Э. В. Ивантер, А. В. Коросов

Введение в количественную биологию

Министерство образования и науки
Российской Федерации
Государственное образовательное учреждение
высшего профессионального образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

Э. В. Ивантер
А. В. Коросов

Введение в количественную биологию

Рекомендовано
Министерством образования Российской Федерации
в качестве учебного пособия
для студентов биологических специальностей

Петрозаводск
Издательство ПетрГУ
2011

ББК 28.08:22.172
И 228
УДК 578.087.1

Рецензенты:
профессор, доктор биологических наук *Н. М. Окулова;*
доцент, доктор биологических наук *Н. С. Ростова*

Печатается по решению редакционно-издательского совета Петрозаводского государственного университета

Ивантер Э. В., Коросов А. В.
И 228 Введение в количественную биологию : учеб. пособие / Э. В. Ивантер, А. В. Коросов. — Петрозаводск : Изд-во ПетрГУ, 2011. — 302 с.
ISBN 978-5-8021-1231-1

Книга служит элементарным пособием для практического применения вариационной статистики в биологических исследованиях.

В краткой, доступной форме на конкретных примерах рассмотрены приемы количественной обработки материалов биологических наблюдений и экспериментов. Приводятся алгоритмы статистических расчетов, показаны принципы биологической интерпретации математических показателей, раскрыты основы статистического оценивания, проверки гипотез, применения методов корреляционного, регрессионного, дисперсионного, дискриминантного, кластерного анализов, метода главных компонент. Рассмотрен общедоступный метод имитационного моделирования в среде Excel. Книгой можно пользоваться, не имея специальной математической подготовки и не прибегая к более сложным руководствам по биометрии. Она содержит справочные таблицы и рекомендации по выполнению статистического анализа на ЭВМ с помощью пакетов Excel и StatGraphics.

Книга рассчитана на биологов различного профиля, научных и практических работников, студентов, аспирантов, преподавателей вузов и школ, специалистов сельского и лесного хозяйства, здравоохранения и ветеринарии.

ББК 28.08:22.172
УДК 578.087.1

ISBN 978-5-8021-1231-1

ВВЕДЕНИЕ

В процессе любых научных, особенно экспериментальных, исследований, как и во всех областях прикладной биологии (медицине, агробиологии, селекции, охотоведении, лесоводстве, биотехнологии и т. д.), мы всегда имеем дело с цифрами – данными о размерах, весе, возрасте, плодовитости организмов, продуктивности экосистем, урожайности сортов, соотношении между признаками, дозами факторов, различными диагностическими и иными тестами и прочими количественными показателями и числовыми характеристиками. За многообразим этих цифр прячутся конкретные закономерности, которые требуют объективной оценки и научного объяснения. И здесь самое широкое применение находят приемы биометрии – пограничной дисциплины, призванной с помощью соответствующего математического аппарата оценить разнообразные связи, зависимости и отношения между биологическими явлениями, объектами и процессами и показать реальность их существования.

Биометрия представляет собой инструмент, способный измерить значимость и надежность полученных результатов, заранее рассчитать и спланировать необходимую численность объектов для того или иного эксперимента, оценить достоверность проверяемой в эксперименте гипотезы, по части охарактеризовать целое, получить точную количественную характеристику изменчивости исследуемого показателя, определить степень и характер различий между признаками и процессами, выделить из множества воздействующих на явление факторов наиболее важные, измерить силу их влияния. Методологией количественной биологии является отделение случайного от закономерного, доказательство существования закономерного в видимом хаосе изменчивости. Это достигается посредством множества методов прикладного статистического анализа, основанных на знании закономерностей поведения случайных величин.

Игнорирование и недооценка статистической обработки полученного исследователем материала может свести на нет результаты многих важных опытов, привести к необоснованным или даже ошибочным заключениям. Напротив, умелое применение биометрических методов увеличивает информативную ценность проведенного иссле-

дования, обогащает экспериментатора новыми знаниями, помогает правильно планировать постановку опытов, глубоко разбираться в полученных данных, объективно оценивать результаты массовых наблюдений, выявлять скрытые закономерности и правильно их трактовать, что в конечном итоге делает биологию точной наукой.

При этом следует иметь в виду, что сама по себе статистическая обработка данных, как бы ни была она совершенна с точки зрения математики, не может служить гарантией качественности выполненного биологом исследования и не способна обеспечить надежности полученных им результатов, если само исследование проведено неправильно или использованные данные ошибочны. Более того, формальное применение математических методов, без понимания их сути и слепое использование ее, даже когда в этом нет никакой необходимости, может принести только вред. В работе биолога одинаково недопустимы как математический фетишизм, подмена биологических методов математическими, так и недооценка статистических приемов обработки.

Составляя настоящее руководство, мы попытались в возможно более простой форме изложить элементарные основы количественной биологии, разъяснить суть и назначение вариационно-статистической обработки количественных данных, помочь начинающему исследователю, не имеющему специальной математической подготовки, сознательно применять общедоступные методы биометрического исследования, познакомить его с порядком и способами расчета основных статистических показателей и принципами их биологической интерпретации. В книге обсуждаются возможности и перспективы применения различных статистических приемов, их достоинства и формы использования в повседневной практике биологических исследований. Сознательно отказавшись от строгого изложения математических аспектов теории биометрии, подробного объяснения и вывода сложных расчетных формул, мы сконцентрировали внимание на необходимом минимуме *статистических идей*, помогающих понять принципы биометрического анализа массовых явлений и характерных биологических задач, и прежде всего *на технике вычислений*. Рассмотрены только те статистические методы, которые авторы достаточно широко применяли в своих биолого-экологических исследованиях и на личном опыте убедились в их эффективности. Другие ме-

тоды статистического исследования приведены в специальных пособиях по вариационной статистике; некоторые из них указаны в списке рекомендуемой литературы (приемы описания биоразнообразия, анализ временных рядов и многомерное шкалирование рассмотрены в книге: Коросов, 2007).

Для каждого метода приведены алгоритмы ручного счета и примеры использования с этой целью пакета Microsoft Excel. Наша книга во многом ориентирована на использование этого пакета и содержит примеры работы в среде MS Excel. Для решения более сложных задач требуются специальные пакеты статистических расчетов, такие как StatGraphics (часть задач решалась в этой среде) или Statistica (см. пособие: Коросов, Горбач, 2010).

В конце книги приведены справочные таблицы, необходимые для статистической обработки данных, и предметный указатель.

Поскольку книга выполняет роль учебного пособия, вводимые понятия постепенно усложняются и главы лучше читать по порядку. Вместе с тем многие положения разделов 1 и 2 полезно перечитывать по мере овладения новыми методами расчетов. Эти главы содержат больший «методологический заряд», чем можно освоить при единственном прочтении в начале освоения курса. Многие мысли первых разделов становятся понятными только после приобретения некоторого опыта выполнения расчетов и начнут помогать только после нескольких повторных прочтений.

В новом издании заново отредактирован текст, исправлены обнаруженные ошибки, расширен круг примеров, изменены некоторые иллюстрации. Мы признательны всем читателям, приславшим свои замечания к рукописи, и с благодарностью примем новые. Наш адрес: korosov@psu.karelia.ru

1
ПРИНЦИПЫ КОЛИЧЕСТВЕННОЙ БИОЛОГИИ

Основные задачи количественной биологии

Биометрия – это инструмент эмпирического познания природы, в отличие от математической биологии, исследующей теоретические проблемы с помощью аналитического моделирования.

Методы количественной биологии (биометрия) призваны конкретизировать отображение биологических фактов, придать строгость биологическим выводам и прогнозам, способствовать целенаправленному исследованию биологических феноменов. Можно говорить о четырех основных задачах количественной биологии.

1. Задача количественного представления биологических фактов (измерение и сокращение размерности) – выразить свойства *отдельного* биологического объекта измерения в виде числа, варианты, значения переменной.
2. Задача обобщенного описания множества фактов (статистическое оценивание) – рассчитать показатели, параметры, которые полноценно отражают свойства *множества* однотипных объектов измерения, свойства выборки.
3. Задача поиска закономерностей (проверка статистических гипотез) – доказать неслучайность отличий между сравниваемыми совокупностями, объектами, показать реальность зависимости их характеристик от неких внешних или внутренних причин.
4. Задача исследования процессов (динамическое имитационное моделирование) – объяснить ход природного процесса множеством специфических отношений (выраженных уравнениями) между переменными биологического объекта и среды.

Для решения каждой из этих задач предлагаются достаточно простые, но эффективные способы, рассмотренные ниже.

Модель

Математическая статистика предлагает исследователю различные модели действительности, с помощью которых можно решать биометрические задачи разной сложности. В слове «модель» заключено только одно содержание: все, что мы думаем о действительности, есть ее отражение в нашем сознании, слепок, подобие. Мысль о природе есть ее модель.

Число – это тоже модель, способ мышления о существенных чертах объектов, отбор из бесчисленного множества его свойств лишь некоторых с указанием того или иного числового значения.

Модели в виде простой формулы часто используются в иллюстративных целях для краткого выражения неких общих мыслей. Таковы рассмотренные ниже понятийные *модели варианты*, на которых основаны разного рода статистические методы.

Для строгого описания действительности статистическая теория предлагает множество математических моделей. Центральной моделью выступает «закон нормального распределения» – функция, описывающая специфическое соотношение между значениями непрерывной случайной величины (t) и частотой (вероятностью) встречаемости ее значений (p):

$$p = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}$$

(формула плотности вероятности непрерывной случайной величины).

Когда говорят, что данный признак имеет нормальное распределение, подразумевается, что «стохастическое поведение» этой случайной величины очень хорошо описывается (аппроксимируется) приведенной формулой. Практика показывает, что эта формула подходит к очень большому числу количественных характеристик. Модель нормального распределения чаще других используют для описания случайных событий. Ее применение (предположение о «нормальности» изучаемых признаков) дает в руки исследователя-биолога множество полезных и удобных инструментов решения биологических задач. Это и интервальная оценка для прогноза ожидаемых значений случайной величины, и метод расчета наиболее теоретически обоснованных общих характеристик выборки (средних, дисперсий) и

показателей сопряженной изменчивости разных признаков (корреляции), и пр. На идее нормального распределения базируются конструкции всевозможных статистических критериев для сравнения параметров разных выборок и проверки статистических гипотез. Кроме нормального закона статистической наукой обнаружены другие виды поведения случайных величин, которые основаны либо на том или ином допущении о нарушении условий формирования нормального закона, либо на специфическом преобразовании случайной величины, исходно распределенной нормально.

Этапы биометрического исследования

Биология по большей части остается наукой эмпирической: сбор фактов в поисках закономерностей проявления природных феноменов доминирует над объяснением существа этих процессов, построением теории (особенно количественной) и прогноза. Поиски «закономерностей» в биологии явно превалируют над поисками «законов», в первом случае говорят об эмпирической (индуктивной) науке, во втором – о теоретической (дедуктивной). Методы, рассмотренные в книге, обслуживают потребности первого научного направления. При этом сохраняется надежда, что из обнаруженных закономерностей со временем «вырастут» биологические законы.

Математическая статистика, исследующая массовые проявления, служит средством доказательства существования той или иной закономерности, причинной обусловленности серии фактов. Факт сам по себе, раз случился, достоверен. Доказывать приходится достоверность существования причин, вызвавших факты к жизни и тем самым обеспечивающих их общность. Если наличие некоей причины обуславливает однотипность протекания биологических процессов, вызывает повторные появления сходных результатов, говорят о обнаружении закономерности. *Закономерное – это повторяющееся*, причем в зависимости от известных условий (причин). Биометрия представляет способы доказательства реальности эмпирических закономерностей. Они служат необходимым средством достижения биологом своих целей, установленных исходя из существа биологической проблемы. В этом смысле для биометрического исследования очень важна точная формулировка биологического вопроса.

Мало обнаружить закономерность, необходимо еще и показать ее реальность, а для этого следует оценить ее количественно. Статистический анализ как раз и служит этой двойной задаче: во-первых, численно охарактеризовать биологический объект, явление или процесс, его масштабы и тенденции и, во-вторых, доказать объективность его существования, достоверность отличия от других явлений или процессов. Опираясь на полученный научный материал, статистика способна доказать несостоятельность выдвинутых гипотез, отделить, как зерна от плевел, истинные отличия от случайных, привнесенных неучтенными факторами, вычленить реальную закономерность из обилия сырого экспериментального материала.

К сожалению, исследователи зачастую подменяют цели исследования средствами их решения, что понятно из такого типичного вопроса: «Вот мои данные, как их нужно статистически обработать?» Конструктивный диалог может начаться только после ответа на другой вопрос, зачем эти данные нужно как-то обрабатывать, зачем вообще они были собраны? Нам кажется, что такой диалог должен быть внутренним и обязан предварять не столько обработку, сколько сбор данных. Как писал отец эмпирической науки Ф. Бэкон, «правильно поставленный вопрос есть половина ответа». Цель исследования организует его. Спланировать способ обработки нужно перед сбором фактических данных!

Ввиду очевидной сложности этого процесса рассмотрим его основные этапы эмпирического исследования.

1. Определить объект исследования. Объект исследования – это не вид животного или растения, это исследуемый феномен со всеми относящимися к делу внешними компонентами, включая пространство (распространение) и время (динамика). Объектом биологии выступает жизнь – процессы жизнедеятельности, функционирования биосистем. Объектом частного биологического исследования выступает ограниченная во времени и пространстве биосистема. В частности, даже «фауна *N*-го района» – понятие динамическое.

2. Определить проблему и актуальность исследования. Проблема («Что плохо?») в научном плане есть отсутствие знаний об объекте исследования в определенной области его биологии. Потребность в недостающей информации появляется в том случае, когда уже имеются некоторые данные, обрисовывающие границы известного и

обнажающие края неизвестного. Актуальность формулируется в терминах уже известного по отношению к еще неизвестному знанию. Так, приступая к исследованию фауны некоей территории, можно предположить, что она населена, но кем именно и в каких количествах? – эта загадка и составляет проблему.

3. Определить цель исследования. Цель («Чего хочется?») в обобщенном виде характеризует итог исследования. Например, изучить видовой состав и численность животных на определенной территории в определенный временной промежуток есть общая цель фаунистического исследования. Только на этом фоне возможны обобщения на больших территориях и временах, т. е. обнаружение неких общих закономерностей. Научная деятельность не может не быть целесообразной, она должна вести к определенной цели. Она определяет шаги исследования, выбор средств и методов, планирование трудовых и финансовых затрат. Цель служит постоянным критерием эффективности выполненных действий, основой рефлексии, ограничителем.

4. Определить задачи исследования. Задачами («Что сделать?») отмечаются шаги к цели, это мост между ней и конкретными средствами ее достижения. Задачи могут быть как научного толка (тогда они предписывают конкретные действия, позволяющие решить частный вопрос специфическими методами), так и методические (определяющие пути разработки недостающих методических приемов работы или развитие инструментальной базы). Задачи – это руководства к действию, указания, как делать и что будет получено в результате, если предпринять такие-то действия.

Именно на этом этапе становится ясным, какими должны быть массивы собираемой количественной информации, вид количественных характеристик (переменных), их число, способы регистрации статуса объектов измерения и факторов среды, схемы опытов и т. п. Знание этих частностей необходимо, чтобы запланировать использование того или иного статистического анализа, предъявляющего свои требования к исходным данным. Точнее всего работают параметрические методы, но они требуют регистрации количественной информации в форме рациональных или натуральных чисел. Если же запланировать получение характеристик объектов в приблизительных полуколичественных шкалах (баллы, ранги) или вообще с помощью

только качественных признаков, то следует помнить, что в конце концов придется пользоваться более грубыми непараметрическими методами статистики.

Понятно, что разработка задач требует от автора предметного знания и опыта аналогичной работы. В реальности практически никогда не удается сделать все, что запланировано, но часто удается получить помимо требуемых важные побочные результаты. Это заставляет переформулировать дефиниции проблемы, целей и задач, увязывая части исследования в целостную систему. Подобная итерация, повторное переосмысление и переработка теоретических и методических основ исследования – норма научной работы.

5. Сбор и накопление данных, изучение биологического явления. При сборе данных важно помнить правило «единообразия и равновероятности» собираемых выборок, чтобы свести к минимуму субъективные и систематические ошибки, уменьшающие точность измерений. Это условие относится к способу формирования выборок, суть которого заключается в создании одинаковых условий наблюдения и обеспечении равной вероятности получаемых результатов: каждая варианта должна иметь возможность представлять весь спектр действующих факторов без ограничений; в противном случае состав выборки будет не гомогенным, и статистические законы будут проявляться «неправильно», что сделает невозможным применение точных статистических критериев.

6. Решение биометрической задачи. Статистические методы исследования требуют жесткой определенности формулировок. Чтобы добиться требуемой строгости, исходно рыхлое словесное описание биологического вопроса предварительно необходимо перевести на формальный язык статистики. После этого выполняются расчетные процедуры и отыскивается требуемый ответ. Можно говорить о следующих семи этапах решения биометрической задачи:

- *конкретизация*,
- *формализация*,
- *выбор вида статистической задачи*,
- *выдвижение нулевой гипотезы*,
- *решение по алгоритму*,
- *статистический вывод*,
- *ответ на вопрос*.

Конкретизация. Формулирование биологической задачи, требующей статистического решения, определения объекта исследования, характеристика условий (факторов, методов) получения выборки, определение численно выраженных свойств и признаков, явное определение отдельной варианты (объекта измерения) и всей выборки вариант. Подготовка данных для последующей обработки.

Формализация. Этот этап требует несколько отойти от биологического содержания задачи и дать ответы на два вопроса общего характера «Что доказать?» и «Что описано?», которые предшествуют выбору конкретного статистического метода.

Ответ на вопрос «Что доказать?» помогает явно назвать один из *четырех типов биометрических задач*: доказать *чужеродность* варианты (принадлежность к классу вариант), доказать *отличие двух выборок*, доказать *влияние фактора* (отличие нескольких выборок), доказать *зависимость признаков*.

Ответ на вопрос «Что описано?» заставляет сделать выбор того обобщенного показателя, который интересует исследователя: описание может касаться *величины* признака (оценивается средней), его *изменчивости* (оценивается дисперсией), *распределения* частот (выражается вариационным рядом), *выборки в целом* (выражается совокупностью ранжированных вариант).

Выбор вида статистической задачи. В зависимости от характера имеющихся данных, способа описания и установленной задачи подбирается тот или иной статистический метод. Именно здесь отчетливее всего проявляются уровень биометрической подготовки исследователя, его профессионализм и мастерство, наконец, чутье на адекватный статистический метод. В этом смысле биометрия выступает как своеобразное искусство постановки статистической задачи в отношении биологических проблем. Вместе с тем многие биолометрические задачи решаются по принципу аналогии. Это позволяет предложить «*Определитель статистического метода*», несколько формальных критериев подбора адекватного статистического приема (табл. 1.1), включая как раз те распространенные статистические приемы, что рассмотрены в настоящем пособии. С помощью этой таблицы можно предварительно подобрать метод, способный решить поставленную задачу, а затем уже непосредственно перейти к вычислительным процедурам по приведенным в книге алгоритмам.

Таблица 1.1

Что доказать?	**Что описано?**	**Метод**
Чужеродность варианты в выборке	Величина	Сравнение средней и варианты
Достоверность отличия двух выборок	Величина	Сравнение средних арифметических
	Изменчивость	Сравнение дисперсий
	Распределение частот	Сравнение эмпирического и теоретического распределений
		Сравнение двух эмпирических распределений
	Выборка в целом	Сравнение двух наборов значений
Достоверность отличия нескольких выборок	Величина	Дисперсионный анализ
	Изменчивость	Сравнение серии дисперсий
	Распределение частот	Сравнение нескольких эмпирических распределений
	Выборка в целом	Непараметрический дисперсионный анализ
Достоверность влияния фактора на признак	Величина	Дисперсионный анализ
Достоверность влияния признака на признак	Величина	Регрессионный анализ
Достоверность сопряженности варьирования двух признаков	Величина	Корреляционный анализ

Выдвижение нулевой гипотезы. Этот этап призван дать четкую статистическую формулировку поставленного вопроса. Нулевая гипотеза – это предположение об отношениях объектов, выраженное в терминах статистики и предназначенное для дальнейшей статисти-

ческой проверки. Во введении уже упоминалось, что математическая статистика изучает случайные события, процессы и явления, поведение случайных величин. При этом она пытается отделить случайность от закономерности, случайные причины от систематических, доминирующих. С позиций случайного, вероятностного характера явлений исходит и нулевая гипотеза.

В самой общей форме эта гипотеза звучит так: «Отличия недостоверны». Согласно ей, например, наблюдаемые отличия двух выборок являются случайными (различия между выборочными параметрами есть ошибки репрезентативности); в действительности обе выборки вместе составляют один и тот же однородный материал и принадлежат к одной генеральной совокупности. В процессе статистического анализа нулевая гипотеза либо отвергается (опровергается, отклоняется), и тогда различия считаются достоверными, либо принимается (сохраняется). Последнее, однако, не означает доказательства отсутствия различий, а лишь говорит о том, что при данном объеме и качестве материала различия остаются недоказанными. Опираясь на полученный в процессе научной работы материал, статистика способна лишь доказать выдвинутые гипотезы или же отсеять и отвергнуть те предположения, для которых недостаточно информации, отделить истинные отличия от случайных, привнесенных неучтенными факторами, вычленить реальную закономерность из обилия сырого экспериментального материала.

Решение по алгоритму. Реализация одного из алгоритмов статистических расчетов. Приведенные в книге алгоритмы вычислений, как правило, снабжены числовыми примерами, и их использование не должно вызывать особых затруднений. Однако при «ручном счете» возможны небольшие технические ошибки, способные, тем не менее, привести к неправильным результатам. Чтобы избежать таких ошибок или, по крайней мере, не допустить их при вычислениях, необходимо придерживаться нескольких правил. Так, арифметические ошибки нетрудно выявить, если еще до начала расчетов ориентировочно прикинуть ожидаемый результат. Полезно дважды пересчитывать рабочие формулы, меняя местами слагаемые и сомножители. При использовании стандартных формул целесообразно вначале выписать их в символьной форме и лишь затем подставлять числовые значения. Очень важно также не путать сумму квадратов (Σx^2) с квад-

ратом суммы ($(\Sigma x)^2$) вариант, объем выборки (n) с числом градаций или групп (k). Вероятность правильного ответа увеличится, если формировать таблицы вычислений по приведенному в книге алгоритму полностью. При этом полезно проверять схождение сумм по строкам и столбцам, а вычисленных величин – по модели анализа. Например, при вычислении критерия хи-квадрат сумма частот эмпирического распределения должна точно совпадать с суммой теоретических частот. На ошибку в расчетах, как правило, указывает большое различие эмпирических и теоретических частот распределения, а также несовпадение величины исходного признака с рассчитанным по регрессионной модели. Кроме того, подозрение на допущенную ошибку должны вызывать отрицательные суммы квадратов (за исключением регрессионного и корреляционного анализов) и минусовые значения критерия Стьюдента (его всегда берут по модулю), а также величины, в десятки и сотни раз превышающие табличные. Наконец, следует помнить, что если распределение количественных признаков приближается к нормальному, то стандартное отклонение примерно равно четверти от всего размаха выборки: $S \approx (\max - \min)/4$. Только распределение Пуассона имеет равные среднюю и дисперсию ($M \approx S^2$). Эффективен контроль за результатами и с помощью графических возможностей Excel. В частности, для контроля правильности применения критерия хи-квадрат необходимо сравнивать гистограммы эмпирических и теоретических частот.

Статистический вывод. Статистический вывод служит главным результатом статистического анализа – это заключение о справедливости или опровержение нулевой гипотезы. Строится он на основе сравнения полученной (эмпирической) величины статистического критерия с табличной (теоретической). Если вычисленные значения критерия больше табличного, говорят о достоверном отличии (влиянии, исключении), если же меньше, то нулевая гипотеза остается в силе. Это позволяет использовать статистический критерий для опровержения нулевой гипотезы. Когда статистический вывод отвергает нулевую гипотезу, отличия выборок считаются доказанными, если же не отвергает, то отсутствие отличий доказанным не считается. На практике для правильного статистического вывода можно воспользоваться упрощенной схемой сравнения эмпирических значений критерия с табличными (рис. 1.1). Числа 0.95 и 0.05 – это доверитель-

ная вероятность и уровень значимости (вероятность правильности или неправильности вывода). Разместив в этой схеме табличные и эмпирические значения критериев, нетрудно заметить случаи, когда вычисленная величина лежит правее табличной, в критической области; это говорит о достоверности отличий сравниваемых параметров.

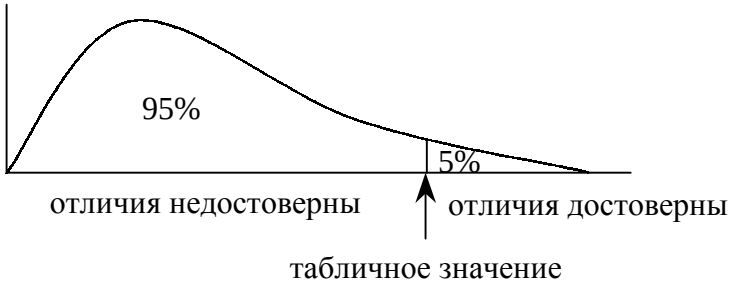


Рис. 1.1. *Схема использования критериев. Отмечены критические зоны для уровней значимости α = 0.05 и α = 0.01 (доверительные вероятности P = 0.95 и P = 0.99). Границами зон служат значения критериев из таблиц Приложения при данном уровне значимости. Если вычисленные величины критерия попадают в критическую зону (правее табличных), значит, отличие сравниваемых параметров достоверно (выборочные параметры оценивают разные генеральные параметры)*

Сказанное можно проиллюстрировать следующим примером. Пусть при сравнении двух средних арифметических нулевая гипотеза состояла в том, что отличие средних арифметических случайно. В расчетах было получено значение критерия $T = 3.5$. Табличная величина для этого случая равна $T = 2.1$. Поскольку полученное значение критерия (3.5) больше табличного (2.1), можно утверждать, что эти средние арифметические достоверно отличаются. Слово «достоверно» означает «статистически доказано»: отличие двух сравниваемых средних и без того бросалось в глаза, но лишь статистическое доказательство позволило распространять конкретный вывод на все явление – достоверность различия выборочных средних означает, что эти выборочные средние оценивают *разные* генеральные средние. Критерий доказал, что отличие средних не случайно, а закономерно.

Какую роль играют отмеченные на схеме значения вероятностей? Это станет ясным из следующих рассуждений. Статистический вывод можно сделать с разной степенью достоверности, иначе говоря, – с разной степенью уверенности, или вероятности. Можно быть уверенным в правильности вывода на 95% (тогда доверительная вероятность $P = 0.95$) или на 99% ($P = 0.99$). *Доверительная вероятность – это вероятность правильности статистического вывода.* Аналогично говорят о степени «неуверенности», иначе – об *уровне значимости*. Его значения обычно берут равными 5%, 1%, 0.1% (или соответственно $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$). Если точность проведения наблюдений или экспериментов невысока, если требуемый вывод не нуждается в особой точности (обычные условия проведения биологических исследований), то выбирается уровень значимости $\alpha = 0.05$. В таблицах *Приложения* приведены значения критериев при разных уровнях точности и числе степеней свободы. Чем выше требуется точность вывода, тем выше берут табличное значение критерия. Это понятно: чем точнее и ответственнее должен быть вывод, тем жестче требования к критерию. Подробнее статистический смысл уровня значимости объясняется в специальных математических руководствах. Для практического же понимания достаточно знать, что *уровень значимости – это вероятность ошибочности наших выводов.* С этой позиции 5% – достаточно мало.

Понятие числа степеней свободы – это число вариант (градаций, групп, случаев, т. е. объем выборки) без числа ограничивающих условий – конкретнее будет рассмотрено ниже.

Ответ на вопрос. Формулируется биологическое утверждение, доказанное статистически. Если удалось доказать достоверность неких отличий, то для биолога принципиально важна их направленность, не только факт отличий, например, средних арифметических, но и как именно они отличаются, какая величина превышает другую. Биологический ответ есть, по существу, перифраза статистического вывода, «одетого» в биологические термины и поэтому приобретающего биологический смысл и содержание.

7. Интерпретация результатов обработки. Биологическая интерпретация основывается на полученном статистическом выводе. Если он не отвергает нулевую гипотезу, то важных с биологической точки зрения заключений сделать нельзя. Дело в том, что, несмотря

на сохранение гипотезы о случайности отличия (влияния) показателей, мы не можем быть в этом полностью уверены. Возможно, в нашем распоряжении просто оказалось недостаточно данных, чтобы получить точный показатель и сделать достоверный вывод. В этой ситуации остается продолжить исследование, которое, впрочем, может быть спланировано более оптимальным образом.

Если же статистический анализ выявил достоверность отличия, влияния или необходимость выбраковки варианты из совокупности, то это дает основание сформулировать более содержательное и убедительное биологическое заключение, в частности, рассматривать выявленные отличия как результат действия какого-то систематического фактора, интерпретировать зависимость как биологическую закономерность, говорить об особых свойствах «выпадающей» из совокупности объекта, варианты.

Решить статистическую задачу, т. е. доказать достоверность отличий статистических параметров, не так уж и сложно, достаточно грамотно сформулировать ее условия и провести соответствующие вычислительные процедуры. Труднее установить, за счет чего эти различия возникли. Действительно ли это следствие объективной биологической закономерности или же результат неточно проведенного опыта, неконтролируемых (и неучтенных) условий, разных навыков у исполнителей и т. д. Для выяснения данного вопроса приходится контролировать всю информационную «атмосферу» в момент получения данных, как теоретические посылки, так и условия, при которых данные были получены. В этом случае удается правильно понять причины варьирования признаков, направления их изменчивости и, в конечном итоге, объяснить биологическое содержание формальных статистических выводов. Отсюда следует важная практическая рекомендация – если не контролировать, то хотя бы регистрировать факторы среды, в которой формируется выборка для исследования.

2
ВЫБОРКА И ЕЕ СТАТИСТИЧЕСКОЕ ОПИСАНИЕ

Биометрическое исследование в центр внимания всегда ставит выборку. В статистическом смысле выборка – набор чисел, множество значений случайной величины, совокупность вариант; отдельная варианта – это число. С предметной стороны варианта предстает как объект, носитель числа, а выборка – как группа объектов. В процессе формирования выборки участвует несколько агентов, которые необходимо иметь в виду для правильной интерпретации различий между выборками. Важная особенность выборки как множества значений случайной величины – это отличие отдельных вариант друг от друга, явление *изменчивости*.

Процесс формирования выборки

В поисках причин варьирования детально рассмотрим отдельную варианту, единичное значение – *число*.

Для понимания структурно-логической сущности *числа* в биометрическом исследовании требуется привлечение как минимум таких четырех понятий, как *объект, признак, фактор, метод*; вместе они образуют элементарный фрейм, логическую структуру минимального размера, необходимую для понимания существа процесса появления выборки.

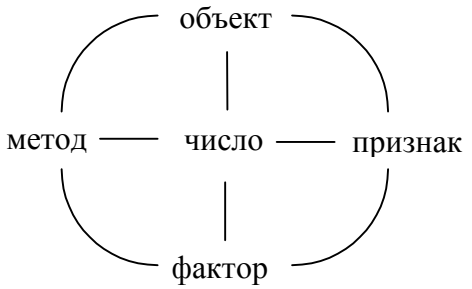


Рис. 2.1. *Элементарный фрейм биометрического исследования*

Число есть количественное выражение *признака* некоего *объекта*, полученного при данном уровне *фактора* внешней среды вполне определенным *методом*. С помощью этого фрейма очень

просто показать основные направления *тиражирования* чисел, т. е. набора множества вариант, формирующих выборки, а также основные трудности, с этим связанные.

Метод

Процедура получения чисел (вариант), включающая субъект, методику, инструмент их измерения и регистрацию. Простейший способ получения выборки – использование разных методов измерения одного и того же объекта. В этом случае отличия повторных примеров будут характеризовать разнокачественность применяемых методик, инструментов или уровни навыка участвующих исполнителей. При этом разные методы обладают разной способностью сообщить 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 выборки методически однородные. Рассмотренная тема приводит к очевидной рекомендации – для формирования сравнимых выборок использовать единую методику, одинаковый инструмент, «одни руки»; это, впрочем, далеко не всегда возможно.

Приступая к составлению выборки, метод ее получения следует соотнести с теми статистическими методами, что планируются для анализа количественных материалов, – не исключено, что выбранная процедура измерений не годится для формирования корректных выборок. Приемы грубого (чувственного) описания позволяют дать только грубые оценки – качественные, или баллы; точные инструментальные методы позволяют получать гораздо более эффективные характеристики в форме непрерывных признаков, дробных чисел. Так, балльные оценки можно статистически исследовать только с помощью непараметрических методов, тогда как для непрерывных количественных признаков можно использовать, кроме того, точные и высокоэффективные параметрические методы.

Важно отметить, что точность инструмента измерения и точность метода измерения – разные понятия. В первом случае говорят о технической характеристике. Под точностью *метода* подразумевается понятие точности (погрешности) измерительной процедуры,

т. е. возможность воспроизведения тех же результатов при повторном измерении одного и того же объекта. Помимо точности (состояния) прибора здесь фигурируют еще и навыки исследователя, и точность инструкции, и особенности условий проведения измерений (влажность, радиация и др.). Можно поэтому утверждать, что точность метода всегда ниже, чем точность инструмента. Это значит, что биологам нет смысла проводить измерения очень точными приборами, если сама процедура измерения предполагает широкое варьирование. В частности, длина тела мелких млекопитающих многими зоологами измеряется штангенциркулем. Во время измерения зверек лежит на столе. При этом у зверьков, попавших в давилки недавно, еще не проходит трупное окоченение, и их позвоночник физически невозможно «распрямить», тогда как мышцы немного «лежалых» зверьков расслабляются и позвоночник выпрямить просто. Промеры зверьков «разной свежести» обязательно дадут отличающиеся результаты с погрешностью 1–2 мм. Зачем в таком случае использовать штангенциркуль с ценой деления 0.1 мм, если удобнее (проще и быстрее) проводить измерения этих мелких животных на миллиметровой бумаге? На наш взгляд, точность измерительного инструмента (и трудоемкость измерения) должна быть соотнесена с погрешностью самой процедуры измерения. В любом случае выбор в пользу того или иного метода регистрации вариант (чисел) требует предварительной оценки их погрешности (причем разными исполнителями, дабы не превращать науку в искусство).

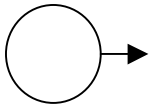
Признак

Признак (свойство, показатель, величина, характеристика, переменная) – 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.* Несмотря на то что точное описание поведения случайной величины получить нельзя, статистика способна выполнить *вероятностное* описание, позволяющее за множеством частных случаев увидеть их единство и дать довольно точные *интервальные* предсказания, ре-

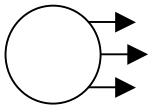
шить поставленные биологией вопросы. Максимально эффективно это можно сделать, если не упускать из вида требования к формированию выборок.

На этапе выбора (конструирования) признака следует иметь в виду ряд обстоятельств. Число свойств (признаков) любого объекта бесконечно, поэтому выбор того или иного признака должен хорошо соответствовать цели исследования. Довольно часто в биометрических исследованиях используются традиционные, общепринятые признаки («стандартные промеры»), что само по себе не гарантирует адекватности рассматриваемого признака целям данного исследования или планируемого способа статистической обработки. Например, традиционные зоологические промеры «длина тела», *Lt*, и «длина хвоста», *Lc*, имеют общую опорную точку на теле животного – передний край анального (клоакального) отверстия. Во время измерения кожа неизбежно натягивается и эта точка всегда смещается относительно тела, что одновременно сказывается на обоих названных промерах, причем прямо противоположным образом. Если по выборке таких промеров оценить средние, они будут адекватно реальности отражать обобщенное свойство выборки животных и могут быть использованы для статистических сравнений с другими выборками. Если же использовать методы, изучающие зависимости признаков (корреляционный, регрессионный), то обозначенная методическая погрешность синхронного искажения промеров будет приводить к появлению *ложной корреляции*, тем более сильной, чем «чище» выборка, чем более сходны животные друг с другом (например, группа одновозрастных однополых особей). В соответствии с биологическим смыслом корреляция между размерами тела и хвоста должна быть положительной (чем больше животное, тем больше у него хвост). Однако ложная корреляция будет отрицательной (чем больше промер тела, тем меньше длина хвоста), она будет вычитаться из общей и тем самым искажать представления об истинной зависимости между признаками. Избежать таких проблем можно, используя видоизмененные признаки, например сумму длины тела и хвоста, т. е. признак «длина позвоночника».

Подходя к формированию выборки, нужно определиться с числом регистрируемых признаков; если их будет несколько, каждая варианта (объект) окажется носителем нескольких значений.



Варианта с одним признаком



Варианта с тремя признаками

Увеличивая число зарегистрированных свойств, мы получаем возможность усложнять методы статистической обработки и от одномерных методов (описательная статистика) переходить к поиску зависимостей между двумя характеристиками (дисперсионный, регрессионный, корреляционный анализы) и многомерному анализу (кластерный, дискриминантный, компонентный анализы). Обычно регистрация нескольких признаков предполагает последующее применение корреляционного анализа. В этом случае имеет смысл позаботиться о том, чтобы признаки были одного вида (лучше, чтобы они были непрерывными).

Вариационная статистика может дать биологу множество эффективных *способов количественного описания* наблюдаемых явлений, которые позволяют с наименьшими ошибками получить точное статистическое (доказательное) суждение в рамках соответствующего статистического метода. Эти рекомендации относятся как к выбору статистического параметра, соответствующего цели, так и к способу количественного описания фактов.

Существует целый ряд методов регистрации признаков биологических объектов.

Качество (нечисловой дискретный признак) – простой, непосредственный, чувственный способ регистрации фактов; это статус, сезон, таксон, цвет, плотность, тип действия и пр. Значения таких признаков выражаются словами или символами, они не имеют количественного содержания и выражают принадлежность данного объекта к определенной обширной группе объектов (зеленый, январь, ♀, ♪).

Для обработки с помощью количественных статистических методов таким признакам придают количественное содержание разными способами. Простейший прием состоит в подсчете частоты встречаемости объектов разного качества в выборке. Так можно оценить соотношение числа особей разного пола в популяции, соот-

ношение объемов возрастных групп, видовое разнообразие в экосистеме.

Многие из качественных признаков оказываются следствием использования грубых (прикидочных, визуальных, чувственных) методов исследования, но их в принципе можно перевести в количественные показатели с помощью соответствующих процедур и приборов (это третий способ). Так, зоны загрязнения можно охарактеризовать в единицах концентрации вредных веществ, измеренных химическими или физическими методами; цвета спектра выражают в единицах длины волны электромагнитного излучения, ноты (звуки) – частотой колебаний в герцах и т. д.

Другой способ состоит в переводе качественных характеристик в полуколичественные, в ранги и баллы.

Ранг (номер) – дискретный полуколичественный признак, выражающий особенности объекта измерения относительно соседних с ним объектов другого качества. Процедура ранжирования применяется в алгоритмах непараметрической статистики. Ранжирование вариант – это присвоение порядкового номера (ранга) объектам, упорядоченным по степени увеличения или снижения выраженности какого-либо качества, воспринимаемого органами чувств. Ранг позволяет говорить только о факте отличия сравниваемых объектов, но не о степени этих отличий.

Например, серия генераций разновозрастных животных может быть обозначена как 1, 2, 3,… Если, в соответствии с этой шкалой, одной особи будет присвоен ранг 1, а другой – 3, то это означает всего лишь, что вторая особь старше первой, но вовсе не в три раза. Другой пример относится к косвенной (полуколичественной) характеристике зон загрязнения вокруг промышленного предприятия. Обычно по мере удаления от источника выбросов уровень загрязнения среды снижается. Это можно выразить, ранжируя некоторые зоны в порядке ослабления влияния как 1, 2, 3 и т. д.

Если же помимо общих соображений есть некие данные о степени загрязнения (по интенсивности запыления, угнетения растительности или другим признакам), то зоны загрязнения могут получить *балльную оценку*, например, 10, 2, 1.

Балл (оценка) – дискретный полуколичественный признак, численная характеристика объекта, присвоенная в соответствии с внешней заранее принятой шкалой (Перегудов, Тарасенко, 1981;

Зайцев, 1990). Вначале разрабатывается некая шкала баллов, учитывающая весь возможный диапазон изменчивости регистрируемых (чаще всего чувственно) качественных признаков и снабженная точными критериями различения объектов разного статуса, соответствующих разным баллам. Во время оценки объект соотносится с этими критериями и ему присваивается соответствующий балл. В отличие от рангов баллы сообщают не только порядок, но и чувственно различимую степень отличия градаций изучаемой характеристики. В нашем примере первая зона загрязнена существенно сильнее по сравнению со второй, чем вторая по сравнению с первой.

В качестве примера рассмотрим шкалу балльной оценки проективного покрытия чем-нибудь какой-либо поверхности. Зрительно человек хорошо отличает отсутствие покрытия (0 баллов – 0%) от единичных объектов (1 – 1–5%), единичные – от слабого покрытия (2 – 5–30%), слабое – от сильного (3 – 40–70%), сильное – от сплошного (4 – 90–100%). По этой причине соотношение между балльными и прямыми количественными оценками не прямо пропорциональное, а имеет степенное выражение (рис. 2.2).

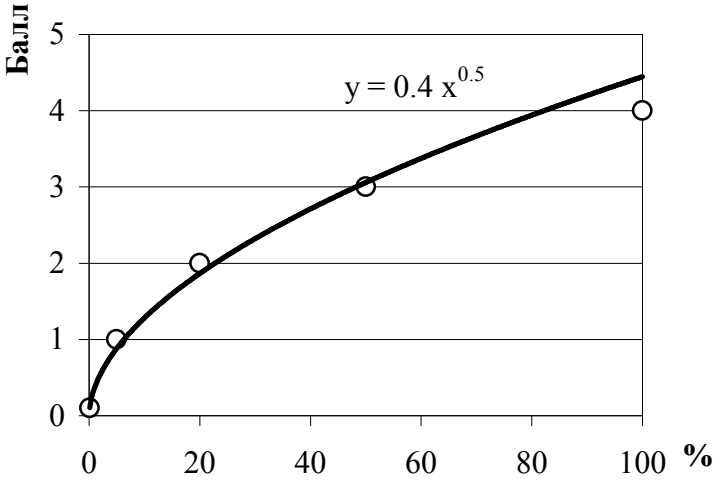


Рис. 2.2. *Соотношение между оценками и баллами*

Это значит, что баллы 2 и 4 не обладают свойствами чисел 2 и 4, в частности, балл 4 не в два раза больше балла 2, для них ариф-

метические и алгебраические операции применять нельзя, только логические операции сравнения.

По этой причине для статистической обработки балльных оценок требуются специальные, непараметрические, методы. Это значит, что для рангов и баллов нельзя обычными методами рассчитывать многие выборочные параметры, например средние и дисперсии. Точнее говоря, их рассчитывать можно, например, для иллюстративных целей. Но эти величины не будут обладать свойствами статистических параметров, в частности, их нельзя статистически сравнивать (с помощью критериев Стьюдента или Фишера). Корректно будет характеризовать выборки балльных оценок с помощью частотных распределений, моды, размаха изменчивости. Балльные оценки оказываются грубыми и приблизительными. В соответствии с этим и методы непараметрической статистики могут иметь только невысокую точность статистических выводов.

Известным хорошим компромиссом оказывается так называемая «шкала желательности», процедура преобразования качественных признаков в количественные с возможностью последующей обработки точными статистическими методами. Шкала желательности позволяет установить «соответствия между физическими и психологическими критериями» (Адлер и др., 1976, с. 36). С ее помощью любые характеристики среды (количественные или качественные) получают субъективную оценку исследователя, выраженную, тем не менее, числами в диапазоне от 0 до 1. В отличие от баллов *функция желательности* (*d*) является непрерывной величиной. Выраженность качества объектов наблюдения соотносят с заранее определенной целью или разной ролью значений изучаемых признаков в достижении этой цели. Чем более важно данное значение на пути к этой цели, тем более высокую оценку желательности оно получит.

При формировании шкалы функции желательности для отдельного признака неким стандартом служит шкала из 5 интервалов (Адлер и др., 1976, с. 36) (табл. 2.1). Каждому интервалу функции ставят в соответствие определенные уровни выраженности свойств объектов измерений. Характеристика выраженности признака в ключевых точках (0.2, 0.37, 0.63, 0.80) должна быть как можно более точной. В качестве примера приведена шкала желательности для оценки качества воды водоема в целях рекреации (Калинкина, 1989).

Таблица 2.1

Желательность	**Диапазон значений функции желательности**	**Пример шкалы желательности качества воды**
Очень хорошо	1.00–0.80	Чистая, прозрачная вода
Хорошо	0.80–0.63	Чистая, слегка желтоватая вода
Удовлетворительно	0.63–0.37	Темная вода или замутненная взвесью
Плохо	0.37–0.20	Мутная вода с легким неприятным запахом
Очень плохо	0.20–0.00	Грязная, пахнущая вода

После разработки шкалы с ее помощью можно проводить количественные оценки качества объектов. Полученный таким образом количественный признак оказывается непрерывным. Это свойство используется для объединения нескольких признаков, оцененных в разных шкалах желательности, в *обобщенную функцию желательности* (среднее геометрическое из *n* частных функций): $D = \sqrt[n]{d_1 \cdot d_1 \cdot .... \cdot d_n}$. В результате мы получаем интегральную характеристику, учитывающую значимость всех регистрируемых признаков. Продолжая наш пример, можно оценить рекреационное качество среды в целом, учитывая не только желательные характеристики воды, но и почвы (берега), воздуха, ландшафта, растительности и пр. Используя такой емкий показатель, можно гораздо точнее формулировать приоритеты научно-практической деятельности.

Сходный метод построения количественных шкал оценок *относительной важности* разных видов деятельности разработан в рамках метода анализа иерархий (см. подробнее: Коросов, 2007).

В заключение отметим, что показатели желательности или относительной важности являются близкими аналогами обобщающих характеристик, используемых в многомерных методах анализа (см. раздел 9).

Количество (число) – дискретный количественный признак (число натурального ряда), характеризующий множество однородных объектов, черт, деталей строения, состав (например, число эмбрионов у самки, число жаберных тычинок у рыб, число тычинок в

цветке, число деревьев на пробной площадке). Отдельную варианту получают, подсчитав число неких дискретных черт строения у отдельного объекта в пространстве ограниченного объема, а также в отдельной *пробе*. Это очень важное понятие. Оно дает одну из возможностей перевода качественных признаков в количественные и, кроме того, раскрывает смысл формирования частотных распределений разного типа. Для иллюстрации понятия «проба» рассмотрим умозрительный пример описания полового состава популяции животных. Если просто подсчитать число самок и самцов, то мы получим два числа, которые можно свести к одному – доле самок в процентах. Если же брать пробы, к примеру, по 10 особей, то число самок в разных пробах будет широко варьировать, создавая тем самым выборку различных вариант. Поскольку для чисел натурального ряда выполняются все операции арифметики, количественные признаки можно обрабатывать всеми параметрическими методами статистики. Для такой выборки можно рассчитать статистические параметры и проводить сравнение с параметрами других выборок.

Промер (ряд дробных, рациональных, чисел) – непрерывный (мерный) количественный признак, характеризующий свойства объектов с помощью различных дополнительных количественных шкал – температурной, весовой, размерной, объемной и т. п. Отдельная варианта получает количественную характеристику выраженности данного признака у данного объекта (в пределах точности метода): температуру тела, его размеры, уровень глюкозы в крови и т. д. Большинство методов статистики разработано для исследования именно таких непрерывных признаков (параметрические методы).

Объект

Объект – биологический феномен, на который направлено внимание исследователя. Здесь важно различать два понятия. *Объект исследования* – это общее понятие, обозначающее биологический предмет (организм, популяция, экосистема) или биологическое явление (размножение, динамика численности, сукцессия). В результате научной деятельности исследователь получает знание об объекте исследования. *Объект измерения* – это конкретный представитель объекта исследования (особь, группа особей в данной ме-

стности, результаты отловов, временные ряды), который непосредственно (материально) измеряется с помощью инструмента (органа чувств или прибора). В результате исследователь получает число, варианту. В дальнейшем объект измерения фигурирует как варианта. Варианта – это категория, обозначающая некий объект, отдельные свойства которого качественно или количественно охарактеризованы. Варианта может «нести» одно значение (объект охарактеризован одним признаком), два (учтены два свойства) и несколько (оценены несколько качеств).

Сформировать выборки можно, специально организуя процедуру регистрации все новых вариант. Один из приемов получения множества чисел – это отбор и измерение более или менее однородных объектов измерения, представляющих некий объект исследования. Отличие между такими вариантами имеет внутренний, эндогенный, источник – индивидуальные отличия по *статусу* и по *состоянию*. Например, животные одного возраста различны индивидуально, генетически, т. е. по статусу. Кроме того, каждое из них в разные годы, сезоны, время суток имеет разные морфофизиологические характеристики, т. е. отличается по состоянию. Следует отметить, что между статусом и состоянием нет непроходимой границы, как ее нет между объектом исследования и объектом измерения, – познание причин изменчивости живого и есть «погружение» во все более глубокие и тонкие отличия индивидуумов, в придании все более узкой специфике объекта измерения статуса объекта исследования. Важно отметить, что наиболее продвинутые в этом отношении науки (токсикология, биохимия, молекулярная биология) стремятся с помощью химической чистоты постановки опытов и выведения чистых линий подопытных животных убрать все мешающие причины «избыточного» варьирования.

Второй метод получения выборок – наблюдение объектов в разных условиях; источник отличий при этом внешний, определяющий разную реакцию на него изучаемых объектов (см. раздел *Фактор*).

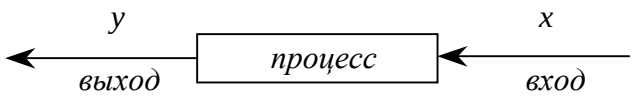
Зная природу объекта, можно правильно оценить соответствие изучаемой выборки требованиям статистической процедуры. Статистические методы ориентированы на изучение случайных величин разного типа. Желательно, чтобы они подчинялись нормальному закону распределения (непрерывные признаки) или биноми-

альному закону (дискретные признаки) (подробнее законы распределения рассмотрены ниже). Зачастую отклонение поведения случайных величин от этих законов связано с непродуманным способом получения выборок, с методическими ошибками и неточностями (хотя многие биологические признаки исходно имеют иные типы распределения). Для приведения распределения к более «чистому» виду нужно выявить, учесть, изучить факторы, влияющие на изменчивость вариант, и самые сильные из них по возможности ликвидировать. Тогда выборка будет точнее соответствовать объекту исследования. Унификацию выборки можно проводить, например, путем формирования из одной выборки двух и более выборок (с вариантами отчетливо разного статуса). Так, в морфологических исследованиях считается очевидно необходимым разделение животных по видам. Но не менее важно отдельно характеризовать самок и самцов, разновозрастных животных и даже представителей разных генераций с разными сроками рождения. Ликвидируя сильные причины варьирования, мы будем формировать выборки со все более «правильными» («нормальными») распределениями. Теоретически мыслима ситуация, когда учтены 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 факторов (внешних и внутренних по отношению к объекту исследования), практически не удается получить абсолютную повторяемость вариант. В лучшем случае множество несущественных причин обеспечит «хорошее» нормальное распределение.

Фактор

Фактор – это условия проведения наблюдений, среда существования объекта, возможная причина, определяющая текущее состояние объекта. Часто выделяют факторы эндогенные, внутренние (статус, способ существования объекта) и экзогенные, внешние (среда, условия существования объекта). Иными словами, граница между *признаком* (статусом) и *фактором* (эндогенным) достаточно условна. Разграничение этих понятий важно только с точки зрения организации вычислительной статистической процедуры. Фактор всегда есть активное, действующее начало, признак – его результат,

последствие. В кибернетической схеме, методической основе построения моделей, фактор есть вход, переменная *x*, признак – выход, переменная *y*: $y = f(x)$.



В биологической системе, содержащей много компонентов, где выход одного процесса является входом для следующего, понятия фактора и признака теряют свою определенность, поэтому их называют переменными, или же потоками. Тем не менее нам важно сохранить в нашем учебнике биометрии эти термины, поскольку они явным образом ориентируют на поиск причинной обусловленности явлений и вполне адекватно соответствуют биоэкологической терминологии.

С методической точки зрения есть факторы контролируемые и неконтролируемые. В первом случае степень проявления фактора точно устанавливается, а затем организуется получение выборок при разных заданных уровнях. Таковы условия проведения лабораторных экспериментов, когда имеется возможность сразу получать «чистые» выборки, не загрязненные эффектами действия посторонних факторов. Во втором случае, таковы натурные наблюдения, значения факторов неподвластны исследователю. Некоторые факторы он в состоянии регистрировать, другие – нет. Эта ситуация наиболее обычна для экологии, и важно понимать, как здесь может помочь статистика. Оказывается, что существуют методы, которые (при достаточно большом числе наблюдений в разных условиях) позволяют из всего многообразия эффектов действия факторов выделять интересующие исследователя (особенно эффективны дисперсионный и компонентный анализы). Самым важным при этом оказывается обязательная *регистрация* максимально возможного числа факторов (как внешних, так и внутренних), тогда появляется возможность исследовать их раздельное действие на объект. Современный путь биометрии – измерение большого числа признаков объектов и факторов их среды с последующим изучением зависимостей между ними методами многомерной статистики. Отсюда следует общая рекомендация при составлении выборки – определение по возможности всех условий ее получения. Например, для морфофизиологи-

ческого исследования нужно знать пол особи, ее год и месяц рождения, фазу динамики численности популяции, когда проводились отловы.

Важно помнить, что цель любого биометрического исследования всегда состоит в том, чтобы доказать достоверность действия какого-либо фактора, определить, влияет ли изменение дозы (силы действия) данного фактора на изменение значений данного признака. Так, сравнение двух выборок уже есть задача сравнения двух доз некоего фактора, представленных соответственно двумя группами вариант. Если групп вариант (доз, градаций фактора) несколько, то мы имеем возможность решить задачу оценки интенсивности этого влияния (дисперсионный анализ), а также задачу оценки характера влияния (регрессионный анализ).

Если фактор однозначно влияет на признак, то он называется систематическим или доминирующим, если влияет спорадически, он должен быть определен как случайный. Эти рассуждения приводят к первой *модели варианты*:

$$x_i = x_{di} \pm x_{ri},$$

где x_i – измеренное значение варианты,
i – индекс варианты ($i = 1, 2, \ldots, n$),
x_{di} – суммарный вклад j доминирующих факторов (*dominant*),
$x_d = \Sigma x_{dj}$,
x_{ri} – суммарный вклад k случайных факторов (*random*),
$x_{ri} = \Sigma x_{rk}$.

Вопрос о влиянии может быть поставлен только в отношении контролируемого или регистрируемого фактора, о неучтенных факторах (которых всегда достаточно много) нельзя сказать почти ничего. Такие неучтенные факторы, о которых нет информации, также относят к случайным. В то же время практика показывает, что эти факторы реально существуют и действуют, вызывая варьирование. Значит, общей перманентной задачей биометрического исследования остается поиск способа регистрации неучтенных факторов и доказательство не случайности их влияния.

Построение вариационного ряда

Мы считаем, что любое статистическое исследование должно начинаться с установления характера распределения изучаемых признаков. *Распределение – это соотношение между значениями случайной величины и частотой их встречаемости.* Статистическая теория началась с идеи подсчитать, как часто случается то или иное событие. Бóльшая повторяемость одних значений по сравнению с другими заставляет задумываться о причинах, о закономерностях наблюдаемых процессов. В качестве первичного описания любого явления может выступить частотное распределение. Если значения признака откладывать по оси абсцисс, а частоты их встречаемости по оси ординат, то можно построить *гистограмму*, частотную диаграмму, удобную для целей иллюстрации и исследования.

Основой для построения гистограммы служит вариационный ряд – представленный в виде таблицы ряд значений изучаемого признака (x), расположенных в порядке возрастания с соответствующими им частотами их встречаемости в выборке (a).

Начнем с примера изучения плодовитости серебристо-черных лисиц, которое дало следующие результаты (число щенков на самку): 5565564445646646455853655555636464625653763468635 56543847543165345674465646 5.

Для дискретного признака (плодовитость – число щенков у одной самки) построение вариационного ряда обычно не представляет сложности, достаточно подсчитать встречаемость конкретных значений.

Плодовитость, *x*	**Частота, *a***
1	1
2	1
3	8
4	16
5	23
6	21
7	3
8	3

Гистограмма, построенная по данным о плодовитости лисиц (рис. 2.3), сразу же обнаруживает характерное поведение случайной величины – высокие частоты встречаемости значений в центре распределения и низкие по периферии.

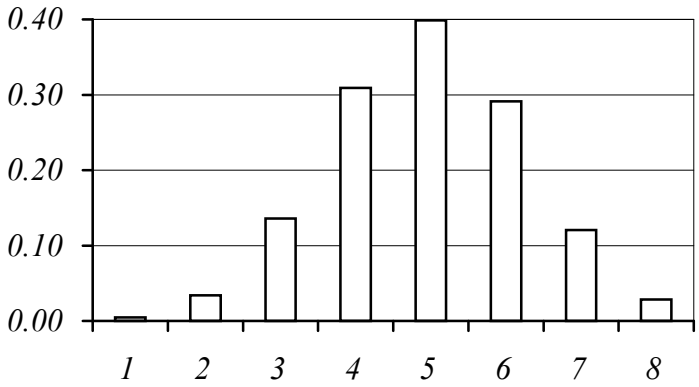


Рис. 2.3. *Распределение плодовитости лисиц*

Если же изучаемый признак непрерывен (таковы размерно-весовые характеристики), то для построения вариационного ряда сначала весь диапазон изменчивости признака разбивается на серию равных интервалов (классов вариант), затем подсчитывают, сколько вариант попало в каждый интервал. Число классов для больших выборок (n>100) должно быть не менее 7 и не более 12, их оптимальное число можно приблизительно определить по эмпирической формуле:

$k = 1+3.32 \cdot \lg(n)$, где n – объем выборки.

Составим для примера вариационный ряд для непрерывного признака – по данным о весе 63 взрослых землероек (г):

9.2	11.6	8.1	9.1	10.1	9.6	9.3	9.7	9.9	9.9	9.6
7.6	10.0	9.7	8.4	8.6	9.0	8.8	8.6	9.3	11.9	9.3
9.2	10.2	11.2	8.1	10.3	9.2	9.8	9.9	9.3	9.1	9.4
9.6	7.3	8.3	8.8	9.2	8.0	8.6	8.8	9.0	9.5	9.1
8.5	8.8	9.7	11.5	10.5	9.8	10.0	9.4	8.7	10.0	7.9
8.6	8.7	9.1	8.2	9.2	9.4	8.8	9.8			

	A
1	63
2	9.2
3	7.6
4	9.2
5	9.6
6	8.5
7	8.6

1) Все операции могут быть выполнены как вручную, так и с помощью функций Excel. Предвидя расчеты, на листе Excel данные лучше всего разместить в столбце (например, в блоке A2:A64). Далее следует определить объем выборки *n*, введя формулу в ячейку A1 и задав мышью диапазон: A1 =СЧЁТ(A2:A64).

2) Рассчитаем пределы размаха изменчивости значений, *лимит* (разность между максимальным и минимальным значением):

$Lim = Y_{max} - Y_{min} = 11.9 - 7.3 = 4.6$,

B1 =МАКС(A2:A64)–МИН(A2:A64).

3) Найдем число классов вариационного ряда по формуле:

$k = 1+3.32*\lg(63) = 6.973811 \approx 7$,

C1 =1+3.32*LOG10(A1).

4) Найдем длину интервала *dx* (допустимо округление):

$dx = Lim/\ k = 4.6/\ 7 \approx 0.7$,

D1 =B1/C1.

D2 =ОКРУГЛ(D1,1).

D	E
0.659611	
0.7	
7	
7.7	7.35
8.4	8.05
9.1	8.75
9.8	9.45
10.5	10.15
11.2	10.85
11.9	11.55
12.6	12.25

5) Установим границы классов; в качестве первой границы имеет смысл взять округленное минимальное значение (D3 = 7). Для расчетов на листе Excel удобно к значениям предыдущей границы прибавлять значение ширины интервала: D4 =D3+0.7 (или D4 =D3+$D$2); далее формулу следует ввести еще в семь ячеек, удобнее всего с помощью приема «*автозаполнение*»: D5 =D4+0.7 … (блок D5:D11).

6) Вычислить центральное значение признака в каждом классе. На листе Excel вычисления аналогичны рассмотренным в п. 4; исходным берется значение центра первого интервала:

E4 =СРЗНАЧ(D4:D3) , E5 =E4+0.7, …, E10 =E9+0.7.

7) Произвести разноску вариант в соответствующие классы с подсчетом их числа методом конверта (табл. 2.2):

1 2 3 4 5 6 7 8 9 10.

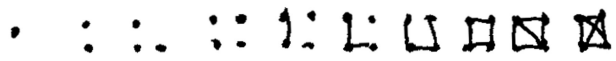


Таблица 2.2

Классы	**Центр классового интервала**	**Подсчет частот**	**Частоты, a**
7–7.7	7.35	:	2
7.8–8.4	8.05	𝍸	7
8.5–9.1	8.75	𝍸𝍸	18
9.2–9.8	9.45	𝍸𝍸:	22
9.9–10.5	10.15	𝍸	10
10.6–11.2	10.85	.	1
11.3–11.9	11.55	:.	3
Сумма			63

Для подсчета частот на листе Excel следует вызвать программу (макрос) построения вариационного ряда командой меню Сервис\ Анализ данных\ Гистограмма и заполнить окно. Каждое действие выполняется в два приема. Сначала нужно установить курсор в нужное окошко, щелкнув туда мышкой, затем мышкой же выделять соответствующие диапазоны ячеек листа Excel, нажимая левую кнопку над первой ячейкой диапазона и отпуская над последней (см. руководства к пакету Excel).

В качестве «Входного интервала» задать массив ячеек, содержащих исходные значения вариант (A2:A64). «Интервал карманов» – это блок значений правых границ классовых интервалов (D3:D11). Для «Выходного интервала» достаточно указать мышью одну ячейку (F3), это будет верхняя левая ячейка для блока результатов подсчета частот. После этого нажать ОК. Если все сделано правильно, появятся результаты, совпадающие с табл. 2.2. Однако необходимо помнить, что на листе Excel значения частот ставятся в соответствие не центрам классовых интервалов, но их правым (большим) границам.

Чтобы в дальнейшем не путаться, можно сразу переместить значения центров интервалов на место соответствующих карманов.

Для этого выделим диапазон E3:E11, перетащим на место F3:F11, подтвердив замену содержимого ячеек (рис. 2.4). Пустая ячейка E3 нужна для упрощения операции автоматического построения диаграммы – значения для оси абсцисс (первый столбец) не должны быть подписаны, а ячейка над значениями для оси ординат (второй столбец) должна содержать надпись.

	A	B	C	D	E	F	G
1	63	4.6	6.973810624	0.659611			
2	9.2	11.9		0.7			
3	7.6	7.3		7			*Частота*
4	9.2	9.6		7.7	7.35	7.35	2
5	9.6	1.15		8.4	8.05	8.05	7
6	8.5			9.1	8.75	8.75	18
7	8.6			9.8	9.45	9.45	22
8	11.6			10.5	10.15	10.15	10
9	10			11.2	10.85	10.85	1
10	10.2			11.9	11.55	11.55	3
11	7.3			12.6	12.25	12.25	0
12	8.8					Еще	0

Рис. 2.4. *Построение вариационного ряда в среде Excel*

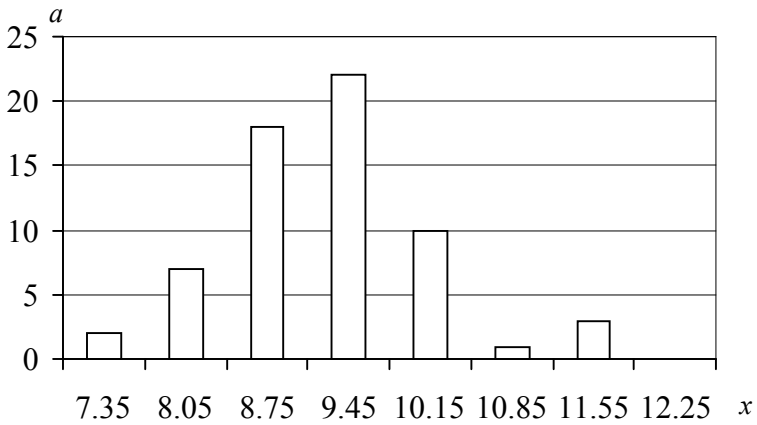


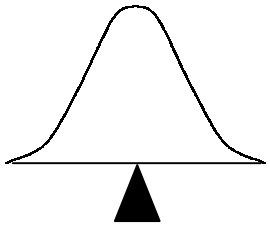
Рис. 2.5. *Распределение бурозубок по весу тела*

Теперь данные можно представить графически, в виде полигона частот (ломаной кривой) или гистограммы (столбиками). Выделим диапазон E3:F10 и с помощью Мастера диаграмм или кнопки Тип диаграммы построим Гистограмму или График (рис. 2.5). Отметим, что шкалирование осей диаграммы прошло автоматически.

Средняя (характеристика величины признака)

Одной из важнейших обобщающих характеристик вариационного ряда является средняя величина признака (часто обозначается буквой *M*). Существует несколько видов средних (средняя арифметическая – простая и взвешенная, средняя гармоническая, средняя квадратичная), но в практике биологических исследований наибольшее значение имеет средняя арифметическая, величина, вокруг которой «концентрируются» варианты.

Физической аналогией может послужить такой образ средней арифметической для признака с нормальным распределением: средняя – это та точка *вырезанного из картонки распределения*, опираясь



на которую левая и правая симметричные половинки уравновешивают друг друга.

Общая формула для определения величины средней арифметической – это отношение суммы значений всех вариант (x_i) выборки к их числу (объему выборки, n):

$$M = \frac{\sum x_i}{n}.$$

Средняя арифметическая характеризует действие систематических факторов, дающих равный вклад в каждую варианту выборки, исходя из рассмотренной модели:

$$x_i = x_{di} \pm x_{ri}.$$

Выполняя суммирование

$\Sigma x_i = \Sigma x_{di} + \Sigma(\pm x_{ri})$,

можно увидеть, что сумма случайных отклонений влево и вправо от средней в силу симметричности нормального распределения обращается в нуль ($\Sigma(\pm x_{ri}) = 0$). Значит, сумма вариант есть сумма эффектов действия только доминирующего фактора, одинакового для всех вариант ($\Sigma x_i = \Sigma x_{di}$). Средняя арифметическая есть поэтому характеристика действия доминирующего фактора на одну варианту: $M = \Sigma x_i / n = \Sigma x_{di}/n$. Модель варианты преобразуется: $x_i = M \pm x_{ri}$.

В среде Excel значение средней арифметической вычисляет функция =СРЗНАЧ(диапазон). Диапазоном может быть как один столбец, так и несколько. Для нашего примера с бурозубками средний вес составит С3 =СРЗНАЧ(А2:А64), M = 9.298412698. При расчетах статистических параметров следует помнить, что большое количество значащих цифр, рассчитанных ЭВМ, обычно не имеет никакого биологического смысла. Записывая такие статистические параметры, как средняя и стандартное отклонение, следует оставлять в лучшем случае на одну значащую цифру больше, чем имели значения вариант, а оценки ошибок – на две значащих цифры. Поскольку масса тела бурозубок колебалась от 7.3 до 11.9 г, средняя с учетом округления должна иметь вид: M = 9.3 г.

В биологических исследованиях зачастую встречается ситуация, когда требуется первичная статистическая обработка большого числа выборок, но необязательно с большой точностью. Это может понадобиться для предварительного рассмотрения и оценки материала, в частности для оперативного выявления общих тенденций его изменчивости, с тем, чтобы в дальнейшем перейти к специальным методам статистического анализа. Таковы, например, параметры многочисленных почвенных проб, результаты лабораторных анализов, морфологические характеристики разных групп животных, органов растений, физиолого-биохимические показатели и др. В этих случаях вычисление средней арифметической по предложенной формуле неоправданно из-за большой трудоемкости и неадекватной задачам исследования избыточной точности. Между тем знание закона нормального распределения позволяет предложить простой экспресс-метод расчета средней арифметической с использованием полученного для данной выборки размаха значений (Lim).

В случае нормального распределения средняя арифметическая находится точно по центру (совпадает со значением медианы), т. е. левая и правая границы распределения (с любой принятой вероятностью) находятся на одинаковом расстоянии от средней. В выборке объемом $n>30$ крайние значения удалены от средней на расстояние $2S$ (с вероятностью $P = 95\%$): $x_{min} = M–2S$, $x_{max} = M+2S$, и среднюю арифметическую можно рассчитать по формуле медианы:

$$M \approx Me = \frac{x_{min} + x_{max}}{2}.$$

Для бурозубок эта средняя составит $M = (7.3+11.9)/2 = 9.6$ г, что вполне соответствует действительности.

В случаях, когда необходимо объединить результаты расчетов по нескольким выборкам и на этой основе найти общую среднюю, характеризующую весь изученный материал, пользуются *взвешенной средней*, которая учитывает объемы частных выборок:

$$M = \frac{\sum n_j \cdot M_j}{\sum n_j},$$

где M_j – значение частной средней,

n_j – условные «веса» частного значения, объемы выборок.

Чтобы рассчитать среднюю взвешенную, необходимо значение средней арифметической помножить на его «вес», все эти произведения сложить и сумму разделить на сумму весов. Получены результаты определения средней величины выводка у рыжих полевок (экз. / самку) по месяцам: май 5.0, июнь 5.4, июль 6.2, август 6.0, сентябрь 4.5, причем известно число определений (самок) для каждого месяца: 22, 43, 103, 33 и 5. Средняя взвешенная составит:

$M = (5\cdot22+5.4\cdot43+6.2\cdot103+6\cdot33+4.5\cdot5)/ (22+43+ 103+33+5) = 5.8.$

Вычисление общей средней арифметической обычным способом дает в этом случае заниженную характеристику:

$M = (5+5.4+6.2+6+4.5)/5 = 5.4.$

Помимо средней арифметической важную область применения находит и *средняя квадратичная*. Ее употребляют при вычислении средних площадей, диаметров, радиусов, например, при расчете среднего размера клеток микроскопических водорослей, диаметра эритроцитов, величины листовой пластинки у растений, размеров колоний микробов и т. д. Средняя квадратичная равняется корню

квадратному из суммы квадратов вариант, отнесенной к их общему числу, и рассчитывается по формуле:

$$M = \sqrt{\frac{\sum x^2}{n}}.$$

Применение этой величины оправдано тем, что указанные признаки имеют несимметричное нормальное распределение, но обладают резко выраженной асимметрией. Возведение в квадрат сильнее сказывается на больших значениях, по сравнению с меньшими, частоты больших значений повышаются, распределение становится более симметричным, близким к нормальному, а средняя арифметическая для квадратов делит распределение пополам. Поэтому средние квадратичные обладают свойствами полноценных средних, тогда как простые средние арифметические, рассчитанные по таким данным, дают смещенные оценки.

