STATISTICAL METHODS FOR RESEARCH WORKERS

by

R. A. FISHER, Sc. D., F. R. S.

TWELEFTH EDITION — REVISED

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 98 GREAT RUSSEL STREET, W. C. 1954

ОГЛАВЛЕНИЕ

Р. А. Фишер

СТАТИСТИЧЕСКИЕ МЕТОДЫ ДЛЯ ИССЛЕДОВАТЕЛЕЙ

Редактор *А. И. Латышев* Художник *И. Д. Богачев*
Техн. редактор *Н. Д. Пятакова* Корректор *М. И. Илларионова*

Сдано в набор 11/X 1957 г. Подписано к печати 7/V 1958 г.
Бумага 60×92¹/₁₆ Бум. л. 12,5. Печ. л. 16,75. Уч.-изд. л. 28,83.
А-06605 Индекс ГЛ-28
Москва, Госстатиздат, Кирова, 39
Заказ № 1161. Цена 9 р. 90 к.

Типография № 8 УПП Ленсовнархоза. Ленинград, Прачечный пер., д. 6.

ОТ ИЗДАТЕЛЬСТВА

Труды Р. А. Фишера оказали большое влияние на развитие математической статистики. Поэтому Госстатиздат выпускает книгу Р. А. Фишера «Статистические методы для исследователей». Автор этой книги является видным английским теоретиком математической статистики. Он в течение многих лет работал руководителем статистической лаборатории Ротамстедской опытной станции (Англия). Данная книга, как указывает автор в предисловии, создавалась в результате его сотрудничества с биологами-экспериментаторами.

Следует учитывать, что Фишеру, как теоретику современной буржуазной статистики, свойственны буржуазная узость и формализм во взглядах. По его концепции количественный анализ является универсальным и абсолютным статистическим средством нашего познания. Он фактически полностью игнорирует качественную сторону явлений. Достаточно указать на его утверждение о том, что социальные учения могут подняться до уровня действительных наук только в той мере, в какой они используют аппарат статистики и строят свои выводы на статистической аргументации. При этом подразумевается математическая статистика, которую автор рассматривает как универсальную науку.

Для Р. А. Фишера характерна обычная для буржуазных ученых двойственность: с одной стороны, подчиняясь объективной необходимости, они делают ценные наблюдения и выводы, с другой стороны, находясь под влиянием идеалистических мировоззрений, они придают полученным ими результатам соответствующую этим взглядам окраску.

Если рассматривать описываемые им методы, то нельзя отрицать их логичность. Если же обратиться к конкретным примерам, которые приведены для иллюстрации этих методов, то здесь встречаемся с некоторыми ненаучными положениями в трактовке социальных вопросов. Так, из одного взятого им примера вытекает, что если один из двух (монозиготных) близнецов оказался преступником, то почти наверняка и второй из близнецов также будет преступником. Отсюда получается, что преступность уже была заложена в том оплодотворенном яйце, из которого в по-

следующем развились эти близнецы. Ясно, что такая концепция «биологического» происхождения преступности совершенно игнорирует социальные условия жизни людей, от которых абстрагируются буржуазные социологи и экономисты.

По своему изложению книга имеет неровный характер: в одних местах автор излагает вопрос весьма детально, вплоть до подробного описания всех расчетов, в других же местах, он дает только общий набросок проблемы. Описание автором статистических методов нельзя считать простым и полностью популярным, знакомство с этой книгой потребует от читателя известного напряжения.

Данная книга хотя и адресована исследователям — биологам и агрономам, однако она может представлять известный интерес и для статистиков-экономистов. Использовав правильные положения математической статистики, советские читатели отбросят все выводы и рассуждения, неприемлемые для подлинной статистической науки.

ПРЕДИСЛОВИЕ АВТОРА

За несколько лет до подготовки к изданию книги автору пришлось работать в тесном сотрудничестве с рядом биологических отделов Ротамстедской опытной станции; книга явилась прямым результатом такого сотрудничества. Повседневное соприкосновение со статистическими проблемами в том их виде, в каком они возникают в лабораторных работах, служило известным стимулом для ряда чисто математических исследований, которые в дальнейшем легли в основу новых методов статистического анализа. В результате этой работы выявилось, что традиционный статистический механизм, установленный биометрической школой, не соответствует условиям и потребностям практических исследований. Бесполезная и в то же время трудоемкая работа по вычислению бесчисленных коэффициентов корреляции и отход от реальных трудностей выборочного метода, выразившийся в полном игнорировании малых выборок, — все это не оправдывало слишком больших претензий биометрической школы. Приемы, рекомендуемые ею, были не только неверно направлены, но они при всей их трудоемкости не были даже достаточно точными. По мнению автора, только взявшись всерьез за проблемы малой выборки, можно найти точные критерии, соответствующие практическим задачам. При поддержке моих коллег и при весьма ценной помощи со стороны ныне покойного В. С. Госсета («Стьюдента»), его ассистента Э. Самерфильда и мисс В. А. Макензи было подготовлено первое издание книги, успешно выдержавшее строгую критику, неизбежную во всяком новом деле.

В данный момент точные критерии существенности уже не нуждаются в защите. Тот факт, что в течение длительного периода непрерывно возрастал спрос на книгу, предназначенную только для узкого круга читателей, уже сам по себе оправдывает по крайней мере некоторые нововведения в ее построении, которые сначала могли показаться спорными (определение степеней свободы, применение фиксированных уровней вероятности при составлении таблиц, предназначенных для оценки существенности, дисперсионный анализ, необходимость рэндомизации при проведении опытов и т. д.). Автор показал практическое значение многих из ранее известных математических положений, которые другим исследователям казались просто академическими тонкостями. При написании этой книги автор питал надежду, что лица, обладающие некоторым опытом исследовательской работы, оценят книгу, которая, не касаясь самой математической теории статистических методов, все же дает новейшие достижения теории в форме практических приемов, соответствующих материалу, с которым исследователи встречаются в действительности. Ведь практическое применение общих теорем — это совсем иное дело, чем их установление при помощи математических доказательств. Все, что требуется для практического применения, — это достаточно

глубокое понимание самого смысла этих теорем, что вполне доступно и без знакомства с доказательствами. Для поддержания изложения на современном уровне науки, в каждое последующее издание книги вводится новый материал, отражающий развитие теории, которая в этих случаях выражалась в практических решениях.

В основном новые методы были направлены на упрощение обработки статистических данных. Главным препятствием к внедрению точных статистических методов до сих пор является известный консерватизм некоторых общих курсов по элементарной статистике, дающих стереотипные и ненужные приближенные решения задач в то время, когда имеются вполне точные их решения. При знакомстве с книгой читатель должен иметь в виду, что отход от этих традиций отнюдь не является капризом автора; только отрешившись от этого, он получит определенную пользу от книги.

Особенности расположения материала все время сохранялись такими, какими они были при первом издании этой книги. В более поздних изданиях некоторых общих курсов по элементарной статистике дисперсионный анализ по педагогическим соображениям дается на более ранней стадии, чем это сделано у меня, и ему уделяется гораздо больше места. Следовательно, авторы этих курсов идут дальше меня в том отношении, что подходят к проблеме дисперсионного анализа непосредственно и без связи его с корреляцией. В оправдание моего отказа взять на себя труд по коренной переработке книги сошлюсь на то, что для полного понимания данного предмета имеет определенное значение знакомство с постановкой вопросов у авторов более ранних работ, а также возможность свести эти проблемы к системе идей, дающих наиболее простое и понятное их решение. Поэтому я здесь довольствуюсь рассмотрением дисперсионного анализа в качестве второго, альтернативного подхода к уже решенным ранее в параграфах 24 и 24.1 задачам.

С точки зрения лучшего построения книги, материал от параграфа 30 до 40, относящийся к вопросу о корреляции, я должен был бы теперь перенести несколько дальше и не рассматривать его, пока не будет накоплен опыт по дисперсионному анализу, и уже после этого рассмотреть учения о корреляции и частной корреляции в соответствии с их значением в биометрической литературе, находящейся под непосредственным влиянием этой теории.

Выявившаяся потребность в обобщающем и детальном изложении принципов, лежащих в основе статистических оценок, заставила ввести во второе издание настоящей книги новую, девятую главу. В первом издании этот вопрос рассматривался только в самой общей форме и поэтому, несмотря на его практическое значение, не обратил на себя внимание преподавательских кругов и не заставил их изменить порочную практику обучения студентов неправильным методам статистического анализа. Эта новая глава заменила собой параграф 6 и пример 1 из первого издания.

В третьем издании в нее введено два новых параграфа (57.1 и 57.2), дающих более широкое представление о применении метода максимального правдоподобия и об определении количества информации. Позднее в превосходной книге К. Матера «The Measurement of Linkage in Heredity» были подробно описаны приемы, относящиеся к весьма разнообразному кругу генетических задач.

В связи с выявившейся в некоторых исследованиях потребностью вычисления полиномов выше пятого порядка в параграф 27 в третьем издании книги было сделано добавление об общих принципах построения ортогональных полиномов. Мисс Ф. Э. Аллан дала простое и прямое обоснование методов, изложенных в параграфах 28 и 28.1.

В четвертом издании было полностью переработано посвященное системе обозначений приложение к главе III, так как к этому времени выяснилось, что недостатки системы моментов оказались бо́льшими, чем те преимущества, которые ей ранее незаслуженно приписывались. В этом же издании был дан и принципиально новый материал, характеризующий более широкое использование ковариационного анализа; этому вопросу был посвящен параграф 49.1. В связи с тем, что некоторые авторы испытывали в этом случае затруднения при оценке существенности отклонений от линии регрессии, в пятом издании этот параграф был несколько расширен.

Другими новыми параграфами в пятом издании были: 21.01, посвященный поправке на непрерывность, предложенной Ф. Иейтсом, и 21.02, где установлен точный критерий существенности для таблицы 2×2. Научных работников, имеющих дело с уравнениями регрессии при большом числе переменных, заинтересует параграф 29.1, в котором даны относительно простые поправки, которые вводятся в уравнение регрессии при исключении одной или нескольких независимых переменных. Имеющаяся возможность исключать переменные без проведения заново всех весьма трудоемких вычислений позволяет исследователю включать в число независимых переменных более широкий круг величин, чем тот, который впоследствии удастся использовать в уравнении регрессии. Параграф 5, прежде дававший перечень таблиц, применяемых при оценке существенности, теперь отведен для справки исторического характера об авторах, разрабатывавших вопрос о логической основе статистических выводов.

В шестом издании пример 15.1 параграфа 22 знакомит с новым критерием однородности при последовательном многократном подразделении данных на группы. В этом же издании была рассмотрена формула Уэркинга и Хотеллинга для выборочной ошибки значений зависимой переменной, вычисленных по уравнению регрессии. В параграфе 29.2 дано применение способа последовательного суммирования к вычислению полиномов.

По совету доктора В. Э. Деминга я расширил таблицу z, включив в нее уровень существенности 0,1 процента. Такие высокие

уровни существенности нужны, когда следует сделать выбор одного из нескольких критериев, которые *априорно* близки друг к другу.

Следует отметить два наиболее важных изменения, включенные в седьмое издание. Во-первых, в параграфе 27 дано введение в общую теорию ортогональных полиномов на основе ортогональных сравнений между наблюдениями, так как этот подход к указанной теории для многих исследователей кажется более доступным. В этом случае получается более простая арифметическая конструкция, поддающаяся обобщению при помощи не столь сложных алгебраических выражений. В настоящее время в «Статистических таблицах» даны ряды независимых сравнений вплоть до пятой степени. Во-вторых, в это издание был добавлен параграф 49.2, дающий представление о новом и важном использовании совокупности показателей для построения по ним наилучшей дискриминантной функции. Для этого случая пока найдены только приближенные критерии существенности, и данный вопрос подлежит дальнейшему изучению. Надо сказать, что представляется просто удивительным разнообразие тех проблем, которые разрешаются при помощи этого метода.

В девятом издании новый параграф дает критерий однородности наблюдений, полученных из разных источников; этот вопрос фактически и логически является дополнением к описанным в предыдущих параграфах методам объединения независимых данных. В десятом издании распространено применение критерия t на случай определения доверительных пределов для отношения средних или коэффициентов регрессии (параграф 26.2).

Принципы экспериментирования, изложенные в ряде параграфов главы VIII, даны в столь кратком виде, что по этому изложению нельзя составить полного представления о самом предмете; этот вопрос рассмотрен только с узко статистической точки зрения, так как его подробное изложение дано в другой специальной книге автора «Планирование опытов» (изд. Оливер и Бойд, 1935, 1937, 1942 и 1946 гг.). Точно так же опубликованы в виде отдельной книги «Статистические таблицы» (изд. Оливер и Бойд, 1938, 1943 и 1946 гг.) те таблицы, которые даны в настоящей книге, с присоединением к ним некоторых других таблиц, предназначенных для различных статистических целей, и в сопровождении примеров их применения. Благодаря этим двум публикациям имеется возможность не увеличивать объема книги до размера, который затруднил бы ее использование в качестве целого и компактного курса. Читатель может быть уверен, что и эти книги вполне для него доступны. Следует отметить, что номера параграфов, таблиц и примеров при включении нового материала не менялись, так что ссылки на них, но отнюдь не на страницы, остаются в силе, к какому бы изданию они не относились.

ОТДЕЛ СЕЛЕКЦИИ, КЭМБРИДЖ.

ГЛАВА ПЕРВАЯ

ВВЕДЕНИЕ

1. Предмет статистики

Статистика, как наука, является одним из разделов прикладной математики и ее можно рассматривать как математику, применяемую при разработке результатов массового наблюдения. В статистике, как и в других математических науках, одна и та же формула в одинаковой мере относится к самым различным материальным объектам. В ней мы имеем дело с абстрактными представлениями относительно этих объектов, так как при разработке математических основ этой науки приходится отвлекаться от многих сторон изучаемого материала. В связи с этим мы сначала рассмотрим предмет статистики в самом общем виде и с трех различных точек зрения, а далее в более строгой математической форме покажем, что именно эти типы проблем встречаются во всех отдельных случаях. Статистику можно рассматривать как: 1) учение о *совокупностях*, 2) учение о *вариации* и 3) учение о методах *приведения данных к компактной форме*.

Слово «статистика» в прямом смысле означает учение о народонаселении, объединенном в некоторую политическую единицу. Однако статистические методы, собственно говоря, имеют дело не с самими совокупностями людей и не с социальными группами, так как в действительности у нас никогда не бывает данных, полностью и во всех отношениях характеризующих человека, в результате чего получаемая нами совокупность всегда является в какой-то мере абстрактной. Если мы, положим, имеем данные о росте 10 000 рекрутов, то это скорее всего совокупность ростов, а не совокупность рекрутов. Но вместе с тем, статистика все же изучает совокупности или объединения объектов, но отнюдь не отдельные индивидуальности. Научные теории, которые имеют дело с большими агрегатами объектов и не рассматривают свойства отдельных объектов в их изолированном виде, как, например, кинетическая теория газов, теория селекции или химическая теория действия масс, по существу, основаны на статистической аргументации и обычно впадают в ошибки, как только оставляют эту позицию. Это особенно отчетливо видно на примере современ-

ной квантовой теории. Статистические методы являются существенным элементом в социальных науках и в основном именно с помощью этих методов социальные учения могут подняться до уровня наук. Эта зависимость социальных наук от статистических методов служит, в частности, источником того неправильного представления, что будто бы статистика является ветвью экономики, тогда как фактически статистические методы столь же тесно связаны с обработкой экономических данных, как с биологическими и другими научными результатами.

Представление о совокупности применимо не только к собранию живых или просто материальных объектов, но имеет и более широкое истолкование. Если такое наблюдение, как измерение какого-либо предмета, будет повторено определенное число раз, то эта масса результатов также будет составлять совокупность измерений. Совокупности этого рода являются предметом отдельного учения, называемого *теорией ошибок,* — одного из старейших и наиболее изученных разделов статистики. Подобно тому, как одно наблюдение может рассматриваться в качестве отдельной единицы совокупности, а его повторение — как образование совокупности, точно так же результат некоторого осуществленного эксперимента может рассматриваться как один из возможных результатов, относящихся к совокупности таких экспериментов. Существующий полезный обычай повторения экспериментов или проведения повторных наблюдений указывает на тот, хотя и не всегда вполне осознанный, факт, что предметом нашего изучения в данном случае является не отдельный результат, а совокупность возможностей, которую и репрезентируют наши эксперименты или наблюдения. Вычисление таких статистических показателей, как средние величины и средние квадратические ошибки, является первым шагом при получении некоторых сведений о такой совокупности.

Определение статистики, как учения о вариации, вполне естественно вытекает из определения статистики, как науки о совокупностях; совокупность объектов, которые полностью и во всех отношениях идентичны между собой, может быть описана целиком, если дать описание единичного ее члена и указать численность всех этих членов. Совокупности, которые являются предметом статистического изучения, всегда характерны тем, что они изменчивы в одном или нескольких отношениях. Этим утверждением, что статистика изучает вариацию, подчеркивается существенное различие между целями современной статистики и задачами ее предшественницы. До недавнего времени многие видные исследователи в этой области не видели никакой иной задачи, кроме простого объединения и усереднения статистических данных. Вариация, взятая сама по себе, не была предметом изучения, и на нее смотрели разве только как на досадное обстоятельство, приводящее к снижению точности средней величины. Например, кривая ошибок средней в выборках из нормальной совокупности известна уже столетие, а такая же кривая для среднего квадратического отклонения стала предметом исследования только с 1915 г. С современной точки зрения изучение причин изменчивости любой переменной величины, начиная с урожайности пшеницы и кончая интеллектом человека, должно производиться на основе измерения и анализа вариации, которая и сама по себе имеет большое значение.

Учение о вариации приводит непосредственно к концепции *распределения численностей.* Эти распределения могут быть различного вида; число классов, по которым распределена некоторая совокупность, может быть конечным или бесконечно большим; при количественной переменной интервалы этих классов могут быть конечными или бесконечно малыми. В простейшем из возможных случаев, когда имеется только два класса, как например мужской и женский пол новорожденного, распределение просто определяется той пропорцией, в которой встречаются члены, принадлежащие к тому и другому классу. Например, установлено, что встречается примерно 51% рождений мальчиков и 49% рождений девочек. В других случаях изменчивость может быть прерывной при неопределенном заранее числе классов, как например при учете количества детей в различных семьях; здесь распределение будет определяться числом семей с 0, 1, 2, ... детьми, число же классов в этом случае должно быть таким, чтобы были включены и самые большие семьи. Варьирующая величина (например, число детей в семье) называется *случайной переменной*; распределение численностей характеризует, как часто случайная переменная принимает каждое из возможных своих значений. Наконец, в ряде случаев случайная переменная (например, рост человека) может принимать любые промежуточные значения внутри интервала ее варьирования; такая случайная переменная называется непрерывной; соответствующее распределение численностей может быть выражено в форме математической функции случайной переменной, причем возможны два случая: 1) определение той доли всей совокупности, в которой случайная переменная имеет значения меньшие, чем некоторая заданная величина, и 2) установление при помощи математического дифференцирования функции распределения той (бесконечно малой) доли совокупности, которая соответствует некоторому бесконечно малому элементу в интервале изменения случайной переменной.

Представление о распределении численностей в одинаковой мере применимо как в случае, когда оно относится к совокупностям, ограниченным по своему объему, так и к бесконечным совокупностям, но случаи, относящиеся к этому последнему виду распределений, имеют более важное значение. Конечная совокупность может быть подразделена только на несколько ограниченных частей и не может выражать собой непрерывную изменчивость. Кроме того, в большинстве случаев только бесконечная совокупность может точно отразить всю массу подлежащих на-

шему изучению возможностей. Фактические же наблюдения, число которых всегда является конечным, будут в этом случае представлять собой только выборку из этой массы возможностей. Распределение бесконечной совокупности определяет доли совокупности, относящиеся к отдельным классам, причем в этом случае может быть: 1) конечное число долей, составляющих единицу, как это, например, имеет место в распределении численностей по Менделю, 2) бесконечный ряд конечных долей, составляющих единицу и 3) математическая функция, определяющая долю целого для каждого из бесконечно малых элементов, на которые может быть разделена область изменения случайной переменной. Последний случай может быть представлен в виде кривой распределения; для ее построения значения переменной откладываются по горизонтальной оси, а относительные частоты — по вертикальной оси. Доля совокупности, соответствующая определенному интервалу изменения случайной величины, будет представлена площадью кривой, опирающейся на данный отрезок горизонтальной оси. Следует заметить, что эти положения, относящиеся к кривым распределения, приложимы только к таким бесконечным совокупностям, переменная величина которых изменяется непрерывно.

Изучение вариации не сводится только к простому измерению ее размера, но включает в себя и проблемы качественного характера, относящиеся к типам и формам изменчивости. Большое значение имеет учение о совместной вариации двух или большего числа переменных. Это учение, возникшее из работ Гальтона и Пирсона, известно под названием теории *корреляции,* но его следовало бы более точно назвать учением о сопряженной вариации.

Третий аспект, в котором следует рассматривать задачи статистики, обусловлен тем обстоятельством, что на практике всегда возникает необходимость свести первоначальную массу данных к небольшому числу показателей. Всякий исследователь, который проводил систематические и обширные наблюдения, вероятно, знаком с настойчивой необходимостью привести свои результаты к компактной и удобной форме. Никакой человеческий ум не способен вместить в себя все содержание более или менее значительного количества числовых данных. Поэтому мы всегда стремимся к тому, чтобы отразить в относительно небольшом числе сводных показателей наиболее важную и существенную информацию, содержащуюся в данной массе наблюдений. Все это является простой практической необходимостью и теория статистики с этим должна считаться. В некоторых случаях представляется возможным исчерпать при помощи одного или небольшого числа показателей всю информацию, содержащуюся в наблюдениях, но, как правило, исследователя интересуют только основные результаты, которым и надлежит придать простую числовую форму, способную осветить стоящие перед исследователем вопросы. Число независимых фактов, содержащихся в первоначальных данных, обычно гораздо больше количества тех фактов, которые заслуживают внимания, и, следовательно, бóльшая часть сведений, заключенных во всем объеме фактических наблюдений, является второстепенной и несущественной. Поэтому процесс статистического исследования имеет целью исключить эту второстепенную информацию и выявить основные и существенные сведения, имеющиеся в наблюдениях.

2. Общий метод, определение сводных показателей — статистик

Различие между той информацией, которая названа второстепенной, и основной информацией, состоит в следующем. Даже в простейших случаях статистического исследования наблюдения (или ряды наблюдений), с которыми мы имеем дело, рассматриваются в качестве выборки из некоторой гипотетической совокупности таких же наблюдений, возможных при сохранении данных условий. Распределение такой совокупности может быть выражено в виде некоторого математического закона, формула которого содержит в себе определенное, обычно небольшое, число *параметров,* или «констант». Эти параметры являются характеристиками данного генерального распределения. Если бы мы знали точные значения параметров распределения, то тем самым нам было бы известно все (и даже более этого), что нам может сообщить любая выборка из этой совокупности. Фактически мы не имеем возможности определить эти параметры точно, но мы можем произвести их приближенную оценку. Такие оценки, называемые *статистиками,* определяются, конечно, по выборочным данным как некоторые сводные показатели. Если мы установим математическую форму распределения, соответствующего нашим наблюдениям, и на основе этих последних вычислим наилучшие из возможных оценок для искомых параметров, то очевидно, остающиеся после этого данные содержат в себе немного или даже не содержат ничего из того, что они могли бы сверх этого сообщить нам о генеральной совокупности. Таким образом, с помощью этих оценок мы извлекаем из наблюдений всю основную и существенную информацию относительно этой совокупности.

Значение подобного рода оценок в большой мере увеличивается, если мы сможем определить природу и размер ошибок, которыми они сопровождаются. В тех случаях, когда имеется возможность основываться на математической форме генеральной совокупности, указанная задача определения ошибок оценки сводится просто к математическому выводу из этой формы закона распределения каждой из таких статистик. Этот тип статистических задач, который еще до недавнего времени оставался без внимания, является основой для создания критериев существенности, позволяющих установить, соответствуют или не соответ-

ствуют наблюденные данные некоторой определенной гипотезе. Частным случаем является задача, в которой необходимо установить адекватность гипотетической формы генеральной совокупности и наблюденных результатов.

Из сказанного следует, что задачи, возникающие в связи с анализом результатов наблюдений, можно подразделить на три типа.

1. Проблема *спецификации,* которая состоит в выборе математической формы генеральной совокупности.

2. После того, как задача спецификации решена, возникает проблема *оценки.* Она заключается в том, что следует установить способ вычисления по данной выборке статистики, пригодной для оценки неизвестного параметра генеральной совокупности.

3. Проблема *распределения* состоит в выводе точной математической формы распределения наших оценок в случайных выборках и в определении других статистик, предназначенных для проверки пригодности произведенной ранее спецификации (критерии *согласия*).

Таким образом, статистическая обработка некоторой массы наблюдений логически содержит в себе то же чередование индуктивного и дедуктивного методов, которое вообще свойственно науке. Сначала со всей тщательностью формулируется некоторая гипотеза; из нее дедуктивным путем выводятся логические следствия; эти следствия сравниваются с надлежащими наблюдениями. Если эти последние находятся в полном соответствии с дедуктивными выводами, то гипотеза считается подтвержденной, по крайней мере до тех пор, пока не будут получены новые и более точные данные. В своей работе «Планирование опытов» (The Design of Experiments) я дал в этом аспекте более подробное изложение логических основ построения экспериментов.

Наличие в приведенной выше общей схеме статистического исследования дедуктивных выводов, относящихся к выборкам и покоящихся на допущении существования генеральных совокупностей, из которых эти выборки взяты, определяет собой то особое положение, которое занимает в статистике классическая *теория вероятностей.* Если дана некоторая генеральная совокупность, то мы имеем возможность определить вероятность появления данной выборки и вместе с этим вероятность (если эта задача имеет более или менее простое математическое решение) появления данного значения статистики, исчисленной по этой выборке. Указанная выше проблема распределения может рассматриваться, как приложение и соответствующее развитие теории вероятностей. Из распределений, рассматриваемых в теории вероятностей, мы будем иметь дело с биноминальным распределением Бернулли, с нормальным распределением Лапласа и распределением Пуассона. В течение длительного периода времени, примерно полтора столетия, делались попытки распространить понятие вероятности на обратный переход от предполагаемых (или наблюденных) выборок к соответствующим генеральным совокупностям. Такие выводы обратного характера обычно выделялись под рубрикой *обратных вероятностей* и были одно время общепринятыми. Здесь нет места входить в детали длительной полемики по этому поводу; в данном изложении основ статистической науки достаточно будет повторить, не вдаваясь в подробности, мою личную точку зрения, состоящую в том, что теория обратной вероятности основана на ошибке и должна быть полностью отброшена. Вывод, относящийся к генеральной совокупности, из которой взяты имеющиеся у нас выборки, не может быть на основе этой теории выражен в терминах вероятностей, исключая разве тривиальный случай, когда сама генеральная совокупность является также выборкой из сверхсовокупности, форма которой нам известна.

Используемые далее для индуктивных выводов критерии существенности, обозначаемые через t и z, имеют другую вероятностную основу, в значительной степени отличающуюся от концепции обратных вероятностей, и поэтому они освобождены от недостатков этой последней. Эти критерии позволяют строить вероятностные суждения относительно генеральных совокупностей на основе нового и неизвестного классическим авторам понятия о вероятности. В отличие от суждений, основанных на классическом понятии вероятности, эти новые суждения, связанные с применением указанных критериев, называются выводами, основанными на доверительной вероятности.

Исходя из теории обратных вероятностей, одно время делался неправильный вывод о том, что мы не можем, основываясь на данной выборке, делать какие-либо заключения о соответствующей генеральной совокупности. Эта точка зрения полностью отрицала правомерность применения статистических методов во всех экспериментальных науках. Как теперь установлено, математическая концепция вероятности в большинстве случаев не дает чувства уверенности или неуверенности в таких индуктивных выводах, и математическая величина, пригодная в данном случае для обоснованного выбора из всех возможных и различных генеральных совокупностей той, которая соответствует имеющейся выборке, фактически не имеет свойств вероятности. Для отличия этой величины от вероятности я назвал ее «правдоподобием» [1], так как оба слова «правдоподобие» и «вероятность» в разговорном языке используются без четкого их разграничения для обозначения, по существу, одного и того же понятия.

3. Классификация статистик

Решения проблем, относящихся к распределениям (эти проблемы могут просто рассматриваться как дедуктивные задачи из теории вероятностей), дают нам возможность найти критерии

[1] Правдоподобие под названием «функция мощности» находит свое применение в одном специальном случае, а именно при сравнении своего рода чувствительности различных критериев существенности.

для установления существенности тех или иных статистических результатов и адекватности гипотетических генеральных совокупностей, положенных в основу наших выводов, с имеющимися данными. Но, кроме этого, они дают реальную основу для выбора из числа статистик наиболее подходящих для оценки неизвестных параметров. Статистики, служащие для этой цели, могут быть подразделены на несколько классов в соответствии с особенностями их распределения при больших выборках.

Если некоторая статистика, например средняя, вычислена на основе большой выборки, то, как правило, мы ей припишем большую точность. Действительно, довольно часто, но отнюдь не всегда, будет верным то положение, что если найти некоторое количество таких статистик и сравнить их между собой, то различия между ними будут становиться все меньше и меньше по мере того, как выборки, на основе которых они получены, будут делаться все большими и большими по своему объему. При неограниченном увеличении объема выборок такие статистики обычно стремятся к определенному значению, характерному для данной генеральной совокупности, и, следовательно, становятся в связь с параметрами этой совокупности. Поэтому, если такая статистика используется для оценки параметров генеральной совокупности, то имеется только одна функция этих параметров, которая на законном основании может быть оценена при помощи данной статистики. Если же наша статистика будет соответствовать некоторой другой функции параметров, то она даже при бесконечно большом объеме выборки не даст правильного значения интересующей нас функции параметров. Действительно, в этом случае она будет стремиться к некоторому пределу, но этот последний отличается от той цели, которую мы имеем в виду. Такие статистики называются *несостоятельными статистиками*. Исключая те случаи, когда соответствующая ошибка крайне незначительна, как, например, при введении поправок Шеппарда, несостоятельные статистики не должны применяться для оценки неизвестных параметров.

С другой стороны, все *состоятельные* статистики по мере возрастания выборки все более и более приближаются к определенным значениям интересующих нас функций параметров; во всяком случае, если они стремятся к некоторому фиксированному значению, то оно не может быть неправильным. В тех простейших случаях, с которыми мы здесь будем иметь дело, состоятельные статистики не только стремятся к истинному значению функции параметров, но вместе с этим они при выборках определенного размера будут иметь ошибки, распределение которых, в свою очередь, будет приближаться к хорошо известному закону распределения, называемому нормальным законом распределения ошибок, или, более просто, *нормальным распределением* (более подробно см. в главе III). В этих условиях система ошибок может быть охарактеризована при помощи среднего квадрата из всех этих

ошибок, который обычно называется *дисперсией*. В тех случаях, с которыми мы здесь будем встречаться, при увеличении объема выборок дисперсия убывает обратно пропорционально размеру выборки.

Иногда при оценке некоторого параметра, как например центра распределения, можно построить несколько различных статистик, например: среднюю, медиану и др. Может случиться так, что все эти статистики будут в одинаковой мере состоятельными в указанном выше смысле и им будут при больших выборках соответствовать дисперсии, изменяющиеся обратно пропорционально размеру выборки. Однако при больших выборках некоторого фиксированного размера дисперсии таких статистик будут, вообще говоря, различны. Поэтому особый практический интерес для нас будет представлять более узкая группа статистик, у которых распределение ошибок по мере увеличения объема выборки будет стремиться к нормальному распределению, имеющему наименьшую из возможных дисперсий. Это позволяет нам выделить из общей массы состоятельных статистик класс статистик, обладающих указанным выше свойством; эти статистики называются *эффективными*.

Введение этого термина можно обосновать таким примером. Если для большой выборки, скажем в 1000 наблюдений, вычислена эффективная статистика А и вторая состоятельная статистика В, имеющая дисперсию в два раза большую, чем у А, то, хотя В и является статистикой, вполне пригодной для оценки искомого параметра, все же она в отношении своей точности определенно занимает более низкое место, чем статистика А. При использовании статистики В необходимо иметь выборку в 2000 наблюдений, чтобы получить столь же хорошую оценку параметра, какую можно получить с помощью статистики А при выборке только в 1000 наблюдений. Основываясь на этом, можно сказать, что статистика В использует только 50% информации, содержащейся в данных наблюдениях, или, более кратко, что ее *эффективность* составляет 50%. Это понятие эффективности допускает существование статистик, эффективность которых составляет не более 100%.

В некоторых случаях, однако, на вполне законном основании могут быть использованы и статистики, эффективность которых ниже 100%. Например, мыслимы такие случаи, когда увеличение числа наблюдений требует меньших затрат труда, чем применение трудоемкого расчета статистики с 100%-ой эффективностью. Может быть и так, что неэффективная статистика вполне достаточна для ответа на частный вопрос исследования. Но все же имеется одно ограничение для широкого применения неэффективных статистик, на которое следует обратить внимание. В тех случаях, когда приходится иметь дело с критериями согласия, следует видеть различие между *ошибками случайного отбора и ошибками, связанными с подбором закона распределения*. Статистики,

применяемые для указанного выше подбора, должны быть не только состоятельными, но и эффективными на все 100%. Это обстоятельство служит серьезным ограничением для применения неэффективных статистик и, кроме того, является источником известных затруднений, так как иногда при исследовании некоторой массы данных возникает необходимость одновременной проверки нескольких гипотез, для чего необходимы в каждом отдельном случае специальные эффективные статистики.

В последующих главах будут даны примеры числовых расчетов для определения различных, преимущественно эффективных, статистик. При построении эффективных статистик во вновь возникающих задачах приходится проводить специальные математические исследования. Изучение этого вопроса привело автора к выводу, что эффективная статистика всегда может быть найдена при помощи *метода максимального правдоподобия*. Учитывая наличие некоторых математических трудностей при решении подобного рода задач, следует обратить внимание на то благоприятное обстоятельство, что в большинстве случаев эффективные статистики могут появиться так же в результате и приближенного решения задачи нахождения максимального правдоподобия. В последней главе вместе с различными способами решения генетических вопросов приводятся и некоторые простые примеры применения к этим вопросам метода максимального правдоподобия.

С практической точки зрения, вообще говоря, нет необходимости развития статистических методов далее того пункта, когда выделены эффективные статистики. Как можно показать, все эффективные статистики при увеличении объема выборки стремятся к одному и тому же значению, т. е. к равенству друг с другом. Поэтому для практики нет особых неудобств в том, что они все же разные. Однако имеется одна категория статистик (к которой относятся и некоторые из наиболее употребительных статистических показателей), обладающая тем замечательным свойством, что такая статистика даже при малой выборке содержит в себе всю основную информацию, заключенную в данных наблюдениях. Такие статистики называются *достаточными*. Эти статистики имеют при малой выборке определенное преимущество перед другими эффективными статистиками. Примером достаточной статистики может служить средняя арифметическая, вычисленная по выборке из нормальной совокупности или из распределения Пуассона.

Тот факт, что для этих двух важнейших типов распределения установлены указанные выше достаточные статистики, придает арифметической средней большое теоретическое значение. Метод максимального правдоподобия приводит именно к достаточным статистикам, если они, конечно, существуют. В тех же случаях, когда достаточные статистики сами по себе не существуют, возможно использование некоторых специальных, но не общих функциональных соотношений, позволяющих производить исчерпывающую оценку при помощи так называемых *служебных статистик*.

Широкое практическое использование эффективных статистик в особенности при больших выборках само по себе не вызывает никакого сомнения и неопределенности, однако во всех случаях следует ясно видеть разницу между параметром распределения, подлежащим оценке, и фактическим значением статистики, используемой в качестве оценки этого параметра; в связи с этим следует довести до сведения читателя, что в дальнейшем он встретится с большим разнообразием приемов, предназначенных для оценки параметров.

4. Содержание настоящей книги

Основная цель, которую ставил перед собой автор настоящей книги, это дать в руки исследователям и главным образом биологам средства для правильного применения статистических критериев к тому цифровому материалу, который собран в их лабораториях или взят из литературных источников. Эти критерии установлены в процессе решения ряда задач, относящихся к распределениям, большинство из которых ранее излагалось в специальных математических работах, но теперь получило более широкое распространение. Математическая сложность этих проблем заставляет ограничиться здесь только: 1) указанием на тип задачи, возникающей при решении того или иного вопроса; 2) приведением цифровых иллюстраций, при помощи которых можно освоить процесс соответствующего статистического исследования; 3) приведением таблиц, позволяющих использовать критерии без алгебраических преобразований и расчетов.

Может показаться, что нет никакой возможности дать все методы, пригодные для весьма большого разнообразия практических задач, но, к счастью, одно благоприятное обстоятельство разрешает эту проблему; дело в том, что в ряде случаев одно и то же математическое решение получается повторно для вопросов, которые на первый взгляд кажутся совсем различными. Например, решение Хельмерта в 1875 г. задачи о распределении суммы квадратов отклонений от средней оказалось фактически однозначным решению задачи о распределении χ^2, найденному К. Пирсоном в 1900 г. Оно было снова и независимо от других найдено Стьюдентом в 1908 г. как распределение дисперсии выборки из нормальной совокупности. То же самое распределение было найдено и автором этой книги для показателя рассеяния малой выборки, взятой из распределения Пуассона. Замечательным является то обстоятельство, что хотя Пирсон в своей работе в 1900 г. допустил серьезную ошибку, которая привела к тому, что до 1921 г. критерий согласия применялся неправильно, все же исправление этой ошибки после того, как были разработаны более эффективные методы оценки, привело не к изменению самого

этого распределения, а только к некоторому уменьшению числа единиц одного из параметров, входящих в формулу этого распределения.

Точно так же благоприятным оказалось и то обстоятельство, что распределение t, впервые установленное в 1908 г. Стьюдентом при изучении вопроса о вероятной ошибке средней, оказалось применимым не только в этом случае, но и в более сложном и даже более часто встречающемся случае сравнения двух средних. Дальнейшее развитие этого метода привело к решению задачи об ошибках выборки для весьма широкого класса статистик, известных под названием коэффициентов регрессии.

При исследовании ряда теоретических распределений в процессе решения некоторых других задач, таких, как изложенный в этой книге вопрос о внутриклассовой корреляции, вопросы об определении линии регрессии, корреляционном отношении и множественной корреляции, автор пришел к одному и тому же третьему распределению, которое названо распределением z и которое, как это можно было ожидать и как это оказалось на самом деле, связано с распределениями, выведенными Пирсоном и Стьюдентом. Таким образом, все распределения, необходимые при исследовании весьма широкого круга вопросов, могут быть подразделены на эти три категории, и, что также очень важно, оказалось возможным иметь для практической работы только небольшое число таблиц. Таблицы же, необходимые для более широкого круга задач, с соответствующими примерами читатель найдет в специальных изданиях.

Изложение в настоящей книге ведется так, что изучающий ее может познакомиться с этими тремя главными распределениями в порядке их логической связи и путем перехода от более простых к более сложным случаям. Методы, описанные в последних главах, в большинстве случаев являются обобщениями более элементарных способов, изложенных в начале книги. Если читатель изучит эту работу методически, в последовательном изложении темы, то можно надеяться, что он не упустит из внимания общее единство методов, лежащих в основе обработки данных, взятых из самых различных областей. После этого читатель, вероятно, почувствует себя подготовленным для того, чтобы со знанием дела разобраться во многих аналогичных проблемах, которые здесь не представилось возможным затронуть. Вместе с этим можно предвидеть, что некоторые исследователи пожелают использовать эту книгу в качестве пособия при обработке своих лабораторных данных, а не как последовательный курс. Такое использование книги вполне допустимо, но только при условии, что читатель постарается проработать полностью и со всеми числовыми расчетами ряд подходящих примеров для того, чтобы быть уверенным не только в том, что его данные соответствуют избранной системе обработки, но и в том, что он при этом знакомстве с нашими примерами получит вполне правильное представление о содержании, заложенном как в схеме разработки, так и в ее результатах.

Здесь, пожалуй, следует возразить против того взгляда, что в элементарном изложении, где нельзя дать математических доказательств и которое предназначено для читателя, не имеющего специального математического образования, нельзя дать многое из того, что относится к последним достижениям науки и что раньше не включалось в популярную статистическую литературу. По этому поводу автор должен высказать следующие соображения.

1. Читателям-нематематикам во многих случаях достаточно уметь правильно использовать специально составленные статистические таблицы; использование точных таблиц не вызывает никаких добавочных затруднений (хотя их построение может быть довольно трудоемким), по сравнению с неточными таблицами, приведенными в более ранних популярных работах.

2. Процесс исчисления вероятности или вероятной ошибки по одной из установленных формул отнюдь не дает определения самого распределения случайных выборок, а только позволяет нам иметь в руках критерий существенности, который мы можем использовать при предположении, что фактическое распределение довольно близко к нормальному и что поэтому допустимо основываться на таблице отклонений нормальной кривой. Вопрос о том, может ли эта схема исследования использоваться в каждом конкретном случае, решается не при помощи математических навыков у исследователя, а на основе ясного представления о способности этой схемы дать правильный ответ на исследуемый вопрос. То положение, что такая схема, основанная на новейших достижениях математиков, применима к разработке и анализу весьма обширного материала, отнюдь не позволяет считать, что она достаточно точна и при лабораторном исследовании, где приходится искать правильное решение вопроса на основе небольшой группы предварительных данных.

3. Распределения, которые рассматриваются в настоящей книге, фактически установлены в связи с исследованиями в области биологии и агрономии. Это положение относится не только к исследованиям самого автора, но имеет своим началом критическое рассмотрение вопроса о статистических распределениях в работе Стьюдента (1908 г.).

Большая часть этой книги отведена числовым примерам, причем количество этих примеров возрастает там, где приходится иллюстрировать новую точку зрения. При выборе этих примеров автору пришлось считаться с тем, что бесполезна всякая попытка отразить в примерах все разнообразие объектов исследования, допускающих применение того или иного метода. Здесь нет примеров из астрономической статистики, которой за последнее время посвящен ряд интересных работ, дано только небольшое

число примеров из социальной статистики, и даже примеры из биологии, которым отведено основное место, приведены без всякой системы. Эти примеры выбирались прежде всего для того, чтобы иллюстрировать тот или иной частный прием, и весьма редко они касались существа самого вопроса. Для полного освоения описываемых приемов, читатель должен научиться так ставить вопросы, относящиеся к его материалу, чтобы эти приемы могли бы дать вполне определенный ответ; вместе с тем, и это является не менее важным, он должен и будущие наблюдения планировать с учетом этой постановки вопросов. Учитывая эту цель наших примеров, читатель должен иметь в виду, что они не могут служить материалом для обсуждения общих научных вопросов, которые требуют исследования гораздо большего объема данных и привлечения ряда других доказательств. Эти примеры имеют только одну задачу: дать критический анализ представленных в них частных цифровых данных.

5. Некоторые замечания исторического характера

Чтобы читатель смог увидеть современное состояние нашей науки в историческом освещении, мы дадим в этом параграфе краткие сведения об основоположниках статистической теории. Это необходимо в связи с тем, что, с одной стороны, проявлен известный интерес к историческому развитию теории, положенной в основу излагаемых в этой книге методов, и, с другой стороны, тем, что время от времени в печати встречаются неправильные утверждения, приписывающие оригинальность некоторым из излагаемых автором методов, хотя они были хорошо известны предшествующим исследователям, и, наоборот, приписывающие его предшественникам современные достижения теории, о которых они не имели никакого представления.

Томас Байес, прославившийся опубликованной в 1763 г. статьей, известен в связи с первой попыткой использовать теорию вероятностей как средство для построения индуктивных выводов, т. е. для перехода от частного к общему или, конкретно, от выборки к генеральной совокупности. Эта работа была опубликована только после его смерти, и мы не знаем, каковы были взгляды Байеса в этом вопросе при жизни. Наоборот, нам известно, что причиной его нерешительности в отношении опубликования этой работы, была его недостаточная уверенность в правильности постулата, положенного в основу известной «теоремы Байеса». Хотя мы в настоящее время должны отвергнуть его постулат, однако вместе с тем следует признать и значение Байеса, заключающееся в том, что он осознал самую проблему, подлежащую решению, что он сделал попытку найти это решение и что, наконец, он более ясно, чем многие последующие авторы, понимал недостатки этой попытки.

В то время как Байес обладал превосходной логической проницательностью, Лаплас (1820 г.) был несравненным мастером аналитического метода. Он ввел в свои исследования принцип обратной вероятности без какой-либо его критики. С другой стороны, мы ему обязаны введением принципа, согласно которому всякое распределение случайной величины может быть разложено на составляющие его независимые элементы, которые характеризуются целым рядом показателей, таких, как средняя, дисперсия и другие кумулянты (см. стр. 65). Это положение было впоследствии и независимо от него установлено Тилем (1889 г.), но математические методы были более эффективными у Лапласа, чем у Тиля, и оказали большое влияние на развитие этой теории во Франции и Англии. Непосредственным следствием изучения Лапласом распределения значений переменной величины, относящихся к множеству независимых друг от друга случаев, было открытие нормального закона распределения ошибок, который обычно приписывается, и не без некоторого основания, его великому современнику Гауссу.

Гаусс, так сказать, эмпирически приблизился к проблеме статистических оценок, распространив вопрос об оценках не только на вероятности, но и на другие количественные параметры. Он уже осознавал возможность применения для этих целей метода максимального правдоподобия, но стремился обосновать этот метод на принципе обратной вероятности. Далее Гаусс дал законченную разработку вопроса об отыскании уравнения простой и множественной регрессии при помощи метода наименьших квадратов, который в тех случаях, когда он применим, является частной формой метода максимального правдоподобия.

Первое из распределений, определяющих современные критерии существенности, хотя и известное со времен Хальмерта, было снова открыто К. Пирсоном в 1900 г. при решении задачи об измерении отклонений наблюденного распределения от гипотетического. Это распределение χ^2. Я думаю, что это был самый большой вклад в сокровищницу статистических методов, свидетельствующий о непревзойденной силе научной мысли проф. Пирсона. Критерий χ^2 дает точную и объективную меру совокупности отклонений некоторого числа нормально распределенных и взаимосвязанных переменных от их ожидаемых значений. В его первоначальном виде в применении к численностям, которые являются дискретными переменными, это распределение по необходимости было только грубо приближенным, но когда были исключены малые численности, это приближение стало вполне удовлетворительным. Распределение χ^2 становится точным в ряду других задач, решенных впоследствии. Когда χ^2 находится для численностей, которые были использованы при расчете соответствующих кривых распределения, то в этом случае часто преувеличивается степень согласия между наблюденными и ожидаемыми частотами в связи с тем, что здесь включаются в рассмотрение пустые или почти пустые классы, которые не увеличивают

или почти не увеличивают наблюденное значение χ^2, но повышают его ожидаемое значение. В результате этого наши суждения относительно параметров распределения, исчисленных по выборочным данным, становятся не вполне правильными. Необходимость введения соответствующих поправок длительное время игнорировалась, а после этого долго обсуждалась, но теперь, как я надеюсь, она признана всеми. Главным источником ошибок, снижающих ценность критерия согласия, является применение неэффективных методов подбора кривых распределений (см. главу IX). Это ограничение едва ли можно было предвидеть в 1900 г., когда не были известны даже начальные положения теории оценок.

Изучение точных выборочных распределений статистик начинается в 1908 г. с работы Стьюдента «Вероятная ошибка средней» (The Probable Error of a Mean). Как только была установлена истинная природа проблемы, сразу же удалось найти математическое решение множества других задач, относящихся к выборочным распределениям. Сам Стьюдент в этой и в последующих своих работах дал правильные решения для трех следующих задач: распределение дисперсии, распределение средней, деленной на ее среднюю квадратическую ошибку, и распределение коэффициента корреляции между независимыми переменными. Этого было достаточно, чтобы установить особое значение для теории выборок распределений χ^2 и t, хотя понадобилась еще дальнейшая работа для того, чтобы доказать, что и многие другие задачи нахождения критериев существенности приводят к этим двум формам, а также к еще более общему критерию z. Работа Стьюдента была не сразу замечена (фактически ее полностью игнорировали в журнале, где она была помещена), и одной из главных задач настоящей книги, начиная с ее первого издания, было ознакомление широких статистических кругов с результатами его исследований, а также с математическими работами, последовавшими за ним. Эта популяризация велась мной, с одной стороны, по линии критики традиционной доктрины в теории ошибок и, с другой стороны, по линии упрощения математических расчетов, необходимых при разработке данных.

ГЛАВА ВТОРАЯ

ДИАГРАММЫ

7. Диаграммы облегчают предварительное изучение первоначальных данных. Они сами по себе ничего не доказывают, но дают возможность в наглядной форме познакомиться с особенностями материала; поэтому они не могут заменить собой тех количественных критериев, которые необходимы при анализе данных, но они ценны в том отношении, что способны подсказать, какие критерии в данном случае применимы, и оказать помощь при истолковании выводов, полученных на основе этих критериев.

8. Графики временных рядов, прирост и коэффициент роста

Графики временных рядов чаще всего отражают изменение во времени изучаемой переменной величины, например, веса животного, или образцов растений различного возраста, или численности населения через следующие один за другим интервалы времени. Следует делать различие между теми случаями, когда, как это бывает в опытах с кормлением животных, один и тот же объект (животное) наблюдается повторно через определенные промежутки времени, и такими случаями, обычно встречающимися в опытах по физиологии растений, когда один и тот же объект (растение) не может быть использован для наблюдения дважды и когда приходится для каждого момента времени брать самостоятельные образцы. То же самое различие встречается при подсчете микроорганизмов: в одних случаях подсчет производится каждый раз внутри одной и той же пробы, в других же случаях — в различных параллельных пробах культуры. Если перед нами стоит задача получить общую форму кривой роста, то второй случай имеет то преимущество, что отклонения от ожидаемой теоретической кривой, относящиеся к последовательным промежуткам времени, будут определяться независимыми наблюдениями; использование же все время одного и того же материала не обеспечивает такой независимости наблюдений. С другой стороны, когда основной интерес сосредоточивается на приросте, имеет преимущество уже использование одного и того же

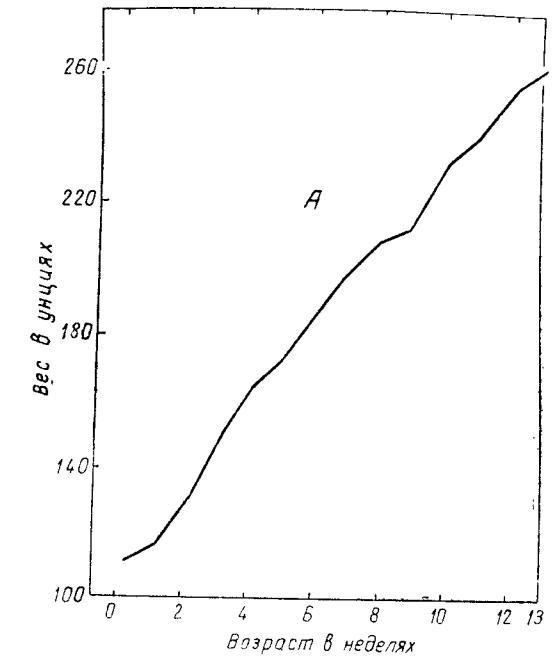
материала, так как только этим путем можно установить фактическое увеличение веса или численности. Оба эти аспекта в известной мере совмещаются друг с другом при повторении наблюдений в каждый момент времени. Так, при измерении нескольких животных для каждого из вариантов кормления становится возможным на основе индивидуального взвешивания, но отнюдь не на основе среднего веса, определить, будет ли кривая роста находиться в согласии с определенной теоретической кривой роста или же они будут довольно значительно отличаться друг от друга. Подобно этому, если некоторое число растений внутри каждого образца будут взвешены по отдельности, то в результате этого будут определены приросты, сопровождающиеся некоторой вероятной ошибкой, что позволяет произвести оценку интересующих нас сравнений.

На рис. 1 дан вес ребенка в унциях через каждую неделю после рождения, а в табл. 1 даны расчеты абсолютного и относительного прироста за день. Абсолютный прирост, представляющий собой средний фактический прирост, полученный из прибавки в весе за неделю, определяется путем вычитания из веса на определенную дату веса на предшествующую дату и делением этой разности на продолжительность периода в днях. Относительные приросты показывают увеличение веса не только в пересчете на единицу времени, но также и на единицу достигнутого веса. Используя тот математический факт, что

$$\frac{1}{m}\frac{dm}{dt} = (\log_e m),$$

можно установить, что средние значения относительного прироста для некоторого периода времени получаются из натуральных логарифмов соответствующих весов, в то время как абсолютный прирост — из непосредственных весов. Такие относительные приросты удобно выражать в процентах, умножая на 100. Если же такой относительный прирост рассчитать путем деления абсолютного прироста на вес начального момента данного периода, то получаются несколько более высокие величины, в связи с тем, что фактический вес ребенка в продолжении некоторого периода времени обычно несколько выше, чем вес на начало этого периода. Эта погрешность становится большей по мере увеличения относительного роста между двумя последовательными взвешиваниями.

На рис. 1А дано изменение абсолютного веса во времени; средний наклон линии на диаграммах этого вида характеризует средний темп роста. На данной диаграмме точки, соответствующие весу в различные моменты времени, располагаются примерно по прямой линии; это указывает на то, что абсолютный прирост является здесь почти постоянным и составляет около 1,66 унции в день. На рис. 1Б изображена динамика натурального логарифма веса; наклон линии в каждой данной точке характеризует



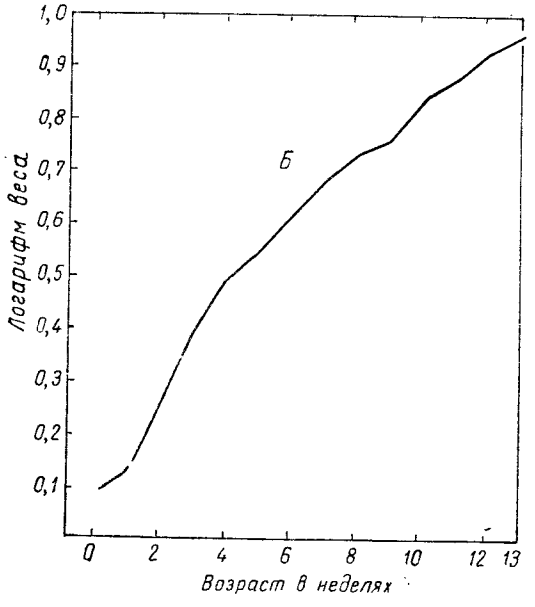


Рис. 1.

Таблица 1

Возраст в неделях $\frac{t}{7}$	Вес в унциях m	Прирост δm	Прирост в день в унциях $\frac{\delta m}{\delta t}$	Натуральный логарифм веса $\log \frac{m}{100}$	Прирост $\delta \log m$	Прирост в процентах в день $\frac{\delta}{\delta t} \log m$
0	110			0,0953		
		4	0,57		0,0357	0,51
1	114			0,1310		
		14	2,00		0,1159	1,66
2	128			0,2469		
		19	2,71		0,1384	1,98
3	147			0,3853		
		16	2,29		0,1033	1,47
4	163			0,4886		
		9	1,29		0,0537	0,77
5	172			0,5423		
		14	2,00		0,0783	1,12
6	186			0,6206		
		12	1,71		0,0625	0,89
7	198			0,6831		
		10	1,43		0,0493	0,70
8	208			0,7324		
		5	0,71		0,0237	0,34
9	213			0,7561		
		19	2,71		0,0855	1,22
10	232			0,8416		
		8	1,14		0,0339	0,48
11	240			0,8755		
		14	2,00		0,0567	0,81
12	254			0,9322		
		7	1,00		0,0272	0,39
13	261			0,9594		

относительный прирост, который, если исключить первую неделю, неуклонно падает по мере увеличения возраста ребенка. Такие кривые с внешней стороны будут выглядеть лучше, если масштабы двух осей выбрать так, чтобы график располагался примерно под одинаковыми углами к той и другой оси; при размещении графика вблизи горизонтальной или вертикальной оси изменения наклона его воспринимаются не столь хорошо.

График относительного прироста можно построить с большей быстротой и удобством, если воспользоваться бумагой со специальным графлением, на которой по горизонтали нанесена логарифмическая шкала с указанием на ней самих чисел (рис. 5). Этот способ позволяет избежать логарифмирования, но таблицы логарифмов все же остаются необходимыми для работы, если, кроме графика, нужно знать и отдельные значения относительного прироста. Чтобы получить первое предварительное представление о согласии наблюдений с некоторым законом роста,

желательно подыскать такое измерение изучаемой величины, чтобы этот закон выражался прямой линией. Таким образом, рис. 1А пригоден для проверки гипотезы о константности абсолютного прироста; допущение же постоянства относительного прироста явным образом опровергается диаграммой на рис. 1Б. При наличии других гипотетических кривых роста следует использовать иные способы преобразования переменной. Например, в случае так называемой «автокаталитической», или «логистической», кривой относительный прирост убывает пропорционально фактическому весу на данный момент времени. Поэтому, если относительные приросты расположить в зависимости от значений фактического веса, то при наличии этого «автокаталитического» закона все точки должны расположиться на прямой линии. В этом случае против каждого наблюденного веса следует отметить точкой среднюю из двух соседних относительных приростов. Применение этого способа построения графика по данным табл. 1 предоставим читателю; здесь можно нанести двенадцать точек, соответствующих весу от 114 до 254 унций. Если построить этот график, то можно убедиться, что даже после усереднения соседних приростов попарно не получается никакого регулярного изменения и, следовательно, нет ясных указаний на то, что наблюдения соответствуют этому закону. В случае, если бы для наших данных была бы найдена прямая линия и если допустить, что данный закон роста остается неизменным в любом возрасте, то пересечение этой прямой с горизонтальной осью показывало бы такой вес, при котором прекращается всякий его прирост.

9. Диаграммы корреляционной связи

Широкое распространение имеют диаграммы, на которых некоторая неконтролируемая, т. е. подверженная ошибкам наблюдений, величина располагается в соответствии с периодами времени или какой-либо другой контролируемой величиной (например, концентрация солей в растворе, температура и пр.). Но еще более широкое применение могут иметь диаграммы корреляционной связи, в которых один неконтролируемый фактор находится в связи с другим, тоже неконтролируемым фактором. Эти графики строятся как точечные диаграммы, в которых каждая точка представляет результат отдельного эксперимента или наблюдения над двумя переменными величинами. По такой диаграмме легко установить, имеется ли более или менее заметная связь между этими величинами. Когда такая точечная диаграмма строится на основе относительно небольшого числа наблюдений, то она довольно часто может подсказать нам, следует ли затрачивать время на накопление материала этого рода. Ценность и значение нашего опыта или наблюдений становятся в этом случае видимыми для глаза и этим путем можно обнаружить такие связи, которые вполне заслуживают того, чтобы наблюдения были бы продолжены.

В тех случаях, когда наблюдений так много, что трудно построить точечную диаграмму, следует все поле диаграммы разделить на квадраты и вписать в каждый из них соответствующие частоты. Такая диаграмма с обозначением частот является уже корреляционной таблицей.

На рис. 2 дана точечная диаграмма, характеризующая зависимость урожаев пшеницы, полученных на опытных делянках (делянки, удобренные навозом, с участка Бродболк Ротамстедской опытной станции), от количества осадков, наблюденных в различные годы. Эти делянки одинаково обрабатывались в течение всего периода с 1854 по 1888 г. Тридцать пять наблюдений над урожайностью и осадками, обозначенные 35-ю точками, указывают на хорошее соответствие между ростом урожайности и уменьшением осадков. Даже в случаях, когда имеется небольшое число наблюдений, точечная диаграмма может натолкнуть мысль на такие связи, которые ранее и не подозревались, или, наоборот, и это не менее важно, покажет отсутствие такой связи, в наличии которой мы прежде были вполне уверены. Общее значение этих диаграмм состоит в том, что они дают наметку наших выводов до того, как эти последние будут строго сформулированы, и направляют нас к предположениям, которые в дальнейшем могут быть проверены более точным статистическим или экспериментальным исследованием.

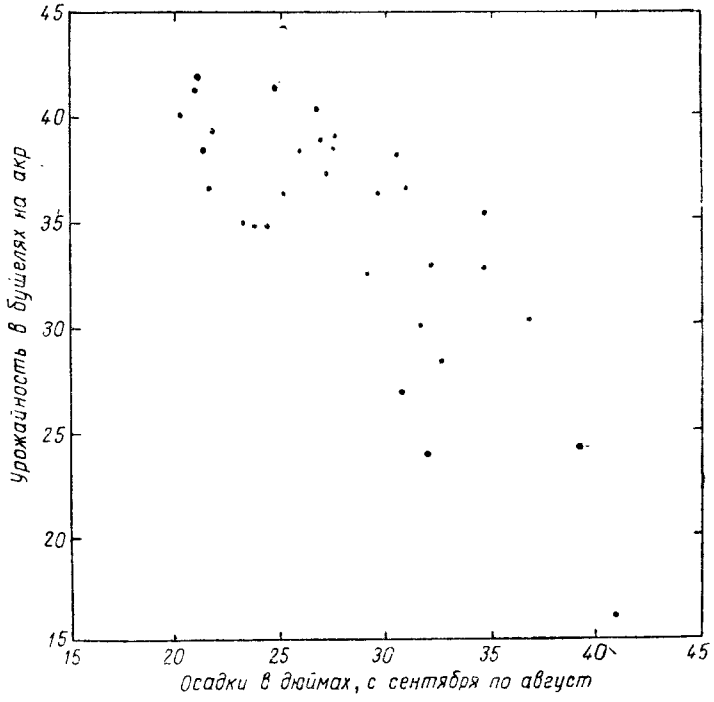


Рис. 2. Урожай пшеницы и осадки за 35 лет — с 1854 по 1888 г.

Кроме построения точечной диаграммы, иногда применяется другой способ изображения связи, при котором значения одной переменной располагаются в порядке их возрастания, а соответствующие значения второй переменной отмечаются точками. Если линия, полученная этим путем, имеет более или менее заметный наклон или общую тенденцию к этому, то можно говорить о наличии связи между этими переменными. На рис. 3 дана такая

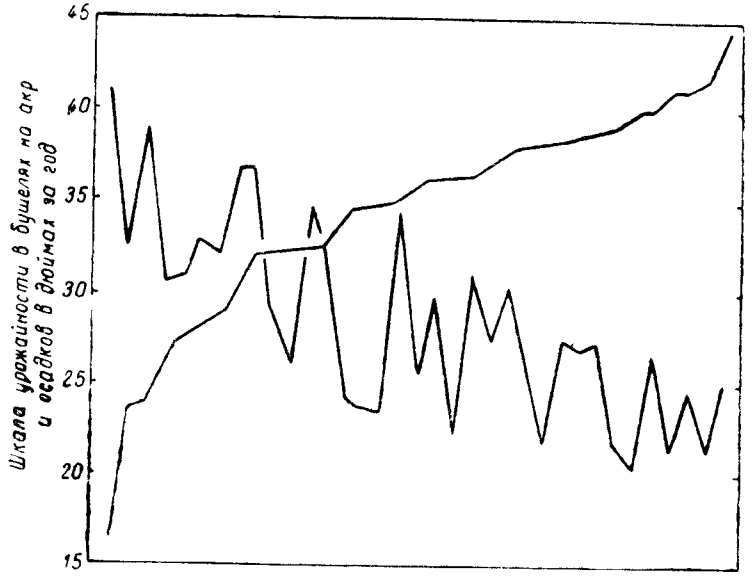


Рис. 3. Осадки и урожаи за 35 лет, расположенные в порядке возрастания урожая.

линия для осадков, когда годы расположены в порядке возрастания урожая пшеницы. Такие корреляционные диаграммы обычно менее показательны, чем точечные, и они иногда могут скрыть сведения, которые выявляются при помощи последних. Вместе с этим точечная диаграмма обладает еще тем преимуществом, что при небольшом числе наблюдений она полностью соответствует корреляционной таблице, а при большом числе наблюдений легко преобразуется в эту таблицу.

При построении корреляционной таблицы значения обеих переменных группируются по классам, причем интервалы группировки внутри каждой переменной должны быть одинаковыми. Например, урожаи пшеницы мы можем подразделить на группы с интервалом в один бушель на акр, а осадки — на группы с интервалом в один дюйм. Поэтому соответствующая корреляционная диаграмма будет делиться на квадраты, в каждый из которых вписывается количество наблюдений. Корреляционная

таблица полезна в трех отношениях. Она дает наглядное представление обо всей массе наблюдений и при небольшом числе наблюдений столь же обозрима, как и точечная диаграмма. Она является компактной регистрацией обширного ряда наблюдений, причем при двух переменных эта регистрация становится исчерпывающей. При большем числе переменных одна корреляционная таблица уже не исчерпывает всю связь; в этом случае строятся корреляционные таблицы для каждой пары переменных. Данное обстоятельство лишает исследователя возможности получить целостное изображение исходного материала, но следует отметить, что, к счастью, в подавляющем большинстве статистических задач ряд таких попарных распределений дает довольно полное освещение изучаемого вопроса. При числе переменных больше двух исходные данные в целях различного рода справок удобно наносить на карточки, имея для каждого отдельного случая самостоятельную карточку, на которой значение каждой переменной заносится в отведенное для нее место. Правда, публикация таких полных данных встречает известные затруднения, но эти затруднения все же не являются непреодолимыми, так как наиболее существенные сведения о материале могут быть даны в компактной форме корреляционных таблиц.

Третья важная особенность корреляционной таблицы состоит в том, что в ней исходные данные представлены в форме, удобной для непосредственного применения статистических методов обработки материала. На основе корреляционной таблицы могут быть легко вычислены все важнейшие статистики, а именно: средние, дисперсии и ковариация. Пример корреляционной таблицы дан на стр. 148, 149 (табл. 31).

10. Диаграммы частот

Если у большого числа объектов наблюдается некоторый признак, каким, например, может быть размер, вес, цвет, плотность и т. д., то появляется возможность составить примерное представление о той *генеральной совокупности*, по отношению к которой наши данные следует рассматривать как выборочные. При этом мы получаем возможность сравнивать ее с другими генеральными совокупностями, отличающимися по своему происхождению или в связи с изменением окружающей обстановки. Так, например, расы, рассматриваемые как совокупности людей, могут значительно различаться друг от друга, хотя отдельные лица, относящиеся к этим расам, могут оказаться во всех отношениях одинаковыми. То же самое и в условиях эксперимента: наблюдение над совокупностью объектов может обнаружить влияние окружающих условий на размер, смертность и т. д., но это влияние может остаться незамеченным при наблюдениях над отдельными объектами. Наглядное представление о результатах измерения некоторого признака у большого числа объектов дает

диаграмма частот. При построении этой диаграммы измеряемый признак располагается на оси абсцисс, т. е. его значения наносятся на горизонтальной оси графика, а по ординате, т. е. по вертикали, откладываются *численности*, соответствующие каждому интервалу.

Рис. 4 представляет собой диаграмму частот, дающую распределение 1375 женщин по их росту (несколько измененные данные Пирсона и Ли). Вся эта выборка была разделена на части, соответствующие последовательным интервалам роста в 1 дюйм. Равные площади на диаграмме соответствуют одинаковым чис-

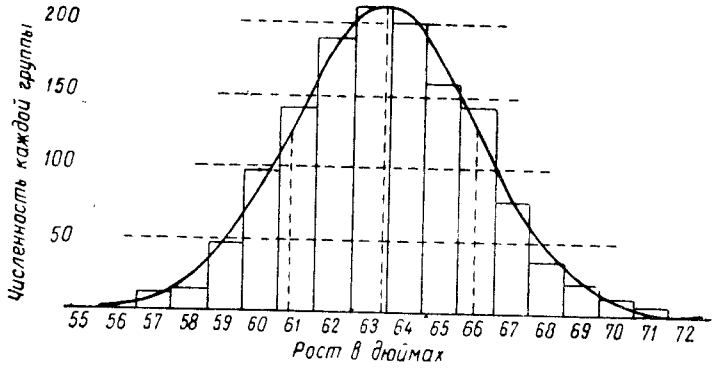


Рис. 4.

ленностям; в случае группировки материала по неравным интервалам изучаемого признака диаграмму следует строить так, чтобы площади были пропорциональны численностям.

Интервал, содержащий наибольшее число наблюдений, называется модальным интервалом. На рис. 4 модальный интервал составляет от 62,5 до 63,5 дюйма. В тех весьма частых случаях, когда переменная изменяется непрерывно, так что возможны все промежуточные значения, выбор размера интервалов и их границ произволен, а это может привести к заметным различиям во внешнем виде диаграммы. Однако возможные границы групп обычно определяются наименьшей из единиц, в которых выражены результаты измерения. Если, например, измерения роста были сделаны с точностью до одной четверти дюйма и все значения между $66\frac{7}{8}$ и $67\frac{1}{8}$ дюйма оказались приравненными к 67 дюймам, все значения между $67\frac{1}{8}$ и $67\frac{3}{8}$ приравнены к $67\frac{1}{4}$ дюйма и т. д., то мы не имеем другого выбора, как взять в качестве единицы группировки 1, 2, 3, 4 и т. д. четверти дюйма при границах интервалов, кратных $\frac{1}{8}$ дюйма. Для статистических расчетов выгодней брать меньшие единицы группировки, так как это дает большую точность, но для графического изображения преимущество имеет более грубая группировка.

На рис. 4 взята единица группировки в 1 дюйм, удобная при большом числе наблюдений; при меньших выборках обычно приходится брать более грубую группировку, чтобы включить в каждую группу достаточное число наблюдений.

Во всех случаях, когда переменная величина изменяется непрерывно, диаграмма частот должна иметь форму гистограммы, в которой площадь прямоугольника, построенного на любом интервале, должна соответствовать численности наблюдений в этом интервале. На практике иногда применяется и другой вид диаграмм, получаемый путем соединения ординат, которые восстанавливаются из центра каждого интервала. В результате получается диаграмма, имеющая известное сходство с непрерывной кривой. Однако преимущество этой формы иллюзорно не только потому, что ее вид иногда может ввести в заблуждение, но и потому, что всегда следует делать различие между бесконечной генеральной совокупностью и фактической выборкой, взятой из этой совокупности. Концепция непрерывной кривой численностей, строго говоря, приложима только к первой, т. е. к генеральной совокупности, и нет никакой необходимости только ради иллюстрации игнорировать указанное выше различие.

