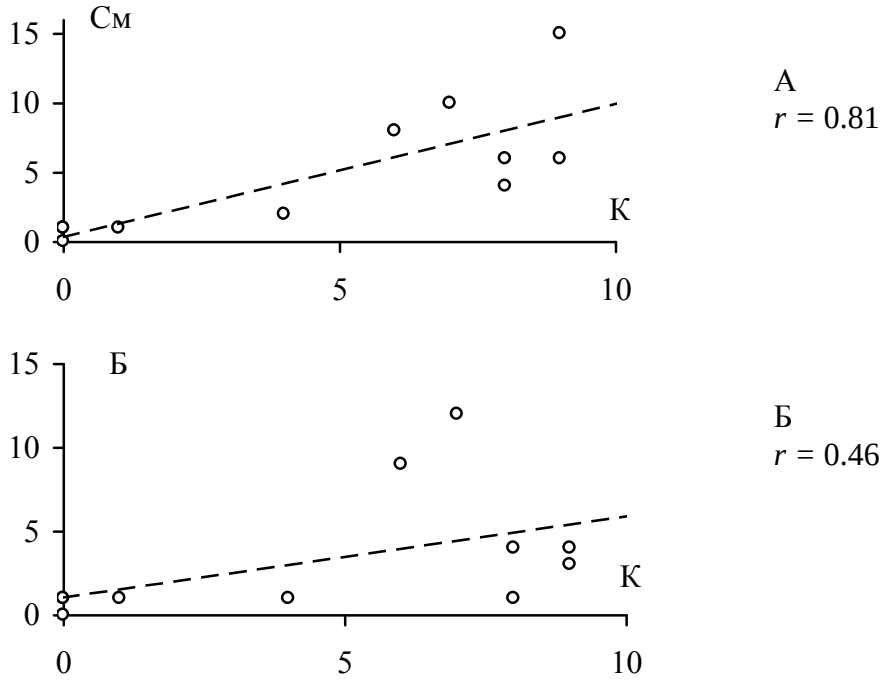


Рис. 6.2.5. Соотношение численности 11 видов между кедровником и смешанным лесом (А), кедровником и березняком (Б)

Метод корреляционных плеяд во многом аналогичен очень эффективному методу главных компонент (п. 8.1). Он также использует последовательную декомпозицию структуры корреляционных связей и позволяет выявлять «чистые» направления изменчивости в изучаемой выборке, соответствующие биологическим закономерностям поведения признаков. Это открывает дорогу для последующего количественного описания явлений в виде модели.

6.3. RMA-регрессия

Регрессионный анализ, имеющий обширную сферу применения в биологии, тем не менее, идеологически ограничен. Благодаря процедуре расчета коэффициентов, уравнение регрессии ориентировано в основном на отображение зависимости одной переменной от другой, на выражение связей вида «причина-следствие», «аргумент–функция», «фактор–признак», «стимул–реакция», «доза–эффект». Ось ординат является главной. Именно на нее ориентирован метод наименьших квадратов: в формулах расчета *лучшего* уравнения регрессии ($\hat{y} = ax + b$) заложено условие минимизации суммы квадратов отклонений ординат y_i от линии регрессии: $\sum(y_i - \hat{y}_i)^2 \rightarrow \min$. Следствием такого подхода (статистически наиболее обоснованного) оказывается невозможность оперировать с уравнением регрессии, как с алгебраическим уравнением, то есть запрещение обратного прогноза значений x по величине y. Для этих целей требуется рассчитывать второе уравнение регрессии, принимая за главную уже ось абсцисс: ($\hat{x} = cy + d$). Эти уравнения в общем случае соответствуют разным линиям, и их коэффициенты взаимно не обратны: $a \neq 1 / c$. Относительно эллипса рассеяния вариант эти линии регрессии, пересекаясь в центре, расходятся. Лишь при функциональной линейной зависимости они совпадают (см. рис. 6.1.3, 6.1.4).

Помимо отмеченных видов отношений биология рассматривает разнообразные случаи взаимозависимости пар переменных, среди которых невозможно выделить «ведущий» признак. Примерами подобных равноценных переменных могут служить характеристики (масса, размеры, состав, число структур) разных органов выборки особей, оценки таксономического состава и численности разных видов в серии биоценозов, демографические характеристики популяции, концентрации разных веществ в воде водоемов и пр. Для этих случаев желательно иметь биологически оправданные обобщенные показатели взаимосвязи, «симметричные» относительно обоих признаков. Этому требованию отвечает коэффициент корреляции, но он не позволяет рассчитать значения переменных для точек, где данные отсутствуют (выполнить интерполяцию или экстраполяцию). Выходом из положения могло бы стать создание аппарата построения уравнения «биологической» связи, строящего ли-

нии зависимости, проходящие вдоль по оси эллипса рассеяния, независимо от величины случайного варьирования.

Один из подходов к построению уравнения связи между двумя показателями основан на использовании третьего (реперного) признака, относительно которого можно разделить сопряженные дисперсии, слитно представленные в ковариации, отдельно для каждого из двух признаков, и считать их отношение (коэффициент пропорциональности) показателем истинной биологической связи (Смирнов, 1979).

Много более прост с вычислительной точки зрения другой прием – метод «ликвидации главной оси» – **r**educed **m**ajor **a**xis regression (RMA-регрессия) (Полищук, Цейтлин, 2001). Принцип поиска коэффициентов уравнения подобен методу наименьших квадратов (МНК) – линия регрессии должна так проходить между эмпирическими точками, чтобы расстояние от них до линии было минимальным. Однако в случае RMA *минимизируются* не расстояния линии до точек относительно признака *y*, а абсолютные расстояния – сумма нормалей (перпендикуляров), опущенных из каждой точки на линию.

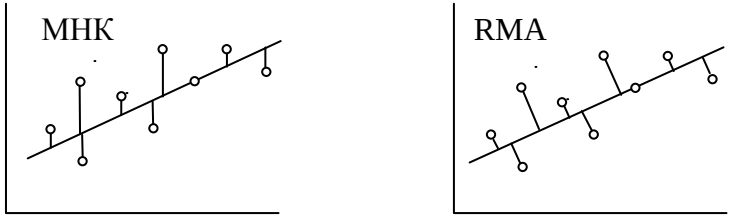


Рис. 6.3.1. Методы измерения отклонений точек от линии

По существу дела, RMA-метод стремится выразить общую тенденцию изменения обоих признаков независимо от реальной силы связи между ними, без учета сопряженной изменчивости. В этом случае для расчета коэффициента регрессии можно использовать обобщенное отношение отклонений от своих средних значений обоих признаков $v = \sqrt{\frac{\sum(y - My)^2}{\sum(x - Mx)^2}} = \frac{S_y}{S_x}$, которое равно простому отношению стандартных отклонений.

Если предварительно по набору вариант было рассчитано уравнение регрессии ($\hat{y} = ax + b$) и коэффициент корреляции (r) (см. п. 8.5), то можно использовать и другую формулу: $v = a / r$. Очень важно, что эта формула сообщает коэффициенту знак (такой же, как и у коэффициента корреляции). Свободный член вычисляется по общему принципу: $u = My - v \cdot Mx$.

В результате имеем уравнение линии RMA-регрессии: $\hat{y} = vx + u$. Важнейшим его свойством оказывается точное совпадение с аналогичным уравнением и линией регрессии относительно переменной x: $\hat{x} = px + q$; их коэффициенты обратно пропорциональны: $v = 1 / p$. Это свойство делает ненужным расчет второго уравнения и позволяет обращать первое. Ошибка коэффициента RMA-регрессии принимается равной ошибке коэффициента обычной регрессии: $m_v \approx m_a$.

При большой доле случайной изменчивости ($r = 0$) коэффициент RMA-регрессии не определен. По этой причине проверка гипотезы Но: $v = 0$ лишена смысла. В то же время, если нулевая гипотеза о равенстве нулю исходного коэффициента регрессии ($a = 0$) опровергнута с помощью критерия Стьюдента ($t = a / m_a \geq t_{(0.05,\ n-2)}$), то коэффициент v также считается значимым, что дает возможность пользоваться уравнением RMA-регрессии.

Для иллюстрации этих идей используем результаты расчета уравнения регрессии (п. 6.1). Исходя из полученных параметров $r = 0.975$, $a = 0.0657$, $S_y = 5.2$, $S_x = 77.1$, $M_y = 32$, $M_x = 450$, новые коэффициенты равны:

$$v = \frac{a}{r} = \frac{0.657}{0.975} = 0.0674 \text{ или } v = \frac{S_y}{S_x} = \frac{5.2}{77.1} = 0.0674,$$

$$u = M_y - a \cdot M_x = 32 - 0.0674 \cdot 450 = 1.67.$$

В силу высокой корреляции между признаками полученное уравнение RMA-регрессии $\hat{y} = 0.0674\ x + 1.67$ несильно отличается от уравнения МНК-регрессии $\hat{y} = 0.0657x + 2.419$.

Глава 7
ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ

Моделирование пока не столь широко распространено, как того требуют сложные задачи современной биологии, особенно экологии. На наш взгляд, одним из препятствий этому служит распространенное мнение, что «полноценными» могут быть лишь дающие прогноз аналитические модели; сопряженные с этим сложности построения системы дифференциальных уравнений и их решения оказываются серьезным препятствием для большинства биологов. Однако изучаемые экологические явления сначала нужно понять, дать им объяснение, а уж затем, при необходимости, и прогнозировать. Мы предлагаем давать *количественное объяснение* с помощью имитационного моделирования – составлять модели, основанные на простейших (линейных) алгебраических уравнениях, и определять значения их параметров посредством внешних *процедур оптимизации*.

Вместо составления и решения дифференциальных уравнений следует *составлять программы* и *настраивать параметры* имитационных моделей. Обе эти проблемы довольно просто решаются с помощью табличных процессоров, например Microsoft Excel.

Способ построения моделей на листе Excel отличается от традиционных приемов программирования (алгоритмического, структурного или объектного) – это *табличное программирование*. На листе Excel модель предстает во всех своих деталях, как таблица, ячейки которой заполнены формулами, имитирующими либо выборку вариант (*статические, описательные модели*), либо ход процесса (*динамические модели*). Каждая ячейка содержит формулу, которая вычисляет соответствующее «модельное» значение варианты или характеристику системы на очередном временно́м шаге. Поскольку «объясненные» значения модельных переменных должны более или менее совпадать с реальными наблюдениями, организуется процедура поиска таких (оптимальных) значений модельных параметров, которые делают отличия между моделью и реальностью наименьшими, минимизируют «функцию отличий» («функцию невязки»). Эта процедура оптимизации выполняется с помощью отдельной программы Поиск решения, встроенной в пакет Excel.

(Ответственное отношение к моделированию требует понимания существа процедуры настройки, подробнее см.: Коросов, 2002). Помимо программирования самой модели и настройки ее параметров требуется доказать значимость модельных параметров или адекватность модели.

Для решения всех рассмотренных задач на листе Excel приходится конструировать *имитационную систему*, включающую следующие элементы:

- блок исходных данных, состоящий из массива независимых и зависимых переменных,
- блок расчета модельных данных, собственно имитационная модель,
- блок параметров, участвующих в расчете модели и изменяемых в процессе настройки,
- блок расчета отличий реальных и расчетных значений переменных,
- значение функции невязки (сумма отличий между моделью и реальностью), которое минимизируется в процессе настройки,
- блок процедуры настройки (программа «Поиск решения»),
- блок графического представления результатов,
- блок статистической оценки результатов.

В форме имитационной системы мы обладаем очень гибким инструментом описания действительности. В потенциях имитационной модели стать сложной и детализированной или, напротив, простой и обобщающей, выражающей законы, управляющие миром.

Процесс построения модели проходит обычно следующие этапы:

1) Определение конкретной цели моделирования (создание количественного описания природных процессов и явлений с точки зрения исследователя).

2) Описание структуры данных (экземпляры, показатели, переменные, параметры) и выборок (объемы, условия, методы).

3) Построение блок-схемы по правилам: *стрелки* отображают потоки, которые имеют единицы изменения; *прямоугольники* отображают функции преобразования потоков, своих единиц измерения не имеют.

4) Математическое описание модели, то есть представление всех связей (преобразований) переменных в форме уравнений, выявление обратных связей, нелинейности.

5) Программирование модели: организация дискретного хода счета, описание переменных, параметров, программирование (ввод формул) и отладка блоков.

6) Настройка (параметров) модели: подбор данных, организация имитационной системы (в первую очередь блоков функции невязки для последующей настройки параметров.

7) Верификация, оценка значимости модельных параметров.

8) Эксперименты с моделью при разных режимах, условиях, заданиях; статистическое описание и оценка результатов.

7.1. Составление формул имитационной модели

Перед составлением системы модельных уравнений следует тщательно описать входящие в их состав численные показатели. Все количественные характеристики модели делятся на переменные величины и константы (параметры).

Переменные характеризуют среду, окружающую изучаемую систему, изменчивое состояние самой системы в целом и ее частных элементов. В терминах блок-схемы переменные – это «потоки», количественное проявление способа существования статических компонентов системы. Отдельный элемент может быть описан несколькими переменными. Поэтому говорят о разных языках описания системы: каждый язык относится к одному виду потоков данного уровня иерархии. Потоки преобразуются в процессе функционирования системы: либо изменяют свою величину, либо трансформируются в другие потоки.

Параметры количественно выражают «режимы» (скорость, интенсивность) преобразования потоков (изменения значений переменных). В отличие от переменных величин параметры обычно задаются неизменными, во всех модельных расчетах они остаются независимыми от состояния системы (они изменяются только в процессе *настройки* модели).

Цель имитационной модели состоит в том, чтобы рассчитать параметры механизма конкретного явления, поэтому формулы имитационных моделей ориентированы на фактический материал. Ими-

тационная модель должна не просто математически описать объект, но описать его таким образом, чтобы вычислялись те же самые переменные, что наблюдались. Модельная формула должна, во-первых, выражать суть механизма явления, во-вторых, рассчитывать аналог матрицы исходных данных. В этом и заключается главное правило составления модельных формул: *выразить известные переменные* (Y_j) *через неизвестные параметры* (a_j).

Формулы имитационной модели вычисляют значения модельных переменных, для которых имеются (это явные переменные) или отсутствуют (это скрытые переменные) аналоги в массиве исходных данных (среди зависимых переменных). Существует два принципиально разных пути модельного отображения реальности (два типа имитационных моделей): описание процесса в целом (составление описательной модели) и характеристика скорости процесса (построение динамической модели).

Описательная модель выражается одним уравнением, которое охватывает процесс целиком и позволяет непосредственно вычислить значения переменных при любых значениях независимой переменной (в любой момент времени). По своим результатам такое моделирование подобно регрессионному анализу. По нему можно сразу рассчитать функцию для данного значения аргумента. В зависимости от целей и наличных данных можно выделить две конструкции обобщенной модели. Обобщенное описание *зависимости процесса от времени* можно выразить формулой: $Y = f(A, T)$ (например, $y_i = a \cdot i + b$; $i = 1, 2 \ldots, T$). Обобщенное описание *зависимости одной переменной от другой* можно выразить формулой: $Y = f(A, X)$ (например, $y_i = a \cdot x + b$). Виды уравнений (линейное, степенное, полиномиальное, логистическое) рассматриваются в биометрических пособиях (Коросов, 2002; Ивантер, Коросов, 2003).

Динамическая модель служит для описания скорости и результата процесса, она характеризует причинный механизм его протекания, поскольку дает возможность сконцентрировать внимание на исследовании каждого момента осуществления процесса. В среде Excel динамические имитационные модели представлены как минимум двумя уравнениями. Одно из них позволяет рассчитать сиюминутный результат процесса для каждого шага. Второе аккумулирует (суммирует) частные результаты, полученные на предыдущих шагах процесса. Для расчета значения функции в заданный конкретный

момент времени (на данном шаге, в отдельной ячейке Excel) требуется предварительно воссоздать ход всего процесса с самого начала. В зависимости от целей исследования конструкции динамических моделей различаются.

Модель автономного процесса (текущее значение переменной определяется ее предыдущим значением) комплектуется из двух формул: во-первых, это частный результат процесса на данном шагу (за данный промежуток времени, скорость процесса):

$$dy_i = a \cdot y_{i-1} + b;$$

во-вторых, интеграция всех предыдущих частных результатов в один общий:

$$y_i = y_{i-1} + dy_i,$$

где y_i – текущее значение переменной (характеристика состояния), y_{i-1} – значение переменной на предыдущем (i–1-м) временно́м шаге (характеристика предыдущего состояния), dy_i – прирост переменной за i-й временной шаг, a, b – коэффициенты пропорциональности, $i = 1, 2 \ldots, T$ – счетчик временных шагов модели.

Модель обусловленности средой описывает иную ситуацию, когда текущее значение модельной переменной определяется как предыдущим значением (предыдущим состоянием моделируемой системы), так и значением внешней переменной (x). Число формул увеличивается:

$$d_1y_i = a \cdot y_{i-1} + b;$$
$$d_2y_i = c \cdot x_{i-1} + d,$$
$$dy_i = d_1y_i + d_2y_i,$$
$$y_i = y_{i-1} + dy_i,$$

где d_1y_i, d_2y_i – приросты переменной за счет внутренних потенций системы и в результате внешнего влияния, a, b, c, d – коэффициенты пропорциональности.

Модель *взаимной обусловленности*, *обратной связи* рассматривает ситуацию, когда текущее значение одной модельной переменной определяется предыдущим значением другой переменной и одновременно текущее значение второй переменной определяется предыдущим значением первой:

$$dx_i = a \cdot y_{i-1} + b,$$
$$dy_i = c \cdot x_{i-1} + d,$$
$$y_i = y_{i-1} + dy_i,$$
$$x_i = x_{i-1} + dx_i.$$

В этих примерах мы привели линейные уравнения для выражения скорости, но можно использовать любую другую формулу из имеющегося арсенала (степенное, показательное).

Табличное программирование

В рамках имитационной системы выбранные уравнения вводятся на лист электронной таблицы Excel. Такое программирование можно назвать *табличным*. Оно обладает рядом черт, отличительных от алгоритмического, блокового, структурного и объектного.

1. Программировать в электронной таблице *очень просто*, это доступно любому пользователю, в общих чертах знакомому с пакетом Excel. Ввод формул в ячейки листа Excel и есть программирование. Для этого достаточно только:

- знать правила ввода формул (проще всего создавать формулы с помощью мыши и затем, при необходимости, редактировать),
- знать правила формирования ссылок на ячейки (абсолютные ссылки, которые не изменяются при автозаполнениях, снабжены значками валюты $),
- уметь среди множества функций листа Excel выбрать нужную (они вызываются кнопкой *f()*), обычно требуется не больше 10 функций (СУММ(), СТАНДОТКЛОН(), СЛЧИС(), СЧЁТ(), ...).

2. Программируя на электронном листе, не нужно заботиться об организации ввода-вывода и интерфейсе, поскольку Excel реализует эти функции самым «дружелюбным» образом. Не только данные, но и вся имитационная система с моделью и ее графическим отображением сохраняются в одном файле. Понятно, что в случае необходимости объем данных можно изменить, увеличив длину ряда. При этом важно следить за тем, чтобы диапазоны ссылок в формулах (суммирование и др.) включали бы и новые данные.

3. На листе Excel не нужно конструировать циклы, необходимые в других пакетах для организации хода времени или для перебора вариант выборки. Табличная форма предполагает «разворачивание» временно́й последовательности на пространстве электронного листа, например, сверху вниз. Тогда ячейки одной *строки* содержат значения переменных, характеризующих состояние исследуемого объекта в *определенный момент времени*, а набор строк от-

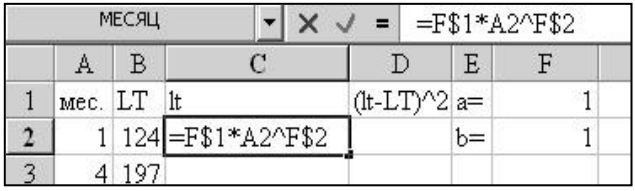
ражает последовательную смену состояния объекта во времени. Большое число строк листа Excel (65563) позволяет отобразить сколь угодно продолжительную динамику (или сколь угодно большую выборку). Модель на листе Excel предстает как множество стереотипных формул, рассчитывающих значения всех переменных для каждого момента времени. Ввод множества формул не представляет проблемы, благодаря предусмотренной в Excel процедуре «автозаполнение». При этом остается только следить за правильностью ссылок на параметры (они должны быть абсолютными).

4. Программа на листе Excel оказывается абсолютно прозрачной, поскольку все данные и результаты счета моментально визуализируются в численной форме и их можно охватить буквально одним взором. Сложные формулы можно разделять на ряд простых и на листе Excel отражать все промежуточные результаты счета (промежуточные переменные). Быстро ориентироваться в структуре перекрестных ссылок позволяют команды панели “Зависимости”, связывающие стрелками ячейки вызываемые и вызывающие ячейки.

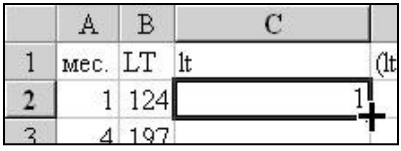
7.2. Статические модели (аппроксимация)

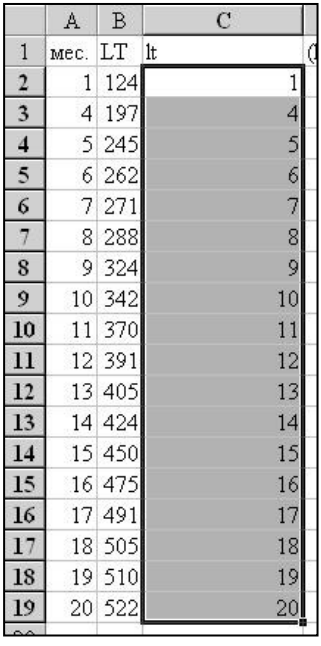
Статические модели похожи на регрессионные; они выполняют расчет модельных значений, опираясь на исходные реальные выборочные данные. Однако процедура расчета модельных параметров меняется: вместо метода наименьших квадратов используется итеративная подгонка результатов счета под исходные значения. Вследствие этого расширяются возможности для *аппроксимации* (приблизительного количественного описания) зависимостей с помощью уравнений любой сложности, например, гиперболы $y = a / x + b$, логистической кривой $y = C + A/[1 + e^{a*x+b)}]$, снимаются ограничения на пропуски в данных, а для случая криволинейных зависимостей результаты расчетов уточняются. Действия по предлагаемому ниже образцу позволят быстро получить навык применения предлагаемого метода имитационного моделирования.

Рассмотрим динамику роста самца обыкновенной гадюки в условиях неволи. Длину тела (LT, мм) измеряли раз в месяц (мес.), данные занесли на лист Excel. Далее называем поля (расчетные значения размеров lt, параметры a, b), вводим в ячейки F1 и F2 первоначальные условные значения параметров, равные единице. Вводим первую формулу уравнения степенной функции (которая лучше других описывает рост) в формате Excel: вместо значений указываем ссылки, причем ссылки на ячейки с параметрами должны быть абсолютными, то есть содержать префикс $.

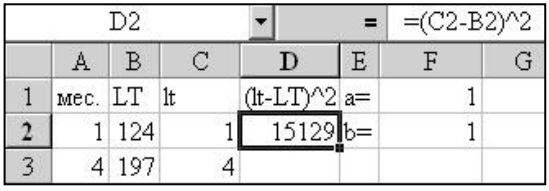


Введенная формула рассчитает «теоретическое» значение длины тела гадюки в возрасте 1 мес. ($1 \cdot 1^1 = 1$), но пока явно неверное. Скопируем с помощью процедуры «автозаполнение» эту формулу в ячейки, соответствующие всем другим месяцам жизни гадюки. Для этого наводим мышку на правый нижний угол ячейки с формулой (курсор из белого крестика становится черным) и с помощью левой кнопки тащим уголок до последней строки данных.





Определить меру отличия модельных расчетов (lt) от реальных данных (LT) позволяет квадрат разности между этими значениями (lt-LT)^2. Вводим соответствующую формулу на лист Excel, для первого месяца получаем величину $(124-1)^2 = 15129$. Путем автозаполнения копируем эту формулу в остальные ячейки столбца.

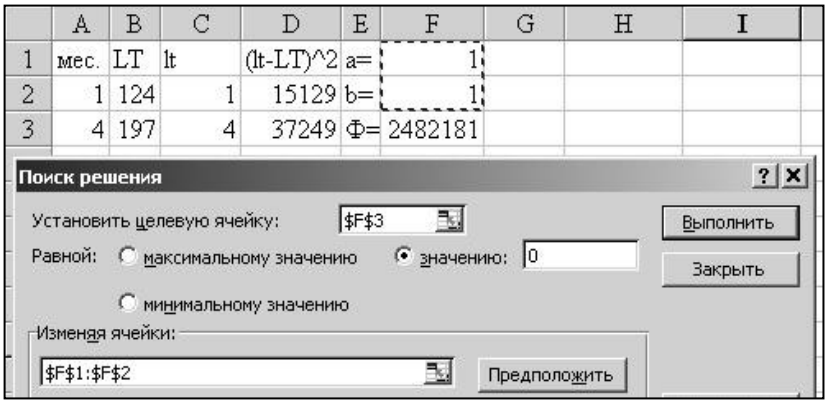


Подсчитываем сумму отличий, невязку Ф. Она оказалась очень большой (Ф = 2482181) и свидетельствует о плохом соответствии модели и реальности.

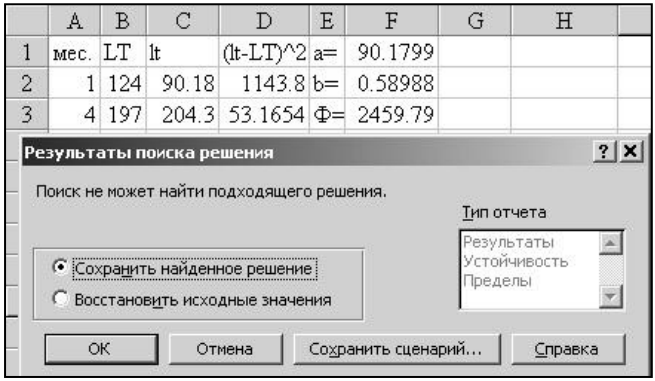
F3 = =СУММ(D2:D19)

	A	B	C	D	E	F	G
1	мес.	LT	lt	(lt-LT)^2	a=	1	
2	1	124	1	15129	b=	1	
3	4	197	4	37249	Ф=	2482181	
4	5	245	5	57600			

Теперь отыщем параметры. Вызываем программу настройки командой главного меню: Сервис \ Поиск решения.



В появившемся окне мышкой устанавливаем целевую ячейку со значением невязки Ф ($F$3), Равной значению 0, Изменяя ячейки $F$1:$F$2, Выполнить. После расчетов появится окно результатов, говорящее о том, что решение не найдено. Это не удивительно, ведь была поставлена задача *обнулить* сумму отличий модели от реальности, а его удалось лишь уменьшить (в тысячу раз).



Для биологических целей этого обычно бывает достаточно (проверка статистической значимости составляет отдельную тему: см.: Коросов, 2002).

Нажимаем OK, рассматриваем результаты. Теоретическая линия ($lt = 90.2 \cdot i^{0.589}$, i – возраст, мес.) прошла очень близко к реальным промерам. Смысл параметров a = 90.2 и b = 0.59 состоит в следующем. Старт роста начинается с величины, близкой к 90 мм и каждый месяц размеры увеличиваются все в меньшей степени (равномерный рост имел бы степенной коэффициент b = 1). Аналогичное уравнение для самок $lt = 110.7 \cdot i^{0.538}$ показывает, что они изначально имеют большие размеры и растут быстрее, чем самцы.

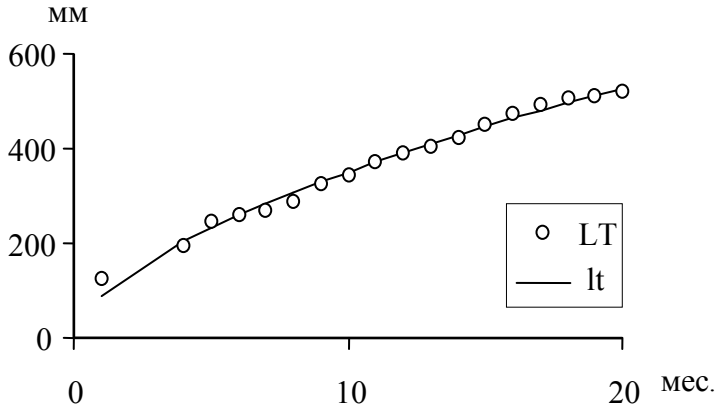


Рис. 7.2.1. Статическая модель динамики роста самцов гадюки

7.3. Динамические модели процессов

Детали хода исследуемого процесса удобно изучать с помощью динамических имитационных моделей. Они структурно более сложны, чем статические модели, поскольку призваны отображать как скорость процесса (локальный результат процесса в единицу времени), так и интеграцию всех частных достижений. Для расчета зависимых переменных, характеризующих результат процесса, достигнутый к определенному моменту времени, приходится восстанавливать всю предшествующую ему динамику. В этом смысле динамическая имитационная модель «живет своей жизнью», а параметры скоростных уравнений характеризуют ее механизм.

Рассмотрим пример расчета абсолютной численности островной популяции травяной лягушки. На протяжении двух лет подсчитывалось число кладок икры (равное числу самок) в репродуктивных водоемах. Оценки совпали $N = 3467$ и $N = 3649$, что позволяет оценить абсолютную численность взрослых особей в 7000 экз.

Для расчета численности всей популяции необходимо реконструировать недостающие значения по молодым лягушкам. Предположив постоянство численности островной популяции (на это указывают наши оценки) и опираясь на известные из литературы оценки демографических параметров, можно в среде Excel реконструировать стационарное возрастное распределение и на его основе рассчитать общую абсолютную численность (табл. 7.3.1).

Величина группы годовиков (N_1) рассчитывается как доля (s_0) икринок, из которых развились и успешно перезимовали особи, а число икринок (N_0) – как суммарная плодовитость (e) всех взрослых самок (N_{fad}):

$N_0 = N_{\text{fad}} \cdot e$, B5 = B1 * B2,

$N_1 = N_0 \cdot s_0$, B6 = B1 * C1.

При средней плодовитости $e = 1800$ и общем количестве самок $N_{\text{fad}} = 3500$ имеем $N_0 = 6300000$ икринок. Выживаемость личинок летом и сеголетков зимой сильно варьирует – от долей до нескольких процентов, для первого случая берем округленное модальное значение $s_0 = 1\%$ (C1 = 0.01), отсюда $N_1 = N_0 \cdot s_0 = 63000$ годовиков. В пределах возрастного распределения текущая численность каждой возрастной группы равна численности предыдущей возрастной группы год назад без числа погибших особей, то есть равна численности выживших особей (s_i – выживаемость i-го возрастного класса; $i = 1, 2\ldots, 9$): $N_{i+1} = N_i \cdot s_i$, B7 =B6 * C2… B11 =B10 * C3… Как известно, молодые и взрослые особи имеют разную выживаемость. Это позволяет ввести в модель и оценивать лишь два параметра – выживаемость младших (1–5 лет) и старших (6–8 лет) особей: $s_{1–5} = s_1 = s_2 = s_3 = s_4 = s_5$, $s_{6–8} = s_6 = s_7 = s_8$, ячейки C2 и C3.

Используя приведенные уравнения, уже можно сконструировать модель (столбцы A:B), но еще нельзя приступить к настройке параметров модели (C2:C3). Для этого необходимо рассчитать некие модельные значения переменных, которые следует сравнивать с реальными значениями.

Таблица 7.3.1. Реконструкция возрастной структуры популяции лягушки при разных уровнях смертности личинок и сеголетков

для s_0 = 0.01 %

	A	B	C	D
1		3500	**0.01**	
2		1800	0.37	0.2
3			0.37	1
4	лет	Все особи	Взрослые	
5	0	6300000		
6	1	63000		
7	2	23574		
8	3	8821	1764	
9	4	3301	3301	
10	5	1235	1235	
11	6	462	462	
12	7	173	173	
13	8	65	65	
14				
15		100631	7000	

для s_0 = 0.001 %

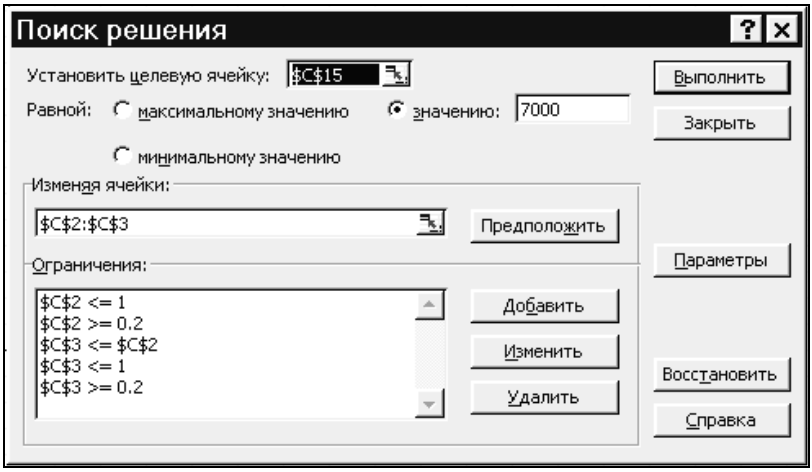
	A	B	C	D
1		3500	**0.001**	
2		1800	0.71	0.2
3			0.71	1
4	лет	Все особи	Взрослые	
5	0	6300000		
6	1	6300		
7	2	4472		
8	3	3174	635	
9	4	2253	2253	
10	5	1599	1599	
11	6	1135	1135	
12	7	806	806	
13	8	572	572	
14				
15		20311	7000	

В нашем распоряжении имеется единственное реальное значение переменной – число всех половозрелых особей, *Nad* = 7000 экз. Необходимо поэтому рассчитать соответствующее модельное значение. Из литературы известно, что в среднем продолжительность жизни травяной лягушки составляет 8 лет; с четырех лет все особи становятся половозрелыми (D3 =1), а среди трехлеток доля половозрелых составляет 20% (D2 =0.2), остальные особи – ювенальные. В модели число взрослых особей рассчитывается как доля (p_i) ото всех особей (столбец C8:C13): $N_{adi} = N_i \cdot p_i$.

Сумма этих значений дает искомое число взрослых особей: C15 =СУММ (C8:C13).

Теперь осталось только вызвать макрос настройки параметров и указать соответствующие ссылки: Установить целевую ячейку C15 Равной значению 7000, Изменяя ячейки

$C$2:$C$3. При этом на возможные значения параметров следует ввести явные ограничения. Так, они не должны превышать 1 (не может выжить более 100% особей) или быть меньше 0.2 (обычно выживаемость выше 20%). Кроме этого, выживаемость амфибий в старших возрастных группах обычно ниже, чем в младших, $C$3<=$C$2. С помощью мыши вводим эти ограничения в окно настройки, используя кнопку Добавить (рис. 7.3.1).



Ошибка!

Рис. 7.3.1. Задание условий настройки модели

В результате настройки получили единственное решение: выживаемость в возрасте 2–5 лет равна $s_{1\text{-}5}$ = 37 %, то же и в возрасте старше 5 лет – $s_{6\text{-}8}$ = 37 %. Общая расчетная численность островной популяции травяной лягушки составила около 100 000 экз.

Представленная модель оказалась очень чувствительной к параметру «выживаемость личинок и сеголетков», s_0. Взяв другое значение, на уровне одной десятой процента s_0 = 0.001 %, в результате новой настройки получаем среднюю выживаемость для других групп s_i = 0.71 %, а общую численность на уровне 20 000 экз. (табл. 7.3.1, рис. 7.3.2).

Как видно из таблицы 7.3.1, результаты двух попыток моделирования довольно сильно различаются. Какая из этих двух моделей данной структуры ближе к действительности, может решить лишь независимая оценка одного из параметров популяции, например, выживаемости взрослых особей или абсолютной численности

одного из младших возрастных классов. Получить такие данные можно только после дополнительных исследований с использованием трудоемкого метода мечения с повторным отловом.

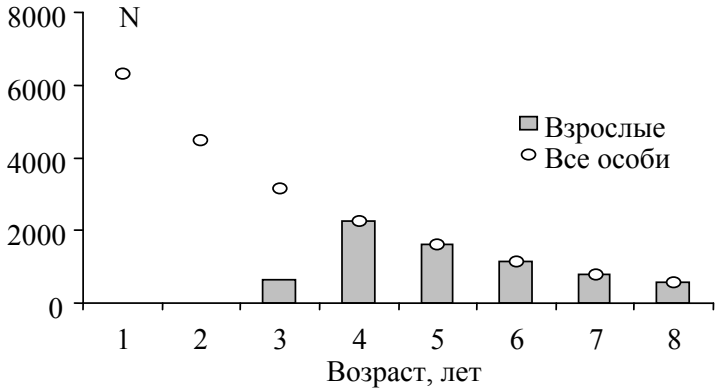


Рис. 7.3.2. Реконструкция возрастного состава популяции лягушек для $s_0 = 0.001$ %

В заключение хотелось бы подчеркнуть гносеологическую роль моделирования. Имитационная модель биосистемы – это не конечный результат исследования, но промежуточный этап процесса исследования. Когда некоторый эмпирический материал получает количественное оформление в виде имитационной модели, открывается несколько путей для продолжения исследований на этой базе. В первую очередь, это экстраполяция на область неизвестных значений зависимых переменных. Здесь экстраполяцию можно рассматривать как способ построения более общих интерполяционных моделей, как очередной шаг моделирования.

Допустим, что исследовалась биосистема на ограниченном (во времени или пространстве) наборе данных и была построена имитационная модель, хорошо их описывающая. По мере накопления новых материалов появлялась возможность проверить работу модели на независимых массивах. Если при этом модельные предсказания уровня зависимых переменных все менее соответствуют наблюдаемым значениям, то есть становились все «хуже», значит (как подсказывает опыт), либо параметры модели не улавливают всего размаха варьирования переменных, либо конструкция модели не учитывает неких новых факторов (независимых переменных),

реально участвующих в динамике изучаемой системы. Такое же положение неопределенности складывается, когда на одних и тех же эмпирических данных настройка модели до уровня адекватности дает разные значения параметров. Если при этом сделать серию прогнозов (рассчитать неизвестные значения модельных переменных, выходящих за известный диапазон изменчивости реальных данных), используя *разные* модельные параметры, то появится перспективная задача выяснить, какой из прогнозов ближе к реальности. В этих случаях модель укажет на идейный изъян, вынудит сконцентрировать внимание на тех моментах анализа природного объекта, которые исходно были плохо структурированы, то есть. будет явно диктовать направление дальнейшей работы.

Построение отдельной модели есть лишь единичный шаг в итеративном процессе исследования. Тем не менее, в публикациях обычно указывается лишь некий окончательный модельный результат длительного пути создания моделей, что создает ложное впечатление о смысле моделирования. Примеры анализа неадекватности моделей в литературе достаточно редки. В этой связи и было интересно отследить динамику исследования, возникновение неадекватных прогнозов и пути их нейтрализации. В рассмотренном примере моделирование показало, что для реконструкции возрастного состава и численности популяции лягушки первичных данных недостаточно, нужны дополнительные факты (а не теоретические посылки). Этот вывод очень важен, поскольку конкретизирует направления поиска. Имитационное моделирование указывает пути дальнейшего исследования.

7.4. Модели динамики систем*

Отдельный интерес для моделирования представляет динамика численности популяции. Такие данные представлены временным рядом: ординаты образуют оценки плотности, по оси абсцисс задано время с равными временными шагами. Колебательный характер кривых изменения численности бросается в глаза, хотя строгая периодичность зачастую отсутствует. Такая динамика напоминает случайные процессы, для описания которых в последнее время привлекают теорию неравновесных систем. Не так давно обнаружены достаточно простые математические системы, которые обладают разнообразным поведением, начиная с поддержания единственного значения и заканчивая хаотической динамикой. В основу таких систем положена идея преемственности, когда следующее состояние системы диктуется ее прошлым. Эти разработки позволяют инвертировать постановку проблемы, ориентируясь на прикладную область: нельзя ли поведение сложной (например, биологической) системы описать с помощью простых математических функций? Интрига в том, чтобы без детальной имитационной модели описать важнейшие демографические свойства популяции.

Цель нашего изложения такова: построить простую нелинейную модель, которая описывала бы динамику природной популяции животных, например, рыжей полевки. Содержание такой работы будет состоять в том, чтобы придать биологический смысл параметрам полученной модели, характеризующим способность системы к автономным реакциям определенного рода, а также в оценке возможного прогнозирования динамики популяции по этой модели.

Нелинейные уравнения

Зададимся вопросом, как перевести на язык биологии термин «модель динамики сложной системы»? Разберемся с особенностями «непериодического» поведения математических объектов и его причинами. Рассматриваемые функции неавтономной динамики организуют преемственность состояний системы, когда значения *функции* (на выходе) служат на следующем временно́м шаге *аргументом* той же функции (на входе): $x_{i+1} = f(x_i, a)$, где x_{i+1}, x_i – значение

* Раздел написан в соавторстве с А. А. Зориной (Коросов, Зорина, 2007).

регистрируемой переменной на разных (i и i+1) шагах модели в отдельные следующие друг за другом моменты времени, $f()$ – непрерывная функция, a – параметр модели.

Один из простых примеров такого рода – *дискретная логистическая модель* (Недорезов, 1997):

$$x_{i+1} = ax_i - ax_i^2 = ax_i(1 - x_i), \quad 0 \le x_i \le 1, 0 \le a \le 1 \ .$$

Уравнение включает квадратичный член, который в принципе трудно биологически интерпретировать, но который сообщает модели характерное поведение в чреде временных шагов. Эта специфика состоит в том, что динамика системы будет неразрывно связана с некоторой *параболой*, описывающей *квадратичную зависимость* состояния системы (x_{i+1}) в следующий (i+1-й) момент времени от предыдущего состояния (x_i). При этом отрицательный знак перед квадратичным членом говорит о том, что ветви параболы направлены вниз (конструкция приведенного уравнения такова, что парабола пересекает ось абсцисс в точках 0 и 1).

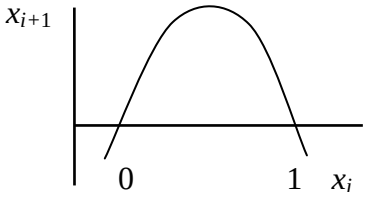


Рис. 7.4.1. Условия реализации модели, заданные параболой

Особенность параболы состоит в том, что при последовательном пересчете значений функции x_{i+1} с использованием значений x_i будет наблюдаться скачкообразная смена значений x: невысокие значения ($x_н$) «отражаются» от ветвей параболы и порождают большие значения ($x_б$) и наоборот – большие порождают невысокие (рис. 7.4.2). В ряду значений x будет наблюдаться цикличность.

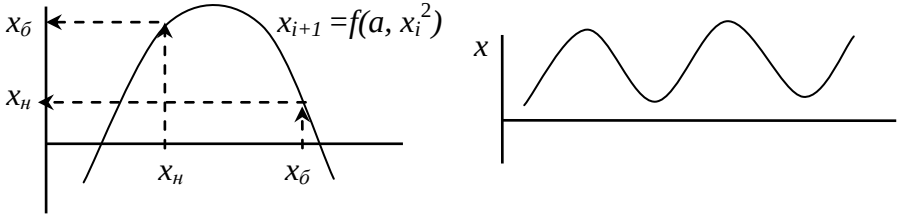


Рис. 7.4.2. Смена значений модели, отраженных от ветвей параболы

Конкретный ход кривой будет зависеть от того, какие были взяты начальные значения (x_0) и каковы значения параметра a. Меняя начальное состояние системы (в диапазоне 0…1), можно получить все полное множество точек, лежащих на линии параболы.

Сложность предстоящего рассмотрения состоит в том, что необходимо выразить зависимость переменной самой от себя, но в предыдущие моменты времени, а именно как зависимость x_i от x_{i-1} и как зависимость x_{i+1} от x_i.

Для понимания хода динамики переменной x в разные моменты времени построим двойной график изучаемой функции:

- зависимость оси ординат от оси абсцисс, $x_i = ax_{i-1} - ax_{i-1}^2$ и
- зависимость оси абсцисс от оси ординат $x_{i+1} = ax_i - ax_i^2$.

Тогда можно на одном графике отобразить каждую пару «шагов» модели: сначала по какому-либо значению x_{i-1} (точка *А*) с помощью нижнего графика найдем значение x_i (точка *Б*), затем по этому значению x_i по графику сбоку слева найти следующее значение x_{i+1} (точка *В*) и т. д. по всему ряду; результатом будет *схождение функции к нулю* (рис. 7.4.3).

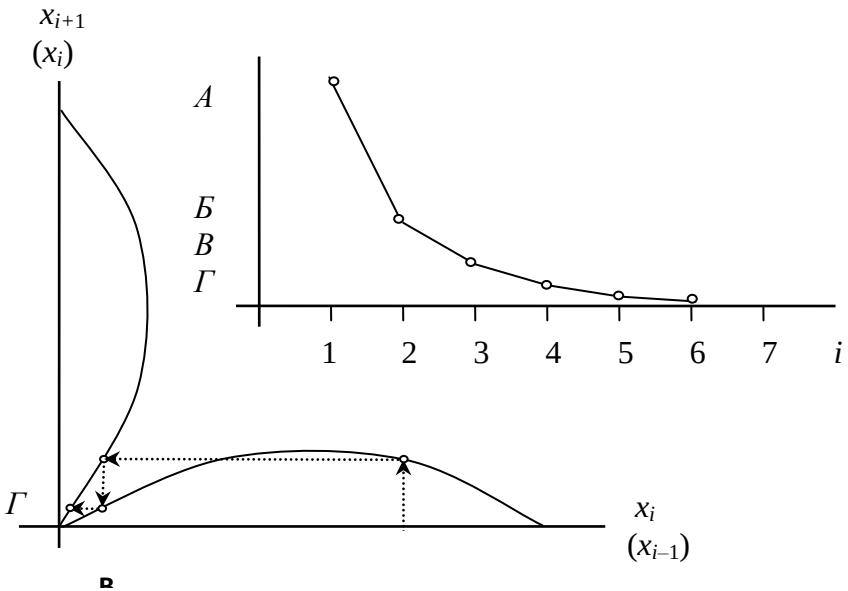


Рис. 7.4.3. Смена значений модели, отраженных от ветвей параболы

Существует более емкий способ (диаграмма Ламерея) для отображения зависимости текущего уровня переменной от предыдущего (x_{i+1} от x_i) – подстановка своеобразного «зеркала», биссектрисы, которая символизирует переход очередного значения параметра с оси ординат на ось абсцисс, «отображение отрезка в себя» (рис. 7.4.4).

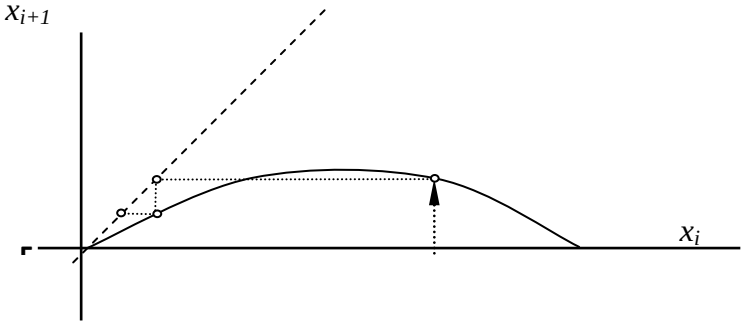
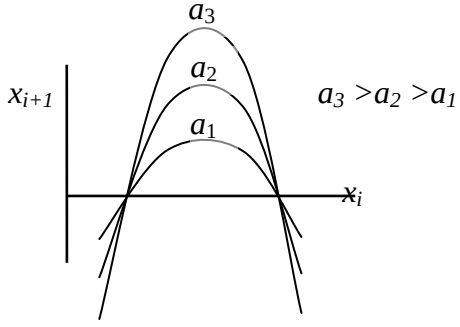


Рис. 7.4.4. Смена значений модели, отраженных от биссектрисы ($0 < a < 1$)

В зависимости от величины параметра a парабола будет менять свою «выгнутость» (рис. 7.4.5), вслед за ней будет меняться и ход кривой динамической переменной x. Как было показано выше, при значениях параметра меньше единицы $a < 1$ парабола будет пологой, а функция будет сходиться к нулю (рис. 7.4.3).