Если, например, отдельные измерения диаметра эритроцитов дали следующие результаты: 7, 8, 10, 8, 11 и 6 мкм, то средний диаметр, найденный как среднее квадратичное, будет:

$$M = \sqrt{7^2 + 8^2 + 10^2 + 11^2 + 6^2)/6} = \sqrt{72.3} = 8.5,$$

тогда как простая средняя арифметическая дает величину 8.3.

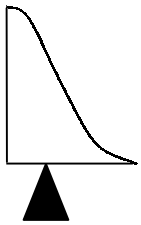
В число прочих констант вариационного ряда входит медиана (*Me*), значение, делящее размах выборки пополам, и мода (*Mo*), класс (или значение), представленный наибольшим числом вариант.

Стандартное отклонение (и другие показатели изменчивости)

Среднее квадратичное отклонение (или стандартное отклонение, *S*) – вторая по значению константа вариационного ряда. Она является мерой разнообразия входящих в группу объектов и показывает, на сколько *в среднем* отклоняются варианты от средней арифметической изучаемой совокупности.

Продолжим рассмотрение физической аналогии, предложенной для средней. Разрежем вырезанное из картонки нормальное распределение по вертикальной линии строго пополам, начиная с точки средней арифметической. Стандартное отклонение для признака с нормальным распределением – это та точка *половинки* выре-

занной из картонки фигуры распределения, опираясь на которую левая и правая *несимметричные* части уравновешивают друг друга.



Стандартное отклонение есть мера изменчивости признаков, обусловленная влиянием на них случайных факторов. Что такое «случайное» при детальном рассмотрении? В формуле модели вариант случайный компонент предстает в виде некой «добавки» к доле варианты, сформированной под действием систематических факторов, $\pm x_{случ.}$. Она, в свою очередь, складывается из эффектов влияния неопределенно большого числа факторов: $x_{случ.} = \Sigma\, x_{случ.j}$.

Каждый из этих факторов может обнаружить свое сильное действие (дать большой вклад), а может почти не участвовать в становлении варианты (слабое действие, незначительный вклад). По этой причине доля случайной «прибавки» для каждой варианты оказывается различной! Рассматривая с большим пристрастием какие-либо характеристики животных, например размеры дафний, можно увидеть, что одна особь крупнее, другая мельче, поскольку одна родилась на несколько часов раньше, другая позже, или одна генетически не вполне идентична прочим, а третья росла в более прогреваемой зоне аквариума и т. д. Если эти частные факторы *не входят в число контролируемых* при сборе вариант, то они, индивидуально проявляясь в разной степени, обеспечивают *случайное* варьирование вариант. Чем больше случайных факторов, чем они сильнее, тем дальше будут разбросаны варианты вокруг средней и тем большим оказывается характеристика варьирования, среднее квадратичное отклонение. Подчеркнем еще раз, что в контексте нашей книги термин «случайное» есть синоним слова «неизвестное», «неподконтрольное». Пока мы каким-либо способом не выразим интенсивность фактора (группировкой, градацией, числом), до тех пор он останется фактором, вызывающим случайную изменчивость.

Рассмотрим путь получения числовой характеристики изменчивости. Исходя из общей модели варианты $x_i = M \pm x_{ri}$, доля случайной изменчивости составит $x_{ri} = x_i - M$.

Простое обобщение (суммирование) эффектов действия случайных факторов для всей выборки невозможно ($\Sigma(\pm x_{ri}) = 0$), поэтому разность возводят в квадрат и затем извлекают из нее корень:

$$\sqrt{\sum x_{ri}^2} = \sqrt{\sum (x_i - M)^2} .$$

Отнеся полученное значение к объему выборки, получаем среднюю долю значения варианты, сформированной под действием всех случайных факторов:

$$S = \sqrt{\frac{\Sigma (x - M)^2}{n}} .$$

Эта формула могла бы служить для вычисления характеристики случайного варьирования, однако, как показано в математической статистике, она дает смещенные оценки, и более правильно применять другую формулу, использующую вместо объема выборки n число степеней свободы n–1.

Итак, величина стандартного отклонения выражается следующей *смысловой формулой*:

$$S = \sqrt{\frac{\Sigma (x - M)^2}{(n - 1)}} ,$$

где x – значение признака у каждого объекта в группе;
M – средняя арифметическая признака;
n – число вариант выборки.

Общая же *рабочая формула* расчета точного значения стандартного отклонения (заложенная и в алгоритм приведенной программы для ЭВМ) имеет следующий вид:

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{(n - 1)}} ,$$

где Σx^2 – сумма квадратов значений признака для всех вариант,
Σx – сумма значений признака,
n – объем выборки.

В среде Excel стандартное отклонение вычисляется с помощью функции =СТАНДОТКЛОН(диапазон). Для примера с массой тела бурозубок стандартное отклонение будет равно:

C4 =СТАНДОТКЛОН(A2:A64), т. е. S = 0.897216496;

после необходимого округления $S = 0.897$ г.

В некоторых случаях бывает необходимо определить *взвешенное среднее квадратичное отклонение* для суммарного распределения, составленного из нескольких выборок, для которых значения стандартных отклонений уже известны. Эта задача решается с помощью формулы:

$$S_{\Sigma} = \sqrt{\frac{\sum S^2(n-1)}{\sum n - k}},$$

где S_{Σ} – усредненная величина среднего квадратичного отклонения для суммарного распределения;

S – усредняемые значения стандартного отклонения;

n – объемы отдельных выборок;

k – число усредняемых стандартных отклонений.

Рассмотрим такой пример. Четыре независимых определения веса печени (мг) у землероек-бурозубок в июне, июле, августе и сентябре дали следующие величины стандартных отклонений: 93, 83, 50, 71 (при n = 17, 115, 132, 140). Подставив в вышеприведенную формулу нужные значения, получим стандартные отклонения для суммарной выборки (для всего бесснежного периода):

$$S_{\Sigma} = \sqrt{\frac{93^2 \cdot 16 + 83^2 \cdot 114 + 503^2 \cdot 131 + 71^2 \cdot 139}{404 - 4}} = 69.9.$$

В случае, если требуется первичная статистическая обработка большого числа выборок, но необязательно с большой точностью, для оценки стандартного отклонения можно воспользоваться экспресс-методом, основанным на знании закона нормального распределения. Как уже отмечалось, крайние значения для выборки (с вероятностью P = 95%) можно считать границами, удаленными от средней на расстояние $2S$: $x_{\min} = M - 2S$, $x_{\max} = M + 2S$. Это значит, что в лимите (Lim), в диапазоне от максимального до минимального выборочного значения, укладываются четыре стандартных отклоне-

ния: $Lim = (M+2S) - (M-2S) = 4S$. Однако этот вывод справедлив только по отношению к выборкам большого размера, тогда как для небольших выборок необходимо делать поправки. Рекомендуется следующая формула приблизительного расчета стандартного отклонения (Ашмарин и др., 1975):

$$S = \frac{x_{max} - x_{min}}{d},$$

где величина d взята из таблицы 2.3 (против соответствующего объема выборки, n).

Таблица 2.3

n	***d***	***n***	***d***	***n***	***d***	***n***	***d***
2	1.128	7	2.704	12	3.258	17	3.588
3	1.693	8	2.847	13	3.336	18	3.640
4	2.059	9	2.970	14	3.407	19	3.689
5	2.326	10	3.079	15	3.472	20	3.735
6	2.534	11	3.173	16	3.532	более	4

Выборочное стандартное отклонение веса тела бурозубок (n = 63), рассчитанное по приведенной формуле, составляет:

S = (11.9–7.3)/4 = 1.15 г,

что достаточно близко к точному значению, S = 0.89 г.

Использование экспресс-оценок стандартного отклонения значительно сокращает время расчетов, существенно не сказываясь на их точности. Отмечается лишь небольшая тенденция к завышению получаемых этим методом значений стандартного отклонения при небольших объемах выборок.

Стандартное отклонение – величина именованная, поэтому с ее помощью можно сравнивать характер варьирования лишь одних и тех же признаков. Чтобы сопоставить изменчивость разнородных признаков, выраженных в различных единицах измерения, а также нивелировать влияние масштаба измерений, используют так называемый *коэффициент вариации (CV)*, безразмерную величину, отношение выборочной оценки S к собственной средней M:

$$CV = \frac{S}{M} \cdot 100\%.$$

В нашем примере с весом тела бурозубок

$$CV = \frac{S}{M} \cdot 100\% = \frac{0.89}{9.3} \cdot 100\% = 9.6\ \%.$$

Индивидуальная изменчивость (варьирование) признаков – одна из наиболее емких характеристик биологической популяции, любого биологического процесса или явления. В связи с этим особенно важно правильно оценивать степень варьирования показателей, что представляется отнюдь не простой задачей, особенно в свете дискуссий о способах измерения и изучения изменчивости. Не затрагивая чисто методических аспектов проблемы и оставляя последнее слово за специалистами-математиками, следует, тем не менее, согласиться с мнением о том, что коэффициент вариации может считаться вполне адекватным и объективным критерием, хорошо отражающим фактическое разнообразие совокупности независимо от абсолютной величины признака. Индекс был создан для унификации показателей изменчивости разных или разноразмерных признаков путем приведения их к одному масштабу. Отнесением квадратичных отклонений к соответствующим средним мы переводим их в соизмеримые показатели и тем самым освобождаем от влияния величины самого признака. Практика показывает, что для многих биологических признаков наблюдается увеличение изменчивости (стандартного отклонения) с ростом их величины (средней арифметической). При этом коэффициент вариации остается примерно на одном и том же уровне 8–15%. За увеличение коэффициента вариации ответственны, как правило, растущие отличия распределения признака от нормального закона.

3
СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

Свойства нормального распределения

Биометрия изучает поведение биологических случайных величин. Начиная биологический эксперимент или приступая к наблюдению, невозможно точно сказать, каков будет результат – уровень численности животных в данном районе, вес еще не отловленных особей, количество сахара в крови через час после введения препарата и т. п. В этом смысле биологические явления случайны, точно не предсказуемы. Однако любому биологу ясно, что случайность эта не абсолютна. Несмотря на сложность точного прогноза, приблизительный результат можно предугадать, в частности, предсказав, что интересующая нас величина будет находиться в пределах некоторого интервала между конкретными минимальными и максимальными значениями. Ясно, например, что рост первого встречного взрослого человека вряд ли превысит два метра или будет меньше полутора метров. Такого рода прогноз можно дать, ориентируясь на повторяемость однотипных наблюдений, на распределение случайных величин. *Распределение – это соотношение между значениями случайной величины и частотой их встречаемости.* Если значения признака откладывать по оси абсцисс, а частоты их встречаемости – по оси ординат, то можно построить *гистограмму*, удобную частотную диаграмму. Изучая такие признаки, как размеры и масса тела, мы наблюдаем повторяемость одних значений и редкость встреч других. Как мы видели, это было характерно для веса тела землероек. При этом числовые значения вариант располагаются в некоторой ограниченной зоне, в центре которой их особенно много, а по краям мало. Такое распределение называют *нормальным*. Его форма помогает строить прогнозы.

Так, в случае продолжения отлова зверьков выше будет вероятность отловить таких новых особей, масса тела которых окажется ближе к центральному значению, чем к крайним. Знание математического закона распределения анализируемого признака позволяет предсказывать значения вариант много точнее – с некоторой вероят-

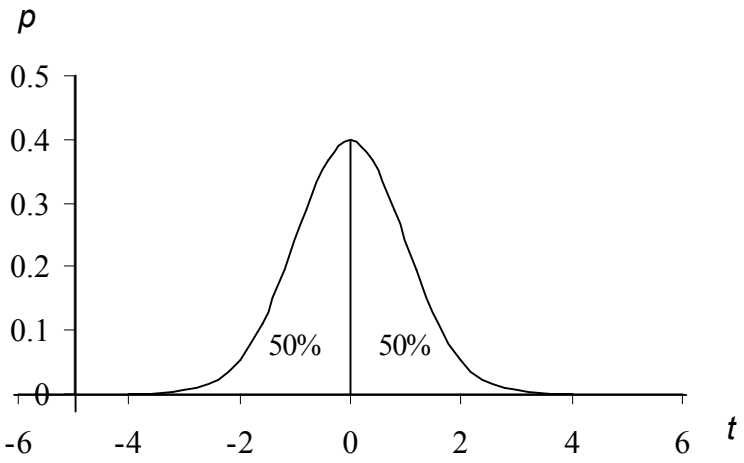
ностью. Закон нормального распределения случайной величины задан уравнением:

$$p = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2},$$

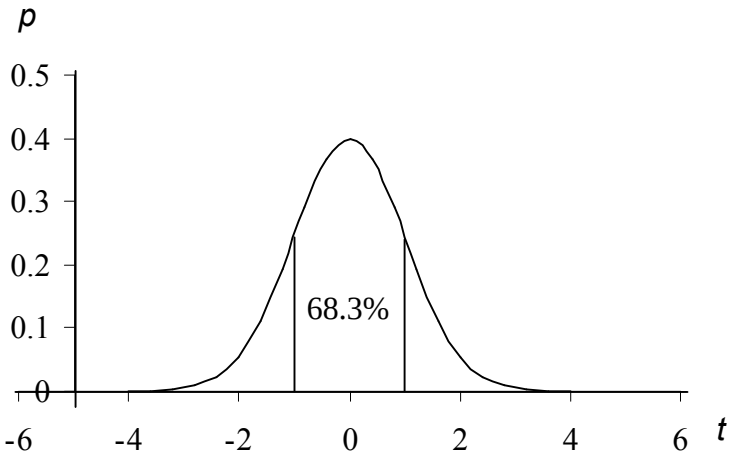
где $t = \frac{(x - M)^2}{S}$ – нормированное отклонение для конкретного признака; M, S – параметры нормального распределения.

Уравнение определяет ход кривой линии, имеющей характерную колоколообразную форму, т. е. позволяет вычислить *ординаты нормальной кривой*, или «плотность вероятности» (p). Вероятность – численная мера возможного; статистическая вероятность определяется как отношение числа вариант определенного вида к общему числу вариант. Поскольку нормальное распределение характерно для непрерывных случайных величин, говорят не о вероятности какого-то определенного значения варианты, но о «плотности вероятности», отражая тем самым плавность изменения вероятности значений для разных значений t, чем ближе к центру распределения, тем плотность вероятности выше. С помощью уравнения плотности вероятности можно рассчитать (интегрируя) вероятность появления нового значения случайной величины в том или ином интервале значений t. Итак, формула количественно выражает вполне определенные свойства поведения случайной величины, из которых можно назвать следующие практически важные следствия:

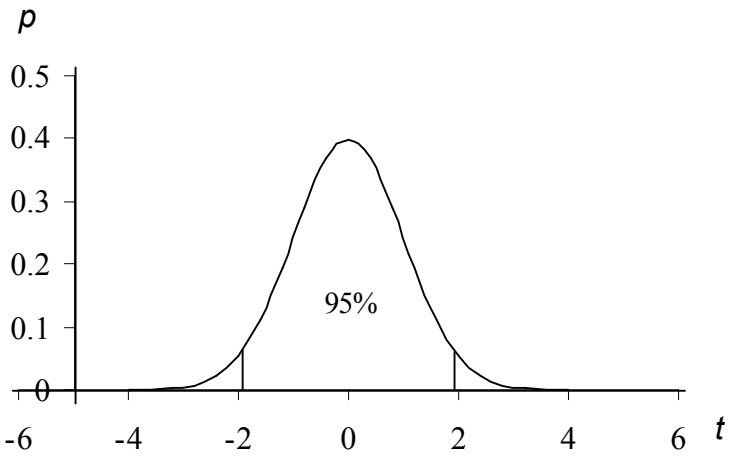
1. Все варианты лежат в интервале плюс-минус бесконечность. Иными словами, с вероятностью $P = 1$ ($P = 100\%$) мы вправе ожидать появление новой варианты в пределах от $-\infty$ до $+\infty$. Слева и справа от средней арифметической лежит по 50% вариант, т. е. с вероятностью $P = 0.5$ (50%) можно предсказать появление новой варианты в интервалах $M-\infty$ и $M+\infty$.



2. В интервале от M–1S до M+1S лежат 68.3% всех вариант; с вероятностью P = 0.683 (P = 68.3%) можно прогнозировать появление новой варианты на расстоянии ±1S от средней, или в диапазоне M±S.



3. Между M–1.96S до M+1.96S лежит 95% вариант. Это позволяет с 95%-ной вероятностью предполагать, что новая варианта окажется в интервале M±1.96S (округленно M±2S – так называемое правило двух стандартных отклонений).



4. С вероятностью $P = 0.99$ значение новой варианты будет заключено в пределах $M \pm 2.58S$ и с вероятностью $P = 0.999$ – в интервале $M \pm 3.3S$.

Исходя из сказанного, можно оценить вероятность появления новых значений признака. В отношении непрерывных случайных величин (метрических признаков) эта процедура сводится к так называемой интервальной оценке. Для полученных ранее характеристик, массы бурозубок, средней $M = 9.26$ и стандартного отклонения $S = 0.79$ (г), находим доверительные интервалы: $M \pm 1S = 9.26 \pm 0.78$, $M \pm 1.96S = 9.26 \pm 1.53$. Новое значение признака с вероятностью $P = 0.68$ ожидается в пределах 8.47–10.6 г., а с вероятностью $P = 0.95$ – между 7.68 и 10.82 г. Предсказание веса землероек, конечно, не имеет большого практического значения и приводится нами исключительно для иллюстрации. Гораздо важнее может быть прогноз численности ценных промысловых видов, сельскохозяйственных вредителей, вспышек болезней, урожая культурных растений и т. п. Эти прогнозы также основаны на оценке доверительной вероятности ожидаемого события.

Важнейшее значение для практического применения имеет «соглашение о 95%». В соответствии с ним совокупности, состоящей из 95% особей (объектов), мы доверяем так же, как и 100%-ной. Высказывание «*доверительная вероятность* $P = 0.95$» означает, что, согласно принятому допущению, 95% вариант достаточно полно характеризуют изучаемое явление (в данном случае изменчивость массы

тела землероек), что позволяет ограничиться рассмотрением вариант в области $M \pm 1.96S$, охватывающей эту 95%-ную совокупность. Так, мы принимаем, что нормальный вес землероек данного вида может изменяться в пределах 7.7–10.8 г, не больше и не меньше. За этими пределами мы обнаруживаем животных иного вида или статуса.

В этом контексте понятие «доверительная вероятность» есть вероятность того, что сделанный нами статистический вывод верен, соответствует действительности. При этом в биометрии обычно довольствуются доверительной вероятностью $P = 0.95$ (уровень значимости $\alpha = 0.05$), хотя в наиболее ответственных исследованиях принимают и более строгие уровни – $P = 0.99$ и $P = 0.999$. Однако это имеет смысл лишь при очень больших выборках исходных данных, точно описывающих закономерности изменчивости признаков. Обычно же выборки не очень велики, что позволяет ограничиться меньшей степенью доверительной вероятности $P = 0.95$.

«Уровень значимости» – понятие, альтернативное доверительной вероятности, это вероятность того, что статистический вывод не верен. Она составляет разность между единицей и доверительной вероятностью ($\alpha = 1–P$). Для доверительной вероятности 0.95 уровень значимости составляет 0.05, а для 0.99 и 0.999 – соответственно 0.01 и 0.001. Уровень значимости, равный 0.05 (5%), можно интерпретировать так: имеется всего 5% шансов, что полученная величина не будет соответствовать изучаемой совокупности. Уровень значимости – это тот теоретический процент вариант нормального распределения, который можно отбросить, не учитывать, дабы с меньшими усилиями получить основную информацию об изучаемом явлении. Можно целую жизнь положить на попытки отловить обыкновенную землеройку-бурозубку весом 2.5 г, но так и не собрать выборку, достаточную по объему, чтобы это реализовать (миллионы особей). Поэтому использование доверительной вероятности и уровня значимости можно назвать средством (теоретической базой) разумного ограничения материала (времени и масштабов исследования), позволяющего получить достоверное общее знание за счет исключения ничтожной доли частной (излишне конкретной) информации. В итоге такой прием дает возможность найти границы нормальной изменчивости изучаемых признаков, отбрасывая ошибочные, наведенные и артефактные значения.

Генеральная совокупность и выборка

Генеральная совокупность – все варианты одного типа. В предметной биологии это понятие можно интерпретировать как мыслимое множество вариант, сформированных при одинаковых (внешних и внутренних) условиях. Например, чистая линия рачков-дафний, выращенных при температуре помещения 20 ºС. При этом вполне может статься, что кроме двух выборок в природе 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. Так, можно вырастить 20 дафний при температуре 30 ºС; они также составят выборку из генеральной совокупности, которой физически не существует. Важнейшее свойство генеральной совокупности состоит в том, что на всех ее вариантах сказываются одни и те же систематические и случайные факторы, их набор уникален для данной генеральной совокупности. Для другой генеральной совокупности он будет другим. Мысленно меняя условия, можно сформировать бесконечное множество бесконечных по объему генеральных совокупностей, отличающихся нюансами условий своего формирования. Вот почему статистические задачи вполне корректно можно ставить и в терминах генеральной совокупности (сравнение выборок из разных генеральных совокупностей), и в терминах факториальной обусловленности (сравнение выборок, сформированных при действии разных факторов).

Кстати сказать, мыслимая бесконечность генеральной совокупности означает, что мы никогда не может познать ее до конца, в действительности мы всегда имеем дело с выборками. Исследовать свойства бесконечного числа значений случайной величины вполне доступно математике, которая на основании открытых законов их поведения предлагает эффективные процедуры для описания и сравнения случайных величин, наблюдаемых в действительности.

Выборочная совокупность, выборка – это множество вариант одного типа, ограниченное способом отбора (методами получения вариант), изъятое из генеральной совокупности.

Отличие выборок от генеральной совокупности состоит не только в разных объемах, или в реальности первых и сюрреалистич-

ности вторых. Дело в том, что в отдельной выборке *в полной мере* не могут проявиться все факторы, действующие в генеральной совокупности. Если доминирующий 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 и т. д. В результате такого влияния варианты, оставаясь в целом единообразными (влияние доминирующих причин), все же будут отличаться друг от друга (влияние случайных причин). При подсчете средней арифметической разнонаправленные случайные воздействия в целом нейтрализуют друг друга, но до конца – никогда. Все равно в разных выборках какие-то случайные факторы будут выражены сильнее, чем остальные.

Каждая новая выборка обязательно будет отличаться от предыдущей *в силу случайности*, варианты новой выборки буду нести одинаковый отпечаток действия доминирующих факторов, но разные следы действия случайных факторов. По этой причине параметры (*M, S*) разных выборок из одной генеральной совокупности никогда не совпадут ни друг с другом, ни со значениями генеральных параметров (обычно обозначаемых буквами μ, σ), они будут немного отличаться, смещаясь относительно друг друга и варьируя вокруг оценок генеральной средней и генерального стандартного отклонения.

Отличие генеральных значений от выборочных оценок состоит в том, что в первом случае они рассчитаны по всем вариантам, а во втором – по ограниченному их числу. Интуитивно понятно, что чем меньше объем выборок, тем менее точными будут выборочные оценки генеральных параметров, и, напротив, чем больше выборка, тем ближе выборочные средние и дисперсии лежат к генеральным значениям. Это явление называется «закон больших чисел»: *с ростом числа наблюдений значения выборочных параметров стремятся воспроизвести генеральные*.

Ошибка репрезентативности выборочных параметров

По части никогда не удается полностью охарактеризовать целое, всегда остается вероятность того, что оценка генеральной сово-

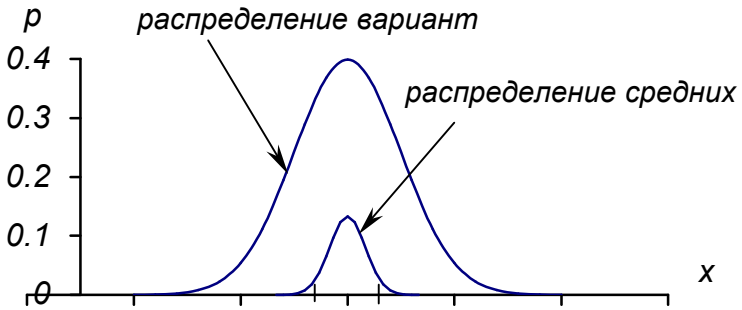
купности на основе выборочных данных недостаточно точна, имеет некоторую большую или меньшую ошибку. Такие ошибки, представляющие собой ошибки обобщения, экстраполяции, связанные с перенесением результатов, полученных при изучении выборки, на всю генеральную совокупность, называются ошибками репрезентативности (репрезентативность – степень соответствия выборочных показателей генеральным параметрам). Отличия значений выборочных параметров от генеральных называются ошибкой репрезентативности данного параметра, или просто (статистической) ошибкой.

Проиллюстрируем это примером. Из лабораторной культуры взяли 8 выборок по 10 одновозрастных дафний. У каждой из них промеряли длину тела. По каждой группе получены следующие средние значения (M): $M_1 = 4.09$, $M_2 = 3.85$, $M_3 = 3.88$, $M_4 = 3.94$, $M_5 = 3.86$, $M_6 = 3.89$, $M_7 = 3.97$, $M_8 = 3.90$ мм.

Общая средняя $M_{общ.} = 3.92$ мм. Несмотря на то что измерялись особи из одной культуры (одинаковые для всех генотипы и условия содержания), получены разные средние величины. Эти отличия и есть ошибки репрезентативности, связанные с неточностью оценок по небольшим выборкам. Если теперь мы найдем среднеквадратичное отклонение этих отдельных средних от общей, оно будет характеризовать средний диапазон отклонения выборочных оценок от генеральных значений. В данном случае показатель изменчивости средних составляет $S_M = 0.078513 \approx 0.08$ мм. Эта величина называется ошибкой средней арифметической (или стандартной ошибкой) и является, по существу, средним квадратичным отклонением множества выборочных средних от генеральной средней. На практике обычно нет возможности делать несколько выборок и вычислять несколько выборочных средних, чтобы по ним проводить расчеты. Статистическая теория показывает, что ошибка средней в $\sqrt{n}$ раз меньше, чем стандартное отклонение. Значит, ошибку можно рассчитать для единичной отдельной выборки по формуле:

$$m = \frac{S}{\sqrt{n}}$$ (S_M обозначается как m).

Используя это уравнение, были рассчитаны ошибки для разных выборок нашего примера, которые принимали значения от 0.05 до 0.11 мм, что оказалось близко к точной величине ошибки $S_M = 0.08$.



Статистические ошибки служат мерой тех пределов, в которых выборочные частные оценки могут отклоняться от параметров генеральной совокупности. Как следует из конструкции расчетной формулы, величина ошибки тем больше, чем больше варьирование признака (S) и чем меньше выборка (n). При увеличении объема выборки ошибки репрезентативности стремятся к нулю (следствие закона больших чисел).

Ошибку репрезентативности имеют все статистические параметры, рассчитанные по выборке: средняя, стандартное отклонение, коэффициент вариации, показатели асимметрии и эксцесса. Для разных типов распределений расчетные формулы могут немного изменяться. Для нормального распределения они имеют следующий вид.

Ошибка средней: $m_M = \dfrac{S}{\sqrt{n}}$,

ошибка стандартного отклонения: $m_S = \dfrac{S}{\sqrt{2 \cdot n}}$,

ошибка коэффициента вариации: $m_{CV} = \dfrac{CV}{\sqrt{2 \cdot n}}$.

Вычисленные значения ошибок подставляют к соответствующим параметрам со знаками плюс-минус (параметр ± ошибка) и в такой форме представляют в научных отчетах и публикациях.

Вернемся к примеру с весом тела бурозубок и определим соответствующие ошибки

средней арифметической: $m_M = \dfrac{0.89}{\sqrt{63}} = 0.113039$, M = 9.3±0.11 г;

стандартного отклонения: $m_S = \frac{0.89}{\sqrt{2 \cdot 63}} = 0.07993$, $S = 0.89 \pm 0.079$ г;

коэффициента вариации: $m_{CV} = \frac{9.6}{\sqrt{2 \cdot 63}} = 0.859613$, $CV = 9.6 \pm 0.9$ %.

Используя понятие ошибки репрезентативности, можно показать, почему в формуле расчета выборочной оценки стандартного отклонения (см. стр. 43) используется число степеней свободы $n-1$ вместо объема выборки n. Выборочная дисперсия S^2 оценивает генеральную дисперсию σ^2 неточно и отличается от нее в среднем на величину ошибки m^2_S : $\sigma^2 = S^2 - m^2_S$. В то же время известно, что ошибка в n раз меньше выборочной дисперсии $m^2_S = S^2/n$. Отсюда

$\sigma^2 = S^2 - S^2/n = n \cdot S^2/n - S^2/n = S^2 \cdot (n-1)/n$,

$\sigma^2 = S^2 \cdot (n-1)/n$ или $\sigma^2 \cdot n = S^2 \cdot (n-1)$.

Иными словами, выборочная дисперсия должна быть несколько больше, чем дает формула без учета ошибки, т. е. формула для ее расчета должна включать в знаменатель число степеней свободы $n-1$ вместо объема выборки n.

Не следует путать статистическую ошибку с методическими ошибками и ошибками точности (точности измерений, анализов, подсчетов и т. д.), хотя методические погрешности и увеличивают ошибку репрезентативности, но другим путем – методические огрехи увеличивают изменчивость признака, стандартное отклонение. Чем лучше взята выборка, чем больше ее размеры, т. е. чем вернее отражает она генеральную совокупность (все явление, весь процесс в полном объеме), тем меньше статистическая ошибка и расхождение между значениями признаков в выборочной и генеральной совокупностях. При всей неизбежности статистической ошибки она может быть сведена к минимуму отбором достаточного числа особей (вариант). С ростом объема выборки оценки параметров стабилизируются, а их ошибки репрезентативности уменьшаются.

Доверительный интервал

При конкретных биологических наблюдениях параметры генеральной совокупности остаются неизвестными, о них судят по выборочным оценкам, используя для этого величину ошибок репрезента-

тивности. *Интервал, в котором с заданной вероятностью ожидается присутствие генерального параметра, называется доверительным интервалом*, а его границы – доверительными. Это *интервальная оценка* генерального параметра, позволяющая предсказывать конкретный диапазон значений, в котором находится генеральный параметр. Теоретические исследования поведения выборочных средних (как случайных величин) показали, что они подчиняются нормальному закону, большинство из них (95%) находится поблизости от генеральной средней – в диапазоне $M_{ген.}$±1.96·m. Это обстоятельство позволяет делать обратное заключение – генеральная средняя арифметическая находится в диапазоне $M_{выбор.}$±1.96·m. В соответствии с законом нормального распределения можно ожидать, что генеральный параметр окажется в интервале

от M–T·m до M+T·m,

где m – ошибка средней арифметической,

T – квантиль распределения Стьюдента (табл. 6*П*) при данном числе степеней свободы (df) и уровне значимости (обычно α = 0.05).

Сказанное можно перефразировать так: с вероятностью P = 0.95 можно ожидать, что генеральная средняя находится в доверительном интервале M±T·m, построенном вокруг выборочной средней арифметической M.

Возвращаясь к примеру о весе землероек-бурозубок, мы теперь можем записать доверительные интервалы при разных уровнях вероятности (граничные значения T взяты для случая n = ∞):

для P = 0.95 M±T·m = 9.3±1.96·0.11 = 9.3±0.21 г;

для P = 0.99 M±T·m = 9.3±2.58·0.11 = 9.3±0.28 г;

для P = 0.999 M±T·m = 9.3±3.30·0.11 = 9.3±0.36 г.

Таким образом, искомая генеральная средняя величина веса землероек с вероятностью P = 95% находится в пределах 9.11–9.53 г, с вероятностью P = 99% – 9.04–9.6, для P = 99.9% – 8.96–9.68 г.

Если объем выборки, для которой были получены параметры и вычислялась ошибка репрезентативности m, был невелик (n<500), то необходимо вводить поправки на объем выборки, расширяя область возможного пребывания генерального параметра. Это понятно, поскольку при дефиците информации любые заключения не могут быть очень точными. Рассчитаем доверительный интервал для тех же данных, но с объемом n = 20 экз.

Ошибка средней арифметической составит

$m_M = \frac{0.89}{\sqrt{20}} = 0.19901$ г, $M = 9.3 \pm 0.2$ г.

При уровне значимости $\alpha = 0.05$ и числе степеней свободы $df = n-1 = 20-1 = 19$ табличная величина статистики Стьюдента равна $T = 2.09$, тогда доверительный интервал составит:

$M \pm T \cdot m = 9.3 \pm 2.09 \cdot 0.2 = 9.3 \pm 0.41$ г – от 8.9 до 9.7 г.

Аналогичным образом можно построить доверительный интервал для стандартного отклонения ($S \pm Tm_S$), коэффициента вариации ($CV \pm Tm_{CV}$), а также других статистических параметров (коэффициентов асимметрии, эксцесса, регрессии, корреляции), рассмотренных в следующих разделах.

Определение точности опыта

Статистическая ошибка позволяет судить о надежности полученных результатов, т. е. о том, достаточное ли количество случаев при данной величине изменчивости было получено, чтобы по части характеризовать целое. В практике биометрического анализа используется относительная ошибка измерений – «показатель точности опыта» или «показатель точности оценки параметров» – отношение ошибки средней к самой средней арифметической, выраженное в процентах:

$$\varepsilon = \frac{m}{M} \cdot 100\% .$$

Чем точнее определена средняя, тем меньше будет ε, и наоборот. Точность считается хорошей, если ε меньше 3%, и удовлетворительной при $3\% < \varepsilon < 5\%$. Если относительная ошибка превышает 5%, то полученные данные следует уточнить (повторить опыт, собрать дополнительный материал и т. д.).

В разобранном выше примере определения характеристик массы тела для выборки бурозубок показатель точности составил $\varepsilon = (0.11/ 9.3) \cdot 100 = 1.2\%$, что говорит о достаточной надежности выборочной оценки.

Оптимальный объем выборки

В биологических исследованиях, при планировании экспериментов, для определении величины подопытных и контрольных групп часто заранее требуется установить число наблюдений, достаточное для получения правильного представления о явлении в целом (получить репрезентативные оценки генеральной совокупности). Можно говорить о двух разных подходах для непрерывных и дискретных признаков.

Идея первого метода состоит в том, чтобы, используя известные соотношения (все формулы были представлены выше) между средней, стандартным отклонением, ошибкой средней, плотностью вероятности распределения Стьюдента (на их основе вычисляется коэффициент вариации, показатель точности оценок, доверительный интервал), найти число степеней свободы, соответствующее доверительному интервалу для средней при уровне значимости $\alpha = 0.05$. Иными словами, решается задача, прямо противоположная рассмотренной в предыдущем разделе.

Объем выборки, достаточной для получения результата заданной точности, находят по формуле:

$$n = \left(\frac{T \cdot CV}{\varepsilon} \right)^2,$$

где n – объем выборки,

T – граничное значение из таблицы распределения Стьюдента (табл. 6*П*), соответствующее принятому уровню значимости при планируемом объеме выборки (в крайнем случае можно взять значение $T = 1.96$ для $df = \infty$),

CV – приблизительное значение коэффициента вариации (%),

ε – планируемая точность оценки (погрешности) (%).

Рассчитаем необходимый объем условной выборки, обеспечивающий хорошую точность $\varepsilon = 3\%$, для уровня значимости $\alpha = 0.05$ ($T = 1.98$, для $df \approx 100$) и для коэффициента вариации $CV = 12\%$ (такова относительная изменчивость многих размерно-весовых признаков животных):

$$n = \left(\frac{1.98 \cdot 12}{3} \right)^2 = 62.726 \approx 63 \text{ экз.}$$

Определим объем выборки, необходимый для оценки среднего веса землероек-бурозубок при условии, что доверительный уровень вероятности должен составить 99% ($\alpha = 0.01$), показатель точности оценки 3% при сохраняющемся уровне изменчивости $CV = 10\%$.

По таблице Стьюдента (табл. 6*П*) в соответствии с заданным уровнем значимости ($\alpha = 0.01$) и для того же числа наблюдений находим $T = 2.62$. Далее вычисляем необходимый объем выборки:

$$n = \left(\frac{2.62 \cdot 10}{3} \right)^2 = 76.27 \approx 76 \text{ экз.}$$

Если исследуется фенотипическое (видовое) разнообразие, то может возникнуть задача определения минимального объема выборки, в которой будет присутствовать хотя бы один экземпляр с определенным фенотипом (Животовский, 1991). 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:

$$N = \frac{\ln(1-P)}{\ln(1-\pi)}.$$

Значение π можно определить приблизительно по имеющимся данным. 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. В частности, при уровне вероятности $P = 0.95$ и предположительной частоте фенотипа в популяции $\pi = 0.05$ потребуется

$$N = \frac{\ln(1-0.95)}{\ln(1-0.05)} = 58.4 \approx 58 \text{ экз.},$$

чтобы отловить хотя бы одну особь с этим дискретным признаком. Нетрудно рассчитать небольшую таблицу, содержащую оценки необходимых объемов выборок для разной вероятности отлова и различных частот фенотипов в популяции.

		Частота признака в популяции, π							
		0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001
Вероятность прогноза, P	0.95	4	6	8	13	28	58	298	2994
	0.99	7	9	13	21	44	90	458	4603

В случае исследования нескольких фенотипов существует достаточно сложный метод расчета объема выборки, в которой с вероятностью P присутствуют все фенотипы, частоты которых в популяции не меньше, чем π_{min}. Фрагмент такой таблицы может оказаться полезным для примерного определения нужных объемов выборок.

		Частота признака в популяции, π_{min}						
		0.5	0.4	0.3	0.2	0.1	0.05	0.01
Вероятность прогноза, P	0.95	6	7	11	21	51	117	754
	0.99	8	10	15	28	66	149	916

Асимметрия и эксцесс

В практике биологических исследований нередки случаи, когда числовые значения признаков дают распределения, в той или иной мере отличающиеся от нормального. Иногда обнаруживается асимметричное, в других сериях – эксцессивное распределение (рис. 3.1, 3.2). Для асимметричных вариационных кривых характерно появление «хвоста» – сдвиг частот от средних значений вправо или влево. В распределении эксцессивных признаков наблюдается чрезмерное накапливание или, наоборот, снижение частот в центральных классах вариационного ряда, вследствие этого вершина кривой распределения либо сильно поднимается и заостряется (положительный эксцесс), либо, напротив, опускается, приобретая вид широкого плато (отрицательный эксцесс, туповершинность) или даже седловины между двумя боковыми вершинами. Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю.

При расчете коэффициентов асимметрии и эксцесса используются следующие базовые формулы:

$$A = \frac{1}{n} \cdot \frac{\sum (x - M)^3}{S^3},$$

$$E = \frac{1}{n} \cdot \frac{\sum (x - M)^4}{S^4} - 3.$$

Однако для ответственных случаев разработаны более сложные формулы, дающие несмещенные оценки:

$$A = \frac{n}{(n-1) \cdot (n-2)} \cdot \frac{\sum (x - M)^3}{S^3},$$

$$E = \frac{n \cdot (n+1)}{(n-1) \cdot (n-2) \cdot (n-3)} \cdot \frac{\sum (x - M)^4}{S^4} - \frac{3 \cdot (n-1)^2}{(n-2) \cdot (n-3)}.$$

Эти последние формулы реализованы в среде Excel в виде функций: для оценки асимметрии =СКОС(диапазон) и для оценки эксцесса =ЭКСЦЕСС(диапазон). Диапазон ячеек должен содержать все значения изучаемой выборки.

Конфигурация кривых распределения отражает существенные биологические особенности изучаемых процессов и явлений и потому заслуживает специальной оценки и анализа. Как показали исследования академика С. С. Шварца и его учеников, характер кривой распределения одного из жизненно важных признаков изучаемой популяции животных может указать на определенную тенденцию в действии и направлении естественного отбора.

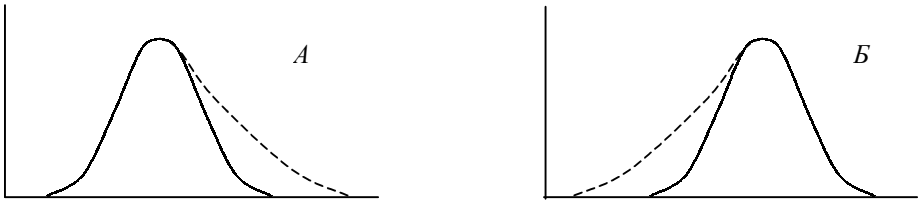


Рис. 3.1. *Асимметрия распределения (обозначена пунктиром относительно нормальной кривой): А – положительная (правосторонняя), Б – отрицательная (левосторонняя)*

Если вариационные кривые, характеризующие изменчивость отдельных признаков, асимметричны, это может означать стремление отбора изменить среднюю норму изменчивости популяции путем

преимущественной элиминации худших (в данных условиях) особей (ведущий отбор). Если же фиксируемое состояние популяции на данном этапе ее развития стабильно (поддерживается стабилизирующим отбором), то изменчивость отдельных признаков у особей данной популяции должна подчиняться закону нормального распределения: отклонения от средней в сторону плюс- и минус-вариант должны встречаться одинаково часто. Асимметричности в этом случае, естественно, не наблюдается.

Рис. 3.2. *Эксцесс распределения (обозначен пунктиром относительно нормальной кривой): А – положительный (островершинность), Б – отрицательный (туповершинность)*

Не меньшую информативную нагрузку несет и характер эксцесса вариационной кривой. Ярко выраженный отрицательный эксцесс распределения однородного по возрастному и половому составу материала может свидетельствовать о действии на популяцию дизруптивного отбора и о тенденции изучаемого вида образовывать не только обычные, типичные формы, но и давать в повышенном количестве новые для него вариации, сильно уклоняющиеся от нормы. При сильном положительном эксцессе не исключено ужесточение стабилизирующего отбора.

Показатели асимметрии и эксцесса используются как тесты для проверки соответствия эмпирического распределения нормальному и биномиальному законам. Значимость этих показателей говорит о нарушении нормальной формы кривой распределения. Критерии Стьюдента для $df = \infty$ проверяют нулевую гипотезу Но: «коэффициент асимметрии (эксцесса) существенно от нуля не отличается, следовательно, асимметрия (эксцесс) достоверно не выражена»:

$$T_A = \frac{A-0}{m_A},$$

$$T_E = \frac{E-0}{m_E},$$

где m_E – статистическая ошибка соответствующего коэффициента.

Точная и приближенная формулы для расчета статистической ошибки показателя асимметрии и эксцесса составляют:

$$m_A = \sqrt{\frac{6 \cdot n \cdot (n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}} \approx \sqrt{\frac{6}{n}},$$

$$m_E = \sqrt{\frac{24 \cdot n \cdot (n-1)^2}{(n-3) \cdot (n-2) \cdot (n+3) \cdot (n+5)}} \approx \sqrt{\frac{24}{n+5}} \approx 2 \cdot \sqrt{\frac{6}{n}}.$$

Проведем вычисление коэффициентов асимметрии и эксцесса для данных по массе бурозубок:

A =СКОС(A2:A64) = 0.608664019,

m_A = КОРЕНЬ(6* 63*(63-1)/((63-2)*(63+1)*(63+3))) = 0.301588566,

T_A = 2.018193285,

E =ЭКСЦЕСС(A2:A64) = 1.130099794,

m_E = КОРЕНЬ(24*C8*(C8-1)^2/((C8-3)*(C8-2)*(C8+3)*(C8+5))) =

= 0.594840621,

T_E = 1.899836283.

Табличное значение критерия Стьюдента составляет $T_{(0.05,\infty)}$ = 1.96. Поскольку полученное значение T_A = 2.02 больше табличного (1.96), коэффициент асимметрии значимо отличается от нуля. По массе тела бурозубки распределены с правосторонней асимметрией, что понятно, поскольку в летних отловах большинство животных – разновозрастные молодые (с нормальным распределением), а меньшинство – зимовавшие (более тяжелые и скопившиеся справа). Поскольку полученное значение T_E = 1.90 меньше табличного (1.96), коэффициент эксцесса значимо от нуля не отличается. Распределение в целом близко к нормальному, что подтверждает предыдущее объяснение асимметрии.

Вычислить все рассмотренные параметры вариационного ряда можно в среде Excel с помощью макроса, который вызывается командой меню Сервис\ Анализ данных\ Описательная статистика. Например, обработка данных по массе бурозубок дает следующие результаты:

Столбец 1	
Среднее	9.298413
Стандартная ошибка	0.113039
Медиана	9.2
Мода	9.2
Стандартное отклонение	0.897216
Дисперсия выборки	0.804997
Эксцесс	1.1301
Асимметричность	0.608664
Интервал	4.6
Минимум	7.3
Максимум	11.9
Сумма	585.8
Счет	63
Уровень надежности (95.0%)	0.225961

Основные типы распределения биологических признаков

Известно несколько типов таких распределений – нормальное, биномиальное, Пуассона и некоторые другие. Зная тип распределения, можно воспользоваться разработанными специально для него приемами математической обработки и получить максимальную, а главное, достоверную информацию о явлении, сделать более точный прогноз, правильнее оценить различия между параметрами разных выборок.

Нормальное распределение

Наиболее характерный тип распределения непрерывных случайных величин, из него можно вывести (к нему сводятся) все остальные. Распределение симметрично, причем крайние значения (наибольшие и наименьшие) появляются редко, но чем ближе значения признака к центру (к средней арифметической), тем оно чаще встречается.

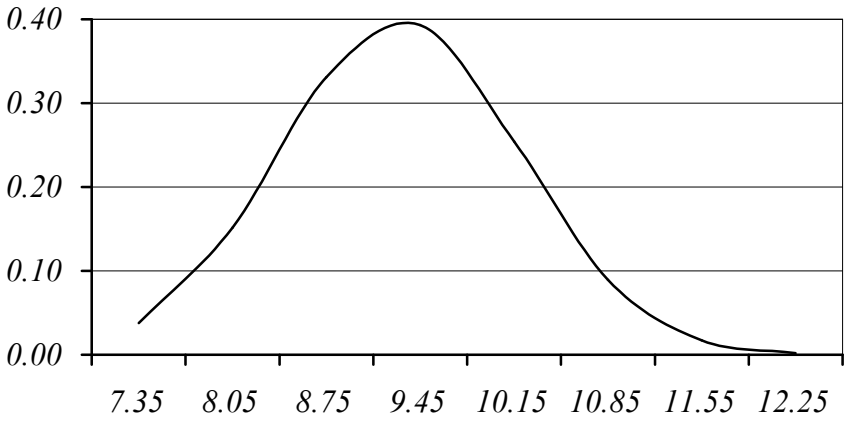


Рис. 3.3. *Нормальное распределение с параметрами n = 63, M = 9.3, S = 0.79. По оси абсцисс – вес тела землероек-бурозубок, по оси ординат – табличные значения для нормального распределения. Рассчитать ординаты нормальной кривой для конкретного значения* x_i

можно по формуле: $p_i = (1/\sqrt{2\pi}) \cdot e^{-(x_i - M)^2 / 2 \cdot S^2}$

Если откладывать на оси абсцисс результаты измерений, а на оси ординат число случаев (частоту получения данного результата измерений), то образуется кривая нормального распределения (кривая Гаусса), характеризующаяся симметричной колоколообразной формой (рис. 3.3). Точная формула кривой плотности вероятности нормального распределения приведена выше. Одно из важных его свойств состоит в том, что среднее квадратичное отклонение примерно 4 раза укладывается в размахе изменчивости признака и по величине значительно уступает средней. Геометрически стандартное отклонение равно расстоянию от центра кривой распределения до точки перегиба кривой. Примеры расчета параметров нормального распределения (средней *M*, дисперсии S^2, асимметрии *A* и эксцесса *E*) приведены выше.

Биномиальное распределение

Во многом близко к нормальному. Отличие состоит лишь в том, что оно характеризует поведение дискретных признаков (выра-

женных целыми числами). Как правило, для описания биологических признаков подходит симметричное биномиальное распределение, у которого дисперсия много меньше средней.

Примерами описания признаков с помощью биномиального распределения могут служить 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 и т. п. В основе биномиального распределения лежит альтернативное проявление изучаемого признака: он может присутствовать у единичного объекта или отсутствовать, проявиться или нет. Отдельный корнеплод может быть больным или здоровым (признак качественный), тогда проба из нескольких корнеплодов будет содержать некоторое *число* здоровых корнеплодов (признак количественный), а множество равнообъемных проб образует уже выборку чисел. (Кстати отметим, что подсчет *числа* однотипных объектов в *пробе* есть эффективный способ перевода качественных признаков в количественные.)

Главной «организующей силой» такого распределения является способ, с помощью которого получают значения случайной величины – это отбор проб. Пробой называют фиксированное (объемом m) множество объектов, которые могут быть только двух типов (например, белое или черное, ● и ■, ♀ и ♂, 🚹 и 🚺 , 0 и 1). Получается, что каждая проба объединяет несколько (m) простых случайных величин.

Для формирования дискретных распределений используются большие группы объектов, либо территория (маршрут), либо процесс, которые разбиваются на пробы – соответственно на порции (группы), участки (площадки), отрезки (этапы), затем идет подсчет числа проб, содержащих то или иное число известных объектов. Значение отдельной варианты представляет собой число объектов определенного качества в отдельной пробе.

Если при этом вероятности появления объектов разного качества приблизительно равны (например, когда общее число поврежденных примерно равно числу здоровых корнеплодов), то биномиальное распределение имеет симметричную, колоколообразную (но ступенчатую) форму, подобную нормальному распределению. Большие отклонения от условия равенства вероятностей элементарных событий порождают асимметричное распределение, и такое поведе-

ние случайной величины лучше описывать с помощью закона распределения Пуассона (см. ниже).

Одна из обычных и важных характеристик популяций животных, плодовитость самок, подчиняется биномиальному закону. Единичное событие – это появление или «непоявление» детеныша. Тогда пробой будет «число потенциальных детенышей у одной самки», объем которой равен значению максимальной плодовитости. Самка с наибольшей плодовитостью реализовала все возможные потенции. Самка с меньшим числом детенышей как бы не смогла реализовать потенциал полностью. При этом вероятность реализации отдельного события «детеныш появится» составляет p, а вероятность события «детеныш не появится» равна $q = 1–p$. В том случае, когда для отдельного детеныша вероятность появиться равна вероятности не появиться, $p = q$, самок с бóльшим числом детенышей и вовсе почти без детенышей будет мало, в большинстве своем самки будут приносить около половины потенциально возможных детенышей; распределение будет строго симметричным. В случае неравенства вероятностей будет наблюдаться та или иная степень асимметрии.

Рассмотрим результаты изучения плодовитости серебристо-черных лисиц (число щенков на самку) (см. данные на стр. 33). Для построения вариационного ряда берем 8 классов, классовый интервал для этого дискретного признака составит $dx = 1$.

Плодовитость, *x*	**Частота, число самок, *a***	**Число эмбрионов, *x·a***
1	1	1
2	1	2
3	8	24
4	16	64
5	23	115
6	21	126
7	3	21
8	3	24
n	76	377

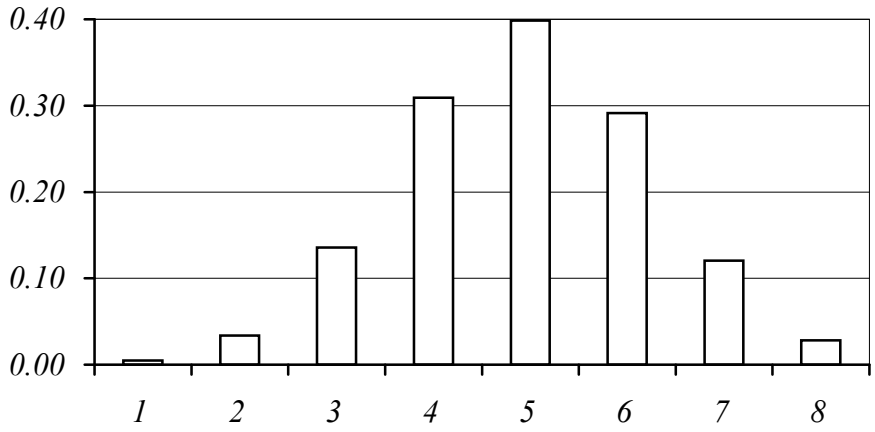


Рис. 3.4. *Биномиальное распределение с параметрами* n = 76, *M = 4.95, S = 1.33. По оси абсцисс – число щенков лисицы на одну самку, по оси ординат – частости (относительные частоты)*

Поместив значения на лист Excel, нетрудно рассчитать (предлагаем читателю это сделать самостоятельно) все основные параметры распределения по рассмотренным выше формулам с использованием функций листа Excel =СРЗНАЧ и =СТАНДОТКЛОН:

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

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

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

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

Рассчитаем вероятности элементарных событий (появления-непоявления детенышей). Для этого сначала рассчитаем общее число появившихся детенышей. Оно равно сумме всех произведений числа детенышей в отдельном выводке (x) на частоту их встречаемости (a):

$N_Д = \Sigma(x \cdot a) = 377$. Затем рассчитаем общее число потенциальных зародышей (суммарный объем всех проб с одинаковым объемом $m = 8$): $N_П = n \cdot 8 = 76 \cdot 8 = 608$ экз. Наконец, определим долю «реализовавшихся» детенышей среди потенциальных:

$p = N_Д / N_П = 377/ 608 = 0.62$

и долю «не реализовавшихся»:

$q = 1 - p = 1 - 0.62 = 0.38$.

Эти ключевые характеристики сущности наблюдаемого процесса размножения могут говорить о том, что вероятность рождения отдельного детеныша ($p = 0.62$) превышает вероятность его нерождения ($q = 0.38$), что в выборке присутствуют животные с ненормально высокой плодовитостью, т. е. об асимметричности распределения. Проверить это предположение можно с помощью соответствующих показателей асимметрии и эксцесса, которые рассчитываются также проще, чем для непрерывного признака:

$$A = \frac{q - p}{\sqrt{n \cdot p \cdot q}} = \frac{0.38 - 0.62}{\sqrt{76 \cdot 0.62 \cdot 0.38}} = -0.05675,\ m_A = 0.28,\ T_A = 0.2 < T_{(0.05,\infty)}.$$

$$E = \frac{\frac{1}{p \cdot q} - 6}{n} = \frac{\frac{1}{0.62 \cdot 0.38} - 6}{76} = -0.07088,\ m_E = 0.56,\ T_E = 0.1 < T_{(0.05,\infty)}.$$

В данном случае, несмотря на некоторое отличие базовых вероятностей, асимметрию нельзя считать существенной.

Найденные выше вероятности p и q позволяют рассчитать параметры биномиального распределения по другим, более простым формулам:

$M = m \cdot p = 8 \cdot 0.62 = 4.96$ экз./самку,

$S = \sqrt{m \cdot p \cdot q} = \sqrt{8 \cdot 0.62 \cdot 0.38} = 1.37$ экз./самку.

Результаты оказываются идентичными с точностью до величины ошибки округления.

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

Это вариант описания стохастического поведения дискретных признаков для случаев, когда базовая вероятность элементарных альтернативных событий неодинакова, одно из них наблюдается заметно чаще другого ($p<<q$) (классический пример – попадание гитлеровских авиационных бомб в разные кварталы Лондона). Закон Пуассона описывает редкие события, происходящие 1, 2, 3 и т. д. раз на сотни и тысячи обычных событий. Поведение биологических объектов, соответствующее закону Пуассона, наблюдается в том случае, когда по пробам случайно распределены редкие объекты. Примеры таких явлений – 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.

Как и в случае с биномиальным распределением, случайная величина, распределенная по закону Пуассона, определяется подсчетом числа элементарных событий *в пробе* (в группе, в навеске, на участке, на этапе). Распределение Пуассона резко асимметрично, причем дисперсия равна средней арифметической, что может служить критерием для оценки характера распределения изучаемого признака (рис. 3.5).

В течение одного года (1946) пометили кольцами и выпустили на волю 32 буревестника.

Число повторных отловов, *x*	**Число отловленных животных, *a***	**Число случаев повторного отлова, *x·a***
0	15	0
1	7	7
2	7	14
3	2	6
4	1	4
n	32	31

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):

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

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

$S^2 \approx M$.

По крайней мере, проверка по критерию Фишера не выявила достоверных отличий между ними (подробнее см. с. 112):

$F = 1.257/0.968 = 1.157 < F_{(0.05,31,31)} = 1.8$.

Эта близость свидетельствует о соответствии наблюдаемого распределения закону Пуассона.

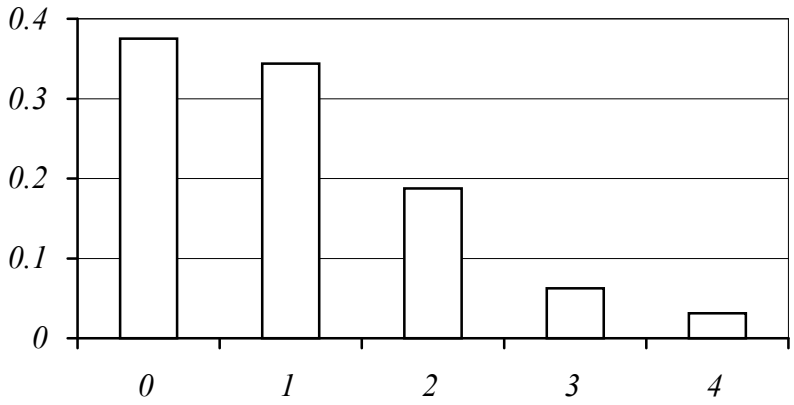


Рис. 3.5. *Распределение Пуассона с параметрами* n = 32, $M \approx S^2 = 0.968$. *По оси абсцисс – число повторных отловов, по оси ординат – частости (относительные частоты)*

Доверительный интервал для параметров распределения Пуассона определить несколько сложнее, чем для других типов. В связи с асимметричностью распределения Пуассона для расчета левой и правой доверительных границ средней арифметической и дисперсии (как известно, эти значения равны) используют специальные формулы с участием статистики хи-квадрат Пирсона (Браунли, 1977):

левая граница $x_{лев.} = 0.5 \cdot \chi^2_{(P,\ df_1)}$,

правая граница $x_{прав.} = 0.5 \cdot \chi^2_{(\alpha,\ df_2)}$,

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 хи-квадрат по таблице 9*П*.

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.

Альтернативное распределение

Распределение дискретной случайной величины, имеющей лишь два противоположных (разнокачественных) значения ($k = 2$). В одной пробе (в одном наблюдении) содержится одна варианта, одно из двух возможных значений. Вероятности каждого из них могут быть равны ($p = q$) либо не равны ($p < q$; $p > q$). В выборке варианты разделены по некоторому признаку на два класса в соответствии с наличием у них одного из двух значений. Примеры: самцы и самки в одной выборке, больные и здоровые организмы, сработавшие и пустые ловушки на одной учетной линии, два варианта аллельных признаков, вакцинированные и невакцинированные пациенты среди заболевших и др. (рис. 3.6). Вычисления констант достаточно просты и не требуют построения вариационного ряда.

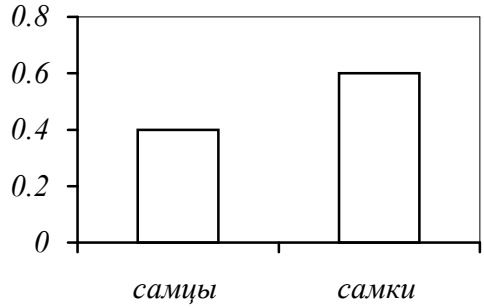


Рис. 3.6. *Альтернативное распределение, представленное двумя классами вариант. По оси ординат – частости (доли) этих групп*

Важнейшей характеристикой является доля (p) вариант определенного вида (А), представленных общим числом n_A в пределах выборки объемом n:

$$p = \frac{n_A}{n},$$

В том случае, если исходы отдельных испытаний характеризуются числами 0 или 1, доля вариант со значением 1 численно совпадает со средней арифметической, рассчитанной для всех значений:

$$M = \frac{\sum x}{n}.$$

Ключевым моментом здесь является соображение, имеем ли мы право придавать испытаниям вид чисел 0 и 1 или это просто удобная форма отображения качественных признаков? Обращаясь к теории формирования сложных распределений дискретной величины, мы можем и для альтернативного распределения воспользоваться идеей отбора вариант группами, принципом отбора *проб*. Только для альтернативного распределения объем одной пробы равен единице (в пробу входит одна варианта), $m = 1$. 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 таком 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 $n = 200$:
00000000000000000000000000000000000000000000000000000000000000000000
11111111111111111111111111111111111111111111111111111111111111111111
11111111111111111111111111111111111111111111111111111111.

Доля вариант со значением 1 составляет:

$$p = \frac{n_A}{n} = \frac{120}{200} = 0.6,$$

что совпадает со средней арифметической для всего ряда:

$$M = \frac{\sum x}{n} = \frac{120}{200} = 0.6.$$

Для альтернативного распределения могут применяться те же формулы расчета выборочных параметров, что и для биномиального распределения.

Средняя (доля самок):

$M = m \cdot p = 1 \cdot 0.6 = p = 0.6$.

Стандартное отклонение (при $m = 1$):

$$S = \sqrt{m \cdot p \cdot q} = \sqrt{p \cdot q} = \sqrt{0.6 \cdot 0.4} = 0.24.$$

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 таблице 10*П*. 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 таблицы 10*П* переводят найденные значения обратно в проценты.

Найдем доверительные границы для доли самок полевок $p = 0.6$ при уровне значимости $\alpha = 0.05$. Используя таблицу 10*П* и проводя 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%.

Полиномиальное распределение

Наблюдается для качественных признаков, имеющих не два альтернативных свойства, но несколько возможных проявлений качества ($k>2$). Примеры полиморфизма популяций – как раз 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 (иначе – мультиномиального) распределения может служить встречаемость 4 фенов головы живородящей ящерицы – 4 вариантов контакта лобно-носового, предлобных и лобного щитков (рис. 3.7).

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: $m_\mu = \sqrt{\dfrac{\mu \cdot (k - \mu)}{n}}$.

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.

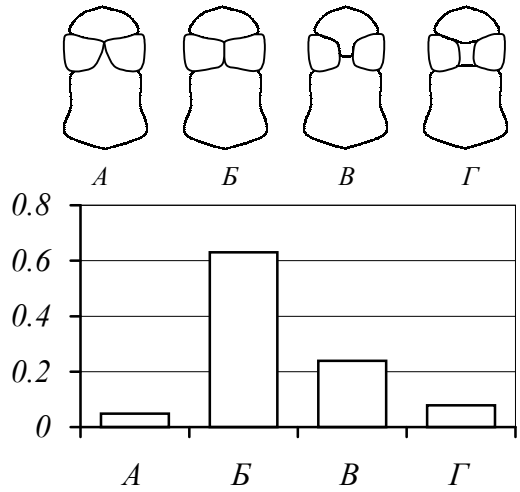


Рис. 3.7. *Полиномиальное распределение (4 фена головы живородящей ящерицы). По оси ординат – частости фенов среди 64 сеголетков, отловленных под Петрозаводском*

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 признака ($p = q$ 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 генов и 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. 3.8).

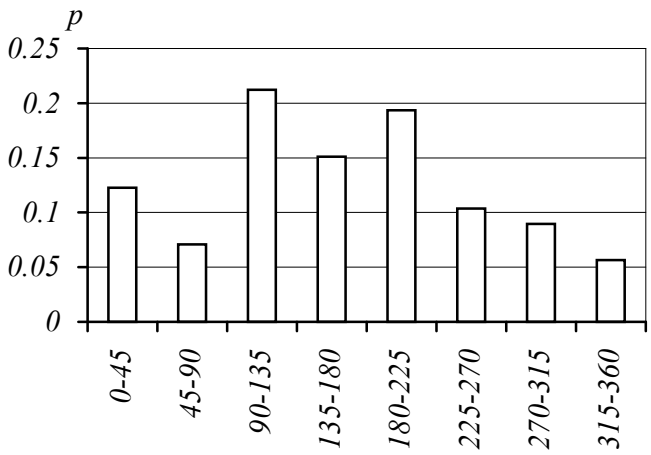


Рис. 3.8. *Предположительно равномерное распределение числа ветвей ели по секторам азимута (°)*

Помимо рассмотренных четырех типов распределения для описания эмпирических совокупностей предложено множество других моделей, основанных на других принципах и дающих нередко более точные оценки параметров.

Для описания природных явлений более реалистичные основания, чем биномиальное, имеет распределение *гипергеометрическое*, оно не предполагает возврата объектов каждой пробы обратно в изучаемую совокупность. Распределение *негативное биномиальное* подходит для случая, когда вероятности элементарных событий (p и q) не постоянны, в отличие от биномиального распределения. Распределения *Максвелла* и *Рэлея* имеют умеренную правостороннюю асимметрию и описывают поведение непрерывных положительных случайных величин. Распределения *Парето* и *показательное* пригодны для описания резко правосторонне асимметричных вариационных рядов с перепадом частот. Распределение *логнормальное*, или логарифмически нормальное, характеризуется тем, что логарифмы исходных значений выборки образуют нормальное распределение; эта модель подходит для описания признаков, имеющих распределения с умеренной правосторонней асимметрией, это, в первую очередь, концентрации веществ в различных средах, т. е. гидрохимические, физиологические и биохимические показатели.

4
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Проверка статистических гипотез – это путь установления биологических закономерностей. Закономерное означает не только повторяемое, но – в связи с известной причиной. «Установить закономерность» значит «установить причинную связь явлений», когда степень выраженности изучаемого свойства объекта определяется внешними или внутренними факторами. Однородность действия факторов определяет и сходство отклика, изменение факторов влечет за собой изменение характеристик наблюдаемого объекта. Параллельное изменение свойств объекта и его среды (признака и фактора) эмпирической наукой принимается как свидетельство зависимости признака от фактора, как закономерная связь явлений.

Для регистрации такого рода зависимости необходимы наблюдения хотя бы двух состояний объекта – при разных уровнях действия фактора. Со статистической точки зрения это должны быть две группы значений (две выборки), характеризующие устойчивость реакции признака на разные уровни действия фактора (разные дозы). Так формируется первая задача поиска закономерностей – сравнение между собой групп вариант, полученных при разной силе воздействия внешних (или внутренних) условий. Если наличие фактора существенно для проявления признака, параметры выборок будут различаться, если фактор безразличен (не действует на объект), отличий между группами не будет.

В терминах математической статистики эта задача формулируется как вопрос о принадлежности выборок к общей генеральной совокупности. Если выборки взяты из одной генеральной совокупности, то между ними не должно быть существенных отличий – только небольшие и случайные (ошибки репрезентативности). Если же выборки взяты из разных генеральных совокупностей, то отличия между выборками должны быть закономерными – достоверными, значимыми. Решая эту задачу, изначально предполагают, что «выборки взяты из одной и той же генеральной совокупности, отличия между средними незначимы». Отличия между выборочными параметрами рассматриваются как отличия по случайным

причинам, как ошибки репрезентативности. Чтобы в этом убедиться, по всему объему данных вычисляются ошибки репрезентативности и затем различия между выборочными параметрами сравниваются с ошибками репрезентативности этих параметров; обычно это частное от деления «отличия»/«ошибка». Такое математическое выражение носит название *статистического критерия*. Если «отличия» немногим больше «ошибки» (небольшая величина критерия), то считается, что параметры действительно не отличаются друг от друга. Если же разность между параметрами много больше величины ошибки репрезентативности (высокое значение критерия), то признают, что это не случайность, но результат действия фактора.

Для разных статистических параметров разработаны соответствующие методы их сравнения с ошибками репрезентативности (критерии). Общим остается принцип формулирования *статистического вывода*: если величина критерия превышает некое «критическое» значение, то нулевая гипотеза отвергается, и тем самым признается – выборки взяты из разных генеральных совокупностей. Это значит, что некий фактор влияет на изменение признака, что удалось установить реальное (закономерное) биологическое явление. Если же величина критерия ниже критической, отличие между выборками признается несущественным, недоказанным.

При всем кажущемся многообразии вариантов проявления различного рода закономерностей, можно выделить всего 4 класса статистических задач вида «доказать отличия»:

1. *Доказать чужеродность варианты в выборке*
   (или «классифицировать объекты»).
2. *Доказать отличие двух выборок*.
3. *Доказать отличие нескольких выборок*
   (или «доказать влияние фактора на признак»).
4. *Найти зависимость между признаками*
   (или «доказать сопряженность варьирования признаков»).

По своей статистической сути все многообразные методы количественной биологии не выйдут за рамки представленного списка задач, хотя в зависимости от конкретной постановки биологического вопроса, типа данных, метода их сбора, способа представления и пр. конкретные алгоритмы могут существенно отличаться. Приемы решения частных задач рассмотрены далее.

5
ЗАДАЧА «ДОКАЗАТЬ ЧУЖЕРОДНОСТЬ ВАРИАНТЫ»

В биологии часто встречается ситуация, когда одна из полученных вариант сильно отличается от остальных. Эти отклонения могли возникнуть в результате неточности измерений, ошибок внимания, методических погрешностей и т. д. Можно ли такие резко выделяющиеся значения использовать при дальнейших расчетах?

С помощью этой редко возникающей задачи о принадлежности данной варианты к данной выборке мы сделаем необходимый переход от практики статистического оценивания к практике проверки статистических гипотез.

Любая статистическая задача – суть вопрос о принадлежности разных вариант к единой генеральной совокупности, о том, что сравниваемые выборочные варианты испытывают на себе действие одних и тех же доминирующих и случайных факторов. В терминах математической статистики поставленный вопрос звучит так: *относится ли данная варианта вместе с другими вариантами изучаемой выборки к одной и той же генеральной совокупности или – к разным?* Его можно сформулировать и по-другому: сформировано ли данное значение варианты под действием тех же доминирующих и случайных факторов, что и все остальные варианты данной выборки, или это были иные факторы? Здесь возможны два ответа:

1. Факторы те же, т. е. все варианты взяты из одной и той же генеральной совокупности.

2. Факторы иные, т. е. особенная варианта и выборка порознь взяты из разных генеральных совокупностей.

Ответ на этот вопрос можно получить с использованием рассмотренных выше свойств нормального распределения. Так, если все варианты были взяты из одной генеральной совокупности, значит, поведение их должно быть однородным, они должны отличаться только в силу случайных причин и (с вероятностью $P = 0.95$) находиться в диапазоне $M \pm 2 \cdot S$ (см. с. 50). Иными словами, по случайным причинам варианты достаточно большой выборки отклоняются влево или вправо от средней арифметической не более чем на $2 \cdot S$:

$x - M < 2 \cdot S$ или $(x - M)/S < 2$.

Общепринятой безразмерной характеристикой отклонения отдельной варианты от средней арифметической служит *нормированное отклонение*, оно показывает, на сколько стандартных отклонений отклоняется та или иная варианта от среднего уровня варьирующего признака, и выражается формулой:

$$t = \frac{x - M}{S} \sim t_{табл.},$$

где t – критерий выпада (исключения);

x – выделяющееся значение признака;

M – средняя величина для группы вариант;

$t_{табл.}$ – стандартные значения критерия выпадов, определяемые свойствами нормального распределения, их можно найти по табл. 5*П* для трех уровней вероятности (для больших выборок обычно пользуются значением $t_{табл.} = 2$ при $P = 0.95$, или $\alpha = 0.05$)

~ – значок можно прочитать как «не больше».