Все сказанное не должно препятствовать наложению кривой распределения, рассчитанной по наблюдениям, на гистограмму (как это сделано на рис. 4). На таких совмещенных диаграммах выявляется различие между гистограммой, представляющей выборку, и непрерывной кривой, дающей примерное изображение формы генеральной совокупности. В этом случае мы можем непосредственно обнаружить серьезное различие между фактическим и гипотетическим распределениями. Однако не следует преувеличивать возможностей для решения на основе такой глазомерной оценки вопроса о том, будет ли это различие более того, которое можно ожидать при условии случайного отбора. Для решения этого вопроса существуют точные статистические критерии, описание которых будет дано в следующих главах.

Если переменная изменяется прерывно, например, принимает только целые значения, то, собственно говоря, нет оснований для изображения соответствующего распределения в виде гистограммы, так как здесь нет изменения величины внутри интервала. С другой стороны, в этом случае нет оснований и для построения непрерывной кривой распределения. Однако представление таких данных в виде гистограммы не является чем-то незаконным; оно вполне допустимо, если мы будем рассматривать дискретную изменчивость, как изменчивость непрерывной величины, значения которой округляются до целых чисел.

10.1. Преобразованные частоты

В ряде случаев при проверке согласия наблюдений с некоторым специальным законом распределения численность, как и всякая другая переменная величина, может быть преобразована в

логарифмическую форму; это можно достигнуть путем размещения на графике логарифмов численностей или самих численностей на логарифмической бумаге. На рис. 5 этим способом нанесены на логарифмическую бумагу числа цветков лютика, имеющих от 5 до 10 лепестков (данные Пирсона); этим графиком облегчается сравнение наблюденных численностей с гипотетическими, осно-

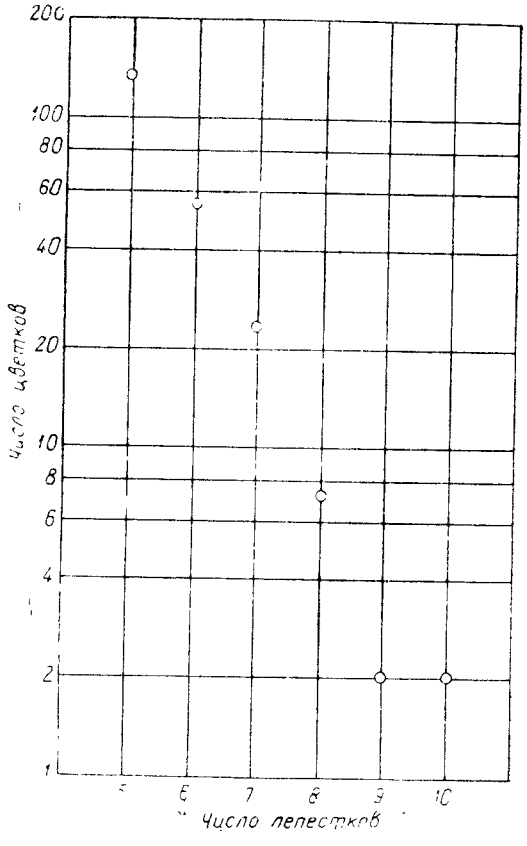


Рис. 5.

ванными на предположении, что численности растений с числом лепестков, большим пяти, уменьшаются в геометрической прогрессии. Такой график, собственно говоря, не является диаграммой частот, хотя в нем используются частоты, так как здесь не соблюдено условие, по которому одинаковые частоты должны быть представлены равными площадями.

Графики, подобные приведенному выше, обычно применяются для сравнения коэффициентов смертности и продолжительности жизни в различных группах населения. В этом случае против определенного возраста ставится точка, соответствующая числу лиц,

доживших до него. Так как коэффициент смертности определяется скоростью уменьшения логарифма для числа лиц, доживших до определенного возраста, то равным наклонам таких кривых соответствуют одинаковые коэффициенты смертности. Поэтому такие диаграммы весьма удобны для наблюдения над увеличением коэффициента смертности по мере увеличения возраста и для сравнения смертности в различных группах населения. Диаграммы этого типа менее отзывчивы на небольшие флюктуации, чем соответствующие диаграммы численностей, дающие распределение населения по возрастам, в которых наступила смерть. Поэтому рассматриваемые здесь диаграммы более пригодны в тех случаях, когда такие небольшие флюктуации обусловлены влиянием ошибок случайного отбора, которые в более чувствительных диаграммах могут завуалировать многие характерные особенности сравнений. Следует, кстати, заметить, что выбор надлежащих методов статистической обработки материала ни в какой мере не зависит от выбора формы для графического изображения.

Часто в генетике, а иногда и в других науках приходится иметь дело с некоторыми отношениями численностей, которые устанавливаются для каждой выборки, взятой из серии их. Эти выборки по данному признаку могут быть однородными или, наоборот, неоднородными. Классификация выборок, таких, как потомство растений или животных, в соответствии с величиной отношения численностей в них, а также установление степени однородности этих выборок являются в этих науках весьма важными вопросами, требующими при своем решении применения точных критериев. Эти последние будут даны в главе IV. Форма же графического изображения таких наблюдений обычно выбирается в соответствии с изучаемым частным вопросом, причем стремятся придать этому изображению наиболее воспринимаемую форму.

В частном случае, когда численности соответствуют двум альтернативам, эти численности для каждой отдельной выборки могут представлять координаты некоторой точки, а множество таких точек будет характеризовать отношения их в выборках. На рис. 5.1 применен более удобный способ, при котором размещаются не сами численности, а их квадратные корни. В этом случае точки, относящиеся к выборкам из n наблюдений, будут попадать внутрь квадранта круга, радиус которого $\sqrt{n}$. Выборки, в которых отношение численностей равно $p:q$, если $p+q=1$, будут попадать на радиус — вектор, образующий с осью угол φ, так что

$$\sin^2\varphi = p, \quad \cos^2\varphi = q.$$

Этот способ построения диаграммы позволяет охватить гораздо более широкий интервал размера выборок и более широкий интервал отношений численностей, чем это возможно при других условиях. Мостеллер и Такей на этом принципе изготовили специально графленую бумагу, которая теперь стала доступной.

Учитывая, что средняя квадратическая ошибка случайного по-

явления φ при данном n пропорциональна $\frac{1}{\sqrt{n}}$ и не зависит от φ, можно прийти к заключению, что разбросанность наблюденных точек по ту и другую сторону соответствующего луча должна быть во всех частях диаграммы одинаковой. В связи с этим по-

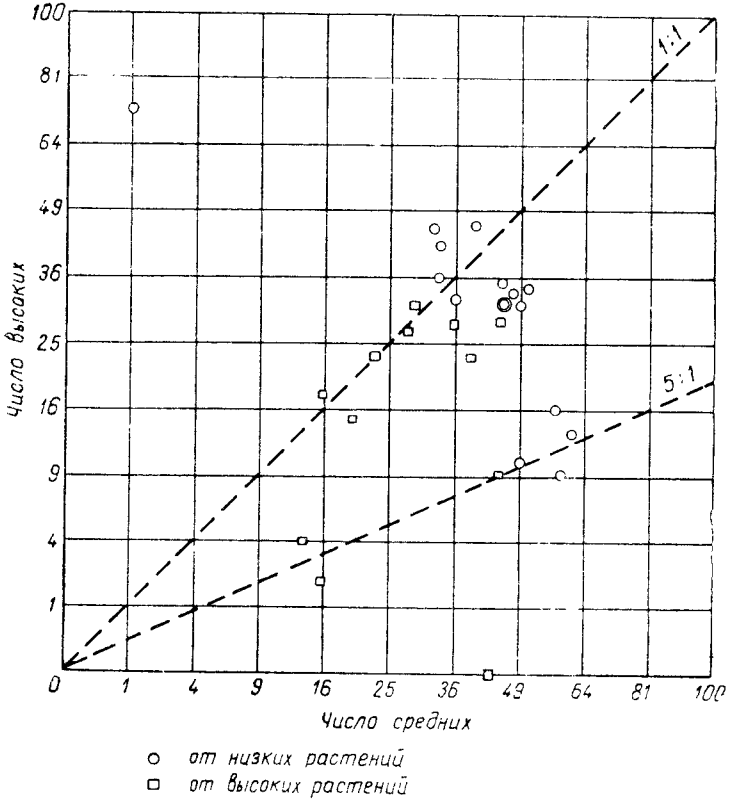


Рис. 5.1. Численности, расположенные в виде диаграммы квадратных корней.

является возможность производить глазомерную оценку однородности групп.

В материале, относящемся к Lythrum salicaria и представленном на рис. 5.1, в соответствии с ожидаемым было обнаружено три класса, представленных 1, 19 и 7 семьями. Одна семья, состоящая из 41 растений (причем все растения имели пестик средней длины), которая очевидно принадлежит к четвертому классу, явилась некоторой неожиданностью: дальнейшие опыты показали, что она содержит три доминантные гена «мид» благодаря тому, что в предшествующем мейозисе произошло двойное редукционное деление и что в результате этого процесса она дала в более обширном опыте 2% «лонг».

ГЛАВА ТРЕТЬЯ

РАСПРЕДЕЛЕНИЯ

11. В основе всех статистических исследований лежит понятие о бесконечной *генеральной совокупности* одной или нескольких переменных, имеющей определенное *распределение плотности вероятности*. Например, основываясь на ограниченном числе наблюдений над отдельными представителями некоторого биологического вида или над элементами погоды определенной местности, мы можем мысленно представить себе некоторую гипотетическую бесконечную совокупность, из которой взята наша выборка наблюдений, а также допустить существование вероятностной природы у будущих выборок, на которые и распространяются наши выводы. Если вторая выборка противоречит этому допущению, то мы делаем вывод, что она, говоря статистическим языком, взята из другой генеральной совокупности, т. е. что условия, в которых находилась эта вторая группа организмов, были существенно иными, чем у первой, или что климат (или методы его определения) также в значительной мере изменился. Критерии, позволяющие решать такие вопросы, могут быть названы критериями существенности; имея подходящий данному случаю критерий, мы можем установить, отличается ли существенно вторая выборка от первой, или такого различия нет.

Величину, определяемую по выборочным данным для характеристики генеральной совокупности, к которой принадлежит наша выборка, будем называть *статистикой*. Например, такой статистикой является средняя некоторого числа наблюдений $x_1, x_2, \ldots x_n$, определяемая уравнением

$$\bar{x} = \frac{1}{n} S(x),$$

где S означает суммирование всех выборочных данных (символ S будет в этом смысле использоваться нами и в дальнейшем) и n — число наблюдений. Конечно, такого рода статистики меняют свои значения от выборки к выборке, и теория распределений определяет специальную характеристику изменчивости этих значений. Если мы точно знаем распределение генеральной совокупности, то теоретически возможно, хотя это часто сопряжено с большими математическими затруднениями, определить распределение той или иной статистики, полученной на основе выборки данного размера. Практическое использование такой статистики и закона ее распределения зависит от особенностей распределения генеральной совокупности, в связи с чем разработка соответствующих приближенных или точных методов оказалась доступной только в очень небольшом числе случаев. На практике чаще всего основываются на том факте, что распределения многих статистик имеют тенденцию приближаться к *нормальному* распределению по мере увеличения объема выборки. В связи с этим нормальная форма распределения применяется к широкому кругу случаев, относящихся к так называемой «теории больших выборок». Основываясь на этой теории, допускают, что такие статистики нормально распределены, в соответствии с чем представляется возможным рассматривать средние квадратические ошибки в качестве приближенных показателей изменчивости таких статистик.

В настоящей главе мы рассмотрим три основных вида распределений: 1) нормальное распределение, 2) распределение Пуассона и 3) биноминальное распределение. Читателю необходимо иметь общее представление об этих трех распределениях, о математических формулах, которыми они характеризуются, об условиях их возникновения и о статистических методах установления тех случаев, когда они могут встретиться. В отношении последней темы нам придется предвосхитить методы, более полное изложение которых будет дано в главах IV и V.

12. Нормальное распределение

Переменная величина распределена нормально, если она может принимать значения от $-\infty$ до $+\infty$ с частотами, подчиненными определенному математическому закону, а именно: логарифм частоты на некотором расстоянии d от центра этого распределения меньше логарифма частоты в центре распределения на величину, пропорциональную d^2. Следовательно, это распределение симметричное, с наибольшей частотой в центре. Хотя переменная величина в этом случае изменяется неограниченно, однако частоты на некотором более или менее значительном расстоянии от центра становятся очень малыми, ибо большие отрицательные логарифмы соответствуют весьма малым числам. На рис. 6 дана кривая нормального распределения, обозначенная В. При нормальном распределении частота, соответствующая некоторому бесконечно малому интервалу dx, определяется формулой

$$df = \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2}\cdot\frac{(x-\mu)^2}{\sigma^2}} dx,$$

где $x - \mu$ является расстоянием наблюденного значения переменной x от центра распределения μ, а σ, называемая *средним квадра-*

тическим отклонением, измеряет рассеяние отдельных значений переменной около центра. Геометрически σ определяет точки перегиба кривой, т. е. расстояния по обе стороны от центра тех точек, в которых кривая изменяется наиболее резко.

Потребность в определении частоты для того или иного значения x на практике возникает не столь часто; чаще всего встречается задача определения суммарной численности для всех значений, превосходящих данное значение x. Эта численность геометрически будет представлена площадью «хвостовой» части кривой, отсеченной в данной точке. Существуют специальные таблицы, дающие такие суммарные численности, т. е. интегралы вероятностей. По этим таблицам можно для любого значения $(x-\mu)/\sigma$

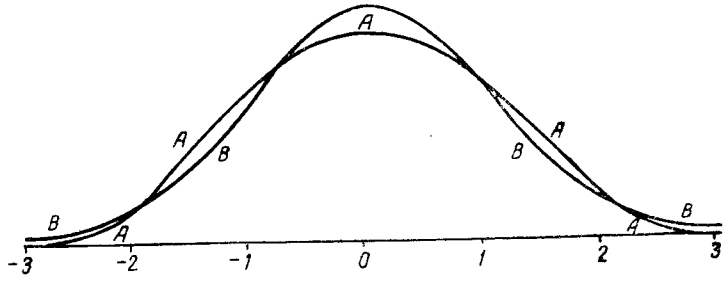


Рис. 6. Сравнение симметричной кривой и кривой нормального распределения.
А — плосковершинная кривая (γ_2 — отрицательное). В — нормальная кривая ($\gamma_2 = 0$).

определить долю совокупности, соответствующую отклонениям большим, чем данное. Другими словами, можно определить вероятность того, что значение переменной, отобранное случайно, будет иметь отклонение от центра большее, чем данное. Таблицы I и II, приведенные в конце настоящей главы, построены так, что определяют отклонения, соответствующие различным значениям вероятности. Рассматривая эти таблицы, можно видеть, как быстро убывает вероятность по мере увеличения расстояния от центра; отклонение, превосходящее среднее квадратическое отклонение σ, встречается примерно один раз в трех испытаниях; удвоенное среднее квадратическое отклонение может быть превзойдено примерно только один раз из 22 испытаний; утроенное среднее квадратическое отклонение — один раз из 370 испытаний. Как показывает таблица II, шестикратное превышение среднего квадратического отклонения встречается всего навсего в одном из миллиарда случаев. Значению вероятности $P = 0{,}05$, т. е. отношению шансов 1 : 20, соответствует отклонение 1,96, или, округляя, 2; представляется удобным взять это значение в качестве критерия для суждений о существенности или несущественности данного отклонения. В этом случае отклонение, превосходящее

удвоенное среднее квадратическое отклонение, рассматривается как существенное. Используя этот критерий в неограниченном числе случаев и руководствуясь при этом только одними этими статистиками, мы придем к неправильным заключениям только в одном из 22 испытаний. Следует отметить, что на этот критерий оказывает влияние недостаточная численность наблюдений, но это дополнительное условие не должно сказываться на выбранном уровне существенности.

Иногда может возникнуть некоторое противоречие, связанное с тем обстоятельством, что в одних случаях нас может интересовать вероятность того, что положительное отклонение превосходит некоторое наблюденное значение переменной, а в других случаях рассматриваются отклонения, которые в одинаковой мере могут быть как положительными, так и отрицательными. Вероятность последнего случая всегда равна половине вероятности первого случая. Так, из таблицы I находим, что нормальное отклонение окажется вне пределов $\pm$ 1,598193 в 11% всех случаев и, следовательно, оно будет превосходить + 1,598193 только в 5,5% случаев.

Значение того отклонечия, выше которого лежит половина всех наблюдений, носит название *квартильного* расстояния; оно составляет 0,67449 часть среднего квадратического отклонения. В недалеком прошлом широко применялась *вероятная ошибка*, которая получалась путем умножения средней квадратической ошибки на этот множитель. Таким образом, вероятная ошибка представляет собой около двух третей средней квадратической ошибки; утроенная вероятная ошибка в качестве критерия существенности равноценна удвоенной средней квадратической ошибке. Применение вероятной ошибки ограничено тем обстоятельством, что таблицы (например, таблицы I и II) построены так, что для нахождения критических значений необходимо иметь отклонения, выраженные в долях средней квадратической ошибки, т. е. отношение $(x-\mu)/\sigma$.

Некоторые дополнительные таблицы, относящиеся к нормальному распределению, приведены в «Статистических таблицах» (Statistial Tables) табл. IX и X и в «Таблицах Шеппарда» (Sheppard's Tables, 1938 г.).

13. Расчет характеристик нормального распределения

Неизвестные характеристики нормального распределения — среднюю и среднее квадратическое отклонение — можно *оценить* при помощи двух легко вычисляемых статистик. Если имеется выборка, состоящая из n наблюдений, то наилучшей оценкой средней μ будет $\bar{x}$, т. е. выборочная средняя, определяемая по формуле

$$\bar{x} = \frac{1}{n} S(x),$$

а наилучшей оценкой σ будет s, вычисляемая по формуле

$$s^2 = \frac{1}{n-1} S(x - \bar{x})^2.$$

Эти две статистики определяются по суммам первых двух степеней наблюденных значений x (см. приложение на стр. 63) и имеют ту специфическую особенность, характерную только для нормального распределения, что они сосредоточивают в себе всю информацию относительно генеральной совокупности, содержащуюся в данной выборке. Расчет сумм различных степеней x, в особенности для построения систем таких статистик, которые известны под названием *моментов*, находит свое широкое применение и в случае скошенных (асимметричных) и других отклоняющихся от нормального распределений. Однако эти последние, вообще говоря, не обладают тем свойством нормального распределения, согласно которому две первые степени x достаточны для полной характеристики этого распределения. Более или менее значительное отклонение распределения от нормальной формы делает недостаточным знание этих двух статистик ($\bar{x}$ и s) для характеристики этого распределения.

Пример 2. *Расчет нормального распределения при большой выборке.* Чтобы вычислить статистику s по данным выборки большого размера, нет необходимости вычислять отдельные отклонения каждого из наблюдений от средней и возводить их в квадрат. В данном случае можно применить сокращенный способ вычисления путем образования групп с одинаковыми интервалами. Этот способ дан в табл. 2, где обработаны измерения роста 1164 мужчин.

Первая графа дает серединные значения для каждой группы (рост в дюймах); следующая графа заполнена соответствующими частотами. Серединное значение центральной группы (68,5 дюйма) выбрано за «рабочую среднюю», или «условное начало». Для получения следующей графы умножают частоты на числа 1, 2, 3 и т. д., соответственно расстоянию этих частот от условного начала: этот процесс повторен и для следующей, четвертой графы. В этой последней графе суммирование производится сверху вниз по способу нарастающего итога, но в третьей графе, в верхней части которой находятся отрицательные, а в нижней — положительные отклонения, суммирование производится раздельно для этих двух частей, после чего из последней суммы вычитается первая. В нашем примере эта разность указывает на то, что вся совокупность 1164 наблюдений в целом на 167 дюймов выше той суммы, которая была бы при условии, что рост всех мужчин был бы равен точно 68,5 дюймов. Делением этого результата (167 дюймов) на численность выборки (1164) получаем ту величину, на которую фактический средний рост превышает рабочую

среднюю, равную 68,5 дюйма. В данном случае средний рост равен 68,6435 дюйма.

Для того чтобы перейти от отклонений от условного начала к отклонениям от фактической средней, следует в сумму четвертой графы ввести поправку путем вычитания из нее произведения итога третьей графы (167 дюймов) и полученной из этого итога средней 0,1435 дюйма. Исправленная сумма квадратов, деленная на 1163, т. е. на численность выборки, уменьшенную на единицу, будет 7,3861 квадратных дюйма. Эта величина является оценкой дисперсии и лежит в основе всех последующих расчетов.

Извлечение квадратного корня из этой величины приводит к оценке среднего квадратического отклонения. Так, извлекая квадратный корень из 7,3861, мы получаем 2,7177 в качестве оценки среднего квадратического отклонения. Однако эта величина вычислена на основе сгруппированных данных, а группировка влечет за собой некоторую погрешность, которая обусловлена допущением, что все значения от $-1/2$ до $+1/2$ интервала группировки имеют одинаковые частоты. Влияние группировки выражается в том, что дисперсия генеральной совокупности и в среднем дисперсия выборки увеличивается на $\frac{1}{12} = 0{,}0833$ величины интервала. Эта величина вычитается из оценки дисперсии и называется поправкой Шеппарда. При введении этой поправки в нашем случае получаем 7,3028 квадратных дюйма для дисперсии и 2,702 дюйма для среднего квадратического отклонения.

Описанные здесь расчеты в одинаковой мере применимы при любом интервале, взятом в качестве единицы группировки, причем вся обработка данных ведется в этих условных единицах и только конечный результат перечисляется в исходные единицы измерения. Так, если в нашем случае требуется перейти от измерения роста в дюймах к измерению в сантиметрах, то просто следует умножить среднюю и среднее квадратическое отклонение на соответствующий переводной коэффициент. Следует только учитывать, что единица группировки должна быть кратной единице измерения; если бы в нашем случае рост измерялся с точностью до десятой доли дюйма, то можно было бы взять ширину интервала, например, 0,6 или 0,7 дюйма.

Если вычисленные выше показатели рассматривать в качестве оценок средней и среднего квадратического отклонения нормальной генеральной совокупности, то на их величину будут оказывать влияние ошибки случайного отбора; это означает, что вторая выборка из этой совокупности не даст точно такие же значения этих показателей. Но эти меняющиеся значения показателей для различных (больших) выборок данного объема n будут распределены примерно по нормальному закону распределения, что позволяет выразить точность любой из таких оценок через ее среднюю квадратическую ошибку. Эти средние квадратические ошибки должны быть вычислены на основе дисперсии, взятой без по-

Таблица 2

Рост в дюймах	Численности мужчин	Произведение численностей на отклонения	Произведения численностей на квадрат отклонения	Численности женщин
52,5	—	—	—	0,5
53,5	—	—	—	0,5
54,5	—	—	—	—
55,5	—	—	—	1
56,5	—	—	—	5
57,5	—	—	—	15
58,5	—	—	—	15,5
59,5	1	−9	81	52
60,5	2,5	−20	160	101
61,5	1,5	−10,5	73,5	150
62,5	9,5	−57	342	199
63,5	31	−155	775	223
64,5	56	−224	896	215
65,5	78,5	−235,5	706,5	169,5
66,5	127	−254	508	151,5
67,5	178,5	−178,5	178,5	81,5
68,5	189	−1143,5		40,5
69,5	137	137	137	19,5
70,5	137	274	548	10
71,5	93	279	837	5
72,5	52,5	210	840	—
73,5	39	195	975	1
74,5	17	102	612	—
75,5	6,5	45,5	318,5	—
76,5	3,5	28	224	—
77,5	1	9	81	—
78,5	2	20	200	—
79,5	1	11	121	—
	1164	1310,5 +167	8614	1456
Средняя		+0,1435		
Поправка для средней .		$167^2 : 1164$	23,96	
Исправленная сумма квадратов			8590,04	

	Оценка дисперсии	Среднее квадратическое отклонение
	7,3861	2,7177
Выборочная дисперсия средней . .	0,006345	0,0797
Выборочная дисперсия из совокупности дисперсий	0,09382	0,3063
Поправка на группировку	0,0833	—
Исправленная дисперсия	7,3028	2,7024

правки Шеппарда, в связи с чем именно эта дисперсия лежит в основе всех расчетов при большой выборке.

Формулы для дисперсий выборочной средней и дисперсии, полученных в порядке случайного отбора из нормальной совокупности, таковы (см. приложение стр. 63):

$$\frac{\sigma^2}{n} \text{ и } \frac{2\sigma^4}{n-1}.$$

Подставляя в эти формулы значение $k_2 = 7{,}3861$ вместо σ^2, мы найдем, что наша оценка средней имеет дисперсию, характеризующую ошибку выборки, равную 0,006345 квадратных дюйма; извлекая отсюда квадратный корень, получим соответствующую среднюю квадратическую ошибку в 0,0797 дюйма. На основе этих данных можно сделать заключение, что наша выборка существенно отклоняется (на $\pm$ удвоенную среднюю квадратическую ошибку) от любой совокупности, средняя которой лежит вне интервала от 68,48 до 68,80 дюйма. Отсюда следует, что в аспекте теории доверительных оценок вполне возможно, что средняя генеральной совокупности, из которой взята наша выборка, лежит именно в этом интервале. Подобно этому наша дисперсия имеет свою выборочную дисперсию 0,09382 или среднюю квадратическую ошибку 0,3063 и, следовательно, мы имеем достаточно оснований считать, что сгруппированная по выбранным нами интервалам генеральная совокупность, из которой взята наша выборка, имеет дисперсию в пределах от 6,773 до 7,999 квадратных дюйма. Для определения же ошибки дисперсии несгруппированной генеральной совокупности мы должны от обоих пределов отнять 0,083.

Возникает вопрос: не наносит ли более или менее заметный ущерб точности оценок широко применяемая на практике группировка данных? Группировка в конце концов сводится к тому, что вместо фактических данных подставляются условные, соответствующие серединам интервалов. Очевидно, что грубая группировка (по относительно крупным интервалам) может привести к серьезным искажениям. Доказано, что при оценке параметров нормальной совокупности потеря информации, обусловленная группировкой, составляет меньше 1%, если интервал группировки не превосходит четвертую часть среднего квадратического отклонения. Группировка в приведенном выше примере, где ширина интервала равна одному дюйму при среднем квадратическом отклонении 2,7024, несколько более груба: здесь потеря информации при оценке среднего квадратического отклонения составляет 2,28%, что равносильно потере 27 наблюдений из 1164; потеря же информации в отношении средней составляет половину этой величины. При надлежащем подборе интервала группировки ущерб в отношении точности незначителен, но этим путем удается в большой степени сократить затраты труда на обработку данных.

Возможна и другая постановка вопроса о потере информации,

возникающей в связи с применением группировки. В этом случае ставится вопрос, в какой мере оценки средней и среднего квадратического отклонения, полученные при группировке данных, будут близки к таким же оценкам, вычисленным без группировки. В соответствии с этой постановкой вопроса можно определить среднюю квадратическую *ошибку группировки*, которую не следует смешивать с средней квадратической ошибкой случайного отбора, измеряющей отклонение выборочного значения показателя от соответствующего показателя в генеральной совокупности. Эта обусловленная группировкой данных средняя квадратическая ошибка для средней и среднего квадратического отклонения в обоих случаях одна и та же и равна

$$\frac{1}{\sqrt{12n}}$$

интервала группировки; в нашем случае она равна 0,0085 дюйма. Для вполне доброкачественной группировки данных необходимо, чтобы эта ошибка не превышала одной десятой средней квадратической ошибки случайного отбора.

При обработке данных большой выборки довольно часто в качестве оценки дисперсии берут величину

$$\frac{1}{n} S(x-\bar{x})^2,$$

которая отличается от предыдущей формулы (см. стр. 44) тем, что в ней делитель является n вместо $(n-1)$. При большой выборке различие между двумя этими формулами незначительно. Использование в качестве делителя n может иметь некоторое теоретическое основание, когда такую дисперсию применяют при вычислении кривой распределения, вообще же говоря, лучше всегда применять делитель $(n-1)$. Для малых выборок это различие все еще мало по сравнению с вероятной ошибкой, но оно становится значительным в случаях, когда дисперсия получается в качестве усередненной оценки, полученной из серии малых выборок. Если, например, проведен ряд экспериментов в одинаковых общих условиях, причем каждый из них содержит в себе шесть наблюдений, то для получения так называемой несмещенной оценки дисперсии мы должны вычислить средний квадрат по формуле

$$\frac{1}{n-1} S(x-\bar{x})^2 = \frac{1}{5} S(x-\bar{x})^2.$$

Если же сумму $S(x-\bar{x})^2$ разделить на 6, то мы получим преуменьшенную оценку дисперсии.

14. Оценка отклонения данного распределения от нормальной формы

В ряде случаев возникает необходимость определить, в какой мере данная выборка близка к нормальному распределению. При решении этого вопроса приходится исчислять третьи, а иногда

и четвертые степени x; на основе тех и других можно вычислить величины g, средние значения которых в условиях нормального распределения равны нулю и которые при больших выборках распределены нормально с средними квадратическими ошибками, определяемыми размером выборки. Величина g_1, определяемая по третьим степеням x, является мерой асимметрии. Параметр γ_1, оценкой которого она является, может быть приравнен к величине $\pm\sqrt{\beta_1}$, введенной в статистику Пирсоном, хотя сам Пирсон использует символ β_1 и для обозначения другой статистики, которая не эквивалентна величине g_1[2]. Величина g_2, вычисляемая по четвертым степеням x, также характеризует приближение к нормальности. В этом случае величина g_2, отличная от нуля, характеризует симметричную кривую распределения, у которой вершина и концы возвышаются над нормальной кривой, а промежуточная часть находится ниже этой последней; или же, наоборот, вершина и концы принижены, а промежуточная часть выходит из пределов нормальной кривой, так что получается относительно пологая кривая (см. рис. 6, стр. 42).

Пример 3. *Вычисление критериев нормальности.* Отклонения от нормальной формы распределения, если только они не представляются явными без всякой оценки, могут быть обнаружены только в случае большой выборки; при малых же выборках оказывается невозможным определение сколько-нибудь надежных статистических критериев для этих отклонений. Ниже приводится пример (табл. 3) обработки 90 наблюдений годовых осадков в Ротамстеде. Процесс вычислений здесь подобен ходу вычислений средней и среднего квадратического отклонения, но только он продолжен на два этапа, где вычисляются суммы третьих и четвертых степеней x.

Формулы, по которым здесь вычислены средняя и статистики k и g, приведены в приложении на стр. 63. Для статистик k получены следующие значения (ширина интервала принята за единицу):

$$k_1 = 0{,}62; \quad k_2 = 23{,}2715; \quad k_3 = -25{,}76; \quad k_4 = -162{,}49;$$

отсюда находим

$$g_1' = k_3' / (k_2')^{3/2} = -0{,}231 \quad \text{и} \quad g_2' = k_4' / (k_2')^2 = -0{,}302.$$

В приложении (стр. 63) даны формулы дисперсий для величин g_1 и g_2 при нормальном распределении, а в табл. 3 приведены численные значения средних квадратических ошибок этих величин. Сопоставление g_1 и g_2 с их ошибками убеждает в том, что значения g_1 и g_2 не больше своих ошибок и, следовательно, они не отклоняются сколько-нибудь существенно от нуля.

Отрицательное значение параметра γ_1 (хотя его существенность в нашем случае не установлена) характеризует асимметричность распределения, которая проявляется в том, что умеренно сухие и весьма влажные годы встречаются соответственно менее часто, чем умеренно влажные и очень сухие годы.

Таблица 3

Исследование вопроса о приближении распределения сумм годовых осадков к нормальной форме

Годовые осадки в дюймах	Частоты				
16	1	—12	144	—1728	20 736
17	—	—	—	—	—
18	—	—	—	—	—
19	3	—27	243	—2187	19 683
20	2	—16	128	—1024	8 192
21	3	—21	147	—1029	7 203
22	—	—	—	—	—
23	3	—15	75	—375	1 875
24	2	—8	32	—128	512
25	12	—36	108	—324	972
26	4	—8	16	—32	64
27	7	—7	7	—7	7
28	4	—	—	—	—
29	8	8	8	8	8
30	9	18	36	72	144
31	6	18	54	162	486
32	7	28	112	448	1 792
33	4	20	100	500	2 500
34	4	24	144	864	5 184
35	4	28	196	1372	9 604
36	3	24	192	1536	12 288
37	3	27	243	2187	19 683
38	—	—	—	—	—
39	1	11	121	1331	14 641
s	90	56	2106	1646	125 574
Поправки на среднюю			—34,84	—3931,2 +43,4	—4 096,2 +4 892,2 —40,5
S	90	—	2071,16	—2241,8	126 329,0
k	—	0,62	23,2715	—25,761	—162,487
Поправка	—	—	—0,0833	—	+0,008
k'		0,62	23,1882	—25,761	—162,479
g				—0,231	—0,302
Средняя квадратическая ошибка				±0,254	±0,503

15. Дискретные распределения

В некоторых случаях переменная x не может иметь все возможные значения, а принимает только некоторые из них, например, только целые значения. Это последнее положение возникает обычно тогда, когда переменная x сама является частотой, полученной в результате подсчета. Например, число клеток, наблюдаемых при анализе крови, или число колоний какой-либо культуры на некоторой площади питательной среды. В то время как нормальное распределение является наиболее важным среди непрерывных распределений, столь же большое значение среди дискретных распределений занимает распределение Пуассона. Если переменная может принимать только значения 1, 2, ... x ... с относительными частотами, соответствующими последовательным членам ряда

$$e^{-m}\left(1,\ m,\ \frac{m^2}{2!},\ \ldots\ \frac{m^x}{x!},\ \ldots\right),$$

где $x!$ является «факториалом x», т. е. $x! = x(x-1)(x-2)\ldots 1$, то тогда говорят, что x имеет распределение Пуассона. Общая сумма этих относительных частот равна единице, так как

$$e^m = 1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots.$$

Нормальное распределение, как мы знаем, определяется двумя параметрами μ и σ, распределение же Пуассона имеет только один параметр m. Этот параметр может быть оценен при помощи средней, исчисленной по наблюденным данным, причем эта средняя в распределении Пуассона имеет тот же смысл, что и в нормальном распределении. Теоретически может быть доказано, что если вероятность появления некоторого факта очень мала, то частота, с которой этот факт будет встречаться, должна быть распределена по закону Пуассона. Например, вероятность смерти человека вследствие того, что его лягнула лошадь, в каждый отдельный день, безусловно, ничтожна, но если распространить наблюдения, положим, на целый армейский корпус и на целый год, то могут встретиться случаи, один или даже несколько, смерти вследствие указанной выше причины. Приводимые далее данные (Борткевич) были получены на основе отчетов десяти армейских корпусов за двадцать лет; всего имелось двести наблюдений:

Таблица 4

Число смертей за год в корпусе	Наблюденные частоты	Вычисленные частоты
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
5	—	0,08
6	—	0,01

Здесь средняя $\bar{x}$ равна 0,61. Если взять ее в качестве оценки параметра m и произвести расчет ожидаемых частот, то получим

ряд, хорошо согласующийся с рядом наблюденных частот. Особенно большое значение распределение Пуассона имеет в биологических исследованиях, где оно было впервые использовано при обработке наблюдений при помощи гемацитометра. Теоретически было установлено, что если техника наблюдения достаточно совершенна, то число посторонних клеток в каждом квадрате поля наблюдения должно быть распределено по закону Пуассона. Дальнейшее изучение этого вопроса позволило прийти к выводу, что при указанных условиях это распределение осуществляется на практике с большой точностью. Иллюстрацией этого может служить приводимая ниже таблица (данные Стьюдента), где дано распределение дрожжевых клеток в 400 квадратах размером в один квадратный миллиметр. Общее число обнаруженных клеток составляет 1872 и, следовательно, среднее число их в одном квадрате 4,68. Ожидаемые частоты, вычисленные на основе этой средней, довольно близки к наблюденным. С методами оценки этого приближения мы познакомимся в главе IV.

Таблица 5

Число клеток	Наблюденные частоты	Вычисленные частоты
0	—	3,71
1	20	17,37
2	43	40,65
3	53	63,41
4	86	74,19
5	70	69,44
6	54	54,16
7	37	36,21
8	18	21,18
9	10	11,02
10	5	5,16
11	2	2,19
12	2	0,86
13	—	0,31
14	—	0,10
15	—	0,03
16	—	0,01
Итого . . .	400	400,00

Если изучаемая величина определяется как сумма ряда других величин, каждая из которых подчинена распределению Пуассона, то эта суммарная величина также распределена по тому же самому закону. Поэтому, например, общее число обнаруженных грибковых клеток (1872) может рассматриваться как выборка из такого распределения, в котором m примерно равно 1872. Дисперсия распределения Пуассона равна, так же как и средняя, параметру m, а при таких больших значениях m, как 1872, распределение выборочных его значений становится довольно близким к нормальному. Следовательно, в нашем случае мы можем сопроводить результат подсчета — 1872 клетки — соответствующей средней квадратической ошибкой $\pm\sqrt{1872} = \pm 43{,}26$. Таким образом, плотность клеток в данной суспенсии определена со средней квадратической ошибкой, равной 2,31%. Если, допустим, вторая параллельная проба отличается от данной на 7%, то возникает сомнение в отношении техники отбора проб.

16. Малые выборки из распределения Пуассона

Те же самые принципы, которые были применены выше к вопросу об оценке точности подсчета с помощью гемацитометра, применимы и при подсчете бактериальных колоний. В этом случае применяется метод ослабления концентрации, причем техника этого ослабления должна обеспечивать полную случайность распределения организмов и условия, чтобы последние развивались на данной поверхности, не оказывая друг на друга никакого влияния. В данном случае степень согласованности наблюдений с распределением Пуассона становится критерием того, в какой мере удовлетворительны техника исследования и подбор питательной среды, подобно тому как ранее она служила критерием совершенства микроскопического исследования крови. Однако фактически имеет место довольно большое различие между тем и другим случаем: в то время как при микроскопическом анализе крови обычно имеют дело с довольно большим числом квадратов, внутри каждого из которых обнаруживается только небольшое число интересующих нас клеток, при подсчете бактерий обычно имеется только 5 параллельных пластинок, на каждой из которых может расселиться до 200 колоний. При единичной выборке в 5 наблюдений нет никакой возможности определить, подчиняется ли данное распределение закону Пуассона, но когда имеется довольно большое число таких малых выборок, полученных в сравнимых условиях, то появляется возможность использовать то обстоятельство, что в распределении Пуассона дисперсия всегда численно равна средней.

После определения средней $\bar{x}$ по числу колоний $x_1, x_2, \ldots x_n$ для каждого ряда параллельных пластинок можно построить следующий показатель рассеяния:

$$\chi^2 = \frac{S(x-\bar{x})^2}{\bar{x}}.$$

Доказано, что в выборках из распределения Пуассона величина χ^2, вычисленная указанным образом, распределена по определенному закону, который отражен в таблице III (стр. 96). Для нахождения табличного значения χ^2 следует взять за n число на единицу меньшее, чем число параллельных пластинок. При малых

выборках допустимый размах варьирования χ^2 весьма широк; так, при пяти пластинках, т. е. при $n = 4$, величина χ^2 в 10% случаев будет меньше 1,064 и в 10% случаев она будет больше 7,779; таким образом, единичная выборка в 5 наблюдений дает нам очень небольшую информацию. Но если мы имеем 50 или 100 таких выборок, то появляется возможность произвести достаточно точную проверку согласия данного распределения с законом Пуассона.

Пример 4. *Оценка согласия некоторого числа малых выборок с распределением Пуассона.* В результате 100 подсчетов бактерий, выращенных на чистом сахаре, были получены следующие данные (табл. 6). Здесь в каждом отдельном случае было взято 6 пластинок, заселенных бактериями, и поэтому из таблицы χ^2 брались значения, соответствующие $n = 5$.

Очевидно, что фактические частоты весьма значительно отличаются от ожидаемых; здесь бросается в глаза превышение первых над последними в первом интервале (χ^2 от 0 до 0,554) и при значениях χ^2, больших 15; с другой стороны, при относительно небольших значениях χ^2, от 2 до 15, не наблюдается больших нарушений ожидаемых соотношений между частотами, что подтверждается последним столбцом таблицы, где произведен расчет 43% от ожидаемых частот. Это обстоятельство указывает на то, что в данном случае достаточно было бы иметь около половины наблюдений; но наряду с этим здесь имеется около 10% слишком больших значений χ^2 и примерно в 45% случаев наблюдается ненормально малая изменчивость.

Таблица 6

χ^2	Ожидаемые частоты	Наблюденные частоты	Ожидаемые в 43% случаев
0	1	26	0,43
0,554	1	6	0,43
0,732	3	11	1,29
1,145	5	7	2,15
1,610	10	7	4,3
2,343	10	2	4,3
3,000	20	12	8,6
4,351	20	7	8,6
6,064	10	3	4,3
7,289	10	4	4,3
9,236	5	1	2,15
11,070	3	3	1,29
13,388	1	0	0,43
15,086	1	11	0,43
Итого . . .	100	100	43,00

В некоторых случаях возникает потребность в оценке размера варьирования при таких обстоятельствах, когда имеется не серия наблюдений с одинаковым числом пластинок, которые можно было бы объединить в одну большую выборку, а некоторое число серий с различным количеством параллельных пластинок. В этом случае нельзя проверить точность отображения теоретического распределения, но имеется возможность определить степень согласованности общего размера варьирования с теоретически ожидаемым размером. Такая проверка основывается на том факте, что распределение суммы некоторого числа независимых величин χ^2 подчинено тому же самому закону, который дан в таблице χ^2, только теперь следует вместо n, определяемого повторностью отдельного эксперимента, брать сумму $S(n)$. Так, для шести экспериментов, по четыре пластинки каждый, было найдено общее значение $\chi^2 = 13{,}85$; соответствующее значение $S(n)$ будет $6 \times 3 = 18$. По таблице χ^2 находим, что значение $\chi^2 = 13{,}85$ при $n = 18$ будет наблюдаться между 70 и 80% всех случаев и, следовательно, оно не является маловероятным.

В другом исследовании были получены следующие результаты:

Таблица 7

Число пластинок в серии	Число серий	$S(n)$	Суммарное значение χ^2
4	8	24	27,31
5	36	144	133,96
9	1	8	8,73
Всего . . .	—	176	170,00

Таким образом, нам предстоит установить, будет ли выходить из нормы величина $\chi^2 = 170$ при $n = 176$. Таблица χ^2 составлена только до $n = 30$, но можно произвести расчет χ^2 и для более высоких значений n, воспользовавшись тем обстоятельством, что распределение χ в этих случаях становится почти нормальным. Хорошее приближение дает также величина $(\sqrt{2\chi^2} - \sqrt{2n-1})$, которая распределена нормально около нуля со средним квадратическим отклонением, равным единице. Если эта величина превосходит 2 или даже 1,645, что соответствует 5% уровню вероятности, то величина χ^2 существенно превосходит теоретически ожидаемое значение. В нашем примере мы имеем:

$$2\chi^2 = 340; \quad \sqrt{2\chi^2} = 18{,}44$$
$$2n - 1 = 351; \quad \sqrt{2n-1} = 18{,}73$$
$$\text{Разность} = -0{,}29.$$

Следовательно, в этих 45 экспериментах различия между повторными пластинками весьма близки к тем, которые ожидаются теоретически. В результате этого анализа можно сделать вывод, что техника экспериментов была удовлетворительной.

17. Наличие или отсутствие организмов в пробе

Если условия, при которых взяты пробы, находятся в соответствии с условиями возникновения рядов Пуассона, то, как мы видели, число проб, содержащих 0, 1, 2 ..... организмов, определяется средним числом организмов в пробе. Если же организмы подвижны, а также в других случаях, когда нет условий для образования дискретных колоний, то среднее число организмов, которое не может быть определено непосредственно, находится по доли фертильных культур, возникших из одного способного к развитию организма. Если m — среднее число организмов в пробе, то доля проб, не содержащих ничего, т. е. доля стерильных проб, будет e^{-m}. На основании этого можно вычислить среднее число организмов, соответствующее 10%, 20% и т. д. фертильных проб. Результаты расчетов приведены в табл. 8.

Таблица 8

Процент фертильных проб	10	20	30	40	50	60	70	80	90
Среднее число организмов	0,1054	0,2231	0,3567	0,5108	0,6932	0,9163	1,2040	1,6094	2,3026

Пользуясь приведенной выше таблицей, нельзя основываться на том, что при данном числе изучаемых проб наибольшая точность отношения числа фертильных проб к числу стерильных будет достигнута при 50% фертильных, так как здесь речь идет не об оценке этого отношения, а о минимальной ошибке исчисления числа организмов, которое определяется с наибольшей точностью около 80% фертильных проб, или около 1,6 организмов в среднем на пробу. В этом последнем случае для получения средней квадратической ошибки в 10% достаточно иметь около 155 проб, в то время как для получения той же точности при 50% фертильных проб необходимо взять уже 208 проб (см. «Планирование опыта» — Design of Experiments, параграф 68).

Ряды Пуассона позволяют также определить, какой процент полученных фертильных культур возник от единичных организмов, так как процент нечистых культур, т. е. таких, которые образовались от размножения двух и более организмов, может быть определен из процентного состава фертильных культур. Если e^{-m} — стерильные, то me^{-m} — чистые культуры, а остальные будут нечистыми культурами. Следующая таблица дает процент фертильных культур и процент фертильных культур, относящихся к нечистым:

Таблица 9

Среднее число организмов в пробе	0,1	0,2	0,3	0,4	0,5	0,6	0,7
Процент фертильных культур	9,52	18,13	25,92	32,97	39,35	45,12	50,34
Процент фертильных нечистых культур . .	4,92	9,67	14,25	18,67	22,92	27,02	30,95

Для того чтобы чистые культуры получались бы с высокой степенью уверенности, необходимо брать столь малые концентрации, чтобы по крайней мере девять десятых проб были стерильными.

18. Биноминальное распределение

Первым из теоретически установленных распределений было биноминальное распределение; оно было найдено Бернулли в конце XVII в. Это распределение возникает в том случае, когда производится n случайных испытаний, в которых вероятность появления некоторого события равна p, а вероятность его непоявления $q=1-p$. Теоретически частоты, с которыми данное событие в n испытаниях появляется 0, 1, 2, ... n раз, равны соответствующим членам разложения бинома

$$(q+p)^n.$$

Этот закон распределения является частным случаем более общего закона, в котором рассматривается не просто альтернатива — появление или непоявление события, — а случай, когда само событие может иметь s различных значений или форм, появлению которых соответствуют вероятности $p_1, p_2, \ldots p_s$. Можно показать, что в этом случае вероятность появления в испытаниях первой формы a_1 раз, второй формы a_2 раз и т. д. равна

$$\frac{n!}{a_1!\,a_2!\ldots a_s!}\,p_1^{a_1}p_2^{a_2}\ldots p_s^{a_s}.$$

Это выражение является общим членом разложения полинома

$$(p_1+p_2+\ldots+p_s)^n.$$

Пример 5. *Биноминальное распределение, полученное в результате подсчета очков на игральных костях.* При выбрасывании идеально правильной игральной кости вероятность выпадения более 4 очков равна $1/3$. Если выбрасывается сразу 12 костей, то выпадениям 5 или 6 очков будут соответствовать теоретические частоты, определяемые членами разложения

$$\left(\frac{2}{3}+\frac{1}{3}\right)^{12}.$$

Если одна или несколько из этих костей не будут правильными и если все эти неправильности остаются неизменными в течение всего эксперимента, то соответствующие частоты могут быть приближенно определены из разложения бинома

$$(q + p)^{12},$$

где p —вероятность, определяемая по фактическим результатам. В процессе 26 306 выбрасываний 12 игральных костей были найдены следующие частоты (данные Уэлдона):

Таблица 10

Число костей с 5 и 6 очками	Наблюденные частоты	Ожидаемые частоты на правильных костях	Ожидаемые частоты на неправильных костях	Значения $\frac{x^2}{m}$	
				правильные кости	неправильные кости
0	185	202,75	187,38	1,554	0,030
1	1 149	1216,50	1146,51	3,745	0,005
2	3 265	3345,37	3215,24	1,931	0,770
3	5 475	5575,61	5464,70	1,815	0,019
4	6 114	6272,56	6269,35	4,008	3,849
5	5 194	5018,05	5114,65	6,169	1,231
6	3 067	2927,20	3042,54	6,677	0,197
7	1 331	1254,51	1329,73	4,664	0,001
8	403	392,04	423,76	0,306	1,017
9	105	87,12	96,03	3,670	0,838
10	14	13,07	14,69		
11	4	1,19	1,36	0,952	0,222
12	—	0 05	0,06		
	26 306	26306,02	26306,00	35,491 $n = 10$	8,179 $n = 9$

Из этой таблицы следует, что наблюдения не согласуются с допущением о правильности костей. При правильных костях число случаев для 0, 1, 2, 3 и 4 должно быть большим, чем наблюдалось фактически, а для 5, 6, ... 11 очков — меньшим. Этот вывод подтверждается пятой графой, где приведены отношения $\frac{x^2}{m}$, в которых m — ожидаемые частоты, а x — разность между ожидаемыми и фактическими частотами. Суммируя эти отношения, получаем величину χ^2, которая измеряет отклонение наблюденного ряда данных от ожидаемого в целом. Значение χ^2 превосходит 35,49, откуда следует, что гипотеза правильности костей имеет вероятность, равную только 0,001 (см. параграф 20).

Общее число появления 5 или 6 очков фактически составляет 106 602 из 315 672 всех выпадений, а ожидаемое число их появления равно 105 224. На основе первых двух цифр можно вычислить вероятность p; она равна 0,3376986. Результаты расчета ожидаемых частот при этом значении p приведены в четвертой

графе табл. 10. Эти величины значительно более близки к фактическому ряду частот, откуда следует, что в данном случае условия эксперимента таковы, что хотя и образуется здесь биноминальное распределение, но оно имеет p, отличающееся от $^1/_3$.

Дисперсия биноминального распределения равна pqn. Поэтому при правильных костях и при 315 672 испытаниях, когда ожидаемое число выпадений 5 или 6 очков равно 105 224, дисперсия будет равна 70149,3, а средняя квадратическая ошибка будет 264,9. Фактическое число этих выпадений на 1378 превосходит ожидаемое, что в 5,20 раз больше его средней квадратической ошибки. В данном случае применен наиболее точный метод оценки, имеющий здесь вполне законную силу, так как при такой большой выборке биноминальное распределение весьма близко к нормальному. Из таблицы интеграла вероятностей видно, что отклонение, превосходящее свою среднюю квадратическую ошибку в 5,2 раза, может встретиться только один раз в 5 миллионах испытаний.

Тот факт, что этот последний критерий дает более жесткую оценку, чем критерий χ^2, объясняется тем, что при помощи распределения χ^2 оценивается система отклонений, в то время как фактически мы имеем дело с единичным отклонением p от $^1/_3$, т. е. с условием более общего характера. Когда производится непосредственная оценка этого отклонения, то его существенность обнаруживается легче и значительно раньше становится очевидной.

Пример 6. *Соотношение численностей полов и биноминальное распределение.* Некоторые биологические наблюдения во многих отношениях аналогичны с рассмотренным выше экспериментом с игральными костями. В качестве примера можно взять данные Гейслера о соотношении полов в германских семьях. Известно, что рождение ребенка мужского пола встречается несколько чаще, чем рождение ребенка женского пола. Поэтому, если рассматривать семью с 8 детьми в качестве выборки из генеральной совокупности, то число мальчиков в таких семьях должно быть распределено согласно разложению бинома

$$(q + p)^8,$$

где p — доля мальчиков в генеральной совокупности. Если, однако, различие семей в этом отношении обусловлено не только случайными моментами, но в какой-либо степени связано с тенденцией появления у некоторых родителей только мальчиков или только девочек, то распределение численности мальчиков в семье должно отклоняться от общего соотношения полов так, что будет образовываться некоторый недостаток семей с равным или почти равным делением по признаку пола. Табл. 11 показывает, что действительно наблюдается такое превышение над нормой количества семей с резко выраженным делением на неравные части.

Приведенный в этой таблице ряд фактических частот заметно отличается от ряда теоретически ожидаемых частот, причем это

различие характерно в двух отношениях: во-первых, оно состоит в преувеличенном числе семей, делящихся на неравные части, и, во-вторых, в отклонениях частот центральной части распределения в сторону, благоприятную для четных вариантов. Для последнего факта нельзя найти никакого биологического объяснения и поэтому мы не будем его принимать во внимание. Отклонения для крайних типов семей могут быть более детально изучены путем сопоставления наблюденной и ожидаемой дисперсии. Ожидаемая дисперсия $npq = 1{,}99828$, в то время как вычисленная по фактическим данным равна 2,06745, что дает превышение на 0,06917, или на 3,46%. Выборочная дисперсия этой оценки для наблюденной дисперсии определяется выражением (см. стр. 67).

$$\frac{2\varkappa_2^2}{N-1}+\frac{\varkappa_4}{N},$$

где N число семей, а $\varkappa_2$ и $\varkappa_4$ — второй и четвертый кумулянты теоретического распределения, которые в нашем случае равны

$$\varkappa_2 = npq = 1{,}99828;$$

$$\varkappa_4 = npq\,(1-6pq) = -0{,}99656.$$

Эти показатели вычислены по значению p, равному доли мальчиков в данной выборке. Средняя квадратическая ошибка этой дисперсии, которая, как легко видеть, близка к $\sqrt{7/N}$, равна 0,01141. Следовательно, отклонение наблюденной дисперсии от дисперсии биноминального распределения превосходит свою ошибку в 6 раз.

Таблица 11

Число мальчиков в семье	Фактическое число семей	Ожидаемое число семей	Отклонение x	$\frac{x^2}{m}$
0	215	165,22	+49,78	14,998
1	1 485	1401,69	+83,31	4,952
2	5 331	5202,65	+128,35	3,166
3	10 649	11034,65	−385,65	13,478
4	14 959	14627,60	+331,40	7,508
5	11 929	12409,87	−480,87	18,633
6	6 678	6580,24	+97,76	1,452
7	2 092	1993,78	+98,22	4,839
8	342	264,30	+77,70	22,843
	53 680	53680,00		91,869

Одна из возможных причин такой несколько увеличенной дисперсии состоит в том, что при рождении близнецов существует тенденция к появлению детей одного и того же пола. В данном случае рождения близнецов не выделены из общего числа рождений, но некоторое представление о влиянии этого фактора можно получить из других данных, относящихся к Германии. Установлено, что двойни рождаются примерно у 12 женщин из тысячи, причем доля одинаковых полов составляет $^5/_8$, а доля разных полов — $^3/_8$. Следовательно, около четвертой части рождений двоен, т. е. 3 случая на тысячу, может считаться количественным эффектом указанной выше тенденции. Это даст примерно 6 детей на тысячу. Подобное же влияние рождения троен и т. д. будет столь незначительным, что его можно не принимать во внимание. Если бы мы имели дело с совокупностью только таких «несбалансированных» рождений двоен одного и того же пола, то, как легко догадаться, теоретическая дисперсия этой совокупности была бы вдвое большей. Поэтому, если объяснять определенное выше увеличение дисперсии на 3,46% наличием «несбалансированных» двоен, то в исследуемой здесь совокупности семей следовало бы ожидать 3,46% таких двоен, что больше чем в 5 раз превосходит фактическую долю их (6 на 1000, или 0,6%). И хотя следует ожидать, что в больших семьях с 8 детьми доля двоен большая, чем во всей совокупности, однако и этот факт не столь значителен, чтобы он мог объяснить увеличение дисперсии только наличием двоен.

19. Малые выборки из биноминального распределения

Когда приходится иметь дело с малыми выборками, обычно встречающимися в экспериментальных исследованиях, то по одной такой выборке не представляется возможным судить с достаточной точностью о степени соответствия фактического распределения с биноминальным законом. Но в этом случае все же возможно проверить, в какой мере изменчивость данного материала близка к той, которую следует ожидать теоретически. Эта проверка может быть проведена при помощи критерия χ^2 по образцу той, которая применялась выше в отношении распределения Пуассона.

Пример 7. *Точность определения доли зараженных семян.* Доля зараженных мухой-зеленоглазкой зерен ячменя определяется путем отбора 100 зерен и подсчета числа зараженных. При повторных пробах получаются различающиеся доли этих зерен. Если материал однороден, то эта доля должна быть распределена по биноминальному закону

$$(q+p)^{100},$$

где p — доля зараженных семян во всей совокупности и q — число здоровых зерен. Следующие данные являются результатом 10 таких наблюдений, относящихся к одной и той же опытной делянке (данные Фру):

16, 18, 11, 18, 21, 10, 20, 18, 17, 21; средняя 17,0.

Спрашивается: можно ли эту изменчивость считать результатом случайного отбора, т. е. можно ли считать материал одно-

родным? Эти данные отличаются от данных, сравниваемых с распределением Пуассона, в том отношении, что каждые отобранные 100 зерен подразделяются только на два класса: зараженные и незараженные; поэтому при изучении изменчивости числа зараженных зерен вместе и одновременно с этим изучается изменчивость численности и другого класса незараженных зерен. Видоизменение показателя рассеяния данных χ^2, соответствующее биномиальному распределению, будет выражаться формулой

$$\chi^2 = \frac{S(x-\bar{x})^2}{npq} = \frac{S(x-\bar{x})^2}{\bar{x}q}$$

и отличаться от формы этого критерия, относящейся к распределению Пуассона, наличием добавочного делителя q, который в нашем случае равен 0,83. В данном случае находим $\chi^2=9{,}21$; это значение, как видно из таблицы χ^2 при $n=9$ ($=10-1$, т. е. на единицу меньше, чем число наблюдений), является допустимым при гипотезе об однородности материала.

Такое исследование результатов единичной выборки не может быть основой для более или менее достоверных выводов, так как сам критерий χ^2 в этом случае изменяется в довольно широких размерах. Однако, если имеется несколько таких малых выборок, хотя бы и полученных с делянок с различной степенью зараженности, то можно, как это имело место и в случае распределения Пуассона, сравнить общую фактическую изменчивость с той, которая должна существовать при биномиальном распределении. Так, при обследовании 20 опытных делянок получено $\chi^2=193{,}64$ при $S(n)=180$. Производя те же вычисления, что и на стр. 55, находим:

$\sqrt{387{,}28}$	$=19{,}68$
$\sqrt{359}$	$=18{,}95$
Разность	$=+0{,}73$

Эта разность меньше единицы, и поэтому мы делаем вывод, что в данном случае фактическая изменчивость в целом не отклоняется существенно от той, которую следует ожидать, исходя из биномиального распределения. Различие между методом, примененным в данном случае, когда сама выборка мала (10 наблюдений), но когда каждое значение доли признака получено на основе довольно большого числа наблюдений (100 зерен), и предыдущим случаем, когда изучалось распределение детей по полу и когда у нас было много семей, но имелось только 8 наблюдений внутри каждой семьи, сводится к исключению члена, входящего в среднюю квадратическую ошибку и связанного с величиной

$$\varkappa_4 = npq(1-6pq).$$

Когда $n=100$, эта величина очень мала по сравнению с $2n^2p^2q^2$ и поэтому оценка через χ^2 становится достаточно точной и при столь малом объеме выборки, как 10.

ПРИЛОЖЕНИЕ

ОБОЗНАЧЕНИЯ И ФОРМУЛЫ

А. Статистики, получаемые путем возведения в степень и суммирования

Пусть n — число наблюденных значений переменной x и требуется определить суммы целых степеней этих значений; в этом случае мы будем записывать:

$$s_1 = S(x); \quad s_2 = S(x^2);$$
$$s_3 = S(x^3); \quad s_4 = S(x^4)$$

и т. д. По этим данным можно вычислить суммы степеней для отклонений от средней, используя такие уравнения:

$$S_2 = s_2 - \frac{1}{n}s_1^2;$$

$$S_3 = s_3 - \frac{3}{n}s_2s_1 + \frac{2}{n^2}s_1^3;$$

$$S_4 = s_4 - \frac{4}{n}s_3s_1 + \frac{6}{n^2}s_2s_1^2 - \frac{3}{n^3}s_1^4.$$

Целый ряд статистических оценок строится на основе именно этих величин:

1. Моменты относительно условного начала $x=0$; эти показатели получаются простым делением соответствующих сумм s на численность выборки. Если p принимает значения 1, 2, 3, 4 ....., то эти моменты определяются формулой

$$m_p' = \frac{1}{n}s_p.$$

Очевидно, что m_1' является средней арифметической, обычно обозначаемой $\bar{x}$.

2. Моменты, не зависящие от выбранного условного начала и имеющие более близкое отношение к характерным особенностям выборочной совокупности, называются «моментами относительной средней», или «центральными моментами». Эти величины находят широкое применение в статистике и определяются делением сумм S отклонений от средней, возведенных в соответствующие степени, на численность выборки. Так, при $p=2, 3, 4\ldots$

$$m_p = \frac{1}{n}S_p.$$

Этим путем вычисляются моменты, которые были бы получены, если средняя арифметическая была бы взята за условное начало отсчета значений x, что арифметически обычно не совсем удобно.

3. В настоящее время применяется более совершенная система показателей, которая имеет большое теоретическое значение и которая получается объединением средней и моментов относительно средней в единую систему так называемых k-статистик:

$$k_1 = \frac{1}{n} s_1;$$

$$k_2 = \frac{1}{n-1} S_2;$$

$$k_3 = \frac{n}{(n-1)(n-2)} S_3;$$

$$k_4 = \frac{n}{(n-1)(n-2)(n-3)} \left\{(n+1) S_4 - 3\frac{n-1}{n} S_2^2\right\}.$$

Легко установить следующие соотношения, определяющие моменты m через k-статистики:

$$m_1' = k_1;$$

$$m_2 = \frac{n-1}{n} k_2;$$

$$m_3 = \frac{(n-1)(n-2)}{n^2} k_3;$$

$$m_4 = \frac{n-1}{n^2(n+1)} \{(n-2)(n-3) k_4 + 3(n-1)^2 k_2^2\}.$$

4. В настоящее время статистики, введенные Тиле и названные семи-инвариантами, представляют только исторический интерес. Эти показатели образуются из моментов m' и m точно так же, как кумулянты (см. ниже раздел Б) определяются через характеристики генеральной совокупности μ' и μ. Так, если h_1, h_2, h_3, ... обозначает ряд последовательных семи-инвариантов, то можно написать:

$$h_1 = m_1'; \quad h_2 = m_2; \quad h_3 = m_3;$$
$$h_4 = m_4 - 3m_2^2; \quad h_5 = m_5 - 10m_3 m_2$$

и т. д. Тиле применял одно и то же наименование «семи-инварианты» как для обозначения параметров генеральной совокупности, так и для статистик, которые являются только оценками этих первых, подобно тому как К. Пирсон и его последователи применяли термин «моменты» также в этих двух смыслах. Благодаря этому смешению понятий кумулянты часто рассматриваются как семи-инварианты генеральной совокупности, и даже k-статистики иногда неправильно называются выборочными семи-инвариантами. Семи-инварианты, впервые введенные Тиле, в настоящее время потеряли свое значение, и о них мы говорим здесь только для того, чтобы избежать неопределенности в терминах.

Б. Моменты и кумулянты теоретических распределений

Каждая из перечисленных выше систем статистик, полученная на основе сумм степеней x, может рассматриваться как система оценок соответствующих параметров теоретических распределений, к которым эти оценки обычно стремятся по мере беспредельного увеличения объема выборки. Эти истинные или генеральные значения параметров будут обозначаться нами греческими буквами; так, m_4' является оценкой четвертого момента относительно условного начала в генеральной совокупности μ_4'; m_4 является оценкой четвертого центрального момента μ_4; k_4 является оценкой $\varkappa_4$ — четвертого кумулянта генеральной совокупности. Соотношения между этими параметрами генеральной совокупности подобны соотношениям между выборочными m и k, т. е.

$$\mu_1' = \varkappa_1; \quad \mu_2 = \varkappa_2; \quad \mu_3 = \varkappa_3;$$
$$\mu_4 = \varkappa_4 + 3\varkappa_2^2; \quad \mu_5 = \varkappa_5 + 10\varkappa_3\varkappa_2$$

и т. д. Общее правило, по которому образуются числовые коэффициенты, входящие в эти равенства, сводится к следующему: коэффициент 3 является числом способов, которыми четыре элемента могут быть подразделены на две группы по два элемента в каждой группе; коэффициент 10 является числом способов, которыми пять элементов могут быть подразделены на две части соответственно с двумя и тремя элементами.

Что же касается соотношения между оценками и соответствующими им параметрами, то, оставаясь в плоскости элементарного изложения, отметим только, что в то время как средние значения m', полученные по выборкам объема n, и средние значения $\varkappa$, полученные тем же порядком, равны соответственно μ' и $\varkappa$, центральные моменты m_2, m_3, m_4 и т. д. уже не обладают этим свойством, ибо

$$\overline{m}_2 = \frac{n-1}{n} \mu_2;$$

$$\overline{m}_3 = \frac{(n-1)(n-2)}{n^2} \mu_3;$$

$$\overline{m}_4 = \frac{n-1}{n^3} \{(n^2 - 3n + 3) \mu_4 + 3(2n-3) \mu_2^2\}.$$

Этот ряд формул с достаточной очевидностью показывает практические неудобства применения центральных моментов и те более сложные по сравнению с k-статистиками алгебраические выкладки, с которыми связано их использование.

Теми же недостатками обладают и семи-инварианты Тиле; хотя они и могут рассматриваться в качестве оценок кумулянтов $\varkappa$, однако их средние значения, исчисленные по данным

конечных выборок, не равны соответствующим значениям $\varkappa$. Действительно, формулы:

$$\overline{h_2} = \frac{n-1}{n}\,\varkappa_2;$$

$$\overline{h_3} = \frac{(n-1)(n-2)}{n^2}\,\varkappa_3;$$

$$\overline{h_4} = \frac{n-1}{n^3}\left\{(n^2-6n+6)\,\varkappa_4 - 6n\varkappa_2^2\right\}$$

показывают, что вычисление более высоких членов этого ряда сопряжено с тем же самым затруднением, что и вычисление центральных моментов.

В нижеприведенной таблице даны четыре кумулянта для трех рассматриваемых в этой главе распределений, причем кумулянты здесь выражены через параметры соответствующих распределений:

	Символ	Распределение		
		Нормальное	Пуассона	Биноминальное
Средняя	$\varkappa_1$	μ	m	np
Дисперсия	$\varkappa_2$	σ^2	m	npq
Третий кумулянт .	$\varkappa_3$	0	m	$-npq\,(p-q)$
Четвертый кумулянт	$\varkappa_4$	0	m	$npq\,(1-6pq)$

В. Выборочная дисперсия статистик, определенных по выборке численностью N

Знание выборочных дисперсий статистик прежде всего необходимо для оценки существенности этих показателей. В отношении суммы степеней, если степень больше двух, этот вопрос исследован только для нормального распределения. Отклонение распределения от нормальной формы определяется при помощи статистик g_1 и g_2, построенных на основе третьих и четвертых степеней x:

$$g_1 = k_3 / k_2^{3/2};\quad g_2 = k_4 / k_2^2.$$

Следует заметить, что в первых трех изданиях настоящей книги для обозначения этих статистик были употреблены символы γ_1 и γ_2, но греческие буквы лучше применять не для статистик, а для параметров, которые они оценивают; поэтому мы

вводим обозначения g_1 и g_2. Выборочные дисперсии k-статистик приводятся ниже:

Дисперсия величины	Общая формула	Нормальное распределение
k_1	$\frac{\varkappa_2}{N}$	$\frac{\sigma^2}{N}$
k_2	$\frac{\varkappa_4}{N} - \frac{2\varkappa_2^2}{N-1}$	$\frac{2\sigma^4}{N-1}$
g_1	—	$\frac{6N(N-1)}{(N-2)(N+1)(N+3)}$
g_2	—	$\frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}$

Г. Поправки на группировку данных

В тех случаях, когда суммы степеней x вычислены по сгруппированным данным, могут быть введены поправки, устраняющие возможное влияние таких группировок. Поправки этого рода для моментов, называемые поправками Шеппарда, вводятся только в суммы четных степеней отклонений x от средней. Если интервал группировки принят за единицу, то, обозначая исправленные значения второй и четвертой k-статистик через k_2' и k_4', можно написать:

$$k_2' = k_2 - \frac{1}{12};\quad k_4' = k_4 + \frac{1}{120}.$$

Эти поправки можно использовать только для оценки, но не для определения существенности показателей. Так, k_2' будет лучшей оценкой дисперсии, чем k_2, но выборочная дисперсия или средняя квадратическая ошибка как средней, так и дисперсии должна вычисляться по неисправленному значению k_2.

ТАБЛИЦА I

Таблица x — отклонений в нормальном распределении, выраженных в долях среднего квадратического отклонения

	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
0,00	2,575829	2,326348	2,170090	2,053749	1,959964	1,880794	1,811911	1,750686	1,695398	1,644854
0,10	1,598193	1,554774	1,514102	1,475791	1,439521	1,405072	1,372204	1,340755	1,310579	1,281552
0,20	1,253565	1,226528	1,200359	1,174987	1,150349	1,126391	1,103063	1,080319	1,058122	1,036433
0,30	1,015222	0,994458	0,974114	0,954165	0,934589	0,915365	0,896473	0,877896	0,859617	0,841621
0,40	0,823894	0,806421	0,789192	0,772193	0,755415	0,738847	0,722479	0,706303	0,690309	0,674490
0,50	0,658838	0,643345	0,628006	0,612813	0,597760	0,582841	0,568051	0,553385	0,538836	0,524401
0,60	0,510073	0,495850	0,481727	0,467699	0,453762	0,439913	0,426148	0,412463	0,398855	0,385320
0,70	0,371856	0,358459	0,345125	0,331853	0,318639	0,305481	0,292375	0,279319	0,266311	0,253347
0,80	0,240426	0,227545	0,214702	0,201893	0,189118	0,176374	0,163658	0,150969	0,138304	0,125661
0,90	0,113039	0,100434	0,087845	0,075270	0,062707	0,050154	0,037608	0,025069	0,012533	0

Значение P для каждой клетки находится путем суммирования цифры, стоящей слева, и цифры, стоящей в заголовке. Соответствующее значение x является таким отклонением, что вероятность наблюдения выйти из интервала от $-x$ до $+x$ равна P. Например, $P = 0{,}03$ для $x = 2{,}170090$; следовательно, 3% значений переменной будут иметь положительные или отрицательные отклонения, превосходящие среднее квадратическое отклонение в 2,170090 раза.

ТАБЛИЦА II

Значения x для малых значений P

P	0,001	0,0001	0,00001	0,000001	0,0000001	0,00000001	0,000000001
x	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

ГЛАВА ЧЕТВЕРТАЯ

КРИТЕРИИ СОГЛАСИЯ, НЕЗАВИСИМОСТИ И ОДНОРОДНОСТИ, ОСНОВАННЫЕ НА РАСПРЕДЕЛЕНИИ χ^2

20. Распределение χ^2

В предыдущей главе мы познакомились с некоторыми приложениями распределения χ^2 при изучении согласия наблюдений с определенной гипотезой. В настоящей главе мы рассмотрим в более общем виде весьма широкий круг проблем, которые решаются при помощи этого распределения.