Рис. 7.4.5. Изменение формы параболы в зависимости от параметра a в модели

$$x_{i+1} = ax_i - ax_i^2 = ax_i(1 - x_i)$$



Другая ситуация наблюдается, когда параметр принимает значение в диапазоне $1 < a < 2$. Парабола оказывается более выпуклой и пересекается биссектрисой. В этом случае каждое следующее значение переменной x становится все больше, пока не достигнет

некоторого предельного значения (места пересечения биссектрисы с параболой, $x = 0.5$).

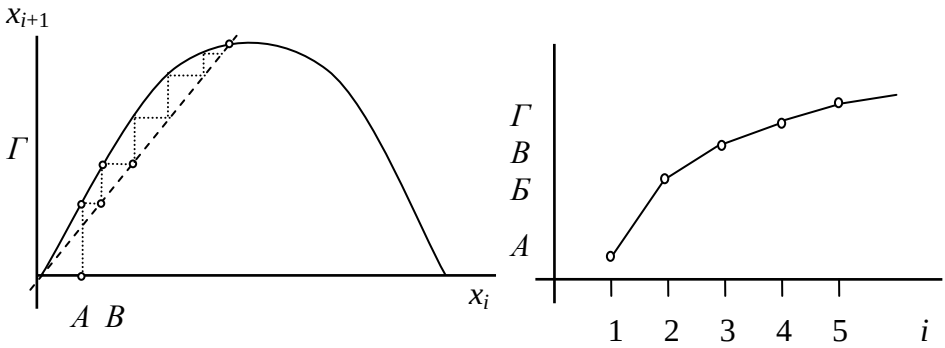


Рис. 7.4.6. Смена значений модели, отраженных от биссектрисы ($1 < a < 2$)

Наиболее интересна ситуация, когда изучаемая переменная x_i попеременно принимает то высокие, то низкие значения. Такое поведение характерно для нее при определенной величине параметра, например, при $a = 1+\sqrt{5}$. В этом случае устойчивыми оказываются всего два значения $x_1 = 0.5$ и $x_2 = 0.809$ («сверхустойчивый цикл»). Эту ситуацию можно исследовать на листе Excel (рис. 7.4.7).

B3 =E$1*B2-E$1*(B2^2)

	A	B	C	D	E	F	G	H
1	i	x(i)	x(i+1)	a=	3.23606798			
2	0	0.5	0.809					
3	1	0.809	0.5					
4	2	0.5	0.809					
5	3	0.809	0.5					
6	4	0.5	0.809					
7	5	0.809	0.5					
8	6	0.5	0.809					
9	7	0.809	0.5					
10	8	0.5	0.809					
11	9	0.809	0.5					
12	10	0.5	0.809					
13	11	0.809	0.5					
14	12	0.5	0.809					
15	13	0.809	0.5					
16	14	0.5	0.809					
17	15	0.809	0.5					

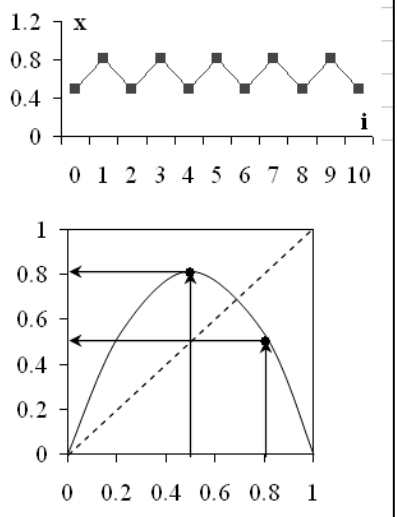
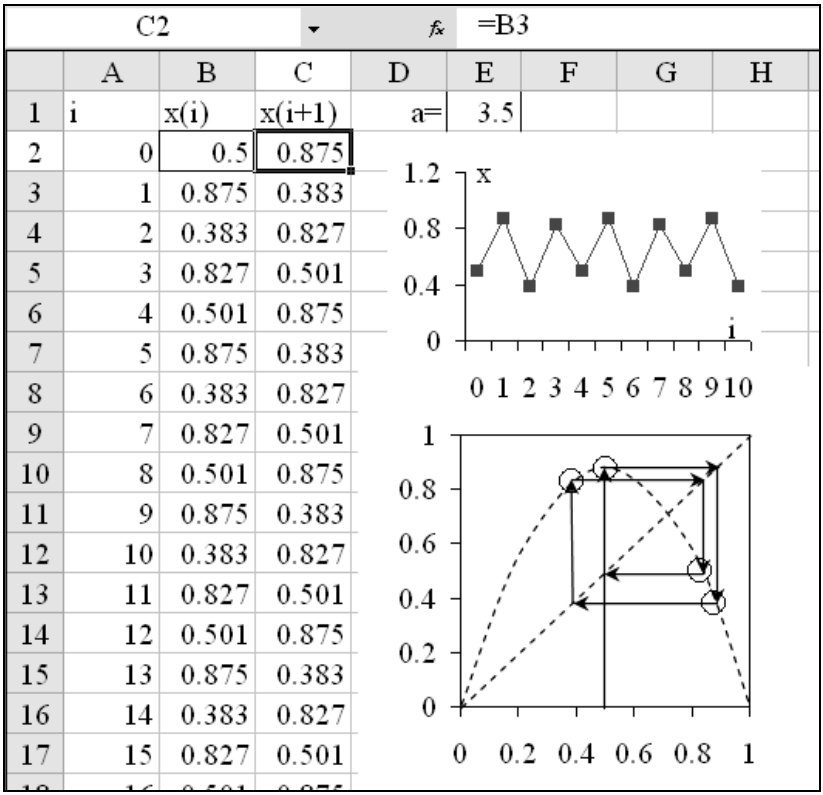


Рис. 7.4.7. Устойчивый цикл при $a = 1+\sqrt{5}$

При других значениях *a* динамика модели существенно усложняется – возникает несколько чередующихся циклов с разной амплитудой (рис. 7.4.8).



	A	B	C	D	E	F	G	H
1	i	x(i)	x(i+1)	a=	3.5			
2	0	0.5	0.875					
3	1	0.875	0.383					
4	2	0.383	0.827					
5	3	0.827	0.501					
6	4	0.501	0.875					
7	5	0.875	0.383					
8	6	0.383	0.827					
9	7	0.827	0.501					
10	8	0.501	0.875					
11	9	0.875	0.383					
12	10	0.383	0.827					
13	11	0.827	0.501					
14	12	0.501	0.875					
15	13	0.875	0.383					
16	14	0.383	0.827					
17	15	0.827	0.501					

Рис. 7.4.8. Устойчивый двойной цикл при *a* = 3.5

С ростом модельного параметра *a* поведение системы все более усложняется, пока оно не становится совершенно хаотическим.

Для нас здесь важно то, что модель с очень простым строением имеет сложное поведение, в том числе демонстрирует регулярную циклику. Не менее важно и то, что эта модель не соответствует ни одному из проявлений динамики численности реальных популяций. В этом нетрудно убедиться, если построить график зависимо-

сти численности животных рыжей полевки в текущем году (N_i) в зависимости от численности в предыдущем году (N_{i-1}) (рис. 7.4.9).

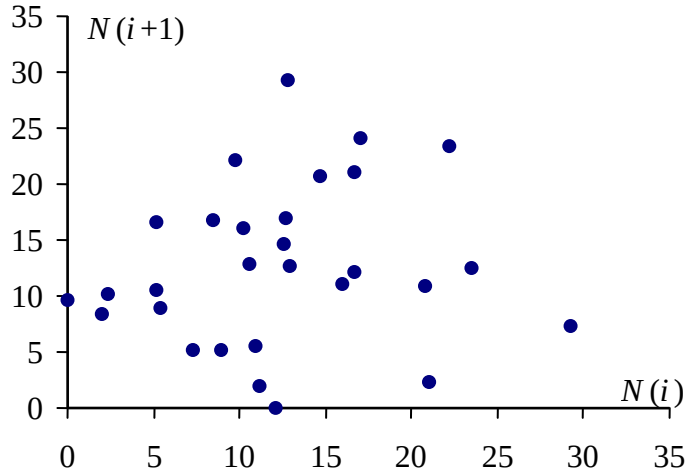


Рис. 7.4.9. Соотношение осенней численности рыжей полевки в заповеднике «Кивач» (данные А. П. Кутенкова)

Диаграмма совершенно не напоминает параболу или какую-либо другую простую линию, соответствующую конструкции рассмотренной модели. Она подобна облаку стохастического шума. Значит ли это, что в динамике численности полевок нет периодической закономерности? Или просто наш способ модельного представления не адекватен действительности? Этот вопрос относится уже к компетенции биологии и должен быть переформулирован в ее терминах, а именно: каковы могут быть общие факторы динамики численности и как их следует соотносить с нашей моделью?

Хорошо известно подразделение этих факторов на внутренние и внешние по отношению к популяции. Изменение рассмотренной выше моделируемой системы целиком и полностью определялось заданным уравнением ($x_i = ax_{i-1} - ax_{i-1}^2$); этот момент может быть понят как моделирование динамики популяции под действием только внутрипопуляционных факторов. В природе же такого положения вещей почти никогда не встречается, внешние условия всегда накладывают мощный отпечаток на жизнь популяции. Обычно говорят о двух группах воздействий – о периодических (сезонных) и о

непериодических (часто катастрофических). Модель должна учитывать эти причины.

Рассмотрим один из интересных и перспективных способов исследования регулярно действующих факторов – моделирование с помощью *функций последования*. Что означает смена сезонов в умеренной полосе для популяции животных, например, грызунов? – это совершенно разные *режимы функционирования*. В холодный период (осень–зима–весна) особи напряженно борются с суровыми условиями и гибнут в большом количестве. В теплый период (весна–лето–осень) происходит их активное размножение. Не рассматривая пока более дробное подразделение состояний, мы выделяем два сезона для популяции грызунов: сезон *вымирания* («зима») и сезон *воспроизводства* («лето»). К концу каждого этапа популяция приходит с тем или иным «достижением» – определенным уровнем численности ($N_{весной}$ и $N_{осенью}$), который выступает в роли текущей характеристики состояния системы (популяции).

В отношении моделирования это рассуждение приводит к важному выводу: при описании многолетней динамики численности невозможно обойтись одной функцией, нужны, как минимум, две функции, которые бы описывали:

– зависимость численности популяции весной от численности популяции осенью предыдущего года (функция зимней **г**ибели), $N_{весной} = f_г(N_{осенью}, a)$,
– зависимость численности популяции осенью от численности популяции весной текущего года (функция летнего **в**оспроизводства) $N_{осенью} = f_в(N_{весной}, b)$.

Каждая функция по отдельности призвана выразить способность популяции реагировать на специфические условия разных сезонов. Иными словами, две модели описывают проявления действия внутрипопуляционных факторов в двух режимах действия внешних факторов. Для определения хода временно́го ряда динамики численности были предложены (Саранча, 1995) особые формы этих двух функций (рис. 7.4.10).

Используем эти графики для построения кривой многосезонной динамики численности. Зная весеннюю численность года (абсцисса N_{i-1}, точка *А*), по функции $f_в$ находим, какой будет осенняя численность текущего года (ордината N_i, точка *Б*).

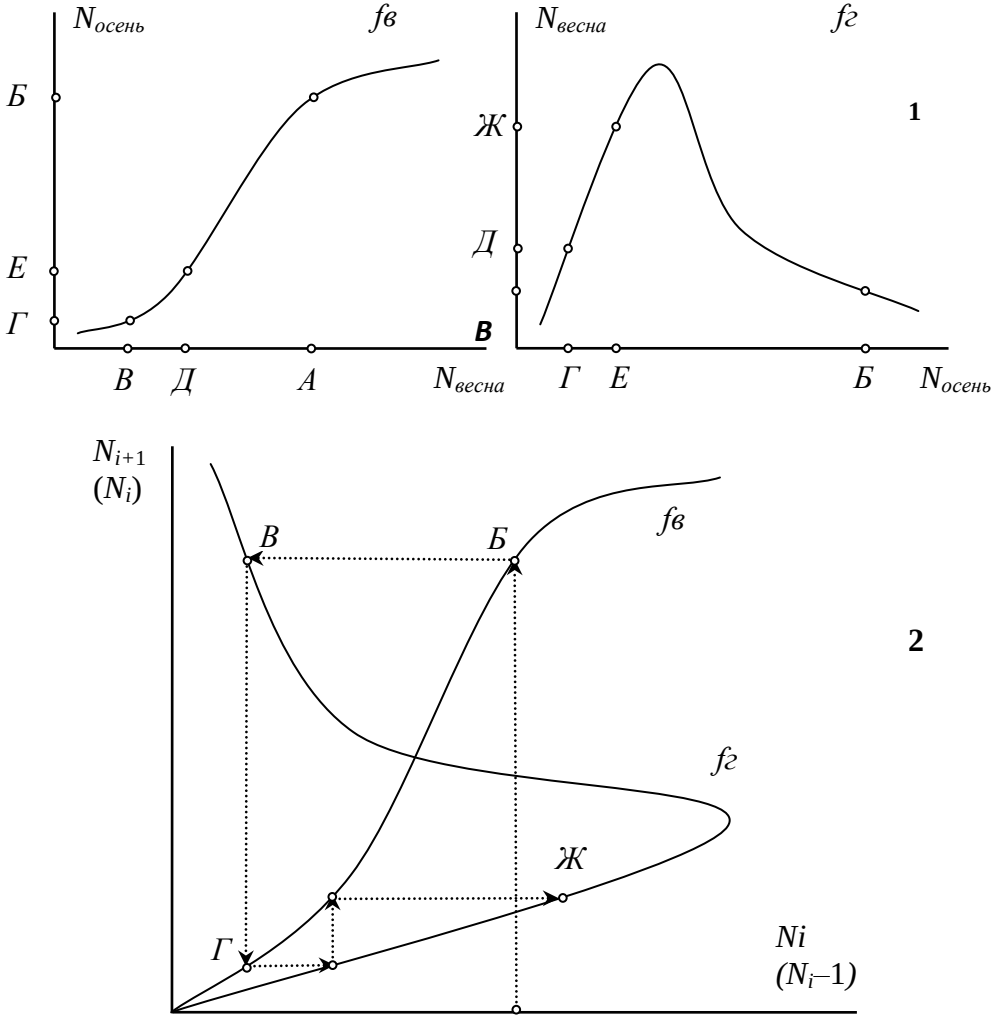


Рис. 7.4.10. Функции последования для характеристики зимней гибели и летнего воспроизводства грызунов, представленные по отдельности (**1**) и совместно (**2**)

Далее, используя то же значение весенней численности (ординату N_i), уже по кривой $f_г$ находим, какой должна быть весенняя численность в следующем году (абсцисса N_{i+1}, точка B), и т. д.

В этой процедуре нет ничего нового, поскольку точно таким же способом мы получали временные ряды для простой функции

(рис. 7.4.3). Отличие состоит только в том, что вместо одной и той же функции используется две разных.

Множество значений, определяемых последовательно попеременно то по функции вымирания, то по функции воспроизводства, производит временной ряд динамики численности (рис. 7.4.11).

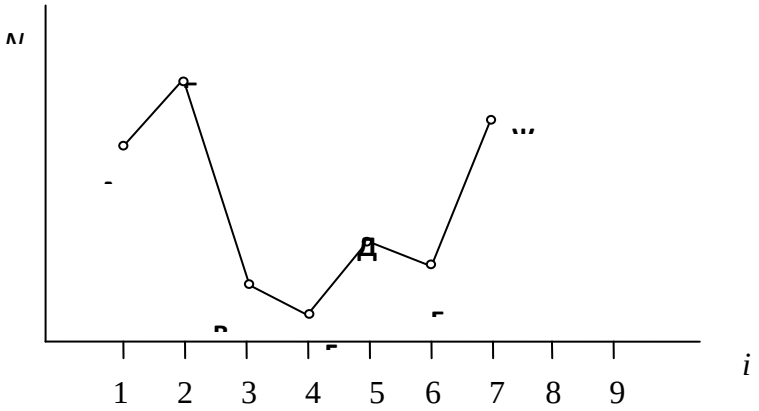


Рис. 7.4.11. Динамика численности, предписанная двумя функциями последования (по рис. 7.4.9)

Каждая сезонная кривая (модель $f_в$ и $f_с$) описывает внутренние характеристики популяции, реализацию внутрипопуляционных факторов *при данных специфических внешних условиях*. Это модели описания разных *режимов функционирования популяции* в контексте той или иной природной обстановки. Разделив жизнь популяции в течение года на два отрезка, в пределах каждого из них удается довольно адекватно описать проявления внутрипопуляционных факторов. Возможно, этот путь (деление длительного временно́го отрезка на краткосрочные) позволит найти оптимальное соотношение сложности и точности получаемых описаний.

Перейдем к рассмотрению модели динамики конкретной популяции рыжей полевки в Карелии (биостанция КарНЦ РАН «Гомсельга») (Коросов, Зорина, 2007). По эмпирическим материалам были выполнены следующие действия:

– построен временной ряда изменения численности животных по сезонам (N_i) (рис. 7.4.12),
– построены графики зависимости N_i от N_{i-1} и N_{i+1} от N_i (зависимости последования) (рис. 7.4.13),

– рассчитаны коэффициенты функций последования $f_г$ и $f_в$ с помощью регрессионного анализа,
– построена имитационная модель динамики численности на основе модельных функций последования (рис. 7.4.14),
– определены окончательные параметры функций последования, ориентируясь на отличия динамики модельных и реальных значений численности с помощью процедуры оптимизации Поиск решения среды Excel (рис. 7.4.15).

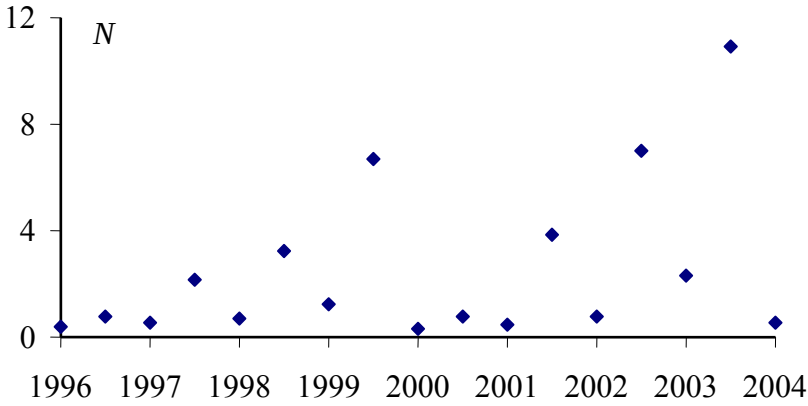


Рис. 7.4.12. Оценки численности популяции рыжей полевки

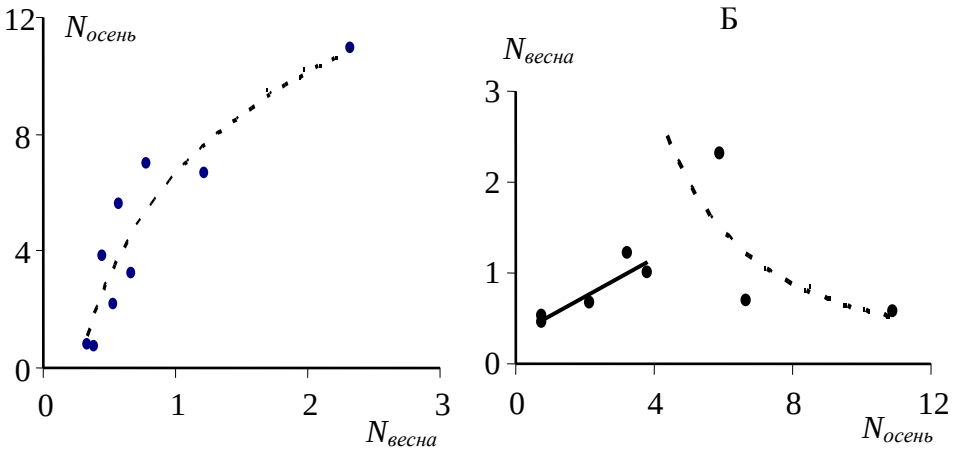


Рис. 7.4.13. Функция летнего воспроизводства (А) и функция зимней гибели, представленная двумя уравнениями (Б) популяции рыжей полевки, построенные по данным рис. 7.4.12

Базовые уравнения функций последования находили с помощью регрессионного анализа. Для функции воспроизводства $f_в$ (зависимость осенней численности от весеннего уровня) хорошо подходит логарифмическое уравнение $N_{осень} = 5.04 \cdot \ln(N_{весна}) + 6.61$, $R^2 = 0.85$. Для определения функции зимнего вымирания $f_з$ используются два уравнения, линейное и степенное, сходящихся в «переломной» величине осенней численности. Для тех лет, когда осенняя численность была ниже критического значения $N_{осень\,(кр)} = 4$, наблюдалась прямая пропорция: $N_{весна} = 0.16 \cdot N_{осень} + 0.38$, $R^2 = 0.57$.

В тех случаях, когда численность была выше «переломного» значения $N_{осень\,(кр)} = 4$, зависимость была обратно пропорциональной и выражалась степенным уравнением $N_{весна} = 3.87 \cdot N_{осень}^{-0.77}$, $R^2 = 0.07$. Найденные на основе регрессионного анализа функции последования $f_з$ и $f_в$ используются при реконструкции сезонного и многолетнего изменения численности с помощью динамической имитационной модели, организованной в нескольких столбцах Excel (рис. 7.4.14).

D5 fx =ЕСЛИ(E4=0;0;(ЕСЛИ(E4>$D$1;G5;F5)))

	A	B	C	D	E	F	G	H	I
1	$R^2_{мод}$	$R^2_{весна}$	$R^2_{осень}$	4	5,04	0,16	3,87		*Ф общ*
2	*0,49*	*0,02*	*0,13*		6,61	0,38	-0,77		*93,59*
3	*год*	*N весна*	*N осень*	*N' весна*	*N' осень*	*N' весна (линия)*	*N' весна (гипербола)*	*Ф весна*	*Ф осень*
4	1996	0,38	0,75	0,38	1,79			0,00	1,09
5	1997	0,53	2,15	**0,67**	4,57	**0,67**	2,47	0,02	5,84
6	1998	0,67	3,22	**1,20**	7,53	1,11	**1,20**	0,29	18,59
7	1999	1,22	6,67	**0,82**	5,59	1,59	**0,82**	0,16	1,15
8	2000	0,33	0,78	**1,03**	6,75	1,27	**1,03**	0,48	35,66
9	2001	0,44	3,83	**0,89**	6,02	1,46	**0,89**	0,20	4,81
10	2002	0,78	6,99	**0,97**	6,46	1,34	**0,97**	0,04	0,28
11	2003	2,32	10,91	**0,92**	6,19	1,41	**0,92**	1,96	22,33
12	2004	0,57	5,61	**0,95**	6,36	1,37	**0,95**	0,14	0,56

Рис. 7.4.14. Построение имитационной модели динамики численности полевок на основе регрессионных функций последования

Для вычисления очередного значения численности попеременно используем то одну, то другую функцию, подставляя в формулы значения численности, рассчитанные на предыдущем шаге. За начальное значение (ячейка D4) берем численность полевок весной 1996 г. $N_{весна1996} = 0.38$. Осенью этого же года численность составит (E4): $N_{осень1996} = 5.04 \cdot \ln(0.38) + 6.61 = 1.79$. Поскольку осенняя численность оказалась меньше переломного значения $N_{осень1996} < 4$, для вычислений выбирается линейное уравнение; весной следующего года численность равна (F5): $N_{весна1997} = 0.16 \cdot 1.79 + 0.38 = 0.67$ (для прочих лет было выбрано уравнение гиперболы G6:G12).

Дисперсионный анализ показал (расчеты не приводятся), что полный ряд модельных значений (n = 17) в целом адекватен исходным данным (R^2 = 0.49, $p < 0.05$), так как улавливает сезонные перепады численности. Если же сопоставлять модельные и реальные оценки численности только для весны (n = 9) или для осени (n = 8), соответствующие коэффициенты детерминации оказываются очень низкими R^2 = 0.02, R^2 = 0.13 $(p > 0.05)$, то есть по отдельности модели не адекватны исходным данным.

Для сближения модельных значений (N') с реальной динамикой численности (N) необходимо таким образом переопределить коэффициенты уравнений (модельные параметры), чтобы суммарные отличия между модельными и реальными значениями (значения невязки) обнулились (расчеты в блоке H4:I12):

$$\Phi = \Phi_{весна} + \Phi_{осень} = \sum (N'_{весна} - N_{весна})^2 + \sum (N'_{осень} - N_{осень})^2 \to 0.$$

Эту задачу решаем с помощью макроса Поиск решения: Установить целевую ячейку $I$2, Равной значению 0, Изменяя ячейки $E$1:$G$2 (с модельными параметрами). После настройки полного обнуления функции невязки не произошло ($\Phi = 52$), но полученная модель лучше соответствует исходным данным (R^2 = 0.69); коэффициенты детерминации между прямыми и расчетными оценками численности существенно возросли: для весны до уровня R^2 = 0.74, для осени – до R^2 = 0.54. Теперь модель демонстрирует четырехлетний периодизм подъемов численности, близкий к исходным данным (рис. 7.4.16).

I2 =СУММ(H4:I12)

	A	B	C	D	E	F	G	H	I
1	$R^2_{мод}$	$R^2_{весна}$	$R^2_{осень}$	4	5,08	0,45	2,11		Ф общ
2	0,69	0,74	0,54		5,51	0,51	-0,78		51,85
3	год	N весна	N осень	N' весна	N' осень	N' весна (линия)	N' весна (гипербола)	Ф весна	Ф осень
4	1996	0,38	0,75	0,38	0,65			0,00	0,01
5	1997	0,53	2,15	**0,80**	4,39	**0,80**	2,94	0,07	4,99
6	1998	0,67	3,22	**0,67**	3,44	2,48	**0,67**	0,00	0,05
7	1999	1,22	6,67	**2,05**	9,16	**2,05**	0,80	0,69	6,22
8	2000	0,33	0,78	**0,37**	0,52	4,62	**0,37**	0,00	0,06
9	2001	0,44	3,83	**0,74**	4,00	**0,74**	3,50	0,09	0,03
10	2002	0,78	6,99	**0,72**	3,81	2,30	**0,72**	0,00	10,13
11	2003	2,32	10,91	**2,22**	9,55	**2,22**	0,74	0,01	1,85
12	2004	0,57	5,61	**0,36**	0,36	4,80	**0,36**	0,04	27,60
13	2005	1,10		**0,67**		0,67	**0,67**		

Рис. 7.4.15. Имитационная модель динамики численности полевок после настройки параметров

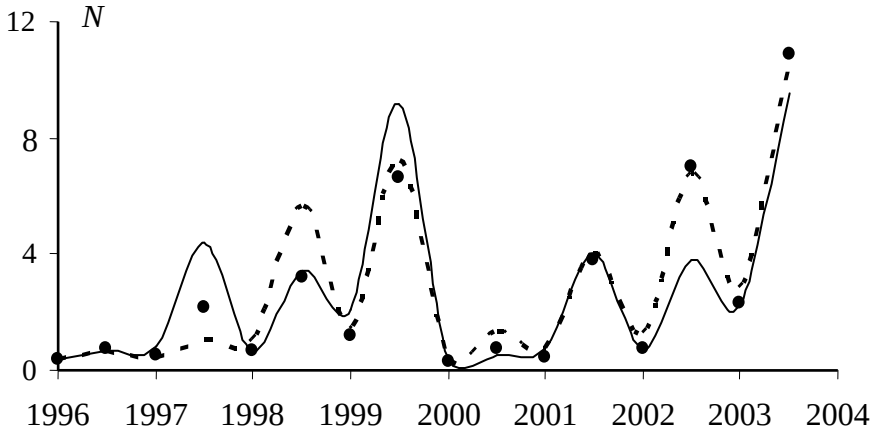


Рис. 7.4.16. Динамики численности рыжей полевки: наблюдения (точки), модель без учета внешних факторов (линия, рис. 7.4.15); модель с учетом влияния температуры (пунктир, рис. 7.4.17)

Как представляется, с помощью рассмотренной процедуры можно строить модели, которые будут характеризовать некую видовую или популяционную «норму динамики численности» в контексте *периодических* сезонных изменений внешних факторов. Однако

реакции популяции на внешние *непериодические* факторы остаются неучтенными. В то же время они очень важны, поскольку нелинейные системы могут быть очень чувствительными к небольшим сдвигам характеристик системы (начальным значениям), что для популяций животных выражается в сбоях «нормальной» динамики численности, так что ее модельный прогноз не будет совпадать с реальным ходом. Как можно учесть и изучить спонтанные воздействия внешних причин и их значимость для популяции полевок?

Один из путей апробируется регрессионным анализом и состоит в том, что значения внешних факторов включаются в модель в качестве независимых переменных, определяющих некоторый «дефицит» численности за счет негативных внешних воздействий:

$$N_i = N_i - dN_i,$$

$$dN_i = f\,(C, Z),$$

где Z – уровни неблагоприятных внешних факторов, C – коэффициенты пропорциональности.

Рассматриваемая нами популяция рыжих полевок обитает на северной периферии ареала вида, где значительную роль в динамике численности играют метеорологические условия в переходные (весенний и осенний) периоды (Ивантер, 1975). Располагая данными по сумме положительных температур ($t_{сум}$, блок B3:B11) за период наблюдений, в модель ввели поправочный компонент – уравнение параболической зависимости «дефицита численности весной» $dN_{весна}$ от переменной $t_{сум.}$с тремя коэффициентами (рис. 7.4.17, I14:J16).

В результате настройки новой модели значение функции невязки снизилось; существенно выросли и коэффициенты детерминации: адекватность весенних и осенних оценок составила $R^2 = 0.95$ и $R^2 = 0.92$, модели в целом – $R^2 = 0.95$ (рис. 7.4.17).

Параболическая зависимость весенней численности от $t_{сум}$ может получить осмысленную интерпретацию. Средние значения суммы температур соответствуют обычному для Карелии и губительному для полевок ходу весны с возвратами холодов, дождями и заморозками. Теплая весна способствует их выживанию; высокая численность зверьков сохраняется и при минимальной сумме положительных температур вследствие позднего разрушения защитного снегового покрова.

G14 =ЕСЛИ(G5>0;G5+($J$14*B5^2-$J$15*B5+$J$16);0)

	A	B	C	D	E	F	G	H	I	J
1	$R^2_{мод}$		$R^2_{вес}$	$R^2_{осень}$	4	5,08	0,45	2,11		*Ф общ*
2	*0,95*		*0,95*	*0,92*		5,51	0,51	-0,78		*8,83*
3	*год*	*t сум*	*N весн*	*N осень*	*N' весна*	*N' осен*	*N' весна (линия)*	*N' весна (гипербола)*	*Ф весн*	*Ф осень*
4	1996	3,88	0,38	0,75	0,38	0,65			0,00	0,01
5	1997	4,13	0,53	2,15	**0,42**	1,04	0,80	2,94	0,01	1,23
6	1998	5,15	0,67	3,22	**1,03**	5,67	0,98	2,04	0,13	5,99
7	1999	2,72	1,22	6,67	**1,41**	7,27	3,05	0,55	0,04	0,36
8	2000	5,08	0,33	0,78	**0,44**	1,37	3,77	0,45	0,01	0,35
9	2001	4,40	0,44	3,83	**0,74**	4,00	1,12	1,65	0,09	0,03
10	2002	5,57	0,78	6,99	**1,29**	6,82	2,30	0,72	0,27	0,03
11	2003	6,49	2,32	10,91	**2,84**	10,81	3,57	0,47	0,27	0,01
12							*N'' весна (лин-dN_i)*	*N'' весна (гип-dN_i)*		
13										
14							**0,42**	2,55	c_1=	0,55
15							**1,03**	2,10	c_2=	4,62
16							3,92	**1,41**	c_3=	9,41
17							3,76	**0,44**		
18							**0,74**	1,27		
19							2,88	**1,29**		
20							5,93	**2,84**		

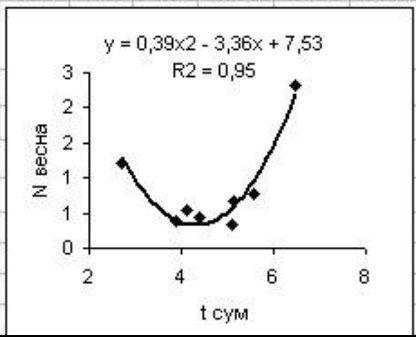


Рис. 7.4.17. Имитационная модель динамики численности полевок с учетом влияния хода весенних событий

В силу небольшой длины ряда данных, этот вопрос требует дополнительной проработки, в то же время представленный способ исследования внешних и внутренних факторов и механизмов популяционной цикличности открывает дорогу для описания специфики динамики численности разных видов, одного вида в разных местах обитания (с разным антропогенным прессом), в разных биотопах. При этом функции последования интерпретируются как характеристики *популяционных потенций* в любых условиях существования.

Опираясь на эту идеологию, можно давать прогнозы численности. Внешние факторы делятся на два класса – обычные (составляющие большинство наблюдений) и редкие экстремальные влияния. Функции последования описывают реакцию популяции на обычные уровни факторов, поэтому с их помощью прогноз может

быть дан только для «нормального» хода фенологических процессов, причем – интервальный. Построим диаграмму: по оси абсцисс отложим значения численности популяции в дискретные моменты времени, по оси ординат – частоты их встречаемости. Стохастическая реализация *циклической* динамики систем выглядит на диаграмме как серия островков с повышенной частотой наблюдений. При строгой регулярности процесса получим несколько отдельных ровных столбиков (рис. 7.4.18, А). При полном хаосе частоты всех численностей будут равны и сформируют равномерное распределение. Для естественных систем более характерен «шумящий цикл», который на диаграмме выглядит как серия отдельных трансгрессирующих выборок, переходящих друг в друга (рис. 7.4.18, Б).

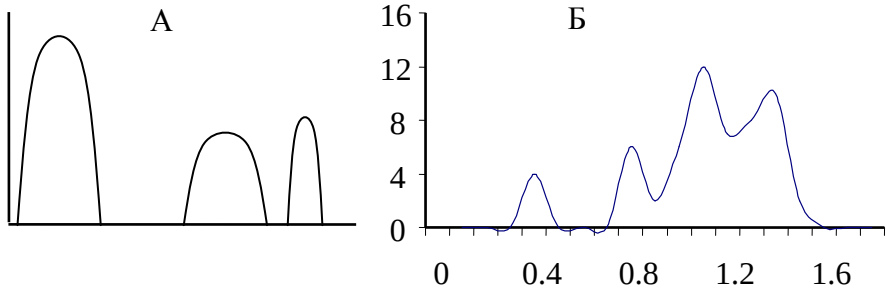


Рис. 7.4.18. Теоретическое распределение значений временно́го ряда с выраженными циклами (*А*) и распределение значений ряда логарифмов численности рыжей полевки длиной 25 лет (*Б*)

Здесь важно, будет ли сохраняться *порядок прохождения* популяцией отдельных «островков» гистограммы? Если да, то, зная этот порядок, прогноз сделать просто. Действие же экстремальных факторов состоит в разрушении последовательности прохождения популяций этих особенных уровней численности, «островков». Когда произойдет такое влияние, предсказать невозможно. Значит, прогноз *сбоя* популяционного ритма, прогноз катастрофы с этих позиций сделать нельзя.

Глава 8
ИЗУЧЕНИЕ МНОГОМЕРНЫХ ДАННЫХ

Любой набор многомерных биологических данных (как, например, морфологическое описание группы особей), как правило, структурирован, несет информацию о сходстве и различии объектов, о взаимозависимости их признаков. Только матрица случайных чисел однородна. Реальные же выборки гетерогенны. Вследствие действия внешних или внутренних факторов объекты агрегируются в кластеры, а признаки – в плеяды. Например, возраст как стадия онтогенеза выступает сильной причиной дифференциации представителей одного вида на крупных и мелких особей. Аналогично дело обстоит с возникновением корреляционных плеяд (см. п. 6.2). Изучение плеядно-кластерной организации многомерных эмпирических данных, по существу, означает поиск факторов, ответственных за ее формирования.

8.1. Анализ (метод) главных компонент

Компонентный анализ как один из многомерных методов является отражением двух основных тенденций развития современной биометрии. С одной стороны, это стремление к более полному (многоплановому, многомерному) изучению действительности, что требует количественной оценки большого числа свойств исследуемых объектов. С другой стороны, это формирование все более наглядного, интегрированного, обобщенного представления об огромных массивах информации; полученные данные «сворачиваются» до размеров, которые в состоянии охватить мысль (и перевести их в категории номинальной шкалы, см. п. 2.2).

Наблюдения за эколого-биологических феноменами дают выборки из n объектов (например, особей), охарактеризованных набором из m признаков (промеры, масса, температура, число морфологических элементов, концентрация веществ и пр., $m \sim 5–15$). Для таких данных предложен эффективный прием сокращения размерности – введение новых показателей, рассчитанных на основе реальных значений, – линейных индексов (Животовский, 1984):

$$C_j = a_{1j} \cdot x_1 + a_{2j} \cdot x_2 + ... + b_{ij} \cdot x_i + ... + b_{mj} \cdot x_m .$$

Такие расчетные величины оказываются информативнее исходных признаков, поскольку каждое значение индекса характеризует объект с учетом нескольких значений реальных признаков. Вследствие обобщения информации число новых индексов (k) обычно невелико, $k < m$; в рамках компонентного анализа они называются *главными компонентами*.

Геометрическая интерпретация

Смысл главных компонент как неких новых расчетных показателей качества объектов проще раскрывать на примере двумерного распределения. Положим, что некие два биологических признака (x_1 и x_2) коррелируют: увеличение одного в целом связано с увеличением другого и наоборот. Область рассеивания множества вариант (точек x_{1j} и x_{2j}) вытянута в эллипс (рис. 8.1.1). При этом оказывается, что каждый признак несколько односторонне оценивает отличия объектов (вариант), измеряет не полную длину эллипса рассеивания, а какую-то ее боковую проекцию (Lx_1). Более полную характеристику распределения вариант можно получить, если измерять расстояния между объектами в направлении их максимальной изменчивости, то есть вдоль главной оси эллипса. Новая ось z_1 будет полнее, чем исходные признаки, измерять существенные отличия между объектами. Ось может быть определена как поворот исходной оси x_1 на угол α (образованный осью x_1 и главной диагональю).

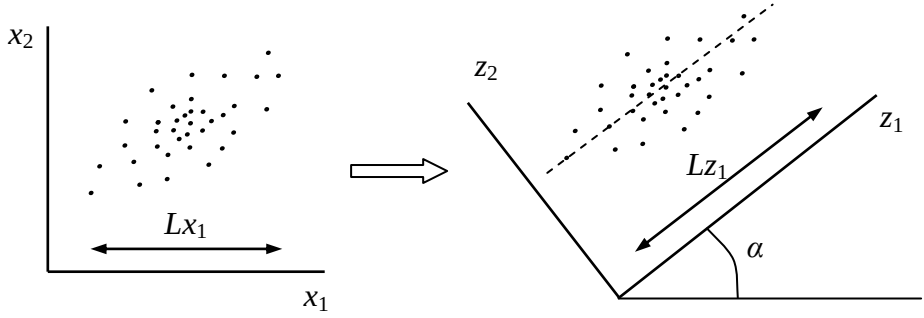


Рис. 8.1.1. Переход от исходных признаков к главным компонентам

Рассмотрим смысл процедуры на примере одной точки j, имеющей координаты x_{1j} и x_{2j}. Заметим, что значение x_{1j} есть расстояние, на которое точка j удалена от начала координат в направлении оси признака x_1. Выражение «повернуть ось x_1» для точки j

означает: вычислить значение z_{1j}, показывающее, на каком расстоянии от начала координат находится точка j *в направлении главной оси* эллипса рассеивания (рис. 8.1.2). Вычисления выполняются с помощью тригонометрических функций. Зная катет ($b = x_{1j}$) и острый угол прямоугольного треугольника (α), требуется найти гипотенузу (c).

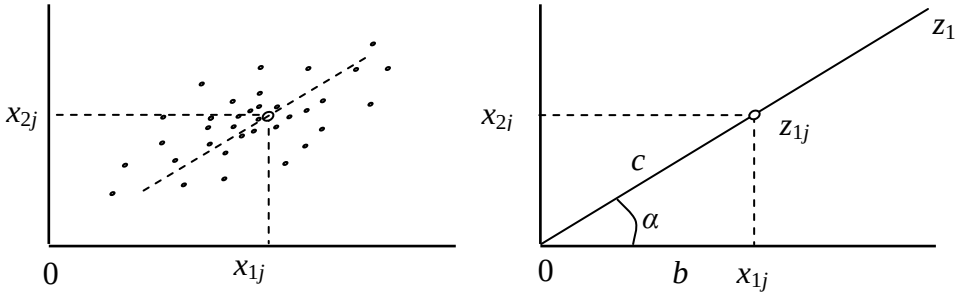


Рис. 8.1.2. Поиск новой координаты для точки j

Поскольку синус угла равен отношению прилежащего катета к гипотенузе ($\sin\alpha = b/c$), гипотенуза равна отношению катета к синусу $c = b/\sin\alpha = \dfrac{1}{\sin\alpha} b$. Обозначив $p_1 = 1 / \sin\alpha$, в наших терминах получаем формулу «поворота» оси для точки j: $z_{1j} = p_1 \cdot x_{1j}$.

С точкой j нам «повезло» – она лежит на главной оси эллипса рассеивания, проходящей через начало координат. Другие точки (i) отстоят от оси (рис. 8.1.3) и для расчета их координат z_{1i} требуется другое уравнение, включающее признак x_2 (Голикова и др., 1981): $z_{1i} = \sin\alpha \cdot x_{1i} + \cos\alpha \cdot x_{2i}$ или $z_{1i} = a_1 \cdot x_{1i} + a_2 \cdot x_{2i}$, где a_1 и a_2 – тригонометрические функции.

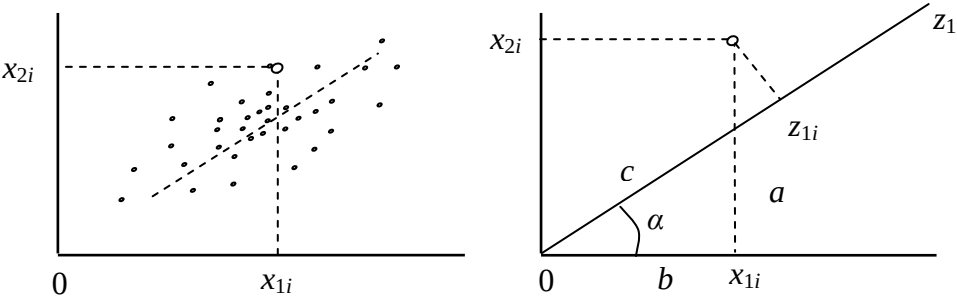


Рис. 8.1.3. Поиск новой координаты для точки i

Помимо координат точек на одной новой оси z_1 аналогичным образом можно вычислить координаты тех же точек на другой оси, z_2, перпендикулярной к первой: $z_{2i} = b_1 \cdot x_{1i} + b_2 \cdot x_{2i}$ (рис. 8.1.1, 8.1.4). Ось нужна для того, чтобы не утратилась информация об отличиях объектов в поперечном направлении эллипса рассеивания.

В результате таких расчетов получаем две новые оси, или главные компоненты, основные свойства которых явствуют из приведенной иллюстрации:

– значения компонент *нормированы*, средние равны нулю (начало координат находится в центе эллипса рассеивания);

– оси компонент перпендикулярны друг другу, то есть компоненты не коррелируют (свойство *ортогональности*);

– информативность компонент падает: первая учитывает максимальную изменчивость объектов (имеет *максимальную дисперсию* S_1^2), вторая учитывает следующее направление существенной изменчивости (имеет вторую по величине дисперсию S_2^2) и т. д.: $S_1^2 > S_2^2 > ... > S_m^2$.

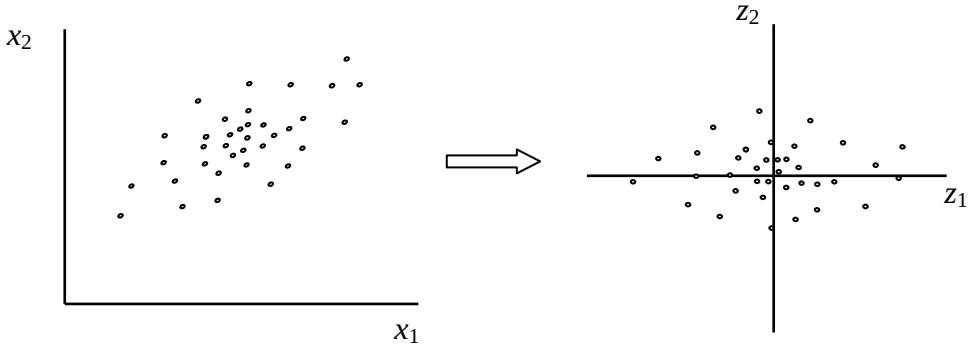


Рис. 8.1.4. Ординация объектов в осях главных компонент

Структурная интерпретация

Результатом эколого-биологического наблюдения выступает выборка объектов (например, особей), охарактеризованных набором из m признаков (промеры, масса, температура, число морфологических элементов, концентрация веществ и пр., m ~ 5–15). Каждый объект отличается от других по значениям этих показателей. При-

чины отличия могут быть случайными или оказаться результатом воздействия на объекты неких общих сильных факторов, таких как различие по возрасту, полу, состоянию здоровья, родство, абиотические условия среды, деятельность человека (факторы *a, b, c… g*). В этой интерпретации значение каждого признака у каждой особи формируется как бы из несколько частей, каждая из которых вызвана к жизни действием определенной причины:

$$x_i = x_{ia} + x_{ib} + ... + x_{ig} \ (i = 1, 2… m).$$

Компонентный анализ стремится сконцентрировать информацию об отличиях объектов в гораздо меньшем числе расчетных индексов ($k < m$), которые названы *главными компонентными* (другое название анализа – метод главных компонент, МГК, principal component analysis). Новые показатели собирают со всех исходных признаков и отображают в себе нечто общее, в силу чего группы признаков изменяются параллельно. Главные компоненты (*ГК*, или *PC, C*) выступают характеристиками причин варьирования сразу нескольких признаков. Отдельная компонента аккумулирует в себе только те доли значений исходных признаков, которые были связаны с воздействием определенного фактора изменчивости. У отдельного объекта значение главной компоненты есть как бы сумма долей (нормированных) значений разных признаков, связанных с воздействием одной общей причины:

$C_a = x_{1a} + x_{2a} + ... + x_{ia} + ... + x_{ma}$ – сумма вкладов фактора *a* в значения всех *m* признаков, где x_{ia} – вклад фактора *a* в значение варианты *i*-го признака данного объекта.

Для другого процесса (фактор *b*) имеем:

$C_b = x_{1b} + x_{2b} + ... + x_{ib} + ... + x_{mb}$ и т. д. для прочих факторов.

Эффект действия *j*-го фактора на признак можно выразить долей от общего значения варианты: $x_{ij} = a_i \cdot x_i$, где a_i – относительный вклад данного фактора в конечное значение варианты *i*-го признака, x_i – «полное» значение варианты признака *i*. Тогда уравнение первой главной компоненты примет вид:

$$C_a = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_i \cdot x_i + ... + a_m \cdot x_m .$$

Аналогично уравнение второй главной компоненты составит:

$C_b = b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_i \cdot x_i + ... + b_m \cdot x_m$ и т. д.

Здесь возникает вопрос, какое содержание имеют коэффициенты a_j, b_j … в уравнениях компонент, как их можно определить? Влияние одного фактора на все признаки может проявиться только в их синхронном изменении от объекта к объекту; так фактор «возраст» влияет на размеры, массу, зрелость и другие признаки особи. Зависимость между парой признаков принято выражать коэффициентом корреляции. Коэффициенты a_j, b_j…, стоящие перед значениями признаков (x_1, x_2…) в уравнениях компонент и есть (условные) коэффициенты корреляции данного признака с общим фактором; они называются *факторные нагрузки*. Большое значение факторной нагрузки в *j*-й компоненте будет иметь *i*-й признак, сильно зависящий от *j*-го фактора. Низкое значение нагрузки говорит о независимости данного признака от выявленного фактора.