Используя этот показатель, можно утверждать, что для вариант, принадлежащих к данной достаточно большой выборке, нормированное отклонение меньше двух (с вероятностью $P = 0.95$):

$t < 2$.

Если же на отдельную варианту действовал какой-либо новый фактор, который вызвал дополнительное, т. е. не случайное, отклонение от средней, то такая варианта окажется за пределами указанного диапазона $M \pm 2S$, а ее нормированное отклонение будет равно или больше двух: $t \geq 2$.

Нормированное отклонение есть простейший статистический критерий, который помогает определять так называемые «выскакивающие» варианты и решать вопрос о возможности их отбрасывания как артефактов (исключать из дальнейшей обработки). Смысл критерия «исключения» состоит в том, чтобы определить, находится ли данная варианта в интервале, характерном для большинства членов выборки, или же вне его. Если значение критерия больше табличного, то это означает, что данное значение не относится к анализируемой совокупности, а есть проявление каких-то особых закономерностей, ошибок и пр. и должно быть поэтому исключено из рассмотрения (отброшено). При этом иногда рекомендуют значения параметров (M, S) рассчитывать без учета «подозрительной» варианты. После такой «чистки» параметры выборки должны быть рас-

считаны заново. К оценке чужеродности вариант, как и к другим методам статистики, нельзя подходить формально; цель биометрического исследования всегда состоит в том, чтобы понять специфику явления. В частности, «отскакивающая» варианта может быть следствием того, что признак имеет иное, *не*-нормальное распределение.

Рассмотрим работу критерия на примере. При измерении длины черепа взрослых самцов обыкновенной землеройки-бурозубки получены выборки с такими параметрами: $M = 18.8$, $S = 0.3$ мм. Общее число животных $n = 85$. Вызывают сомнения два слишком больших значения 19.2 и 21.0. Определим для них критерии выпада:

$$t_1 = \frac{19.2 - 18.8}{0.3} = 1.3 < 2, \qquad t_2 = \frac{21.0 - 18.8}{0.3} = 7.3 > 2.$$

Согласно таблице 5*П*, критическое значение нормированного отклонения для уровня значимости $\alpha = 0.05$ и $n = 85$ равно $t = 2.0$. Поскольку первое полученное значение (1.3) меньше табличного (2), первый из сомнительных результатов исключать не следует, а второй должен быть отброшен – критерий выпада (7.3) превышает табличное значение (2).

Понятие «нормированное отклонение» позволяет точнее определить важнейшее понятие статистического критерия. *Статистический критерий* – безразмерная случайная величина, которая имеет известный закон распределения и используется для проверки статистических гипотез.

Нормированное отклонение есть статистический критерий. Во-первых, это безразмерная величина, поскольку единицы измерения числителя ($x_i–M$) и знаменателя (S) взаимно уничтожаются. Во-вторых, оно имеет вполне определенное распределение (в случае непрерывных признаков – нормальное) со своими параметрами: средняя равна нулю $M_t = t_M = (M–M)/S = 0$, а стандартное отклонение равно единице $S_t = t_S = (S–M)/S = (S–0)/S = S/S = 1$. Последний тезис стоит рассмотреть более предметно, поскольку он имеет большое практическое значение.

Рассмотрим на примере конкретных данных, почему нормированное отклонение имеет такие параметры. Значения длины хвоста (Lc, мм) для выборки из $n = 9$ гадюк дают среднюю $M = 73.1$, стандартное отклонение $S = 11.7$ мм.

										M	*S*
x_{Lc}	58	59	75	93	65	85	79	68	76	73.1	11.7
t_{Lc}	–1.29	–1.2	0.16	1.69	–0.69	1.01	0.50	–0.44	0.25	–0	1

Рассчитаем для каждого значения нормированное отклонение, например, для $x = 59$ $t = (x–M)/S = (59–73.1)/11.7 = –1.20$, а для $x = 93$ $t = (93–73.1)/11.7 = 1.69$. Нетрудно подсчитать, что для полученного ряда нового расчетного признака t средняя по всему ряду составит $M_t = –2 \cdot 10^{-16} \approx 0$, стандартное отклонение $S_t = 1$.

Здесь важно подчеркнуть, что нормированное отклонение – универсальная величина. Какой бы признак (имеющий нормальное распределение) мы ни брали, его значения можно выразить в виде расстояния от центра в единицах стандартного отклонения, т. е. на сколько S данное значение x отклонилось от M. При этом, как следует из свойств нормального распределения, крайние значения в 95% случаев не будут принимать значения меньше –2 и больше 2 (рис. 5.1).

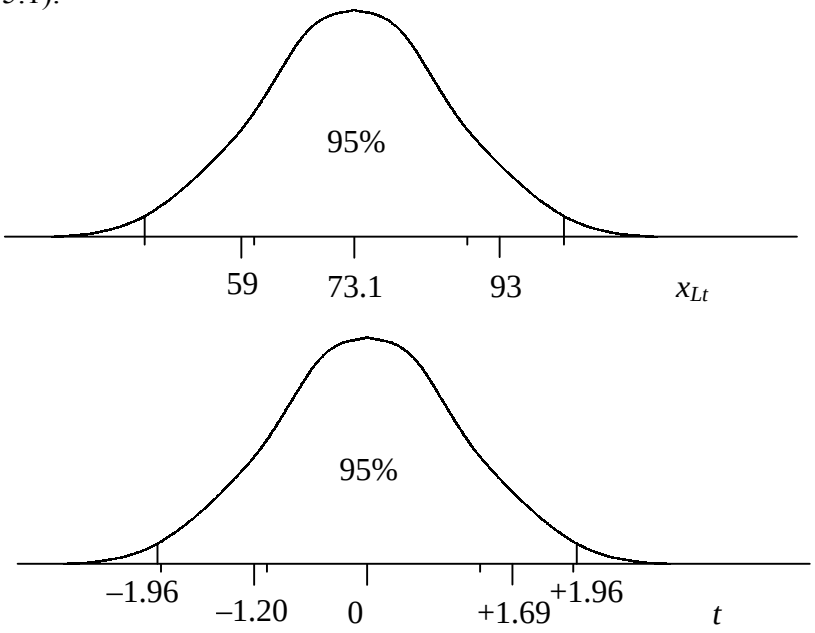


Рис. 5.1. *Переход от реального признака x к нормированному отклонению t*

С помощью нормированного отклонения можно, например, сравнивать объекты разного качества (организмы разных видов, разных пород и сортов, разных возрастов) – по разным свойствам (признакам).

Так, промеры длины хвоста (*Lc*, мм) и длины тела (*Lt*, см) у выборки гадюк разного пола позволяют увидеть, что самец № 5 при средних размерах тела (x_{Lt} = 0.03) обладает относительно небольшим хвостом (t_{Lc} = –0.69), а самец № 6 при такой же длине имеет существенно более длинный хвост (t_{Lc} = 1.01).

№	1	2	3	4	5	6	7	8	9		
Пол	*f*	*f*	*m*	*m*	*m*	*m*	*f*	*m*	*f*	*M*	*S*
x_{Lc}	58	59	75	93	65	85	79	68	76	73.1	11.7
t_{Lc}	–1.29	–1.20	0.16	1.69	–0.69	1.01	0.50	–0.44	0.25	0.00	1.00
x_{Lt}	45	46	48	49	50	50	53	53	55	49.9	3.3
t_{Lt}	–1.47	–1.17	–0.57	–0.27	0.03	0.03	0.93	0.93	1.53	0.00	1.00

Нормированное отклонение можно использовать и для сравнительной оценки разных индивидов по одному и тому же признаку. Например, если сопоставляемые по относительному весу сердца молодая и взрослая землеройки-бурозубки демонстрируют одинаковые показатели (10.5 мг%), то это, тем не менее, не означает их сходства по изучаемому признаку. Используя известную информацию (у молодых средний индекс сердца равен *M* = 10.0 при стандартном отклонении *S* = 1.3, у взрослых – *M* = 11.8, *S* = 1.1), рассчитаем нормированное отклонение для молодого зверька

$$t_1 = \frac{10.5 - 10}{1.3} = 0.3$$ и для взрослого $$t_2 = \frac{10.5 - 11.8}{1.1} = -1.2.$$

Налицо существенное различие: взрослый зверек имеет относительно низкий показатель сердечного индекса, а молодой близок по этому признаку к видовой норме.

Наибольшее развитие такой подход получает в процедурах обработки многомерных данных, при исследовании объектов, охарактеризованных по многим признакам, методом корреляций, главных компонент, при их кластеризации и т. п. Во многих случаях обработка многомерного массива начинается с *нормирования* данных, с помощью формулы нормированного отклонения.

6
ЗАДАЧА «ДОКАЗАТЬ ОТЛИЧИЕ ДВУХ ВЫБОРОК»

Сравнение двух выборок не может быть самоцелью биологического исследования, поскольку современную биологию интересуют не просто факты, но их подоплека, не столько конкретное биологическое явление, сколько причина его возникновения. В этом ключе сравнение двух выборок выступает в роли метода поиска отличий в причинах, обеспечивших существование двух групп объектов (вариант) разного качества; в конце концов, это поиск влияния фактора, поиск закономерности. В свете рассмотренного ранее фрейма формирования выборок источник отличий между выборками следует усматривать в различии методик сбора данных, различиях объектов исследования по статусу или состоянию или в различиях условий существования объектов. Переводя эти случаи в форму статистического вопроса, можно спросить, сравниваемые выборки взяты из одной или разных генеральных совокупностей? Поскольку выборки могут быть охарактеризованы несколькими обобщающими параметрами, то и сравниваться они будут с помощью разных статистических методов. Сравнение двух выборок есть развитие задачи сравнения варианты с выборкой, это своеобразный поиск «чужеродности» всех вариант одной выборки по отношению к другой.

Ранее было показано, что специфику выборки можно охарактеризовать с разных сторон, используя разные способы описания выборок; это порождает целое подсемейство методов оценки различия выборок (табл. 6.1).

Сравнение двух выборок по величине признака

Биологический смысл процедуры сравнения двух выборок по уровню развития признака (по средним арифметическим) состоит в том, чтобы определить, действовал ли в одной из выборок новый систематический фактор по сравнению с другой выборкой, поскольку, как было показано выше, средняя арифметическая характеризует действие систематических факторов, дающих равный вклад в каждую варианту выборки.

Таблица 6.1

Задача	**Содержание задачи**	**Методы**
Доказать различие двух средних арифметических для одного признака	Отличаются доминирующие факторы, формирующие выборки	Критерий Стьюдента
Доказать различие нескольких пар средних арифметических	Отличаются доминирующие факторы	Метод попарных сравнений Шеффе (см. однофакторный дисперсионный анализ)
Доказать различие двух стандартных отклонений	Отличаются случайные факторы, участвующие в формировании выборки	Критерий Стьюдента
Доказать различие двух дисперсий	Отличаются случайные факторы	Критерий Фишера
Доказать различие двух коэффициентов вариации	Отличаются случайные факторы	Критерий Стьюдента
Доказать различие между эмпирическим и теоретическим частотными распределениями	Поведение случайной величины соответствует какому-либо закону распределения	Критерий Пирсона хи-кадрат
Доказать различие двух эмпирических частотных распределений	Отличаются в целом любые факторы	Критерий Пирсона хи-кадрат
Доказать различие двух выборок в целом	Отличаются в целом любые факторы	Непараметрические критерии Уилкоксона, Уайта, Розенбаума
Доказать различие двух выборок по силе корреляции по двум признакам	Отличается сила сопряжения двух признаков в разных выборках	Метод z-преобразования Фишера и критерий Стьюдента
Доказать различие двух выборок по характеру регрессионной зависимости между двумя признаками	Отличается характер сопряжения двух признаков в разных выборках	Сравнение линий регрессии по критериям Фишера и Стьюдента

Если условия формирования выборок были одинаковы, варианты в обеих выборках будут отличаться друг от друга только по случайным причинам, и средние для этих выборок будут характеризовать одну и ту же главную причину. Если же какой-либо фактор действовал на разные выборки по-разному, статистический критерий покажет достоверное отличие средних арифметических. Дело в том, что дополнительный систематический фактор во время набора выборки сообщает каждой варианте некую прибавку ($_{\Delta}x$) к обычному значению. Понятно, что для выборки, испытавшей такое действие, мы получим среднюю арифметическую, смещенную именно на эту дополнительную величину ($M_1 = M_2 + {}_{\Delta}x$).

Например, если измерять размеры тела двух групп разновозрастных животных, все особи старшей группы (росшие дольше молодых), будут немного крупнее последних и индивидуально, и в среднем. Пример из области токсикологии: добавление токсиканта вредно влияет на всех подопытных животных и может снизить, например, плодовитость – как индивидуально, так и в среднем. В общем дополнительными существенными причинами могут выступить разный статус и состояние объектов, разные условия их существования, отличие методов формирования выборок и т. п. Дело биолога понять, какие из причин существенны в данном случае.

Вместе с этим на величине выборочных средних арифметических будут сказываться и случайные факторы. Поскольку действие этих факторов на каждую варианту различен, и нет двух одинаковых вариант, то нет и двух одинаковых наборов вариант, двух идентичных выборок. В полной мере случайные факторы проявляются только в бесконечной генеральной совокупности, а в ограниченных выборках их действие проявляется неполно, по-разному.

По этой причине выборочные средние арифметические всегда будут, во-первых, отличаться друг от друга, во-вторых, – от генеральной средней. Ограниченные объемы выборок недостаточны для того, чтобы полностью воспроизвести условия формирования генеральной средней; между генеральной и выборочными средними всегда будет отличие, ошибка «воспроизводимости», *ошибка репрезентативности*.

Сравнение средних арифметических по критерию Т Стьюдента

Задача сравнения выборочных средних – это вопрос о том, действовал ли в одной из выборок новый систематический фактор по сравнению с другой выборкой? В терминах статистики отличия между средними могут иметь два противоположных источника:

1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано в основном действием разных доминирующих факторов (а также и случайно).

Статистическая задача состоит в том, чтобы сделать обоснованный выбор. Исходно предполагается (Но): «достоверных отличий между средними нет».

Отличить закономерное от случайного можно только на основе знания законов поведения случайной величины. Для исключения чужеродных («выскакивающих») вариант мы применяли закон нормального распределения: в диапазоне четырех стандартных отклонений, $M \pm 1.96 \cdot S$, отклонение вариант от средней происходит по случайным причинам; за границами этого диапазона лежат чужеродные для данной выборки значения. Поскольку выборочные средние имеют нормальное распределение (см. раздел **Ошибка репрезентативности выборочных параметров**, с. 53), критерий отличия двух выборочных средних также базируется на свойствах *нормального распределения*: в границах $M_{общ.} \pm 1.96 \cdot m$ (или приблизительно $M_{общ.} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам. Критерий отличия средних формируется по типу критерия «исключения», если одну из выборочных средних (M_1) принять в качестве генеральной средней, другую взять как «подозрительную» варианту (M_2), а роль характеристики варьирования играет обобщенная ошибка репрезентативности (m_d):

$$t = \frac{x - M}{S} \Rightarrow t = \frac{M_1 - M_2}{m_d}.$$

Обобщенная ошибка получена объединением двух ошибок, рассчитанных по сравниваемым выборкам (для случая, когда выборочные дисперсии отличаются несильно):

$$m_d = \sqrt{m_1^2 + m_2^2},$$

которые, в свою очередь, определены рассмотренным выше соотношением:

$$m = \frac{S}{\sqrt{n}}.$$

Тогда рабочая формула для *T* критерия отличия средних будет:

$$T = \frac{|M_1 - M_2|}{\sqrt{m_1^2 + m_2^2}} \sim T_{(\alpha,\, df)}.$$

Следует помнить, что разность средних нужно брать по модулю, т. е. без учета знака. Полученное этим способом значение критерия *T* Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0.05$) и числе степеней свободы (объемы выборок без числа ограничений, $df = n_1+n_2-2$). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы достоверны. Если же полученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами недостоверны. Последнее говорит о том, что различия случайны, никакого определенного вывода сделать нельзя, нулевая гипотеза остается не опровергнутой.

Табличные значения критерия следует брать из таблицы Стьюдента (табл. 6*П)*. Обычно эта статистика соответствует нормальному распределению, но в случае небольших выборок дает необходимую поправку на объем выборки, предупреждает возможность сделать слишком жесткий вывод по недостаточным данным. По этой причине критерий различия средних арифметических носит название критерия Стьюдента. Одно из необходимых требований к применению этого критерия – это уверенность в том, что изучаемые признаки имеют распределение, в целом соответствующее нормальному. Если такой уверенности нет, для сравнения средних арифметических лучше воспользоваться непараметрическими критериями.

Рассмотрим такой пример. В процессе специальных исследований было установлено, что у стариков (20 человек) до лечения инсулином среднее содержание белков в крови составляло 81.04±1.7,

а после лечения – 79.33±1.6. Можно видеть, что полученные величины неодинаковы. Но достоверно ли это различие, закономерно ли оно? Можно ли утверждать, что лечение инсулином понижает содержание белков в крови? Согласно общей нулевой гипотезе, средние не отличаются. Проверим ее с помощью критерия Стьюдента:

$$T = \frac{M_1 - M_2}{\sqrt{m_1^2 + m_2^2}} = \frac{81.04 - 79.33}{\sqrt{1.7^2 + 1.6^2}} = 0.7.$$

По таблице граничных значений критерия (табл. 6*П)* находим, что для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = 20+20-2 = 38$ величина критерия составляет $T_{(0.05,39)} = 2.03$. Поскольку полученное значение (0.7) меньше табличного (2.03), нулевая гипотеза сохраняется, различия между средними величинами статистически недостоверны (незначимы). Следовательно, влияние инсулина на содержание белков в крови приведенными выше данными не подтверждается и остается недоказанным, возможно, из-за недостаточного числа определений.

В среде Excel определить величину *T* можно с помощью двух функций. Первая из них имеет формат:

=ТТЕСТ(массив1;массив2;хвосты;тип),

где массив1 – диапазон со значениями вариант первой выборки,

массив2 – диапазон со значениями вариант второй выборки,

хвосты – число, определяющее, какой критерий используется, односторонний или двухсторонний; обычно неизвестно, какая из средних величин должна быть больше, поэтому ставим 2 (двухсторонний),

тип – число, определяющее тип выполняемого теста, мы рассматривали двухвыборочный с равными дисперсиями, ставим 2 (двухпарный).

Результатом выполнения этой функции оказывается уровень значимости, соответствующий степени различия средних, т. е. вероятность того, что различия средних недостоверны. Поскольку обычно в биологии принимают в качестве границы уровень значимости $\alpha = 0.05$, все значения функции =ТТЕСТ, меньшие 0.05, будут свидетельствовать о достоверных отличиях сравниваемых средних арифметических. Для рассмотренного выше случая оценки действия инсулина функция показала:

=ТТЕСТ(диапазон1;диапазон2;2;2) = 0.492876.

Вероятность того, что отличия недостоверны, очень высока (α = 0.49)! Расчетные уровни значимости можно перевести в привычную форму T критерия Стьюдента с помощью второй функции:

=СТЬЮДРАСПОБР(вероятность;степени_свободы),

где вероятность – уровень значимости, рассчитанный функцией =ТТЕСТ, т. е. ссылка на ячейку, содержащую формулу этой функции,

степени_свободы – число степеней свободы $df = n_1+n_2-2$.

В нашем случае =СТЬЮДРАСПОБР(0.492876;38) = 0.7.

Если объемы сравниваемых выборок существенно отличаются ($n_1 \neq n_2$) или их дисперсии далеко не равны ($S^2_1 \neq S^2_2$), то для оценки достоверности отличий двух выборочных средних следует пользоваться другой, более точной, рабочей формулой:

$$T = \frac{\left|M_1 - M_2\right|}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\cdot\frac{(n_1-1)\cdot S_1^2+(n_2-1)\cdot S_2^2}{n_1+n_2-2}}}.$$

Сравним самцов и самок гадюки (см. данные в табл. 6.3 на стр. 108) по средней длине хвоста (M_1 = 81.6, M_2 = 65.1 мм), объемы выборок примерно одинаковы (n_1 = 8, n_2 = 9), зато дисперсии отличаются (S^2_1 = 24.8, S^2_2 = 6.9). Величина критерия составит:

$$T = \frac{\left|81.6-65.1\right|}{\sqrt{\left(\frac{1}{8}+\frac{1}{9}\right)\cdot\frac{(8-1)\cdot 24.8+(9-1)\cdot 6.9}{8+9-2}}} = 8.7.$$

Отличие средних достоверно, поскольку рассчитанное значение превышает табличное $T_{(0/05,15)}$ = 2.13.

Для этого случая при вычислениях в среде Excel следует использовать третий тип критерия – двухпарный с неравными отклонениями: =ТТЕСТ(диапазон1;диапазон2;2;3) = 0.00000627,

и далее =СТЬЮДРАСПОБР(0.00000627;15) = 6.8.

Значения 8.7 и 6.8 немного отличаются, поскольку формула критерия для функции Excel несколько отличается от приведенной и более чувствительна к отличию дисперсий. Обычно расчеты по обоим формулам совпадают.

Когда исследуемые признаки подчиняются другому закону распределения, к ним могут быть применены другие критерии. Рас-

смотрим случай с *распределением Пуассона*. Как уже говорилось, для признаков, подчиняющихся этому закону, характерно совпадение по величине средней арифметической и дисперсии. Это позволяет проводить сравнение и средних арифметических, и дисперсий по критерию *F* Фишера (подробнее см. ниже) и строить выводы одновременно и по различию средних, и по различию дисперсий.

$$F = \frac{S_1^2}{S_2^2 + 1} = \frac{M_1}{M_2 + 1} \sim F_{(\alpha,\, df1,\, df2)}.$$

Полученное значение сравнивается с табличным (табл. 7*П)* при выбранном уровне значимости ($\alpha = 0.05$) и степенях свободы $df_1 = 2 \cdot M_2 + 2$, $df_2 = 2 \cdot M_1$.

Рассмотрим случай сравнения частоты встречаемости растений (фиалка) на нескольких пробных площадках двух типов лугов. Для каждого луга получили средние значения 1.5 и 14.2 экз. на 1 площадку. Нулевая гипотеза состоит в том, что плотность данного вида на лугах одинакова. Критерий Фишера дает:

$$F = \frac{M_1}{M_2 + 1} = \frac{14.2}{1.5 + 1} = 5.68;\ df_1 = 2 \cdot (1.5+1) = 5,\ df_2 = 2 \cdot 14.2 = 28.$$

Значение $F = 5.68$ больше табличного $F_{(0.05,5,28)} = 5.67$; нулевую гипотезу можно отбросить и считать доказанным, что плотность растений на лугах разного типа достоверно отличается.

При сравнении достоверности различия долей (*p*) *альтернативных признаков* применяют критерий Фишера с φ-преобразованием. Вместо процентов берут фи-значения ($\varphi = \arcsin\sqrt{p}$ или по таблице 10*П)* и подставляют их в формулу:

$$F = \frac{(\varphi_1 - \varphi_2)^2 \cdot n_1 \cdot n_2}{n_1 + n_2} \sim F_{(\alpha,\, df1,\, df2)},$$

где φ_1 и φ_2 – преобразованные доли, n_1 и n_2 – объемы выборок.

Полученное значение сравнивают с табличным в соответствии с заданным уровнем значимости, $\alpha = 0.05$, и числом степеней свободы: $df_1 = 1$, $df_2 = n_1 + n_2 - 2$.

Например, в процессе учетов мелких млекопитающих в двух разных биотопах, где стояло по 200 ловушек, попалось соответственно 5 и 15 зверьков. Отличается ли численность животных на этих площадках? Если рассматривать ловушку как варианту, способную

принимать два значения – «пустая» и «сработавшая» (со зверьком), то получаем выборку вариант (ловушек) с альтернативным распределением. Число пойманных особей можно пересчитать в процент сработавших ловушек:

M_1 = 100%·5/200 = 2.5%, M_1 = 100%·15/200 = 7.5%. По таблице 10*П* находим значения φ и вычисляем значение критерия:

$$F = \frac{(0.318 - 1.555)^2 \cdot 200 \cdot 200}{200 + 200} = 5.62.$$

Полученная величина (5.62) больше критической $F_{(0.05,\ 1,\ 398)} = 3.9$, значит, численность мелких млекопитающих во втором биотопе достоверно выше, чем в первом.

Сравнение двух выборок по изменчивости признака

При сравнении двух выборок статистические критерии позволяют оценить достоверность отличий стандартных отклонений, дисперсий и коэффициента вариации, характеризующих степень разнородности вариант двух выборок. Здесь могут возникнуть сомнения, как можно ставить вопрос о достоверности различий показателей, выражающих действие случайных причин, как можно говорить о *неслучайном* отличии проявлений *случайности*? Казалось бы, случайное не может отличаться от случайного! Парадокс легко разрешается, если вспомнить, что влияющих на признак случайных причин множество: $x_{случ.} = \Sigma x_{случ.j}$ (см. раздел **Стандартное отклонение**, с. 41). Во-первых, на варианты разных выборок может действовать различное число случайных факторов, во-вторых, случайные факторы могут быть разного качества (сильные, слабые). На одну выборку по сравнению с другой может действовать *не случайно больше* случайных факторов или они могут быть *не случайно сильнее*. Чем больше более сильных случайных факторов будут вносить большую прибавку к значениям вариант, тем в большей степени одни варианты будут отличаться от других, тем выше будет изменчивость в такой выборке и тем выше будут ее оценки – дисперсия, стандартное отклонение и коэффициент вариации. Соответственно, если в сравниваемых выборках «действовали» случайные факторы, отличные по числу или качеству, две выборочные дисперсии будут отличаться достоверно.

Сравнение стандартных отклонений по критерию T Стьюдента

Существует два распространенных подхода к установлению достоверности отличий между выборочными дисперсиями, хотя нулевая гипотеза (Но) в обоих случаях одинакова: сравниваемые выборки взяты из одной генеральной совокупности, т. е. выборочные дисперсии служат отражениями одной и той же генеральной дисперсии. Стандартные отклонения можно сравнить с помощью критерия Стьюдента:

$$T = \frac{S_1 - S_2}{\sqrt{m_{s_1}^2 + m_{s_2}^2}} \sim T_{(0.05,\, n1+n2-2)},$$

оценив ошибки по формуле: $m_S = \frac{S}{\sqrt{2 \cdot n}}$.

Сравнение дисперсий по критерию F Фишера

Наиболее точным методом определения достоверности различий между выборочными дисперсиями служит критерий *F* Фишера, который представляет собой отношение дисперсий (большее значение должно стоять в числителе):

$$F = \frac{S_1^2}{S_2^2} \sim F_{(\alpha,\, df1,\, df2)},$$

где $S_1 > S_2$, $df_1 = n_1 - 1$, $df_2 = n_2 - 1$.

Если полученная величина *F* больше табличного значения при принятом уровне значимости (табл. 7*П* для $\alpha = 0.05$ и табл. 8*П* для $\alpha = 0.01$) и числе степеней свободы (df_1 и df_2), то различие между дисперсиями признается достоверным; если она меньше, то расхождение между ними может считаться несущественным, случайным, т. е. нулевая гипотеза не отвергается.

Рассмотрим такой пример. При сравнении по показателю плодовитости (число эмбрионов на самку) двух популяций красной полевки с разным уровнем численности (у первой, горной, популяции плотность населения в два раза выше, чем у равнинной) оказалось, что при очень близких средних арифметических (соответственно $M_1 = 5.8$ и $M_2 = 5.4$, разница статистически недостоверна)

стандартные отклонения значительно различаются: $S_1 = 1.82$, $S_2 = 0.52$ (при $n_1 = 27$, $n_2 = 12$). Отсюда

$$F = \frac{S_1^2}{S_2^2} = \frac{3.3124}{0.2704} = 12.25 .$$

Полученное значение критерия (12.2) больше табличного $F_{(0.05,26,11)} = 2.6$, следовательно, нулевую гипотезу о случайности отличий можно отбросить, сделав вывод о том, что показатели изменчивости плодовитости в разных по численности популяциях достоверно отличаются. С биологических позиций это понятно, поскольку генетические отличия между особями практически по всем признакам, включая плодовитость, в больших популяциях выше, чем в малых. Новым фактором, усиливающим изменчивость особей в выборке, становится возможность появления аберрантных форм в условиях более свободной панмиксии.

В среде Excel определить величину F можно с помощью двух функций. Первая из них имеет формат:

=ФТЕСТ(массив1;массив2),

где массив1 – диапазон со значениями вариант первой выборки,

массив2 – диапазон со значениями вариант второй выборки.

Результатом выполнения этой функции оказывается уровень значимости, соответствующий степени различия дисперсий, т. е. вероятность того, что различия дисперсий недостоверны. Поскольку обычно в биологии принимают в качестве границы уровень значимости $\alpha = 0.05$, все значения функции ФТЕСТ, меньшие 0.05, будут свидетельствовать о достоверных отличиях между выборочными дисперсиями. Для рассмотренного выше случая функция показала:

=ФТЕСТ(диапазон1;диапазон2) = 0.000058.

Расчетные уровни значимости можно перевести в привычную форму F критерия Фишера с помощью второй функции:

=FРАСПОБР(вероятность;степени_свободы1;степени_свободы2),

где вероятность – уровень значимости, рассчитанный функцией ФТЕСТ, или ссылка на ячейку, содержащую формулу этой функции,

степени_свободы1 – число степеней свободы для выборки с большей дисперсией, $df_1 = n_1 – 1$,

степени_свободы2 – число степеней свободы для выборки с меньшей дисперсией, $df_2 = n_2 – 1$.

В нашем случае =FРАСПОБР(0.000058;26;11) = 12.28.

Сравнение коэффициентов вариации по критерию Т Стьюдента

Коэффициенты вариации не имеют единиц измерения, поэтому их можно использовать для сравнения изменчивости разных показателей. Достоверность отличий коэффициентов оценивается с помощью критерия Стьюдента по формуле:

$$T = \frac{CV_1 - CV_2}{\sqrt{m_1^2 + m_2^2}} \sim T_{(0.05,n1+n2-2)},$$

где CV_1, CV_2 – значения коэффициентов вариации,

m_1, m_2 – ошибки коэффициентов вариации.

Вывод о достоверности отличий делается в том случае, если рассчитанное значение превысит табличное при заданном уровне значимости $\alpha = 0.05$ и числе степеней свободы $df = n_1 + n_2 - 2$. Сравним по критерию Стьюдента изменчивость веса тела землероек и плодовитости лисиц: $CV_1 = 8.6 \pm 0.77\%$, $n_1 = 63$; $CV_2 = 26.7 \pm 2.2\%$, $n_2 = 76$, отсюда

$$T = \frac{8.6 - 26.7}{\sqrt{0.77^2 + 2.2^2}} = 7.76.$$

Поскольку полученное значение (7.8) больше табличного ($T_{(0.05,\ 137)} = 1.96$), изменчивость плодовитости лисиц достоверно выше, чем изменчивость веса тела землероек.

Сравнение двух выборок в целом (непараметрические критерии)

Описанные выше статистические критерии (*T, F* и др.) относятся к параметрическим, так как используют стандартные параметры распределений (*M, S, n*). Они связаны с законом нормального распределения и применяются для оценки расхождения между генеральными параметрами по выборочным показателям сравниваемых совокупностей. Существенным достоинством параметрических критериев служит их большая статистическая мощность, т. е. широкие разрешающие возможности, а недостатком – трудоемкость расчетов, неприменимость к распределениям, сильно отклоняющимся от нормального, а также при исследовании качественных признаков.

Поэтому, наряду с параметрическими критериями, для ориентировочной оценки расхождений между выборками (особенно небольшими) применяются так называемые непараметрические критерии, ориентированные в первую очередь на исследование соотношений *рангов* исходных значений вариант (рассмотрение всех видов непараметрических критериев не входит в наши задачи). Они позволяют сравнивать выборки по качественным признакам, значения которых не имеют числового представления, но которые можно ранжировать. *Ранг – это число натурального ряда, которым обозначается порядковый номер каждого члена упорядоченной совокупности вариант.* К рангам неприложимы обычные арифметические действия, поэтому вычисления конструкции непараметрических критериев отличаются простотой.

Вся 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. При этом с параметрическими критериями их роднит общая идеологическая подоплека. Нулевая гипотеза, как правило, состоит в том, что сравниваемые выборки взяты из одной и той же генеральной совокупности, значит, характер распределения вариант в этих выборках должен быть сходным. Поскольку вместо самих значений вариант используются ранги, все непараметрические методы исследуют один вопрос, насколько равномерно варианты разных выборок «перемешаны» между собой. Если варианты разных выборок более или менее регулярно чередуются в общем упорядоченном ряду, значит, они распределены сходным образом и отличий между совокупностями нет. Если же выборки пересекаются не полно (смешиваются только краями распределений, либо одна 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 (довольно точный, но не очень простой для вычислений), критерий T Уайта (менее точный, но более простой), критерий λ (ламбда) Колмогорова – Смирнова (ориентирован на сравнение больших выборок) и критерий Q Розенбаума (самый простой для расчетов, но и не очень точный).

Критерий U Уилкоксона – Манна – Уитни

Этот метод сравнения двух выборок признается наиболее чувствительным, мощным и в то же время достаточно простым для расчетов. Согласно нулевой гипотезе, сравниваемые совокупности имеют одинаковые распределения.

Расчеты начинаются с ранжирования – варианты выборки упорядочиваются по возрастанию или убыванию и каждому значению присваивают ранг, равный порядковому номеру. Группам одинаковым (повторяющимся) значениям присваивают средний арифметический ранг. После этого ранги вариант суммируют отдельно по каждой выборке:

$R_1 = \Sigma r_i$, $R_2 = \Sigma r_j$, $i = 1, 2, \ldots, n_1$, $i = 1, 2, \ldots, n_2$, $n = n_1 + n_2$

и вычисляют величину критерия:

$$T = \frac{U - 0.5 \cdot n_1 \cdot n_2}{\sqrt{(n_1 \cdot n_2 \cdot (n+1)/12)}},$$

где $U = \max(U_1, U_2)$ – максимальное значение из двух величин:

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

$$U_2 = n_1 \cdot n_2 + 0.5 \cdot n_2 (n_2 + 1) - R_2.$$

Если выборка достаточно велика (n>20), то значение T сравнивается с табличным значением критерия Стьюдента для $df = \infty$ и $\alpha = 0.1$ (т. е. только для верхней 95% области нормального распределения). Считается, что метод хорошо работает для выборок объемом больше 10. В случае с меньшими выборками нужно пользоваться таблицами Уилкоксона – Манна – Уитни (табл. 11*П*).

В качестве примера сравним две выборки балльных оценок активности щелочной фосфатазы в лейкоцитах крови (E) самок старых (24 мес.) крыс, содержавшихся при естественном (ест.) и постоянном (пост.) освещении (данные из: Л. Б. Узенбаева и др. Успехи геронтологии. 2008. № 3).

$E_{\text{ест.}}$	3	1	3	3	3	3	3	3	3
$E_{\text{пост.}}$	4	4	3	4	3	3	3	4	4

Поскольку выборки баллов нельзя сравнивать с использованием параметрических критериев, воспользуемся критерием Уилкоксона. Ранжируем всю совокупность; упорядочим значения баллов по возрастанию. Затем упорядочим все значения вместе, но так,

чтобы баллы каждой выборки располагались в двух отдельных рядах ($E_{ест.}$, $E_{пост.}$). Такое расположение упрощает назначение рангов (ряды r_5, r_{35}; для парных значений 31 назначен средний ранг 2.5) и суммирование рангов (R):

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	R
$E_{ест.}$	1	3	3	3					3	3	3	3					
$E_{пост.}$					3	3	3	3					4	4	4	4	
$r_{ест.}$	1	7	7	7					7	7	7	7					50
$r_{пост.}$					7	7	7	7					14.5	14.5	14.5	14.5	86

$$U_1 = 9 \cdot 9 + 0.5 \cdot 9 \cdot (9+1) - 50 = 76,$$
$$U_2 = 9 \cdot 9 + 0.5 \cdot 9 \cdot (9+1) - 86 = 40,$$
$$U = \max(U_1, U_2) = 66.5,$$
$$n = n_1 + n_2 = 9 + 9 = 18,$$
$$T = \frac{76 - 0.5 \cdot 9 \cdot 9}{\sqrt{(9 \cdot 9 \cdot 19/12)}} = 3.1.$$

Полученное значение (3.1) больше табличного ($T_{(0.1,\infty)} = 1.65$; табл. 6*П*), т. е. активность щелочной фосфатазы имеет более высокий уровень у крыс, выросших в экстремальных условиях постоянного освещения. Поскольку выборки малы, воспользуемся точными таблицами Уилкоксона – Манна – Уитни (табл. 11*П*). Получаем $T_{(0.05,n1,n2)} = T_{(0.05,9,9)} = 63$. Полученное значение (76) больше табличного (51), следовательно, различия между выборками достоверны.

Критерий Т Уайта

Этот критерий применяется для проверки нулевой гипотезы о сходстве двух независимых распределений. Этот критерий более грубый, чем предыдущий, зато почти не требует вычислений. Техника расчетов аналогична; результатом первого этапа обработки должны стать два значения суммы рангов по выборкам, из которых выбирается меньшее значение: $T = min(R_1, R_2)$.

Достоверность отличий выборок оценивается с помощью критерия T Уайта по специальной таблице 12*П*. Полученная величина T сравнивается с табличным значением критерия с учетом объ-

ема сравниваемых совокупностей для принятой доверительной вероятности: $P = 0.95$ или $P = 0.99$ (т. е. для уровня значимости $\alpha = 0.05$ и $\alpha = 0.01$). Если расчетное значение T меньше табличного числа ($T_Э < T_Т$), значит, обнаружены достоверные отличия между выборками, и нулевая гипотеза (о том, что распределения одинаковы) отвергается. Если же фактическая величина критерия T больше или равна табличной ($T_Э \geq T_Т$), нулевая гипотеза сохраняется и различие между выборками считается статистически недостоверным. Следует обратить внимание на то обстоятельство, что для многих непараметрических статистик вывод о достоверности отличий делается в случае, если расчетное значение критерия *меньше* табличного, тогда как параметрические статистики дают заключения о значимости различий, когда расчетная величина критерия *больше* табличной.

Используем этот метод для примера, рассмотренного выше. Суммы рангов для каждой совокупности составили: $R_1 = 76$, $R_2 = 40$.

Меньшую сумму $T = 40$ сравниваем с табличным значением критерия для $n_1 = 9$ и $n_2 = 9$ ($T_{(0.05,9,9)} = 63$). Поскольку полученное значение (40) меньше табличного (63), наблюдаемые различия в активности щелочной фосфатазы крови крыс из разных условий содержания носят неслучайный характер, т. е. статистически достоверны, нулевая гипотеза о сходстве выборок отклоняется.

Это заключение соответствует статистическому выводу, сделанному в предыдущем разделе. Ясно, что наблюдаемое отличие пока имеет отчетливое выражение и для заключения было достаточно и столь незначительного объема статистического материала.

Критерий Q Розенбаума

Этот критерий, как и предыдущие, оценивает достоверность различий двух эмпирических распределений, но в отличие от них почти не требует вычислений. Сравним два ряда цифр, характеризующих привесы (г) барашков одного возраста при добавлении в корм специальной подкормки (234, 277, 214, 201, 174, 167, 184, 157, 196, 173, 190, 191, 141, 150, 191) и без нее (183, 154, 175, 159, 157, 189, 198, 165, 176, 124, 173, 182, 204, 151, 147). Устанавливаем максимальные (277 и 204) и минимальные (141 и 124) значения и определяем порядковый номер сравниваемых совокупностей. В качестве

первой следует принять выборку с *наибольшей вариантой* 277.

Далее находим число значений первой выборки, *превышающих* максимальное значение второй выборки (204): $Q_1 = 3$ (234, 277, 214). Затем определяем число вариант второй выборки, *уступающих* по величине минимальному значению первой выборки (141): $Q_2 = 1$ (124). Далее определяем критерий Розенбаума как сумму полученных чисел: $Q = Q_1+Q_2 = 3+1 = 4$. По таблице 13*П* находим критическое значение $Q_{(0.05,15,15)} = 6$. Поскольку эмпирическое значение (4) меньше табличного (6), приходим к выводу об отсутствии достоверного отличия выборок друг от друга, а значит, и влияния подкормки на привесы барашков. Следует все же иметь в виду, что возможности этого метода ограничены, он дает лишь прикидочный результат и оказывается эффективным только в случае сравнительно больших различий между выборками.

Сравнение двух выборок по силе корреляции двух признаков

Изложенный здесь материал следует читать после раздела 8. До сих пор были рассмотрены выборки вариант, несущих по одному значению. Корреляционный анализ изучает выборки, в которых каждая варианта охарактеризована двумя признаками (x, y). В центр внимания ставится не просто варьирование, но сопряженное варьирование значений двух признаков. При сравнении двух выборок исследуется вопрос о том, из одной ли генеральной совокупности они извлечены и не одинакова ли в них сила сопряженного варьирования признаков x и y. Нулевая гипотеза предполагает общность происхождения выборок, т. е. два коэффициента корреляции (для двух выборок) оценивают один и тот же характер зависимости между признаками генеральной совокупности, но искажены случайностью.

Для получения сопоставимых случайных величин используют z-преобразование коэффициентов корреляции: $z = 0.5 \cdot \ln\left(\frac{1+r}{1-r}\right)$ (или по табл. 14*П*; знак сохраняется) и сравнивают полученные величины с помощью критерия Стьюдента:

$$T = \frac{|z_1 - z_2|}{m_z} \sim T_{(0.05,\, n1+n2-4)},$$

где m_z – обобщенная ошибка преобразованных коэффициентов:

$$m_z = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}},$$

n_1, n_2 – объемы сравниваемых выборок.

Полученное значение сравниваем с табличным при принятом уровне значимости и числе степеней свободы $df = n_1 + n_2 - 4$.

Сравним корреляционное сходство между тремя биотопами (**с**основник, **е**льник, **л**иственный лес) по численности (n) обитающих там 13 видов мелких млекопитающих (исходные данные по отловам в канавки представлены в табл. 6.2).

Таблица 6.2

Вид	**Численность, *n***				***N***		
	n_C	n_E	$n_Л$	M	$n_C - M$	$n_E - M$	$n_Л - M$
Обыкн. бурозубка	3.9	7.2	6.0	5.700	–1.800	1.500	0.300
Средняя бурозубка	1.8	1.1	0.5	1.133	0.667	–0.033	–0.633
Малая бурозубка	1.9	2.0	1.6	1.833	0.067	0.167	–0.233
Равнозубая бурозубка	0.01	0.2	0.1	0.103	–0.093	0.097	–0.003
Крошечная бурозубка	0.04	0.04	0	0.027	0.013	0.013	–0.027
Водяная кутора	0.04	0.06	0.4	0.167	–0.127	–0.107	0.233
Лесная мышовка	0.6	0.3	0.7	0.533	0.067	–0.233	0.167
Лесной лемминг	0.2	0	0.05	0.083	0.117	–0.083	–0.033
Мышь-малютка	0.04	0	0	0.013	0.027	–0.013	–0.013
Рыжая полевка	1.5	0.8	0.8	1.033	0.467	–0.233	–0.233
Красная полевка	0.06	0.6	0.02	0.227	–0.167	0.373	–0.207
Темная полевка	0.2	0	0.7	0.300	–0.100	–0.300	0.400
Полевка-экономка	0	0.2	0.2	0.133	–0.133	0.067	0.067

По исходным данным корреляция между ельником и сосняком составила $r_{EC} = 0.916$, между ельником и лиственным лесом – $r_{СЛ} = 0.981$. Спрашивается, действительно ли население ельников более сходно с населением лиственных лесов, чем с населением сосняков ($r_{СЛ} > r_{EC}$)? В таблице 14*П* находим: $z_{СЛ} = 1.5275$, $z_{СЛ} = 2.2976$. Вычисляем критерий Стьюдента:

$$m_z = \sqrt{\frac{1}{13 - 3} + \frac{1}{13 - 3}} = \sqrt{0.2} = 0.447214,$$

$$T = \frac{|1.5375 - 2.2976|}{0.4472} = 0.06.$$

Полученное значение критерия (0.06) меньше табличного $T_{(0.05,13+13-4)} = 2.07$, значит, отличие коэффициентов корреляции незначимо. Сосняк и березняк неотличимы от ельников по соотношениям численности разных видов мелких млекопитающих. Причина этого, на первый взгляд, странного вывода состоит в том, что корреляция между биотопами характеризует не видовые предпочтения, но структуру доминирования. Средние многолетние показатели попадаемости бурозубок в любом биотопе на 2–3 порядка выше, чем остальных видов, что обеспечивает сильную вытянутость эллипсу рассеяния и автоматически высокий уровень корреляции. Очевидно, что в таком виде корреляционная мера сходства оказалась неудачной.

Для изучения биотопической приверженности видов (а это реальный факт) следовало бы изучить зависимость показателей встречаемости, центрированных (или нормированных) на общую численность (M) для всего района работ:

$N_{ij} = n_{ij} - M_j$,

где j – индекс биотопа, всего m биотопов ($j = 1, 2 \ldots m$),

i – индекс вида, всего k видов ($i = 1, 2 \ldots k$).

Теперь корреляция между ельником и сосняком составила: $r_{EC} = -0.883$, между ельником и лиственным лесом – $r_{CЛ} = 0.168$. По таблице 14*П* находим: $z_{EC} = -1.3758$, $z_{CЛ} = 0.1614$. Критерий Стьюдента (при той же ошибке) равен:

$$T = \frac{|-1.3758 - 0.1614|}{0.4472} = 2.71.$$

Полученное значение больше табличного $T_{(0.05,13+13-4)} = 2.07$, коэффициенты корреляции между разными биотопами (для удельных показателей численности) отличаются достоверно. Между сосняком и ельником наметился определенный антагонизм, а сходство ельника с лиственным лесом оказалось несущественным.

Сравнение двух линий регрессии

Изложенный здесь материал следует читать после ознакомления с разделом 8. Регрессионный анализ, рассматривая зависи-

мость между признаками, выражает ее специфическим образом – через уравнения регрессии. Линейные уравнения вида $Y = ax+b$ содержат два коэффициента регрессии, характеризующие степень сопряжения и пропорциональность изменения признаков (коэффициент a отражает силу связи, т. е. наклон линии) и место пересечения оси ординат (коэффициент b определяет место положения линии в осях координат). Когда ставится вопрос о сходстве характера связи между признаками, то в отношении линии регрессии он распадается на три отдельных вопроса:

- одинаков ли характер распределения признаков?
- одинаков ли наклон линий регрессии?
- одинаково ли положение линий регрессии относительно осей координат?

1) Для того чтобы решить вопрос о сходстве угла наклона линий регрессии, необходимо убедиться в том, что обе линии характеризуются одной и той же случайной дисперсией, сходным характером рассеяния вариант вокруг линий, т. е. сходными значениями случайной дисперсии, Но: $S^2_{остат.1} = S^2_{остат.2}$. Эта первая гипотеза проверяется с помощью F критерия Фишера:

$$F = \frac{S^2_{остат.1}}{S^2_{остат.2}} \sim F_{(\alpha,\, df1,\, df2)},\ S^2_{остат.} = \sum^{n} (y_x - Y_x)^2 /(n-2),$$

где $S^2_{остат.}$ – остаточная дисперсия, сумма квадратов отклонения исходных значений (y_x) от рассчитанных по уравнению регрессии (Y_x), нормированная на число степеней свободы (n–2). Это значение получают из таблицы дисперсионного анализа регрессионной модели («Остаток»).

2) Если остаточные дисперсии для разных линий значимо не отличаются, то можно приступать к сравнению коэффициентов регрессии, определяющих характер зависимости между признаками, т. е. ответственных за угол наклона прямых. Этой цели служит T критерий Стьюдента:

$$T = \frac{a_1 - a_2}{m_{a1,2}} \sim T_{(\alpha,\, df)},$$

где a_1, a_2 – коэффициенты регрессии сравниваемых уравнений,
$m_{a1,2}$ – обобщенная ошибка коэффициентов регрессии.

Для выборок одинакового объема обобщенная ошибка рассчитывается по формуле:

$$m_{a1,2} = \sqrt{m_{a1}^2 + m_{a2}^2},$$

где m_{a1}, m_{a2} – ошибки коэффициентов регрессии:

$$m_a = \sqrt{\frac{(1-r^2)}{n-2}} \cdot \frac{S_y}{S_x},$$

Sy, *Sx* – стандартные отклонения, рассчитанные по всему объему выборки *n*,

r – коэффициент корреляции между признаками *x* и *y*.

Для выборок, имеющих разный объем, обобщенная ошибка репрезентативности коэффициентов регрессии вычисляется более сложным путем:

$$m_{a1,2} = S_{остат.} \cdot \sqrt{\frac{1}{C_{x1}} + \frac{1}{C_{x2}}},$$

где $S_{остат.}$ – обобщенная остаточная дисперсия, вычисленная по формуле:

$$S_{остат.} = \sqrt{\frac{(n_1-2)\cdot S_{остат.1}^2 + (n_2-2)\cdot S_{остат.2}^2}{n_1+n_2-4}},$$

C_{x1}, C_{x2} – суммы квадратов отклонений значений признака *x* от своих средних (M_X) в двух выборках:

$$C_x = \sum^{n} (x - M_x)^2,$$

$S_{остат.1}^2, S_{остат.2}^2$ – остаточные дисперсии (см. выше).

Различие между коэффициентами регрессии a_1 и a_2 считается значимым, если расчетное значение критерия Стьюдента превосходит табличное значение при заданном уровне значимости и числе степеней свободы $df = n_1+n_2-4$.

3) Если критерий Стьюдента не показал отличий коэффициентов регрессии, то проверяется, наконец, третья гипотеза – об одинаковом положении линий регрессии (т. е. гипотеза о полном совпадении линий) – с помощью *T* критерия Стьюдента:

$$T = \frac{b_1 - b_2 - a \cdot (M_{x1} - M_{x2})}{S_{остат.} \cdot \sqrt{\left[\frac{\frac{1}{n_1} + \frac{1}{n_2} + (M_{x1} - M_{x2})^2}{C_{x1} + C_{x2}} \right]}} \sim T_{(\alpha,\, df),}$$

где a – усредненный коэффициент корреляции

$$a = \frac{C_{x1} \cdot a_1 + C_{x2} \cdot a_2}{C_{x1} + C_{x2}},$$

M_{x1}, M_{x2} – средние для признака x в двух выборках,

Различие между коэффициентами регрессии b_1 и b_2 считается значимым, если расчетное значение критерия Стьюдента превосходит табличное значение при заданном уровне значимости и числе степеней свободы $df = n_1+n_2–3$.

В качестве примера сравним характер зависимости между длиной хвоста (*Lc*, мм) и длиной тела (*Lt*, см) у самцов (*m*) и самок (*f*) обыкновенной гадюки (табл. 6.3), уравнения регрессии приведены на иллюстрации (рис. 6.2).

1) Найти остаточные дисперсии $S^2_{остат.}$ для каждой выборки проще всего, выполнив полный регрессионный анализ в среде Excel с помощью макроса, вызываемого командой меню Сервис\ Анализ данных\ Регрессия.

Получим: $S^2_{остат.1} = 12.202$, $S^2_{остат.2} = 4.006$,

отсюда $F = \frac{S^2_{остат.1}}{S^2_{остат.2}} = \frac{12.202}{4.006} = 3.046.$

Поскольку полученное значение (3.04) меньше табличного $F_{(\alpha,df1,df2)} = 3.4$, отличия между дисперсиями незначимы. Можно продолжать сравнение линий регрессии.

2) Для проверки различий коэффициентов регрессии требуется найти обобщенную ошибку $m_{a1,2}$, используя значения ошибок из таблиц проведенного ранее регрессионного анализа в среде Excel. Поскольку объемы выборок отличаются не сильно, можно использовать первую формулу:

$$m_{a1,2} = \sqrt{m^2_{a1} + m^2_{a2}} = \sqrt{0.4436^2 + 0.27695^2} = 0.52298.$$

Таблица 6.3

	A	B	C
1		***Lt***	***Lc***
2	m_1	45	77
3	m_2	46	84
4	m_3	47	81
5	m_4	45	76
6	m_5	47	80
7	m_6	50	78
8	m_7	53	90
9	m_8	51	87
10		***Lt***	***Lc***
11	f_9	50	62
12	f_{10}	55	65
13	f_{11}	49	65
14	f_{12}	51	66
15	f_{13}	52	64
16	f_{14}	50.5	64
17	f_{15}	53	68
18	f_{16}	51	62
19	f_{17}	57	70

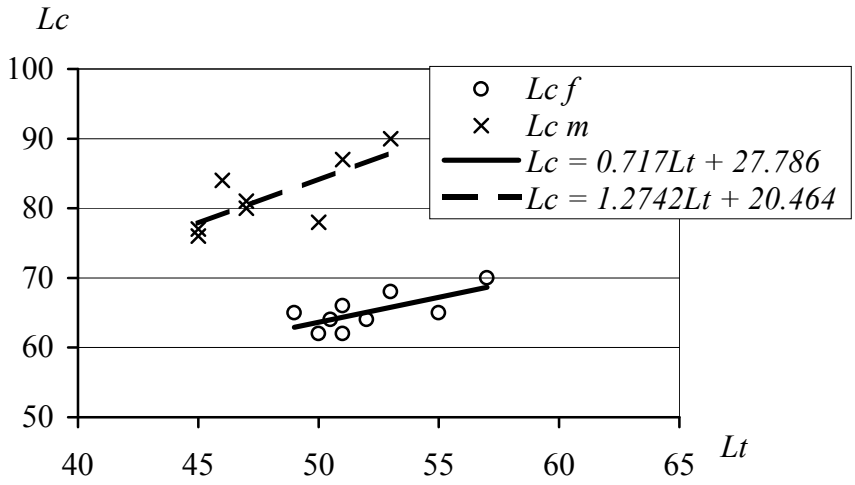


Рис. 6.1. *Регрессия длины хвоста по длине тела у гадюк*

Для целей иллюстрации рассчитаем и более точную оценку. Для этого предварительно нужно найти суммы квадратов отклонений значений независимой переменной *x* (в нашем случае ее роль играет длина тел *Lt*) от своих средних. Найдем величины с помощью функции Excel =КВАДРОТКЛ(диапазон). Для таблицы 6.3 имеем:

C_{x1} =КВАДРОТКЛ(C2:C9) = 62,

C_{x2} =КВАДРОТКЛ(C11:C19) = 52.222.

Поскольку общая остаточная дисперсию $S^2_{остат.}$ равна:

$$S_{остат.} = \sqrt{\frac{(n_1-2)\cdot S^2_{остат.1} + (n_2-2)\cdot S^2_{остат.2}}{n_1+n_2-4}} = \sqrt{\frac{(8-2)\cdot 12.202 + (9-2)\cdot 4.006}{8+9-4}} = 2.7908,$$

обобщенная ошибка коэффициентов регрессии составит:

$$m_{a1,2} = S_{остат.}\cdot\sqrt{\frac{1}{C_{x1}}+\frac{1}{C_{x2}}} = 2.7908\cdot\sqrt{\frac{1}{62}+\frac{1}{52.222}} = 0.52419,$$

т. е. практически не отличается от рассчитанной первым способом. Теперь можно оценить значимость отличий коэффициентов (для $df = n_1+n_2-4 = 8+9-4 = 13$):

$$T = \frac{a_1 - a_2}{m_{a1,2}} = \frac{1.27419 - 0.71702}{0.52419} = 10.76.$$

Полученное значение критерия Стьюдента больше табличного даже для уровня значимости $\alpha = 0.001$($T_{(0.001,13)} = 4.22$), т. е. коэффициенты регрессии не равны.

Итак, результаты сравнения показывают, что линии регрессии имеют разный угол наклона; с увеличением размеров тела длина хвоста у самцов ($a = 1.2$) прирастает быстрее, чем у самок ($a = 0.7$).

Сравнение двух выборок по характеру распределения

Рассмотренные выше методы сравнения двух выборок проверяют предположение либо о действии систематического, контролируемого, фактора (по критерию Стьюдента оценивается различие средних), либо о действии разного набора случайных факторов (критерий Фишера пытается обнаружить отличие дисперсий), либо обеих причин вместе (с помощью непараметрических статистик).

Специфические методы χ^2 Пирсона и λ Колмогорова – Смирнова позволяют проверять гипотезы о соответствии друг другу двух частотных распределений и тем самым улавливать не только отличия в общих тенденциях, но и частные особенности отдельных классов вариант.

Критерий χ^2 Пирсона

Критерий позволяет выяснить, насколько полученный экспериментатором фактический материал подтверждает теоретическое предположение, в какой мере анализируемые данные совпадают с теоретически ожидаемыми. Возникает задача статистической оценки разницы между фактическим и теоретическим распределениями. С формальных позиций сравниваются два вариационных ряда, две выборки: одна – эмпирическое распределение, другая представляет собой выборку с теми же параметрами (*n*, *M*, *S* и др.), что и эмпирическая, но ее частотное распределение построено в точном соответствии с выбранным теоретическим законом (нормальным, Пуассона, биномиальным и др.), которому предположительно подчиняется поведение изучаемой случайной величины.

Нулевая гипотеза предполагает отсутствие различий между сравниваемыми распределениями. Для ее проверки и служит «критерий согласия» χ^2 Пирсона:

$$\chi^2 = \sum \frac{(a - A)^2}{A},$$

где a – фактическая частота наблюдений,

A – теоретически ожидаемая частота для данного класса.

Расчетное значение критерия сравнивают с критическим значением для принятых уровня значимости (α) и числа степеней свободы (df) (табл. 9*П)*. Если вычисленная величина χ^2 равна или превышает табличную $\chi^2_{(\alpha, df)}$, решают, что эмпирическое распределение от теоретического отличается достоверно. Тем самым гипотеза об отсутствии этих различий будет опровергнута. Если же $\chi^2 < \chi^2_{(\alpha, df)}$, то нулевая гипотеза остается в силе. Обычно принято считать допустимым уровень значимости $\alpha = 0.05$, так как в этом случае остается только 5% шансов, что нулевая гипотеза правильна и, следовательно, есть достаточно оснований (95%), чтобы от нее отказаться.

Как и раньше, для определения числа степеней свободы из общего объема выборки нужно вычесть число ограничений (т. е. число параметров, использованных для расчета теоретических частот). Однако необходимо помнить, что в случае с критерием хи-квадрат для определения числа степеней свободы используют не объем выборки n, а число классов частотного распределения k.

Для *альтернативного* распределения ($k = 2$) в расчетах участвует только один параметр, объем выборки, следовательно, число для него $df = k–1 = 2–1 = 1$. Для проверки равномерности распределения результатов дигибридного скрещивания (известно четыре класса) $df = k–1 = 4–1 = 3$. Для проверки соответствия вариационного ряда распределению *Пуассона* используются уже два параметра – объем выборки и среднее значение (численно совпадающее с дисперсией); число степеней свободы $df = k–2$. При проверке соответствия эмпирического распределения вариант *нормальному* или *биномиальному* закону число степеней свободы берется как число фактических классов минус три условия построения рядов – объем выборки, средняя и дисперсия, $df = k–3$. Сразу стоит отметить, что критерий χ^2 работает только для выборок *объемом не менее 25 вариант*, а частоты отдельных классов должны быть *не ниже 4*.

Общий порядок работы таков. Сначала строится вариационный ряд, т. е. частотное (a) распределение для фактических данных. Затем формулируются теоретические соображения о том, какой тип распределения реализуется в изучаемой совокупности. В соответствии с этим выдвигается нулевая гипотеза: «эмпирические частоты соответствуют данному типу распределения» или, что то же самое, «в генеральной совокупности реализован такой-то тип распределения». На следующем этапе формируется «теоретическая выборка». Для этого, во-первых, требуется явно вычислить теоретические частости (p), соответствующие значениям вариационного ряда. Пожалуй, это самый ответственный момент всех расчетов, поскольку ранее высказанная идея воплощается в числа – теоретические частости данного значения. После этого рассчитываются частоты распределения выбранного теоретического типа (A) для конкретных параметров исходной выборки. Завершается процедура расчетом величины критерия хи-квадрат (χ^2), ее сопоставлением с табличным значением ($\chi^2_{(\alpha, df)}$). В итоге формулируется статистический вывод о соответствии или не соответствии эмпирических рядов теоретиче-

скому распределению. Это дает возможность прийти к тому или иному биологическому заключению.

В качестве первого примера решим задачу, соответствует ли **закону Пуассона** распределение числа повторных отловов альбатросов (табл. 6.4). В этом случае рассматривается процесс, этапами которого выступают события «отлов птицы». В чреде таких событий встречаются редкие – «отлов меченной особи». Биологическая подоплека состоит в следующем: случайны ли повторные отловы птиц или есть факторы, ответственные за нарушение случайности? Например, птицы могут приманиваться и стремиться попасть вновь либо могут стараться избежать повторного отлова. В обоих случаях птицы будут «умышленно» попадаться чаще или реже, нарушая *случайность* повторного отлова и искажая тем самым форму распределения, которое будет отходить от формы, предписанной законом Пуассона. Согласно нулевой гипотезе, птицы ведут себя случайно, их встречаемость соответствует этому закону.

Алгоритм расчетов теоретических частот для распределения Пуассона достаточно прост и основан на формулах, не требующих предварительного расчета теоретических частостей *p*:

$$A_0 = \frac{n}{e^M}$$ (частота нулевого класса),

$$A_x = \frac{M}{x} \cdot A_{x-1}$$ (частота прочих классов),

где M – средняя арифметическая ряда,

x – значение ряда (число объектов в пробе),

A_x – теоретическая частота значения x,

n – объем выборки (число проб),

e = 2.7183… – основание натурального логарифма.

Параметры данного вариационного ряда были рассчитаны в разделе **Основные типы распределений**: M = 0.968. Теоретическая частота нулевого значения равна:

$$A_0 = \frac{n}{e^M} = \frac{32}{e^{0.968}} = 11.93803 \approx 12,$$

Таблица 6.4

Число повторных отловов, *x*	**Фактическая частота, *a***	**Теоретическая частота, *A***	$\frac{(a-A)^2}{A}$
0	15	12	0.75
1	7	11	1.45
2	7	6	
3	2 } 10	2 } 9	0.17
4	1	1	
Сумма	$n = \Sigma a = 32$	$n = \Sigma A = 32$	$\chi^2 = 2.31$

частота значения $x = 1$:

$$A_x = \frac{M}{x} \cdot A_{x-1} = \frac{0.968}{1} \cdot 11.93 = 11.55602 \approx 11$$

и т. д. (табл. 6.4, графа 3).

По окончании вычислений получаем два ряда частот, отличия между которыми оцениваются по критерию хи-квадрат.

Перед расчетом значения критерия следует убедиться, что выполнены требования к данным для расчета критерия χ^2:

– объем выборки более 25 вариант, $n>25$,

– суммы эмпирических и теоретических частот равны объему выборки $n = \Sigma a = \Sigma A$ (с точностью не ниже 1–2%),

– все классы эмпирического и теоретического рядов имеют частоты более 4, $a_j>4$; если какие-либо классы имеют меньше 4 вариант (у нас значения 3 и 4 имею частоты 2 и 1), то они должны быть объединены (суммированы) с соседними, что и показано в таблице с помощью фигурных скобок. Далее вычисляем значения критерия: для первой строки

$$\frac{(a-A)^2}{A} = \frac{(15-12)^2}{12} = 0.75$$

и т. д. (графа 4), итого $\chi^2 = 2.31$. Число степеней свободы находим как число окончательных классов (3) минус число ограничений (средняя и объем выборки): $df = k–2 = 3–2 = 1$.

Табличное значение $\chi^2_{(0.05,1)} = 3.84$. Полученная величина (2.31) меньше табличной (3.84), следовательно, нулевая гипотеза не

отвергается: эмпирическое распределение достоверно не отличается от распределения Пуассона. Иными словами, у нас нет оснований утверждать, что вероятность повторного отлова изменяется: нельзя утверждать, что операция отлова птиц привлекает или пугает.

Кстати, соответствие эмпирического ряда *распределению Пуассона* можно проверить и другим способом, сравнив по критерию Фишера величины средней арифметической и дисперсии для числа степеней свободы: $df_1 = n–1$, $df_2 = n–1$. В нашем случае $M = 0.968$, $S^2 = 1.257$; $F = 1.257/0.968 = 1.157$. Поскольку эта величина меньше табличной ($F_{(0.05,31,31)} = 1.84$), сравниваемые показатели достоверно не отличаются, а равенство средней и дисперсии характерно лишь для распределения Пуассона.

В качестве второго примера рассмотрим анализ пространственного размещения особей. Как известно, есть три важнейших типа размещения: регулярное (соответствующее жестким конкурентным отношениям), агрегированное (скученность особей вблизи от источников необходимых ресурсов) и случайное (когда нет острой конкуренции или дефицита ресурсов). Зная тип размещения особей, можно многое сказать об их биологии. Судить о характере пространственного размещения можно по распределению встреч особей по небольшим одинаковым пробным площадкам, на которые разбивается исследуемая территория (рис. 6.2). Равномерное территориальное размещение особей дает унимодальное распределение встреч (одна вершина повышенных частот) (рис. 6.2, *В*). Если наблюдается агрегация, то имеет место бимодальное распределение (много площадок без особей, много площадок с несколькими особями и мало площадок с единичными экземплярами) (рис. 6.2, *Б*). Когда же размещение животных или растений по территории местообитания случайно, при обобщении получается частотное распределение Пуассона (рис.6.2, *А*). Поэтому, проверяя, соответствует ли этому закону эмпирическое распределение особей по площадкам, мы тем самым проверяем гипотезу о случайном размещении организмов в пространстве. Возьмемся проверить, действительно ли на иллюстрации «случайное размещение» из монографии А. М. Гилярова (1990, с. 41, рис. 8) точки размещены случайно? Разбиваем территорию на пробные площадки, нарисовав сетку. Подсчитываем число площадок (a), на которых встретилось разное число точек (x), формируем вариационный ряд (табл. 6.5).

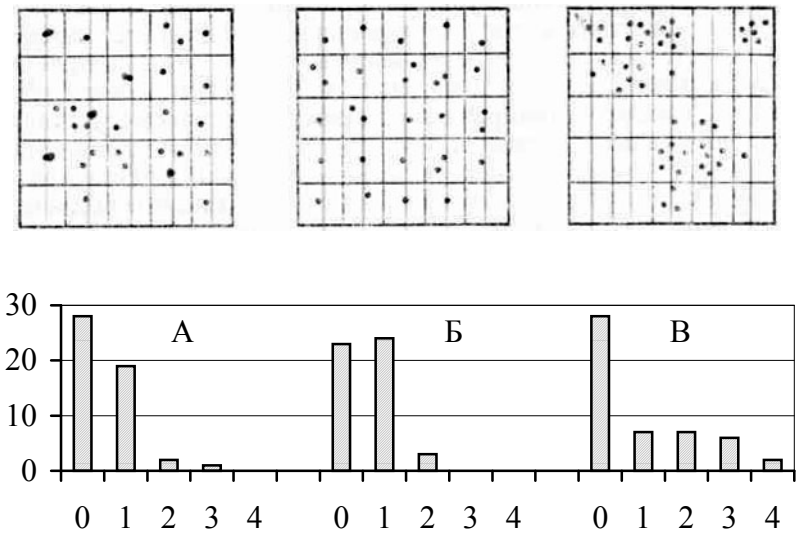


Рис. 6.2. *Территориальное размещение особей и соответствующие распределения*

Таблица 6.5

Число точек на одной площадке, *x*	**Фактическая частота, *a***	**Теоретическая частота, *A***	$\frac{(a-A)^2}{A}$
0	28	29.7	2.98
1	19	15.5	
2	2 } 22	4.0 } 20.3	3.01
3	1	0.7	
4	0	0.1	
Сумма	$n = \Sigma a = 50$	$n = \Sigma A = 50$	$\chi^2 = 5.99$

Определяем объем выборки ($n = 50$), среднюю арифметическую ($M = 0.52$). Предполагая распределение Пуассона, рассчитываем по алгоритму теоретические частоты (A), объединяем классы, где частоты меньше 4, вычисляем χ^2, отыскиваем табличное значение $\chi^2_{(0.05,1)} = 3.84$. Поскольку полученное значение критерия (5.99) больше табличного (3.84), эмпирическое распределение отличается

от распределения Пуассона. На иллюстрации отображено не случайное размещение особей в пространстве, поскольку пустых площадок слишком мало, а единичных слишком много; размещение точек тяготеет к агрегированному. Такому типу лучше соответствует биномиальное распределение с неравными вероятностями исходов.

Теория статистического оценивания строится на идее **нормального распределения**. Многие из параметров и критериев предлагаются ею в предположении, что изучаемые признаки имеют нормальное распределение. По большому счету, используя статистические методы для описания непрерывных признаков, нужно быть уверенным, что они действительно подчиняются нормальному закону, а в случае дискретных признаков – биномиальному. Для такой проверки нулевая гипотеза звучит так: «полученное распределение соответствует нормальному (биномиальному)» или «выборка взята из генеральной совокупности, подчиняющейся закону нормального (биномиального) распределения».

Все вычислительные операции для случаев нормального и биномиального распределений совпадают. Рассмотрим проверку на *не*-нормальность распределения массы тела бурозубок.

Расчеты начинаются с построения вариационного ряда и поиска центральных значений для каждого класса (табл. 6.6 и 6.7). Далее по формуле $t = \frac{|x_j - M|}{S}$ вычисляются нормированные отклонения середины каждого классового интервала (x_j) от общей средней M (S – стандартное отклонение). В нашем случае $M = 9.29$ г, $S = 0.897$ г. Для второго интервала: $t = |8.05–9.27|/0.897 = 1.38$. Далее определяем теоретические частости нормального распределения, или ординаты нормальной кривой (табл. 4*П)*, соответствующие вычисленным нормированным отклонениям. Для $t = 1.38$ находим $p = 0.1539 \approx 0.15$ (табл. 6.6, графа 5). (Следует отметить, что модуль в формуле нормированных отклонений берется потому, что в таблице 6*П* приведены частости p только для положительных значений t.) Следующая операция, вычисление теоретических частот распределения, ведется по формуле:

$$A = c \cdot p,$$

где p – ординаты нормальной кривой;

c – константа ряда, определяемая по формуле $c = \frac{dx \cdot n}{S}$,

dx – классовый интервал (в данном случае он равен 0.7);

n – объем выборки (63).

Для нашего примера $c = \frac{0.7 \cdot 63}{0.897} = 49.16$.

Теоретическая частота для $f = 0.15$ составит:

$A = 49.16 \cdot 0.1539 = 7.55 \approx 8$ (графа 6).

В результате вычислений получаем теоретическую выборку с параметрами $M = 9.29$ г, $S = 0.897$ г, $n = 63$, частоты которой соответствуют нормальному распределению (см. рис. 3.3, с. 63).