В основе всех этих исследований лежит сопоставление фактически наблюденных численностей, относящихся к некоторому числу классов, с численностями, ожидаемыми в соответствии с проверяемой гипотезой. Если m — ожидаемая численность и $m + x$ — фактическая численность некоторого класса, то мы можем вычислить

$$\chi^2 = S\left(\frac{x^2}{m}\right),$$

где суммирование производится по всем классам. Очевидно, что чем больше согласованность фактических численностей с ожидаемыми, тем меньшим должно быть значение χ^2. Но для определения вероятности, соответствующей данному значению χ^2, при помощи специально составленной таблицы χ^2 необходимо знать еще значение n, с которым эта вероятность связана. Общее правило для определения n состоит в том, что n является числом степеней свободы, на которое фактический ряд данных может отличаться от гипотетического ряда. Другими словами, n равно числу классов, частоты которых могут приниматься любые произвольные значения, не связанные с наблюденными частотами. Последующие примеры дадут конкретные иллюстрации этого правила.

Распределение χ^2_n для некоторых значений n, которые, как ясно из предыдущего, должны быть целыми числами, было определено Пирсоном в 1900 г. Оно дает возможность определить ту

долю случаев, в которых наблюденное значение χ^2 может превзойти табличное значение χ^2. Эта доля определяется величиной P, которая является вероятностью того, что χ^2 будет превосходить некоторое заданное значение. Таким образом, каждому значению χ^2 соответствует определенное значение P; когда χ^2 возрастает от нуля до бесконечности, P убывает от 1 до 0. Равным образом, определенному значению P в указанных выше пределах соответствует некоторое значение χ^2. Соотношение между двумя этими величинами довольно сложно, и поэтому для практического применения критерия необходимо иметь соответствующую таблицу.

Таблица этого рода была составлена Эльдертоном и известна под названием: «Таблица для критерия согласия Эльдертона». Эльдертон дал в ней значения P с шестью десятичными знаками; эти P соответствуют целым значениям χ^2 от 1 до 30 и от 30 до 70 через десять единиц. Вместо n в этой таблице было взято $n'=n+1$, так как тогда считали, что этот параметр должен равняться числу классов. Таблица содержит в себе n' от 3 до 30, что соответствует n от 2 до 29. В дальнейшем Юл дал дополнительную таблицу для $n'=2$ или $n=1$. В связи с ограничениями, налагаемыми авторским правом, мы не можем здесь привести эту таблицу Эльдертона, и поэтому вместо нее даем здесь новую таблицу (табл. III на стр. 96) в такой форме, которая, как показывает опыт, более удобна для практического употребления. Вместо того, чтобы давать значения P, соответствующие некоторому ряду значений χ^2, мы даем значения χ^2, соответствующие некоторым специально выбранным вероятностям P. Тем самым мы имеем возможность включить в компактной форме и те части распределений χ^2, которые в прежних таблицах не нашли своего отражения, а именно: в нашей таблице имеются значения χ^2, меньшие единицы, которые довольно часто встречаются при малых значениях n, а также значения χ^2, большие 30, которые обычно встречаются при больших n.

Интересно отметить, что показатель рассеяния Q, введенный в статистику германским экономистом Лексисом, если его правильно рассчитать, равнозначен величине $\frac{\chi^2}{n}$ в наших обозначениях. Мне кажется, что в английской статистической литературе, описывающей метод Лексиса, не был замечен тот факт, что открытие распределения χ^2, по существу, явилось завершением этого метода. Если возникнет потребность перехода к терминам Лексиса, то легко преобразовать нашу таблицу в таблицу показателя Q путем простого деления χ^2 на n.

Составляя нашу таблицу, мы имели в виду, что на практике не столь важно знать точное значение вероятности P, соответствующее данному значению χ^2, как важно определить, в какой мере достоверно наблюденное значение χ^2. Если вероятность P содержится в широком промежутке от 0,1 до 0,9, то у нас не будет

никаких оснований сомневаться в проверяемой гипотезе; если же вероятность P становится, например, ниже 0,02, то это прямо указывает на несостоятельность данной гипотезы. Риск впасть в ошибку не будет слишком большим, если мы проведем пограничную линию у $P=0{,}05$ и будем считать, что значения χ^2, лежащие выше этой линии, указывают на наличие существенных и реальных отклонений.

Оценка существенности χ^2 путем деления ее на «вероятную ошибку» и исчисления по таблице интеграла вероятностей P является ничем иным, как заменой точного распределения χ^2, определяемого нашей таблицей, на неточное (нормальное) распределение.

Применяемый здесь термин «критерий согласия» может иногда служить источником заблуждений, так как он создает впечатление, что наиболее высокие значения P являются наилучшим основанием для принятия данной гипотезы. Между тем вероятность, например, 0,999 будет указывать на то, что значения, равные или меньшие соответствующего χ^2, при правильной гипотезе могут встретиться только один раз на тысячу испытаний. Вообще этот случай указывает на то, что мы использовали для гипотезы неправильную формулу. Однако могут встретиться и случаи, когда имеется ряд чрезвычайно малых значений χ^2, не связанных с какой-либо расчетной формулой, как, допустим, в примере 4 при подсчете бактериальных колоний. В таких случаях рассматриваемая гипотеза будет отвергнута как бы на уровне 0,001.

В тех случаях, когда вычислен довольно большой ряд величин χ^2, имеется возможность обнаружить небольшие отклонения, остающиеся неуловимыми при единичном значении χ^2, так как в этих случаях предоставляется возможным сравнить фактическое распределение χ^2 с теоретически ожидаемым распределением. Для этого достаточно распределить фактические значения χ^2 по тем интервалам, которые даны в таблице χ^2; этот прием применен в примере 4 на стр. 54. Ожидаемые же теоретически частоты для этих интервалов легко определить по значениям P; также, если в этом есть необходимость, можно применить тот же критерий χ^2 для установления степени согласия наблюденных частот с этими теоретическими частотами.

Здесь полезно напомнить, что сумма нескольких величин χ^2 сама распределена по закону распределения χ^2 с параметром n, равным сумме значений n, относящихся к отдельным величинам χ^2. Такой суммарный критерий, естественно, более чувствителен и, пользуясь им, можно обнаружить такие отклонения от нормы, которые при отдельных значениях χ^2 кажутся незначительными или не вполне определенными.

Таблица, которая дана в нашей книге, содержит значения n только до 30; оценку же при больших значениях n можно с достаточным основанием строить на допущении, что величина $\sqrt{2\chi^2}$

распределена нормально около центра $\sqrt{2n-1}$ со средним квадратическим отклонением, равным единице. В порядке проверки этого положения читатель может определить этим способом вероятности P для значений χ^2 при $n=30$ и сравнить эти величины P с табличными значениями. Такое сравнение показывает, что соответствующие погрешности при $n=30$ уже незначительны; они будут еще более уменьшаться по мере возрастания n.

Пример 8. *Сравнение фактических частот с частотами, ожидаемыми согласно закону Менделя.* При скрещивании двух менделевских факторов, если они независимы друг от друга и если они дают одинаково жизнеспособное потомство, в четырех классах гибридов ожидаются частоты, подчиненные соотношению 9 : 3 : 3 : 1. Рассмотрим, находятся ли в согласии с этой гипотезой следующие наблюдения над Primula (данные Уинтона и Бэтсона):

Таблица 12

Частоты	Плоские листья		Сморщенные листья		Итого
	нормальный глазок	бледно-розовый глазок	нормальный глазок	бледно-розовый глазок	
Фактические $(m+x)$	328	122	77	33	560
Ожидаемые (m) .	315	105	105	35	560
$\frac{x^2}{m}$	0,537	2,752	7,467	0,114	10,870

Ожидаемые частоты здесь вычислены на основе общего итога фактических частот, так что общая сумма тех и других одинакова. Поэтому, если ожидаемые частоты трех любых классов могут быть какими угодно, частота четвертого класса может иметь только определенное значение, зависящее от трех первых. Следовательно, здесь $n=3$ и $\chi^2=10{,}87$, а соответствующая вероятность находится между 0,01 и 0,02. Если принять $P=0{,}05$ за демаркационную точку между существенными и несущественными отклонениями, то можно прийти к выводу, что в данном случае наблюдаются довольно существенные отклонения фактических данных от гипотетических.

Теперь рассмотрим другую гипотезу, отличающуюся от первой в том отношении, что теперь мы допустим меньшую жизнеспособность растений с сморщенными листьями. Конечно, полная проверка этой гипотезы требует дополнительных данных, но в нашем случае мы исследуем только частный вопрос: в какой мере эта гипотеза согласуется с имеющимися наблюдениями. Гипотеза, подлежащая проверке, ничего не говорит о том, каков размер уменьшения жизнеспособности у растений с сморщенными листьями; поэтому мы возьмем итоги для групп с плоскими и сморщенными листьями такими, какими они фактически получены, и разложим их на части в отношении 3 : 1.

Таблица 13

Частоты	Плоские листья		Сморщенные листья		χ^2
	нормальный глазок	бледно-розовый глазок	нормальный глазок	бледно-розовый глазок	
Фактические	328	122	77	33	—
Ожидаемые	337,5	112,5	82,5	27,5	—
$\frac{x^2}{m}$	0,267	0,802	0,367	1,100	2,536

Теперь n равно 2, так как только две частоты могут произвольно изменять свои значения; однако величина χ^2 теперь уменьшается столь значительно, что вероятность P увеличивается до 0,2, и, следовательно, отклонения фактических частот от теоретически ожидаемых теперь не могут считаться существенными. Таким образом, значительная доля отклонения от менделевского закона в данном случае обусловлена соотношением между численностями растений с плоскими и сморщенными листьями.

В недалеком прошлом считали, что для отыскания табличного значения χ^2 всегда следует брать n равным числу классов без единицы; однако эта точка зрения неправильна и ей противоречат приведенные выше расчеты. Если основываться на этой старой точке зрения, то любое усложнение гипотезы, такое, например, как введенное выше допущение о различной жизнеспособности потомства, обязательно даст повышение согласованности между наблюдением и гипотезой. Если же принять во внимание изменение величины n, связанное с усложнением гипотезы, то улучшение согласованности становится необязательным. Поэтому повышение правильно рассчитанной вероятности P, как это наблюдалось в последнем примере, дает полное основание считать это результатом улучшения гипотезы, а не следствием простого увеличения числа параметров, рассчитанных по наблюденным данным.

Пример 9. *Сравнение фактических частот с частотами, ожидаемыми согласно закону Пуассона и биномиального распределения.* В табл. 5 на стр. 52 были приведены фактические частоты, ожидаемые при распределении Пуассона. Так как распределение χ^2, вычисленное при наличии в некоторых классах слишком малых частот, не является точным, то при расчете критерия χ^2 рекомендуется производить группировку материала так, чтобы численности в отдельных классах не были меньше 5. Поэтому, объединяя в табл. 5 численности для классов 0 и 1 клетка

и численности для классов 10 и более клеток, мы получим следующие данные:

Таблица 14

Частоты	0 и 1	2	3	4	5	6	7	8	9	10 и более	Итого
Фактические . .	20	43	53	86	70	54	37	18	10	9	400
Ожидаемые	21,08	40,65	63,41	74,19	69,44	54,16	36,21	21,18	11,02	8,66	400
$\frac{x^2}{m}$	0,055	0,136	1,709	1,880	0,005	0,000	0,017	0,477	0,093	0,013	4,385

Используя эти 10 классов, мы находим $\chi^2 = 4{,}385$. При определении значения n необходимо учесть, что ожидаемые частоты рассчитаны не только при неизменном общем итоге (400), но и так, чтобы их средняя совпадала бы с фактической средней. Следовательно, здесь остается 8 степеней свободы и поэтому в данном случае $n = 8$. В таблице распределения χ^2 для этого значения χ^2 находим P между 0,8 и 0,9, что указывает на хорошее согласие между фактическими и ожидаемыми частотами.

В табл. 10 (стр. 58) было получено значение χ^2 для 11 классов при двух гипотезах, а именно при допущении, что кости «правильные» и что они «неправильные». В первом случае ожидаемые частоты рассчитаны только на основе общей суммы фактических частот и поэтому здесь $n = 10$, но при допущении гипотезы о «неправильности» костей при расчете χ^2 была использована и фактическая средняя, вследствие чего n уменьшается еще на единицу и равно 9. В первом случае χ^2 даже выходит из рамок таблицы, что указывает на наличие весьма существенных отклонений фактических частот от теоретически ожидаемых; во втором же случае вероятность P находится между 0,5 и 0,7 и, следовательно, значение χ^2 не выходит из пределов допустимого.

21. Критерии независимости. Таблицы сопряженности признаков

Критерии *независимости* составляют специальный и весьма важный класс критериев, основанных на сопоставлении теоретически ожидаемых и фактических численностей. Когда некоторая группа объектов классифицируется по двум (или более) признакам, то возникает вопрос о том, в какой мере эти две классификации могут считаться не зависящими друг от друга. Примером такой двойной классификации может служить распределение людей по группам: получивших и не получивших некоторую предохранительную прививку, и одновременно с этим группировка тех же людей по группам: заболевших и не заболевших данной болезнью.

В простейшем случае, когда каждый признак имеет два класса, получается таблица 2×2, или, как ее часто называют, четырехклеточная таблица.

Пример 10. Таблица, относящаяся к заболеваниям сыпным тифом, взята у Гринвуда и Юла.

Фактические данные *Таблица 15*

	Заболевшие	Незаболевшие	Итого
Получившие прививку . .	56	6 759	6 815
Не получившие прививку .	272	11 396	11 668
Итого	328	18 155	18 483

Теоретически ожидаемые численности *Таблица 16*

	Заболевшие	Незаболевшие	Итого
Получившие прививку . .	120,94	6694,06	6 815
Не получившие прививку	207,06	11460,94	11 668
Итого	328	18155	18 483

Чтобы установить независимость этих двух признаков, т. е. установить, можно ли считать прививку бесполезной и не предохраняющей от заболевания, следует произвести сравнение фактических численностей с ожидаемыми, исчисленными на основе предположения о пропорциональном расслоении параллельных групп. Так как в данном случае изучается только независимость без каких-либо иных дополнительных гипотез относительно общего числа заболевших или получивших прививку, то «ожидаемые» численности по этим категориям в целом совпадают с фактическими численностями в итогах табл. 15. Следовательно, внутри таблицы ожидаемых численностей должно быть вычислено только одно значение, а именно:

$$\frac{328 \times 6815}{18483} = 120{,}94,$$

остальные же можно определить путем вычитания из соответствующих итогов. Отсюда следует, что фактические численности могут отличаться от ожидаемых только при одной степени свободы и поэтому параметр n при оценке независимости по таблице 2×2 будет всегда равен единице. Так как в данном случае $\chi^2 = 56{,}234$, то наши наблюдения явным образом противоречат гипотезе о независимости этих двух признаков. В таблице 2×2 χ^2 можно определить и без расчета отдельных ожидаемых численностей непосредственно по формуле:

$$\chi^2 = \frac{(ad - bc)^2 (a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)},$$

где a, b, c и d являются четырьмя фактическими численностями.

Вычисление χ^2 может быть упрощено и в случае, когда только один из признаков имеет две градации, а другой — большее число их и когда нет потребности знать отдельные ожидаемые численности. Если a и a' составляют пару фактических численностей, а n и n' являются соответствующими им итогами, то, следуя за Пирсоном, мы можем для каждой такой пары вычислить величину

$$\frac{1}{a+a'}(an' - a'n)^2.$$

Величина χ^2 будет являться суммой таких величин, деленных на nn'.

Другая формула, предложенная Брандтом и Снедекором, имеет то преимущество, что помимо того, что ускоряет расчеты, она тесно связана с широко применяемым методом дисперсионного анализа. В этом случае вычисляется для каждой пары численностей отношение

$$p = \frac{a}{(a+a')},$$

а по итогам величина

$$\bar{p} = \frac{n}{(n+n')},$$

после чего определяется

$$\chi^2 = \frac{1}{\bar{p}\,\bar{q}}\{S(ap) - n\bar{p}\},$$

где $\bar{q} = 1 - \bar{p}$. Другое преимущество этого способа состоит в том, что здесь имеется возможность наблюдать соотношение фактических численностей внутри каждого класса. Если наблюдается более или менее значительное различие между двумя рядами отношений $\left(\text{т. е. отношений } \frac{a}{(a+a')} \text{ и } \frac{a'}{(a+a')}\right)$, то рекомендуется строить расчеты на рядах с меньшими отношениями.

Пример 11. *Критерий независимости при классификации* $2 \times n'$. Шотландские дети, отдельно мальчики и девочки, были распределены по признаку цвета волос следующим образом (данные Токера):

Таблица 17

	Цвет волос					
	Белокурый	Рыжий	Русый	Черный	Очень черный	Итого
Мальчики	592	119	849	504	36	2100
Девочки	544	97	677	451	14	1783
Итого . .	1136	216	1526	955	50	3883
Отношения полов	0,52113	0,55093	0,55636	0,52775	0,72000	0,54082

В последней строке этой таблицы дано отношение полов, выраженное долей мальчиков. Умножая каждое из этих чисел на соответствующую численность мальчиков и вычитая из суммы таких произведений аналогичное произведение, исчисленное по последнему столбцу, получаем остаток 2,603, который при делении на $\bar{p}\bar{q}$ дает $\chi^2 = 10{,}48$.

В этой таблице произвольно и без нарушения итоговых данных можно выбрать четыре величины и поэтому здесь $n = 4$. Соответствующее значение P лежит между 0,02 и 0,05 и, следовательно, различие полов по признаку цвета волос, поскольку об этом можно судить по приведенным здесь данным, является довольно существенным. Следует заметить, что этот способ обработки данных требует вычисления p с достаточно высокой точностью. При определении этих отношений с пятью десятичными знаками величина χ^2 имеет погрешность уже во втором десятичном знаке; поэтому, во избежание недоразумений, рекомендуется вычисление проводить так, чтобы обеспечить точность не меньше, как с двумя десятичными знаками. Нетрудно заметить, что различие между полами в основном состоит в том, что относительно велика доля мальчиков в группе «очень черные».

Пример 12. *Критерий независимости в таблице* 4×4. В качестве примера более сложной таблицы сопряженности признаков рассмотрим данные о скрещивании мышей, когда учитываются два фактора: окраска — черная и бурая, и форма окраски — сплошная и пегая (данные Уоктера):

Таблица 18

	Черная сплошная	Черная пегая	Бурая сплошная	Бурая пегая	Итого
Сцепление:					
F_1 самцы . . .	88 (85,37)	82 (75,24)	75 (70,93)	60 (73,46)	305
F_1 самки . . .	38 (34,43)	34 (30,34)	30 (28,60)	21 (29,63)	123
Отталкивание:					
F_1 самцы . . .	115 (117,00)	93 (103,11)	80 (97,21)	130 (100,68)	418
F_1 самки . . .	96 (100,20)	88 (88,31)	95 (83,26)	79 (86,23)	358
Итого . .	337	297	280	290	1204

В данном случае скрещивание производилось четырьмя путями: во-первых, при условии гетерозиготности (F_1) в отношении обоих факторов родителей мужского и женского пола и, во-вторых, при получении двух доминантных генов только от одного (сцепление) или от каждого из родителей (отталкивание).

Простые соотношения Менделя здесь могут быть нарушены различиями в жизнеспособности потомства, сцеплением генов или наличием летального сцепления. Два последних момента в данном случае вряд ли могли иметь место и поэтому, если считаться

с различной жизнеспособностью четырех генотипов, то они во всех колонках должны встречаться в одинаковой пропорции. Чтобы установить, сколь существенны имеющиеся здесь отклонения от пропорциональности, можно воспользоваться критерием χ^2, вычислив его для таблицы 4×4 в целом. В табл. 18 в скобках даны ожидаемые частоты при гипотезе, что все четыре серии, представленные столбцами, внутри себя однородны. Отдельные слагаемые, образующие величину χ^2 и относящиеся к отдельным клеткам табл. 18, приведены в табл. 19.

Таблица 19

0,081	0,607	0,234	2,466	3,388
0,370	0,442	0,069	2,514	3,395
0,034	0,991	3,047	8,539	12,611
0,176	0,001	1,655	0,606	2,438
0,661	2,041	5,005	14,125	21,832

Таким образом, χ^2 равно 21,832; здесь $n=9$, так как при заданных итогах остаются для свободного выбора три ряда и три строки. Вообще для таблицы сопряженность признаков с r строками и c колонками число степеней свободы $n=(r-1)(c-1)$. Для $n=9$ значение $\chi^2=21{,}832$ соответствует P, меньшему, чем 0,01, и, следовательно, здесь отклонения от пропорциональности не являются случайными. Это нарушение пропорциональности, как легко увидеть, нашло свое выражение в слишком большой фактической численности при буро-пегой окраске в группе F_1 — самцы — отталкивание.

Следует заметить, что описываемые в настоящей главе методы не предназначены для измерения *степени* связанности между двумя классификациями, а дают только возможность определить, не имеют ли наблюдаемые отклонения от независимого действия факторов такой размер, который нельзя приписать случайности. Одна и та же степень зависимости может быть признана существенной при большой выборке и несущественной при малой выборке; если она несущественна, то при наличии имеющихся данных нет оснований считаться с ней и бесполезна любая попытка ее измерения. С другой стороны, если эта зависимость существенна, то величина χ^2 указывает только на наличие этого факта, но отнюдь не измеряет самою силу связи. Поскольку то или иное отклонение признано вполне существенным, практически уже безразлично, будет ли P равно 0,01 или 0,000001; в частности, по этой причине табличные значения χ^2 не выходят за пределы вероятности 0,01. При измерении же степени связанности необходимо опираться на определенную гипотезу относительно того отклонения от независимости, которое подлежит измерению. Так, например, для измерения степени сцепления двух менделевских факторов можно воспользоваться

процентом рекомбинаций; в этом случае доказательство наличия существенного сцепления может быть построено на сравнении разности между процентом рекомбинаций и 50% (что соответствует отсутствию сцепления факторов) с ее средней квадратической ошибкой. Это сравнение, если оно проведено с достаточной точностью, должно дать результаты, вполне согласованные с теми, которые получаются при применении критерия χ^2. Возьмем второй пример. Данные таблицы 2×2 иногда могут рассматриваться как результат деления на четыре части распределения двух коррелированных и нормально распределенных переменных так, что одна из частей распределения переменной лежит ниже, а другая — выше некоторой выбранной линии сечения. Например, распределение роста отцов и сыновей может быть разделено на четыре части сечениями, проходящими через рост в 68 дюймов. В данном случае отклонение от независимости двух переменных может быть выражено через корреляцию между ростом отца и сына. Эту величину можно определить по наблюденным частотам, а сравнение ее с соответствующей средней квадратической ошибкой, если оно проведено с достаточной точностью, должно дать результаты, согласующиеся с теми, которые дает критерий χ^2 в качестве критерия существенности этой связи. Установленная этим путем существенность становится по мере увеличения размера выборки все более и более явной, но обнаруженная в этом случае корреляция, конечно, не будет возрастать, а будет стремиться к некоторому фиксированному значению. Таким образом, критерий χ^2 не предназначен для измерения степени связанности и как критерий существенности не зависит от любых дополнительных гипотез относительно природы такого соответствия.

Критерии *однородности* математически тождественны критериям независимости. Последний пример может в одинаковой мере рассматриваться как с той, так и с другой точки зрения. Примененные в главе III критерии согласия с биноминальным распределением были, по существу, критериями однородности. Например, десять проб по 100 зерен ячменя в каждой (пример 7, стр. 61) можно рассматривать как данные таблицы 2×10. В этом случае показатель рассеяния χ^2 будет эквивалентным χ^2, полученному из таблицы сопряжения признаков. Несмотря на эту тождественность, описываемый в этой главе метод является более общим и применим в таких случаях, когда размер выборок различен.

Пример 13. *Однородность семейств по признаку цвета глаз.* Следующие данные относятся к 33 семействам Gammarus и дают численности черных и красных глаз в каждом семействе (данные Гёксли).

Соотношение итогов 2565 и 772 явным образом не соответствует менделевскому отношению 3 : 1; это отклонение следует приписать сцеплению генов. Теперь перед нами стоит вопрос: можно ли считать, что отношение между численностями черных и красных глаз одно и то же во всех семействах или установлен-

Таблица 20

Черный .	79	120	24	117	62	79	66	45	61	64	208	154	31	158	21	105	28
Красный .	14	31	6	29	17	20	12	11	14	13	52	45	4	45	4	28	7
Итого .	93	151	30	146	79	99	78	56	75	77	260	199	35	203	25	133	35
Черный . .	58	81	25	95	47	67	30	70	139	179	129	44	24	19	45	91	2 565
Красный .	19	27	8	29	16	21	11	28	57	62	44	17	9	8	23	41	772
Итого .	77	108	33	124	63	88	41	98	196	241	173	61	33	27	68	132	3 337

ное нарушение соотношения 3 : 1 обусловлено только отклонениями в отдельных семьях. Для всей таблицы в целом имеем $\chi^2 = 35{,}620$ при $n = 32$. Эта величина выходит за рамки таблицы, и поэтому здесь следует применить способ, описанный на стр. 71.

$$\sqrt{2\chi^2} = 8{,}44$$
$$\sqrt{2n-1} = 7{,}94$$
$$\text{Разность} = +0{,}50 \pm 1$$

Следовательно, эта группа семейств не может считаться существенно разнородной, т. е. в каждом семействе соотношение численностей черных и красных глаз согласуется с соотношением этих численностей, взятым по всем семействам вместе.

Тот же самый прием обработки можно применить, если эти данные рассматривать не в качестве 33 выборок, в которых выделено два класса (черные и красные глаза), а как две выборки с 33 классами (семействами). Вопрос «Относятся ли эти две выборки к одной и той же совокупности?» идентичен вопросу «Является ли соотношение между численностями черных и красных глаз одинаковым во всех семействах?» Идентичность этих вопросов часто упускается из виду, а между тем она имеет большое значение.

21.01. Поправка Иейтса на непрерывность

Распределение χ^2, представленное в таблице III, относится к числу непрерывных распределений. Однако распределение частот всегда дискретно. Следовательно, применение критерия χ^2 при сопоставлении фактических и ожидаемых частот дает только приближенные результаты, ибо непрерывное распределение может в этом случае рассматриваться только в качестве предельного, к которому стремится фактическое дискретное распределение при увеличении размера выборки. Ранее в целях исключения искажающего влияния небольшого числа наблюдений в некоторых группах было рекомендовано объединение таких малочисленных

групп, с тем чтобы ожидаемые для них численности были бы равны не меньше пяти. При этом ограничении можно высказать общее пожелание, чтобы число групп было бы по возможности большим, а соответствующие численности — малыми (но не меньше 5), так как в этом случае распределение в большей степени приближается к табличному непрерывному распределению χ^2.

Особый интерес представляет случай, когда имеется только одна степень свободы и когда, следовательно, величина χ^2 может быть определена по фактической численности единичной группы. Если эта численность очень мала, положим 3, то вероятность, соответствующая этой численности, должна вычисляться не сама по себе, а как сумма вероятностей еще более резких отклонений при численностях 2, 1 и 0. Поэтому, если нам надо установить, будет ли фактическая численность 3 столь малой, чтобы определять собой существенность отклонения от ожидаемой численности, то мы должны будем установить, не будет ли сумма вероятностей, относящихся к 3, 2, 1 и 0, меньше, чем некоторое критическое значение, например, $P = 0{,}05$. Другими словами, мы должны в этом случае определить, не будет ли вероятность, относящаяся к фактическому отклонению и ко всем другим еще бо́льшим отклонениям, столь малой, что у нас не будет оснований считать данное отклонение простой случайностью.

Следовательно, наша задача, если ее сформулировать точно, состоит в определении некоторого конечного числа вероятностей, которые в простейших случаях могут быть определены даже непосредственно. С другой стороны, эта же проблема решается при помощи распределения χ^2, в котором берется определенная площадь, отсекаемая от «хвоста» непрерывной кривой. Но так как эта кривая является сглаженным изображением фактического распределения, то фактической вероятности для наблюденного значения 3 соответствует площадь кривой между $3\frac{1}{2}$ и $2\frac{1}{2}$, а суммарной вероятности значения 3 и меньше соответствует площадь кривой, отсеченная у точки $3\frac{1}{2}$. Таким образом, наша задача может быть решена при помощи распределения χ^2, но для этого следует брать не фактически наблюденную, а увеличенную на половину единицы частоту. Этот корректив предложен Ф. Иейтсом.

Пример 13.1. *Распространение преступности среди лиц, у которых их близнецы братья или сестры — преступники.* Ланге установил, что у 13 преступников, которые имели монозиготных близнецов братьев или сестер, среди последних 10 человек оказались также преступниками и только трое не были уличены в каком-либо преступлении. В случае дизиготных двоен (одного и того же пола) из 17 преступников только двое имели своих близнецов (братьев или сестер) преступников, а у остальных 15 их близнецы не являлись преступниками. Было установлено, что окружающая обстановка для дизиготных и монозиготных двоен

была одинакова и что поэтому наличие у последних повышенного числа случаев, когда антисоциальные поступки совершаются обоими братьями-близнецами или обеими сестрами-близнецами, в основном должно быть отнесено за счет генетических факторов. Можно ли считать, что данные Ланге подтверждают тот факт, что преступность значительно больше распространена среди монозиготных двоен, один из которых стал преступником, чем среди дизиготных двоен-преступников?

На основе этих данных можно построить такую таблицу 2×2:

Таблица 20.1

	Уличенные в преступлении	Не уличенные в преступлении	Итого
Монозиготные двойни . . .	10	3	13
Дизиготные двойни	2	15	17
Итого . . .	12	18	30

Разность $(ad - bc)$ равна 144 и, следовательно, величина

$$\chi^2 = \frac{144^2 \cdot 30}{12 \cdot 18 \cdot 13 \cdot 17} = 13{,}032$$

имеет вполне существенное значение, так как превосходит свою среднюю квадратическую ошибку в 3,61 раза. Вероятность превзойти такое отклонение определяется отношением 1 : 6500.

Для того чтобы применить поправку Иейтса, мы должны переписать эту таблицу, уменьшив большие частоты 10 и 15 на половину единицы и увеличив меньшие частоты 2 и 3 также на половину единицы.

Разность между произведениями крест-накрест $ad - bc$ теперь равна 129, что меньше прежней разности 144 на 15 или на половину общего числа наблюдений (30). Во всех других отношениях вычисления остаются теми же, что и ранее. Новое значение χ^2 равно 10,458, которое все же остается довольно существенным при одной степени свободы, но теперь оно превосходит свою среднюю квадратическую ошибку только в 3,234 раза, что соответствует 1 из 1638 шансов. Точное соотношение шансов, как это будет определено в следующем параграфе, равно 1 к 2150. Следовательно, в данном случае поправка дала несколько преувеличенную вероятность.

21.02. Точная обработка данных таблицы 2×2

Обработка численностей при помощи χ^2 является приближенной и ее преимущество в том, что она относительно проста. Однако в сомнительных случаях следует применять, хотя и более трудоемкий, но точный расчет, который дает возможность уяснить себе природу этой оценки, остающейся неясной, когда применяется метод χ^2.

Пусть p является вероятностью какого-либо события; вероятность того, что оно произойдет в a случаях при $(a + b)$ испытаниях, согласно биноминальному закону распределения, будет

$$\frac{(a+b)!}{a!\,b!} p^a q^b,$$

где $q = 1 - p$. Вероятность того, что при $(c + d)$ испытаниях оно появится c раз, будет

$$\frac{(c+d)!}{c!\,d!} p^c q^d.$$

Следовательно, вероятность наличия в таблице 2×2 частот a, b, c и d будет

$$\frac{(a+b)!\,(c+d)!}{a!\,b!\,c!\,d!} p^{a+c} q^{b+d}.$$

Эта величина остается неопределенной, если неизвестно p. Однако множитель, содержащий p и q, будет одним и тем же во всех таблицах, в которых итоговые частоты те же самые, что и у данной таблицы, т. е. $a + c$, $b + d$, $a + b$, $c + d$; поэтому отдельным членам возможного ряда наблюдений, имеющих одни и те же итоговые частоты, будут соответствовать вероятности, пропорциональные величинам

$$\frac{1}{a!\,b!\,c!\,d!},$$

каким бы ни было значение p.

Сумма величин $\frac{1}{a!\,b!\,c!\,d!}$ во всех выборках с одними и теми же итогами сторон таблицы 2×2, будет равна

$$\frac{n!}{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!},$$

где $n = a + b + c + d$. Поэтому вероятность некоторого определенного ряда наблюдений при данных итоговых частотах будет

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!} \frac{1}{a!\,b!\,c!\,d!}.$$

В примере 13.1 мы имеем ряд:

$$\frac{18!\,12!\,17!\,13!}{30!} \left\{ \frac{1}{2!\,3!\,10!\,15!};\ \frac{1}{2!\,11!\,16!};\ \frac{1}{1!\,12!\,17!} \right\},$$

соответствующий вероятностям при фактических частотах и при двух возможных случаях более резкого отклонения от независимости. Следовательно, данная таблица наблюдений позволяет без какого-либо дополнительного допущения прийти к вполне

точному суждению о существенности противоречия между данными и гипотезой, согласно которой имеет место пропорциональное распределение численностей по классам. Это противоречие будет существенным, если величина

$$\frac{18!\,13!}{30!}(2992+102+1)$$

будет незначительной. Подсчет дает 619/1330665, или около 1 шанса на 2150, откуда следует, что если бы гипотеза о пропорциональном распределении была правильной, то такие наблюдения, как данные, могли бы встретиться только в порядке крайнего исключения.

21.03. Точные оценки, основанные на распределении χ^2

Критерий χ^2 первоначально применялся для сравнения ряда наблюденных частот с частотами, ожидаемыми согласно той или иной гипотезе; в этом случае он является не точным, а приближенным, хотя и имеет в этом виде довольно широкое применение при решении целого ряда практически важных задач. В других случаях, когда вместо численностей участвуют некоторые непрерывно изменяющиеся величины, этот критерий становится точным критерием существенности. Наиболее важными из этих случаев являются:

1) применение этого критерия для установления того, в какой мере согласуются данные, полученные в качестве выборки из нормальной совокупности, с дисперсией, которая ожидается на основе некоторых теоретических соображений;

2) применение этого критерия в сочетании с показаниями, полученными на основе некоторго числа независимых критериев существенности.

Пример 14. *Согласованность наблюдений с дисперсией, ожидаемой при условии нормального распределения.* Если ряд $x_1, x_2, \ldots$ является выборкой из нормальной совокупности, среднее квадратическое отклонение которой равно σ, то величина

$$\frac{1}{\sigma^2}S(x-\overline{x})^2$$

распределена как χ^2 с n, равным числу наблюдений в выборке без единицы. Бисфем дает три ряда экспериментальных значений частного коэффициента корреляции, каждое из которых определено по 30 наблюдениям над тремя переменными. Он указывает, что эти коэффициенты должны быть распределены так, чтобы $\frac{1}{\sigma^2}=29$, но на самом деле следует считать, что $\frac{1}{\sigma^2}=28$. Для трех выборок в 1000, 200 и 100 после группировки данных были получены следующие суммы $S(x-\overline{x})^2$:

35,0279; 7,4573; 3,6146

На основе этого были определены значения χ^2, ожидаемые согласно каждой из этих двух гипотез (табл. 21):

Таблица 21

	Эксперимент 1	2	3	Итого	$\sqrt{2\chi^2}$	Разность
$29S(x-\overline{x})^2$ . . .	1015,81	216,26	104,82	1336,89	51,71	+0,79
$28S(x-\overline{x})^2$ . . .	980,78	208,80	101,21	1290,79	50,81	−0,11
Ожидаемое значение (n) . . .	999	199	99	1297	50,92	—

Отсюда следует, что правильная формула дисперсии дает несколько лучшую согласованность с ожидаемыми значениями n. Однако, как это можно видеть из последних двух колонок таблицы, различие между двумя методами не может в этом случае считаться существенным. Это различие может быть установлено экспериментально только при наличии по крайней мере 6000 наблюдений.

21.1. Объединение вероятностей, соответствующих некоторому числу критериев существенности

Если при помощи критерия существенности χ^2 производится ряд оценок независимости, то иногда может случиться так, что в некоторых случаях χ^2 будет выходить за пределы существенности, в то время как эти критерии χ^2, взятые в своей совокупности, останутся в рамках, допустимых для случайных отклонений. Иногда возникает необходимость определить вероятность только этого последнего факта, не обращая внимания на входящие в него отдельные случаи. Этот общий критерий существенности будет основываться, очевидно, на произведении вероятностей, относящихся к отдельным случаям.

При построении критерия, соответствующего этому обобщению, можно использовать тот факт, что сумма показателей χ^2 сама распределена по закону распределения χ^2 с соответствующим числом степеней свободы. В частности, когда отдельным значениям χ^2 соответствуют две степени свободы ($n=2$), натуральный логарифм вероятности равен $-\frac{1}{2}\chi^2$. Поэтому, если мы найдем натуральный логарифм такой частной вероятности, переменим знак на обратный и удвоим его, то тем самым получим величину χ^2 с двумя степенями свободы. Эти величины можно просуммировать и получить в результате обобщенный критерий существенности χ^2.

Пример 14.1. *Существенность произведения некоторого числа независимых вероятностей.* По трем значениям критерия существенности найдены следующие вероятности: 0,145; 0,263; 0,087. Спрашивается, не приведет ли к существенности объедине-

ние этих трех критериев вместе? В соответствии с выше сказанным, мы имеем

P	$-\log_e P$	Степени свободы
0,145	1,9310	2
0,263	1,3356	2
0,087	2,4419	2
	5 7085	6

$$\chi^2 = 11{,}4170\,.$$

Таким образом, для 6 степеней свободы мы нашли $\chi^2 = 11{,}417$. Для 5%-ного уровня существенности табличное значение равно 12,592, в то время как для 10%-ного уровня — 10,645. Следовательно, вероятность, соответствующая совокупности этих трех критериев существенности, превосходит 0,05 и близка к 0,075.

Следует отметить, что при применении данного метода нам необходимо знать не только то, что частные критерии были существенными или несущественными, но и иметь соответствующие им вероятности, вычисленные с точностью до двух или трех десятичных знаков. Для определения этих последних в большинстве случаев достаточно применить интерполяцию между логарифмами вероятностей P, взятых из таблицы III. В данном случае могут быть использованы как натуральные, так и обычные (десятичные) логарифмы. Покажем схему этих расчетов на примере определения вероятности того, что χ^2 превосходит 11,417 при $n = 6$.

Наше значение χ^2 превосходит табличное значение для 10%-ного уровня на 0,772, в то время как табличное значение при 5%-ном уровне превосходит значение 10%-ного уровня на 1,947. Отсюда находим отношение

$$\frac{0{,}772}{1{,}947} = 0{,}397\,.$$

Разность между десятичными логарифмами для чисел 5 и 10 составляет 0,3010; умножая ее на 0,397, получим 0,119. Следовательно, искомой вероятности соответствует отрицательный логарифм — 1,119, а поэтому сама вероятность будет равна 0,076. Заметим, что вероятность, вычисленная точным методом, равна 0,07631.

22. Разложение χ^2 на компоненты

Вместе с объединением нескольких значений χ^2 и получением благодаря этому более широкого критерия оценки возможна и обратная операция дробления χ^2 на отдельные части с выделением соответствующих им степеней свободы, что позволяет иметь оценки частных различий.

Пример 15. *Разложение χ^2 при проверке закона Менделя.* В таблице, приведенной на стр. 87, дано распределение шестнадцати семей Primula по восьми классам, полученным в резуль-

Таблица 22

Тип	Номер семьи 54	55	58	59	107	110	119	121	122	127	129	131	132	133	135	178	Итого
Ch G W	5	18	17	2	12	17	9	10	24	9	3	16	20	9	11	10	192
Ch G w	10	13	11	12	20	16	10	7	23	3	6	24	18	2	13	12	200
Ch g W	4	10	17	3	14	10	6	8	19	5	5	23	18	10	7	12	171
Ch g w	9	17	11	11	13	13	9	8	9	6	3	12	18	1	9	12	161
ch G W	13	22	20	10	5	5	16	2	30	3	8	21	19	4	9	12	199
ch G w	14	16	18	9	12	6	14	3	16	5	7	13	14	4	13	10	174
ch g W	10	11	12	6	7	3	18	2	11	5	4	14	23	4	6	13	149
ch g w	7	12	16	6	10	8	10	4	23	5	4	22	23	7	8	16	181
Итого	72	119	122	59	93	78	92	44	155	41	40	145	153	41	76	97	1427
χ^2	9,78	7,86	5,48	13,00	12,55	19,23	10,09	12,36	18,06	4,86	4,80	9,21	3,18	14,22	5,05	2,05	151,78

тате скрещивания с тройным рецессивом (данные Уинтона и Бэтсона).

Теоретическая постановка вопроса состоит в следующем: определить, не появляются ли классы с одинаковой частотой в соответствии с гипотезой о том, что в каждом классе аллеломорфы встречаются одинаково часто и что данные три фактора между собой не связаны. Этот вопрос в целом легко решается на основе итоговых данных для всех шестнадцати семей, но у отдельных семей встречаются некоторые неправильности, требующие своего изучения. В последней строке таблицы для каждой отдельной семьи даны значения χ^2, полученные в результате сравнения фактических данных с ожидаемыми согласно указанной выше гипотезе. Каждому из этих значений χ^2 соответствует 7 степеней свободы.

Из таблицы видно, что в 6 случаях из 16 вероятность P меньше 0,1 и в двух случаях она даже меньше 0,02. Это указывает на наличие достаточно выраженной неравномерности распределения внутри семей. Суммарная величина χ^2 (которую не следует смешивать с величиной χ^2, рассчитанной по итогам) равна 151,78 и ей соответствует 112 степеней свободы. Отсюда

$$\sqrt{223} = 14{,}93$$

$$\sqrt{303{,}56} = 17{,}42$$

$$\text{Разность} = +2{,}49$$

и, следовательно, судя по этому значению χ^2, имеются явные отклонения от ожидаемых частот внутри семей.

Фактическое распределение у каждой семьи может отличаться от ожидаемого распределения частот семью независимыми способами. Углубляя наш анализ, мы можем каждое из χ^2 подразделить на семь частей, соответствующих единичным степеням свободы. Математически такое разложение χ^2 может быть проведено многими способами, но из них только один представляет интерес с биологической точки зрения, а именно такой, когда выделение частей связано с неравенством аллеломорфов трех факторов и с тремя возможными связями факторов. Если мы возьмем отдельные частоты со знаками, указанными в табл. 23, то получим семь подразделений, которые в целом независимы друг от друга, так как любые два из них имеют в четырех случаях совпадающие и в четырех случаях несовпадающие знаки. Три первые степени свободы представляют неравенства аллеломорфов у трех факторов *Ch*, *G* и *W*; следующие три степени свободы включают в себя сведения относительно связи трех факторов, взятых попарно, последняя же седьмая степень свободы, хотя и не имеет столь простого биологического истолкования, но она все же необходима для полноты анализа.

Если мы возьмем, например, первую семью и найдем разность между числом растений в группах, содержащих символ *W*,

Таблица 23

	Ch	*G*	*W*	*G W*	*Ch W*	*Ch G*	*Ch G W*
Ch G W	+	+	+	+	+	+	+
Ch G w	+	+	−	−	−	+	−
Ch g W	+	−	+	−	+	−	−
Ch g w	+	−	−	+	−	−	+
ch G W	−	+	+	+	−	−	−
ch G w	−	+	−	−	+	−	+
ch g W	−	−	+	−	−	+	+
ch g w	−	−	−	+	+	+	−

и в группах, содержащих символ *w* (эта разность равна 8), то этой степени свободы будет соответствовать χ^2, которое можно найти, возводя эту разность в квадрат и разделив результат на численность данного семейства, т. е. $\chi^2 = 8^2 : 72 = 0{,}889$. Этим путем можно найти все части общего критерия $\chi^2 = 151{,}78$, соответствующие каждой из 112 степеней свободы в отдельности (табл. 24).

Просматривая в каждой колонке итоговые значения χ^2 и учитывая, что каждому из этих значений χ^2 соответствует $n = 16$, мы видим, что для всех их, за исключением первой колонки, вероятность P содержится между 0,05 и 0,95, в то время как χ^2 первой колонки столь велико, что должно считаться существенным. Это указывает на то, что большую часть отклонений от нормы сле-

Таблица 24

Семейство	*Ch*	*G*	*W*	*GW*	*ChW*	*ChG*	*ChGW*	Итого
54	3,556	2,000	0,889	0,222	2,000	0,889	0,222	9,778
55	0,076	3,034	0,076	3,034	0,412	1,017	0,210	7,859
58	0,820	0,820	0,820	0,295	1,607	0,820	0,295	5,477
59	0,153	0,831	*4,898*	0,017	*6,119*	0,831	0,153	13,002
107	*6,720*	0,269	3,108	1,817	0,097	0,269	0,269	12,549
110	*14,821*	1,282	0,821	0,821	0,205	1,282	0	19,232
119	*6,261*	0,391	0,391	0,174	2,130	0,043	0,696	10,086
121	*11,000*	0	0	0,364	0,818	0,091	0,091	12,364
122	0,161	*6,200*	1,090	1,865	0,523	0,316	*7,903*	18,058
127	0,610	0,024	0,220	0,610	1,195	0,220	1,976	4,855
129	0,900	1,600	0	0,400	0,100	0,900	0,900	4,800
131	0,172	0,062	0,062	0,062	0,062	0,338	*8,448*	9,206
132	0,163	0,791	0,320	0,320	0,059	1,471	0,059	3,183
133	0,220	0,220	*4,122*	0,024	*8,805*	0,220	0,610	14,221
135	0,211	3,368	1,316	0,053	0,053	0	0,053	5,054
178	0,258	0,835	0,093	0,093	0,010	0,258	0,505	2,052
Итого	46,102	21,727	18,226	10,171	24,195	8,965	22,390	151,776

дует приписать поведению фактора *Ch*, что обусловлено связью с рецессивным летальным геном в одном из типов семей. Следует заметить, что значения данных этой первой колонки слишком велики только у четырех семей с № 107 по № 121. Если эти четыре семьи исключить, то χ^2 снижается до 97,545 при $n=84$; в этом случае величина χ^2 примерно равна своей средней квадратической ошибке, и поэтому отклонение от нормы теперь уже не может считаться существенным.

Кроме того, можно видеть, что у оставшихся 12 семей имеются 7 довольно больших значений χ^2, из которых 6 располагаются попарно в трех семьях. Распределение этих оставшихся 12 семей по соответствующим им вероятностям P представлено в табл. 25, из которой видно, что у данного распределения замечается некоторый эксцесс в сторону больших значений χ^2.

Таблица 25

P	1,0	0,9	0,8	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0	Итого
Число семей . . .	1	1	0	4	1	2	0	1	1	1	0		12

Этот результат, как и другие подобные несущественные итоги, рассмотрены здесь только для того, чтобы дать представление о путях более углубленного анализа результатов, получаемых при проверке той или иной гипотезы.

В параграфе 55 будут даны общие положения, относящиеся к разложению χ^2 на составные части.

Пример 15.1. *Сложный критерий однородности данных и его последовательное разложение.* Табл. 25.1 показывает общее число потомков и число рекомбинаций, установленных у 22 типов потомства садового горошка (данные Рейсмуссона). Все это потомство было получено от одного растения при помощи обратного скрещивания. Данное потомство образует 9 родственных семейств; общее число растений и число рекомбинаций указано в таблице. В трех случаях имеются по две группы, являющиеся потомками одного и того же растения; такие парные группы включены для того, чтобы представилась возможность изучить связь генов. Следовательно, представленные здесь 9 групп являются потомством 6 родителей F_3, которые, в свою очередь, происходят от 3 прародительских растений F_2. В результате генерация дала 922 растения, из которых 105 определены как рекомбинации. По этим данным требуется определить, как меняется доля рекомбинаций на четырех стадиях перехода от общей группировки к более детальной.

При анализе подобного рода данных большое удобство представляет метод, предложенный Брандтом и Снедекором. Если

у некоторого типа потомства или у объединенной группы нескольких таких типов, насчитывающей T растений, установлено c рекомбинаций, то для каждой такой группы можно вычислить величину $\frac{c^2}{T}$. Эти отношения даны в табл. 25.2, где они расположены в соответствии с размещением групп и подгрупп в табл. 25.1.

Таблица 25.1

Общее число растений (T) и рекомбинаций (c) у 22 типов потомства и объединение этого потомства в более широкие группы

Потомство																							
единичных растений		братских растений		родительских растений F_3		прародительских растений F_2		Итого	
T	c	T	c	T	c	T	c	T	c
34	3	77	11	171	21				
43	8								
94	10	94	10						
73	3					427	42		
20	2	130	8						
16	2								
21	1			256	21				
31	6								
51	4	126	13						
29	2								
15	1							922	105
64	9	119	22	119	22	119	22		
55	13								
55	7	89	12	89	12				
34	5								
37	3	37	3	37	3				
45	7					376	41		
28	0	108	9						
35	2								
68	8			250	26				
44	5	142	17						
30	4								

Так как доля рекомбинаций, наблюденная в различных подгруппах одной и той же группы, различна, то сумма величин $\frac{c^2}{T}$ по этим подгруппам будет больше, чем такое же отношение для группы в целом. Следовательно, пять итогов табл. 25.2 дадут убывающий ряд чисел; последовательные разности между ними будут измерять неоднородность материала.

Если бы эта схема расчета была бы применена к вполне однородному материалу, то можно было бы получить ряд величин χ^2, распределенных по закону распределения χ^2. Для этого необходимо только разделить каждую из указанных в таблице разностей на pq, если через p обозначить число рекомбинаций, а через q — число прежних комбинаций. Числа степеней свободы для

этих величин χ^2 определяются как разности между количествами групп двух соседних столбцов.

Таблица 25.2

Значения $\frac{c^2}{T}$ для всех групп и подгрупп

Единичные растения	Братские растения	Родительские растения F_3	Прародительские растения F_2	Итого
0,26471	1,57143	2,57895	4,13115	11,95770
1,48837				
1,06383	1,06383			
0,12329				
0,20000	0,49231			
0,25000				
0,04762		1,72266		
1,16129				
0,31373	1,34127			
0,13793				
0,06667				
1,26562	4,06723	4,06723	4,06723	
3,07273				
0,89091	1,61798	1,61798		
0,73529				
0,24324	0,24324	0,24324		
1,08889			4,47074	
0,00000	0,75000			
0,11429		2,70400		
0,94118				
0,56818	2,03521			
0,53333				
14,57110	13,18250	12,93406	12,66912	11,95770
1,38860	0,24844	0,26494	0,71142	Разности
13,760	2,462	2,625	7,050	χ^2
13	3	3	2	n

Данные нижней части табл. 25.2 показывают, что более или менее явная неоднородность в величине связи существует только между растениями F_2, т. е. только на самой ранней стадии появления неоднородности. Здесь величина χ^2 равна 7,050 и имеет две степени свободы; следовательно, в данном случае вероятность лежит между 5 и 2 процентами. Поэтому вполне возможно, что наблюдаемое расслоение потомства под влиянием связи возникло на этой первой стадии, вследствие чего данные, полученные на последней стадии, должны оцениваться с учетом этой неоднородности, возникшей на ранней стадии размножения.

Собственно говоря, в данном случае следовало бы расчет χ^2 произвести по каждой из трех групп растений F_2 в отдельности, применяя соответствующие делители pq. Однако в нашем случае, как можно заметить, различие между такими делителями в первой и третьей группе небольшое: в первой группе доля рекомбинаций составляет 9,836%, а в третьей — 10,904%, т. е. эти группы дают примерно один и тот же размер связи в отличие от второй

группы, где доля рекомбинаций составляет 18,487%. Поэтому мы можем объединить первую и третью группы и произвести расчет χ^2 исходя из отношения 83/803, и расчет χ^2 для потомства второго растения из F_2 на основе отношения 22/119.

Если бы требовалось произвести расчет χ^2 для каждой отдельной части табл. 25.2, то лучше всего начать с вычисления разностей между суммой показателей $\left(\frac{c^2}{T}\right)$ всех подгрупп и таким же показателем группы в целом. В табл. 25.3 иллюстрируется этот прием расчетов; в ней весь ряд данных, относящихся к 21 степени свободы, подразделен на 13 отдельных сравнений. Значение χ^2 каждой из этих частей зависит от величины pq, на которую и следует делить соответствующую разность. Так, для оценки различий между растениями F_2 мы должны взять $p=105:922$, в результате чего получим как и прежде, $\chi^2=7{,}0498$ с двумя степенями свободы. Для потомства первого и третьего растения из F_2 мы возьмем в качестве делителя 0,0926786, а для потомства второго растения — 0,1506956, в результате чего получим величины, приведенные в табл. 25.4.

Таблица 25.3

Разности между показателями $\frac{c^2}{T}$, относящимися к родственным растениям и подгруппам, с указанием соответствующих степеней свободы

Растения F_5	Растения F_4	Растения F_3	Растения F_2
0,18165 (1)	0,05631 (1)	0,17046 (1)	0,71142 (2)
—			
0,12860 (3)	0,11092 (1)		
0,33835 (3)			
0,27112 (1)	—	—	
0,00822 (1)	—		
—	—	0,09448 (2)	
0,45318 (2)	0,08121 (1)		
0,00748 (2)			

Итоги для последующих поколений в табл. 25.4 несколько отличаются от соответствующих итогов табл. 25.2, но они дают лучшую оценку неоднородности на этих последующих стадиях, так как они основаны на допущении, что растения F_2 не были однородными. Следует отметить тот факт, что значение χ^2 при трех степенях свободы, соответствующих различиям у растений F_4, полученных от одних и тех же растений F_3, а также значение χ^2 для различий между растениями F_3, несколько увеличились, несмотря на то, что для расчета показателей для первой и третьей группы растений применялся меньший делитель. Это произошло потому, что в этих колонках увеличение делителя для потомства второго растения не было компенсировано.

Таблица 25.4

Растения F_5	Растения F_4	Растения F_3	Растения F_2
1,9600 (1) — 1,3876 (3) 3,6508 (3) 1,7991 (1) 0,0887 (7) — 4,8898 (2) 0,0807 (2)	0,6076 (1) 1,1968 (1) — — — 0,8763 (1)	1,8393 (1) — 1,0194 (2)	7,0498 (2)
13,8567	2,6807	2,8587	7,0498
13	3	3	2

Отсутствие существенных значений χ^2 в первых трех колонках табл. 25.4 показывает, что здесь нет необходимости проводить более детальный анализ, так как это не даст никаких новых сведений о неоднородности материала.

Таблица

n	$P = 0{,}99$	0,98	0,95	0,90	0,80	0,70
1	0,000157	0,000628	0,00393	0,0158	0,0642	0,148
2	0,0201	0,0404	0,103	0,211	0,446	0,713
3	0,115	0,185	0,352	0,584	1,005	1,424
4	0,297	0,429	0,711	1,064	1,649	2,195
5	0,554	0,752	1,145	1,610	2.343	3,000
6	0,872	1,134	1,635	2,204	3,070	3,828
7	1,239	1,564	2,167	2,833	3,822	4,671
8	1,646	2,032	2,733	3,490	4,594	5,527
9	2,088	2,532	3,325	4,168	5,380	6,393
10	2,558	3,059	3,940	4,865	6,179	7,267
11	3,053	3,609	4,575	5,578	6,989	8,148
12	3,571	4,178	5,226	6,304	7,807	9,034
13	4,107	4,765	5,892	7,042	8.634	9,926
14	4,660	5,368	6,571	7,790	9,467	10,821
15	5,229	5,985	7,261	8,547	10,307	11,721
16	5,812	6,614	7,962	9,312	11,152	12,624
17	6,408	7,255	8,672	10,085	12,002	13,531
18	7,015	7,906	9,390	10,865	12,857	14,440
19	7,633	8,567	10,117	11,651	13,716	15,352
20	8,260	9,237	10,851	12,443	14,578	16,266
21	8,897	9,915	11,591	13,240	15,445	17,182
22	9,542	10,600	12,338	14,041	16,314	18,101
23	10,196	11,293	13,091	14,848	17,187	19,021
24	10,856	11,992	13,848	15,659	18,062	19,943
25	11,524	12,697	14,611	16,473	18,940	20,867
26	12,198	13,409	15,379	17,292	19,820	21,792
27	12,879	14,125	16,151	18,114	20,703	22,719
28	13,565	14,847	16,928	18,939	21,588	23,647
29	14,256	15,574	17,708	19,768	22,475	24,577
30	14,953	16,306	18,493	20,599	23,364	25,508

При значениях n, больших 30, для оценки χ^2 можно воспользоваться

Таблица III

значений χ^2

0,50	0,30	0,20	0,10	0,05	0,02	0,01
0,455	1,074	1,642	2,706	3,841	5,412	6,635
1,386	2,408	3,219	4,605	5,991	7,824	9,210
2,366	3,665	4,642	6,251	7,815	9,837	11,345
3,357	4,878	5,989	7,779	9,488	11,668	13,277
4,351	6,064	7,289	9,236	11,070	13,388	15,086
5,348	7,231	8,558	10,645	12,592	15,033	16,812
6,346	8,383	9,803	12,017	14,067	16,622	18,475
7,344	9,524	11,030	13,362	15,507	18,168	20,090
8,343	10,656	12,242	14,684	16,919	19,679	21,666
9,342	11,781	13,442	15,987	18,307	21,161	23,209
10,341	12,899	14,631	17,275	19,675	22,618	24,725
11,340	14,011	15,812	18,549	21,026	24,054	26,217
12,340	15,119	16,985	19,812	22,362	25,472	27,688
13,339	16,222	18,151	21,064	23,685	26,873	29,141
14,339	17,322	19,311	22,307	24,996	28,259	30,578
15,338	18,418	20,465	23,542	26,296	29,633	32,000
16,338	19,511	21,615	24,769	27,587	30,995	33,409
17,338	20,601	22,760	25,989	28,869	32,346	34,805
18,338	21,689	23,900	27,204	30,144	33,687	36,191
19,337	22,775	25,038	28,412	31,410	35,020	37,566
20,337	23,858	26,171	29,615	32,671	36,343	38,932
21,337	24,939	27,301	30,813	33,924	37,659	40,289
22,337	26,018	28,429	32,007	35,172	38,968	41,638
23,337	27,096	29,553	33,196	36,415	40,270	42,980
24,337	28,172	30,675	34,382	37,652	41,566	44,314
25,336	29,246	31,795	35,563	38,885	42,856	45,642
26,336	30,319	32,912	36,741	40,113	44,140	46,963
27,336	31,391	34,027	37,916	41,337	45,419	48,278
28,336	32,461	35,139	39,087	42,557	46,693	49,588
29,336	33,530	36,250	40,256	43,773	47,962	50,892

величиной $\sqrt{2\chi^2} - \sqrt{2n-1}$, которая имеет нормальное распределение с дис-

ГЛАВА ПЯТАЯ

ОЦЕНКА СУЩЕСТВЕННОСТИ СРЕДНИХ, РАЗНОСТИ СРЕДНИХ И КОЭФФИЦИЕНТОВ РЕГРЕССИИ

23. Средняя квадратическая ошибка средней

В основе статистической оценки средних лежит следующее основное положение: если некоторая величина распределена нормально с дисперсией σ^2, то средняя случайной выборки, состоящей из n наблюдений, распределена нормально с дисперсией $\frac{\sigma^2}{n}$.

Практическое значение этого положения отчасти усиливается еще тем фактом, что даже если исходное распределение не будет точно нормальным, все же распределение средней стремится к нормальной форме при возрастании размера выборки. Таким образом, это положение имеет широкую область применения и остается правомерным и в тех случаях, когда мы не имеем достаточных оснований считать, что исходное распределение нормально, лишь бы при этом у нас была бы уверенность в том, что оно не принадлежит к тому исключительному классу распределений, при которых распределение средней не стремится к нормальному.

Следовательно, если нам известна дисперсия некоторой генеральной совокупности, то мы можем вычислить дисперсию средней случайной выборки любого размера и тем самым получаем возможность проверить, насколько существенно эта средняя отличается от некоторого фиксированного значения. Если эта разность значительно превосходит среднюю квадратическую ошибку средней, то она существенна. Удобно в качестве критического значения существенности условно взять удвоенную среднюю квадратическую ошибку; это примерно соответствует пограничному значению вероятности $P=0{,}05$, которое было уже использовано в распределении χ^2. В таблице IV (стр. 144) в нижней строке приведены отклонения в нормальном распределении, соответствующие приведенным значениям P. Более подробное табулирование нормального распределения было дано в табл. 1.

Пример 16. *Существенность средней при большой выборке.* Для иллюстрации расчетов возьмем эксперимент Уэлдона с игральными костями (пример 5, стр. 57). Переменной величиной здесь является число выпадений 5 или 6 очков при выбрасывании 12 костей. В опыте это число изменяется от нуля до одиннадцати, а фактическая средняя равна 4,0524. Средняя, ожидаемая при предположении, что кости правильные, равна 4, так что отклонение фактической средней от ожидаемой составит 0,0524. Если теперь найти оценку дисперсии всей выборки из 26306 наблюдений, как это было сделано в примере 2 (без поправки Шеппарда, так как здесь обрабатываются несгруппированные данные; но даже при группировке данных, когда средняя подвержена ошибкам группировки, дисперсия этой средней все же должна вычисляться без поправки Шеппарда), то получим:

$$\sigma^2 = 2{,}69826,$$

отсюда

$$\frac{\sigma^2}{n} = 0{,}0001026$$

и

$$\frac{\sigma}{\sqrt{n}} = 0{,}01013.$$

Таким образом, средняя квадратическая ошибка средней примерно равна 0,01, а фактическое отклонение более чем в 5,2 раза превышает эту ошибку. Итак, хотя и несколько иным путем мы приходим к тому же заключению, к которому пришли ранее на стр. 59. Различие между этими двумя методами состоит в том, что настоящая наша оценка средней не зависит от гипотезы о биноминальной форме распределения, но, с другой стороны, мы здесь допускаем, что полученная на основе наблюдений σ правильна. При малой выборке такое допущение было бы неприемлемым. Основная задача этой главы состоит в том, чтобы показать, какие уточнения необходимо ввести в рассматриваемые критерии существенности, если учитывать погрешности нашей оценки среднего квадратического отклонения.

Возвращаясь к теории больших выборок, отметим, что хотя в случаях, подобных рассмотренному выше, довольно часто возникает необходимость сравнить фактическое значение средней с значением, определяемым проверяемой нами гипотезой, однако столь же часто или даже более часто нам приходится сравнивать два экспериментальных значения средней и устанавливать степень их согласованности друг с другом. В этих случаях нам необходимо знать дисперсию разности между такими средними. Для определения ее можно воспользоваться тем положением, что дисперсия разности двух *независимых* переменных равна сумме их дисперсий. Так, если средние квадратические отклонения этих

переменных равны σ_1 и σ_2, а дисперсии — σ_1^2 и σ_2^2, то дисперсия разности будет $\sigma_1^2 + \sigma_2^2$, а среднее квадратическое отклонение ее равно $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Пример 17. *Средняя квадратическая ошибка разности средних при большой выборке.* В табл. 2 были даны распределения по росту группы мужчин и группы женщин; соответствующие средние равны 68,64 и 63,87 дюйма; разность между ними составляет 4,77 дюйма. Дисперсия для роста мужчин равна 7,3861 кв. дюйма; деля ее на 1164, мы находим дисперсию среднего роста 0,006345. Подобно этому, дисперсия роста женщин 6,7832; при делении ее на 1456 получаем дисперсию средней 0,004659. Чтобы определить дисперсию разности этих двух средних, достаточно сложить эти дисперсии, в результате чего получим 0,011004; следовательно, средняя квадратическая ошибка разности между этими средними будет 0,1049 дюйма. Различие в росте мужчин и женщин может быть изображено так:

$$4{,}77 \pm 0{,}105 \text{ дюйма.}$$

Это различие существенное, так как оно превосходит свою среднюю квадратическую ошибку в 45 раз. В данном случае мы можем не только констатировать существенность различия, но с известной степенью уверенности утверждать, что оно находится в интервале между $4^1/_2$ и 5 дюймами. Следует заметить, что мы, подобно многим другим авторам, обрабатывали эти две выборки как *независимые*; однако следует учитывать тот факт, что в эти две группы иногда входят братья и сестры, вместе с тем известно, что имеется определенная тенденция к сходству между братьями и сестрами в отношении роста. Это приводит к тому, что вероятная ошибка разности оказывается завышенной. Наличие связи между членами двух выборок имеет значение и должно учитываться при построении экспериментов. Если говорить языком эксперимента, то сестры являются лучшим «контролем» для их братьев, чем посторонние женщины (см. «Планирование опытов» глава IV). Поэтому различие полов в отношении роста будет более точно определено при сравнении каждого брата с его сестрой. Примерно такой материал использован в следующем примере, построенном на основе корреляционной таблицы для роста братьев и сестер (данные Пирсона и Ли). Он не вполне удовлетворяет указанному выше условию только в том отношении, что в некоторых случаях одно и то же лицо сравнивается не с одной, а с несколькими сестрами или не с одним, а с несколькими братьями.

Пример 18. *Средняя квадратическая ошибка средней разности.* В табл. 26 дается распределение разностей роста братьев и роста сестер; оно включает в себя 1401 попарных сравнений:

Таблица 26

Разность роста в дюймах	−5	−4	−3	−2	−1	0	1	2	3	4	5
Численности	0,25	1,5	1,25	4,5	11,25	27,5	71,75	122,75	171,75	209,75	220,5

Разность роста в дюймах	6	7	8	9	10	11	12	13	14	15	16	Итого
Численности	205,5	148,75	95,75	57	26	11,25	8,5	2,75	1	—	0,75	1401

Обрабатывая эти данные, получим: средняя равна 4,895, дисперсия равна 6,5480, дисперсия средней равна 0,004674 и средняя квадратическая ошибка средней равна 0,0684. Отсюда следует, что разность между ростом мужчин и женщин находится в пределах $4^3/_4$ и 5 дюймов.

В приведенных выше примерах, где было показано применение среднего квадратического отклонения при оценке средних, мы допускали, что дисперсия генеральной совокупности определена нами точно. В 1908 г. Стьюдент показал, что хотя при малых выборках, обычно встречающихся при полевых и лабораторных опытах, дисперсия генеральной совокупности оценивается только грубо, все же ошибки такой оценки поддаются учету и поэтому на основе малых выборок могут быть сделаны достаточно точные заключения.

Если x (например, средняя выборки) распределена нормально около нуля и σ есть ее точная средняя квадратическая ошибка, то вероятность того, что отношение $\frac{x}{\sigma}$ превзойдет некоторое фиксированное значение, находится на основе таблицы нормального распределения. Но если нам неизвестна σ, а вместо нее мы имеем s — оценку σ, то распределение $\frac{x}{s}$ уже не будет нормальным. Правильное значение $\frac{x}{\sigma}$ может быть получено из $\frac{x}{s}$ путем умножения на $\frac{s}{\sigma}$; эта последняя величина и является ошибкой от замены σ на s. В предыдущей главе указывалось, что отношение $\frac{s^2}{\sigma^2}$ распределено в случайных выборках как $\frac{\chi^2}{n}$, где n равно числу степеней свободы, соответствующему s^2. Отсюда следует, что хотя σ нам неизвестна, мы все же можем, основываясь на известном законе распределения $\frac{s}{\sigma}$, воспользоваться доверительным распределением σ при данном s для определения вероятности того, что x в определенное число раз превосходит величину s. Таким образом, знание точного распределения $\frac{x}{s}$ это все, что нам нужно для решения задачи. Это распределение характерно тем, что оно меняется в зависимости от

n — числа степеней свободы, соответствующего оценке s. Такие распределения для различных n были определены в 1908 г. Стьюдентом; в дальнейшем другими авторами были даны более полные таблицы; в конце настоящей главы (стр. 144) приводятся эти распределения в той же форме, что и наша таблица χ^2.

24. Существенность средней в единичной выборке

Если $x_1, x_2, \ldots\ldots x_{n'}$ — члены выборки некоторой переменной x и если эта выборка содержит в себе все доступные нам сведения относительно интересующего нас вопроса, то прежде всего следует установить, сколь существенно отклонение среднего значения x от нуля. Для этого вычисляем статистики:

$$\bar{x} = \frac{1}{n'} S(x);$$

$$\frac{s^2}{n'} = \frac{1}{n'(n'-1)} S(x-\bar{x})^2;$$

$$t = \bar{x} : \sqrt{\frac{s^2}{n'}};$$

$$n = n' - 1.$$

Эти вычисления могут быть упрощены, если воспользоваться тем, что сумма квадратов отклонений от средней равна разности суммы квадратов отклонений от нуля и произведения суммы на среднюю, так как

$$S(x^2) = S(x-\bar{x})^2 + \bar{x}S(x).$$

В данном случае произведено разложение суммы квадратов x на две части, из которых первая характеризует изменчивость внутри выборки, а вторая относится к отклонению средней от нуля. Первая часть имеет $n-1$ степеней свободы, а вторая — только одну степень свободы.

В более сложных случаях анализа, с которыми мы познакомимся в последующих главах, удобно применять таблицу с колонками, куда вносятся последовательно степени свободы, суммы квадратов и сравнимые между собой средние квадраты. Так, в нашем случае эта таблица будет иметь следующий вид:

	Степени свободы	Сумма квадратов	Средний квадрат
Отклонение $\bar{x}$ от 0	1	$\bar{x}S(x)$	t^2s^2
Отклонения внутри выборки .	$n-1$	$S(x-\bar{x})^2$	s^2
Итого . . .	n	$S(x^2)$	—

Средний квадрат каждой строки здесь получен путем деления суммы квадратов на соответствующее число степеней свободы. Отношение средних квадратов в данном случае равно t^2. Эта форма расположения результатов анализа, получившая название таблицы дисперсионного анализа, применяется гораздо чаще, чем выражение этих результатов в форме алгебраических равенств.

Величина t в случайных выборках из нормальной совокупности распределена около нуля по определенному закону, зависящему от n. В таблице значений t (стр. 144) последовательные колонки относятся к значениям 0,9 . . . 0,01 вероятности P, — вероятности выхода за пределы интервала $\pm t$. Так, последняя колонка указывает, что при $n=10$ можно ожидать один процент таких случайных выборок, у которых t будет больше $+3{,}169$ или меньше $-3{,}169$. Если задача исследования требует знания вероятности превзойти некоторое значение t только в положительном (или только в отрицательном) направлении, то значение P следует разделить пополам. При ознакомлении с этой таблицей можно видеть, что для одной и той же степени уверенности приходится брать тем большие значения t, чем меньше степеней свободы n. В нижней строке таблицы, соответствующей бесконечному значению n, приведены значения t, относящиеся к нормальному распределению, т. е. отклонения этого распределения, деленные на среднее квадратическое отклонение σ.