Компонентный анализ позволяет рассчитать несколько главных компонент, каждая из которых связана со своим направлением изменчивости исходных данных (например, различия между особями по возрасту обычно отображает первая главная компонента, а различия по полу, состоянию здоровья – последующие).

Информативность каждого признака, которая содержится в различиях между объектами, выражается дисперсией S_i^2. Если признаки нормировать ($z = (x - M) / S$), то дисперсия отдельного признака будет равна единице $S_i^2 = 1$. Суммарную информативность выборки выражает объединение дисперсий в общую $S^2_{общ.x} = m$ (ее принимают за 100% информации об изменчивости).

Каждая компонента, как и исходные признаки, имеет характеристики своей информативности в виде дисперсии $S^2_{C\,j}$, а общая информативность всех компонент будет равна общей информативности всех исходных признаков и составит: $S^2_{общ.C} = S^2_{общ.x} = m$ (100%). Эффект «концентрирования» информации в начальных главных компонентах состоит в том, что их дисперсии много больше, чем дисперсии исходных признаков. Первые две-три компоненты накапливают в себе основную содержательную информацию об отличиях объектов: $S^2_{C\,1} \approx 50\%$, $S^2_{C\,2} \approx 30\%$, $S^2_{C\,3} \approx 10\%$. На долю прочих остается учет несущественной случайной изменчивости; об этом свидетельствует величина дисперсии ниже дисперсии отдельного признака $S^2_{C\,i} < 1$ до $S^2_{C\,m} \approx 0\%$. Обычно 2–3 компоненты заменяют собой 5–12 исходных признаков.

Смысл компонентного анализа состоит в интерпретации каждой компоненты, поиске и обозначении биологического явления, стоящего за параллельным изменением разных характеристик изучаемых объектов. На отдельный фактор (отображаемый обычно отдельной главной компонентой) реагируют не все, а некоторая часть исходных признаков – плеяды (см. п. 6.2). Какие именно из признаков реагируют и каким образом, показывает большая величина факторных нагрузок и их знак. Характер отличия объектов дает диаграмма их расположения в осях новых признаков (ординация объектов на плоскости главных компонент). Каждую главную компоненту называют, ориентируясь на взаимное расположение объектов и наборы учтенных признаков, стремясь тем самым обозначить выявленное направление биологического различия между объектами.

Математическая интерпретация

Алгоритм оперирует с прямоугольной матрицей данных **X**, состоящей из n строк (n записей, относящихся к отдельным объектам, например, особям, пробам, площадкам) и m столбцов (m полей, признаков, показателей, численно характеризующих каждый объект измерения), причем $n > m$. Эту матрицу предварительно нормируют $z = (x - M) / S$, формируется матрица **Z**. Рассчитывается матрица парных корреляций между всеми m признаками.

В терминах матричной алгебры процедура компонентного анализа есть поиск *собственных векторов* (**a**) и *собственных чисел* (**S**) матрицы *корреляций* (**R**), связанных соотношением: **Ra** = **Sa**.

$$\mathbf{R} = \begin{matrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{matrix}$$

Матрица **R** размерностью $m*m$ – это множество парных коэффициентов корреляции между изучаемыми признаками; диагональные элементы равны единице ($r_{ii} = 1$), матрица симметрична относительно главной диагонали ($r_{ij} = r_{ji}$).

$$\mathbf{a} = \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix}$$

Квадратная матрица **a** собственных векторов (или факторных нагрузок) имеет размерность $m*m$ и состоит из m векторов ($a_{ij} \neq a_{ji}$, $a_{ii} \neq 1$); произведение вектора на самого себя равно единице: $a_i \cdot a_i' = 1$.

$$\mathbf{S} = \begin{matrix} s_{11} & 0 & 0 \\ 0 & s_{22} & 0 \\ 0 & 0 & s_{33} \end{matrix}$$

Квадратная матрица **S** собственных значений имеет размерность $m*m$. Все ее элементы равны нулю $s_{ij} = 0$, кроме диагональных $s_{ii} > 0$, причем наибольшую величину имеет первое значение s_1, следующее по величине – второе s_2 и так далее до s_m: $s_1 > s_3 > s_3$. Значения s_i интерпретируются как дисперсии главных компонент.

Сами значения главных компонент получают перемножением матрицы нормированных данных **Z** на матрицу собственных векторов **a**: **C** = **Za**; для j-й варианты значение первой главной компоненты будет равно: $C_{1j} = a_{11} \cdot z_{1j} + a_{21} \cdot z_{2j} + a_{31} \cdot z_{3j}$. Если рассчитать дисперсию главной компоненты непосредственно по рассчитанным значениям C_1, мы получим величину, равную s_{11}.

R- и Q-техника компонентного анализа

Рассмотрим результаты компонентного анализа, выполненного для данных по размерам гадюки (W – масса тела, Lt – длина тела, Lc – длина хвоста) (табл. 8.1.1) в среде пакета StatGraphics.

Таблица 8.1.1. Промеры гадюк: исходные, нормированные и главные компоненты

	Исходные данные			Нормированные данные			Главные компоненты		
пол	W	Lt	Lc	*норм.* W	*норм.* Lt	*норм.* Lc	$ГК_1$	$ГК_2$	$ГК_3$
m_1	40	45	77	–1.40	–1.52	0.44	–2.027	–0.596	0.024
m_2	43	46	84	–1.25	–1.22	1.20	–2.108	0.220	0.082
m_3	45	47	81	–1.16	–0.93	0.87	–1.715	0.098	–0.095
m_4	48	45	76	–1.01	–1.52	0.34	–1.723	–0.614	0.295
…	…	…	…	…	…	…	…	…	…
f_{14}	82	50.5	64	0.67	0.10	–0.96	0.941	–0.648	0.249
f_{15}	90	53	68	1.06	0.84	–0.52	1.437	0.143	0.145
f_{16}	100	51	62	1.55	0.25	–1.17	1.701	–0.595	0.770
f_{17}	112	57	70	2.14	2.02	–0.31	2.746	1.088	0.226
M	68.5	50.1	72.9	0	0	0	1.08	0.84	0.29
S	20.3	3.39	9.29	1	1	1	1.45	0.84	0.44
S^2	412	11.5	86.3	1	1	1	2.09	0.71	0.19

Цель анализа состояла в том, чтобы оценить степень различия между самками (f, 9 экз.) и самцами (m, 8 экз.) обыкновенной гадюки по пропорциям тела. Вначале выполнили нормирование исходных значений признаков $норм.X = (X - M)/S$, необходимое для расчетов компонент. Например, для первого самца (40 г) получаем нормированную массу:

$$норм.W = (40–68.5)/20.3 = –1.40.$$

Далее рассчитываются стандартные коэффициенты парной корреляции Пирсона (табл. 8.1.2). Можно заметить, что корреляция дины тела с массой положительная ($r_{12} = r_{WLt} = 0.789$), а с хвостом – отрицательная ($r_{13} = r_{WLc} = –0.492$), то есть чем крупнее особь, тем короче у нее хвост.

Таблица 8.1.2. Матрица корреляций между промерами гадюк

	W	*Lt*	*Lc*
W	1.00	0.79	–0.49
Lt	0.79	1.00	–0.33
Lc	–0.49	–0.33	1.00

Обработка корреляционной таблицы с помощью итерационной процедуры поиска собственных векторов и собственных значений дает матрицу факторных нагрузок (табл. 8.1.3), которые показывают корреляционные отношения между признаками «в одной плоскости».

Таблица 8.1.3. Факторные нагрузки (собственные векторы) матрицы корреляции между промерами гадюк

	a_1	a_2	a_3
W	0.644	0.191	0.741
Lt	0.603	0.467	–0.655
Lc	–0.470	0.863	0.186
Дисперсия, S^2	2.10	0.71	0.19
Дисперсия, %	70	24	6

Например, для первой компоненты нагрузки у массы и длины тела велики и положительны (0.644 и 0.603) вследствие их тесной положительной корреляции. Коэффициент у длины хвоста (–0.47) указывает на отрицательную корреляцию данного признака с остальными; значения C_1 отражают диспропорцию «тело–хвост».

В общем случае значимыми считаются компоненты, дисперсии которых близки или превышают единицу $S^2 \geq 1$. В примере имеет смысл как первая, так и вторая компонента.

Для визуализации результатов можно воспользоваться *четырьмя видами иллюстраций*. Факторные нагрузки удобно представить столбчатой диаграммой (рис. 8.1.5), на которой четко видны альтернативные вклады в первую компоненту массы–длины тела и длины хвоста; вторая компонента в основном представляет собой «длину хвоста» ($a_{2\ Lc} = 0.86$).

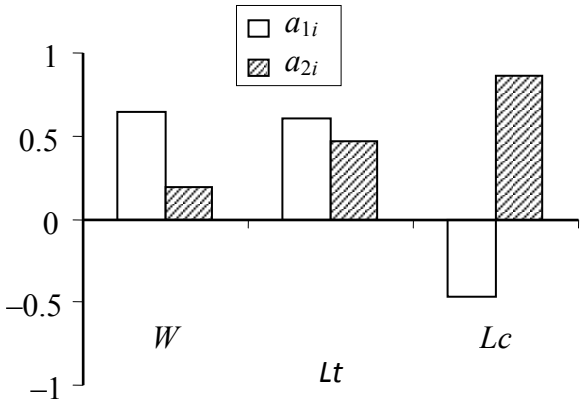


Рис. 8.1.5. Факторные нагрузки первой (a_{1i}) и второй (a_{2i}) компонент

Подставляем нагрузки в уравнение первой компоненты

$$C_1 = 0.644 \cdot норм.W + 0.603 \cdot норм.Lt - 0.470 \cdot норм.Lc$$

и находим значения компонент для конкретных особей (табл. 8.1.1). Для первого самца (m_1) и последней (f_{17}) самки имеем:

$$C_{1_1} = 0.644 \cdot (-1.44) + 0.603 \cdot (-1.52) + 0.470 \cdot (0.44) = -2.026,$$

$$C_{1_{17}} = 0.644 \cdot (2.14) + 0.603 \cdot (2.02) + 0.470 \cdot (-0.31) = 2.75.$$

Аналогично рассчитываем значения второй компоненты.

Разнонаправленная динамика признаков проявляется особенно ярко, если отсортировать объекты по величине первой компоненты и построить линейные диаграммы (рис. 8.1.6).

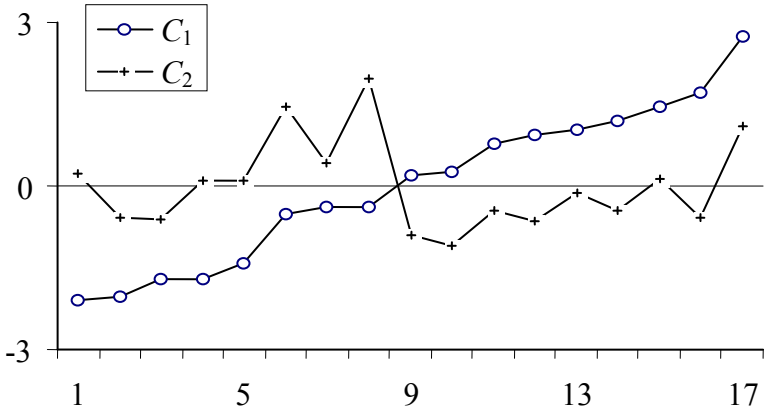


Рис. 8.1.6. Две первые главные компоненты для промеров гадюк

Значения первой компоненты, отражающей противопоставление общих размеров тела (большие положительные корреляции) длине хвоста (отрицательные корреляции) («рост размеров при уменьшении хвоста»), плавно возрастают. На ее фоне вторая компонента («длина хвоста») в середине ряда скачкообразно меняет свои значения, отмечая переход от длиннохвостых самцов к короткохвостым самкам.

Еще более рельефно половой диморфизм проявляется на точечной диаграмме рассеяния объектов в осях двух первых компонент (рис. 8.1.7). Особи разного пола разделились на две изолированные группы: справа внизу – самки, слева вверху – самцы.

Для интерпретации результатов счета полезно наложить на точки диаграммы рассеяния лучи факторных нагрузок (рис. 8.1.7, 8.1.8). Каждый из этих лучей идет от точки пересечения осей (0,0) до точки с координатами факторных нагрузок двух первых компонент (a_{1i},a_{2i}) (i – номер признака): (0.644, 0.191), (0.603, 0.467) и (–0.470, 0.863). Положение лучей имеет ясный теоретический смысл: угол между лучами пропорционален коэффициенту корреляции между соответствующими признаками. Например, узкий угол между векторами 0–*W* и 0–*Lt* пропорционален величине $r_{WLt} = 0.789$.

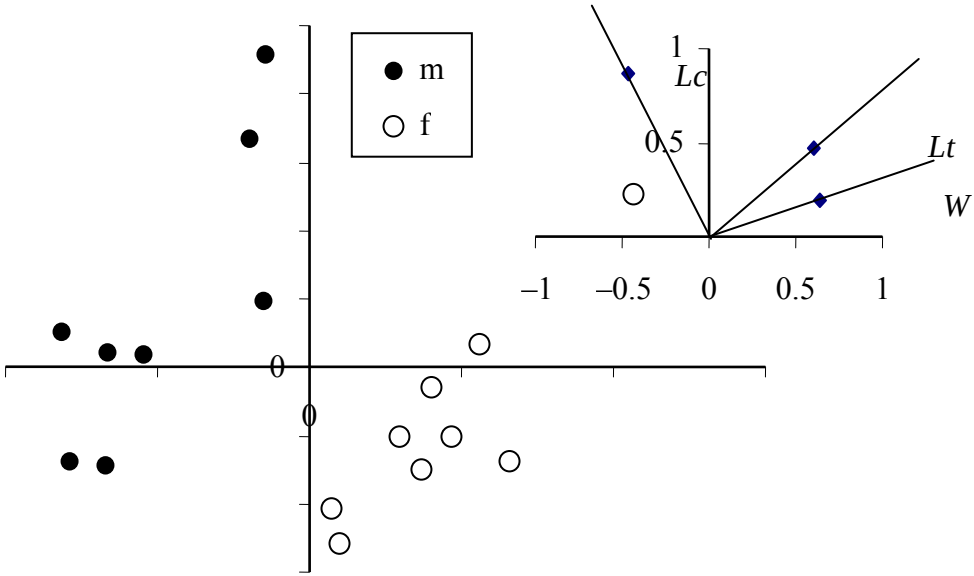


Рис. 8.1.7. Ординация промеров самцов (m) и самок (f) гадюки в осях двух главных компонент (А) и векторы, проведенные из начала координат через точки факторных нагрузок (Б)

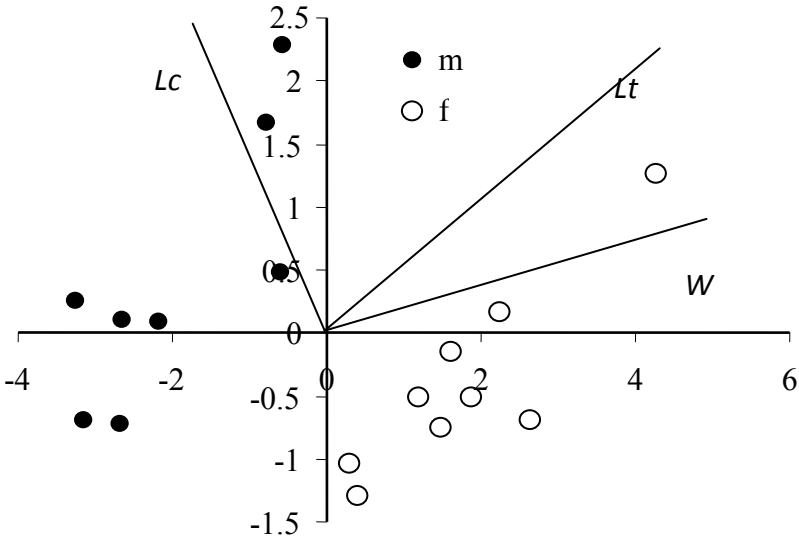


Рис. 8.1.8. Биплот: объединение диаграмм для компонент и факторных нагрузок.

Объединение двух диаграмм главных компонент и факторных нагрузок (рис. 8.1.8) возможно потому, что и компоненты, и факторные нагрузки есть безразмерные величины. Полученный двойственный график (biplot) наглядно показывает направления изменчивости объектов, за которые ответственны определенные признаки. По промерам гадюк видно, что первое направление изменчивости (выявленное первой главной компонентой) определяет отличие особей по массе (*W*) и длине тела (*Lt*) (соответствующие векторы направлены вдоль оси первой компоненты вправо), которым отчетливо противостоит длина хвоста (*Lc*) (вектор направлен влево). Второе направление изменчивости связано в основном с отличиями по длине тела и хвоста, поскольку оба вектора *Lt* и *Lc* направлены вдоль оси второй компоненты вверх, причем в большей мере, чем вектор массы тела.

Q-техника компонентного анализа

Процедура компонентного анализа применима и к транспонированной матрице исходных данных, в которой строки представляют собой значения отдельных признаков (*m* строк), а столбцы соответствуют отдельным объектам (*n* столбцов). Число признаков должно быть сравнимо с числом объектов.

Поскольку парные корреляции отыскиваются между двумя столбцами, то отдельный коэффициент корреляции есть не что иное, как мера коррелятивного сходства двух объектов, а матрица корреляций (размерностью *n***n*) есть матрица сходства всех объектов друг с другом.

Если по этой матрице рассчитать факторные нагрузки (собственные векторы), то они в компактном виде выразят степень сходства между объектами. В этом случае может быть всего *m* значений компонент, которые будут характеризовать относительную информативность изучаемых признаков (в этом и состоит Q-техника). Результаты анализа представленных выше данных по гадюкам максимально наглядно представляет биплот (рис. 8.1.9).

Здесь пунктирные векторы для самок направлены в основном по линии противостояния длины (*Lt*) и массы (*W*) тела, тогда как сплошные векторы нагрузок для самцов вытянуты в основном перпендикулярно этому первому направлению в сторону промеров длины хвоста (*Lc*).

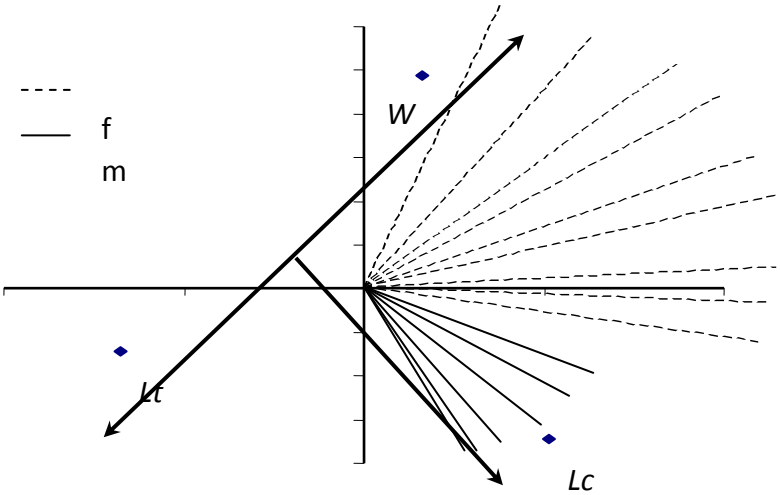
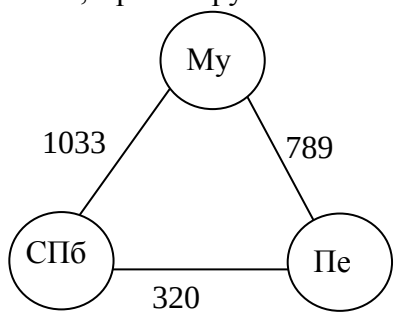


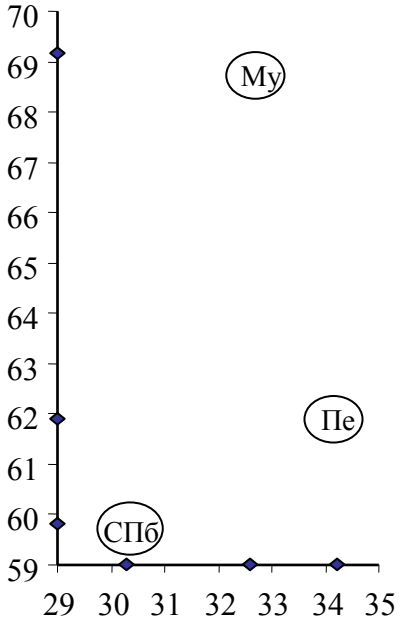
Рис. 8.1.9. Пример биплота для R-техники МГК

8.2. Многомерное шкалирование

Подобно другим многомерным методам шкалирование позволяет рассчитать несколько (немного) характеристик, выражающих основные направления изменчивости (взаимного отличия) изучаемых объектов. Особенность данного метода состоит в том, что для поиска новых осей исходными данными служит множество оценок различия объектов, только одна матрица дистанций.

Хороший интуитивный образ дает задача о *расчете географических координат* трех городов по *измеренным расстояниям* между ними. Например, от Санкт-Петербурга Мурманск отстоит на 1033 км, Петрозаводск – на 320 км, а между этими городами по прямой 789 км. Многомерное шкалирование, ориентируясь на взаимное удаление объектов, позволяет рассчитать некие значения координат (подобные широте и долготе местности). Так, если по широте Санкт-Петербург и Петрозаводск ближе друг к другу (южнее Мурманска), то по долготе сходство между Мурманском и Петрозаводском выше (эти города

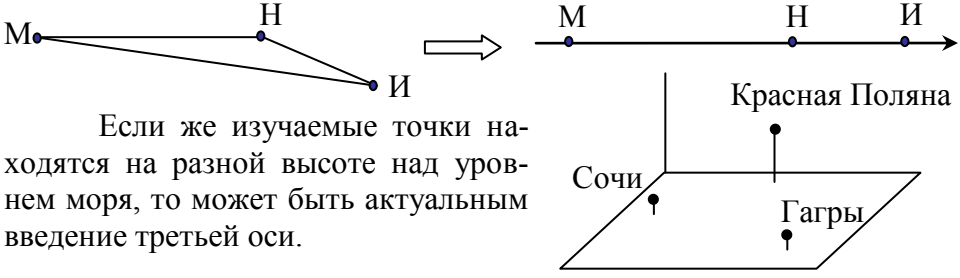




расположены восточнее северной столицы). Новые характеристики (оси, шкалы) выражают тонкие нюансы взаимоотношений между объектами. Многомерное шкалирование можно рассматривать как своеобразное продолжение кластерного анализа, когда полученная матрица расстояний служит не для построения дендрограммы, а для извлечения более детальной информации об отношениях между объектами.

В то же время расчетные оси координат характеризуют направления наиболее существенных отличий между объектами. В этом многомерное шкалирование похоже на метод главных компонент, который позволяет выразить самые важные факторы, влияющие на изменчивость изучаемых признаков. Отличие этих методов состоит, в частности, в том, что один из них изучает матрицу сходства объектов, второй – матрицу различий.

В примере число новых координат (шкал) равно двум, что кажется очевидным для географических задач. Фактически же при многомерном шкалировании размерность нового пространства (число новых осей) определяется лишь структурой отношений между объектами. В некоторых географических задачах для выражения основных особенностей объектов может быть достаточно одной оси, если они расположены примерно на одной линии. Так, расстояние от Москвы до Иркутска (4223 км) почти равно сумме расстояний между этими городами и Новосибирском (2840 и1399 км).



Если же изучаемые точки находятся на разной высоте над уровнем моря, то может быть актуальным введение третьей оси.

Определение минимально необходимой размерности нового пространства – это одна из важных задач, решаемых методом многомерного шкалирования.

Собственно процедура расчетов состоит в следующем. Объектом анализа служит квадратная (n*n) матрица расстояний (**D**) между n объектами. Важное свойство такой матрицы состоит в том, что все диагональные элементы, выражающие степень отличия каждого объекта от самого себя, равны нулю ($D_{ii} = D_{jj} = 0$), различия между любой парой объектов обычно от нуля отличаются $D_{ij} \neq 0$. Для алгоритма многомерного шкалирования безразлично, каким способом получены оценки различий между всеми парами объектов – они могут быть чувственными, балльными, ранговыми, корреляционными, Съёренсена, Чекановского, евклидовы и др. В любом случае некие исходные индивидуальные характеристики объектов (переменные y) используются лишь для оценки дистанций $D_{ij} = f(y_{ip}, y_{jp})$ (i, j – индексы объектов, p – индекс признака) и в дальнейших расчетах не участвуют.

Цель анализа состоит в том, чтобы для каждого из n объектов рассчитать минимально необходимое количество (k) таких новых характеристик (x_1, x_2, ... x_k), чтобы они выражали существенные свойства этих объектов, чтобы и в новых осях $\boldsymbol{x}$ отношения между объектами соответствовали реальности. Говоря конкретно, различия между объектами в новых шкалах $\delta_{ij} = f(x_{ip}, x_{jp})$ должны каким-то образом соответствовать исходным различиям D_{ij} между ними, то есть чтобы матрица расстояний $\boldsymbol{\delta}$ была в некотором смысле эквивалентна (равна, пропорциональна) матрице расстояний **D**.

Разные методы многомерного шкалирования требуют разную степень подобия матриц $\boldsymbol{\delta}$ и **D**. *Метрическое* шкалирование строится на прямой пропорции $\mathbf{D} = \boldsymbol{\delta}$, расстояние между каждой парой объектов в новых осях должно быть численно равно исходно заданной дистанции: $D_{ij} = \delta_{ij}$. *Неметрическое* шкалирование исходит из того предположения, что новые расстояния $\boldsymbol{\delta}$ могут и не быть прямо пропорциональны исходным значениям **D**, но между ними имеется определенное, хотя и менее жесткое, соотношение, заданное некоторой функцией f: $D_{ij} = f(\delta_{ij})$. Наиболее мягкое требование такого рода не предполагает какой-либо численной пропорции между матрицами расстояний, кроме сохранения *условия порядка*: если исход-

ное различие между объектами i и j больше, чем различие между объектами p и q, то это же соотношение должно сохраняться в новых координатах: если $D_{ij} > D_{pq}$, то $\delta_{ij} > \delta_{pq}$.

Расчет координат $\boldsymbol{x}$ выполняется при постоянном контроле за сходством матриц расстояний $\boldsymbol{\delta}$ и **D**. Различие между исходной и расчетной матрицами дистанций носит не очень удачное название *стресс*, который вычисляется с помощью большого набора разнообразных формул. Мы рассмотрим лишь первую формулу Краскала (S_1) и коэффициент отчуждения Гуттмана (k). Чем меньше стресс или коэффициент отчуждения, тем точнее матрица новых различий $\boldsymbol{\delta}$ воспроизводит исходную **D**.

Общая последовательность процедуры многомерного шкалирования такова (существует множество алгоритмов). Вначале (1) выполняются измерения или оценка расстояний между группой объектов с помощью разнообразных мер отличия, результатом чего оказывается квадратная матрица дистанций **D**. Затем (2) назначают некие начальные значения новых переменных $\boldsymbol{x}$. Это могут быть набор случайных чисел или величины, имеющие теоретический смысл, или результаты предварительной обработки матрицы расстояний (суммы значений в строках) и др. Используя эти шкалы, (3) рассчитывается матрица новых расстояний между объектами $\boldsymbol{\delta}$ и затем (4) – стресс, обобщенное различие между **D** и $\boldsymbol{\delta}$. Далее (5) с помощью математических процедур (итерации или оптимизации и математического программирования) идет направленное изменение стартовых значений $\boldsymbol{x}$ с целью снижения стресса. Смысл процедур итерации и оптимизации рассмотрен в литературе (Лоули, Максвелл, 1967; Шуп, 1990; Коросов, 1996, 2002). Важно то, что они могут быть выполнены в среде Excel с помощью встроенной программы Поиск решения. Вариации значений $\boldsymbol{x}$ осуществляются многократно после чего идет возврат на шаг (3) и (4). Процедура завершается, когда при очередных модификациях $\boldsymbol{x}$ существенного изменения стресса больше не происходит. Значения стресса служат и критерием отбора значимых осей x, и показателем качества выполненного анализа. Считается, что при $S_1 < 0.1$ между матрицами **D** и $\boldsymbol{\delta}$ имеется хорошее совпадение, и расчетные значения $\boldsymbol{x}$ выражают главные факторы изменчивости изучаемых объектов.

8.3. Метрическое шкалирование

В основе этого вида многомерного шкалирования лежит предположение о равенстве матриц расстояний: **δ** = **D**, при условии, что **δ** состоит из значений евклидовой меры различий между каждой парой (*i, j*) объектов в *k* новых осях:. $\delta_{ij} = \sqrt{\sum^{k}(x_{ik} - x_{ik})^2}$ Такое требование предполагает, что матрица **D** должна быть получена с опорой на некие признаки ***y***, определенные в интервальных, относительных или частных абсолютных шкалах (таких как масса, температура, численность). Если же признаки ***y*** заданы в порядковых или номинальных шкалах (оценки в баллах), то к матрице **D** следует применить алгоритмы неметрического шкалирования (п. 8.4).

В качестве примера рассмотрим полученную ранее матрицу различия между шестью местообитаниями мелких млекопитающих Прибайкалья (**к**едровник, **п**ихтач, **э**котон, **с**осняк, **б**ерезняк, **л**уг), полученную с помощью меры Чекановского (табл. 5.4.2, рис. 8.3.1). Поскольку видовой состав описан в частной абсолютной шкале (численность), метрическое шкалирование применять можно. Задача состоит в том, чтобы, ориентируясь на матрице сравнения биотопических группировок, объяснить ее структуру с помощью немногих факторов (причин изменчивости).

Метод главных координат

Идея метода метрического шкалирования состоит в том, что в евклидовом пространстве каждый элемент d_{ij} *дважды центрированной матрицы расстояний* есть произведение векторов новых характеристик x_i и x_j: $d_{ij} = x_i x_j$.

(Центрирование гарантирует, что матрица не будет вырожденной и собственные значения не будут содержать больших отрицательных величин.)

Дважды центрированную матрицу **d** получают из исходной матрицы **D**, если от каждого ее элемента, возведенного в квадрат (D_{ij}^2), отнять средний квадрат значений по строкам ($\overline{D}_{i.}^2$), отнять средний квадрат значений по столбцам ($\overline{D}_{.j}^2$), прибавить средний

квадрат значений по всей матрице ($\overline{D}^2_{..}$) и умножить на –1/2 (рис. 8.3.1): $d_{ij} = -(D^2_{ij} - \overline{D}^2_{i.} - \overline{D}^2_{.j} + \overline{D}^2_{..})/2$.

	A	B	C	D	E	F	G	H	I	J
1	D	К	П	Э	С	Б	Л			
2	К	0	0.218	0.338	0.461	0.397	0.626			
3	П	0.218	0	0.259	0.351	0.273	0.603			
4	Э	0.338	0.259	0	0.225	0.116	0.471			
5	С	0.461	0.351	0.225	0	0.187	0.586			
6	Б	0.397	0.273	0.116	0.187	0	0.462			
7	Л	0.626	0.603	0.471	0.586	0.462	0			
8										
9	D^2	К	П	Э	С	Б	Л			
10	К	0.000	0.047	0.115	0.212	0.158	0.391	0.154		
11	П	0.047	0.000	0.067	0.123	0.074	0.363	0.113		
12	Э	0.115	0.067	0.000	0.051	0.013	0.222	=СРЗНАЧ(B12:G12)		
13	С	0.212	0.123	0.051	0.000	0.035	0.343	0.127		
14	Б	0.158	0.074	0.013	0.035	0.000	0.213	0.082		
15	Л	0.391	0.363	0.222	0.343	0.213	0.000	0.256		
16		0.154	0.113	0.078	0.127	0.082	0.256	0.135		
17										
18	d	К	П	Э	С	Б	Л			
19	К	0.086	0.042	-0.009	-0.033	-0.028	-0.058			
20	П	0.042	0.045	-0.006	-0.009	-0.007	-0.065			
21	Э	-0.009	-0.006	0.011	0.010	0.006	-0.012			
22	С	-0.033	-0.009	0.010	0.060	0.020	-0.048			
23	Б	-0.028	-0.007	0.006	=-(E14-$H14-E$16+$H$16)/2					
24	Л	-0.058	-0.065	-0.012	-0.048	-0.005	0.188			

Рис. 8.3.1. Этапы центрирования матрицы расстояний (изображения разных формул совмещены с помощью редактора растра)

Так, квадрат расстояния между кедровником и пихтачом равен $D^2_{\mathbf{КП}} = 0.218^2 = 0.047$. Средний квадрат по первой строке составит $\overline{D}^2_{\mathbf{К}.} = 0.154$, по второму столбцу – $\overline{D}^2_{.\mathbf{П}} = 0.113$, а по всей матрице – $\overline{D}^2_{..} = 0.135$. Отсюда центрированная оценка сходства кедровника с пихтачом равна:

$$d_{\mathbf{КП}} = -(0.047 - 0.154 - 0.113 + 0.135)/2 = 0.042.$$

Теперь по матрице **d** предстоит определить частные шкалы ***x***, измеряющие отличия между объектами. Показано (Дэйвисон, 1988), что значения новых характеристик ***x*** есть *собственные векторы* матрицы **d**, которые получают с помощью *Q*-техники компонентного анализа (в его терминах искомые новые характеристики – это факторные нагрузки, см. п. 8.1). Конкретное значение x_{1i} соответствует положению *i*-го объекта на первой новой оси (факторные нагрузки первой компоненты, PC-1), x_{2i} – координате этого объекта на второй оси (факторные нагрузки PC-2) и т. д. Обычно число новых осей невелико (1–3). Для них выполняются известные условия: *ортогональность* (*главные координаты* взаимно перпендикулярны, то есть новые признаки не коррелируют), *неравенство* дисперсий (наибольшую дисперсию имеет первая координата, вторую по величине – вторая и т. д.); если дисперсия координаты (*собственное число* координаты) меньше средней дисперсии исходного признака, то данная координата считается незначимой и не рассматривается.

К сожалению, распространенные пакеты статистической обработки (StatGraphics, Statistica) не имеют модулей расчета собственных чисел и собственных векторов по матрице расстояний. Для этого следует воспользоваться пакетами математической обработки данных (MatCad, MathLab) или создать собственный макрос на языке Visual Basic в среде Excel, который выполняет расчеты по приведенной программе (табл. 8.3.1).

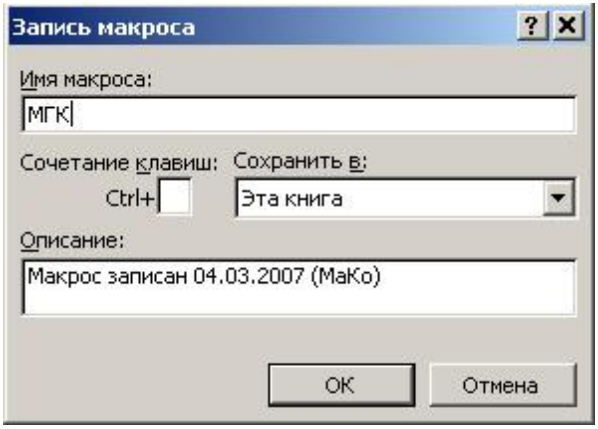
Макрос создается в несколько этапов. Сначала нужно ввести текст программы на лист Блокнота (Notepad, не Word!) и сохранить как файл *.txt.

Табл. 8.3.1. Листинг макроса Visual Basic оценки главных координат

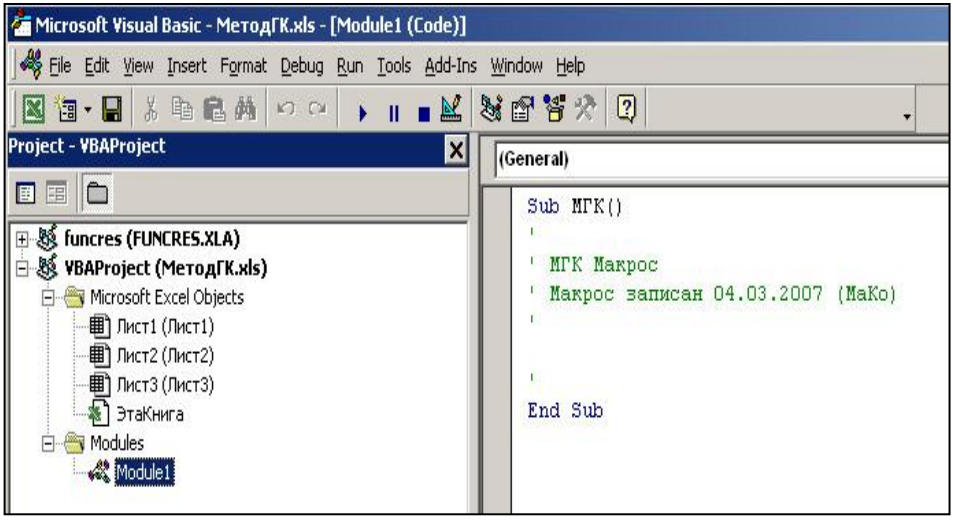
```
Dim NewList As String
Dim NameOfObject(50) As String * 10
Dim r(50, 50), u(50, 50), V(50, 50) As Single
Dim W(50), U1(50), S(50) As Single
Dim mm, i, j, k, l, m, n, p1, uOLD, Ru As Single
Dim SS, S1, DEV, UU, VV, UV As Single
NewList = ActiveSheet.Name
'-------------------------- reading name of object (m)
i = 0
Do : i = i + 1
NameOfObject(i) = Worksheets(NewList).Cells(1, i + 1)
Loop While Trim(NameOfObject(i)) <> ""
m = i - 1
For i = 1 To m : For j = 1 To m
 r(i, j) = Val(Worksheets(NewList).Cells(i + 1, j + 1))
Next j: Next i
Worksheets(NewList).Cells(3 + m, 1).Formula = "PCA"
Worksheets(NewList).Cells(4 + m, 1).Formula = "Loads"
mm = m : p1 = 0.00001 '----error level
'============================== calculating mm vectors
For k = 1 To mm : uOLD = 1
 For i = 1 To m : Ru = 0 : For j = 1 To m:
   Ru = Ru + r(j, i) : Next j: U1(i) = Ru : Next i
KOC:
 S(k) = 0
 For i = 1 To m :
   If Abs(S(k)) < Abs(U1(i)) Then S(k) = U1(i)
 Next i
ss = 1/S(k) : For i = 1 To m : u(i,k) = U1(i)*ss : Next i
DEV = Abs((Abs(S(k))-uOLD)/uOLD)
   If DEV <= p1 Then GoTo KOB
uOLD = Abs(S(k))
For i = 1 To m : Ru = 0 : For j = 1 To m:
  Ru = Ru + u(j,k)*r(j,i): Next j: U1(i) = Ru: Next i
GoTo KOC
'------------earching for of the auxiliary vector
KOB:
 UU = 0:For i = 1 To m: UU = UU + u(i,k)*u(i,k): Next i
 VV = Sqr(Abs(S(k)) / UU) : UV = Sqr(UU)
For i = 1 To m:W(i) =VV*u(i,k):V(i, k) = u(i,k)/ UV:Next i
```

```
'------calculation of the matrix to remaining correlation
 For i = 1 To m : For j = 1 To m
   r(j, i) = r(j, i) - W(j) * W(i) : Next j : Next i
Next k
'---------------------------output of the results
For i = 1 To m
Worksheets(NewList).Cells(5 + i + m, 1) = NameOfObject(i)
  For j = 1 To m :
Worksheets(NewList).Cells(5 + i + m, j + 1) = Str(V(i, j))
  Next j
Next i
Worksheets(NewList).Cells(7 + 2 * m, 1) = "Disp":
Worksheets(NewList).Cells(8 + 2 * m, 1) = "Disp%"
SS = 0: For i = 1 To mm: SS = SS + S(i): Next
For i = 1 To mm: S1 = Int(100 * S(i) / SS)
 Worksheets(NewList).Cells(5 + m, i + 1) = "PC-" & Str(i)
 Worksheets(NewList).Cells(7 + 2 * m, i + 1) = Str(S(i))
 Worksheets(NewList).Cells(8 + 2 * m, i + 1) = Str(S1)
Next
```

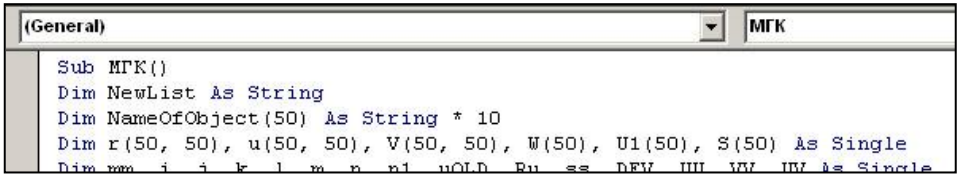
Затем в среде Excel нужно создать новую книгу (Файл \ Создать) и сохранить ее под именем, например МетодГК. Дать команду Сервис \ Макрос \ Начать запись, ввести его имя (например, МГК), ОК. На появившейся иконке нажать синий квадратик, завершив запись.



Далее начинаем редактировать макрос. По команде Сервис \ Макрос \ Редактор Visual Basic открываем редактор. В окне Project - VBAProject находим проект VBAProject (МетодГК), в нем – список модулей (Modules) и среди них – наш модуль (Module1); щелкаем на него мышкой. Далее командой главного меню View \ Code открываем код (программу) записанного ранее модуля. Он состоит пока из двух операторов Sub МГК() и End Sub, а также из следующих за апострофом ‘ ненужных комментариев между ними, которые стоит удалить.



Запускаем редактор Блокнот, открываем файл с набранным кодом макроса (текстом программы), копируем его целиком в буфер обмена и вставляем между операторами Sub МГК() и End Sub.



Сохраняем макрос Excel командой File \ Save МетодКГ и закрываем окно редактора Visual Basic. Чтобы им можно было пользоваться в дальнейшем, в окне, вызываемом командой Сервис \ Макрос \ Безопасность, следует установить хотя бы Средний уровень. Сохраняем файл МетодГК (Файл \ Сохранить).

Используется макрос следующим образом. В верхний левый угол любого листа книги МетодГК с макросом МГК помещаем массив дважды центрированных расстояний **d** (строки и столбцы поименованы). Далее даем команду главного меню Сервис \ Макрос \ Макросы, выбираем Имя макроса МГК, нажимаем кнопку Выполнить. Снизу от данных появляется матрица расчетов значений собственных векторов (факторных нагрузок, Factorial loads), собственных значений (дисперсий, Disp) и относительной величины дисперсий (Disp%) (рис. 8.3.2).

Microsoft Excel - МетодГК.xls

Файл Правка Вид Вставка Формат Сервис Диаграмма Окно Справка

Times New Roman CYR 13 Ж К Ч

Область диаграммы =

	A	B	C	D	E	F	G
1	d	К	П	Э	С	Б	Л
2	К	0.086	0.042	-0.009	-0.033	-0.028	-0.058
3	П	0.042	0.045	-0.006	-0.009	-0.007	-0.065
4	Э	-0.009	-0.006	0.011	0.010	0.006	-0.012
5	С	-0.033	-0.009	0.010	0.060	0.020	-0.048
6	Б	-0.028	-0.007	0.006	0.020	0.015	-0.005
7	Л	-0.058	-0.065	-0.012	-0.048	-0.005	0.188
8							
9	Principal component analysis						
10	Factorial loads						
11		PC- 1	PC- 2	PC- 3	PC- 4	PC- 5	PC- 6
12	К	-0.374	-0.610	-0.501	-0.078	0.254	-0.374
13	П	-0.345	-0.214	0.709	-0.214	-0.346	-0.345
14	Э	-0.025	0.152	-0.165	0.783	-0.411	-0.025
15	С	-0.132	0.632	-0.351	-0.510	-0.183	-0.132
16	Б	0.025	0.302	0.309	0.203	0.778	0.025
17	Л	0.850	-0.262	-0.001	-0.184	-0.092	0.850
18							
19	Disp	0.248	0.127	0.018	0.011	0.002	0.000
20	Disp%	61	31	4	2	0	-1

Рис. 8.3.2. Оценка местообитаний мелких млекопитающих в новых осях координат – метрических шкалах

Поскольку на каждый из шести изученных признаков приходится примерно по 100 / 6 = 17% общей изменчивости, то главные координаты PC, имеющие относительные значения дисперсии (Disp%) меньше этой величины, можно не рассматривать. В нашем примере остаются лишь первая и вторая шкалы, учитывающие существенную долю информации об отличиях объектов (в сумме 82%), прочие улавливают лишь случайные эффекты. Значения собственных векторов (значения полей PC-1 и PC-2) и есть координаты расположения шести объектов на плоскости двух осей (рис. 8.3.3).

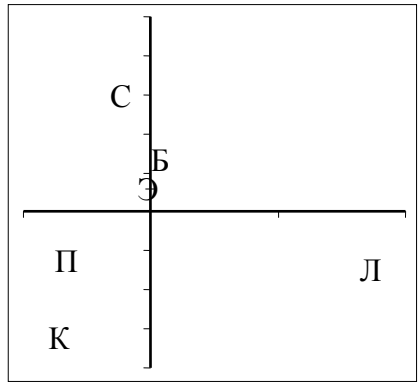


Рис. 8.3.3. Расположение местообитаний мелких млекопитающих на плоскости двух расчетных осей

Заключительный этап шкалирования состоит в том, чтобы дать интерпретацию (названия) новым осям координат, что позволяет лучше понять структуру отношений между рассматриваемыми объектами. Основой для этого служат теоретические соображения о качестве сравниваемых групп и результаты непосредственных замеров их свойств. Ориентируясь на матрицу исходных данных (переменные ***y***) по численности всех видов (табл. 5.4.1), нетрудно увидеть, что противостояние самых заселенных биотопов (кедровника с пихтачом) самому малонаселенному (луг) по первой оси можно обозначить как «общая численность». На роль второй оси напрашивается «видовое богатство», но этому мешает тот факт, что максимальное видовое разнообразие имеет березняк (см. табл. 5.2.1), который на диаграмме (рис. 8.3.3) занимает почти центральное положение. Кроме того, в показателях сходства по выравненности ведущую роль играет не наличие вида, а его численность. Учитывая эту мысль, мы рассчитали соотношение грызунов и насекомоядных. Оказалось, что группировка микромаммалий сосняка отличается от всех прочих самой высокой долей грызунов (67%) вследствие очень низкой численности бурозубок; в кедровнике и пихтаче эта величина много ниже: 49 и 46%, чему хорошо соответствует расположение данных биотопов на диаграмме. Однако отрыв между березняком и

лугом по второй оси слишком велик, хотя они имеют почти равные «доли грызунов» (53 и 52%) (но на лугу их численность ниже). Таким образом, вторую ось можно назвать «доля и обилие грызунов».

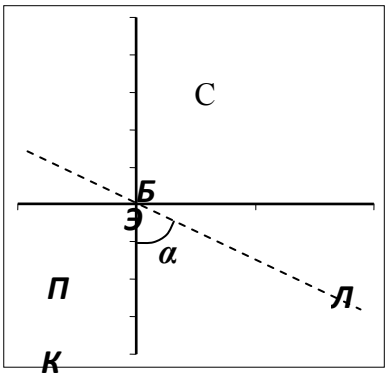
Вторая ось оказалась не вполне самостоятельной, на ней так же, как и на первой оси, сказался фактор численности животных. Это не очень хороший результат, поскольку задача состояла в выявлении факторов различия территориальных группировок «в чистом виде». Есть несколько выходов.