Таблица 6.6

Классовые интервалы	**Центр интервала, x_j**	**Фактическая частота, a**		**Нормированное отклонение, t**	**Ординаты нормальной кривой, p**	**Теоретическая частота, A**		$\frac{(a-A)^2}{A}$
7–7.7	7.35	2	9	2.16	0.04	2	10	0.1
7.8–8.4	8.05	7		1.38	0.15	8		
8.5–9.1	8.75	18		0.60	0.33	16		0.25
9.2–9.8	9.45	22		0.18	0.39	19		0.47
9.9–10.5	10.15	10		0.96	0.25	12		0.33
10.6–11.2	10.85	1	4	1.74	0.09	4	5	0.2
11.3–11.9	11.55	3		2.52	0.02	1		
Σ		$n = \Sigma a = 63$				$n = \Sigma A = 63$		$\chi^2 = 1.36$

Теперь оцениваются отличия частот двух рядов по критерию хи-квадрат. Но перед этим необходимо убедиться в совпадении суммы эмпирических и теоретических частот (по 63 варианты) и в том, что минимальная частота в отдельных классах обоих рядов не ниже 4. Поскольку в крайних классах частоты были ниже, проводим их объединение (отмечено скобками), после чего число классов снизилось до $k = 5$. Далее вычисляем критерий хи-квадрат: для первого класса $(9–10)^2/10 = 0.1$. Значение критерия составило: $\chi^2 = 1.36$. Число степеней свободы (при трех ограничениях и пяти классах) равно: $df = 5–3 = 2$. Табличное значение (табл. 9*П*) $\chi^2_{(0.05,2)} = 5.99$.

Поскольку полученное значение (1.36) меньше табличного (5.99), нулевая гипотеза сохраняется, распределение бурозубок по массе тела достоверно от нормального не отличается.

Аналогичные расчеты для дискретного признака (плодовитость лисиц), имеющего предположительно **биномиальное распределение** (дискретный аналог нормального), представлены в табл. 6.7. Так, при параметрах $M = 5$ экз., $S = 1.33$ экз. для второго интервала получаем: $t = |8–5|/1.33 = 1.5$.

Таблица 6.7

Центр интервала, x_j	**Фактическая частота,** a	**Нормированное отклонение,** t	**Ординаты нормальной кривой,** p	**Теоретическая частота,** A	$\frac{(a-A)^2}{A}$
2	1 } 9	2.26	0.031	2 } 9	0
3	8 }	1.5	0.129	7 }	
4	16	0.75	0.301	17	0.05
5	23	0	0.399	23	0
6	21	0.75	0.301	17	0.94
7	3 } 6	1.5	0.129	7 } 9	1
8	3 }	2.26	0.031	2 }	
Сумма	$n = \Sigma a = 75$			$n = \Sigma A = 75$	$\chi^2 = 2$

Соответствующая ордината нормальной кривой равна: $p = 0.1295$ (графа 4), теоретическая частота составит:

$A = c \cdot p = 56.38 \cdot 0.1295 = 7.3 \approx 7$ (графа 5),

поскольку значение $c = 1 \cdot 75/1.33 = 56.38$. В результате вычислений получаем частоты (A) распределения (с параметрами $M = 5$, $S = 1.33$, $n = 75$), строго соответствующего биномиальному (см. рис. 3.4, с. 69). Объединим классы с частотами менее 4 и рассчитаем значение критерия $\chi^2 = 2$. Число степеней свободы (при трех ограничениях и пяти классах) равно: $df = 5–3 = 2$. Поскольку это значение ($\chi^2 = 2$) меньше критического табличного ($\chi^2_{(0.05,2)} = 5.99$), нулевая гипотеза не может быть отклонена, значит, распределение лисиц по плодовитости достоверно от биномиального не отличается.

В рассмотренных примерах проводилась проверка соответствия эмпирического распределения тому или иному типу распределения, заданному статистическим законом. На основании этого за-

кона и рассчитывались ожидаемые частости *p*. Однако метод χ^2 позволяет проверять гипотезы, диктуемые не только формальными статистическими законами, но и содержательными (биологическими) соображениями. Основанием для подобных гипотез могут быть биологические законы расщепления признаков в гибридных поколениях, представленность морф, соотношение разнополых и разновозрастных групп в популяции, соотношения видов в ценозах и пр. Таким случаям соответствуют признаки с альтернативным и полиномиальным распределением. Для расчета теоретически ожидаемых частостей *p* используются идея о полной группе событий (сумма частостей для всех возможных событий равна 1) и содержательные соображения.

Рассмотрим применение критерия хи-квадрат при анализе **альтернативной изменчивости**. В одном из опытов по изучению наследственности у томатов было обнаружено 3629 красных и 1176 желтых плодов. Теоретическое соотношение частот при расщеплении признаков во втором гибридном поколении должно быть 3:1 (75% к 25%, или в долях: $p_1 = 0.75$, $p_2 = 0.25$). Выполняется ли оно? Иными словами, взята ли данная выборка из той генеральной совокупности, в которой соотношение частот 3:1?

Для того чтобы это проверить, сформируем уже знакомую таблицу (табл. 6.8), заполнение которой аналогично рассмотренным, только для расчета теоретической частоты используется формула:

$$A = n \cdot p,$$

где p – теоретические частости;

n – объем выборки.

Например, $A_2 = n \cdot p_2 = 4805 \cdot 0.25 = 1201.25 \approx 1201$.

Таблица 6.8

Значение (цвет плода), x_j	**Фактическая частота, *a***	**Теоретическая частость, *p***	**Теоретическая частота, *A***	$\frac{(a-A)^2}{A}$
Красный	3629	0.75	3603	0.187621
Желтый	1176	0.25	1201	0.5204
Сумма	$n = \Sigma a = 4805$	1	$n = \Sigma A = 4805$	$\chi^2 = 0.71$

Далее вычисляем хи-квадрат: $\chi^2 = 0.71$ и число степеней свободы (при двух классах и одном ограничении, объеме выборки) $df = k–1 = 2–1 = 1$. По табл. 9*П* находим критическое значение $\chi^2_{(0.05,1)} = 3.84$. Поскольку полученная величина (0.71) меньше табличной (3.84), различия сравниваемых распределений статистически недостоверны. Иначе говоря, фактические частоты хорошо согласуются с теоретически ожидаемыми. По данным первой строки таблицы видно, что полученное значение χ^2 соответствует уровню значимости, большей $\alpha = 0.30$ (напомним, что порогом, как было установлено выше, является $\alpha = 0.05$). Значит, совпадение между фактическими результатами и ожидаемыми достаточно велико. Полученные данные не отвергают принятую гипотезу о том, что в нашем случае имеется отношение 3:1.

Здесь следует еще раз обратить внимание читателей на то обстоятельство, что сохранение нулевой гипотезы нельзя считать доказательством справедливости нулевой гипотезы. Результатами представленных вычислений теория о расщеплении по фенотипам в отношении 3:1 (0.75:0.25) *не доказана*, хотя и не опровергнута. Статистика доказывает только факт отличий, но не их отсутствие. Чтобы доказать теорию, нужно предположить антитеорию (для нашего примера соотношение 1:1) и опровергнуть ее с помощью статистических приемов.

В выборке рыжих полевок, отловленных в первый день учета численности, присутствуют 64 самца и 12 самок. Требуется определить, подтверждают ли эти данные факт преобладания самцов во всей популяции (как генеральной совокупности) или налицо просто случайное отличие значений в данной выборке. Теоретическое соотношение полов в популяции животных составляет 1:1 (или 38:38 экз.). Нарушается ли оно? Иными словами, выдвигается нулевая гипотеза, что данная выборка взята из генеральной совокупности, в которой соотношение полов 1:1.

Таблица 6.9

Пол, x_j	**Фактическая частота, *a***	**Теоретическая частость, *p***	**Теоретическая частота, *A***	$\frac{(a-A)^2}{A}$
Ж	64	0.5	38	17.78947
М	12	0.5	38	17.78947
Сумма	$n = \Sigma a = 76$	1	$n = \Sigma A = 76$	$\chi^2 = 35.57$

Сравнение вычисленного (35.6) и критического значений ($\chi^2_{(0.05,1)}$ = 3.84) явно свидетельствует о существенном отклонении фактического соотношения полов от гипотезы – 1:1. Вероятность правильности нулевой гипотезы (т. е. что в данном случае действительно имеет место численное равенство полов) оказалась много меньше 0.01. Соответственно, доверительная вероятность, т. е. вероятность несоответствия между числом самцов и самок очень велика и составляет более 0.99. Итак, есть все основания говорить о статистически достоверном преобладании самцов среди особей, отловленных в первый день. Из какой же генеральной совокупности они отбираются, если достоверно не из той, где ♀♀:♂♂= 1:1? Видимо, речь идет о группе особей, активно осваивающих территорию. Понятно, что наиболее активными оказались самцы, практически не привязанные, как самки, к гнезду с выводком.

Принципы исследования **полиномиальных распределений** остаются прежними, возрастает число классов и степеней свободы. Метод хи-квадрат позволяет сравнивать между собой не только теоретический и фактический ряды данных, но пару (и более) *эмпирических выборок*. Для ее решения эмпирические частоты каждого ряда сопоставляются со *средними теоретическими частотами*, рассчитанными на основе нулевой гипотезы «все выборки взяты из одной и той же генеральной совокупности», т. е. «все распределения одинаковы», или «доли вариант с данным значением в разных распределениях одинаковы». Этим методом можно сравнивать между собой признаки, имеющие любые типы распределения.

Фактические данные наблюдений группируются в таблицу (a), далее рассчитываются средние теоретические частости (p), затем теоретические частоты (A) и критерий χ^2.

Рассмотрим алгоритм на примере изучения фенетической структуры популяций красной полевки с разным уровнем численности зверьков. Получены частоты встречаемости пяти комплексов фенов от 1 до 5 (признаки: число перфораций черепа в разных областях). Например, первым комплексом фенов обладали 146 особей из первой популяции и 208 из второй (табл. 6.10). Выдвинуто предположение, что различия в частотах фенов случайны. В соответствии с этим допущением частости фенов каждого из пяти типов в двух сравниваемых популяциях должны быть равны.

Сначала определяем усредненные (теоретические) частости

(p_i) для всех фенетических комплексов, поделив суммы особей в группах (Σ_i) на объединенный объем выборок (N = 600): $p_i = \Sigma_i/N$. Так, для второй группы фенов: p_2 = 190/600 = 0.317.

Таблица 6.10

Группы фенов	a_1	A_1	a_2	A_2	Σ_i	p_i	$\frac{(a_1 - A_1)^2}{A_1}$	$\frac{(a_2 - A_2)^2}{A_2}$
1	146	*170*	208	*184*	354	0.59	3.39	3.13
2	112	*91.2*	78	*98.8*	190	0.317	4.74	4.38
3	8	*9.6*	10	*9.4*	18	0.03	0.27	0.04
4	21	*15.8*	12	*17.2*	33	0.055	1.71	1.57
5	1	*2.4*	4	*2.6*	5	0.008	0.82	0.75
$n = \Sigma a = \Sigma A$	288	*288*	312	*312*	600	1	10.9	9.87

Далее находим частоты всех фенов *с поправкой на разные объемы* выборок. Общая формула для вычисления усредненных частот, как и раньше, имеет вид:

$$A_{ji} = n_j \cdot p_i,$$

где A_{ji} – теоретическая частота для i-го значения j-й выборки,

p_i – теоретические (усредненные) частости,

n_j – объем выборки.

Усредненные (теоретические) частоты представлены в таблице 6.10 справа внизу от реальных значений частот. Например, ожидаемая частота второй группы во второй выборке составит $A_{2,2}$ = 0.317·312 = *98.8*, для пятой группы – $A_{2,5}$ = 0.008·312 = *2.6*.

Критерий хи-квадрат вычисляется по обычной формуле. При этом отыскиваются разности только между эмпирическими и теоретическими частотами для каждой выборки отдельно. Например, для первого класса первой выборки имеем:

$$\frac{(146-170)^2}{170} = 3.39.$$

В завершение значения χ^2, полученные для разных выборок, складываются. В нашем случае $\chi^2 = 10.9+9.87 = 20.8$.

Расчет числа степеней свободы производится по формуле $df = (k–1)\cdot(r–1)$, где k – число значений (классов) вариант, в данном случае 5 классов фенов, r – число сравниваемых выборок, в нашем примере их две. Отсюда $df = (5–1)(2–1) = 4$. Табличное значение $\chi^2_{(0.05,4)} = 16.92$. То, что фактическая величина (20.8) больше табличной, позволяет отвергнуть нулевую гипотезу и сделать вывод о том, что частота (распределение) фенов в сравниваемых популяциях достоверно отличается, причем в основном за счет встречаемости первых двух групп фенов. Отмеченные различия обусловлены увеличением фенетического (генетического) разнообразия первой популяции, отличавшейся более высокой численностью.

Критерий λ Колмогорова – Смирнова

Этот критерий, обозначаемый греческой буквой λ (ламбда), можно применять как для оценки расхождения между фактическими и теоретическими распределениями, так и для определения достоверности различий между любыми двумя распределениями частот одного и того же признака, причем даже при неодинаковом числе классов и частот у этих распределений. По своему назначению и возможностям он напоминает описанный выше критерий хи-квадрат, но более прост в применении. Нулевая гипотеза о случайности расхождения между сопоставляемыми распределениями отвергается и различия считаются достоверными, если эмпирическая величина критерия λ превосходит свое критическое значение для принятого порога доверительной вероятности, и наоборот, различия могут считаться случайными (нулевая гипотеза сохраняется), если эмпирический критерий не достигает требуемого значения квантили.

Для сравнения распределений при одинаковом числе классов и одинаковой общей численности групп критерий λ вычисляется по формуле:

$$\lambda = \frac{\max\left|\sum a_1 - \sum a_2\right|}{\sqrt{n}},$$

а при сравнении выборок разного объема:

$$\lambda = \max\left|\frac{\sum a_1}{n_1} - \frac{\sum a_2}{n_2}\right| \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

где max|| – максимальная разность (без учета ее знака) между накопленными частотами в сравниваемых распределениях,

a_1 и a_2 – частоты первого и второго рядов (это могут быть как две выборки, так и эмпирическое и теоретическое распределения);

n – общее число (сумма) всех вариант совокупности;

n_1 и n_2 – объемы сравниваемых выборок.

Критерий λ не требует специальной таблицы для оценки значимости отличий, так как для любого числа классов предельные значения критерия λ, соответствующие трем порогам доверительной вероятности (0.95, 0.99 и 0.999), одинаковы и равны соответственно 1.36, 1.63 и 1.95.

Применение критерия λ можно показать на таком примере. Сравнивается плодовитость зимовавших (a_1) и прибылых (a_2) рыжих полевок, у которых частота встреч выводков разной величины (число эмбрионов на самку, x) отличалась.

Таблица 6.11

x	1	2	3	4	**5**	6	7	8	9	10	11	n_i
a_1	0	0	4	25	53	68	36	17	2	0	5	210
a_2	1	1	6	44	97	89	57	26	6	5	0	332
Σa_1	0	0	4	29	82	150	186	203	205	205	210	
Σa_2	1	2	8	52	149	238	295	321	327	332	332	
$\Sigma a_1/n_1$	0.00	0.00	0.02	0.14	0.39	0.71	0.89	0.97	0.98	0.98	1.00	
$\Sigma a_2/n_2$	0.00	0.01	0.02	0.16	0.45	0.72	0.89	0.97	0.98	1.00	1.00	
Разность	0.003	0.006	0.005	0.019	**0.058**	0.003	0.002	0.000	0.009	0.024	0.000	

Требуется оценить достоверность расхождения между этими распределениями. Ход вычислений показан в таблице 6.11. Сначала получают *накопленные частоты* путем суммирования частот от первого класса до конца вариационного ряда (Σa), затем рассчиты-

вают относительные накопленные частоты ($\Sigma a/n$). После этой процедуры отыскивают максимальную разность (max) относительных частот в каком-либо классе.

В нашем случае максимальная разность между отношениями накопленных частот к объемам выборок составила:

max = |0.39–0.45| = 0.058 (5-й класс),

откуда по формуле для сравнения распределений разного объема находим величину критерия:

$$\lambda = 0.058 \cdot \sqrt{\frac{210 \cdot 332}{542}} = 0.67.$$

Поскольку найденная величина $\lambda = 0.67$ оказалась ниже критического значения даже для первого порога вероятности ($\lambda_{(0.05)} = 1.36$), то нулевая гипотеза не отвергается и, следовательно, расхождения между сопоставляемыми распределениями носят случайный характер. Таким образом, существование возрастных отличий в плодовитости полевок в данном случае остается недоказанным (полученными данными не подтверждается).

Отношения между статистиками *t, T, F* и χ^2

Рассмотренные выше разнообразные критерии используют четыре статистики, поведение которых в своей основе базируется на законе нормального распределения, модифицированном для разных целей. Как указывалось ранее, *нормальное* соответствие относительной частоты (p) значений случайной величины (t) задается уравнением: $p = (1/\sqrt{2\pi}) \cdot \exp(-t^2/2)$. Значение случайной величины хи-квадрат представляет собой сумму нескольких нормально распределенных случайных величин, возведенных в квадрат: $\chi^2 = \sum_{i=1}^{df} t_i^2$ (df – число степеней свободы). По таблицам 4*П* и 9*П* нетрудно убедиться, что для $df = 1$ $\chi^2 = t^2$ и границы критических областей для α=0.05 составляют $\chi^2 = t^2 = 1.96^2 = 3.84$.

Распределение T Стьюдента использует распределение нормальное и хи-квадрат: $T = t/\sqrt{\chi^2/df}$. Распределение F Фишера использует два распределения хи-квадрат с разным числом степеней свободы: $F = (\chi_1^2/df_1)/(\chi_2^2/df_2)$.

7
ЗАДАЧА «ДОКАЗАТЬ ОТЛИЧИЕ НЕСКОЛЬКИХ ВЫБОРОК» («ДОКАЗАТЬ ВЛИЯНИЕ ФАКТОРА»)

При изучении и анализе сложных и многообразных причинно-следственных отношений между объектами и явлениями биологу приходится учитывать целый комплекс внешних и внутренних факторов, от которых в конечном итоге зависят уровень и ход наблюдаемых процессов, те или иные биологические свойства живых организмов, их динамика и разнообразие. При этом зачастую важно оценивать не только роль одного из многочисленных внешних факторов, но и их взаимодействие при констелляционном влиянии на популяцию или организм.

Идейная база для изучения действия факторов содержится уже в методе сравнения двух выборок. Биологическим содержанием операции сравнения двух выборок, в конце концов, выступает поиск факторов, ответственных за смещение средних арифметических или усиление изменчивости признаков. Развивая это направление биометрического исследования, можно не ограничиваться только двумя «дозами» фактора, но изучить серию ситуаций, в которых фактор проявлял разную силу действия на результативный признак – от самого слабого, до самого сильного. При этом каждому уровню фактора будет соответствовать отдельная выборка и общая задача получит формулировку «сравнить несколько выборок». В терминах факториальной биометрии вопрос о влиянии фактора на признак звучит так: сказывается ли отличие условий получения разных выборок на качестве (значениях) вариант? В терминах статистики вопрос звучит несколько иначе: из одной ли генеральной совокупности отобраны все выборки, оценивают ли выборочные средние арифметические одну и ту же генеральную среднюю? Вариантов ответа может быть только два:

1. Все выборки отобраны из одной генеральной совокупности, условия возникновения вариант одни и те же.
2. Выборки отобраны из разных генеральных совокупностей, условия возникновения вариант выборок различаются.

В постановке вопроса можно уловить противоречие. Выше было сказано, что по условию задачи выборки формировались в разных условиях, и тут же предполагается, что условия были одинаковые. На самом деле противоречия нет, поскольку речь идет об определении чувствительности признака к действию фактора. Условия формирования выборок могут отличаться, но они могут никак и не сказаться на величине изучаемого признака, не отразиться на значениях вариант. Смысл статистического сравнения в том и состоит, чтобы оценить эффективность действия фактора на признак, доказать реальность реакции вариант выборок на разные условия их формирования. Круг методов сравнения нескольких выборок довольно широк, их выбор зависит от конкретной задачи (табл. 7.1).

Таблица 7.1

Задача	**Содержание задачи**	**Методы**
Доказать различие нескольких средних (для *одного* признака)	Отличаются доминирующие факторы, формирующие выборки	Однофакторный дисперсионный анализ
Доказать различие нескольких средних (для *нескольких* признаков)	Отличаются доминирующие факторы, формирующие выборки	Двух- и многофакторный дисперсионный анализ
Доказать различие нескольких пар средних в контексте сравнения *нескольких* выборок	Отличаются доминирующие факторы, формирующие две сравниваемые выборки	Метод парных сравнений Шеффе
Доказать различие нескольких дисперсий (для *одного* признака)	Отличаются случайные факторы, формирующие выборки	Метод Бартлетта
Доказать различие нескольких частотных распределений (для *одного* признака)	Факторы, участвующие в формировании выборки, отличаются в целом	Критерий χ^2 Пирсона
Доказать различие нескольких выборок в целом (для *одного* признака)	Факторы, участвующие в формировании выборки, отличаются в целом	Непараметрический дисперсионный анализ

Сравнение нескольких выборок по величине одного признака (однофакторный дисперсионный анализ)

Дисперсионный анализ позволяет оценить достоверность отличия нескольких выборочных средних одновременно, т. е. изучить влияние одного контролируемого фактора на результативный признак путем оценки его относительной роли в общей изменчивости этого признака, вызванной влиянием всех факторов.

Логико-теоретические основы

Задача дисперсионного анализа состоит в том, чтобы охарактеризовать силу и достоверность влияния фактора на признак, причем только на *величину* (средний уровень) признака, но не на его изменчивость. Дисперсионный анализ есть метод сравнения нескольких средних арифметических. В этом смысле он подобен методу сравнения двух средних арифметических с помощью критерия Стьюдента:

$$T = \frac{\text{обобщенный показатель отличия средних}}{\text{обобщенный показатель случайного варьирования}},$$

$T = (M_1 - M_2)/ m_d$, или $T = dM/ m_d$,

где M_1 , M_2 – две выборочные средние,

dM – обобщенный показатель отличия выборочных средних,

m_d – обобщенная ошибка репрезентативности $m_d = \sqrt{m_1^2 + m_2^2}$.

Критерий сравнивает две средние арифметические двух выборок, полученных при разных условиях, при действии двух доз некоего фактора. В числителе этой формулы стоит оценка действия возможного доминирующего фактора, а в знаменателе стоит оценка действия случайных факторов варьирования выборочных значений. Если изучаемый фактор сказывается на значении вариант, то оценка его действия (dM) превысит оценку действия случайных факторов (m_d), хотя бы в 2 раза (критическое значение критерия Стьюдента для репрезентативных выборок $T_{(0.05,30)} \approx 2$). В этом случае говорят о достоверном отличии средних арифметических, о достоверном влиянии на варианты различных условий их формирования.

В дисперсионном анализе использован такой же показатель достоверности влияния фактора, но адаптированный к случаю сравнения нескольких выборок (критерий Фишера):

$F = S^2_{факт.} / S^2_{случ.}$.

В качестве обобщенной *меры отличия нескольких выборочных средних* выступает дисперсия, рассеяние выборочных средних (M_j) вокруг общей средней ($M_{общ.}$):

$$S^2_{факт.} = \sum^{k} (M_j - M_{общ.})^2 / df_{факт.},$$

где $df_{факт.} = k–1$,

$j = 1, 2, \ldots k$,

k – число сравниваемых средних.

В качестве обобщенной *меры случайного варьирования* служит дисперсия вариант (x_i) вокруг средней в каждой градации (M_j):

$$S^2_{случ.} = \sum^{k} \sum^{n_j} (x_{ij} - M_j)^2 / df_{случ.},$$

где $df_{случ.} = n–1$,

$i = 1, 2, \ldots n$, n – число вариант всех выборок.

В этом отношении критерий Фишера, используемый для сравнения нескольких средних арифметических, подобен критерию Стьюдента, служащему для сравнения двух средних:

$$T = \frac{M_1 - M_2}{m_d} \Rightarrow \frac{\text{изменчивость за счет систематических причин}}{\text{изменчивость за счет случайных причин}} \Rightarrow F = \frac{S^2_{факт.}}{S^2_{случ.}}$$

Применяя дисперсионный анализ, это обстоятельство важно всегда иметь в виду: несмотря на то что критерий Фишера использует дисперсии, тем не менее сравниваются друг с другом выборочные средние арифметические!

Техника расчетов

В основе однофакторного дисперсионного анализа (дословно – разложение дисперсий) лежит модель варианты (x_i), которая выражает ее отклонение от общей средней (M) за счет действия кон-

тролируемого фактора ($x_{факт.}$) и действия случайных причин ($x_{случ.}$):

$$x_i = M \pm x_{факт.} \pm x_{случ.}$$

Иными словами, отклонение варианты от общей средней связано с отклонением за счет действия изучаемого фактора и за счет действия прочих неучтенных факторов.

Каждой дозе изучаемого фактора соответствует одна выборка (градация). Поэтому каждая групповая (выборочная) средняя будет характеризовать реакцию объектов на соответствующую дозу изучаемого фактора и эффект изучаемого фактора можно выразить как отклонение групповой средней от общей средней:

$$x_{факт.} = M_j - M.$$

В свою очередь, от групповой средней каждая варианта будет отличаться в силу случайных неучтенных причин, эффект действия случайных факторов можно выразить как отклонение отдельной варианты от данной групповой средней:

$$x_{случ.} = x_i - M_j.$$

Получается, что отклонение варианты от общей средней будет равно отклонению групповой средней от общей средней (эффект учтенного фактора) и отклонению варианты от своей групповой средней (эффект неучтенных факторов). Отсюда

$$(x_i - M) = (M_j - M) + (x_i - M_j).$$

Обобщая эту запись для всех вариант выборки (возведя в квадрат и суммировав), получаем правило разложения общей вариации признака на составные части, отражающие влияние всех названных причин:

$$C_{общ.} = C_{факт.} + C_{случ.}$$

Общая сумма квадратов признака рассчитывается как сумма квадратов отклонений всех вариант (x_i) от общей средней (M):

$$C_{общ.} = \Sigma (x_i - M)^2.$$

Факториальная сумма квадратов рассчитывается как сумма квадратов отклонений частных средних (M_i) для каждой выборки (всего k выборок) от общей средней:

$$C_{факт.} = \Sigma (M_j - M)^2.$$

Остаточная (случайная) сумма квадратов есть сумма квадратов отклонений вариант каждой выборки (x_i) от своей средней (M_j):

$$C_{случ.} = \Sigma (x_i - M_j)^2.$$

Параметры дисперсионного анализа и порядок их вычислений представлены в таблице 7.2.

Отношение сумм квадратов (*SS, sum of squares*) к соответствующему числу степеней свободы дает оценку величины дисперсии, или средний квадрат (*MS, mean square*), иногда ее именуют варианса. Влияние изучаемого фактора отражает факториальная, или межгрупповая, дисперсия $S^2_{факт.}$, а влияние случайных неорганизованных в данном исследовании причин – случайная, или внутригрупповая, остаточная дисперсия $S^2_{случ.}$, или $S^2_{остат}$.

Таблица 7.2

Составляющие дисперсии	**Суммы квадратов (*SS*), *C***	**Сила влияния, η^2**	**Степени свободы, *df***	**Дисперсии (средний квадрат, *MS*), S^2**	**Критерий влияния, *F***
Факториальная	$C_{факт.} = \Sigma (M_j - M)^2$	$\frac{C_{факт.}}{C_{общ.}}$	$k-1$	$S^2_{факт.} = \frac{C_{факт.}}{df_{факт.}}$	$F = \frac{S^2_{факт.}}{S^2_{случ.}}$
Случайная	$C_{случ.} = \Sigma (x_i - M_j)^2$		$n-k$	$S^2_{случ.} = \frac{C_{случ.}}{df_{случ.}}$	
Общая дисперсия	$C_{общ.} = \Sigma (x_i - M)^2$				

Сила влияния фактора определяется как доля частной суммы квадратов в общем варьировании признака. Показатель силы влияния изучаемого фактора составляет: $\eta^2_{факт.} = C_{факт.} / C_{общ.}$, неорганизованных (случайных): $\eta^2_{случ.} = C_{случ.} / C_{общ.}$; сумма этих показателей, естественно, равна единице: $\eta^2_{факт.} + \eta^2_{случ.} = 1$.

В то же время нам кажется, что придавать большое значение этому индексу не стоит. Во-первых, в литературе показано, что он дает не точную характеристику вклада фактора в общую изменчивость и для него приходится рассчитывать некую поправку. Во-вторых, утверждение вроде «фактор влияет с силой 20%» ничего не передает, кроме впечатления о не очень большом влиянии фактора. Гораздо интереснее было бы дать прогноз возможных значений результативного признака при том или ином уровне действия фак-

тора, а это можно сделать только с помощью регрессионного анализа или имитационного моделирования. По этим причинам мы рекомендуем рассматривать показатель $\eta_{\text{факт.}}$ как простую и удобную характеристику влияния фактора на признак, подталкивающую исследователя к решению о необходимости продолжения биометрического исследования в рамках регрессионного анализа. Чем большую долю в общей дисперсии занимает ее факториальная часть, тем большая часть общего разнообразия обусловлена варьированием за счет действия фактора.

Нулевая гипотеза гласит: «влияние фактора на признак отсутствует». Проверяют гипотезу по критерию Фишера:

$$F = S^2_{факт.} / S^2_{случ.} \geq F_{(a,df_1,df_2)},$$

где $df_1 = k-1$, $df_2 = n-k$,

k – число градаций результативного признака,

n – общий объем всех выборок по всем градациям.

Влияние считается достоверным, если величина расчетного критерия равна или превышает свое табличное значение с принятым уровнем значимости (обычно $\alpha = 0.05$) (F определяется по табл. 7*П*).

Дисперсионный анализ для количественных признаков

Однофакторным называется анализ, изучающий действие на результативный признак только одного организованного фактора *A*. Для примера оценим влияния растворенного в воде вещества на плодовитость дафний, используемых в качестве тест-объектов в водно-токсикологических экспериментах. В ходе предварительного исследования были получены четыре выборки, четыре группы значений плодовитости животных, выращенных в средах с разным содержанием химической добавки.

Сначала необходимо сгруппировать выборочный материал в комбинативную таблицу (организовать дисперсионный комплекс). Для этого варианты каждой выборки записываются в отдельные графы, именуемые градациями (табл. 7.3). Результативным признаком служит средняя плодовитость дафний за неделю (для иллюстративности расчетов она дана в целых числах). В нашем примере организованы 4 градации – чистая вода (контроль, градация *A*1; значения плодовитости 6, 5, 5, 7), слабая концентрация вещества (5 мг/л,

*A*2; 8, 7, 6, 6), средняя (15 мг/л, *A*3; 8, 8, 7) и сильная (30 мг/л, *A*4; 8, 7, 9). Предлагаемый ниже алгоритм расчетов позволяет использовать неравное число вариант в градациях. Расчеты несложны и показаны в таблице 7.3.

Таблица 7.3

	Градации фактора									
	*A*1		*A*2		*A*3		*A*4			
	x	x^2	x	x^2	x	x^2	x	x^2		
	6 5 5 7	36 25 25 49	8 7 6 6	64 49 36 36	8 8 7	64 64 49	8 7 9	64 49 81		
									Σ	
Σx^2		135		185		177		194	691	$H1 = \Sigma\Sigma x^2 = 691$
Σx	23		27		23		24		97	$H2 = (\Sigma\Sigma x)^2/n =$
n	4		4		3		3		14	$= (97)^2/14 = 672$
$\Sigma x^2/n$	132		182		176.3		192		682.8	$H3 = \Sigma\Sigma x^2/n =$ $= 682.8$
M	5.8		6.8		7.67		8		6.93	

$$C_{факт.} = H3 - H2 = 682.8 - 672 = 10.76$$
$$C_{случ.} = H1 - H2 = 691 - 672 = 8.17$$
$$C_{общ.} = H1 - H3 = 691 - 682.8 = 18.93$$

Полученные значения позволяют вычислить дисперсии, определить силу влияния фактора и критерий достоверности Фишера.

Составляющие дисперсии	**Суммы квадратов, *C***	**Сила влияния, *η***	**Степени свободы, *df***	**Дисперсии, S^2**	**Критерий, *F***
Факториальная	10.76	57%	3	3.59	4.39
Случайная	8.17		10	0.82	
Общая	18.93			4.39	

Поскольку полученное значение критерия (F = 4.39) больше табличного ($F_{(0.05,3,10)}$ = 3.7) (табл. 7*П*), отличие факториальной и случайной дисперсий достоверно, влияние фактора значимо.

Отсюда следует биологический вывод: стимулирующее влияние изучаемого фактора (вещества) на плодовитость дафний относительно велико (57%) и достоверно (с вероятностью P>0.95).

Выполнить дисперсионный анализ по представленному алгоритму можно и в среде Excel. Для этого введем подписанные метками (*A*1, *A*2…) данные в четыре столбца, отдельно для каждой градации.

	A	B	C	D
1	A1	A2	A3	A4
2	6	8	8	8
3	5	7	8	7
4	5	6	7	9
5	7	6		

Вызовем программу обработки командой Сервис \ Анализ данных…\ Однофакторный дисперсионный анализ, ОК. Заполним окно макроса, выделив блок данных с метками и поставив галочку в поле «Метки в первой строке», ОК. На новом листе (рис. 7.1) появятся результаты расчетов, идентичные приведенным в табл. 7.3. Чтобы все надписи были видны, нужно изменить ширину столбцов. Это можно сделать, нажав на серый квадрат слева вверху листа (над 1, левее А), перевести курсор на границу между любыми столбцами (курсор примет форму креста со стрелками, направленными в стороны) и дважды кликнуть левой кнопкой мыши. Ширина каждого столбца будет автоматически определена по максимально длинному содержимому какой-либо ячейки этого столбца.

С помощью макроса Однофакторный дисперсионный анализ в рамках пакета Excel можно обрабатывать выборки самого разного размера, в том числе очень большого, поэтому мы не приводим специальных алгоритмов для ручного обсчета больших выборок.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	A1	4	23	5.75	0.916667		
6	A2	4	27	6.75	0.916667		
7	A3	3	23	7.666667	0.333333		
8	A4	3	24	8	1		
9							
10							
11	Дисперсионный анализ						
12	*рник вари*	*SS*	*df*	*MS*	*F*	*P-Значение*	*критическое*
13	Между гру	10.7619	3	3.587302	4.392614	0.032353	3.708266
14	Внутри гр	8.166667	10	0.816667			
15							
16	Итого	18.92857	13				

Рис. 7.1. *Дисперсионный анализ в среде Excel*

Парные сравнения выборочных средних методом Шеффе

Дисперсионный анализ позволяет установить достоверность отличия нескольких средних арифметических друг от друга, но он не сообщает, какие именно средние от каких именно средних отличаются. Может статься, например, что действие фактора вызывает более или менее плавное изменение средних без заметных переломов в этой тенденции. При этом биологический вопрос может состоять в том, чтобы определить минимальную дозу фактора, которая по сравнению с контролем значимо влияет, изменяет среднюю для этой градации, т. е. определить «первую действующую концентрацию». Казалось бы, этот вопрос относится к задаче сравнения двух выборок: контрольная выборка поочередно сравнивается с выборками, полученными для все возрастающих доз фактора, а первое достоверное отличие средних как раз и означает, что данная доза уже «действующая». Однако с точки зрения статистики такое сравнение оказывается некорректным и неточным.

При таком «лобовом» попарном сравнении выборок одна из них (для градации контроля) все время участвует в этой процедуре, на основании которой формулируются разные статистические выводы о достоверности отличий средних с той или иной вероятностью. Тем самым эти выводы оказываются зависимыми друг от друга. Доверительная вероятность каждого из этих выводов поэтому от сравнения к сравнению уменьшается! Чем больше выводов сделано на одном и том же материале, тем меньше вероятность их справедливости.

Второе негативное обстоятельство связано с тем, что такая процедура учитывает далеко не всю информацию о явлении. Действительно, изменчивость вариант комплекса выборок (в нашем примере было 14 вариант в 4 выборках) определяется как действием изучаемого фактора, так и множеством других не учитываемых, случайных, причин. При сравнении же всего двух выборок (например, выборок 1 и 4) эта случайная изменчивость представлена не всем объемом информации, но только той частью, что проявилась в рамках этих двух сравниваемых выборок (две выборки содержат лишь 7 вариант). Поэтому оценки случайной изменчивости для двух выборок оказываются не столь точными, как могли бы быть по всем градациям.

Улучшить ситуацию позволяет метод попарного сравнения выборок, проводимый на базе однофакторного дисперсионного анализа (метод Шеффе). Для сравнения двух средних предлагается критерий F Фишера, в числителе которого стоит оценка действия фактора (разность средних) для любых двух сравниваемых градаций, а в знаменателе – оценка случайной изменчивости, общая для всего дисперсионного комплекса:

$$F = \frac{(M_i - M_j)^2}{(k-1)\cdot S^2_{случ.}\cdot\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} \sim F_{(\alpha,df1,df2)},$$

где M – средние арифметические для любых двух (i, j) градаций однофакторного дисперсионного комплекса,

$S^2_{случ.}$ – оценка случайной изменчивости из таблицы дисперсионного анализа,

k – число градаций фактора,

n_i, n_j – объемы выборок сравниваемых градаций,

α – принятый уровень значимости (обычно $\alpha = 0.05$),

df – число степеней свободы $df_1 = k–1$, $df_2 = (k–1)\cdot(n–1)$.

Отличия средних считаются достоверными, если расчетное значение критерия Фишера превысит табличное $F_{(\alpha,df1,df2)}$ (табл. 7*П*).

Сопоставляя выборочные средние для первой и четвертой градаций нашего примера (табл. 7.3), имеем:

$F_{1,4} = (5.8–8)^2/[(4–1)\cdot 0.82\cdot(1/4+1/3)] = 3.37$,

$df_1 = 4–1 = 3$; $df_2 = (4–1)\cdot(14–1) = 39$,

$F_{(0.05,3,39)} = 2.87$.

Полученное значение (3.37) больше табличного (2.87), следовательно, между средними арифметическими первой и последней градаций есть достоверное отличие; разные дозы фактора действительно вызывают изменение плодовитости дафний.

Сравнение выборок первой и второй градаций показывает, что низкие дозы фактора в них не позволяют говорить о существенном влиянии на дафний: для данных объемов выборок полученное значение критерия (0.69) меньше табличного (2.87).

$F = (5.8–6.8)^2/[(4–1)\cdot 0.82\cdot(1/4+1/3)] = 0.69 < 2.87$.

Непараметрический однофакторный дисперсионный анализ

Рассмотренные выше схемы дисперсионного анализа исходили из предположения о нормальном распределении изучаемого результативного признака. Когда для какого-либо признака нет уверенности, что выполняется предположение о нормальном распределении изучаемого признака, когда требуется провести анализ быстро и без особой точности, когда мало данных или они выражены качественными признаками, можно использовать схему непараметрического дисперсионного анализа. Этот метод более неприхотлив, но менее точен, нежели параметрический анализ. Он исследует распределения вариант в нескольких выборках. Нулевая гипотеза состоит в том, что распределения одинаковы, т. е. выборки взяты из одной генеральной совокупности.

Порядок вычислений состоит в том, что все варианты ранжируются в порядке возрастания. Затем суммируются ранги вариант по каждой выборке отдельно и рассчитывается критерий:

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

где n – число всех вариант,
n_j – объем j-й градации фактора,
R_j – сумма рангов для каждой j-й градации фактора,
k – число градаций фактора (j = 1, 2, … k).

При объеме выборок больше 5 вариант статистика H имеет распределение хи-квадрат с $df = k-1$ степенями свободы и сравнивается со значениями из табл. 9*П*.

Применим эту схему (табл. 7.4) к нашим данным из табл. 7.3, расположив их в строку.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
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

Затем упорядочим и ранжируем. Для нескольких одинаковых значений берется средний ранг.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14
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

Наконец, разнесем ранги по градациям и подсчитаем необходимые суммы.

Таблица 7.4

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
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				14.5				27			30.5			33
n				4				4			3			3
R^2/n				52.56				182.3			310.1			363

Объем всей выборки равен: $n = 14$. Величина критерия H составит:

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

$$= 0.065934 \cdot 907.8958 - 45 = 14.86 .$$

По таблице хи-квадрат для $\alpha = 0.05$ и $df = 4–1 = 3$ находим: $\chi^2_{(0.05,3)} = 7.81$. Полученное значение критерия (14.86) больше табличного (7.81), значит, отличие выборочных распределений достоверно. Химическая добавка действительно изменяет плодовитость дафний.

Сравнение нескольких выборок по изменчивости признака

Одна из задач сравнения двух выборок состояла в том, чтобы оценить однородность варьирования значений в их пределах, т. е. чтобы сопоставить множества случайных причин, действовавших при формировании выборок. Для двух выборок задача решалась с помощью метода сравнения двух дисперсий по критерию Фишера. В случае нескольких выборок используется критерий Бартлетта. С его помощью проверяется нулевая гипотеза о равенстве нескольких дисперсий по всем градациям дисперсионного комплекса (Но: $S_1^2 = ... = S_j^2 = … = S_k^2$) – «фактор, действующий на разные выборки, не вызывает изменения характера варьирования».

Существенным ограничением для использования этого критерия является требование соответствия сравниваемых распределений нормальному закону. В другом случае критерий будет фиксировать не отличие дисперсий, но отличие типов распределений. Это значит, что уверенность в «нормальности» распределения должна быть условием выполнения процедуры, рассмотренной ниже.

Метод основан на том известном явлении, что выборочные дисперсии несколько отличаются от генеральной (в силу ошибки репрезентативности), а с ростом объема выборки ошибка репрезентативности уменьшается. Это значит, что для принятой нулевой гипотезы каждая из выборочных дисперсий (S_j^2) может отличаться от общей дисперсии (S^2), рассчитанной по всей совокупности, только случайно. Показано, что сумма отличий выборочных дисперсий от общей есть случайная величина примерно с χ^2-распределением:

$$\chi^2 \approx \sum_{j=1}^{k} \frac{S_j^2}{S^2}.$$

Стабилизировать поведение данной случайной величины позволяют поправки, применение которых дает критерий Бартлетта:

$$\chi^2 = \frac{B}{C} \sim \chi^2_{(\alpha,k-1)},$$

$$B = [\sum (n_j - 1)] \cdot \ln S^2 - \sum (n_j - 1) \cdot \ln S_j^2,$$

$$C = 1 + \frac{1}{3 \cdot (k-1)} \cdot \left[\sum \frac{1}{n_j - 1} - \frac{1}{\sum (n_j - 1)} \right],$$

$$S_j^2 = \frac{\sum (x - M_j)^2}{n_j - 1},$$

$$S^2 = \frac{\sum (x - M)^2}{N - k} = \frac{\sum (n_j - 1) \cdot S_j^2}{N - k},$$

где k – число градаций (сравниваемых выборок),

n_j – объем j-й градации (j = 1, 2,…, k),

S^2 – дисперсия для всей совокупности данных (общая средняя сумма квадратов),

S_j^2 – дисперсии для каждой j-й градации (средняя сумма квадратов по каждой градации),

$\sum = \sum_{j=1}^{k}$ – операция суммирования по всем k градациям.

Рассмотрим пример использования критерия Бартлетта для изучения изменчивости длины тела дафний (данные Н. М. Калинкиной). Животных в течение месяца содержали в пяти разных концентрациях лигнина, основного компонента сточных вод предприятий целлюлозно-бумажной промышленности. К концу опыта размеры рачков (M, мм) в высоких концентрациях стали выше, чем в контроле. Возникает вопрос, не сказалась ли жизнь в загрязненной

среде на изменчивости (*S*, мм) размеров тела дафний? Предварительные расчеты приведены в табл.7.5.

Таблица 7.5

Концентрация лигнина, мг/л	***M***	***n***	**1/(*n*–1)**	S_j	$(n_j-1)\cdot S_j^2$	$(n_j-1)\cdot \ln S_j^2$
0	4.05	8	0.143	0.057	0.0227	–40.106
1	4.08	10	0.111	0.158	0.2247	–33.213
50	4.45	10	0.111	0.126	0.1429	–37.286
100	4.36	10	0.111	0.190	0.3249	–29.893
150	4.40	10	0.111	0.158	0.2247	–33.213
Всего		48	0.587	0.689	0.93988	–173.711

Искомые величины составят:

$$S^2 = \frac{\sum (n_j - 1)\cdot S_j^2}{N - k} = \frac{0.93988}{48 - 5} = 0.02185,$$

$$B = [\sum (n_j - 1)]\cdot \ln S^2 - \sum (n_j - 1)\cdot \ln S_j^2 =$$

$$= 43\cdot \ln 0.02185 - (-173.711) = -164.416 + 173.711 = 9.29816,$$

$$C = 1 + \frac{1}{3\cdot (k-1)}\cdot \left[\sum \frac{1}{n_j - 1} - \frac{1}{\sum (n_j - 1)}\right] =$$

$$= 1 + \frac{1}{3\cdot (5-1)}\cdot \left[0.587 - \frac{1}{43}\right] = 1.04698,$$

$$\chi^2 = \frac{B}{C} = \frac{9.29816}{1.04698} = 8.88.$$

Полученная величина (8.88) не превышает табличное значение критерия $\chi^2_{(0.05,4)} = 9.49$ (табл. 9*П*), следовательно, отличия дисперсий друг от друга недостоверны. Пока не удалось доказать влияние токсиканта на изменчивость длины тела дафний.

Сравнение нескольких выборок по величине двух признаков (двухфакторный дисперсионный анализ)

Двухфакторный дисперсионный анализ исследует влияние на результативный признак двух факторов как порознь, так и совместно. Учет эффекта влияния каждого фактора по отдельности теоретически ничем не отличается от описанных выше схем. И там и тут оценивается изменчивость средних по градациям на фоне случайной изменчивости вариант внутри градаций, с помощью критерия Фишера устанавливается достоверность отличий межгрупповых дисперсий от внутригрупповых.

Важным преимуществом двухфакторного дисперсионного анализа перед однофакторным служит то, что с его помощью в рамках факториальной изменчивости удается определить варьирование по сочетанию градаций $C_{сочет.}$, позволяющее получить новый и весьма ценный в биологическом отношении показатель – оценку влияния сочетанного действия (взаимодействия) факторов.

Логико-теоретические основы

Модель двухфакторного дисперсионного анализа становится сложнее и выражает отклонение варианты (x_i) от общей средней (M) за счет действия двух контролируемых факторов порознь ($x_{фактA.}$, $x_{фактB.}$) и совместно ($x_{сочетAB.}$), а также за счет действия случайных причин ($x_{случ.}$):

$$x_i = M \pm x_{фактA.} \pm x_{фактB.} \pm x_{сочетAB.} \pm x_{случ.}.$$

Правило разложения вариаций предстает как:

$$C_{общ.} = C_A + C_B + C_{AB} + C_{случ.},$$

$$C_{факт.} = C_{общ.} - C_{случ.} = C_A + C_B + C_{AB},$$

где $C_{общ.} = \Sigma(x_i - M)^2$,

$C_{A.} = \Sigma(M_{Aj} - M)^2$, j – число градаций фактора A,

$C_B = \Sigma(M_{Bk} - M)^2$, k – число градаций фактора B,

$C_{случ.} = \Sigma(x_i - M_{xi})^2$,

$C_{AB} = C_{общ.} - (C_A + C_B + C_{случ.})$.

Сочетанное действие (взаимодействие) факторов означает, что каждый из них помимо прямого воздействия на объект исследо-

вания сказывается и на характере влияния на объект и другого фактора, усиливает или ослабляет его. К примеру, неурожай кормов усугубляет негативное действие зимнего холода на численность популяций мелких млекопитающих.

Выделяют три основных вида взаимодействия факторов:

– аддитивное, когда взаимодействия факторов нет, их эффекты просто складываются,

– антагонизм, когда один фактор ослабляет действие другого, и наоборот,

– синергизм, когда наблюдается усиление действия обоих факторов.

Эти эффекты часто встречаются в практике токсикологических исследований. Рассмотрим гипотетические примеры действия на подопытных животных двух веществ, взятых в разных концентрациях. По осям OX и OY диаграммы отложены концентрации этих веществ в диапазоне от 0 до CL_{50}, которые за время опыта вызывают гибель 50% особей (рис. 7.2). Первая иллюстрация показывает точки на осях, в которых концентрации вещества $[A] = 0$ и $[B] = CL_{50} = {}_{B}CL_{50}$, а наблюдаемая гибель составляет 50% особей, то же наблюдается для вещества $[A] = CL_{50} = {}_{A}CL_{50}$ и $[B] = 0$.

Аддитивное действие – простое сложение влияний. Прямая, соединяющая точки $[A] = CL_{50} = {}_{A}CL_{50}$, $[B] = CL_{50} = {}_{B}CL_{50}$, есть множество опытов, в которых токсиканты ведут себя по отношению друг к другу как одно и то же вещество, поскольку эффекты от их доз просто суммируются. Так, половинные по эффекту дозы ${}_{A}CL_{50}/2$ и ${}_{B}CL_{50}/2$ в сумме дают одну «полноценную» совместную CL_{50}. Однако важно отметить, что пропорциональность *эффектов* разных веществ вовсе не означает пропорциональности их *концентраций*.

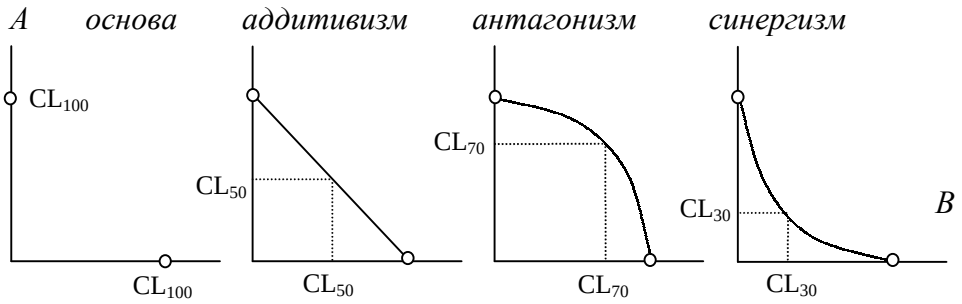


Рис. 7.2. *Виды взаимодействия веществ*

Антагонизм – подавление вредного действия одного вещества другим. Любая точка на выгнутой кривой свидетельствует о том, что для достижения эффекта CL_{50} требуется взять дозы, которые в сумме должны бы превышать эффект CL_{50}. Например, эффект CL_{50} в точке 1 достигается суммой $0.7 \cdot {}_{A}CL_{50} + 0.7 \cdot {}_{B}CL_{50}$. Чисто арифметически (аддитивно) эффект должен был составить $1.4 \cdot CL_{50}$, т. е. 70% гибели тест-объектов.

Синергизм – усиление действия. Точки на вогнутой кривой соответствуют ситуации, когда для достижения эффекта CL_{50} можно взять дозы, суммы которых аддитивно меньше CL_{50}. Так, эффект CL_{50} обнаруживается в точке для суммы $0.4 \cdot {}_{A}CL_{50} + 0.4 \cdot {}_{B}CL_{50}$. Аддитивный эффект должен был составить $0.8 \cdot CL_{50}$, т. е. 40% гибели организмов, но синергизм обеспечивает гибель 50% особей.

Сочетанное действие факторов нельзя смешивать с корреляцией факторов. Взаимодействие осуществляется «внутри» объекта исследования и связано со спецификой реакции биосистемы, а корреляция реализуется «снаружи» и связана как с природой фактора, так и со способом организации наблюдений. Чтобы выявить эффект именно взаимодействия, совместного воздействия, изучаемые факторы должны быть, по возможности, независимы друг от друга.

Кроме этого, имеется ряд условий правильного применения данного метода. Так, дисперсионному комплексу необходима полнота, т. е. второй фактор (B) должен быть представлен в каждой градации первого фактора (A) одинаковым числом градаций.

Ниже рассмотрены алгоритмы, относящиеся лишь к равномерным комплексам, характеризующимся равной численностью групп (в градациях содержатся одинаковое число вариант). Что же касается неравномерных многофакторных комплексов, то их анализ принципиально возможен, но имеет свои особенности, существенно усложняющие технику вычислений.

Если исходные данные представлены по градациям неравномерно, вполне допустимо искусственное превращение их в равномерные комплексы. Для этого нужно составить выборки одинаковой величины, используя часть имеющихся данных. Следует помнить, что такой отбор должен быть не субъективным, но случайным. При организации случайного отбора вариант лучше всего прибегнуть к жеребьевке. Например, убирать из выборки те варианты, номера которых совпадают со значениями случайных чисел (табл. 3*П*). От-

бросив часть вариант, мы лишаемся и части информации о варьировании признаков; избежать неправильных выводов, вызванных методикой формирования выборок, помогает многократный пересчет по схеме дисперсионного анализа с использованием результатов нескольких жеребьевок. Ограниченные рамки настоящего краткого руководства не позволяют остановиться на этом вопросе более подробно, поэтому мы отсылаем заинтересованного читателя к специальным пособиям, где техника дисперсионного анализа неравномерных многофакторных комплексов изложена с исчерпывающей полнотой.

Условием эффективности многофакторного анализа является также выбор схемы организации факторов в градации. Выше был рассмотрен дисперсионный анализ массива данных с повторностями в каждой градации, для которого разложение суммы квадратов соответствует выражению $C_{общ.} = C_A + C_B + C_{AB} + C_{случ.}$ (табл. 7.6).

Таблица 7.6

Двухфакторный дисперсионный комплекс: **c** *градаций фактора A (столбцы) и* **r** *градаций фактора B (ряды) с* **n** *повторениями в каждой градации (l = 1, 2,…, r; j = 1, 2,…, c; i = 1, 2,…, n)*

	***A*1**	**…**	***Aj***	***Ac***
*B*1	x_{111} x_{112} …	…	…	x_{1c1} x_{1c2} …
…	…	…	…	…
Bl	…	…	x_{lji}	…
Br	x_{r11} … x_{r1n}	…	x_{rjn}	x_{rc1} … x_{rcn}

Однако простейшей структурой дисперсионного анализа служит таблица, поля и графы которой характеризуют градации действия двух факторов, а в каждой ячейке содержится лишь одно значение результативного признака (табл. 7.7).

Таблица 7.7

Дисперсионный комплекс для трех градаций без повторений

	***A*1**	***A*2**	***A*3**
*B*1	x_{11}	x_{12}	x_{13}
*B*2	x_{21}	x_{22}	x_{23}
*B*3	x_{31}	x_{32}	x_{33}

Комплексы без повторений в градациях упрощают не только алгоритм обработки, но, к сожалению, и результаты. Сумма квадратов разлагается только на следующие компоненты:

$C_{общ.} = C_A + C_B + C_{остат.}$,

эффект сочетанного действия становится не отличим от случайного варьирования ($C_{остат.} = C_{AB} + C_{случ.}$).

Техника расчетов

Рассмотрим конкретный пример – испытания стимулятора многоплодия при разной полноценности рационов. Полноценность рациона (первый фактор) представлена двумя градациями: *A*1– рацион с недостатком минеральных веществ, *A*2 – рацион, полностью сбалансированный по всем питательным веществам, включая и минеральные. Стимулятор (второй фактор) был испытан в трех дозах: *B*1 – одинарная, *B*2 – двойная, *B*3 – тройная. Результативный признак – плодовитость самок, измерявшаяся числом детенышей в помете. Для каждого сочетания градаций рациона и стимулятора были подобраны три одновозрастные самки.

Комбинативная таблица двухфакторного равномерного дисперсионного комплекса с трехкратной повторностью ($n_i = 3$) включает две градации по фактору *A* и три градации по фактору *B* (табл. 7.8). Варианты размещаются по градациям, определяется объем градации, вычисляются суммы вариант, частные средние, затем вспомогательные величины ($H1$, $H2$, $H3$, H_A, H_B) и суммы квадратов отклонений (дисперсий) по рабочим формулам. В завершение всего заполняют таблицу дисперсионного анализа (табл. 7.9), находят показатель достоверности влияния Фишера и, сопоставляя его с табличным для соответствующих степеней свободы и принятого уровня значимости, делают статистический вывод.

Таблица 7.8

		***A*1**		***A*2**				**Для**	***B***
		x	x^2	x	x^2	Σ	M_B	ΣΣx^2/n	Σ(Σx^2/n)
*B*1		5 6 7	25 36 49	1 4 1	1 16 1				
	Σx^2		110		18	ΣΣx^2 = 128			
	Σx	18		6		ΣΣx = 24	4	96	
	n	3		3		n_{B1} = 6			
	Σx^2/n	108		12		Σ(Σx^2/n)= 120			
*B*2		4 3 5	16 9 25	10 9 11	100 81 121				***H*$_B$** = Σ(Σx^2/n) = 486
	Σx^2		50		302	ΣΣx^2 = 352			
	Σx	12		30		ΣΣx = 42	7	294	
	n	3		3		n_{B2} = 6			
	Σx^2/n	48		300		Σ(Σx^2/n)= 348			
*B*3		2 3 1	4 9 1	7 4 7	49 16 49				
	Σx^2		14		114	ΣΣx^2 = 128			
	Σx	6		18		ΣΣx = 24	4	96	
	n	3		3		n_{B3} = 6			
	Σx^2/n	12		108		Σ(Σx^2/n)=120			
ΣΣ	ΣΣx^2		174		434	***H*1** = ΣΣΣx^2 = 608			
	ΣΣx	36		54		ΣΣΣx = 90	***H*2** = (ΣΣΣx)²/N = 450		
	n_A = Σn	9		9		N = ΣΣn = 18			
	Σx^2/n	168		420		***H*3** = ΣΣ(Σx^2/n) = 588			
Для A	M_A = ΣΣx/n	2		6		c – число градаций фактора А (столбцы) r – число градаций фактора В (ряды)			
	Σx^2/n	144		324					
	***H*$_A$** = Σ(Σx^2/n) = 468								

$C_{общ.} = H1 – H2 = 608 – 450 = 158$
$C_{случ.} = H1 – H3 = 608 – 588 = 20$
$C_{факт.} = C_{A+B+AB} = H3 – H2 = 588 – 450 = 138$
$C_A = H_A – H2 = 468 – 450 = 18$
$C_B = H_B – H2 = 486 – 450 = 36$
$C_{AB} = C_{факт.} – C_A – C_B = 138 – 18 – 36 = 84$

В нашем примере все факториальные влияния оказались достоверными с доверительной вероятностью $P>0.95$. Это позволяет сделать определенные выводы относительно действия стимулятора на плодовитость самок. Влияние каждого фактора в отдельности (качества рациона и дозы стимулятора) и их суммарного эффекта достаточно существенно, но особенно результативно действие стимулятора в сочетании с полноценным рационом (величина η^2_{AB} выше, чем η^2_A и η^2_B). Более того, при недостатке в корме минеральных веществ двукратные и трехкратные дозы стимулятора могут даже снизить плодовитость животных.

Таблица 7.9

Составляющие дисперсии	**Суммы квадратов, C**	**Сила влияния, η^2 (%)**	**Степени свободы, *df***	**Дисперсии, S^2**	**Критерий, *F* ($F_{(\alpha,dfi,dfсл.)}$)**
Фактор *A*	18	11	$c–1 = 1$	18	10.8 (4.7)
Фактор *B*	36	23	$r–1 = 2$	18	10.8 (3.9)
Взаимодействие *AB*	84	53	$df_A \cdot df_B = 2$	42	25.2 (3.9)
Факториальная (всего)	138	87	$c \cdot r – 1 = 5$	27.6	16.5 (3.1)
Случайная	20	13	$N – c \cdot r = 12$	1.67	
Общая	158	100	$N – 1 = 17$		

Таблица двухфакторного дисперсионного анализа имеет ту же структуру, что и таблица для однофакторного анализа, только факториальная дисперсия разложена на три компоненты (для факторов *A*, *B* и их взаимодействия). Для каждой из них требуется вычислить число степеней свободы с учетом числа градаций фактора *A* (*c*, количество столбцов) и числа градаций фактора *B* (*r*, количество

рядов), значения дисперсий, а также критерий Фишера. Поскольку каждому из расчетных значений критерия соответствует свое число степеней свободы, табличные значения окажутся разными.

Дисперсионный анализ в среде Excel

Алгоритм двухфакторного дисперсионного анализа, естественно, требует более сложных вычислительных операций, чем однофакторный, но все они заложены в программе Excel «Двухфакторный дисперсионный анализ с повторениями» (только для равномерных комплексов!), который вызывается командой меню Сервис/ Анализ данных …. Исходные данные следует расположить на листе Excel по схеме 1 (табл. 7.6).

	A	B	C
1		A1	A2
2	B1	5	1
3		6	4
4		7	1
5	B2	4	10
6		3	9
7		5	11
8	B3	2	7
9		3	4
10		1	7
11			

В пункте макроса «Число строк для выборки» следует поставить объем выборки в одной градации, n; для нашего примера $n = 3$. Результаты расчетов (рис. 7.3) помимо общей статистической обработки для каждой градации содержат дисперсионную таблицу, почти идентичную приведенной выше (табл. 7.9). Отличие касается отсутствия строки для учета общей факториальной суммы квадратов ($C_{факт.}$) и дисперсии ($S^2_{факт.}$), а также добавления новых столбцов – табличного значения уровня значимости для рассчитанного критерия F Фишера; табличное значение критерия также приведено для уровня значимости $\alpha = 0.05$.

В среде пакета Excel есть возможность провести и «Двухфакторный дисперсионный анализ без повторений», который вызывает-

ся командой меню Сервис/ Анализ данных …. Как отмечалось выше, эта схема организации данных не позволяет разделить случайное варьирование и взаимодействие факторов. Например, если в качестве исходных данных взять только первые значения из предыдущего набора (табл. 7.8), дисперсионный анализ без повторений даст следующие результаты (рис. 7.4).

	A	B	C	D	E	F	G
21							
22	*Итого*						
23	Счет	9	9				
24	Сумма	36	54				
25	Среднее	4	6				
26	Дисперси	3.75	13.75				
27							
28							
29	Дисперсионный анализ						
30	*очник вари*	*SS*	*df*	*MS*	*F*	*P-Значение*	*критическое*
31	Выборка	36	2	18	10.8	0.002075	3.88529
32	Столбцы	18	1	18	10.8	0.006503	4.747221
33	Взаимодей	84	2	42	25.2	5.06E-05	3.88529
34	Внутри	20	12	1.666667			
35							
36	Итого	158	17				

Рис. 7.3. *Двухфакторный дисперсионный анализ на листе Excel*

Как видно из таблицы анализа, изменчивость, обусловленная взаимодействием факторов, объединена со случайной в строке «Погрешность».

Рассмотренные схемы дисперсионного анализа принципиально соответствуют и более сложным задачам, в частности, многофакторному дисперсионному анализу. Поскольку статистическая обработка многофакторных (особенно неравномерных) комплексов требует значительного увеличения расчетных работ, для таких задач мы рекомендуем использовать не возможности Excel, но специализированные пакеты программ ЭВМ, например StatGraphics.

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
2							
3	*ИТОГИ*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
4	B1	2	6	3	8		
5	B2	2	14	7	18		
6	B3	2	9	4.5	12.5		
7							
8	A1	3	11	3.666667	2.333333		
9	A2	3	18	6	21		
10							
11							
12	Дисперсионный анализ						
13	*очник вари*	*SS*	*df*	*MS*	*F*	*P-Значение*	*критическое*
14	Строки	16.33333	2	8.166667	0.538462	0.65	19.00003
15	Столбцы	8.166667	1	8.166667	0.538462	0.539434	18.51276
16	Погрешно	30.33333	2	15.16667			
17							
18	Итого	54.83333	5				

Рис. 7.4. *Двухфакторный дисперсионный анализ данных без повторений на листе Excel*

Дисперсионный анализ в среде StatGraphics

Рассмотрим использование пакета StatGraphics для проведения двухфакторного дисперсионного анализа по тем же данным. Исходные данные для обработки с помощью пакета StatGraphics лучше всего подготавливать на листе Excel, а затем импортировать в StatGraphics. Среда Excel более «дружелюбна», допускает операции автозаполнения и к тому же при импорте названия переменных назначаются автоматически (см. ниже). Пакет StatGraphics (версия 2.1) разработан для ранних версий Windows, поэтому импорт данных возможен только в старых форматах файлов типа *.dbf (для dBase II, III) или *.xls. (для MS Excel 4.0). Общий порядок операций по обработке данных таков:

- подготовка данных в среде Excel,
- экспорт данных в файле типа *.xls. для MS Excel 4.0,

– импорт данных в среду StatGraphics,
– проведение расчетов.

Подготовим данные из табл. 7.8 для двухфакторного дисперсионного анализа в среде Excel.

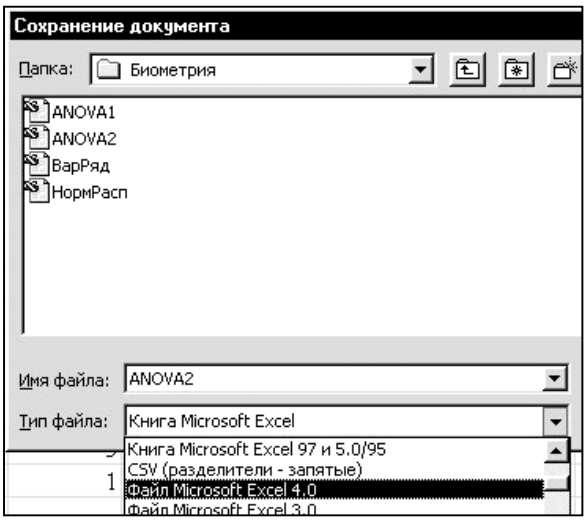
	A	B	C
1	A	B	E
2	1	1	5
3	1	1	6
4	1	1	7
5	1	2	4
6	1	2	3
7	1	2	5
8	1	3	2
9	1	3	3
10	1	3	1
11	2	1	1
12	2	1	4
13	2	1	1
14	2	2	10
15	2	2	9
16	2	2	11
17	2	3	7
18	2	3	4
19	2	3	7
20			

Чтобы StatGraphics мог распознать градации факторов, при которых получены значения результативного признака *E* (плодовитость), нужно ввести коды для доз обоих факторов, причем в форме числовых переменных. Так, первые 9 значений плодовитости получены при действии дозы 1 фактора *A*, следующие 9 значений – при дозе 2; вводим для этих значений признака *E* коды доз 1 (ячейки A2:A10) и 2 (ячейки A11:A19). Каждая из этих градаций включает по три градации фактора *B*, введем их коды в столбец B. Третий столбец образуют собственно значения результативного признака плодовитости (*E*), полученные в соответствующих градациях двух факторов.

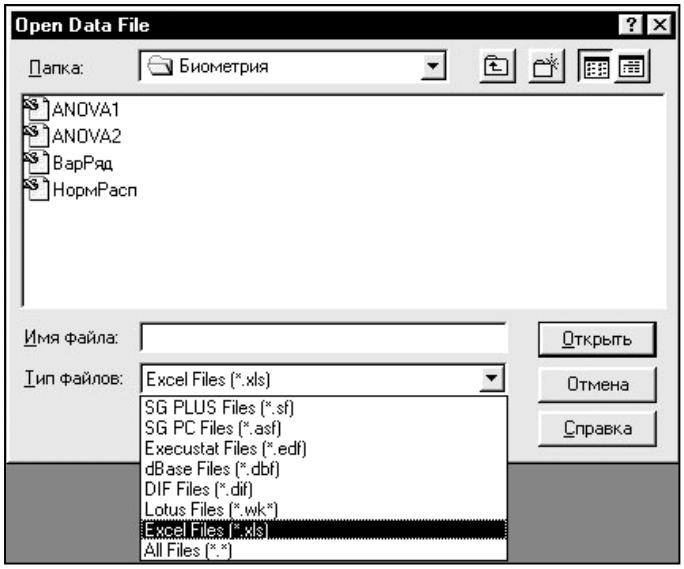
Так, значение $E = 11$ (ячейка C7) получено при дозах $A = 1$, $B = 2$. Если проводится изучение действия более чем двух факторов, на листе организуются все новые и новые столбцы с кодами градаций факторов. При этом важно следить, чтобы были представлены все сочетания градаций. В нашем случае, например, и градация *A*1, и градация *A*2 должны содержать по три градации второго фактора: *B*1, *B*2, *B*3. При этом StatGraphics не требует равного объема выборок для всех градаций.

Экспорт подготовленных данных из среды Excel осуществляется командой меню Файл\ Сохранить как …. В окне Тип файла: следует выбрать Файл Microsoft Excel 4.0. В окне Имя файла: задать новое имя, чтобы не утратить информацию, содержащуюся на дру-

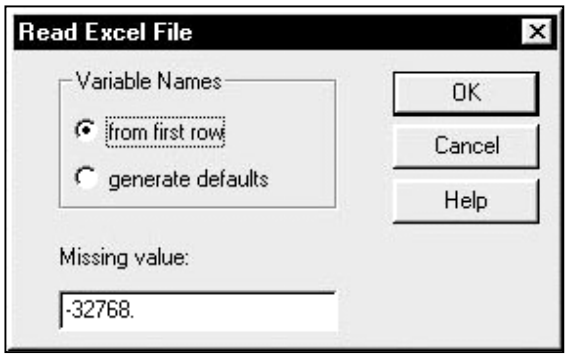
гих листах текущей книги, ОК. Далее, на запрос о сохранении только текущего листа ответить ОК.



Импорт данных в среду StatGraphics осуществляется третьей слева кнопкой панели Toolbar или командой меню File\ Open Data File... . В окне Тип файлов (Files type:) появившегося фрейма выделить Excel Files (*.xls), затем следует указать директорию, содержащую искомый файл, щелкнуть на его имя и Открыть.

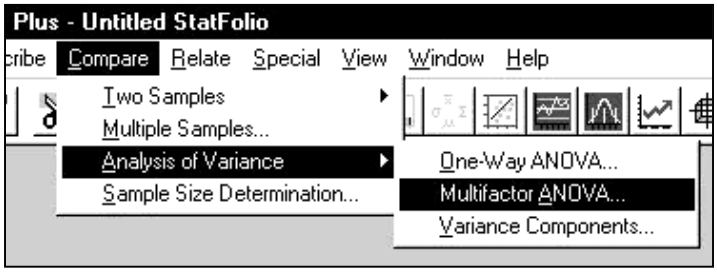


В появившемся окошке Read Excel File указать, что имена переменных Variable Names нужно брать из первого ряда (from first row), OK.



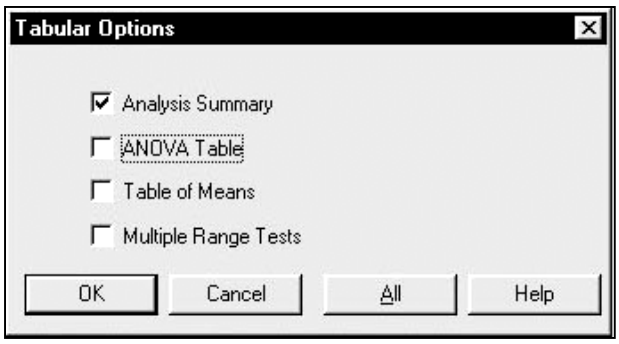
Информация из файла попадет в блок данных, чья свернутая панель расположена слева внизу. Развернуть окно данных можно двойным кликом на «шапке».

Расчеты по схеме двухфакторного дисперсионного анализа запускаем командой меню Compare\ Analysis of Variance\ Multifactor ANOVA.