Пример 19. *Существенность средней при малой выборке.* Следующие данные (Кушни и Пиблса), которые я беру из работы Стьюдента, получены в опыте по изучению влияния двух наркотических снотворных средств А и В на продолжительность сна у 10 пациентов.

Таблица 27

Дополнительные часы сна, полученные в результате применения двух испытываемых снотворных средств

Пациенты	А	В	Разность (В — А)
1	+0,7	+1,9	+1,2
2	−1,6	+0,8	+2,4
3	−0,2	+1,1	+1,3
4	−1,2	+0,1	+1,3
5	−0,1	−0,1	0,0
6	+3,4	+4,4	+1,0
7	+3,7	+5,5	+1,8
8	+0,8	+1,6	+0,8
9	0,0	+4,6	+4,6
10	+2,0	+3,4	+1,4
Средняя ($\bar{x}$) . . .	+0,75	+2,33	+1,58

Последняя колонка дает контрольное сравнение эффективности двух снотворных средств, испытанных на одних и тех же пациентах. Для этого ряда разностей находим

$$\bar{x} = +1{,}58;$$

$$\frac{s^2}{10} = 0{,}1513;$$

$$\frac{s}{\sqrt{10}} = 0{,}3890;$$

$$t = 4{,}06.$$

При $n=9$ только одно значение t из ста может превзойти 3,250, и поэтому эта средняя разность, безусловно, существенна. Если в этом случае применить методы, описанные в предыдущих главах, то мы придем почти с такой же уверенностью к этим же выводам. Так, если бы эти два средства были одинаково эффективны, то в последней колонке положительные и отрицательные знаки должны были бы встречаться одинаково часто. Однако у нас 9 значений положительны и одно нулевое, а из биноминального распределения

$$\left(\frac{1}{2}+\frac{1}{2}\right)^9$$

следует, что имеется только два шанса из 512 за то, чтобы получить все отклонения одного и того же знака. Метод, примененный в настоящей главе, отличается от только что изложенного тем, что в нем использованы абсолютные значения разностей, а не только их знаки. Следовательно, он предназначен для случаев, когда даны количественные значения переменной.

24.1. Сравнение двух средних

В условиях эксперимента очень часто возникает необходимость определить, является ли различие двух выборок в отношении их средних существенным или, наоборот, следует считать, что они происходят из одной и той же генеральной совокупности. В последнем случае не будет существенным и различие между вариантами опыта, связанными с этими выборками.

Если $x_1, x_2, \ldots x_{n_1+1}$ и $x'_1, x'_2, \ldots x'_{n_2+1}$ представляют собой две выборки, то существенность разности между их средними может быть определена путем вычисления следующих статистик:

$$\bar{x} = \frac{1}{n_1+1} S(x); \quad \bar{x}' = \frac{1}{n_2+1} S(x');$$

$$s^2 = \frac{1}{n_1+n_2}\{S(x-\bar{x})^2 + S(x'-\bar{x}')^2\};$$

$$t = \frac{\bar{x}-\bar{x}'}{s}\sqrt{\frac{(n_1+1)(n_2+1)}{n_1+n_2+2}};$$

$$n = n_1 + n_2.$$

Здесь средние вычисляются, как обычно: среднее квадратическое отклонение оценивается путем объединения сумм квадратов обеих выборок и делением на общее число степеней свободы. Если σ будет точным средним квадратическим отклонением, то дисперсия первой средней должна быть $\frac{\sigma^2}{(n_1+1)}$, а второй средней $\frac{\sigma^2}{n_2+1}$ и поэтому дисперсия разности средних будет $\sigma^2\left[\frac{1}{(n_1+1)}+\frac{1}{(n_2+1)}\right]$. Величина t находится делением разности $\bar{x}-\bar{x}'$ на ее среднюю квадратическую ошибку, определенную выше. Ошибка этой оценки разности средних устанавливается при помощи таблицы t, в которой следует брать n, равное числу степеней свободы, относящемуся к оценке s, т. е. $n=n_1+n_2$. Так появляется возможность применить способ обработки Стьюдента, предназначенный для оценки средней, к оценке разности двух выборочных средних.

Рассмотрим вопрос о построении таблицы дисперсионного анализа для этого случая. Если мы возьмем такие таблицы для каждой из двух выборок в отдельности, а потом сложим их по частям, то получим:

	Степени свободы	Сумма квадратов
Отклонения	2	$\bar{x}S(x)+\bar{x}'S(x')$
Внутри выборок	n_1+n_2	$S(x-\bar{x})^2+S(x'-\bar{x}')^2$
Итого . . .	n_1+n_2+2	$S(x^2)+S(x'^2)$

Но если мы обработаем все наблюдения как единую выборку со средней m, то должны получить:

	Степени свободы	Сумма квадратов
Отклонения	1	$mS(x)+mS(x')$
Внутри выборок	n_1+n_2+1	$S(x-m)^2+S(x'-m)^2$
Итого . . .	n_1+n_2+2	$S(x^2)+S(x'^2)$

Здесь произведено два различных разложения одной и той же суммы квадратов; но так как все сравнения внутри отдельных выборок являются в то же время сравнениями внутри объединенной выборки, то путем вычитания одного из другого можно получить:

	Степени свободы	Сумма квадратов
Отклонения	1	$\bar{x}S(x)+\bar{x}'S(x')-mS(x)-mS(x')$
Внутри выборок . .	n_1+n_2	$S(x-\bar{x})^2+S(x'-\bar{x}')^2$
Итого . . .	n_1+n_2+1	$S(x^2)+S(x'^2)-mS(x)-mS(x')$

Каждая часть этого анализа может быть легко вычислена. Читатель может проверить, что величина t^2, полученная при нашем первом расчете, является ничем иным, как отношением средних квадратов из последней таблицы дисперсионного анализа.

Изложенный здесь метод дисперсионного анализа основан на объединении оценок дисперсий, относящихся к двум выборкам. В связи с этим следует отметить, что различие двух генеральных совокупностей, к которым принадлежат эти две выборки, может заключаться только в различии их дисперсий, и это иногда может приводить к выходу значения t за критические границы. Поэтому настоящий критерий, если значение t существенно, разрешает общий вопрос о том, что данные выборки не могут относиться к одной и той же совокупности, и вполне возможен такой случай, когда существенность значения t определяется различием дисперсий, но не средних. Этот теоретически возможный случай в применении к опытным данным не представляет интереса. Но если это необходимо, то в таких случаях всегда может быть применен описанный в параграфе 41 дополнительный критерий, устанавливающий существенность различий между дисперсиями.

Здесь следует еще остановиться на возможном заблуждении, будто бы метод, изложенный в этом параграфе, основывается на «допущении» равенства двух дисперсий. Это неправильно сформулированное положение, так как равенство дисперсий является неотъемлемой частью проверяемой гипотезы о том, что данные две выборки принадлежат к одной и той же совокупности.

Поэтому правомерность критерия t, как критерия именно этой гипотезы, абсолютна и не требует никакого допущения. Конечно, вполне законно требование об отыскании другого критерия существенности, отвечающего на вопрос «Могут ли эти две выборки принадлежать к различным совокупностям, имеющим одинаковые средние?» Эта задача была фактически поставлена и разрешена, но в реальных условиях, встречающихся в биологических исследованиях, этот вопрос представляет собой только академический интерес. Числовые таблицы такого критерия были впервые вычислены Сукхатме; они применяются в тех случаях, когда желают устранить какие-либо сомнения в интерпретации, связанной с нашим критерием существенности, и когда в то же время есть основание ожидать неравенства дисперсий (см. «Статистические таблицы» V1 и V2).

Пример 20. *Существенность разности средних при малых выборках.* Допустим, что данные табл. 27 являются результатом испытания двух снотворных средств не на одних и тех же, а на разных пациентах. В этом случае эксперимент не так хорошо контролируем, и поэтому следует ожидать менее явные результаты при том же числе наблюдений, ибо априорно можно считать, и приведенные ранее данные подтверждают это, что каждое отдельное лицо имеет свою собственную реакцию на снотворное

средство, и эта реакция примерно одинакова в отношении обоих снотворных.

Обрабатывая эти данные так, как будто они были получены при наблюдении над двумя различными рядами пациентов, находим:

$$\bar{x} - \bar{x}' = +1{,}58;$$

$$s^2\left(\frac{1}{10}+\frac{1}{10}\right) = 0{,}7210;$$

$$t = +1{,}861;$$

$$n = 18.$$

Здесь, следовательно, значение P находится между 0,1 и 0,05 и не может считаться существенным. Этот пример со всей очевидностью устанавливает то значение, которое имеет правильное построение экспериментов, в особенности при их небольших размерах, а также показывает, что эффективность такого построения может быть вполне правильно оценена соответствующей статистической обработкой.

Знакомство с распределением Стьюдента дает нам возможность уяснить себе всю ценность повторных наблюдений. Значение повторных наблюдений состоит не только в том, что при этом происходит уменьшение средней квадратической ошибки средней обратно пропорционально корню квадратному из числа параллельных наблюдений, но и в том, что одновременно с этим становится более точной наша оценка этой средней квадратической ошибки. Необходимость дублирования наблюдений достаточно широко известна и общепринята, но в некоторых случаях у исследователей нет достаточно ясного понимания того, что трехкратное повторение эксперимента, если при этом выбран довольно высокий уровень уверенности (положим 0,01), позволяет обнаружить существенные разности, примерно в семь раз меньшие, чем те существенные разности, которые при том же уровне уверенности можно установить при двухкратной повторности эксперимента.

Уверенность, с которой принимается нами тот или иной результат, зависит не только от фактического размера средней, но в той же мере и от согласованности между повторными наблюдениями. Так, если в некотором сельскохозяйственном опыте в первом повторении была получена прибавка урожая в 8 бушелей на акр, а во втором повторении — 9 бушелей, то мы имеем $n=1$ и $t=17$. Этот результат дает основание более или менее уверенно считать, что в опыте был установлен определенный эффект изучаемого мероприятия. Но если во втором повторении была получена прибавка не в 9, а в 18 бушелей, то хотя средняя здесь будет более высокой, все же мы будем иметь меньшую уверенность в том, что изучаемое мероприятие дало положительный эффект, так как теперь t падает до 2,6, т. е. до такого значения, которое при $n=1$ очень часто может быть превзойдено просто в силу слу-

чайности. Этот кажущийся парадокс находит свое объяснение в том, что разность между параллельными наблюдениями в 10 бушелей означает присутствие таких неконтролируемых условий, при которых оба наблюдения могут появиться в качестве случайных. В первом же случае довольно хорошая согласованность параллельных наблюдений показывает, что неконтролируемые факторы не оказали на результаты большого влияния. Особая роль повышенной повторности наблюдений обусловлена тем обстоятельством, что при проведении небольшого числа повторных наблюдений и тем более, когда этих наблюдений только два, наша оценка удельного веса неконтролируемых факторов становится весьма грубой.

В тех случаях, когда между членами двух серий наблюдений имеется строгое соответствие, вполне допустима обработка ряда разностей между соответствующими членами серий, подобно тому, как это было сделано в примере 19. Однако этот путь не всегда может быть приемлемым, так как более высокая точность сравнения получается при данном методе только тогда, когда между двумя рядами наблюдений существует положительная корреляция.

Более того, эта положительная корреляция должна быть такого размера, чтобы компенсировать уменьшение точности, возникающее в связи с наличием меньшего числа степеней свободы, на котором основывается наша оценка дисперсии. Рассмотрим этот вопрос на примере.

Пример 21. *Существенность изменений в численности бактерий.* В следующей таблице приведены средние количества бактериальных колоний на пластинку. Эти данные получены при помощи четырех различных методов взятия почвенных проб соответственно в 4 часа дня и 8 часов вечера (данные Торнтона).

Таблица 28

Метод	в 4 часа дня	в 8 часов вечера	Разность
A	29,75	39,20	+9,45
B	27,50	40,60	+13,10
C	30,25	36,20	+5,95
D	27,80	42,40	+14,60
Средняя	28,825	39,60	+10,775

Обрабатывая ряд разностей, находим $\bar{x}=+10{,}775$; $^1/_4 s^2=3{,}756$; $t=5{,}560$ и $n=3$, откуда по таблице устанавливаем, что P содержится между 0,01 и 0,02. Если же воспользоваться методом из примера 20 и обработать два отдельных ряда наблюдений, то можно найти: $\bar{x}-\bar{x}'=+10{,}775$; $^1/_2 s^2=2{,}188$; $t=7{,}285$ и $n=6$. Здесь получено не только большее значение n, но и большее значение t, которое теперь выходит из рамок таблицы, указывая тем самым на чрезвычайно малое значение вероятности P. В данном случае различия методов внутри каждого ряда наблюдений незначительны и к тому же оказалось, что эти различия не координируются друг с другом, вследствие чего сравнение каждого члена одного ряда с противостоящим ему членом другого ряда приводит только к снижению точности. В случаях, подобных данным, иногда бывает так, что один метод обработки наблюдений не дает существенной разности, в то время как другой приводит к существенной разности. Это противоречие разрешается так: если какой-либо метод определенно указывает на существенность результатов, то его показания нельзя игнорировать, даже если другой метод не обнаруживает этой существенности. Вообще же, если между членами одного ряда данных и членами другого ряда нет никакого соответствия, то применим только второй метод обработки.

25. Коэффициенты регрессии

Методы, рассматриваемые в настоящей главе, применимы не только к средним величинам в узком смысле, но и к более широкому классу статистик, известных под названием коэффициентов регрессии. Понятие о регрессии обычно вводится в связи с теорией корреляции, но на самом деле оно более широкое и более простое. Более того, коэффициенты регрессии представляют интерес и научную ценность во многих случаях, когда коэффициент корреляции превращается в искусственный показатель, не имеющий реального смысла. Общее обсуждение вопроса, к которому мы сейчас перейдем, имеет целью познакомить читателя с учением о регрессии и подготовить почву для рассмотрения числовых примеров.

Известно, что рост ребенка зависит от его возраста, но несмотря на это, зная возраст ребенка, мы все же не можем точно определить его рост. В каждом возрасте рост имеет различные значения, образующие внутри некоторого интервала распределение, характерное для данного возраста. Сводные показатели этого распределения, такие, как средняя, будут являться непрерывной функцией возраста. Функция, которая характеризует зависимость среднего роста от возраста, называется функцией регрессии роста на возраст; графически она может быть изображена в виде кривой, или линии регрессии. В отношении такой линии регрессии возраст выступает в качестве *независимой* переменной, а рост — *зависимой* переменной.

Эти две переменные — зависимая и независимая — имеют различное отношение к линии регрессии. Если при измерении роста будут иметь место случайные ошибки, то это не повлияет на линию регрессии роста на возраст, так как во всех возрастах положительные и отрицательные ошибки встречаются примерно одинаково часто и поэтому они в среднем компенсируют друг друга.

Наоборот, ошибки в возрасте, вообще говоря, окажут свое влияние на регрессию роста на возраст; поэтому, производя записи возраста с погрешностями или допуская слишком грубую группировку по возрасту, мы можем получить неправильное соотношение между средним ростом и возрастом. Следует отметить и второе различие. Функция регрессии не зависит от формы распределения независимой переменной, и поэтому правильная линия регрессии может быть получена даже в том случае, когда возрастные группы выбираются произвольно, как это имеет место при обследовании детей «школьного возраста». Наоборот, такой же выбор групп для зависимой переменной может полностью видоизменить линию регрессии.

Из этих двух положений со всей очевидностью следует, что регрессия роста на возраст во всех отношениях отлична от регрессии возраста на рост и что в то время как одна из них может иметь определенный физический смысл, другая будет иметь исключительно условное содержание, связанное только с ее математическим определением. Но в ряде случаев обе регрессии двух переменных в одинаковой степени содержательны. Так, если мы выразим средний рост сыновей в зависимости от роста отцов, то наблюдения покажут, что каждому добавочному дюйму в росте отцов соответствует примерно полдюйма в среднем росте сыновей. Подобно этому, если мы определим средний рост отцов для сыновей данного роста, то мы найдем, что каждому добавочному дюйму в росте сыновей соответствует полдюйма в среднем росте отцов. Здесь нет оснований для предпочтительного выбора между ростом отцов и ростом сыновей, так как каждая из этих переменных распределена нормально и вместе с другой переменной образует распределение нормального типа. Обе эти линии регрессии прямолинейны и вследствие этого появляется возможность выразить наличие их связи в тех простых соотношениях, которые были приведены выше.

Когда линия регрессии является прямой, или, другими словами, когда функция регрессии линейна, все характерные черты регрессии значительно упрощаются, так как, кроме общей средней, нам следует установить только отношение, в котором находится увеличение средней у зависимой переменной к увеличению независимой переменной. Это отношение называется коэффициентом регрессии. В данном случае функция регрессии имеет вид:

$$Y = a + b(x - \bar{x}),$$

где b — коэффициент регрессии y на x, а Y — ожидаемое значение y для каждого отдельного значения x. Физический смысл коэффициента регрессии зависит от содержания переменных; так, в регрессии роста на возраст коэффициент регрессии выражается в дюймах роста за год и является средней прибавкой роста, а в регрессии роста отцов на рост сыновей коэффициент регрессии выражается отношением полдюйма к дюйму, т. е. он просто является отвлеченным числом $^1/_2$. Конечно, коэффициент регрессии может быть как положительным, так и отрицательным числом.

В общем случае линия регрессии представляет собой кривую линию, которая может быть выражена, например, такой функцией регрессии:

$$Y = a + bx + cx^2 + dx^3,$$

где все четыре коэффициента могут быть названы коэффициентами регрессии, если придать этому термину более широкое толкование. Вообще говоря, могут быть и более сложные функции x, но применение их на практике встречает то затруднение, что не всегда имеются теоретические основания для выбора той или иной формы этой функции; поэтому, как правило, на практике применяются более простые степенные ряды (или полиномы x). В статистической практике наибольшее значение имеет прямолинейная регрессия.

26. Ошибки выборки у коэффициентов регрессии

Уравнение линейной регрессии содержит два параметра, подлежащие оценке на основе данных наблюдения. Если взять это уравнение в форме

$$Y = a + b(x - \bar{x}),$$

то величина a будет просто средней из наблюденных значений зависимой переменной $\bar{y}$. Отсюда следует, что сумма значений $b(x - \bar{x})$ всегда равна нулю, каким бы ни было значение b.

Оценка коэффициента регрессии y на x, т. е. b, получается на основе суммы произведений x и y. При определении оценки дисперсии для одной переменной последовательность действий была такой: сначала определяется сумма квадратов отклонений от средней путем вычитания из суммы квадратов x произведения $n\bar{x}^2$, т. е. по формуле:

$$S\{(x - \bar{x})^2\} = S(x^2) - n\bar{x}^2,$$

после чего эта сумма делится на $(n - 1)$, что и дает оценку дисперсии. Подобно этому, при двух переменных x и y можно получить сумму произведений отклонений от средних путем вычитания $n\bar{x}\bar{y}$, т. е. по формуле:

$$S\{(x - \bar{x})(y - \bar{y})\} = S(xy) - n\bar{x}\bar{y}.$$

Среднее произведение двух переменных, если последние выражены в отклонениях от их средних, называется *ковариацией*, которая, подобно дисперсии одной переменной, определяется путем деления на $(n - 1)$. Сумма произведений, на основе которой определяется ковариация, может быть представлена также в одной из таких двух форм:

$$S\{y(x - \bar{x})\} \text{ или } S\{x(y - \bar{y})\}.$$

Оценкой коэффициента регрессии b будет отношение ковариации двух переменных к дисперсии независимой переменной. Если сократить делитель $(n-1)$, который участвует в обоих членах этого отношения, то эту оценку можно представить такой формулой:

$$b = \frac{S\{y(x-\bar{x})\}}{S\{(x-\bar{x})^2\}}.$$

Таким образом, мы имеем возможность вычислить по наблюденным данным оценки двух параметров, определяющих прямую линию. Точная формула регрессии, которая была бы получена при бесконечно большом числе наблюдений, представляет собой уравнение

$$Y = \alpha + \beta(x-\bar{x}).$$

Разности $a-\alpha$ и $b-\beta$ являются ошибками случайного отбора.

Для определения размера этих ошибок выборки рассмотрим бесконечную совокупность выборок, имеющих одни и те же значения x. Изменчивость наших статистик a и b от выборки к выборке будет определяться тем, что для каждого данного значения x все значения y в совокупности выборок не будут равны друг другу. Если σ^2 является дисперсией y при данном значении x, то очевидно, что ошибкой коэффициента a является просто средняя из n' независимых ошибок, имеющих одну и ту же дисперсию σ^2. Поэтому дисперсия статистики a равна $\frac{\sigma^2}{n'}$. Вторая статистика b также является линейной функцией величины y и ее выборочная дисперсия может быть поэтому получена на основе таких соображений. В этом случае каждое отклонение y от правильной линии регрессии умножается на $x-\bar{x}$, поэтому дисперсия такого произведения равна $\sigma^2(x-\bar{x})^2$, а дисперсия суммы произведений, стоящей в числителе формулы для коэффициента b, должна быть равна:

$$\sigma^2 S(x-\bar{x})^2.$$

Коэффициент b определяется делением указанной выше суммы произведений на $S[(x-\bar{x})^2]$ и, следовательно, дисперсия b может быть получена путем деления дисперсии числителя на $S^2[(x-\bar{x})^2]$, что приводит к выражению

$$\frac{\sigma^2}{S(x-\bar{x})^2}.$$

Следует отметить, что величина, установленная в качестве выборочной дисперсии a, не является просто выборочной дисперсией нашей оценки для средней из значений y, а представляет собой оценку средней из тех значений y, которые относятся к фиксированному значению x, т. е. к такому x, которое остается неизменным от выборки к выборке. Это различие, которое на первый взгляд кажется слишком незначительным и неуловимым, все же следует всегда иметь в виду; например, на основе измерений роста школьников можно произвести оценку среднего роста десятилетнего школьника и средний рост всех школьников; эти оценки будут совпадать только тогда, если средний возраст всех школьников равен точно десяти годам. Первая средняя, как правило, будет иметь более точную оценку, чем вторая, ибо в ней элиминирована изменчивость школьного возраста, что, несомненно, вносит с собой нечто добавочное в изменчивость среднего роста школьников.

Для того чтобы произвести при помощи критерия существенности оценку разности между фактическим значением b и некоторым гипотетическим значением β, следует прежде всего произвести оценку σ^2. В данном случае лучшей оценкой будет

$$s^2 = \frac{1}{n'-2} S(y-Y)^2.$$

Эта оценка находится путем суммирования квадратов отклонений y от значений Y, вычисленных по уравнению регрессии, и делением этой суммы на $(n'-2)$. Основанием того, что делителем является $(n'-2)$ служит то обстоятельство, что по n' значениям величины y уже определены две статистики a и b, которые вошли в формулу Y. Следовательно, ряд разностей $y-Y$ фактически имеет только $n'-2$ степени свободы.

Когда n' мало, то оценка s^2, определенная выше, будет в известной мере ненадежной. Поэтому при сравнении разности $b-\beta$ с ее средней квадратической ошибкой в целях установления существенности этой разности следует воспользоваться методом Стьюдента, взяв $n=n'-2$. Когда же n' велико, то распределение t будет примерно нормальным. Значение t, необходимое для определения по таблице вероятности P, находится путем деления разности $(b-\beta)$ на ее среднюю квадратическую ошибку, т. е.

$$t = \frac{(b-\beta)\sqrt{S(x-\bar{x})^2}}{s}.$$

Для определения же существенности разности между a и некоторым гипотетическим значением α тем же путем находим:

$$t = \frac{(a-\alpha)\sqrt{n'}}{s}; \quad n = n'-2.$$

Этот критерий существенности a будет более правомерен, чем критерий, в котором оставляется без внимания наличие связи, обусловленной тем, что варьирование y в той или иной мере определяется изменчивостью x. В этом случае значение s, определенное на основе регрессии, будет меньше того значения, которое определяется по непосредственным наблюдениям. Но, с другой

стороны, при этом всегда теряется одна степень свободы и поэтому, если b мало, то повышения точности может и не быть.

В общем случае, когда стоит задача оценить значение зависимой переменной, соответствующее любому значению независимой переменной, отличающемуся от средней, то, как показали Уоркинг и Хотеллинг, необходимо определить выборочную дисперсию величины:

$$Y = a + b(x - \bar{x}).$$

Так как выборочные ошибки a и b независимы, то для дисперсии Y имеем

$$V(a) + (x - \bar{x})^2 V(b),$$

где $V(a)$ и $V(b)$ — соответствующие выборочные дисперсии наших оценок a и b.

Следовательно, мы имеем

$$V(Y) = \sigma^2\left(\frac{1}{n'} + \frac{(x - \bar{x})^2}{S(x - \bar{x})^2}\right),$$

где σ^2 — точная дисперсия y для данного x. Если значение x близко к средней и поэтому разность $(x - \bar{x})$ мала, то эта дисперсия будет только незначительно превосходить дисперсию данной средней выборки, но для значений x, более удаленных от центра, второй компонент ошибки, связанный с оценкой b, начнет играть преобладающую роль, вследствие чего точность нашей оценки снизится.

Пример 22. *Влияние азотных удобрений на урожай в многолетнем опыте.* Урожаи зерна пшеницы в бушелях на акр, представленные в табл. 29, были получены с двух делянок участка Бродболк в Ротамстеде в течение тридцати лет. Между этими делянками различие заключалось только в том, что на делянку «9a» была внесена селитра, а на делянку «7b» — эквивалентное количество азота в форме сульфата аммония. В продолжение опыта делянка «9a» перегоняла по урожайности делянку «7b». Можно ли считать факт прогрессирующего различия между урожаями этих двух делянок существенным?

В большинстве случаев направление изменения урожайности по годам у обеих делянок одинаково; следовательно, наиболее отчетливые результаты опыта будут определяться рядом разностей этих урожаев. В данном случае имеется еще одно обстоятельство, упрощающее обработку данных, которое состоит в том, что тридцать значений независимой переменной (времени) образуют ряд с одинаковыми интервалами между последовательными значениями и что каждому из этих значений независимой переменной соответствует только одно значение зависимой переменной. В таких случаях обработка упрощается в связи с тем, что можно воспользоваться формулой

$$S(x - \bar{x})^2 = \frac{n'(n'^2 - 1)}{12},$$

где n' — число значений независимой переменной; у нас $n' = 30$.

Для определения коэффициента b следует вычислить сумму произведений

$$S[y(x - \bar{x})],$$

которая находится в таком же отношении к ковариации двух переменных, в каком находится сумма квадратов к дисперсии при одной переменной. Эту сумму произведений можно определить не-

Таблица 29

Год	9a	7b	9a—7b
1855	29,62	33,00	−3,38
1856	32,38	36,91	−4,53
1857	43,75	44,84	−1,09
1858	37,56	38,94	−1,38
1859	30,00	34,66	−4,66
1860	32,62	27,72	+4,90
1861	33,75	34,94	−1,19
1862	43,44	35,88	+7,56
1863	55,56	53,66	+1,90
1864	51,06	45,78	+5,28
1865	44,06	40,22	+3,84
1866	32,50	29,91	+2,59
1867	29,13	22,16	+6,97
1868	47,81	39,19	+8,62
1869	39,00	28,25	+10,75
1870	45,50	41,37	+4,13
1871	34,44	22,31	+12,13
1872	40,69	29,06	+11,63
1873	35,81	22,75	+13,06
1874	38,19	39,56	−1,37
1875	30,50	26,63	+3,87
1876	33,31	25,50	+7,81
1877	40,12	19,12	+21,00
1878	37,19	32,19	+5,00
1879	21,94	17,25	+4,69
1880	34,06	34,31	−0,25
1881	35,44	26,13	+9,31
1882	31,81	34,75	−2,94
1883	43,38	36,31	+7,07
1884	40,44	37,75	+2,69
Средняя . .	37,50	33,03	+4,47

$$S(x - \bar{x})^2 = \frac{n'(n'^2 - 1)}{12} = 2247{,}5$$

$$b = 0{,}26679$$

$$S(y - \bar{y})^2 = 1020{,}56$$

$$b^2 S(x - \bar{x})^2 = 159{,}97$$

$$S(y - Y)^2 = 860{,}59$$

$$s^2 = 30{,}74$$

$$\frac{s^2}{S(x - \bar{x})^2} = 0{,}013675 = (0{,}11694)^2$$

$$t = 2{,}2814$$

$$n = 28$$

сколькими способами. Во-первых, можно умножить последовательно значения y на $-29, -27, \ldots\ldots +27, +29$, сложить произведения и разделить сумму на 2. Этот метод соответствует приведенной ранее формуле. Тот же результат можно получить путем умножения y на $1,2,\ldots\ldots 30$ и вычитания из этой суммы произведений суммы значений y, умноженной на $15^1/_2 = \left(\frac{n'+1}{2}\right)$. Этот последний метод может быть проведен в более удобной форме последовательного суммирования. Начиная снизу, путем присоединения каждого последующего значения к накопленной уже ранее сумме определяются и выписываются в новую колонку последовательные суммы: 2,69; 9,76; 6,82 и т. д. Эта колонка суммируется и из суммы вычитается сумма предыдущей колонки, умноженная на $15^1/_2$. В результате будет получена искомая величина. В нашем случае имеем 599,615 и, деля на 2247,5, определяем $b=0{,}26679$. Таким образом, делянка «9a» с течением времени перегоняет по своей урожайности делянку «7b» примерно на одну четверть бушеля на акр в год.

Для определения средней квадратической ошибки коэффициента b следует найти сумму квадратов отклонений y от значений Y, устанавливаемых по уравнению регрессии,

$$S(y-Y)^2,$$

т. е. следует оценить остаточную вариацию. Зная коэффициент b, можно вычислить тридцать значений Y по формуле

$$Y = \bar{y} + (x - \bar{x})\, b.$$

При вычислении первого значения Y следует взять $x - \bar{x} = -14{,}5$, а все остальные значения Y можно найти путем последовательного прибавления b. Вычитая каждое значение Y из соответствующего y, возводя эти разности в квадрат и суммируя, можно определить искомую величину непосредственно по указанной выше формуле. Однако, этот путь довольно трудоемок и поэтому на практике предпочитают пользоваться следующим алгебраическим тождеством

$$S(y-Y)^2 = S(y-\bar{y})^2 - b^2 S(x-\bar{x})^2 =$$
$$= S(y^2) - n'\bar{y}^2 - b^2 S(x-\bar{x})^2.$$

Обработка данных в этом случае состоит в следующем: возводятся в квадрат все y и квадраты суммируются, из этой суммы вычитаются две величины, которые определяются по средней y и по величине b. В связи с этим методом упрощенного вычисления следует отметить, что даже небольшие погрешности у средней $\bar{y}$ и коэффициента b могут привести к значительным ошибкам в окончательных результатах, и поэтому здесь необходимо вычислять эти величины по возможности точно и с достаточным числом десятичных знаков. Дело в том, что погрешности, которые при первом методе не оказывают большого влияния на результаты вычислений, при втором методе могут совсем их исказить.

Последующую работу по вычислению средней квадратической ошибки коэффициента b лучше всего провести по схеме, приведенной в табл. 29. Оценка среднего квадратического отклонения равна 0,1169, и поэтому для проверки гипотезы о том, что $\beta=0$, т. е. что делянка «9a» не перегоняет со временем по своей урожайности делянку «7b», следует разделить b на 0,1169, в результате чего получится $t=2{,}2814$. Так как величине s соответствует 28 степеней свободы, то $n=28$, и, следовательно, наше значение t показывает, что P содержится между 0,02 и 0,05.

Этот результат может быть принят в качестве существенного, хотя и с некоторой осторожностью. Рассматривая исходные данные, мы не должны, конечно, игнорировать тот факт, что на этом участке совместно с другими удобрениями селитра значительно лучше поддерживает плодородие почвы, чем сульфат аммония, однако имеющиеся данные все же не дают безусловного подтверждения этого вывода.

Средняя квадратическая ошибка для $\bar{y}$, вычисленная по этим данным, равна 1,012; так что нет никакого сомнения в том, что разность между средними урожаями: существенна.

Если бы мы поставили задачу оценить существенность средней без учета порядка, в каком расположены y, т. е. вычислили бы s^2 делением 1020,56 на 29, то получили бы среднюю квадратическую ошибку 1,083. Следовательно, значение коэффициента регрессии b оказалось достаточно высоким, чтобы уменьшить среднюю квадратическую ошибку. В данном случае вполне возможно, что при определении по этому материалу более сложной линии регрессии вероятная ошибка уменьшится еще более и приведет к тому, что коэффициент b станет существенным без всяких оговорок. Вопрос об определении криволинейной регрессии будет рассмотрен в дальнейшем.

26.1. Сравнение коэффициентов регрессии

Ранее при сравнении двух средних, полученных из выборок различного размера, был применен метод определения средней квадратической ошибки, основанный на объединении сумм квадратов, вычисленных по каждой из этих выборок в отдельности. Этот же прием можно применить и при сравнении коэффициентов регрессии, исчисленных на основе двух выборок, у которых ряды наблюдений над независимой переменной неодинаковы или, как частный случай, у которых эти ряды тождественны.

Пример 23. *Сравнение показателей относительного роста двух культур морской водоросли.* В табл. 30 даны десятичные ло-

гарифмы объемов, занятых в течение следующих друг за другом дней клетками морской водоросли. Наблюдения велись в течение таких периодов, когда относительный рост оставался примерно одинаковым: над культурой А наблюдение велось в течение девяти дней, над культурой же В — в течение восьми дней (данные Бристоль-Рога).

Вычисление $Sy(x-\bar{x})$ по способу нарастающих сумм дано в третьем и четвертом столбцах: первоначальные данные, начиная снизу, складываются, в результате чего получаются последовательные суммы от 6,087 до 43,426 (у культуры А); последняя сумма должна, конечно, совпасть с итогом соответствующих исходных данных. Из итога таких накопленных сумм вычитается сумма исходных данных, умноженная на 5 для культуры А и на $4^1/_2$ для культуры В. Эта разность и будет величиной $Sy(x-\bar{x})$; ее следует разделить на соответствующую сумму $S(x-\bar{x})^2$, т. е. соответственно, для двух наших культур, на 60 и 42, в результате чего и будет получен коэффициент регрессии b, характеризующий относительный рост культуры.

Чтобы произвести оценку существенности разности между вычисленными этим способом коэффициентами регрессии, следует для каждого из двух случаев вычислить $S(y^2)$, откуда вычесть произведение средней на соответствующую сумму и произведение b на $Sy(x-\bar{x})$. В результате таких расчетов получается две величины $S(y-Y)^2$, которые следует сложить и разделить на n,

Таблица 30

	Логарифмы объемов		Накопленные суммы		
	A	B	A	B	
	3,592	3,538	43,426	38,358	$S(y-Y)^2$: A = 0,05089
	3,823	3,828	39,834	34,820	B = 0,07563
	4,174	4,349	36,011	30,992	$ns^2 = 0,12652$
	4,534	4,833	31,837	26,643	$s^2 = 0,009732$
	4,956	4,911	27,303	21,810	$\frac{s^2}{60} = 0,0001622$
	5,163	5,297	22,347	16,899	
	5,495	5,566	17,184	11,602	$\frac{s^2}{42} = \frac{0,0002317}{0,0003939}$
	5,602	6,036	11,689	6,036	
	6,087	—	6,087	—	
Итого . . .	43,426	38,358	235,718	187,160	Ср. кв. ошибка = 0,01985
Средняя . .	4,8251	4,7947	217,130	172,611	$b' - b = 0,0366$
		$Sy(x-\bar{x}) = 18,588$		14,549	$t = 1,844$
		$b = 0,3098$		0,3464	$n = 13$

что и даст, наконец, s^2. Значение n равно сумме 7 степеней свободы для ряда А и 6 степеней свободы для ряда В, т. е. равно 13. Оценки дисперсий этих двух коэффициентов регрессии можно получить делением s^2 соответственно на 60 и 42, а оценка дисперсии их разности будет равна сумме этих двух оценок. Извлекая квадратный корень из этой суммы, получим среднюю квадратическую ошибку 0,01985, откуда следует, что $t = 1,844$. На основе этого показателя приходим к выводу, что хотя разность между коэффициентами регрессии довольно большая, все же она не может считаться существенной, т. е. нельзя утверждать, что скорость роста культуры В выше скорости роста культуры А.

26.2. Отношения средних и коэффициентов регрессии

Пример 23.1. В тех случаях, когда два ряда наблюдений, например, таких, как приведенные в табл. 27 (стр. 103), дают средние, разность которых после обработки признана достаточно существенной, имеется возможность составить некоторое представление о размере истинной разности. Это производится путем определения для нее пределов, которые находятся из установленной разности путем прибавления к ней или вычитания из нее средней квадратической ошибки, умноженной на соответствующий коэффициент. Значение этого коэффициента определяется уровнем существенности, принятым за основу, т. е. вероятностью того, что точная разность находится в установленных пределах. Так, эта вероятность будет равна 0,95, если выбранный нами уровень существенности составляет 0,05; в этом случае указанный выше коэффициент t находится в соответствии с числом степеней свободы. Например, в упомянутом выше примере 27 при 9 степенях свободы $t = 2,262$.

Однако в некоторых случаях разность между средними результатами воздействия двух факторов или мероприятий сама по себе представляет меньший интерес, чем отношение таких результатов. Такое положение, например, возможно, если эффекты от каждого из снотворных средств пропорциональны их количествам. Оно возможно и в тех случаях, когда отношение эффектов теоретически должно быть константным, благодаря чему отношение эффектов при любой дозировке является оценкой одного и того же постоянного отношения. Разность же таких эффектов не остается постоянной, а в значительной мере находится в зависимости от экспериментального материала, от условий эксперимента и от размера выбранной нами дозы. При этих условиях довольно часто возникает задача установления пределов для предполагаемого константного отношения эффектов. Эта задача аналогична той, которая была рассмотрена выше при определении пределов для предполагаемой константной разности.

Положим, что x и y — наблюденные эффекты двух мероприя-

тий A и B, а α является точным отношением эффекта A к эффекту B (в простейшем случае веса A и B равны единице). Возьмем величину

$$z = x - \alpha y$$

в качестве переменной, изменчивость которой определяется на основе экспериментальных данных. Арифметические расчеты, необходимые для решения этой задачи, подобны тем, которые даны в примере 20, только к средним и суммам квадратов для переменных x и y, которые там были вычислены, следует добавить еще сумму их произведений. Если взять данные табл. 27, то для x имеем:

$$\begin{aligned} S(x^2) &= 34{,}43 \\ \bar{x}S(x) &= \ 5{,}625 \\ \hline S(x-\bar{x})^2 &= 28{,}805 \end{aligned}$$

для y:

$$\begin{aligned} S(y^2) &= 90{,}37 \\ \bar{y}S(y) &= 54{,}289 \\ \hline S(y-\bar{y})^2 &= 36{,}081 \end{aligned}$$

и для произведения xy:

$$\begin{aligned} S(xy) &= 43{,}11 \\ \bar{x}S(y) = \bar{y}S(x) &= 17{,}475 \\ \hline S(x-\bar{x})(y-\bar{y}) &= 25{,}635 \end{aligned}$$

Полагая α все еще неизвестным, мы находим

$$S(z) = 7{,}5 - 23{,}3\alpha$$

и

$$S(z-\bar{z})^2 = 28{,}805 - 2\alpha(25{,}635) + \alpha^2(36{,}081).$$

Кроме того, легко видеть, что существенное отклонение от предполагаемого α будет в том случае, если:

$$\frac{9}{10} S^2(z) > t^2 S(z-\bar{z})^2.$$

Если выбрать 5%-ный уровень существенности, то

$$t = 2{,}262 \quad \text{и} \quad t^2 = 5{,}116644,$$

и поэтому мы приходим к уравнению относительно α:

$$303{,}9874\alpha^2 - 26{,}1098(2\alpha) - 96{,}7599 = 0.$$

Отсюда находим два значения α:

$$\alpha = +0{,}6566 \quad \text{и} \quad \alpha = -0{,}4848.$$

Следовательно, отношение между силой действия снотворного средства A и силой действия средства B согласно имеющимся данным не может быть больше, чем 0,6566, или примерно $^2/_3$. Тот

факт, что второе значение α — отрицательное, указывает на то, что при имеющихся данных и при избранном уровне существенности нельзя считать, что средство A вообще обладает положительным снотворным действием, т. е. фактически может оказаться, что оно является не снотворным, а, наоборот, возбуждающим средством. Сила возбуждения этого средства должна быть примерно в два раза меньше, чем снотворная сила средства B, и только тогда разность между эффектами этих средств станет существенной. Тот же метод можно применить и при установлении пределов для тех значений независимой переменной, при которых функция регрессии принимает заданное значение или при которых пересекаются две линии регрессии.

Если α_1 и α_2 являются точными средними, а β_1 и β_2 — точными коэффициентами регрессий двух зависимых переменных, то точка пересечения этих линий регрессии, которую обозначим через X, должна удовлетворять условию:

$$\alpha_1 + (X - \bar{x}_1)\beta_1 = \alpha_2 + (X - \bar{x}_2)\beta_2.$$

Выборочная дисперсия величины

$$a_1 + (X - \bar{x}_1)b_1 - a_2 - (X - \bar{x}_2)b_2$$

будет определяться формулой:

$$s^2\left\{\frac{1}{N_1} + \frac{1}{N_2} + \frac{(X-\bar{x}_1)^2}{S_1} + \frac{(X-\bar{x}_2)^2}{S_2}\right\},$$

где S_1 и S_2 — суммы квадратов отклонений независимой переменной для двух выборок и s^2 — средний квадрат отклонений зависимых переменных от линий регрессии. Отсюда, если выражение

$$\{a_1 - a_2 - b_1\bar{x}_1 + b_2\bar{x}_2 + X(b_1 - b_2)\}^2$$

приравнять к

$$s^2t^2\left\{\frac{1}{N_1} + \frac{1}{N_2} + \frac{\bar{x}_1^2}{S_1} + \frac{\bar{x}_2^2}{S_2} - 2X\left(\frac{\bar{x}_1}{S_1} + \frac{\bar{x}_2}{S_2}\right) + X^2\left(\frac{1}{S_1} + \frac{1}{S_2}\right)\right\},$$

то мы придем к квадратному уравнению относительно X, корни которого и будут пределами, соответствующими уровню существенности, определяемому значением t.

Пример 23.2. *Возраст, при котором девочки по своему росту перегоняют мальчиков.* Карн (1934 г.) приводит данные измерения роста у 4007 школьников Кройдона:

	Число N	Средний возраст x (в годах)	Средний рост a (в дюймах)	Регрессия b (дюйм/год)	$S(x-\bar{x})^2$ (годы)
Мальчики .	1 946	12,2016	56,004	1,60	337,894
Девочки . .	2 061	12,1300	56,550	2,45	382,835

Средний квадрат отклонений от линий регрессии s^2 равен

$$8{,}17915$$

и соответствует 3991 степени свободы. Для пределов при 5%-ном уровне существенности

$$t = 1{,}96.$$

Отсюда

$$s^2t^2 = 31{,}421.$$

Взяв 12 лет в качестве условного начала отсчета x, по остальным данным находим

$$a_2 - a_1 - b_2\bar{x}_2 + b_1\bar{x}_1 = 0{,}55016;$$

$$(b_2 - b_1)X = 0{,}85X;$$

$$\frac{1}{N_1} + \frac{1}{N_2} + \frac{\bar{x}_1^2}{S_1} + \frac{\bar{x}_2^2}{S_2} = 0{,}001163582;$$

$$-\left(\frac{\bar{x}_1}{S_1} + \frac{\bar{x}_2}{S_2}\right) \cdot 2X = -0{,}000936405\,(2X);$$

$$\left(\frac{1}{S_1} + \frac{1}{S_2}\right) \cdot X^2 = 0{,}00557160\,(X^2).$$

На основе этих данных строим квадратное уравнение для X

$$(0{,}547435)X^2 + (0{,}497064) \cdot 2X + 0{,}266122 = 0,$$

корни которого равны

$$-1{,}490 \text{ и } -0{,}326,$$

что соответствует возрастам

$$10{,}510 \text{ и } 11{,}674 \text{ года.}$$

Расчет же точки пересечения фактических линий регрессии дает 11,353 года, что значительно ближе к верхнему пределу. В данном случае все дети имели возраст 11—12 лет, и поэтому для более раннего возраста точность сравнений значительно снижается. Именно в связи с этим нижний предел, установленный ранее, так резко отличается от результата непосредственного определения точки пересечения регрессий. Если бы были измерены дети более раннего возраста или если бы вообще был бы взят более широкий интервал возрастов, то при таком же числе наблюдений можно было бы получить более точные результаты.

27. Вычисление криволинейных регрессий

Теория криволинейной регрессии в значительной мере упрощается, если рассматривать частный и в то же время наиболее важный случай, когда изменчивость зависимой переменной остается одной и той же при всех значениях независимой переменной и вместе с этим подчинена закону нормального распределения. Здесь будет дано подробное изложение техники расчета линии регрессии

$$Y = a + bx + cx^2 + dx^3 + \ldots$$

для того случая, когда последовательные значения x находятся друг от друга на одинаковом расстоянии. Более

сложный общий случай, когда отдельным значениям x соответствует различное число наблюдений y, будет рассмотрен в параграфе 29,2; если же в дополнение к этому интервалы x не равны между собой, то применим еще более общий метод параграфа 29, в котором степени x рассматриваются в качестве независимых переменных.

В том виде, в каком выше приведена формула регрессии, она не вполне удобна для вычислений, так как не позволяет вести расчеты по последовательным этапам. В этом последнем случае речь идет о вычислении в последовательном порядке средней $\bar{y}$, потом линейного уравнения относительно x, далее квадратного уравнения относительно x и т. д., причем каждое последующее уравнение получается из предыдущего путем простого прибавления некоторого нового члена. Все это проводится так, что на каждом этапе применяется одна и та же схема вычислений. Для того чтобы познакомиться с этим методом, рассмотрим уравнение:

$$Y = A + B\xi_1 + C\xi_2 + D\xi_3, + \ldots,$$

где ξ_1, ξ_2, ξ_3 — являются функциями x соответственно 1-й, 2-й и 3-й степени, при раскрытии которых должна получиться формула регрессии.

Эти функции от x выполняют роль коэффициентов в формулах, относящихся к некоторым сравнениям между наблюденными значениями y. Сравнения, о которых здесь идет речь, характерны тем, что они являются компонентами варьирования переменной y. Так ξ_1 представляет собой ни что иное, как разность $x - \bar{x}$; если, например, имеется 7 наблюдений, то значениями ξ_1 будут —3, —2, —1, 0, 1, 2, 3 и сравнением между y, соответствующим функции 1-й степени x, будет:

(I) $$-3y_1 - 2y_2 - y_3 + 0y_4 + y_5 + 2y_6 + 3y_7.$$

Величинами ξ_2 будут служить коэффициенты такого сравнения между y:

(II) $$5y_1 + 0y_2 - 3y_3 - 4y_4 - 3y_5 + 0y_6 + 5y_7.$$

Эти коэффициенты ξ_2 определяются квадратичной формой x:

$$(x - \bar{x})^2 - 4 = \xi_1^2 - 4 = \xi_2.$$

Следует отметить, что сумма этих коэффициентов и сумма произведений их с ξ_1 равны нулю.

Для функций ξ 3-й, 4-й и 5-й степени следует взять выражения:

(III) $$-y_1 + y_2 + y_3 - y_5 - y_6 + y_7 \quad \frac{\xi_1(\xi_1^2 - 7)}{6};$$

(IV) $3y_1 - 7y_2 + y_3 + 6y_4 + y_5 - 7y_6 + 3y_7 \quad \frac{(7\xi_1^4 - 67\xi_1^2 + 72)}{12}$;

(V) $-y_1 + 4y_2 - 5y_3 + 5y_5 - 4y_6 + y_7 \quad \frac{(21\xi_1^5 - 245\xi_1^3 + 524\xi_1)}{60}$.

Отметим, что в каждом отдельном выражении сумма коэффициентов равна нулю и, следовательно, эти выражения определяются только различием значений y. Вместе с этим сумма произведений соответствующих коэффициентов в двух любых из этих выражений также равна нулю, откуда следует, что эти сравнения полностью независимы.

Чтобы определить уравнение криволинейной регрессии, следует в каждую из этих функций y подставить данные наблюдений и разделить результат подстановки на сумму квадратов входящих в эту функцию коэффициентов. Эти частные и являются коэффициентами при соответствующих функциях x. Так, суммы квадратов коэффициентов в приведенных выше пяти выражениях соответственно равны 28, 84, 6, 154 и 84. Таким образом, последовательными членами определяемого уравнения криволинейной регрессии будут выражения:

$$\frac{(-3y_1 - 2y_2 - y_3 + y_5 + 2y_6 + 3y_7)\,\xi_1}{28};$$

$$\frac{(5y_1 - 3y_3 - 4y_4 - 3y_5 + 5y_7)\left(\xi_1^2 - 4\right)}{84};$$

$$\frac{(-y_1 + y_2 + y_3 - y_5 - y_6 + y_7)\left(\xi_1^3 - 7\xi_1\right)}{36};$$

$$\frac{(3y_1 - 7y_2 + y_3 + 6y_4 + y_5 - 7y_6 + 3y_7)\left(7\xi_1^4 - 67\xi_1^2 + 72\right)}{1848};$$

$$\frac{(-y_1 + 4y_2 - 5y_3 + 5y_5 - 4y_6 + y_7)\left(21\xi_1^5 - 245\xi_1^3 + 524\xi_1\right)}{5040}.$$

Первое из этих выражений дает наилучшую прямую линию. Если взять два первых выражения вместе, то получится парабола второго порядка, если взять три выражения, то получится парабола третьего порядка и т. д.

В «Статистических таблицах» даны такие ортогональные полиномы, т. е. независимые сравнения наблюдений y, пригодные для вычисления кривых вплоть до 5-й степени и для числа наблюдений вплоть до $n = 75$. Общие формулы были приведены также в прежних (3-е — 6-е) изданиях настоящей книги. В случае же более обширного ряда наблюдений и при необходимости иметь уравнение регрессии более высокого порядка разработку данных следует вести по методу, описанному в следующем параграфе.

Эти же компоненты могут быть выражены также через конечные разности ряда y_n. Так, сравнения, приведенные выше, могут быть в этом случае выражены следующим образом:

(I) $\Delta(3y_1 + 5y_2 + 6y_3 + 6y_4 + 5y_5 + 3y_6)$;
(II) $\Delta^2(5y_1 + 10y_2 + 12y_3 + 10y_4 + 5y_5)$;
(III) $\Delta^3(y_1 + 2y_2 + 2y_3 + y_4)$;
(IV) $\Delta^4(3y_1 + 5y_2 + 3y_3)$;
(V) $\Delta^5(y_1 + y_2)$,

где Δy_1 означает $y_2 - y_1$ и т. д. Вместо суммы квадратов коэффициентов прежней формулы теперь следует брать квадраты сумм коэффициентов при конечных разностях соответствующего порядка r, деля их на

$$\frac{n(n^2-1)\ldots(n^2-r^2)}{(2\cdot 6)(6\cdot 10)\ldots[(4r-2)(4r+2)]}.$$

Этот способ обработки конечных разностей иногда может дать большую экономию труда, так как эти разности в большинстве случаев являются более простыми числами по сравнению с первоначальными данными, к тому же знаки соответствующих коэффициентов здесь всегда положительны. Сами же эти коэффициенты могут быть для любого порядка r определены путем последовательного умножения, начиная с единицы, на множители:

$$\frac{(r+1)(n-r-1)}{1(n-1)}; \quad \frac{(r+2)(n-r-2)}{2(n-2)}, \ldots$$

Этот метод легко проверить, конструируя данным способом формулу для $n=7$ и $r=4$, приведенную выше; для этого следует только продифференцировать четыре раза выражения коэффициентов, входящих в четвертый компонент.

Хотя для расчетов удобнее иметь приведенные выше выражения, у которых постоянные множители вынесены за скобки, т. е. иметь выражения в их наиболее простой алгебраической форме, все же в некоторых других отношениях представляется удобным иметь коэффициенты полиномов в виде целых чисел. Так, ξ_3, определяемые как $\xi_1^3 - 7\xi_1$, являются целыми числами и равны ушестеренным коэффициентам приведенного ранее выражения. При этих условиях суммы квадратов коэффициентов определяются по формуле

$$\frac{n(n^2-1)\ldots(n^2-r^2)}{12\cdot 15\cdot\ldots\cdot\left(16-\frac{4}{r^2}\right)} = \frac{(n+r)!}{(n-r-1)!}\,\frac{(r!)^4}{(2r)!\,(2r+1)!}.$$

Таким образом, процесс вычисления кривой теперь сводится к уравнениям

$$A = \bar{y} = \frac{1}{n'}S(y);$$

$$B = \frac{12}{n'(n'^2-1)} S(y\xi_1);$$

$$C = \frac{180}{n'(n'^2-1)(n'^2-4)} S(y\xi_2),$$

где коэффициент члена порядка r определяется по общей формуле:

$$\frac{(2r)!\,(2r+1)!}{(r!)^4 n'(n'^2-1)\ldots(n'^2-r^2)} S(y\xi_r).$$

По мере того как определяются и вводятся в уравнение регрессии члены, приводящие к последовательному приближению линии регрессии к фактическим данным, происходит уменьшение суммы квадратов отклонений:

$$S(y-Y)^2.$$

Вычисление этой суммы можно провести без определения отдельных значений Y путем вычитания из $S(y^2)$ последовательно таких величин:

$$n'A^2; \quad \frac{n'(n'^2-1)}{12} B^2; \quad \frac{n'(n'^2-1)(n'^2-4)}{180} C^2;$$

или в более простом выражении

$$AS(y); \quad BS(y\xi_1); \quad CS(y\xi_2)$$

и т. д. Эти величины определяют собой то уменьшение суммы квадратов, которое происходит на каждом этапе развертывания кривой регрессии. Для определения оценки s^2, характеризующей остаточную дисперсию, следует $S(y-Y)^2$ разделить на n — число степеней свободы, оставшихся после установления линии регрессии. Это n находится вычитанием из n' числа констант в формуле регрессии. Так, если определена прямая линия регрессии, то $n = n' - 2$, если же определена кривая пятого порядка, то $n = n' - 6$.

28. Техника вычислений

Основная расчетная работа по определению криволинейной регрессии значительно упрощается, если применить способ повторного суммирования, с которым мы познакомились на примере 23. Допустим, что данные выписаны в порядке возрастания переменной x; тогда каждый этап суммирования y проводится сверху вниз (а не наоборот, как это было в примере 23). Обозначим суммы последовательных столбцов через S_1, S_2, ... Каждая из этих сумм делится на соответствующий делитель, зависящий только от n', в результате чего получается новый ряд величин $a, b, c, \ldots$, определяемых следующим образом:

$$a = \frac{1}{n'} S_1 = \frac{1}{n'} S(y) = \bar{y}$$

$$b = \frac{1 \cdot 2}{n'(n'+1)} S_2$$

$$c = \frac{1 \cdot 2 \cdot 3}{n'(n'+1)(n'+2)} S_3 \text{ и т. д.}$$

На основе этих величин вычисляется третья серия показателей $a', b', c', \ldots$, причем этот переход уже не зависит от n'. Ниже приводятся шесть первых величин a, b, и т. д.; они необходимы при расчете уравнения регрессии пятого порядка:

$$a' = a$$
$$b' = a - b$$
$$c' = a - 3b + 2c$$
$$d' = a - 6b + 10c - 5d$$
$$e' = a - 10b + 30c - 35d + 14e$$
$$f' = a - 15b + 70c - 140d + 126e - 42f$$

Образование входящих сюда коэффициентов производится путем последовательного умножения на

$$\frac{r(r+1)}{1 \cdot 2}; \quad \frac{(r-1)(r+2)}{2 \cdot 3}; \quad \frac{(r-2)(r+3)}{3 \cdot 4} \text{ и т. д.}$$

Эти новые величины a', b' и т. д. пропорциональны искомым коэффициентам уравнения регрессии; для определения последних необходимо каждую из величин a', b' и т. д. разделить на соответствующие числа, указанные ниже:

$$A = a' \qquad B = \frac{6}{n'-1} b'$$

$$C = \frac{30}{(n'-1)(n'-2)} c' \qquad D = \frac{140}{(n'-1)(n'-2)(n'-3)} d'$$

$$E = \frac{630}{(n'-1)(n'-2)(n'-3)(n'-4)} e'$$

$$F = \frac{2772}{(n'-1)(n'-2)(n'-3)(n'-4)(n'-5)} f'$$

Числа, стоящие в числителе этих множителей, могут быть выражены через показатель порядка r:

$$\frac{(2r+1)!}{(r!)^2}.$$

Если определено уравнение регрессии порядка r, то оценки средних квадратических ошибок для всех коэффициентов регрессии основываются на одном и том же значении s^2, определяемом по формуле

$$s^2 = \frac{1}{n'-r-1}\left\{S(y)^2 - n'A^2 - \frac{n'(n'^2-1)}{1 \cdot 2} B^2 - \ldots\right\}.$$

Чтобы получить оценку средней квадратической ошибки коэффициента, положим, при ξ_p, следует s^2 разделить на

$$S(\xi_p^2) = \frac{(p!)^4}{(2p)!(2p+1)!} n' (n'^2 - 1) \ldots (n'^2 - p^2)$$

и из результата извлечь квадратный корень. Число степеней свободы для такой оценки равно $(n' - r - 1)$; оно и должно быть использовано в качестве n при определении вероятности по таблице t.

Материалом, подходящим для применения этого метода, могут служить данные примера 22 (стр. 114), по которым можно вычислить кривую второго или третьего порядка.

28.1. Вычисление полиномиальных значений Y

Методы, изложенные в предыдущих параграфах, приводят к разложению наблюдений на две части, одна из которых состоит из компонентов, определяемых полиномами той или иной степени x, а вторая является остатком, в котором такие отдельные компоненты не выделены. В большинстве случаев, когда такой анализ проводится, не возникает потребности в исчислении отдельных полиномиальных значений Y для каждого значения x. Однако в некоторых случаях возникает необходимость иметь эти значения Y; например, они нужны для построения графика, для определения отклонений от кривой в той или иной интересующей нас области или для того, чтобы произвести более полную проверку достаточности достигнутых на данной стадии результатов.

Сам по себе процесс вычисления отдельных значений Y и вычисления по ним и по соответствующим коэффициентам отдельных полиномиальных значений Y довольно утомителен. Этого можно избежать, если воспользоваться способом, который позволяет установить весь ряд значений Y по их разностям при помощи некоторого непрерывного процесса. Такой способ обычен при прямолинейной регрессии. Если в этом случае в качестве конечного значения Y и в качестве постоянной разности между последовательными значениями Y взять величины

$$Y_1 = a' + 3b';$$
$$\Delta Y_1 = -\frac{6}{n'-1} b',$$

то путем последовательного прибавления этой константной разности можно построить весь ряд значений Y. Этот же прием вполне применим и к полиномам более высокого порядка, и в этих случаях он более чем на 75% сокращает количество труда, затрачиваемого на вычисления. Для кривой второго порядка соответствующие исходные данные таковы:

$$Y_1 = a' + 3b' + 5c'$$

$$\Delta Y_1 = -\frac{6}{n'-1}(b' + 5c');$$

$$\Delta^2 Y_1 = \frac{60}{(n'-1)(n'-2)} c'.$$

Начиная с конечного значения ΔY_1, можно построить ряд разностей первого порядка; для этого следует только последовательно прибавлять константную разность второго порядка $\Delta^2 Y_1$. Затем, начиная с Y_1 и прибавляя последовательно полученные ранее разности первого порядка, можно построить весь ряд Y.

В табл. 30.2 даны для кривых различного порядка коэффициенты при a', b', c' и т. д., которыми следует пользоваться при определении конечных значений и разностей различного порядка. Формулы же для образования участвующих в этих случаях множителей, знак которых попеременно то положительный, то отрицательный, таковы:

$$1;\ \frac{-2\times3}{n'-1};\ \frac{3\times4\times5}{(n'-1)(n'-2)};\ \frac{-4\times5\times6\times7}{(n'-1)(n'-2)(n'-3)};\ \ldots$$

Ниже на данных примера 22 дается иллюстрация этих расчетов по способу последовательного суммирования. В левой части табл. 30.1 дана нижняя часть табл. 29, включающая пять последних наблюдений, а также результаты последовательного суммирования при определении кривой регрессии третьего порядка; в правой же части приведены результаты последовательного суммирования для определения полиномиальных значений Y.

Таблица 30.1

Наблюдения	1-я сумма	2-я сумма	3-я сумма	Полиномиальные значения	1-я разность	2-я разность	3-я константная разность
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
−0,25	117,38	960,77	4440,58	5,86	0,739	−0,1280	
+9,31	127,19	1087,96	5528,54	4,99	0,871	−0,1320	
−2,94	124,25	1212,21	6740,75	3,98	1,008	−0 1361	
+7,07	131,32	1343,53	8084,28	2,84	1,148	−0,1402	
+2,69	134,01	1477,54	9561,82	1,544	1,2919	−0,14423	0,004061
134,01	1477.54	9561,82	39167,21	134,00			
4,467000	3,177505	1,927786	0,957165				
4,467000	1,289495	−1,209431	−0,105995				

В нижних частях первых четырех колонок даны значения a, b, c и d, полученные непосредственно из итогов, и a', b', c' и d', вычисленные на основе этих последних. Для того чтобы определить значения Y с двумя десятичными знаками, следует Y_1 вы-

числить с тремя десятичными знаками, а разности высшего порядка следует определить с точностью на один десятичный знак большей, чем у разностей предшествующего порядка; следовательно, при образовании разностей низшего порядка после окончания вычислений один знак отбрасывается. Принимая все это во внимание, следует для a, b, c и d взять шесть десятичных знаков. Для еще большего повышения точности достаточно оставить у этих величин некоторое количество добавочных знаков. Сумма полиномиальных значений Y должна совпадать с суммой фактических значений y. Это может служить контролем для вычислений, проведенных в последних колонках, но отнюдь не проверкой правильности первоначального суммирования.

Таблица 30.2

Коэффициенты при a', b', c' . . . для определения конечных значений Y и их разностей

1	3	5	7	9	11	13	15	17	19	21
	1	5	14	30	55	91	140	204	285	385
		1	7	27	77	182	378	714	1 254	2 079
			1	9	44	156	450	1 122	2 508	5 148
				1	11	65	275	935	2 717	7 007
					1	13	90	442	1 729	5 733
						1	15	119	665	2 940
							1	17	152	952
								1	19	189
									1	21
										1

29. Регрессия с несколькими независимыми переменными

В ряде случаев возникает необходимость выразить средние значения зависимой переменной y через значения нескольких независимых переменных $x_1, x_2, \ldots x_p$. Например, осадки, выпадающие в некотором районе, могут фиксироваться в ряде пунктов этого района, каждый из которых имеет определенную долготу, широту и высоту над уровнем моря. Поскольку эти три переменные величины влияют на количество осадков, может возникнуть необходимость определения средних эффектов для каждой из них. Когда говорится о том, что долгота, широта и высота местности являются независимыми переменными, то это следует понимать в том смысле, что количество осадков может быть выражено через значения этих переменных, но отнюдь не в том смысле, что они совсем не коррелированы между собой. Вполне может быть так, что более южные станции лежат западнее по сравнению с северными станциями и, следовательно, для этих станций долгота, измеренная в направлении к западу, окажется отрицательно коррелированной с широтой, измеряемой в направлении к северу. Если в этом случае количество осадков повышается к западу, но

не зависит от широты, то, производя простое сопоставление количества осадков с различными широтами, мы все же получим уравнение регрессии, указывающее на то, что осадки уменьшаются по мере продвижения к северу. В связи с этим возникает задача определения уравнения, учитывающего измерения трех переменных на каждой станции и более или менее согласующегося с наблюденными значениями зависимой переменной; такое уравнение называется уравнением *частной* регрессии, а его коэффициенты — частными коэффициентами регрессии.

Для упрощения алгебраических преобразований мы допустим, что y, x_1, x_2 и x_3 измерены путем отсчета от соответствующих средних. Рассмотрим теперь такую форму связи:

$$Y = b_1 x_1 + b_2 x_2 + b_3 x_3.$$

Если через S обозначить суммирование по всем наблюдениям, то можно написать следующие три уравнения:

$$b_1 S(x_1^2) + b_2 S(x_1 x_2) + b_3 S(x_1 x_3) = S(x_1 y);$$
$$b_1 S(x_1 x_2) + b_2 S(x_2^2) + b_3 S(x_2 x_3) = S(x_2 y);$$
$$b_1 S(x_1 x_3) + b_2 S(x_2 x_3) + b_3 S(x_3^2) = S(x_3 y).$$

Девять коэффициентов этих уравнений определяются по наблюденным данным путем непосредственного их перемножения и последующего суммирования или, если число наблюдений велико, путем построения корреляционных таблиц для каждой из шести пар переменных. Для определения неизвестных b_1, b_2 и b_3 решение этих трех уравнений проводится обычным порядком: сначала из первого и третьего уравнений, а потом из второго и третьего уравнений элиминируется b_3, в результате чего получается два уравнения с двумя неизвестными b_1 и b_2; элиминируя из этих уравнений b_2, находим b_1, после чего путем подстановки определяем b_2 и b_3.

В некоторых случаях определенному ряду значений независимых переменных соответствуют значения нескольких зависимых переменных, для которых требуется найти уравнения регрессии. Такое положение возникает, например, в случае, когда у одного и того же ряда метеорологических станций имеются наблюдения над осадками за несколько месяцев или лет. Для таких случаев можно установить довольно простую формулу, пригодную для определения уравнения каждой зависимой переменной, избегая решения системы уравнений в каждом отдельном случае.

Это достигается путем решения трех систем, каждая из которых состоит из трех уравнений:

$$b_1 S(x_1^2) + b_2 S(x_1 x_2) + b_3 S(x_1 x_3) = 1\ 0\ 0;$$
$$b_1 S(x_1 x_2) + b_2 S(x_2^2) + b_3 S(x_2 x_3) = 0\ 1\ 0;$$
$$b_1 S(x_1 x_3) + b_2 S(x_2 x_3) + b_3 S(x_3^2) = 0\ 0\ 1.$$

Три решения этих систем уравнений могут быть обозначены следующим образом:

$$b_1 = c_{11},\quad c_{12},\quad c_{13};$$
$$b_2 = c_{12},\quad c_{22},\quad c_{23};$$
$$b_3 = c_{13},\quad c_{23},\quad c_{33}.$$

После определения этих шести величин легко вычислить частные коэффициенты регрессии для каждого отдельного случая. Для этого следует найти $S(x_1y)$, $S(x_2y)$ и $S(x_3y)$ и подставить эти суммы в формулы:

$$b_1 = c_{11}S(x_1y) + c_{12}S(x_2y) + c_{13}S(x_3y);$$
$$b_2 = c_{12}S(x_1y) + c_{22}S(x_2y) + c_{23}S(x_3y);$$
$$b_3 = c_{13}S(x_1y) + c_{23}S(x_2y) + c_{33}S(x_3y).$$

Величины c, которые известны под названием элементов матрицы ковариаций, необходимы, кроме этого, и при определении точности коэффициентов регрессии, поэтому приведенный выше способ их определения рекомендуется не только в рассмотренном случае, но и вообще.

Метод частной регрессии имеет весьма широкое применение. Следует отметить, что независимые переменные могут определяться самыми различными способами. Так, например, если считается необходимым выразить количество осадков как линейную функцию широты и долготы и как квадратичную функцию высоты, то просто следует квадрат высоты рассматривать в качестве четвертой независимой переменной. Описанный выше процесс вычислений остается тем же самым, только следует считать, что $S(x_3x_4) = S(x_3^2)$, и вычислять эту сумму квадратов по данным о высоте местности.

Изложенный в параграфах 27 и 28 метод анализа, основанный на ортогональных полиномах, может быть распространен и на случай множественной корреляции. В том специальном случае, когда имеется простой ряд наблюдений над зависимой переменной, каждому из значений которой соответствуют равноудаленные значения зависимой переменной (что часто встречается при изучении изменений во времени экономических и социальных показателей), применение ортогональных полиномов представляет очевидное преимущество. Если же количество наблюдений, относящихся к отдельным значениям независимой переменной, различно или интервалы у этой последней неодинакового размера, то метод ортогональных полиномов, хотя и может быть распространен и на эти случаи, все же становится сложным и менее удобным по сравнению с обработкой данных по методу множественной регрессии. Этот метод позволяет определять регрессии, выражаемые не только через различные степени независимых переменных, но и через другие функции, такие, как логарифмы, степенные и тригонометрические функции и пр.

Для оценки выборочных ошибок отдельных коэффициентов регрессии необходимо установить степень приближения вычисленных по уравнению регрессии значений Y к фактическим значениям y. Как и в предыдущих случаях, сумму квадратов отклонений $(y - Y)$ можно вычислить как разность между суммой квадратов $S(y^2)$ и величинами, связанными с коэффициентами регрессии b_1, b_2, ... Так, при трех переменных

$$S(y - Y)^2 = S(y^2) - b_1S(x_1y) - b_2S(x_2y) - b_3S(x_3y).$$

Если имеется n' наблюдений и p независимых переменных, то сначала определяется остаточная дисперсия

$$s^2 = \frac{1}{n' - p - 1} S(y - Y)^2,$$

после чего находится критерий существенности различия между, положим, b_1 и каким-либо гипотетическим значением β_1:

$$t = \frac{b_1 - \beta_1}{s\sqrt{c_{11}}}.$$

На основе этого критерия по таблице значений t определяется вероятность P, причем считается, что $n = n' - p - 1$.

При большом числе переменных рекомендуется применение карточек, на каждую из которых заносятся значения изучаемых переменных величин. Группируя эти карточки по заранее выбранным интервалам группировки некоторой пары переменных, можно построить корреляционную таблицу, по которой вычисляются необходимые для дальнейшей обработки суммы квадратов и произведений.