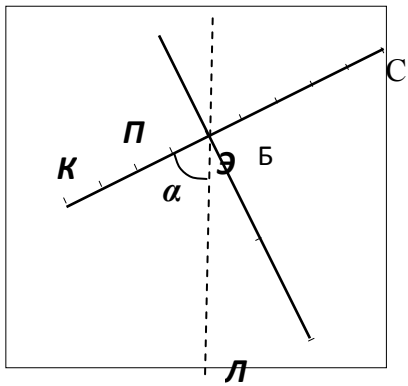
Во-первых, следовало бы уточнить исходные характеристики $\boldsymbol{y}$ с точки зрения соответствия поставленным целям. В нашем случае, например, в расчет брались многолетние средние значения учетов в давилки, которые плохо отлавливают бурозубок. В свою очередь, многолетние средние нивелируют истинные биотопические предпочтения, поскольку сильно зависят от лет с высокой численностью и широким расселением всех видов. Возможно, более уместными характеристиками выступили бы спектры «индексов верности биотопу» или раздельный анализ ситуации в годы с высокой и низкой численностью зверьков. Во-вторых, можно выполнить процедуру переопределения осей с помощью метода наименьших квадратов, рассмотренного в следующем разделе. Наконец, возможен «поворот осей», часто делающий новое положение объектов более «понятным». В рамках факторного анализа проводится автоматический поворот с использованием разных алгоритмов (вариамакс и др.). Мы выполним «ручной поворот».

Идея «ручного» метода состоит в том, чтобы, ориентируясь на необходимый угол поворота α, осуществить тригонометрические преобразования прежних координат (Дейвисон, 1988, с. 72). Как было сказано, соотношение грызунов и бурозубок в березняках и на лугах примерно одинаково, логично было бы для отражения этого факта выровнять эти биотопы относительно оси ординат, то есть повернуть их на угол около $\alpha \approx 60°$ (рис. 8.3.3).

Матрица ортогонального преобразования содержит косинусы углов, на которые нужно выполнить поворот. В нашем случае $\cos 60° \approx 0.5$ и матрица преобразований примет вид: $\begin{pmatrix} 0.5 & -0.5 \\ 0.5 & 0.5 \end{pmatrix}$.

На нее нужно умножить справа матрицу главных координат. Это просто сделать в среде Excel (рис. 8.3.4). Так, первый элемент матрицы равен: (–0.374·0.5) + (–0.610·0.5) = –0.492.



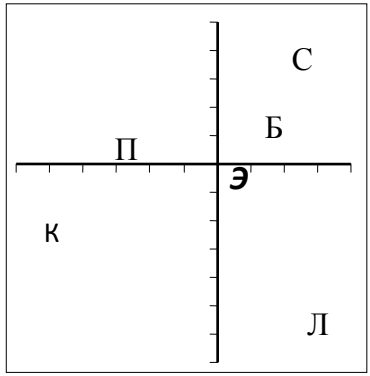


=B6*E$2+C6*E$3

	A	B	C	D	E	F	G	H	I
1		PC- 1	PC- 2			x1	x2		
2	К	-0.374	-0.610	0.5	-0.5	-0.492	-0.118		
3	П	-0.345	-0.214	0.5	0.5	-0.280	0.065		
4	Э	-0.025	0.152			0.064	0.089		
5	С	-0.132	0.632			0.250	0.382		
6	Б	0.025	0.302			0.164	=B6*E$2+C6*E$3		
7	Л	0.850	-0.262			0.294	-0.556		

Рис. 8.3.4. Поворот главных координат

Рис. 8.3.5. Соотношения биотопов после поворота осей



В результате расчетов оси координат повернулись, интерпретация изменилась. Первая координата теперь стала читаться как «доля бурозубок», а вторая приобрела смысл «численность рыжих полевок»: максимальное значение имеет сосняк, минимальное – луг.

Метод наименьших квадратов

Рассмотренный выше метод главных координат использует итерационную процедуру последовательного уточнения значений собственных векторов (новых координат объектов). Для этой же цели можно применить и другой прием – *многомерную оптимизацию*, которая в среде Excel выполняется встроенным макросом Поиск решения.

Ключевые идеи те же: расчеты строятся на условии, что матрица расстояний между объектами в новых осях **δ** должна быть равна исходной матрице **δ** = **D**. В качестве меры различия используется сумма квадратов разности между каждой из *n* комплиментарных пар δ_{ij} и D_{ij} (Справочник…, 1990): $S = \sum^{n}(\delta_{ij} - D_{ij})^2$.

В процессе оптимизации выполняется направленный подбор таких значений новых переменных ***x***, чтобы сумма квадратов *S* стала наименьшей, $S \to 0$. Выполнить это требование в точности не удается, обычно ограничиваются некоторым небольшим порогом. Для унификации результатов счета можно использовать *первую формулу стресса* Краскала (Пузаченко, 2004, с. 219), которая основана на той же сумме квадратов: $S_1 = \sum^{n}(\delta_{ij} - D_{ij})^2 / \sum^{n}\delta_{ij}^2$.

Эта нормированная величина минимальным значением имеет нуль $S_1 = 0$, но никогда его не достигает. При уровне $S_1 < 0.1$ сходство между матрицами считается достаточно хорошим.

Используем эту технологию для нашего примера. Когда число сравниваемых объектов не очень велико ($n < 50$), все расчеты можно организовать на одном листе книги Excel. Он содержит *шесть конструкций* (рис. 8.3.6).

1) Исходная матрица расстояний **D** помещена в блок ячеек A1:G7.

2) Набор новых координат *x* занесен в блок B17:E23. Обычно достаточно всего две, редко три шкалы. В качестве исходного набора значений рекомендуется брать результаты шкалирования методом главных координат. Для унификации новых шкал предлагается *приводить средние к нулевым значениям* и *требовать отсутствия корреляций* между ними (условие ортогональности).

	A	B	C	D	E	F	G	H	I	J	K	L
1	**D**	К	П	Э	С	Б	Л		**δ**	**D**	**(δ-D)^2**	**δ^2**
2	К	0	0.22	0.34	0.46	0.40	0.63		0.40	0.22	0.03	0.16
3	П	0.22	0	0.26	0.35	0.27	0.60		0.84	0.34	0.25	0.70
4	Э	0.34	0.26	0	0.23	0.12	0.47		1.26	0.46	0.65	1.60
5	С	0.46	0.35	0.23	0	0.19	0.59		1.00	0.40	0.36	0.99
6	Б	0.40	0.27	0.12	0.19	0	0.46		1.27	0.63	0.42	1.62
7	Л	0.63	0.60	0.47	0.59	0.46	0		0.49	0.26	0.05	0.24
8									0.87	=C5	0.27	0.76
9	δ	К	П	Э	С	Б	Л		0.64	0.27	0.13	0.40
10	К								1.20	0.60	0.35	1.43
11	П	0.40							0.49	0.23	0.07	0.24
12	Э	0.84	0.49						0.16	0.12	0.00	0.02
13	С	1.26	0.87	0.49					0.97	0.47	0.25	0.94
14	Б	1.00	=КОРЕНЬ(($C$19-$C22)^2+($D$19-$D22)^2+($E$19-$E22)^2)									
15	Л	1.27	1.20	0.97	1.33	1.00			1.33	0.59	0.55	1.76
16									1.00	0.46	0.29	1.00
17			x1	x2	x3		0.19		S1 =	0.56	3.70	12.00
18		К	-0.37	-0.61								
19		П	-0.34	-0.21								
20		Э	-0.02	0.15								
21		С	-0.13	0.63								
22		Б	0.03	0.30								
23		Л	0.85	-0.26								
24		Σ	0.00	=СРЗНАЧ(D18:D23)								
25		*r*		0.00								
26												
27		Stress	0.56									
28												
29												

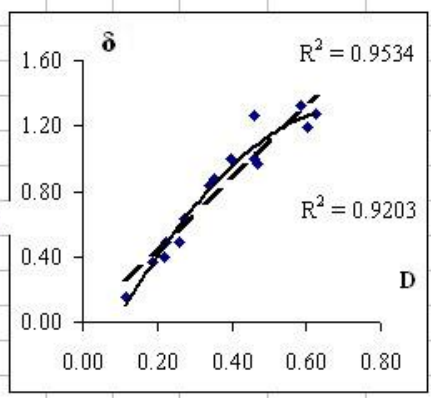


Рис. 8.3.6. Подготовка среды Excel для метрического шкалирования (с помощью редактора растра объединены несколько изображений)

Вводим формулы усреднения в ячейки C24, D24, E24, а в ячейку D25 – формулу корреляции =КОРРЕЛ(C18:C23,D18:D23). Ранее полученные главные координаты (как собственные значения) центрированы и не коррелируют, поэтому введенные нами формулы выдали нуль. При дальнейшей настройке координат от процедуры

оптимизации нужно будет потребовать сохранения в этих ячейках нулевых значений.

3) Расчет матрицы расстояний **δ** между объектами в новых координатах организуется в блоке A9:G15 с помощью евклидовой меры: $\delta_{ij} = \sqrt{\sum(x_{ik} - x_{ik})^2}$. Матрицы расстояний симметричны относительно главной диагонали, поэтому можно рассчитать значения лишь для нижней треугольной части. Вводим формулы, начиная с ячейки B11:
=КОРЕНЬ(($C$18-$C19)^2+($D$18-$D19)^2+($E$18-$E19)^2),
что в численном виде дает:

$$\sqrt{(-0.37+0.34)^2+(-0.61+0.21)^2+(0-0)^2} = 0.40.$$

Префиксы позволяют выполнить автозаполнение одного столбца матрицы расстояний. Для других столбцов ссылки нужно специально поменять мышкой (перетаскивая рамки ссылок в режиме редактирования скопированной формулы). В ячейке C14 имеем:
=КОРЕНЬ(($C$19-$C22)^2+($D$19-$D22)^2+($E$19-$E22)^2)
или $\sqrt{(-0.34-0.03)^2+(-0.21-0.30)^2+(0-0)^2} = 0.64$ и т. д.

4) Расчет функции стресса S_1 организован в блоке I1:L17. Для большей наглядности и чтобы получить возможность строить диаграммы и перестраивать структуру критерия, элементы сравниваемых треугольных матриц **δ** и **D** *копируются* в два столбца I1:I16 и D1:D16 с помощью *простых ссылок* на соответствующие элементы матриц, начиная с I2 =B11 и J2 =B3. Далее находим квадрат разности между каждой парой элементов матриц (столбец K2:K16), например K4 =(J4-I4)^2 или $(1.26 - 0.46)^2 = 0.65$; значения **δ** возводим в квадрат (L2:L16). Для каждого столбца находим сумму значений K17 =СУММ(K2:K16), L17 =СУММ(L2:L16). Стресс равен корню из частного от деления сумм по этим столбцам: D17 =КОРЕНЬ(K17/L17), $S_1 = \sqrt{3.70/12} = 0.56$.

5) Диаграмма призвана отображать соотношения значений δ_{ij} (ось ординат) и D_{ij} (ось абсцисс). Если новые координаты ***x*** будут полноценно характеризовать объекты, то расчетная и исходная матрицы

расстояний между ними совпадут, а все точки (δ_{ij}, D_{ij}) выстроятся в одну линию под углом 45°. Обычно столь идеального отношения не бывает: точки частично распылены и оси не пропорциональны. В примере δ изменяется от 0.16 до 1.33, а D – от 0.12 до 0.63. Если все точки, начиная с левых нижних, соединить одной ломаной линией, мы получили бы *диаграмму Шеппарда* (Дэйвисон, 1988, с. 101), которая позволяет анализировать величину различий между матрицами расстояний в разных областях значений. В целом же оценить характер пропорции между δ_{ij} и D_{ij} позволяет сравнение линейного и параболического трендов, построенных по всем точкам (δ_{ij}, D_{ij}). На нашей диаграмме коэффициенты детерминации линейного (приведен снизу, $R^2 = 0.92$) и параболического (приведен сверху, $R^2 = 0.95$) трендов численно выражают доминирующую тенденцию. Их большие значения свидетельствуют о том, что между матрицами **δ** и **D** имеется очень высокое сходство. Очевидно и то, что в их соотношении есть определенная кривизна (параболический тренд лучше описывает пропорцию, чем линейный, 0.95 > 0.92). Значит, матрицы расстояний совпадают не идеально и координаты ***x*** требуют настройки.

6) Макрос оптимизации вызывается командой меню Сервис \ Поиск решения (рис. 8.3.7). Мышкой следует Установить целевую ячейку со значением стресса S_1 ($J$17) Равной значению 0, Изменяя ячейки $C$18:$D$23 (содержащие обе настраиваемые переменные x_1 и x_2), а также, нажав кнопку Добавить, ввести Ограничения: $C$24=0, $D$24=0 (необязательное требование нулевых средних), $C$25=0 (обязательное требование ортогональности двух новых осей). Заполнив поля, нажимаем Выполнить.

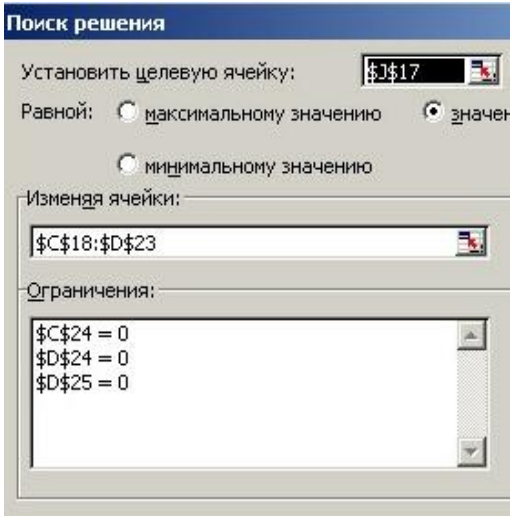


Рис. 8.3.7. Заполнение макроса оптимизации

Если появившееся затем окно Результаты поиска решения содержит сообщение Поиск не может найти подходящего решения, то следует нажать кнопку ОК. Стресс никогда не сводится к нулю (а задача оптимизации сформулирована именно так), но если величина S_1 стала меньше 0.1, настройку можно считать успешной.

В нашем случае (рис. 8.3.8) эта критическая граница оказалась пройденной, $S_1 = 0.06 < 0.1$. Матрицы расстояний **δ** и **D** стали гораздо ближе друг к другу, точки на диаграмме стянулись к линейному тренду, параболический тренд выпрямился, их коэффициенты детерминации стали одинаково высокими $R^2 \approx 0.98$.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**D**	К	П	Э	С	Б	Л		**δ**	**D**	**(δ-D)^2**	**δ^2**
2	К	0	0.22	0.34	0.46	0.40	0.63		0.19	0.22	0.00	0.04
3	П	0.22	0	0.26	0.35	0.27	0.60		0.33	0.34	0.00	0.11
4	Э	0.34	0.26	0	0.23	0.12	0.47		0.50	0.46	0.00	0.25
5	С	0.46	0.35	0.23	0	0.19	0.59		0.41	0.40	0.00	0.17
6	Б	0.40	0.27	0.12	0.19	0	0.46		0.62	0.63	0.00	0.38
7	Л	0.63	0.60	0.47	0.59	0.46	0		0.23	0.26	0.00	0.05
8									0.33	0.35	0.00	0.11
9	**δ**	К	П	Э	С	Б	Л		0.29	0.27	0.00	0.08
10	К								0.64	0.60	0.00	0.41
11	П	0.19							0.23	0.23	0.00	0.05
12	Э	0.33	0.23						0.08	0.12	0.00	0.01
13	С	0.50	0.33	0.23					0.44	0.47	0.00	0.20
14	Б	0.41	0.29	0.08	0.16				0.16	0.19	0.00	0.03
15	Л	0.62	0.64	0.44	0.60	0.45			0.60	0.59	0.00	0.36
16									0.45	0.46	0.00	0.21
17			x1	x2	x3		0.21		S1 =	0.06	0.01	2.44
18		К	-0.17	-0.24								
19		П	-0.21	-0.06								
20		Э	0.00	0.03								
21		С	-0.08	0.24								
22		Б	0.02	0.11								
23		Л	0.43	-0.09								
24		Σ	0.00	0.00	####							
25		*r*		0.00								
26												
27		Stress	0.06									
28												
29												

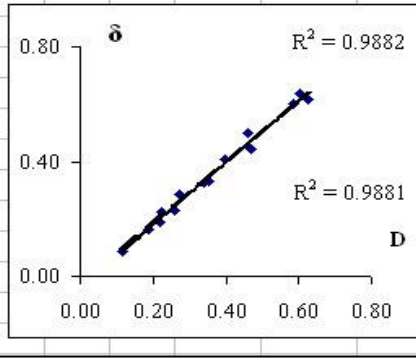


Рис. 8.3.8. Результаты метрического шкалирования

После настройки прежняя картина (см. рис. 8.3.3) лишь немного изменилась (рис. 8.3.9): по первой оси пихтач переместился левее кедровника, что лучше соответствует соотношению «общей численности» в этих местообитаниях. В то же время расположение биотопов по второй оси по-прежнему не поддается однозначной интерпретации.

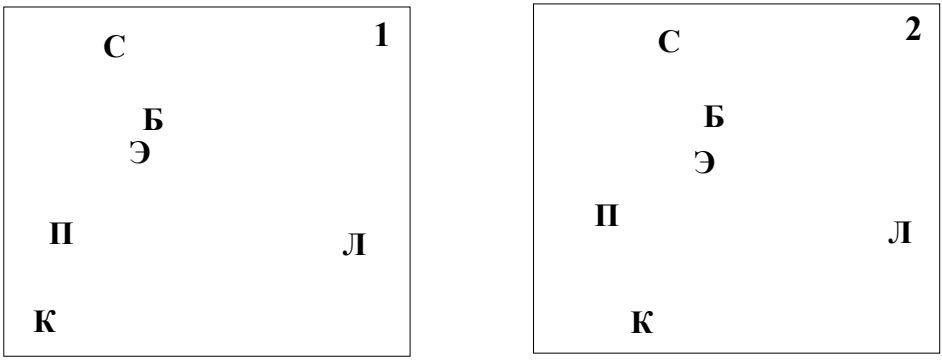


Рис. 8.3.9. Ординация биотопов в двух осях, рассчитанных методом главных компонент (1) и методом наименьших квадратов (2)

Улучшить ситуацию можно разными способами, например, сделать попытку «смыслового поворота осей» или выполнять многократную и поэтапную настройку шкал.

«Поворот осей»

Для улучшения интерпретации новых шкал можно предложить логический аналог процедуры поворота осей в факторном анализе, но с помощью иных алгоритмов. Идея метода такова: в процессе оптимизации новых координат $\boldsymbol{x}$ потребовать, чтобы они сильно коррелировали с предполагаемыми факторами изменчивости объектов (макрос Поиск решения реализует и алгоритмы линейного программирования).

В нашем примере такими факторами выступают обилие мелких млекопитающих. Обратимся к таблице, где представлены корреляции (r_{Nx1}, r_{Nx2}) между ранее рассчитанными новыми осями для ординации биотопов (x_1, x_2) и оценками численности (N, долями p)

разных групп зверьков (табл. 8.3.2). Хорошо видно, что первая главная шкала x_1 сильно и отрицательно коррелирует с оценками численности всех групп ($r_{Nx2} \approx -0.9$), кроме серых полевок – для них наблюдается высокая положительная корреляция ($r_{Nx2} = 0.9$). Вторая шкала сильно и отрицательно коррелирует с обилием мышей и бурозубок ($r_{Nx2} = -0.7$). Получается, что и первая, и вторая оси требуют «улучшения» «смысловых» свойств.

Попробуем повернуть оси так, чтобы одна из них (x_1) ориентировалась на обилие рыжих полевок (составляющих 40% от общей численности всех видов), а вторая ось (x_2) коррелировала с оценками для серых полевок, группой хоть малочисленной, но индицирующей сильную антропогенную трансформацию ландшафтов).

Таблица 8.3.2. Численность и доля групп животных в разных биотопах, две первичные главные шкалы и корреляции между ними (построена по материалам табл. 5.4.1 и рис. 8.3.8)

Биотопы	Все виды	Бурозубки	Все грызуны	Мыши	Серые полевки	Рыжие полевки	Доля бурозубок	Доля серых полевок	x_1	x_2
К	52	27	25	8	0	16	0.52	0	–0.17	–0.24
П	53	29	24	6	1	16	0.55	0.02	–0.21	–0.06
Э	39	14	25	3	3	17	0.36	0.08	0.00	0.03
С	33	11	22	1	0	20	0.33	0	–0.08	0.24
Б	38	18	20	1	2	16	0.47	0.05	0.02	0.11
Л	18	8.4	9.6	0.2	5	3.4	0.47	0.28	0.43	–0.09
r_{Nx1}	–0.9	–0.8	–0.9	–0.7	0.9	–0.9	–0.2	1.0		
r_{Nx2}	–0.3	–0.5	0.0	–0.7	–0.1	0.4	–0.7	–0.2		

Модифицируем имитационную систему (совокупность настраиваемых блоков модели) (рис. 8.3.10): слева от значений координат (x_1 и x_2) поместим показатели «доз фактора» (Ф1 – численность рыжих полевок, Ф2 – численность серых полевок). В ячейках

A25 и B25 рассчитаем корреляции между факторами и осями (Ф1 и x_1, Ф2 и x_2): A25 =КОРРЕЛ(A18:A23,D18:D23),
B25 =КОРРЕЛ(B18:B23,E18:E23).

	Ф1	Ф2		x1	x2	x3
17					1.00	0.46	0.29	1.00
17				S1 =	0.56	3.70	12.00
18	16	0	К	-0.37	-0.61	
19	16	1	П	-0.34	-0.21	
20	17	3	Э	-0.02	0.15	
21	20	0	С	-0.13	0.63	
22	16	2	Б	0.03	0.30	
23	3.4	5	Л	0.85	-0.26	
24			Σ	0	0	0
25	-0.87	-0.11	*r*		0.00	
26						
27		Stress	0.56			
28						
29						

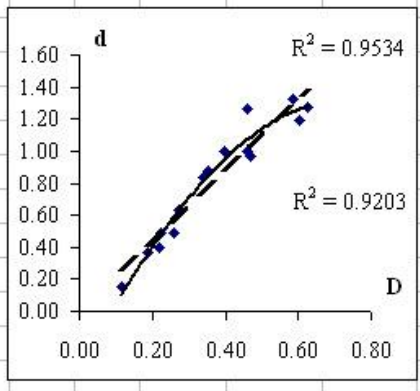


Рис. 8.3.10. Подготовка к «повороту осей»

Теперь можно вызвать программу настройки (Сервиз \ Поиск решения) и задать необходимые условия расчетов (Ограничения). Потребуем, чтобы первая ось x_1 максимально коррелировала с численностью рыжих полевок Ф1 ($A$25=1), а вторая ось x_2 – с численностью серых полевок Ф2 ($B$25=1). Чтобы поиск шел успешнее, в первый прогон следует снять ограничение на ортогональность новых осей, то есть из окна Поиск решения Удалить Ограничение $E$25=0 (остальные настройки сохраняются, рис. 8.3.11, 1) и на-

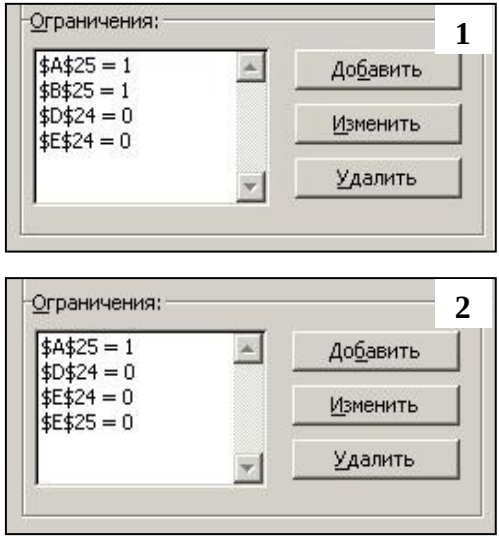


Рис. 8.3.11. Два этапа подготовки к оптимизации

жать кнопку Выполнить.

17	Ф1	Ф2		x1	x2	x3		S1 =	0.18	0.08	2.51
18	16	0	К	-0.02	-0.28						
19	16	1	П	0.00	-0.13						
20	17	3	Э	0.09	0.12						
21	20	0	С	0.23	-0.10						
22	16	2	Б	0.06	0.06						
23	3.4	5	Л	-0.35	0.33						
24			Σ	0	0	0					
25	0.97	0.95	*r*		-0.56						
26											
27		Stress	0.34								
28											
29											

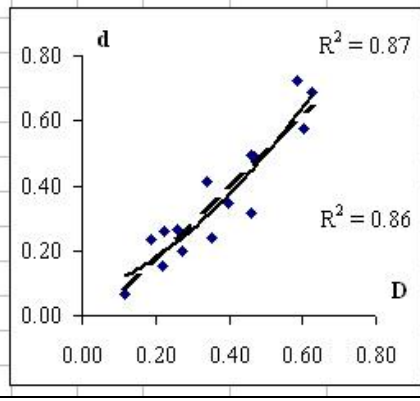


Рис. 8.3.12. Результат первичной настройки

Результаты расчетов с условием сильной корреляции между осями и внешними переменными оказались неудовлетворительными: уровень стресса получился слишком высоким ($S_1 = 0.18 > 0.1$), а оси – существенно неортогональны ($r = -0.56$) (рис. 8.3.12). Изменим условия: снимем требование ко второй оси, но вновь введем условие ортогональности (рис. 8.3.11, 2).

На этот раз уровень стресса опустился ниже критической отметки ($S_1 = 0.09 < 0.1$), то есть для описания изменчивости объектов достаточно всего два фактора. К сожалению, вторая ось x_2 утратила сильное сродство со вторым гипотетическим фактором ($r_{2x\Phi} = 0.29$) (рис. 8.3.13).

Понять содержание второй оси (и уточнить смыл первой) позволяет анализ корреляций между расчетными значениями новых шкал (x) и опорными характеристиками объектов (y) (табл. 8.3.3).

	Ф1	Ф2		x1	x2	x3
17						
18	16	0	К	0.04	-0.29	
19	16	1	П	0.11	-0.14	
20	17	3	Э	0.05	0.06	
21	20	0	С	0.2	0.2	
22	16	2	Б	0.03	0.13	
23	3.4	5	Л	-0.43	0.04	
24			Σ	0	-0	0
25	0.99	0.29	*r*		0.00	
26						
27		Stress	0.09			
28						
29						

S1 =	0.09	0.02	2.45

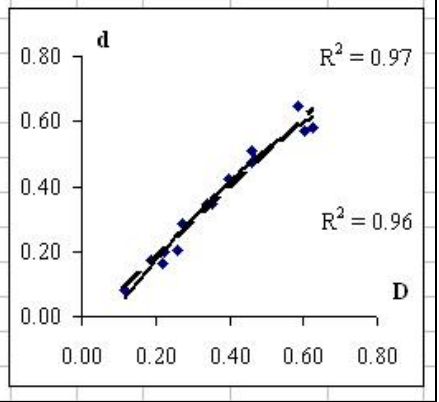


Рис. 8.3.13. Результат вторичной настройки

Таблица 8.3.3. Численность и доля групп животных в разных биотопах, две первичные главные шкалы и корреляции между ними (построена по материалам табл. 5.4.1 и рис. 8.3.13)

Биотопы	Все виды	Бурозубки	Все грызуны	Мыши	Серые полевки	Рыжие полевки	Доля бурозубок	Доля серых полевок	x_1	x_2
К	52	27	25	8	0	16	0.52	0	0.04	–0.29
П	53	29	24	6	1	16	0.55	0.02	0.11	–0.14
Э	39	14	25	3	3	17	0.36	0.08	0.05	0.06
С	33	11	22	1	0	20	0.33	0	0.20	0.20
Б	38	18	20	1	2	16	0.47	0.05	0.03	0.13
Л	18	8.4	9.6	0.2	5	3.4	0.47	0.28	–0.43	0.04
r_{Nx1}	0.7	0.4	0.9	0.4	–0.9	1.0	–0.2	–1.0		
r_{Nx2}	–0.7	–0.8	–0.3	–0.9	0.3	0.1	–0.7	0.2		

Ось x_1, как и планировалось, отражает численность рыжих полевок ($r_{Nx1} = 1$), но также и численность серых полевок (с обратным знаком). Это значит, что в область наименьших значений x_1 попадут биотопы, населенные в основном серыми полевками, а в область наибольших x_1 – биотопы, населенные только рыжими полевками (рис. 8.3.14). Таким образом, для первой оси подходит название «численность полевок».

Со второй осью наибольшие (отрицательные) корреляции имеет численность мышей ($r_{Nx2} = –0.9$) и бурозубок ($r_{Nx2} = –0.8$). Иными словами, биотопы, заселенные мышами и бурозубками, будут иметь низкие значения x_2, а местообитания, которых эти зверьки избегают, получат высокие значения показателя x_2. Вторую ось можно обозначить, как «отсутствие бурозубок и мышей».

Теперь пространственное размещение местообитаний мелких млекопитающих (рис. 8.3.14) можно охарактеризовать в немногих емких терминах. Луг отличается от основной массы исследованных местообитаний высокой численностью серых полевок и отсутствием рыжих полевок. Только для коренных местообитаний (кедровник и пихтач) характерна высокая численность мышей и бурозубок, а во вторичных стациях (сосняк, березняк) их мало.

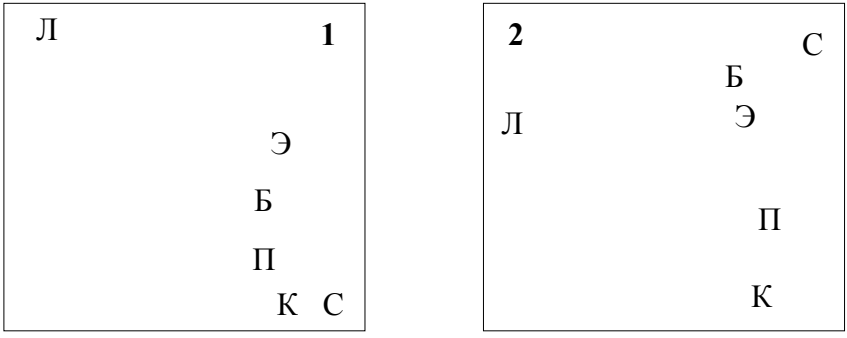


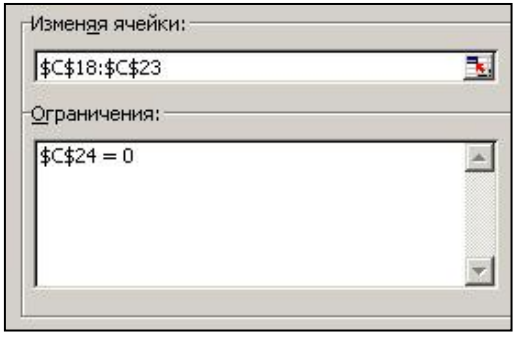
Рис. 8.3.14. Ординация биотопов в осях гипотетических факторов Ф1 и Ф2 (1) и в осях x_1 и x_2 после «поворота» (2)

Текущие результаты по сравнению с предыдущими отличаются гораздо большей определенностью интерпретации, более точным соответствием интуитивным ожиданиям от шкалирования. Взяв другие возможные факторы отличия биотических группиро-

вок, можно придти к другой ординации биотопов в новых осях. Но сама ориентация на некие определенные факторы дает возможность осям обрести точный смысл.

Поэтапное определение шкал

Расчет характеристик объектов в новых осях ***x*** имеет смысл выполнять последовательно, особенно в целях определения размерности нового пространства признаков (достаточного числа новых шкал). Вначале отыскиваются значения первой шкалы x_1, затем второй x_2 и т. д. до тех пор, пока значение стресса не снизится до уровня $S_1 = 0.1$.



Для настройки первой оси вводим значения главных координат в столбец x_1 (C18:C23; столбцы под другие шкалы очищаем), вводим формулу расчета средней (в ячейке C24), вызываем Поиск решения, добавляем Ограничения: $C$24 = 0, нажимаем кнопку Выполнить.

		x1	x2	x3
17				
18	К	-0.36		
19	П	-0.23		
20	Э	0.04		
21	С	-0.03		
22	Б	0.09		
23	Л	0.49		
24	Σ	0		
25	*r*			
26				
27	Stress	0.25		
28				
29				

S1 =	0.25	0.16	2.59

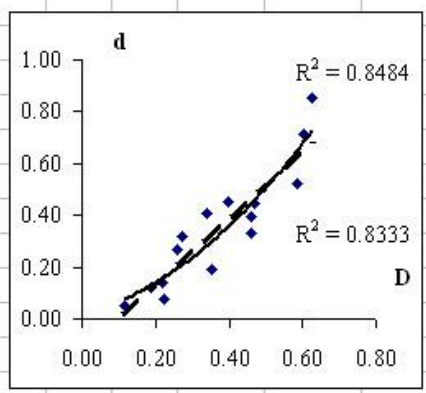
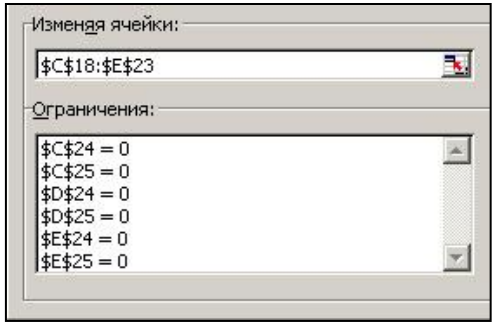


Рис. 8.3.15. Первый прогон шкалирования

Результаты пока неудовлетворительны – уровень стресса высок ($S_1 = 0.25$), матрица новых расстояний между объектами **δ** не очень хорошо соответствует матрице исходных расстояний **D** ($R^2 = 0.84$).

На втором этапе рассчитываем вторую ось x_2, которая должна быть ортогональна первой оси. Реализация этой задачи показана на рис. 8.3.7. и 8.3.8. Оценка стресса опустилась ниже уровня 0.1 и коэффициенты детерминации трендов выросли ($R^2 = 0.98$). На этом можно было бы заканчивать процедуру, но для полноты картины построим и третью ось x_3.

Для этого нужно ввести формулы расчета коэффициента корреляции между третьей осью и первой (ячейка D25), между третьей и второй (ячейка E25) и в окне Поиск решения ввести требование их независимости (условие ортогональности) D25 = 0, E25 = 0. Результат оптимизации оказался еще более точным: матрицы **δ** и **D** фактически совпали ($S_1 = 0.02$, $R^2 = 0.997$) (рис. 8.3.16).



		x1	x2	x3
К		-0.19	-0.22	0.06
П		-0.17	-0.08	-0.1
Э		-0.01	0.05	0.06
С		-0.08	0.24	0.02
Б		0.02	0.10	-0.04
Л		0.43	-0.09	-0.01
Σ		0	0	0
r		0.00	0.00	0.00
Stress		0.25	0.06	0.02

S1 =	0.02	0.00	2.43

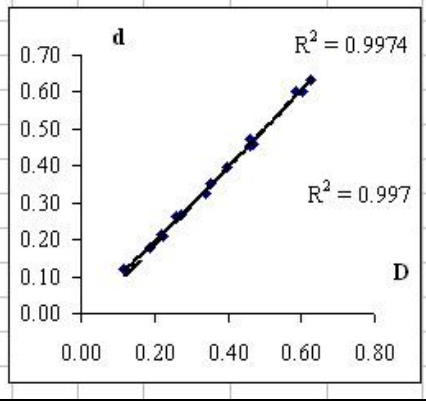


Рис. 8.3.16. Третий прогон шкалирования

Однако третья ось оказалась не доступной для интерпретации (рис. 8.3.17), в частности, среди оценок корреляции этой оси с исходными признаками не было значимых коэффициентов. Видимо, она уловила стохастический шум. Первые же две оси показали результат, аналогичный полученному прямым путем, – как по величине стресса, так и по характеру расположения объектов на их плоскости (рис. 8.3.17), даже несмотря на то, что численно значения x_1 и x_2 не совпадают с рассчитанными ранее. Причина состоит в том, что в качестве начальных величин в этих вариантах брались разные значения. Даже при небольших различиях алгоритм оптимизации будет находить разные, хотя и эквивалентные решения. Повторяющиеся результаты можно получить, если ориентироваться на некий внешний фактор, устанавливая условия для «поворота осей».

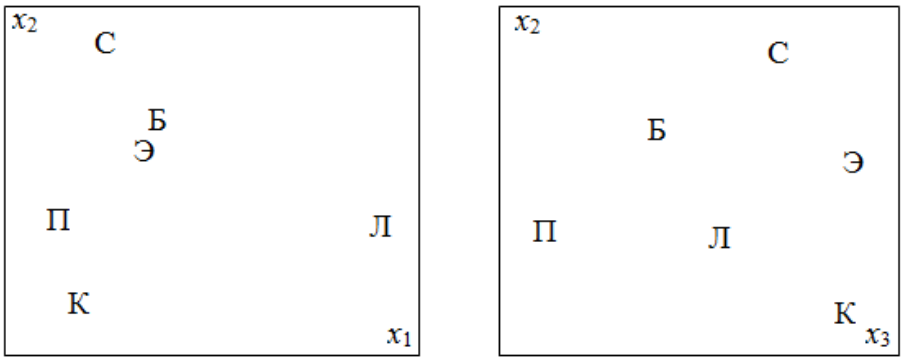


Рис. 8.3.17. Размещение биотопов в осях x_1 и x_2, x_3 и x_2

8.4. Неметрическое шкалирование

Технология метрического шкалирования строится на том предположении, что матрица исходных различий между объектами **D** определена в евклидовом пространстве и поэтому может в точности совпадать с матрицей расстояний в новых шкалах **δ** (для создания которой используется именно евклидова метрика). Такая ситуация наблюдается, если исходные характеристики объектов ***y*** выражались в интервальных или относительных шкалах, то есть когда имелась возможность инструментально и точно определить признаки сравниваемых объектов.

Однако многие эколого-биологические (особенно социально-педагогические) показатели основаны на чувственной или экспертной оценке, поэтому реальные отличия объектов и различия между их балльными оценками могут не совпадать. Например, характеризуя проективное покрытие «на глаз», легко отличить одно растение от десяти, но трудно – 91 от 100 экземпляров. Участок с одним растением получит 1 балл, с десятью – 2 балла, а с 91 и 100 растениями – одинаковый балл 5. *На левой и правой ветвях слабой порядковой шкалы одинаково сходные объекты получат разные балльные характеристики*! Отличить удовлетворительные знания учащихся от неудовлетворительных легче, чем отличные – от хороших. Насыщенность видовых списков разных территорий во многом определяется продолжительностью или масштабностью исследования. Фактор «изученность» будет вносить существенные искажения в матрицу дистанций между коллекциями. Выходом из проблемной ситуации оказывается использование сильных шкал (переход к определению вероятности обнаружения данного вида, к показателям выравненности). Другое решение состоит в попытке «исправить» расстояния между объектами, например «растянуть» короткие и «сократить» длинные дистанции. Это путь *неметрического шкалирования*.

Исправление пропорций

Предположив диспропорцию между матрицами **D** и **δ,** неметрическое шкалирование предлагает метод ее исправления – подогнать значения евклидовой матрицы **δ** к значениям неевклидовой матрицы **D** с помощью некой функции: $D_{ij} = f(\delta_{ij})$. Способы исправления диспропорций данных известны из регрессионного анализа (Ивантер, Коросов, 2003), например большие величины при логарифмировании сильнее уменьшаются, а при возведении в квадрат – сильнее увеличиваются, чем малые. Возможно конструирование и специального уравнения, преобразующего значения δ_{ij}; его параметры определяются одновременно с поиском значений новых координат $\boldsymbol{x}$. Из простых функций удобнее пользоваться параболой.

Рассмотрим пример исследования структуры биотопических группировок мелких млекопитающих, охарактеризованных присутствием видов по учетам канавками (см. табл. 5.2.1). Мерой расстояния между группировками выступила дополненная до единицы метрика Съёренсена $D_{ij} = 1 - C_{ij}$ (см. табл. 5.2.2).

С технической точки зрения базовую имитационную систему (см. рис. 8.3.6) следует дополнить блоком перерасчета матрицы **δ** с помощью полинома 2-й степени со своими параметрами (блок G20:G22) (рис. 8.4.1): $f(\delta) = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2^2$.

K5 = =G$20+G$21*J5+G$22*J5^2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**D**	К	П	Э	С	Б	Л	Г		**δ**	**f(δ)**	**D**	**(f(δ)-D)^2**	**f(δ)^2**
2	К	0.00	0.17	0.13	0.16	0.26	0.28	0.52		0.80	0.80	0.17	0.40	0.64
3	П	0.17	0.00	0.04	0.24	0.10	0.19	0.48		0.36	0.36	0.13	0.05	0.13
4	Э	0.13	0.04	0.00	0.20	0.14	0.15	0.45		0.70	0.70	0.16	0.29	0.49
5	С	0.16	0.24	0.20	0.00	0.33	0.36	0.44		0.55	0.55	0.26	0.08	0.30
6	Б	0.26	0.10	0.14	0.33	0.00	0.07	0.31		0.32	0.32	0.28	0.00	0.10
7	Л	0.28	0.19	0.15	0.36	0.07	0.00	0.25		0.52	0.52	0.52	0.00	0.27
8	Г	0.52	0.48	0.45	0.44	0.31	0.25	0.00		0.44	0.44	0.04	0.16	0.19
9														0.01
10	δ	К	П	Э	С	Б	Л							0.06
11	К													0.23
12	П	0.80												0.08
13	Э	0.36	0.44											0.12
14	С	0.70	0.10	0.34										0.04
15	Б	0.55	0.25	0.19	0.15									0.00
16	Л	0.32	0.48	0.04	0.38	0.23								0.03
17	Г	0.52	0.28	0.16	0.18	0.03	0.20							0.02
18														0.14
19	Ф1		x1	x2		**f()**								0.03
20	11	К	0.05			a0=	0.00							0.05
21	13	П	0.85			a1=	1.00			0.03	0.03	0.31	0.08	0.00
22	12	Э	0.41			a2=	0.00			0.20	0.20	0.25	0.00	0.04
23	8	С	0.75								S1 =	0.71	1.49	2.98
24	16	Б	0.6											
25	14	Л	0.37			Stress	0.71							
26	9	Г	0.57											
27		Σ	0.51	####										
28	-0.06	*r*		####										

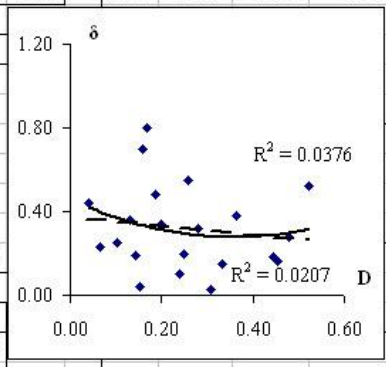
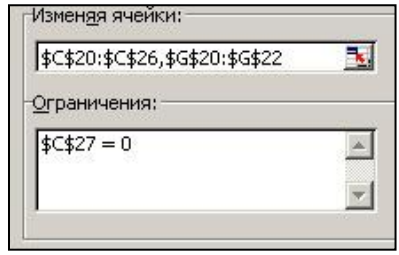


Рис. 8.4.1. Подготовка к неметрическому шкалированию

Исходно параметры задаются как 0, 1, 0. Например, [K5] =G$20+G$21*J5+G$22*J5 = 0 + 1·0.55+0·$0.55^2$ = 0.55. Вначале будем строить одну шкалу. В качестве исходных значений x_1 возьмем случайные числа (=СЛЧИС()). Для объяснения результата вычислений в графу Ф1 внесем число видов в биотопах и формулу расчета корреляции между этими значениями и первой осью:

[A28] =КОРРЕЛ(A20:A26,C20:C26).



Запускаем программу настройки и в область изменяемых параметров вносим, кроме диапазона новых переменных x_1 ($C$20:$C$26), диапазон значений параметров функции преобразования **δ** ($G$20:$G$22) (через запятую). Задаем условие обнуления средней ($C$27 = 0).

18							0.61	0.37	0.36	0.00	0.14
19	Ф1		x1	x2	f()		1.08	0.51	0.44	0.00	0.26
20	11	К	0.28		a0=	0.00	0.08	0.06	0.07	0.00	0.00
21	13	П	0.08		a1=	0.78	0.55	0.34	0.31	0.00	0.12
22	12	Э	0.11		a2=	-0.29	0.47	0.30	0.25	0.00	0.09
23	8	С	0.44					S1 =	0.11	0.02	1.74
24	16	Б	-0.1								
25	14	Л	-0.2		Stress	0.11					
26	9	Г	-0.6								
27		Σ	0.00	####							
28	-0.13	*r*		####							
29											
30											
31											
32											
33											
34											
35											

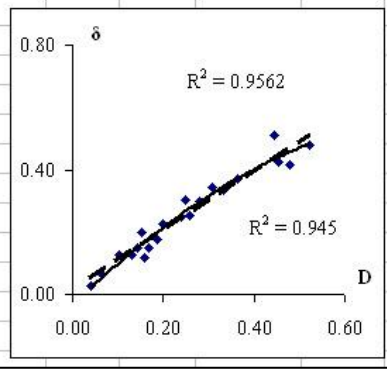


Рис. 8.4.2. Результат построения первой шкалы x_1

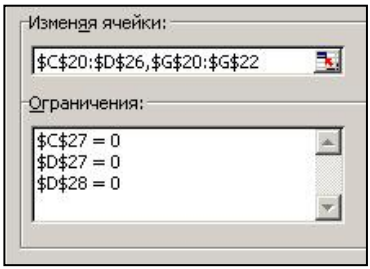
После настройки изменились как значения шкалы x_1, так и параметры полинома a_i. Например, последнее значение матрицы расстояний **δ** стало равно $\delta_{ГЛ} = 0.47$, но в пересчете по уравнению полинома обрело значение

$$f(\delta_{ГЛ}) = 0.00 + 0.78 \cdot x_1 - 0.29 \cdot x_2^2 = 0.00 + 0.78 \cdot 0.47 - 0.29 \cdot 0.47^2 = 0.30,$$

которое сблизилось со значением исходной матрицы $D_{ГЛ} = 0.25$.

Значения новой шкалы x_1 могут быть поняты как «отсутствие синантропных видов грызунов», поскольку на новой оси биотоп «город», населенный крысами, домовыми мышами и обыкновенной полевкой, занимает крайнее левое положение, а сосняк, где живут только обитатели тайги, – крайнее правое.

Величина стресса оказалась довольно большой $S_1 = 0.11$, значит, нужна настройка второй оси x_2. Зададим ее случайными числами, а в условия подгонки добавим обнуление средней и ортогональность осей ($D$27 = 0; $D$28 = 0). Запускать макрос настройки лучше три раза. Сначала следует задать только изменение осей ($C$20:$D$26). Затем задать только изменение параметров полинома ($G$20:$G$22). И лишь в третий раз через запятую включить в настройку оба блока переменных ($C$20:$D$26, $G$20:$G$22).



18								0.32	0.35	0.36	0.000	0.12
19	Ф1		x1	x2	f()			0.38	0.45	0.44	0.000	0.20
20	11	К	-0	0.16	a0=	0.04		0.05	0.07	0.07	0.000	0.00
21	13	П	0.08	0.08	a1=	0.46		0.30	0.32	0.31	0.000	0.10
22	12	Э	0.05	0.08	a2=	1.56		0.25	0.26	0.25	0.000	0.07
23	8	С	-0.2	0.1					S1 =	0.05	0.00	1.72
24	16	Б	0.1	-0.1								
25	14	Л	0.08	-0.1	Stress	0.05						
26	9	Г	-0.1	-0.3								
27		Σ	0.00	0.00								
28	0.94	*r*		0.00								
29												
30												
31												
32												
33												
34												
35												
36												

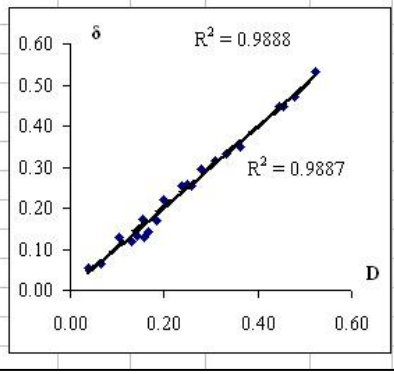


Рис. 8.4.3. Результат построения двух шкал x_1 и x_2

После расчета новых шкал стресс уменьшился до приемлемой величины $S_1 = 0.05 < 0.1$. Коэффициент корреляции между числом видов в биотопе (колонка Ф1: A19:A26) и первой осью (x_1) приблизился к единице ($r_{\Phi x1} = 0.94$); отсюда выводим ее название: «общее число видов». Не менее четкое содержание обрела и вторая ось – это «число таежных видов».