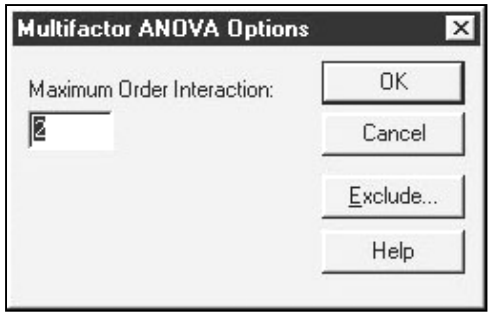


В появившемся окне Multifactor ANOVA результативный признак *E* заносим в графу «зависимая переменная» (Dependent Variable:), т. е. выделяем имя мышкой и нажимаем на кнопку стандартного отклонения стрелкой. Оба фактора заносим в графу Factors:, OK. Сразу же все расчеты будут выполнены, но отобразится только одна панель с общим описанием переменной и факторов.

Для отображения главных результатов, в первую очередь таблицы дисперсионного анализа, нужно нажать на вторую слева желтую кнопку (Tabular options), в новом окне отметить галочкой ANOVA Table или нажать кнопку All, OK.



Чтобы раскрыть новое окно `Analysis of variance for E`, следует на нем дважды кликнуть. Раскроется таблица дисперсионного анализа, рассчитанная по схеме «без повторений» и не содержащая оценку взаимодействия факторов. Рассчитать этот эффект можно, изменив установки анализа. Правой кнопкой мыши нужно щелкнуть на поле дисперсионной таблицы и выбрать из контекстного меню пункт Analysis options, после чего в окошке Multifactor ANOVA options указать, что число взаимодействующих факторов равно 2, OK.



Дисперсионная таблица сразу приобретет строку учета взаимодействия (`INTERACTIONS AB`). С помощью этой опции можно эффективно регулировать «глубину» учета взаимодействий, когда исследуется несколько факторов.

```
Analysis of Variance for E - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:A                             18.0      1           18.0
 B:B                             36.0      2           18.0      10.80     0.0021

INTERACTIONS
 AB                              84.0      2           42.0      25.20     0.0001

RESIDUAL                         20.0     12        1.66667
--------------------------------------------------------------------------------
TOTAL (CORRECTED)               158.0     17
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.


The StatAdvisor
---------------
   The ANOVA table decomposes the variability of E into contributions
due to various factors.  Since Type III sums of squares (the default)
have been chosen, the contribution of each factor is measured having
removed the effects of all other factors.  The P-values test the
statistical significance of each of the factors.  Since 2 P-values are
less than 0.05, these factors have a statistically significant effect
on E at the 95.0% confidence level.
```

Результаты дисперсионного анализа полностью идентичны табл. 7.8 и табл. 7.9. Важно отметить, что итог всех вычислений в среде StatGraphics сопровождается комментариями о методах расчета, а также статистическими выводами. Текст комментариев можно скопировать в буфер обмена из окна StatAdvisor.

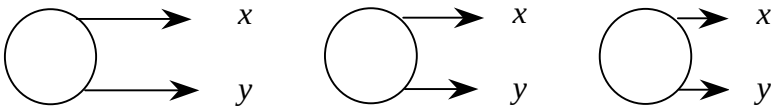
8
ЗАДАЧА «НАЙТИ ЗАВИСИМОСТЬ МЕЖДУ ДВУМЯ ПРИЗНАКАМИ»

Изложенные выше методы статистического анализа дают возможность изучать изменчивость биологических объектов по отдельным признакам – весу, размерам, плодовитости, физиологическим показателям и др. Однако в ряде случаев важно знать, какова зависимость между вариацией двух или нескольких признаков, изменяются ли две переменные самостоятельно, независимо друг от друга, или изменчивость одного признака в какой-то степени связана с изменчивостью другого. В качестве второй переменной часто выступает какой-либо фактор среды.

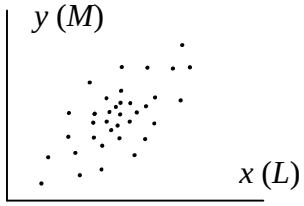
Эту задачу можно рассматривать как развитие метода дисперсионного анализ, решающего задачу сравнения нескольких выборок (изучения влияния фактора на признак). Техника дисперсионного анализа имеет две особенности. Во-первых, фактор (факториальный признак) задан дискретно, в виде градаций, или «доз». Когда исследуется фактор, заданный качественно, то градации оказываются очень эффективным способом его превращения в подобие количественно заданного фактора. Вместе с тем фактор, выраженный количественной величиной, имеет большее число значений, чем число градаций. Тогда в грубой градуальной схеме дисперсионного анализа утрачивается часть информации, имеющейся в исходных выборках. Кроме этого, дисперсионный анализ явным образом не учитывает тенденции изменения среднего уровня признака при изменении уровня фактора, не содержит показателя динамики зависимости признака от фактора.

Сделать необходимые дополнения позволяет исследование сопряженной (взаимозависимой) изменчивости признаков в рамках регрессионного и корреляционного анализов. Способ представления отдельных наблюдений здесь меняется: каждая варианта рассматривается как носитель двух численных характеристик объекта измерения, двух *зависимых* значений случайной величины. Если выше мы отождествляли отдельное значение с отдельной вариантой, то теперь мы рассматриваем варианту как некоторое тело, объект, обла-

дающий, как минимум двумя зарегистрированными качествами, различными у разных вариант:



Например, для любого животного можно определить массу (*M*) и длину (*L*) тела; отдельная варианта будет нести два значения (*L*, *M*). При этом множество вариант выборки можно отобразить графически как точки на плоскости осей двух признаков *M* и *L*.



Вся выборка предстанет в виде множества точек на плоскости (двумерное рассеяние). Как видно на диаграмме, «облако» вариант вытянуто в направлении диагонали облака точек. Справа вверху находятся варианты с высокими значениями и размеров, и массы тела, в левом нижнем углу – с наименьшими значениями. В центре находятся варианты с промежуточными, средними значениями. В первом приближении двумерное распределение – это простая ординация вариант на плоскости осей двух признаков.

Помимо рассеяния на плоскости в определение двумерного распределения входит и частота встречаемости отдельных вариант. В соответствии с идеологией регрессионного анализа признаки *x* и *y* должны подчиняться нормальному закону. Значит, для каждого значения *x* признак *y* дает множество нормально распределенных значений; то же и для каждого значения признака *y* (для случая математической совокупности бесконечного объема) (рис. 8.1). Скопление вариант в трех осях (оси признаков *x, y* и частоты *a*) образует весьма странный «бугор», растянутое в пространстве трехмерное нормальное распределение. Однако в реальности такой идеальной картины получить никогда не удается, приходится ориентироваться только на плоскую фигуру рассеяния немногочисленных вариант.

Если область, занятую вариантами, очертить по периферии плавной линией, мы получим вытянутую фигуру, эллипс, ограничи-

вающий область рассеяния вариант, эллипс рассеяния. Эллипс рассеяния – это область распространения вариант одной совокупности.

Можно видеть, что в нашем случае признаки связаны друг с другом – есть общая тенденция: чем больше длина тела, тем больше вес, хотя эта зависимость и не очень жесткая, но размыта индивидуальными особенностями.

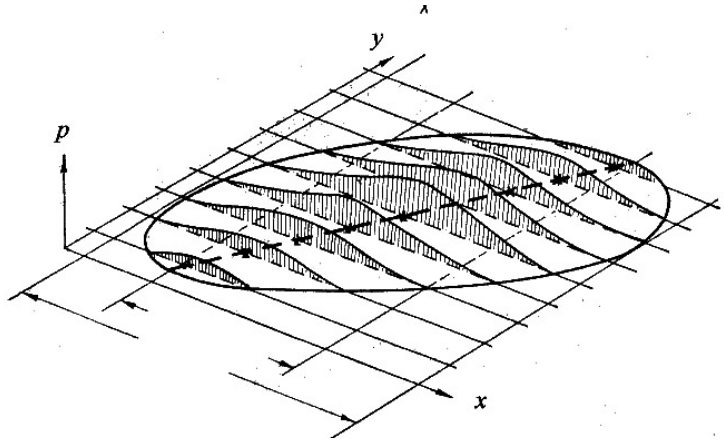


Рис. 8.1. *Двумерное распределение*

Таблица 8.1

Задача	**Содержание задачи**	**Методы**
Доказать зависимость одного признака от другого	Признак *x* служит доминирующим фактором для признака *y*	Регрессионный, дисперсионный и корреляционный анализы
Доказать зависимость одной переменной от нескольких других	Переменные x_1, x_2, … влияют на признак *y*	Множественная корреляция, регрессия
Доказать взаимозависимость двух признаков	Признак *x* служит доминирующим фактором для признака *y*, и наоборот	Корреляционный анализ
Доказать связь двух признаков, исключив влияние третьего	Признак *z* служит доминирующим фактором для признаков *x* и *y*	Метод частной корреляции
Доказать зависимость неколичественных признаков	Изменчивость признаков сопряжена	Коэффициент Спирмена

Итак, в двумерном распределении проявляются два эффекта: синхронное изменение двух признаков и размывание этой синхронности, т. е. действие факторов доминирующих и случайных: доминирующий фактор (фактор сопряжения признаков) действует вдоль оси эллипса, случайные факторы – поперек оси, размывая взаимозависимость *y* и *x*. Проблема изучения зависимости распадается на ряд частных задач (табл. 8.1).

Регрессионный анализ зависимости двух признаков

Регрессионный анализ изучает эффект влияния одного признака на другой, зависимость признака от фактора, зависимость результативного признака от факториального. Его основные результаты таковы:

1. Таблица дисперсионного анализа, в которой показаны сила и достоверность влияния на признак изучаемого фактора или другого признака (таблица разложения общего варьирования результативного признака на компоненты и соотнесение их друг с другом).

2. Уравнение регрессии, выражающее пропорциональность сопряженного изменения признаков, тенденции их взаимосвязанной изменчивости или динамики.

3. Оценки значимости параметров регрессионного уравнения.

Логико-теоретические основы

Регрессионный анализ методически односторонне ориентирован на изучение зависимости одного признака от другого (зависимость *y* от *x* или, напротив, зависимость *x* от *y*), хотя может применяться к случаям, когда фактически имеется взаимозависимость двух переменных. В свою очередь, обобщенная зависимость исследуется «симметричным» методом – корреляционным анализом.

Судить о том, как меняется одна величина по мере изменения другой, позволяет коэффициент регрессии (*a*), показывающий, на какую величину в среднем изменяется один признак (*y*) при изменении другого (*x*) на единицу измерения:

$$y - Y = a \cdot (x - X).$$

Простые преобразования:

$$y = a \cdot x + Y - a \cdot X,$$

$b = Y - a \cdot X$

приводят к уравнению линейной регрессии:

$y = ax + b$.

Возможность получить уравнение зависимости признаков позволяет важная смена идеологии: регрессионный анализ сравнивает друг с другом не выборки, разнесенные по градациям фактора, но отдельные варианты, т. е. изучает характер рассеяния вариант в осях двух изучаемых признаков, сопряженную изменчивость признаков.

Основную тенденцию взаимосвязанного изменения двух признаков можно отобразить с помощью простого графического приема. Разобьем ось *x* на несколько интервалов. Найдем для каждого из них среднее (M_y) значение признака *y*. Теперь проведем через эти средние точки ломаную линию. Это будет линия регрессии *Y* по *x*. *Регрессия* – изменение среднего уровня одного признака при изменении другого (рис. 8.2).

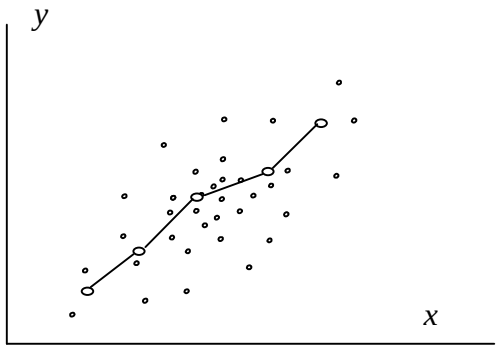


Рис. 8.2. *Эмпирическая линия регрессии*

К сожалению, ход ломаной линии нельзя передать простым уравнением, к тому же на нем сказываются способ интервального разбиения оси абсцисс, а также уровень репрезентативности в разных областях распределения. В этом смысле предпочтительнее была бы единственная прямая линия регрессии, подчеркивающая основные тенденции зависимости признаков и выраженная простым уравнением:

$Y = ax + b$

(заменив символ для обозначения зависимого признака с *y* на *Y*, мы

подчеркиваем, что на базе признака x уравнение позволяет рассчитать теоретическое, среднее, значение признака Y, в общем не равное ни одному наблюдаемому значению y).

Грубо регрессионную линию можно построить, взяв всего две точки – средний уровень признаков в верхней и нижней половинках эллипса (рис. 8.3).

Гораздо точнее определить и уравнение регрессии, и ход графика прямой линии можно в том случае, если учесть информацию по всем вариантам изучаемой совокупности. Для этой цели разработан *метод наименьших квадратов*, основная идея которого состоит в том, чтобы линия регрессии прошла на наименьшем удалении от каждой точки, т. е. чтобы сумма квадратов расстояний от всех точек до прямой линии была наименьшей. В математической статистике показано, что для случая двумерного нормального распределения лучшей (эффективной, несмещенной и пр.) линией, описывающей зависимость одного признака от другого, может быть только линия средних арифметических. Линия регрессии признака y по признаку x – это множество частных средних Y_i, соответствующих определенным значениям x_i.

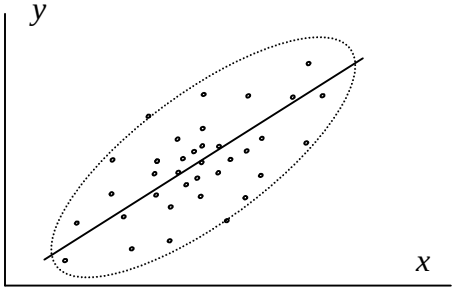


Рис. 8.3. *Примерная прямолинейная регрессия*

Используя метод наименьших квадратов, вычислить коэффициенты линейной регрессионной модели можно по следующему алгоритму.

Сначала найдем вспомогательные величины:

$Cx = \Sigma x^2 - (\Sigma x)^2/n$, $Cy = \Sigma y^2 - (\Sigma y)^2/n$, $Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y)/n$, $M_y = \Sigma y/n$, $M_x = \Sigma x/n$.

Затем рассчитаем коэффициенты:

$a = Cxy/Cx$, $b = M_y - a \cdot M_x$.

Существо коэффициента регрессии ***a*** состоит в том, что он призван выражать пропорцию изменения признака *y* при изменении признака *x*:

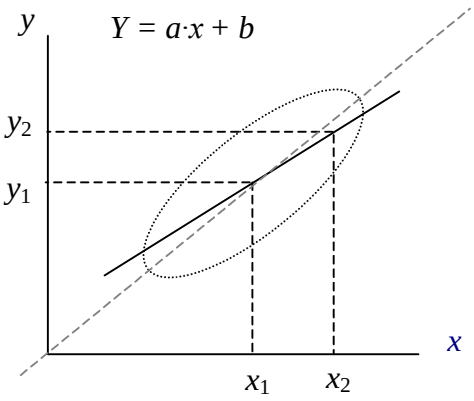
$$y - Y = a\cdot(x - X) \text{ или } a = \frac{y - M_y}{x - M_x},$$

но обобщенно для всех вариант выборки:

$$a = \frac{\sum (y - M_y)(x - M_x)}{\sum (x - M_x)^2} = \frac{Cxy}{Cx}.$$

В этой формуле числитель характеризует только сопряженную изменчивость обоих признаков, знаменатель – квадрат общей изменчивости признака *x*; в итоге имеем показатель пропорции изменения одного признака при изменении другого. Однако это не «чистая» пропорция, но искаженная случайными факторами. Здесь уместно обратиться к истории.

Термин «регрессия» предложил Ф. Гальтон. Анализируя зависимость роста сыновей (*y*) от роста отцов (*x*), он обнаружил, что в соответствии с линейным графиком у низкорослых отцов сыновья должны иметь более высокий рост, чем отцовский. Напротив, у более высоких отцов сыновья должны быть менее высоки, чем они сами ($x_2 - x_1 > y_2 - y_1$). Вместо интуитивно ожидаемой прямой пропорции между ростом отцов и детей (отмечена серым пунктиром, это



ось эллипса рассеяния) наблюдается определенное *возвращение* к среднему уровню, «регрессия», как ее назвал исследователь.

Причины такого явления состоят в том, что в случае стохастической зависимости для предсказания значений одного признака

по значениям другого требуется показатель, который наиболее обоснован со статистической точки зрения. Таким показателем является средняя арифметическая (точнее, условная средняя, линия регрессии), но ее значения не лягут точно на ось эллипса рассеяния, кроме центральной точки (M_y, M_x). Однако истинную зависимость (пропорцию) не дает точно охарактеризовать случайная изменчивость. Поэтому чем больше величина случайной составляющей общей изменчивости (Cx) по сравнению с сопряженной (Cxy), тем сильнее линия регрессии будет отклоняться от оси эллипса, т. е. чем больше знаменатель, тем ближе к нулю величина коэффициента регрессии.

Построить регрессионное уравнение – это еще даже не полдела, важнее оценить значимость зависимости признаков, реальность их взаимодействия, т. е. установить, что признак x является существенным, «доминирующим» фактором, сказывается на изменчивости признака y.

Сходную задачу о достоверном влиянии фактора мы решали с помощью критерия исключения выскакивающих вариант. При этом изучаемая выборка состояла из двух частей – некоего «ядра», внутри которого варианты отличаются друг от друга по случайным причинам, и периферических вариант, которые отклонились от «ядра» за счет действия каких-то новых (доминирующих) факторов. Границы области случайного варьирования определялись по «соглашению 95%» и составляли $M \pm 2S$. Чем больше выборка, тем более точно определяются эти границы.

Перенесем эту логику на случай двумерного нормального распределения. Это значит, что всю область рассеяния вариант можно разбить на две зоны. Во-первых, это «ядро», в котором варианты отличаются друг от друга только по случайным причинам, т. е. факториальный признак x не влияет на результативный признак y. На плоскости двух осей граница области случайного варьирования будет иметь форму окружности, случайный разлет вариант от средней возможен, естественно, во все стороны. Во-вторых, по периферии будут располагаться варианты, отклонившиеся от «ядра» за счет действия доминирующего фактора, т. е. за счет взаимодействия признаков. Такое положительное влияние x на y означает, что чем больше будет значение признака x, тем больше будет и значение признака y, а чем меньше x, тем меньше y. Получается, что вариан-

ты, не случайно отклонившиеся от общей средней (от центра), будут накапливаться вверху справа и внизу слева от круглого «ядра». Область рассеяния вариант сформирует эллипс.

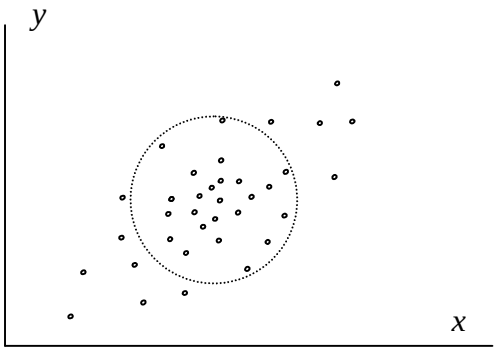


Рис. 8.4. *Взаимодействие признаков есть «растягивание» окружности в эллипс*

Оценка достоверности взаимодействия признаков есть задача описания пропорций эллипса рассеяния: достаточно ли много вариант выходят за границы случайного рассеяния (за границы круга), чтобы с уверенностью говорить о реальности связи признаков *x* и *y*. Для этой цели используется общая идея статистического оценивания – соотнести отклонения под действием доминирующего фактора с отклонениями по случайным причинам.

Лучшим показателем взаимосвязи является линия регрессии (динамика среднего уровня), которая пытается показать только взаимозависимое изменение признаков и вовсе не рассматривает независимое варьирование каждого из них. В свою очередь, характеристикой чисто случайного варьирования выступает отклонение отдельных вариант от линии регрессии.

Эта идея позволяет построить базовую модель варианты в регрессионном анализе (рис. 8.5):

$$y_i = M_y \pm y_x \pm y_{сл.},$$

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.

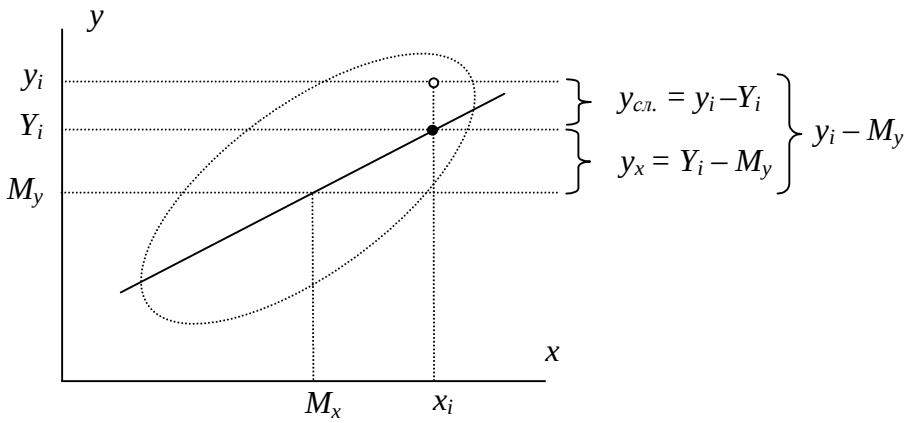


Рис. 8.5. *Модель варианты в регрессионном анализе*

Таким образом, отклонение варианты от общей средней арифметической связано с действием факториального признака и с действием случайных причин:

$$(y_i - M_y) = (y_i - Y_i) + (Y_i - M_y),$$

где $y_i - M_y$ – общее отклонение варианты от средней,

$y_{сл.} = y_i - Y_i$ – отклонение варианты от линии регрессии, отклонение по случайным причинам,

$y_x = Y_i - M_y$ – отклонение линии регрессии (для точки x_i) от средней, т. е. отклонение под действием факториального признака x.

Представленная модель позволяет подойти к количественной оценке достоверности связи признаков в целом. Для этого нужно все рассмотренные отклонения объединить по всем вариантам выборки, причем чтобы суммы отклонений не обратились в нуль, возвести их в квадрат. Таким образом мы получаем оценки факториальной и остаточной сумм квадратов, т. е. можем построить таблицу дисперсионного анализа, аналогичную рассмотренной выше (однофакторный дисперсионный анализ): изменчивость признака y складывается из варьирования, учтенного регрессионной моделью, и из варьирования по случайным причинам, т. е. остаточного.

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. Оста-

точную сумму квадратов ($C_{остат.} = \Sigma(y_i - Y_i)^2$) находят также непосредственно как сумму квадратов отличий между значением y_i для каждой варианты и значением, предварительно рассчитанным по уравнению регрессии $Y_i = ax_i + b$ (для соответствующих значений x_i). Модельную сумму квадратов ($C_{мод.} = \Sigma(Y_i - M_y)^2$) рассчитывают как разность между общей и остаточной ($C_{мод.} = C_{общ.} - C_{остат.}$).

Таблица 8.2

Составляющие дисперсии	**Суммы квадратов, *C***	**Формулы расчета сумм квадратов**	***df***	**S^2**	**F**
Наклон модельной линии	$C_{мод.} = \Sigma(Y_i - M_y)^2$	$C_{общ.} - C_{остат.}$	1	$S^2_{мод.} = \frac{C_{мод.}}{df_{мод.}}$	$\frac{S^2_{мод.}}{S^2_{остат.}}$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma(y_i - Y_i)^2$		$n-2$	$S^2_{остат.} = \frac{C_{остат.}}{df_{остат.}}$	$F_{(0.05,1,n-2)}$
Общая (всего)	$C_{общ.} = \Sigma(y_i - M_y)^2$	$(\Sigma y_i^2 - \Sigma y_i)^2/n = C_y$			

На этом этапе можно 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.

На основе полученных сумм квадратов рассчитываем модельную и остаточную дисперсии. Число степеней свободы для остаточной дисперсии берут равным $df = n-2$, поскольку в расчетах теоретических значений принимают участие два параметра – a и b. В тех случаях, когда свободный член (b) значимо от нуля не отличается, расчеты теоретических значений проводятся при одном коэффициенте (a) и число степеней свободы берут $df = n-1$.

После предварительных расчетов с помощью критерия Фишера можно проверить нулевую гипотезу Но: предсказания модели

в целом неадекватно описывают исходные данные, зависимости между признаками нет. Конструкция критерия исследует вопрос, превышает ли варьирование, учтенное моделью, случайное (остаточное) варьирование? Критерий Фишера вычисляется как отношение модельной и остаточной дисперсии:

$F = S^2_{мод.}/S^2_{остат.} \sim F_{(0.05,1,n-2)}$.

Если значение критерия окажется выше табличного, значит, дисперсия реального признака y приближается по величине к дисперсии модельного признака Y, 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.

Помимо дисперсионного анализа и критерия Фишера существуют другие способы доказательства влияния признака x на y, например, критерий T Стьюдента, проверяющий нулевую гипотезу Но: $a = 0$, коэффициент регрессии значимо от нуля не отличается. С этой целью рассчитывается ошибка коэффициента регрессии m_a и вычисляется величина

$T = (a-0)/ m_a = a/ m_a \sim T_{(0.05,\, n-2)}$.

Смысл этого критерия состоит в следующем. Коэффициент регрессии a характеризует сопряженность пропорционального изменения двух признаков, т. е. отвечает за то, что линия регрессии имеет некоторый угол относительно оси абсцисс. Значение $a = 0$ означает, что линия регрессии идет параллельно оси OX, что при изменении признака x признак y не меняется, что y не зависит от x. Значения $a > 0$ или $a < 0$ говорят о том, что взаимосвязь признаков имеет место.

Поскольку значение коэффициента регрессии оценивается по выборке, может статься, что a будет отличаться от нуля в силу случайных причин, вследствие ошибок репрезентативности (в действительности связи нет, а в выборке сочетание вариант дало слабый эффект). Иными словами, если при исследовании одного и того же явления получить множество выборок и для каждой из них рассчитать уравнение регрессии, то возможны два случая:

1. Для каждой повторной выборки мы будем получать устойчивые и сходные значения коэффициента регрессии, отличные от

нуля, т. е. зависимость между признаками действительно есть (рис. 8.6, А).

2. Для каждой повторной выборки мы будем получать варьирующие значения коэффициента регрессии, близкие к нулю, т. е. зависимость между признаками отсутствует (рис. 8.6, Б).

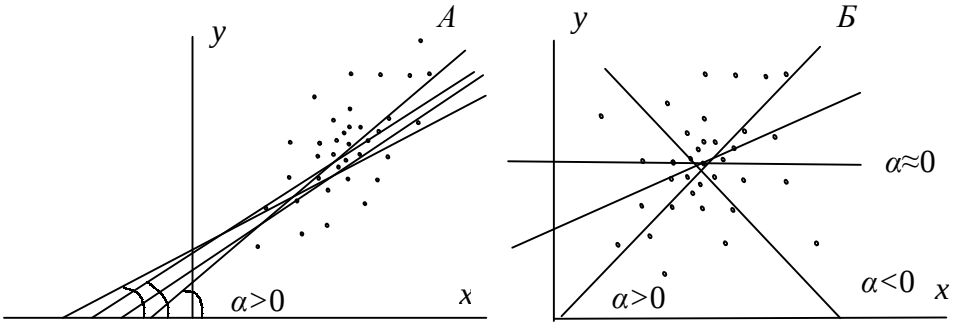


Рис. 8.6. *Варианты хода линии регрессии*

Коэффициенты регрессии, рассчитанные по разным выборкам, будут отличаться друг от друга и от генеральных значений. Соответственно, выборочные линии регрессии будут иметь разные углы наклона. Межвыборочную изменчивость коэффициентов регрессии можно охарактеризовать стандартным отклонением, названным ошибкой (репрезентативности) коэффициента регрессии (m_a). Понятно, что она будет характеризовать варьирование этого параметра по случайным причинам. В свою очередь, как показано выше, наклоненность линии регрессии обеспечена не случайными причинами. Поэтому отличие коэффициента регрессии от нуля ($a–0$), или просто величина a, оценивает силу связи между изучаемыми признаками. Если эта связь не случайна, то сопряженное варьирование двух признаков будет сильнее их свободного варьирования, тогда и отношение коэффициента регрессии к своей ошибке превысит критический уровень T статистики Стьюдента ($T_{(0.05,\, n–2)}$):

$T = (a–0)/ m_a = a/ m_a$.

Если же связи нет или она сильно загрязнена стохастическим шумом, то линия регрессии скроется в облаке возможных случайных траекторий, критерий даст значение ниже табличного.

Говоря о технической стороне, важно отметить, что рассчитать ошибку коэффициента регрессии можно и по одной единственной выборке (используя промежуточные величины, показанные выше):

$$m_a = \frac{S_y}{S_x} \cdot m_r ,$$

где S_x, S_y – стандартные отклонения для признаков,

$$S_x = \sqrt{Cx/(n-1)} , S_y = \sqrt{Cy/(n-1)} ,$$

m_r – ошибка коэффициента корреляции,

$$m_r = \sqrt{\frac{1-r^2}{n-2}} ,$$

r – коэффициент корреляции,

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} ,$$

n – объем выборки.

Оценка значимости коэффициентов регрессии особенно важна для случая множественной регрессии, когда оценивается зависимость результативного признака от нескольких факториальных. С помощью этой процедуры удается отделить существенные факторы влияния от малозначимых.

Наряду с первым коэффициентом линейной регрессии можно проверить значимость и второго коэффициента, b. Идеология метода не меняется, но рассматривается другая гипотеза Но: $b = 0$, т. е. проходит ли линия регрессии через начало осей координат, через нуль.

Здесь возможны те же варианты: либо линия регрессии проходит через нуль, и тогда выборочные коэффициенты регрессии случайно варьируют около этого значения (рис. 8.7, А), либо линия регрессии не проходит через точку пересечения осей координат, и выборочные коэффициенты регрессии действительно отличны от нуля (рис. 8.7, Б).

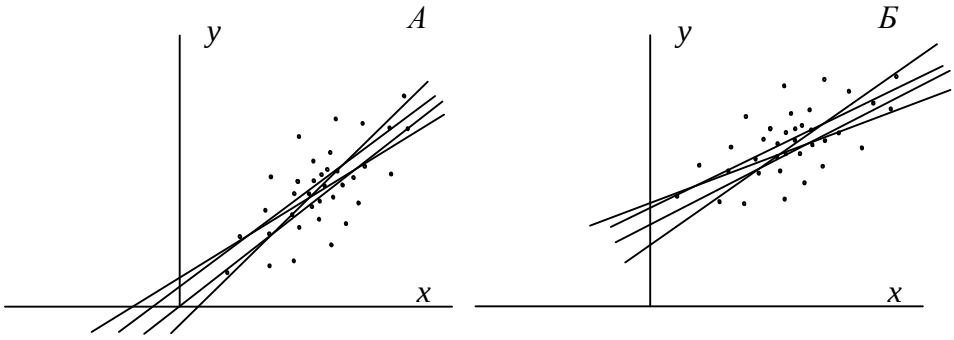


Рис. 8.7. *Варианты хода линии регрессии*

Проверяется эта гипотеза с помощью критерия Стьюдента, меняется только метод расчета ошибки второго коэффициента регрессии:

$$T = (b–0)/m_b = b/m_b \sim T_{(0.05,\ n-2)},$$

где $$m_b = m_y \cdot \sqrt{\frac{1}{n} + \left(\frac{M_x}{Cx}\right)^2},$$

n – объем выборки,

Cx – вспомогательная величина для признака x,

$Cx = \Sigma x^2 - (\Sigma x)^2/n$,

m_y – ошибка регрессионной средней или остаточное стандартное отклонение, может вычисляться по разным формулам:

$m_y = S_y \cdot \sqrt{1-r^2}$ (упрощенная формула для больших выборок),

$m_y = S_y \cdot \sqrt{\frac{(n-1)\cdot(1-r^2)}{n-2}}$ (точная формула для небольших выборок),

$m_y = \sqrt{\frac{\sum^{n}(y_i - Y_i)^2}{n-2}} = \sqrt{\frac{C_{остат.}}{n-2}} = \sqrt{S^2_{остат.}}$ (общая точная формула),

r – коэффициент корреляции,

M_x, M_y, S_y – средняя арифметическая и стандартное отклонение для рядов значений x и y,

$C_{остат.} = \sum^{n}(y_i - Y_i)^2$ – сумма квадратов отклонений расчетных (Y_i) от реальных значений признака y (остаточная сумма квадратов из таблицы дисперсионного анализа).

Если свободный член, коэффициент b, значимо от нуля не отличается, т. е. линия регрессии проходит через начало осей координат, следует пересчитать первый коэффициент регрессии a. Формула расчета коэффициента регрессии при этом упрощается:

$a = \Sigma(x \cdot y)/ \Sigma x^2$.

Регрессионная модель примет вид: $Y = ax$.

Ошибки коэффициентов регрессии позволяют рассчитать для каждого из них доверительные интервалы, ограничивающие область возможного варьирования с принятым уровнем значимости (значение $T_{(\alpha, n-2)}$ берется по таблице Стьюдента): $a \pm T \cdot m_a$, $b \pm T \cdot m_b$.

Варьирование коэффициентов a и b означает, что выборочная линия регрессии может иметь иной угол наклона, нежели генеральная, проходить в окрестностях несколько выше или несколько ниже центра, образуя целый «букет» из возможных случайно наклоненных выборочных линий регрессии (рис. 8.6). В силу нормального распределения признаков их множество укладывается в область сложной конфигурации с «перетяжкой» в окрестностях центра распределения. Этот феномен достаточно просто объяснить, имея в виду форму двумерного нормального распределения частот (рис. 8.1).

Точнее всего выборочные линии регрессии «угадывают» положение центра распределения (точки, соответствующей средним M_y, M_x), поскольку в этой области концентрация вариант наиболее велика, значит, и средняя оценивается с наименьшей ошибкой. Обычно линия регрессии пересекает этот центр. Напротив, по краям двумерного распределения частоты уменьшаются, варианты разрежены. Поэтому на периферии эллипса рассеяния ошибки определения среднего уровня результативного признака увеличены и выборочные линии регрессии могут далеко отклоняться от генеральной линии регрессии. По этой причине доверительный интервал, или доверительная зона линии регрессии, имеет не простую, не линейную конфигурацию (рис. 8.8).

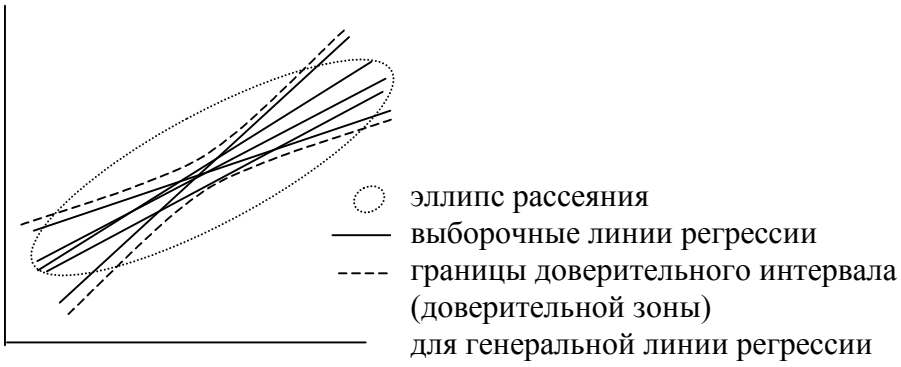


Рис. 8.8. *Доверительный интервал линии регрессии*

Теория двумерного нормального распределения предлагает методы расчета значений изменяющихся ошибок репрезентативности линии регрессии (m_Y), а также доверительного интервала (в котором с той или иной вероятностью находится генеральная линия регрессии); он задается границами:

$$Y_i \pm T \cdot m_Y = Y_i \pm T \cdot m_y \cdot \sqrt{\frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}},$$

где m_Y – ошибка линии регрессии (ошибка прогноза регрессионных средних Y_i),

Y_i – значение, рассчитанное по регрессионной модели для x_i,

T – величина нормированного отклонения из таблицы Стьюдента (табл. 6*П*), выбранная для данного числа степеней свободы ($df = n–2$) и уровня значимости α,

$S_{остат.} = \sqrt{S^2_{остат.}}$ – стандартное отклонение для случайных отклонений исходных значений y от теоретических Y,

n – объем выборки,

$(x_i – M_x)^2$ – мера отклонения значения x_i от средней M_x,

$Cx = \sum^{n} (x - M_x)^2$ – сумма квадрата отклонений всех значений x от своей средней M_x; рассчитывается по рабочей формуле: $Cx = \Sigma x^2 – (\Sigma x)^2/n$.

Как следует из формул, чем дальше значение x_i находится от средней арифметической M_x, тем больше числитель подкоренного выражения, т. е. тем больше для этого значения получится ошибка линии регрессии m_Y и тем шире будет доверительный интервал линии регрессии, т. е. интервал для предсказанного среднего значения признака Y_i для очередного наблюдаемого значения x_i.

Кроме этого, в рамках регрессионного анализа можно рассчитать *интервал прогноза* новых наблюдений:

$$Y_i \pm T \cdot S_Y = Y_i \pm T \cdot m_y \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}},$$

где S_Y – расчетное стандартное отклонение для предсказанных значений признака y.

Если *доверительный интервал* линии регрессии ($Y_i \pm T \cdot m_Y$) характеризует область ожидания генеральной линии регрессии (для средних), то *интервал прогноза* ($Y_i \pm T \cdot S_Y$) характеризует область, в которой с заданной вероятностью ожидается появление новых значений признака y (вариант) в случае продолжения наблюдений. Вероятность (уровень значимости), с которой в данном интервале ожидается появление варианты или среднего прогноза, задается соответствующей табличной величиной критерия Стьюдента $T_{(\alpha,\, n\text{-}2)}$.

Техника расчета линейной регрессии

Судить о том, на какую величину в среднем изменяется один признак (Y) при изменении другого (x) на единицу измерения, позволяет уравнение линейной регрессии: $Y = ax + b$.

Термин «линейная» относится к методу оценки коэффициентов регрессии (a, b), это *метод наименьших квадратов*, дающий уравнение *линии*, удаленной от всех точек двумерного распределения на наименьшее расстояние.

Способ вычисления уравнения регрессии показан в таблице 8.3 на примере зависимости между живым весом коров и их приплода (кг). По таблице рассчитываются квадраты вариант и их произведения, а также суммы вариант, квадратов и произведений. Вычисления ведутся по точным рабочим формулам.

Таблица 8.3

i	y	x	y^2	x^2	$x \cdot y$	Y	$(y-Y_i)^2$	$T \cdot m_Y$	min$Y$	max$Y$
1	25	352	625	123904	8800	25.6	0.31	2.0	23.6	27.5
2	26	376	676	141376	9776	27.1	1.29	1.7	25.5	28.8
3	31	402	961	161604	12462	28.8	4.65	1.4	27.4	30.2
4	32	453	1024	205208	14496	32.2	0.04	1.2	31.0	33.4
5	34	484	1156	234256	16456	34.2	0.06	1.3	32.9	35.5
6	38	528	1444	278784	20064	37.1	0.76	1.7	35.4	38.9
7	38	555	1444	308025	21090	38.9	0.81	2.1	36.8	41.0
Σ	224	3150	7330	1453158	103144		7.92			

Проведем последовательные расчеты. Сначала определим вспомогательные величины:

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y)/n = 103144 - 3150 \cdot 224/7 = 2344,$

$C_{общ.} = Cy = \Sigma y^2 - (\Sigma y)^2/n = 7330 - 224^2/7 = 162,$

$Cx = \Sigma x^2 - (\Sigma x)^2/n = 1453158 - 3150^2/7 = 35658,$

$C_{остат.} = 7.92,$

$C_{мод.} = 162 - 7.92 = 154.08;$

затем – параметры:

$M_y = \Sigma y/n = 224/7 = 32,$

$M_x = \Sigma x/n = 3150/7 = 450,$

$$S_y = \sqrt{\frac{\sum y^2 - \frac{(\sum y)^2}{n}}{(n-1)}} = \sqrt{\frac{Cy}{n-1}} = \sqrt{\frac{162}{6}} = 5.2,$$

$$S_x = \sqrt{\frac{Cx}{n-1}} = \sqrt{\frac{35658}{6}} = 77.1,$$

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$$

$$a = \frac{Cxy}{Cx} = \frac{2344}{35658} = 0.0657,$$

$b = M_y - a \cdot M_x = 32 - 0.0657 \cdot 450 = 2.419.$

Получено уравнение линейной регрессии $Y = 0.0657x + 2.419$, которое позволяет рассчитать теоретические значения Y_i (табл. 8.3) и провести дисперсионный анализ (табл. 8.4).

Расчетное значение F (97.3) превышает табличное (6.0), следовательно, модель адекватна реальности. Судя по коэффициенту детерминации, «сила влияния» веса коров на вес плода велика:

$$R^2 = \frac{154.08}{162} = 0.951.$$

Далее найдем ошибки параметров:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-0.975^2}{7-2}} = 0.099,$$

$$m_a = \frac{S_y}{S_x} \cdot m_r = \frac{5.2}{77.1} \cdot 0.099 = 0.00667,$$

$$m_y = S_y \cdot \sqrt{\frac{(n-1)\cdot(1-r^2)}{n-2}} = 5.2 \cdot \sqrt{\frac{(7-1)\cdot(1-0.975^2)}{7-2}} = 1.2582,$$

или $m_y = \sqrt{S^2_{остат.}} = \sqrt{1.5832} = 1.2582,$

$$m_b = m_y \cdot \sqrt{\frac{1}{n} + \frac{(M_x)^2}{Cx}} = 1.2582 \cdot \sqrt{\frac{1}{7} + \frac{(450)^2}{35658}} = 3.0359$$

и, наконец, критерий T Стьюдента для проверки значимости коэффициентов: $T_a = a/ m_a = 0.0657/ 0.00667 = 9.84$,

$T_b = b/ m_b = 2.419/ 3.0359 = 0.7968$.

Для уровня значимости α=0.05 и числа степеней свободы $df = n–2 = 5$ находим табличное значение критерия Стьюдента $T_{(0.05,5)} = 2.57$.

Таблица 8.4

Составляющие дисперсии	***C***		***df***	***S²***	***F***
Наклон модельной линии	$C_{мод.} = \Sigma (Y_i - Y)^2$	154.08	1	$S^2_{мод.} = 154.08$	$F = \frac{154.08}{1.58} = 97.3$
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma (y_i - Y_{xi})^2$	7.92	5	$S^2_{остат.} = 1.58$	$F_{(0.05,1,5)} = 6.6$
Общая (всего)	$C_{общ.} = \Sigma (y_i - Y)^2$	162			

Полученная величина (9.84) значительно превышает табличную (2.57), что говорит о высокой статистической значимости первого коэффициента регрессии (a), о достоверности его отличия от нуля. Масса тела теленка действительно возрастает вслед за ростом массы тела коровы.

Рассчитаем доверительный интервал, в котором с той или иной вероятностью заключено теоретическое значение веса новорожденных. Примем уровень значимости $\alpha = 0.05$, тогда для числа степеней свободы $df = n–1 = 6$ критерий Стьюдента (нормированное отклонение) составит 2.45. Далее находим границы. Так, для значения $x = 352$ кг прогноз равен $Y = 25.56$, отклонение составит:

$$T \cdot m_Y = T \cdot m_y \cdot \sqrt{\frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}} = 2.45 \cdot 1.2582 \cdot \sqrt{\frac{1}{7} + \frac{(352-450)^2}{35658}} =$$

$$= 2.45 \cdot 0.81 = 1.98.$$

Отсюда находим границу доверительного интервала (табл. 8.3):

верхнюю: $\max Y = Y_i + T \cdot m_Y = 25.56 + 1.98 = 27.54$

и нижнюю: $\min Y = Y_i - T \cdot m_Y = 25.56 - 1.98 = 23.58.$

Для найденного значения доверительный интервал имеет границы 25.56 ±1.98 кг, или от 23.58 до 27.58 кг. Именно в этом весовом диапазоне с вероятностью $P = 0.95$ должен находиться средний вес новорожденных телят, рожденных от коров весом 352 кг.

Интервал прогноза рассчитывается аналогично. Так, для тех же значений $x = 352$ кг и $Y_{352} = 25.56$ кг отклонение составит:

$$T \cdot S_Y = T \cdot m_y \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_i - M_x)^2}{Cx}} = 2.45 \cdot 1.258 \cdot \sqrt{1 + \frac{1}{7} + \frac{(352-450)^2}{35658}} =$$

$$= 3.66.$$

Отсюда находим границы интервала прогноза:

верхнюю: $Y_i + T \cdot S_Y = 25.56 + 3.66 = 29.22$

и нижнюю: $Y_i - T \cdot S_Y = 25.56 - 3.66 = 21.89.$

Для найденного значения 25.56 кг зона прогноза имеет границы 25.56 ±3.66 кг, или от 21.89 до 29.22 кг. В таком диапазоне с вероятностью $P = 0.95$ должен находиться вес очередного новорожденного от коровы массой 352 кг. Доверительные интервалы и интервалы прогноза, рассчитанные для других значений, отображены на диаграмме (табл. 8.3, рис. 8.9). В пределах доверительной зоны с

вероятностью $P = 0.95$ находится генеральная (истинная) линия регрессии, в пределах зоны прогноза ожидаются новые значения вариант.

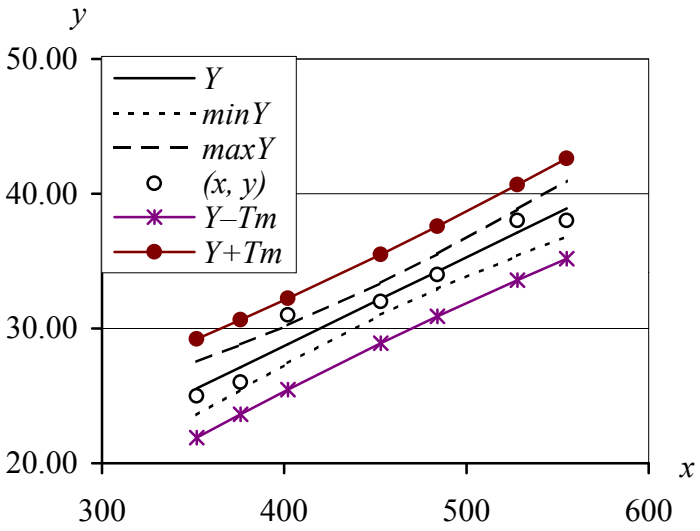


Рис. 8.9. *Линия регрессии, ее доверительный интервал и интервал прогноза для модели* $Y = 0.0657 \cdot x + 2.1347$

Итак, расчетное уравнение регрессии приняло вид: ($Y = a \cdot x + b$): $Y = 0.0657 \cdot x + 2.1347$. Однако анализ показал, что критерий Стьюдента для второго коэффициента (свободного члена уравнения) (2.13) оказался ниже табличного значения (2.57), т. е. коэффициент b значимо от нуля не отличается (на данном объеме собранных материалов). Это позволяет пересчитать коэффициент регрессии: $a = \Sigma(x \cdot y)/\Sigma x^2 = 0.071$.

Отсюда уравнение регрессии ($Y = a \cdot x$) будет иметь вид:

$Y = 0.071 \cdot x$.

Подставляя в него любые значения x, мы получим соответствующие теоретические (т. е. средние) значения Y и таким образом сможем построить на графике линию регрессии. Например, при массе тела коровы $x = 376$ кг масса теленка должна составить $Y = 0.071 \cdot 376 = 26.7$ кг, а при $x = 555$ $Y = 39.4$. Соединив на графике точки с этими координатами, получаем линию регрессии, весьма наглядно иллюстрирующую характер изучаемой связи (рис. 8.10).

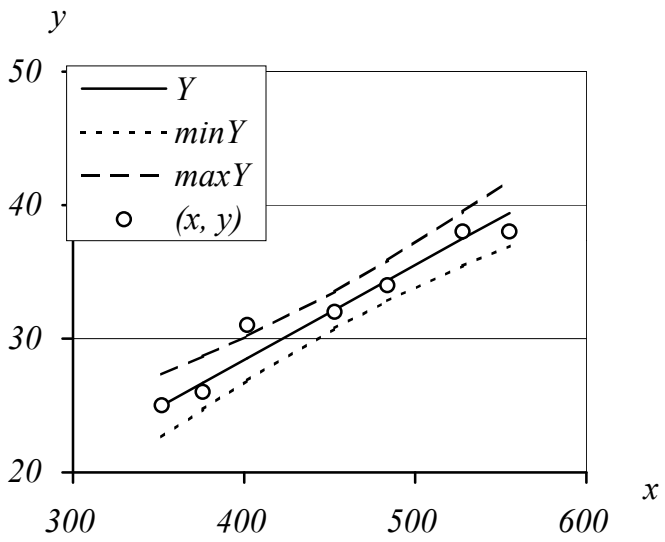


Рис. 8.10. *Линия регрессии, ее доверительный интервал и интервал прогноза для модели* $Y = 0.071 \cdot x$

В заключение оценим адекватность полученной модели исходным данным с помощью дисперсионного анализа. Для этого следовало бы вновь найти оценку остаточной суммы квадратов отклонений реальных значений от новых прогнозных, $\Sigma(y-Y)^2$, затем оценить регрессионную сумму квадратов, найти дисперсии и рассчитать критерий Фишера (кстати, число степеней свободы для остаточной дисперсии берется как $df = n-1 = 6$, раз в расчетах участвует только один коэффициент (a)). Для этой цели воспользуемся программой, встроенной в пакет Excel. Она вызывается командой меню Сервис\ Анализ данных\ Регрессия.

Дисперсионный анализ (табл. 8.5) показал, что расчетное значение ($F = 102.9$) выше табличного (6.0), т. е. регрессионная дисперсия существенно превышает остаточную, иначе говоря, исходные данные и модельные расчеты хорошо согласуются друг с другом, модель адекватна реальности. Коэффициент детерминации указывает, что «сила влияния» веса коров на вес плода очень велика:

$$R^2 = \frac{153.08}{162} = 0.945.$$

Таблица 8.5

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0.9720							
R-квадрат	0.9449							
Нормир R-квадрат	0.7782							
Стандартн	1.2193							
Наблюден	7							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*Значимость F*			
Регрессия	1	153.079	153.079	102.958	0.00015			
Остаток	6	8.92085	1.48680					
Итого	7	162						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95.0%*	*Верхние 95.0%*
Y-пересеч	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
x	0.0709	0.00101	70.1713	5.6E-10	0.06850	0.07345	0.06850	0.07345

В окне макроса нужно указать диапазоны ячеек, содержащих ряды значений изучаемых признаков (не перепутав *x* и *y*), желательно сразу с метками этих рядов (в этом случае нужно поставить галочку в окне Метки), ОК. Результаты будут выведены на новый автоматически созданный лист книги Excel. Помимо описательной статистики они содержат таблицу дисперсионного анализа, а также коэффициенты регрессии с их ошибками и оценкой статистической значимости по Стьюденту. Если при первом прогоне программы оказалось, что свободный член значимо от нуля не отличается, при втором прогоне макроса в окне Константа-ноль следует поставить галочку.

Выполнение регрессионного анализа с помощью пакета StatGraphics показано в следующем разделе.

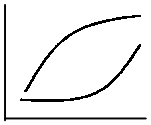
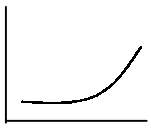
Криволинейная регрессия

Рассмотренный выше метод линейной регрессии позволяет описывать и прогнозировать явления и процессы, при которых зависимость между изучаемыми признаками приближается к линейной, простой пропорции. Таковы, например, зависимость веса сердца разных видов млекопитающих от массы их тела или экстраполяция данных о численности популяции, полученных на пробных площадях, на всю занимаемую ею территорию и т. п. Однако в большинстве случаев связь биологических признаков не бывает линейной и они изменяются с разной скоростью (и в разных масштабах). Соответственно на графике форма такой связи отображается не прямой, а кривой линией. Примерами могут служить геометрическая прогрессия роста численности популяции в оптимальных условиях, характерное для теплокровных животных изменение метаболизма – невысокий уровень в оптимуме, ускоренно возрастающий при смене условий, рост числа видов, попавших в описание, по мере увеличения площади обследованной территории, а также различие скоростей роста разных частей тела, определяющее аллометрический характер зависимости признаков. Так, увеличение массы тела опережает по темпам весовой рост сердца и других внутренних органов, лицевой отдел черепа растет более интенсивно, чем мозговой; с разной скоростью растут листья на одном и том же побеге.

В подобных случаях использование уравнения прямой линии ($y = ax + b$) неэффективно: теряются многие детали процесса, коэффициенты корреляции и регрессии получаются заниженными, а результаты анализа – приблизительными, недостаточно точными. Решить эту проблему можно с помощью уравнений кривых линий. В практике биологических исследований в число наиболее часто используемых входят следующие пять видов криволинейной зависимости (табл. 8.6).

Существуют два достаточно простых пути подгонки уравнений под конкретные данные (аппроксимации данных – кривой), два способа оценки коэффициентов в уравнениях кривых – это настройка параметров модели с помощью макроса «Поиск решения» (этот путь рассмотрен в разделе **Имитационное моделирование**) и расчет коэффициентов методом наименьших квадратов.

Таблица 8.6

Название зависимости	Уравнение	График
Степенная (аллометрическая) (multiplicative)	$y = bx^a$	
Гиперболическая (reciprocal)	$y = \frac{a}{x} + b$	
Показательная (экспоненциальная, exponencial)	$y = be^{ax}$ и $y = b^{ax}$	
Параболическая (polynomial)	$y = c+bx+ax^2$	
Логистическая (кривая Ферхюльста) (logistic)	$y = \frac{A}{1+10^{ax+b}} + C$	

Поскольку метод наименьших квадратов исходно ориентирован на линию (поиск уравнения линии, наименее удаленной ото всех эмпирических точек), прямой расчет уравнений кривых в рамках регрессионного анализа невозможен. Натурные данные необходимо предварительно «выпрямить», т. е. сделать возможным вычисление *линейного уравнения регрессии* с тем, чтобы потом из него получить уравнение криволинейной связи. Общий порядок регрессионного анализа для криволинейной зависимости следующий:

- преобразование исходных данных, «выпрямляющее» зависимость,
- расчет коэффициентов линейной регрессии преобразованных данных,
- проведение дисперсионного анализа, оценка значимости коэффициентов регрессии,

– обратное преобразование коэффициентов линейной регрессии для конструирования уравнения криволинейной регрессии.

Рассмотрим процесс поиска уравнения криволинейной регрессии на примере изучения зависимости веса печени прыткой ящерицы от длины ее тела (рис. 8.11).

Рассчитанное по исходным данным уравнение линейной регрессии имеет вид:

$y = 107.9x - 404.2$.

И хотя коэффициент регрессии достоверен ($T = 7.6$, $\alpha < 0.05$) и коэффициент детерминации высок $R^2 = 0.866$, это уравнение весьма приблизительно описывает зависимость признаков – для наименьших наблюдаемых значений длины тела оно дает абсурдное (отрицательное) значение массы печени ($107.9 \cdot 3.4 - 404.2 = -37.3$ мг). Итак, линейная модель не годится даже для интерполяции изучаемых данных. Гораздо успешнее справляется с подобной задачей степенная (аллометрическая) функция $y = bx^a$.

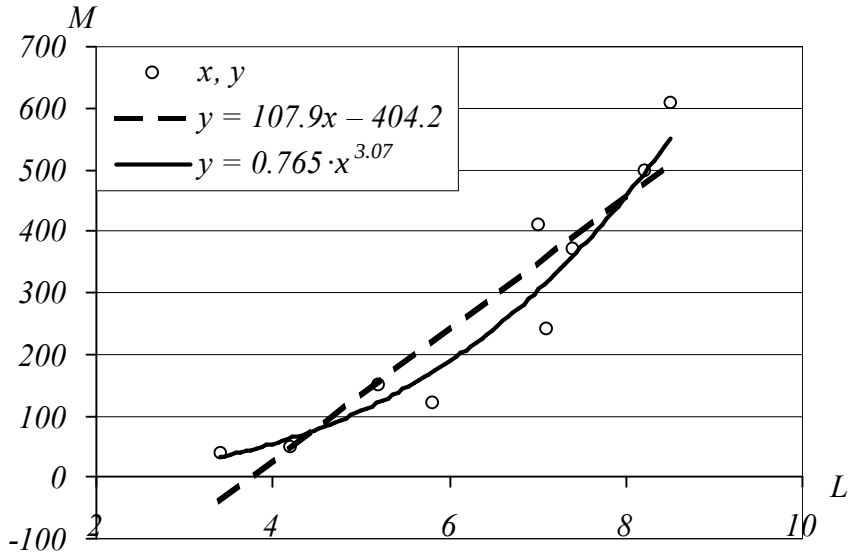


Рис. 8.11. *Зависимость веса печени (M, мг) от длины тела (L, мм) у ящериц*

Для вычисления коэффициентов этого уравнения воспользуемся преобразованием: $Y = \lg y$, $X = \lg x$, $B = \lg b$. После логарифмирования степенное уравнение приняло линейный вид: $\lg y = \lg b + a \cdot \lg x$ или $Y = B + aX$. Теперь остается отыскать коэффициенты уравнения B и a, используя алгоритм метода наименьших квадратов (табл. 8.7).

Таблица 8.7

№	x	y	$X = \lg x$	$Y = \lg y$	X^2	Y^2	$X \cdot Y$	Y'	$(Y'-Y)^2$	y'
1	3.4	40	0.531	1.60	0.282	2.567	0.85	1.517	0.00718	33
2	4.2	50	0.623	1.69	0.388	2.886	1.06	1.799	0.01009	63
3	5.2	150	0.716	2.18	0.513	4.735	1.56	2.085	0.00838	121
4	5.8	120	0.763	2.08	0.583	4.323	1.58	2.23	0.02284	170
5	7.1	240	0.851	2.38	0.725	5.665	2.03	2.5	0.01442	316
6	7.0	410	0.845	2.61	0.714	6.827	2.21	2.481	0.01728	303
7	7.4	370	0.869	2.57	0.756	6.596	2.23	2.556	0.00016	359
8	8.2	500	0.914	2.69	0.835	7.284	2.47	2.693	0.00004	493
9	8.5	610	0.929	2.78	0.864	7.758	2.59	2.741	0.00201	550
Σ	56.8	2490	7.043	20.6	5.66	48.64	16.6		0.08239	

Для преобразования исходных данных ($Y = \lg y$, $X = \lg x$) можно воспользоваться функцией =log10(ячейка) среды Excel.

Далее рассчитаем суммы, необходимые промежуточные значения и коэффициенты (приведены округленные значения с листа Excel): $\Sigma Y = \Sigma \lg y = 20.6$, $\Sigma Y^2 = \Sigma(\lg y)^2 = 48.64$, $\Sigma X = \Sigma \lg x = 7.043$,

$\Sigma X^2 = \Sigma(\lg x)^2 = 5.659$, $\Sigma XY = \Sigma(\lg x \cdot \lg y) = 16.577$,

$M_Y = \Sigma Y/n = 20.6/9 = 2.289$, $M_X = \Sigma X/n = 7.043/9 = 0.7826$,

$C_{XY} = \Sigma XY - (\Sigma X)\cdot(\Sigma Y)/n = 16.572 - 7.043 \cdot 20.602/9 = 0.45542$,

$C_X = \Sigma X^2 - (\Sigma X)^2/n = 5.655 - (7.04)^2/9 = 0.14816$,

$C_Y = \Sigma Y^2 - (\Sigma Y)^2/n = 48.638 - (20.601)^2/9 = 1.4823$,

$$S_Y = \sqrt{C_Y/(n-1)} = \sqrt{1.4823/8} = 0.4305,$$

$$S_X = \sqrt{C_X/(n-1)} = \sqrt{0.14816/8} = 0.1361,$$

$$r = C_{XY} \cdot \sqrt{C_X \cdot C_Y} = 0.45542 \cdot \sqrt{0.14816 \cdot 1.34823} = 0.9718,$$

$a = C_{XY}/C_X = 0.45541/0.14815 = 3.0739$,

$B = M_Y - aM_X = 2.289 - 3.0739 \cdot 0.7826 = -0.11643$.

Линейное уравнение для преобразованных данных имеет вид:

$\lg y = 3.07 \cdot \lg x - 0.116$ или $Y' = 3.07 \cdot X - 0.116$.

Оно дает возможность рассчитать теоретические значения признака Y' (теоретические значения логарифмов массы печени), а также квадраты отклонений прогнозных значений от реальных: $(Y'–Y)^2$ и их сумму $\Sigma(Y''–Y)^2 = 0.08239$.

Эта величина есть остаточная сумма квадратов; вместе с общей суммой квадратов $Cy = C_{общ.} = 1.4823$ она позволяет сформировать таблицу дисперсионного анализа (табл. 8.8):

$$C_{мод.} = C_{общ.} – C_{остат.} = 1.4823 – 0.08239 = 1.39993.$$

Таблица 8.8

Составляющие дисперсии	***C***		***df***	S^2	***F***
Наклон модельной линии	$C_{мод.} = \Sigma (Y'_i – M_Y)^2$	1.39993	1	$S^2_{мод.} =$ 0.39993	$F = \frac{1.39993}{0.01177} =$ = 118.9377
Отклонения вариант от линии регрессии	$C_{остат.} = \Sigma (y_i – Y'_i)^2$	0.08239	6	$S^2_{остат.} =$ 0.01177	$F_{(0.05,1,7)} =$ = 5.6
Общая (всего)	$C_{общ.} = \Sigma (y_i – M_Y)^2$	1.482322			

Полученное значение $F = 118$ больше табличного (5.6), следовательно, дисперсия, обусловленная регрессией, достоверно больше случайной, т. е. признак Y действительно зависит от признака X, и линия регрессии адекватна исходным данным. Коэффициент детерминации больше, чем у линейной регрессии, и составляет: $R^2 = C_{мод.}/C_{общ.} = 1.39993/1.482322 = 0.944417$.

Ошибка коэффициента криволинейной регрессии равна:

$$m_a = \frac{S_y}{S_x} \cdot \sqrt{\frac{1-r^2}{n-2}} = \frac{0.430}{0.136} \cdot \sqrt{\frac{1-0.9718^2}{9-2}} = 0.281,$$

а критерий Стьюдента, проверяющий гипотезу Но: $a = 0$, составляет:

$$T = a/\, m_a = 3.0739/0.281 = 10.9.$$

Полученное значение больше табличного ($T_{(0.05,8)}$ = 2.31 для уровня значимости α = 0.05 и числа степеней свободы $df = n–2 = 8$) ($\alpha<<0.05$), зависимость признака Y от X есть, причем очень тесная. Следует помнить, что при расчете ошибки коэффициента криволинейной регрессии используются стандартные отклонения для преобразованных (у нас – прологарифмированных) значений признаков.

В завершение выполним обратное преобразование второго коэффициента регрессии, свободный член равен:

$b = 10^B = 10^{-0.11643} = 0.764839$.

Теперь уравнение регрессии принимает вид степенной зависимости:

$y' = 0.765 \cdot x^{3.07}$.

Теоретические значения y', рассчитанные по этому уравнению, гораздо ближе к исходным данным, что хорошо видно и на графике (рис. 8.11), и по большей величине коэффициента детерминации (0.94>0.87) (читателю несложно будет проделать все вычисления в среде Excel с помощью программы Регрессия – как для исходных, так и для преобразованных данных).

Аллометрическое уравнение ($y' = 0.77x^{3.1}$) не только лучше описывает зависимость между сравниваемыми признаками в статистическом плане, но и придает ей более ясный биологический смысл (масса печени = 0.77·длина тела$^{3.1}$). Как известно, объемные величины (объем, масса тела) пропорциональны кубу линейных промеров (длина тела). В свою очередь, вес печени и вес тела связаны прямой пропорциональной зависимостью. Так становится понятной наблюдаемая пропорциональность веса печени кубу длины тела.

Когда зависимость между изучаемыми признаками имеет иную форму, чем может описать степенное уравнение, пользуются другими способами преобразования данных (табл. 8.9).

Выбрать, какой из видов описания лучше подходит к эмпирическим данным, можно, ориентируясь на величину коэффициента детерминации или корреляции. Чем ближе линия проходит к эмпирическим точкам, тем меньше остаточная сумма квадратов, тем больше коэффициент детерминации. Существуют и другие уравнения для описания криволинейных зависимостей (например, очень интересна парабола).

Таблица 8.9

Название уравнения зависимости	**Линейный вид криволинейной зависимости** $Y = B + AX$	**Необходимое преобразование исходных значений переменных** x, y	**Обратное преобразование коэффициентов**
Степенное $y = bx^a$	$\lg y = \lg b + a \cdot \lg x$	$Y = \lg y$, $X = \lg x$	$b = 10^B$
Гипербола $y = \frac{a}{x} + b$	$y = aX + b$	$X = 1/x$	–
Показательное $y = be^{ax}$ или $y = b^{ax}$	$\lg y = \lg b + \lg a \cdot x$	$Y = \lg y$	$a = 10^A$, $b = 10^B$
Логистическая кривая $y = \frac{A}{1+10^{ax+b}} + C$	$\lg\left(\frac{A}{y-C} - 1\right) = ax + b$	$Y = \lg\left(\frac{A}{y-C} - 1\right)$	–

Самый простой способ расчета уравнений регрессии в среде Excel реализуется программой Добавить линию тренда. Для того чтобы построить линию и рассчитать уравнение регрессии между двумя *столбцами* данных (x и y), следует сначала построить точечную диаграмму (чтобы получить зависимость $y = f(x)$, столбец x должен быть первым, y – вторым). На построенной диаграмме должны присутствовать точки только одного цвета, наличие точек двух цветов говорит о том, что диаграмма построена неверно.

Далее нужно один раз щелкнуть мышкой по какой-либо точке (x, y) диаграммы. При этом точки ряда окрасятся другим цветом, а в главном меню появится новый пункт Диаграмма (справа от Сервис). Он позволяет построить линию регрессии с помощью команды Диаграмма\ Добавить линию тренда…

В открывшемся окне (вкладка Тип) будет предложено на выбор пять моделей (линейная, логарифмическая, полиномиальная, степенная, экспоненциальная) и сглаживание по средним, с помощью которых можно дать обобщенное описание данных. На вкладке Параметры следует поставить галочку, как минимум, в одном поле – Показывать уравнение на диаграмме, ОК. На диаграмме появится черная жирная линия регрессии. Изменить установки можно в окне настройки, которое появляется после двойного клика по линии.

Когда исходные данные содержат нулевые значения, их преобразование (логарифмирование) для «выпрямления» зависимости становится невозможным; в этом случае на вкладке Тип будут высвечиваться не все виды уравнений криволинейной регрессии. Ситуацию удается исправить, если нули исключить из рассмотрения или заменить правдоподобными малыми числами, следя за тем, чтобы основную роль в расчете уравнения играли реальные значения.

Регрессионный анализ в среде StatGraphics

Обширный список криволинейных функций предлагает пакет StatGraphics. Для выбора лучшего уравнения организуется таблица, сравнивающая результаты разных способов аппроксимации.

Сначала необходимо ввести данные на лист StatGraphics (один из простейших способов – простое копирование данных с листа Excel через буфер обмена). Для расчета разных видов парной регрессии нужно дать команду меню Relate\ Simple Regression…, выбрать переменные, выбрать все позиции Tabular options и Graphics options. Исходно в появившихся окнах будет рассчитана линейная регрессия. В окошке Comparison of Alternative Models будут отображены результаты 12 способов аппроксимаций, ранжированных по величине коэффициентов детерминации. Увидеть результаты расчетов для других видов уравнений можно, щелкнув правой кнопкой мыши и выбрав в окне Analysis Options нужный вид модели (Type of Model). Найти уравнение полиномиальной зависимости (параболы) можно, дав команду Relate\ Polynomial Regression… Рассчитать линейную регрессию без свободного члена позволяет команда Relate\ Multiple Regression…, если в окне контекстного меню Analysis Options убрать галочку из рубрики Constant in Model.

Корреляционный анализ

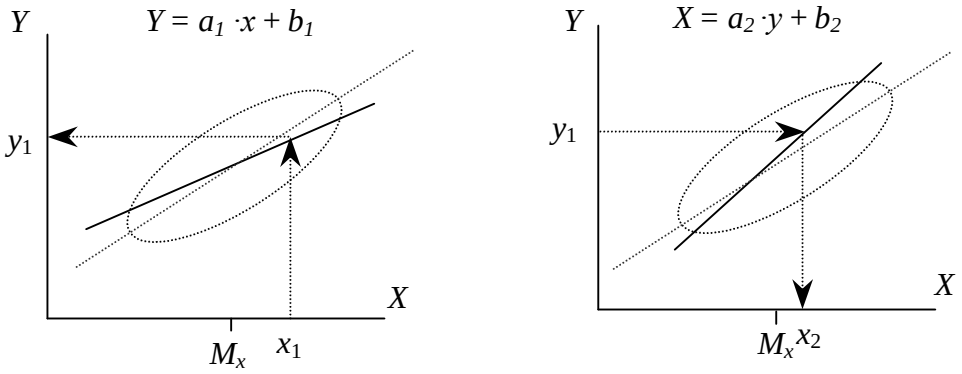
Взаимная связь (взаимная зависимость) двух признаков при их изменчивости, т. е. сопряженность их вариации, называется корреляцией. Корреляция имеет место в тех случаях, когда признаки изменяются не автономно, а согласованно. Если с увеличением одного признака происходит соответствующее увеличение другого, говорят о положительной корреляции, и коэффициент корреляции имеет в этом случае положительный знак (+). Если же по мере увеличения первого признака второй уменьшается, то это отрицательная корреляция, и коэффициент корреляции пишется со знаком минус (–).

Полная положительная корреляция выражается единицей $r = 1$, полная отрицательная $r = -1$. В природе такая ситуация встречается редко, и степень связи выражается той или иной долей единицы. При этом о тесной (сильной) корреляции обычно говорят в тех случаях, когда коэффициент корреляции не ниже ±0.6; значения ниже ±0.6 указывают на среднюю связь, а ниже ±0.3 – на слабую.

Логико-теоретические основы

Рассмотренный выше регрессионный анализ изучает изменение среднего уровня одного признака при изменении другого, т. е. ориентирован асимметрично на один из признаков. Однако по любому массиву значений двух сопряженных признаков (x и y) можно рассчитать два уравнения регрессии и построить две линии регрессии зависимости y от x и зависимости x от y:

$$Y = a_1 \cdot x + b_1,\ X = a_2 \cdot y + b_2.$$



При этом не только уравнения содержат разные коэффициенты пропорциональности, но и линии регрессии не совпадают, как и прогнозы по ним ($x_1 > x_2$). Как указывалось выше, причина того, что линии регрессии не совпадают в осью эллипса рассеяния, а значит, и друг с другом, состоит в том, что случайная изменчивость признаков не дает точно определить коэффициенты пропорциональности (регрессии) и, следовательно, точно охарактеризовать взаимозависимое изменение обоих признаков.

В то же время по графикам видно, что каждый коэффициент регрессии неточен по-своему, в результате чего линии регрессии лежат по разные стороны оси эллипса. Возникает вопрос, нельзя ли вычислить некий усредненный показатель взаимосвязи, в котором свойства коэффициентов регрессии обобщаются? Такой характеристикой (средней геометрической) для линейной зависимости выступает коэффициент корреляции:

$$r = \sqrt{a_1 \cdot a_2}\ .$$

Корреляционный анализ, состоящий в расчете и оценке значимости коэффициента корреляции, держит в поле зрения в равной мере оба изучаемых признака – как их сопряженную, так и общую изменчивость. Коэффициент корреляции призван численно выражать долю сопряженной вариации двух признаков в общей их вариации:

$$r = \sqrt{\frac{ковариация}{изменчивость}} = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum (y - M_y)(x - M_x)}{\sqrt{\sum (y - M_y) \cdot \sum (x - M_x)}},$$

где Cxy – характеристика сопряженной изменчивости признаков,
Cx, Cy – характеристика общей изменчивости признаков.