Пример 24. *Зависимость количества осадков от местоположения и высоты пункта наблюдения.* Расположение 57 метеорологических станций Гертфортшира характеризуется средней долготой 12′,4 W и средней широтой 51°48′,5 N, а их средняя высота над уровнем моря составляет 302 фута. За единицы измерения этих переменных взяты две минуты долготы, одна минута широты и 20 футов высоты. При этих условиях были получены следующие суммы квадратов и произведений отклонений от соответствующих средних:

$$S(x_1^2) = 1934{,}1;\quad S(x_2x_3) = +119{,}6;$$
$$S(x_2^2) = 2889{,}5;\quad S(x_3x_1) = +924{,}1;$$
$$S(x_3^2) = 1750{,}8;\quad S(x_1x_2) = -772{,}2.$$

Чтобы получить коэффициенты, которые можно было бы использовать при любых частных показателях погоды на этих станциях, решим следующую систему уравнений:

$$1934{,}1c_{11} - 772{,}2c_{12} + 924{,}1c_{13} = 1;$$
$$-772{,}2c_{11} + 2889{,}5c_{12} + 119{,}6c_{13} = 0;$$
$$+924{,}1c_{11} + 119{,}6c_{12} + 1750{,}8c_{13} = 0.$$

Элиминируя c_{13} из первого и третьего, а также из второго и третьего уравнений, имеем

$$2532{,}3\ c_{11} - 1462{,}5\ c_{12} = 1{,}7508$$
$$-1462{,}5 c_{11} + 5044{,}6 c_{12} = 0.$$

Исключая отсюда c_{12}, находим

$$10635{,}5 c_{11} = 8{,}8321.$$

Отсюда следует, что

$$c_{11} = 0{,}00083044;\ c_{12} = 0{,}00024075;\ c_{13} = -0{,}00045477.$$

Последние две величины определены здесь при помощи последовательной подстановки.

Поскольку соответствующие уравнения c_{12}, c_{22} и c_{23} отличаются от предыдущих только иным расположением в правой части чисел 1,0, 0, то можно сразу написать:

$$-1462{,}5 c_{12} + 5044{,}6 c_{22} = 1{,}7508.$$

Подставляя сюда уже найденное ранее c_{12}, получим

$$c_{22} = 0{,}00041686 \text{ и } c_{23} = -0{,}00015555.$$

Наконец, для определения c_{33} следует ранее известные c_{13} и c_{23} подставить в уравнение

$$924{,}1 c_{13} + 119{,}6 c_{23} + 1750{,}8 c_{33} = 1,$$

в результате чего получим:

$$c_{33} = 0{,}00082183.$$

Чтобы избежать ошибок округления, рекомендуется, как это сделано у нас, производить вычисления с точностью на один десятичный знак больше, чем это требуется для окончательного результата.

После того, как определены коэффициенты c, весьма просто определить регрессию любого показателя погоды на эти три переменные. Например, в январе 1922 г. средние осадки, отмеченные на этих станциях, составляли 3,87 дюйма, а суммы произведений отклонений осадков от их средней с этими тремя независимыми переменными (в которых единицей измерения осадков было 0,1 дюйма) были: $S(x_1 y) = +1137{,}4$; $S(x_2 y) = -592{,}9$; $S(x_3 y) = +891{,}8$.

Умножая эти суммы сначала на c_{11}, c_{12} и c_{13} и суммируя полученные произведения, находим коэффициент регрессии на долготу:

$$b_1 = 0{,}39624.$$

Используя в том же порядке множители c_{12}, c_{22} и c_{23}, получим коэффициент регрессии на широту:

$$b_2 = -0{,}11204$$

и, наконец, при помощи c_{13}, c_{23} и c_{33}, получим регрессию на высоту:

$$b_3 = 0{,}30788.$$

Если теперь принять во внимание те условные единицы измерения переменных, которые ранее были приняты, то можно сказать, что в этом месяце количество осадков увеличивалось на 0,0198 дюйма на каждую минуту долготы в направлении на запад, уменьшалось на 0,0112 дюйма на каждую минуту широты в направлении на север и возрастало на 0,00154 дюйма на каждый фут высоты.

Теперь определим влияние ошибок выборки на коэффициент регрессии высоты. Взяв те же самые, что и ранее, условные единицы измерения, находим сумму квадратов 57 отклонений осадков на отдельных станциях от их средней

$$S(y^2) = 1786{,}6.$$

Используя известные значения b_1, b_2 и b_3, определяем по указанной ранее формуле:

$$S(y - Y)^2 = 994{,}9.$$

Чтобы получить s^2, следует эту остаточную сумму квадратов разделить на число степеней свободы, оставшихся после определения регрессии на три переменных, т. е. на число 53. В результате получим

$$s^2 = 18{,}772.$$

Умножая эту величину на c_{33} и извлекая квадратный корень, найдем

$$s\sqrt{c_{33}} = 0{,}12421.$$

Так как $n = 53$ достаточно велико, то мы, не рискуя сделать грубую ошибку, можем сказать, что регрессия осадков на высоту в условных единицах составляет 0,308 и ее средняя квадратическая ошибка равна 0,124, или, если перейти к фактическому измерению осадков в дюймах, а высоты в сотнях футов, то этот коэффициент регрессии будет равен 0,154 дюйма и его средняя квадратическая ошибка 0,062 дюйма.

Значение описанного здесь метода состоит, во-первых, в его общности и, во-вторых, в том, что он представляет собой единый процесс последовательного определения: а) наилучшего уравнения регрессии заданного типа и б) остаточной дисперсии и точности коэффициентов регрессии.

Рассмотренный сейчас пример дает возможность составить представление о той роли, которую выполняют величины c; мы видели, что выборочная дисперсия коэффициента регрессии, по-

ложим, b_1 определяется путем умножения остаточной дисперсии s^2 на величину c_{11}, вычисляемую на основе наблюденных значений независимых переменных. Во многих других случаях применения метода регрессий множители c применяются при оценке связи между двумя коэффициентами регрессии; например, ковариация для b_1 и b_2 может быть определена умножением все той же остаточной дисперсии s^2 на величину c_{12}. Таким образом, появляется возможность вести широкое изучение полученных результатов без проведения каких-либо повторных и дополнительных расчетов. Хотя в условиях приведенных выше примеров метеорологических наблюдений это не имеет никакого значения, однако в целом ряде других исследований часто возникает необходимость сравнить значения двух коэффициентов регрессии, т. е. установить, будет ли, например, b_1 существенно отличаться от b_2. При решении этого вопроса следует сопоставить разность $b_1 - b_2$ с ее средней квадратической ошибкой. Эта последняя находится как корень квадратный из величины:

$$s^2(c_{11} - 2c_{12} + c_{22}),$$

так как дисперсия разности двух величин всегда равна сумме двух соответствующих дисперсий минус удвоенная ковариация, что следует из простого алгебраического тождества

$$(x - y)^2 = x^2 - 2xy + y^2.$$

Следовательно, определив величины c, мы имеем возможность произвести оценку существенности суммы, или разности, или вообще какой-либо иной линейной функции двух или бо́льшего числа коэффициентов регрессии. Для этого следует вычислить соответствующую среднюю квадратическую ошибку и найти отношение значения данной линейной функции к этой ошибке; полученной величине t соответствует число степеней свободы остаточной дисперсии.

29.1. Исключение независимой переменной

В некоторых случаях, после того, как определено уравнение регрессии, одна из переменных может оказаться излишней, не представляющей для нас никакого интереса. В этих условиях, пожалуй, имеет смысл освободиться от нее и определить новую регрессию для оставшихся независимых переменных. Но это потребует самостоятельного решения системы уравнений, составленной на основе сумм квадратов и произведений этих переменных. Однако если применить специальный прием исключения переменных, то можно сократить затраты труда, необходимые для такого пересчета. Исключение одной из переменных будет всегда сопровождаться увеличением числа остаточных степеней свободы

на единицу; вместе с этим и остаточная сумма квадратов увеличится на величину, относящуюся к этой степени свободы. Если x_3 является переменной, которая подлежит исключению, то, как известно, дисперсия соответствующего коэффициента регрессии b_3 будет равна $\sigma^2 c_{33}$. Поэтому дисперсией величины $\frac{b_3}{\sqrt{c_{33}}}$ будет σ^2, а величина

$$\frac{b_3^2}{c_{33}}$$

будет приращением суммы квадратов при исключении переменной x_3.

Подобно этому, если мы желаем заменить в уравнении регрессии b_3 на некоторое теоретическое значение β_3, отличное от нуля, то соответствующее приращение остаточной суммы квадратов будет

$$\frac{(b_3 - \beta_3)^2}{c_{33}}.$$

После исключения одной из переменных может возникнуть необходимость исправить коэффициенты у остальных переменных. Эти коэффициенты, хотя они и были вычислены ранее, но теперь, когда одна из переменных, предположим по-прежнему x_3, исключена, будут иметь уже иные значения. Это легко сделать путем простого вычитания, например, из b_1 величины

$$\frac{c_{13}}{c_{33}} b_3.$$

Такие же поправки вводятся и в другие коэффициенты b.

Я весьма обязан профессору Х. Шульцу из Чикаго за сообщение о возможностях более широкого применения этого метода по сравнению с тем, что было дано в пятом издании настоящей книги. При исключении переменной преобразуется матрица величин c при помощи подстановки

$$c'_{12} = c_{12} - \frac{c_{13}c_{23}}{c_{33}}.$$

Величины c' образуют ту матрицу, которая должна получиться после исключения x_3. Эти величины позволяют определить дисперсии и ковариации исправленных коэффициентов, а также дают возможность вычислить дальнейшие поправки, вводимые при исключении второй переменной.

Таким образом, если будет определена регрессия некоторой зависимой переменной от значительного числа, например шести или более переменных, в отношении которых предполагается, что они оказывают влияние на зависимую переменную, то в случае, если у одной из этих переменных не обнаружено явного действия, всегда можно при очень небольших затратах труда перейти от

этого первоначального уравнения к такому, которое было бы получено, если бы эта исключаемая переменная не участвовала в наших вычислениях. Мы оставляем в стороне более сложный вопрос об исключении нескольких переменных.

29.2. Полиномиальное выравнивание при неодинаковых частотах

Изложенные в параграфах 28 и 28.1 приемы вычислений могут применяться еще в тех случаях, когда определяется уравнение регрессии при неодинаковом числе значений зависимой переменной, соответствующих различным значениям независимой переменной. Здесь мы не будем касаться вопроса об образовании последовательного ряда полиномов различного порядка и также не будем ставить перед собой задачу оценки существенности каждого отдельного коэффициента, а рассмотрим только вопрос об определении регрессии заданного порядка. Соответствующий данному случаю способ расчета со всеми его деталями будет показан на примере вычисления кривой третьего порядка, в котором зависимой переменной является время пробега 100 ярдов, а независимой переменной — возраст мальчика, сделавшего этот пробег. Данные относятся к 988 мальчикам различного возраста от 9,25 до 19,25 года (данные Грея).

Таблица 30.3

Сокращенный процесс суммирования частот

Возраст	Частоты	0	1	2	3	4	5	6
9,25	6	6	6	6	6	6	6	
9,75	8	14	20	26	32	38	44	
10,25	10	24	44	70	102	140	184	
10,75	28	52	96	166	268	408	592	
11,25	29	81	177	343	611	1 019	1 611	
11,75	46	127	304	647	1 258	2 277	3 888	
12,25	40	167	471	1 118	2 376	4 653	8 541	
12,75	53	220	691	1 809	4 185	8 838	17 379	
13,25	54	274	965	2 774	6 959	15 797	33 176	
13,75	66	340	1 305	4 079	11 038	26 835	60 011	
14,25	87	648						
14,75	71	561	2 296					
15,25	98	490	1 735	4 975				
15,75	84	392	1 245	3 240	7 398			
16,25	85	308	853	1 995	4 158	7 961		
16,75	67	223	545	1 142	2 163	3 803	6 309	
17,25	65	156	322	597	1 021	1 640	2 506	
17,75	44	91	166	275	424	619	866	
18,25	25	47	75	109	149	195	247	
18,75	16	22	28	34	40	46	52	
19,25	6	6	6	6	6	6	6	
	S	988	991	9 054	—3 640	34 796	—53 702	129 109

В данном случае процесс суммирования производится в отдельности как по частотам, так и по суммарному времени пробега. Табл. 30.3 дает частоты для 21 группы возрастов с интервалами через полугодие. Для определения уравнения регрессии 3-й степени эти частоты суммируются семь раз (соответствующие колонки занумерованы от 0 до 6). Результаты последнего суммирования могут не выписываться: достаточно подсчитать только их итог. Работа в значительной мере может быть упрощена, если выбрать «условное начало», которое в нашем случае равно 14,25 года. При этих условиях сплошное суммирование производится только в группах с возрастом меньшим, чем 14,25. При первом суммировании снизу (колонка 0) включается частота, относящаяся к условному началу, в последующих же колонках суммирование снизу каждый раз уменьшается на один шаг. В колонках с четным номером итоги суммирования сверху и снизу складываются, а в нечетных колонках итоги верхних частей вычитаются из итогов нижних частей колонок. Окончательные итоги обозначим $S_0, S_1, \ldots S_6$.

Таблица 30.4

Суммирование общего времени пробега

Возраст	Суммарное время	0	1	2	3
9,25	101,4	101,4	101,4	101,4	
9,75	127,2	228,6	330,0	431,4	
10,25	167,0	395,6	725,6	1 157,0	
10,75	445,2	840,8	1 566,4	2 723,4	
11,25	475,6	1 316,4	2 882,8	5 606,2	
11,75	713,0	2 029,4	4 912,2	10 518,4	
12,25	612,0	2 641,4	7 553,6	18 072,0	
12,75	800,3	3 441,7	10 995,3	29 067,3	
13,25	810,0	4 251,7	15 247,0	44 314,3	
13,75	943,8	5 195,5	20 442,5	64 756,8	
14,25	1 209,3	8 354,9			
14,75	958,5	7 145,6	28 656,9		
15,25	1 303,4	6 187,1	21 511,3	61 169,6	
15,75	1 075,2	4 883,7	15 324,2	39 658,3	
16,25	1 088,0	3 808,5	10 440,5	24 334,1	
16,75	830,8	2 720,5	6 632,0	13 893,6	
17,25	780,0	1 889,7	3 911,5	7 261,6	
17,75	541,2	1 109,7	2 021,8	3 350,1	
18,25	297,5	568,5	912,1	1 328,3	
18,75	198,4	271,0	343,6	416,2	
19,25	72,6	72,6	72,6	72,6	
		13 550,4	8 214,4	125 926,4	—86 433,4

Такой же процесс вычислений, но только с четырьмя этапами суммирования (от 0 до 3), проводится и по суммарному времени пробега по группам. В табл. 30.4 даны результаты этих вычислений при том же условном начале.

В предыдущих параграфах, где уравнение регрессии устанавливалось на основе полиномов специально упрощенной формы, коэффициенты регрессии определялись также весьма простыми уравнениями. Подобно этому и в данном случае соответствующие уравнения имеют упрощенную форму. Чтобы определить регрессию третьего порядка, необходимо вычислить четыре коэффициента.

В уравнениях, из которых определяются эти коэффициенты, в правых частях будут стоять итоги табл. 30.4, а коэффициенты при неизвестных в левых частях будут определяться по итогам S_0, S_1, ... S_6 табл. 30.3, в соответствии со следующей схемой:

S_0	S_1	S_2	S_3
S_1	$2S_2+S_1$	$3S_3+2S_2$	$4S_4+3S_3$
S_2	$3S_3+2S_2$	$6S_4+6S_3+S_2$	$10S_5+12S_4+3S_3$
S_3	$4S_4+3S_3$	$10S_5+12S_4+3S_3$	$20S_6+30S_5+12S_4+S_3$

При определении кривой более высокого порядка эта система уравнений соответствующим образом расширяется. Дополнение ее двумя или тремя строками и столбцами для получения кривой более высокого порядка может служить хорошим упражнением для сообразительности и мы предоставим это читателю. Итак, в нашем случае мы получим такие уравнения

$$\begin{aligned}
998A+\ 991B+\ 9054C-\ 3640D&=\\
991A+19099B+\ 7188C+\ 128264D&=\\
9054A+\ 7188B+195990C-\ 130388D&=\\
-3640A+128264B-130388C+1385032D&=
\end{aligned}$$

	Контроль
=	13550,4 20943,4
=	8214,4 163756,4
=	125926,4 207770.4
=	— 86433,4 1292834,6

Здесь неизвестные A, B, C и D являются соответственно полиномиальным значением Y, отсчитанным от условного начала, и его первыми тремя разностями возрастающего порядка. Решение этих уравнений полностью дано ниже. Так как коэффициенты левой части образуют симметричную матрицу, то можно из двух одинаковых коэффициентов оставить только один. Обработка этого примера проведена с контролем вычислений, который располагается в последней колонке, где даны суммы чисел каждого ряда, стоящего в обеих частях уравнения. Над этими числами контрольного столбца производятся те же действия, которые производятся над коэффициентами соответствующего ряда. Расчет

приведен в табл. 30.5, где расположение данных приспособлено к работе на счетной машине.

Таблица 30.5

Последовательные этапы при решении четырех уравнений

Коэффициенты уравнений				Правые части уравнений	Контрольный столбец
988	991	9 054	—3 640	13550,4	20943,4
	19 099	7 188	128 264	8214,4	163756,4
		195 990	—130 388	125926,4	207770,4
			1 385 032	—86433,4	1292834,6
1,355162	1,839448	12,06547		18,45312	33,71320
	10,00107	26,67970		22,46350	60,98372
		254,4514		163,1422	456,3388
1,992473	1,461470			27,27035	30,72429
	18,32980			13,63284	33,42411
34,385737					

После того как путем исключения неизвестных число уравнений сведено к одному, находится значение A, далее при подстановке его во второе уравнение определяется B и находится новое значение A из последних двух уравнений; этим же путем вычисляются C, B и A из трех уравнений и D, C, B и A из четырех первоначальных уравнений.

Таблица 30.6

Исправленные решения каждого уравнения

A	B	C	D
13,95742			
42	—0,3690990		
42	90	0,01802630	
42	83	49	0,01015438

Такая система контроля позволяет избежать все возможные арифметические ошибки и в то же время дает представление о достигнутой в процессе решения точности результатов.

Чтобы получить полиномиальные значения Y с двумя десятичными знаками, необходимо сохранить соответственно 3, 4, 5 и 6 знаков после запятой в значениях A, B, C и D. Построение полиномов производится путем последовательного наращивания разностей, как это показано в табл. 30.7. Очевидно, что при образовании вторых разностей на счетной машине следует визировать 6 знаков, хотя выписываются из них только четыре, полученные в результате округления. В остальном эта таблица не требует пояснений.

Таблица 30.7

Фактическое среднее время	Вычисленные полиномиальные значения	Разности		
		1-я	2-я	3-я
16,9	16,40			
		+0,009		
15,9	16,41		−0,0835	
		−0,074		
16,7	16,34		−0,0734	
		−0,148		
15,9	16,19		−0,0632	
		−0,211		
16,4	15,98		−0,0530	
		−0,264		
15,5	15,72		−0,0429	
		−0,307		
15,3	15,41		−0,0327	
		−0,340		
15,1	15,07		−0,0226	
		−0,362		
15,0	14,71		−0,0124	
		−0,375		
14,3	14,33		−0,0023	
		−0,377		
13,9	13,957		+0,0079	
		−0,3691		
13,5	13,59		0,01803	0,010154
		−0,351		
13,3	13,24		0,0282	
		−0,323		
12,8	12,91		0,0383	
		−0,285		
12,8	12,63		0,0485	
		−0,236		
12,4	12,39		0,0586	
		−0,178		
12,0	12,21		0,0688	
		−0,109		
12,3	12,11		0,0790	
		−0,030		
11,9	12,08		0,0891	
		+0,059		
12,4	12,14		0,0993	
		+0,159		
12,1	12,29			

Сумма квадратов, соответствующая полиномиальным значениям Y, взвешенным по относящимся к ним частотам, определяется обычным порядком, т. е. путем умножения коэффициентов регрессии на суммы, стоящие в правой части системы уравнений, определяющей эти коэффициенты. Так как в данном случае уравнения регрессии выражены в натуральных единицах измерения, а не в отклонениях от средней, то вычисленная по A сумма квадратов будет суммой квадратов не отклонений от средней, а отклонений от нуля. Для того чтобы перейти к отклонениям от средней, следует из прежней суммы квадратов, относящейся к полиномиальным значениям, вычесть величину $\frac{(13550,4)^2}{988}$, в результате чего получится сумма квадратов регрессии 1645,58 с тремя степенями свободы. Вычитая эту сумму из суммы квадратов с 20 степенями свободы, характеризующую различия между группами, получим остаток 31,24, относящийся к отклонениям от линии регрессии.

Вопрос о том, в какой мере выбранная нами форма кривой отражает фактическое изменение средней, может быть разрешен сравнением среднего квадрата, характеризующего отклонения от регрессии, со средним квадратом, относящимся к сравнениям внутри возрастных групп. Грей приводит средние квадратические отклонения для каждой возрастной группы. Вычисленная по этим данным общая сумма квадратов для сравнений внутри этих групп равна 1620,27. Итак, общая вариация данного материала может быть разложена на те три части, которые приведены в таблице дисперсионного анализа.

	Степени свободы	Сумма квадратов	Средний квадрат
Регрессия	3	1645,58	548,53
Остаточная вариация	17	31,24	1,838
Вариация внутри возрастных групп	967	1620,27	1,676
Итого	987	3297,09	—

Так как остаточный средний квадрат довольно близок к среднему квадрату внутри групповых сравнений, то отсюда следует, что для этих данных нельзя подобрать более подходящую кривую. Применяя здесь этот критерий, мы предвосхитили метод, который будет изложен в параграфе 44.

Таблица IV

Таблица значений t

n	$P=0,9$	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01
1	0,158	0,325	0,510	0,727	1,000	1,376	1 963	3,078	6 314	12,706	31,821	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2 132	2,776	3,747	4,604
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,394	0 538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1 313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
∞	0,12566	0,25335	0,38532	0,52440	0,67449	0,84162	1,03643	1,28155	1,64485	1,95996	2,32634	2,57582

ГЛАВА ШЕСТАЯ

КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

30. Среди статистических показателей нет ни одного, который бы более соответствовал биологическим задачам, чем коэффициент корреляции, и, пожалуй, нет такого статистического метода, который столь широко применялся бы к самым разнообразным данным, как корреляционный метод. В частности, данные наблюдения, если на них сказывается влияние самых разнообразных причин, не поддающихся регулированию, приобретают при помощи этого метода совсем иной смысл и значение. В экспериментальных работах он имеет менее важное значение; здесь он находит себе применение обычно на предварительных стадиях исследования, как например, тогда, когда два фактора, казавшиеся ранее независимыми, при ближайшем и специальном рассмотрении оказываются в определенной связи друг с другом. Однако в условиях эксперимента редко возникает необходимость выразить наши выводы в терминах теории корреляции.

Одним из наиболее ранних и удачных приложений корреляционного метода было применение его в учении о наследственности. В то время, когда ничего не было известно о механизме наследственности и о структуре зародышевых клеток, оказалось возможным при помощи этого метода показать наличие наследственности и «измерить ее силу» на организме, в отношении которого безусловно невозможно такое экспериментирование, а именно на человеке. Путем сравнения показателей, полученных в результате различного рода измерений человека, с результатами таких же измерений других организмов было установлено, что природа человека не в меньшей мере подчиняется законам наследственности, чем остальной животный мир. Эта аналогия была в дальнейшем расширена констатацией того факта, что примерно такие же корреляционные связи были получены как для характеристики умственного и морального склада человека, так и для его физических измерений.

Эти результаты имеют основное значение не только при исследовании наследственности человека, которая не поддается экспериментальному изучению, и не только для создания мето-

дов проверки умственных способностей, даже при отсутствии данных об имеющихся предрасположениях, но и для изучения организмов, в отношении которых возможны эксперименты. В этом последнем случае имеется то благоприятное обстоятельство, что некоторые факторы, определяющие модификационную изменчивость, могут разлагаться на части и их влияние может быть изучено на основе учения Менделя. Эта флюктуация, приблизительно подчиняющаяся нормальному распределению, является характерной чертой большинства полезных признаков у культурных растений и животных, и хотя имеются полные основания думать, что наследственность в таких случаях подчинена закону Менделя, все же биометрический метод изучения этого вопроса в настоящее время, пожалуй, является единственно способным внушить надежду на прогресс этого учения.

Тот факт, что в этом методе центральное положение занимает коэффициент корреляции, придает этому показателю особое значение даже для тех исследователей, которые намерены вести свой анализ несколько иными путями.

Табл. 31 является примером корреляционной таблицы. Она содержит представленные в компактной форме данные о росте 1376 отцов и дочерей (данные Пирсона и Ли). Эти данные разнесены по группам с интервалом в один дюйм; в тех случаях, когда рост выражался в целых дюймах, т. е. находился на границе двух групп, соответствующие численности дробились пополам, каждая из этих половин относилась к той и к другой из этих групп; так, например, рост отца в 67 дюймов отмечался $^1/_2$ в группе 66,5 и $^1/_2$ в группе 67,5. Подобный прием применялся и к росту дочерей и, следовательно, когда оба роста, т. е. рост отца и рост дочери, выражались целыми числами, то они отмечались по $^1/_4$ в каждом из четырех соседних квадратов. Это придает таблице несколько странный вид, поскольку она содержит в себе дроби, хотя по своему смыслу она должна давать распределение численностей, т. е. целых чисел. Применение такого дробления частот не является обязательным. Некоторое раздумье при выборе пределов группировки всегда поможет избежать всяких условностей такой группировки. Если дроблению подвергается большое число частот, то поправки Шеппарда уже не приводят к увеличению точности.

Первое, что бросается в глаза при рассмотрении этой таблицы, это то, что совсем не встречаются случаи, когда очень высокий отец имел бы дочь очень малого роста или, наоборот, у отца низкого роста была бы дочь очень высокого роста. Действительно, верхний правый и нижний левый углы таблицы остались не заполненными цифрами; это говорит за то, что такие случаи, если и можно наблюдать при выборке в 1400 наблюдений, то только весьма и весьма редко. Все наблюдения располагаются примерно внутри эллипса, размещенного по диагонали таблицы. Если мы наметим ту область, в которой частоты больше 10, то

окажется, что и эта область, если не считаться с вполне естественным нарушением этого правила, в целом подобна указанному эллипсу и имеет подобное же расположение. Частоты, взятые в любом направлении, увеличиваются по мере приближения к центральной области таблицы, где наблюдаются самые большие частоты, превосходящие 30. Линии одинаковых частот в целом образуют подобные и расположенные подобным образом эллипсы. В периферийной зоне наблюдения встречаются только случайно и поэтому здесь нет никакой регулярности; эта область может быть охвачена наблюдениями только при значительно большей выборке.

Наша корреляционная таблица разделена на четыре квадранта: линии сечения проведены от центральных значений той и другой переменной; эти значения равны 67,5 дюйма для роста отцов и 63,5 дюйма для роста дочерей; они близки к соответствующим средним. После такого деления корреляционной таблицы на четыре части становится отчетливо видно, что нижний правый и верхний левый квадранты более заполнены, чем два других, причем различие заключается не только в том, что первые имеют большее число заполненных клеток, но и в том, что соответствующие частоты являются более высокими. Это указывает на то, что мужчины высокого роста имеют высоких дочерей чаще, чем мужчины низкого роста, и наоборот. Корреляционный метод и служит для измерения степени подобного рода соответствия.

Итоговые графы этой корреляционной таблицы дают распределение численностей соответственно для отцов и для дочерей. В обоих случаях получены примерно нормальные распределения, что является частым случаем при работе с биологическим материалом, взятым без какого-либо преднамеренного отбора. В этом состоит часто встречающееся различие между биометрическим и экспериментальным материалом. При изучении рассматриваемого здесь вопроса экспериментатор, наверное, постарался бы иметь дело с двумя контрастными группами отцов, например, с ростом в 63 и 72 дюйма и отобрал бы отцов, принадлежащих только к этим крайним группам. В этом случае коэффициент корреляции, если экспериментатор решит им воспользоваться, не будет иметь никакого смысла. Такой «эксперимент» может служить только для установления регрессии роста дочери на рост отца и вместе с этим для определения влияния на рост дочерей того отбора, который был применен по отношению к отцам, но он не даст нам коэффициента корреляции, предназначенного для описания наблюдаемой совокупности такой, как она есть. Это описание может быть полностью испорчено, если произведен преднамеренный отбор данных.

Подобно тому, как нормальное распределение одной переменной может быть выражено при помощи формулы, в которой логарифм численности является квадратичной функцией этой переменной, также и в случае двух переменных численность может

Таблица 31

Рост дочерей в дюймах	Рост отцов в дюймах									
	58,5	59,5	60,5	61,5	62,5	63,5	64,5	65,5	66,5	67,5
52,5	—	—	—	—	0,25	0,25	—	—	—	—
53,5	—	—	—	—	0,25	0,25	—	—	—	—
54,5	—	—	—	—	—	—	—	—	—	—
55,5	—	—	—	—	—	—	—	1	—	—
56,5	0,25	0,25	—	0,25	1,25	0,5	—	1	0,5	0,5
57,5	0,25	0,25	0,5	1,5	4,5	1	1,5	1,5	2,5	—
58,5	0,25	0,75	0,5	0,75	0,75	1	1,75	1,25	5	1
59,5	0,5	1	2	—	6	4,75	5	6,25	11,75	4,5
60,5	0,75	0,75	—	2,5	8	6,25	12,5	18,25	20,25	14,5
61,5	—	0,5	1,75	2	9,75	11,5	13	23,75	23,75	15,5
62,5	—	1	2,25	2	4,5	12	22,75	26	33	48,5
63,5	—	—	0,25	2	6	8,25	11	27,25	35,75	37,25
64,5	—	—	0,25	2,5	1,75	3,25	9,25	23	18,75	28,5
65,5	—	—	—	0,5	1	0,5	11	12,25	9,25	19,75
66,5	—	—	—	0,5	0,5	1,5	3,25	7,25	8,75	16
67,5	—	—	—	—	—	—	1	5,75	7	4
68,5	—	—	—	—	0,25	0,25	0,25	0,25	1,5	3
69,5	—	—	—	—	0,25	0,25	0,25	0,25	0,25	0,25
70,5	—	—	—	—	—	—	—	—	—	—
71,5	—	—	—	—	—	—	—	—	—	—
72,5	—	—	—	—	—	—	—	—	—	—
Итого	2	4,5	7,5	14,5	45	51,5	92,5	155	178	175

68,5	69,5	70,5	71,5	72,5	73,5	74,5	75,5	Итог
—	—	—	—	—	—	—	—	0,5
—	—	—	—	—	—	—	—	0,5
—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	1
—	—	—	—	—	—	—	—	4,5
0,5	0,5	—	—	—	—	—	—	14,5
0,5	0,25	—	—	—	—	—	—	15,5
3,5	2	1,75	0,5	—	—	—	—	48,5
9	4,75	2,5	1,25	1,25	—	—	—	99
16,5	10,25	4,25	3	1,25	—	—	—	141,5
24,75	14,25	13,75	4,75	0,75	0,5	—	—	190,5
31,5	26,25	16,25	7,75	1,5	0,75	0,25	—	212
33	34,25	24,5	11,75	5,5	1	0,25	1	198,5
30	26,5	22,25	15	4,75	3,75	2	1	159,5
26,25	26,75	20,5	18,5	7,75	4,25	0,25	0,5	142,5
14,25	13,25	12	11,25	4,5	3,75	0,75	—	77,5
5,5	4,25	5,75	5,25	3,75	2,5	1,5	2	36
1	2,5	6,5	2,25	2,75	2	1	—	19,5
0,75	0,25	4,5	0,75	1,25	0,75	0,25	—	9,5
0,5	—	0,5	0,5	1,5	0,75	0,25	—	4
1	—	—	—	—	—	—	—	1
199,5	166	135	82,5	36,5	20	6,5	4,5	1376

быть выражена через квадратичную функцию теперь уже двух переменных. В этом случае мы будем иметь нормальную корреляционную поверхность, для которой частоты определяются формулой

$$df = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2}-\frac{2\rho xy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right\}} dx\,dy.$$

Здесь x и y являются отклонениями двух переменных от их средней, σ_1 и σ_2 — средние квадратические отклонения и ρ — корреляция между x и y. Корреляция ρ может быть положительной или отрицательной, но по своему абсолютному значению она не может быть больше единицы; она является отвлеченным числом и не имеет никакой физической меры. Если $\rho = 0$, то приведенное выше выражение превращается в произведение двух величин

$$\frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} dx \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_2^2}} dy.$$

Из этого следует, что нормальная поверхность двух переменных при отсутствии корреляции распадается на два распределения переменных x и y, изменяющихся в целом независимо друг от друга. При другом крайнем условии, когда ρ равно $+1$ или -1, изменения двух переменных находятся в строгом соответствии друг с другом, так что значение одной из них может быть абсолютно точно определено по значению второй. Другими словами, в данном случае мы имеем дело не с двумя переменными, а просто с двумя измерениями одной и той же варьирующей величины.

Если мы выделим случай, когда одна из переменных принимает некоторое фиксированное значение, то будем говорить о строе; столбцы и ряды корреляционной таблицы, если отвлечься от варьирования фиксированной переменной внутри интервалов, можно считать такими строями. В случае нормальной корреляции вариация внутри того или иного строя может быть определена по приведенной выше общей формуле, если в нее вместо x подставить фиксированное значение, скажем, a и разделить на суммарную численность, соответствующую этому значению a. В этом случае имеем:

$$df = \frac{1}{\sigma_2 \sqrt{2\pi}\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\rho\frac{a\sigma_2}{\sigma_1}\right)^2} dy.$$

Отсюда следует: 1) варьирование y внутри строя нормально, 2) среднее значение y для данного строя равно $\rho a\sigma_2/\sigma_1$, т. е. регрессия y на x линейна и имеет коэффициентом регрессии величину

$$\rho\frac{\sigma_2}{\sigma_1}$$

и 3) дисперсия y внутри строя равна $\sigma_2^2(1-\rho^2)$ и остается одной и той же в каждом строе. Это последнее положение можно сформулировать так: общая дисперсия переменной y в своей доле, равной $(1-\rho^2)$, не зависит от x, в то время как в другой оставшейся доле, равной ρ^2, она определяется значением x и может быть вычислена по этому значению.

Эти положения с соответствующей перестановкой символов полностью относятся и к переменной x: регрессия x на y линейна и имеет коэффициент регрессии $\rho\frac{\sigma_1}{\sigma_2}$. Следовательно, коэффициент корреляции ρ является средней геометрической из двух коэффициентов регрессии. Две линии регрессии, представляющие средние значения x соответственно отдельным значениям y и средние значения y соответственно отдельным значениям x, не могут совпадать, если только ρ не равно $+1$. Вариация x внутри строя, в котором y фиксировано, нормальна и характеризуется дисперсией $\sigma_1^2(1-\rho^2)$; поэтому можно сказать, что вариация x в своей доле, равной $(1-\rho^2)$, не зависит от y и в оставшейся доле, равной ρ^2, определяется значением y.

Таковы формальные математические следствия из определения нормальной корреляции. Многие биометрические данные определенно указывают на общую согласованность наблюдений с положениями, вытекающими из предложения о наличии нормальной корреляции; но я не хочу быть опрометчивым и утверждать, что этот вопрос не должен быть предметом критического исследования. Приближенная согласованность — это, возможно, все, что необходимо для более или менее обоснованного использования корреляции в качестве величины, характеризующей генеральную

совокупность. Ее значение в этом отношении несомненно, и нет ничего удивительного в том, что в ряде случаев она вместе с соответствующими средними и дисперсиями дает, по существу, полное описание совместного варьирования двух переменных.

31. Статистическая оценка корреляции

Дисперсия нормальной совокупности одной переменной величины имеет достаточную оценку, образующуюся из суммы квадратов отклонений от средней. Подобно этому достаточная оценка ковариации, при условии нормального распределения двух переменных, находится из суммы произведений. Достаточной же оценкой корреляции будет отношение ковариации к средней геометрической из двух дисперсий. Пусть x и y представляют собой отклонения двух переменных от их средних; в этом случае можно вычислить три статистики: s_1, s_2 и r следующим образом:

$$ns_1^2 = S(x^2); \quad ns_2^2 = S(y^2); \quad nrs_1s_2 = S(xy).$$

Здесь s_1 и s_2 — оценки средних квадратических отклонений σ_1 и σ_2, а r — оценка корреляции ρ. Эта оценка называется *коэффициентом корреляции*, или *совокупным моментом корреляции*, причем последний термин отмечает то положение, что в третье уравнение входит произведение xy. Величина n, естественно, должна являться числом степеней свободы, т. е. она равна числу парных наблюдений без единицы. Однако, как только величина n определена для r, она уже, собственно говоря, не имеет никакого отношения к расчету этого коэффициента, так как вычисления обычно более удобно проводить на основе суммы произведений без деления на n.

Этот метод расчета основан на том положении, что коэффициент корреляции генеральной совокупности является средней геометрической из двух коэффициентов регрессии; наши оценки этих двух регрессий следующие:

$$\frac{S(xy)}{S(x^2)} \quad \text{и} \quad \frac{S(xy)}{S(y^2)}$$

и поэтому оценка ρ будет:

$$r = \frac{S(xy)}{\sqrt{S(x^2)\cdot S(y^2)}}.$$

Пример 25. *Корреляция между ростом отцов и дочерей.* В табл. 32 дан конкретный пример расчета коэффициента корреляции. Первые восемь колонок не требуют объяснений, поскольку они повторяют обычный процесс вычисления средних и дисперсий для двух итоговых распределений. В данном случае нет необходимости вычислять фактическую среднюю, например, делением итога 3-й колонки 480,5 на 1376, так как всю работу можно провести, пользуясь неделенными итогами. Поправка на

то, что условное начало не совпадает с точной средней, вводится в 4-ю колонку вычитанием частного $(480{,}5)^2 : 1376$; подобная же поправка вводится в 8-й колонке, а также в последней колонке. Эта последняя поправка, вводимая в сумму произведений, вычисляется как частное $(480{,}5 \times 250{,}5) : 1376$; она может быть как положительной, так и отрицательной: если суммы отклонений для двух переменных противоположны по знаку, то она положительна, если одинаковы, то отрицательна. Внизу таблицы приведены суммы квадратов с поправкой Шеппарда (1376 : 12) и фактические, т. е. без этой поправки; в сумму произведений эта поправка не вводится.

Девятая колонка дает сумму отклонений от условного начала роста дочерей для каждого из 18 столбцов табл. 31. При небольших числах эти итоги легко определить в уме, в данном же случае, когда частоты довольно велики и их объединение усложняется еще дроблением их на четыре части, требуется известная осторожность при этих подсчетах. Итог 9-й колонки проверяет итог 3-й колонки. Чтобы этот контроль был возможен, следует в центральную часть 9-й колонки поставить +15,5, хотя это число при построении колонки «произведения» не участвует. Каждая цифра 9-й колонки умножается на соответствующее отклонение от условного начала роста отцов, что дает 10-ю колонку. В нашем случае все цифры этой 10-й колонки положительны, однако в этой колонке, вообще говоря, возможны как положительные, так и отрицательные числа. В таких случаях рекомендуется строить две колонки: одна для положительных произведений, другая для отрицательных. Для проверки вычислений лучше всего повторить работу, относящуюся к двум последним колонкам, заменяя переменные друг другом, т. е. произвести определение суммарных отклонений для роста отцов по каждому ряду и умножить их на отклонения роста дочерей. В обоих случаях должен получиться один и тот же итог 5136,25. Эта проверка особенно нужна для небольшой таблицы, когда быстрое проведение в уме вычислений двух последних колонок может часто сопровождаться ошибками.

Величина коэффициента корреляции с поправкой Шеппарда в данном случае определяется делением 5045,28 на среднюю геометрическую из 9209,0 и 10392,5 и равна +0,5157. Если же поправку Шеппарда не применять, то этот коэффициент будет равен +0,5097. В нашем случае различие между этими результатами невелико по сравнению с ошибкой случайной выборки. Поскольку никогда не может быть учтен весь эффект поправки Шеппарда на распределение коэффициента корреляции в случайных выборках, возникает определенное сомнение в том, что введение этой поправки дает улучшенную оценку корреляции и что ее необходимо вводить в этот коэффициент. Вместе с тем распределение в случайных выборках неисправленного значения коэффициента корреляции проще и с ним легче иметь дело. Поэтому при оценке существенности коэффициента корреляции, при

Таблица 32

Дочери: отклонения	Дочери: численность				Отцы: отклонения	Отцы: численность			Итоги для роста дочерей	Произведения
−11	0,5		5,5	60,5	−9	2	18	162	−8,75	+78,75
−10	0,5		5	50	−8	4,5	36	288	−15,25	+122
−9	—		—	—	−7	7,5	52,5	367,5	−19	+133
−8	1		—	64	−6	14,5	87	522	−23	+138
−7	4,5		31,5	220,5	−5	45	225	1125	−108,75	+543,75
−6	14,5		87	522	−4	51,5	206	824	−81	+324
−5	15,5		77,5	387,5	−3	92,5	277,5	832,5	−76,25	+228,75
−4	48,5		194	776	−2	155	310	620	−88,5	+177
−3	99		297	891	−1	178	178	178	−131,25	+131,25
−2	141,5		283	566	0	175	—	—	+15,5	—
−1	190,5		190,5	190,5		−1390				
0	212		−1179	—	1	199,5	199,5	199,5	+183,25	+181,25
1	198,5		198,5	198,5	2	166	332	664	+197,25	+394,5
2	159,5		319	638	3	135	405	1215	+245	+735
3	142,5		427,5	1282,5	4	82,5	330	1320	+174,75	+699
4	77,5		310	1240	5	36,5	182,5	912,5	+105,25	+526,25
5	36		180	900	6	20	120	720	+71,5	+429
6	19,5		117	702	7	6,5	45,5	318,5	+25,25	+176,25
7	9,5		66,5	465,5	8	4,5	36	288	+14,5	+116
8	4		32	256						
9	1		9	81						
	1376		+1659,5			1376	+1650,5		480,5	
			−1179				−1390			
	Итого.		+480,5	9491,5		Итого.	+260,5	10556,5	Итого	+5136,25
Поправка на среднюю				−167,8	Поправка на среднюю			−49,3	Поправка на среднюю	−90,97
				9323,7				10507,2		
Поправка Шеппарда				114,7	Поправка Шеппарда			114,7		
				9209				10392,5		+5045,28

которой, вообще говоря, следовало бы учитывать эту поправку, все же лучше основываться на неисправленном по Шеппарду значении этого показателя. Когда предполагается произвести оценку существенности, лучше не пользоваться грубыми группировками и тогда надобность в поправках Шеппарда сама собой отпадет. Уже один тот факт, что в малых выборках коэффициент корреляции, вычисленный с поправкой Шеппарда, может быть больше единицы, говорит о той путанице, которая связана с этой поправкой.

32. Частные корреляции

Значительное расширение понятия о корреляции связано с рассмотрением вопроса относительно связи нескольких переменных. Если взять три переменные величины и определить корреляцию между каждой парой из них, то появляется возможность исключить одну из этих переменных и тем самым определить корреляцию двух других переменных, которая соответствует совокупности, отобранной так, что эта третья переменная оставалась бы константной.

Если *на основе всего объема данных* будут установлены оценки для трех корреляций, то процесс элиминирования, описанный ниже, приведет к оценке частной корреляции, точно соответствующей той, которая могла бы быть определена непосредственно.

Пример 26. *Элиминирование возраста из корреляции между антропологическими показателями детей.* Для группы мальчиков различного возраста была установлена корреляция между *ростом стоя* и *окружностью груди*, равная +0,836 (данные Мамфорда и Янга). Можно ожидать, что некоторая часть этого соответствия между двумя показателями возникла благодаря связи между возрастом и ростом. Однако для решения этого вопроса фактически имеется только относительно небольшое число мальчиков одинакового возраста. Даже если мы возьмем такой широкий возрастной интервал, как целый год, все же у нас будет намного меньше наблюдений, чем при учете всех возрастов, взятых вместе. Если определить корреляцию между ростом стоя и возрастом, а также между окружностью груди и возрастом, то оказывается возможным решить вопрос на основе всего материала, не прибегая к его дроблению. В данном случае эти необходимые нам два коэффициента даны: 0,714 и 0,708.

Общая формула для вычисления частных коэффициентов корреляции имеет такой вид:

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}\cdot r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}.$$

Здесь три переменные занумерованы числами 1, 2 и 3. Наша задача состоит в том, чтобы определить корреляцию между переменными 1 и 2 при исключении переменной 3. Это будет «частная» корреляция между 1 и 2; соответствующий коэффициент частной корреляции обозначается $r_{12\cdot 3}$, где дается указание на исключение переменной 3. Символы r_{12}, r_{13}, r_{23} относятся к корреляциям, определяемым непосредственно между каждой парой переменных; эти корреляции в отличие от частных называются «общими».

Подставляя в приведенную выше формулу наши числовые значения общих коэффициентов корреляции, получим $r_{12\cdot 3} = 0{,}668$, откуда следует, что при элиминировании возраста корреляция, хотя и остается еще значительной, однако она все же заметно снижается. Упоминавшиеся ранее авторы вычислили среднее значение из коэффициентов корреляции, определенных по отдельным группам мальчиков с одним и тем же годом рождения, и оно оказалось равным 0,653, т. е. весьма близким к нашему $r_{12\cdot 3} = 0{,}668$.

Этим же путем могут быть последовательно элиминированы две или более переменных. Так, при четырех переменных можно сначала исключить переменную 4 путем трехкратного применения приведенной выше формулы для определения $r_{12\cdot 4}$, $r_{13\cdot 4}$ и $r_{23\cdot 4}$. После этого, применяя эту же самую формулу снова к этим трем коэффициентам, получим:

$$r_{12;34} = \frac{r_{12\cdot 4} - r_{13\cdot 4}\cdot r_{23\cdot 4}}{\sqrt{(1-r_{13\cdot 4}^2)(1-r_{23\cdot 4}^2)}}.$$

Следует отметить, что объем работы очень быстро возрастает по мере увеличения числа исключаемых переменных. При элиминировании s переменных число операций, каждая из которых состоит из вычислений по указанной выше формуле, будет составлять $\frac{1}{6}\, s(s+1)(s+2)$. Для значений s от 1 до 6 потребуется соответственно 1, 4, 10, 20, 35 и 56 операций.

Работа значительно упрощается, если воспользоваться таблицей $\sqrt{1-r^2}$, составленной Майнером.

Ранее указывалось, что независимые переменные уравнения регрессии не должны обязательно иметь точное или хотя бы приближенное нормальное распределение. То же самое относится и к элиминируемым переменным в корреляционном анализе. Вместе с тем, и это очень часто упускается из виду, случайные ошибки элиминируемых переменных ведут к систематическим ошибкам результатов элиминирования. Например, если частная корреляция переменных 1 и 2 фактически равна нулю, т. е. если r_{12} фактически равно $r_{13}\cdot r_{23}$, то случайные ошибки измерения или оценки переменной 3, имеющие тенденцию снизить абсолютные значения r_{13} и r_{23}, приведут к тому, что произведение этих коэффициентов чаще всего будут меньше r_{12}. Следовательно, случайные ошибки третьей переменной будут создавать кажущуюся корреляцию между первыми двумя переменными.

Необходимо хорошо понять сущность коэффициента корреляции. Например, задача измерения «силы наследственности», разрешаемая корреляционным методом, явным образом основывается на предположении о том, что целый ряд факторов определяет подобие родственных организмов в противоположность несходству неродственных индивидуумов и образует то, что мы называем наследственностью. Эта предпосылка, я думаю, вполне допустима при решении всех практических задач, но корреляция отнюдь не говорит нам, что это именно так и есть, она просто указывает на степень сходства, например между отцами и дочерями, в данной, фактически взятой совокупности тех и других. Она устанавливает, в какой мере рост отца содержит в себе информацию относительно роста дочери, или, в иной интерпретации, она указывает на относительную роль факторов, которые действуют одинаковым образом на рост отцов и дочерей, по сравнению со всей массой факторов, воздействующих на данный признак. Если мы знаем, что B зависит от A и от ряда других факторов, не зависящих от A, и что B не влияет на A, то корреляция между A и B указывает на то, сколь значительно по отношению к влиянию других воздействующих факторов это влияние фактора A. Если же у нас нет этих сведений, то корреляция не может сказать нам, будет ли A причиной B, или будет ли B причиной A, или же оба эти влияния находят себе место под воздействием каких-либо общих причин.

Все сказанное равным образом относится и к частной корреляции. Если нам известно, что феномен A сам по себе не оказывает влияния на некоторые другие феномены B, C, D..., а наоборот, сам находится под их непосредственным воздействием, то, вычисляя частные корреляции A с каждым феноменом из B, C, D... и каждый раз элиминируя остальные, мы тем самым произведем достаточно глубокий анализ причинности этого A. Если же, наоборот, мы возьмем группу, например, социальных феноменов, не зная априорно ничего о форме причинной связи между ними и даже о наличии такой казуальной связи, то вычисление как общего, так и частного коэффициента корреляции не продвинет нас ни на шаг вперед в отношении оценки удельного веса отдельных причин явления.

Корреляция между A и B в удобной форме измеряет удельный вес факторов, воздействующих сходным образом на A и B, среди прочих факторов, от которых связь A и B не зависит. Если мы элиминируем некоторую третью переменную C, то тем самым выводим из сравнений те факторы, которые становятся недействующими, когда C фиксировано. Если группа этих последних состоит только из таких факторов, от которых не зависят A и B, то корреляция между A и B, будет ли она положительной или отрицательной, численно увеличится. Здесь мы исключаем не имеющие прямого отношения к A и B факторы, затемняющие картину, и в соответствии с этим получаем как бы более контролируемый эксперимент. Вместе с тем возможны и такие случаи, когда элиминируется переменная C, факторы которой воздействуют одинаково или в противоположном направлении на две коррелируемые переменные. В этих случаях изменчивость C фактически маскирует собой ту связь между A и B, которую мы исследуем, и поэтому элиминирование C делает эту связь более явной. В третьих, C может быть одним из промежуточных звеньев той цепи явлений, посредством которой A воздействует на B или наоборот. Та роль, которую выполняет C в качестве канала, пропускающего через себя это влияние, может быть оценена посредством элиминирования C. В качестве примера для этого случая можно взять факт незначительного влияния скрытых факторов наследственности у человека, определяемого корреляцией между внуками и дедами при элиминировании родителей, стоящих между ними. Однако мы не можем с полной уверенностью предугадать, будет ли польза от элиминирования некоторых переменных, до тех пор, пока мы не узнаем соответствующую данному случаю качественную схему казуальности или, по крайней мере, пока не будем иметь основания предположить такую схему. Конечно, для чисто описательных целей, т. е. при установлении особенностей совокупности нескольких переменных, как частная, так и общая корреляция всегда сохраняют свое значение и в этом аспекте корреляция любого типа может представлять несомненный интерес.

В качестве иллюстрации ко всему сказанному рассмотрим, в каком смысле коэффициент корреляции является мерой «силы наследственности», допуская, что только наследственность сосредоточивает в себе причины сходства между родственниками, т. е. что любые другие сопутствующие эффекты распределены чисто случайно. Прежде всего следует отметить, что если влияние таких сопутствующих условий возрастает по своей силе, то корреляция будет уменьшаться. Так, одна и та же совокупность будет давать более высокую корреляцию, если она, говоря генетическим языком, воспитана при относительно однородном режиме питания, чем в том случае, когда этот режим весьма разнообразен. Однако генетический процесс, интересующий нас, будет в обоих случаях одинаков. Во-вторых, если влияние сопутствующих условий вообще значительно, то все же в смешанной совокупности с весьма разнообразной силой наследственности мы можем наблюдать более высокую корреляцию, чем это будет при более однородной совокупности. В-третьих, хотя, например, влияние отца на дочь является в определенном смысле непосредственным, а именно в том смысле, что он вносит свою долю в образование зародыша дочери, однако нельзя считать, что только от этого факта полностью зависит корреляция в целом, ибо, как это было доказано, муж и жена по своему росту также обнаруживают значительное сходство, и поэтому у высоких отцов, как правило, будут высокие дочери, отчасти и потому, что у них были жены высокого роста. По этой причине, например, следует ожидать определенно положительную

корреляцию между падчерицами и отчимами. Следует также ожидать, что при элиминировании роста матери частная корреляция между ростом отца и ростом дочери будет несколько меньшей, чем общая корреляция. Такое подробное обсуждение данного вопроса имеет большое значение для определения смысла, который следует вкладывать в несколько неясный термин «сила наследственности» в тех случаях, когда это явление изучается корреляционным методом. Необходимо полное понимание того, что и в других, менее изученных, случаях подобное обсуждение вопроса имеет такое же значение и должно проводиться со всей возможной скрупулезностью и полнотой.

33. Точность коэффициента корреляции

При большой выборке и при умеренной или низкой корреляции коэффициент корреляции, вычисленный по выборке, содержащей n парных наблюдений, распределен нормально около точного значения ρ с дисперсией

$$\frac{(1-\rho^2)^2}{n-1}.$$

Поэтому, произведя обычную замену ρ на выборочное значение r, мы получим среднюю квадратическую ошибку

$$\frac{(1-r^2)}{\sqrt{n-1}} \quad \text{или} \quad \frac{(1-r^2)}{\sqrt{n}}.$$

В малых выборках значение r может довольно сильно отклоняться от точной величины ρ и поэтому разность $1-r^2$ будет содержать в себе более или менее крупную ошибку. Мало того, распределение r в этих случаях значительно отличается от нормального, в связи с чем оценка существенности, если пользоваться формулой для большой выборки, в этих случаях часто становится весьма обманчивой. Так обстоит дело при малых выборках (меньших 100), обычно встречающихся в практике исследовательской работы при изучении различного рода связей. Эти затруднения можно преодолеть при помощи более точных методов оценки, описание которых дается ниже.

Эти методы в одинаковой мере применимы как к общей, так и частной корреляции; следует только учитывать следующее различие между этими двумя случаями: при расчете числа степеней свободы следует вычитать столько дополнительных единиц, сколько элиминировано переменных. Так, частная корреляция, найденная путем исключения трех переменных при 13 наблюдениях, имеет такое же распределение, какое имеет общая корреляция того же размера при 10 наблюдениях.

34. Существенность коэффициента корреляции

Определяя существенность коэффициента корреляции, мы должны найти вероятность появления такой корреляции в случайной выборке из некоррелированной совокупности. Если эта вероятность достаточно мала, то мы можем считать корреляцию существенной. Для такой оценки существенности можно воспользоваться таблицей значений t, приведенной в предыдущей главе (стр. 144). Если n' — число парных наблюдений, на основе которых вычислен коэффициент корреляции r (без поправки Шеппарда), то можно найти:

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n'-2}$$

$$n = n' - 2.$$

Можно показать, что распределение вычисленного таким образом t соответствует закону распределения, представленному в указанной таблице.

Можно видеть, что этот критерий, как того и следовало ожидать, идентичен с рассмотренным в предыдущей главе критерием, устанавливающим существенность отклонения от нуля наблюденного коэффициента линейной регрессии.

Таблица V A (стр. 171) дает этот критерий в такой форме, которая позволяет определить существенность r непосредственно по самому значению r. В этой таблице дано четыре уровня существенности, представленных вероятностями P, равными 0,10; 0,05; 0,02 и 0,01, и значения n сплошь от 1 до 20 и далее от 20 до 100 с некоторыми интервалами.

Пример 27. *Существенность коэффициента корреляции между урожаем пшеницы и осенними осадками.* По наблюдениям, проведенным в Восточной Англии в течение двадцати лет — с 1885 по 1904 г., — была определена корреляция между валовым урожаем пшеницы и количеством осенних осадков. Коэффициент корреляции оказался равным — 0,629. Можно ли считать это значение существенным? Для решения вопроса находим последовательно:

$$1 - r^2 = 0{,}6044$$

$$\sqrt{1-r^2} = 0{,}7774$$

$$\frac{r}{\sqrt{1-r^2}} = -0{,}8091$$

$$t = -3{,}433.$$

Взяв $n = 18$, устанавливаем, что P меньше 0,01 и что, следовательно, корреляция определенно существенна. К тому же выводу можно прийти и без вычислений, если воспользоваться таблицей VA при $n = 18$.

Если же мы вычислим и применим для оценки существенности среднюю квадратическую ошибку

$$\sigma_r = \frac{1 - r^2}{\sqrt{n' - 1}},$$

то получим

$$t = \frac{r}{\sigma_r} = \frac{r}{1 - r^2}\sqrt{n' - 1} = -4{,}536,$$

т. е. значительно больше, чем вычисленная ранее величина t, в связи с чем у нас получится сильно завышенная оценка существенности. Мало того, сделав при этой оценке допущение о нормальном распределении r, т. е. допустив, что $n = \infty$, мы произвели дальнейшее завышение существенности. Этот последний расчет здесь приведен для того, чтобы показать, в какой мере при малых выборках является обманчивым метод оценки при помощи средней квадратической ошибки σ_r, базирующейся на допущении о нормальном распределении коэффициента корреляции. Но без этого допущения нет никаких оснований для применения с этой целью средней квадратической ошибки. Этот вводящий в заблуждение характер последней формулы для t еще более усиливается довольно часто практикуемой заменой $n' - 1$ на n'. Если судить по отклонению $t = 4{,}536$ при условии нормального распределения, то следует считать, что корреляция такого или большего размера может появиться в случайных выборках из фактически некоррелированной совокупности в 6 из миллиона случаев. На самом же деле корреляция большего размера, чем наблюденная, может встретиться в 3000 случаев из миллиона, т. е. в 500 раз более часто, чем это считалось при указанном допущении.

Читатель должен со всей серьезностью отнестись к этому предупреждению относительно вводящего в заблуждение характера средней квадратической ошибки коэффициента корреляции, определенного по небольшой выборке, так как использование коэффициента корреляции при вычислении его средней квадратической ошибки, по существу, является применением его в условиях, о которых мало что известно и которые характеризуются весьма скудными данными. При наличии достаточно обширного материала, обычного при биометрических исследованиях, вообще говоря, нет большой опасности, производя оценку существенности коэффициента корреляции этим методом, прийти к неправильным выводам, но при сравнительно небольшом числе наблюдений, на которых, как правило, основывается экспериментатор, механическое перенесение методов, оправдавших себя в биометрике, может столь часто приводить к заблуждениям, что это подорвет доверие и к более совершенным средствам исследования. Совершенно неправильно, и это доказывается предыдущим примером, будто основываясь на малой выборке, нельзя сделать никаких ценных выводов. Если при определении вероятности будут применены

точные методы, полностью учитывающие размер выборки, то на наше суждение окажет влияние только величина этой вероятности. При правильном методе вычисления вероятности находит свое правильное отражение и то возрастание уверенности, которое обусловлено увеличением числа наблюдений.

Пример 28. *Существенность частного коэффициента корреляции.* Исследуя данные 32 обществ помощи бедным, Юл нашел, что за период времени с 1881 по 1891 г. относительное изменение процента рабочих, получающих пособие, было коррелировано с изменением отношения числа получателей пособия на дому к числу лиц, помещенных в работных домах. При этом были элиминированы две переменные: изменение процента лиц старше 65 лет и размер самой совокупности.

Вычисленная Юлом корреляция после исключения двух переменных оказалась равной $+0{,}457$. Такая корреляция называется частной корреляцией второго порядка. Определим ее существенность.

Ранее указывалось, что распределение частных коэффициентов корреляции можно получить из распределения общих коэффициентов корреляции путем простого вычитания из численности выборки числа исключенных переменных. Вычитая 32—2, получим 30 в качестве условного объема выборки, отсюда:

$$n = 28.$$

Вычисляя, как и раньше, величину t, получим

$$t = 2{,}719.$$

По таблице значений t находим, что P лежит между 0,02 и 0,01. Поэтому данная корреляция существенна. Здесь, конечно, как и в других случаях, имеет место допущение, что коррелируемые переменные (но не обязательно элиминируемые) распределены нормально. Экономические показатели редко имеют нормальное распределение, но то обстоятельство, что в данном случае мы имеем дело со скоростью изменения, во много раз увеличивает возможность нормального распределения. Вероятности из таблицы VA для $n = 25$ и $n = 30$ также указывают на существенность этого коэффициента.

35. Преобразование корреляции

Кроме рассмотренной выше задачи об оценке существенности коэффициента корреляции, которая сводится к решению вопроса о наличии факторов, определяющих сопряженность признаков, довольно часто возникает необходимость произвести одну или несколько из перечисленных ниже операций, каждая из которых требует применения среднего квадратического отклонения в нормальном распределении. Мы видели, что при большой выборке, исключая случаи, когда корреляция близка к ± 1, распределение коэффициента корреляции может считаться нормальным и по-

этому все вопросы оценки в этих случаях могут быть решены при помощи средней квадратической ошибки коэффициента корреляции. Но в случае малых выборок, часто встречающихся на практике, приходится применять специальные методы, позволяющие получить реальные и освобожденные от грубых систематических ошибок результаты. Эти методы направлены на решение следующих вопросов:

1. Определение того, что наблюденная корреляция существенно отличается от некоторого теоретического значения.

2. Определение того, что две наблюденные корреляции существенно отличаются друг от друга.

3. Определение некоторой обобщенной оценки корреляции, когда имеется некоторое число независимых оценок ее.

4. Применение критериев пп. 1 и 2 к такой усередненной оценке.

Способ решения всех этих задач аналогичен тому, при помощи которого была решена задача оценки существенности отдельного коэффициента корреляции. В этом последнем случае по данному значению r определялась величина t, распределение которой известно и отражено в соответствующей таблице. Таким образом, преобразование r привело нас к хорошо изученному распределению, позволяющему найти точное значение вероятности P. Подобно этому и преобразование, которое будет сейчас рассмотрено, также приводит, хотя только с известным приближением, к хорошо изученному нормальному распределению, основываясь на котором, не представляет никаких трудностей провести исследование указанных выше вопросов.

Возьмем величину:

$$z = \frac{1}{2}[\log_e(1+r) - \log_e(1-r)] = r + \frac{1}{3}r^2 + \frac{1}{5}r^5 + \dots$$

Когда r меняется от 0 до 1, величина z принимает значения от 0 до ∞. Для небольших значений r величина z близка к r, но по мере приближения r к единице величина z возрастает беспредельно. Преимущество этого преобразования r в величину z определяется особенностями распределений этих двух величин в случайных выборках. Среднее квадратическое отклонение r зависит от точного значения корреляции ρ, как это следует из формулы

$$\sigma_r = \frac{1-\rho^2}{\sqrt{n'-1}}.$$

Так как ρ неизвестно, то вместо этой величины мы подставляем фактическое значение r, но при малой выборке r не является достаточно точной оценкой ρ. Средняя же квадратическая ошибка z примерно равна

$$\sigma_z = \frac{1}{\sqrt{n'-3}}$$

и практически не зависит от величины корреляции в генеральной совокупности, к которой принадлежит данная выборка.

Кроме этого, распределение r в малых выборках не является нормальным и даже в больших выборках при высоких значениях корреляции это распределение значительно отклоняется от нормального. Распределение же величины z, хотя и не является строго нормальным, все же с большой скоростью приближается к нормальному по мере увеличения размера выборки при любых значениях корреляции в генеральной совокупности. Вопрос об отклонениях распределения z от нормального будет рассмотрен в дальнейшем.

Наконец, форма распределения r довольно быстро меняется с изменением ρ и, следовательно, в этом случае нельзя построить некоторое стандартное распределение, пригодное для всех случаев. Наоборот, распределение z всегда имеет примерно одинаковую форму и поэтому путем введения небольшой поправки на отклонение его от нормального распределения можно получить вполне точные критерии. Надо сказать, что эта поправка столь мала, что не представляет практического значения, и нам не придется даже иметь с ней дело. Во всех обычных случаях вполне достаточно считать, что распределение z нормально.

Рис. 7 и 8 иллюстрируют те преимущества, которые имеет распределение z перед распределением r. На рис. 7 даны фактические распределения r при 8 парах наблюдений и при условии, что в генеральной совокупности корреляция равна 0 и 0,8. На рис. 8 даны соответствующие кривые распределения z. Две кривые на рис. 7 резко отличаются друг от друга по своим модальным высотам; обе они явно не являются нормальными кривыми; по своей форме они также различны: одна из них симметрична, другая асимметрична. Наоборот, на рис. 8 обе кривые не отличаются значительно друг от друга по своей высоте; хотя по форме они и не являются строго нормальными, однако они столь близки к этой форме, что даже при малой выборке, содержащей только 8 парных наблюдений, на глаз нельзя установить это различие; это приближение к нормальной форме распределения сохраняется даже при крайних значениях $\rho = \pm 1$. Следует обратить внимание на один факт, который обнаруживается при знакомстве с рис. 8. Распределение z для $\rho = 0{,}8$ представляет собой симметричную кривую, характерную тем, что ордината нулевой ошибки не проходит через центральную точку этого распределения, а отклонена несколько влево. Это обусловлено тем, что при оценке коэффициента корреляции возникает некоторое небольшое смещение. Это смещение в дальнейшем будет нами учитываться. В следующей главе мы встретимся с подобным же смещением, возникающим при определении внутриклассовой корреляции.

Для облегчения преобразования r в z мы даем таблицу VB (стр. 172—173) значений r, соответствующих значениям z от 0 до 3

через 0,01. При знакомстве с этой таблицей можно видеть, что в начальной ее части наблюдаются незначительные различия между r и z, но при более высоких значениях корреляции небольшим

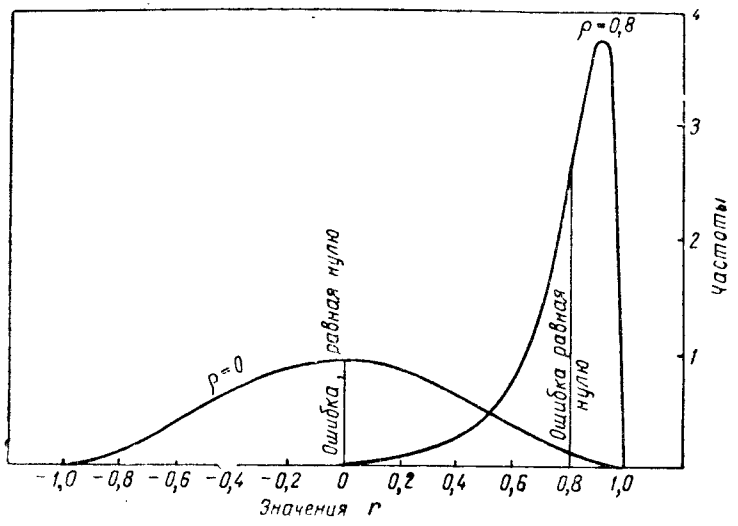


Рис. 7.

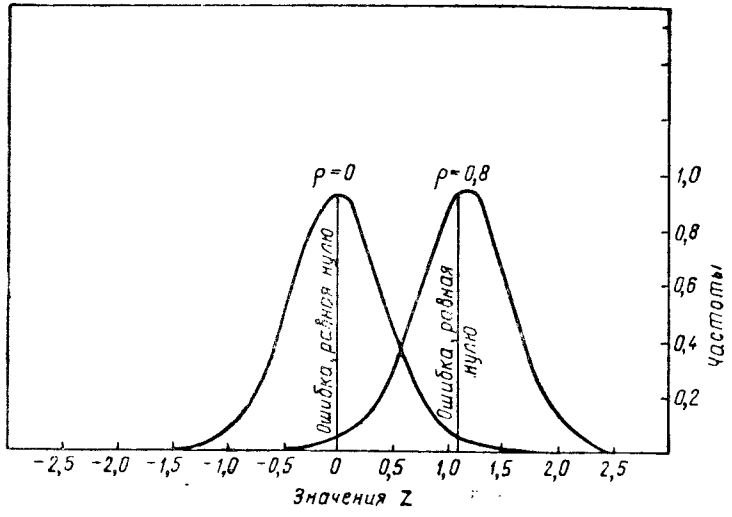


Рис. 8.

приростам r соответствуют уже относительно большие приросты z. Так, например, различие между корреляциями 0,99 и 0,95, взятое по шкале z, оказывается бо́льшим, чем различие по этой же шкале между корреляциями 0 и 0,6. Вообще, величина z дает

более правильное представление о соотношениях связей различного уровня, чем это доступно для величины r.

Для того чтобы по этой таблице определить z, соответствующее данному значению r (положим, 0,6), следует сначала внутри таблицы найти два ближайших к 0,6 числа и определить соответствующие z. Таким образом, можно видеть, что z лежит в этом случае между 0,69 и 0,70. После этого интервал между этими числами следует разделить на части, пропорциональные недостатку и избытку табличных значений r по отношению к 0,6, и первую часть добавить к 0,69. Так, в нашем случае имеем $20/64 = 0{,}31$ и, следовательно, $z = 0{,}6931$. При определении r по данному z, (предположим, $z = 0{,}9218$) можно сразу найти, что r лежит между 0,7259 и 0,7306. Разность между этими значениями равна 0,0047, а 18% от нее составит 0,0008. Добавляя это последнее число к меньшему из значений r, получим окончательно $r = 0{,}7267$. Следовательно, одна и та же таблица VB позволяет перейти от r к z и обратно от z к r.

Пример 29. *Проверка нормальности распределения z.* Чтобы показать ту степень точности, которую дает критерий z, возьмем данные, обработанные ранее в примере 27 при помощи точного метода. В этом случае корреляция, равная минус 0,629, получена на основе 20 парных наблюдений. Оценим ее существенность методом z.

Для $r = -0{,}629$ при помощи таблицы натуральных логарифмов или при помощи таблицы VB находим $z = -0{,}7398$. Деление этой величины на ее среднюю квадратическую ошибку равносильно умножению ее на $\sqrt{17}$. Это дает $-3{,}050$ и должно рассматриваться как отклонение в нормальном распределении. Из таблицы для нормального распределения видно, что отклонения, большие данного, могут встретиться примерно в 23 случаях из 10 000. Точное значение этого шанса для $r = -0{,}629$, как это было определено в примере 27, составляет 30 на 10 000 случаев. Таким образом, полученная здесь погрешность только незначительно преувеличивает существенность данного коэффициента корреляции.

Пример 30. *Еще одна проверка нормальности распределения z.* Частный коэффициент корреляции $+0{,}457$ был установлен на основе выборки из 32 наблюдений и путем элиминирования двух переменных. Отличается ли существенно этот коэффициент от нуля? Здесь $z = 0{,}4935$; после исключения двух переменных, эффективный, т. е. учитываемый объем выборки, становится равным 30, а средняя квадратическая ошибка равна $1/\sqrt{27}$. Умножая z на $\sqrt{27}$, получим нормированное отклонение 2,564. Таблица I (или последняя строка таблицы IV) показывает, что P примерно равно 0,01. Здесь опять имеет место небольшое преувеличение существенности, но оно еще меньше, чем в предыдущем примере.

Эти примеры показывают, что преобразование r приводит к

переменной z, которая в большинстве практических случаев может считаться распределенной по нормальному закону. В этом случае также можно применить и более простой способ оценки существенности, основанный на таблице t. Но в последующих примерах этот последний метод уже не может применяться и только применение метода z дает возможность быстро получить достаточную для практических целей точность оценки.

Пример 31. *Существенность отклонения фактического значения коэффициента корреляции от теоретически ожидаемого его значения.* В выборке, состоящей из 25 параллельных обмеров родителей и детей, был установлен коэффициент корреляции 0,60. Можно ли считать, что это значение совместимо с допущением, что точная корреляция изучаемых признаков равна 0,46?

Сначала находим разность между соответствующими значениями z; эта разность определена в табл. 33.