Теперь структура населения мелких млекопитающих Прибайкальской равнины легко читается (рис. 8.4.4). Коренные темно-

хвойные леса (**к**едровник) населены не очень большим числом (средние значения x_1) исключительно таежных видов (высокое x_2), многие из которых расселяются в окрестные переходные (**п**ихтач, **э**котон) и вторичные ценозы (**с**осняк, **б**ерезняк, **л**уг). Урбанизированные (**г**ород) зоны заселены небольшим числом (низкое x_1) в основном синантропных лесостепных видов (низкое x_2), которые расселяются и по окрестностям (**л**уга и **б**ерезняк). В результате такого смешения наиболее богатый видовой состав имеют соседние биотопы (**б**ерезняк и **п**ихтач с максимальными x_1), а наиболее бедный – вторичный сосняк и город.

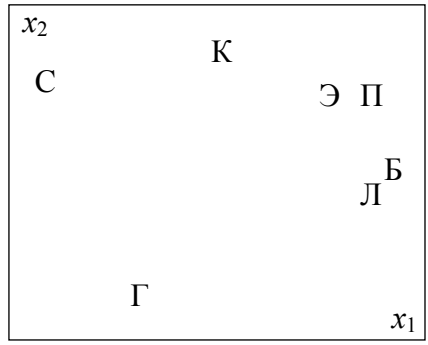


Рис. 8.4.4. Ординация биотопических группировок на плоскости двух осей, полученных неметрическим шкалированием видовых списков

Ранговая согласованность

Принятая за основу в неметрическом шкалировании формула $D_{ij} = f(\delta_{ij})$ вовсе не означает, что соотношение $f()$ между исходной и расчетной матрицей расстояний обязательно должно иметь вид уравнения, строго связывающего величины D_{ij} и δ_{ij}. Иногда матрица **D** настолько не соответствует евклидовой структуре матрицы **δ**, что от их элементов можно требовать лишь выполнения *условия порядка* (или *монотонности*): если исходное различие между объектами i и j больше, чем различие между объектами p и q, то это же соотношение должно сохраняться в новых координатах: для $D_{ij} > D_{pq}$ $\delta_{ij} > \delta_{pq}$. Иными словами, в процессе определения новых шкал необходимо следить за тем, чтобы в упорядоченных рядах значений δ и D ранги расстояний δ_{ij} по возможности совпадали с рангами расстояний D_{ij}.

Например, как ни сильно отличается каждая пара значений z_1 и z_2, их ранги (r_1 и r_2) в упорядоченных рядах совпадают.

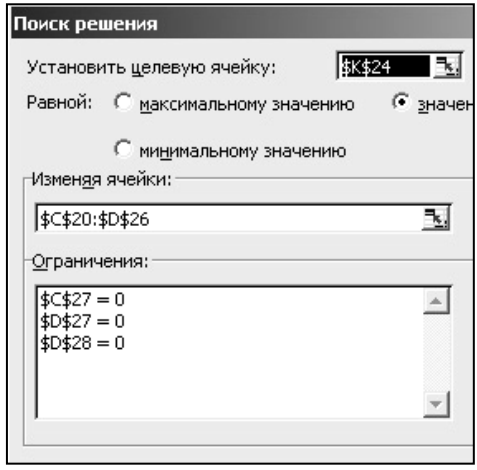
z_1	1	10	100	100	10000
z_2	1.8	2	588	589	590
r_1	1	2	3	4	5
r_2	1	2	3	4	5

Наиболее простая и удобная в использовании *мера ранговой согласованности* – это *коэффициент монотонности* (Дэйвисон, 1988): $\mu = \frac{\sum \delta_{ij} D_{ij}}{\sqrt{\sum \delta_{ij}^2 \sum D_{ij}^2}}$, на основе которого строится *коэффициент отчуждения* матриц **D** и **δ**: $k = \sqrt{1-\mu^2}$.

Как и стресс, коэффициент отчуждения тем ниже, чем более похожи сравниваемые матрицы расстояний, точнее, чем ближе ранги каждой пары δ_{ij} и D_{ij}. Мера μ есть, по сути, коэффициент корреляции между исходной и расчетной матрицами расстояний, а коэффициент k – корреляционная дистанция.

Взяв за основу предыдущую имитационную систему (рис. 8.4.1), организуем на листе Excel столбцы, где вычисляются соответствующие квадраты (δ^2 и D^2) и произведения ($\delta \cdot D$). Тогда формулы искомых коэффициентов примут вид:
[K23] =N23/КОРЕНЬ(L23*M23) и [K24] =1-K23^2.

Используя случайные числа в качестве исходных значений для осей x_1 и x_2, вызываем макрос настройки Поиск решения, задаем ограничения (нулевые суммы, ортогональность осей; также можно потребовать высокую коррелированность с контрольным фактором $A$28 = 1), запускаем программу кнопкой Выполнить.



Результаты расчетов (рис. 8.4.5, 8.4.6) оказываются практически идентичными рассмотренным выше (рис. 8.4.3, 8.4.4), хотя ось x_2 оказалась инвертирована и называется «число синантропных видов». Отличие численных результатов состоит только в том, что абсолютные значения расчетной матрицы расстояний **δ** не равны соответствующим парным значениям матрицы исходных расстояний **D**, хотя направление отличий (большие δ_{ij} соответствуют большим D_{ij}, а малые – малым) в высокой степени совпадает (коэффициент монотонности почти равен единице μ = 0.998).

K23 = =N23/КОРЕНЬ(L23*M23)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**D**	К	П	Э	С	Б	Л	Г		**δ**	**D**	**δ^2**	**D^2**	**δ*D**
2	К	0.00	0.17	0.13	0.16	0.26	0.28	0.52		0.72	0.17	0.51	0.03	0.12
3	П	0.17	0.00	0.04	0.24	0.10	0.19	0.48		0.59	0.13	0.35	0.02	0.08
4	Э	0.13	0.04	0.00	0.20	0.14	0.15	0.45		0.72	0.16	0.51	0.02	0.11
5	С	0.16	0.24	0.20	0.00	0.33	0.36	0.44		1.23	0.26	1.52	0.07	0.32
6	Б	0.26	0.10	0.14	0.33	0.00	0.07	0.31		1.41	0.28	1.99	0.08	0.40
7	Л	0.28	0.19	0.15	0.36	0.07	0.00	0.25		2.43	0.52	5.91	0.27	1.27
8	Г	0.52	0.48	0.45	0.44	0.31	0.25	0.00		0.21	0.04	0.04	0.00	0.01
9										1.22	0.24	1.49	0.06	0.29
10	δ	К	П	Э	С	Б	Л	Г		0.61	0.10	0.37	0.01	0.06
11	К									0.87	0.19	0.75	0.03	0.16
12	П	0.7								2.09	0.48	4.36	0.23	1.00
13	Э	0.6	0.2							1.02	0.20	1.04	0.04	0.20
14	С	0.7	1.2	1.0						0.65	0.14	0.42	0.02	0.09
15	Б	1.2	0.6	0.7	1.5					0.86	0.15	0.74	0.02	0.13
16	Л	1.4	0.9	0.9	1.5	0.3				2.01	0.45	4.05	0.21	0.91
17	Г	2.4	2.1	2.0	2.2	1.6	1.3			1.50	0.33	2.24	0.11	0.50
18										1.55	0.36	2.39	0.13	0.56
19	Ф1		x1	x2						2.20	0.44	4.83	0.20	0.98
20	11	К	-0.1	-0.87						0.29	0.07	0.08	0.00	0.02
21	13	П	0.4	-0.38						1.56	0.31	2.45	0.09	0.48
22	12	Э	0.19	-0.38						1.28	0.25	1.63	0.06	0.32
23	8	С	-0.8	-0.61						μ =	0.998	37.66	1.71	8.02
24	16	Б	0.44	0.22						k =	0.005			
25	14	Л	0.3	0.48										
26	9	Г	-0.4	1.55										
27		Σ	0.00	0.00										
28	0.93	r		0.00										

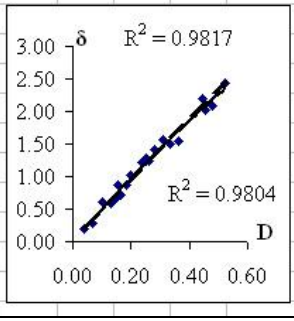


Рис. 8.4.5. Построение двух шкал по критерию ранговой согласованности

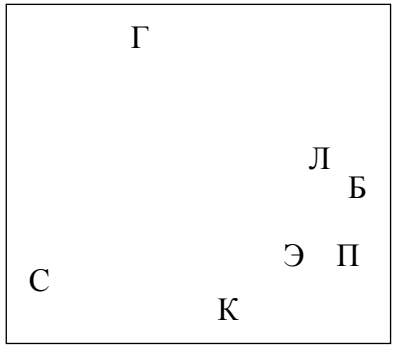


Рис. 8.4.6. Ординация биотопических группировок на плоскости двух осей, полученных неметрическим шкалированием видовых списков по критерию ранговой согласованности

Глава 9
ИЗУЧЕНИЕ РЯДОВ

Рассмотренные выше статистические методы изучают множества вариант, которые характеризуют свойства объекта исследования независимо друг от друга, при этом порядок получения вариант никак не учитывается. Однако во многих случаях информация о *последовательности появления новых значений* изучаемой случайной величины представляет большую ценность, поскольку позволяет сформировать представление как об общих закономерностях наблюдаемого явления, так и о механизмах его осуществления. Совокупность данных, расположенных в порядке получения, называется *временны**ы**м рядом*. Несмотря на название, ряд может быть сформирован и в том случае, если за шкалу отсчета взять пространственные единицы, например расстояния (так образуется *регионализированная*, или пространственная, *переменная*). Аналогично ряд можно построить, учитывая градацию другого признака, фактора среды (дозы, экспозиции). Для обозначения упорядоченных последовательностей данных используется общий термин *ряд*.

Примерами таких последовательностей могут служить наблюдения за токами сердца (кардиограмма), измерения подвижности животных (суточная активность), оценки численности популяций (волны жизни), значения спектральной яркости в полоске пикселей космического снимка (изменение отражательной способности растений и других объектов на профиле поверхности Земли), серия расстояний от центра округлого объекта до его поверхности (радиальный рельеф), показатели видового разнообразия биоценозов при разном поражении природы по мере удаления от источника выбросов (деградация, сукцессия). Одним из вариантов представления информации о природе является космический снимок, поверхность территориально упорядоченных значений, состоящая из множества рядов.

Из широкого спектра задач, решаемых методами исследования временных рядов (Дженкинс, Ваттс, 1971; Отнес, Эноксон, 1982; Максимов, Ердаков, 1985; Голиков и др., 1986; Пузаченко, 2004), мы рассмотрим следующие:

1. *Статистическая характеристика выборочного* ряда (средняя, дисперсия);
2. *Описание* монотонных направленных изменений, *трендов* (уравнения линейной и полиномиальной регрессии);
3. Выявления основных тенденций изменения рядов путем *сглаживания* (скользящая средняя, фильтр, сплайн);
4. *Выявление* однородных *областей* и *перепадов* в значениях ряда (производные, расщепляющие окна, полувариограмма);
5. *Оценка* повторяемости значений ряда, т. е. *периодичности* процесса (автокорреляция, компонентный анализ);
6. *Выделение* периодических *слагаемых* изучаемого ряда (разложение Фурье, спектральный анализ).

Для реализации этих замыслов будут использованы разнообразные алгоритмы вычислений, которые можно найти или сконструировать в пакетах обработки данных Excel, StatGraphics, Statistica.

9.1. Структура ряда

Временной ряд состоит из множества значений некой переменной величины (функции y), измеренной обычно через равные промежутки времени $\Delta = t_i - t_{i-1}$; где t_i – момент замера под номером i. Время t выступает в роли аргумента функции $y(t)$ или y_t. Функцию y часто называют *сигналом*. Единицы измерения оси абсцисс для характеристики времени выражаются в секундах, часах, годах и пр., для отображения пространства используются метры, км, пиксели снимка, для отображения градиента – конкретные концентрации, дозы. Общее количество значений ряда (объем выборки, или число *отсчетов*) составляет n (во многих руководствах обозначается буквой L).

В качестве первого примера изучим наблюдения за динамикой численности (экз./ 100 ловушко-суток) популяции рыжей полевки в Подмосковье (Европейская рыжая полевка, 1981) (рис. 9.1.1).

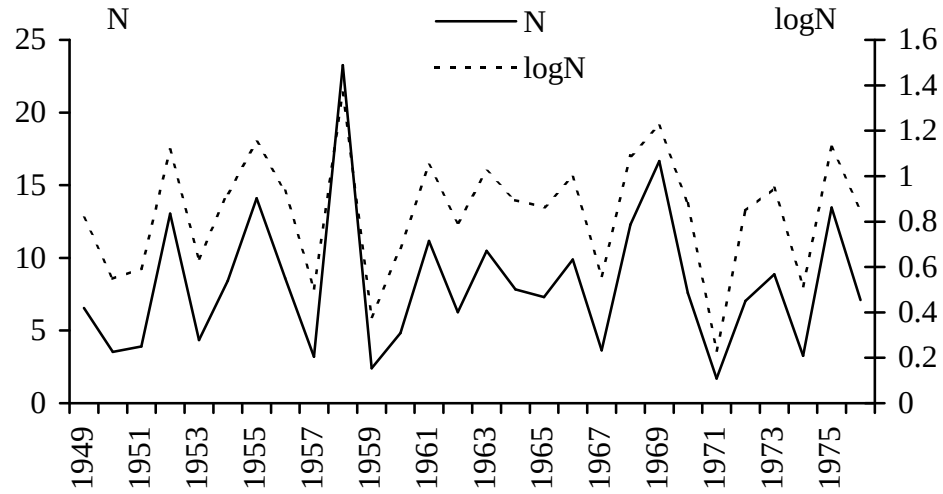


Рис. 9.1.1. Средняя годовая численность рыжей полевки

Вообще говоря, характер отношений между изучаемой переменной y и временем t может быть различным – от строго функциональной зависимости (линейный рост или периодические измерения) до абсолютно случайной вариации. Чаще наблюдается один из промежуточных вариантов, когда некое периодическое изменение сигнала y искажено трендом и размыто случайным варьированием. Какое именно значение примет переменная y в следующий момент времени, определяется той или иной вероятностью. Поэтому говорят, что y – величина случайная. Упорядоченное множество случайных величин y_i и связанных с ним распределений вероятностей называется *случайным процессом*. Естественно, что для изучения временных рядов применимы статистические процедуры.

Составить первое впечатление об особенностях наблюдаемого случайного процесса можно, рассматривая *распределение случайной величины* y (по всем значениям наблюдаемого ряда). В примере распределение оценок численности демонстрирует отчетливо выраженную правостороннюю асимметрию, что характерно для логнормального закона (п. 3.2). Поскольку базовые статистические методы предполагают нормальное распределение признака, перед обработкой рядов численностей рекомендуется их прологарифмировать (заменить y на $\log y$), взяв любое основание – 10, 2 или e. Распреде-

ление логарифмов более симметричное, гораздо ближе к нормальному (рис. 9.1.2) и допускает параметрическую обработку данных.

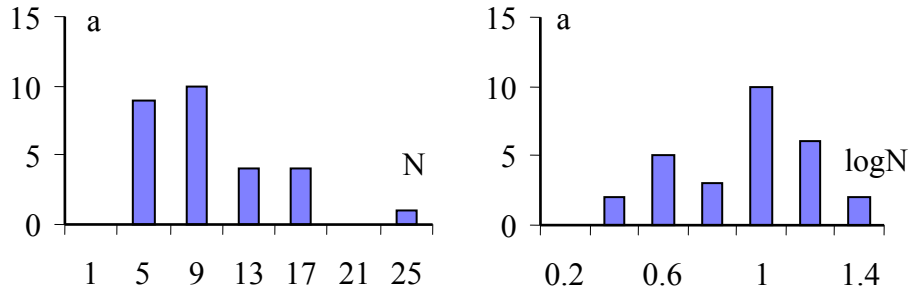


Рис. 9.1.2. Распределение частот (*a*) исходных значений численности (*N*) и их логарифмов (log*N*)

Наиболее общими и простыми характеристиками случайной величины выступают *средняя арифметическая* (*M*), которая показывает, вокруг какого значения варьирует изучаемая переменная, и *стандартное отклонение* (*S*) или *дисперсия* (S^2), характеризующее среднюю величину отклонения множества значений переменной *y* от своего среднего уровня. Большое значение в анализе имеет размах изменчивости, или *амплитуда*, – диапазон значений между минимальным и максимальным значениями величины: $A = y_{max} - y_{min}$. В нашем примере эти величины составили $M = 0.25$, $S = 0.272$, $A = 1.3$.

Для придания биологического смысла рассчитанные параметры из формы десятичного логарифма следует перевести в исходные единицы, вычислив степень для десяти: $M = 10^{0.25} = 1.78$ (эта величина фактически есть средняя геометрическая для исходных значений).

Важным этапом анализа является поиск и исправление «выскакивающих» значений, связанных обычно с ошибками наблюдений или записи сигнала. Быстро выявить артефакты позволяет сравнение средней арифметической с медианой, которая является «робастой» оценкой средней. Медиана есть значение, которое расположено посередине ранжированного (т. е. упорядоченного по величине значений) ряда, она делит его пополам (п. 4.1). В среде Excel для оценки параметра служит одноименная функция =МЕДИАНА(). Если распределение изучаемой переменной строго симметрично, то значения медианы и средней совпадут. При появлении в выборке

нескольких сильно отклоняющихся (высоких) значений средняя смещается (увеличивается), тогда как медиана почти не меняется. По величине отличий этих параметров можно судить о наличии в ряду «выскакивающих» значений и о возникающей по этой причине асимметрии распределения. Для ряда логарифмов численности медиана равна: $Me = 0.21$, ее отклонение от средней арифметической ($M = 0.25$) составляет 19%. Это указывает на повышенную долю высоких значений y, что связано, скорее всего, не с артефактами, а с методикой отлова животных. В то же время сравнение средней и медианы для исходных оценок численности ($M = 2.16$, $Me = 1.61$) дает большее расхождение – 26%; это свидетельствует о полезности процедуры логарифмирования, хотя она и не привела распределение к необходимо симметричной форме.

Модель варианты временно́го ряда

Изменение значений y временного ряда происходит по многим причинам. На ней отражаются как некие постоянно действующие факторы, обеспечивающие поддержание определенного уровня величины, так и время от времени «подключающиеся» причины, направленно смещающие значения от средней, а также множество неопределенных обстоятельств, создающих случайный «шум» в передаче сигнала. Каждое значение изучаемой величины y_i несет на себе отпечаток нескольких разнородных воздействий и поэтому может быть представлено в виде *суммы вкладов* постоянных (c), периодических (p) и случайных (r) факторов в общий результат (y_c, y_p, y_r – доли конкретного значения функции, связанные с действием разных факторов): $y_i = \sum y_c + \sum y_p + \sum y_r$.

Это выражение представляет собой общую модель варианты, отдельного значения ряда. Компонент одного типа может быть несколько (на это указывает значок Σ).

По существу, количественное *исследование временно́го ряда* направлено на то, чтобы разделить все значения на эти компоненты и тем самым оценить их роль в формировании особенностей динамики изучаемой функции y. В результате *анализа* из каждой варианты «извлекают» долю, связанную с действием того или иного фактора, поэтому весь исходный ряд распадается на множество рядов, зависящих от параллельно идущих процессов. Биологический смысл такого исследования состоит в поиске внешних и внутренних фак-

торов, ответственных за каждую из слагаемых. С технической стороны мы *из одного ряда получаем несколько рядов*, которые в сумме дают исходный (рис. 9.1.3).

Случайная компонента y_r представляет собой ряд независимых друг от друга значений и называется «*белый шум*». Частота появления тех или иных значений целиком определяется видом заданного распределения вероятностей. В биологической практике белый шум обычно имеет нормальное распределение. В некоторых ситуациях исследование «чистых» случайных рядов имеет смысл, если на разных участках ряда изменяются вид или параметры распределения изучаемой случайной величины. Однако обычная цель биометрического анализа рядов состоит в том, чтобы всеми путями избавиться от подобных флюктуаций, отфильтровать, освободить от него содержательные регулярные компоненты и тренд.

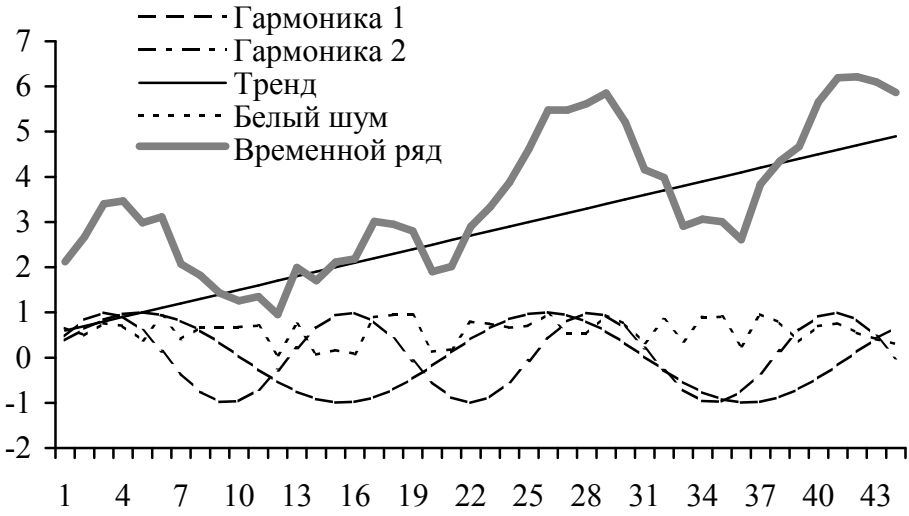


Рис. 9.1.3. Типичные компоненты временно́го ряда

Периодическая компонента y_p – это ряд, значения которого через некоторые промежутки времени (периоды, T) повторяются (причем независимо от характера их изменения в пределах этих промежутков). Это явление выражают уравнением:

$f(t) = f(t+aT)$ или $y_t = y_{t+aT}$,

где t – некий момент времени, T – *период*, отрезок времени, через который значения ряда повторяются, a – целое число, показывающее, что повторение значений происходит многократно, $f(t) = y_t$ – способы написания изучаемой функции y, зависящей от времени t.

Аналогичное условие периодичности ($y_x = y_{x+aX}$) можно сформулировать и для случая, когда ось абсцисс формируется не временем, а характеристиками пространства или выраженностью фактора x. В этом случае расстояние X между двумя смежными точками, в которых обнаруживаются пики (одинаковые значения) функции y, правильнее называть *длиной волны*, оставив термин *период* только для обозначения того временно́го интервала, через который процесс формирования функции y производит одинаковые значения. Для простоты мы не будем следовать этой демаркации.

Очищенная от стохастического шума и тренда регулярная компонента ряда может иметь различную форму, которая, как правило, определяется тем обстоятельством, что сама она представляет собой сумму (суперпозицию) нескольких слагающих ее гармоник (*гармоника* – элементарная периодическая составляющая временно́го ряда). Обычно отдельную гармонику представляют как синусоиду (или косинусоиду) в виде графика волнообразной кривой линии. Поскольку эта функция применяется для анализа рядов, рассмотрим ее характеристики подробнее.

Синус и косинус используются в тригонометрии для выражения отношений длины сторон прямоугольных треугольников. Запись $\sin\alpha$ служит для обозначения отношения длины катета, лежащего против угла α, к длине гипотенузы: $\sin\alpha = bc / ab$ (рис. 9.1.4). Множество треугольников разной формы дают ряд значений синуса, которые можно соотнести с углом α (угол измеряется между прилежащим катетом и гипотенузой).

Чем меньше угол α, то есть чем меньше противолежащий катет, тем меньше и величина синуса: $\sin0° = 0$. Напротив, при углах, близких к прямому ($\alpha = 90°$), длина противолежащего катета приближается к длине гипотенузы, $bc \approx ab$, поэтому $\sin90° = 1$. Таким образом, при увеличении угла α с 0 до 90° увеличивается и его синус с 0 до 1.

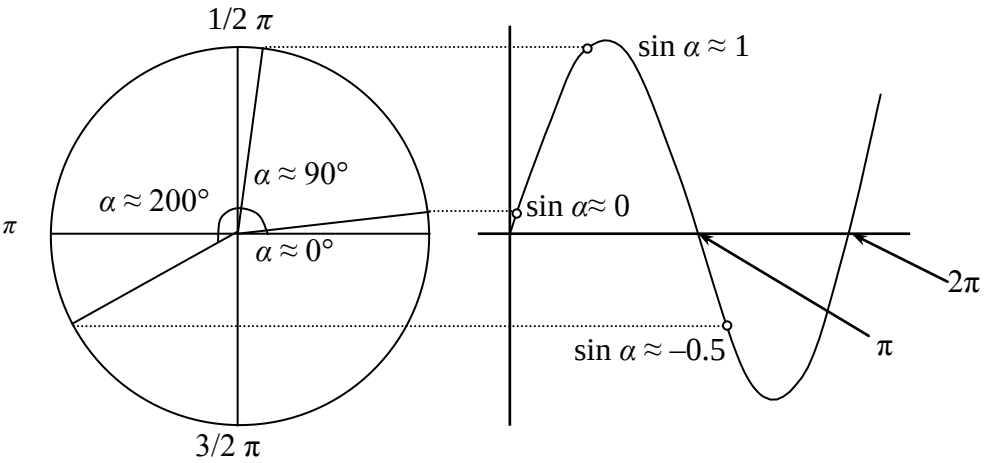


Рис. 9.1.4. Функция синуса

В прямоугольных декартовых координатах угол можно и дальше увеличивать, при этом сам треугольник как бы переходит в область отрицательных значений оси абсцисс, а синус уменьшает свои значения с 1 ($\alpha = 90°$) до 0 ($\alpha = 180°$). При дальнейшем росте угла ($\alpha > 180°$) треугольник расположится в области отрицательных значений обеих осей, поэтому значения синуса станут отрицательными, $\sin 270° = -1$, а затем, при полном обороте вокруг начала координат, вновь обнуляются: $\sin 360° = \sin 0° = 0$. Мы обрисовали ситуацию, когда гипотенуза длиной ab совершила полный оборот вокруг точки начал координат и сыграла тем самым роль радиуса, наметив линию окружности. Это значит, что для идентификации значений синуса можно пользоваться не только величиной угла α (в градусах), но и длиной соответствующей дуги окружности (в радианах) (табл. 9.1.1).

Таблица 9.1.1. Некоторые соотношения между величиной синуса, угла и длиной дуги

α (°)	0	90	180	270	360
Длина дуги (радианы)	0	$\pi/2$	π	$3\pi/2$	2π
sin	0	1	0	–1	0

При дальнейшем увеличении величины угла α (360, 450, ...) значения функции синуса будут повторяться с периодом $T = 360°$ (длина волны $T = 2\pi$) (рис. 9.1.5): $\sin\alpha = \sin(\alpha + 360°) = \sin(t + 2\pi)$.

График функции косинуса (отношения прилежащего катета к гипотенузе) имеет такую же форму и период, но смещен относительно синуса на $\pi/2$ (рис. 9.1.5).

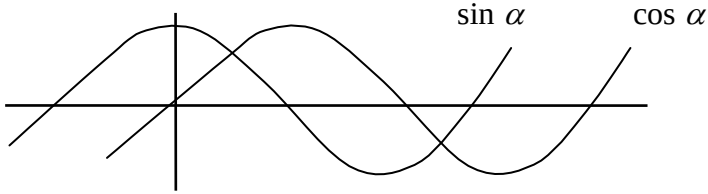


Рис. 9.1.5. Функция синуса и косинуса

Постоянные компоненты ряда y_c можно определить как доли значений признака, дающих одинаковый или монотонно изменяющийся вклад в величину *каждой* варианты ряда. Возможно существование нескольких постоянных слагаемых. Первая компонента есть *средняя арифметическая* ряда (M), величина, вокруг которой наблюдается варьирование данных. Вторая – это *тренд*, прямая или кривая линия, выражающая изменение (увеличение или уменьшение) изучаемой переменной. Линейный тренд обеспечивает однонаправленное изменение функции, криволинейные тренды меняют свое направление на разных отрезках ряда. Временно́й ряд, который не содержит трендов, то есть не изменяет своих статистических свойств во времени, называется *стационарным*. В общем случае стационарный ряд можно получить, если от исходных значений переменной отнять значения, соответствующие тренду.

9.2. Выявление тренда

Задача определить тренд означает превращение исходного ряда варьирующих значений в серию величин, образующих плавную (лучше – прямую) линию, исключив все случайные и периодические составляющие, вызывающие варьирование. Делается это для того, чтобы общая тенденция изменения значений стала очевидной и поддавалась биологической интерпретации. Найденные значения тренда можно вычесть из каждого значения временно́го ряда (тогда

из остатков сформируется стационарный ряд) и затем приступить к исследованию его периодических и случайных составляющих. Найти тренд можно разними средствами, в первую очередь, с помощью простого регрессионного анализа. Более сложные методы (гармонический анализ, спектральный анализ, компонентный анализ, имитационное моделирование) «попутно» выявляют и тренд, хотя предназначены для других целей.

Обычно генеральную тенденцию изменения величины *y* выражают с помощью регрессионного уравнения и графика *линейного тренда* вида $y' = at + b$. Достаточно быстро это можно сделать с помощью Excel (рис. 9.2.1). Вводим наши данные в столбцы A2:A29 и B2:B29 значения временно́й оси *t* и значения функции *y*. Далее строим *точечную диаграмму* (в качестве независимого признака берем *t*, а в качестве зависимого – *y*). Затем, выделив несколькими щелчками мыши точки данных на диаграмме, даем команду Добавить линию тренда (из контекстного меню или пункта главного меню Диаграмма); при этом на вкладке Параметры нужно поставить галочку в окне Показывать уравнение на диаграмме. Рассчитать значимость коэффициентов позволяет регрессионный анализ (в среде Excel запускается по команде Сервис \ Анализ данных \ Регрессия). Расчеты уравнения линейного тренда в изменении численности полевок, показали отсутствие значимости коэффициента регрессии *a* (полученное значение $t_a = 0.196$ меньше табличного $t_{(0.05,\ df\ =\ 20)} = 2.01$, уровень значимости $\alpha = 0.84 >> 0.05$). Направленных изменений в динамике популяций доказать не удалось.

Уравнение *криволинейного тренда* можно получить с помощью той же процедуры добавления линии тренда к заранее построенной диаграмме (рис. 9.2.1). Однако для оценки значимости коэффициентов приходится прибегать к внешней программе статистической обработки, например StatGraphics. У нас в уравнении параболы (полином второй степени) $y' = -0.0127 \cdot t^2 + 49.768 \cdot t - 48848$ оба коэффициента регрессии оказались незначимыми, то есть криволинейный тренд также не прослеживается.

Для отображения *однотипной* устойчивой тенденции используют полиномы не выше второй степени. Если же задача состоит в том, чтобы отобразить более частные особенности хода изучаемого процесса, можно пойти по пути увеличения длины (порядка) полинома, вводя в уравнение члены высоких степеней.

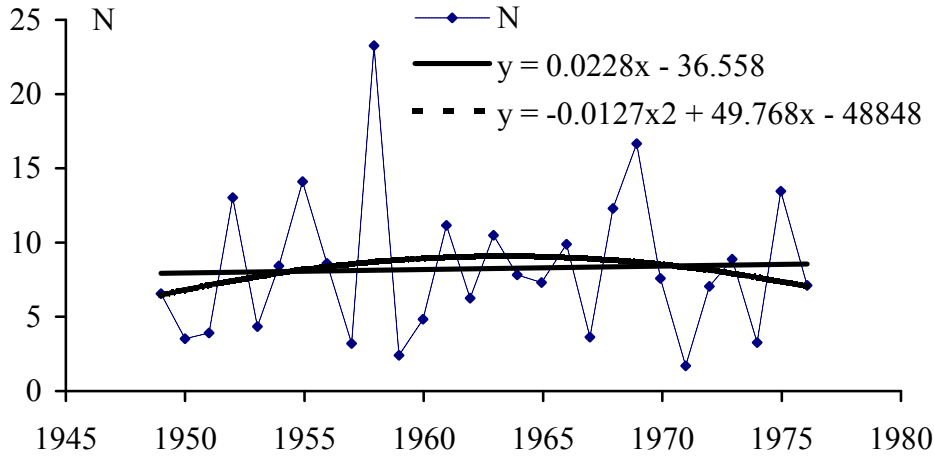


Рис. 9.2.1. Линейный и полиномиальный тренды временно́го ряда численности полевки

Чем выше степень полинома, тем более «извилистым» становится его график, учитывающий все более частные изменения изучаемой функции. Как известно, линия полинома $k-1$ степени пройдет через все точки ряда, состоящего из k вариант. При кажущейся пластичности метод имеет существенные недостатки. Полином высоких степеней невозможно биологически интерпретировать, поскольку в конструкции уравнения не заложено никаких особенных теоретических предположений; это просто математический способ *аппроксимации* (приблизительного описания) множества эмпирических точек. Если ряды достаточно длинные, то полином не годится даже для сглаживания, т. к. либо уравнение становится слишком громоздким, либо сглаживание получается слишком грубым. Для выявления частных тенденций динамики ряда и его сглаживания пользуются другими методами, например *сплайном*.

Простой плавный тренд выявляется для того, чтобы дать ему причинное биологическое (содержательное) объяснение. Обычно наличие тренда обусловлено градиентом некоего фактора среды, монотонно возрастающего (снижающегося) со временем или в пространстве. Так, рост общего антропогенного пресса на какой-либо территории (вырубки, гари, застройка, дороги, беспокойство от посещений и пр.) всегда приводит к снижению численности лесных

обитателей (при сохранении сезонной и эндогенной ритмики жизни популяций). Другим ярким примером может служить распространение промышленных выбросов. По мере удаления от точечного источника концентрация загрязнителя снижается (пространственный градиент), а при длительных наблюдениях в одной точке – возрастает (временно́й градиент). В соответствии с уровнем загрязнения будет ухудшаться и состояние биотических компонентов природы.

Объяснив тренд, можно приступать к исследованию других (периодических и случайных) составляющих временно́го ряда. При этом часто практикуется процедура вычитания тренда $y_{ост.} = y_i - y'$ из ряда, а затем выполняется *анализ остатков*.

9.3. Сглаживание и фильтрация

Помимо «жестких» линейных и криволинейных тенденций изменения функции *y* со временем, большой интерес представляют характерные черты ее плавного хода на отдельных участках ряда, «незашумленного» случайным варьированием. Предварительное выявление основных локальных тенденций, в том числе и периодичности во временно́м ряду, составляет предмет *разведочного анализа*. Основными инструментами выявления частных особенностей процесса (после вычитания тренда) служат сглаживание, фильтрация, сплайн. Аналогично тому, как общая средняя арифметическая (M), рассчитанная по всем значениям, представляет величину функции *y* для всего ряда, можно рассчитать *локальные средние* для нескольких соседних значений ряда (M_i), которые будут характеризовать значение функции в ограниченный период времени (или в ограниченной зоне пространства). Последовательно рассчитывая локальные средние для соседних участков одинаковой длины (k), мы получаем множество локальных значений M_i, которые названы *скользящими средними*.

В расчетах скользящих средних участвуют наборы из небольшого нечетного числа вариант исходного ряда (k = 3, 5, 7), результат усреднения приписывается моменту времени, соответствующего центральной варианте. Каждый новый набор из k значений получают, смещаясь от начала предыдущего набора на один временно́й шаг, т. е. исключив из прежнего набора одну левую варианту и добавив одну правую варианту.

Можно, например, получить первое значение, усредняя первые три варианты ряда $(y_1 + y_2 + y_3) / 3$, второе получают от усреднения второй, третьей и четвертой вариант: $(y_2 + y_3 + y_4) / 3$ и т. д. Сглаженные ряды скользящих средних оказываются короче, чем исходные, поскольку для расчета крайних значений (у нас y_i и y_n) не хватает информации. Общая формула определения нового объема ряда такова: $n^* = n - (k - 1)$; при сглаживании «по тройкам» имеем $n^* = 30 - (3 - 1) = 28$.

В среде Excel создать линию скользящих средних можно, если построить диаграмму ряда, выделить точки и добавить линию тренда с помощью контекстного меню, выделив на вкладке Тип картинку Линейная фильтрация (рис. 9.3.1).

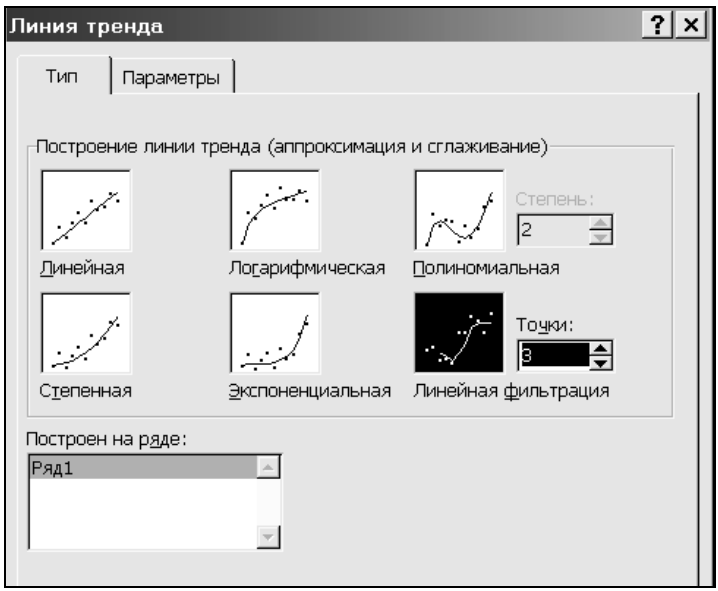


Рис. 9.3.1. Меню линии тренда в Excel

Для того чтобы подчеркнуть важность центрального момента времени на каждом интервале сглаживания, в формулы скользящих средних вводят *весовые коэффициенты*; в этом случае сумму произведений весов на значения вариант из локального набора делят на сумму весов. Распространены формулы сглаживания «по тройкам» и «по пятеркам»:

$M_i = (y_{i-1} + 3y_i + y_{i+1}) / 5$; $M_i = (y_{i-2} + 3y_{i-1} + 7y_i + 3y_{i+1} + y_{i+2}) / 15$.

Ниже показаны и другие формулы с большим числом членов. Иногда весовые коэффициенты заранее делят на их сумму, тогда они принимают вид дробей. Например, те же формулы сглаживания по тройкам и пятеркам примут следующий вид:

$M_i = 0.2 \cdot y_{i-1} + 0.6 \cdot y_i + 0.2 \cdot y_{i+1}$;

$M_i = 0.07y_{i-2} + 0.2y_{i-1} + 0.47y_i + 0.2y_{i+1} + 0.07y_{i+2}$.

Чем длиннее локальный сегмент сглаживания (окно) и чем меньше весовые коэффициенты, тем более пологий ход покажет скользящая средняя (рис. 9.3.2).

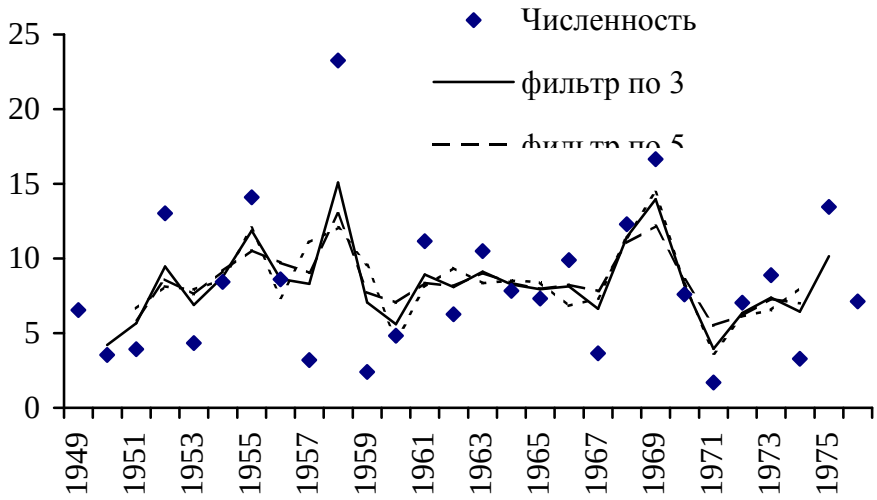


Рис. 9.3.2. Результаты сглаживания
по тройкам $M_i = 0.2 \cdot y_{i-1} + 0.6 \cdot y_i + 0.2 \cdot y_{i+1}$,
по пятеркам $M_i = 0.07y_{i-2} + 0.2y_{i-1} + 0.47y_i + 0.2y_{i+1} + 0.07y_{i+2}$
и с помощью фильтра Шеппарда–Тодда
$M_i = -0.0857y_{i-2} + 0.3428y_{i-1} + 0.4857y_i + 0.3428y_{i+1} - 0.0857y_{i+2}$

Необходимо иметь в виду, что скользящие средние M_i так специфически *искажают* исходные значения y_i, что это приводит к *смещению* сглаженного ряда вправо относительно исходного.

Прием сглаживания также носит название *фильтрация*. Термин «фильтр» позаимствован из электротехники: это аналоговые устройства, которые усиливают одни виды электромагнитных волн (полезные сигналы) и задерживают (гасят) другие виды (помехи,

шум). *Математические фильтры* также используются для выявления во временных рядах локальных «всплесков» (волны) значений функции *y*. В общем это и есть *наборы весовых коэффициентов* для расчета скользящих средних. Таков, например, известный фильтр Шеппарда–Тодда:

$M_i = (-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2}) / 35$ или

$M_i = -0.0857y_{i-2} + 0.3428y_{i-1} + 0.4857y_i + 0.3428y_{i+1} - 0.0857y_{i+2}$.

Главная особенность математических фильтров состоит в том, что весовые коэффициенты для их формул строятся, исходя из определенных теоретических соображений с помощью определенных функций. Например, фильтр Шеппарда–Тодда есть квадратичное уравнение (*полином второй степени*), натянутое на пять смежных точек (это парабола, расположенная ветвями вниз) (рис. 9.3.3).

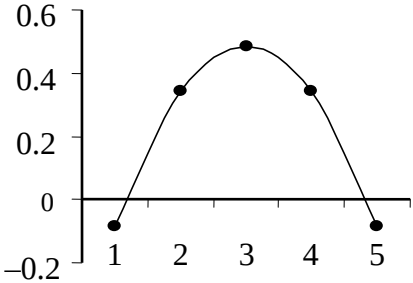


Рис. 9.3.3. Форма фильтра Шеппарда–Тодда

Двигая фильтр вдоль ряда *входных значений* со смещением на 1 шаг («прикладывая» его к последовательной череде сегментов, например, длиной 5 значений от y_{i-2} до y_{i+2}), вычисляют *выходные значения* нового ряда M_i. Функция данного фильтра состоит в усилении волн некоторого типа. Наибольшее значение *на выходе* фильтра будет получено в том случае, когда пять значений *на входе* будут по пропорциям повторять соотношения весовых коэффициентов, то есть когда форма волны в пределах данного отрезка ряда повторит форму фильтра (всплеск) (в примере это наблюдается, например, для значений за 1967–1971 гг., рис. 9.3.2). Напротив, минимальное значение фильтр выдаст в области вогнутости на графике (данные за 1969–1973 гг.). Другие соотношения смежных значений будут давать промежуточные сглаженные величины.

Важно подчеркнуть, что рассмотренный фильтр Шеппарда–Тодда усиливает лишь довольно короткие и пологие, *пятичленные*, волны. Для выявления более широких или более крутых волн предложены разнообразные фильтры, разработанные в том числе и на основе полиномов более высоких (третьей – пятой) степеней. Некоторые из них приведены в таблице 9.3.1, другие можно найти в ли-

тературе (Дэвис, 1990, с. 305). Входной сегмент ряда, который используется для расчета одного выходного сглаженного значения, называют апертурой или *окном*, а длину этого сегмента – *шириной окна*, размером апертуры, размерностью фильтра. Многие из фильтров обладают выдающимися качествами, широко используются и несут свои названия, например, фильтры Гаусса, Тьюки, Бартлетта, «ящик», «пила» и др.

Таблица 9.3.1. Полиномиальные весовые коэффициенты для сглаживания ряда

Тип полинома	Ширина окна (длина входного сегмента ряда)	Весовые коэффициенты
Квадратичный	5	-3 12 17 12 -3
Квадратичный	9	-36 9 44 69 84 89 84 69 44 9 -36
Кубический	9	15 -55 30 135 179 135 30 -55 15
Кубический	13	110 -198 -160 110 390 600 677 600 390 110 -160 -198 110

Организовать сглаживание ряда с помощью какого-либо сложного фильтра достаточно просто в среде Excel; на тех же данных по численности полевок рассмотрим применение фильтра Шеппарда–Тодда. Поместив значения численности в блок B2:B29, введем в ячейку E4 формулу
=-0.0857*B2+0.3428*B3+0.4857*B4+0.3428*B5+0.0857*B6.

Выделим мышкой ячейку E4, щелкнем дважды по черному квадратику в правом нижнем углу: произойдет автозаполнение столбца E4:E29 формулами и рассчитанными значениями (рис. 9.3.4).

	A	B	C	D	E	F	G	H	I	J
1		Числ	фильтр по	фильтр по	фильтр Шеппарда-Тодда					
2	1949	6.6								
3	1950	3.5	4.213504							
4	1951	3.9	5.662378	5.915811	=-0.0857*B2+0.3428*B3+0.4857*B4+0.3428*B5+-0.0857*B6					
5	1952	13	9.469513	8.612999	8.133752					
6	1953	4.3	6.896087	7.592867	7.921136					
7	1954	8.4	8.746077	9.163852	8.561669					
8	1955	14	11.8635	10.55889	12.03736					

Рис. 9.3.4. Расчеты сглаженного ряда в среде Excel

Возможен и другой формат организации временно́го ряда, который будет полезен для некоторых методов обработки последовательностей (рис. 9.3.5). На листе Excel поместим годы отсчетов в столбец A, ряд значений изучаемой переменной (с меткой Численность) – в столбец B; данные по численности заняли ячейки от B2 до B29. Выделим этот блок, скопируем в буфер обмена (Ctrl–C), щелкнем мышкой в ячейку C3 и вставим блок (Ctrl–V); выполним эту вставку еще 3 раза, начиная с ячеек D4, E5, F6.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Численность					фильтр Шеппарда						
2	1949	6.6											
3	1950	3.5	6.6										
4	1951	3.9	3.5	6.6									
5	1952	13	3.9	3.5	6.6								
6	1953	4.3	13	3.9	3.5	6.6	6.65						
7	1954	8.4	4.3	13	3.9	3.5	8.13						
8	1955	14	8.4	4.3	13	3.9	=-0.0857*B8+0.3428*C8+0.4857*D8+0.3428*E8-0.0857*F8						
9	1956	8.6	14	8.4	4.3	13	8.56						

Рис. 9.3.5. Расчеты сглаженного ряда по матрице, составленной из одного ряда

В результате этих действий образовался массив, состоящий из 5 вертикально расположенных рядов. При этом горизонтальный блок B6:F6 содержит первые 5 значений ряда (с 1-го по 5-е), в блок B7:F7 включена пятерка со смещением на 1: со 2-го по 6-е, и т. д.

Таблица (матрица) значений в блоке B6:F29 по-прежнему содержит все данные ряда, но «упакованные» по пятеркам. Теперь можно рассчитать сглаженные значения по другой формуле. Введем первую формулу в ячейку D6:

[D6] =-0.0857*B6+0.3428*C6+0.4857*D6+0.3428*E6-0.0857*F6.

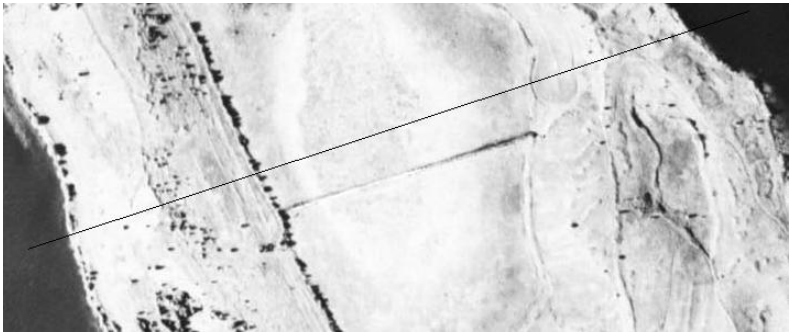
Скопируем ее в блок D7:D29 (лучше методом автозаполнения). Результаты расчетов должны совпасть с первым вариантом.