Рабочая формула для расчетов имеет вид:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{\sum xy - (\sum x \cdot \sum y)/n}{\sqrt{(\sum x^2 - (\sum x)^2/n) \cdot (\sum y^2 - (\sum y)^2/n)}}.$$

Когда степень сопряженной изменчивости признаков велика, коэффициент корреляции имеет большую величину, вплоть до $r = \pm 1$ – при функциональной зависимости. Если признаки варьируют независимо друг от друга и сопряженная изменчивость отсутствует, выборочный коэффициент корреляции приближается к нулю, хотя практически никогда не имеет арифметических нулевых значений. В любом случае для доказательства существования зависимости

между признаками необходимо проверить статистическую гипотезу Но: «коэффициент корреляции значимо от нуля не отличается», $r = 0$, т. е. «в генеральной совокупности изучаемые признаки не зависят друг от друга». Значимость отличия коэффициента корреляции от нуля оценивается с помощью критерия Стьюдента:

$$T = (r–0)/\, m_r = r/\, m_r \sim T_{(0.05,\, n–2)},$$

где $m_r = \sqrt{\frac{1-r^2}{n-2}}$.

Из приведенной формулы следует, что ошибка репрезентативности выборочного коэффициента корреляции определяется только объемом выборки и величиной самого показателя. Это позволяет предложить таблицу значимых коэффициентов корреляции (табл. 16*П)*, в которой приведены минимальные *значимые* (достоверно отличные от нуля) коэффициенты корреляции при разных объемах выборок. Если коэффициенты корреляции выше табличных, то они также значимы, если ниже, то от нуля отличаются недостоверно.

Как статистический параметр, выборный коэффициент корреляции в той или иной степени соответствует генеральному параметру. Определить диапазон возможных значений генерального коэффициента корреляции можно с помощью доверительного интервала, хотя его *нельзя* построить непосредственно как для других выборочных параметров: $r \pm T_{(\alpha,df)} \cdot m_r$. Дело в том, что область изменений коэффициента ограничена рамками ±1, поэтому распределение выборочных коэффициентов корреляции в общем не соответствует нормальному (для него нужен диапазон изменчивости ±∞). Поэтому перед расчетом коэффициент корреляции переводят в величину, имеющую нормальное распределение по формуле: $z = 0.5 \cdot \ln\left(\frac{1+r}{1-r}\right)$ (или по табл. 14*П*, знак сохраняется), затем вычисляют ошибку коэффициентов: $m_z = \sqrt{\frac{1}{n-3}}$. Теперь доверительный интервал принимает вид: $z \pm T_{(\alpha,df)} \cdot m_z$. Далее отыскиваются границы интервала:

верхняя: $_{\max}z = z + T_{(\alpha,df)} \cdot m_z$

и нижняя: $_{\min}z = z - T_{(\alpha,df)} \cdot m_z$.

После этого значения $_{\max}z$ и $_{\max}z$ с помощью таблицы 15*П* переводят-

ся обратно, в прежние единицы $_{max}r$, $_{min}r$; это и будут границы доверительного интервала для генерального значения коэффициента корреляции.

Биологическая интерпретация коэффициента корреляции

Понятие «корреляция» имеет длительную историю использования в биологии. Важно различать два понимания этого термина – статистическое и биологическое. Корреляционный анализ как статистический метод призван лишь установить факт сопряженного варьирования двух величин. Он ничего не сообщает о каузальной обусловленности изменения одного признака при изменении другого. Причинно-следственный характер этих объективных отношений устанавливает биолог. Можно говорить о трех классах биологической корреляции – это влияние, взаимовлияние и «наведение».

Влияние – это тот случай, когда величина одного признака действительно определяется величиной другого. Число видов и численность животных зависят от благоприятных экологических условий – климата, обеспечения кормами. Например, в Карелии продолжительность безморозного периода снижается к северу, что позволяет размножаться живородящим видам почти на всей территории республики, а яйцекладущим – только в южной части; число видов рептилий увеличивается к югу. Для исследования влияний корреляционный анализ очень удобен; изучение криволинейной зависимости требует предварительного «исправления» данных.

Говоря о *взаимовлиянии*, подразумевают прямую и обратную связь между переменными: один признак зависит от другого, изменение которого, в свою очередь, сопряжено с первым. Самые яркие примеры этого – физиологические реакции организма и экологические отношения, например, между популяциями паразита и его хозяина. Естественный рост численности хозяина непосредственно обеспечивает рост численности паразита, который, в свою очередь, может негативно сказываться на состоянии особей хозяина, вызывая их преждевременный выход из процесса размножения и смерть, т. е. приводить к снижению численности популяции хозяина. Обратная связь – это и есть взаимовлияние. Исследовать такие отношения с помощью корреляционного анализа неэффективно, поскольку один коэффициент не в состоянии учесть двойственную природу явления.

Обратные связи наиболее эффективно можно исследовать с помощью динамических имитационных моделей (см. раздел 10).

Если величина обоих изучаемых признаков определяется внешней причиной, «*наводится*» ею извне, то между признаками можно обнаружить корреляцию в силу синхронности их реакций на этот фактор. Так, в годы роста численности рыжей полевки увеличивается и численность обыкновенной бурозубки, в другие (неблагоприятные) годы наблюдается депрессия обоих видов. Корреляция между этими показателями отражает вовсе не симбионтные (ценотические) отношения видов, но их сходную реакцию на одинаковые условия среды, не взаимное влияние видов друг на друга, а сходство видовых потребностей, причем опосредованно, – через реакцию на среду. В онтогенезе особи наблюдаются аналогичные отношения между признаками, связанными со степенью развития эмбриона. Оба признака выступают по отношению друг к другу индикаторами действия третьей силы. В этом случае корреляционный анализ также уместен.

В природе обычно наблюдается более сложная картина – величина изучаемых переменных определяется не только их связью друг с другом, но и одновременным влиянием внешних факторов. Например, развитие органов особи в онтогенезе зависит как от соседних органов (морфогенетические корреляции), так и от организма в целом (геномные, эргонтические корреляции); численность видов в ценозе определяется и общими (абиотическими, биокосными) условиями жизни в данных местообитаниях (зонах), и обилием других сочленов сообщества (объектов питания, конкурентов, хищников); токсичность стоков-загрязнителей зависит не только от их объема, происхождения, типа природной воды, но и от взаимодействия (антагонизм, синергизм) их компонентов. В процессе интерпретации биологических корреляций приходится декомпозировать сложные случаи, явно выделять направления функциональной («влияние») и косвенной («наведение») связи. Для этого следует, во-первых, контролировать (или хотя бы регистрировать) условия наблюдения и эксперимента. Во-вторых, важно осознанно формировать выборку для анализа, исходя из цели исследования, а не из имеющихся данных. В-третьих, распознать причины наблюдаемых корреляций можно, применив «сильные» статистические методы, такие как частная корреляция и компонентный анализ.

Направление изменчивости

Термин «направление изменчивости» характеризует способ формирования выборки для изучения зависимости между признаками. Во многом именно этот способ определяет, в какой мере объекты будут отличаться друг от друга по серии признаков, а значит, и степень коррелированности этих признаков. Обычно при исследовании зависимости биологических признаков их изменчивость не учитывается специально. Справедливо считается, что свойство «случайно варьировать» и свойство «сопряженно варьировать» (коррелировать) – разные свойства: если признаки не зависят друг от друга, то сколько не увеличивай их изменчивость, корреляции не добиться. При этом упускается из вида, что если признаки все же объективно взаимосвязаны, то выборочная мера связи, коэффициент корреляции, будет очень чувствителен к степени разнородности вариант в изучаемой выборке. Опыт свидетельствует: чем более однотипны объекты в выборке, тем ниже корреляция между их признаками (и даже случается смена знака коэффициента корреляции), но чем сильнее объекты отличаются, тем корреляция выше.

Тем не менее для исследования корреляций часто выдвигается требование «единообразия» вариант, например, чтобы особи в выборках были «одновозрастными». Тогда коэффициент корреляции принимают за оценку биологических взаимозависимостей, характерных для объектов данного типа. Этот подход как будто бы позволяет сопоставлять коэффициенты, полученные для разных групп. Так возникают похожие выводы: «скоррелированность признаков растущего листа... в среднем значительно выше, чем у листа закончившего свой рост» или «у бобров старшей возрастной группы... наблюдается ослабление значительного числа связей». Если бы авторы обратили внимание на принципиальное отличие критерия «одновозрастные» для молодых и старых организмов, то их выводы могли оказаться иными. Дело в том, что выборки, составленные с обычной методической погрешностью в определении числа прожитых дней (месяцев, лет) особей разного возраста, будут представлять различные по длительности отрезки онтогенеза. За те же 10 дней, когда старый лист никак не изменится, молодой вырастет на 30 %. В течение полугода взрослый бобр наберет лишь 5% «размера тела», а молодой – 70%. Выборки, составленные из «методически»

одновозрастных молодых особей, фактически будут представлять разновозрастных особей (по масштабу ростовых процессов). Выборки же взрослых, действительно, будут однородны. В первом случае облако рассеяния в пространстве признаков примет форму сильно вытянутого эллипса, во втором – близкого к окружности. Понятно, что и корреляции между признаками в группе молодых должны быть много выше, чем в группе старых. Однако вряд ли можно на этом основании делать вывод такого рода: «в ходе онтогенеза... имеет место частичная дезинтеграция», т. е. принимать особенность выборок за биологическое свойство. Аналогичные проблемы могут возникать в тех случаях, когда по уровню коррелированности сравниваются выборки объектов из дикой природы и с плантаций, из лаборатории: изменчивость (а значит, и показатель коррелированности) природных объектов всегда выше, чем у контролируемых человеком.

Помимо рассмотренного приема предлагается так подбирать выборку, «чтобы индивидуальные различия были как можно бóльшими». Но он также не лишен недостатков, поскольку при резком отличии значений вариант коэффициенты корреляции приближаются к единице, ничего не сообщая исследователю о специфике взаимоотношений разных признаков.

Видимо, полная унификация правил составления сравниваемых выборок никогда не может быть достигнута. Единственным средством формирования адекватных выводов может быть специальный учет условий, при которых данные корреляции были получены. Для характеристики этих условий мы предлагаем термин «направление изменчивости», который явно указывает на источник возникновения разнокачественных объектов. Рассмотрим основные причины появления различных значений случайных величин.

Исходной иллюстрацией является условный математический пример, когда случайная изменчивость одной переменной не сказывается на (случайной) изменчивости другой. Корреляция между переменными близка к нулю, ее направленность не выражена. Двумерное распределение имеет форму окружности, а не эллипса, не ориентировано.

В остальных случаях можно отметить три основных направления изменчивости в выборке, связанные с отличиями объектов *во времени* (онтогенез, этап, стадия), *в пространстве* (расстояние, уда-

ление, условия) и *по статусу* (габитус, зрелость, качество).

Пусть выборка составлена из ряда пар значений, полученных при наблюдении процесса через некоторые (равные или неравные) промежутки времени, как, например, серия все увеличивающихся значений размеров разных частей особей (длина и ширина листовой пластинки растения) в онтогенезе. Коэффициент корреляции будет отражать здесь связь динамики признаков во времени, т. е. *временно́е* (динамическое) направление. На графике двумерного распределения объекты (лист в отмеченный день наблюдений) будут ориентированы вдоль оси этого направления – от объектов меньшего размера (младшие) – к крупным объектам (старшие): наименьшие размеры пластинка имеет на ранних стадиях роста, наибольшие – на последних стадиях. Если весь период онтогенеза листа разбить на две равные части (начальную и заключительную), то корреляция между промерами на первом отрезке времени будет больше, чем на втором.

Второй случай – это изучение пространственного распределения объектов и оценка связи их признаков. Например, с севера на юг, от района к району Карелии продолжительность морозного периода уменьшается, а сумма летних температур параллельно увеличивается. Корреляция велика и достоверна: $r = –0.85$. Интерпретация связи признаков должна учесть эффект неоднородности факторов среды (условия инсоляции на разных широтах), т. е. *пространственную* (географическую, факториальную) направленность связи признаков. Важно отметить, что если выборку ограничить лишь пятью северными районами, то коррелированность между факторами среды (вместе с изменчивостью) уменьшится.

Часто формируется выборка объектов разного *статуса*, когда своеобразие их «внутреннего» качества нельзя явно связать с каким-либо отличием в пространственном размещении или стадии развития. Для организмов – это различия по полу, степени зрелости, заболеванию, генотипу; для популяций, ценозов – по области распространения, параметрам структуры, стадии сукцессии, для экосистем – по типу трофности, деградации и т. п. Например, на прибайкальской равнине антропогенная трансформация коренных кедровых лесов привела к возникновению серии вторичных биотопов. Для них выявлена высокая корреляция численности двух таежных обитателей – азиатской лесной мыши и красной полевки. Это свиде-

тельствует о резком различии условий обитания в разных ценозах: животные обоих видов предпочитают хвойные леса и избегают открытых стаций. Налицо *экологическая* направленность корреляции – от биотопов, не подходящих для мышей и полевок, – к благоприятным. Еще один пример демонстрирует статику (!) развития беременности, зафиксированную в выборке перезимовавших самок красной полевки. В связи с интенсификацией экскреторной и регуляторной функций в период развития плода масса печени и надпочечников параллельно увеличивается, достигая на поздних стадиях максимального развития. *Физиологическая* (не динамическая!) направленность корреляции очевидна.

Чаще всего, конечно, встречается *смешанный* случай, когда о статусе, а также о пространственном и временно́м распределении объектов мало что известно. Это наименее информативная для эколога выборка, ибо причины зависимости признаков оказываются скрытыми. Так, в случайной выборке животных из природы всегда можно обнаружить и крупных и мелких, что проявляется в ощутимой корреляции размеров и массы. В то же время эта корреляция не будет «истинной», «видовой» характеристикой, поскольку объединяет несколько направлений сопряженного изменения признаков – и онтогенетическое (особи разного реального возраста есть и в «одновозрастной» природной выборке – в пределах точности метода возрастной диагностики), и физиологические (особи разного пола или зрелости, объединенные из-за неточности определения), и генетические (индивидуальные, популяционные, расовые особенности), и экологические (отличие условий роста и жизни в разных местообитаниях, в разные годы). В любой выборке, имеющей близкое к эллипсу двумерное распределение, можно выделить несколько направлений коррелятивной изменчивости.

Разобраться в пересечении разных направлений изменчивости признаков, выяснить причины их сопряженного изменения можно, если специально разрабатывать метод сбора биологических данных, стремясь учесть все источники варьирования. К этому нужно подходить ответственно.

Термин «направление изменчивости», или «направление корреляции», заставляет рассматривать коэффициент корреляции как не абсолютную, а контекстуальную характеристику связи признаков, проявившуюся именно в данной совокупности.

Техника расчета линейного коэффициента корреляции

Часто ее наличие пытаются оценить на глаз с помощью графиков. Однако даже если и удается определить сам факт корреляционной взаимосвязи, то степень ее остается неизвестной. Корреляционный анализ призван количественно выразить связь и определить ее достоверность.

Конструкция коэффициента корреляции в своей основе имеет линейную математическую модель (метод наименьших квадратов). Поэтому единичное значение коэффициент корреляции принимает лишь тогда, когда все точки графика зависимости переменных лежат на одной прямой линии. Во всех остальных случаях он будет отличаться от единицы.

Способ вычисления коэффициента корреляции показан на примере исследования зависимости между живым весом коров и их приплода (кг) (табл. 8.3, стр.175). Рассчитываются квадраты вариант и их произведения, а также суммы значений, их квадратов, произведений, другие вспомогательные величины:

$Cxy = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y)/n = 103144 - 3150 \cdot 224/7 = 2344$

$Cy = \Sigma y^2 - (\Sigma y)^2/n = 7330 - 224^2/7 = 162,$

$Cx = \Sigma x^2 - (\Sigma x)^2/n = 1453158 - 3150^2/7 = 35658.$

Затем вычисляется коэффициент корреляции:

$$r = \frac{Cxy}{\sqrt{Cx \cdot Cy}} = \frac{2344}{\sqrt{35658 \cdot 162}} = 0.975,$$

его ошибка:

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.975^2}{7 - 2}} = 0.099$$

и критерий Стьюдента, проверяющий нулевую гипотезу Но: «коэффициент корреляции достоверно от нуля не отличается», $r = 0$.

$T_r = r/m_r = 0.975/0.099 = 9.84.$

То, что эта величина значительно превышает табличную (для уровня значимости $\alpha = 0.05$ и числе степеней свободы $df = n - 2 = 5$ $T_{(0.05,5)} = 2.57$), говорит о высокой статистической значимости полученного коэффициента корреляции.

По таблице 6*П* определяется уровень значимости коэффициента корреляции. Полученное значение критерия $T_r = 9.84$ превышает порог даже для уровня значимости $\alpha = 0.001$, т. е. шанс ошибоч-

ного заключения даже ниже 1 на 1000, иначе вероятность справедливости заключения очень высока, $P>0.999$.

Оценить достоверность отличия коэффициента корреляции от нуля можно, не прибегая к вычислению ошибки и критерия Стьюдента. Для этого служит специальная таблица 16*П*, в которой указаны минимальные значимые значения коэффициента корреляции при разных объемах выборок и уровне значимости. Чтобы полученный коэффициент корреляции можно было считать достоверным, он должен превышать табличное значение при данном n. В нашем случае ($n = 7$, $\alpha = 0.05$) достоверно уже значение $r = 0.666$, полученный коэффициент корреляции ($r = 0.975$) превышает табличное, следовательно, также значим.

Доверительный интервал для нашего случая ($r = 0.975$, $\alpha = 0.05$, $n = 7$, $df = n–2 = 5$, $T_{(0.05,5)} = 2.57$) рассчитывается так. Преобразуем r:

$$z = 0.5 \cdot \ln\left(\frac{1+0.975}{1-0.975}\right) = 2.184724 \text{ (по таблице 14}П\ z = 2.0923).$$

Ошибка составит: $m_z = \sqrt{\frac{1}{7-3}} = 0.5.$

Определяем верхнюю границу:

$$_{\max}z = z+T_{(\alpha,df)}\cdot m_z = 2.09+2.57\cdot 0.5 = 3.375,$$

нижнюю границу:

$$_{\min}z = z+T_{(\alpha,df)}\cdot m_z = 2.09–2.57\cdot 0.5 = 0.805.$$

Обратное преобразование (по табл. 15*П*) дает: $_{\max}r \approx 1.00$, $_{\min}r \approx 0.67$. Истинный коэффициент корреляции находится в диапазоне от $r = 0.67$ до $r = 1.00$.

В среде Excel существует несколько путей поиска корреляций. Отдельный коэффициент корреляции между двумя переменными проще всего определить с помощью статистической функции = КОРРЕЛ(диапазон*X*;диапазон*Y*). Аналогичный результат дает регрессионный анализ с помощью макроса, вызываемого командой меню Сервис\ Анализ данных\ Регрессия. Когда изучаются два признака, Множественный R на самом деле является парным коэффициентом корреляции между ними. Для расчета корреляций между несколькими переменными можно использовать программу, вызываемую командой меню Сервис\ Анализ данных\ Корреляция. Результатом ее работы оказывается матрица коэффициентов корреляции.

Ложная корреляция

Когда величина коэффициента корреляции определяется в первую очередь способом подбора вариант в выборку, а не реальной зависимостью между изучаемыми признаками, то говорят о «ложной корреляции».

Величина коэффициента корреляции зависит от вытянутости эллипса рассеяния: чем больше длина главной оси эллипса отличается от сечения, тем выше значение коэффициента. Случайные единичные, а тем более парные значения могут резко повысить показатель силы связи признаков. Особенно чувствителен коэффициент корреляции к нулям, которые могут попасть в исходную матрицу при переносе данных между электронными таблицами.

Явление ложной корреляции возникает и в том случае, когда исследуемые показатели имеют в сумме постоянное значение, например 100%. Рассмотрим соотношение численности грызунов и насекомоядных в разных биотопах (табл. 8.10). Представители и первого, и второго отрядов чаще встречаются в коренных хвойных лесах, нежели в антропогенных стациях, тем более в агроценозах. Синхронность их реакции на трансформацию ландшафтов выражается высоким коэффициентом корреляции их численности: $r = 0.85$.

Таблица 8.10

Биотоп	**Численность (экз./100 конусо-суток)**			**Доля (%)**		
	*б*урозубок *Nб*	*г*рызунов *Nг*	*о*бщая *No*	*б*урозубок *Nб*/ *No*	*г*рызунов *Nг* / *No*	*о*бщая *No* / *No*
Кедровник	25	29	54	0.46	0.54	1
Смешанный	25	32	57	0.44	0.56	1
Экотон	23	21	44	0.52	0.48	1
Сосняк	22	16	38	0.58	0.42	1
Березняк	20	23	43	0.47	0.53	1
Луг	10	9	19	0.53	0.47	1
r	0.85			–1.00		

Если же оценить зависимость между долей грызунов (*Рг = Nг/No*) и долей бурозубок (*Рб = Nб/No*) в этих стациях (между индексами доминирования), она составит: $r = -1.00$. Дело в том, что эти показатели рассчитываются относительно общей суммы, поэто-

му доля полевок составляет разницу между 1 и долей бурозубок: $Рг = 1–Рб$. По существу, мы имеем уравнение строго функциональной обратной регрессии ($y = 1–1 \cdot x$), которому соответствует, естественно, максимальный отрицательный коэффициент корреляции. Требование неизменности суммы двух показателей (1 или 100 %), принятое для вычисления процентов, оказывается причиной постоянной обратной пропорции между этими показателями. Такая корреляция должна быть названа ложной, потому что характеризует не биологическую зависимость показателей, а способ их расчета. Когда общую сумму образуют три и более признаков, ложная корреляция будет отличаться от $r = –1$, но от этого не утратит своей природы математического артефакта.

При обработке массивов данных с большим числом производных признаков (индексы доминирования видов в сообществе, морфофизиологические индикаторы) нетрудно пропустить еще один вид ложной корреляции, которая наблюдается между двумя признаками, отнесенными к общей для них третьей переменной.

По неосмотрительности коэффициенты связи между индексами можно воспринять как оценку зависимости между признаками. Такие корреляции, бессознательно наведенные третьим фактором, также можно назвать ложными.

Безусловно, содержательную интерпретацию можно дать как корреляции признаков, так и корреляции индексов, но они будут кардинально отличаться. Например, для нескольких видов куньих (от ласки до барсука) коэффициент корреляции ($r = 0.96$) между длиной тонкого и толстого отделов кишечника отражает простые морфологические пропорции: у крупного животного кишечник длиннее, чем у мелкого. Однако корреляция между индексами этих органов (размеров, отнесенных к длине тела особи) характеризует уже отличия диеты разных видов ($r = 0.78$): кишечник относительно меньше у облигатных хищников, нежели у полифагов. Однако в большом массиве производных значений такие отношения между индексами могут восприниматься как зависимости между признаками, что неизбежно приведет к ложным выводам.

Чтобы уйти от подобной двусмысленности, к обработке желательно привлекать только предварительно выверенные реальные исходные показатели, а не доли, проценты или индексы.

Метод множественной корреляции

Разобранные выше примеры корреляционных зависимостей касались главным образом взаимосвязи двух сопряженных процессов, явлений или варьирующих признаков. Между тем в практике биологических исследований нередко приходится сталкиваться с более сложными случаями, например, когда сопряжены не два, а три или более изменчивых фактора (признака). В такой ситуации возникает необходимость изучить множественные связи между большим числом взаимодействующих переменных, выступающих как в виде целой системы коррелированных признаков организма, так и в форме совместного влияния сложной совокупности факторов на определенное явление. Корреляционная зависимость нескольких переменных носит название множественной корреляции и оценивается коэффициентом, определяемым на основе корреляций между всеми парами признаков. Например, коэффициент множественной корреляции между тремя признаками *A*, *B и C* вычисляется по формуле:

$$r_{A.BC} = \sqrt{\frac{r_{AB}^2 + r_{AC}^2 - 2 \cdot r_{AB} \cdot r_{AC} \cdot r_{BC}}{1 - r_{AB}^2}}\ .$$

Полученная величина характеризует связь первого признака (*A*) с двумя другими (*B* и *C*). Покажем этот способ на примере совокупного действия двух факторов, *B и C* (температуры и влажности), на суточную активность травяных лягушек (*A*). Определение парных корреляций дало следующие результаты (n = 110): r_{AB} = +0.58; r_{AC} = +0.80; r_{BC} = –0.45. Отсюда

$$r_{A.BC} = \sqrt{\frac{0.58^2 + 0.8^2 - 2 \cdot 0.58 \cdot 0.8 \cdot 0.45}{1 - 0.45^2}} = 0.86.$$

Сводный коэффициент корреляции оказался довольно высоким и, как показывает его сопоставление со стандартным значением по таблице 16*П*, вполне достоверным (при α=0.001).

С другой стороны, если обнаружена значительная корреляция между признаками *A* и *C* и между *B* и *C*, то не исключена возможность мнимой корреляционной зависимости между *A* и *B*, которая создается за счет одновременного влияния на них третьего признака *C*. Например, установленная по исследованиям в Карелии корреляция между численностью лесных полевок и урожаем семян сосны, скорее всего, объясняется не значением последних в питании

грызунов (т. е. прямой причинной связью), а тем, что оба эти явления (численность полевок и урожай семян) контролируются одними и теми же экологическими факторами (прежде всего метеорологическими) и поэтому изменяются параллельно, хотя непосредственно между собой не связаны.

В этом и подобных случаях (например, когда настоящие зависимости между признаками животных маскируются влиянием возраста или когда связи между отдельными промерами организма создаются за счет влияния живого веса и т. д.) возникает задача изучить корреляцию между двумя признаками (*A* и *B),* исключив влияние на эту связь третьего признака (С), как бы элиминировав его.

Метод частной корреляции

Этой цели служит коэффициент частной корреляции, оценивающий связь между первым и вторым признаками при постоянных значениях третьего и вычисляемый по формуле:

$$r_{A(BC)} = \frac{r_{AB} - r_{AC} \cdot r_{BC}}{\sqrt{(1 - r_{AC}^2) \cdot (1 - r_{BC}^2)}},$$

где *A* и *B* – факторы, связь которых требуется изучить;

C – фактор, влияние которого необходимо исключить из корреляционной зависимости между *A* и *B* (реперный признак);

r_{AB}, r_{AC}, r_{BC} – соответствующие парные коэффициенты корреляции, вычисляемые обычным способом;

$r_{A(BC)}$ – искомый коэффициент частной корреляции, показывающий связь между двумя признаками при исключении влияния третьего.

Этот же метод можно применить и для элиминации двух факторов при четырех переменных и т. д. Формула для расчетов примет в этом случае следующий вид:

$$r_{AB(BD)} = \frac{r_{AB(C)} - r_{AC(B)} \cdot r_{BC(D)}}{\sqrt{(1 - r_{AC(D)}^2) \cdot (1 - r_{BC(D)}^2)}}.$$

Рассмотрим нахождение коэффициента частной корреляции на упрощенном примере (взятом из книги П. Ф. Рокицкого). Получены данные о корреляции между давлением крови (*A*), содержанием в ней холестерина (*B*) и возрастом (*C*) у 142 женщин. Соответст-

вующие коэффициенты корреляции следующие: $r_{AB} = +0.25$; $r_{AC} = +0.33$; $r_{BC} = 0.51$.

Известно, что повышенное артериальное давление может быть связано с высоким содержанием холестерина в стенках кровеносных сосудов, однако и давление крови, и концентрации холестерина увеличиваются с возрастом. Поэтому возникает вопрос, создается ли корреляция между давлением крови и содержанием в ней холестерина за счет их общей связи с возрастом, или же она реально существует для каждого возраста (и независимо от него). Элиминируя эффект возраста по приведенной выше формуле, получим:

$$r_{A(BC)} = \frac{0.25 - 0.33 \cdot 0.51}{\sqrt{(1 - 0.33^2) \cdot (1 - 0.5^2)}} = 0.12.$$

По таблице 16*П* можно установить, что при $n = 150$ для достоверности коэффициента корреляции даже при уровне значимости $\alpha = 0.05$ его величина должна быть не меньше 0.16. В данном же случае полученное значение меньше табличного и, следовательно, коэффициент корреляции от нуля достоверно не отличается. Таким образом, внутри отдельных возрастных групп корреляционной связи между давлением крови и содержанием холестерина, по крайней мере на изученном материале, не обнаруживается. Пока нет оснований отбрасывать нулевую гипотезу.

Второй пример демонстрирует использование коэффициента частной корреляции для более глубокого проникновения в структуру нескольких факторов наведения. Рассмотрим выборку объектов разного статуса (11 видов мелких млекопитающих), взяв в качестве признаков их численность в семи биотопах прибайкальской равнины. Реперным признаком послужила суммарная численность вида во всех биотопах. Здесь коэффициент корреляции отражает сходство между биотопами по соотношениям численности 11 видов. Например, оказалось, что между березняком и экотоном (граница между березняком и коренными лесами) и общая корреляция ($r = 0.92$), и частная ($r = 0.64$) высока и положительна. Можно утверждать, что население животных этих биотопов почти идентично.

В свою очередь, корреляция между кедровником и лугом не проявилась ($r = –0.08$), но коэффициент частной корреляции был велик и отрицателен ($r = –0.43$). Этим оттеняется тот факт, что виды, отсутствующие на лугу, многочисленны в кедровнике (красная полевка, мышь), а обычные в агроценозе – крайне редки в тайге (серые

полевки). Частная корреляция не просто показала, что население биотопов не сходно, но и что во многом диаметрально противоположно.

Тем самым удалось выявить два уровня факторов наведения. Первый из них хорошо известен – это расселение таежных видов в другие биотопы, в том числе на луга. В результате сезонных миграций видовой состав тайги и луга меняется несогласованно, без определенной направленности (одни виды идут из тайги в агроценозы, другие – в противоположном направлении), отличия по численностям всех видов получаются стохастические $r = –0.08$.

Частная корреляция устраняет эффект прироста численности за счет иммигрантов и выдвигает на первый план контраст остаточной численности. Понятно, что ее формируют в первую очередь характерные обитатели биотопов: в тайге это лесные полевки, на лугу – серые. Так проявляется второй фактор «наведения»: отличие качества среды в разных биотопах. Он обеспечивает формирование принципиально несходных зооценозов, что и выявляется высокой частной корреляцией $r = –0.43$.

Корреляционное отношение и критерий линейности

Для измерения силы связи между переменными величинами при криволинейных зависимостях, т. е. когда равномерному изменению первого признака соответствуют определенные неравномерные изменения второго, коэффициент корреляции подходит плохо. В таких случаях применяется корреляционное отношение, обозначаемое греческой буквой η (эта), причем оно описывает взаимосвязь между переменными двусторонне – как y по x ($\eta_{y/x}$), так и x по y ($\eta_{x/y}$). Значения корреляционных отношений, показывающие зависимость изменения первого признака от второго и второго от первого, тем более сходны по величине, чем сильнее связь и чем она ближе к линейной. При линейной зависимости корреляционное отношение совпадает по величине с коэффициентом корреляции (который служит равнозначной мерой связи признаков), а при криволинейной – отличается от него: одно из значений оказывается больше, другое меньше коэффициента корреляции.

В природе редко встречаются случаи двусторонних причинных зависимостей между двумя переменными, чаще наблюдается

односторонняя зависимость. Например, если плодовитость животных зависит от кормовых условий, то последние, естественно, от плодовитости животных не зависят.

Корреляционное отношение есть отношение дисперсии предсказанных значений одного из признака к его общей дисперсии (сокращая число степеней свободы, имеем отношение сумм квадратов):

$$\eta_{x/y} = \sqrt{\frac{(x'-Mx)^2}{(x-Mx)^2}},\ \eta_{y/x} = \sqrt{\frac{(y'-My)^2}{(y-My)^2}}.$$

Значимость величин оценивается по критерию Стьюдента:

$$T = \eta / m_\eta,\ \text{где}\ m_\eta = \frac{1-\eta^2}{\sqrt{n}}.$$

Ход вычислений можно показать на примере из раздела **Криволинейная регрессия**. Сначала рассчитываются два уравнения *линейной* регрессии

$H' = 107.88 \cdot Lt - 404.15$, $Lt' = 0.008 \cdot H + 4.0896$

и теоретические значения каждого из признаков (табл. 8.11).

Таблица 8.11

№	*H*	*Lt*	*H'*	*Lt'*
1	3.4	40	4.4	–37.4
2	4.2	50	4.5	48.9
3	5.2	150	5.3	156.8
4	5.8	120	5.0	221.6
5	7.1	240	6.0	361.8
6	7.0	410	7.4	351.0
7	7.4	370	7.0	394.2
8	8.2	500	8.1	480.5
9	8.5	610	9.0	512.8
M	6.3	276.7	6.3	276.7
$\Sigma(x-M)^2$	25.1	336800	21.6	291754.1

Затем рассчитываем средние (M), суммы квадратов отклонения от них отдельных вариант ($\Sigma(x-M)^2$) и сами корреляционные отношения:

$$\eta_{x/y} = \sqrt{\frac{(x'-Mx)^2}{(x-Mx)^2}} = \sqrt{\frac{291754.1}{336800}} = 0.931,$$

$$\eta_{y/x} = \sqrt{\frac{(y'-My)^2}{(y-My)^2}} = \sqrt{\frac{21.6}{25.1}} = 0.927.$$

$$m_\eta = \frac{1-\eta^2}{\sqrt{n}} = \frac{1-0.931^2}{3} = 0.044,$$

$T = \eta / m_\eta = 0.927 / 0.044 = 20.9.$

Полученная эмпирическая величина (20.9) много больше табличной для $\alpha = 0.05$ и $df = 9–2 = 7$ $T_{(0.05,7)} = 2.37$ (табл. 6*П*). Таким образом, сомневаться в достоверности отличия от нуля вычисленных коэффициентов нет оснований.

В данном случае значения корреляционных отношений почти совпадают как друг с другом, так и с коэффициентом корреляции (0.931, 0.927 и 0.931 соответственно), что характерно для случая линейной зависимости между переменными.

Высказанное предположение можно проверить с помощью *критерия линейности*. В соответствии с простейшим из них связь считается криволинейной, если разность квадратов корреляционного отношения и коэффициента корреляции превышает 0.1:

$\eta^2 – r^2 > 0.1.$

Этот критерий показывает, что в нашем случае линия хорошо описывает зависимость веса печени от размеров тела:

$0.930727^2–0.930693^2 = 0.866253–0.8661902 = 0.00006 < 0.1.$

Более точные оценки, учитывающие объем выборки, дает критерий Блекмана, согласно которому связь считается криволинейной, если произведение разности квадратов корреляционного отношения и коэффициента корреляции на объем выборки превышает 11.37:

$n\cdot(\eta^2 – r^2) > 11.37.$

И этот критерий говорит о линейности изучаемой связи:

$9\cdot0.00006 = 0.00054 < 11.37.$

Существуют и другие, более точные критерии линейности. Тем не менее, для оценки степени криволинейности связи лучше пользоваться более точным методом – дисперсионным анализом и более простым показателем – коэффициентом детерминации R^2, к тому же их расчеты автоматизированы в среде Excel.

Ранговый коэффициент корреляции Спирмена

Помимо рассмотренных выше параметрических показателей связи в биометрии применяются и непараметрические. Обычно их используют при сильных отклонениях изучаемого распределения от нормального (или сомнениях на этот счет), а также в тех случаях, когда требуется оценить зависимость между качественными или полуколичественными признаками, точное количественное измерение которых затруднено (оценки в баллах или других условных единицах). Если варианты выборки могут быть упорядочены по степени выраженности их свойств, для измерения степени сопряженности между ними можно воспользоваться непараметрическим показателем связи – ранговым коэффициентом корреляции Спирмена:

$$r_S = 1 - \frac{6 \cdot \sum d^2}{n \cdot (n^2 - 1)},$$

где d – разность между рангами сопряженных значений признаков x и y;

n – объем выборки.

Этой формулой следует пользоваться в тех случаях, когда выборки не содержат повторяющихся вариант, когда все ранги выражены разными целыми числами. Если же исходные ряды содержат одинаковые значения, расчет корреляции приходится вести по другой формуле, включающей поправку на повторы (при этом одинаковым вариантам присваивается средний ранг):

$$r_S = \frac{\frac{(n^3 - n)}{6} - (T_x + T_y) - \sum d^2}{\sqrt{\left(\frac{(n^3 - n)}{6} - 2 \cdot T_x\right)\left(\frac{(n^3 - n)}{6} - 2 \cdot T_y\right)}},$$

где T_x, T_y – поправки на серии повторов для каждой выборки:

$$T_x = \frac{\sum^{k} (t_x{}^3 - t_x)}{12},$$

t – число членов в каждой группе одинаковых вариант.

Поправки T_x, T_y учитывают k групп повторяющихся вариант.

Рассмотрим технику вычислений на примере изучения связи между оцененными в баллах численностью лисицы (x) и обилием мышевидных грызунов (y) (по годам наблюдений):

	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966
x	2.6	2.1	2.3	2.3	1.6	2.2	3.0	2.1	1.5	2.2
y	3.0	2.4	3.6	2.9	3.7	3.3	4.0	2.1	1.0	3.5

Чтобы проверить наличие и определить силу этой связи, нужно упорядочить значения сопряженных признаков по степени их выраженности, затем присвоить им ранги, обозначив значения порядковыми числами натурального ряда, и рассчитать коэффициент корреляции. Техника вычислений показана в таблице 8.12.

Таблица 8.12

Численность лисицы в баллах, *x*	**Обилие грызунов в баллах, *y***	**Ранги вариант**		**Разность между рангами, *d***	d^2
		R_x	R_y		
1.5	1.0	1	1	0	0
1.6	3.7	2	6	–4.0	16.00
2.1	2.4	3.5	3	+0.5	0.25
2.1	2.1	3.5	2	+1.5	2.25
2.2	3.3	5.5	7	–1.5	2.25
2.2	3.6	5.5	8.5	–3.0	9.00
2.3	3.6	7.5	8.5	–1.0	1.00
2.3	2.9	7.5	4	+3.5	12.25
2.6	3.0	9	5	+4.0	16.00
3.0	4.0	10	10	0	0
					Σ = 59

В ряду значений признака x есть три пары одинаковых вариант, поэтому поправка будет равна:

$$T_x = \frac{(2^3 - 2) + (2^3 - 2) + (2^3 - 2)}{12} = 1.5.$$

В ряду признака y всего одна пара одинаковых значений, поправка составит:

$$T_y = \frac{(2^3 - 2)}{12} = 0.5.$$

Величина $\frac{(n^3-n)}{6}=\frac{(10^3-10)}{6}=165.$

Коэффициент ранговой корреляции равен:

$$r_S=\frac{165-(1.5+0.5)-59}{\sqrt{(165-2\cdot1.5)(165-2\cdot0.5)}}=0.638.$$

Если воспользоваться формулой без поправок, результат будет несколько иным:

$$r_S=1-\frac{6\cdot\sum d^2}{n\cdot(n^2-1)}=1-\frac{6\cdot59}{10\cdot(10^2-1)}=0.642.$$

Статистическая ошибка и критерий достоверности отличия коэффициента корреляции от нуля вычисляются по формулам:

$$m_r=\sqrt{\frac{1-r_S^2}{n-2}}=\sqrt{\frac{1-0.638^2}{10-2}}=0.272,$$

$T_r = r_S/m_r = 0.638/0.272 = 2.34.$

Величина критерия равна несколько выше критического значения (2.31) для уровня значимости $\alpha = 0.05$ и числа степеней свободы $df = n–2 = 8$ (табл. 6*П)*. Казалось бы, это дает основание отвергнуть нулевую гипотезу $r_S = 0$ и с вероятностью $P = 95\%$ констатировать достоверность установленной связи. Однако в связи с тем, что при небольших выборках статистические свойства коэффициента Спирмена еще «хуже», чем коэффициент Пирсона, для оценки значимости корреляции лучше воспользоваться специально подготовленной таблицей 17*П*, аналогичной рассмотренной выше таблице 16*П*. Чтобы полученный коэффициент можно было считать достоверно отличным от нуля, он должен превышать табличное значение при данном n. В нашем случае ($n = 10$, $\alpha = 0.05$) коэффициент $r = 0.638$ ниже табличного $r = 0.64$, следовательно, значимо от нуля не отличается. Зависимость численности лисицы и грызунов по приведенным данным достоверно не прослеживается.

Корреляция между качественными признаками

Степень сопряженности (сочетаемость) двух возможных состояний двух качественных признаков можно измерить с помощью особого коэффициента корреляции – коэффициента контингенции Шарлье.

У каждой особи отмечают два альтернативных признака, и вся выборка разбивается на четыре части:

a – число особей, имеющих оба признака (+ +),

b – число особей, имеющих первый признак, но не имеющих второго (+ –);

c – число особей, не имеющих первого признака, но имеющих второй (– +);

d – число особей, не имеющих обоих признаков (– –).

На схеме это выглядит как четырехклеточная корреляционная решетка:

Признак 2 / Признак 1	Присутствует (+)	Отсутствует (–)	Σ
Присутствует (+)	*a*	*c*	*a* + *c*
Отсутствует (–)	*b*	*d*	*b*+ *d*
Σ	*a* + *b*	*c* + *d*	*n* = *a* + *b* + *c* + *d*

Степень взаимосвязи определяется по формуле:

$$r = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}}.$$

При вычислении коэффициента корреляции между двумя альтернативными признаками выясняется вопрос о том, чаще ли оба признака одновременно присутствуют или отсутствуют у варианты, чем это могло бы быть по случайным причинам. Достоверность отличия от нуля оценивается по критерию Стьюдента:

$$T_r = r / m_r,$$

где $$m_r = \frac{1 - r^2}{\sqrt{n-1}}.$$

При проверке влияния перекрытий на оплодотворяемость самок песцов получены первичные материалы о численности родивших (+) и неродивших (–) самок из числа хотя бы дважды перекрытых (+) и неперекрытых (–).

Признак 2 / Признак 1	Родившие (+)	Неродившие (–)	Σ
Перекрытые (+)	370	90	460
Неперекрытые (–)	100	120	220
Σ	470	210	n = 680

Коэффициент контингенции равен:

$$r = \frac{370 \cdot 120 - 100 \cdot 90}{\sqrt{470 \cdot 210 \cdot 460 \cdot 220}} = 0.35.$$

Ошибка коэффициента составит:

$$m_r = \frac{1 - 0.35^2}{\sqrt{680 - 1}} = 0.0327,$$

а критерий Стьюдента T_r = 0.35/0.0327 = 10.7.

Полученная величина (10.7) настолько велика, что превышает табличное даже для доверительной вероятности выше P = 0.999 (уровень значимости α<0.001). Влияние повторных покрытий на оплодотворяемость самок песцов несомненно.

При исследовании связи между белой мастью и красными глазами у кроликов получены следующие данные.

При подстановке всех значений сумм из таблицы в формулы получим: r = 0.76, m = 0.04, T = 19. Достоверность связи не вызывает сомнений.

	Красные глаза	Некрасные глаза	Σ
Белая шерсть	29	11	40
Окрашенная шерсть	1	59	60
Σ	30	70	100

9
ЗАДАЧА «КЛАССИФИЦИРОВАТЬ ОБЪЕКТЫ»

Методы многомерного анализа

Методы многомерной статистики – своеобразный отклик математики на запрос современной науки обеспечить, с одной стороны, более полное (многоплановое) количественное описание биологических объектов и окружающей среды (с помощью большого числа переменных), а с другой стороны – представить огромные массивы информации в более наглядном, интегрированном, обобщенном виде. Поиск максимально полной, но интегрированной характеристики каждого объекта привел к идее *рассчитывать* небольшое число новых *признаков*, вбирающих в себя почти всю информацию от исходных характеристик; в результате полученные данные «сворачиваются» до размеров, которые в состоянии охватить мысль. Так решается «задача сокращения размерности».

Теоретической основой для методов многомерной статистики служит понятие гиперпространства, или многомерного пространства. В отличие от привычного физического трехмерного пространства, имеющего три ортогональных (взаимно перпендикулярных) оси, многомерное пространство имеет множество осей координат, в качестве которых выступают признаки (переменные) изучаемых объектов. Отдельный объект, охарактеризованный по нескольким признакам, рассматривается как отдельная точка, а множество объектов – как облако точек. Если объекты (особи разного возраста, пола, органы, пробы, даты, разные популяции, виды, биотопы, местообитания и т. п.) отличаются друг от друга по разным признакам, то они будут занимать разное положение в многомерном пространстве; объекты оказываются рассеянными в нем.

Главной характеристикой объектов становится расстояние между ними в этом гиперпространстве, а главной особенностью всей выборки – форма облака рассеяния со своими пустотами и сгущениями объектов. Методы многомерной статистики изучают информацию, «закодированную» в порядке расположения объектов друг относительно друга. Например, взрослая особь по множеству

размерных признаков превосходит молодую. Она будет находиться в зоне особей с большими размерами, тогда как молодая – в зоне мелких. Исследование относительного места расположения особи в «облаке» других особей раскрывает, «расшифровывает» ее биологический статус.

В кластерном анализе вычисляется один новый признак (абсолютное расстояние между объектами), многомерные отношения объектов нанизываются на одну ось. В дискриминантном и компонентном анализах можно рассчитывать несколько новых признаков, рассматривающих пространственные отношения объектов с разных точек зрения. Суммарное отличие объектов друг от друга, т. е. их дисперсия, становится важнейшей характеристикой информационной насыщенности массива данных.

Основы кластерного анализа

Классификация, кластеризация – методы, широко используемые в современной таксономии, – позволяют наглядно представить сходство или различие биологических объектов, охарактеризованных по многим параметрам. Эти подходы можно применять в самых разных областях биологии, в частности, для сравнения условий среды в сериях местообитаний, выявления различий и сходства между сообществами живых организмов, отдельными их популяциями, группами, особями и т. п. Кластерный анализ, как и многие другие многомерные статистические приемы, не имеет достаточно разработанного математического аппарата для статистического оценивания полученных данных; его основная функция – выявление скрытой структуры биологического материала, что позволяет затем целенаправленно ставить и решать конкретные биометрические задачи с помощью простых статистических методов (регрессионного, корреляционного, дисперсионного и др.).

Суть кластерного анализа состоит в

1) определении «расстояний» (меры различия) между объектами по всей совокупности признаков,
2) группировании сходных объектов в кластеры (*кластеризация*),
3) графическом изображении сети (или древа) расстояний между всеми объектами.

Речь, следовательно, идет о формировании *одного нового признака* (относительного расстояния) на основании нескольких исходных.

В качестве меры расстояния может выступить любой признак. Так, разность между значениями длины тела двух полевок есть не что иное, как расстояние между ними по одному признаку. Достоинство кластерного анализа заключается в том, что он позволяет получить обобщенную меру расстояния между объектами по всему множеству анализируемых признаков.

Один из вариантов такой меры основан на коэффициенте сходства Съёренсена, который используется для сравнения многовидовых сообществ. «Расстояние» вычисляется по формуле:

$$S = 1 - \frac{2 \cdot A}{B + C},$$

где B и C – число видов в двух сравниваемых сообществах,

A – число общих видов.

Рассмотрим в качестве примера анализ биоценотических группировок мелких млекопитающих Приладожья. Видовой состав изучен в 7 основных биотопах: **л**ишайниковых **с**осняках (А)**, с**осняках-**з**еленомошниках (Б), **е**льниках (В), **с**пелых **л**иственных и смешанных лесах (Г), **л**иственном **м**елколесье (Д), молодых зарастающих **в**ырубках (Е) и по границе сеяного **л**уга и **о**льшаника (Ж). Встречаемость (по принципу отсутствие – присутствие) и относительная численность (число особей на 100 ловушко-суток) 14 видов землероек и грызунов показаны в таблицах 9.1 и 9.2. Дальнейшая процедура сводится к следующему.

По данным таблицы 9.1 рассчитывается матрица расстояний между разными биотопами. Например, в ельниках (В) отмечено 12 видов мелких млекопитающих, а на вырубках (Е) – 5; из них общих для обоих биотопов – 5. Отсюда расстояние:

$S = 1–10/17 = 0.41$.

Смысл следующей операции – собственно кластеризации (от слова «кластер» – гроздь, группа) – состоит в последовательном объединении объектов в кластеры, в группы, внутри которых сходство между объектами выше, чем с другими объектами или кластерами. Вначале объединяются наиболее сходные объекты (с наименьшим расстоянием между собой), затем приближающиеся к ним

по этому показателю и так далее до момента слияния всех объектов в один общий кластер. При этом на промежуточных этапах могут образовываться несколько отдельных кластеров. Уровень каждого объединения фиксируется и затем отображается на графике.

Таблица 9.1

Вид	**Биотопы**						
	ЛС	**СЗ**	**Е**	**СЛ**	**ЛМ**	**В**	**ЛО**
	А	Б	В	Г	Д	Е	Ж
Обыкновенная бурозубка	1	1	1	1	1	1	1
Средняя бурозубка	0	1	1	1	1	0	0
Малая бурозубка	0	1	1	1	1	1	1
Равнозубая бурозубка	0	0	1	1	0	0	0
Крошечная бурозубка	0	0	1	0	0	0	0
Водяная кутора	0	0	1	1	0	0	0
Лесная мышовка	1	1	1	1	1	1	0
Лесной лемминг	0	0	0	1	0	0	0
Полевая мышь	0	0	1	1	0	0	1
Мышь-малютка	0	0	1	0	0	0	1
Рыжая полевка	1	1	1	1	1	1	1
Красная полевка	1	1	1	1	1	0	0
Темная полевка	0	1	1	1	1	1	1
Полевка-экономка	0	0	0	1	1	0	1
Число видов	4	7	12	12	8	5	7

Таблица 9.2

ЛС	СЗ	Е	СЛ	**ЛМ**	В	ЛО	
А	0.27	0.5	0.5	0.33	0.34	0.64	ЛС
	Б	0.26	0.26	**0.07**	0.17	0.42	**СЗ**
		В	0.17	0.3	0.41	0.58	Е
			Г	0.21	0.41	0.47	СЛ
				Д	0.23	0.47	ЛМ
					Е	0.33	В
						Ж	ЛО

Существует множество вариантов процедуры кластеризации, из них наиболее простым считается метод «ближайшего соседа», не требующий обязательного использования ЭВМ. Сначала по матрице расстояний (табл. 9.2) отыскиваются ближайшие соседи для всех объектов и заносятся в таблицу наименьших расстояний (табл. 9.3). Так, к лишайниковому сосняку (А) ближе всего сосняк-зеленомошник (Б): $S_{АБ} = 0.27$, а к сосняку-зеленомошнику (Б) – лиственное мелколесье (Д): $S_{БД} = 0.07$, (минимальное расстояние из всех изученных биотопов).

Таблица 9.3

Сосед 1	**Сосед 2**	**Расстояние, S**	**Кластер**	**Сосед 2**	**Расстояние, S**	**Кластер**
А	Б	0.27	1			
Б	Д	**0.07**	1			
В	Г	0.17	2	Б	0.26	
Г	В	0.17	2	Д	0.21	3
Д	Б	0.07	1			
Е	Б	0.17	1			
Ж	Е	0.33	1			

Заполнив четыре первые графы, приступают к построению предварительного дендроида расстояний (рис. 9.1, *А*). Для этого на график наносят индексы объектов и расстояния между ними, соединяют их прямыми линиями. В нашем случае сначала образовалось два отдельных кластера (АБДЕЖ и ВГ), но их может быть и больше. Теперь вновь возвращаемся к таблицам 9.2 и 9.3. В пятой графе против объектов из меньшего кластера следует отметить индекс ближайших объектов из большего кластера, а в шестой – расстояние между ними. Далее выбираем звено наименьшей протяженности – это спелые лиственные леса (Г) и молодняки (Д): $S_{ГД} = 0.21$. На рис. 9.1 соединим кластеры пунктирной линией, и кластеризация завершена.

Последний этап – построение окончательного варианта дендрограммы. Здесь также есть несколько возможностей. Представленное на рис. 9.1, *Б* «древо минимальной протяженности» строится с учетом единственного условия – соблюдения пропорций расстояний между биотопами-соседями.

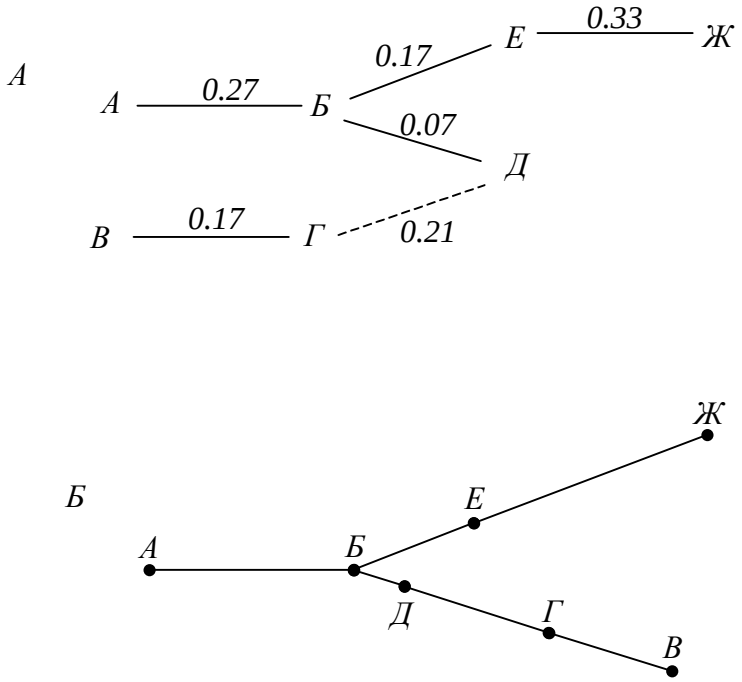


***Рис. 9.1**. А – схема поэтапной кластеризации; Б – «древо минимальной протяженности»; А–Ж – индексы биотопов*

Классический вариант дендрограммы приведен на рис. 9.2. По оси абсцисс размещаются объекты в том порядке, который продиктован логикой их связей и субъективными вкусами исследователя, отдельные ветви «древа» при этом не должны пересекаться. По оси ординат откладывается расстояние между ближайшими соседями (рис. 9.2).

Интерпретация полученных результатов подчеркивает достоинства дендрограммы как емкой иллюстрации обобщающих характеристик. Так, в данных по Приладожью кластерный анализ выделил группы биотопов, наиболее близких по условиям обитания и видовому составу зверьков. Наиболее богаты видами еловые и смешанные леса (В и Г). Обедненными териокомплексами, представленными в основном политопными видами, характеризуются сосняки-зеленомошники, вырубки и лиственное мелколесье (Б, Е, Д). Население сосняков (Б и А) в общем сходно (табл. 9.1), но в лишайни-

ковых сосняках видов очень мало. Наконец, наиболее обособленное положение занимает биотопический комплекс экотона – границы между лугом и лесом (Ж), включающий представителей смежных биотопов.

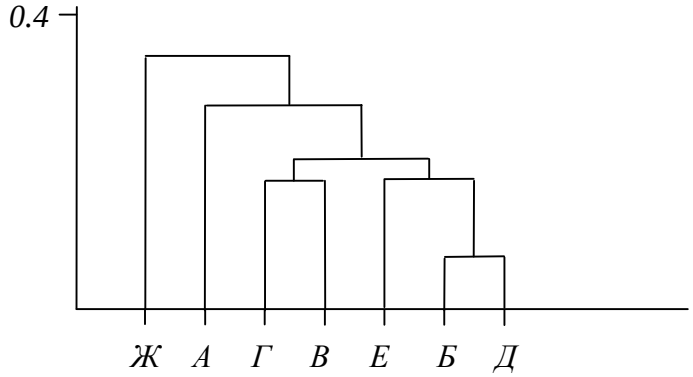


Рис. 9.2. *Дендрограмма сходства биотопов по видовому составу мелких млекопитающих*

При использовании в кластерном анализе количественных признаков применяют евклидову меру расстояния:

$$d_{ji} = \sqrt{\frac{\sum (x_{kj} - x_{ki})^2}{m}}$$

где x_{kj}, x_{ki} – значения k-го признака у j-го и i-го объектов,

m – число учитываемых признаков.

Рассчитав матрицу расстояний между объектами по комплексу количественных признаков, проводят кластеризацию и построение дендрограмм по описанному выше методу. Рассмотрим эту процедуру на уже знакомом примере многовидовых группировок мелких млекопитающих в трех типах биотопов Приладожья (Б – сосняки-зеленомошники, В – ельники, Д – мелколесье), но по данным количественных учетов канавками (экз. на 10 канавко-суток; табл. 9.4).

Рассчитаем евклидово расстояние сначала между двумя биотопами – сосняком (Б) и ельником (В):

$$d_{ji} = \sqrt{\frac{(3.9-7.2)^2 + (1.8-1.1)^2 + ... + (0-0.2)^2}{13}} = \sqrt{\frac{12.252}{13}} = 0.971.$$

Таблица 9.4

Вид	**Численность, экз. на 10 канавко-суток**		
	сосняк-зеленомошник (Б)	**ельник (В)**	**лиственное мелколесье (Д)**
Обыкновенная бурозубка	3.9	7.2	6.0
Средняя бурозубка	1.8	1.1	0.5
Малая бурозубка	1.9	2.0	1.6
Равнозубая бурозубка	0.01	0.2	0.1
Крошечная бурозубка	0.04	0.04	0
Водяная кутора	0.04	0.06	0.4
Лесная мышовка	0.6	0.3	0.7
Лесной лемминг	0.2	0	0.05
Мышь-малютка	0.04	0	0
Рыжая полевка	1.5	0.8	0.8
Красная полевка	0.06	0.6	0.02
Темная полевка	0.2	0	0.7
Полевка-экономка	0	0.2	0.2
Всего	10.3	12.9	10.9

Повторив эту процедуру для других пар биотопов, получим: $d_{БД} = 0.741$ и $d_{ВД} = 0.417$. Сведем полученные данные в матрицу расстояний:

Б	0.97	0.74
	В	0.42
		Д

Сосед 1	Сосед 2	Расстояние, *d*
Б	Д	0.74
В	Д	0.42
Д	В	0.42

Дендрограмма приведена на рис. 9.3. По сравнению с предыдущим случаем она выявляет новые нюансы отношений между биоценотическими комплексами млекопитающих. Если по видовому

составу лиственные леса (Д) были ближе к соснякам (Б) (в отличие от ельников и там и тут встречались лесной лемминг и темная полевка), то по уровню численности лиственные леса ближе к ельникам (в отличие от сосняков в этих биотопах существенно больше обыкновенных бурозубок и рыжих полевок).

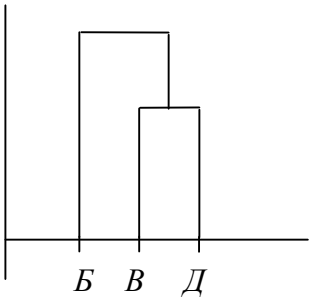


Рис. 9.3. *Дендрограмма сходства биотопов по численности мелких млекопитающих*

Когда изучаемые признаки имеют разную размерность (экз./га, кг, мм, %), то вместо таблицы исходных данных (см. табл. 9.4) для вычисления евклидовой меры расстояния следует подготовить таблицу нормированных значений. Для этого по каждой строке первичной таблицы рассчитываются средняя арифметическая (M_j) и стандартное отклонение (S_j), а затем – нормированные значения каждой варианты из этой строки:

$$t = \frac{x - M_j}{S_j},$$

где x – исходные значения вариант 1-й строки (i-го признака).

Например, для первой строки таблицы 9.4 $M_1 = 5.7$, $S_1 = 1.67$. Новые значения строки будут равны: $t_{11} = (3.9–5.7)/1.67 = –1.078$, $t_{12} = (7.2–5.7)/1.67 = 0.89$, $t_{13} = (6.0–5.7)/1.67 = 0.18$.

Полученная таким образом таблица используется для вычисления евклидовой меры расстояния между объектами по рассмотренному выше алгоритму.

Кроме рассмотренных мер расстояния для кластерной классификации объектов исследования используют коэффициент корреляции (r) в форме коэффициента «не-корреляции»: $d_{ji} = 1– r_{ji}$. При этом следует использовать нормированные значения признаков.

В этом случае матрица расстояний формируется по предварительно рассчитанной корреляционной матрице. Поскольку метод корреляционного анализа рассмотрен нами выше, а дальнейшие процедуры несложны и очевидны, мы не иллюстрируем этот прием конкретным примером.

В среде Excel нет программы для проведения кластерного анализа. Но его можно выполнить с помощью пакета StatGraphics.

Основы дискриминантного анализа

Этот метод многомерной статистики служит для дискриминации, т. е. различения (дифференциации) и диагностирования (распознавания) биологических объектов и явлений, отличия между которыми неочевидны. В медицине этот метод используется для идентификации заболевания по ряду показателей (характерных симптомов), а в биологии – для установления групповой принадлежности отдельных особей (объектов). Иными словами, общая задача дискриминантного анализа заключается в том, чтобы определить, к какой из двух известных групп объектов принадлежит изучаемый объект. Как и в кластерном анализе, исследуемые объекты представлены несколькими численными признаками и (в простейшем случае) требуется сформировать *один расчетный признак*, однозначно характеризующий каждый объект. Однако задачи дискриминантного анализа прямо противоположны кластерному: не выделить из множества объектов группы близких, а отнести тот или иной объект к определенной, априорно выделенной группе. Эта идентификация (дискриминация) объекта выполняется с помощью уравнения дискриминации (дискриминантной функции), которое воплощает в себе максимальное отличие между предварительно заданными группами (дискриминация «с обучением»).

Рассмотрим общий принцип использования этого метода на примере определения пола у пеночек-весничек. Визуально молодые самцы и самки этого вида не различаются, а распределения морфологических признаков (длина крыла, хвоста, цевки) у них довольно сильно перекрываются, что не позволяет с уверенностью диагностировать пол этих птиц. Например, для длины крыла степень трансгрессии составляет 20%, а длины цевки – 90%. Между тем дискриминантный анализ в силах справиться с подобной задачей.

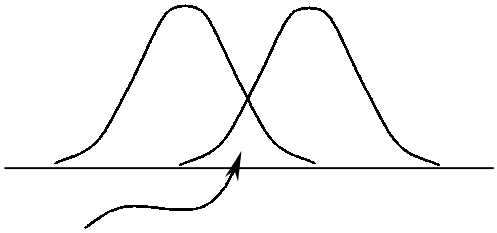


Рис. 9.4. *Зона трансгрессии – наложение «хвостов» распределений*

На основе реальных признаков птиц математически рассчитывается искусственный и единственный признак, учитывающий все незначительные морфологические отличия полов в целом по всем признакам. Эти расчеты проводятся с условием, чтобы различия между группами самцов и самок стали наиболее выраженными, а трансгрессия между их распределениями – наименьшей. Так удается свести к минимуму долю животных неопределенного пола и с высокой степенью достоверности предсказывать пол по морфологическим признакам.

В основе дискриминантного анализа лежит дискриминантная функция; для двух признаков она имеет такой вид:

$$Z = a \cdot x + b \cdot y - H - \Delta Z \ .$$

Как можно видеть, признаки x и y, объединяясь, дают один признак Z. Если в анализ будут включены размерные признаки, такие как длина крыла и длина хвоста, то новый признак можно назвать «относительные размеры тела».

Коэффициенты a и b оценивают «вклад» каждого из признаков в диагностические возможности функции. На первых этапах работы в расчеты обычно вовлекается большое число реальных признаков, многие из которых никак не влияют на диагностические возможности дискриминантной функции, для них дискриминантные коэффициенты близки к нулю. Такие признаки исключают из рассмотрения, а дискриминантую функцию рассчитывают заново. Формальным критерием для отбраковки «неинформативных» признаков служит аналог критерия Стьюдента для оценки значимости коэффициентов регрессии, который мы не рассматриваем.

Коэффициент H – это граничная величина между значениями Z для самцов и самок. Свободный член уравнения ΔZ – поправка на разные объемы выборок.

Теперь сформулируем задачу более конкретно. У 10 самцов и

10 самок погибших по разным причинам пеночек-весничек (их пол был определен путем вскрытия) взяли промеры длины крыла и хвоста. По этим данным вычислены необходимые для дальнейших расчетов величины (суммы значений, их квадратов и произведений), сведенные в таблицу 9.5:

$\Sigma x = 1218$; $\Sigma x^2 = 74324\ 5$; $\Sigma y^2 = 9275$; $\Sigma y^2 = 43087.25$;$\Sigma(x \cdot y) = 56564.5$.

Таблица 9.5

Самцы (1)					
65	50	4225	2500	3250	0.14
61	47	3721	2209	2867	0.01
64	48	4096	2304	3072	0.09
63.5	51	4032.25	2601	3238.5	012
63	47	3969	2209	2961	0.05
62	46	2844	2116	2852	0.02
63	48	3969	2304	3024	0.07
63.5	48	4032.25	2304	3048	0.08
62	47	3844	2209	2914	0.03
64	46	4096	2116	2944	0.06
$\Sigma_1 = 631$	478	39828.5	22872	301705	–

Самки (2)					
59	44	3481	1936	2596	–0.08
59	46	3481	2116	2714	–0.05
54	45	2916	2025	2430	–0.17
57.5	43	3306.25	1849	2472.5	–0.12
61	46.5	3721	2162.25	2836.5	0.0004
60.5	46	3660.25	2116	2783	–0.01
57.5	45	3306.25	2025	2722.5	–0.09
58	44	3364	1936	2552	–0.10
60.5	45	3660.25	2025	2610	–0.03
60	45	3600	2025	2700	–0.04
$\Sigma_2 = 587$	449.5	34496	20215.25	26416.5	–
$\Sigma = 221218$	927.5	74324.5	43087.25	56564.5	–

Теперь определим средние арифметические:

$M_{x1} = 631/10 = 63.1$; $M_{y1} = 478/10 = 47.8$; $M_{x2} = 58.7$; $M_{y2} = 44.95$

и их разности: $d_x = 63.1–58.7 = 4.4$; $d_y = 47.8–44.95 = 2.85$.

Находим также вспомогательные величины:

$$C_x = \sum x^2 - \frac{(\sum x)^2}{N} = 7432.5 - \frac{1218^2}{20} = 148.3,$$

$$C_y = \sum y^2 - \frac{(\sum y)^2}{N} = 43087.25 - \frac{927^2}{20} = 74.44,$$

$$C_{xy} = \sum (x \cdot y) - \frac{(\sum x \cdot \sum y}{N} = 56564.5 - \frac{927.5 \cdot 1218}{20} = 79.75.$$

Наконец, для определения коэффициентов *a* и *b* необходимо решить следующую систему уравнений:

$$C_x \cdot a + C_{xy} \cdot b = d_x$$

$$C_{xy} \cdot a + C_y \cdot b = d_y.$$

Ее корнями будут:

$$a = \frac{C_{xy} \cdot d_y - C_y \cdot d_x}{C_{xy} \cdot C_{xy} - C_x \cdot C_y} = \frac{79.5 \cdot 2.85 - 74.44 \cdot 4.4}{79.75^2 - 148.3 \cdot 74.44} = 0.021423,$$

$$b = \frac{d_x - C_x \cdot b}{C_{xy}} = \frac{4.4 - 148.3 \cdot 0.021423}{79.75} = 0.015335.$$

Теперь найдем средние значения признака Z для самцов и самок:

$Z_1 = a \cdot M_{x1} + b \cdot M_{y1} = 0.021423 \cdot 63.1 + 0.015335 \cdot 47.8 = 2.0848,$

$Z_2 = a \cdot M_{x2} + b \cdot M_{y2} = 0.021423 \cdot 58.7 + 0.015335 \cdot 44.95 = 1.9468.$

Определяем разность между этими средними, или центроидами:

($D = Z_1 - Z_2$): $D = 2.0848 - 1.9468 = 0.138$.

Найдем границу между группировками самцов и самок:

$H = Z_2 + D/2 = 1.9468 + 0.138/2 = 2.0158$.

Так получен третий член уравнения дискриминации. Что касается четвертого, поправки на объем выборки, то он определяется по формуле:

$$\Delta Z = \frac{\ln(n_{max} / n_{min})}{D},$$

где n_{max} – объем большей,

n_{min} – объем меньшей выборки объектов разного качества.

В нашем случае поправка равна 0, так как группы имеют одинаковый объем (по 10). Теперь можно записать уравнение дискриминации в полном виде:

$Z = 0.021423 \cdot x + 0.015335 \cdot y - 2.0158$.

Рассчитаем с его помощью значения нового признака «относительные размеры тела» для конкретных особей. Для первого самца величина разницы составит:

$Z_{11} = 0.021423 \cdot 65 + 0.015335 \cdot 50 - 2.0158 = 0.14.$

Значения для всех остальных особей занесены в таблицу 9.5, из которой видно, что самцы имеют положительные, а самки (кроме одной) – отрицательные значения функции *Z*. Распределения нового признака перекрываются на одну двадцатую часть, всего на 5%. По исходным данным видно, что трансгрессия по признаку *x* составила 10% (значение 61), а по признаку *y* – 25% (значения 46 и 46.5). Таким образом, рассчитанный признак характеризуется меньшей трансгрессией по сравнению с реальными признаками, т. е. позволяет снизить число неверных определений пола у живых птиц. Дальнейшие операции, связанные с использованием дискриминантной функции, вполне очевидны. Для особи с неизвестным полом, но известными промерами частей тела (когда птица после отлова и взятия промеров отпускается живой) вычисляется значение функции. Если оно больше 0, значит, это самец, если меньше – самка.

Заключительный этап – оценка достоверности уравнения по критерию Фишера:

$$F = \frac{(N-3) \cdot n_1 \cdot n_2}{2 \cdot (N-2) \cdot n} \cdot D \sim F_{(\alpha,\, 2,\, n-3)}.$$

В нашем случае

$$F = \frac{17 \cdot 20 \cdot 20}{2 \cdot 18 \cdot 20} \cdot 0.138 = 0.32.$$

Полученное значение критерия Фишера (0.32) меньше табличного (табл. 7*П)* для $\alpha = 0.05$ и $df_1 = 2$, $df_2 = 20–3 = 17$ $F_{(\alpha,\, 2, n-3)} = 3.6$, значит, уравнение недостоверно. Это объясняется небольшим объемом выборки в нашем примере: для исходных данных из 50 экз. птиц каждого пола (обычный объем зоологического материала) критерий Фишера был равен: $F = 4.2$ при $F_{(\alpha,\, 2,\, 47)} = 3.1$. Отсюда следует, что уравнение дискриминации для 50 особей достоверно и вполне пригодно для прижизненного определения пола пеночек-весничек.

Уверенность в результатах анализа может придать оценка работоспособности дискриминантной функции на независимой проверочной выборке особей с известным статусом.