Таблица 33

	r	z
Выборочное значение	0,60	0,6931
Значение в генеральной совокупности	0,46	0,4973
Разность		0,1958

Чтобы определить нормированное отклонение, умножаем эту разность на $\sqrt{22}$, в результате получим 0,918. Это отклонение меньше единицы и поэтому фактическое значение корреляции находится в достаточном согласии с нашей гипотезой.

Пример 32. *Существенность различия между двумя коэффициентами корреляции.* В двух выборках, из которых в первой было 20, а во второй 25 парных наблюдений, получены коэффициенты корреляции соответственно 0,6 и 0,8. Можно ли считать существенным различие этих значений?

В этом случае необходимо определить не только разность значений z, но и среднюю квадратическую ошибку этой разности. Дисперсия разности z равна сумме чисел, обратных 17 и 22. Обработка данных приведена в табл. 34.

Таблица 34

	r	z	$n'-3$	Обратное число
1-я выборка . . .	0,60	0,6931	17	0,05882
2-я выборка . . .	0,80	1,0986	22	0,04545
Разность	—	$0,4055 \pm 0,3230$	Сумма	0,10427

Средняя квадратическая ошибка, поставленная около разности со знаками плюс и минус, получена как корень квадратный из

дисперсии, вычисленной в последней колонке. Разность не превосходит свою ошибку в два раза и поэтому она не может считаться существенной. Таким образом, у нас нет никаких оснований отрицать возможность появления этих двух выборок из совокупностей, имеющих одинаковую корреляцию.

Пример 33. *Объединение корреляций, вычисленных на основе малых выборок.* Положим, что две выборки последнего примера были получены из совокупностей с одинаковой корреляцией; определим оценку такой обобщенной корреляции.

В этом случае два значения z должны быть взвешены обратно пропорционально их дисперсиям. Поэтому умножим первое значение на 17, а второе на 22, сложим произведения и результат разделим на 39. Это и будет оценкой обобщенной величины z; соответствующее же значение r может быть найдено по таблице VB.

Таблица 35

	r	z	$n'-3$	$(n'-3)\,z$
1-я выборка	0,60	0,6931	17	11,7827
2-я выборка	0,80	1,0986	22	24,1692
	0,7267	0,9218	39	35,9519

Взвешенная средняя величина z равна 0,9218; этому соответствует $r=0,7267$. Найденное этим способом z распределено с дисперсией, равной $1/39$. Поэтому ее точность эквивалентна такой точности, которая была бы в выборке из 42 парных наблюдений. Следовательно, к этой усредненной величине z можно применить тот же критерий существенности, который применялся ранее к оценке отдельных значений z.

36. Систематические ошибки

При объединении корреляций, определенных по малым выборкам, выступают на сцену два типа систематических ошибок, которые имеют весьма малое значение в каждом отдельном случае, но приобретают уже известный вес, когда производится усреднение показателя связи для некоторого числа выборок.

Значение z, полученное из некоторой выборки, является оценкой некоторой точной величины ζ, относящейся к генеральной совокупности, подобно тому, как выборочное значение r является оценкой ρ в генеральной совокупности. Если бы процесс определения корреляции был бы свободен от смещения, то значения z были бы нормально распределены около средней $\bar{z}$, которая совпадала бы с ζ. Фактически же имеет место некоторое небольшое смещение, которое делает среднюю величину z численно не-

сколько большей, чем ζ, поэтому корреляция как положительная, так и отрицательная всегда будет слегка преувеличенной. Это смещение может быть корректировано путем вычитания из z поправки

$$\frac{\rho}{2(n'-1)}.$$

Когда берется одна выборка, эта поправка не имеет никакого значения, так как она обычно мала по сравнению со средней квадратической ошибкой величины z. Например, если $n'=10$, то средняя квадратическая ошибка z равна 0,378, в то время как поправка равна $\rho/18$, т. е. не больше 0,056. Однако, если $\bar{z}$ является средней из 1000 таких отдельных значений z, каждое из которых получено по выборке из 10 наблюдений, то средняя квадратическая ошибка в этом случае будет только 0,012 и поправка, которая не меняется при усреднении z, приобретает уже большой вес.

Второй вид систематических ошибок связан с применением поправок Шеппарда. При вычислении z мы всегда должны основываться на значении r без поправок Шеппарда, так как последние усложняют форму распределения z. Но исключение поправок Шеппарда вводит некоторую систематическую ошибку, которая направлена в противоположную сторону по отношению к предыдущей систематической ошибке. Хотя эта ошибка по своей величине очень мала, однако она быстро возрастает по мере уменьшения размера выборки. В тех случаях, когда производится обобщение корреляций при грубой группировке малых выборок, среднее значение z должно определяться по значениям r без поправок Шеппарда, после чего можно в окончательный результат ввести поправку, представляющую усредненный эффект поправок Шеппарда.

37. Корреляция между рядами наблюдений

В данном случае речь идет об установлении связи между рядами наблюдений (например, ежегодных данных) произведенных через равные промежутки времени. Здесь мы встречаемся с особым случаем применения частной корреляции, хотя соответствующий материал может быть обработан и по методу криволинейной регрессии, описанному в параграфе 27 (стр. 122).

Если, например, мы имеем данные о количестве смертных случаев от некоторой болезни за ряд следующих друг за другом лет и желаем установить, находится ли эта смертность в связи с метеорологическими условиями или с распространением какой-либо другой болезни, или со смертностью в некоторой другой возрастной группе и т. д., то серьезным затруднением при непосредственном применении коэффициента корреляции является то, что число этих смертных случаев будет, вероятно, меняться в течение данного периода в какой-либо прогрессии и под влиянием постоянно действующих факторов. Эти изменения могут произойти или под влиянием изменения самой численности населения, среди которого наблюдается смертность от данной болезни (безразлично, будет ли это население некоторой области или численность той или иной группы этого населения), или под влиянием изменений в санитарных условиях жизни этого населения и усовершенствования медицинского обслуживания, или, наконец, под влиянием изменений в расовом и генетическом составе населения. В ряде случаев наблюдается, что такая изменчивость все же остается и после того, когда вместо абсолютного числа смертных случаев берется относительный показатель смертности, в котором учтено изменение размера совокупности, так как этим путем исключается только одна сторона изменения совокупности.

Если такое прогрессивное изменение может быть достаточно хорошо изображено прямой линией, то в этом случае можно считать *время* третьей переменной и исключить его путем вычисления соответствующего частного коэффициента корреляции. Однако чаще всего это изменение не имеет столь простой формы и поэтому возникает необходимость ввести в ее выражение вторую или еще более высокую степень времени. В этом случае частная корреляция должна быть найдена не только элиминированием t, но и элиминированием t^2, t^3, t^4, ..., рассматриваемых в качестве отдельных переменных, ибо если мы элиминируем эти степени вплоть, скажем, до 4-й, то тем самым получаем возможность косвенным образом элиминировать из корреляции любую функцию времени в 4-й степени, включая и ту, которая наилучшим образом отражает в себе данное прогрессивное изменение.

Эту частную корреляцию можно определить непосредственно по коэффициентам регрессии, определяемым согласно параграфу 28 (стр. 126). Если y и y' — две коррелируемые величины, то можно найти для y коэффициенты A, B, C... и для y' коэффициенты A', B', C'..... Суммы квадратов отклонений этих переменных от линий регрессии, как и прежде, определяются по равенствам:

$$S(y-Y)^2 = S(y^2) - n'A^2 - \frac{n'(n'^2-1)}{12}B^2 -$$

$$S(y'-Y')^2 = S(y'^2) - n'A'^2 - \frac{n'(n'^2-1)}{12}B'^2 -$$

Сумма же произведений определяется по аналогичной формуле

$$S\{(y-Y)(y'-Y')\} = S(yy') - n'AA' - \frac{n'(n'^2-1)}{12}BB' \ldots$$

Отсюда находится частная корреляция:

$$r = \frac{S[(y - Y)(y' - Y')]}{\sqrt{S(y - Y)^2 \cdot S(y' - Y')^2}}.$$

Число элиминируемых переменных здесь равно степени t, до которой доведено развертывание уравнения регрессии. Обе переменные y и y' должны быть выражены регрессиями одинаковой степени, даже если одна из них довольно хорошо может быть охарактеризована уравнением более низкого порядка, чем другая.

Таблица VA

Значения коэффициента корреляции для различных уровней существенности

n	$P = 0,1$	0,05	0,02	0,01
1	0,98769	0,996917	0,9995066	0,9998766
2	0,90000	0,95000	0,98000	0,990000
3	0,8054	0,8783	0,93433	0,95873
4	0,7293	0,8114	0,8822	0,91720
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3884	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0.2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Для общего коэффициента корреляции n на 2 единицы меньше, чем число парных наблюдений в выборке; для частного коэффициента корреляции, кроме этого, следует вычесть число элиминируемых переменных.

Таблица r для значений z от 0 до 3

z	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	0,10
0,0	0,0100	0,0200	0,0300	0,0400	0,0500	0,0599	0,0699	0,0798	0,0898	0,0997
0,1	0,1096	0,1194	0,1293	0,1391	0,1489	0,1586	0,1684	0,1781	0,1877	0,1974
0,2	0,2070	0,2165	0,2260	0,2355	0,2449	0,2543	0,2636	0,2729	0,2821	0,2913
0,3	0,3004	0,3095	0,3185	0,3275	0,3364	0,3452	0,3540	0,3627	0,3714	0,3800
0,4	0,3885	0,3969	0,4053	0,4136	0,4219	0,4301	0,4382	0,4462	0,4542	0,4621
0,5	0,4699	0,4777	0,4854	0,4930	0,5005	0,5080	0,5154	0,5227	0,5299	0,5370
0,6	0,5441	0,5511	0,5580	0,5649	0,5717	0,5784	0,5850	0,5915	0,5980	0,6044
0,7	0,6107	0,6169	0,6231	0,6291	0,6351	0,6411	0,6469	0,6527	0,6584	0,6640
0,8	0,6696	0,6751	0,6805	0,6858	0,6911	0,6963	0,7014	0,7064	0,7114	0,7163
0,9	0,7211	0,7259	0,7306	0,7352	0,7398	0,7443	0,7487	0,7531	0,7574	0,7616
1,0	0,7658	0,7699	0,7739	0,7779	0,7818	0,7857	0,7895	0,7932	0,7969	0,8005
1,1	0,8041	0,8076	0,8110	0,8144	0,8178	0,8210	0,8243	0,8275	0,8306	0,8337
1,2	0,8367	0,8397	0,8426	0,8455	0,8483	0,8511	0,8538	0,8565	0,8591	0,8617
1,3	0,8643	0,8668	0,8692	0,8717	0,8741	0,8764	0,8787	0,8810	0,8832	0,8854
1,4	0,8875	0,8896	0,8917	0,8937	0,8957	0,8977	0,8996	0,9015	0,9033	0,9051
1,5	0,9069	0,9087	0,9104	0,9121	0,9138	0,9154	0,9170	0,9186	0,9201	0,9217
1,6	0,9232	0,9246	0,9261	0,9275	0,9289	0,9302	0,9316	0,9329	0,9341	0,9354
1,7	0,9366	0,9379	0,9391	0,9402	0,9414	0,9425	0,9436	0,9447	0,9458	0,94681
1,8	0,94783	0,94884	0,94983	0,95080	0,95175	0,95268	0,95359	0,95449	0,95537	0,95624
1,9	0,95709	0,95792	0,95873	0,95953	0,96032	0,96109	0,96185	0,96259	0,96331	0,96403
2,0	0,96473	0,96541	0,96609	0,96675	0,96739	0,96803	0,96865	0,96926	0,96986	0,97045
2,1	0,97103	0,97159	0,97215	0,97269	0,97323	0,97375	0,97426	0,97477	0,97526	0,97574
2,2	0,97622	0,97668	0,97714	0,97759	0,97803	0,97846	0,97888	0,97929	0,97970	0,98010
2,3	0,98049	0,98087	0,98124	0,98161	0,98197	0,98233	0,98267	0,98301	0,98335	0,98367
2,4	0,98399	0,98431	0,98462	0,98492	0,98522	0,98551	0,98579	0,98607	0,98635	0,98661
2,5	0,98688	0,98714	0,98739	0,98764	0,98788	0,98812	0,98835	0,98858	0,98881	0,98903
2,6	0,98924	0,98945	0,98966	0,98987	0,99007	0,99026	0,99045	0,99064	0,99083	0,99101
2,7	0,99118	0,99136	0,99153	0,99170	0,99186	0,99202	0,99218	0,99233	0,99248	0,99263
2,8	0,99278	0,99292	0,99306	0,99320	0,99333	0,99346	0,99359	0,99372	0,99384	0,99396
2,9	0,99408	0,99420	0,99431	0,99443	0,99454	0,99464	0,99475	0,99485	0,99495	0,99505

Для получения большей точности и для более высоких значений z применяются формулы: $r = (e^{2z} - 1) : (e^{2z} + 1)$; $z = \frac{1}{2}[(\log_e (1 + r) - \log_e (1 - r)]$.

ГЛАВА СЕДЬМАЯ

ВНУТРИКЛАССОВАЯ КОРРЕЛЯЦИЯ И ДИСПЕРСИОННЫЙ АНАЛИЗ

38. Некоторые довольно часто встречающиеся статистические данные, с одной стороны, могут быть обработаны методами, близкими к методу корреляции, и, с другой стороны, их более удобно обрабатывать по методу дисперсионного анализа, который позволяет отделить дисперсию, обусловленную одной группой причин, от дисперсии, приписываемой другой группе причин. В этой главе мы сначала рассмотрим применяемые в биометрике методы, аналогичные корреляционному методу, изложенному в предыдущей главе, и уже после этого перейдем к более общим случаям, встречающимся по преимуществу при обработке экспериментальных данных, где применение корреляционного метода представляется не вполне уместным, в то время как применение дисперсионного анализа дает полное освещение изучаемых нами вопросов. Сравнение этих двух методов обработки является прекрасной иллюстрацией того часто упускаемого из виду обстоятельства, что критерии существенности, если они правильно обоснованы, находятся между собой в полном согласии, независимо от того, каким путем они выведены.

Если, положим, имеются результаты измерения некоторого признака у n' пар братьев, то корреляцию между ними можно определить двумя несколько отличающимися друг от друга способами. Во-первых, можно разделить этих братьев на два класса, например на старших и младших братьев, и найти корреляцию между этими классами, подобно тому как это делалось ранее по отношению к родителям и их детям. Приняв этот путь, следует найти среднее значение признака у старших братьев и отдельно такое же среднее значение у младших братьев. Точно так же следует определить для этих двух классов и самостоятельные средние квадратические отклонения от соответствующих средних. Определенная этим путем корреляция, являющаяся корреляцией между двумя классами объектов, называется *междуклассовой*. Этот путь будет обязательным, если коррелируемыми признаками являются, предположим, возрасты на определенную дату или

какие-либо иные показатели, тесно связанные с возрастом. Но, с другой стороны, мы в каждом отдельном случае можем и не знать, какое значение признака принадлежит старшему и какое значение этого признака принадлежит младшему брату, или, положим, такая классификация братьев совсем не представляет интереса в свете стоящих перед нами задач. В таких случаях представляется более правомерным взять общую среднюю, полученную сплошь по всем наблюдениям, а также и общее среднее квадратическое отклонение от этой средней. Если x_1, x_1'; x_2, x_2' ... $x_{n'}$, $x_{n'}'$ являются попарными значениями изучаемого признака, то можно определить:

$$\bar{x} = \frac{1}{2n'} S(x + x')$$

$$s^2 = \frac{1}{2n} [S(x - \bar{x})^2 + S(x' - \bar{x})^2]$$

$$r = \frac{1}{ns^2} S[(x - \bar{x})(x' - \bar{x})]$$

Корреляция, вычисленная этим способом, в отличие от прежней называется *внутриклассовой* корреляцией, так как здесь все братья считаются принадлежащими к одному и тому же классу, имеют одну и ту же среднюю и одно и то же среднее квадратическое отклонение. Внутриклассовая корреляция, применяемая без учета таких различий, как возрастные различия у братьев, вообще говоря, дает более точную оценку фактического значения корреляции, чем та, которую дает междуклассовая корреляция, исчисленная по тому же самому материалу, так как в первом случае оценки средней и среднего квадратического отклонения основываются на $2n'$ данных вместо n'. Таким образом, коэффициент внутриклассовой корреляции не является оценкой, эквивалентной коэффициенту междуклассовой корреляции; напротив, он является несколько более точной оценкой. Как будет видно из дальнейшего изложения, распределение ошибок указывает и на другую сторону различий, которая также требует самостоятельного рассмотрения вопроса о внутриклассовой корреляции.

Вычисление внутриклассовой корреляции производится в так называемой симметричной корреляционной таблице, которая в известном отношении отличается от обычной корреляционной таблицы. Вместо внесения в корреляционную таблицу каждой пары наблюдений один раз, теперь это наблюдение при перестановке данных вносится дважды, например (x_1, x_1') и (x_1', x_1). Общее число внесенных таким образом данных составит теперь $2n'$, а итоговые распределения по столбцам и строкам таблицы будут в этом случае одинаковыми, т. е. будут представлять одно и то же распределение $2n'$ наблюдений. Остальная техническая работа по вычислению внутриклассовой корреляции во всем по-

добна работе по вычислению междуклассовой корреляции. Из способа вычисления симметричной таблицы вытекает, что хотя внутриклассовая корреляция и является более точной, однако она в этом отношении не может конкурировать с междуклассовой корреляцией, вычисленной по $2n'$ наблюдениям.

Различие между этими двумя корреляциями становится еще более заметным, когда обрабатываются не наблюдения, составляющие пары, а ряды с тремя или большим числом наблюдений, например при изучении какого-либо признака у трех братьев каждой семьи. В таких случаях также может быть построена симметричная корреляционная таблица. Каждые три брата дают возможность составить три пары, каждая из которых вносится в корреляционную таблицу дважды, так что любым трем братьям будет соответствовать 6 внесений в таблицу. Вычисление коэффициента корреляции в такой таблице будет производиться по формулам:

$$\bar{x} = \frac{1}{3n'} S(x + x' + x'')$$

$$s^2 = \frac{1}{3n} S[(x - \bar{x})^2 + (x' - \bar{x})^2 + (x'' - \bar{x})^2]$$

$$r = \frac{1}{3ns^2} S[(x - \bar{x})(x' - \bar{x}) + (x - \bar{x})(x'' - \bar{x}) + (x' - \bar{x})(x'' - \bar{x})]$$

В ряде случаев внутриклассовая корреляция определяется по большому числу наблюдений в каждой «семье», т. е. группе. Такое положение встречается, например, при изучении сходства между листьями одного и того же дерева, для чего с каждого из нескольких деревьев берется, положим, по 26 листьев, или, другой пример, когда с каждого из нескольких растений берется по 100 бобов для изучения их сходства. Если k является числом наблюдений в каждой «семье», т. е. классе, то это дает $k(k-1)$ попарных внесений в симметричную корреляционную таблицу. Это может привести к громоздкой корреляционной таблице, что, в свою очередь, затруднит всю вычислительную работу. Для устранения этого затруднения Гаррис ввел упрощенный метод расчета, который позволяет определить коэффициент внутриклассовой корреляции непосредственно, без составления симметричной таблицы. Для этого определяются два распределения: 1) распределение всей массы kn' наблюдений, на основе которого вычисляется $\bar{x}$ и s; 2) распределение n' средних по классам. Если $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_{n'}$ представляют собой такие средние, то выражение:

$$kS(\bar{x}_p - \bar{x})^2 = ns^2[1 + (k-1)r]$$

является уравнением, по которому может быть определено r, т. е. то значение внутриклассового коэффициента корреляции, которое было бы получено из симметричной корреляционной таблицы.

читателю предлагается проверить это, основываясь на формулах, приведенных выше для случая $k=3$.

Установленное выше соотношение позволяет выявить один замечательный факт: сумма квадратов и, следовательно, вся левая часть приведенного выше уравнения является всегда положительной. Поэтому r не может иметь отрицательного значения, меньшего, чем $-1/(k-1)$. Для положительного же значения r такого предела не существует и здесь возможны все значения, вплоть до $+1$. Более того, если число членов класса k по своему смыслу не ограничено каким-либо определенным значением, то корреляция r в генеральной совокупности вообще не может быть отрицательной. Например, в карточной игре, где число мастей ограничено четырьмя, корреляция между числом карт с различной мастью, имеющихся на одних руках, может иметь отрицательное значение вплоть до $-1/3$. Но этого не будет в генеральной совокупности листьев или детей, о которых говорилось ранее, так как число и тех и других в классе или семье не является обязательно меньше некоторого фиксированного числа. При отсутствии такого обязательного ограничения нельзя ожидать отрицательной внутриклассовой корреляции. В этом заключается явное и весьма существенное различие между внутриклассовой и междуклассовой корреляциями. Очевидно, что вследствие этого распределение внутриклассовой корреляции в случайных выборках также будет иным, чем у междуклассовой корреляции, так как в этих двух случаях и сами пределы варьирования различны.

39. Выборочные ошибки внутриклассовой корреляции

В случае, когда $k=2$, т. е. в случае, наиболее близком к междуклассовой корреляции, можно провести преобразование, подобное тому, которое было применено к этой последней, а именно:

$$z = \frac{1}{2}[\log(1+r) - \log(1-r)].$$

Здесь z также имеет распределение, очень близкое к нормальному; это распределение не зависит от значения корреляции ρ в генеральной совокупности, из которой взята данная выборка. Дисперсия z зависит только от размера выборки и определяется формулой

$$\sigma_z^2 = \frac{1}{n' - 3/2}.$$

Это преобразование обладает известным преимуществом по сравнению с таким же преобразованием для междуклассовой корреляции, так как внутриклассовая корреляция имеет более высокую точность по сравнению с междуклассовой корреляцией, основанной на том же количестве парных наблюдений. Эта более высокая точность связана с заменой делителя $n' - 3$ на $n' - 3/2$, что

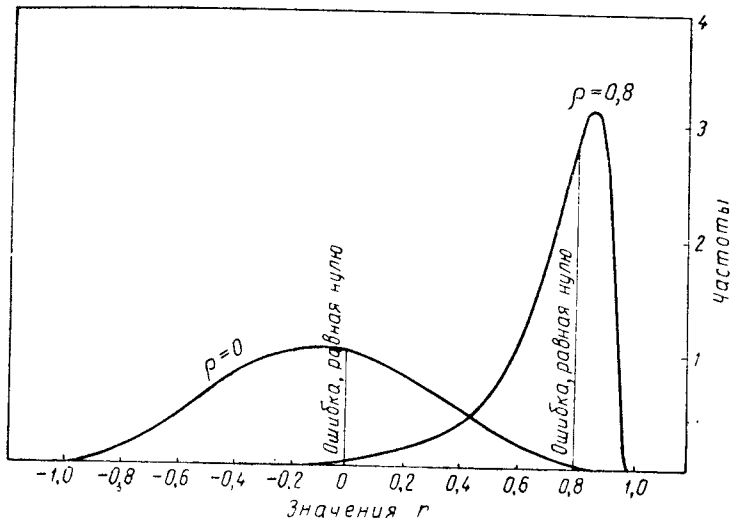


Рис. 9.

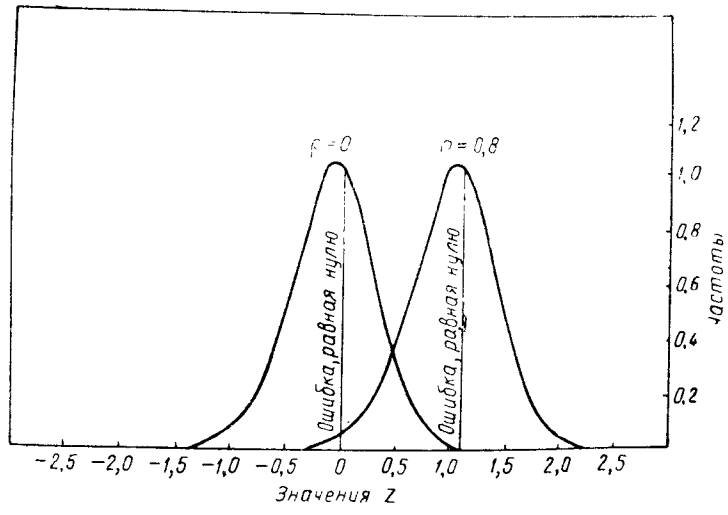


Рис. 10.

равноценно $1^1/_2$ добавочным наблюдениям. Второе различие этих двух видов корреляции касается характера смещения у соответствующих распределений z. При междуклассовой корреляции значение z, полученное из выборки, будет ли оно положительным или отрицательным, всегда является преувеличенным и требует поправки

$$-\frac{\rho}{2(n'-1)}.$$

В случае же внутриклассовой корреляции аналогичное смещение будет всегда отрицательным и не зависящим от ρ. Поэтому в данном случае поправка равна $+\frac{1}{2}\log\frac{n'}{n'-1}$, или, приближенно, $+\frac{1}{2n'-1}$. Это смещение характерно для внутриклассовой корреляции при любом значении k и обусловлено тем, что симметричная корреляционная таблица не способна дать наилучшую оценку корреляционной связи.

Из сравнения рис. 9 и 10 можно установить то влияние, которое оказывает это преобразование на кривую ошибок. На рис. 9 даны фактические кривые ошибок коэффициента корреляции r, когда r вычисляется по симметричной таблице, образованной на основе 8 парных наблюдений, при условии, что корреляция в генеральной совокупности равна 0 и 0,8. На рис. 10 изображены соответствующие кривые ошибок для величины z. Те три главные преимущества распределения z, о которых говорилось при сравнении рис. 7 и 8, остаются теми же самыми и при сравнении рис. 9 и 10: кривые с весьма неодинаковой вариацией заменены на кривые с одинаковой дисперсией; асимметричные кривые заменены нормальными кривыми; кривые, форма которых резко различна, заменены на кривые, одинаковые по своей форме. Но в одном отношении результат этого преобразования при внутриклассовой корреляции более точен, чем при междуклассовой корреляции. Дело в том, что хотя в обоих случаях преобразованные кривые не являются строго нормальными, однако при внутриклассовой корреляции они полностью константны как в отношении дисперсии, так и формы; при междуклассовой же корреляции эти кривые несколько меняют и свою дисперсию и свою форму по мере того, как изменяется величина корреляции ρ в генеральной совокупности. На рис. 10 хорошо показано то смещение, которое присуще распределению коэффициента корреляции, определяемого по симметричной корреляционной таблице; это смещение вместе с другими характерными особенностями этих кривых остается абсолютно константным на всей шкале величины z.

Пример 34. *Точность внутриклассового коэффициента корреляции.* По 13 парным наблюдениям был определен коэффициент внутриклассовой корреляции, оказавшийся равным 0,6000. Следует произвести оценку корреляции в той генеральной сово-

купности, из которой взята выборка, а также найти пределы, в которых лежит эта корреляция при заданной вероятности.

Располагая величины r и z в параллельных столбцах, находим:

Таблица 36

	r	z
Вычисленное значение	+0,6000	+0,6931
Поправка	—	0,0400
Оценка	+0,6250	+0,7331
Средняя квадратическая ошибка	—	±0,2949
Верхний предел	+0,8675	+1,3229
Нижний предел	+0,1423	+0,1433

Сначала все вычисления проводятся в столбце z и уже после этого по таблице VB находятся соответствующие значения r. Исходное значение r получено из симметричной корреляционной таблицы, а соответствующее значение z определено по указанной таблице. Эти значения имеют небольшое отрицательное смещение, которое исключается путем введения в z надлежащей поправки. Таким образом, несмещенная оценка z равна 0,7331, а соответствующее несмещенное значение r равно 0,6250. Чтобы определить пределы ожидаемой точной корреляции, вычисляется средняя квадратическая ошибка z, после чего ее удвоенное значение прибавляется к полученной ранее оценке и вычитается из нее. В результате получаются верхний и нижний пределы z. По этим последним определяются соответствующие значения r. Фактическое значение коэффициента корреляции в данном случае может считаться существенным, так как нижний предел здесь положителен. Опасность впасть в ошибку очень мала, если сделать вывод, что точный коэффициент корреляции не меньше 0,14 и не больше 0,87.

Когда k больше 2, то выборочные ошибки лучше всего определить при помощи дисперсионного анализа, но в то же время, если есть необходимость сделать это в терминах теории корреляции, можно произвести некоторое преобразование, имеющее силу при любом значении k. Это преобразование производится по формуле:

$$z = \frac{1}{2}\log\frac{1+(k-1)r}{1-r}.$$

В частном случае при $k=2$ оно приводится к тому, которым мы уже пользовались. Теперь, когда рассматриваются выборки групп, содержащих k наблюдений, распределение ошибок z также не зависит от точного значения ρ и приближается к нормальному по мере увеличения n', хотя и не с такой скоростью, как это было

при $k=2$. Дисперсия z, если n' достаточно велико, будет приблизительно равна

$$\frac{k}{2(k-1)(n'-2)}.$$

При определении r по данному значению z в этом случае может быть использована та же таблица VB.

Пример 35. *Применение таблицы VB в более общем случае.* Требуется определить значение r, соответствующее $z=+1{,}0605$ при $k=100$.

Сначала вычтем из данного значения z половину натурального логарифма от $(k-1)$, найдем по этой разности взятой в качестве z, по таблице VB значение r и умножим его на k, затем прибавим $(k-2)$ и разделим на $2(k-1)$. Ниже приводятся соответствующие числовые расчеты.

Таблица 37

z	+1,0605
$\frac{1}{2}\log(k-1) = \frac{1}{2}\lg 99$	2,2975
„z“	−1,2370
„r“	−0,8446
k „r“ $= 100$ „r“	−84,46
$k-2$	98
$2r(k-1) = 198r$	13,54
r	+0,0684

Пример 36. *Существенность внутриклассовой корреляции при большой выборке.* Корреляция между «пустыми гнездами» в различных бобах, взятых с одного дерева Cercis Canadensis оказалась равной +0,0684, причем с каждого из 60 деревьев было взято по 100 бобов (данные Гарриса). Будет ли эта корреляция существенна?

Согласно результатам предыдущего примера $z=1{,}0605$, а средняя квадратическая ошибка z равна 0,0933. Значение z превосходит свою среднюю квадратическую ошибку в 11 раз; следовательно, корреляция, несомненно, существенна.

Когда n' достаточно велико и если r не очень близко к +1, то, как мы видели, можно считать, что междуклассовая корреляция в случайных выборках распределена нормально со средней квадратической ошибкой

$$\frac{1-\rho^2}{\sqrt{n'-1}}.$$

Аналогичная формула для внутриклассовой корреляции при k наблюдениях внутри класса будет:

$$\frac{(1-\rho)[1+(k-1)\rho]}{\sqrt{\frac{1}{2}k(k-1)n'}}.$$

Практическое применение этой формулы ограничено даже еще более жесткими рамками, чем формулы для междуклассовой корреляции, так как n' чаще всего величина небольшая. Вместе с этим область, в которой данная формула не применима даже при больших n', теперь находится не в окрестностях ± 1, а в окрестностях $+1$ и $-\frac{1}{k-1}$. Когда k велико, последняя величина близка к нулю, вследствие чего имеет место резкая асимметрия распределения не только при высокой, но также и при очень низкой корреляции. Поэтому в последнем случае для непосредственной оценки существенности коэффициента корреляции нет никакой достаточно точной формулы. Эта ненормальность распределения r все же является лишь частностью, так как она возникает только в ближайшей окрестности нуля и ей здесь противостоит известный выигрыш в точности при больших значениях k. Вблизи нуля, как показывает приведенная выше формула, точность внутриклассовой корреляции при большой выборке эквивалентна точности корреляции при $\frac{1}{2}k\ (k-1)n'$ независимых парных наблюдениях. Это при большом k дает громадный выигрыш в точности. При корреляции около 0,5, сколь бы ни было велико k, точность не будет выше той, которая получится при $9n'/2$ парных наблюдениях. Вблизи же $+1$ она не может быть большей, чем при n' парных наблюдениях.

40. Внутриклассовая корреляция как один из случаев дисперсионного анализа

Вопрос о внутриклассовой корреляции значительно упрощается, если воспользоваться тем обстоятельством, что в этом случае корреляция просто-напросто измеряет относительную роль двух групп факторов, обусловливающих вариацию данных. Мы видели, что в ходе вычисления внутриклассовой корреляции необходимые для этого величины kns^2 и $ns^2\{1+(k-1)r\}$ соответственно приравниваются к

$$\sum_{1}^{kn'}(x-\bar{x})^2 \quad \text{и} \quad k\sum_{1}^{n'}(\bar{x}_p-\bar{x})^2.$$

Первая из этих величин является суммой kn' квадратов отклонений всех наблюдений от их общей средней, а вторая является умноженной на k суммой квадратов n' отклонений средней каждого класса от общей средней. Вместе с тем легко видеть, что

$$\sum_{1}^{kn'}(x-\bar{x})^2 = k\sum_{1}^{n'}(\bar{x}_p-\bar{x})^2+\sum_{1}^{kn'}(x-\bar{x}_p)^2.$$

Здесь последний член является суммой квадратов отклонений каждого отдельного наблюдения от средней того класса, к которому оно принадлежит. Приводимая ниже таблица систематизирует эти соотношения и дает числа степеней свободы, соответствующие каждой из сумм квадратов. В последнем столбце таблицы дана интерпретация каждой суммы квадратов в терминах внутриклассовой корреляции, определяемой по симметричной корреляционной таблице.

Таблица 38

	Степени свободы	Сумма квадратов	
Внутри классов . .	$n'(k-1)$	$\sum_{1}^{kn'}(x-\bar{x}_p)^2$	$ns^2(k-1)(1-r)$
Между классами .	$n'-1$	$k\sum_{1}^{n'}(\bar{x}_p-\bar{x})^2$	$ns^2[1+(k-1)r]$
Итого . . .	$n'k-1$	$\sum_{1}^{kn'}(x-\bar{x})^2$	ns^2k

Можно видеть, что величина z предыдущего параграфа, если оставить в стороне некоторую константу, является половиной разности логарифмов тех двух частей, на которые разлагается общая сумма квадратов. Тот факт, что форма распределения z не зависит от величины корреляции в генеральной совокупности, является следствием того, что отклонения отдельных наблюдений от средней соответствующего класса *не зависят* от отклонений таких средних от общей средней. В результате получаются независимые оценки двух дисперсий. Если эти дисперсии равны, то корреляция равна нулю; если эти оценки не отличаются существенно друг от друга, то и корреляция будет несущественной. Когда же эти дисперсии существенно различаются, то можно, если в этом возникает потребность, этот факт выразить в терминах корреляции.

Интерпретация такого неравенства дисперсий в терминах корреляции может быть лучше понятой, если воспользоваться приводимыми ниже построениями, которые вместе с этим убеждают в том, что при расчете корреляции по симметричной корреляционной таблице возникает некоторая неточность. Допустим, что некоторая величина состоит из двух частей, распределение каждой из которых нормально и независимо; пусть дисперсия первой части A, а второй B; ясно, что дисперсия всей изучаемой величины в целом $A+B$. Возьмем выборку n' значений первой части и к каждому из этих значений добавим выборку из k значений второй части. В этом случае мы будем иметь n' классов с k наблюдениями в каждом классе. В бесконечной генеральной совокупности, из которой эти выборки взяты, корреляция между попарными членами одного и того же класса будет

$$\rho = \frac{A}{A+B}.$$

На основе ряда из kn' наблюдений мы можем построить оценки для величин A и B или, другими словами, мы можем разложить общую дисперсию на части, относящиеся к двум группам причин варьирования. Внутриклассовая корреляция является просто долей в общей дисперсии, которая обусловлена причинами вариации внутри классов. Величина B может быть оценена непосредственно, так как вариация внутри каждого класса определяется только одними этими причинами, и, следовательно:

$$\sum_{1}^{kn'} (x - \bar{x}_p)^2 = n'(k-1)B.$$

Средняя из наблюдений некоторого класса подразделяется на две части, первая из которых характеризуется дисперсией A и вторая, являющаяся средней из k значений второй части отдельного наблюдения, характеризуется поэтому дисперсией B/k. Следовательно, вариация средних по классам будет

$$k\sum_{1}^{n'} (\bar{x}_p - \bar{x})^2 = (n'-1)(kA+B).$$

В связи с этим следует изменить табл. 38, в последнем столбце s^2 выражая через A и B, и считая r уже несмещенной оценкой корреляции.

Таблица 39

	Степени свободы	Сумма квадратов	
Внутри классов	$n'(k-1)$	$\sum_{1}^{kn'} (x-\bar{x}_p)^2$	$n'(k-1)B = n's^2(k-1)(1-r)$
Между классами	$n'-1$	$k\sum_{1}^{n'}(\bar{x}_p-\bar{x})^2$	$(n'-1)(kA+B)=(n'-1)s^2[1+(k-1)r]$
Итого . .	$n'k-1$	$\sum_{1}^{kn'}(x-\bar{x})^2$	$(n'-1)kA+(n'k-1)B=s^2[n'k-1-(k-1)r]$

Сравнивая последний столбец этой таблицы с соответствующим столбцом табл. 38, можно видеть, что различие возникает только вследствие замены n на n' в первой строке и n на $n'-1$ во второй. В связи с этим отношение между суммами квадратов меняется, как n': $(n'-1)$, в результате чего элиминируется отрицательное смещение наблюденного значения z, имеющее место при предыдущем методе. Ошибка этого последнего метода состоит в допущении, что общая дисперсия, полученная для n' рядов связанных наблюдений, оценивается вполне точно путем приравнивания суммы квадратов отклонений всех наблюдений от их средней к величине ns^2k, как будто все они независимы. Эта ошибка не имеет большого значения, когда n' велико, как это

обычно бывает при $k=2$, но при большом значении k, когда число наблюдений может быть для практики слишком большим даже при малом n', неисправленные значения z могут привести к серьезным ошибкам.

Оценка существенности внутриклассовой корреляции в такой таблице дисперсионного анализа может быть произведена и без фактического вычисления r. Если корреляция отсутствует, то A не будет существенно отличаться от нуля; следовательно, между классами в этом случае не будет никаких различий, кроме тех, которые обусловлены влиянием случайного отбора внутри классов. Здесь все наблюдения, взятые в целом, составляют однородную группу с дисперсией B.

41. Критерий существенности для разности дисперсий

Критерий существенности для внутриклассовой корреляции является одним из членов более широкой группы критериев существенности, применяемых в дисперсионном анализе. Все эти критерии относятся к одной проблеме, заключающейся в установлении того, в какой мере одна оценка дисперсии, основанная на n_1 степенях свободы, существенно больше второй оценки, основанной на n_2 степенях свободы. Эта проблема в ее простейшей форме приводит к определению величины z, которая равна половине разности между натуральными логарифмами этих оценок дисперсии или разности натуральных логарифмов соответствующих средних квадратических отклонений. Если обозначить через P вероятность случайного появления значения больше данного z, то на основе предыдущего предоставляется возможность определить значения z, соответствующие различным значениям P, n_1 и n_2.

Построение полной таблицы такого рода, включающей в себя три переменных величины, было бы очень громоздким. Поэтому здесь дается таблица только для трех особо важных значений P и для некоторого числа комбинаций n_1 и n_2, достаточных для определения других, опущенных здесь сочетаний этих величин (таблица VI, стр. 198—200). В дальнейшем будут даны самые разнообразные примеры применения этой таблицы. Когда n_1 и n_2 велики и даже при умеренных их значениях, если они равны или примерно равны между собой, распределение z довольно близко к нормальному, что позволяет оценку этой величины производить на основе ее среднего квадратического отклонения, которое определяется по формуле

$$\sqrt{\frac{1}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

В частности, сюда относится и случай внутриклассовой корреляции при $k=2$, так как, если имеется n' парных наблюдений, то дисперсия между классами будет основываться на $n'-1$ сте-

пенях свободы, а дисперсия внутри классов — на n' степенях свободы, и, следовательно,

$$n_1 = n' - 1 \text{ и } n_2 = n'.$$

Уже при относительно небольших значениях n', как было установлено ранее, можно считать, что величина z распределена нормально. Когда же k больше 2, то мы имеем

$$n_1 = n' - 1 \quad \text{и} \quad n_2 = (k-1)\,n'.$$

Эти числа, если только n' не бесконечно большое число, значительно отличаются друг от друга. Поэтому в данном случае распределение z будет заметно асимметричным и его среднее квадратическое отклонение не может быть применено для построения достаточно точного критерия существенности.

Пример 37. *Различие в вариации роста у лиц разного пола.* По 1164 измерениям роста мужчин была определена сумма квадратов отклонений и получено число 8590; по 1456 измерениям роста женщин получена сумма квадратов 9870. Можно ли считать существенным различие в вариации роста у мужчин и женщин?

Таблица 40

	Степени свободы	Сумма квадратов	Средний квадрат	log среднего квадрата	$\frac{1}{n}$
Мужчины .	1163	8 590	7,386	1,9996	0,0008598
Женщины .	1455	9 870	6,783	1,9145	0,0006873
			Разность	0,0851	Сумма = 0,0015471

Средние квадраты получены делением соответствующих сумм квадратов на числа степеней свободы; разность логарифмов равна 0,0851 и, следовательно, $z = 0{,}0426$. Дисперсия z равна половине суммы последнего столбца, так что среднее квадратическое отклонение z равно 0,02781. Таким образом, хотя различие в вариации кажется достаточно большим, однако при этих данных оно не может считаться существенным.

Пример 38. *Однородность малых выборок.* При исследовании точности подсчета почвенных бактерий образец почвы делился на четыре части. После разбавления каждой из них было приготовлено 7 пластинок для микроскопического подсчета. Ниже приводятся количества колоний на каждой пластинке. Можно ли считать случайными различия в результатах подсчета у этих четырех образцов? Другими словами, следует установить, является ли этот ряд из 28 наблюдений однородным или здесь имеется более или менее значительная внутриклассовая корреляция.

Таблица 41

	Образцы			
	I	II	III	IV
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	62
6	53	58	56	58
7	51	52	56	54
Сумма . .	426	461	450	440
Средняя .	60,86	65,86	64,28	62,86

По этим данным составляем таблицу дисперсионного анализа:

Таблица 42

	Степени свободы	Сумма квадратов	Средний квадрат	Среднее квадратическое отклонение	Логарифм среднего квадратического отклонения
Внутри классов .	24	1446	60,25	7,762	2,0493
Между классами .	3	94,96	31,65	5,626	1,7274
Итого . . .	27	1540,96	57,07	7,55	−0,3219
					Разность = z

Вариация внутри классов оказалась фактически больше вариации между классами и, следовательно, если и существует какая-либо корреляция, то она может быть только отрицательной. Так как числа степеней свободы небольшие и не равные друг другу, то для оценки различий необходимо использовать таблицу VI. Эта таблица составлена так, что за n_1 следует брать число степеней свободы большей дисперсии. В нашем случае $n_1 = 24$, а $n_2 = 3$. Эта таблица для 5%-ного уровня существенности дает 1,0781, откуда следует, что фактическая разность 0,3219 невелика и несущественна. В целом этот ряд из 28 наблюдений должен считаться однородным; его дисперсия равна 57,07.

Следует заметить, что если бы было только два образца, то данный критерий существенности был бы эквивалентен критерию t, с которым мы познакомились в главе V. Действительно, при $n_1 = 1$ табличные значения z (стр. 198) представляют собой ничто иное, как логарифмы значений t для $P = 0{,}05$ и 0,01

(стр. 144). Точно так же табличные значения z для $n_2=1$ в таблице VI являются логарифмами обратных величин, которые в таблице IV составляют графы $P=0{,}95$ и $P=0{,}99$. Описанный здесь метод может считаться обобщением метода главы V при сравнении нескольких средних; он может рассматриваться и как обобщение методов главы IV: если n_2 будет бесконечным, то z будет равняться $\frac{1}{2}\log(\chi^2/n)$ из таблицы III для $P=0{,}05$ и $0{,}01$; если же n_1 будет бесконечным, то z будет равен $-\frac{1}{2}\log(\chi^2/n)$ для $P=0{,}95$ и $0{,}99$. Следовательно, критерии согласия, выборочная дисперсия которых не дана априори, а должна оцениваться на основе наблюдений, могут быть определены уже не по таблице III, а по таблице VI (см. главу VIII).

Пример 39. *Сравнение внутриклассовых корреляций.* В различных бобах для числа гнезд одного и того же дерева (Cercis Canadensis) были установлены следующие корреляции (данные Гарриса), причем с каждого дерева было взято по 100 бобов:

Мерамек, Шотландия 60 деревьев $+0{,}3527$

Лауренс, Канзас 22 дерева $+0{,}3999$

Можно ли считать, что корреляция в Лауренсе существенно выше, чем в Мерамеке?

Сначала находим z для каждого из этих случаев по формуле

$$z=\frac{1}{2}[\log(1+99r)-\log(1-r)]$$

(см. стр. 180), в результате получаем $z=2{,}0081$ для Мерамека и $2{,}1071$ для Лауренса. Так как эти результаты получены из симметричной корреляционной таблицы, то следует ввести в них небольшие поправки $1/(2n'-1)$, после чего получим $2{,}0165$ для Мерамека и $2{,}1304$ для Лауренса, т. е. те значения, которые были бы получены по методу дисперсионного анализа.

Определение ошибок этих показателей начнем с корреляции в Лауренсе, которая вычислена только по 22 наблюдениям и поэтому имеет бо́льшую ошибку. Здесь мы имеем $n_1=21$ и $n_2=22\times99=2178$. Таких значений в таблице VI нет, но при $n_1=24$ и $n_2=\infty$ положительные ошибки, превосходящие $0{,}2085$, как видно из этой же таблицы, встречаются чаще, чем в 5% выборок. Уже один этот расчет дает отрицательный ответ на вопрос о существенности различия, так как значение z для Лауренса превосходит значение z для Мерамека только на $0{,}1139$.

В тех случаях, когда необходима большая точность, можно воспользоваться тем обстоятельством, что в таблице z пять значений 6, 8, 12, 24 и ∞ выбраны так, чтобы получалась гармоническая прогрессия. Это значительно облегчает интерполяцию, если взять в качестве переменной величины $\frac{1}{n}$. Если приходится интерполировать в обоих направлениях, т. е. n_1 и n_2, то эта операция производится в три этапа. В нашем случае мы сначала найдем

значения z для $n_1=12$ и $n_2=2178$, далее для $n_1=24$ и $n_2=2178$ и, наконец, перейдем к заданным $n_1=21$ и $n_2=2178$.

Чтобы определить значение z для $n_1=12$ и $n_2=2178$, сначала вычисляем

$$\frac{60}{2178}=0{,}0275;$$

далее находим по таблице для $n_2=\infty$ значение $z=0{,}2804$, но для $n_2=60$ значение z больше предыдущего на $0{,}0450$; отсюда находим $0{,}2804+0{,}275\times0{,}0450=0{,}2816$, что и является приближенным значением z для $n_2=2178$.

Подобно этому, для $n_1=24$ находим:

$$0{,}2085+0{,}0275\times0{,}0569=0{,}2101.$$

По этим двум значениям z следует определить z для $n_1=21$. Сначала находим

$$\frac{24}{21}=1+\frac{1}{7};$$

следовательно, мы должны прибавить к значению z для $n_1=24$ одну седьмую разности между этим значением и значением z при $n_1=12$; это дает

$$0{,}2101+\frac{0{,}0715}{7}=0{,}2203.$$

Полученный результат и будет искомым положительным отклонением z, которое может быть превзойдено в 5% всех случайных выборок.

Тем же способом, которым было определено вычисленное выше положительное отклонение для 5%-ного уровня существенности, можно найти и отрицательное значение отклонения для того же 5%-ного уровня; для этого следует только переменить местами n_1 и n_2. В нашем случае такой расчет дает $z=0{,}2978$. Если мы допустим, что наблюденное значение z не выходит за эти 5%-ные границы в обоих направлениях, т. е. что оно лежит в центральной части распределения, составляющей девять десятых всего объема совокупности, то можно будет сказать, что истинное значение z для Лауренса лежит между $1{,}9101$ и $2{,}4282$. Эти доверительные пределы получены путем вычитания из наблюденного значения z положительного отклонения, в одном случае, и путем добавления к наблюденному z отрицательного отклонения, в другом случае.

Тот факт, что эти два отклонения — положительное и отрицательное — резко отличаются друг от друга, что всегда наблюдается, когда n_1 и n_2, не будучи большими числами, не равны друг другу, указывает на асимметричность распределения z и на невозможность произвести оценку z на основе средней квадратической ошибки этой величины.

Как мы видели, путем интерполяции можно определить z с достаточно высокой точностью, но надо заметить, что такая интерполяция не применима для того угла таблицы значений z,

где n_1 больше 24 и n_2 больше 30. В этих случаях можно воспользоваться специальной формулой, которая с точностью до одной сотой дает z для 5%-ного уровня существенности. Если обозначить через h среднюю гармоническую из n_1 и n_2, т. е. если

$$\frac{2}{h} = \frac{1}{n_1} + \frac{1}{n_2},$$

то можно написать

$$z = \frac{1{,}6449}{\sqrt{h-1}} - 0{,}7843\left(\frac{1}{n_1} - \frac{1}{n_2}\right).$$

Для 1%-ного уровня существенности имеется подобная приближенная формула

$$z = \frac{2{,}3263}{\sqrt{h-1{,}4}} - 1{,}235\left(\frac{1}{n_1} - \frac{1}{n_2}\right).$$

Наконец, для 0,1%-ного уровня можно взять формулу:

$$z = \frac{3{,}0902}{\sqrt{h-2{,}1}} - 1{,}925\left(\frac{1}{n_1} - \frac{1}{n_2}\right).$$

Замена корня $\sqrt{h-1}$, который входит в формулу 5%-ного уровня, на $\sqrt{h-1{,}4}$ и $\sqrt{h-2{,}1}$ при двух других уровнях предложена Кочрэном.

Применим этот способ для определения 5%-ного уровня z по данным Мерамека, где $n_1 = 59$ и $n_2 = 5940$.

Таблица 43

$1/n_1 = 0{,}01695$ $1/n_2 = 0{,}00017$ $2/h = 0{,}01712$ $1/h = 0{,}00856$ $h = 116{,}8$	$\sqrt{h-1} = 10{,}76$ $\frac{1}{\sqrt{h-1}} = 0{,}09294$ $\frac{1}{n_1} - \frac{1}{n_2} = 0{,}01678$	Первый член = 0,15288 Второй член = 0,01316 Разность = 0,1397 Сумма = 0,1660

Таким образом положительное отклонение для 5%-ного уровня будет равно 0,1397, а отрицательное 0,1660. Основываясь на этих результатах, можно считать, что для Мерамека значение z лежит между 1,8768 и 2,1825 при доверительной вероятности 0,90. То, что этот доверительный интервал большей своей частью накладывается на интервал для Лауренса, показывает, что корреляции, относящиеся к этим двум районам, не отличаются существенно друг от друга.

42. Дисперсионный анализ для случая, когда дисперсия разлагается не на две, а на большее число частей

В ряде случаев возникает необходимость в разложении общей дисперсии на число частей, большее двух. Как в опытах, так и при наблюдениях встречаются такие случаи, когда данные могут быть классифицированы в нескольких направлениях и когда каж-

Таблица 44

Часы	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь	Ноябрь	Декабрь	Итого
1	19	16	34	17	21	26	18	18	11	27	31	28	266
2	27	16	32	21	18	25	17	24	14	37	27	29	287
3	27	19	27	22	21	31	15	32	16	36	27	33	306
4	19	20	23	25	26	28	15	31	20	39	28	34	308
5	23	18	23	19	28	31	23	23	16	42	28	31	305
6	20	23	33	19	22	29	15	22	17	42	24	26	292
7	20	22	35	25	19	33	15	18	20	36	19	23	285
8	22	22	28	26	24	23	15	18	20	32	21	31	282
9	22	20	25	21	24	28	11	22	16	35	24	28	276
10	21	19	19	15	18	22	16	19	16	31	20	24	240
11	17	19	23	18	26	25	17	13	14	33	22	21	248
12	20	21	31	19	25	27	24	25	20	36	31	25	304
13	16	18	35	28	23	32	24	25	18	29	29	28	305
14	21	20	35	28	23	32	20	27	15	28	24	28	301
15	17	18	39	32	22	27	25	30	20	28	31	32	321
16	18	31	38	38	31	32	31	28	15	37	27	34	360
17	26	29	37	30	22	32	24	31	20	38	24	38	351
18	21	25	41	24	24	28	25	27	21	37	26	33	332
19	26	18	35	25	23	30	25	27	20	36	26	31	322
20	24	18	30	21	28	25	28	18	16	34	30	25	297
21	23	20	27	23	18	21	25	20	14	35	27	39	292
22	25	16	33	22	29	23	22	22	14	22	26	36	290
23	22	15	33	18	24	23	21	19	11	31	28	28	273
24	22	18	30	19	26	27	18	22	14	22	28	27	273
Итого	518	481	746	555	565	660	489	561	398	803	628	712	7116

дое наблюдение относится к одному из классов типа A и в то же время к одному из классов типа B и т. д. В таких случаях имеется возможность найти отдельно дисперсию между классами A, между классами B и т. д. Остаток общей дисперсии, после исключения из нее предыдущих дисперсий, представляет собой дисперсию внутри каждого субкласса или характеризует добавочный эффект взаимодействия между факторами варьирования, которое выражается в том, что изменения изучаемой величины по классам типа A не будут одинаковыми во всех классах типа B. Наличие или отсутствие такого взаимодействия может быть проверено, если в субклассах имеется достаточное число наблюдений, так как для этих субклассов можно непосредственно определить фактическую дисперсию и в то же время установить теоретически ожидаемую дисперсию; сравнение этих дисперсий и решает вопрос о взаимодействии.

Пример 40. *Изменчивость дневных и годовых осадков.* Наблюдения, проводившиеся в течение 10 лет в Ричмонде, дали следующие (табл. 44) данные о частоте выпадения осадков в различные часы, систематизированные по месяцам (данные Шоу с двумя исправлениями в итогах).

В этом случае анализ дисперсии будет таким:

Таблица 45

	Степени свободы	Сумма квадратов	Средний квадрат
Месяцы	11	6568,58	597,144
Часы	23	1539,33	66,928
Остаток	253	3819,58	15,097
Итого	287	11927,50	

Средняя из 288 данных, приведенных в табл. 44, равна 24,7. Если бы исходные данные характеризовали независимую выборочную изменчивость, то следовало бы ожидать, что остаточный средний квадрат будет примерно таким же или даже больше среднего квадрата, относящегося к распределению осадков в течение дня. Можно видеть, что остаточная дисперсия ниже этой нормы и что явной причиной этого служит то обстоятельство, что выпадение осадков в тот или иной час дня является фактом, зависящим от наличия осадков в предшествующий час того же дня. В связи с этим, отдельный случай выпадения осадков при данном учете может фиксироваться дважды и даже большее число раз, что приводит к положительной корреляции между данными, относящимися к смежным часам дня. Благодаря этому в изменчивость,

приписываемую месяцам года, входит большая доля случайной вариации и по всей вероятности эта последняя обусловливает ту нерегулярность, которая наблюдается в месячных итогах. Дисперсия же между почасовыми осадками значительно больше, чем остаточная дисперсия; это указывает на то, что дождливые часы в целом примерно одинаковы в различных месяцах и что имеется определенное влияние времени внутри дня на осадки. По этим данным не представляется возможным оценить влияние времен года, а также исследовать, является ли влияние времени внутри дня одним и тем же для всех месяцев.

Пример 41. *Дисперсионный анализ в полевом опыте.* В табл. 46 показаны урожаи картофеля в фунтах на делянку в полевом опыте Ротамстедской опытной станции. Участок, удобренный в целом навозом, был разделен на 36 делянок, на которых было размещено 12 сортов картофеля; три делянки каждого сорта были без всякой системы, т. е. в рэндомизированном порядке разбросаны по всей площади участка. Каждая делянка была разделена на три части, одна из которых не получила никакого добавочного удобрения и была только известкована, в то время как две другие получили соответственно сульфат калия и хлорид калия.

Эти данные могут дать самую различную информацию. Итоговые урожаи 36 делянок дают нам 35 степеней свободы, из которых 11 соответствуют различиям по урожайности 12 сортов и 24 представляют различия между делянками, относящимися к одному и тому же сорту картофеля. Сравнивая дисперсию этих двух видов, можно определить существенность межсортовых различий урожайности при данных почвенных и климатических условиях эксперимента. Дополнительные 72 степени свободы, связанные с урожайностями отдельных частей делянок, содержат 2 степени свободы, относящиеся к различиям между удобрениями, причем эти 2 степени свободы можно подразделить на одну, характеризующую различие между калийным удобрением и известкованием, и вторую, характеризующую различие между удобрением сульфатом калия и хлоридом калия; остальные же 70 степеней свободы в основном характеризуют различия, наблюдаемые в отзывчивости на удобрения различных делянок. Эти 70 степеней свободы в дальнейшем могут быть подразделены на 22 степени свободы, характеризующие различия в отзывчивости различных сортов на удобрения, и 48 степеней, характеризующие эти же различия на разных делянках одного и того же сорта.

Для установления существенности эффективности удобрений следует сравнить дисперсию двух степеней свободы, относящихся к удобрениям с остаточной дисперсией при 48 степенях свободы.

Для оценки различий в отзывчивости сортов на удобрение следует сравнить дисперсию с 22 степенями свободы с теми же 48 степенями свободы, а для оценки существенности различий

Таблица 46

Сорт	Сульфат калия			Хлорид калия			Известкование		
Ajax	3,20	4,00	3,86	2,55	3,04	4,13	2,82	1,75	4,71
Arran Comrade	2,25	2,56	2,58	1,96	2,15	2,10	2,42	2,17	2,17
British Queen	3,21	2,82	3,82	2,71	2,68	4,17	2,75	2,75	3,32
Duke of York	1,11	1,25	2,25	1,57	2,00	1,75	1,61	2,00	2,46
Epicure	2,36	1,64	2,29	2,11	1,93	2,64	1,43	2,25	2,79
Great Scot	3,38	3,07	3,89	2,79	3,54	4,14	3,07	3,25	3,50
Iron Duke	3,43	3,00	3,96	3,33	3,08	3,32	3,50	2,32	3,29
K. of K.	3,71	4,07	4,21	3,39	4,63	4,21	2,89	4,20	4,32
Kerr's Pink	3,04	3,57	3,82	2,96	3,18	4,32	2,00	3,00	3,88
Nithsdale	2,57	2,21	3,58	2,04	2,93	3,71	1,96	2,86	3,56
Tinwald Perfection	3,46	3,11	2,50	2,83	2,96	3,21	2,55	3,39	3,36
Up-to-Date	4,29	2,93	4,25	3,39	3,68	4,07	4,21	3,64	4,11

Таблица 47

Сорт	Удобрения			Итого	Деляики		
	сульфат	хлорид	известь		I	II	III
Ajax	11,06	9,72	9,28	30,06	8,57	8,79	12,70
Arran Comrade	7,39	6,21	6,76	20,36	6,63	6,88	6,85
British Queen	9,85	9,56	8,82	28,23	8,67	8,25	11,31
Duke of York	4,61	5,32	6,07	16,00	4,29	5,25	6,46
Epicure	6,29	6,68	6,47	19,44	5,90	5,82	7,72
Great Scot	10,34	10,47	9,82	30,63	9,24	9,86	11,53
Iron Duke	10,39	9,73	9,11	29,23	10,26	8,40	10,57
K. of K.	11,99	12,23	11,41	35,63	9,99	12,90	12,74
Kerr's Pink	10,43	10,46	8,88	29,77	8,00	9,75	12,02
Nithsdale	8,36	8,68	8,38	25,42	6,57	8,00	10,85
Tinwald Parfection	9,07	9,00	9,30	27,37	8,84	9,46	9,07
Up-to-Date	11,47	11,14	11,96	34,57	11,89	10,25	12,43
Итого	111,25	109,20	106,26	326,71	—	—	—

в урожайности того же самого сорта на различных делянках следует сравнить 24 степени свободы, относящиеся к сравнениям делянок, занятых одним и тем же сортом, с 48 степенями свободы, характеризующими отзывчивость на удобрения различных делянок, занятых одним и тем же сортом.

Для всех этих оценок необходимо определить общий урожай по каждому сорту, общий урожай на каждой из трех делянок, общий урожай на делянках, удобрявшихся одним видом удобрений, и, наконец, общий урожай по видам удобрений при объединении всех сортов вместе.

Все эти данные приведены в табл. 47.

Сумма квадратов отклонений от средней для всех 108 данных равна 71,699; разбивая эти данные в соответствии с числом делянок на 36 классов по 3, получаем сумму квадратов для 36 делянок, равную 61,078; разбивая эти 36 данных снова уже в соответствии с числом сортов на 12 классов по 3, получаем сумму квадратов для 12 сортов, равную 43,638. Эти результаты можно представить в виде табл. 48.

Здесь значение z, определяемое как разность логарифмов, приведенных в последнем столбце, равно 0,8484; 1%-ному уровню существенности соответствует табличное значение $z = 0{,}564$; следовательно, различия между урожаями сортов весьма существенны.

Таблица 48

	Степени свободы	Сумма квадратов	Средний квадрат	Логарифм среднего квадратического отклонения
Между сортами	11	43,6384	3,967	0,6890
Между делянками сорта	24	17,4401	0,727	—0,1594
Внутри делянок	72	10,6204	—	—
Итого . . .	107	71,6989	—	—

Вариация внутри делянок может быть подразделена в соответствии с указанными ранее двумя сравнениями видов удобрения. Рассмотрим итоги трех видов удобрений: сумма квадратов отклонений этих итогов от их средней, деленная на 36, равна 0,3495; в этой величине содержится 0,0584 — квадрат разности между итогами двух видов калийного удобрения, разделенный на 72, и остаток 0,2911, являющийся квадратом разности между средней из итогов этих двух видов и итогом известкованных делянок, деленным на 54. Однако возможно, что по этим данным не будет выявлена эффективность удобрений в целом по всем сортам, так как, если сорта имеют разнонаправленную отзывчивость на удобрение, то итоги по удобрениям окажутся примерно одинаковыми и не будет выявлена эффективность удобрений. Остальные 70 степеней свободы также не будут однородными; 36 чисел, относящихся к каждому сорту и к каждому удобрению, дают 35 степеней свободы, из которых 11 представляют различия сортов, 2 — различия удобрений и остальные 22 относятся к отзывчивости сортов на удобрения. Ниже приводится анализ этой части опыта.

Таблица 49

Дисперсия, обусловленная	Степени свободы	Сумма квадратов	Средний квадрат
Калийным удобрением	1	0,2911	0,2911
Различием сульфата и хлорида . . .	1	0,0584	0,0584
Различиями в отзывчивости сортов .	22	2,1911	0,0996
Различиями в отзывчивости делянок с одним и тем же сортом картофеля	48	8,0798	0,1683
Итого . . .	72	10,6204	—

Чтобы произвести оценку существенности дисперсии, относящейся к урожаям делянок с одним и тем же сортом картофеля, следует сравнить 0,727 — число, полученное ранее для 24 степеней свободы, с величиной 0,1683, относящейся к 48 степеням свободы. Значение z — половина разности логарифмов этих чисел — равно 0,7316, в то время как табличное значение для 1%-ного уровня равно 0,394. Отсюда следует, что первое впечатление о неодинаковом плодородии делянок подтверждается. Здесь мы встречаем обычное для полевых опытов затрудняющее работу обстоятельство, состоящее в наличии некоторого разнообразия в характере и глубине пахотного слоя, оказывающего свое влияние на урожайность. В данном случае эта невыравненность условий была, возможно, частично связана с неравномерностью распределения и заделки навоза, а также с неодинаковым его качеством.

В данном опыте не обнаруживается различие в отзывчивости сортов на удобрения; действительно, различие между частями делянок с разными сортами даже меньше, чем различие между частями делянок внутри одного и того же сорта. Однако разность между этими значениями не является существенной: $z = 0{,}2623$, в то время как для 5%-ного уровня z равен около 0,33.

Эффективность удобрений оказалась весьма малой, хотя различия, обусловленные калийными удобрениями, превышают ту

Таблица VI

5%-ный уровень в распределениях z

Значения n_2	Значения n_1									
	1	2	3	4	5	6	8	12	24	∞
1	2,5421	2,6479	2,6870	2,7071	2,7194	2,7276	2,7380	2,7484	2,7588	2,7693
2	1,4592	1,4722	1,4765	1,4787	1,4800	1,4808	1,4819	1,4830	1,4840	1,4851
3	1,1577	1,1284	1,1137	1,1051	1,0994	1,0953	1,0899	1,0842	1,0781	1,0716
4	1,0212	0,9690	0,9429	0,9272	0,9168	0,9093	0,8993	0,8885	0,8767	0,8639
5	0,9441	0,8777	0,8441	0,8236	0,8097	0,7997	0,7862	0,7714	0,7550	0,7368
6	0,8948	0 8188	0,7793	0,7558	0,7394	0,7274	0,7112	0,6931	0,6729	0,6499
7	0,8606	0.7777	0,7347	0,7080	0,6896	0,6761	0,6576	0,6369	0,6134	0,5862
8	0,8355	0,7475	0,7014	0,6725	0,6525	0,6378	0,6175	0,5945	0,5682	0,5371
9	0,8163	0,7242	0,6757	0,6450	0,6238	0,6080	0,5862	0,5613	0,5324	0,4979
10	0,8012	0,7058	0,6553	0,6232	0,6009	0,5843	0,5611	0,5346	0,5035	0,4657
11	0,7889	0,6909	0,6387	0,6055	0,5822	0,5648	0,5406	0,5126	0,4795	0,4387
12	0,7788	0,6786	0,6250	0,5907	0,5666	0,5487	0,5234	0,4941	0,4592	0,4156
13	0,7703	0,6682	0,6134	0,5783	0,5535	0,5350	0,5089	0,4785	0,4419	0,3957
14	0,7630	0,6594	0,6036	0,5677	0,5423	0,5233	0,4964	0,4649	0,4269	0,3782
15	0,7568	0,6518	0,5950	0,5585	0,5326	6,5131	0,4855	0,4532	0,4138	0,3628
16	0,7514	0,6451	0,5876	0,5505	0,5241	0,5042	0,4760	0,4428	0,4022	0,3490
17	0,7466	0,6393	0,5811	0,5434	0,5166	0,4964	0,4676	0,4337	0,3919	0,3366
18	0,7424	0,6341	0,5753	0,5371	0,5099	0,4894	0,4602	0,4255	0,3827	0,3253
~~29~~	0,7386	0,6295	0,5701	0,5315	0,5040	0,4832	0,4535	0,4182	0,3743	0,3151
20	0,7352	0,6254	0,5654	0,5265	0,4986	0,4776	0,4474	0,4116	0,3668	0,3057
21	0,7322	0,6216	0,5612	0,5219	0,4938	0,4725	0,4420	0,4055	0,3599	0,2971
22	0,7294	0,6182	0,5574	0,5178	0,4894	0,4679	0,4370	0,4001	0,3536	0,2892
23	0,7269	0,6151	0,5540	0,5140	0,4854	0,4636	0,4325	0,3950	0,3478	0,2818
24	0,7246	0,6123	0,5508	0,5106	0,4817	0,4598	0,4283	0,3904	0,3425	0,2749
25	0,7225	0,6097	0,5478	0,5074	0,4783	0,4562	0,4244	0,3862	0,3376	0,2685
26	0,7205	0,6073	0,5451	0,5045	0,4752	0,4529	0,4209	0,3823	0,3330	0,2625
27	0,7187	0,6051	0,5427	0,5017	0,4723	0,4499	0,4176	0,3786	0,3287	0,2569
28	0,7171	0,6030	0,5403	0,4992	0,4696	0,4471	0,4146	0,3752	0,3248	0,2516
29	0,7155	0,6011	0,5382	0,4969	0,4671	0,4444	0,4117	0,3720	0,3211	0,2466
30	0,7141	0,5994	0,5362	0,4947	0,4648	0,4420	0,4090	0,3691	0,3176	0,2419
60	0,6933	0,5738	0,5073	0,4632	0,4311	0,4064	0,3702	0,3255	0,2654	0,1644
∞	0,6729	0.5486	0,4787	0,4319	0,3974	0,3706	0,3309	0,2804	0,2085	0

1%-ный уровень в распределениях z

1	4,1535	4,2585	4,2974	4,3175	4,3297	4,3379	4,3482	4,3585	4,3689	4,3794
2	2,2950	2,2976	2,2984	2,2988	2,2991	2,2992	2,2994	2,2997	2,2999	2,3001
3	1,7649	1,7140	1,6915	1,6786	1,6703	1,6645	1,6569	1,6489	1,6404	1,6314
4	1,5270	1,4452	1,4075	1,3856	1,3711	1,3609	1,3473	1,3327	1,3170	1,3000
5	1,3943	1,2929	1,2449	1,2164	1,1974	1,1838	1,1656	1,1457	1,1239	1,0997
6	1,3103	1,1955	1,1401	1,1068	1,0843	1,0680	1,0460	1,0218	0,9948	0,9643
7	1,2526	1,1281	1,0672	1,0300	1,0048	0,9864	0,9614	0,9335	0,9020	0,8658
8	1,2106	1,0787	1,0135	0,9734	0,9459	0,9259	0,8983	0,8673	0,8319	0,7904
9	1,1786	1,0411	0,9724	0,9299	0,9006	0,8791	0,8494	0,8157	0,7769	0,7305
10	1,1535	1,0114	0,9399	0,8954	0,8646	0,8419	0,8104	0,7744	0,7324	0,6816
11	1,1333	0,9874	0,9136	0,8674	0,8354	0,8116	0,7785	0,7405	0,6958	0,6408
12	1,1166	0,9677	0,8919	0,8443	0,8111	0,7864	0,7520	0,7122	0,6649	0,6061
13	1,1027	0,9511	0,8737	0,8248	0,7907	0,7652	0,7295	0,6882	0,6386	0,5761
14	1,0909	0,9370	0,8581	0,8082	0,7732	0,7471	0,7103	0,6675	0,6159	0,5500
15	1,0807	0,9249	0,8448	0,7939	0,7582	0,7314	0 6937	0,6496	0,5961	0,5269
16	1,0719	0,9144	0,8331	0,7814	0,7450	0,7177	0,6791	0,6339	0,5786	0,5064
17	1,0641	0,9051	0,8229	0,7705	0,7335	0,7057	0,6663	0,6199	0,5630	0,4879
18	1,0572	0,8970	0,8138	0,7607	0,7232	0,6950	0,6549	0,6075	0,5491	0,4712
19	1,0511	0,8897	0,8057	0,7521	0,7140	0,6854	0,6447	0,5964	0,5366	0,4560
20	1,0457	0,8831	0,7985	0,7443	0.7058	0,6768	0,6355	0,5864	0,5253	0,4421
21	1,0408	0,8772	0,7920	0,7372	0,6984	0,6690	0,6272	0,5773	0,5150	0,4294
22	1,0363	0,8719	0,7860	0,7309	0,6916	0,6620	0,6196	0,5691	0,5056	0,4176
23	1,0322	0,8670	0,7806	0,7251	0,6855	0,6555	0,6127	0,5615	0,4969	0,4068
24	1,0285	0,8626	0,7757	0,7197	0,6799	0,6496	0,6064	0,5545	0,4890	0,3967
25	1,0251	0,8585	0,7712	0,7148	0,6747	0,6442	0,6006	0,5481	0,4816	0,3872
26	1,0220	0,8548	0,7670	0,7103	0,6699	0,6392	0,5952	0,5422	0,4748	0,3784
27	1,0191	0,8513	0,7631	0,7062	0 6655	0,6346	0,5902	0,5367	0,4685	0,3701
28	1,0164	0,8481	0,7595	0,7023	0,6614	0,6303	0,5856	0,5316	0,4626	0,3624
29	1,0139	0,8451	0,7562	0,6987	0,6576	0,6263	0,5813	0,5269	0,4570	0,3550
30	1,0116	0,8423	0,7531	0,6954	0,6540	0,6226	0,5773	0,5224	0,4519	0,3481
60	0,9784	0,8025	0,7086	0,6472	0,6028	0,5687	0,5189	0,4574	0,3746	0,2352
∞	0,9462	0,7636	0,6651	0,5999	0,5522	0,5152	0,4604	0,3908	0,2913	0

0,1 %-ный уровень в распределениях z

Значения n_2	Значения n_1									
	1	2	3	4	5	6	8	12	24	∞
1	6,4577	6,5612	6,5966	6,6201	6,6323	6,6405	6,6508	6,6611	6,6715	6,6819
2	3,4531	3,4534	3,4535	3,4535	3,4535	3,4535	3,4536	3,4537	3,4536	3,4536
3	2,5604	2,5003	2,4748	2,4603	2,4511	2,4446	2,4361	2,4272	2,4179	2,4081
4	2,1529	2,0574	2,0143	1,9892	1,9728	1,9612	1,9459	1,9294	1,9118	1,8927
5	1,9255	1,8002	1,7513	1,7184	1,6964	1,6808	1,6596	1,6370	1,6123	1,5845
6	1,7849	1,6479	1,5828	1,5433	1,5177	1,4986	1,4730	1,4449	1,4134	1,3783
7	1,6874	1,5384	1,4662	1,4221	1,3927	1,3711	1,3417	1,3090	1,2721	1,2296
8	1,6177	1,4587	1,3809	1,3332	1,3008	1,2770	1,2443	1,2077	1,1662	1,1169
9	1,5646	1,3982	1,3160	1,2653	1,2304	1,2047	1,1694	1,1293	1,0830	1,0279
10	1,5232	1,3509	1,2650	1,2116	1,1748	1,1475	1,1098	1,0668	1,0165	0,9557
11	1,4900	1,3128	1,2238	1,1683	1,1297	1,1012	1,0614	1,0157	0,9619	0,8957
12	1,4627	1,2814	1,1900	1,1326	1,0926	1,0628	1,0213	0,9733	0,9162	0,8450
13	1,4400	1,2553	1,1616	1,1026	1,0614	1,0306	0,9875	0,9374	0,8774	0,8014
14	1,4208	1,2332	1,1376	1,0772	1,0348	1,0031	0,9586	0,9066	0,8439	0,7635
15	1,4043	1,2141	1,1169	1,0553	1,0119	0,9795	0,9336	0,8800	0,8147	0,7301
16	1,3900	1,1976	1,0989	1,0362	0,9920	0,9588	0,9119	0,8567	0,7891	0,7005
17	1,3775	1,1832	1,0832	1,0195	0,9745	0,9407	0,8927	0,8361	0,7664	0,6740
18	1,3665	1,1704	1,0693	1,0047	0,9590	0,9246	0,8757	0,8178	0,7462	0,6502
19	1,3567	1,1591	1,0569	0,9915	0,9442	0,9103	0,8605	0,8014	0,7277	0,6285
20	1,3480	1,1489	1,0458	0,9798	0,9329	0,8974	0,8469	0,7867	0,7115	0,6086
21	1,3401	1,1398	1,0358	0,9691	0,9217	0,8858	0,8346	0,7735	0,6964	0,5904
22	1,3329	1,1315	1,0268	0,9595	0,9116	0,8753	0,8234	0,7612	0,6828	0,5738
23	1,3264	1,1240	1,0186	0,9507	0,9024	0,8657	0,8132	0,7501	0,6704	0,5583
24	1,3205	1,1171	1,0111	0,9427	0,8939	0,8569	0,8038	0,7400	0,6589	0,5440
25	1,3151	1,1108	1,0041	0,9354	0,8862	0,8489	0,7953	0,7306	0,6483	0,5307
26	1,3101	1,1050	0,9978	0,9286	0,8791	0,8415	0,7873	0,7220	0,6385	0,5183
27	1,3055	1,0997	0,9920	0,9223	0,8725	0,8346	0,7800	0,7140	0,6294	0,5066
28	1,3013	1,0947	0,9866	0,9165	0,8664	0,8282	0,7732	0,7066	0,6209	0,4957
29	1,2973	1,0903	0,9815	0,9112	0,8607	0,8223	0,7679	0,6997	0,6129	0,4853
30	1,2936	1,0859	0,9768	0,9061	0,8554	0,8168	0,7610	0,6932	0,6056	0,4756
60	1,2413	1,0248	0,9100	0,8345	0,7798	0,7377	0,6760	0,5992	0,4955	0,3198
∞	0,1910	0,9663	0,8453	0,7648	0,7059	0,6599	0,5917	0,5044	0,3786	0

эффективность, которую теперь мы можем назвать случайным варьированием ($0{,}2911 > 0{,}1683$), однако здесь z равно только 0,3427, тогда как при существенном уровне требуется около 0,7. При отсутствии общей отзывчивости на удобрение следует, пожалуй, ожидать (хотя это отнюдь не обязательно) несущественности отзывчивости на удобрения и по отдельным сортам. По-видимому, растения на известкованных делянках имели достаточное количество калия и поэтому его добавление не сказалось на урожае и не выявило различия в действии сульфата и хлорида калия.