В некоторых случаях стандартных полиномиальных фильтров может оказаться недостаточно для выявления характерных пропорций между значениями ряда. Тогда фильтры специально изготавливают, находя нужные веса методом наименьших квадратов (предполагающим составление и решение уравнений) или прямой подгонкой (с помощью процедуры оптимизации).

Рассмотренный выше (п. 9.2) аппарат полиномиальной аппроксимации лежит в основе еще одного эффективного, но «безыдейного» метода сглаживания: *сплайн* – это кривая линия, огибающая экспериментальные точки (аппроксимирующая изучаемую функцию), составленная из отрезков полиномов 3–5-го порядка, последовательно построенных на коротких сегментах ряда. Сплайн построен при условии, что концы частных полиномов смыкаются, образуя непрерывную линию. Меняя степень полиномов, можно добиться большей общности (грубости) линии или большей точности (адекватности всем точкам). Кусочно-полиномиальный сплайн не имеет статистического смысла, он всего лишь позволяет ликвидировать избыточное варьирование, обеспечивая регулируемый уровень сглаживания данных и тем самым облегчая поиск содержательных тенденций изменения функции *y*. Один из простых вариантов построения сплайна состоит в том, чтобы в среде Excel построить линейную диаграмму, дважды кликнуть на линии и на вкладке Вид панели Формат ряда данных поставить галочку в поле Сглаженная линия. Также можно построить сплайн, воспользовавшись пакетом Statistica, выбрав пункт меню Graphs \ Stats 2D Graphs \ Lane Plots. Затем в появившемся окне назначить переменные Variabiles X: *t*, Y: *y*, указать Graph Type: XY Trace, FIT: Spline, OK.

9.4. Выявление однородных областей и границ

Если ряд образован значениями пространственной переменной, то очередной задачей анализа может стать выделение однородных территорий. В пределах однотипных выделов изучаемая характеристика не должна испытывать существенных изменений (отсутствуют как тренд, так и периодические флуктуации), но между территориями различного качества обнаруживаются резкие перепады значений показателя – границы. Результатом зонирования должно стать разделение последовательности данных на относительно однородные сегменты. Рассмотрим двухмерный случай прямоугольных координат, когда ось абсцисс задана расстоянием между точками пространства, ось ординат характеризует выраженность изучаемой переменной, например, значения яркости (*B*, выражена целыми числами от 0 до 255.) полоски пикселей, вырезанной из аэрофотоснимка (рис. 9.4.1).



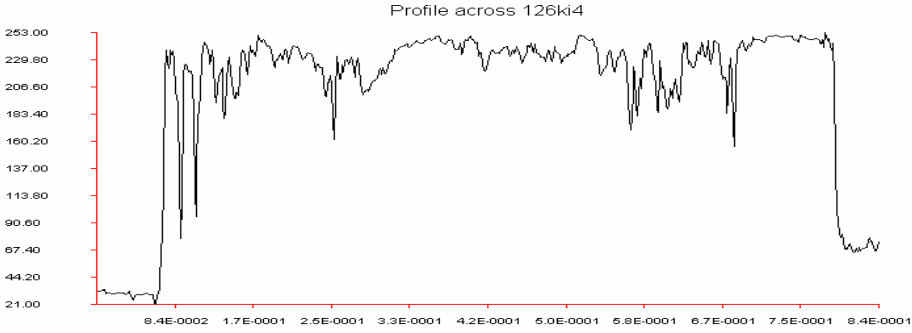


Рис. 9.4.1. Фрагмент аэрофотоснимка острова и профиль оптической яркости цепочки вырезанных пикселей

Окно производной

Один из способов отыскания границ основан на свойствах производной, которая выражает скорость изменения первообразной функции. Как известно, при относительном постоянстве значений какой-либо функции (горизонтальная линия графика), ее производная равна нулю, но скачкам изменения наклона соответствуют большие положительные (при взлетах) или отрицательные (при падениях) значения производной. Если по данным исходного ряда построить график производной, то отдельные вершины (и провалы) будут указывать на перепады значений исходной функции, то есть на границы между относительно однородными областями.

Достаточно простой путь построения подобного графика состоит в том, чтобы сглаживать исходный ряд с помощью производной какого-либо базового фильтра. Выше было показано, что одной из эффективных форм для фильтра служит парабола, позволяющая отыскивать «холмы» и «впадины» «временно́го рельефа». Производная от параболы (рис. 9.4.5) представляет прямую линию, наклоненную вправо: высокая скорость нарастания первообразной (левая ветвь параболы) постепенно снижается, пока не станет нулевой (на вершине параболы), а затем начинает возрастать скорость снижения значений параболы (правая ветвь).

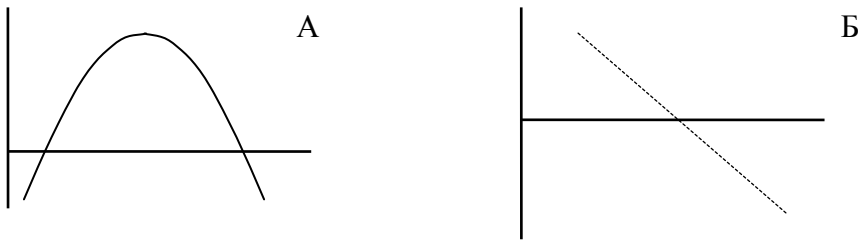


Рис. 9.4.2. Параболический фильтр (А) и производный от параболы фильтр (Б)

Что означает «получить производную фильтра»? Это значит получить весовые коэффициенты в формуле сглаживания, пропорциональные нисходящей линии. Если, например, от известного параболического пятичленного фильтра Шеппарда–Тодда

$$M_i = (-3y_{i-2} + 12y_{i-1} + 17y_i + 12y_{i+1} - 3y_{i+2}) / 35$$

взять производную, мы получим пятичленный фильтр следующего вида: $M_i = -2y_{i-2} - 1y_{i-1} + 0y_i + 1y_{i+1} + 2y_{i+2}$.

Новые веса (–2, –1, 0, 1, 2) в сумме равны нулю и не требуют деления результата на их сумму. В процессе сглаживания производный фильтр позволяет отыскивать фрагменты ряда с быстрым возрастанием значений (в этих местах *скользящая производная* принимает самые высокие значения) и с быстрым снижением значений (здесь получаются минимальные значения). График такой производной будет представлять собой неровную линию с отдельными гребнями и провалами. Перед работой рекомендуется исходный ряд предварительно сгладить скользящей средней; тогда экстремумы производной буду заметны более отчетливо.

При расширении окна фильтра отдельные всплески производной (возможно, имеющие случайную природу) могут заменяться возвышенностями с широким основанием, что более надежно характеризует перепад значений первообразной в координатах вершины. Производная параболического одиннадцатичленного фильтра имеет следующие весовые коэффициенты: –5, –4, –3, –2, –1, 0, 1, 2, 3, 4, 5. Отсюда нетрудно получить веса для разной ширины окна фильтрации.

Скорость изменения исходных значений может быть выше, чем это предусмотрено параболической производной. В таких случаях лучше пользоваться более крутыми производными кубических фильтров. Кубические фильтры имеют следующие производные веса: пятичленный: 1, –8, 0, 8, –1, одиннадцатичленный: 30, –142, –193, –126, 0, 126, 193, 142, –86. Другие варианты можно найти в литературе (Дэвис, 1990).

Анализ профиля яркости пикселей на аэрофотоснимке острова (рис. 9.4.3) показал перепады значений на границе вода / суша (точки 1, 5), луг / болото (2, 3), склон западной / восточной экспозиции (4).

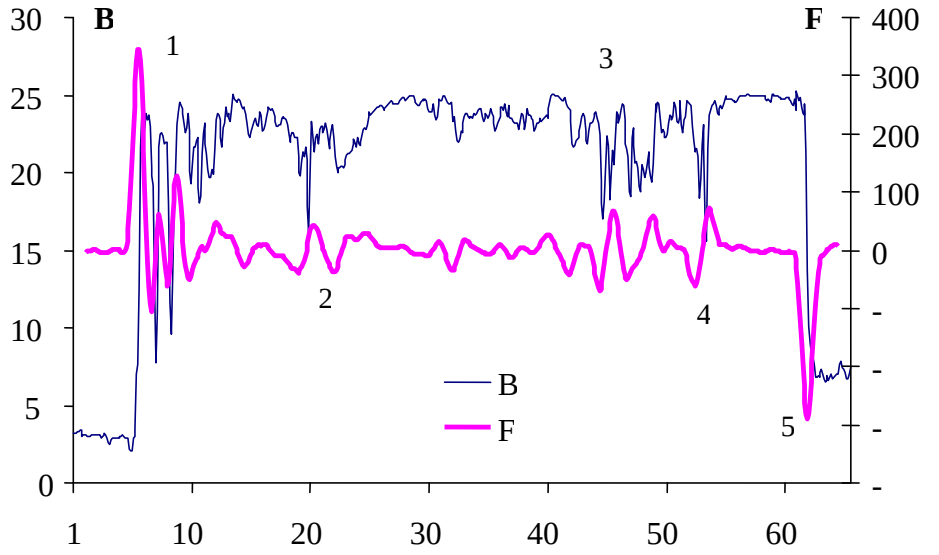


Рис. 9.4.3. Сглаживание профиля оптической яркости пикселей (*B*) с помощью 13-членного производного параболического фильтра (*F*)

Расщепляющее окно

Рассмотренный выше производный фильтр как бы состоит из двух частей: левой, содержащей в основном отрицательные члены, и правой, содержащей положительные члены. Отфильтрованное значение, приписанное центральной точке, имеет смысл количественной оценки сравнения двух половинок окна фильтрации. Эта идея в более контрастном стиле воплощена в формуле расщепляющего скользящего окна: $D^2 = \dfrac{(M_л - M_п)^2}{S_л^2 + S_п^2}$, где $M_л$, $M_п$,– средние арифметические для левого и правого сегментов ряда, $S_л$, $S_п$ – стандартные отклонения для левого и правого сегментов ряда.

При оценке параметров центральное значение окна входит и в левую и в правые части окна. Если применить формулу к данным по численности полевок (рис. 9.4.4), для 1953 г. имеем значение: $D^2 = \dfrac{(M_л - M_п)^2}{S_л^2 + S_п^2} = \dfrac{(7.1 - 8.96)^2}{5.1 + 4.9} = 0.0686.$

	A	B	C	D	E	F	G	H	I	J	K	L
1		t	N	D(3)								
2	1949	1	6.6									
3	1950	2	3.5									
4	1951	3	3.9	0.2021								
5	1952	4	13	0.0658								
6	1953	5	4.3	=((СРЗНАЧ(C4:C6)-СРЗНАЧ(C6:C8))^2)/(ДИСП(C4:C6)+(ДИСП(C6:C8)))								
7	1954	6	8.4	0.1074								
8	1955	7	14	0.0020								
9	1956	8	8.6	0.0146								

Рис. 9.4.4. Формула расщепляющего окна (n = 5) в среде Excel

Продолжая анализ поверхности острова по аэрофотоснимку, можно отметить, что расщепляющее окно более определенно, чем производный фильтр, указывает на границы более или менее однородных областей (рис. 9.4.5). В частности, сформировался пик, характеризующий излом рельефа в левой части острова (точка 2), проявилась неоднородность болота, обнаружилась его внутренняя область (двойные пики границ 3 и 4).

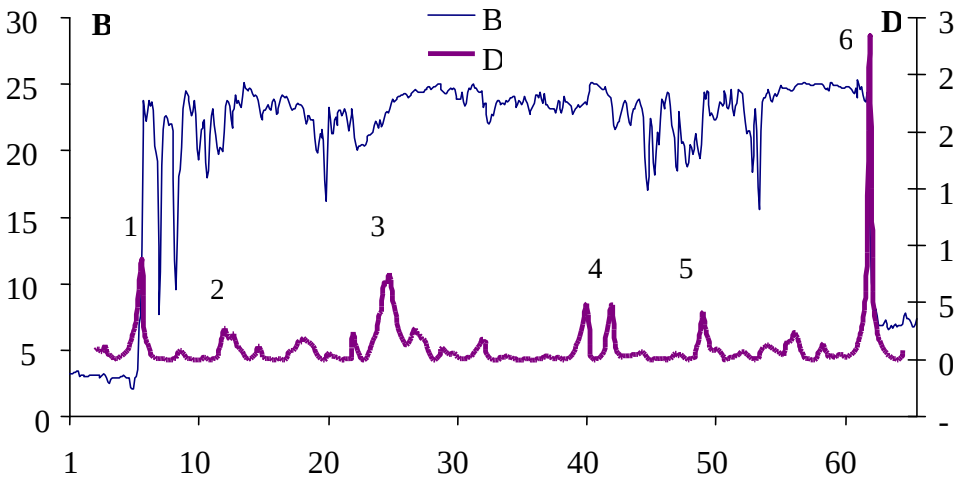


Рис. 9.4.5. Сглаживание профиля оптической яркости пикселей (B) с помощью 40-членного (n = 40) расщепляющего окна (D)

9.5. Изучение поверхности: вариограмма

Подходя к описанию периодичности, значения ряда можно рассматривать с точки зрения их сходства, близости друг к другу (автокорреляция, гармонический анализ п. 9.6–9.9), можно оценить и обобщенные различия между ними. Эта идея лежит в основе расчета *полудисперсии* (s_L), корня из суммы квадратов отклонений значений ряда, отстоящих друг от друга на один и тот же лаг (L), деленной на число сравниваемых пар (Дэвис, 1990):

$$s_L = \frac{1}{2}\sqrt{\frac{\sum_{i=1}^{n^*}(y_i - y_{i+L})^2}{n}},$$

где $n^* = n - L$ – число пар значений ряда длиной n, отстоящих друг от друга на лаг L.

Полудисперсия (она же полувариансa) характеризует не все отличия между значениями выборки (это задача дисперсии), но различия между несколькими парами значений. Величина удаления вариант называется *лагом*. При лаге $L = 0$ отыскивается разница каждого значения ряда с самим собой; естественно, $s_0 = 0$. При $L = 1$ вычисляются суммы квадратов различий между всеми парами непосредственно соседних значений: $(y_1 - y_2)^2 + (y_2 - y_3)^2 + \ldots$. При $L = 2$ берут значения через одно $(y_1 - y_3)^2 + (y_2 - y_4)^2 + \ldots$ и т. д.

В названии этого показателя приставка «полу» добавлена потому, что в форме имеется константа ½. Она имеет следующее происхождение: если рассчитать две оценки изменчивости, используя сумму квадратов отклонений вариант выборки от своей средней (стандартное отклонение) $S^2 = \frac{\sum(x_i - M)^2}{n-1}$ и используя сумму квадратов отклонений вариант друг от друга (попарное отклонение) $\Delta^2 = \frac{\sum(x_i - x_j)^2}{n}$, то эти показатели будут отличаться в два раза $S^2 = \frac{1}{2}\Delta^2$. Поскольку полудисперсия использует именно разности между вариантами, для приведения ее значений к величине дисперсии результат делится на два.

Описанным методом получают множество оценок полудисперсии, в обобщенной форме характеризующих изменение отличий значений, удаленных друг от друга на разные временные промежутки или расстояния. Графическое представление полудисперсий получило название *полувариограмма* (или вариограмма) Различают три их формы.

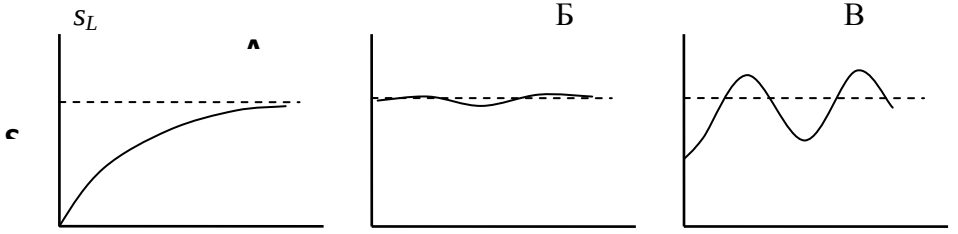
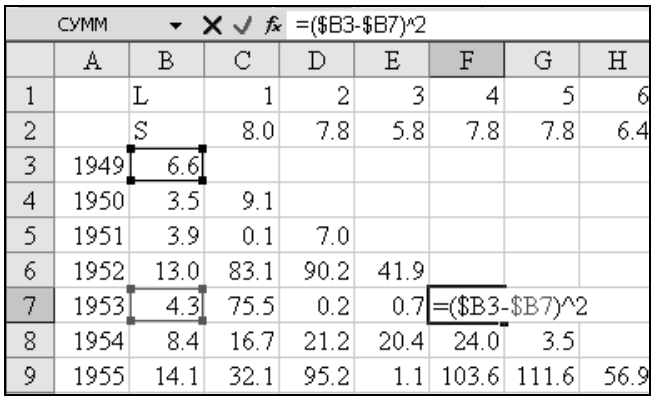


Рис. 9.5.1. Формы полувариограмм

Для однородного временно́го ряда с трендом характерно высокое сходство соседних значений (небольшие отличия) и постепенное возрастание значений полудисперсии (s_L) до уровня дисперсии всей выборки (S) (рис. 9.5.1, А). Стационарный белый шум имеет постоянно высокий уровень изменчивости ($s_L \approx S$) (рис. 9.5.1, Б). Ряд, содержащий периодические компоненты, демонстрирует регулярное изменение величины полудисперсии (рис. 9.5.1, В). По смыслу ряд полудисперсий подобен зеркальному отражению автокорреляционной функции (численно равен $1 - r_a$ при нормировании значений ряда) (см. п. 9.6). Как и автокорреляция, полудисперсия позволяет судить о структуре временно́го и пространственного рядов. Полувариограмма используется в картографировании для построения изолиний, ограничивающих однородные зоны.

Стандартные пакеты статистических расчетов, как правило, не имеют функции расчета полудисперсий. Это можно сделать в среде Excel. Внедрив исходные данные (численность полевок в Подмосковье) на электронный лист (блок A3:B30), отобразим в первой строке (B1:T1) значения лага, а вторую строку (B2:T2) зарезервируем для значений полудисперсии. Начиная с ячейки C4, вводим формулы для расчета квадрата отличий между значениями ряда C4 =(B3-B4)^2, автозаполняем остальные ячейки столбца. Смещаясь вправо и вниз на величину лага, аналогично вводим формулы для расчета квадрата отличий между значениями, разнесенными на

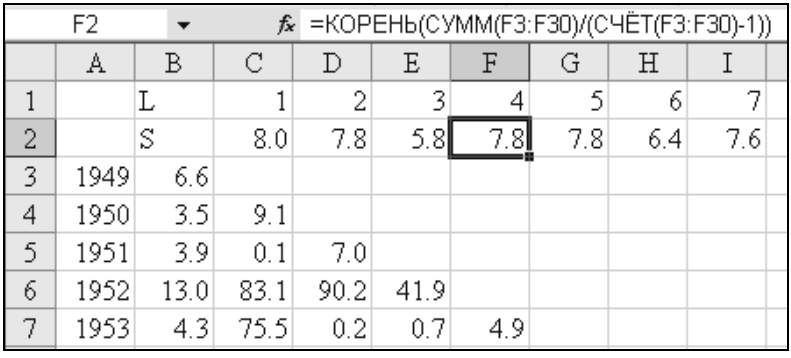
2, 3… *L* шагов (рис. 9.5.2). Для вычислений вариограммы актуальны те же ограничения, что и для автокорреляции: ряд оценок не должен быть больше двух третьих от длины всего ряда (2/3 *n*), в нашем случае он составит 2·28/ 3 ≈ 18.



	A	B	C	D	E	F	G	H
1		L	1	2	3	4	5	6
2		S	8.0	7.8	5.8	7.8	7.8	6.4
3	1949	6.6						
4	1950	3.5	9.1					
5	1951	3.9	0.1	7.0				
6	1952	13.0	83.1	90.2	41.9			
7	1953	4.3	75.5	0.2	0.7	=($B3-$B7)^2		
8	1954	8.4	16.7	21.2	20.4	24.0	3.5	
9	1955	14.1	32.1	95.2	1.1	103.6	111.6	56.9

Рис. 9.5.2. Расчет квадрата отличий для лага *L* = 4

Далее рассчитанные квадраты суммируются, делятся на число степеней свободы и из них извлекается корень (рис. 9.5.3); для *L* = 4 имеем: F2 =КОРЕНЬ(СУММ(F3:F30)/(СЧЁТ(F3:F30)-1)).



	A	B	C	D	E	F	G	H	I
1		L	1	2	3	4	5	6	7
2		S	8.0	7.8	5.8	7.8	7.8	6.4	7.6
3	1949	6.6							
4	1950	3.5	9.1						
5	1951	3.9	0.1	7.0					
6	1952	13.0	83.1	90.2	41.9				
7	1953	4.3	75.5	0.2	0.7	4.9			

Рис. 9.5.3. Расчет полудисперсии для лага *L* = 4

Наши расчеты дали вариограмму третьего типа (рис. 9.5.4), исходный ряд содержит периодические компоненты. Падение значений функции приходится на лаг 3 и 6 лет, то есть выявляется оп-

ределенный трехлетний периодизм, который на бóльших отрезках времени ослабевает.

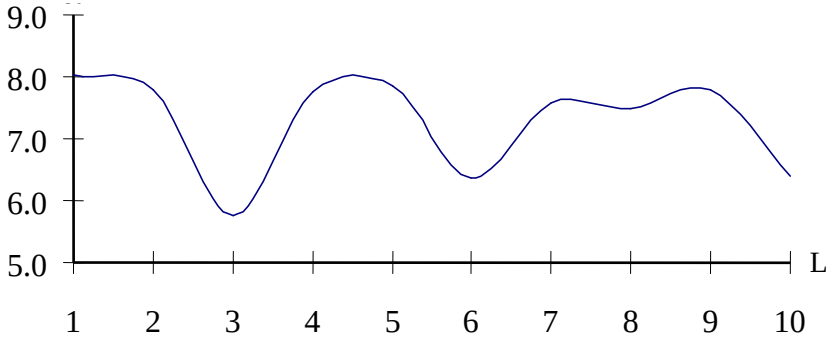


Рис. 9.5.4. Полувариограмма для ряда численности рыжей полевки

9.6. Автокорреляционный анализ

Одним из простых методов выявления регулярных составляющих временнóго ряда служит корреляционный анализ. Можно воспользоваться двумя вариантами расчетов, которые отличаются по способу формирования выборок.

В первом случае используются значения всего ряда. Как было показано выше, коэффициент корреляция рассчитывается между двумя характеристиками (x, y) и выражает силу их сопряженной изменчивости. Если в качестве таких характеристик взять множество соседних значений (y_i и y_{i+1}), то коэффициент корреляции покажет степень согласованности их изменения на протяжении всего периода наблюдений. Такой показатель называется *автокорреляцией*, поскольку оценивается взаимное сопряжение одних и тех же чисел (хотя и смещенных на один шаг). В принципе можно рассчитать коэффициент корреляции между значениями, отстоящими друг от друга не только на 1 шаг, но и на 2 (y_i и y_{i+2}), 3 (y_i и y_{i+3}) и более шагов. Для одного временнóго ряда таким образом можно получить серию коэффициентов корреляции – *автокорреляционную функцию*. С технической точки зрения массив исходных данных формируется следующим образом. Расположим значения ряда вертикально (табл. 9.6.1).

Таблица 9.6.1. Подготовка рядов для расчета автокорреляционной функции ряда численности рыжей полевки за 16 лет

	Лаг L															
Год	0		1		2		3		4		5		6		7	
1949	**10**	**10**	10		10		10		10		10		10		10	
1950	**5**	**5**	**5**	**10**	5		5		5		5		5		5	
1951	**6**	**6**	**6**	**5**	**6**	**10**	6		6		6		6		6	
1952	**22**	**22**	**22**	**6**	**22**	**5**	**22**	**10**	22		22		22		22	
1953	**6**	**6**	**6**	**22**	**6**	**6**	**6**	**5**	**6**	**10**	6		6		6	
1954	**14**	**14**	**14**	**6**	**14**	**22**	**14**	**6**	**14**	**5**	**14**	**10**	14		14	
1955	**27**	**27**	**27**	**14**	**27**	**6**	**27**	**22**	**27**	**6**	**27**	**5**	**27**	**10**	27	
1956	**12**	**12**	**12**	**27**	**12**	**14**	**12**	**6**	**12**	**22**	**12**	**6**	**12**	**5**	**12**	**10**
1957	**6**	**6**	**6**	**12**	**6**	**27**	**6**	**14**	**6**	**6**	**6**	**22**	**6**	**6**	**6**	**5**
1958	**40**	**40**	**40**	**6**	**40**	**12**	**40**	**27**	**40**	**14**	**40**	**6**	**40**	**22**	**40**	**6**
1959	**2**	**2**	**2**	**40**	**2**	**6**	**2**	**12**	**2**	**27**	**2**	**14**	**2**	**6**	**2**	**22**
1960	**10**	**10**	**10**	**2**	**10**	**40**	**10**	**6**	**10**	**12**	**10**	**27**	**10**	**14**	**10**	**6**
1961	**20**	**20**	**20**	**10**	**20**	**2**	**20**	**40**	**20**	**6**	**20**	**12**	**20**	**27**	**20**	**14**
1962	**11**	**11**	**11**	**20**	**11**	**10**	**11**	**2**	**11**	**40**	**11**	**6**	**11**	**12**	**11**	**27**
1963	**20**	**20**	**20**	**11**	**20**	**20**	**20**	**10**	**20**	**2**	**20**	**40**	**20**	**6**	**20**	**12**
1964	**14**	**14**	**14**	**20**	**14**	**11**	**14**	**20**	**14**	**10**	**14**	**2**	**14**	**40**	**14**	**6**
				14		20		11		20		10		2		40
						14		20		11		20		10		2
								14		20		11		20		10
										14		20		11		20
												14		20		11
														14		20
																14
n	16		15		14		13		12		11		10		9	
r	1.00		0.39		–0.19		**0.55**		–0.28		–0.20		0.31		–0.33	
m_r			0.26		0.28		0.25		0.30		0.33		0.34		0.36	
t			2.16		2.18		2.20		2.23		2.26		2.31		2.36	
$+ tm_r$			0.55		0.62		0.55		0.68		0.74		0.78		0.85	
$- tm_r$			–0.55		–0.62		–0.55		–0.68		–0.74		–0.78		–0.85	

Поскольку коэффициент корреляции рассчитывается между двумя столбцами значений, достроим по соседству с первым столбцом второй с тем же набором значений, $N = 16$. Вычислим коэффициент корреляции, который, естественно, будет равен единице. В следующей паре рядов сместим второй относительно первого на 1 шаг (год) вниз. Величина смешения называется *лагом* (здесь $L = 1$). Рассчитаем коэффициент корреляции по известному алгоритму. Повторим операции для возрастающих значений лага.

Стандартные статистические ошибки коэффициентов рассчитываются по формуле: $m_r = \sqrt{(1-r^2)/(n-2)}$. Для оценки значимости показателей автокорреляции строится доверительный интервал $r \pm tm_r$, где t – табличное значения распределения Стьюдента для $\alpha = 0.05$, $df = n - 2$ (используется z-преобразование, см. Ивантер, Коросов, 2003).

Важно иметь в виду, что формирование все новых выборок связано с уменьшением их размера и сопровождается снижением значимости полученных коэффициентов. Рекомендуется ограничивать минимальный объем на уровне четверти ($n_{\min} = N / 4$) и даже трети от длины исходного ряда.

Форма корреляционной функции позволяет делать выводы о периодизме изучаемого процесса и о его общих статистических свойствах. Большие положительные значения функции r свидетельствуют о большом сходстве динамики переменной в обоих рядах. На рис. 9.6.1 и по табл. 9.6.1 хорошо видно, что смещение на 3 года обеспечивает довольно высокую и значимую коррелированность ($r_3 = 0.55$, $\alpha = 0.05$). Иными словами, через каждые три года ряд достаточно точно повторяет, *копирует самого себя*, то есть периодичен, и величина этого периода составляет $T = 3$ года. В свою очередь, высокие отрицательные значения коэффициента корреляции свидетельствуют о противоположном ходе динамики, о *противофазе* (это наблюдается при смещении ряда на 1/2 периода). Для полевки наибольшие отрицательные коэффициенты вызываются смещениями на iT+1 год (1, 4, 7). Это понятно – спад численности на следующий после пика (iT) год резко ему противоположен. Значения, близкие к нулю, говорят либо об отсутствии корреляции, либо об относительном смещении рядов на 1/3 периода.

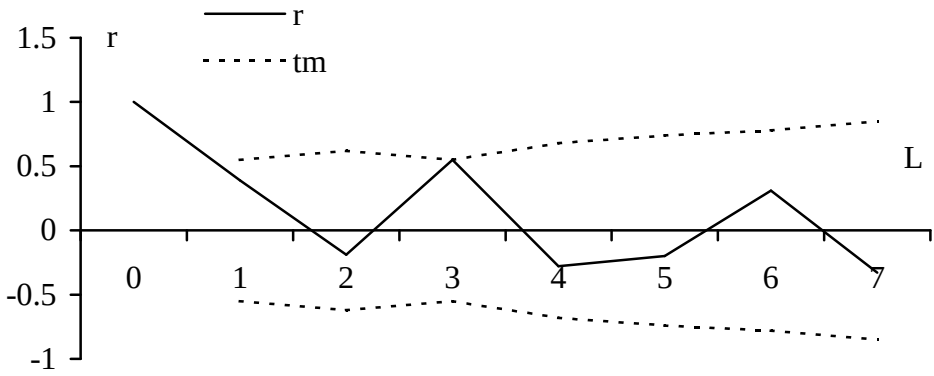


Рис. 9.6.1. Автокорреляционная функция (r) для ряда динамики численности и границы доверительной зоны (± tm)

Второй способ расчета автокорреляции аналогичен первому с той лишь разницей, что вначале выбирается небольшой (базовый) отрезок временно́го ряда определенной длины *n*, а затем коэффициенты корреляции рассчитываются между ним и соответствующим ему по длине сегментом этого ряда.

«Скольжение» (с шагом 1) базового отрезка вдоль ряда позволяет рассчитать серию коэффициентов *локальной автокорреляции*. Произвольно выберем для нашего примера базовый отрезок длиной 7 лет (это больше одной четверти от длины ряда) и выполним расчеты (табл. 9.6.2).

Автокорреляционная функция показывает ту же периодичность динамики численности, 3 года, но коэффициенты оказались много выше, чем в первом случае. Причина состоит в некотором изменении смысла полученного показателя. Если полная автокорреляционная функция отражает реализацию периодичности на протяжении большей части ряда, то локальная автокорреляция – лишь на ограниченном его отрезке. Коэффициент *r* есть показатель соответствия динамики переменной на базовом отрезке и на соответствующем интервале основного ряда. Так, высокие значения коэффициентов для смещений 3 и 6 лет ($r_3 = 0.86$, $r_6 = 0.63$) говорят о том, что локальное сродство с базовым сегментом в начале ряда много выше, чем в его конце при смещении на 9 лет, ($r_9 = 0.21$).

Таблица 9.6.2. Подготовка данных и расчет локальной функции корреляции с базовым отрезком ряда численности рыжей полевки

	Лаг L									
Год	0	1	2	3	4	5	6	7	8	9
1949	**10 10**	10	10	10	10	10	10	10	10	10
1950	**5 5**	**5 10**	5	5	5	5	5	5	5	5
1951	**6 6**	**6 5**	**6 10**	6	6	6	6	6	6	6
1952	**22 22**	**22 6**	**22 5**	**22 10**	22	22	22	22	22	22
1953	**6 6**	**6 22**	**6 6**	**6 5**	**6 10**	6	6	6	6	6
1954	**14 14**	**14 6**	**14 22**	**14 6**	**14 5**	**14 10**	14	14	14	14
1955	**27 27**	**27 14**	**27 6**	**27 22**	**27 6**	**27 5**	**27 10**	27	27	27
1956	12	**12 27**	**12 14**	**12 6**	**12 22**	**12 6**	**12 5**	**12 10**	12	12
1957	6	6	**6 27**	**6 14**	**6 6**	**6 22**	**6 6**	**6 5**	**6 10**	6
1958	40	40	40	**40 27**	**40 14**	**40 6**	**40 22**	**40 6**	**40 5**	**40 10**
1959	2	2	2	2	**2 27**	**2 14**	**2 6**	**2 22**	**2 6**	**2 5**
1960	10	10	10	10	10	**10 27**	**10 14**	**10 6**	**10 22**	**10 6**
1961	20	20	20	20	20	20	**20 27**	**20 14**	**20 6**	**20 22**
1962	11	11	11	11	11	11	11	**11 27**	**11 14**	**11 6**
1963	20	20	20	20	20	20	20	20	**20 27**	**20 14**
1964	14	14	14	14	14	14	14	14	14	**14 27**
n	7	7	7	7	7	7	7	7	7	7
r	**1.00**	–0.12	–0.43	**0.83**	–0.24	–0.60	0.63	–0.35	–0.15	0.21

Это значит, что до шестидесятых годов трехлетний периодизм динамики численности рыжей полевки проявляется отчетливо, а в начале шестидесятых произошел «сбой» этого ритма.

Для расчетов был выбран начальный семилетний отрезок ряда (от x_1 до x_7). В принципе можно было выбрать любой другой отрезок и построить аналогичную функцию локальной корреляции, которая отражала бы сходство динамики на участках ряда с другим базовым сегментом. Перебирая варианты, составим матрицу, включающую все возможные отрезки длинной $n = 7$ (табл. 9.6.3), и рассчитаем все парные коэффициенты корреляции между ними (табл. 9.6.4).

Таблица 9.6.3. Матрица для расчета всех локальных автокорреляционных функций

Лаг									
0	1	2	3	4	5	6	7	8	9
10	5	6	22	6	14	27	12	6	40
5	6	22	6	14	27	12	6	40	2
6	22	6	14	27	12	6	40	2	10
22	6	14	27	12	6	40	2	10	20
6	14	27	12	6	40	2	10	20	11
14	27	12	6	40	2	10	20	11	20
27	12	6	40	2	10	20	11	20	14

Отдельная строка в этой матрице (табл. 9.6.4) выражает сродство данного отрезка к серии последовательно перебираемых отрезков ряда. Каждый коэффициент показывает, насколько динамика численности, «захваченная» данным фрагментом, походит на динамику другого участка ряда, отстоящего от него на 1, 2 и т. д. шагов. Иными словами, каждая строка матрицы корреляций есть *локальная автокорреляционная функция*, ориентированная на «свой» базовый сегмент, смещенный относительно предыдущего на 1 шаг.

Таблица 9.6.4. Корреляции между всеми отрезками временно́го ряда (петитом выделены значимые коэффициенты, $\alpha < 0.05$)

r	0	1	2	3	4	5	6	7	8	9
0	**1.00**	–0.12	–0.43	**0.83**	–0.24	–0.60	0.63	–0.35	–0.15	0.21
1	–0.12	**1.00**	–0.16	–0.38	**0.80**	–0.28	–0.62	0.73	–0.37	–0.20
2	–0.43	–0.16	**1.00**	–0.52	–0.14	**0.78**	–0.36	–0.46	0.64	–0.51
3	**0.83**	–0.38	–0.52	**1.00**	–0.62	–0.33	0.60	–0.26	–0.21	0.29
4	–0.24	**0.80**	–0.14	–0.62	**1.00**	–0.46	–0.34	0.58	–0.29	–0.12
5	–0.60	–0.28	**0.78**	–0.33	–0.46	**1.00**	–0.51	–0.23	0.54	–0.40
6	0.63	–0.62	–0.36	0.60	–0.34	–0.51	**1.00**	–0.53	–0.22	0.53
7	–0.35	0.73	–0.46	–0.26	0.58	–0.23	–0.53	**1.00**	–0.54	–0.10
8	–0.15	–0.37	0.64	–0.21	–0.29	0.54	–0.22	–0.54	**1.00**	–0.62
9	0.21	–0.20	–0.51	0.29	–0.12	–0.40	0.53	–0.10	–0.62	**1.00**

Поскольку наибольшее сходство у каждого фрагмента с самим собой, в матрице корреляций мы наблюдаем диагональный ряд высоких значений $r_0 = 1.00$. Следующие «волны» сходств проявляются через два года на третий (r_3: 0.83, 0.80, 0.78, 0.60, 0.58, 0.54, 0.53), на шестой год (r_3: 0.63, 0.73, 0.64, 0.29). Матрица корреляций уже выявляет не только две явные волны трехлетней цикличности процесса, но и тенденции в степени выраженности этой периодичности в разных частях временно́го ряда: она более четко выражена в первой его половине (до середины пятидесятых годов) и «размывается» во второй половине (начало шестидесятых). Причины регулярных колебаний лежат в хорошо изученном качественном преобразовании популяции на каждой фазе динамики численности в оптимальных условиях (авторегуляция), а сбой ритма связан, вероятно, с возрастающим внешним антропогенным прессом.

При внимательном рассмотрении таблицы и диаграммы заметно, что ряды автокорреляций с лагом 0 и 1 во многом противостоят друг другу, положительным коэффициентам одного ряда соответствуют отрицательные коэффициенты другого (хотя и не самые большие). Складывается впечатление, что первый ряд в большей степени выражает периодизм пиков численности, а второй – довольно регулярное чередование депрессии численности. Тогда третий ряд (с лагом $L = 2$) можно интерпретировать как повторение фазы роста популяции.

Довольно эффектным приемом визуализации результатов обработки может стать совместное отображение двух автокорреляционных функций, например, для отрезков с лагом $L = 0$ и $L = 1$ (рис. 9.6.2). Отдельная точка на диаграмме соответствует двум коэффициентам корреляции каждого сегмента ряда с двумя базовыми – первым (смещение 0) и вторым (смещение на 1 шаг). Например, корреляция первого сегмента ($L = 0$) с самим собой составляет $r_{00} = 1.00$, а со вторым $r_{01} = –0.12$; это дает самую правую точку на диаграмме (подпись: 0). Корреляция первого сегмента со вторым ($L = 1$) составляет $r_{10} = –0.12$, а второго со вторым (с самим собой) $r_{11} = 1.00$; имеем самую верхнюю точку на диаграмме (1). Корреляция первого сегмента с третьим ($L = 2$) составляет $r_{20} = –0.43$, а второго с третьим – $r_{21} = –0.16$; точка расположилась в левой части (2).

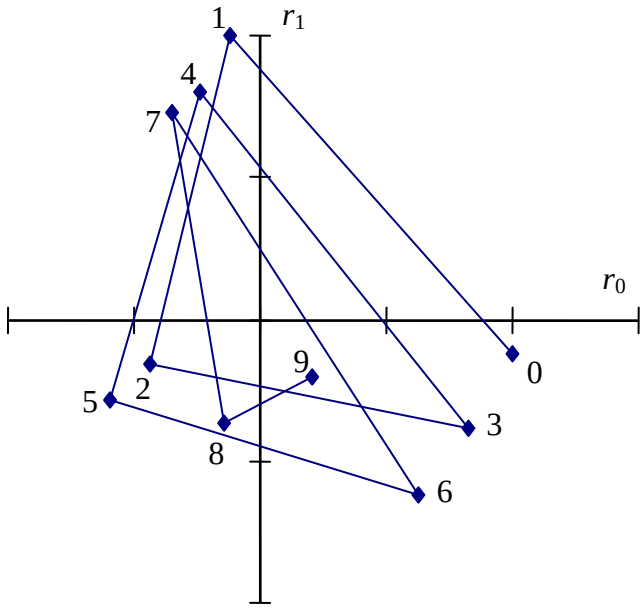


Рис. 9.6.2. Совместное изображение двух локальных автокорреляционных функций для отрезков с лагом $L = 0$ ($r0$) и $L = 1$ ($r1$); 0–9 – лаг смещения

Все множество точек довольно тесно сконцентрировалось в трех областях. Справа расположились значения, подчеркивающие сродство разных отрезков ряда с первым сегментом (для лагов 0, 3, 6, 9). Вверху сгруппировались сегменты, хорошо соответствующие второму отрезку (лаг 1, 4, 7). Слева внизу скопились фрагменты, плохо коррелирующие как с первым, так и со вторым сегментом. Поскольку таких скоплений сформировалось всего лишь три, можно говорить, что в изучаемом ряду реализовались трехлетний тип изменения численности, соответствующий качественным градациям «много, немного, немного» (или «пик, падение, рост»), и многократно повторяющийся.

Возможность декомпозиции, детализации, увеличения информации о временно́м ряде следует считать сильной стороной второго метода автокорреляции. Полученная таблица корреляций является базовой для анализа ряда с помощью компонентного анализа.

9.7. Компонентный анализ периодичности

Этим методом обрабатывается прямоугольная таблица, содержащая множество объектов (строки, всего n), численно оцененных по нескольким признакам (всего m); признакам соответствуют столбцы (Ефимов и др., 1988). Изучая периодичность, компонентный анализ использует множество равных по длине фрагментов временно́го ряда, «нарезанного» описанным выше способом (п. 9.6).

Таблица 9.7.1. Данные для расчета локальных автокорреляционных функций и главных компонент

Лаг	0	1	2	3	4	5	6	7	8	9
Год	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961
1	10	5	6	22	6	14	27	12	6	40
2	5	6	22	6	14	27	12	6	40	2
3	6	22	6	14	27	12	6	40	2	10
4	22	6	14	27	12	6	40	2	10	20
5	6	14	27	12	6	40	2	10	20	11
6	14	27	12	6	40	2	10	20	11	20
7	27	12	6	40	2	10	20	11	20	14

Компонентный анализ изучает связи между признаками. Коэффициент корреляции между фрагментами ряда (столбцами значений) есть мера их сходства, это локальная автокорреляционная функция (п. 9.6). Каждый коэффициент в строке показывает, насколько динамика численности, «захваченная» данным объектом-фрагментом, походит на динамику другого участка ряда, отстоящего от него на 1, 2 и т. д. шагов.

При анализе таблиц корреляции бросается в глаза высокая степень соответствия между фрагментами ряда, отстоящими друг от друга на 3 года, например, с лагом 0 и 3 ($r_{03} = 0.83$), 0 и 6 ($r_{03} = 0.63$), 3 и 6 ($r_{36} = 0.60$), 6 и 9 ($r_{03} = 0.53$) и др. Следует отметить, что в силу симметричности корреляционной матрицы, она одинаково читается и по горизонтали, и по вертикали. Компонентный анализ эти оценки корреляций между признаками выражает (после соответствующих вычислений) в виде факторных нагрузок – набора обобщенных ко-

эффициентов взаимной зависимости (табл. 9.7.3). Максимальные веса имеют наиболее тесно связанные признаки.

Таблица 9.7.2. Корреляций между всеми отрезками временного ряда (петитом выделены значения больше +0.5)

Лаг L	0	1	2	3	4	5	6	7	8	9
0	**1.00**	−0.12	−0.43	**0.83**	−0.24	−0.60	**0.63**	−0.35	−0.15	0.21
1	−0.12	**1.00**	−0.16	−0.38	**0.80**	−0.28	−0.62	**0.73**	−0.37	−0.20
2	−0.43	−0.16	**1.00**	−0.52	−0.14	**0.78**	−0.36	−0.46	**0.64**	−0.51
3	**0.83**	−0.38	−0.52	**1.00**	−0.62	−0.33	**0.60**	−0.26	−0.21	0.29
4	−0.24	**0.80**	−0.14	−0.62	**1.00**	−0.46	−0.34	**0.58**	−0.29	−0.12
5	−0.60	−0.28	**0.78**	−0.33	−0.46	**1.00**	−0.51	−0.23	**0.54**	−0.40
6	**0.63**	−0.62	−0.36	**0.60**	−0.34	−0.51	**1.00**	−0.53	−0.22	**0.53**
7	−0.35	**0.73**	−0.46	−0.26	**0.58**	−0.23	−0.53	**1.00**	−0.54	−0.10
8	−0.15	−0.37	**0.64**	−0.21	−0.29	**0.54**	−0.22	−0.54	**1.00**	−0.62
9	0.21	−0.20	−0.51	0.29	−0.12	−0.40	**0.53**	−0.10	−0.62	**1.00**

Таблица 9.7.3. Факторные нагрузки анализа динамики численности рыжей полевки

Год	Лаг	a_1	a_2	a_3	a_4	a_5	a_6
1952	0	**0.93**	−0.06	−0.85	0.01	0.74	0.11
1953	1	−0.44	**0.97**	−0.43	−0.09	1.00	0.25
1954	2	−0.78	−0.71	−0.03	0.50	0.98	−0.63
1955	3	**0.98**	−0.24	−0.36	−0.94	0.22	−0.07
1956	4	−0.41	**0.94**	−0.25	1.00	−0.09	−0.07
1957	5	−0.75	−0.74	0.40	−0.70	0.47	−0.11
1958	6	**1.00**	−0.28	0.11	0.76	−0.42	−0.81
1959	7	−0.32	**1.00**	0.10	−0.85	−0.62	−0.31
1960	8	−0.50	−0.84	−0.60	0.29	−0.74	1.00
1961	9	**0.75**	0.22	1.00	0.36	0.62	0.90
	S^2	3.9	3.4	1.1	0.9	0.4	0.2
	S^2, %	39	34	11	8	4	2
	$\sum S^2$, %	39	73	84	92	96	98

Как видно из таблицы 9.7.3, самые большие факторные нагрузки в первой компоненте (a_{1i}) соответствуют четырем сегментам ряда, отстоящим друг от друга на 3 года: наиболее тесно коррелируют друг с другом сегменты под номерами 0, 3, 6, 9 (для них факторные нагрузки равны $a_{10} = 0.93$, $a_{13} = 0.98$, $a_{16} = 1.00$, $a_{19} = 0.75$). То же показывали локальные автокорреляционные функции для этих сегментов в матрице корреляций (см. столбцы или строки № 0, 3, 6, 9), но факторные нагрузки выражают эту согласованность в более емкой форме – всего в одной колонке коэффициентов.

Судя по величине дисперсии первой компоненты, рассмотренные факторные нагрузки отображают около 39% имеющейся информации о варьировании оценок численности. Бóльшую часть оставшейся информации (34%) объясняет вторая главная компонента. Ее факторные нагрузки выражают высокую взаимную коррелированность других фрагментов ряда: второго, пятого и восьмого ($a_{21} = 0.97$, $a_{24} = 0.94$, $a_{27} = 1.00$); все они также отстоят друг от друга на 3 года. Отметим, что первая плеяда, отображенная первой компонентой, содержит 4 коррелирующих сегмента и она крупнее второй, имеющей 3 коррелирующих сегмента. Итак, две первые компоненты объяснили бóльшую часть общей изменчивости ряда (73%) наличием трехлетнего периодизма.

Поскольку факторные нагрузки выступают в роли коэффициентов линейных индексов, они участвуют в расчете значений главных компонент. В нашем случае стоит рассматривать лишь две компоненты ($ГК_1$ и $ГК_2$), имеющие высокие дисперсии (39 и 34%).

Таблица 9.7.4. Главные компоненты временнóго ряда динамики численности рыжей полевки −

	$ГК_1$	$ГК_2$	$ГК_3$	$ГК_4$	$ГК_5$	$ГК_6$
1	4.1	–0.2	3.2	0.0	−0.4	0.8
2	−4.9	−4.4	−0.7	1.2	−1.9	0.3
3	−2.3	5.8	0.2	−2.0	−1.2	−0.8
4	5.1	−1.8	−0.4	1.9	0.2	−1.4
5	−5.2	−3.3	0.8	−1.3	2.1	−0.3
6	−1.7	5.3	−0.9	2.5	1.1	0.7
7	4.7	−1.3	−2.3	−2.3	0.2	0.7

Для интерпретации результатов нужно вспомнить, что главные компоненты являются обобщением информации, содержащейся в исходных данных. Значения главной компоненты можно воспринимать как усредненные (но относительные) значения численности на нескольких сильно коррелирующих сегментах временно́го ряда. Первая компонента есть результат обобщения фрагментов ряда с лагом 0, 3, 6 и 9, имеющих в ней наиболее высокие нагрузки (рис. 9.7.1). Динамику численности полевок на этих отрезках можно качественно описать как всплеск, падение, всплеск, падение, всплеск. Вторая компонента – это нечто среднее между сегментами с лагом 1, 4, 7. Для них динамика выглядит немного смещенной: падение, всплеск, падение, всплеск, падение.