Основы метода главных компонент

Метод главных компонент реализует идеологию многомерной статистики – стремление дать максимально полную характеристику каждому объекту измерения с помощью минимального числа неких расчетных признаков. Компонентный анализ позволяет вместо многочисленных исходных характеристик объектов исследования рассчитать несколько новых признаков, *линейных индексов* (названных главными компонентами), т. е. максимально эффективно справляется с задачей сокращения размерности. С вычислительной точки зрения количество главных компонент может быть равно числу исходных (m), но обычно основная доля информации об отличиях объектов «концентрируется» в гораздо меньшем числе компонент, которые и рассматриваются как полноценная характеристика всех объектов.

Главные компоненты как факторы

Зачем же делать такую подмену одних признаков другими? Дело в том, что новые показатели – это не совсем «признаки», характеристики объектов. С бóльшим основанием их можно назвать «явлениями»; это отображения неких процессов (или факторов), затрагивающих сразу группы признаков объектов измерения.

Взять, к примеру, индивидуальный рост животных, который сказывается и на размерах тела, и на массе особи, ее внутренних органов, степени развития генеративных органов, интенсивности обменных процессов и т. д. Опыт показывает, что в выборке разновозрастных животных одна из главных компонент формируется при участии всех этих признаков и поэтому может быть названа «возрастные изменения», т. е. как явление, а не признак. Что же могут представлять из себя другие главные компоненты, какие явления они могут описывать, какие общие направления изменчивости? Таким направлением может быть, например, половой диморфизм по многим признакам: самки отличаются от самцов и массой, и размерами, и пропорциями, и степенью гипертрофии органов при беременности и т. д. Это вторая причина изменчивости затрагивает те же признаки, что и онтогенез, но «в другом направлении». Наконец, если рост и развитие разных особей проходили в разных условиях

(разные сезоны, районы ареала, антропогенной пресс), они не могли не сказаться на морфологии, но своим, особенным образом – третья причина.

Эта логика приводит нас к двойственному заключению:

– каждый реальный признак характеризует только какую-то одну сторону явлений, которыми захвачены особи,
– каждое из этих явлений (факторов) сказывается на многих признаках.

Получается, что в значении каждой отдельной варианты воплощается реализация нескольких разнородных процессов; модель значения варианты любого *исходного* признака имеет вид:

$$x = x_a + x_b + x_c + \ldots,$$

где x – исходное значение какого-либо признака x,

x_a – выражение процесса ***a*** в формировании варианты x,

x_b – роль процесса ***b*** в формировании значения варианты x.

Понятно, что разные факторы будут оказывать на варианты разное влияние, одни более сильное, другие более слабое. Например, из рассмотренных выше возможных отличий вариант, воплощенных в конечном признаке каждой особи (пусть это будет масса тела), наибольшую роль сыграет возраст, а также половой диморфизм, затем условия развития, индивидуальные отличия и пр., т. е. $a > b > c > \ldots$

Если попытаться выразить массу какого-либо мелкого животного (например, обыкновенной гадюки) предложенным способом, получим:

$$W_i = W_{вид} \pm W_{пол} \pm W_{возраст} \pm W_{плод} \pm W_{сезон} \pm \ldots,$$

где W_i – значение массы отдельной i-й особи,

$W_{вид}$ – вклад в значение массы видовой нормы (средней) (примерно 50 г),

$W_{пол}$ – вклад в значение массы половых отличий (±50 г),

$W_{возраст}$ – вклад в значение массы этапа онтогенеза (±80 г),

$W_{плод}$ – вклад в значение массы наличие эмбрионов (±60 г),

$W_{сезон}$ – вклад в значение массы сезона (упитанности, развития) (отличия до 50 г).

Так, для молодого половозрелого самца гадюки летом имеем примерно

$W = 50 + 40 – 20 + 0 + 0 = 70$ г,

для старой беременной самки летом
$W = 50 + 100 + 100 + 30 – 20 = 260$ г,
для трехлетней ювенальной особи весной
$W = 50 + 0 – 30 + 0 + 0 = 20$ г.

Пример показывает, благодаря действию каких причин отличаются размеры животных, какие *направления изменчивости* реализованы в этих вариантах; в порядке возрастания значимости это:

– отличия по возрасту,
– отличия по полу,
– отличия по участию в размножении,
– отличия по сезону (упитанности).

Важно указать, что «видовая норма», определенная комплексом процессов, определяющих типичные для вида размерные характеристики (условная средняя), дает одинаковый вклад во все значения вариант; вклады остальных причин для каждой особи различны.

Аналогичным образом можно расписать влияние названных причин на любой другой признак, например, на линейные размеры тех же гадюк:

$$Lt_i = Lt_{вид} \pm Lt_{пол} \pm Lt_{возраст} \pm Lt_{плод} \pm Lt_{сезон} \pm \ldots \text{ и т. п.}$$

Итак, одни и те же процессы (факторы) сказываются на разных количественных характеристиках изучаемых объектов, при этом на разные варианты один и тот же фактор воздействует с разной силой. Сила действия данного фактора может быть, видимо, определена по величине соответствующей «добавки» к значению варианты.

Такой «декомпозирующий» взгляд на числа в матрице исходных данных позволяет предложить принцип поиска и количественной характеристики общих причин, ответственных за отличия объектов выборки. Используя информацию, заключенную в исходной матрице данных, в рамках компонентного анализа предлагается количественно выразить факторы, ответственные за отличия объектов. Данный l-й фактор можно представить как сумму всех эффектов (x_{lj}) его воздействия во все изучаемые признаки ($x_1,\ldots x_j,\ldots x_m$), т. е. как сумму всех «добавок» данного фактора во все значения отдельных признаков отдельной особи:

$x_1 = x_{a1} + x_{b1} + x_{c1} \ldots$ (x_1 как сумма вкладов разных факторов в первый признак)

$x_j = x_{aj} + x_{bj} + \ldots + x_{lj} \ldots,$

$x_m = x_{am} + x_{bm} + \ldots + x_{mj}$ (x_m как сумма вкладов разных факторов в m-й признак)

$ГК_a = x_{a1} + \ldots + x_{aj} + \ldots + x_{am}$ – сумма вкладов одного фактора в значения всех признаков,

где $ГК_a$ – значение **г**лавной **к**омпоненты, характеризующей действие одного из процессов формирования вариант (фактор ***a***),
x_{aj} – вклад фактора ***a*** в значение варианты j-го признака данного объекта.

Для другого процесса (фактор ***b***) имеем:

$$ГК_b = x_{b1} + x_{b2} + \ldots + x_{bj} + \ldots + x_{bm}$$

и т. д. для всех прочих факторов.

Например, как показывает практика, первой главной компонентой в выборке животных обычно оказывается фактор возрастных отличий, что позволяет записать примерное выражение:

$$ГК_{возраст} = норм.W_{возраст} + норм.Lt_{возраст} + норм.Lc_{возраст} + \ldots.$$

Иными словами, главная компонента, характеризующая действие возраста, представляет собой сумму соответствующих долей вариант по всем признакам.

Конечно, странно и неправильно было бы складывать граммы с миллиметрами и миллиграммами, поэтому в уравнении присутствует префикс *норм.*, говорящий о том, что в расчетах принимают участие значения, предварительно преобразованные к виду, позволяющему проводить такие операции. Эти значения центрированы (к средней) и нормированы (на стандартное отклонение):

$$норм.\ x_{ji} = z_{ji} = (x_{ji} - M_j) / S_j,$$

где $норм.\ x_{ji}$, или z_{ji}, – нормированное i-е значение j-го признака,
M_j, S_j – средняя и стандартное отклонение j-го признака по всей выборке,
i – индекс объекта, особи,
j – индекс признака.

После нормирования признаки утрачивают единицы измерения и складывать их значения вполне допустимо.

Требование максимума дисперсии

Представленным выше способом формируется столько главных компонент, сколько существенных причин участвовало в формировании вариант. Теперь можно детальнее показать, почему количество расчетных признаков (главных компонент) должно быть меньше, чем число исходных переменных. На выборке объектов можно часто наблюдать, как от объекта к объекту разные признаки изменяются чуть ли не синхронно, т. е. сходным образом реагируют на одни и те же факторы. Факт корреляции между признаками означает, что они содержат много общей информации о действующих факторах. При этом каждый отдельный фактор влияет на несколько признаков. Главные компоненты как раз и выражают эти немногие причины изменчивости, которых всегда меньше, чем исходных признаков.

Получается, что 100% информации об изменчивости вариант, заключенной в исходной матрице данных, перераспределяется между компонентами по-иному, чем между признаками. Например, когда изучается 10 признаков, можно условно принять, что каждый из них привносит по 10% информации. Пусть при этом *половина* значения каждой варианты каждого признака будет изменяться у разных особей под действием одной причины (например, возраста), тогда на долю главной компоненты, которая уловит эти возрастные отличия, придется 50% общей информации; она будет в пять раз более информативна, чем любой исходный признак. Аналогично можно представить, что на половые отличия придется 30% информации (изменчивости значений вариант), на отличия по срокам наблюдения – 10%, а на все прочие более слабые причины – оставшиеся 10%. Эти 10–20% относятся, как правило, к стохастическому шуму (слабые несущественные факторы, ошибки измерения), их обычно не рассматривают. В итоге можно увидеть, что вместо 10 признаков львиную долю общей изменчивости вариант отобразили, «объяснили» всего 3 главных компоненты.

В рамках компонентного анализа «сила» каждой компоненты (характеристики некоего фактора) оценивается как доля дисперсии

данной компоненты в общей дисперсии признаков, составляющей 100% (этот принцип, по существу, заимствован из дисперсионного анализа). Как уже говорилось, количество информации в многомерной статистике выражается степенью отличия объектов друг от друга, т. е. общей дисперсией их значений ($S^2_{ГКj}$). Эта общая по всем признакам дисперсия перераспределяется между разными компонентами. (В публикациях можно найти выражения вроде «доля дисперсии первой главной компоненты составляет 34%»; буквально это означает, что относительная сила влияния некоего фактора, выраженного этой компонентой, составляет 34%.) Процедура расчета главных компонент организована таким образом, что первыми описываются самые сильные влияния, действие самого сильного фактора, т. е. чтобы дисперсия первой компоненты имела наибольшее значение. Затем вычисляются оценки действия второго по значимости фактора, с меньшей дисперсией, и так далее в порядке уменьшения величины дисперсии главных компонент:

$$S^2_{ГК1} > S^2_{ГК2} > ... > S^2_{ГКl} > ... > S^2_{ГКm}.$$

Факторные нагрузки

Подходя к рассмотрению техники расчетов главных компонент, выразим их модель с использованием не абсолютных значений вкладов разных признаков, но относительных. Если значение отдельной варианты есть сумма вкладов разных факторов $x_j = x_{aj} + x_{bj} + x_{cj}$, то величина вклада в значение варианты отдельного фактора составит некую долю от общего значения варианты:

$$x_{aj} = a_j \cdot x_j,$$

где a_j – относительный вклад данного фактора в конечное значение варианты,

x_j – значение варианты признака j.

Используя это преобразование, а также исходную формулу $ГК_a = x_{a1} + ... + x_{aj} + ... + x_{am}$, получаем уравнение первой главной компоненты:

$$ГК_a = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_i \cdot x_j + ... + a_m \cdot x_m,$$

второй:

$$ГК_b = b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_i \cdot x_j + ... + b_m \cdot x_m$$ и так далее.

Общая модель компонентного анализа примет вид:

$$ГК_l = a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_i \cdot x_i + ... + a_m \cdot x_m ,$$

где l – номер компоненты, l = 1, 2,… k (значимых компонент всегда меньше, чем признаков, $k \leq m$).

Как же практически можно определить, какую долю каких факторов содержит в себе каждое значение исходных признаков, т. е. чему равны конкретные значения коэффициентов a_j (факторных нагрузок) и как их вычислить? Для упрощения объяснения на первых порах придется несколько пожертвовать строгостью понятий.

Сначала зададимся более простым вопросом – как определить долю участия некоего внешнего фактора в каких-либо двух изучаемых признаках (например, масса и размеры особи)? Если некий фактор будет действовать на оба признака одновременно, это значит, что изменения значений вариант от объекта к объекту будут происходить более или менее синхронно, сопряженно. Поскольку известно, что сопряженное варьирование двух признаков лучше всего оценивать с помощью корреляционного анализа, значит, коэффициент корреляции и покажет, чтó в варьировании двух признаков есть общего и какова степень этой общности. Корреляция на уровне r = 1 свидетельствует о том, что оба изучаемых признака абсолютно детерминированы друг другом или единственной внешней причиной. Говоря упрощенно, коэффициент корреляции r = 0.5 свидетельствует, что примерно половинная доля значений каждой из вариант обоих признаков определяется действием некоего *общего фактора*, а другие «половинки значений» сформированы под влиянием иных обстоятельств. Такой уровень корреляции как раз характерен для связи вес – размеры особи. Любой коэффициент корреляции будет отражать то общее, что есть между каждой парой изучаемых признаков, что заставляет их сопряженно изменяться от варианты к варианте.

Коэффициенты в уравнениях главных компонент – это аналоги коэффициентов корреляции между признаками, они названы *факторными нагрузками* (отличия между коэффициентами корреляции и факторными нагрузками показаны ниже). Это удачное название показывает, во-первых, какой эффект данный l-й фактор оказал на данный j-й исходный признак, а также, во-вторых, какой вклад вносит данный признак в значение данной главной компонен-

ты. Итак, факторные нагрузки есть *аналоги* коэффициентов корреляции между признаками (например, между первым признаком и всеми остальными, r_{1i}); это позволяет записать примерную формулу:

$$КК_j \approx r_{11} \cdot x_1 + r_{12} \cdot x_2 + ... + r_{1i} \cdot x_i + ... + r_{1m} \cdot x_m .$$

Расчет корреляционных компонент

Используя наши данные по гадюкам (*W* – масса тела, *Lt* – длина тела, *Lc* – длина хвоста), рассмотрим расчет таких «корреляционных компонент», аналогов главных компонент (табл. 9.6, 9.7). Если рассчитать корреляции между тремя признаками, получим всего шесть коэффициентов (включая автозависимости, $r = 1.00$).

Таблица 9.6

Матрица корреляций			
	W	*Lt*	*Lc*
W	1.00	0.79	-0.49
Lt	0.79	1.00	-0.33
Lc	-0.49	-0.33	1.00

Возьмем в качестве факторных (точнее, «корреляционных») нагрузок первый столбец коэффициентов, выражающий сопряжение трех признаков с массой тела змеи: $r_{11} = r_{WW} = 1.00$, $r_{12} = r_{WLt} = 0.789$, $r_{13} = r_{WLc} = –0.492$. Тогда уравнение первой корреляционной компоненты примет вид:

$$КК_1 = 1.00 \cdot W + 0.789 \cdot Lt - 0.492 \cdot Lc .$$

Рассчитаем значения компонент, новых признаков, для конкретных особей (для простоты обойдемся без нормирования); для первого самца:

$$КК_{1_1} = 1.00 \cdot 40 + 0.789 \cdot 45 - 0.492 \cdot 77 = 37.6 ,$$

для последней самки

$$КК_{1_{17}} = 1.00 \cdot 112 + 0.789 \cdot 57 - 0.492 \cdot 70 = 122.6 .$$

Таблица 9.7

Исходные данные					**«Очищенные» данные**			
пол	***W***	***Lt***	***Lc***	$\boldsymbol{KK_1}$	$\boldsymbol{W-KK_1}$	$\boldsymbol{Lt-KK_1}$	$\boldsymbol{Lc-KK_1}$	$\boldsymbol{KK_2}$
m_1	40	45	77	37.6	2.4	7.4	39.4	48.7
m_2	43	46	84	38.0	5.0	8.0	46.0	58.5
m_3	45	47	81	42.3	2.7	4.7	38.7	45.9
m_4	48	45	76	46.1	1.9	–1.1	29.9	30.5
m_5	53	47	80	50.7	2.3	–3.7	29.3	27.7
m_6	65	50	78	66.1	–1.1	–16.1	11.9	–4.8
m_7	68	53	90	65.6	2.4	–12.6	24.4	14.4
m_8	70	51	87	67.5	2.5	–16.5	19.5	5.8
f_9	60	50	62	69.0	–9.0	–19.0	–7.0	–33.8
f_{10}	61	55	65	72.4	–11.4	–17.4	–7.4	–35.0
f_{11}	68	49	65	74.7	–6.7	–25.7	–9.7	–41.0
f_{12}	77	51	66	84.8	–7.8	–33.8	–18.8	–59.0
f_{13}	82	52	64	91.6	–9.6	–39.6	–27.6	–75.0
f_{14}	82	50.5	64	90.4	–8.4	–39.9	–26.4	–73.1
f_{15}	90	53	68	98.4	–8.4	–45.4	–30.4	–82.4
f_{16}	100	51	62	109.8	–9.8	–58.8	–47.8	–114.1
f_{17}	112	57	70	122.6	–10.6	–65.6	–52.6	–126.3
M	68.47	50.15	72.88					
S	20.31	3.39	9.29					

Что же дают нам эти первые результаты? Значения главных компонент, новых признаков, обозначает одно общее направление изменчивости, характерное для всех морфологических признаков – это увеличение размеров тела с возрастом. Ход графика первой корреляционной компоненты (KK_1) в общих чертах совпадает с ходом графика изменения массы (*W*) и длины (*Lt*) тела; эта компонента, по существу, *подменяет* собой два исходных признака, ее можно назвать общим термином «размеры особи». Факторные нагрузки (табл. 9.6) для этих двух признаков велики и положительны. Третий же признак дает отрицательный вклад в первую компоненту, отделяя себя от прочих. Есть все основания считать, что он характерен для какого-то иного направления изменчивости. (В нашем примере

– это половые отличия: у самок хвосты короче.) Таким образом, на первом этапе удалось выделить одно направление изменчивости и наметить другое. Конкретизируем его с помощью второй главной компоненты.

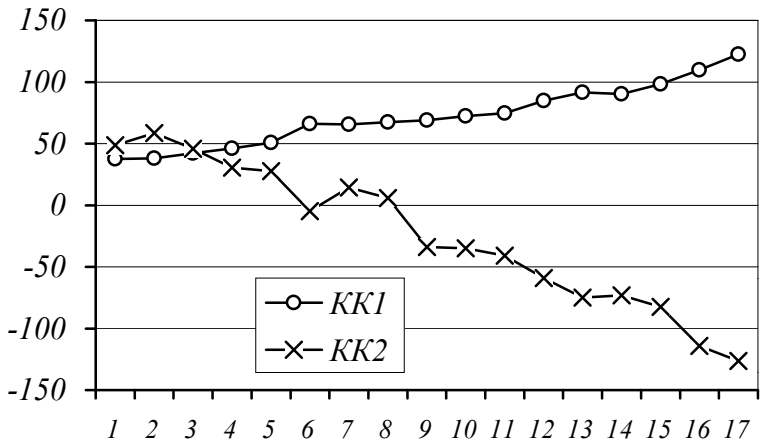


Рис. 9.5. *Корреляционные компоненты*

Откуда же взять значения факторных нагрузок во второй и следующих компонентах? Ведь они должны быть другими, поскольку, по определению, следующие компоненты должны характеризовать другие направления изменчивости вариант, другие факторы!

Здесь компонентный анализ идет по пути расчета частных коэффициентов корреляции. Общий коэффициент корреляции отражает сопряженное варьирование признаков только относительно самого сильного общего фактора, тогда как эффекты действия более слабых факторов (иных направлений изменчивости) затушевываются. Чтобы выявить оставшиеся направления изменчивости, нужно удалить эффект главного фактора! Для этого из всех значений вариант следует, условно говоря, «вычесть» долю, обусловленную этим самым сильным фактором. Для нашего примера попробуем поступить грубо и от значений исходных признаков непосредственно вычтем значение первой главной компоненты: $x'_{ij} = x_{ij} - ГК_1$.

Оставшаяся часть значения каждого признака будет отражать действие всех прочих причин, кроме первой. Если теперь рассчитать корреляцию для вариант, «очищенных» от влияния первого фактора, то корреляция между признаками должна показать их сопряженное изменение относительно другого, второго по силе фактора. Понятно, что корреляционная структура «очищенной» матрицы данных будет совершенно другой, нежели у исходной: все зависимости оказались высокими и положительными ($r > +0.9$) (табл. 9.8).

Для расчета значений второй компоненты в качестве факторных нагрузок возьмем коэффициенты корреляции с опорой на признак ($Lc - KK_1$) (табл. 9.8).

Эти новые коэффициенты корреляции сыграют роль факторных нагрузок для уравнения второй корреляционной компоненты:

$$KK_2 = 0.976 \cdot (W - KK_1) + 0.923 \cdot (Lt - KK_1) + 1.00 \cdot (Lc - KK_1);$$

расчеты значений этой компоненты для конкретных особей приведены в табл. 9.7.

Таблица 9.8

Матрица корреляций			
	$W - KK_1$	$Lt - KK_1$	$Lc - KK_1$
$W - KK_1$	1.00	0.822	0.976
$Lt - KK_1$	0.822	1.00	0.923
$Lc - KK_1$	0.976	0.923	1.00

Судя по графику хода второй компоненты (рис. 9.5), она в первую очередь «пытается» отследить и усилить второе направление изменчивости данных – отличие самцов (особи № 1–8) и самок (особи № 9–17) по длине хвоста: у самок хвост короче, чем у самцов. Как показывают факторные нагрузки, признаку «длина хвоста» (1.00) в этом помогают переменные «масса» (0.976) и «длина тела» (0.923). Итак, вторая компонента обозначила другой внутренний фактор отличия особей, изменчивость по длине хвоста, половой диморфизм.

Требование ортогональности компонент

В рамках компонентного анализа рассмотренная процедура «вычитания» информации о влиянии отдельного фактора из общей информации об изменчивости вариант имеет одно важное условие, специально оговоренное и обязательно выполняемое при вычислениях. Компоненты должны быть ортогональны, т. е. вовсе не должны коррелировать друг с другом:

$$r_{ГК_{il}} = 0.$$

Идеологически это понятно: исходные значения должны полностью утратить «след» первого учтенного фактора, чтобы можно было оценивать роль второго; информация, которая воплотится в следующую компоненту, должна быть полностью независима от предыдущей компоненты. Обязательное отсутствие корреляции между компонентами гарантирует, что каждая из главных компонент содержит уникальную информацию об обособленном направлении изменчивости признаков.

Поскольку в проведенных выше расчетах это условие специально не выдерживалось, оказалось, что корреляционные компоненты существенно не ортогональны: $r_{КК_{12}} = -0.98$. Судя по высокому отрицательному коэффициенту, здесь явно проявляется ложная корреляция (см. раздел 8, с. 201) как результат вычитания общего значения ($-K_1$).

Процедура компонентного анализа не имеет такого недостатка, поскольку «вычленение» информации, учтенной главной компонентой, выполняется непосредственно из матрицы коэффициентов корреляции: из каждого общего коэффициента корреляции вычисляется коэффициент корреляции между теми долями вариант, которые сформированы под действием первого фактора. Детали этого вычислительного процесса не так и важны, главное, что обновленная матрица корреляций полностью утрачивает информацию о первом факторе и никаких ложных корреляций не появляется.

Компонентный анализ

Рассмотрим результаты собственно компонентного анализа (табл. 9.9, 9.10), выполненного для исходных данных по размерам гадюки (табл. 9.7) в среде пакета StatGraphics. Условие ортогональности выполнено, компоненты независимы (с точностью до ошибки округления): $r_{ГК_{12}} = -0.00002$.

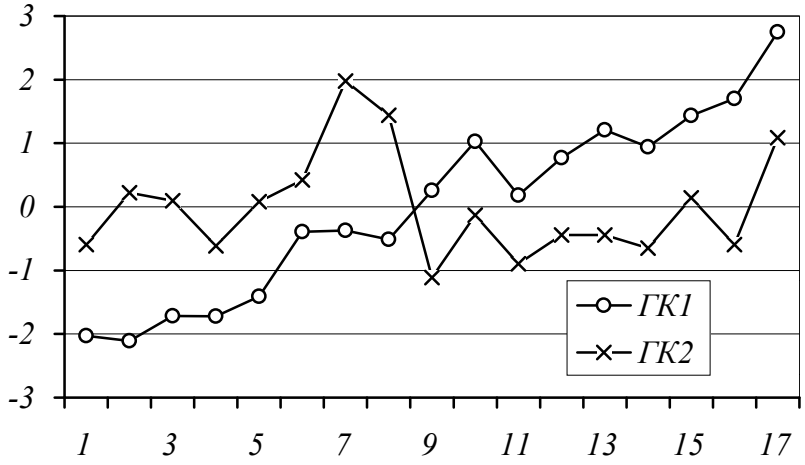


Рис. 9.6. *Главные компоненты*

Таблица 9.9

Факторные нагрузки			
	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	2.10	0.71	0.19
Дисперсия, %	70	24	6

Исходя из полученных факторных нагрузок, уравнение первой главной компоненты имеет вид:

$$ГК_1 = 0.644 \cdot норм.W + 0.603 \cdot норм.Lt - 0.470 \cdot норм.Lc\,.$$

Используя его, рассчитаем значения компонент, новых признаков, для конкретных особей, помня, что вместо исходных значений берутся нормированные. Таблица 9.10 содержит нормированные значения (*норм.X = (X–M)/S*); в частности, для первого самца (40 г) получаем *норм.W* = (40–68.47)/20.31 = –1.40). Параметры *M, S* взяты из таблицы 9.7. Значения первой компоненты составят:
для первого самца

$$ГК_{1_1} = 0.644 \cdot (-1.44) + 0.603 \cdot (-1.52) + 0.470 \cdot (0.44) = -2.026,$$

для последней самки

$$ГК_{1_{17}} = 1.00 \cdot (2.14) + 0.789 \cdot (2.02) - 0.492 \cdot (-0.1) = 2.74.$$

Таблица 9.10

***пол*№**	***норм.W***	***норм.Lt***	***норм.Lc***	***ГК*1**	***ГК*2**	***ГК*3**
m_1	–1.40	–1.52	0.44	–2.027	–0.596	0.024
m_2	–1.25	–1.22	1.20	–2.108	0.220	0.082
m_3	–1.16	–0.93	0.87	–1.715	0.098	–0.095
m_4	–1.01	–1.52	0.34	–1.723	–0.614	0.295
m_5	–0.76	–0.93	0.77	–1.411	0.081	0.177
m_6	–0.17	–0.04	0.55	–0.395	0.422	0.004
m_7	–0.02	0.84	1.84	–0.372	1.978	–0.218
m_8	0.08	0.25	1.52	–0.513	1.442	0.175
f_9	–0.42	–0.04	–1.17	0.255	–1.110	–0.498
f_{10}	–0.37	1.43	–0.85	1.026	–0.132	–1.354
f_{11}	–0.02	–0.34	–0.85	0.179	–0.894	0.044
f_{12}	0.42	0.25	–0.74	0.770	–0.441	0.011
f_{13}	0.67	0.55	–0.96	1.208	–0.441	–0.037
f_{14}	0.67	0.10	–0.96	0.941	–0.648	0.249
f_{15}	1.06	0.84	–0.52	1.437	0.143	0.145
f_{16}	1.55	0.25	–1.17	1.701	–0.595	0.770
f_{17}	2.14	2.02	–0.31	2.746	1.088	0.226
M	0	0	0	0	0	0
S^2	1	1	1	2.0997	0.711	0.189

Для первой компоненты «корреляционные» (табл. 9.6) и факторные нагрузки (табл. 9.9) очень близки и отражают рассмотренное

явление – противопоставление общих размеров тела (большие положительные корреляции) длине хвоста (большие отрицательные корреляции). График первой главной компоненты (рис. 9.6) также похож на график первой корреляционной компоненты (рис. 9.5) и характеризует «общие размеры тела» (объединяя два признака – *W* и *Lt*). В то же время достаточно высокий вклад переменной «длина хвоста» (–0.47) заставляет включить и этот признак в название компоненты, обозначая направление изменчивости «рост размеров при уменьшении хвоста».

Вторая главная компонента отличается от своего корреляционного аналога. Нагрузка для переменной «длина хвоста» остается высокой (0.863), но для первых двух признаков значения факторных нагрузок существенно ниже корреляционных (0.191 и 0.467 против 0.976 и 0.923). Эти небольшие коэффициенты свидетельствуют о том, что половой диморфизм сказывается и на общих размерах тела, но в меньшей степени, чем размер хвоста. Причины несовпадения коэффициентов корреляции и факторных нагрузок состоят в том, что первичные коэффициенты корреляции отражают, вообще говоря, действия всего множества факторов сопряженного варьирования исходных признаков, «смесь». Сильные факторы определяют уровень коррелированности в большей мере, слабые – в меньшей. Факторные же нагрузки вычленяют эффект действия своего фактора «в чистом виде». Изменчивость второй главной компоненты менее определена, чем второй корреляционной компоненты. Однако вместе с первой они хорошо дифференцируют особей разного пола на две изолированные группы: в осях двух главных компонент самки «расположены» справа внизу, самцы – слева вверху.

Информативность и значимость компонент

Следует отметить, что участие двух компонент в дифференциации объектов неодинаково. Первая компонента имеет наибольшую дисперсию (2.1) и на 70% исчерпала информацию об изменчивости признаков (табл. 9.9), тогда как на долю второй приходится всего 24%. Получается, что роль этой компоненты ниже, чем роль любого из исходных признаков (на каждый из них приходится по 33%), и вторая компонента (как и третья) не должна бы участвовать в дальнейшем рассмотрении. В компонентном анализе обычно ис-

пользуется содержательный критерий значимости: компоненты с дисперсией менее 1 не рассматриваются.

Это справедливо для небольших объемов выборок (десятки объектов), но для обширных выборок может оказаться неверным. Для этого предлагается формальный критерий оценки значимости компонент, проверяющий нулевую гипотезу о равенстве дисперсий k компонент:

$S^2_i = S^2_{i+1} = \ldots = S^2_k$.

Если дисперсии компонент равны, значит, они не используют общей информации о коррелированности исходных признаков, не являются общими факторами, не сказываются на признаках, т. е. незначимы. Критерий имеет χ^2-распределение с $df = \frac{k \cdot (k+1)}{2} - 1$ степенями свободы:

$$\chi^2 = -(n-1) \cdot \sum_{j=i}^{k} \ln S_j + k \cdot (n+1) \cdot \ln \frac{\sum_{j=i}^{k} S_j}{k} \sim \chi^2_{(\alpha,df)},$$

где n – объем выборки,

k – число рассматриваемых компонент,

i – номер начальной учитываемой компоненты,

S^2 – дисперсия компоненты.

Проверим гипотезу о равенстве трех главных компонент. Для уровня значимости $\alpha = 0.05$ имеем: $i = 1$, $k = 3$, $df = 7$, $n = 17$, $\chi^2_{(0.05,7)} = 4.07$, $S^2_1 = 2.1$, $S^2_2 = 0.71$, $S^2_3 = 0.19$,

$$\chi^2 = -(17-1) \cdot \sum_{j=1}^{3} \ln S^2_j + 3 \cdot (17+1) \cdot \ln \frac{\sum_{j=1}^{3} S^2_j}{3} = 20.18.$$

Полученное значение (20.18) больше табличного (14.07), следовательно, дисперсии трех компонент отличаются достоверно.

Проверяя гипотезу о равенстве второй и третьей компонент, получим: $i = 1$, $k = 2$, $df = 2$, $n = 17$, $\chi^2 = 5.99$, $S^2_2 = 0.71$, $S^2_3 = 0.19$,

$$\chi^2 = -(17-1) \cdot \sum_{j=2}^{2} \ln S^2_j + 2 \cdot (17+1) \cdot \ln \frac{\sum_{j=2}^{2} S^2_j}{2} = 4.9.$$

На сей раз полученное значение (4.9) меньше табличного (5.99), дисперсии второй и третьей компонент отличаются недостоверно.

Вывод очевиден: первая (значимая) компонента выделяется среди прочих (незначимых) компонент по информационной насыщенности. Специфика исходных трех переменных воплотилась в единственный расчетный признак – первую главную компоненту.

Этапы компонентного анализа

Метод главных компонент достаточно сложен, но это самая эффективная процедура разведочного анализа любой многомерной совокупности данных, имеющая примерно семь крупных шагов:

1) организация массива данных с метками объектов и именами переменных,
2) изучение направлений изменчивости исходных признаков,
3) выполнение расчетов в среде специальных пакетов (StatGraphics),
4) изучение факторных нагрузок,
5) изучение ординации объектов в осях значимых главных компонент,
6) присвоение названий значимым компонентам,
7) вывод об основных направлениях (факторах) изменчивости данных,
8) отсев или отбор признаков и повторение расчетов; итерации позволяют глубже понять структуру связей между признаками.

Поэтапно проанализируем данные по динамике снеготаяния на прибайкальской равнине в зоне действия Байкальского целлюлозно-бумажного комбината, который имеет большие объемы пылегазовых выбросов.

1) Глубину снега (h, см) измеряли в 9 точках Прибайкальской равнины вдоль побережья оз. Байкал 4 раза за сезон с 21 апреля по 18 мая 1986 г. (табл. 9.11).

2) Данные показывают, что с запада на восток уровень снега в среднем постепенно повышается, достигая в некоторых точках (85 км) глубины $h_{21.4.86}$ = 110 см. При этом для начала весны (21.4.86) отмечается плавное повышение уровня снега с запада на

восток, а к концу (18.5.86) становятся заметны резкие перепады между отдельными точками.

Таблица 9.11

Расстояние запад – восток, км	**21.04.86**	**02.05.86**	**11.05.86**	**18.05.86**	**Средняя**	$ГК_1$	$ГК_2$
0	5	0	0	0	1.3	–3.2	0.7
20	55	40	25	20	35.0	–1.3	0.2
32	55	35	10	1	25.3	–1.9	–0.3
39	95	80	70	30	68.8	0.7	–0.3
33 (БЦБК)	75	55	15	0	36.3	–1.2	–0.8
35	105	95	85	70	88.8	1.8	0.4
45	125	110	85	75	98.8	2.4	0.01
75	110	80	60	60	77.5	1.2	0.04
85	110	85	70	65	82.5	1.5	0.2

3) Порядок расчетов в StatGraphics рассмотрен на с. 251.

4) В результате расчетов получены коэффициенты линейных индексов (факторные нагрузки) (табл. 9.12), позволяющие рассчитать значения главных компонент по формулам вида:

$$ГК_1 = 0.49 \cdot h_{21.04.86} + 0.51 \cdot h_{2.05.86} + 0.50 \cdot h_{11.05.86} + 0.49 \cdot h_{18.05.86},$$

$$ГК_2 = -0.55 \cdot h_{21.04.86} - 0.38 \cdot h_{2.05.86} + 0.26 \cdot h_{11.05.86} + 0.69 \cdot h_{18.05.86}$$ и т. д.

Таблица 9.12

Дата	a_1	a_2	a_3	a_4
21.04.86	0.49	–0.55	–0.38	0.54
02.05.86	0.51	–0.38	0.13	–0.76
11.05.86	0.50	0.26	0.75	0.33
18.05.86	0.49	0.69	–0.52	–0.11
S^2	3.741	0.191	0.059	0.008
S^2, %	93.5	4.8	1.5	0.2

Первая главная компонента имеет бóльшую дисперсию (3.7 из 4), т. е. забирает на себя бóльшую часть информации (93.5%); остатки почти целиком приходятся на вторую компоненту (4.8%).

Очевидно, что при данном количестве наблюдений вторая компонента незначима, тем не менее мы ее рассмотрим подробнее.

В первой компоненте факторные нагрузки («веса») разных признаков почти равны (по 0.5), это значит, что чем больше будут значения всех промеров, тем больше будет и значение компоненты.

Во второй главной компоненте достаточно большие факторные нагрузки имеют только первая (21.04.86) и последняя (18.05.86) даты (–0.55 и 0.69 соответственно), причем с разными знаками. Вторая компонента как бы противопоставляет зимние и весенние глубины. Максимальные значения этой компоненты будут наблюдаться для точек, в которых зимой был наименьший уровень снега, а весной – наибольший, т. е. там, где уровень снега почти не менялся. Минимальные значения должны наблюдаться для тех точек, где зимой снега было много, а весной – мало, т. е. где снег быстро сошел.

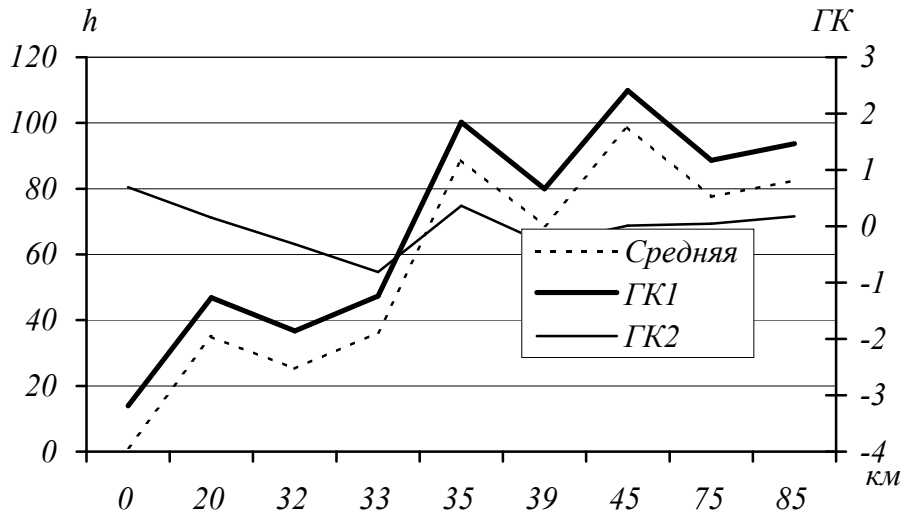


Рис. 9.7. *Компонентный анализ динамики снеготаяния*

5) Отследим значения главных компонент для отдельных точек. Значения первой компоненты велики для самых восточных точек (1.5), где максимальны все промеры снега, и минимальны для западных (–3.2), где снега почти нет. Значения второй компоненты высоки для многих пунктов промера (где снег сходил более или менее равномерно), а минимальны только для точки 33 км: здесь наблюдается резкий перепад глубин между отдельными промерами.

6) Ход первой компоненты подобен средней арифметической по всем промерам (рис. 9.7), ее можно назвать «запасы снега». Высокие значения второй компоненты выявляют зоны медленного схода снега, а низкие – быстрого, поэтому ее можно назвать «устойчивость снегового покрова весной».

7) Рассматривая явление в новых терминах, можно сказать, что в общем запасы снега на Прибайкальской равнине плавно увеличиваются с запада на восток. Для окрестностей БЦБК характерна средняя мощность, но низкая устойчивость снегового покрова. Как показали специальные исследования, причина этого явления – загрязнение поверхности пылевыми частицами, которые способствуют его нагреванию под лучами солнца и быстрому таянию.

Варианты представления результатов

Для представления результатов компонентного анализа часто используются три разных способа выражения величины факторных нагрузок.

При *первом* из них, показанном выше (табл. 9.9, 9.13), в качестве векторов факторных нагрузок выступают так называемые *собственные векторы* (техника и теория их расчетов приведена во многих пособиях, например: Коросов, 1996).

Таблица 9.13

Факторные нагрузки			
	a_1	a_2	a_3
W	0.644	0.191	0.741
Lt	0.603	0.467	–0.655
Lc	–0.47	0.863	0.186
Дисперсия, S^2	2.09	0.71	0.19

Для собственных векторов выполняется важное условие: произведение вектора на самого себя дает единицу. Так,

$$(0.644\ 0.603\ -0.470)\cdot\begin{pmatrix}0.644\\0.603\\-0.47\end{pmatrix} = 0.644\cdot 0.644 + 0.603^2 + (-0.47)^2 = 1.$$

На основании этих факторных нагрузок рассчитываются те значения главных компонент, дисперсии которых, S^2, представлены в нижней строке таблицы с результатами (табл. 9.10, 9.13). Кстати сказать, сумма дисперсий всех компонент равна числу изучаемых признаков, m (2.09+0.71+0.19 = 3). В такой форме результаты анализа выдает пакет StatGraphics.

Несмотря на прозрачный математический смысл, интерпретировать такие факторные нагрузки неудобно из-за какой-то непонятной «абсолютности» собственных векторов.

Второй способ позволяет более эффективно сопоставлять нагрузки, относящиеся к каждой компоненте *по отдельности*. Для этого все нагрузки делят на модуль максимального значения.

Так, для первого вектора $_{\max}a = 0.644$; нормированная нагрузка для признака *W* составит: 0.644/0.644 = 1.000, а для признака *Lt* – 0.603/0.644 = 0. 936 и т. д. (табл. 9.14).

Таблица 9.14

Факторные нагрузки			
	a_1	a_2	a_3
W	1.000	0.221	1.000
Lt	0.936	0.541	–0.884
Lc	–0.730	1.000	0.251
Дисперсия, S^2	2.09	0.71	0.19

В результате факторные нагрузки обретают значения от –1 до +1, их становится легче сравнивать друг с другом в контексте одной компоненты. При этом, правда, свойства векторов нагрузок меняются и их произведение на себя уже не дает значения 1. В то же время новая относительная величина позволяет применить простой критерий оценки достоверности отличия нагрузки от нуля, для этого она

должна быть по модулю больше 0.7: $|a|>0.7$. Получается, что большие коэффициенты нагрузки как бы приравниваются к единице (полный учет признака), а остальные – к нулю (признак не участвует в компоненте). Такой прием во многом облегчает первый шаг в интерпретации главных компонент. Провести рассмотренные преобразования можно в среде Excel.

Третий способ презентации результатов МГК позволяет сравнивать факторные нагрузки одного признака в *разных* главных компонентах. В качестве основания для нормирования такого рода служит стандартное отклонение конкретной компоненты *S*, на величину которой умножаются факторные нагрузки (табл. 9.15). Например, нагрузка признака *Lt* теперь составит для второй компоненты: 0.467·0.843 = 0.394, для третьей компоненты: –0.655·0.435 = –0.286.

Таблица 9.15

Факторные нагрузки			
	a_1	a_2	a_3
W	0.933	0.161	0.322
Lt	0.874	0.394	–0.286
Lc	–0.681	0.728	0.081
Дисперсия, S^2	2.09	0.71	0.19
Стандартное отклонение, *S*	1.449	0.843	0.435

Такое преобразование позволяет оценить относительную роль признака в той или иной компоненте: несмотря на относительно высокое значение исходной нагрузки признака *Lt* в третьей компоненте (-0.655) по сравнению со второй (0.467) (табл. 9.13), его рассмотрение в контексте общего варьирования говорит от обратном: фактическое влияние признака на изменчивость третьей компоненты (–0.28) меньше, чем влияние на вторую компоненту (0.394) (табл. 9.15).

Более того, новое преобразование позволяет точно вычислить, какую долю своей изменчивости каждый признак сообщает

каждой компоненте (иначе, какую долю изменчивости признака учитывает та или иная компонента). Поскольку факторные нагрузки можно рассматривать как аналоги коэффициентов корреляции, то их квадраты могут играть роль коэффициентов детерминации, выражающих как раз долю варьирования за счет действия фактора в общем варьировании признака. Как известно, общая дисперсия отдельного *нормированного* признака равна единице (см. табл. 9.10), поэтому квадраты факторных нагрузок будут представлять собой искомые доли (табл. 9.16). Так, из общей дисперсии признака *Lt* компоненты «забрали» такие доли: первая – 0.764, вторая – 0.155, третья – 0.082; при этом 0.764 + 0.155 + 0.082 = 1, или 76 + 16 + 8 = 100%. Как видно, в наибольшей степени длина тела учтена в первой компоненте.

Таблица 9.16

Признаки	a_1	a^2_1	a_2	a^2_2	a_3	a^2_3	$\sum a^2_{1\text{-}3}$
W	0.933	0.870	0.161	0.026	0.323	0.104	1.0
Lt	0.874	0.764	0.394	0.155	-0.286	0.082	1.0
Lc	-0.681	0.464	0.727	0.528	0.081	0.006	1.0
Сумма = дисперсия, S^2		2.098		0.71		0.19	3.0

Произведения преобразованных векторов факторных нагрузок на самих себя не равны единице, но – величине дисперсии (что вытекает из процедуры получения векторов):

$$(0.933\ 0.874\ -0.681)\cdot\begin{pmatrix}0.933\\0.874\\-0.681\end{pmatrix} = (0.933^2 + 0.874^2 + (-0.681)^2) = 2.098.$$

Значения главных компонент, рассчитанные с помощью преобразованных факторных нагрузок, будут отличаться от тех, что высчитаны по первой схеме. Если их дисперсии вычислить непосредственно, то они будут равны единице, а не той величине, что указа-

на в нижней строке таблицы. В такой форме результаты компонентного анализа представлены в пакете Statistica.

В заключение следует отметить, что каким бы способом представления факторных нагрузок мы ни пользовались, как бы не трансформировались значения главных компонент, все равно ординация (взаиморасположение) объектов в осях главных компонент *не меняется*! Это позволяет правильно интерпретировать компонентный анализ изменчивости признаков независимо от метода отображения его результатов.

Резюме

Компонентный анализ позволяет рассчитывать *линейные индексы* исходных признаков (главные компоненты), используя в качестве коэффициентов пропорциональности *факторные нагрузки*. Процедура расчетов линейных индексов, главных компонент, строится на выполнении следующих условий:

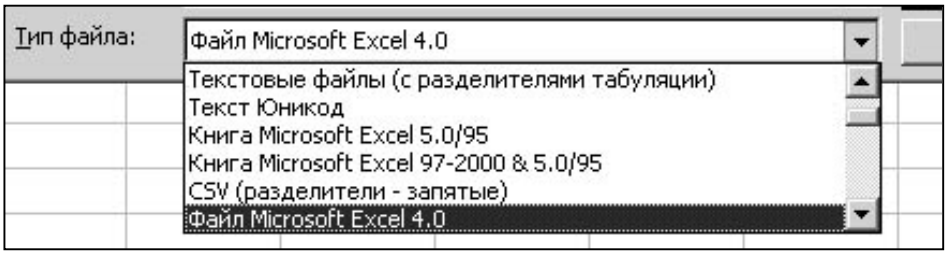
- факторные нагрузки отражают корреляцию исходных признаков,
- компоненты ортогональны, т. е. не коррелируют друг с другом,
- дисперсия следующей компоненты меньше, чем предыдущей.

Выполнение этих требований достигается в процессе многократно повторяющейся (итеративной) процедуры «подгонки» результатов вычислений под выдвинутые требования.

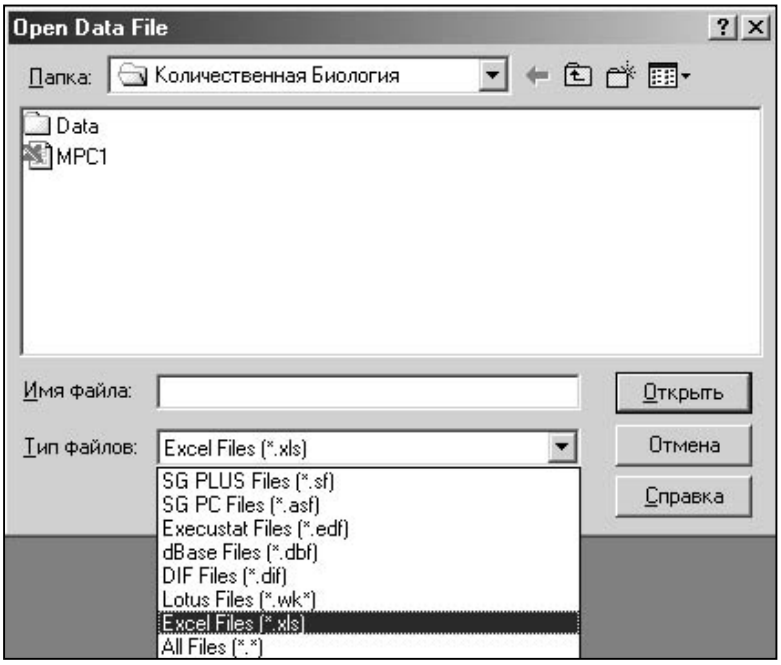
Компонентный анализ дает два основных итога. Во-первых, с его помощью удается выяснить тонкую структуру зависимостей переменных друг от друга и от общих причин, т. е. установить *состав корреляционных плеяд признаков*. Во-вторых, этот метод позволяет количественно оценить обобщенные отличия между всеми объектами, отделить несходные и объединить сходные, т. е. выявить *кластерную структуру объектов*. Обозначив плеяды признаков и кластеры объектов, компонентный анализ заставляет исследователя задуматься над причинами наблюдаемой структурированности, выйти за рамки известного, направляет дальнейший научный поиск.

Компонентный анализ в среде StatGraphics

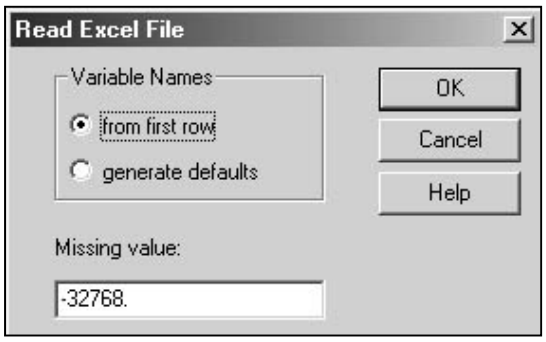
Для проведения расчетов в среде StatGraphics нужно занести данные на электронный лист, например, скопировать через буфер обмена с листа Excel. Лучший вариант – сохранение данных в формате листа Excel ранних версий. Рассмотрим ключевые этапы работы для примера с морфологической изменчивостью гадюк.



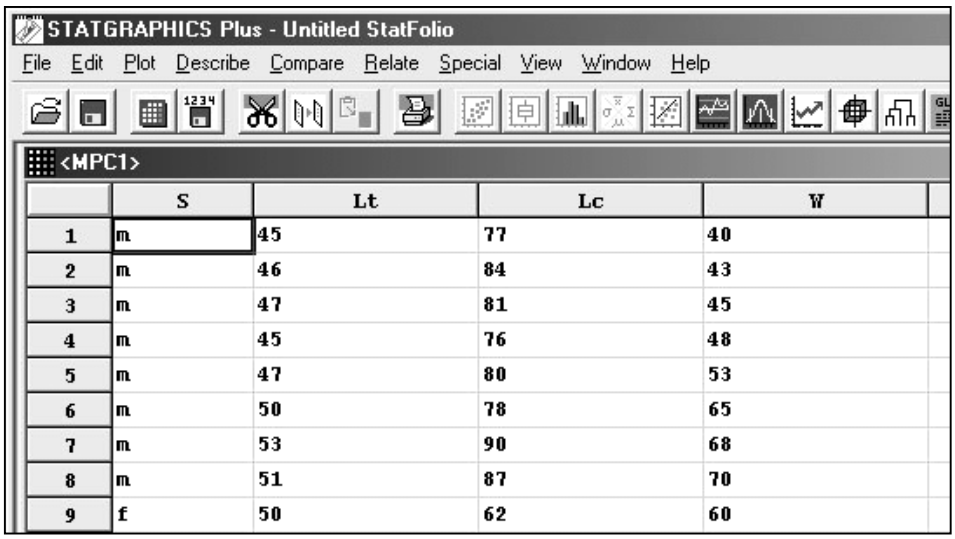
Открыть в среде StatGraphics файл следует командой меню или кнопкой Open Data File.



Чтобы имена переменных, назначенных в Excel, автоматически становились именами столбцов, они должны даваться латиницей; в окошке запроса отметить, что имена переменных в первом ряду есть.

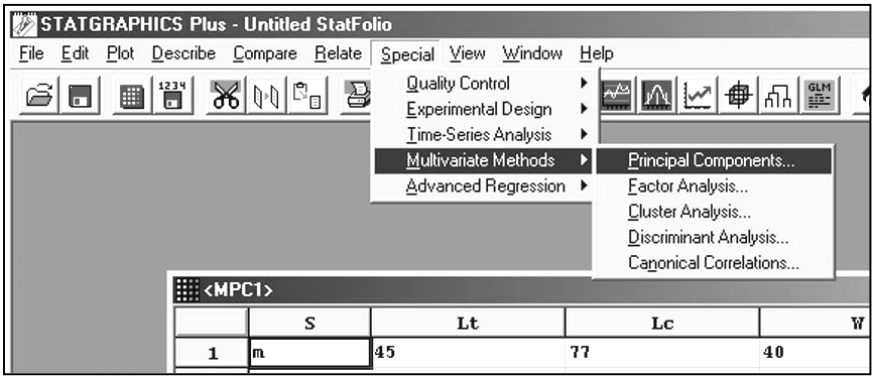


Результаты экспорта данных можно посмотреть в окне данных, специально распахнув окно иконки, лежащей на сером поле слева в нижнем углу.

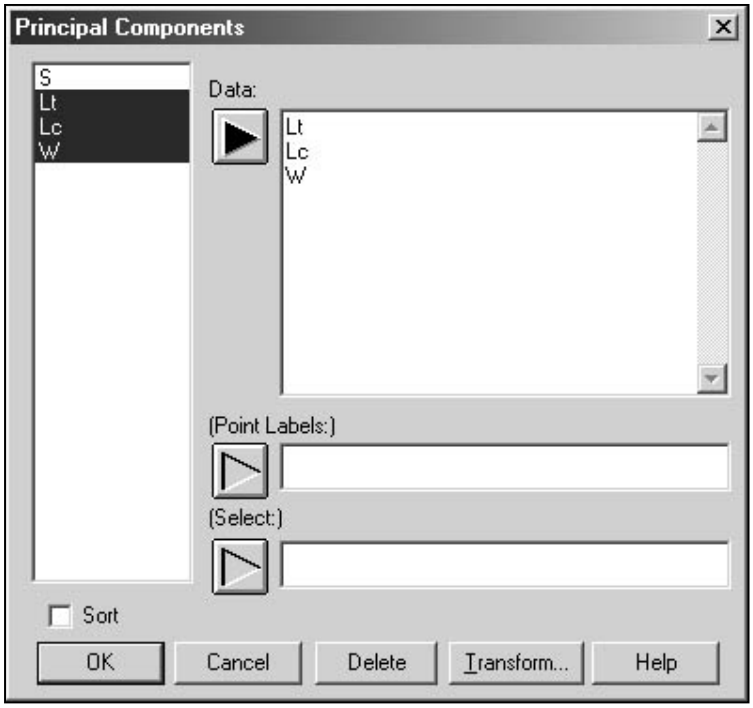


	S	Lt	Lc	W
1	m	45	77	40
2	m	46	84	43
3	m	47	81	45
4	m	45	76	48
5	m	47	80	53
6	m	50	78	65
7	m	53	90	68
8	m	51	87	70
9	f	50	62	60

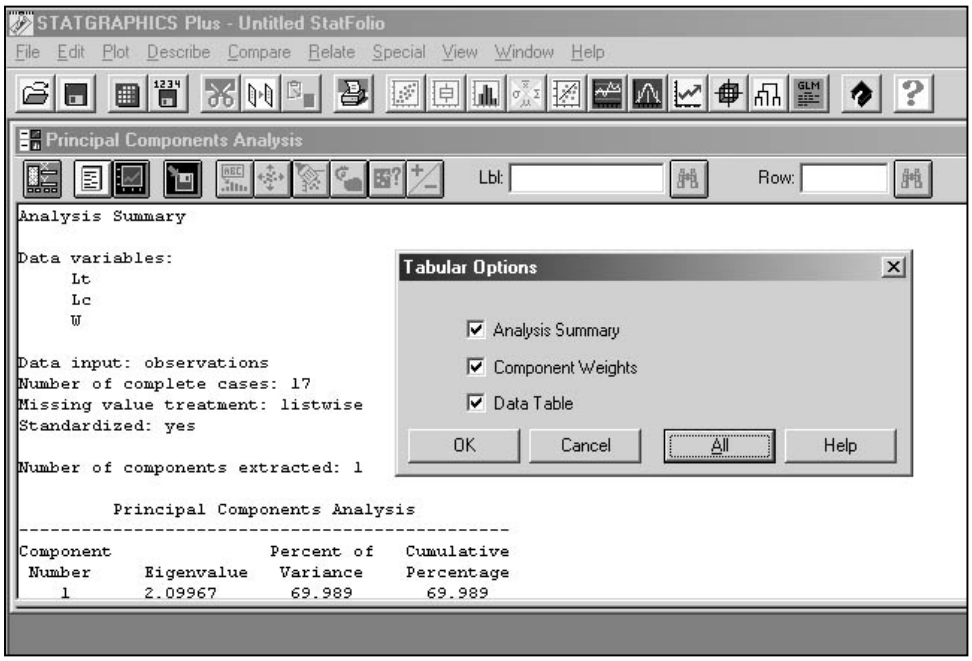
Запустить программу компонентного анализа можно только командой меню Special\ Multivariate Methods\ Principal Components.



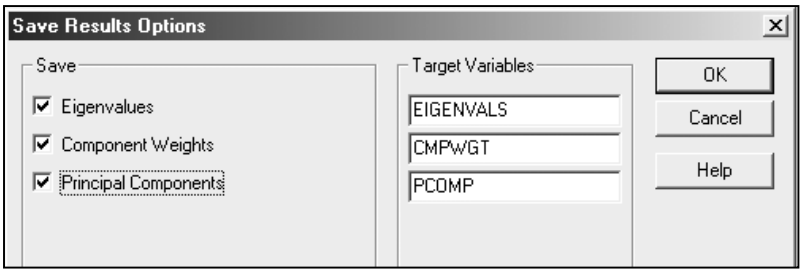
Выбрав мышкой имена нужных переменных, кнопкой Data: их нужно скопировать в правое окно, ОК. Для дальнейшей идентификации объектов их метки следует поместить в окно Point Labels:.

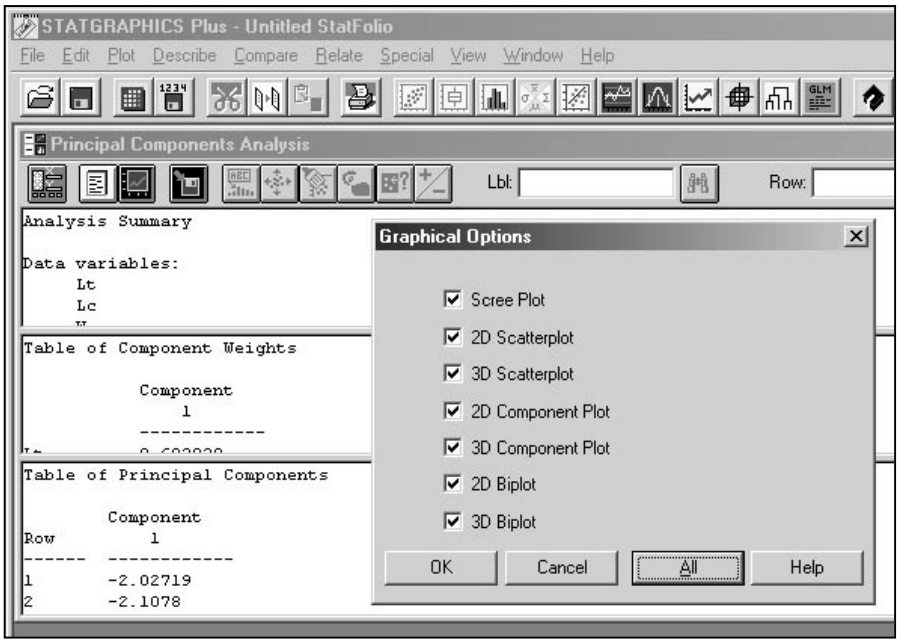


В появившемся окне Principal Component Analysis четыре кнопки играют важную роль. Первая слева кнопка Input Dialog позволяет вернуться на предыдущий шаг и переопределить список анализируемых переменных. Кнопка Tabular Options обеспечивает доступ ко всем результатам анализа (All, OK). Окно Analysis Summary выводит значения дисперсий главных компонент, окно Table of Component Weights дает значения факторных нагрузок, в окно Table of Principal Components выведены значения главных компонент.

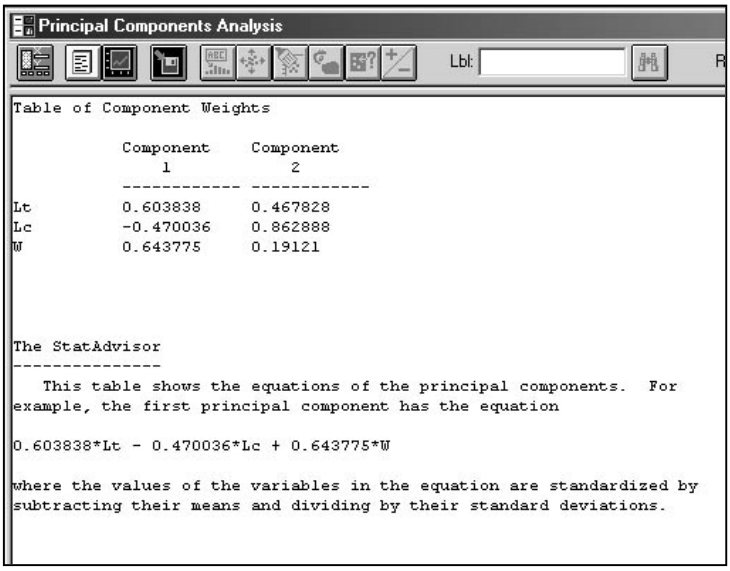


Кнопка Graphical Options раскрывает окна с графическими иллюстрациями (All, OK).





Все окно результатов компонентного анализа предстает в виде десяти небольших окошек; распахнуть любое из них позволяет двойной клик левой кнопкой мыши.



Полнота результатов вычислений во многом определяется установками в окне Principal Components Options, которое вызывается командой контекстного меню Analysis options… (правый клик на любом окне анализа). Минимально необходимый объем информации появляется, если в блоке Extract by … Number of Components задать число 2 (т. е. выводить результаты для двух компонент); кроме того, можно задать иное минимальное значение дисперсии главной компоненты (Eigenvalue), чем принятое по умолчанию значение 1. В результате на графиках и в таблицах будут отображаться данные по компонентам, дисперсия которых превышает заданный уровень.

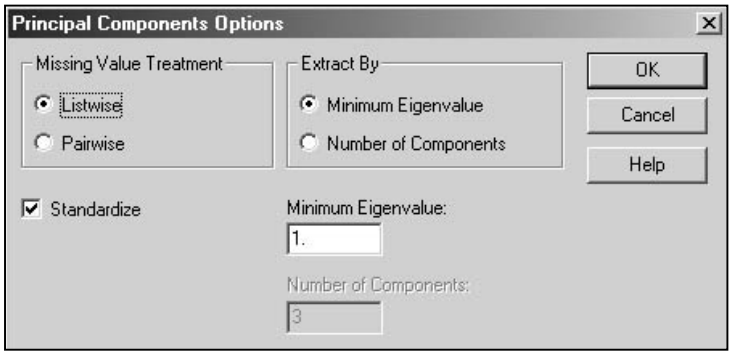
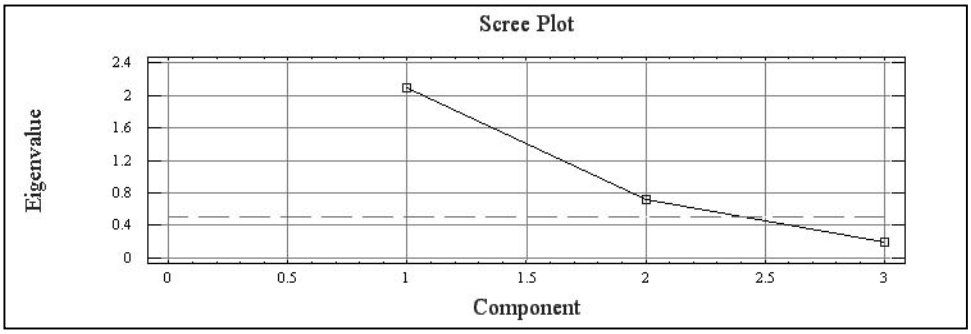
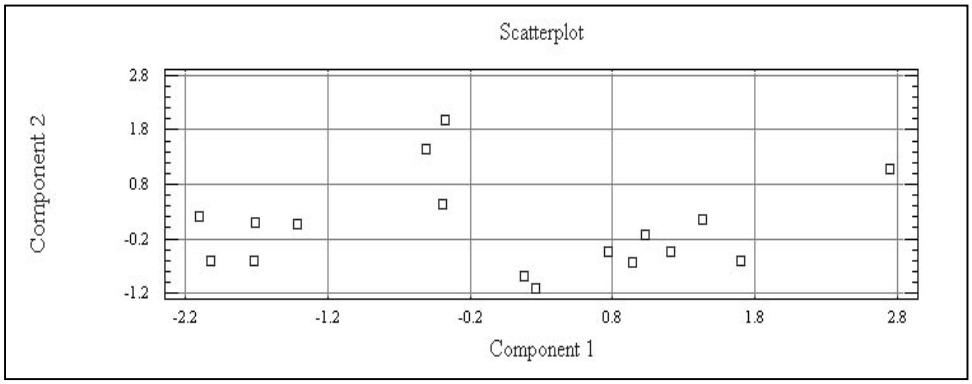


Диаграмма факторных нагрузок (Plot of Component Weights) копирует таблицу Table of Component Weights и призвана наглядно представить степень коррелированности соответствующих признаков.

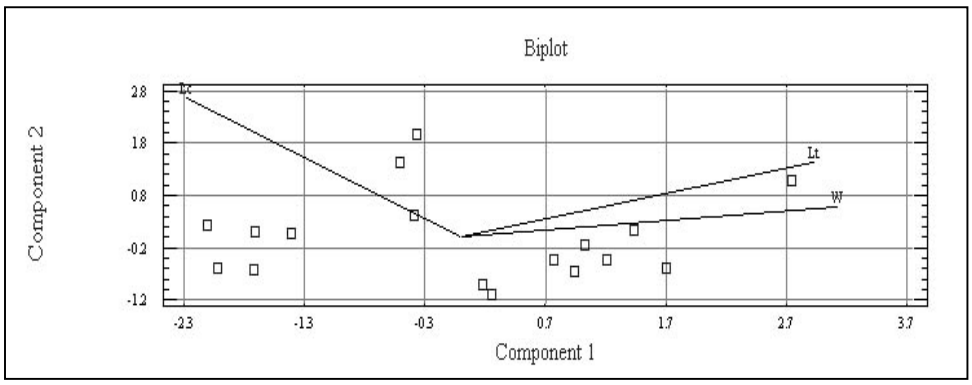
График Scree Plot отражает изменение дисперсий компонент и (пунктиром) минимальный уровень значимых компонент.



Наиболее интересна диаграмма Scatterplot, где представлена ординация объектов в осях компонент,

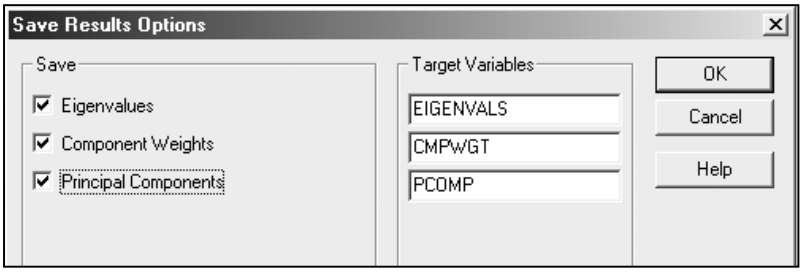


а также Biplot, где к диаграмме Scatterplot добавлена диаграмма Plot of Component Weights в форме лучей.



Каждый из этих лучей построен по двум опорным точкам: в месте пересечения осей компонент (0,0) и в точке с координатами факторных нагрузок двух первых компонент (a_{1j},a_{2j}) (здесь j – номер соответствующего признака). Это возможно потому, что и компоненты, и факторные нагрузки есть безразмерные признаки. Биплот наглядно показывает направления изменчивости данных, 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*), а второе (вторая компонента) связано в основном с отличиями по длине хвоста (Lc).

Результаты расчетов можно поместить на электронный лист (с помощью кнопки Save results, поставив галочки в нужных окошках), через буфер обмена скопировать на лист Excel и воспользоваться его богатыми графическими возможностями.

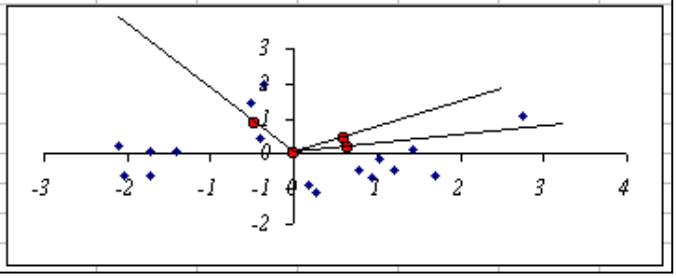


<MPC1>

	W	EIGENVALS	CMPWGT_1	CMPWGT_2	PCOMP_1	PCOMP_2
1	40	2.09967	0.603838	0.467828	-2.02719	-0.59622
2	43	0.711266	-0.470036	0.862888	-2.1078	0.219521
3	45	0.189064	0.643775	0.19121	-1.71467	0.0979875
4	48				-1.72312	-0.613703
5	53				-1.41059	0.0805049
6	65				-0.394844	0.421885
7	68				-0.371925	1.97755
8	70				-0.513151	1.44202
9	60				0.255382	-1.10976

В частности, чтобы понять принцип построения биплота, следует объединить (копированием) две точечные диаграммы, построенные раздельно по значениям главных компонент и факторных нагрузок, соединив лучами точки нагрузок с пересечением осей.

	A	B	C	D	E	F	G
1	0.603838	0.467828	-2.02719	-0.59622			
2	-0.47004	0.862888	-2.1078	0.219521			
3	0.643775	0.19121	-1.71467	0.097988			
4	0	0	-1.72312	-0.6137			
5			-1.41059	0.080505			
6							
7							
8							
9							
10							
11							
12							
13							
14							



10
ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ В СРЕДЕ *EXCEL*

Вообще говоря, любая мысль об окружающем мире есть его модель. Имитационная модель – это компьютерная программа, которая служит для количественного отображения поведения реальных объектов в разных условиях. Смысл построения имитационных моделей состоит, во-первых, в том, чтобы установить (выразить уравнением) количественные закономерности протекания явлений природы, во-вторых, – оценить модельные параметры (коэффициентов пропорциональности между переменными уравнений). Параметры моделей часто имеют биологический смысл, поскольку выражают существо отношений между характеристиками объектов исследования.

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 предлагается *составлять программы* и *настраивать параметры* имитационных моделей. Обе эти проблемы оптимально решаются в среде пакета Microsoft Excel.

Способ построения моделей на листе Excel отличается от традиционных способов программирования (алгоритмического, структурного или объектного) – это *табличное программирование*. На листе Excel модель предстает в всех своих деталях, как таблица,

ячейки которой заполнены формулами, имитирующими либо выборку вариант (статические модели), либо ход процесса (динамические модели). Каждая ячейка содержит формулу, которая вычисляет соответствующее «модельное» значение варианты или характеристику системы на очередном временно́м шагу. Поскольку «объяснительные» 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» («невязку»). Эта процедура оптимизации выполняется с помощью отдельной программы «Поиск решения», встроенной в пакет Excel. (Ответственное отношение к моделированию требует понимания существа процедуры настройки! (см.: Коросов, 2002).)