ГЛАВА ВОСЬМАЯ

РАЗЛИЧНЫЕ ПРИЛОЖЕНИЯ ДИСПЕРСИОННОГО АНАЛИЗА

43. В этой главе будут даны примеры более широкого использования дисперсионного анализа, который в предыдущей главе рассматривался только в связи с внутриклассовой корреляцией. Теперь же он будет рассматриваться как самостоятельный статистический метод. Недостаток места не дает нам возможности привести здесь примеры всех разнообразных приложений этого метода, поэтому мы ограничимся рассмотрением только тех наиболее важных для практики случаев, когда имеет место применение неправильных методов или когда исследование не может быть проведено никаким иным методом, кроме метода дисперсионного анализа.

44. Определение формы уравнения регрессии

При анализе экспериментальных данных прежде всего возникает вопрос, в какой мере они согласуются с некоторой рабочей гипотезой. В предыдущих главах были даны главным образом критерии, основанные на численностях определенных случаев; к таким гипотезам относятся, например, гипотеза Менделя, гипотеза о линейном расположении некоторого числа связанных групп, гипотеза о независимости или коррелированности переменных и пр. Однако более часто возникает необходимость проверить гипотезы, касающиеся формы регрессии. Например, требуется установить, подчиняется ли рост животного, растения или народонаселения некоторому определенному закону, т. е., положим, увеличивается ли он во времени по арифметической или геометрической прогрессии или он подчиняется так называемому «автокаталитическому» или «логистическому» закону роста. В другом случае следует, положим, установить, не находится ли в соответствии с некоторым гипотетическим законом рост урожая, обусловленный повышением доз удобрения. Такие вопросы возникают не только при проверке строгих законов, имеющих широкое распространение, но и в случае проверки соотношений, имеющих эмпирический характер; эти вопросы имеют значение не только на окончательной стадии изучения законов природы, но и на более ранних стадиях, например при оценке точности экспериментальной техники. Методами, предназначенными для решения этих вопросов, мы уже пользовались ранее при определении критериев согласия наблюденных данных с заданной линией регрессии. Эти методы, с одной стороны, направлены на упрощение вычислений путем приведения их к стандартной схеме, позволяющей произвести точную оценку коэффициентов регрессии, и, с другой стороны, они требуют такого построения всего процесса разработки материала, которое полностью соответствовало бы поставленной задаче.

Допустим, что для каждого из нескольких значений независимой переменной x произведено некоторое число наблюдений над зависимой переменной y; пусть число значений x равно a, тогда a является числом строев. Обозначим некоторый частный строй подстрочным показателем p, число наблюдений в этом строе — n_p и соответствующую среднюю — $\bar{y}_p$. Общую среднюю для всех значений y обозначим $\bar{y}$. В этом случае, каким бы ни был закон изменения наших данных, можно написать чисто алгебраическое тождество:

$$S(y-\bar{y})^2 = S\{n_p(\bar{y}_p-\bar{y})^2\} + SS(y-\bar{y}_p)^2 .$$

Это тождество показывает, что сумма квадратов отклонений всех значений y от их общей средней может быть разложена на две части, одна из которых представляет собой сумму отклонений средних по строям от общей средней, причем каждый из этих квадратов отклонений умножается на число наблюдений в данном строе, а вторая часть является суммой квадратов отклонений каждого наблюдения от средней того строя, к которому это наблюдение принадлежит. Этот анализ сходен с тем, который был применен для определения внутриклассовой корреляции, но только теперь число наблюдений в каждом классе различно. Отклонения от средних внутри строев обусловлены такими причинами варьирования, как ошибки группировки, погрешности наблюдений и пр., которые не зависят от значений x. Таким образом, среднее квадратическое отклонение, обусловленное этими причинами, дает основу для определения того, в какой мере существенными являются отклонения средних по строям от тех их значений, которые ожидаются согласно проверяемой гипотезе.

Положим, что Y_p является таким ожидаемым значением для строя номер p, тогда величина

$$S\{n_p(\bar{y}_p - Y_p)^2\}$$

будет измерять различие между фактической и гипотетической средней. При сравнении этой величины с вариацией внутри строев, конечно, следует учитывать число степеней свободы, принадлежа-

щих системе отклонений фактических средних от гипотетических их значений. В некоторых относительно редких случаях гипотеза вполне точно определяет ожидаемое для каждого строя значение средней. В таких случаях указанной выше сумме квадратов соответствует a степеней свободы, равное числу строев. Однако более часто гипотеза определяет только форму линии регрессии, имеющей один или несколько параметров, которые должны определяться по наблюдениям, как например в том случае, когда требуется проверить, не является ли регрессия прямолинейной, и когда эта гипотеза прямолинейности может быть проверена только после вычисления по имеющимся данным соответствующих коэффициентов регрессии. В таких случаях число степеней свободы определяется как разность числа строев a и числа параметров, определенных по наблюдениям.

Пример 42. *Определение того, в какой мере регрессия может считаться линейной.* Следующие данные взяты из работы Херша, в которой изучалось влияние температуры на число глазных фасеток у Drosophila melanogaster в различных монозиготных и гетерозиготных фазах фактора «bar». Здесь взяты самки

Таблица 50

Температура С	15°	17°	19°	21°	23°	25°	27°	29°	31°	Итого
+8,07	3	1	1	—	—	—	—	—	—	5
+7,07	5	2	5	1	—	—	—	—	—	13
+6,07	13	7	3	—	—	—	—	—	—	23
+5,07	25	9	2	1	—	—	—	—	—	37
+4,07	22	10	16	—	—	2	—	—	—	50
+3,07	12	10	12	6	1	3	—	—	—	44
+2,07	7	5	14	16	2	2	—	—	—	46
+1,07	3	4	14	21	8	9	—	—	—	59
+0,07	—	3	7	26	7	19	1	—	—	63
−0,93	—	1	7	12	11	24	3	1	—	59
−1,93	—	—	1	9	14	22	8	6	—	60
−2,93	—	2	1	5	12	15	15	4	—	54
−3,93	—	—	—	2	19	18	44	10	1	94
−4,93	—	—	—	1	4	4	26	6	6	47
−5,93	—	—	—	—	2	2	19	14	13	50
−6,93	—	—	—	—	2	—	11	28	9	50
−7,93	—	—	—	—	3	1	8	8	8	28
−8,93	—	—	—	—	1	—	2	5	5	13
−9,93	—	—	—	—	—	—	—	4	4	8
−10,93	—	—	—	—	—	—	—	10	2	12
−11,93	—	—	—	—	—	1	—	1	2	4
−12,93	—	—	—	—	—	—	—	0,5	1,5	2
−13,93	—	—	—	—	—	—	—	0,5	0,5	1
−14,93	—	—	—	—	—	—	—	—	—	—
−15,93	—	—	—	—	—	—	—	—	1	1
Итого	90	54	83	100	86	122	137	98	53	823

гетерозиготные в отношении «full» и «double — bar», число фасеток измерялось в факториальных единицах, что равносильно логарифмической шкале. Можно ли влияние температуры на число фасеток, измеряемое в указанных единицах, представить в виде прямой линии?

Здесь имеется 9 строев, соответствующих 9 различным температурам. Взяв в качестве условного начала — 1,93, вычисляем для каждого строя и для всех строев, взятых в целом, суммы и средние отклонения от условного начала. Каждая средняя получена делением суммы отклонений на численность соответствующего строя; при этих расчетах у средней по строям достаточно определить три знака после запятой, увеличив для обобщенной средней число знаков до четырех. В результате получим:

Таблица 51

Строй	15	17	19	21	23
Сумма отклонений	583	294	367	225	−43
Средние отклонения . . .	6,478	5,444	4,422	2,250	−0,500

Строй	25	27	29	31	В целом
Сумма отклонений	+37	−369	−463,5	−306,5	+324
Средние отклонения . . .	+0,303	−2,693	−4,730	−5,783	+0,3937

Сумма произведений этих чисел, взятых попарно, после вычитания из нее произведения двух последних чисел дает величину

$$S\{n_p(\bar{y}_p - \bar{y}^2\} = 12\,370.$$

В то же время из распределения всех наблюдений y можно найти

$$S(y - \bar{y})^2 = 16\,202.$$

Теперь можно построить такую таблицу:

Таблица 52

Дисперсия	Степени свободы	Сумма квадратов	Средний квадрат
Между строями	8	12 370	—
Внутри строев	814	3 832	4,708
Итого . . .	822	16 202	—

Таким образом, дисперсия внутри строев равна только 4,7. Дисперсия же между строями должна быть разложена на части, одна из которых относится к линейной регрессии, а вторая — к системе отклонений фактических средних по строям от прямой линии. Для определения той части, которая представляет линейную регрессию, вычисляем:

$$S(x-\bar{x})^2 = 4742{,}21$$

$$S(x-\bar{x})(y-\bar{y}) = -7535{,}38,$$

причем последняя величина может быть получена путем умножения приведенных выше сумм отклонений на соответствующие $x-\bar{x}$. Вычисляя после этого

$$\frac{(7535{,}38)^2}{4742{,}21} = 11{,}974,$$

можно построить такую таблицу дисперсионного анализа:

Таблица 53

Дисперсия между строями, обусловленная	Степени свободы	Сумма квадратов	Средний квадрат
Линейной регрессией	1	11 974	—
Отклонениями от регрессии	7	396	56,6
Итого . . .	8	12 370	—

Сумму квадратов 396 можно проверить путем определения соответствующих отклонений $\bar{y}_p$ от значений Y_p, вычисленных по формуле регрессии, и дальнейшего расчета по формуле:

$$S\{n_p(\bar{y}_p - Y_p)^2\}.$$

Этот контроль полезен и в том отношении, что он дает возможность видеть, в каких строях имеют место наибольшие отклонения. В нашем случае это происходит при температуре 23° и 25°С.

Отклонения от линейной регрессии по сравнению с вариацией внутри строев, очевидно, больше тех, которые следовало бы ожидать при прямолинейной регрессии. Для расчета критерия z имеем табл. 54.

Табличное значение z для 1%-ного уровня существенности составляет около 0,488. Здесь, таким образом, не может быть никакого сомнения в статистической достоверности отклонений от прямой линии, несмотря на то, что последняя обусловливает собой большую часть вариации между строями.

Таблица 54

Степени свободы	Средний квадрат	Натуральный логарифм	$\frac{1}{2}\log_e$
7	56,6	4,0360	2,0180
814	7,708	1,5493	0,7746
	Разность (z)		1,2434

Отметим, что при вычислении этого критерия поправка Шеппарда не применялась; конечно, некоторая часть вариации внутри строев и отклонений от линии регрессии определяется ошибками группировки, но исключение в каждом из этих случаев средней ошибки, связанной с этой причиной, приведет к излишней контрастности сравнения и к некоторой потере точности нашего критерия существенности.

В качестве дополнительного материала для применения изложенного в этой главе критерия может служить пример расчета регрессии в параграфе 29.2.

45. Корреляционное отношение η

Ранее мы видели, что из суммы квадратов отклонений всех наблюдений от общей средней может быть выделена часть, характеризующая различия между отдельными строями. Отношение этой части к целому обычно обозначается символом η^2, так что

$$\eta^2 = S[n_p(\bar{y}_p - \bar{y})^2] : S(y-\bar{y})^2$$

и квадратный корень из этой величины называется корреляционным отношением y на x. Подобно этому, если значения Y являются ординатами гипотетической функции регрессии, то можно определить R, как отношение:

$$R^2 = S[n_p(Y-\bar{y})^2] : S(y-\bar{y})^2.$$

Здесь R коэффициент корреляции между y и Y; в частности, если регрессия прямолинейна, то $R^2 = r^2$, где r — коэффициент корреляции между x и y. Из этих соотношений видно, что η больше R и, следовательно, η определяет собой верхний предел в том смысле, что не существует такой функции регрессии, корреляция которой с y была бы выше η.

В качестве самостоятельной статистики корреляционное отношение имеет весьма ограниченное применение. Следует отметить, что число степеней свободы в числителе η зависит от числа строев; в примере 42 значение η зависит не только от интервала фиксированных температур, но и от числа температур, отмеченных внутри этого интервала.

Критерий для определения того, будет ли некоторое полученное из наблюдений значение корреляционного отношения существенным, является в то же время критерием того, что можно ли считать вариацию между строями существенно большей, чем это можно ожидать на основе данной вариации внутри строев. Этот критерий можно получить из таблицы дисперсионного анализа (табл. 52) и сравнить его с табличным значением z. Были сделаны попытки построить критерий существенности для корреляционного отношения на основе его средней квадратической ошибки, но при этом упускался из виду тот факт, что даже при сколь угодно больших выборках распределение η для строев, между которыми фактически нет различий, не имеет тенденции приближаться к нормальному распределению, если только и число строев также не возрастает беспредельно. Напротив, при очень больших выборках величина $N\eta^2$, где N — общее число наблюдений, стремится к распределению χ^2 при числе степеней свободы n, равном $a - 1$, т. е. на единицу меньше, чем число строев.

46. Критерий Блекмана

В той же мере, в какой η^2 измеряет различия между строями, величина $(\eta^2 - R^2)/(1 - R^2)$ измеряет обобщенное отклонение средних по строям от гипотетической линии регрессии. Попытка получить критерий того, что данная регрессия линейна, путем сравнения этой величины с ее средней квадратической ошибкой привела к известному критерию Блекмана. В этом критерии попрежнему не учитывается число строев и, следовательно, он не дает даже приближенной оценки допустимых значений разности $\eta^2 - r^2$. Подобно тому, как это было с величиной η^2 при нулевой регрессии, распределение величины $\eta^2 - r^2$ при беспредельном увеличении числа наблюдений не приближается к нормальному распределению, но распределение величины $\frac{N(\eta^2 - r^2)}{1 - r^2}$ стремится к распределению χ^2, в котором $n = a - 2$. Среднее значение этой величины равно $(a - 2)$; следовательно, если не принимается во внимание число строев a, то тем самым упускается из виду главная характеристика в ее выборочном распределении.

В примере 42 было установлено, что при 9 строях отклонение от линейной регрессии было существенным, но легко видеть, что если бы здесь было 90 строев при тех же самых значениях η^2 и r^2, то отклонение от линейной регрессии было бы даже меньше того, которое ожидается на основе вариации внутри строев. Пользуясь же критерием Блекмана, мы не смогли бы различить эти два противоположные друг другу случая.

Как и во всех других случаях применения критерия согласия, при оценке того, в какой мере линия регрессии согласуется с имеющимися данными, существенным является то положение, что если по наблюденным данным устанавливаются некоторые параметры уравнения регрессии, то этот процесс обязательно должен найти свое отражение в соответствующем критерии согласия.

Некоторые сведения относительно методов решения такого рода задач были даны в главе V. Вообще же, если иметь в виду более сложные случаи, не затронутые в этой книге, то указанное выше условие находит свое осуществление в способе разработки статистического материала, известном под названием метода наименьших квадратов.

В этом случае сводится к минимуму влияние гипотетических условий, связанных с выбранной формой Y, на величину

$$S[n_p(\bar{y}_p - Y_p)^2],$$

характеризующую отклонение от линии регрессии.

Случаи, когда может быть применен метод наименьших квадратов, являются специальными случаями приложения метода максимального правдоподобия, из которого первый метод может быть получен в качестве частного случая. Метод максимального правдоподобия будет рассмотрен в главе IX.

47. Существенность множественного коэффициента корреляции

Если, как это было в параграфе 29 (стр. 130), регрессия зависимой переменной y на некоторое число независимых переменных, положим x_1, x_2, x_3, выражена в форме уравнения

$$Y = b_1x_1 + b_2x_2 + b_3x_3,$$

то корреляция между y и Y будет больше, чем корреляция y с любой другой линейной функцией этих независимых переменных и, следовательно, эта корреляция измеряет в той мере, в какой величина y зависит от этого, совместную вариацию независимых переменных. Определенная таким образом корреляция, обозначаемая R, может быть вычислена по формуле

$$R^2 = [b_1S(x_1y) + b_2S(x_2y) + b_3S(x_3y)] : S(y^2).$$

Этот множественный коэффициент корреляции R отличается от коэффициента корреляции при одной независимой переменной тем, что он всегда положителен. Кроме этого, установлено, что распределение множественного коэффициента корреляции в случайных выборках зависит от числа независимых переменных. Обработка данных в этом случае полностью сходна с той, которая была дана для корреляционного отношения (параграф 45), и приводит к аналогичному дисперсионному анализу.

При построении дисперсионного анализа мы будем основываться на том обстоятельстве, что

$$S(y^2) = S(y - Y)^2 + [b_1S(x_1y) + b_2S(x_2y) + b_3S(x_3y)].$$

Если n' число наблюденных значений y и p — число независимых переменных, то этим трем суммам будут соответствовать сте-

пени свободы: $n'-1$, $n'-p-1$ и p. Поэтому дисперсионный анализ будет иметь такой общий вид:

Таблица 55

Дисперсия, обусловленная	Степени свободы	Сумма квадратов
Функцией регрессии	p	$b_1 S(x_1 y) + \ldots$
Отклонениями от функции регрессии . .	$n'-p-1$	$S(y-Y)^2$
Итого . . .	$n'-1$	$S(y^2)$

Здесь считается, что y измеряется в отклонениях от средней. Если в действительности нет никакой связи между независимыми переменными x и зависимой переменной y, то числа в колонке «сумма квадратов» будут примерно пропорциональны числам степеней свободы; если же имеет место некоторая реальная связь между этими переменными, то p степеней свободы, относящихся к функции регрессии, получат значительно большую долю из суммы квадратов $S(y^2)$. Оценка того, будет ли R существенным, фактически является оценкой того, будет ли средний квадрат, относящийся к функции регрессии, существенно большим, чем средний квадрат отклонений от функции регрессии. Эта оценка может быть проведена, как и в других подобных случаях, при помощи таблицы z.

Пример 43. *Существенность множественной корреляции.* Для иллюстрации этого анализа произведем по данным примера 24 оценку существенности связи между осадками и местоположением станции, определяемым долготой, широтой и высотой над уровнем моря. По результатам, полученным в этом примере, можно построить следующую таблицу:

Таблица 56

Дисперсия, обусловленная	Степени свободы	Сумма квадратов	Средний квадрат	$\frac{1}{2}\log_e$
Формулой регрессии .	3	791,7	263,9	2,7878
Отклонениями	53	994,9	18,77	1,4661
Итого . . .	56	1786,6	—	—

Таким образом, фактическое значение z равно 1,3217, в то время как для 1%-ного уровня оно равно 0,714; это указывает на

явную существенность множественной корреляции. Сам коэффициент корреляции может быть вычислен по данным этой же таблицы

$$R^2 = 791{,}7 : 1786{,}6 = 0{,}4431$$
$$R = 0{,}6657.$$

Этот расчет имеет самостоятельное значение и не является необходимым при оценке существенности корреляции.

48. Техника полевого экспериментирования

Статистический метод дисперсионного анализа имеет существенное значение для понимания принципов, лежащих в основе современных способов построения полевых сельскохозяйственных опытов. Этот и два последующих параграфа будут посвящены приложению этого метода к полевому экспериментированию. Данный вопрос тесно связан с быстро развивающейся теорией построения опыта; более полное изложение принципов и логики экспериментирования можно найти в моей работе «The Design of Experiments».

Первое требование, которое предъявляется всякому хорошо поставленному опыту, состоит в том, что он должен давать возможность не только проводить сравнение различных удобрений, способов обработки, сортов и т. д., но и возможность производить оценку существенности таких различий. Следовательно, все варианты опыта должны быть по меньшей мере дублированы для того, чтобы расхождения повторных результатов могли бы быть в качестве своего рода стандарта, с которым можно было бы сравнить наблюдаемые различия между вариантами. Вообще же лучше иметь более высокую повторность опыта. Это требование является общим для всех различного вида экспериментов; но в сельскохозяйственных полевых опытах приходится иметь дело еще с некоторым дополнительным и достаточно важным обстоятельством. Как установлено на основании многочисленных дробных учетов, участок, предназначенный для размещения на нем опытных делянок, обычно весьма неоднороден в том отношении, что плодородие его изменяется в какой-то мере систематически, хотя это происходит довольно сложным образом. Для того чтобы наша оценка существенности была бы правильной, необходимо, чтобы различия в плодородии между повторными делянками варианта имели бы ту же репрезентативность, как и различия в плодородии между делянками, предназначенными для различных вариантов опыта. Это условие не будет удовлетворено, если опытные делянки будут подчинены некоторой специально выбранной системе расположения, так как систематическое расположение будет, и это доказано на основе изучения данных дробных учетов, в целом отражать систематическую изменчивость плодородия участка, и поэтому критерий существенности в этом случае станет недействительным.

Пример 44. *Точность при случайном распределении вариантов опыта.* Непосредственный путь преодоления отрицательного влияния систематического изменения плодородия участка состоит в том, что варианты опыта размещаются на делянках по жребию, т. е. случайно. Например, если 20 удлиненных делянок предназначаются для изучения 5 различных вариантов опыта, каждый в четырехкратной повторности, то для установления порядка расположения вариантов можно взять 20 карточек, выписать на них названия вариантов и, тщательно перемешав их, принять тот порядок, который сложился в результате их перемешивания. Допустим, получен порядок табл. 57:

Таблица 57

B	C	A	C	E	E	E	A	D	A
3504	3430	3376	3334	3253	3314	3287	3361	3404	3366
B	C	B	D	D	B	A	D	C	E
3416	3291	3244	3210	3168	3195	3330	3118	3029	3085

Здесь буквы A, B и т. д. обозначают 5 различных вариантов опыта; под каждой из них дан вес корней кормовой свеклы, полученный Мерсером и Галлом при дробном учете участка, разрезанного на 20 полос.

Отклонения суммарных урожаев каждого варианта от средней из этих суммарных урожаев таковы:

A	B	C	D	E
+290	+216	−59	−243	−204

В таблице дисперсионного анализа сумма квадратов, стоящая в строке «варианты», определена как четвертая часть суммы квадратов этих отклонений. Так как сумма квадратов всех 20 отклонений от общей средней равна 289766, то можно составить следующую таблицу дисперсионного анализа:

Таблица 58

Дисперсия, обусловленная	Степени свободы	Сумма квадратов	Средний квадрат	Среднее квадратическое отклонение
Вариантами	4	58 726	14 681	121,1
Экспериментальной ошибкой	15	231 040	15 403	124,1
Итого . . .	19	289 766	15 251	123,5

Отсюда следует, что средняя квадратическая ошибка в пересчете на единичную делянку равна 124,1, в то время как известное нам в данном случае ее точное значение составляет 123,5. Здесь имеет место весьма хорошее приближение их друг к другу,

что указывает на то, что рассматриваемый способ чисто случайного размещения делянок обеспечивает несмещенность вычисленной экспериментальной ошибки относительно реально существующей ошибки.

Пример 45. *Ограничения случайного размещения вариантов.* Придерживаясь того весьма существенного условия, что ошибки, свойственные результатам опыта, должны быть ничем иным, как случайной выборкой таких ошибок, которая должна дать правильную оценку ошибки опыта, мы все же можем специальным построением опыта исключить из ошибки опыта довольно значительную часть влияния, оказываемого почвенной неоднородностью. Это увеличение точности опыта достигается путем наложения известных ограничений на порядок расположения делянок. Рассмотрим в качестве иллюстрации широко распространенного способа закладки опытов схему опыта, в котором 20 делянок разделены на 4 блока (повторения), а размещение вариантов подчинено условию, чтобы каждый вариант в каждом блоке встречался бы только один раз. В этом случае дисперсию можно разложить на три части, представляющие: 1) локальные различия между блоками, 2) различия, обусловленные вариантами, 3) экспериментальные ошибки. Если 5 вариантов внутри блоков размещены случайно, то наша оценка экспериментальной ошибки будет несмещенной оценкой фактических ошибок у различных вариантов. В качестве примера случайного размещения вариантов, подчиненного этому ограничению, рассмотрим следующую схему опыта:

AECDB CBEDA ADEBC CEBAD

Проведя анализ на основе тех же числовых данных, что и ранее, и выделяя из общей суммы квадратов части, относящиеся к локальным различиям между блоками и к различиям между вариантами, находим:

Таблица 59

Дисперсия, обусловленная	Степени свободы	Сумма квадратов	Средний квадрат	Среднее квадратическое отклонение
Локальными различиями . . .	3	154 483	51 494	—
Вариантами	4	40 859	10 215	—
Экспериментальной ошибкой .	12	94 424	7 869	88,7
Вариантами + ошибкой . . .	16	135 283	8 455	92,0

Локальные различия между блоками оказались весьма существенными, в результате чего точность опыта значительно возросла, на что указывает снижение остаточной дисперсии почти на 55% ее прежнего значения. Это расположение вариантов внутри блоков, проведенное по жребию, случайно оказалось не-

сколько неудачным в том отношении, что ошибки, принадлежащие вариантам, оказались несколько завышенными, вследствие чего средняя квадратическая ошибка опыта оказалась равной 88,7 вместо точного значения 92,0. Такого рода отклонения от точного значения не представляют собой нечто неожиданное, и именно с учетом возможности появления их строится критерий существенности.

Может показаться, что более рационально расположить варианты в систематическом порядке, например в таком:

ABCDE EDCBA ABCDE EDCBA

Если взять рассматриваемый здесь числовой пример, то такое размещение вариантов приведет к меньшим ошибкам по строке «варианты», так как в этом случае делянки каждого варианта лучше охватывают различные градации плодородия участка. Однако при таком расположении опыта мы не имеем гарантии того, что оценка средней квадратической ошибки будет правильно репрезентировать различия, возникшие между отдельными вариантами, и, следовательно, в данном случае этой средней квадратической ошибке нельзя доверять, а в связи с этим мы лишаемся возможности опираться на достаточно надежный критерий существенности.

Та часть фактора плодородия, которая не оказалась включенной в различия блоков, может быть частично элиминирована, если рассматривать местоположение делянок внутри блоков, т. е. целые числа 1, 2, 3, 4, 5 или, что более удобно, числа —2; —1; 0, 1, 2, в качестве значений независимой переменной, характеризующей это местоположение делянок, а в качестве зависимой переменной взять урожай и если применить здесь метод регрессий. Ковариационный анализ указанных целых чисел (x) и урожаев делянок (y), подобный анализу параграфа 49.1, дает следующие результаты:

Таблица 59.1

Ковариационный анализ урожая (y) и порядковых номеров внутри блоков (x)

	Степени свободы	x^2	xy	y^2	y_x^2	Средний квадрат
Блоки	3	0	0	154 483	—	—
Варианты	4	5,5	—268,25	40 859	28 678	7169,5
Ошибка	12	34,5	—1206,75	94 424	52 214	4746,7
Варианты+ошибка	16	40,0	—1475,00	135 283	80 892	5392,8

Из этой таблицы видно, что точность опыта увеличилась. Средний квадрат по строке «варианты + ошибка» теперь равен 5393 вместо 8455 в опыте с блоками и 15 251 в опыте без образования блоков.

Возьмем в качестве условной единицы, измеряющей познавательную ценность эксперимента, такой опыт, который дает урожаи со средней квадратической ошибкой, составляющей 10% от средней. В этом случае значение любого эксперимента, поставленного так же, как у нас в четырехкратной повторности, может быть определено следующим образом: возводим в квадрат одну десятую среднего урожая, взятого по всему опыту в целом, и умножаем на четыре (в нашем случае это дает 431 816), после этого делим найденное число на средний квадрат, полученный при том или ином построении опыта. Для случайного расположения без блоков мы имеем

$$\frac{431816}{15250{,}8} = 28{,}18 \text{ единиц информации.}$$

Для рэндомизированных блоков мы имеем 51,07, а при введении поправки на размещение вариантов внутри блоков на основе регрессии мы повысим значение опыта до 80,07 единиц информации.

В данном случае, как и во многих других подобных случаях, понижение среднего квадрата сопровождается потерей некоторого числа степеней свободы. Это приводит к тому, что применяемые в этих случаях критерии существенности становятся менее строгими. Если n — число степеней свободы для ошибки, то уменьшение объема информации, возникающее по этой причине, примерно составляет долю $\frac{2}{(n+3)}$ (см. Design of Experiments). Эта поправка в нашем случае дает такие результаты:

Таблица 59.2

Подсчет количества информации при различных методах

	Степени свободы	Единиц информации	
		при первом подсчете	при точном подсчете
Рэндомизация 20 делянок	15	28,18	25,05
Рэндомизация в четырех блоках	12	51,07	44,26
Элиминирование порядка расположения в блоках	11	80,07	68,63

В нашем случае даже после введения этой поправки на уменьшение числа степеней свободы получается выгода от использования блоков и от учета местоположения вариантов внутри блоков. Последнее, конечно, обусловлено наличием систематического изменения плодородия участка, что и выявляется методом регрессий.

49. Латинский квадрат

Этот прием наложения ограничений на случайное размещение вариантов и исключения соответствующих степеней свободы из дисперсии получает дальнейшее развитие в специальном постро-

нии опыта, известном под названием латинского квадрата. В блоке из 25 делянок, размещенных в 5 рядов и 5 столбцов, в котором изучается 5 вариантов, можно осуществить такое размещение вариантов, чтобы каждый вариант встречался один раз в каждом ряду и один раз в каждом столбце. В остальном, кроме этих ограничений, размещению вариантов предоставляется полная свобода. В этом случае из 24 степеней свободы 4 будут относиться к вариантам, 8 степеней, характеризующих различия в плодородии между рядами и столбцами, могут быть элиминированы, а остальные 12 будут относиться к ошибке опыта. Эти 12 степеней могут дать несмещенную оценку ошибок сравнения между вариантами опыта, основанную на том, что в каждом варианте одинаковым образом представлены как все ряды, так и все столбцы.

Пример 46. *Расположение вариантов с двойным ограничением.* Мерсером и Галлом были получены следующие урожаи кормовой свеклы с 25 делянок. Имея эти данные, мы распределили буквы латинского алфавита, представляющие условно 5 различных вариантов, в случайном порядке, но так, чтобы каждая из них встречалась один раз в каждом ряду и один раз в каждом столбце.

Таблица 60

						Итого по рядам
	D 376	E 371	C 355	B 356	A 335	1 793
	B 316	D 338	E 336	A 356	C 332	1 678
	C 326	A 326	B 335	D 343	E 330	1 660
	E 317	B 343	A 330	C 327	D 336	1 653
	A 321	C 332	D 317	E 318	B 306	1 594
Итого	1 656	1 710	1 673	1 700	1 639	8 378

Выделяя из дисперсии части, относящиеся к рядам, столбцам и вариантам, находим:

Таблица 61

Различия между	Степени свободы	Сумма квадратов	Средний квадрат	Средние квадратические отклонения
Рядами	4	4240,24	—	—
Столбцами	4	701,84	—	—
Вариантами	4	330,24	130,3	11,41
Остаточные различия . . .	12	1754,32		
Итого . . .	24	7024,64	292,8	17,11

Благодаря исключению почвенных различий между рядами и столбцами средний квадрат уменьшился больше чем наполовину и познавательная ценность опыта увеличилась по этой причине более чем вдвое. Этот прием элиминирования различий между рядами и столбцами может быть объединен с приемом элиминирования различий по повторениям; в связи с этим можно получить довольно высокую точность опыта, если его разместить в нескольких блоках, состоящих, положим, из 5 рядов и 5 столбцов. Такое построение опыта применимо даже в случае изучения только трех вариантов. Так как этот способ пригоден какими бы ни были и отчего бы ни происходили различия в фактическом плодородии почвы, то тот же статистический прием исключения различий между рядами и столбцами можно применить и тогда, когда, положим, 25 удлиненных делянок идут одна за другой, составляя полосу. Рассматривая каждый блок из 5 таких делянок в качестве последовательных столбцов латинского квадрата, можно элиминировать не только различия между блоками, но и такие различия, которые обусловлены градациями плодородия, связанными с порядковым номером делянки внутри блока. Следовательно, когда число делянок опыта равно квадрату числа вариантов, варианты могут быть не только *сбалансированы* в отношении повторений, но и *уравнены* в отношении порядкового места внутри повторения. В этом случае, по аналогии с латинским квадратом, следует ожидать уменьшения средней квадратической ошибки опыта даже после исключения соответствующего числа степеней свободы. Такое двойное элиминирование по рядам и столбцам будет особенно выгодным, когда повторения совпадают с размещением некоторого физического фактора пестроты плодородия, такого, как, например, последовательность вспашки, которая часто создает характерную периодичность в изменении плодородия в связи с изменением глубины пахотного слоя, дренажа и других подобных моментов.

Суммируем все сказанное. Следует избегать систематического расположения вариантов опыта, так как в этом случае оценка ошибки опыта может быть получена несколькими различными способами, каждый из которых дает совсем иные результаты и зависит от целого ряда допущений относительно распределения естественного плодородия. При случайном расположении вариантов без каких-либо ограничений ошибка опыта, хотя и может быть правильно определена, но все же обычно бывает слишком высокой. В хорошо спланированном опыте на случайное размещение делянок обычно наложено некоторое ограничение, причем так, что все еще остается возможность правильной оценки точности опыта и в то же время исключается большая часть неоднородности участка.

Следует подчеркнуть, что когда при помощи подходящего способа расположения вариантов мы снижаем среднюю квадратическую ошибку наполовину, то эффективность опыта возрастает

по крайней мере в четыре раза, так как только при четырехкратном повторении опыта в его первоначальной форме можно получить такую точность. Следует сказать, что эта аргументация скорей недооценивает значение таких точных опытов, так как в условиях сельскохозяйственного экспериментирования опыт нельзя повторить при неизменных условиях почвы и климата.

49.1. Ковариационный анализ

В предыдущем изложении было установлено, что точность опыта может быть значительно повышена путем выравнивания некоторых возможных источников ошибок, включенных в различия сравниваемых вариантов. Так, при делении участка, предназначенного для некоторого сельскохозяйственного опыта, на блоки, в каждом из которых в одинаковой мере представлены все варианты, различия в плодородии между разными блоками, которые могли бы быть одним из источников экспериментальной ошибки, исключаются из сравнений. Одновременно они исключаются из нашей оценки этой ошибки, получаемой при дисперсионном анализе. В латинском квадрате различия между рядами и между столбцами исключаются из сравнений и из ошибки, так что реальная и исчисленная точность таких сравнений становится такой, какой она была бы, если бы опытный участок имел одинаковый уровень плодородия по всем рядам и по всем столбцам.

Аналогичное этому выравнивание находит широкое применение во всех видах экспериментальной работы. Так, в опытах, в которых изучаются варианты кормления животных, скорость роста самцов и самок может быть различной, но в то же время оба пола могут быть в одинаковой мере пригодны для определения преимуществ одного рациона перед другим. В этом случае влияние пола на скорость роста может быть элиминировано из ошибки опыта путем выделения для каждого варианта одинаковой доли самцов и, чем часто пренебрегают, путем исключения средней разности между полами. Различная реакция на кормление часто наблюдается также и у разных пород животных; поэтому каждая из имеющихся пород должна быть одинаково представлена во всех вариантах опыта. В этом случае влияние пород будет исключено из сравнений вариантов и в соответствии с этим должно быть элиминировано при помощи дисперсионного анализа и из ошибки опыта. Иногда считают, что при эксперименте все животные должны быть обязательно одной и той же породы. Это при недостатке необходимого количества однородного материала не позволяет иметь достаточную повторность опыта. Приведенные выше примеры убеждают, что такое жесткое требование является излишним и не приводит к уточнению опыта. Действительно, не содействуя повышению точности, это направление вместе с тем определенно ограничивает область, к которой могут быть отнесены результаты опыта, так как результаты, относящиеся к нескольким породам, очевидно, приложимы к более широкому кругу условий, чем результаты, установленные только для одной породы. Работая с весьма выравненным материалом, мы подвергаемся определенной опасности прийти к выводам, которые будут весьма ненадежными, если к ним подходить как к выводам, покоящимся на широкой индуктивной основе.

Вместе с этим имеется множество факторов, влияющих на точность сравнений, которые, хотя и не могут быть полностью выравнены, однако могут быть учтены и с тем или иным основанием приняты в расчет. Такими факторами будут, например, возраст и вес животного; в частности, начальный вес имеет большое значение в опытах по изучению скорости роста при различных рационах питания. В полевых опытах с корнеплодами урожай делянок очень часто явным образом зависит от числа растений на делянках. Если в этом случае можно не считаться с влиянием изучаемых вариантов на густоту стояния растений, то лучше производить сравнения вариантов, приводя их к одинаковому числу растений. Далее, хотя мы не можем полностью выравнить плодородие делянок, предназначенных для различных вариантов, однако можно предварительно изучить опытный участок путем дробного учета, проведенного в предшествующий закладке опыта год. Урожаи делянок с таким предварительным учетом имеют определенное значение при изучении экспериментальных урожаев. Это положение, в частности, имеет большое значение в опытах с многолетними культурами, так как здесь остаются в какой-то мере постоянными не только почвенные условия, но и сами растения, выращенные при этих условиях. Такой предварительный учет позволяет повысить точность опыта и облегчает анализ результатов опыта, заложенного на новой и еще неиспользованной плантации.

Пример 46.1. *Ковариация урожаев чая в последовательные периоды.* Т. Иден приводит данные об урожае шестнадцати делянок чайного листа в опыте на Цейлоне, относящиеся к последовательным периодам, в каждом из которых произведено четырнадцать сборов чайного листа. Урожаи даны в процентах к средней для каждого периода, но процесс описываемой ниже обработки будет тем же, если даны не относительные, а абсолютные урожаи. Ниже (табл. 61.1 и 61.2) приводятся данные за второй и третий периоды, которые здесь будут рассматриваться соответственно как урожаи, определенные с помощью предварительного и экспериментального учетов. Эти шестнадцать делянок размещены в квадрате 4×4.

Допустим, что этот участок в период эксперимента отведен под опыт с четырьмя вариантами, поставленный по схеме латинского квадрата. Из 15 степеней свободы 6 степеней, соответствующие различиям между рядами и столбцами, будут элиминированы, а оставшиеся 9 должны быть подразделены на две части: одна, состоящая из 3 степеней свободы, относится к различиям

между вариантами и вторая, состоящая из 6 степеней свободы, предназначается для оценки ошибки опыта. Так как здесь фактические варианты отсутствуют, то для оценки ошибки опыта следует использовать все 9 степеней свободы. Результаты обработки условно экспериментальных урожаев приведены в табл. 61.3.

Таблица 61. 1

Предварительный поделяночный учет урожая чайного листа

88	102	91	88	369
94	110	109	118	431
109	105	115	94	423
88	102	91	96	377
379	419	406	396	1 600

Таблица 61. 2

Экспериментальные поделяночные урожаи чайного листа

90	93	85	81	349
93	106	114	121	434
114	106	111	93	424
92	107	92	102	393
389	412	402	397	1 600

Таблица 61. 3

Дисперсионный анализ для экспериментальных урожаев

	Степени свободы	Сумма квадратов	Средний квадрат
Ряды	3	1095,5	—
Столбцы	3	69,5	—
Ошибка	9	875,0	97,22
Итого . . .	15	2040,0	136,00

Даже после исключения очень большой дисперсии, относящейся к рядам, остаточная дисперсия все же еще велика и составляет 97,22. Следовательно, средняя квадратическая ошибка единичной делянки составляет примерно 9,86% от среднего урожая, а средняя квадратическая ошибка урожая варианта, определенного по четырем делянкам, равна примерно 4,93%.

Но если сравнить данные предварительного и экспериментального периодов, то легко увидеть, что большая часть этой вариации урожая в экспериментальном периоде обусловлена варьированием урожая в предыдущем периоде. При беглом просмотре соответствующих таблиц видно, что из девяти делянок, которые дали в экспериментальный период урожаи выше среднего, семь имели урожаи выше среднего и в предшествующем периоде. Если бы мы подобрали ряды, каждый состоящий из четырех делянок, так, чтобы в первый период суммарные урожаи этих рядов были бы одинаковыми, и отвели бы каждый такой ряд делянок для каждого отдельного варианта опыта, то тем самым мы смогли бы значительно снизить ошибки сравнений наших вариантов. Такое уравнивание суммарных урожаев в предшествующий период желательно, но оно редко осуществимо на практике по причинам, которые станут ясными из дальнейшего изложения. Общий смысл того положения, что ряды делянок, давшие одинаковые урожаи в предшествующий период, при одинаковой обработке в экспериментальном периоде должны дать одинаковые урожаи, заключается в предположении, что ожидаемый в последующем урожай некоторой делянки хорошо репрезентируется урожаем предыдущего периода при помощи линейной функции регрессии. Значение этого положения состоит в том, что поправки, соответствующие некоторой формуле регрессии (линейная форма которой обычно имеет наибольшее применение), могут быть введены в непосредственные результаты опыта без проведения в предварительном периоде какого-либо распределения делянок между будущими вариантами. Этот метод регрессий позволяет также освободиться от двух затруднений, которые неизбежны при уравнивании урожаев предшествующего периода. Первое затруднение состоит в том, что необходимость элиминирования между рядами и столбцами (или блоками) будет часто препятствовать такому уравниванию; второе обстоятельство состоит в том, что такое уравнивание, если оно и возможно, будет всегда только приблизительным и не вполне точным.

Поправка, вводимая в разность между урожаями двух делянок, у которых известны урожаи предыдущего периода, очевидно, является разностью урожаев этих делянок, которая ожидается в последующем периоде на основе различий одинаково обработанных делянок. Соответствующий коэффициент линейной регрессии получается как отношение ковариации к дисперсии независимой переменной, которая в этом случае является дисперсией урожаев предшествующего периода, относящейся при данном расположении вариантов к ошибке опыта. Для того чтобы определить эту дисперсию независимой переменной, следует урожаи предшествующего периода подвергнуть тому же самому анализу, какой применялся к урожаям экспериментального периода (табл. 61.4). Третья таблица, предназначенная для анализа ковариации урожаев во время предварительного и экспериментального периодов, строится по той же системе, но вместо квадратов урожаев теперь берутся произведения урожаев этих двух периодов:

Таблица 61. 4

Анализ урожаев предварительного учета

	Степени свободы	Сумма квадратов
Ряды	3	745,0
Столбцы	3	213,5
Ошибка	9	567,5
Итого	15	1526,0

При образовании этих двух таблиц применяется один и тот же арифметический процесс. Чтобы это было ясным, рассмотрим образование величин, которые во всех трех таблицах стоят под рубрикой «столбцы». В табл. 61.1 и 61.2 средняя из итогов по столбцам составляет 400; обозначим отклонения итогов отдельных столбцов от этой средней через x и y; так, в табл. 61.1 отклонение первого столбца $x=-21$, а в табл. 61.2 отклонение этого столбца $y=-11$. Выпишем квадраты и произведения каждой пары в виде таких параллельных рядов.

Таблица 61. 5

x^2	xy	y^2
441	+231	121
361	+228	144
36	+12	4
16	+12	9
854	+483	278

Разделив каждый из этих итогов на 4 (т. е. на число делянок, объединенных в каждом из этих итогов), мы получим соответствующие данные для рубрики «столбцы» в следующей тройной таблице:

Суммы квадратов и произведений *Таблица 61. 6*

	Степени свободы	x^2	xy	y^2
Ряды	3	745,0	837,0	1095,5
Столбцы	3	213,5	120,75	69,5
Ошибка	9	567,5	654,25	875,0
Итого	15	1526,0	1612,00	2040,0

В параллельных столбцах этой таблицы дан анализ дисперсий и ковариации двух переменных x и y.

Связь этих переменных, которая может быть выражена в форме регрессии или корреляции, в данном случае определяется независимо по каждой отдельной строке. В частности, нас особенно интересует отношение 654,25/567,5, представляющее регрессию y и x для однородных делянок, полученных в результате элиминирования различий между рядами и между столбцами. Это отношение, очевидно, и является той поправкой, которую следует ввести в экспериментальные урожаи на каждую единицу отклонения соответствующего x от средней.

Эта поправка может быть применена к отдельным делянкам или к составным суммам, представляющим ряды, столбцы или варианты. Результат введения поправок и дисперсионный анализ таких исправленных урожаев можно получить и непосредственно из представленного выше анализа сумм квадратов и произведений. Так, если обозначить коэффициент регрессии через b, то сравнение исправленных урожаев по своему существу будет сравнением разностей $y - bx$. Кроме того, можно написать

$$(y - bx)^2 = b^2x^2 - 2bxy + y^2.$$

Следовательно, для определения суммы квадратов исправленных урожаев по той или другой строке табл. 61.6 нужно только умножить данные этой таблицы соответственно на b^2, $-2b$ и 1 и сложить эти произведения.

В настоящем примере $b=1{,}1529$; $b^2=1{,}3291$, откуда:

Таблица 61. 7

Анализ исправленных урожаев

	Степени свободы	Сумма квадратов	Средний квадрат
Ряды	3	155,8	51,93
Столбцы	3	74,8	24,93
Ошибка	8	120,7	15,09
Итого	14	351,3	25,09

Следует отметить, что общее число степеней свободы здесь с 15 снизилось до 14 в связи с тем, что по этим данным была определена одна константа в формуле регрессии. Эта одна степень свободы вычитается по той частной строке, по которой определялось численное значение коэффициента регрессии b, исходя из равенства

$$bS(x^2) = S(xy).$$

Следовательно, по этой строке мы имеем

$$S(y-bx^2)=S(y^2)-\frac{S^2(xy)}{S(x^2)},$$

откуда вытекает, что данные по этой строке теряют одну степень свободы. По другим строкам в связи с введением поправок данные могут как уменьшиться, так и увеличиться.

Величина b, применяемая для исправления урожаев вариантов, является статистическим показателем, на который влияют ошибки случайного отбора. Следовательно, хотя величины $y-bx$ являются вполне подходящими оценками исправленных урожаев, однако они, как это было показано в параграфе 26, все же имеют варьирующую точность. Поэтому суммы квадратов исправленных урожаев по той строке, по которой величина b не вычислялась, не могут лежать в основе оценки существенности отклонений от регрессии.

В нашем случае такой критерий существенности потребовался бы, если бы в опыте изучались действительно различающиеся друг от друга варианты; в этом случае для каждой переменной x и y мы имели бы 3 степени свободы для вариантов и только остальные 6 — для ошибки. Именно по этим 6 степеням свободы и должна быть вычислена величина b. В нашем примере нет фактических вариантов, и поэтому покажем применение критерия существенности, применив его к рядам, хотя различия между ними фактически не являются следствием эксперимента.

Взяв из табл. 61.7 те части, которые относятся только к рядам и ошибке, вычислим новые значения сумм квадратов зависимой переменной y соответственно для ошибки и итога путем вычитания в каждом случае из $S(y^2)$ величины

$$\frac{S^2(xy)}{S(x^2)},$$

рассчитанной для данной строки. Для ошибки это дает, как показывает табл. 61.7, уменьшенную сумму квадратов 120,7; для итога же мы имеем 1970,5—1694,3=276,2 и соответственно этому 11 степеней свободы. Вычитая первое число из второго, находим сумму квадратов 155,5, приписываемую 3 степеням свободы рядов. Эта величина сравнима с уменьшенной суммой квадратов ошибки и может быть положена в основу точного критерия существенности. В целом этот процесс расчета представлен в табл. 61.71 (см. также табл. 59.1). Из нашего примера видно, что сумма квадратов значений $y-bx$ указывает на хорошее приближение к линии регрессии. Однако эта сумма в известной, а иногда и в значительной мере испытывает на себе влияние выборочных ошибок коэффициента b. Тем не менее не представляет никаких затруднений применить в данном случае вполне точный критерий, учитывающий именно эти выборочные ошибки.

Таблица 61. 71

Критерий существенности, основанный на остаточной дисперсии

	Степени свободы	x^2	xy	y^2	Степени свободы	Остаточное y^2	Средний квадрат
Ряды . .	3	745,0	837,0	1095,5	3	155,5	51,83
Ошибка	9	567,5	654,25	875,0	8	120,7	15,09
Итого	12	1312,5	1491,25	1970,5	11	276,2	—

Сравнивая этот анализ исправленных урожаев с тем, который был ранее получен без учета предыдущих сборов чайного листа, можно видеть, что средний квадрат на делянку снизился с 97,22 до 15,09; несмотря на уменьшение числа степеней свободы, точность опыта здесь выросла примерно в 6 раз. Следует отметить также и второй момент: большое различие между урожаями разных рядов, которое было обнаружено в первоначальном анализе, теперь снизилось до одной седьмой своей первоначальной величины. Это указывает на то, что при благоприятных условиях эта часть почвенной пестроты может быть элиминирована путем учета данных предварительного учета. Однако это ни в коем случае не уменьшает значения элиминирования различий между большими частями опытного участка, такими, как блоки, ряды, столбцы и т. д., даже при наличии предварительного учета. Действительно, в нашем случае элиминирование рядов и столбцов имеет большее значение, когда оно при исправленных урожаях снижает средний квадрат с 25,09 до 15,09, чем тогда, когда оно при неисправленных урожаях снижает средний квадрат с 136 до 97,2. Если, например, взять эксперимент с 10%-ной ошибкой средней за условную единицу познавательной ценности эксперимента, то элиминирование рядов и столбцов при неисправленных урожаях повысит эту ценность опыта только с 2,94 до 4,12, т. е. даст выигрыш в 1,18 условных единиц; если же то же самое элиминирование произвести в отношении исправленных урожаев, то ценность эксперимента возрастет с 16,61 до 26,51 единиц, т. е. получится выигрыш в 9,90 условных единиц, или примерно в 9 раз больше. Такие сравнения количеств реализуемой информации практически следует проводить, особенно при небольшом числе степеней свободы, с учетом числа степеней свободы, как это было сделано в табл. 59.2.

Знакомство с процессом анализа, примененным в данном примере, позволяет установить, что в нем объединены преимущества и требования двух широко известных методов, таких, как регрессия и дисперсионный анализ. Если хорошо усвоен способ построения таблицы ковариационного анализа, то не представляет никаких затруднений распространить этот метод на три и более пере-

менных, построить ряд соответствующих ковариаций и тем самым принять в расчет одновременно два или более учитываемых, но не контролируемых условий, сопутствующих нашим наблюдениям. Эти наблюдения в таком случае рассматриваются в качестве значений зависимой переменной, варьирование которой может быть частично выражено через изменчивость сопутствующих условий. Так, если мы намерены изучить влияние некоторых агротехнических мероприятий на чистый выход сахара, извлекаемого из сахарной свеклы, то переменные, которые будут определять сопутствующую вариацию, будут: а) процентное содержание сахара и б) вес корней. Анализ ковариации, примененный к этим трем переменным, — чистый выход, процент сахаристости, вес корней, учтенных по отдельным делянкам, — даст нам возможность изучить влияние испытываемых мероприятий только на один признак — чистый выход сахара, т. е. исключив всякое влияние на этот признак веса корней и концентрации сахара в свекле, и все это без фактического уравнивания этих двух показателей по вариантам опыта.

В исследованиях такого рода открывается даже возможность элиминировать, например, не просто средний вес корней на делянке, но также и квадрат веса, используя в этом случае нелинейную регрессию на вес корня. Далее, если исследователь имеет данные не только о среднем весе корней, но и отдельные их значения, по которым исчислена эта средняя, то он может получить данные о чистом выходе сахара с поправками, которые адекватны уравниванию по всем делянкам как среднего веса, так и дисперсии веса корней.

Рассматривая результаты анализа данных, в которых внесены поправки указанного вида, можно установить влияние этих поправок на остаточную дисперсию .Так, если в примере 46.1 сравнить табл. 61.3 и 61.7, то можно видеть, что 9 степеней свободы ошибки при неисправленных урожаях в результате введения поправок делятся на две части, из которых одна содержит в себе 1 степень свободы, относящуюся к уравнению регрессии, а вторая, состоящая из 8 степеней свободы, остается после введения поправок в урожаи, основанных на уравнении регрессии. Этот анализ ошибки приведен в табл. 61.8.

Большой эффект введения поправок здесь обусловлен тем, что у 1 степени свободы, относящейся к уравнению регрессии, сосредоточивается основная часть остаточной суммы квадратов. В данном случае роль регрессии столь велика, что нет необходимости оценивать ее существенность перед тем, как вводить соответствующие поправки в урожаи вариантов, но такая оценка была бы необходимой, если бы существенность регрессии не была бы столь очевидной.

Хотя ковариационный анализ дает, как мы видели, возможность извлекать из имеющегося объема данных как можно больше сведений, однако его основное значение состоит в том, что он

Таблица 61.8

Анализ остаточной дисперсии

	Степени свободы	Сумма квадратов	Средний квадрат
Регрессия	1	754,3	754,3
Ошибка исправленных урожаев	8	120,7	15,09
Ошибка неисправленных урожаев . . .	9	875,0	—

дает возможность планировать программу наблюдений и отбора из множества сопутствующих условий, таких, которые надлежит учитывать. Так, на примере опыта с чаем было показано, что в данном случае ценность плантации для экспериментальных целей была увеличена в шесть раз за счет сравнительно небольшой дополнительной работы по учету урожая по отдельным делянкам в предшествующий опыту период. В опытах с однолетними культурами введение предварительного учета примерно удваивает работу по проведению опыта; но в данном случае особенно важно то, что обработка почвы, производимая в экспериментальный год заново, вызывает сомнение относительно пригодности данных предварительного учета. Ковариационный анализ урожаев последовательных лет, примененный к дробному учету однолетних культур, показывает, что хотя при учете урожая в предшествующем опыту году и повышается ценность эксперимента, но это повышение редко превосходит 60%. Отсюда следует, что опыт с однолетними культурами выгоднее произвести при увеличенной в два раза повторности на неизученном предварительно участке, чем растягивать его во времени и затрачивать работу на предварительное изучение пестроты плодородия опытного поля.

Следует отметить, что в опытной работе часто остаются неиспользованными многие возможности значительного повышения точности опыта при помощи сравнительно простых дополнительных наблюдений и что такие возможности достаточно обоснованного использования дополнительных наблюдений мало знакомы широкому кругу исследователей. В то же время в этой области весьма велики возможности значительного усовершенствования методов экспериментирования путем повышения точности или пропорционального уменьшения трудовых затрат.

Ковариационный анализ, помимо соотношений между зависимой и независимой переменными, включает в себя некоторую первоначальную классификацию (по блокам, вариантам и пр.). Иногда эта классификация может быть сложной, как например при последовательной группировке, производимой при помощи трех или большего числа следующих друг за другом подразделений.

Точно так же возможно наличие не одной, а нескольких зависимых переменных. Примеры, относящиеся к этим усложнениям, а также многие детали такого материала можно найти в библиографии (см. работу автора совместно с Деем, 1937 г.).

49.2. Установление различий групп на основе ряда показателей

Одним из имеющих большое практическое значение приложений системы расчетов, применяющихся во множественной регрессии, является отыскание из числа всех возможных линейных соединений нескольких показателей такого, который наилучшим образом характеризует различие между двумя группами объектов. Например, нижняя челюсть или челюстная кость человека может быть найдена при таких обстоятельствах, что пол человека, которому принадлежала эта кость, нельзя установить иначе, как на основе изучения особенностей этой кости. В той мере, в какой это при данных обстоятельствах возможно, антрополог ставит перед собой задачу с достаточной достоверностью определить пол человека, к которому относится такая находка.

Если он имеет некоторое число челюстей, относящихся к людям известного пола, то ряд показателей, полученных при их измерении, может дать ключ к решению вопроса о поле. Некоторые из этих показателей будут у разных полов довольно резко различаться по своей величине, но, как это чаще всего бывает, они будут находиться друг с другом в определенной связи и поэтому не могут считаться самостоятельными и независимыми друг от друга. По этой же причине другие признаки, которые сами по себе не могут служить средством распознания полов, все же могут с учетом их связи с другими показателями оказать свою помощь при исследовании этого вопроса. Только после того, как будет определена некоторая линейная функция этих показателей, которая лучше, чем всякая другая линейная функция, определяет различие нижних челюстей у двух полов, можно установить, что некоторые показатели бесполезны, в то время как другие имеют реальное значение при определении пола человека по его челюсти.

Чтобы показать формальную тождественность этого метода с методом множественной регрессии, допустим, что имеется N_1 мужских и N_2 женских нижних челюстей, по каждой из которых определены показатели $x_1, x_2, \ldots x_p$. Средние разности этих показателей (мужская челюсть — женская челюсть) обозначим через $d_1, d_2, \ldots, d_p$. Возьмем далее суммы квадратов и произведения этих показателей, игнорируя пол, и обозначим

$$S_{ij} = S(x_i - \bar{x}_i)(x_j - \bar{x}_j).$$

В этом случае можно показать, что корни $b_1, b_2, \ldots b_p$ таких уравнений:

$$\begin{aligned} S_{11}b_1 + S_{12}b_2 + \ldots + S_{1p}b_p &= d_1 \\ S_{1p}b_1 + S_{2p}b_2 + \ldots + S_{pp}b_p &= d_p \end{aligned}$$

будут пропорциональны коэффициентам такого линейного уравнения:

$$X = b_1x_1 + b_2x_2 + \ldots + b_px_p.$$

Это уравнение, поскольку об этом позволяют судить имеющиеся данные, будет наиболее полно определять различие между челюстями двух полов.

Если мы введем некоторую условную y, равную $\frac{N_1}{N_1+N_2}$ для всех мужских челюстей, $\frac{-N_1}{N_1+N_2}$ — для всех женских челюстей, то уравнения для коэффициентов множественной регрессии y на $x_1, x_2, \ldots x_p$ будут фактически отличаться от уравнений, приведенных ранее, только множителями $\frac{N_1N_2}{N_1+N_2}$ в правой части. Следовательно, коэффициент множественной регрессии y на $x_1, x_2, \ldots x_p$ будет определяться формулой

$$R^2 = \frac{N_1N_2}{N_1+N_2}(b_1d_1 + \ldots + b_pd_p).$$

Хотеллинг (1931 г.) показал, что если переменные x внутри групп распределены нормально, то существенность этой корреляции может быть определена при помощи дисперсионного анализа, в котором отдельным частям соответствуют p и $n - p + 1$ степеней свободы, причем здесь n — число степеней свободы внутри групп. Таким образом,

$$e^{2z} = \frac{n-p+1}{p} \cdot \frac{R^2}{1-R^2}.$$

Это выражение и служит основой для определения, является ли существенным то или иное различие между группами. Может возникнуть также задача, в которой требуется определить, будет ли некоторая гипотетическая дискриминантная функция

$$X' = \beta_1x_1 + \ldots + \beta_px_p,$$

определяемая отношениями между коэффициентами, но не их абсолютными значениями, совместимой с наблюденными фактами. Можно показать, что решение этого вопроса сводится просто к определению коэффициента корреляции между X и X' внутри групп. Если этот коэффициент корреляции обозначить через r, то значение R^2 в критерии Хотеллинга следует умножить на $(1 - r^2)$ и, как обычно, уменьшить на одну степень свободы ту часть, которая относится к оцениваемой специфической форме дискриминантной функции.

Теперь значение z может быть определено из формулы:

$$e^{2z} = \frac{n-p+1}{p-1} \cdot \frac{R'^2}{1-R'^2},$$

где $R'^2 = R^2(1 - r^2)$, а степени свободы соответственно равны: $n_1 = p - 1$ и $n_2 = n - p + 1$. Таким образом, этот критерий отвергает любую предложенную дискриминантную формулу, дающую r

столь малое, что определенная выше величина z становится несущественной.

Этот метод может быть распространен и на случай регрессии средних нескольких выборок на некоторую переменную, связанную с этими выборками. Так, Бернард применил регрессии средних некоторых показателей у черепов египтян на примерную дату захоронения для установления того, какая из линейных функций показателей черепа обнаруживает наиболее явственное изменение с течением времени. Весьма важное применение этого метода в селекции растений дал Смит, определив, какие из характеристик исходного растения следует комбинировать для получения некоторого частного результата.

Если в некотором сортоиспытании, поставленном с определенной повторностью, признаки $x_1, x_2, \ldots . x_p$ учитываются на каждой отдельной делянке, то можно получить суммы квадратов и произведений, во-первых, для сортов, которые мы обозначим через t_{ij}, и, во-вторых, для ошибок e_{ij}. Вычитая вторые из первых, мы получим несмещенные оценки сортовых эффектов $g_{ij} = t_{ij} - e_{ij}$. Если теперь определить достоинства некоторого сорта, для которого значения $x_1, x_2, \ldots x_p$ известны, при помощи формулы

$$a_1 x_1 + a_2 x_2 + \ldots + a_p x_p,$$

где коэффициенты a могут быть как положительными, так и отрицательными, то мы можем вычислить

$$A_i = a_1 g_{1i} + a_2 g_{2i} + \ldots + a_p g_{pi}$$

для каждого i. Соответствующие коэффициенты $b_1, b_2, \ldots b_p$ для установления селекционных достоинств некоторого сорта в этом случае будут определяться из такой системы уравнений:

$$b_1 t_{11} + \ldots b_p t_{1p} = A_1$$
$$b_1 t_{12} + \ldots b_p t_{2p} = A_2$$

и т. д.

Решая эти уравнения, мы можем установить ценность каждого сорта при помощи сводного показателя

$$X = b_1 \bar{x}_1 + b_2 \bar{x}_2 + \ldots + b_p \bar{x}_p,$$

где X будет функцией наблюдаемых показателей, наилучшим образом характеризующей действительную ценность сорта.

Предшествующие примеры являлись иллюстрацией общего принципа, согласно которому мы определяем ряд дополнительных коэффициентов таким образом, чтобы получить максимальное отношение одного из выбранных компонентов, полученного в результате анализа дисперсии, к сумме квадратов других компонентов этого анализа. Этот же принцип может быть применен к приведению к максимуму отношения, в котором сравнивается сумма квадратов для n_1 степеней свободы с суммой квадратов остаточных степеней свободы. После введения поправки при установлении максимального отношения, включающего в себя p дополнительных констант, мы должны оценить существенность $n_1 + p$ степеней свободы по сравнению с $n_2 - p$ степеней свободы.

Когда при отыскании максимума такого отношения берется только один компонент, то соответствующие уравнения линейны и может быть применен метод множественной регрессии. Другие случаи могут привести к уравнениям более высоких степеней. Так, располагая таблицей с двумя входами для неколичественных наблюдений, мы можем задаться вопросом, какие значения или условные величины следует им приписать для того, чтобы эти наблюдения дополняли друг друга в той мере, в какой это возможно.

Пример 46.2. *Построение системы взаимно-дополняющих показателей на основе серологических данных.* Результаты, приведенные в табл. 61.9 (данные Тейлора, Гальтоновская лаборатория), относятся к изучению реакции двенадцати проб крови человека на двенадцать различных сывороток. Реакции обозначены пятью символами: —, ?, w, (+) и +.

Таблица 61.9

Таблица неколичественных серологических показаний

№ сыворотки	1	2	3	4	5	6	7	8	9	10	11	12
1	w	w	w	(+)	w	(+)	?	w	w	(+)	w	w
2	?	w	?	w	w	w	?	w	w	w	w	?
3	w	w	w	w	w	w	w	w	w	w	w	w
4	w	w	w	w	w	w	—	w	w	w	w	?
5	w	(+)	w	(+)	w	w	?	(+)	w	(+)	w	w
6	w	w	(+)	(+)	w	w	?	w	w	(+)	w	w
7	(+)	(+)	(+)	(+)	+	+	w	(+)	w	(+)	(+)	w
8	w	+	(+)	(+)	w	(+)	w	(+)	w	(+)	(+)	w
9	w	(+)	(+)	(+)	w	(+)	w	(+)	w	(+)	w	w
10	?	?	w	w	w	w	?	w	w	w	w	?
11	w	w	(+)	w	w	w	?	w	w	w	w	w
12	w	(+)	+	(+)	(+)	(+)	w	(+)	w	(+)	+	w

Если ввести условные численные значения 0 для символа —, 1 для символа +, то значения для символов ?, w и (+) могут быть представлены буквами x, y и z. Подсчитывая после этого числа различного рода символов в каждом ряду и столбце, можно определить сумму квадратов, относящуюся к рядам и столбцам, которую удобно представить в виде симметричной квадратной матрицы 4×4:

Таблица 61.901

Матрица для рядов и столбцов

	x	y	z	1
x	718	2	−672	−106
y	2	1 630	−1 416	−218
z	−672	−1 416	1 944	216
1	−106	−218	216	118

Таким, образом, коэффициент при x^2 равен 718, в то время как коэффициенты при xy и yx оба равны 2, а соединение двух членов вместе дает $4xy$. Чтобы избавиться от дробей, умножаем на 144. Подобно этому, общая сумма квадратов для 143 степеней свободы находится из матрицы:

Таблица 61.902

Матрица для итога

	x	y	z	1
x	1 703	−1 157	−468	− 65
y	−1 157	4 895	−3 204	−445
z	−468	−3 204	3 888	−180
1	−65	− 445	−180	695

Чтобы найти такие значения x, y и z, которые должны дать по возможности высокое отношение первого из этих выражений ко второму, необходимо решить уравнение 4-й степени. Если из каждого элемента первой матрицы вычесть соответствующий элемент второй, умноженный на Θ, то приравнивание полученного детерминанта нулю даст соответствующее уравнение. Следовательно, задача сводится к нахождению решения этого уравнения.