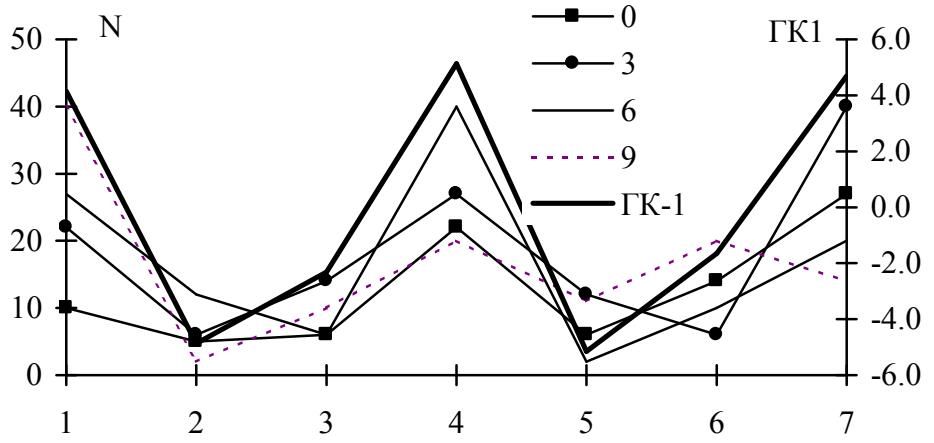


Рис. 9.7.1. Значения численности на фрагментах ряда с лагом 0, 3, 6, 9 (левая ось ординат) и значения первой главной компоненты (правая ось ординат)

По существу дела, каждая компонента представляет собой расчетный, дополнительный, *реперный* фрагмент ряда. С этой точки зрения факторные нагрузки выступают в роли коэффициента корреляции между данной компонентой и десятью сегментами ряда, а множество факторных нагрузок – суть локальная автокорреляционная функция данной компоненты (рис. 9.7.2). Здесь важно определиться, к какому моменту времени (году) относится каждое значения факторной нагрузки? Если оно имеет смысл коэффициента корреляции между компонентой и реальным фрагментом, то его следу-

ет соотносить, видимо, с центральной датой конкретного фрагмента. Например, корреляция первой компоненты с первым фрагментом (оценки численности с 1949 по 1955 гг.) оценивается значением нагрузки $a_{11} = 0.93$. Центральной датой является 1952 г., ему и приписываем значение $a_{1,1952} = 0.93$ (рис. 9.7.3).

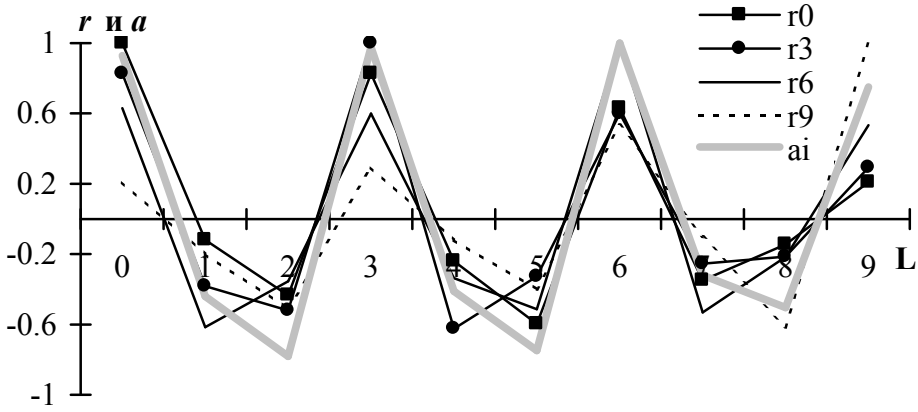


Рис. 9.7.2. Значения факторных для нагрузок первой главной компоненты (a_i) и локальных автокорреляционных функций (r) для четырех сегментов с лагом 0, 3, 6, 9

Как следует из диаграммы и таблиц, первая компонента сильно коррелирует с сегментами для лага 0, 3, 6, 9, то есть с теми, обобщением которых она является (табл. 9.7.3, рис. 9.7.1). С другими сегментами ее корреляция либо невелика, либо отрицательна (противофаза). Вторая компонента хорошо соответствует фрагментам с лагом 1, 4, 7 (табл. 9.7.3) и дает для этих точек высокие значения факторных нагрузок.

Важной особенностью компонентного анализа является его установка на выявление ортогональных (независимых) направлений изменчивости данных. Во временных рядах он выявляет последовательности значений, на которых изменения переменной осуществляются «вразнобой», не синхронно и не асинхронно, и поэтому эти ряды не коррелируют. Для нашего примера две первые компоненты имеют близкие периоды ($T \approx 3$ года), но они антисинхронны (взаимно смещены на 1/3 периода), и коэффициент корреляции между ними равен нулю. Это значит, что $ГК_1$ и $ГК_2$ несут разную информацию о временно́м ряде и могут дополнять друг друга.

Одним из эффектных приемов представления результатов анализа является «фазовый портрет» временно́го ряда в осях главных компонент или в осях факторных нагрузок (рис. 9.7.3).

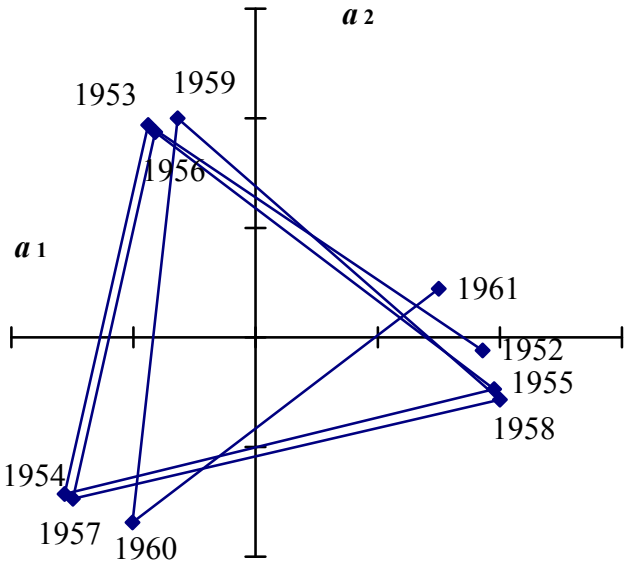


Рис. 9.7.3. «Фазовый портрет» динамики численности рыжей полевки в осях факторных нагрузок двух главных компонент

На диаграмме отдельная точка соответствует значениям факторной нагрузки в разных компонентах для одного и того же конкретного фрагмента ряда. Например, соответствие первого сектора (1949–1955 гг., центр – 1952 г.) первой компоненте оценивается величиной $a_{1,1952} = 0.93$, а его соответствие второй компоненте оценено как $a_{2,1952} = -0.06$. Соответствующая точка расположилась справа вблизи от оси абсцисс (1952). Раз диаграмма отображает корреляционные функции одновременно двух главных компонент к одним и тем же сегментам исходного ряда, можно сказать, что «фазовый портрет» есть отражение сродства к обеим главным компонентам всех последовательно перебираемых фрагментов ряда.

При этом происходит агрегация фрагментов со сходной динамикой в разных частях диаграммы. Справа собрались фрагменты, похожие на первую главную компоненту (их центры соответствуют

1952, 1955, 1958 и 1961 гг. Вверху расположились фрагменты, похожие на вторую главную компоненту с центрами в 1953, 1956, 1959 гг.). Область слева внизу занимают сегменты, не похожие ни на первую, ни на вторую компоненты (1954, 1957, 1960 гг.). Интересно, что динамика численности на фрагментах этой третьей группы довольно сильно асинхронна как динамике первой компоненты, так и второй: соответствующие факторные нагрузки варьируют от –0.5 до –0.8 (табл. 9.7.3).

На основании этой диаграммы можно прийти к тем же выводам, что были сделаны по диаграмме двух автокорреляционных функций (рис. 9.7.2), но с бóльшими основаниями, поскольку компоненты обобщают информацию по всем фрагментам. В ряду динамики численности полевки мы наблюдаем трехлетний тип ее изменения, соответствующий трем градациям («пик, падение, рост») и многократно повторяющийся.

Компонента как фильтр

Можно провести аналогию между векторами факторных нагрузок и фильтром, который усиливает совпадающие и гасит несовпадающие значения временно́го ряда.

Факторные нагрузки выступают в роли «весов» для соседних значений ряда, с помощью которых вычисляется главная компонента (аналог скользящей средней значения для каждого фрагмента):

$$ГК_l = a_1 \cdot z_1 + a_2 \cdot z_2 + ... + a_j \cdot z_j + ... + a_m \cdot z_m ,$$

где $z_{ji} = (x_i - M_j) / S_j$ – нормированное отклонение значения численности для каждого фрагмента относительно своих средних и стандартного отклонения, a_i – факторные нагрузки.

Разбитый на отрезки ряд проходит сквозь фильтры разных компонент, освобождаясь от «примесей» иных гармоник, кроме «заказанной» данным фильтром. В случае ясно выраженного тренда главная компонента обретает содержание именно как скользящей средней. Когда тренд отсутствует, первая компонента может восприниматься как усредненный ход динамики в центре ряда. Понятно, что широта охвата соседних значений «правилом сглаживания» зависит от выбранной «ширины окна» – от длины базового фрагмента. Вектор факторных нагрузок выгодно отличается от обычных жестких приемов сглаживания, поскольку формирует это привило

самостоятельно, исходя из структуры каждого отдельного ряда, подчеркивая именно его индивидуальные особенности.

Второй вектор факторных нагрузок характеризует ход динамики на каком-нибудь другом фрагменте ряда, смещенного по фазе на 1/3 периода, в силу чего между компонентами нет корреляции.

Рассмотрим последовательность этапов анализа на примере.

1) *Организация массива данных*. Изучали увеличение длины тела (мм) самки обыкновенной гадюки в возрасте 4–20 месяцев.

4	5	6	7	8	9	10	11	12	13	14	15	16	17	15	16	17
211	262	281	297	320	337	385	419	451	463	470	483	498	505	483	498	505

Из этих значений рассмотренным выше способом (см. табл. 9.6.2, 9.6.3) была сформирована матрица данных для компонентного анализа (табл. 9.7.5).

Таблица 9.7.5. Матрица, сформированная из временно́го ряда роста гадюки, и ее главные компоненты

Средняя дата	Лаг			Компоненты	
	0	1	2	$ГК_1$	$ГК_2$
5	211	262	281	–3.00	0.059
6	262	281	297	–2.50	–0.172
7	281	297	320	–2.10	–0.122
8	297	320	337	–1.80	–0.096
9	320	337	385	–1.20	0.123
10	337	385	419	–0.60	0.280
11	385	419	451	0.09	0.200
12	419	451	463	0.56	0.060
13	451	463	470	0.86	–0.109
14	463	470	483	1.06	–0.090
15	470	483	498	1.28	–0.019
16	483	498	505	1.49	–0.052
17	498	505	520	1.72	–0.039
18	505	520	532	1.93	0.009
19	520	532	540	2.14	–0.032

Объектом классификации оказывается строка из трех смежных промеров (лаги 0, 1, 2), которые уместно обозначить по срединной дате. Поэтому первая строка (локальный фрагмент ряда) получает название «5-й месяц», вторая строка – «6-й месяц» и т. д.

2) *Изучение направлений изменчивости исходных признаков*. График (рис. 7.2.1) показывает, что размеры тела увеличиваются.

3) *Выполнение расчетов*. Были получены матрицы коэффициентов корреляции, факторных нагрузок (табл. 9.7.6) и значений двух главных компонент (табл. 9.7.5; рис. 9.7.4).

4) *Изучение факторных нагрузок и ординации объектов в осях значимых главных компонент*. Все факторные нагрузки первой компоненты оказались высокими и почти равными, то есть значение первой компоненты есть примерно половина суммы трех соседних (*нормированных*) ежемесячных промеров. По существу, эта компонента представляет собой сглаженный (по тройкам) средний промер (рис. 9.7.4), он плавно возрастает. Этот положительный тренд отчетливо доминирует над остальными направлениями изменчивости данных: дисперсия (S^2) первой компоненты забирает («объясняет») 99.3% общей вариации.

Таблица 9.7.6. Факторные нагрузки двух компонент

	a_1	a_2
0	0.58	–0.70
1	0.58	–0.02
2	0.58	0.72
S^2	2.98	0.02
S^2, %	99.3	0.6

Факторные нагрузки второй компоненты противопоставляют друг другу краевые значения в тройке (a_{20} = –0.7 против a_{23} = +0.7) (табл. 9.7.6). Следовательно, значения второй компоненты есть *оценки сродства* всех фрагментов ряда *к тенденции увеличения длины тела*. Так, тройка промеров, соответствующая десятому месяцу (337, 385, 419), ярче других демонстрирует рост значений (перепад составляет 419 – 337 = 82), поэтому для этой даты имеем наиболее высокое значение второй компоненты ($ГК_{2.10}$ = 0.28, табл. 9.7.5).

Другая тройка, относящаяся к 13-му месяцу (451, 463, 470), имеет меньший перепад значений (470 – 451 = 19) и получает относительно высокое значения второй компоненты ($ГК_{2.11}$ = –0.11, табл. 9.7.5). Отслеживая динамику второй компоненты, можно видеть рост градиента в конце первого года жизни гадюки в неволе, снижение в начале второго года и последующее сохранение тенденции.

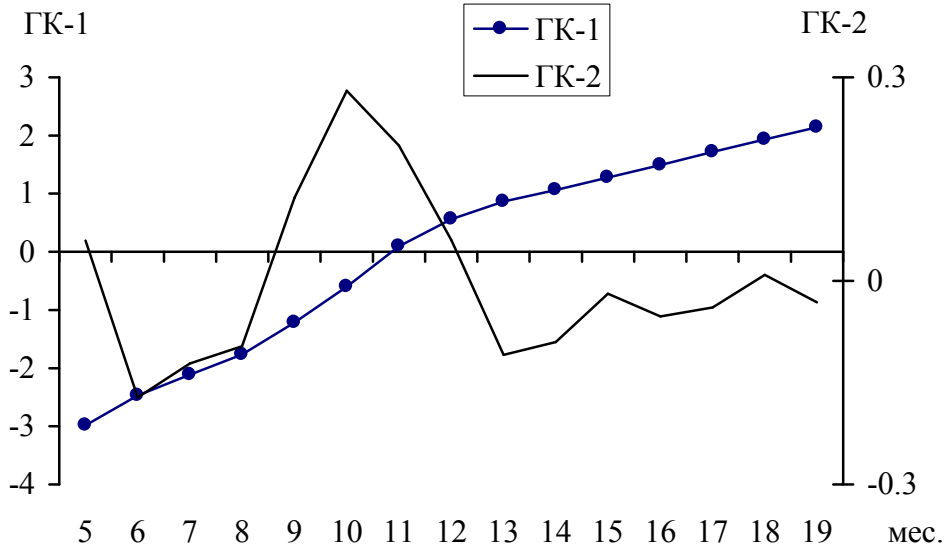


Рис. 9.7.4. Динамика двух первых компонент

5) *Присвоение названий компонентам*. Судя по результатам, первая компонента может быть названа как «средние размеры тела», а вторая – «скорость роста».

6) *Вывод об основных направлениях (факторах) изменчивости данных.* Наблюдения показывают, что с течением времени «средние размеры тела» гадюки увеличиваются, но не равномерно; «скорость роста» данной особи в конце первого года жизни повысилась, затем стабилизировалась.

Несмотря на то, что вторая компонента оказалась «слабой» (S^2 < 1%), обнаруженная плавная тенденция может иметь биологический смысл и должна быть проверена на других особях.

9.8. Анализ Фурье: вычленение гармоник

Продолжим анализ ряда динамики численности рыжей полевки (п. 9.2). После удаления избыточного стохастического шума выявлены некие колебания уровня изучаемого признака (п. 9.3). Выясним, в какой степени они периодичны.

Формализуем описание. Изучаемый признак y задан дискретно как множество отдельных значений, количество которых ограничено и равно $n = 28$ лет. Интервал (шаг) между отдельными наблюдениями составляет $\Delta = 1$ год. Общая задача состоит в том, чтобы установить повторяемость значений функции y через какой-либо период T (лет), то есть обнаружить выполнение условия $y_t = y_{t+aT}$ ($y_0 = y_{0+T} = y_{0+2T} = ...$).

С первого взгляда на диаграмму (рис. 9.2.1) заметно, что пики (и спады) численности рыжей полевки повторяются через 2–4 года: «всплески» наблюдались в 1949, 1952, 1956, 1958.... Можно предположить, что *период процесса* равен $T = 3$ (года). Цикличность можно выразить и через *частоту* $f = 1/T$, единицы измерения которой – «число периодов, приходящихся на единицу времени». Частота – это относительная единица, с помощью которой удобно сравнивать разные процессы. Для гипотетического периода 3 года частота составит $f = 1/3 = 0.3$ (год$^{-1}$), то есть за один год выполняется треть одного периода.

Самый короткий период, который только можно зафиксировать у любого ряда дискретных значений, составляет $T = 2$, то есть повторения значений случаются через 1 шаг. (Если значения повторяются на каждом шагу, то они равны друг другу в каждой точке, что соответствует прямой линии, не содержащей изменения значений функции.) Наименьшему периоду $T = 2$ соответствует *максимальная частота* $f = 1/2 = 0.5$; она названа по имени исследователя этого вопроса *частотой Найквиста*. Наибольший период, который может отразить ряд длиной n, имеет величину $T = n$; ему соответствует *минимальная частота* $f_1 = 1/n$, названная *основной частотой*. В примере основная частота могла бы составить $f_1 = 1/28 = 0.0357$; буквально это значит, что в течение одного года реализуется примерно 0.04 часть периода длиной $T_1 = 28$ лет (самого продолжительного возможного периода для данного ряда).

Обычно используются разные методы поиска регулярности в значениях ряда. В одних случаях достаточно установить наличие периодизма явления и его параметры (период, частоту), в других требуется получить конкретное уравнение, служащее для прогноза. Отметим, что любой тренд, включая линейный, рассматривается как процесс с периодом, превышающим длину имеющегося ряда.

Гармонический анализ

С помощью этого вида математических исследований можно получить серию уравнений, каждое из которых описывает какую-либо одну периодическую компоненту из объединенных в изучаемом ряду. Метод предложен Ж. Б. Фурье и носит соответствующее название – *разложение Фурье*, или *гармонический анализ*. Каждое значение эмпирического временного ряда можно представить как сумму вкладов отдельных гармонических процессов (рис. 9.8.1):

$$y = y_0 + y_1 + ... y_i + ... + y_k,$$

где y – любое значение исходного ряда, y_0 – среднее значение для всего ряда, y_1, ..., y_i, ..., y_k – вклады, которые каждая из k гармонических слагаемых вносит в каждое значение ряда, i = 1, 2,... k – номера гармоник.

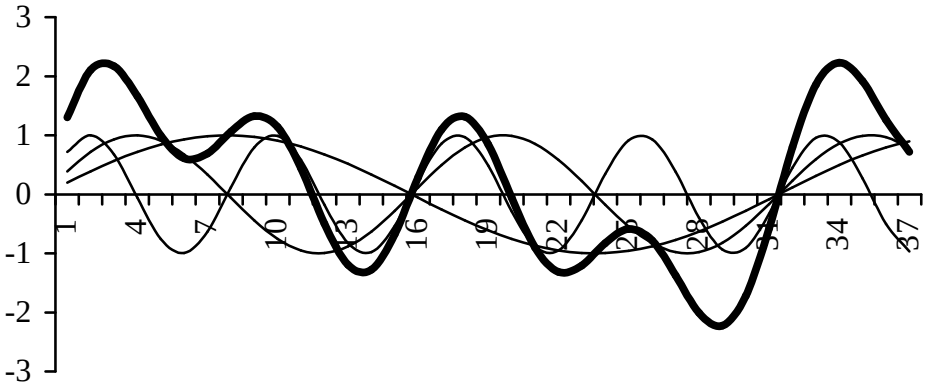


Рис. 9.8.1. Три первые гармонические слагаемые некоего гипотетического ряда

Множество значений каждой гармоники есть ряд значений, повторяющихся с правильной периодичностью в форме синусоиды с определенной частотой. Каждое слагаемое задано уравнением синусоиды: $y_i = A_i \sin(f_i t + \varphi_i)$,

где A_i – *амплитуда* колебаний, f_i – частота i-й гармоники, φ_i – *фазовый сдвиг* i-й гармоники (рис. 9.8.2).

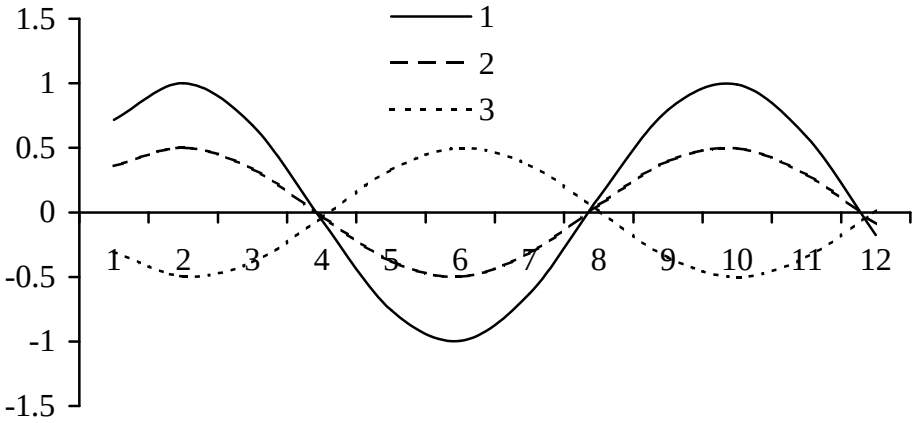


Рис. 9.8.2. Гармоники с разной амплитудой ($A_1 = 2 \cdot A_2$) и фазовым сдвигом (ряд 3 смещен относительно 1 и 2 на полупериод $T / 2$)

Добавив свободный член $y_0 = a_0 / 2$ ($a_0 = A_0 \cos\varphi_i$,) и объединив уравнения для всех гармоник, получаем формулу разложения Фурье: $y_i = a_0 / 2 + A_i \sum \sin(f_i t + \varphi_i)$.

Ортогональность гармоник

Важнейшей особенностью анализа Фурье является *ортогональность* разных периодических слагаемых. Иными словами, коэффициенты корреляции между любыми двумя гармоническими рядами (y_1 и y_2, y_1 и y_3, y_2 и y_3 и т. д.) должны быть равны нулю. Эта установка позволяет извлекать из эмпирических данных всю уникальную информацию (без дублирования в разных уравнениях) с помощью минимального числа показателей. (Аналогичное условие характерно и для других многомерных подходов, например для метода главных компонент, п. 8.1). Ортогональность обеспечивается тем, что частота каждой гармоники пропорциональна основной частоте f_1 и одновременно кратна своему номеру i: $f_i = i \cdot f_1$. Иными словами, период i-й гармоники в i раз короче наибольшего периода, равного n: $T_i = T_1 / i$. По этой причине кривые гармоник, даже начиная с одинаковых позиций, под конец ряда противостоят друг другу и не коррелируют. Для нашего примера наибольший возможный пе-

риод равен $T_1 = 28$ лет и основная частота равна $f_1 = 1/28 = 0.0357$. Отсюда находим период второй гармоники $i = 2$: $T_2 = 28/2 = 14$ лет, ее частота, соответственно, будет в два раза выше основной $f_2 = 2 \cdot f_1 = 2 \cdot 0.0357 = 0.0714$, то есть в течение одного года реализуется примерно 0.07 часть периода длиной 14 лет. Аналогично частота третьей гармоники составит $f_3 = 3 \cdot 0.0357 = 0.107$ (для периода 9.33) и т. д. Анализ Фурье позволяет изучить $k = n/2$ гармоник ограниченного ряда (в примере $28/2 = 14$).

С вычислительной точки зрения анализ Фурье состоит в отыскании двух коэффициентов A_i и φ_i в уравнении каждой гармоники $y_i = A_i \cdot \sin(f_i t + \varphi_i)$. Зная амплитуду A_i и фазовый сдвиг φ_i, можно построить график соответствующей гармоники. По существу мы отыскиваем уравнение регрессии со своими коэффициентами.

Несколько забегая вперед, следует отметить, что в поиске коэффициента A_i (амплитуды) состоит главный смысл исследования. По аналогии с регрессионным анализом, если параметр A_i оказывается достаточно большой (и статистически значимой) величиной, можно говорить о существовании соответствующей i-й гармоники, о том, что периодические изменения величины y с частотой f_i – реальность. Если же величина A_i окажется незначимо отличной от нуля, значит, периодизм с частотой f_i не характерен для изучаемого процесса (гармоника с нулевой амплитудой есть прямая линия, реализующая одно значение – среднее). Таким образом, рассчитав (и статистически оценив) значения A_i для всех гармоник, мы узнаем, какие из них действительно воплотились в структуре изучаемого ряда. Конкретный алгоритм и пример оценки значимости гармоник приведен ниже.

Расчет коэффициентов гармоник

Для удобства дальнейших расчетов в исходное уравнение $y_i = a_0/2 + A_i \sum \sin(f_i t + \varphi_i)$ вводят члены $a_i = A_i \cos\varphi_i$, $b_i = A_i \sin\varphi_i$; получаем следующее выражение: $y_i = a_0/2 + \sum(a_i \sin tf_i + b_i \cos tf_i)$.

Для расчета тригонометрических функций sin и cos прежние единицы измерения частоты («число периодов на шаг») переводятся в радианы («число циклов на шаг») путем умножения основной частоты f_1 на 2π: $f_1 \rightarrow 2\pi f_1 = \dfrac{2\pi}{n} = \dfrac{\pi}{m}$,

где m – половина выборки: $m = n/2$ для четных и $m = (n-1)/2$ для нечетных рядов.

Частоты других гармоник, кратные их номерам i, становятся равными $i\frac{\pi}{m}$. Тогда аргументы тригонометрических функций принимают следующий вид: $t\frac{i\pi}{m}$, а выражение Фурье в целом:

$$y_i = a_0/2 + \sum(a_i \sin \frac{it\pi}{m} + b_i \cos \frac{it\pi}{m}).$$

Рабочие формулы для расчета искомых значений коэффициентов гармоник определены с помощью модифицированного Бесселем метода наименьших квадратов:

$$a_0 = \frac{1}{m}\sum_{t=0}^{n-1} y_t ,$$

$$a_i = \frac{1}{m}\sum_{t=0}^{n-1} y_t \cos\frac{it\pi}{m},$$

$$b_i = \frac{1}{m}\sum_{t=0}^{n-1} y_t \sin\frac{it\pi}{m},$$

$$A_i = \sqrt{a_i^2 + b_i^2} \ , \ \varphi_i = arctg\,\frac{a_i}{b_i},$$

где m – половина выборки, t – текущий момент времени или номер варианты ряда ($t = 0, 1, ..., n–1$), y_t – значение функции в t-й момент наблюдений, i – номер гармоники ($i = 1, 2, ..., k$).

Выполнить все расчеты можно в среде Excel. Перед вычленением гармоник следует избавиться от линейного тренда, то есть из значений функции y_t вычесть значения, рассчитанные по уравнению линейной регрессии Y_t, полученного на предыдущих этапах исследования. (Это стоит сделать даже в том случае, если параметры линейного уравнения оказались незначимы). Для удобства расчетов отсчет времени следует начать с момента $t = 0$ и вести до момента $n - 1$. Значения времени можно задать как $j = \frac{t\pi}{m}$, тогда аргументы тригонометрических функций запишутся как ij.

Для выполнения вычислений организуется таблица с полями: t, $j = \frac{t\pi}{m}$, y_t, Y_t, $u_t = y_t - Y_t$, $\sin(ij)$, $\cos(ij)$, $u_t \cdot \sin(ij)$, $u_t \cdot \cos(ij)$. Таблица дает возможность определить коэффициенты любой гармоники $U_t = a_t \cdot \sin(ij) + b_t \cdot \cos(ij) + \ldots$. Рассмотрим алгоритм получения первой гармоники изучаемого ряда (табл. 9.8.1).

Как было установлено ранее, длина изучаемого ряда составляет $n = 28$; $m = n/2 = 14$. Для второго момента времени (1950 г., $t = 1$) значение численности было равно $y_t = 4.954$, по уравнению линии получена средняя численность: $Y_t = 0.0999 \cdot t - 181.75 = 0.0999 \cdot 1950 - 181.75 = 13.05$. За вычетом тренда имеем остаток $u_t = y_t - Y_t = 4.954 - 13.05 = -8.1$. В соответствии с приведенными формулами, значение момента времени $t = 1$ в радианах равно: $j = \frac{t\pi}{m} = \frac{1 \cdot 3.14}{14} = 0.224$, что позволяет найти косинус и синус: $\cos(ij) = \cos(1 \cdot 0.224) = 0.975$, $\sin(ij) = \sin(1 \cdot 0.224) = 0.222$,
а также оба произведения этих функций на остаток:
$u_1 \cdot \cos(ij) = -8.1 \cdot 0.975 = -7.898$, $u_1 \cdot \sin(ij) = -8.1 \cdot 0.222 = -1.8$.
Подсчет сумм подобных произведений для всех членов временно́го ряда дает значения:

$$\sum u_t \cdot \cos(ij) = \sum_{t=0}^{n-1} u_t \cos\frac{it\pi}{m} = -27.45, \quad \sum u_t \cdot \sin(ij) = \sum_{t=0}^{n-1} u_t \sin\frac{it\pi}{m} = 13.92.$$

Теперь можно рассчитать коэффициенты:

$$a_0 = \frac{1}{m}\sum_{t=0}^{n-1} u_t = 0.628 / 14 = 0.045,$$

$$a_1 = \frac{1}{m}\sum_{t=0}^{n-1} u_t \cos\frac{1t\pi}{m} = [\sum u_t \cdot \cos(1j)] / m = -27.45 / 14 = -1.961,$$

$$b_1 = \frac{1}{m}\sum_{t=0}^{n-1} u_t \sin\frac{1t\pi}{m} = [\sum u_t \cdot \sin(1j)] / m = 13.92 / 14 = 0.994.$$

Уравнение для первой гармоники (в отсутствие тренда) примет вид (рис. 9.8.3): $U_1 = a_0/2 + \sum(a_i \sin if_i t + b_i \cos if_i t)$ или
$U_1 = a_0/2 + a_1 \sin(1j) + b_1 \cos(1j) = 0.022 - 1.96 \cdot \sin j + 0.994 \cdot \cos j$.
Амплитуда первой гармоники равна:

$$A_1 = \sqrt{a_1^2 + b_1^2} = \sqrt{(-1.961)^2 + (0.994)^2} = 2.199.$$

Таблица 9.8.1. Расчет первой гармоники Фурье

	t	j	y_t	Y_t	u_t	cos(1j)	sin(1j)	$u$cos (1j)	$u$sin (1j)	U_1
1949	0	0	8.88	12.9	–4.1	1	0	–4.073	0	–1.96
1950	1	0.22	4.95	13.0	–8.1	0.975	0.222	–7.898	–1.8	–1.74
1951	2	0.44	5.85	13.1	–7.3	0.901	0.434	–6.578	–3.17	–1.52
1952	3	0.67	22.9	13.2	9.67	0.782	0.623	7.562	6.027	–1.34
1953	4	0.89	6.45	13.3	–6.9	0.624	0.782	–4.308	–5.4	–1.18
1954	5	1.12	14.6	13.4	1.23	0.434	0.901	0.536	1.111	–1.06
1955	6	1.34	27.0	13.5	13.5	0.223	0.975	3.014	13.16	–0.99
1956	7	1.57	11.8	13.6	–1.8	8E–04	1	–0.001	–1.82	–0.96
1957	8	1.79	5.18	13.7	–8.6	–0.22	0.975	1.899	–8.35	–0.99
1958	9	2.01	39.6	13.8	25.8	–0.43	0.901	–11.18	23.28	–1.06
1959	10	2.24	3.36	13.9	–10.	–0.62	0.783	6.589	–8.28	–1.18
1960	11	2.46	8.75	14.0	–5.3	–0.78	0.624	4.139	–3.31	–1.34
1961	12	2.69	19.7	14.1	5.59	–0.9	0.435	–5.035	2.433	–1.52
1962	13	2.91	10.9	14.2	–3.4	–0.97	0.224	3.264	–0.75	–1.73
1963	14	3.14	19.3	14.3	5.04	–1	0.002	–5.042	0.008	–1.95
1964	15	3.36	12.5	14.4	–1.9	–0.98	–0.22	1.826	0.413	–2.18
1965	16	3.58	13.8	14.5	–0.7	–0.9	–0.43	0.657	0.315	–2.39
1966	17	3.81	18.3	14.6	3.69	–0.78	–0.62	–2.889	–2.3	–2.57
1967	18	4.03	5.83	14.7	–8.9	–0.63	–0.78	5.572	6.958	–2.73
1968	19	4.26	23.2	14.8	8.40	–0.44	–0.9	–3.662	–7.56	–2.85
1969	20	4.48	31.8	14.9	16.9	–0.22	–0.97	–3.803	–16.5	–2.92
1970	21	4.71	13.9	15.0	–1.1	–0	–1	0.003	1.117	–2.95
1971	22	4.93	3.06	15.1	–12	0.22	–0.98	–2.661	11.79	–2.93
1972	23	5.15	13.0	15.2	–2.2	0.432	–0.9	–0.964	2.016	–2.85
1973	24	5.38	15.4	15.3	0.12	0.621	–0.78	0.077	–0.1	–2.74
1974	25	5.60	3.24	15.4	–12	0.78	–0.63	–9.524	7.64	–2.58
1975	26	5.83	23.7	15.5	8.19	0.9	–0.44	7.368	–3.58	–2.39
1976	27	6.05	13.2	15.6	–2.4	0.974	–0.23	–2.34	0.542	–2.18
			Σ =		0.63			–27.45	13.92	
			Σ / m =		0.04			–1.961	0.994	
							A =	2.199		

Расчет уравнения и амплитуды второй гармоники отражены в таблице 9.8.2 (аналогично можно получить параметры всех последующих компонент). Особенность этих вычислений состоит в том, что вторая гармоника имеет номер $i = 2$; этот множитель подставляется в выражение аргумента для синусов и косинусов. Таблица для расчетов должна иметь следующие поля:

$t,\ j = \dfrac{t\pi}{m},\ y_t,\ Y_t,\ u_t = y_t - Y_t,\ \sin(2j),\ \cos(2j),\ u_t{\cdot}\sin(2j),\ u_t{\cdot}\cos(2j).$

После необходимых вычислений и суммирования результирующее уравнение для двух гармоник примет вид:

$$U_{1,2} = a_0/2 + a_1{\cdot}\sin(j) + b_1{\cdot}\cos(j) + a_2{\cdot}\sin(2j) + b_2{\cdot}\cos(2j) =$$
$$= 0.022 - 1.96{\cdot}\sin j + 0.994{\cdot}\cos j - 2.23{\cdot}\sin 2j + 1.234{\cdot}\cos 2j.$$

Амплитуда у второй гармоники, как и у первой, невелика:

$$A_2 = \sqrt{a_2^2 + b_2^2} = \sqrt{(-2.23)^2 + (1.234)^2} = 2.545$$

и так же указывает на несущественный ее вклад в общую изменчивость всего ряда. Продолжая вычисления, мы в принципе можем получить ряд Фурье, состоящий из 14 слагаемых. Уравнение, например, восьми первых гармоник имеет вид:

$$U_{1\text{-}8} = 0.022 - 1.96{\cdot}\sin j + 0.994{\cdot}\cos j - 2.23{\cdot}\sin 2j + 1.234{\cdot}\cos 2j + 0.966{\cdot}\sin 3j - 0.7{\cdot}\cos 3j + + 0.914{\cdot}\sin 4j - 2.52{\cdot}\cos 4j - 3.2{\cdot}\sin 5j - 1.52{\cdot}\cos 5j + 0.67{\cdot}\sin 6j + 0.72{\cdot}\cos 6j - 1.03{\cdot}\sin 7j + + 0.08{\cdot}\cos 7j - 3.97{\cdot}\sin 8j - 5.58{\cdot}\cos 8j.$$

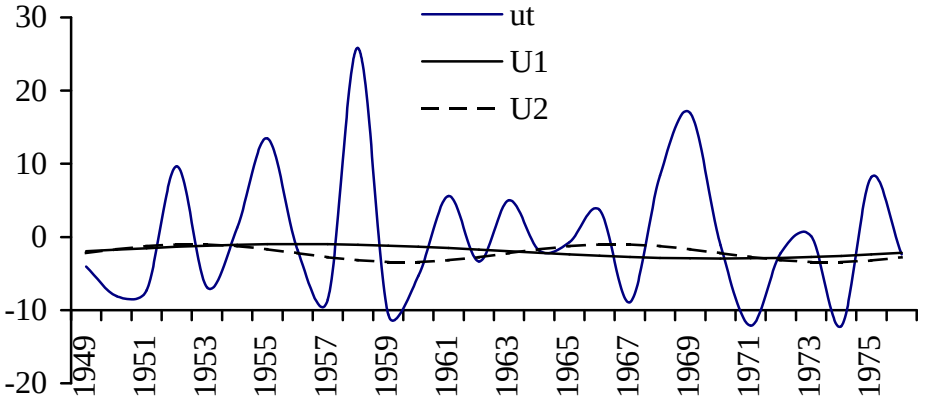


Рис. 9.8.3. Ряд остатков (u_t) и две первые гармоники (U_1 и U_2)

Таблица 9.8.2. Расчет второй гармоники Фурье для ряда значений численности полевок

	t	j	y_t	Y_t	u_t	cos(2j)	sin(2j)	$u$cos(2j)	$u$sin(2j)	U_2	$U_{1,2}$
1949	0	0	8.88	12.9	–4.0	1.0	0.0	–4.07	0	–2.22	–4.16
1950	1	0.22	4.95	13.0	–8.1	0.901	0.434	–7.3	–3.51	–1.69	–3.41
1951	2	0.44	5.85	13.1	–7.3	0.624	0.782	–4.55	–5.71	–1.26	–2.77
1952	3	0.67	22.9	13.2	9.7	0.223	0.975	2.16	9.43	–1.02	–2.34
1953	4	0.89	6.45	13.3	–6.9	–0.22	0.975	1.53	–6.73	–1.02	–2.18
1954	5	1.12	14.6	13.4	1.23	–0.62	0.783	–0.77	0.96	–1.26	–2.3
1955	6	1.34	27.0	13.5	13.5	–0.9	0.435	–12.2	5.87	–1.68	–2.66
1956	7	1.57	11.8	13.6	–1.8	–1.0	0.002	1.82	–0	–2.22	–3.17
1957	8	1.79	5.18	13.7	–8.5	–0.9	–0.43	7.725	3.7	–2.75	–3.73
1958	9	2.01	39.6	13.8	25.8	–0.63	–0.78	–16.1	–20.2	–3.18	–4.23
1959	10	2.24	3.36	13.9	–10	–0.22	–0.97	2.38	10.31	–3.42	–4.59
1960	11	2.46	8.75	14.0	–5.3	0.22	–0.98	–1.17	5.17	–3.42	–4.75
1961	12	2.69	19.7	14.1	5.59	0.621	–0.78	3.47	–4.38	–3.19	–4.7
1962	13	2.91	10.9	14.2	–3.3	0.9	–0.44	–3.01	1.46	–2.76	–4.48
1963	14	3.14	19.3	14.3	5.04	1.0	–0.0	5.04	–0.02	–2.22	–4.17
1964	15	3.36	12.5	14.4	–1.8	0.902	0.431	–1.69	–0.81	–1.69	–3.85
1965	16	3.58	13.8	14.5	–0.7	0.626	0.78	–0.46	–0.57	–1.26	–3.63
1966	17	3.81	18.3	14.6	3.69	0.226	0.974	0.83	3.59	–1.02	–3.58
1967	18	4.03	5.83	14.7	–8.9	–0.22	0.976	1.95	–8.7	–1.02	–3.74
1968	19	4.26	23.2	14.8	8.40	–0.62	0.785	–5.21	6.59	–1.25	–4.09
1969	20	4.48	31.8	14.9	16.9	–0.9	0.438	–15.2	7.41	–1.68	–4.59
1970	21	4.71	13.9	15.0	–1.1	–1.0	0.005	1.12	–0.01	–2.21	–5.15
1971	22	4.93	3.06	15.1	–12	–0.9	–0.43	10.92	5.19	–2.75	–5.66
1972	23	5.15	13.0	15.2	–2.2	–0.63	–0.78	1.40	1.74	–3.18	–6.02
1973	24	5.38	15.4	15.3	0.12	–0.23	–0.97	–0.03	–0.12	–3.42	–6.15
1974	25	5.60	3.24	15.4	–12	0.217	–0.98	–2.65	11.92	–3.43	–5.99
1975	26	5.83	23.7	15.5	8.19	0.619	–0.79	5.07	–6.43	–3.19	–5.57
1976	27	6.05	13.2	15.6	–2.4	0.898	–0.44	–2.16	1.06	–2.76	–4.93
			Σ =		0.63			–31.2	17.27		
			Σ/m =		0.045			–2.23	1.234		
							A =	2.545			

9.9. Спектральный анализ

По уравнению Фурье уже можно судить о том, какие из гармоник в основном образуют изучаемый временно́й ряд – их коэффициенты имеют большие значения. Анализ показал, что коэффициенты восьмой и десятой гармоник выше, чем у других, значит, они играют роль важных источников варьирования изучаемой переменной *y*. Тем не менее результаты гармонического анализа эффективнее выражать таблицей, содержащей значения амплитуды для соответствующих периодов и частот, а еще лучше – диаграммой, по оси ординат которой представлены амплитуды гармоник, по оси абсцисс – соответствующие им частоты (или периоды). Такая диаграмма называется *спектр амплитуд*. График спектра, построенный для дискретных временных рядов (с которыми обычно и приходится иметь дело), носит синонимичное название *периодограмма*. Поскольку амплитуда A_i представляет собой размах колебаний доли значений ряда, связанной с реализацией данной *i*-й гармоники, анализ ломаной линии спектра амплитуд приводит к выводам о составе регулярных составляющих исходного ряда. Большие значения амплитуды (ординаты) A_i, соответствующие той или иной гармонике, свидетельствуют о действительной реализации колебаний с данной частотой. Значения амплитуды, близкие к нулю, говорят о том, что с данной частотой не наблюдается периодической повторяемости значений ряда, то есть свидетельствуют о нулевом вкладе данной гармоники в общую картину варьирования показателя.

Периодограмма

Рассчитать периодограмму в среде пакета StatGraphics можно с помощью процедуры Special \ Time-Series Analysis \ Descriptive Methods. На вкладке Periodogram для каждой гармоники с соответствующей частотой (Frequency) и периодом (Period) представлены значения, характеризующие амплитуды (Ordinate). В пакете StatGraphics и Statistica для характеристики амплитуды используется не само значение A_i, а сумма квадратов отклонений значений данной гармоники от общей средней $CK_i = m \cdot A^2{}_i = n \cdot S^2{}_i$. Это делается для того, чтобы воспользоваться свойством аддитивности сумм квадратов и рассчитать относительный вклад отдельных гармоник, а также накопление их вклада в исходные значения временно́го ряда.

Гармонический (теперь уже *спектральный*) анализ динамики численности полевки (табл. 9.9.1, графа A_i, рис. 9.9.1) показал относительно высокие значения амплитуды для восьмой и десятой гармоник ($A_8 = 6.85$, $A_{10} = 6.77$) с частотами $f_8 = 0.29$ (период $T_8 = 3.5$ лет) и $f_{10} = 0.38$ ($T_{10} = 2.8$ года). Это обстоятельство позволяет предполагать наличие трехлетней цикличности, хотя иногда популяционные циклы проходят за 2, иногда за 4 года.

Таблица 9.9.1. Амплитуды и дисперсии гармоник

i	T_i	f_i	A_i	A^2_i	S^2_i	$m \cdot A^2_i = CK_i$	CK_i, %	$\Sigma\, CK_i$	g_i
0	–	0	0	0	0	0	0	0	0
1	28	0.04	2.2	4.83	2.42	67.7	3	3	0.02
2	14	0.07	2.54	6.48	3.24	90.7	4	7	0.02
3	9.33	0.11	1.19	1.42	0.71	19.9	1	8	0
4	7	0.14	2.68	7.21	3.6	101	4	12	0.02
5	5.6	0.18	3.54	12.6	6.28	176	8	20	0.04
6	4.67	0.21	0.99	0.98	0.49	13.7	1	21	0.01
7	4	0.25	1.04	1.08	0.54	15.1	1	21	0
8	**3.5**	**0.29**	6.85	46.9	**23.4**	656	29	51	0.14
9	3.11	0.32	1.95	3.81	1.9	53.3	2	53	0.01
10	**2.8**	**0.36**	6.77	45.8	**22.9**	642	28	81	0.14
11	2.55	0.39	3.59	12.9	6.44	180	8	89	0.04
12	2.33	0.43	2.77	7.7	3.85	108	5	94	0.02
13	2.15	0.46	2.7	7.27	3.63	102	5	99	0.02
14	2	0.5	1.44	2.09	1.04	29.2	1	100	0.01
Σ						2255	100		

Оценка относительной величины вклада той или иной гармоники выполняется с помощью другого показателя, служащего для характеристики варьирования и более привычного для практики статистического анализа, – с помощью дисперсии значений данной гармоники, S^2_i. Между дисперсией и амплитудой каждой гармоники

имеется простое соотношение: $S^2_i = A^2_i / 2$, поскольку амплитуда есть не что иное, как характеристика разброса значений по обеим сторонам от оси гармоники.

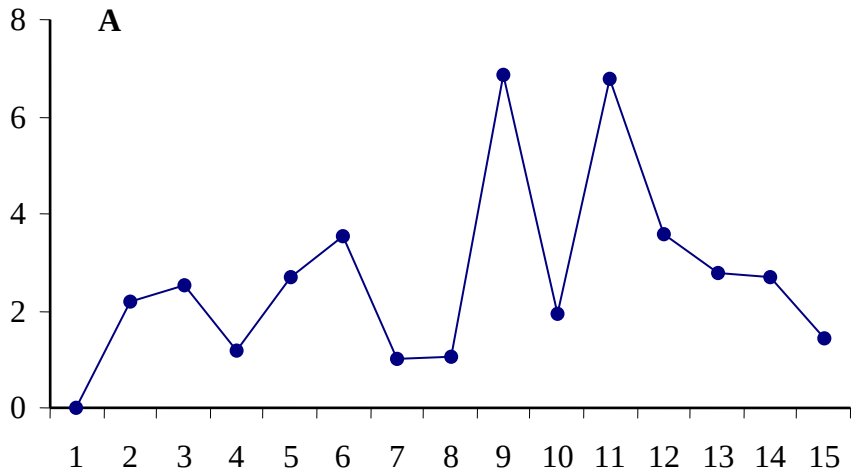


Рис. 9.9.1. Периодограмма: амплитуды гармоник для ряда динамики полевок

Заменив понятия, можно обсуждать спектральный анализ ряда в терминах дисперсионного (или регрессионного) анализа. Равно как общая амплитуда временного ряда складывается из амплитуд отдельных гармоник, также и общая изменчивость всех данных временно́го ряда складывается из варьирования значений, принадлежащих отдельным гармоникам (за вычетом тренда и шума). Это дает возможность оценить величину *относительного вклада* каждой периодической компоненты в общую дисперсию, сопоставляя суммы квадратов отклонений гармоник от средней $CK_i = m \cdot A^2_i = n \cdot S^2_i$ с общей суммой квадратов $CK_{общ.} = \sum CK_i = (n-1) \cdot S^2_u$.