Помимо программирования самой модели и настройки ее параметров требуется доказать значимость модельных параметров, адекватность модели. Для решения этой задачи на листе Excel приходится конструировать целую *имитационную систему*, состоящую из следующих компонентов:

- блок исходных данных, зачастую состоящий из массива независимых и зависимых переменных;
- блок расчета модельных данных, собственно имитационная модель, состоящая из уравнений; осуществляет расчет явных переменных и скрытых переменных;
- блок параметров, участвующих в расчете модельных данных и изменяемых в процессе настройки;
- блок расчета отличий реальных и расчетных значений переменных;
- значение суммы отличий между моделью и реальностью (значение функции невязки); оно минимизируется в процессе настройки;
- блок процедуры настройки (программа «Поиск решения»);
- блок графического представления результатов;
- блок статистической оценки результатов.

В результате несложных действий мы получаем очень гибкий инструмент описания действительности. В потенциях имитационной модели стать сложной и детализированной или, напротив, простой и обобщающей, выражающей законы, управляющие миром.

Задача аппроксимации данных (статические модели)

Статические модели похожи на регрессионные; они выполняют расчет модельных значений, опираясь на исходные реальные выборочные данные. Меняется лишь процедура расчета модельных параметров (коэффициентов в уравнениях) – вместо метода наименьших квадратов используется итеративная процедура подгонки результатов счета под исходные значения. Вследствие этого расширяются возможности для аппроксимации любых зависимостей, возможности для описания наблюдаемых явлений с помощью уравнений самого разного вида и сложности, снимаются ограничения на пропуски в данных, а для случая криволинейных зависимостей результаты расчетов уточняются.

Моделирование – это не только инструмент преобразования нагромождения фактов в емкую модельную формулу. Более интересна возможность анализировать сложное явление, представляя его как объединение более простых взаимодействующих компонентов. С помощью имитационной модели любая сложная кривая зависимости может быть представлена как сумма более простых кривых, параметры уравнений которых приобретают биологический смысл. Одна из часто встречающихся сложных кривых – параболическая зависимость. Многие явления в жизни природы подчинены цикличности, когда постепенное увеличение уровня изучаемого признака сменяется плавным его падением. При этом за восходящую ветку кривой, как правило, ответственны одни процессы, а за нисходящую – другие (смертность). Результирующая параболическая кривая оказывается следствием синтеза двух противоположных тенденций.

Рассмотрим пример модельного исследования одного сложного процесса – изменения плодовитости рачка дафнии (Daphnia magna Straus) во все возрастающих концентрациях лигнина, основного компонента сточных вод целлюлозно-бумажных комбинатов (Калинкина, 1993). Опыт состоял в следующем. Дафний одного возраста (7 дней) помещали в растворы с разными концентрациями лигнина. Ежедневно в течение месяца собирали живую молодь. Графическое изображение результатов опыта представляет собой близкую к параболической зависимость плодовитости дафний (E) от все возрастающих концентраций лигнина ($[Lig]$) (рис. 10.1).

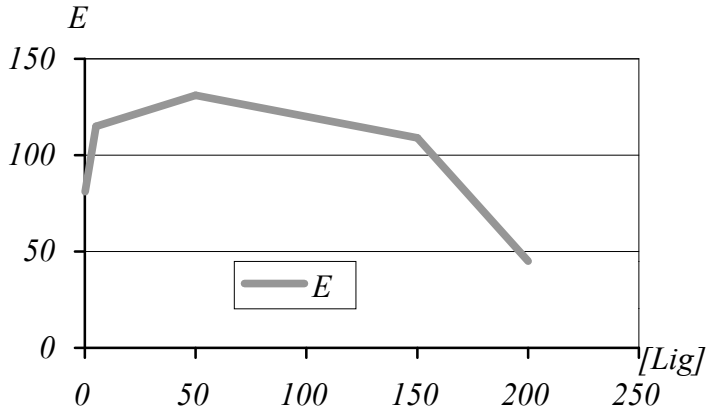


Рис. 10.1. *Зависимость плодовитости дафний* (*E*) *от лигнина* (*[Lig]*)

Сначала решим простую задачу аппроксимации (приблизительного описания) исходных данных параболической кривой.

Введем на лист Excel данные, названия переменных (включая модельную плодовитость), а также обозначения и предварительные значения параметров (*a*, *b*, *c*), коэффициентов уравнения. Следующая задача состоит в том, чтобы по формуле параболы рассчитать теоретические значения плодовитости (*Emod*), соответствующие данным концентрациям лигнина: $E = a \cdot [Lig]^2 + b \cdot [Lig] + c$.

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	*ф*		
3	0.1	81			*a*=	1
4	5	115			*b*=	1
5	50	131			*c*=	1
6	150	109				
7	200	45				
8						

Введем эту формулу в ячейку C3 для расчета первого модельного значения. На листе Excel она представляет собой набор

ссылок, во-первых, на ячейки, содержащие значения аргументов (столбец A), во-вторых, на ячейки, содержащие значения параметров модели ($F$3, $F$4, $F$5). Ссылки на параметры должны быть абсолютными (содержать префиксы $), поскольку в расчетах разных значений модели используются одни и те же значения параметров.

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	*ф*		
3	0.1	81	=$F$3*A3^2+$F$4*A3+$F$5			
4	5	115			*b*=	1
5	50	131			*c*=	1
6	150	109				
7	200	45				

Далее следует скопировать введенную формулу в остальные ячейки, например, с помощью операции «автозаполнение». Для этого нужно навести мышь на правый нижний угол выделенной ячейки с формулой, чтобы появился курсор в виде черного крестика,

и, нажав левую кнопку, «протянуть» выделение по нескольким ячейкам (C3:C7). Весь блок заполнится модельными формулами.

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	*ф*		
3	0.1	81	1.11		*a*=	1
4	5	115	31		*b*=	1
5	50	131	2551		*c*=	1
6	150	109	22651			
7	200	45	40201			

Правильность формул легко проверить, щелкнув мышкой на любой ячейке блока и нажав функциональную клавишу F2; так же выверяется правильность ссылок.

КВАДРОТКЛ | =$F$3*A6^2+$F$4*A6+$F$5

	A	B	C	D	E	F	G	H
1	Модель плодовитости дафний от лигнина							
2	*[Lig]*	*E*	*Emod*	*ф*				
3	0.1	81	1.11		*a*=	1		
4	5	115	31		*b*=	1		
5	50	131	2551		*c*=	1		
6	150	109	=$F$3*A6^2+$F$4*A6+$F$5					
7	200	45	40201					

Рассчитанные значения плодовитости (от 1 до 40201) сильно отличаются от реальных; предварительно заданные параметры явно не подходят. Для «настройки» модели на действительность необходимо рассчитать обобщенное отличие между ними; этой цели служит сумма квадратов разности между всеми значениями. Вводим формулу определения различий: $\phi = (x_{mod} - x)^2$.

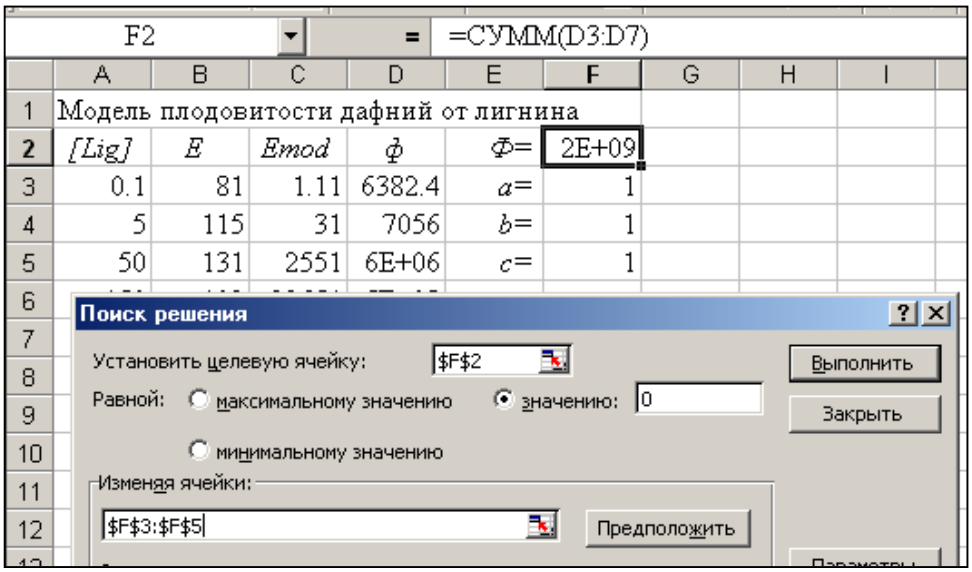
ДИСПА | =(C3-B3)^2

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	*ф*		
3	0.1	81	1.11	=(C3-B3)^2		1
4	5	115	31		*b*=	1
5	50	131	2551		*c*=	1
6	150	109	22651			
7	200	45	40201			
8						

Автозаполнив ею ячейки блока D3:D7, рассчитаем суммарное отличие Ф, например, в ячейке F2 =СУММ(D3:D7).

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	*ϕ*	*Φ=*	2E+09
3	0.1	81	1.11	6382.4	*a=*	1
4	5	115	31	7056	*b=*	1
5	50	131	2551	6E+06	*c=*	1
6	150	109	22651	5E+08		
7	200	45	40201	2E+09		

Теперь можно провести настройку модели, изменяя ее параметры с помощью макроса оптимизации, который вызывается командой меню Сервис\ Поиск решения.



В появившемся окне нужно с помощью мыши:

- Установить целевую ячейку (с суммой разности квадратов),
- Равной значению 0,
- Изменяя ячейки (со значениями параметров).

После этого нажать кнопку Выполнить, и в появившемся окне Результаты поиска решения выбрать Сохранить результаты и рассмотреть их.

	A	B	C	D	E	F
1	Модель плодовитости дафний от лигнина					
2	*[Lig]*	*E*	*Emod*	ϕ	Φ=	418.16
3	0.1	81	94.332	177.74	*a*=	-0.007
4	5	115	99.666	235.14	*b*=	1.1234
5	50	131	133.28	5.1781	*c*=	94.22
6	150	109	108.69	0.0979		
7	200	45	45.043	0.0018		

Итак, для описания эмпирической зависимости настройка предлагает следующие коэффициенты модели (рис. 10.2):

$$Emod = -0.007 \cdot [Lig]^2 + 1.1234 \cdot [Lig] + 94.22.\ 7$$

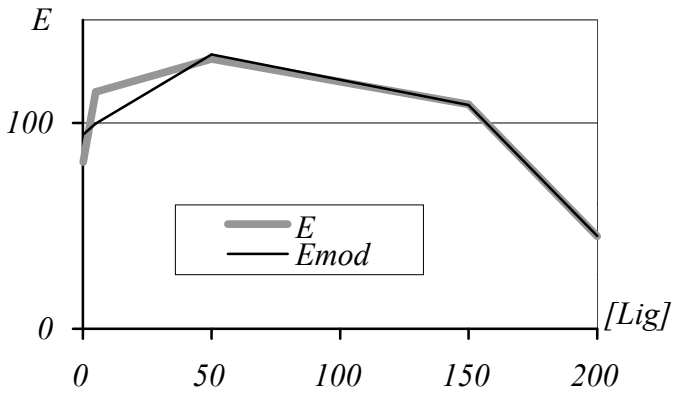


Рис. 10.2. *Модель зависимости плодовитости* (*E*) *от лигнина* (*[Lig]*)

Корреляция между реальными и модельными значениями составила: $r_{E,Emod} = 0.953$. Полученная модель хорошо описывает правые точки (50–200 мг/л), но для небольших концентраций (5 мг/л) ее прогноз оказался заниженным (100 против 115 экз.).

Такое же уравнение в среде Excel можно построить с помощью программы «Добавить линию тренда».

Преимущество имитационных моделей перед описательными регрессионными состоит в том, что они с легкостью могут послужить средством анализа сложной зависимости, разложения ее на составные компоненты, т. е. для получения моделей, более понятных и адекватных исходным данным.

Продолжим исследование реакции дафний на лигнин с помощью имитационного моделирования с целью объяснить существо наблюдаемого процесса.

Помимо регистрации живой молоди проводились и другие наблюдения. В частности, в высоких концентрациях на дне сосудов наблюдались скопления останков (карапаксов) погибшей молоди, не вошедшей в учет. Кроме того, к концу опытов даже невооруженным глазом было видно, что дафнии в контроле существенно мельче дафний из опытных сред.

Эти наблюдения позволяют определить два возможных фактора, ответственных за изменение плодовитости. Во-первых, лигнин явно стимулировал рост животных и крупные особи стали давать больше молоди; для Cladocera это явление широко известно. Во-вторых, в высоких концентрациях лигнина возросла смертность новорожденных, что и снизило общую плодовитость; этот эффект также обычен для интоксицированных дафний.

Параболическая кривая плодовитости, следовательно, образована обобщением двух простых монотонно возрастающих кривых: это рост потенциальной плодовитости с ростом размеров тела и рост смертности молоди в возрастающих концентрациях лигнина. Определение параметров этих уравнений и составляет очередную задачу моделирования.

Решить ее можно даже в отсутствие данных по размерам тела, напрямую связав рост плодовитости рачков с увеличением концентрации лигнина (мг/л). Тогда модель должна быть представлена тремя уравнениями:

– связь потенциальной плодовитости (экз./самку/30 дней) и размеров тела, зависящих от концентрации лигнина: $Ep = a \cdot [Lig]^{b}$,

– зависимость смертности от концентрации лигнина: $Ed = c \cdot [Lig]^{d}$,

– результирующая, наблюдаемая плодовитость: $Emod = Ep - Ed$,

где a, b, c, d – модельные параметры.

Здесь две модельные переменные, потенциальная плодовитость (*Ep*) и смертность (*Ed*), оказываются скрытыми (латентными), они используются для расчета одной явной переменной – результирующей плодовитости (*Emod*). При этом обе скрытые переменные определяются одной общей для них независимой переменной *[Lig]*.

Имитационную систему на листе Excel следует дополнить этими уравнениями (табл. 10.1), которые в формате Excel примут такой вид (например, для ряда 4):

C4 =$H$3*A4^$H$4,
D4 =$H$5*A4^$H$6,
E4 =C4-D4.

Таблица 10.1

	A	B	C	D	E	F	G	H
1	Модель плодовитости дафний от лигнина							
2	*[Lig]*	*E*	*Ep*	*Ed*	*Emod*	*ф*	*Ф=*	892.31117
3	0.1	81	63.546	1E-05	63.546	304.63	*a=*	80.0000000
4	5	115	93.97	0.025	93.945	443.33	*b=*	0.1000000
5	50	131	118.3	2.5	115.8	231.02	*c=*	0.0010000
6	150	109	132.04	22.5	109.54	0.2895	*d=*	2.0000000
7	200	45	135.89	40	95.892	2590		
8								

Для того чтобы модельные уравнения правильно отображали тенденции изменения соответствующих переменных, значения их параметров необходимо задать, хотя бы приблизительно, еще перед настройкой (метод биологического правдоподобия). Так, рост плодовитости может начаться только со значения 81 экз. (плодовитость в контроле); примем начальную величину первого параметра на уровне $a = 80$. Поскольку рост плодовитости идет не очень интенсивно, примем степенной параметр равным $b = 0.1$ (с этими параметрами первые значения потенциальной и реальной плодовитости почти совпадут). Смертность, в отличие от плодовитости, напротив, возрастает очень резко (примем $d = 2$), начавшись с низкого уровня (например, $c = 0.01$). Подставив эти примерные величины параметров в имитационную систему, получим приблизительное описание изменения плодовитости дафний в разных концентрациях раствора.

Далее, используя макрос оптимизации («Поиск решения»), удается получить хорошее решение, тесное совпадение модельных и реальных данных (табл. 10.2, рис. 10.3).

Таблица 10.2

	A	B	C	D	E	F	G	H
1	Модель плодовитости дафний от лигнина							
2	*[Lig]*	*E*	*Ep*	*Ed*	*Emod*	*ф*	*Ф=*	9.430131
3	0.1	81	82.616	7E-09	82.616	2.611	*a=*	99.6447672
4	5	115	113.59	0.0012	113.59	1.9883	*b=*	0.0813911
5	50	131	137	1.4758	135.53	20.512	*c=*	0.0000084
6	150	109	149.82	43.808	106.01	8.9277	*d=*	3.0862867
7	200	45	153.37	106.45	46.919	3.682		
8								

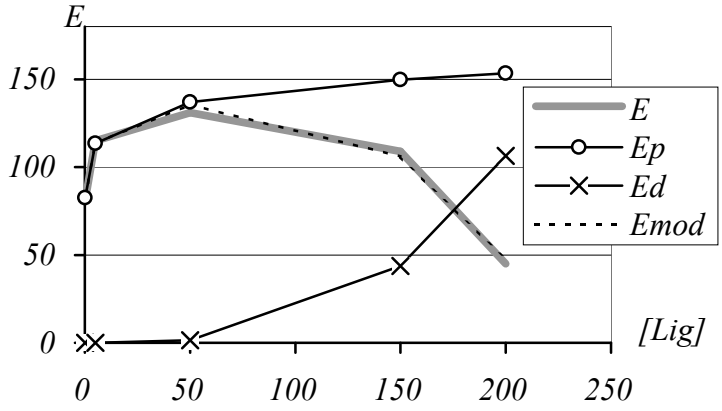


Рис. 10.3. *Рост потенциальной плодовитости и смертности молоди в возрастающих концентрациях лигнина.*

Корреляция между реальными и модельными значениями составила: $r_{E,Emod} = 0.996$, стала немного выше, чем для первой модели. Новая модель в целом лучше описывает процесс, в частности, *Emod* ближе к *E* для низких концентраций. Объясняется это тем, что «ответственность» за восхождение и падение кривой общей плодовитости разделили между собой два уравнения, которые с помощью разных наборов параметров описывают частные механизмы изучаемого явления.

Задача изучения процессов (динамические модели)

Динамические модели структурно более сложны, чем статические, поскольку призваны отображать ход исследуемых процессов. Они во многом напоминают процедуру интегрирования (выполненного для дискретного случая на ограниченном временно́м промежутке), когда имеется известное уравнение, выражающее скорость процесса (аналог производной), и для каждого временно́го интервала идет расчет (интеграция, суммирование) результатов процесса, достигнутых к данному шагу. Получается, что для расчета зависимых переменных, характеризующих результат процесса, достигнутый к определенному времени, приходится восстанавливать всю динамику, предшествующую этому моменту. В этом смысле динамическая имитационная модель «живет своей жизнью», и параметры скоростных уравнений характеризуют механизм этого процесса.

Рассмотрим этапы моделирования процесса на примере. В течение последних 10 лет на о. Кижи (Онежское озеро, Карелия) изучалась популяция обыкновенной гадюки. Животных метили, определяли встречаемость меченых особей (m) в повторных пробах разного объема (n). Так, из 158 гадюк, помеченных в 1994 г., проба 1995 г. (n = 365 экз.) содержала m = 18 особей (табл.10.3, графы n, m).

Таблица 10.3

	A	B	C	D	E	F	G	H	I
1	Год	n	m	N'	d'	M'	m'		Φ
2	1994	158		5000		158			
3	1995	365	18	5000	0.1	142	10		58
4	1996	273	10	5000	0.1	128	7		9.1
5	1997	214	10	5000	0.1	115	5		26
6	1998	238	9	5000	0.1	104	5		17
7									
8		C =	4.2		N =	5000		$C_{ост.}$ =	109
9		df =	3		Nd =	500		$S_{ост.}$ =	36
10		$C_{общ.}$ =	53		Nb =	500		F =	-1.6
11		$S_{мод.}$ =	-57		d% =	10			

Положим целью моделирования определение ежегодной численности (N) и смертности (Nd) в островной популяции гадюк (при отсутствии массовых миграций). Обычные методы расчетов здесь не работают, т. к. в данном случае не выполняются важные требования (отсутствие смертности, только троекратный отлов и т. д.). Для иллюстрации работы метода имитации покажем решение упрощенной задачи, приняв ежегодную численность и смертность в островной популяции гадюки неизменной:

$N = N_i$ = const (i = 1994, … 1998), Nd = const.

Главный момент имитационного моделирования состоит в том, чтобы выразить известные переменные через неизвестные параметры. Имитационная модель должна вычислять те же величины, что наблюдаются в природе, опыте. Тогда появляется возможность, перебирая возможные значения параметров, найти такие, при которых модельные значения переменных совпадут с реальными. В этом случае можно обсуждать найденные значения параметров как характеристику механизма наблюдаемого явления. Для популяции гадюки нам известны следующие переменные: число одноразово меченых животных (M), объемы повторных отловов (n), число повторно отловленных особей в каждой новой пробе (m). Неизвестными остаются общая численность (N), число ежегодно гибнущих особей (Nd) и объем пополнения (Nb) популяции. Три последних значения и требуют оценки, но их необходимо задать сразу же в первом приближении. Разместим их на электронном листе Excel (табл. 10.3) в отдельном блоке: F8 =5000, F9 =500, F10 =F9, D2 =F8.

В реальной популяции численность ежегодно поддерживается балансом процессов гибели и пополнения:

$N_{i+1} = N_i - Nd + Nb$.

Эта динамика в формате Excel примет вид:

D3 =D2-$F$9+$F$10, D4 =D3-$F$9+$F$10, …, D6 =D5-$F$9+$F$10

(табл. 10.3, столбец D).

Несмотря на множество формул, их ввод не составляет проблемы, достаточно одну формулу ввести вручную, а остальные – с помощью операции «автозаполнение» (см. инструкцию к Excel). При этом важно следить за тем, чтобы ссылки на общие параметры были абсолютными, т. е. содержали префиксы $, например $F$9.

После ввода всех формул в таблице Excel отображаются результаты расчетов; в данном случае численность сохраняется неизменной $N'_1 = 5000$ экз. (табл. 10.3, графа N').

Ежегодная смертность, в том числе среди меченых, составит:

$d'_i = Nd/N'_i$,

или в формате Excel: E3 =$F$9/D3, … (графа d').

Число погибших меченых особей составит:

$dM = d'_i \cdot M$,

а число выживших меченых будет равно:

$M'_{i+1} = M'_i - d'_i \cdot M$,

или F3 =F2-F2*E2, … (графа M').

Как видно из табл. 10.3, число меченых гадюк со временем сокращается. Сокращаться должно и число повторно отловленных меток (m'). Поскольку концентрация меченых особей равна

$pM'_i = M'_i / N'_i$,

то число меченых в пробе объемом n составит:

$m'_i = n_i \cdot pM'_i = n_i \cdot M'_i / N'_i$,

или G3 =B3*F3/D3, …, G6 =B6*F6/D6 (графа m').

Модельное число повторно отловленных гадюк (m') уменьшается, но сильно отличается от наблюдаемых значений (m). Это говорит о том, что произвольно взятые величины N и Nd не соответствуют реальности. Для расчета степени отличия модели от натурных наблюдений используем формулу:

$d_i = (m_i - m_i')^2$, или I3 =(C3-G3)^2,…(табл. 10.3, графа Φ).

Общее отличие есть сумма всех частных отличий: I8 =СУММ(I3:I6).

В нашем случае это обобщенное отличие (функция невязки) равно $\Phi = 109$. Понятно, что если бы модель абсолютно точно описывала реальность, то функция невязки была бы равна нулю.

С помощью макроса «Поиск решения» пытаемся выполнить это условие. После вызова макроса остается заполнить его окно, т. е. указать, что целевой ячейкой выступает ячейка I8 (со значением функции невязки), что она должна быть равной значению 0, что для этого можно изменять значения в ячейках F8:F9. Нажимаем кнопку «Выполнить», вслед за этим появляется окно «Результаты поиска решения», выбираем «Сохранить результаты», ОК. Для нашего примера они представлены в табл. 10.4

Таблица 10.4

	A	B	C	D	E	F	G	H	I
1	Год	n	m	N'	d'	M'	m'		Φ
2	1994	158		3086		158			
3	1995	365	18	3086	0.07	146	17		0
4	1996	273	10	3086	0.07	135	12		4
5	1997	214	10	3086	0.07	125	9		2
6	1998	238	9	3086	0.07	116	9		0
7									
8					N =	3086		$C_{ост.}$ =	6
9		df =	3		Nd =	228		$D_{ост.}$ =	2
10		$C_{общ.}$ =	53		Nb =	228		F =	23
11		$D_{мод.}$ =	47		d% =	7.4			

Как видно из табл. 10.4, при численности островной популяции обыкновенной гадюки, равной $N = 3086$ экз., и смертности $d = 7.4\%$ модельная динамика снижения числа меченых животных оказалась почти такой же, что наблюдалась и в поле. «Почти», потому что функция невязки так и не обнулилась, после настройки $\Phi = 6$.

Для решения вопроса, соответствует ли модель реальности, предлагается несколько способов; используем дисперсионный анализ линейной регрессии. В соответствии с рассмотренной выше схемой модель считается адекватной, если модельная дисперсия достоверно больше остаточной; в этом случае критерий Фишера $F = S^2_{мод.}/S^2_{ост.}$ превысит табличное значение $F_{(0.05,df.мод.,df.остат.)}$.

Для расчета дисперсий нужно определить модельную и остаточную сумму квадратов. Остаточная сумма квадратов и есть функция невязки, получаем: $S^2_{ост.} = C_{ост.}/(n–1)$, или I9 =I8/C9.

Модельная дисперсия равна модельной сумме квадратов, поскольку число степеней свободы для ее расчета, $S^2_{мод.} = C_{мод.}/df_{мод.}$, равно единице $df_{мод.} = 1$. Модельную сумму квадратов находят как разность между общей и остаточной:

$C_{мод.} = C_{общ.} – C_{ост.}$ или C11=C10–I8.

В свою очередь, общую сумму квадратов можно определить с помощью особой функции листа Excel, подсчитывающей сумму

квадратов отклонения вариант от средней C10 =КВАДРОТКЛ(C3:C6).

Величина критерия Фишера составит:

$F = S^2_{мод.} / S^2_{ост.}$, или H10 =C11/I9.

В нашем случае значение критерия (23) превышает табличное $F_{(0.05,1,3)} = 6.6$; модель в целом адекватна наблюдаемым данным. Видимо, численность наблюдаемой островной популяции гадюки действительно приближается к 3000 экз.

Рассмотренный метод оценки адекватности можно использовать и для статических имитационных моделей.

ПРИЕМЫ РАБОТЫ В *EXCEL*

Программе Excel посвящена обширная литература. Однако за детальностью изложения многие важные возможности остаются невостребованными, их просто не удается найти среди множества советов. Многолетняя практика работы с пакетом позволяет предложить несколько советов.

Организация банка данных в Excel

Числовые данные, полученные во время натурных наблюдений, перед статистической обработкой попадают на лист Excel. При этом далеко не безразлично, как они будут структурированы. Часто правильная организация банка данных предопределяет тот или иной вид статистического анализа. Не упустить полезную информацию и эффективно подготовить числовой массив для последующей обработки поможет выполнение нескольких правил размещения данных в электронной таблице Excel.

	A	B	C	D	E	F	G
1	№	Вид	Дата	Место	Пол	W	Lt
2	1	S. sp.	03.03.1990	Петроз	f	0.53	0.9
3	2	S. sp.	04.03.1990	Питкяр	m	0.42	0.8
4	3	S. spp	08.03.1990	Питкяр	m	0.85	0.1
5	4	S. spp	08.03.1990	Кончез	f	0.58	0.9

Под каждое свойство (характеристику, показатель, признак, переменную) объектов измерения следует отводить отдельный столбец (так называемое "поле"). Названия полей должна содержать первая (единственная!) строчка (ряд 1); это упростит операции сортировки и экспорта данных в другие пакеты, например в StatGraphics, Access. Названия следует формировать из букв и чисел, избегая пробелов и спецсимволов (- ^ : , . ' + = …), лучше латиницей.

Каждый объект измерения (препарат, особь, маршрут, биоценоз…) представлен рядом ячеек (начиная с ряда 2), которые содержат числовые и текстовые характеристики. Множество рядов (строк) соответствует множеству изученных объектов. Иногда возникают проблемы с тем, чтобы правильно поместить в базу объекты, обладающие общей информацией (одного статуса, собранных в

одном месте или в одни и те же сроки). Так и хочется блок таких объектов озаглавить: пробы такие-то, место такое-то… Этого делать, естественно, нельзя, все записи (ряды) должны быть однотипными. Для правильного отражения общей информации организуется новое поле (столбец), по которому однотипные объекты получают одинаковые значения. В нашем примере таковы были виды, даты, места работы, пол особи. Небольшая избыточность информации в дальнейшем с лихвой компенсируется простотой доступа к данным, возможностью сортировки и фильтрации.

Быстрая команда "автозаполнения"

В соответствующих разделах показана процедура "автозаполнения" с протяжкой мыши. Еще быстрее можно выполнить эту операцию, дважды кликнув левой кнопкой на черном крестике. Формула или значения из текущей ячейки автоматически заполнят столько нижележащих ячеек, сколько заполненных ячеек содержится в соседнем столбце (приоритет у левого столбца).

Разделители

Обычно в качестве разделителя целой и дробной частей числа устанавливается запятая. Это очень неудобно при экспорте данных в пакет StatGraphics, который числа с запятыми воспринимает как текст. Не менее несуразно использование "по умолчанию" точки с запятой в качестве разделителя в тех же экспортных списках. Заменить запятую на точку, а точку с запятой на запятую можно в окне Языки и стандарты вкладки Числа. Для этого нажмите кнопку Пуск панели Windows, выберите Настройка \ Панель управления \ Языки и стандарты.

Специальные символы и другие полезные кнопки

Большая часть функций, с помощью которых создаются формулы в ячейках Excel, имеет кириллическое написание. Поэтому удобнее работать с клавиатурой в русском регистре. При этом для ввода некоторых важных символов (знак валюты $, возведение в степень ^) приходится часто переключаться в латинский регистр и обратно, что увеличивает число ошибок. Упростить ситуацию можно, если вывести указанные символы как кнопки на панель инструментов. Тогда для ввода символа в нужное место формулы (с позиции курсора) достаточно щелкнуть соответствующей кнопкой.

Поместить новые кнопки на главную панель можно следующим образом. Установите курсор мыши на любой панели, правой

кнопкой откройте контекстное меню, выберите Настройка, вкладка Команды. В Категории Вставка найдите значки Знак возведения в степень и Знак доллара. Мышью перетащите значки на главную панель, закройте окно настройки. Как минимум еще одна очень полезная кнопка должна быть представлена на панели инструментов. Это кнопка Тип диаграммы. Открыв то же окно настройки, выберите категорию команд Диаграмма. Почти в самом конце списка команд найдите кнопку (не строку меню) Тип диаграммы и поместите ее на панель, например, рядом с кнопкой Мастер диаграмм. Кнопка Тип диаграммы позволяет не только быстро изменять тип диаграммы, но и моментально строить диаграмму нужно типа. Достаточно отметить область данных, после чего, нажав кнопку, выбрать нужный тип диаграммы и отпустить кнопку мыши.

Вставка и форматирование таблицы Excel в среде Word

Выделяете нужный блок ячеек в таблице Excel, копируете в буфер обмена, переключаетесь в Word, устанавливаете курсор в нужном месте, вставляете блок из буфера. Проблемы возникают с расположением и форматом таблицы и чисел в ней: вдруг оказывается, то таблица расположена не по центру, имеет разномастные столбцы, слишком длинную дробную часть чисел, "прилипших" к правому краю ячеек.

Для придания таблице необходимого формата часть операций следует выполнить перед экспортом в среде Excel, часть – в Word. Быстрый способ отформатировать ширину столбца по размеру содержимого ячеек состоит в следующем. Установите курсор мыши между буквенными названиями столбцов, из белого креста он превратится в черный со стрелками по бокам. Теперь двойной клик изменит ширину столбца точно по ширине наибольшей записи, содержащейся в какой-либо ячейке левого от курсора столбца. Для форматирования чисел, выделив нужный блок ячеек в таблице Excel, откройте окно Формат \ Формат ячейки. На вкладке Число выберите из Числовых форматов Числовой и установите нужное число десятичных знаков. Чтобы не загромождать таблицу, устанавливайте минимальное число знаков после запятой. Если исходные данные имеют xx значащих цифр, значения средней арифметической и стандартного отклонения могут содержать на одну значащую цифру больше (xxx), значение ошибки средней – на две значащих цифры (xxxx). Например, если промеры имеют порядок 3.8, 5.6, 8.0,

средняя может быть 5.32, стандартное отклонение – 1.23, ошибка – 0.412. Чтобы шрифт копируемого текста из Excel был одинаков со шрифтом Word, Times New Roman Cyr 12, сделайте изменения на панели Сервис \ Параметры \ Общие.

Обозначения всех граф и настройку их ширины удобнее выполнять в среде Excel, чем в Word. Таблица из Excel вставляется в Word выровненной по левому краю. Для центрирования таблицы на листе Word сначала следует ее правильно выделить. Для этого подведите курсор мыши к верхнему краю первого столбца таблицы, где он преобразуется в черную вертикальную стрелку. Нажав левую кнопку мыши, ведите ее направо, выделяя все столбцы, и далее – за правую границу таблицы; отпустите кнопку мыши. Помимо почерневших столбцов справа от таблицы появится колонка добавочных черных квадратиков. В этом режиме щелкните кнопку выравнивания "по центру" панели Форматирование. Таблица переместится в центр листа. Теперь нажмите Shift и один раз стрелку влево. Дополнительные черные квадратики исчезнут. Вновь щелкните кнопку выравнивания "по центру" панели Форматирование. Текст в каждой ячейке отформатируется по центру.

Вставка и форматирование диаграмм Excel в среде Word

Построенные в среде Excel диаграммы, внедряются в среду Word посредством OLE-связей, позволяющих эффективно редактировать диаграммы непосредственно на листе документа Word. Это достигается посредством внедрения в файл Word'а дополнительных программных кодов. При этом объем файлов, содержащих документ, резко возрастает. Однако отчетные научные документы должны снабжаться иллюстративным материалом в виде диаграмм (графиков, гистограмм и пр.), пригодном для распространения по электронной почте или на флеш-картах, т. е. иметь небольшой объем. Выходом из ситуации является разрыв OLE-связей и представлением внедренных иммюстраций в виде векторных рисунков. В готовом документе нет необходимости что-либо исправлять, а размер файла может многократно уменьшиться. Для этой цели следует выделить внедренную диаграмму, дать команду Главного меню Формат \ Объект \ Положение \ Перед текстом \ Ок. Затем вызвать контекстное меню и дать сначала команду Группировка \ Разгруппировать, а потом – Группировка \ Группировать.

ПРИЕМЫ РАБОТЫ В *EXCEL*

Программе Excel посвящена обширная литература. Однако наша многолетняя практика работы с пакетом позволяет предложить несколько полезных советов.

Организация банка данных в Excel

Числовые данные, полученные во время натурных наблюдений, перед статистической обработкой попадают на лист Excel. При этом далеко не безразлично, как они будут структурированы. Часто правильная организация банка данных предопределяет тот или иной вид статистического анализа. Не упустить полезную информацию и эффективно подготовить числовой массив для последующей обработки поможет выполнение нескольких правил размещения данных в электронной таблице Excel.

	A	B	C	D	E	F	G
1	№	Вид	Дата	Место	Пол	W	Lt
2	1	S. sp.	03.03.1990	Петроз	f	0.53	0.9
3	2	S. sp.	04.03.1990	Питкяр	m	0.42	0.8
4	3	S. spp	08.03.1990	Питкяр	m	0.85	0.1
5	4	S. spp	08.03.1990	Конче	f	0.58	0.9

Под каждое свойство (характеристику, показатель, признак, переменную) объектов измерения следует отводить отдельный столбец (так называемое «поле»). Названия полей должна содержать первая (единственная!) строчка (ряд 1); это упростит операции сортировки и экспорта данных в другие пакеты, например в StatGraphics, Access. Названия следует формировать из букв и чисел, избегая пробелов и спецсимволов (- ^ : , . ' + = …), лучше латиницей.

Каждый объект измерения (препарат, особь, маршрут, биоценоз…) представлен рядом ячеек (начиная с ряда 2), которые содержат числовые и текстовые характеристики. Множество рядов (строк) соответствует множеству изученных объектов. Иногда возникают проблемы с тем, чтобы правильно поместить в базу объекты, обладающие общей информацией (одного статуса, собранные в одном месте или в одни и те же сроки). Так и хочется блок таких объектов озаглавить: пробы такие-то, место такое-то… Этого делать, естественно, нельзя, все записи (ряды) должны быть однотип-

ными. Для правильного отражения общей информации организуется новое поле (столбец), по которому однотипные объекты получают одинаковые значения. В нашем примере таковы были виды, даты, места работы, пол особи. Небольшая избыточность информации в дальнейшем с лихвой компенсируется простотой доступа к данным, возможностью сортировки и фильтрации.

Быстрая команда «автозаполнения»

На с. 263 показана процедура «автозаполнения» с протяжкой мыши. Еще быстрее можно выполнить эту операцию, дважды кликнув левой кнопкой на черном крестике. Ячейка автоматически скопируется в столько же нижележащих ячеек, сколько заполненных ячеек содержится в соседнем столбце (приоритет у левого столбца).

Разделители

Обычно в качестве разделителя целой и дробной частей числа устанавливается запятая. Это очень неудобно при экспорте данных в пакет StatGraphics, который числа с запятыми воспринимает как текст. Использование «по умолчанию» точки с запятой в качестве разделителя также неудобно. Заменить запятую на точку, а точку с запятой на запятую можно в окне, которое вызывается командой Пуск \ Настройка \ Панель управления \ Языки и стандарты \ Числа.

Специальные символы и другие полезные кнопки

Бóльшая часть функций, с помощью которых создаются формулы в ячейках Excel, имеет кириллическое написание. Поэтому удобнее работать с клавиатурой в русском регистре. При этом для ввода некоторых важных символов (знак валюты $, возведение в степень ^) приходится часто переключаться в латинский регистр и обратно, что увеличивает число ошибок. Упростить ситуацию можно, если вывести указанные символы как кнопки на панель инструментов. Тогда для ввода символа в нужное место формулы (с позиции курсора) достаточно щелкнуть соответствующей кнопкой.

Поместить новые кнопки на главную панель можно следующим образом. Установите курсор мыши на любой панели, правой кнопкой откройте контекстное меню, выберите Настройка, вкладка Команды. В Категории Вставка найдите значки Знак возведения в степень и Знак доллара. Мышью перетащите значки на главную панель, закройте окно настройки. Как минимум еще одна очень полезная кнопка должна быть представлена на панели инструментов. Это кнопка Тип диаграммы. Открыв то же окно настройки, выбери-

те категорию команд Диаграмма. Почти в самом конце списка команд найдите кнопку (не строку меню) Тип диаграммы и поместите ее на панель, например, рядом с кнопкой Мастер диаграмм. Кнопка Тип диаграммы позволяет не только быстро изменять тип диаграммы, но и моментально строить диаграмму нужно типа. Достаточно отметить область данных, после чего, нажав кнопку, выбрать нужный тип диаграммы и отпустить кнопку мыши.

Вставка и форматирование таблицы Excel в среде Word

Выделяете нужный блок ячеек в таблице Excel, копируете в буфер обмена, переключаетесь в Word, устанавливаете курсор в нужном месте, вставляете блок из буфера. Проблемы возникают с расположением и форматом таблицы и чисел в ней: вдруг оказывается, то таблица расположена не по центру, имеет разномастные столбцы, слишком длинную дробную часть чисел, «прилипших» к правому краю ячеек.

Для придания таблице необходимого формата часть операций следует выполнить перед экспортом в среде Excel, часть – в Word. Быстрый способ отформатировать ширину столбца по размеру содержимого ячеек состоит в следующем. Установите курсор мыши между буквенными названиями столбцов, из белого креста он превратится в черный со стрелками по бокам. Теперь двойной клик изменит ширину столбца точно по ширине наибольшей записи, содержащейся в какой-либо ячейке левого от курсора столбца. Для форматирования чисел, выделив нужный блок ячеек в таблице Excel, откройте окно Формат \ Формат ячейки. На вкладке Число выберите из Числовых форматов Числовой и установите нужное число десятичных знаков. Чтобы не загромождать таблицу, устанавливайте минимальное число знаков после запятой. Если исходные данные имеют xx значащих цифр, значения средней арифметической и стандартного отклонения могут содержать на одну значащую цифру больше (xxx), значение ошибки средней – на две значащих цифры (xxxx). Например, если промеры имеют порядок 3.8, 5.6, 8.0, то средняя может быть 5.32, стандартное отклонение – 1.23, ошибка – 0.412. Чтобы шрифт копируемого текста из Excel был одинаков со шрифтом Word, Times New Roman Cyr 12, сделайте изменения на панели Сервис \ Параметры \ Общие.

Обозначения всех граф и настройку их ширины удобнее выполнять в среде Excel, чем в Word. Таблица из Excel вставляется в

Word выровненной по левому краю. Для центрирования таблицы на листе Word сначала следует ее правильно выделить. Для этого подведите курсор мыши к верхнему краю первого столбца таблицы, где он преобразуется в черную вертикальную стрелку. Нажав левую кнопку мыши, ведите ее направо, выделяя все столбцы, и далее – за правую границу таблицы; отпустите кнопку мыши. Помимо почерневших столбцов справа от таблицы появится колонка добавочных черных квадратиков. В этом режиме щелкните кнопку выравнивания «по центру» панели Форматирование. Таблица переместится в центр листа. Теперь нажмите Shift и один раз стрелку влево. Дополнительные черные квадратики исчезнут. Вновь щелкните кнопку выравнивания «по центру» панели Форматирование. Текст в каждой ячейке отформатируется по центру.

Вставка диаграмм Excel в среду Word

Построенные в среде Excel и скопированные в среду Word диаграммы можно эффективно редактировать прямо на листе документа. Это достигается посредством внедрения в файл Word'а дополнительных программных кодов (OLE-связей). При этом после вставки диаграммы объем файла резко возрастает. В то же время иллюстрированные научные документы должны быть пригодными для распространения по электронной почте или на флеш-картах, т. е. иметь небольшой объем. Выходом из ситуации является разрыв OLE-связей и представление внедренных иллюстраций в виде векторных рисунков (в готовом документе нет необходимости что-либо исправлять). Размер файла с простыми векторными рисунками много меньше, чем с диаграммой Excel.

Выделим внедренную диаграмму, дадим команду Главного меню Формат \ Объект \ Положение \ Перед текстом \ ОК. Затем правым кликом мыши (курсор находится на диаграмме) вызовем контекстное меню и дадим сначала команду Группировка \ Разгруппировать, а потом (не смещая курсора) – Группировка \ Группировать. Диаграмма станет векторным рисунком.

Если форматирование диаграммы полностью выполнено в среде Excel, то ее можно сразу вставить на лист Word как рисунок. Для этого копируем диаграмму с листа Excel (выделить, Вид \ Копировать). Переходим на лист Word'а, даем команду Главного меню: Правка \ Специальная вставка \ Метафайл Windows (EMF) \ ОК.

СПИСОК ЛИТЕРАТУРЫ

Адлер Ю. П., Макарова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий. М.: Наука, 1976. 280 с.

Андреев В. А. Классификационные построения в экологии и систематике. М.: Наука, 1980.

Ашмарин И. П. и др. Быстрые методы статистической обработки и планирования экспериментов. Л.: Изд-во ЛГУ, 1975.

Безель В. С. Популяционная экотоксикология млекопитающих. М.: Наука, 1987. 130 с.

Бейли Н. Статистические методы в биологии. М.: Мир, 1964.

Браунли К. А. Статистическая теория и методология в науке и технике. М.: Наука, 1977.

Гиляров А. М. Популяционная экология. М.: Изд-во МГУ, 1990. 191 с.

Гроссман С., Терней Дж. Математика для биологов. М.: Высшая школа, 1983.

Гублер Е. В., Генкина А. А. Применение непараметрических критериев статистики в медико-биологических исследованиях. Л.: Медицина, 1973.

Джефферс Дж. Введение в системный анализ: применение в экологии. М.: Мир, 1981.

Дэвис Дж. С. Статистический анализ данных в геологии: В 2 кн. М.: Недра, 1990.

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.

Иванищев В. В., Михайлов В. В., Тубольцева В. В. Инженерная экология. Л.: Наука, 1989. 144 с.

Калинкина Н. М. Оценка пригодности участка водоема для рекреации// Экологические исследования природных вод Карелии. Петрозаводск, 1989. С. 87–89.

Коли Г. Анализ популяций животных. М.: Мир, 1979. 364 с.

Коросов А. В. Имитационное моделирование в среде MS Excel (на примерах из экологии). Петрозаводск, 2002. 212 с.

Коросов А. В. Экологические приложения компонентного анализа. Петрозаводск, 1996. 152 с.

Коросов А. В. Специальные методы биометрии. Петрозаводск, 2007. 364 с.

Коросов А. В., Горбач В. В. Компьютерная обработка биологических данных: Методическое пособие. Петрозаводск: Изд-во ПетрГУ, 2010. 84 с.

Лакин Г. Ф. Биометрия. М.: Высшая школа, 1973. 343 с.

Методы математической биологии. Математические решения задач биологии и медицины на ЭВМ. Киев: Выща шк., 1984. Т. 8. 344 с.

Моисеев Н. Н. Математические задачи системного анализа. М.: Наука, 1981. 487 с.

Перегудов Ф. И., Тарасенко Ф. П. Введение в системный анализ. М.: Высшая школа, 1989. 367 с.

Плохинский Н. А. Биометрия. М.: Изд-во МГУ, 1970.

Поллард Дж. Справочник по вычислительным методам статистики. М.: Финансы и статистика, 1982.

Пэнтл Р. Методы системного анализа окружающей среды. М.: Мир, 1979. 214 с.

Розенберг Г. С. Модели в фитоценологии. М.: Наука, 1984. 265 с.

Рокицкий П. В. Биологическая статистика. Минск: Вышейшая школа, 1973.

Саати Т., Кернс К. Аналитическое планирование. Организация систем. М.: Радио и связь, 1991. 224 с.

Страшкраба М., Гнаук А. Пресноводные экосистемы. Математическое моделирование. М.: Мир, 1989. 376 с.

Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. М.: ИНФРА, 1998. 528 с.

Урбах В. Ю. Биометрические методы. М.: Наука, 1964. 415 с.

Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях. М.: Медицина, 1975. 294 с.

Фишер Р. Статистические методы для исследователей. М.: Госстатиздат, 1958.

Юл Дж. Э., Кендэл М. Дж. Теория статистики. М.: Госстатиздат, 1960. 779 с.

Яковлев Е. И. Машинная имитация. М.: Наука, 1975. 158 с.

Sokal R., Rohlf J. Biometry. Principles and practice of statistics in biological research. 3-th ed. N.-Y., 1995. 888 p.

СПРАВОЧНЫЕ ТАБЛИЦЫ

Таблица 1*П*

Квадраты и квадратные корни для чисел 1…99

x	x^2	$\sqrt{x}$	x	x^2	$\sqrt{x}$	x	x^2	$\sqrt{x}$
1	1	1.000	34	1156	5.831	67	4489	8.185
2	4	1.414	35	1225	5.916	68	4624	8.246
3	9	1.732	36	1296	6.000	69	4761	8.307
4	16	2.000	37	1369	6.083	70	4900	8.367
5	25	2.236	38	1444	6.164	71	5041	8.426
6	36	2.449	39	1521	6.245	72	5184	8.485
7	49	2.646	40	1600	6.325	73	5329	8.544
8	64	2.828	41	1681	6.403	74	5476	8.602
9	81	3.000	42	1764	6.481	75	5625	8.660
10	100	3.162	43	1849	6.557	76	5776	8.718
11	121	3.317	44	1936	6.433	77	5929	8.775
12	144	3.464	45	2025	6.708	78	6084	8.832
13	169	3.606	46	2116	6.782	79	6241	8.888
14	196	3.742	47	2209	6.856	80	6400	8.944
15	225	3.873	48	2304	6.928	81	6561	9.000
16	256	4.000	49	2401	7.000	82	6724	9.055
17	289	4.123	50	2500	7.071	83	6889	9.110
18	324	4.243	51	2601	7.141	84	7056	9.165
19	361	4.359	52	2704	7.211	85	7225	9.220
20	400	4.472	53	2809	7.280	86	7396	9.274
21	441	4.583	54	2916	7.348	87	7569	9.327
22	484	4.690	55	3025	7.416	88	7744	9.381
23	529	4.796	56	3136	7.483	89	7921	9.434
24	576	4.899	57	3249	7.550	90	8100	9.487
25	625	5.000	58	3364	7.616	91	8281	9.539
26	676	5.099	59	3481	7.681	92	8464	9.592
27	729	5.196	60	3600	7.746	93	8649	9.644
28	784	5.292	61	3721	7.810	94	8836	9.695
29	841	5.385	62	3844	7.874	95	9025	9.747
30	900	5.477	63	3969	7.937	96	9216	9.798
31	961	5.568	64	4096	8.000	97	9409	9.849
32	1024	5.657	65	4225	8.062	98	9604	9.899
33	1089	5.745	66	4356	8.124	99	9801	9.950

Таблица 2П

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 З*П*

Значения случайных чисел, равномерно распределенных на интервале (0, 1)

10097	32533	76520	13586	34673	64876
37542	04865	64894	74296	24805	24037
08422	68953	19645	09303	23209	02560
99019	02529	09376	70715	38311	31165
12807	99970	80157	36147	64032	36653
80969	09117	39292	74945	66065	74717
20636	10402	00822	91665	31060	10805
15953	34764	35080	33606	85269	77602
88676	74397	04436	27659	63573	32135
98951	16877	19171	78833	73796	45753
34072	76850	36697	36170	65813	39885
45571	82406	35303	42614	86779	07439
02051	65692	68665	74818	73053	85247
05325	47048	90553	57548	28468	28709
03529	64778	35808	34282	60935	20344
11199	29170	98520	17767	14905	68607
23403	09732	11805	05431	39808	27732
18623	88579	83452	99634	06288	98083
83491	25624	88685	40200	86507	58401
35273	88435	99594	67348	87517	64960
52109	40555	60970	93433	50500	73998
50725	68248	29405	24201	52775	67851
13746	70078	18475	40610	68711	77817
36766	67951	90364	76493	29609	11062
91826	08928	93785	61368	23478	34113
65481	17674	17468	50950	79335	51748
80124	35635	17727	08015	82391	90324
74350	99817	77402	77214	50024	23356
69915	26803	66252	29148	24892	09994
09883	20505	14225	68514	83647	76938

Таблица 4П

Ординаты нормальной кривой

(значения функции $f(t)=\frac{1}{\sqrt{2\pi}}\cdot e^{\frac{t^2}{2}}$ **)**

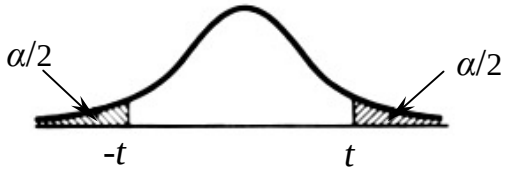
t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3825	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3726	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2987	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1456	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0191	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0053	0.0051	0.0050	0.0048	0.0047	0.0046
3	0.0044	0.0043	0.0042	0.0041	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006

Таблица 5*П*

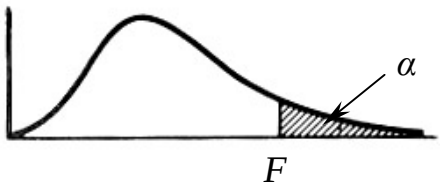
Значение критерия *t* для отбраковки «выскакивающих» вариант

n	***α***			***n***	***α***		
	0.05	**0.01**	**0.001**		**0.05**	**0.01**	**0.001**
5	3.04	5.04	9.43	20	2.15	2.93	3.98
6	2.78	4.36	7.41	25	2.11	2.85	3.82
7	2.62	3.96	6.37	30	2.08	2.80	3.72
8	2.51	3.71	5.73	35	2.06	2.77	3.65
9	2.43	3.54	5.31	40	2.05	2.74	3.60
10	2.37	3.41	5.01	45	2.04	2.72	3.57
11	2.33	3.31	4.79	50	2.03	2.71	3.53
12	2.29	3.23	4.62	60	2.02	2.68	3.49
13	2.26	3.17	4.48	70	2.01	2.67	3.46
14	2.24	3.12	4.37	80	2.00	2.66	3.44
15	2.22	3.08	4.28	90	2.00	2.65	3.42
16	2.20	3.04	4.20	100	1.99	2.64	3.41
17	2.18	3.01	4.13	0	1.96	2.58	3.29
18	2.17	2.98	4.07				

Пороговые значения распределения *T* Стьюдента; *α* для двустороннего критерия



Пороговые значения распределения *F* Фишера



Пороговые значения распределения χ^2 Пирсона

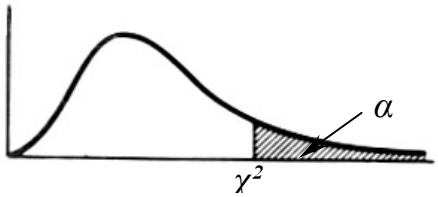


Таблица 6*П*

Значения критерия Стьюдента

Число степеней свободы, *df*	Доверительная вероятность (*P*) Уровень значимости (α)		
	P=0.095 α=0.05	*P*=0.099 α=0.01	*P*=0.0999 α=0.001
2	4.303	9.925	31.598
3	3.182	5.841	12.941
4	2.776	4.604	8.610
5	2.571	4.032	6.859
6	2.447	3.707	5.959
7	2.365	3.499	5.405
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
22	2.074	2.819	3.792
25	2.060	2.787	3.725
30	2.042	2.750	3.646
35	2.030	2.724	3.591
40	2.021	2.704	3.551
45	2.014	2.690	3.520
50	2.008	2.678	3.496
55	2.004	2.669	3.476
60	2.000	2.660	3.460
70	1.994	2.648	3.435
80	1.989	2.638	3.416
90	1.986	2.631	3.402
100	1.982	2.625	3.390
120	1.980	2.617	3.373
>120	1.960	2.5758	3.2905

Таблица 7П

Значения критерия Фишера F при уровне значимости $\alpha = 0.05$ (число степеней свободы указано для дисперсии знаменателя – в строке, для дисперсии числителя – в столбце)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	161.0	200.0	216.0	225.0	230.0	234.0	237.0	239.0	241.0	242.0	244.0	246.0	248.0	250.0	254.0
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.4
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.9	8.8	8.8	8.7	8.7	8.7	8.6	8.5
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	5.9	5.9	5.9	5.8	5.8	5.6
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7	4.7	4.6	4.6	4.5	4.4
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
7	5.6	4.7	4.4	4.1	4.0	3.9	3.8	3.7	3.7	3.6	3.6	3.5	3.4	3.4	3.2
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
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	3.1	3.0	2.9	2.9	2.7
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
11	4.8	4.0	3.6	3.4	3.2	3.1	3.0	3.0	2.9	2.9	2.8	2.7	2.7	2.6	2.4
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
13	4.7	3.8	3.4	3.2	3.0	2.9	2.8	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.2
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

Таблица 8*П*

Значения критерия Фишера *F* при уровне значимости $\alpha = 0.01$ (число степеней свободы указано для дисперсии знаменателя – в строке, для дисперсии числителя – в столбце)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6261	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5
3	31.4	30.8	29.5	28.7	28.4	27.9	27.7	27.5	27.3	27.2	27.0	26.9	26.7	26.5	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.0	10.0	9.7	9.5	9.4	9.0
6	13.7	10.9	9.8	9.1	8.7	8.5	8.3	8.1	8.0	7.9	7.7	7.6	7.4	7.2	6.9
7	12.3	9.5	8.5	7.8	7.5	7.2	7.0	6.8	6.7	6.6	6.5	6.3	6.2	6.0	5.6
8	11.3	8.7	7.6	7.0	6.6	6.4	6.2	6.0	5.9	5.8	5.7	5.5	5.4	5.2	4.9
9	10.6	8.0	7.0	6.4	6.1	5.8	5.6	5.5	5.3	5.3	5.1	5.0	4.8	4.6	4.3
10	10.0	7.6	6.5	6.0	5.6	5.4	5.2	5.1	4.9	4.8	4.7	4.6	4.4	4.2	3.9
11	9.7	7.2	6.2	5.7	5.3	5.1	4.9	4.7	4.6	4.5	4.4	4.2	4.1	3.9	3.6
12	9.3	6.9	5.9	5.4	5.1	4.8	4.6	4.5	4.4	4.3	4.2	4.0	3.9	3.7	3.4
13	9.1	6.7	5.7	5.2	4.9	4.6	4.4	4.3	4.2	4.1	4.0	3.8	3.7	3.5	3.2
14	8.9	6.5	5.6	5.0	4.7	4.5	4.3	4.1	4.0	3.9	3.8	3.7	3.5	3.3	3.0
15	8.7	6.4	5.4	4.9	4.6	4.3	4.1	4.0	3.9	3.8	3.7	3.5	3.4	3.2	2.9
16	8.5	6.2	5.3	4.8	4.4	4.2	4.0	3.9	3.8	3.7	3.5	3.4	3.3	3.1	2.7
17	8.4	6.1	5.2	4.7	4.3	4.1	3.9	3.8	3.7	3.6	3.5	3.3	3.2	3.0	2.6
18	8.3	6.0	5.1	4.6	4.2	4.0	3.8	3.7	3.6	3.5	3.4	3.2	3.1	2.9	2.6
19	8.2	5.9	5.0	4.5	4.2	3.9	3.8	3.6	3.5	3.4	3.3	3.1	3.0	2.8	2.5
20	8.1	5.8	4.9	4.4	4.1	3.9	3.7	3.6	3.5	3.4	3.2	3.1	2.9	2.8	2.4
21	8.0	5.8	4.9	4.4	4.0	3.8	3.6	3.5	3.4	3.3	3.2	3.0	2.9	2.7	2.4
22	7.9	5.7	4.8	4.3	4.0	3.8	3.6	3.4	3.3	3.3	3.1	3.0	2.8	2.7	2.3
23	7.9	5.7	4.8	4.3	3.9	3.7	3.5	3.4	3.3	3.2	3.1	2.9	2.7	2.6	2.3
24	7.8	5.6	4.7	4.2	3.9	3.7	3.5	3.4	3.3	3.2	3.0	2.9	2.7	2.6	2.2
26	7.7	5.5	4.6	4.1	3.8	3.6	3.4	3.3	3.2	3.1	3.0	2.8	2.7	2.5	2.1
28	7.6	5.4	4.6	4.1	3.7	3.5	3.4	3.2	3.1	3.0	2.9	2.7	2.6	2.4	2.1
30	7.6	5.4	4.5	4.0	3.7	3.5	3.3	3.2	3.1	3.0	2.8	2.7	2.5	2.4	2.0
40	7.3	5.2	4.3	3.8	3.5	3.3	3.1	3.0	2.9	2.8	2.7	2.5	2.4	2.2	1.8
60	7.1	5.0	4.1	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.5	2.3	2.2	2.0	1.6
120	6.8	4.8	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.3	2.2	2.0	1.9	1.4
∞	6.6	4.6	3.8	3.3	3.0	2.8	2.6	2.5	2.4	2.3	2.2	2.5	1.9	1.7	1.0

Таблица 9*П*

Значения критерия χ^2

df	Уровень значимости, *α*				
	0.95	**0.75**	**0.25**	***0.05***	**0.01**
1	–	0.10	1.32	3.84	6.63
2	0.10	0.58	2.77	5.99	9.21
3	0.35	1.21	4.11	7.81	11.34
4	0.71	1.92	5.39	9.49	13.28
5	1.15	2.67	6.63	11.07	15.09
6	1.64	3.45	7.84	12.59	16.81
7	2.17	4.25	9.04	14.07	18.48
8	2.73	5.07	10.22	15.51	20.09
9	3.33	5.90	11.39	16.92	21.67
10	3.94	6.74	12.55	18.31	23.21
11	4.57	7.58	13.70	19.68	24.72
12	5.23	8.44	14.85	21.03	26.22
13	5.89	9.30	15.98	22.36	27.69
14	6.57	10.17	17.12	23.68	29.14
15	7.26	11.04	18.25	25.00	30.58
16	7.96	11.91	19.37	26.30	32.00
17	8.67	12.79	20.49	27.59	33.41
18	9.39	13.68	21.60	28.87	34.81
19	10.12	14.56	22.72	30.14	36.19
20	10.85	15.45	23.83	31.41	37.57
21	11.59	16.34	24.93	32.67	38.93
22	12.34	17.24	26.04	33.92	40.29
23	13.09	18.14	27.14	35.17	41.64
24	13.85	19.04	28.24	36.42	42.98
25	14.61	19.94	29.34	37.65	44.31
26	15.38	20.84	30.43	38.89	45.64
27	16.15	21.75	31.63	40.11	46.96
28	16.93	22.66	32.62	41.34	48.28
30	18.49	24.48	34.80	43.77	50.89
40	26.51	33.66	45.62	55.76	63.69
50	34.76	42.94	56.33	67.50	76.15
60	43.19	52.29	66.98	79.08	88.38
70	51.74	61.70	77.58	90.53	100.42
80	60.39	71.14	88.13	101.88	112.33
90	69.13	80.62	98.64	113.14	124.12
100	77.93	90.13	109.14	124.34	135.81

Таблица 10*П*

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
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

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**
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
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

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**
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
100	3.142									

Таблица 11*П*

Значения критерия *U* Уилкоксона – Манна – Уитни

Уровень значимости α = 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 α = 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 12*П*

Значения критерия *T* Уайта при уровне значимости α = 0.05

Большее число наблюдений	Меньшее число наблюдений													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
4			11											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	56	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												

Таблица 13*П*

Значения критерия *Q* Розенбаума

n_2 \ n_1	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
Уровень значимости α = 0.05																
11	6															
12	6	6														
13	6	6	6													
14	7	7	6	6												
15	7	7	6	6	6											
16	7	7	7	7	6	6										
17	7	7	7	7	7	7	7									
18	7	7	7	7	7	7	7	7								
19	7	7	7	7	7	7	7	7	7							
20	7	7	7	7	7	7	7	7	7	7						
21	8	7	7	7	7	7	7	7	7	7	7					
22	8	7	7	7	7	7	7	7	7	7	7	7				
23	8	8	7	7	7	7	7	7	7	7	7	7	7			
24	8	8	8	8	8	8	8	8	8	8	7	7	7	7		
25	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	
26	8	8	8	8	8	8	8	8	8	8	7	7	7	7	7	7
Уровень значимости α = 0.01																
11	9															
12	9	9														
13	9	9	9													
14	9	9	9	9												
15	9	9	9	9	9											
16	9	9	9	9	9	9										
17	10	9	9	9	9	9	9									
18	10	10	9	9	9	9	9	9								
19	10	10	10	9	9	9	9	9	9							
20	10	10	10	10	9	9	9	9	9	9						
21	11	10	10	10	9	9	9	9	9	9	9					
22	11	11	10	10	10	9	9	9	9	9	9	9				
23	11	11	10	10	10	10	9	9	9	9	9	9	9			
24	12	11	11	10	10	10	10	9	9	9	9	9	9	9		
25	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9	
26	12	12	11	11	10	10	10	10	10	9	9	9	9	9	9	9

Таблица 14*П*

Значения величины $z = 0.5 \cdot \ln\frac{1+r}{1-r}$

r	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0100	0.0200	0.0300	0.0400	0.0501	0.0601	0.0701	0.0802	0.0902
0.1	0.1003	0.1105	0.1206	0.1308	0.1409	0.1511	0.1614	0.1717	0.1820	0.1923
0.2	0.2027	0.2132	0.2237	0.2342	0.2448	0.2554	0.2661	0.2769	0.2877	0.2986
0.3	0.3095	0.3206	0.3317	0.3428	0.3541	0.3654	0.3769	0.3884	0.4001	0.4118
0.4	0.4236	0.4356	0.4477	0.4599	0.4722	0.4847	0.4973	0.5101	0.5230	0.5361
0.5	0.5493	0.5627	0.5763	0.5901	0.6042	0.6184	0.6328	0.6475	0.6625	0.6777
0.6	0.6931	0.7089	0.7250	0.7414	0.7582	0.7753	0.7928	0.8107	0.8291	0.8480
0.7	0.8673	0.8872	0.9076	0.9287	0.9505	0.9730	0.9962	1.0203	1.0454	1.0714
0.8	1.0986	1.1270	1.1518	1.1881	1.2212	1.2562	1.2933	1.3331	1.3758	1.4219
0.9	1.4722	1.5275	1.5890	1.6584	1.7380	1.8318	1.9459	2.0923	2.2976	2.6467

Таблица 15*П*

Значения *r* для *z* от 0.00 до 2.99

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.000	0.010	0.020	0.030	0.040	0.050	0.060	0.070	0.080	0.090
0.1	0.100	0.110	0.119	0.129	0.139	0.149	0.159	0.168	0.178	0.188
0.2	0.197	0.207	0.217	0.226	0.236	0.245	0.254	0.264	0.273	0.282
0.3	0.291	0.300	0.310	0.319	0.328	0.336	0.345	0.354	0.363	0.371
0.4	0.380	0.389	0.397	0.405	0.414	0.422	0.430	0.438	0.446	0.454
0.5	0.462	0.470	0.478	0.485	0.493	0.501	0.508	0.515	0.523	0.530
0.6	0.537	0.544	0.551	0.558	0.565	0.572	0.578	0.585	0.592	0.598
0.7	0.604	0.611	0.617	0.623	0.629	0.635	0.641	0.647	0.653	0.658
0.8	0.664	0.670	0.675	0.681	0.686	0.691	0.696	0.701	0.706	0.711
0.9	0.716	0.721	0.726	0.731	0.735	0.740	0.744	0.749	0.753	0.757
1.0	0.762	0.766	0.770	0.774	0.778	0.782	0.786	0.790	0.793	0.797
1.1	0.801	0.804	0.808	0.811	0.814	0.818	0.821	0.824	0.828	0.831
1.2	0.834	0.837	0.840	0.843	0.846	0.848	0.851	0.854	0.857	0.859
1.3	0.862	0.864	0.867	0.869	0.872	0.874	0.876	0.879	0.881	0.883
1.4	0.885	0.888	0.890	0.892	0.894	0.896	0.898	0.900	0.902	0.903
1.5	0.905	0.907	0.909	0.910	0.912	0.914	0.915	0.917	0.919	0.920
1.6	0.922	0.923	0.925	0.926	0.928	0.929	0.930	0.932	0.933	0.934
1.7	0.935	0.937	0.938	0.939	0.940	0.941	0.943	0.944	0.945	0.946
1.8	0.947	0.948	0.949	0.950	0.951	0.952	0.953	0.954	0.955	0.955
1.9	0.956	0.957	0.958	0.959	0.960	0.960	0.961	0.962	0.963	0.963

Значения *r* для *z* от 0.00 до 2.99

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.964	0.965	0.965	0.966	0.967	0.967	0.968	0.969	0.969	0.970
2.1	0.970	0.972	0.972	0.972	0.973	0.973	0.974	0.974	0.975	0.975
2.2	0.976	0.976	0.977	0.977	0.978	0.978	0.979	0.979	0.979	0.980
2.3	0.980	0.981	0.981	0.981	0.982	0.982	0.982	0.933	0.983	0.983
2.4	0.984	0.934	0.984	0.985	0.935	0.935	0.936	0.986	0.986	0.986
2.5	0.987	0.987	0.987	0.987	0.988	0.988	0.938	0.988	0.989	0.989
2.6	0.989	0.939	0.989	0.990	0.990	0.990	0.990	0.990	0.991	0.991
2.7	0.991	0.991	0.991	0.992	0.992	0.992	0.992	0.992	0.992	0.993
2.8	0.993	0.993	0.993	0.993	0.993	0.993	0.994	0.994	0.994	0.994

Таблица 16*П*

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	**16**	0.468	0.59	**40**	0.304	0.393
2	0.95	0.99	**17**	0.456	0.575	**45**	0.288	0.372
3	0.878	0.959	**18**	0.444	0.561	**50**	0.273	0.354
4	0.811	0.917	**19**	0.433	0.549	**60**	0.25	0.325
5	0.754	0.874	**20**	0.423	0.537	**70**	0.232	0.302
6	0.707	0.834	**21**	0.413	0.526	**80**	0.217	0.283
7	0.666	0.798	**22**	0.404	0.515	**90**	0.205	0.267
8	0.632	0.765	**23**	0.396	0.505	**100**	0.195	0.254
9	0.602	0.735	**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
11	0.553	0.684	**26**	0.374	0.478	**200**	0.138	0.181
12	0.532	0.661	**27**	0.367	0.47	**300**	0.113	0.148
13	0.514	0.641	**28**	0.361	0.463	**400**	0.098	0.128
14	0.497	0.623	**30**	0.349	0.449	**500**	0.088	0.115
15	0.482	0.606	**35**	0.325	0.418	**1000**	0.062	0.081

Таблица 17*П*

Минимальные значения коэффициента ранговой корреляции Спирмена, достоверно отличные от нуля (*df* = *n*–2)

n	***α* = 0.05**	***α* = 0.01**	***n***	***α* = 0.05**	***α* = 0.01**	***n***	***α* = 0.05**	***α* = 0.01**
5	0.94		17	0.48	0.62	29	0.37	0.48
6	0.85		18	0.47	0.60	30	0.36	0.47
7	0.78	0.94	19	0.46	0.58	31	0.36	0.46
8	0.72	0.88	20	0.45	0.57	32	0.36	0.45
9	0.68	0.83	21	0.44	0.56	33	0.34	0.45
10	0.64	0.79	22	0.43	0.54	34	0.34	0.44
11	0.61	0.76	23	0.42	0.53	35	0.33	0.43
12	0.58	0.73	24	0.41	0.52	36	0.33	0.43
13	0.56	0.70	25	0.40	0.51	37	0.33	0.42
14	0.54	0.68	26	0.39	0.50	38	0.32	0.41
15	0.52	0.66	27	0.38	0.49	39	0.32	0.41
16	0.50	0.64	28	0.38	0.48	40	0.31	0.40

ТЕМАТИЧЕСКИЙ УКАЗАТЕЛЬ

СОДЕРЖАНИЕ

Учебное издание

Ивантер Эрнест Викторович
Коросов Андрей Викторович

Введение в количественную биологию

Редактор *О. В. Обарчук*

Рисунок на обложке – *Ю. М. Коросова*
Компьютерная верстка – *А. В. Коросов*

Подписано в печать 01.03.2011. Формат 60 x 84 $^{1}/_{16}$.
Бумага офсетная. Гарнитура Академическая.
18 уч.-изд. л. Тираж 300 экз. Изд. № 4.

Государственное образовательное учреждение
высшего профессионального образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Отпечатано в типографии Издательства ПетрГУ
185910, Петрозаводск, пр. Ленина, 33

Ивантер Эрнест Викторович – член-корреспондент РАН, доктор биологических наук, профессор, заведующий кафедрой зоологии и экологии, декан эколого-биологического факультета Петрозаводского государственного университета, главный научный сотрудник Института биологии КарНЦ РАН. Автор более 330 печатных работ, в том числе более 30 монографий, учебных пособий и научно-популярных книг о природе и животном мире Карелии. Э. В. Ивантер – вице президент Всероссийского териологического общества, действительный член Российской академии естественных наук. Заслуженный деятель науки Российской Федерации и Республики Карелия.

Коросов Андрей Викторович – профессор, доктор биологических наук, профессор кафедры зоологии и экологии Петрозаводского государственного университета. Автор более 120 печатных работ, в том числе более 10 монографий, учебных пособий и научно-популярных книг о животном мире Карелии. А. В. Коросов – член Всероссийского герпетологического общества и Всероссийского териологического общества. Соросовский доцент. Заслуженный деятель науки Республики Карелия.