Вычисление коэффициентов этого уравнения не является обязательным. Обычно более удобно производить отыскание значения Θ путем подбора с последующим уточнением при помощи метода конечных разностей. В следующей таблице даны значения детерминанта, для упрощения разделенные на 3456, которые получены для шести значений Θ:

Таблица 61.91

Значения детерминанта и их конечные разности

Θ	Детерминант	Первая разность	Вторая разность	Третья разность	Четвертая разность
0	429 106				
0,2	49 982,376	−1 895 618,12			
0,4	−598 370,560	−3 241 764,68	−3 365 366,4		
0,6	−1 536 668,072	−4 691 487,56	−3 624 307,2	−431 568	
0,8	4 123 941,552	28 303 048,12	82 486 339,2	143 517 744	179 936 640
1,0	30 181 877	130 289 677,24	254 966 572,8	287 467 056	179 936 640

Второй столбец определяется путем деления последовательных разностей первого столбца на 0,2, т. е. на интервал между соседними значениями Θ; третий столбец подобным же образом найден по данным второго столбца, причем делителем здесь будет разность значений Θ, взятых через интервал, что в данном случае составляет 0,4. Так как у любого выражения 4-й степени четвертые конечные разности константы, то этим можно воспользоваться для проверки расчетов, взяв достаточное число значений Θ.

Из табл. 61.91 видно, что искомое значение Θ лежит между 0,6 и 0,8. Так как четвертые разности константны как при равных, так и неравных, а также как при положительных, так и отрицательных интервалах, то решение детерминанта можно свести к дальнейшему подбору значений Θ так, чтобы он оказался равным нулю с достаточной для нас точностью. Так, для вычисления его при 0,7 добавим новую строку, в которой третья разность повышается на константную четвертую разность, умноженную на 0,3; этот множитель является просто разностью между этим новым значением $\Theta=0,7$ и четвертым, начиная снизу, значением в табл. 61.91, т. е. $\Theta=0,4$. Эта новая третья разность умножается на 0,1, т. е. на разность с третьим снизу значением Θ, и прибавляется ко второй разности. Множитель, на который умножается эта новая вторая разность, будет $-0,1$, так как $\Theta=0,7$ на 0,1 меньше второго снизу значения $\Theta=0,8$. Наконец новая первая разность умножается на $-0,3$ и прибавляется к значению детерминанта при 1,0, что и дает значение этого детерминанта при $\Theta=0,7$. В табл. 61.92 дан расчет этой строки с принятой ранее точностью. Эту точность следует сохранить и в последующих строках таблицы. Отметим, что значение Θ в третьей строке отличается от точного значения только на 3 единицы в пятом знаке после запятой, тем не менее имеется, как мы видим, возможность продолжать работу по уточнению Θ. Задаваясь более низкой точностью, можно в каждом столбце брать меньшую точность, да и сам процесс вычислений может иметь меньшее число этапов. Однако при наличии счетной машины работа и с большим числом

знаков не вызывает затруднений, в то же время эта схема расчетов позволяет избежать алгебраических преобразований детерминанта.

Определенное таким путем значение Θ является фактически долей, которую составляет сумма квадратов, приписываемая рядам и столбцам в общей сумме квадратов, при том условии, что эта доля приведена к максимуму.

Таблица 61.92

Этапы решения алгебраического уравнения методом конечных разностей. Четвертая разность везде равна 179 936 640

Θ	Детерминант	Первая разность	Вторая разность	Третья разность
1,0	30 181 877	130 289 677,24	254 966 572,8	287 467 056
0,7	−231 684,844	101 378 539,48	289 111 377,6	341 448 048
0,708	−18 463,0046	26 652 729,92	255 910 306,7	360 881 205
0,70869	+860,1817	28 004 617,84	155 568 230,5	344 451 190
0,7086593	−2,7600	28 108 851,19	158 096 991,4	292 028 323
0,7086593982	−0,0002	28 104 007,21	158 290 581,8	293 586 466

Чтобы найти значения соответствующих условных показателей x, y и z, следует умножить матрицу итоговых сумм квадратов на Θ и вычесть это произведение из матрицы для рядов и столбцов. В результате получится следующая система уравнений:

$$\begin{aligned} -488{,}8470\,x + 821{,}9189\,y - 340{,}3474\,z &= 59{,}9371 \\ 821{,}9189\,x - 1838{,}8878\,y + 854{,}5447\,z &= -97{,}3534 \\ -340{,}3474\,x + 854{,}5447\,y - 811{,}2677\,z &= -343{,}5587 \end{aligned}$$

Отсюда находим решения:

$$\begin{aligned} x &= 0{,}192959 \\ y &= 0{,}584453 \\ z &= 0{,}958163, \end{aligned}$$

соответствующие символам ?, w и (+) при условии, что символу «—» соответствует нуль, а символу «+» соответствует единица. Рассматривая эти данные (где следует считаться только с двумя первыми цифрами), можно видеть, что все они лежат между 0 и 1, в порядке возрастания изучаемой реакции. Такой результат отнюдь не является следствием способа, принятого для определения этих характеристик, а полностью определяется особенностями изучаемого материала.

Оценка существенности различий по строкам и столбцам может быть произведена непосредственно по значению Θ, т. е. без

численного определения условных показателей x, y и z, ибо для этой оценки надо знать только отношение между суммами квадратов, которое и определяется величиной Θ. Вычисление x, y и z приводит к переходу 3 степеней свободы, соответствующих этим величинам, из остатка к тем 22 степеням свободы, которые соответствуют рядам и столбцам.

Таблица 61.93

Дисперсионный анализ для данных таблицы 61.9

	Степени свободы	Сумма квадратов	Средний квадрат	$^1/_2 \log_e$
Ряды и столбцы	25	0,70866	0,028346	1,6723
Остаток	118	0,29134	0,002469	0,4519
Итого	143	1,00000	—	$z = 1{,}2204$

Таким образом, различия между рядами и между столбцами явно существенны. Поэтому можно сделать вывод о том, что здесь имеются большие различия в силе действия сывороток или в силе реакции различных образцов крови. Эта оценка и вывод из нее достаточно важны, ибо только в этом случае определенные выше условные показатели приобретают практическое значение.

49.3. Точность оценки условных показателей

Числовые значения условных показателей, конечно, подчинены выборочным ошибкам. Однако нельзя определить ошибку каждого отдельного условного показателя, так как этот отдельный показатель нельзя взять без его связи с другими такими же показателями, составляющими систему, в которой значениям двух условных показателей придан произвольный размер. Это затруднение может быть преодолено только путем построения критерия для оценки того, в какой мере вся система условных показателей существенно отличается от некоторой заданной системы условных показателей. Так, оставляя значение 0 для символа —, мы можем задаться значениями 0,25; 0,50; 0,75 и 1,00 или равносильными значениями 1, 2, 3 и 4 соответственно для символов ?, w, (+) и +. В этом случае критерий существенности полностью аналогичен критерию, который применялся ранее для оценки того, будет ли некоторая система показателей, взятая в целом, определять существенность различий между рядами и столбцами.

Чтобы построить такой критерий, относящийся к тому типу критериев, которые основываются на данных, составляющих не одну, а несколько таблиц, мы введем новую переменную ξ. В этом случае в табл. 61.901 и 61.902 можно образовать новый столбец

путем умножения имеющихся там четырех столбцов на 1, 2, 3 и 4 с последующим суммированием этих произведений. В результате получится два ряда таких величин:

Ряды и столбцы	Итого
—1 718	—2 275
—1 858	—2 759
3 192	4 068
578	1 285

. Если мы умножим эти четыре строки соответственно на 1, 2, 3 и 4 и эти произведения сложим, то получим дисперсионный анализ для ξ. Если же умножить эти же строки на систему ранее определенных условных показателей, то получится ковариационный анализ для X и ξ, где X обозначает вышеуказанную систему условных показателей. Этим же путем строится и дисперсионный анализ для X. В результате этих расчетов получаем:

Таблица 61.94

Дисперсионный и ковариационный анализ гипотетических и эмпирических значений условных показателей

	$S(\xi^2)$	$S(\xi X)$	$S(X^2)$
Ряды и столбцы	6 454	2219,039	770,496
Остаток	3 097	912,280	316,762
Итого . .	9 551	3131,319	1087,258

Если теперь, в соответствии с общими принципами анализа, мы элиминируем ξ по строкам «остаток» и «итог», для чего вычтем из $S(X^2)$ квадрат $S(\xi X)$, деленный на $S(\xi^2)$, то после этого путем вычитания из результата, полученного для «итога», результата, полученного для «остатка», можно определить результат для строки «ряды и столбцы».

Здесь число степеней свободы для рядов и столбцов уменьшено на единицу, так как после элиминирования свободно изменяться могут только два из трех условных показателей. Значение z превосходит 20%-ный уровень существенности, но находится ниже 5%-ного уровня. Следовательно, наша таблица наблюдений при избранных условных значениях для символов «—» и «+» не противоречит существенно типотезе о том, что система условных показателей составляет линейную серию.

В данной таблице дисперсионного анализа, как и в табл. 61.93, критерий z не является вполне точным, хотя его приближение

Таблица 61.95

Дисперсионный анализ эмпирических условных показателей после элиминирования гипотетических их значений

	Степени свободы	Сумма квадратов	Средний квадрат	$^1/_2$ log
Ряды и столбцы	24	12,615	0,5256	0,8297
Остаток	118	48,033	0,4071	0,6982
Итого . .	142	60,648	z	0,1315

в обоих этих случаях достаточно для ответа на вопрос, являющийся предметом обсуждения. В табл. 61.93 распределение доли общей суммы квадратов, которая оказалась равной 0,70866, зависит от трех параметров при 22 и 121 исходных степенях свободы и при 4 степенях свободы, соответствующих 4-й степени уравнения, определяющего Θ, что на единицу больше числа условных показателей. В табл. 61.95 соответствующее отношение 0,2080 подобным же образом зависит от чисел 22, 120 и 3. Общее решение этой проблемы уже найдено, но пока нет соответствующих точных таблиц, пригодных для практического использования.

Метод, изложенный в этом параграфе, приложим к весьма широкому кругу практических задач. Часто бывает так, что статистик имеет дело с объединенными данными, которые требуется разместить по различным категориям. Так, например, мы можем иметь данные о расходной части бюджета в различных семьях, но без указания о том, как эти расходы распределены между мужем и женой или между детьми различного возраста. Если расходы каждой семьи известны, то удельный вес каждого класса потребителей может быть определен путем приведения к минимуму отклонения между фактическими расходами и расходами, ожидаемыми на основе состава семьи. В тех случаях, когда имеют дело с непрерывными переменными, например с возрастом человека, то для каждого возраста не представляется возможным ввести самостоятельный условный показатель, но в этих случаях все же можно ввести показатель возраста, если в качестве независимой переменной взять его квадрат, а если возможно, то и куб или более высокие степени его, подобно тому, как это было при отыскании криволинейной регрессии. Так, Дей из Службы лесоводства Соединенных Штатов добился успеха при распределении стоимости перевозки бревен различного диаметра по данным о составе и о стоимости доставки каждой из нескольких партий бревен. В этом случае для кривой стоимости было признано достаточным квадратное уравнение.

ГЛАВА ДЕВЯТАЯ

ОСНОВНЫЕ ПРИНЦИПЫ СТАТИСТИЧЕСКИХ ОЦЕНОК

50. Практическое значение теоретически обоснованных методов проведения статистических оценок, с одной стороны, и широкое использование в статистической практике показателей, которые в свете положений параграфа 3 являются неэффективными статистиками, с другой стороны, настоятельно требуют, чтобы исследователь при интерпретации своих результатов и при изучении результатов других экспериментаторов умел видеть различие между заключениями, соответствующими природе наблюдаемых фактов, и такими выводами, которые обусловлены просто применением неправильных методов оценки.

Пример 47. В качестве примера, в котором легко выявляются основные принципы теории оценок и который не требует такого объема данных, который затруднял бы показ применения различных методов, рассмотрим здесь оценку генетической связи у потомства самоопыляющихся гетерозиготов. Для двух генетических факторов кукурузы («Крахмальная» и «Сахарная») и цвета основных листьев («Зеленый» или «Белый») имеются такие данные подсчета ростков (данные Карвера):

Таблица 62

„Крахмальная"		„Сахарная"		Итого
„Зеленый"	„Белый"	„Зеленый"	„Белый"	
1 997	906	904	32	3 839

51. Существенность проявления связи

Прежде чем производить статистическую оценку такого факта, как интенсивность связи между генами, следует сначала установить вообще, имеет ли место то, что подлежит такой оценке. Поэтому нам прежде всего надлежит определить, не встречаемся ли

мы здесь с простой независимостью наследования этих двух факторов. Если бы здесь встретился именно этот случай, то каждый из факторов разделился бы в отношении 3 : 1, а в целом мы имели бы четыре комбинации с отношениями 9 : 3 : 3 : 1, т. е. с такими ожидаемыми численностями, которые даны в табл. 63, где приведен и соответствующий расчет величины χ^2.

Таблица 63

Ожидаемые численности (m)	2159,4	719,8	719,8	239,9	—
Разности (d)	− 162,4	+186,2	+184,2	−207,9	—
$\frac{d^2}{m}$	12,21	48,17	47,14	180,17	287,69

Так как для 3 степеней свободы при 1-%ном уровне существенности величина χ^2 равна только 11,34, то следует признать, что фактические численности явно не согласуются с теоретически ожидаемыми. Однако этот результат может появиться под влиянием двух причин: или он обусловлен наличием связи между генами, или отклонением от теоретического отношения 3 : 1. Этот вопрос можно выяснить при помощи специального разложения величины χ^2 на ее компоненты по примеру параграфа 22. Для этого обозначим четыре фактические численности через a, b, c и d, а их сумму через n; отклонения от теоретически ожидаемого отношения для численностей «Крахмальной» и «Сахарной» кукурузы будут определяться разностью

$$x = (a + b) - 3(c + d) = +95,$$

а для численностей в зависимости от другого фактора — разностью:

$$y = (a + c) - 3(b + d) = +87.$$

В то же время для анализа в целом мы имеем:

$$z = a - 3b - 3c + 9d = -3145.$$

Разделив квадрат каждого из этих отклонений на соответствующую ему дисперсию, а именно на $3n$ для x и y и на $9n$ для z, мы получим компоненты χ^2:

$$\begin{aligned} x^2 : 3n &= 0{,}784 \\ y^2 : 3n &= 0{,}657 \\ z^2 : 9n &= 286{.}273 \\ \text{Итого} &= 287{,}714 \end{aligned}$$

Этот итог в пределах погрешности вычисления вполне согласуется с прежним значением χ^2. Отсюда можно прийти к выводу, что ни одно из отношений, принадлежащих отдельному фактору, не является отклоняющимся от нормы, и что основная часть отклонений от нормы должна быть приписана связи между генами. В параграфе 55 будут с бóльшей подробностью выяснены принципы построения величин x, y и z.

52. Распределение потомства по классам при наличии связи между факторами

Когда, как в настоящем случае, результаты наблюдений интерпретируются с точки зрения определенной теории, распределение по классам является простым следствием этой теории. Теория, рассматриваемая здесь, состоит в том, что хотя для обоих — мужских и женских — гаметопроизводителей в целом имеются одинаковые шансы произвести крахмальный или сахарный ген, а также ген зеленого или белого основного листа, все же родительские комбинации «Крахмальная» — «Белый» и «Сахарная» — «Зеленый» более продуктивны, чем комбинации «Крахмальная» — «Зеленый» и «Сахарная» — «Белый». Если вероятности для двух последних классов будут p для женского и p' для мужского гаметопроизводителя, то вероятности для четырех типов женских яйцеклеток и пыльцы будут такими:

Таблица 64

	„Крахмальная"		„Сахарная"	
	„Зеленый"	„Белый"	„Зеленый"	„Белый"
Яйцеклетки	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$
Пыльца	$\frac{1}{2}p'$	$\frac{1}{2}(1-p')$	$\frac{1}{2}(1-p')$	$\frac{1}{2}p'$

Далее рассматриваемая теория утверждает, что каждое зернышко пыльцы с одной и той же вероятностью оплодотворяет любую яйцеклетку и что все семена и всходы будут одинаково жизнеспособны. В этом случае вероятность того, что зародыш будет двойным рецессивом «Сахарная» — «Белый» (что может произойти только в том случае, если пыльца и яйцеклетки относятся к этому типу), будет равна $\frac{1}{4}pp'$. Вероятности каждого из остальных трех классов зародышей определяются совсем просто, так как вероятность двух классов «Сахарная», взятых вместе, независимо от наличия связи между генами составляет $\frac{1}{4}$; отсюда следует, что для класса «Сахарная»—«Зеленый» вероятность будет $\frac{1}{4}(1-pp')$. Подобно этому вероятность для класса «Крахмальная» — «Белый» также равна $\frac{1}{4}(1-pp')$, откуда следует, что для класса «Крахмальная» — «Зеленый» вероятность составит $\frac{1}{4}(2+pp')$.

Так как эти вероятности включают в себя только pp', то имеющиеся данные позволяют произвести оценку только этого произведения, но не p и p' в отдельности. Поэтому в данном случае наша иллюстрация способов оценки будет представлять собой различного рода оценку именно этой величины pp', которую мы обозначим теперь через θ. Если p и p' равны, то $\sqrt{\theta}$ будет давать долю рекомбинаций у обоих полов; если же они не равны, то $\sqrt{\theta}$ будет являться средней геометрической из этих долей. Однако имеющиеся у нас данные не дают возможности выяснить эти детали и пригодны для оценки только величины θ. Можно видеть, что в случае оплодотворения, когда оба доминантные гены получены из одного и того же прародительского растения, применяется та же классификация, в которой произведена только замена $\sqrt{\theta}$ на $1-\sqrt{\theta}$; эта последняя интерпретируется как доля рекомбинаций.

Теперь статистическая задача принимает вполне определенную форму: вероятности четырех возможностей

$$\frac{1}{4}(2+\theta);\quad \frac{1}{4}(1-\theta);\quad \frac{1}{4}(1-\theta);\quad \frac{1}{4}\theta$$

должны дать оценку значения параметра θ на основе фактических численностей a, b, c и d.

53. Множественность состоятельных статистик

Прежде всего напомним, в чем сущность методов оценок. Основной задачей в этом случае является установление принципов, руководствуясь которыми можно определить различие между достаточными и недостаточными методами оценок. Дальнейшее углубление вопроса требует сопоставления нескольких различного рода статистик, которые имеют определенную самостоятельность. Здесь мы возьмем пять таких статистик.

Можно видеть, что в нашем случае при возрастании θ вероятности первого и четвертого классов увеличиваются, а вероятности двух других классов уменьшаются. Поэтому выражение

$$a-b-c+d$$

может быть использовано в качестве оценки θ. Для того чтобы на этой основе получить состоятельную статистику, подставим ожидаемые значения

$$\frac{n}{4}(2+\theta,\ 1-\theta,\ 1-\theta,\ \theta)$$

вместо фактических a, b, c и d, в результате чего получим $n\theta$; обозначим нашу первую оценку θ через T_1; она будет определяться уравнением:

$$nT_1 = a-b-c+d.$$

С другой стороны, мы можем построить величину z по примеру параграфа 51, которая будет мерой связи между генами и удобна для оценки существенности этой связи. Подставляя, как и ранее, ожидаемые теоретические значения, мы получим $n(4\theta - 1)$, что дает возможность определить новую оценку T_2 параметра θ, определяемую уравнением:

$$n(4T_2 - 1) = a - 3b - 3c + 9d$$

или

$$4nT_2 = 2a - 2b - 2c + 10d.$$

Обычно этим способом можно образовать любое число подобных оценок. Вместо того, чтобы рассматривать сумму ожидаемых крайних численностей a и d, можно взять их произведение. Отношение произведения ad к произведению bc возрастает по мере увеличения θ. При подстановке ожидаемых значений получается уравнение, определяющее третью оценку:

$$\frac{\theta(2+\theta)}{(1-\theta)^2} = \frac{ad}{bc}.$$

Положительным решением этого уравнения будет T_3.

В качестве четвертой статистики можно взять величину, определяемую при помощи метода максимального правдоподобия. Этот метод состоит в том, что логарифмы ожидаемых для каждого класса численностей умножаются на фактические численности и эти произведения суммируются по всем классам, после чего находится то значение θ, которое приводит эту сумму к максимуму.

Путем дифференцирования выражения

$$a\log(2+\theta) + b\log(1-\theta) + c\log(1-\theta) + d\log\theta$$

можно установить, что оно приводится к минимуму при условии

$$\frac{a}{2+\theta} + \frac{d}{\theta} = \frac{b+c}{1-\theta}.$$

Отсюда получаем квадратное уравнение относительно θ:

$$n\theta^2 - (a - 2b - 2c - d)\theta - 2d = 0.$$

Положительный корень этого уравнения и будет статистикой T_4, удовлетворяющей условию максимального правдоподобия.

Наконец, оценку величины θ можно найти путем сравнения фактических частот с ожидаемыми и определения различия между ними при помощи χ^2. В данном случае χ^2 выражается формулой

$$\chi^2 = \frac{4}{n}\left(\frac{a^2}{2+\theta} + \frac{b^2}{1-\theta} + \frac{c^2}{1-\theta} + \frac{d^2}{\theta}\right) - n.$$

Значение θ, при котором χ^2 достигает своего минимума, будет являться положительным корнем такого уравнения 4-й степени:

$$\frac{a^2}{(2+\theta)^2} + \frac{d^2}{\theta^2} = \frac{b^2+c^2}{(1-\theta)^2}.$$

Эту статистику обозначим через T_5.

54. Сравнение статистик при помощи критерия χ^2

Все эти статистики, исключая последнюю, вычисляются весьма просто. Читатель может самостоятельно вычислить первые четыре из них и проверить, что и значение пятой статистики, приводимое ниже, с достаточной точностью удовлетворяет соответствующему уравнению. На основе каждой из этих статистик можно вычислить ожидаемые в каждом классе численности и сравнить их с фактическими численностями. Все это представлено в табл. 65, в которой приведены и значения χ^2, соответствующие этим сравнениям.

Сопоставляя значения оценок θ, можно видеть, что первые три метода довольно сильно различаются друг от друга, но если взять последние три, то, наоборот, они весьма близки друг к другу; они так близки, что ожидаемые значения третьего и пятого методов отличаются от значения четвертого метода примерно только на одну пятнадцатую растения в каждом классе. При сравнении ожидаемых значений с фактическими наблюдаются большие различия в четвертом классе, где второй метод дает большое, а первый метод даже очень большое отклонение. Различия между первыми тремя методами в отношении значений χ^2 очень велики. Для двух степеней свободы (но отнюдь не для трех, ибо при определении степени связи одна степень свободы должна быть исключена) значение χ^2, равное 9,21, может встретиться только один раз из тысячи случаев. Значение χ^2 при втором методе само по себе не является существенным, но так как это значение примерно в два раза больше, чем значение χ^2, при третьем, четвертом и пятом методах, то мы можем с уверенностью считать, что критерий χ^2, если он правилен при последних методах, то при втором методе он

Таблица 65

Сравнение пяти статистических оценок связи генов

Метод	1	2	3	4	5	
T	0,057046	0,045194	0,035645	0,035712	0,035785	—
Рекомбинация в процентах	23,88	21,26	18,880	18,898	18,917	Фактические
Ожидаемые численности	1974,25	1962,875	1953,711	1953,775	1953,845	1997
	905,00	916,375	925,539	925,475	925,405	906
	905,00	916,375	925,539	925,475	925,405	904
	54,75	43,375	34,211	34,275	34,345	32
χ^2	9,717	3,860	2,0158	2,0154	2,0153	—

столь же ошибочен, как и при первом методе. Общее положение, которое находит здесь место, состоит в том, что критерий согласия χ^2 имеет силу только при условии проверки гипотезы с помощью эффективных статистик; в данном случае, как это будет показано в следующем параграфе, третий, четвертый и пятый методы являются эффективными, а первый и второй методы не относятся к числу таковых.

55. Выборочные дисперсии статистик

Более обоснованное исследование достоинств различных статистик можно произвести при помощи определения выборочной дисперсии каждой из них. Так как определение выборочных дисперсий часто сопряжено с применением довольно сложных математических приемов, то мы дадим здесь некоторое число довольно простых формул, которые в большинстве практических случаев достаточны для определения дисперсии той или иной статистики.

Во-первых, если x является линейной функцией фактических численностей, например:

$$k_1a + k_2b + k_3c + k_4d,$$

то, обозначая теоретическую вероятность некоторого класса через p, получим такое среднее значение x:

$$nS(pk).$$

После этого дисперсия x в случайных выборках определяется формулой

$$\frac{1}{n} V(x) = S(pk^2) - S^2(pk), \qquad (A)$$

а если среднее значение x равно нулю, то дисперсия x принимает еще более простую форму

$$nS(pk^2).$$

Во-вторых, если имеется вторая линейная функция y тех же численностей, определяемая коэффициентами k', то ковариация x и y будет

$$nS(pkk').$$

После знакомства с этими положениями выбор линейных функций, которыми мы пользовались при разложении χ^2 в параграфе 51, уже не будет казаться произвольным. Вместе с этим теперь легко провести и вычисление их выборочных дисперсий. Так, для значений p имеем:

$$\frac{1}{16}(9,\ 3,\ 3,\ 1).$$

и для x значения k следующие:

$$1,\ 1,\ -3,\ -3,$$

отсюда

$$S(pk) = 0, \quad S(pk^2) = 3.$$

Следовательно, искомое значение дисперсии x будет $3n$. Очевидно, что для y мы получим эти же значения только с тем дополнительным условием, что среднее значение xy равно нулю. Для z имеем

$$S(pk) = 0, \quad S(pk^2) = 9$$

и вместе с тем каждое из средних значений xz и yz также равно нулю. При разложении χ^2 на его компоненты всегда берутся линейные функции численностей так, чтобы средние значения этих функций были равны нулю и чтобы все ковариации между ними отсутствовали, т. е. чтобы тоже были бы равны нулю.

В нашем случае следует заметить, что среднее значение xy будет равно нулю только при отсутствии связи между генами. Когда же имеется связь между генами, то новые значения p:

$$\frac{1}{4}(2+\theta,\ 1-\theta,\ 1-\theta,\ \theta)$$

дают для ковариации x и y величину:

$$nS(pkk') = n(4\theta - 1),$$

а кореляция между ними будет

$$\frac{1}{3}(4\theta - 1).$$

Некоторые статистики, применяемые в качестве оценок, могут не иметь формы линейной функции численностей, но они могут стремиться к некоторому конечному значению по мере увеличения объема выборки, однако мы будем довольно часто встречаться именно с линейной формой

$$T = \frac{1}{n}(k_1a + k_2b + k_3c + k_4d),$$

как например в случае наших статистик T_1 и T_2.

В таких случаях удобна формула:

$$nV(T) = S(pk^2) - \theta^2 \qquad (B)$$

при допущении, что T — состоятельная статистика. Так как при определении T_1 коэффициенты k равны $+1$ или -1, то можно сразу найти:

$$V(T_1) = \frac{1-\theta^2}{n}.$$

В формуле T_2 коэффициенты $k = \frac{1}{2}$; $-\frac{1}{2}$; $-\frac{1}{2}$ и $2\frac{1}{2}$ и $p = \frac{1}{4}(2+\theta,\ 1-\theta,\ 1-\theta,\ \theta)$, поэтому

$$V(T_2)=\frac{1+6\theta-4\theta^2}{4n}.$$

Эти две выборочные дисперсии значительно отличаются друг от друга. Если θ мало (явной связи генов не обнаруживается), то дисперсия T_2 составляет только одну четвертую часть дисперсии T_1 и в этом случае можно сказать, что T_2 использует в четыре раза большую информацию, чем T_1. Это различие весьма значительно, но оно существует только при отсутствии связи, так как уже при $\theta=\frac{1}{4}$ эти дисперсии относятся как 5 : 3. Данные дисперсии становятся равными при $\theta=\frac{1}{2}$, когда значение рекомбинаций отталкивания $1-\sqrt{\theta}=0{,}29$, а для еще более тесной связи генов T_1 становится статистикой лучшей, чем T_2.

Средняя квадратическая ошибка каждой из оценок T определяется обычным путем, т. е. извлечением квадратного корня из дисперсии. Наибольший практический интерес представляет определение средней квадратической ошибки для доли рекомбинаций $\sqrt{\theta}$. Для этой цели приведенные выше дисперсии делятся на 4θ, после чего извлекается квадратный корень. Подставляя в эти дисперсии, например, значение $\theta=0{,}0357$, можно получить такие оценки процента рекомбинаций:

$$23{,}88\pm4{,}268 \text{ и } 21{,}26\pm2{,}348.$$

В первом случае можно считать, что процент рекомбинаций лежит примерно между 15,3 и 32,4, а во втором — в более узких пределах от 16,6 до 26,0.

Для любой функции численностей, будет ли численность выборки n строго определенной или неопределенной, можно найти приближенное значение выборочной дисперсии, основываясь на теории больших выборок, которая дает формулу

$$\frac{1}{n}V(T)=S\left[p\left(\frac{\partial T}{\partial a}\right)^2\right]-\left(\frac{\partial T}{\partial n}\right)^2. \qquad (C)$$

Эта формула получается путем дифференцирования по каждой фактической численности a и по общему итогу n той функции, которая определяет данную статистику. После дифференцирования вместо каждой численности a подставляется ее теоретически ожидаемое значение np. Если применим формулу (C) к функции

$$F=\log(ad)-\log(bc)=\log[T_3(2+T_3)]-2\log(1-T_3),$$

то значения $\frac{\partial F}{\partial a}$ соответственно будут

$$\frac{1}{a},\quad -\frac{1}{b},\quad -\frac{1}{c},\quad \frac{1}{d},$$

в то время как $\frac{\partial F}{\partial n}=0$, потому что значение n здесь точно не

определено. Следовательно, подставляя pn вместо a и выражая известные уже нам значения p через θ, можно найти:

$$\frac{n}{4}V(F)=\frac{1}{2+\theta}+\frac{2}{1-\theta}+\frac{1}{\theta}=\frac{2(1+2\theta)}{\theta(1-\theta)(2+\theta)}.$$

Чтобы определить дисперсию T_3, следует это выражение разделить на квадрат dF/dT_3 и подставить θ вместо T_3. Но

$$\frac{dF}{dT_3}=\frac{1}{2+T_3}+\frac{2}{1-T_3}+\frac{1}{T_3},$$

откуда

$$nV(T_3)=\frac{2\theta(1-\theta)(2+\theta)}{1+2\theta}.$$

При определении дисперсии той статистики, которая найдена из условия максимального правдоподобия, представляется более удобным применить довольно простой и непосредственный метод. В параграфе 53 было определено выражение для T_4, полученное путем дифференцирования с последующим приравниванием нулю:

$$\frac{a}{2+\theta}-\frac{b+c}{1-\theta}+\frac{d}{\theta}.$$

Если его продифференцировать по θ еще раз и вместо a, b, c и d подставить ожидаемые значения, то получим:

$$-\frac{n}{4}\left(\frac{1}{2+\theta}+\frac{2}{1-\theta}+\frac{1}{\theta}\right).$$

Эта величина равна $-\frac{1}{V(T_4)}$. Следовательно,

$$nV(T_4)=\frac{2\theta(1-\theta)(2+\theta)}{1+2\theta},$$

т. е. та же самая величина, которая была ранее получена для дисперсии T_3. Это положение имеет большое значение для обсуждаемой здесь проблемы, так как оно вытекает из положения, согласно которому не может быть такой статистики, которая имела бы при большой выборке дисперсию, меньшую, чем дисперсия статистики, исчисленной на основе принципа максимального правдоподобия. Такая группа статистик (к их числу относится и та, которая находится из условия минимума χ^2), которые имеют дисперсии, одинаковые с дисперсией статистики, найденной на основе условия максимального правдоподобия, обладает поэтому особой ценностью и составляет группу эффективных статистик. Это следует понимать в том смысле, что при больших выборках такие статистики используют целиком всю информацию, заключающуюся в фактических данных, в то время как менее эффективные статистики, такие, как T_1 и T_2, используют только часть этой информации.

Выражение для минимальной дисперсии

$$\frac{2\theta(1-\theta)(2+\theta)}{(1+2\theta)n}$$

представляет поэтому специфическое свойство, присущее данным, безотносительно к тому, каким методом производится сама оценка θ.

При больших выборках мы можем интерпретировать обратную величину этого выражения:

$$I=\frac{(1+2\theta)n}{2\theta(1-\theta)(2+\theta)},$$

как меру общего количества информации о величине θ, содержащейся в данной выборке. Каждое из наблюдений, очевидно, содержит в себе определенное количество информации, измеряемое величиной

$$\frac{1+2\theta}{2\theta(1-\theta)(2+\theta)}.$$

Это положение служит известной основой при разработке проблемы оценок и при малых выборках; если мы имеем возможность определить общее количество информации, содержащееся в данных, то, вообще говоря, мы также можем, хотя практически это может оказаться довольно трудным делом, определить то количество информации, которое извлекается применяемым методом анализа. Сравнение этих двух количеств дает возможность установить объективную меру эффективности применяемого метода в отношении сохранения имеющейся в используемых данных информации.

Та часть информации большой выборки, которая используется той или иной неэффективной статистикой, может быть определена как отношение выборочной дисперсии эффективных статистик к дисперсии данной статистики. Так, для T_1 мы имеем:

$$E(T_1)=V(T_4):V(T_1)=\frac{2\theta(2+\theta)}{(1+2\theta)(1+\theta)}.$$

Эффективность статистики T_1 достигает единицы при $\theta=1$, но она меньше единицы при всех других значениях θ. Для T_2 имеем:

$$E(T_2)=V(T_4):V(T_2)=\frac{8\theta(1-\theta)(2+\theta)}{(1+2\theta)(1+6\theta-4\theta^2)}.$$

Эффективность достигает единицы при $\theta=\frac{1}{4}$ и падает до нуля при $\theta=0$ и $\theta=1$.

Рис. 11 показывает изменение этих отношений, выраженных в процентах, для всех значений рекомбинаций, т. е. для $\sqrt{\theta}$ при отталкивании и $1-\sqrt{\theta}$ при сцеплении генов. Из этого графика видно, что при фактическом значении θ, равным в нашем примере

около 19% в области отталкивания, эффективность T_1 составляет около 13%, в то время как эффективность T_2 — около 44%. В случае применения T_1 в качестве оценки θ растрачивается совершенно напрасно около семи восьмых той информации, которая используется статистиками T_3, T_4 и T_5. В случае же применения T_2 не используется несколько более половины этой информации. Другими словами, T_1 дает такую оценку θ, при которой из 3839 фактических наблюдений как бы используется только 503 наблюдения; статистика же T_2 дает оценку θ, основанную на 1661 наблюдении из того же числа фактических наблюдений.

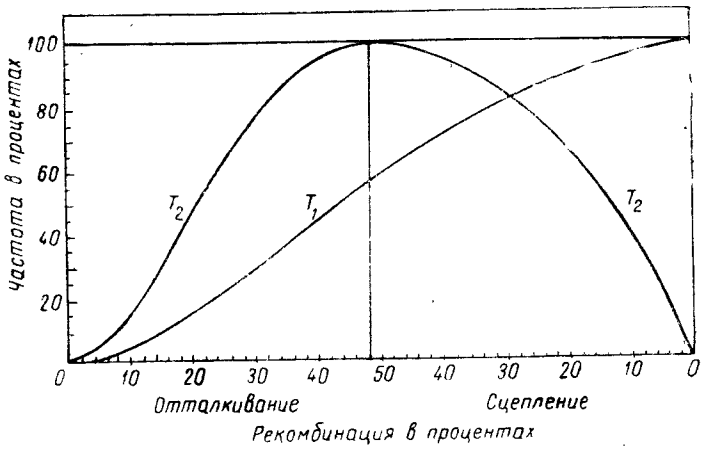


Рис. 11. Эффективность T_1 и T_2 для всех значений θ. T_3, T_4 и T_5, имеющие 100%-ную эффективность во всем интервале, представлены верхней линией.

Средняя квадратическая ошибка эффективных оценок рекомбинации составляет 1,545%, что дает вероятные пределы для истинного значения этой величины от 15,8 до 22,0%. Следовательно, применение неэффективных статистик приводит не только к более худшим оценкам искомой величины, но и к таким оценкам, которые явным образом противоречат тем данным, на основе которых они получены. Значение 23,88%, полученное для T_1, отличается от значений эффективных статистик более, чем на утроенную среднюю квадратическую ошибку этих последних. В данном случае применение неэффективной статистики вводит нас в заблуждение, которое уже само по себе доказывает ошибочность этой оценки.

Второй момент, обусловливающий собой порочность неэффективных статистик, связан с применением критерия согласия χ^2. Опираясь на T_1, мы должны были бы прийти к выводу, что хотя гипотеза о наличии связи в целом согласуется с имеющимися данными, однако ее следует дополнить добавочным допущением о раз-

личной жизнеспособности отдельных групп растений. Найдя только 32 случая двойной рецессивности при 55 ожидаемых случаях, мы, вполне естественно, должны были бы заключить, что этот генотип имеет пониженную жизнеспособность. Но правильно интерпретируемые на основе эффективных статистик данные не приводят к таким выводам. С другой стороны, будет ли это различие объясняться различной жизнеспособностью растений или как-либо иначе, все равно уже одно наличие этого отклонения явится довольно серьезной причиной для сомнения в отношении размера связи между генами; если же, напротив, будут применены эффективные методы оценки, то очевидно, что нет оснований для такого рода сомнений.

56. Сравнение эффективных статистик

Мы видели, что три рассмотренные выше эффективные статистики дали весьма сходные результаты. Это обстоятельство находится в полном согласии с общей теоремой о том, что корреляция между любыми двумя эффективными статистиками стремится к $+1$ по мере увеличения объема выборки до бесконечности. Поэтому значения параметра, полученные на основе этих эффективных статистик, обычно будут одними и теми же. Из рис. 11 видно, что при некоторых значениях θ и статистики T_1 и T_2 также становятся эффективными.

T_2 эффективна при $\theta = \frac{1}{4}$, т. е. при отсутствии связи. Это находится в согласии с результатом применения в параграфе 51 величины z для проверки гипотезы об отсутствии связи. Критерий T_2 в этом случае тождествен простой констатации того, что величина z^2 не будет превосходить, допустим, $36\,n$, т. е. учетверенную дисперсию z. Вообще любой критерий, основанный на сравнении эффективной оценки связи генов с ее средней квадратической ошибкой, должен находиться в согласии с приведенной выше оценкой z. Таким образом, могут существовать такие статистики, как T_2, которые в одних случаях приводят к превосходным критериям существенности, но в других случаях они становятся совсем не эффективными для оценки существенности. Примером этого могут служить моменты третьего и четвертого порядка, применяемые для определения отклонения кривой распределения от нормальной кривой. Эти моменты приводят к точным критериям существенности отклонения от нормальной кривой, но так бывает не всегда, и, например, когда распределение относится к одному из типов Пирсона, значительно отличающемуся от нормального распределения, третий и четвертый моменты являются уже весьма неэффективными статистиками. Это обстоятельство заслуживает особого внимания, так как метод моментов имеет широкое применение именно для такого рода оценок. Является твердо установленным фактом, что эффективность каждой из этих двух статистик достигает 100% только в частном случае, когда распределение нормально, подобно тому как и эффективность статистики T_2 достигает 100% только в частном случае при отсутствии связи. Эффективность этих моментов зависит от формы кривой, подобно тому как эффективность T_2 зависит от значения θ, и быстро падает, как только мы выходим из области высокой эффективности.

Статистика T_1 полностью эффективна при $\theta = 1$, т. е. при весьма сильной связи в фазе сцепления, и поэтому при больших выборках она в этом случае даст оценку, равнозначную оценкам T_3, T_4 и T_5. Этот крайний случай $\theta = 1$ интересен как предельный в теории больших выборок, обращение к которой иногда дает особое освещение вопроса.

Для этой теории важно, чтобы ни одно из чисел a, b, c и d не было бы слишком малым, но при высокой связи в области сцепления типы b и c могут быть представлены только весьма незначительным числом наблюдений. Правда, при любой доле кроссинговера, сколь мала бы она ни была, теоретически всегда можно взять столь большую выборку, что b и c будут все же достаточно большими числами; в таких случаях, конечно, теория больших выборок имеет законную силу. Но также верно и то, что в выборке какого-либо определенного размера связь может быть столь сильной, что число растений типов b и c будет совсем незначительным. Можно видеть, что в этом случае некоторые из эффективных статистик могут потерять свое значение. Если, например, b или c равны нулю, то T_3 обязательно будет равна единице, что будет указывать на полную связь, в то время как два или три растения в другом классе рекомбинаций будут указывать на то, что кроссинговер фактически все же имеет место. В этом случае и статистика T_5 также теряет свое значение; она приводит к тому, что доля рекомбинаций пропорциональна $\sqrt{b^2 + c^2}$, в то время как из T_1 и T_4 следует, что она пропорциональна сумме $b + c$. Вообще, приведение к минимуму величины χ^2, как это следует из конструкции этой величины, не будет достаточным, когда некоторые из классов имеют слишком малое число наблюдений. Поэтому данный метод теряет свое значение, если число классов является неопределенным, как это обычно бывает при распределении непрерывных переменных. Две остальные эффективные при $\theta = 1$ статистики T_1 и T_4 дают одинаковые оценки доли рекомбинаций;

$$\frac{b + c}{n}.$$

Конечно при несколько неполной связи эффективность T_1, как это следует из рис. 11, немного ниже 100%, и поэтому точному значению T_4 в этом случае следует отдать предпочтение; но все же T_1, если b и c малы, дает заметно лучшую оценку, чем T_3 и T_5.

57. Интерпретация отклонения χ^2

Статистика, полученная при помощи метода максимального правдоподобия, находится в определенном соотношении с мерой отклонений χ^2. Исследование этого отношения может служить иллюстрацией метода, основанного на анализе степеней свободы, который был изложен в главе IV и который нашел себе применение во всех частях этой книги.

В предыдущем изложении было установлено, что хотя наблюдения распределяются по четырем классам при 3 степенях свободы, все же, поскольку при решении данной задачи ожидаемые численности вычисляются по фактическим при помощи некоторого параметра θ, остается только 2 степени свободы, относящиеся к различиям между фактическими и гипотетическими численностями. Следовательно, значение χ^2, вычисленное в этом случае, сравнимо с его теоретическим распределением при двух степенях свободы. Этот принцип уже был обсужден и общие положения, на которых он основывается, получили уже достаточно полное теоретическое обоснование. Рассмотрим конкретное содержание этих двух степеней свободы в данном примере. Число наблюдений каждого класса будет полностью определено, если мы знаем:

1) численность выборки;

2) соотношение числа растений, отнесенных к типу крахмальных, к числу растений типа сахарных;

3) соотношение растений с зелеными основными листьями к числу растений с белыми листьями;

4) силу связи этих признаков.

Если ожидаемые численности полностью удовлетворяют пунктам 1 и 4, то они могут отличаться от фактических численностей только в отношении пунктов 2 и 3. Эти различия полностью определяются двумя величинами x и y, имеющими такую форму

$$x = a + b - 3c - 3d$$
$$y = a - 3b + c - 3d$$

Оба эти выражения являются линейными функциями численностей, соответствующими указанным в пп. 2 и 3 соотношениям.

Средние значения x и y равны нулю, а дисперсия случайных выборок для каждой из них равна $3n$. При отсутствии связи между признаками их отклонения от средней будут независимы, но если признаки связаны, то среднее значение xy будет равно.

$$-3n\,\frac{1-4\theta}{3},$$

а корреляция между x и y в этом случае будет

$$\rho = -\frac{1-4\theta}{3}.$$

Совместное отклонение x и y от нуля будет поэтому определяться величиной (см. параграф 30):

$$Q^2 = \frac{1}{1-\rho^2}\left[\frac{x^2 - 2\rho xy + y^2}{3n}\right] =$$

$$= \frac{3}{8n(1-\theta)(1+2\theta)}\left|x^2 + y^2 + \frac{2}{3}(1-4\theta)xy\right|.$$

Это выражение, зависящее от θ, является квадратической функцией численностей; внешне оно напоминает χ^2, но, производя почленное сравнение этих двух выражений, можно установить, что

$$\chi^2 = Q^2 + \frac{1}{I}\left[\frac{a}{2+\theta} - \frac{b+c}{1-\theta} + \frac{d}{\theta}\right]^2,$$

где I — количество информации, содержащееся в фактических данных. Эта величина была определена в параграфе 55.

Из этого равенства вытекают два важных следствия. Во-первых, равенство $\chi^2 = Q^2$ осуществляется только при частном значении θ, определяемым уравнением максимального правдоподобия, и ни при каких иных значениях θ. Отсюда следует, что в этом частном случае даже при конечных выборках отклонения фактических численностей от ожидаемых состоят только из отклонений тех двух соотношений, которые обусловлены двумя факторами, взятыми в отдельности.

Во-вторых, для любого значения θ величина χ^2 является суммой двух положительных величин, из которых одна является Q^2, а другая измеряет отклонение данного значения θ от того значения, которое является решением уравнения максимального правдоподобия. Эта последняя часть является результатом включения в χ^2 ошибок оценки; отклонения же фактических численностей от гипотетических, определенных на основе любого значения θ, измеряются только величиной Q^2.

На рис. 12 даны значения χ^2 и Q^2 в той области, в которой содержатся наши три эффективные статистики.

Соприкосновение графиков в точке, соответствующей методу максимального правдоподобия, подтверждает, что решение, основанное на приведении к минимуму χ^2, не может представлять особый интерес, хотя сама по себе величина χ^2 является вполне обоснованной мерой различия между фактическими и гипотетическими численностями. По мере того, как меняется гипотетическое значение θ, меняется и величина Q^2, и хотя это изменение очень небольшое, все же оно приводит к линии с достаточно выраженной кривизной.

Если отвлечься от части, приписываемой ошибкам оценки, которая при выборе достаточно эффективного метода оценки сводится к небольшой величине, то можно сказать, что мера отклонений χ^2 в рассматриваемой нами задаче просто измеряет отклонения от ожидаемых значений тех указанных ранее соотношений,

которые обусловлены взятыми в отдельности двумя признаками. Поэтому о существенности χ^2 можно судить, сравнивая ее значение с табличным значением χ^2 при двух степенях свободы. Такое сравнение будет являться объективным критерием, зависящим только от имеющихся данных и не зависящим от принятого метода их обработки, однако при условии, что погрешность этой оценки по сравнению с методом максимального правдоподобия достаточно мала. Конечно, эти условия осуществляются, когда

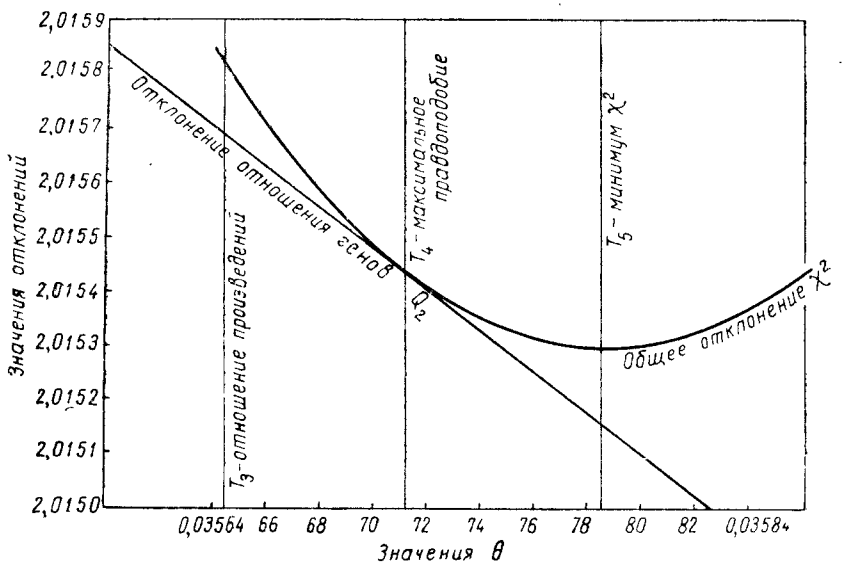


Рис. 12. Графики χ^2 и Q^2 для различных θ в окрестности эффективных оценок.

для оценки параметра при больших выборках применяется та или иная эффективная статистика; вообще все сказанное всегда и безусловно будет иметь силу только при применении метода максимального правдоподобия.

57.1. Обработка фрагментарных данных

В практике статистической работы довольно часто случается так, что одна часть выборки может быть точно классифицирована, в то время как некоторые ее члены имеют ту или иную степень неопределенности в отношении их классификации. Так как разработка таких данных сопряжена с большими трудностями, то обычно прилагаются все меры к тому, чтобы иметь дело по возможности с законченной классификацией материала. Однако в ряде случаев та или иная степень некомплектности является неизбежной, и в связи с этим возникает задача установления надлежащей системы статистической обработки, которая позволяла бы использовать всю фактически доступную информацию. Здесь мы покажем, что если к этому вопросу подходить определенным образом и если основываться на общей теории оценок, то такие задачи не будут казаться совершенно непреодолимыми. В качестве примера мы опять возьмем задачу оценки связи признаков, но следует иметь в виду, что подобного рода затруднения встречаются не только в генетике, но и в других статистических исследованиях.

Пример 48. Тедин, исследуя два взаимносвязанных фактора у Pisum, обозначаемыми *Ar* и *Oh*, получил потомство от 216 растений, которые были классифицированы так: 99 — *OhAr*, 71 — *ohAr* и 46 — *ar*. Фактор *Oh* не мог быть установлен в последней группе растений, и поэтому имеющиеся данные об этом потомстве дают очень мало сведений относительно силы связи, что неизбежно при очень небольшом числе наблюдений и при большой силе связи. У 63 растений группы *OhAr* потомки были получены самоопылением, что дает возможность классифицировать их родителей: 3 были гомозиготными по *Ar*, но не по *Oh*, 8 — гомозиготны по *Oh*, но не по *Ar*, и 52 были гетерозиготны по обоим факторам. Далее, все эти 52 растения указали на отталкивание генов. Наконец, 47 растений группы *ohAr* произвели потомство, давшее только 3 гетерозиготных по *Ar*, остальные же 44 были гомозиготны.

Теперь эти 110 полностью классифицированных растений распределим по классам табл. 66 и рядом построим таблицу, дающую относительные частоты, с которыми полностью классифицированные растения должны попадать в отдельные классы, если доля рекомбинаций равна p.

Таблица 66

	Ar *Ar*	*Ar* *ar*	*ar* *ar*
OhOh	0	8	—
Ohoh	3	0 52	—
ohoh	44	3	—

Таблица 67

	Ar *Ar*	*Ar* *ar*	*ar* *ar*
OhOh	p^2	$2p(1-p)$	$(1-p)^2$
Ohoh	$2p(1-p)$	$2p^2\ 2(1-p)^2$	$2p(1-p)$
ohoh	$(1-p)^2$	$2p(1-p)$	p^2

Далее мы имеем 60 растений, от которых не было получено потомства, но которые были классифицированы по их внешнему виду следующим образом:

Таблица 68

	Ar Ar / *Ar ar*	*ar* / *ar*
OhOh	36	0
Ohoh	24	0
ohoh		

Таблица 69

	Ar Ar / *Ar ar*	*ar* / *ar*
OhOh	$2+p^2$	$1-p^2$
Ohoh	$1-p^2$	p^2
ohoh		

Наконец имеются 46 растений, для которых классификация является еще менее полной:

Таблица 70

	Ar Ar / *Ar ar*	*ar* / *ar*
OhOh	0	46
Ohoh		
ohoh		

Таблица 71

	Ar Ar / *Ar ar*	*ar* / *ar*
OhOh	3	1
Ohoh		
ohoh		

Если теперь допустить, что те растения, которые внутри некоторого класса не получили полного определения, составляют случайную выборку из членов этого класса, то представляется возможность применить по примеру параграфа 53 метод максимального правдоподобия. Для этого следует умножить логарифм теоретически ожидаемой численности каждого класса на число наблюдений в этом классе и сложить все классы вместе, безотносительно к тому, в какой мере классификация была полной. Когда ожидаемые численности у двух классов одинаковы, их можно объединить. В результате мы получим:

$$8+3(1+3)\log[2p(1-p)]+52\log[2(1-p)^2]+44\log(1-p)^2+36\log(2+p^2)+24\log(1-p^2)+46\log(1).$$

Это будет логарифм правдоподобия, который и следует привести к минимуму. Постоянными множителями в выражениях ожидаемых численностей можно пренебречь, так как они дают величины, не зависящие от p. В частности, ожидаемая численность в классе *arar* полностью не зависит от p, а число, стоящее в этом классе, ничего не дает нам для установления наличия или отсутствия связи между признаками и поэтому этот класс в целом может быть отброшен. При этих упрощениях, а также основываясь на том, что логарифм произведения равен сумме логарифмов сомножителей, можно получить такое выражение, подлежащее приведению к минимуму

$$14\log p+206\log(1-p)+36\log(2+p^2)+24\log(1-p^2).$$

После дифференцирования этого выражения по p получим следующее уравнение максимального правдоподобия:

$$\frac{14}{p}-\frac{206}{1-p}+\frac{72p}{2+p^2}-\frac{48p}{1-p^2}=0.$$

Первые два члена этого выражения относятся к полностью классифицированным растениям и характеризуют основной объем искомой информации, последние же два члена содержат дополнительную информацию, которую дают 60 растений класса *Ar*, имеющих менее подробную классификацию. Если основываться на первых двух членах, то следует считать, что p близко к отношению 14 : 220, т. е. оно находится между 6 и 7%. Более точную оценку p можно быстро найти путем подстановки в уравнение приближенных значений p и интерполирования. Так, полагая p равным 0,06 и 0,07, находим:

Таблица 72

	$p = 0{,}06$	$p = 0{,}07$	$p = 0{,}0638$
$\frac{14}{p}$	233,33	200,00	219,436
$-\frac{206}{1-p}$	−219,15	−221,51	−220,038
$\frac{72p}{2+p^2}$	2,16	2,51	2,292
$-\frac{48p}{1-p^2}$	−2,89	−3,38	−3,075
Итого $\frac{\partial L}{\partial p}$	+13,45	−22,38	−1,385

В результате подстановки $p = 0{,}06$ получено 13,45, а при подстановке $p = 0{,}07$ получено — 22,38; более точное значение p, приводящее уравнение к нулю, будет примерно $0{,}06+0{,}01\times$ $\times (13{,}45 : 35{,}83)$, т. е. 0,0638. Результат подстановки этого значения приведен в последнем столбце табл. 72, который служит как для проверки предыдущих вычислений, так и исходным пунктом для отыскания, если в этом есть потребность, более точного решения. Это более точное решение равно 0,06345, исходя из которого читатель в порядке упражнения может найти еще более точное значение p.

57.2. Подсчет количества информации: план и точность

Средняя квадратическая ошибка такой оценки p получается непосредственно из количества информации, содержащейся в данных. В тех случаях, когда данные фрагментарны, следует, как обычно, произвести дифференцирование левой части уравнения максимального правдоподобия и переменить знаки на обратные у всех членов полученного таким образом выражения. Но при подстановке ожидаемых численностей вместо фактических следует учитывать ту основу, на которой покоятся эти ожидаемые численности. Так, в классификации первого года ожидаемые численности при 216 растениях составят $54(2+p^2)$ для *OhAr* и $54(1-p^2)$ для *ohAr*, что дает следующее количество информации:

$$54(2+p^2)\left(\frac{2p}{2+p^2}\right)^2+54(1-p^2)\left(\frac{-2p}{1-p^2}\right)^2$$

или

$$216p^2\left[\frac{1}{2+p^2}-\frac{1}{1-p^2}\right]=216\,\frac{3p^2}{(2+p^2)(1-p^2)}\,. \qquad (A)$$

Эта легко вычисляемая величина представляет собой количество информации, получаемой из классификации первого года.

Если теперь взять 47 растений класса *ohAr*, имевших потомство, то ожидаемыми численностями будут $47(1-p)^2:(1-p^2)$ для класса *ArAr* и $47\times 2p(1-p):(1-p^2)$ для класса *Arar*. Дополнительная информация, которая здесь содержится, будет:

$$47\,\frac{1-p}{1+p}\left(\frac{-1}{1-p}\right)^2+94\,\frac{p}{1+p}\left(\frac{1}{p}\right)^2-47\left(\frac{1}{1+p}\right)^2,$$

где ожидаемые численности каждой части умножены на квадрат дифференциала ее логарифма. После этого можно произвести объединение всех частей, что дает:

$$47\left[\frac{1}{1-p^2}+\frac{2}{p(1+p)}-\frac{1}{(1+p)^2}\right]=47\,\frac{2}{p(1-p)(1+p)^2}\,.$$

Поэтому дополнительная информация на одно растение в этой группе составит:

$$\frac{2(1-p)}{p(1-p^2)^2}\,. \qquad (B)$$

Наконец, рассматривая распределение по классам 63 растений *ArOh*: 52 класса *ohAr|Ohar*, 11 класса *OhOh|Arar* или *Ohoh|ArAr* и 0 класса *OhOh|ArAr* или *OhAr|ohar*, придем к таким ожидаемым численностям:

$$\frac{63}{(2+p^2)}\,[2(1-p^2),\ 4p(1-p),\ 3p^2].$$

Дополнительная информация на одно растение в этой группе будет:

$$\frac{1}{2+p^2}\Big[2(1-p)^2\left(\frac{-2}{1-p}\right)^2+4p(1-p)\left(\frac{1}{p}-\frac{1}{1-p}\right)^2+$$
$$+3p^2\left(\frac{2}{p}\right)^2\Big]-\left(\frac{2p}{2+p^2}\right)^2$$

или

$$\frac{1}{2+p^2}\left[8+\frac{4(1-2p)^2}{p(1-p)}+12\right]-\frac{4p^2}{(2+p^2)^2}\,,$$

что можно окончательно представить в виде

$$\frac{4(2+2p-p^2)}{p(1-p)(2+p^2)^2}\,. \qquad (C)$$

При $p = 6{,}345\%$ числовые значения информации на одно растение для (A), (B) и (C) соответственно будут равны 0,006051; 29,76 и 35,58. Таким образом, классификация второго года дает примерно в 5—6 тысяч раз большую информацию на

одно растение, чем классификация первого года. Общее количество информации в данном случае равно 3642. Обратная ее величина 0,0002746 является дисперсией доли рекомбинаций, 2,746 — дисперсией процента рекомбинаций и 1,657% — средней квадратической ошибкой его.

Определение количества информации, получаемой на каждой стадии эксперимента, имеет большое значение, так как точность, которой удается достигнуть в опыте, обычно лимитируется количеством земли, труда и возможностью наблюдения; на основе изучения, из каких частей складывается общее количество информации, представляется возможным получить известный выигрыш путем лучшего использования указанных ресурсов. Например, в рассматриваемом эксперименте, вероятно, следует отдать предпочтение потомству растений класса OhAr перед потомством класса *ohAr*.

Если же наша задача состоит просто в определении средней квадратической ошибки для некоторого частного результата, то можно приближенно определить количество информации путем непосредственного дифференцирования уравнения максимального правдоподобия. В нашем случае получаем

$$\frac{14}{p^2}+\frac{206}{(1-p)^2}-\frac{72\,(2-p^2)}{(2+p^2)^2}+\frac{48\,(1+p^2)}{(1-p^2)^2}\ ,$$

отсюда находим 3725, т. е. общее количество информации, на которой основывается наша оценка p, а также определяем среднюю квадратическую ошибку этой величины, выраженной в процентах, равную 1,638. Следует заметить, что полученную этим путем оценку нельзя считать худшей по сравнению с оценкой, построенной на основе вышеприведенных теоретических соображений; она только не дает указаний в отношении мероприятий по улучшению эксперимента. Тот факт, что последний расчет дал несколько большее количество информации, чем при прежней конкретной классификации, следует считать чистой случайностью.

Различие между количеством информации, фактически содержащейся в изучаемых данных, и средней величиной, ожидаемой при некотором определенном ряде наблюдений, представляет собой чисто теоретический интерес и поэтому довольно редко возникает необходимость в тех несколько более точных расчетах, которые приведены выше. В целях же простой оценки точности полученного результата можно применить упрощенный метод расчета. Из данных табл. 72 следует, что при изменении p на 0,01 величина $\frac{\partial L}{\partial p}$ уменьшается на 35,83. Отсюда сразу можно определить, что количество информации составляет 3583 единицы и что средняя квадратическая ошибка равна 1,67%; для большинства практических случаев это вполне удовлетворительная оценка.

В некоторых случаях это довольно грубое приближение может быть недостаточным. Оно фактически определяет количество информации, соответствующее значению p около 6,5%, т. е. середине интервала между двумя выбранными значениями 6 и 7%. Фактическое же значение, полученное для p, равно 6,345%, т. е. меньше 6,5%. Более точную цифру для количества информации можно найти на основе трех опорных значений p. Для интервала $p=0{,}06$ и $p=0{,}0638$ мы имеем:

$$\frac{13{,}45\times 1{,}385}{0{,}0038}=3904,$$

что соответствует значению $p = 0{,}0619$. Для интервала $p = 0{,}0638$ и $p = 0{,}07$ получим:

$$\frac{-1{,}385\times 22{,}38}{0{,}0062}=3386,$$

что соответствует $p = 0{,}0669$.

Отсюда для $p = 0{,}06345$ можно найти

$$\frac{0{,}00155\times 3386+0{,}00345\times 3904}{0{,}005}=3743.$$

Этому значению соответствует средняя квадратическая ошибка 1,635%, что является достаточно точным результатом, полученным без построения формулы для количества информации.

57.3. Критерий однородности данных, используемых для построения оценок

В ряде случаев, когда на основе разнообразных данных определяется то, что с теоретической точки зрения является одной и той же величиной, приходится, как это было в последнем параграфе, объединять данные, получаемые из различных источников, с целью получить единую оценку, основанную на всей массе наблюдений. Потребность в соответствующих этой задаче статистических методах вряд ли недооценивается, однако на практике в одинаковой, если не большей, мере важно иметь критерий того, что эти различные источники информации полностью совместимы друг с другом для данной цели или, наоборот они дают только механическое соединение наблюдений, которые сами по себе совершенно различны. Здесь рассмотрим вопрос о применении в этих случаях критерия однородности χ^2, произведя те же вычислительные операции, которые применяются для нахождения такой объединенной оценки.

При тетрасомной наследственности каждая хромосома способна соединиться с любой другой из четырех хромосом гомологичного ряда, к которому она принадлежит. Если различные части ее соединяются с различными партнерами, то возможно получение полностью идентичных двух гомологичных генов, переносимых некоторой единичной гаметой. Обозначим долю таких гамет, относящихся к некоторому фактору, через α. Таким образом, для

растения, содержащего один доминантный ген из имеющихся четырех, возникает возможность передать два таких гена в одну и ту же гамету. Частоты, с которыми происходит передача 0,1 и 2 доминантных генов, будут равны $2+\alpha$, $2-\alpha$ и α, а вместе 4. Соответствующие частоты для дуплексного растения (несущего два доминантных гена) будут $1+2\alpha$, $4-4\alpha$ и $1+2\alpha$, а вместе 6.

Для гена, определяющего отмирание ботвы у картофеля, привитого подвоем; который заражен вирусом *X*, Кадман взял данные из четырех источников: backcross и intercross потомки simplex растений и backcross и intercross потомков duplex растений. Эти данные приведены в табл. 73.

Они позволяют оценить размер величины α и установить однородность этих данных. Схема расчетов приведена в табл. 74.

В табл. 74 взяты значения $\alpha = 0{,}120$ и $0{,}122$ как наиболее подходящие для определения оценки α, которая оказалась равной 12,16%, а также для расчета количества информации. Этот расчет в табл. 74 произведен по частям.

Таблица 73

		Отмершая ботва	Сохранившаяся ботва	Итого
Simplex растения	Backcross	762	842	1604
	Intercross	122	41	163
Duplex растения	Backcross	144	38	182
	Intercross	122	10	132

Величина *I* определяется довольно точно делением разности между двумя оценками при $\alpha = 0{,}120$ и $\alpha = 0{,}122$ на 0,2.

Количество информации, подсчитанное при $\alpha = 12{,}16\%$, довольно близко к точному значению. Для точного значения α, как установлено в параграфе 57, величина χ^2 равна D^2/I; в нашем случае суммированы данные отдельных частей и из этой суммы вычтено D^2/I, исчисленное для всего материала в целом. В нашей таблице значения *D* взяты для $\alpha = 0{,}122$; читатель может самостоятельно определить этот критерий для $\alpha = 0{,}120$.

Следует отметить, что точные расчеты, описание которых дано в параграфе 57, требуют, чтобы *I* было вычислено самостоятельно по каждой отдельной группе. Так, для группы simplex backcross *I* будет равно $\frac{1604}{(4-\alpha^2)}$. Этот расчет дает количество информации, несколько иное, чем то, которое было получено в табл. 74, а именно для четырех отдельных частей 402,5; 56,71; 123,0 и 61,302, а в целом 643,5. Соответствующие значения $\frac{D^2}{I}$ будут в этом случае 0,1992; 0,7204; 0,0023 и 3,4457 с общей суммой $\chi^2 = 4{,}3675$. Эти последние значения точно соответствуют распре-

Таблица 74

	$\alpha = 0{,}120$	$\alpha = 0{,}122$	*I*	$\frac{D^2}{I}$	$\frac{D^2}{I}$
Simplex backcross					
$842/(2+\alpha)$	397,1698	396,7955			
$-762/(2-\alpha)$	−405,3191	−405,7508			
D	−8,1493	−8,9553	403,0	0,1990	0,5131
Simplex intercross					
$82\cdot(2+\alpha)/[16-(2+\alpha)^2]$	38,6792	38,6428			
$-244\cdot(2+\alpha)/[16-(2+\alpha)^2]$	−44,9590	−45,0346			
D	−6,2798	−6,3918	56,0	0,7296	
Duplex backcross					
$76/(1+2\alpha)$	61,2903	61,0932			
$-288/(5-2\alpha)$	−60,5042	−60,5551			
D	+0,7861	+0,5381	124,0	0,0023	1,0984
Duplex intercross					
$40\cdot(1+2\alpha)/[36-(1+2\alpha)^2]$	32,2581	32,1543			
$-488\cdot(1+2\alpha)/[36-(1+2\alpha)^2]$	−17,5588	−17,6206			
D	+14,6993	+14,5337	82,8	2,5511	
Итого	+1,0563	−0,2753	665,8	−0,0001	−0,0001
			χ^2	3,4819	1,6114
			n	3	1

делениям χ^2, основанным на числе наблюдений в каждом классе. Различие между этими двумя значениями χ^2, полученными разными методами расчета, имеет своим источником ошибки случайной выборки и оно приближается к нулю по мере увеличения размера выборки. Поэтому оба метода при однородном материале в пределе, т. е. при безграничном увеличении объема выборки, дают одно и то же теоретическое распределение χ^2 и, следовательно, нет никаких оснований отдать предпочтение одному методу перед другим. Однако метод, рассматриваемый в настоящем параграфе, более универсален, так как он применим и тогда, когда оценка производится на основе величин, не являющихся численностями, так что численности не являются исключительными величинами, по которым только и может быть произведен этот расчет.

Критерий однородности χ^2 здесь может быть применен в двух вариантах: во-первых с тремя степенями свободы для оценки различия между четырьмя группами данных и, во-вторых, с одной степенью свободы для оценки различия между простыми и дуплексными родительскими растениями. Если основываться на обоих этих критериях, то однородность материала можно считать установленной, хотя, пожалуй, желательно повторить исследование при большем количестве информации по сравнению с теми 665,8 единицами, которые были использованы в данном случае.

58. Обзор принципов

При решении любой задачи, относящейся к области статистических оценок, можно применить бесчисленное количество методов, каждый из которых имеет тенденцию дать правильные результаты по мере увеличения объема данных до бесконечности. Каждый из этих методов дает соответствующую формулу, по которой на основе имеющихся данных можно вычислить статистику, являющуюся оценкой неизвестной статистической характеристики. Такие статистики, оценивающие одну и ту же величину, могут иметь самые различные численные значения.

Изучение пяти таких статистик при решении приведенной выше простой генетической задачи показало, что имеется некоторая группа статистик, дающих весьма близкие друг к другу значения, в то время как другие оценки довольно сильно отклоняются от этих значений. Эти отклонения иногда являются источником погрешностей критерия χ^2.

Исследование ошибок выборки показало, что группа согласованных друг с другом статистик при большой выборке имеет дисперсию, равную той, которая получается при решении уравнения максимального правдоподобия, т. е. такую малую дисперсию, какая только возможна. Это будут эффективные статистики. Дисперсии же неэффективных статистик будут большими и могут даже быть столь большими, что значения этих статистик не будут согласовываться с данными, на основе которых они получены.

Эффективные статистики дают почти одинаковые результаты, если выборки достаточно велики, но в тех случаях, когда теория больших выборок теряет свою силу, эти статистики, кроме той, которая получена по методу максимального правдоподобия, могут оказаться неправильными.

Меру отклонений χ^2 можно разложить на две части, одна из которых измеряет реальные различия между фактическими и гипотетическими данными, а вторая измеряет только различие между значением примененной в данном случае статистики и тем значением, которое получается по методу максимального правдоподобия. Основываясь на этом, следует считать, что при оценке однородности материала, полученного из разных источников, необходимо применять именно этот последний метод.

Количество информации, содержащееся в данных, можно определить совершенно точно, а на основе этого можно вычислить и ту часть информации, которая используется той или иной неэффективной статистикой. Этот же прием, хотя и в более усложненной форме, можно применить и при сравнении эффективных статистик при малых выборках.

Метод максимального правдоподобия может быть применен и в случае разработки фрагментарных данных, когда одна часть их имеет менее полную классификацию, чем другая. В этом случае каждая часть материала вносит в общее количество информации свою долю, соответствующую той полноте, с которой в ней проведена классификация. Знание количества информации, содержащегося в различных частях материала, может оказать помощь при планировании затрат труда и других ресурсов, необходимых для проведения различного рода наблюдений.

Следует иметь в виду, что то подробное исследование, которое мы провели на трех несколько упрощенных примерах, взятых из генетики, совсем не является необходимым во всех практических случаях. Наша задача здесь состояла в том, чтобы дать изложение принципов, приложимых при решении всех задач, относящихся к проблеме статистических оценок. В большинстве же случаев бывает достаточно решить с удовлетворительным приближением уравнение максимального правдоподобия и вычислить выборочную дисперсию соответствующей оценки.

ИМЕННОЙ И ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