В нашем случае дисперсия первой гармоники составила $S^2_1 = A^2_1/2 = 2.199^2/2 = 2.42$, сумма квадратов $CK_1 = n \cdot S^2_1 = 28 \cdot 2.42 = 67.7$. Дисперсия для исходных значений u равна $S^2_u = 82.99$, сумма квадратов с точностью до ошибки округлений составит $CK_{общ} = \sum CK_i = (n-1) \cdot S^2_u = 27 \cdot 82.99 = 2241$.

Очевидно, что вклад первой гармоники в общую изменчивость ($100 \cdot CK_1/CK_{общ.} = 67.7 / 2241 \approx 3\%$) очень мал, то есть исходный ряд данных фактически не содержит гармонических компонент

с таким длинным периодом $T = 28$ и такой низкой (основной) частотой $f_1 = 1/28 = 0.0357$. Относительная роль восьмой и десятой гармоник, напротив, весьма велика – 29 и 28%.

Статистическая значимость вклада отдельной гармоники определяется с помощью критерия, похожего на известный критерий Фишера: $g_i = \dfrac{S_i^2}{2S_{общ.}^2}$. Для каждой гармоники нашего примера значения критерия рассчитаны в графе g_i табл.9.9.1.

Критическое значение для принятого уровня доверительной вероятности $P = 0.95$ вычисляется по формуле:

$g \approx 1 - \exp\left(\dfrac{\ln P - \ln m}{m-1}\right)$, где $m = n/2$ для четного числа значений ряда и $m = (n-1)/2$ для нечетного n.

В примере имеем: $g \approx 1 - \exp\left(\dfrac{\ln 0.95 - \ln 14}{13}\right) = 0.186.$

Проверим значимость первой и восьмой гармоник. Полученное значение $g_1 = 0.06$ меньше критического $g = 0.186$, следовательно, при доверительной вероятности $P = 0.95$ участие первой гармоники в формировании значений изученного временно́го ряда от нуля не отличается. То же приходится констатировать и для остальных гармоник, даже для восьмой и десятой:

$g_8 = 0.14 < 0.186$ и $g_{10} = 0.14 < 0.186$.

Таким образом, для данного ряда значимых периодических компонент обнаружить не удалось. Гипотеза о трехлетнем цикле численности полевок не подтвердилась. Причина состоит, видимо, в невысокой репрезентативности данных. Нетрудно подсчитать значения критерия для рядов разной длины: для $n = 20$ $g = 0.15$, для $n = 30$ $g = 0.11$, для $n = 100$ $g = 0.05$. Иными словами, восьмая гармоника, возможно, могла бы стать значимой при несколько большей длине изучаемого ряда ($n \approx 35$–50).

Непрерывный спектр

Периодичность природных явлений никогда не бывает строгой, однотипные состояния биосистем повторяются через *неодинаковые* отрезки времени. Например, дата выпадения первого снега в г. Петрозаводске варьирует в пределах более месяца (от 20 сентября

до 20 ноября), то есть имеет период от 11 до 13 месяцев и частоту от 1.09 до 0.92 год$^{-1}$. Помимо сезонных явлений, подчиненных строгим законам астрономической цикличности, нечеткий ритм в большей степени характерен для эндогенных биологических ритмов (например, для динамики популяций рыжей полевки). Обнаруженные высокие значения амплитуд для двух гармоник (с периодом 3.5 и 2.8 г.) могут свидетельствовать лишь о том, что «всплески» численности повторяются один раз в 2–4 года. Для таких нестрого периодических процессов не имеет смысла проводить гармонический анализ, ориентированный на выявление уравнений *отдельных гармоник* с определенной частотой. Скорее, следует говорить о выявлении *полосы частот*, соответствующей целому «кусту» (семейству) циклических компонент, множеству реализаций нестрого циклической компоненты. Так, явление первого снегопада характеризуется полосой частот от 0.92 до 1.09, но не единственной частотой 1 год$^{-1}$, равно как и динамика численности полевки существенную выраженность имеет лишь полоса частот 0.28–0.43 (период 3.5–2.3 года) или еще шире.

Существенно осложняет интерпретацию периодограммы то обстоятельство, что значения ординат дисперсии (амплитуды) в пределах одной полосы частот оказываются неодинаковыми, здесь часто наблюдаются «провалы». Например, для ряда численности полевок период 3.1 года ($f_9 = 0.32$) имеет низкую амплитуду ($S_9 = 1.9$), тогда как окружающие ее частоты (f_8 и f_{10}) имеют высокие значения ординаты ($S_8 = 23.4$, $S_{10} = 22.9$). В чем тут дело? Причина, как правило, достаточно тривиальна: за резкие перепады *близких* частот ответственны случайные факторы. Вследствие того, что объемы изучаемых выборок ограничены, для получения «хорошей» выраженности *всех* близких частот данной полосы не хватает данных. Однако если хотя бы несколько соседствующих частот представлены достаточно большими ординатами, мы можем, во-первых, утверждать, что периодичность имеет место, а во-вторых, получить усредненную информацию по близким значениям спектра путем его *сглаживания*. Процедура фильтрации дает *непрерывный спектр*, или *спектральную плотность* (spectral density). Метод вычислений, дающий спектральную плотность, называется *спектральным анализом*. Общий смысл непрерывного спектра таков же, что и у периодограммы – ординаты представляют собой значения амплитуды, но они сглажены и соответствуют не отдельным ортогональным гар-

моникам, а полосам частот. Рост значений амплитуды (где график непрерывного спектра образует «горб») указывает на периодизм процесса с частотой, близкой к центру полосы.

Для сглаживания значений спектра разработаны разнообразные математические фильтры (*окна*), частично рассмотренные выше. Конструкции фильтров (*формы окон*) основаны на разных теоретических соображениях; они могут иметь различное число (k) весовых коэффициентов (разную *ширину окна*) и разные их значения. Так, окно Даниеля представляет собой простую скользящую среднюю (весовой коэффициент у каждого члена равен единице); усреднение нескольких соседних значений амплитуды с приписыванием результата центральному значению этого сегмента. Прочие окна (Бартлетта, Тьюки, Хэмминга и др.) назначают весовые коэффициенты по форме полиномов разных степеней. Окно Бартлетта обеспечивает сильный приоритет центральных значений сегмента над периферическими (для окна шириной 5 весá такие: 0, 0.25, 0.5, 0.25, 0). Окно Хэмминга использует более пологую параболу, что придает периферическим значениям бóльшие веса (для ширины 5 – 0.035, 0.241, 0.464, 0.241, 0.035). Форма и ширина окна фильтрации во многом определяют вид кривой спектральной плотности. Выбор этих параметров и, следовательно, эффективность выполненного анализа во многом зависят от опыта исследователя. Общей рекомендацией может быть апробация обработки одного стандартизированного ряда разными метриками, что позволит лучше понять работоспособность в данной ситуации той или иной характеристики фильтра.

Построить как периодограмму, так и график спектральной плотности с использованием разных фильтров позволяет пакет Statistica. Возможен следующий порядок работы. Сначала нужно организовать таблицу с данными: создать новую таблицу (File \ New), увеличить число записей до размеров изучаемого ряда (Cases \ Add), ввести значения ряда или с помощью буфера обмена скопировать данные из среды Excel и вставить в таблицу пакета Statistica (можно импортировать и файл типа *.xls). Далее вызываем программу обработки временных рядов: Analysis \ Other Statistics \ Time Series/Forecasting (или Analysis \ Resume Analysis, если эта панель уже была ранее открыта). Выбираем переменную для анализа (кнопка Variables, выбор, OK). Далее нажимаем кнопки Spectral

(Fourier) analysis и в новом окне – Single Series Fourier Analysis. Появившаяся панель позволяет построить периодограмму (кнопка Periodogram), график непрерывного спектра (Spectral density), откладывая по оси абсцисс (Plot by) значения частот (Frequency), периодов (Periods) или их логарифмов (Log (Period)). В блоке настроек окна фильтрации (Data windows for spectral estimates) можно выбрать вид фильтра (Daniell, Hamming, Bartlett, Tukey, Parsen) или ввести собственный формат (Use-definede), а также назначить необходимую ширину окна (Width of data window). Нажав кнопку Summary, все результаты расчетов выводим в численном виде в электронную таблицу, после чего их можно распечатать или через буфер обмена скопировать в другие документы (табл. 9.9.2).

Таблица 9.9.2. Гармонический и спектральный анализы в среде Statistica

No. of cases: 28							Spectral
	Frequn	Period	Cosine	Sine	Period	Densi-	Bartlett Weigths
1	0.036	28.000	-1.960	1.012	68.125	52.727	0.111111
2	0.071	14.000	-2.231	1.241	91.257	61.041	0.222222
3	0.107	9.333	0.966	-0.679	19.526	76.164	0.333333
4	0.143	7.000	0.935	-2.508	100.29	88.460	0.222222
5	0.179	5.600	-3.187	-1.547	175.67	87.736	0.111111
6	0.214	4.667	0.659	0.733	13.598	131.15	0
7	0.250	4.000	1.039	0.112	15.291	179.84	
8	0.286	3.500	-3.907	-5.629	657.40	307.11	
9	0.321	3.111	0.946	-1.743	55.053	327.92	
10	0.357	2.800	3.277	5.899	637.58	349.88	
11	0.393	2.545	-3.408	1.143	180.92	243.46	
12	0.429	2.333	2.465	1.248	106.84	173.12	
13	0.464	2.154	1.063	2.517	104.51	96.759	
14	0.500	2.000	-1.441	0.000	29.075	79.888	

В примере (рис. 9.9.2, А, Б) сглаживание выполнено с помощью окна Бартлетта шириной 7. Как можно заметить, пики в окрестностях частоты $f_9 \approx 0.33$ (период $T_9 \approx 3$ года) слились, образовав одну доминирующую вершину.

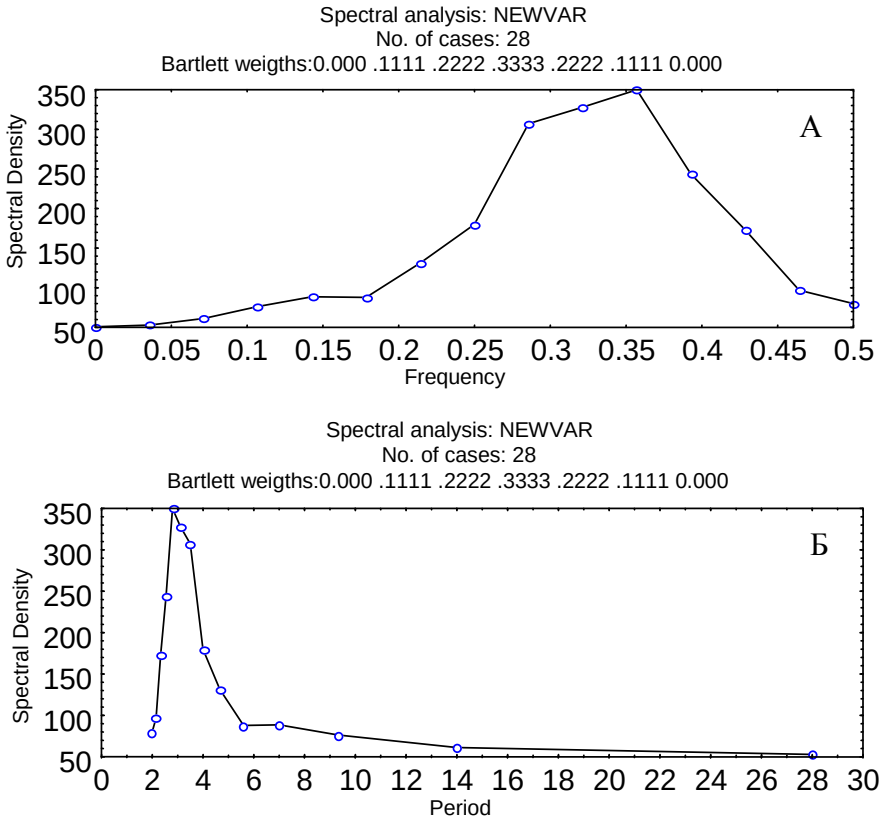


Рис. 9.9.2. Спектральная плотность для ряда численности рыжей полевки; по оси абсцисс отложены частоты (А) и периоды (Б)

Информация, сконцентрированная в кривой спектральной плотности, указывает на 2–4-летний периодизм рассматриваемого ряда, то есть большей строгости описания получить не удалось. Следовательно, процесс динамики популяции вовсе не однозначен, и единственной оценки, характеризующей численность за весь год, видимо, недостаточно, чтобы объяснить ритм смены фаз роста и депрессии популяции. Направление дальнейшего анализа состоит в детализации исходной информации и усложнении метода исследования. Каждую среднегодовую оценку численности следует заменить хотя бы на две, связанные с явлениями смены фаз размножения и вымирания. Эту идею реализуют с помощью функций последования (п. 7.4).

СПИСОК ЛИТЕРАТУРЫ

Базовая литература

Айвазян С. А., Бухштабер В. М., Энюков М. С., Мешалкин Л. Д. Прикладная статистика. Классификации и снижение размерности. М.: Финансы и статистика,1989. 607 с.

Браунли К. А. Статистическая теория и методология в науке и технике. М.: Наука, 1977. 408 с.

Гроссман С., Терней Дж. Математика для биологов. М.: Высшая школа, 1983. 383 с.

Джефферс Дж. Введение в системный анализ: применение в экологии. М.: Мир, 1981. 253 с.

Дэвис Дж. Статистический анализ данных в геологии. М.: Недра, 1977. 573 с.

Дэвис Дж. Статистический анализ данных в геологии: В 2 кн. М.: Недра, 1990. Кн. 1. 346 с.

Животовский Л. А. Популяционная биометрия. М.: Наука, 1991. 271 с.

Зайцев Г. Н. Математический анализ биологических данных. М.: Наука, 1981. 183 с.

Зайцев Г. Н. Математика в экспериментальной ботанике. М.: Наука, 1990. 267 с.

Ивантер Э. В., Коросов А. В. Основы биометрии. Петрозаводск: Изд-во ПетрГУ, 1992. 165 с.

Ивантер Э.В., Коросов А. В. Введение в количественную биологию, учебное пособие. Петрозаводск: Изд-во ПетрГУ, 2003. 304 с.

Лакин Г. Ф. Биометрия. М.: Высшая школа, 1973. 343 с.

Методы математической биологии. Т. 1: Общие методы анализа биологических систем. Киев: Вища школа, 1980. 239 с.

Методы математической биологии. Т. 8. Математические решения задач биологии и медицины на ЭВМ. Киев: Вища школа, 1984. 344 с.

Плохинский Н. А. Алгоритмы биометрии. М., 1986. 199 с.

Поллард Дж. Справочник по вычислительным методам статистики. М.: Финансы и статистика, 1982. 344 с.

Пузаченко Ю. Г. Математические методы в экологических и географических исследованиях. М.: Академия, 2004. 416 с.

Справочник по прикладной статистике. М.: Финансы и статистика, 1990. Т. 2. 526 с.

Терентьев П. В., Ростова Н. С. Практикум по биометрии. Л., 1977. 151 с.

Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981. 694 с.

Урбах В. Ю. Биометрические методы. М.: Наука, 1964. 415 с.

Юл Дж. Э., Кендэл М. Дж. Теория статистики. М.: Госстатиздат, 1960. 779 с.

Дополнительная литература

Андреев В. А. Классификационные построения в экологии и систематике. М.: Наука, 1980.

Ашмарин И. П. Васильев Н. Н., Амбросов В. А. Быстрые методы статистической обработки и планирования экспериментов. Л.: Изд-во ЛГУ, 1975. 78 с.

Василевич В. И. Статистические методы в геоботанике. М.: Наука, 1969. 232 с.

Голиков А. П., Черванев И. Г., Трофимов А. М. Математические методы в географии. Харьков: Вища школа, 1986. 144 с.

Голикова Т. И., Никитина Е. П., Терехин А. Т. Математическая статистика. М.: Изд-во МГУ, 1981. 185 с.

Гублер Е. В., Генкин А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л.: Медицина, 1973. 142 с.

Гублер Е. В. Информатика в патологии, клинической медицине и педиатрии. Л.: Медицина. 1990. 176 с.

Дженкинс Г., Ваттс Д. Спектральный анализ и его приложение. Вып. 1. М.: Мир, 1971. 317 с.

Дэйвисон М. Многомерное шкалирование: методы наглядного представления данных. М.: Финансы и статистика, 1988. 254 с.

Дюк В. А. Компьютерная психодиагностика. СПб.: «Братство», 1994. 364 с.

Енюков И. С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА. М.: Финансы и статистика, 1986. 232 с.

Ефимов В. М., Галактионов Ю. К., Шушпанова Н. Ф. Анализ и прогноз временных рядов методом главных компонент. Новосибирск: Наука, 1988. 71 с.

Иберла К. Факторный анализ. М.: Статистика, 1980. 367 с.

Казенс Дж. Введение в лесную экологию. М.: Лесная промышленность, 1982. 144 с.

Кноринг Л. Д., Деч В. Н. Геологу о математике. Л.: Недра, 1989. 208 с.

Коросов А. В. Экологические приложения компонентного анализа. Петрозаводск: Изд-во ПетрГУ, 1996. 152 с.

Коросов А. В. Имитационное моделирование в среде MS Excel. Петрозаводск: Изд-во ПетрГУ, 2002. 212 с.

Котов В. Н. Применение теории измерений в биологических исследованиях. Киев: Наукова думка, 1985. 100 с.

Лоули Д. И., Максвелл А. Э. Факторный анализ как статистический метод. М.: Мир, 1967. 144 с.

Минцер О. П., Угаров Б. Н., Власов В. В. Методы обработки медицинской информации. Киев: Вища школа, 1982. 160 с.

Недорезов Л. В. Лекции по математической экологии. Новосибирск, 1997. 161 с.

Нивергельт Ю., Фаррар Дж., Рейнгольд Э. Машинный подход к решению математических задач. М.: Мир, 1977. 352 с.

Окунь Я. Факторный анализ. М.: Статистика,1974. 200 с.

Отнес Р., Эноксон Л. Прикладной анализ временных рядов. Основные методы. М.: Мир, 1982. 428 с.

Перегудов Ф. И., Тарасенко Ф. П. Введение в системный анализ. М.: Высшая школа, 1989. 367 с.

Песенко Ю. А. Принципы и методы количественного анализа в фаунистических исследованиях. М.: Наука, 1982. 286 с.

Розенберг Г. С. Модели в фитоценологии. М.: Наука, 1984. 265 с.

Саати Т., Кернс К. Аналитическое планирование. Организация систем. М.: Радио и связь, 1991. 224 с.

Саранча Д. А. Биомоделирование. М., 1995. 139 с.

Саранча Д. А. Количественные методы экологии. Биофизические аспекты и математическое моделирование. М.: МФТИ, 1996. 252 с.

Страшкраба М., Гнаук А. Пресноводные экосистемы. Математическое моделирование. М.: Мир, 1989. 376 с.

Стройк Д. Я. Краткий очерк истории математики. М.: Наука, 1990. 256 с.

Терехина А. Ю. Анализ данных методами многомерного шкалирования. М.: Наука, 1986. 168 с.

Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981. 694 с.

Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. М.: ИНФРА, 1998. 528 с.

Шитиков В. К., Розенберг Г. С., Зинченко Т. Д. Количественная гидроэкология: методы системной идентификации. Тольятти, 2003. 463 с.

Специальная литература

Адлер Ю. П., Макарова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 1976. 280 с.

Антипина Г. С. Урбанофлора Карелии. Петрозаводск, 2002. 200 с.

Антомонов Ю. Г. Моделирование биологических систем: Справочник. Киев: Наукова думка, 1977. 260 с.

Безель В. С. Популяционная экотоксикология млекопитающих. М.: Наука, 1987. 130 с.

Боровиков В. П., Боровиков И. Д. STATISTICA – статистический анализ и обработка данных в среде WINDOWS. М., 1997. 608 с.

Вейль Г. Математическое мышление. М.: Наука, 1989. 400 с.

Гелашвили Д. Б., Якимов В. Н., Логинов В.В., Епланова Г. В. Статистический анализ флуктуирующей асимметрии билатеральных признаков разноцветной ящурки // Актуальные проблемы герпетологии и токсинологии: Сб. науч. тр. Вып. 7. Тольятти, 2004. С. 45–59.

Гильдерман Ю. И. Лекции по высшей математике для биологов. Новосибирск: Наука, 1974. 411 с.

Гиляров А. М. Популяционная экология. М.: Изд-во МГУ, 1990. 191 с.

Дженкинс Г., Ваттс Д. Спектральный анализ и его приложение. Вып. 1. М.: Мир, 1971. 317 с.

Европейская рыжая полевка. М.: Наука, 1981. 352 с.

Ердаков Л. Н. Организация ритмов активности грызунов. Новосибирск: Наука, 1984. 182 с.

Животовский Л. А. Интеграция полигенных систем в популяции. М.: Наука, 1984. 183 с.

Захаров В. М. Асимметрия животных (популяционно-феногенетический подход). М.: Наука, 1987. 216 с.

Захаров В. М. Онтогенез и популяция (стабильность развития и популяционная изменчивость) // Экология. 2001. № 3. С. 164–168.

Иванищев В. В., Михайлов В. В., Тубольцева В. В. Инженерная экология. Л.: Наука, 1989. 144 с.

Ивантер Э. В. Популяционная экология мелких млекопитающих таежного Северо-Запада СССР. Л.: Наука, 1975. 240 с.

Карнап Р. Философские основания физики. Введение в философию науки. М.: Прогресс, 1971. 391 с.

Клайн М. Математика. Поиск истины. М.: Мир, 1988. 295 с.

Клейн Ф. Элементарная математика с точки зрения высшей: В 2 т. Т. 1.: Арифметика. Алгебра. Анализ. М.: Наука, 1987. 432 с.

Коли Г. Анализ популяций животных. М.: Мир, 1979. 364 с.

Коросов А. В., Зорина А. А. Исследование динамики численности рыжей полевки с помощью функций последования // Экология. 2007. № 1. С. 49–54.

Лебедева Н. В., Дроздов Н. Н., Криволуцкий Д. А. Биологическое разнообразие. М.: ВЛАДОС, 2004. 432 с.

Максимов А. А. Многолетние колебания численности животных, их причины и прогноз. Новосибирск, Наука, 1984. 250 с.

Максимов А. А. Ердаков Л. Н. Циклические процессы в сообществах животных. Новосибирск: Наука, 1985. 236 с.

Методические рекомендации по выполнению оценки качества среды по состоянию живых существ (оценка стабильности развития живых организмов по уровню асимметрии морфологических структур). Распоряжение Росэкология от 16.10.2003 № 460-р.

Миркин Б. М., Розенберг Г. С., Наумова Л. Г. Словарь понятий и терминов современной фитоценологии. М.: Наука, 1989. 223 с.

Михеев В. И. Моделирование и методы теории измерений в педагогике. М.: Едиториал УРСС, 2004. 200 с.

Моисеев Н. Н. Математические задачи системного анализа. М.: Наука, 1981. 487 с.

Наумов Н. Н. Очерки сравнительной экологии мышевидных грызунов. М ;Л.: Изд-во АН СССР, 1948. 204 с.

Недорезов Л. В. Лекции по математической экологии. Новосибирск, 1997. 161 с.

Никитина Н. А. Рыжие полевки // Итоги течения млекопитающих. М.: Наука, 1980. С.189-219.

Окулова Н. М., Баженова А. Ф. Задачи по биометрии для зоологов. Иваново: Изд-во ИГУ, 1993. 88 с.

Полищук Л. В., Цейтлин В. Б. Масса тела, плотность популяции и число потомков у млекопитающих // Журн. общ. биол. 2001. Т. 62. № 1. С. 3–24.

Прицкер А. Введение в имитационное моделирование и язык СЛАМ. М.: Мир, 1987. 644 с.

Пфанцагль И. Теория измерений. М.: Мир, 1976. 248 с.

Пэнтл Р. Методы системного анализа окружающей среды. М.: Мир, 1979. 214 с.

Реймерс Н. Ф. Популярный биологический словарь. М.: Наука, 1991. 544 с.

Реньи А. Трилогия о математике. М.: Мир, 1980. 376 с.

Саймон Г. Науки об искусственном. М.: УРСС, 2004. 144 с.

Селиванов М. Н., Фридман А. Э., Кудряшова Ж. Ф. Качество измерений: Метрологическая справочная книга. Л.: Лениздат, 1987. 295 с.

Сена Л. А. Единицы физических величин и их размерности. Учебно-справочное руководство. М,: Наука, 1988. 432 с.

Смирнов В. С. Определение характера связи между признаками и вычисление сопряженной вариабельности // Применение количественных методов в экологии. Свердловск, 1979. С. 3–17.

Соколов Г.А. Млекопитающие кедровых лесов Сибири. Новосибирск: Наука, 1979. 256 с.

Терентьев П. В. Дальнейшее развитие метода корреляционных плеяд // Применение математических методов в биологии. Л., 1960. С. 27–36.

Терентьев П. В. Метод корреляционных плеяд // Вестник ЛГУ. 1959. Вып. 2. N 9. С. 135–141.

Уилкокс Б. А. Островная экология и охрана природы // Биология охраны природы. М.: Мир, 1980. С. 117-142.

Шварц С. С., Смирнов В. С., Добринский Л. Н. Метод морфофизиологических индикаторов в экологии наземных позвоночных. Свердловск, 1968. 387 с.

Шуп Т. Е. Прикладные численные методы в физике и технике. М.: Высшая школа, 1990. 254 с.

Экоинформатика. Теория. Практика. Методы и системы. СПб.: Гидрометеоиздат, 1992. 520 с.

СПРАВОЧНЫЕ ТАБЛИЦЫ

Таблица 1С. Перевод календарных дат в непрерывный ряд

Месяцы											
III	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**I**	**II**
1	32	62	93	123	154	185	215	246	276	307	338
2	33	63	94	124	155	186	216	247	277	308	339
3	34	64	95	125	156	187	217	248	278	309	340
4	35	65	96	126	157	188	218	249	279	310	341
5	36	66	97	127	158	189	219	250	280	311	342
6	37	67	98	128	159	190	220	251	281	312	343
7	38	68	99	129	160	191	221	252	282	313	344
8	39	69	100	130	161	192	222	253	283	314	345
9	40	70	101	131	162	193	223	254	284	315	346
10	41	71	102	132	163	194	224	255	285	316	347
11	42	72	103	133	164	195	225	256	286	317	348
12	43	73	104	134	165	196	226	257	287	318	349
13	44	74	105	135	166	197	227	258	288	319	350
14	45	75	106	136	167	198	228	259	289	320	351
15	46	76	107	137	168	199	229	260	290	321	352
16	47	77	108	138	169	200	230	261	291	322	353
17	48	78	109	139	170	201	231	262	292	323	354
18	49	79	110	140	171	202	232	263	293	324	355
19	50	80	111	141	172	203	233	264	294	325	356
20	51	81	112	142	173	203	234	265	295	326	357
21	52	82	113	143	174	205	235	266	296	327	358
22	53	83	114	144	175	206	236	267	297	328	359
23	54	84	115	145	176	207	237	268	298	329	360
24	55	85	116	146	177	208	238	269	299	330	361
25	56	86	117	147	178	209	239	270	300	331	362
26	57	87	118	148	179	210	240	271	301	332	363
27	58	88	119	149	180	211	241	272	302	333	364
28	59	89	120	150	181	212	242	273	303	334	365
29	60	90	121	151	182	213	243	274	304	335	(366)
30	61	91	122	152	183	214	244	275	305	336	
31		92		153	184		245		306	337	

Таблица 2С. Ординаты нормальной кривой: $f(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{t^2}{2}}$

t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.399	0.399	0.399	0.399	0.399	0.398	0.398	0.398	0.398	0.397
0.1	0.397	0.397	0.396	0.396	0.395	0.395	0.394	0.393	0.383	0.392
0.2	0.391	0.390	0.389	0.389	0.388	0.387	0.386	0.385	0.384	0.383
0.3	0.381	0.380	0.379	0.378	0.377	0.375	0.374	0.373	0.371	0.370
0.4	0.368	0.367	0.365	0.364	0.362	0.361	0.359	0.357	0.356	0.354
0.5	0.352	0.350	0.349	0.347	0.345	0.343	0.341	0.339	0.337	0.335
0.6	0.333	0.331	0.329	0.327	0.325	0.323	0.321	0.319	0.317	0.314
0.7	0.312	0.310	0.308	0.306	0.303	0.301	0.299	0.297	0.294	0.292
0.8	0.299	0.287	0.285	0.283	0.280	0.278	0.276	0.273	0.271	0.269
0.9	0.266	0.264	0.261	0.259	0.257	0.254	0.252	0.249	0.247	0.244
1	0.242	0.240	0.237	0.235	0.232	0.230	0.228	0.225	0.223	0.220
1.1	0.218	0.216	0.213	0.211	0.208	0.206	0.204	0.201	0.199	0.197
1.2	0.194	0.192	0.190	0.187	0.185	0.183	0.180	0.178	0.176	0.174
1.3	0.171	0.169	0.167	0.165	0.163	0.160	0.158	0.156	0.154	0.152
1.4	0.150	0.148	0.146	0.144	0.142	0.139	0.137	0.135	0.133	0.132
1.5	0.130	0.128	0.126	0.124	0.122	0.120	0.118	0.116	0.115	0.113
1.6	0.111	0.109	0.107	0.106	0.104	0.102	0.101	0.099	0.097	0.096
1.7	0.094	0.093	0.091	0.089	0.088	0.086	0.085	0.083	0.082	0.080
1.8	0.079	0.078	0.076	0.075	0.073	0.072	0.071	0.069	0.068	0.067
1.9	0.066	0.064	0.063	0.062	0.061	0.060	0.058	0.057	0.056	0.055
2	0.054	0.053	0.052	0.051	0.050	0.049	0.048	0.047	0.046	0.045
2.1	0.044	0.043	0.042	0.041	0.040	0.040	0.039	0.038	0.037	0.036
2.2	0.036	0.035	0.034	0.033	0.033	0.032	0.031	0.030	0.030	0.029
2.3	0.028	0.028	0.027	0.026	0.026	0.025	0.025	0.024	0.024	0.023
2.4	0.022	0.022	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018
2.5	0.018	0.017	0.017	0.016	0.016	0.015	0.015	0.015	0.014	0.014
2.6	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011	0.011
2.7	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008
2.8	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006	0.006
2.9	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.005
3	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003

Таблица 3С. Значения $\varphi = 2\arcsin\sqrt{p}$

***p*, %**	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
	0.000	0.063	0.089	0.110	0.127	0.142	0.155	0.168	0.179	0.190
1	0.200	0.210	0.220	0.229	0.237	0.246	0.254	0.262	0.269	0.277
2	0.284	0.291	0.298	0.304	0.311	0.318	0.324	0.330	0.336	0.342
3	0.348	0.354	0.360	0.363	0.371	0.376	0.382	0.387	0.392	0.398
4	0.403	0.408	0.413	0.418	0.423	0.428	0.432	0.437	0.442	0.448
5	0.451	0.456	0.460	0.465	0.469	0.473	0.478	0.482	0.486	0.491
6	0.495	0.499	0.503	0.507	0.512	0.516	0.520	0.524	0.528	0.532
7	0.536	0.539	0.543	0.546	0.551	0.555	0.559	0.562	0.566	0.570
8	0.574	0.577	0.581	0.584	0.588	0.592	0.595	0.599	0.602	0.606
9	0.609	0.613	0.616	0.620	0.623	0.627	0.630	0.633	0.637	0.640
10	0.644	0.647	0.650	0.653	0.657	0.660	0.663	0.666	0.670	0.673
11	0.676	0.679	0.682	0.686	0.689	0.692	0.695	0.698	0.701	0.704
12	0.707	0.711	0.714	0.717	0.720	0.723	0.726	0.729	0.732	0.735
13	0.738	0.741	0.744	0.747	0.750	0.752	0.755	0.758	0.761	0.764
14	0.767	0.770	0.773	0.776	0.778	0.781	0.784	0.787	0.790	0.793
15	0.795	0.798	0.801	0.804	0.807	0.809	0.812	0.815	0.818	0.820
16	0.823	0.826	0.828	0.831	0.834	0.837	0.839	0.842	0.845	0.847
17	0.850	0.853	0.855	0.858	0.861	0.863	0.866	0.868	0.871	0.874
18	0.876	0.879	0.881	0.884	0.887	0.889	0.892	0.894	0.897	0.900
19	0.902	0.905	0.907	0.910	0.912	0.915	0.917	0.920	0.922	0.925
20	0.927	0.930	0.932	0.935	0.937	0.940	0.942	0.945	0.947	0.950
21	0.952	0.955	0.957	0.959	0.962	0.964	0.967	0.969	0.972	0.974
22	0.976	0.979	0.981	0.984	0.986	0.988	0.991	0.993	0.996	0.998
23	1.000	1.003	1.005	1.007	1.010	1.012	1.015	1.017	1.019	1.022
24	1.024	1.026	1.029	1.031	1.033	1.036	1.038	1.040	1.043	1.045
25	1.047	1.050	1.052	1.054	1.056	1.059	1.061	1.063	1.066	1.068
26	1.070	1.072	1.075	1.077	1.079	1.082	1.084	1.086	1.088	1.091
27	1.093	1.095	1.097	1.100	1.102	1.104	1.106	1.109	1.111	1.113
28	1.115	1.117	1.120	1.122	1.124	1.126	1.129	1.131	1.133	1.135
29	1.137	1.140	1.142	1.144	1.146	1.148	1.151	1.153	1.155	1.157
30	1.159	1.161	1.164	1.166	1.168	1.170	1.172	1.174	1.177	1.179
31	1.182	1.183	1.185	1.187	1.190	1.192	1.194	1.196	1.198	1.200
32	1.203	1.205	1.207	1.209	1.211	1.213	1.215	1.217	1.220	1.222

p, %	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
33	1.224	1.226	1.228	1.230	1.232	1.234	1.237	1.289	1.241	1.243
34	1.245	1.247	1.249	1.251	1.254	1.256	1.258	1.260	1.262	1.264
35	1.266	1.268	1.270	1.272	1.274	1.277	1.279	1.281	1.283	1.285
36	1.287	1.289	1.291	1.293	1.295	1.297	1.299	1.302	1.304	1.306
37	1.308	1.310	1.312	1.314	1.316	1.318	1.320	1.322	1.324	1.326
38	1.328	1.330	1.333	1.335	1.337	1.339	1.341	1.343	1.345	1.347
39	1.349	1.351	1.353	1.355	1.357	1.359	1.361	1.363	1.365	1.367
40	1.369	1.371	1.374	1.376	1.378	1.380	1.382	1.384	1.346	1.388
41	1.390	1.392	1.394	1.396	1.398	1.400	1.402	1.404	1.406	1.408
42	1.410	1.412	1.414	1.416	1.418	1.420	1.422	1.424	1.426	1.428
43	1.430	1.432	1.434	1.436	1.438	1.440	1.442	1.444	1.446	1.448
44	1.451	1.453	1.455	1.457	1.459	1.461	1.463	1.465	1.466	1.469
45	1.471	1.473	1.475	1.477	1.479	1.481	1.483	1.485	1.487	1.489
46	1.491	1.493	1.495	1.497	1.499	1.501	1.503	1.505	1.507	1.509
47	1.511	1.513	1.515	1.517	1.519	1.521	1.523	1.525	1.527	1.529
48	1.531	1.533	1.535	1.537	1.539	1.541	1.543	1.545	1.547	1.549
49	1.551	1.553	1.555	1.557	1.559	1.561	1.563	1.565	1.567	1.569
50	1.571	1.573	1.575	1.577	1.579	1.581	1.583	1.585	1.587	1.589
51	1.591	1.593	1.595	1.597	1.599	1.601	1.603	1.605	1.607	1.609
52	1.611	1.613	1.615	1.617	1.619	1.621	1.623	1.625	1.627	1.629
53	1.631	1.633	1.635	1.637	1.639	1.641	1.643	1.645	1.647	1.649
54	1.651	1.653	1.655	1.657	1.659	1.661	1.663	1.665	1.667	1.669
55	1.671	1.673	1.675	1.677	1.679	1.681	1.683	1.685	1.687	1.689
56	1.691	1.693	1.695	1.697	1.699	1.701	1.703	1.705	1.707	1.709
57	1.711	1.713	1.715	1.717	1.719	1.721	1.723	1.725	1.727	1.729
58	1.731	1.734	1.736	1.738	1.740	1.742	1.744	1.746	1.748	1.750
59	4.752	1.754	1.756	1.758	1.760	1.762	1.764	1.766	1.768	1.770
60	1.772	1.774	1.776	1.778	1.780	1.782	1.784	1.786	1.789	1.791
61	1.793	1.795	1.797	1.799	1.801	1.803	1.805	1.807	1.809	1.811
62	1.813	1.815	1.817	1.819	1.821	1.823	1.826	1.828	1.830	1.832
63	1.834	1.836	1.838	1.840	1.842	1.844	1.846	1.848	1.850	1.853
64	1.855	1.857	1.859	1.861	1.863	1.865	1.867	1.869	1.871	1.873
65	1.875	1.878	1.880	1.882	1.884	1.886	1.888	1.890	1.892	1.894
66	1.897	1.899	1.901	1.903	1.905	1.907	1.909	1.911	1.913	1.916
67	1.918	1.920	1.922	1.924	1.926	1.928	1.930	1.933	1.935	1.937

p, %	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
68	1.939	1.941	1.943	1.946	1.948	1.950	1.952	1.954	1.956	1.958
69	1.961	1.963	1.965	1.967	1.969	1.971	1.974	1.976	1.978	1.980
70	1.982	1.984	1.987	1.989	1.991	1.993	1.995	1.998	2.000	2.002
71	2.004	2.006	2.009	2.011	2.013	2.015	2.018	2.020	2.022	2.024
72	2.026	2.029	2.031	2.033	2.035	2.038	2.040	2.042	2.044	2.047
73	2.049	2.051	2.053	2.056	2.058	2.060	2.062	2.065	2.067	2.069
74	2.071	2.074	2.076	2.078	2.081	2.083	2.085	2.087	2.090	2.092
75	2.094	2.097	2.099	2.101	2.104	2.106	2.108	2.111	2.113	2.115
76	2.118	2.120	2.122	2.125	2.127	2.129	2.132	2.134	2.136	2.139
77	2.141	2.144	2.146	2.148	2.151	2.153	2.156	2.158	2.160	2.163
78	2.165	2.168	2.170	2.172	2.175	2.177	2.180	2.182	2.185	2.187
79	2.190	2.192	2.194	2.197	2.199	2.202	2.204	2.207	2.209	2.212
80	2.214	2.217	2.219	2.222	2.224	2.222	2.229	2.231	2.234	2.237
81	2.240	2.242	2.245	2.247	2.250	2.262	2.255	2.258	2.260	2.263
82	2.265	2.268	2.271	2.273	2.276	2.278	2.281	2.284	2.286	2.289
83	2.292	2.294	2.297	2.300	2.302	2.305	2.308	2.310	2.313	2.316
84	2.319	2.321	2.324	2.327	2.330	2.332	2.335	2.338	2.341	2.343
85	2.346	2.349	2.352	2.355	2.357	2.360	2.363	2.366	2.369	2.372
86	2.375	2.377	2.380	2.383	2.386	2.389	2.392	2.395	2.398	2.402
87	2.404	2.407	2.410	2.413	2.416	2.419	2.422	2.425	2.428	2.431
88	2.434	2.437	2.440	2.443	2.447	2.450	2.453	2.456	2.459	2.462
89	2.465	2.469	2.472	2.475	2.478	2.482	2.485	2.488	2.491	2.495
90	2.498	2.501	2.505	2.508	2.512	2.515	2.518	2.522	2.525	2.529
91	2.532	2.536	2.539	2.543	2.546	2.550	2.554	2.557	2.561	2.564
92	2.568	2.572	2.575	2.579	2.583	2.587	2.591	2.594	2.598	2.600
93	2.606	2.610	2.614	2.618	2.622	2.626	2.630	2.634	2.638	2.642
94	2.647	2.651	2.655	2.659	2.664	2.668	2.673	2.677	2.638	2.642
95	2.691	2.695	2.700	2.705	2.709	2.714	2.719	2.724	2.729	2.734
96	2.739	2.744	2.749	2.754	2.760	2.765	2.771	2.776	2.782	2.788
97	2.793	2.799	2.805	2.811	2.818	2.824	2.830	2.837	2.844	2.851
98	2.858	2.865	2.872	2.880	2.888	2.896	2.904	2.913	2.922	2.931
99	2.941	2.952	2.963	2.974	2.987	3.000	3.015	3.032	3.052	3.078

Таблица 4С. Значения критерия t Стьюдента

Число степеней свободы, *df*	Доверительная вероятность (*P*) Уровень значимости (*α*)		
	***P* = 0.9** ***α* = 0.1**	***P* = 0.95** ***α* = 0.05**	***P* = 0.99** ***α* = 0.01**
2	2.920	4.303	9.925
4	2.132	2.776	4.604
6	1.943	2.447	3.707
8	1.860	2.306	3.355
10	1.712	2.228	3.169
12	1.782	2.179	3.055
14	1.761	2.145	2.977
16	1.746	2.120	2.921
18	1.734	2.101	2.878
20	1.725	2.086	2.845
25	1.708	2.060	2.787
30	1.697	2.042	2.750
35	1.690	2.030	2.724
40	1.684	2.021	2.704
50	1.676	2.008	2.678
60	1.671	2.000	2.660
70	1.667	1.994	2.648
90	1.664	1.986	2.631
100	1.660	1.982	2.625
120	1.658	1.980	2.617
∞	1.645	1.960	2.576

Пороговые значения распределения *t* Стьюдента; *α* для двустороннего критерия

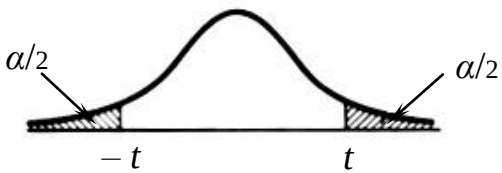


Таблица 5С . Значения критерия Фишера F при уровне значимости α=0.05 (df_1 – для числителя, df_2 – для знаменателя)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
6	6.0	5.1	4.7	4.5	4.4	4.3	4.2	4.2	4.1	4.1	4.0	4.0	3.9	3.8	3.7
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3	3.3	3.2	3.2	3.1	3.0
10	5.0	4.1	3.7	3.5	3.3	*3.2*	3.1	3.1	3.0	3.0	2.9	2.9	2.8	2.7	2.5
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.9	2.8	2.8	2.7	2.6	2.5	2.5	2.3
14	4.6	3.7	3.3	3.1	3.0	2.9	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.3	2.1
15	4.5	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5	2.5	2.4	2.3	2.2	2.1
16	4.5	3.6	3.2	3.0	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.3	2.3	2.2	2.0
17	4.4	3.6	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.4	2.3	2.2	2.1	2.0
18	4.4	3.5	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.3	2.3	2.2	2.1	1.9
19	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.5	2.4	2.4	2.3	2.2	2.2	2.1	1.9
20	4.3	3.5	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3	2.3	2.2	2.1	2.0	1.8
21	4.3	3.5	3.1	2.8	2.7	2.6	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.0	1.8
22	4.3	3.4	3.0	2.8	2.7	2.5	2.5	2.4	2.3	2.3	2.2	2.1	2.1	2.0	1.8
23	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3	2.2	2.1	2.0	1.9	1.8
24	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.0	1.9	1.7
26	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2	2.1	2.1	2.0	1.9	1.7
28	4.2	3.3	2.9	2.7	2.6	2.4	2.4	2.3	2.2	2.2	2.1	2.0	2.0	1.9	1.6
30	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2	2.1	2.0	1.9	1.8	1.6
40	4.1	3.2	2.8	2.6	2.4	2.3	2.2	2.2	2.1	2.1	2.0	1.9	1.8	1.7	1.5
60	4.0	3.1	2.8	2.5	2.4	2.2	2.2	2.1	2.0	2.0	1.9	1.8	1.7	1.6	1.4
120	3.9	3.1	2.7	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.8	1.7	1.7	1.6	1.2
∞	3.8	3.0	2.6	2.4	2.2	2.1	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.5	1.0

Пороговые значения распределения *F* Фишера

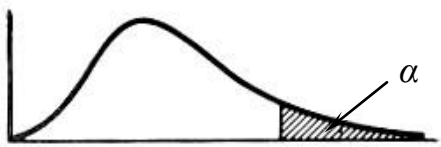


Таблица 6С. Значения критерия χ^2

df	Уровень значимости, α		
	0.25	**0.05**	**0.01**
1	1.32	3.84	6.63
2	2.77	5.99	9.21
3	4.11	7.81	11.34
4	5.39	9.49	13.28
5	6.63	11.07	15.09
6	7.84	12.59	16.81
8	10.22	15.51	20.09
10	12.55	18.31	23.21
12	14.85	21.03	26.22
14	17.12	23.68	29.14
16	19.37	26.30	32.00
18	21.60	28.87	34.81
20	23.83	31.41	37.57
22	26.04	33.92	40.29
24	28.24	36.42	42.98
26	30.43	38.89	45.64
28	32.62	41.34	48.28
30	34.80	43.77	50.89
40	45.62	55.76	63.69
50	56.33	67.50	76.15
60	66.98	79.08	88.38
70	77.58	90.53	100.42
80	88.13	101.88	112.33
90	98.64	113.14	124.12
100	109.14	124.34	135.81

Пороговые значения распределения χ^2 Пирсона

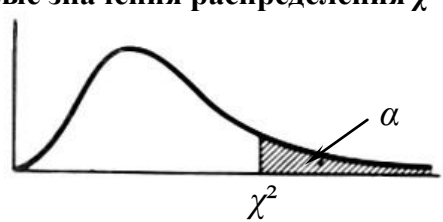


Таблица 7С. Критерий U Уилкоксона – Манна – Уитни

Уровень значимости $\alpha = 0.05$							
n	**4**	**5**	**6**	**7**	**8**	**9**	**10**
4	10	11	12	13	14	15	15
5		17	18	20	21	22	23
6			26	27	29	31	32
7				36	38	40	42
8					49	51	53
9						63	65
10							78

Уровень значимости $\alpha = 0.01$							
n	**4**	**5**	**6**	**7**	**8**	**9**	**10**
4			10	10	11	11	12
5		15	16	17	17	18	19
6			23	24	25	23	27
7				32	34	35	37
8					43	45	47
9						56	58
10							71

Таблица 8С. Минимальные значения коэффициента корреляции *r*, достоверно отличные от нуля ($df = n - 2$)

	α			α			α	
df	**0.05**	**0.01**	***df***	**0.05**	**0.01**	***df***	**0.05**	**0.01**
1	0.997	1	**18**	0.444	0.561	**50**	0.273	0.354
2	0.95	0.99	**19**	0.433	0.549	**60**	0.25	0.325
3	0.878	0.959	**20**	0.423	0.537	**70**	0.232	0.302
4	0.811	0.917	**21**	0.413	0.526	**80**	0.217	0.283
5	0.754	0.874	**22**	0.404	0.515	**90**	0.205	0.267
6	0.707	0.834	**23**	0.396	0.505	**100**	0.195	0.254
8	0.632	0.765	**24**	0.388	0.496	**125**	0.174	0.228
10	0.576	0.708	**25**	0.381	0.487	**150**	0.159	0.208
12	0.532	0.661	**26**	0.374	0.478	**200**	0.138	0.181
14	0.497	0.623	**27**	0.367	0.47	**300**	0.113	0.148
15	0.482	0.606	**28**	0.361	0.463	**400**	0.098	0.128
16	0.468	0.59	**30**	0.349	0.449	**500**	0.088	0.115
17	0.456	0.575	**40**	0.304	*0.393*	**1000**	0.062	0.081

УКАЗАТЕЛЬ ТЕРМИНОВ

Учебное издание

Коросов Андрей Викторович

Специальные методы биометрии

Редактор *О. В. Обарчук*

Рисунок на обложке: *Ю. М. Матросова*
Компьютерная верстка: *А. В. Коросов*

Подписано в печать 30.06.07. Формат 60 x 84 $^{1}/_{16}$.
Бумага газетная. Гарнитура Академическая.
22.0 уч.-изд. л.
Тираж 400 экз. Изд. № 126.

Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования

ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Отпечатано в типографии Издательстве
Петрозаводского государственного университета
185910, Петрозаводск, пр. Ленина, 33