СТАТИСТИЧЕСКИЕ МЕТОДЫ
в экспериментальной физике

Перевод с английского В. С. КУРБАТОВА.

Под редакцией профессора
А. А. ТЯПКИНА

МОСКВА АТОМИЗДАТ 1976

УДК 519.25 : 539.12

STATISTICAL METHODS IN EXPERIMENTAL PHYSICS

BY W. T. EADIE, D. DRYARD, F. E. JAMES, M. ROOS*, B. SADOULET

CERN, Geneva

* and University of Helsinki

NORTH—HOLLAND PUBLISHING COMPANY · AMSTERDAM · LONDON, 1971

Статистические методы в экспериментальной физике. Пер. с англ. Под ред. А. А. Тяпкина. М, Атомиздат, 1976, с. 335. (Авт.: Идье В., Драйард Д., Джеймс Ф., Рус М., Садуле Б.)

В книге изложены методы математической статистики, применяемые в экспериментальной физике. Авторами рассмотрены все аспекты обработки экспериментальных данных: методы оценки неизвестных параметров, критерии проверки гипотез, способы принятия вероятностно-достоверных решений и др. Излагаемый материал иллюстрируется многочисленными примерами из физики высоких энергий. Затрагиваемые в книге сложные вопросы математической статистики изложены в форме, доступной для самого широкого круга физиков-экспериментаторов.

Книга может быть рекомендована для физиков-экспериментаторов, работающих в области физики высоких энергий, а также для всех специалистов по ядерной физике, связанных с обработкой результатов наблюдений. Кроме того, она может быть полезна преподавателям и студентам старших курсов физических факультетов.

С $\frac{20402-013}{034(01)-76}$ 13—76

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Предлагаемая читателю в русском переводе книга «Статистические методы в экспериментальной физике» написана специалистами Европейского центра ядерных исследований (ЦЕРН, Женева). Она весьма полно охватывает круг задач математической статистики, с которыми сталкиваются физики-экспериментаторы при обработке результатов измерений и планировании экспериментов, а также физики-теоретики при анализе экспериментальных данных и сопоставлении их с теоретическими построениями.

Основное внимание при изложении материала авторы уделили прикладной стороне вопроса. Приведенных в книге основных положений теории вероятности и кратких сведений из области обоснования описанных методов математической статистики вполне достаточно для того, чтобы позволить читателю самостоятельно ориентироваться в выборе правильных статистических критериев и оптимальных методов решения различных задач, встречающихся в современной экспериментальной физике.

Практическому усвоению материала книги способствует большое число приведенных примеров применения статистических методов к решению конкретных физических задач. Правда, эти примеры выбраны авторами в основном из одной области физики — из арсенала задач физики высоких энергий. Но этот выбор, хотя он и требует от специалистов других областей физики несколько бо́льшего труда для уяснения сути разбираемых примеров, имеет определенное оправдание в том, что позволяет наилучшим образом продемонстрировать обоснованность и плодотворность применения довольно сложных методов математической статистики на самых ярких примерах косвенных измерений.

В физике элементарных частиц, как ни в одной другой области науки, величины, подлежащие определению в эксперименте, удалены от величин, непосредственно измеряемых макроприборами. Эта область экспериментальной физики в изобилии дает, казалось бы, парадоксальные примеры определения разности масс частиц ($\sim 6 \cdot 10^{-39}$ *г*) без их взвешивания, определения временных интервалов ($\sim 10^{-23}$ *сек*) и пространственных промежутков ($\sim 10^{-15}$ *см*) без непосредственного измерения времени и пространства. Научно обоснованные выводы в этой области экспериментальных исследо-

ваний были бы просто невозможны без привлечения всего арсенала методов математической статистики. Можно без преувеличения утверждать, что проникновение в микромир стало возможным в XX в. не только благодаря развитию техники, но и созданию новых методов математической статистики. Подобно тому, как астрономия и геодезия в прошлом представляли основные области для применения классической теории ошибок, так и физика элементарных частиц и атомного ядра в настоящее время для многих разделов математической статистики стала основной ареной апробации практического использования в научных исследованиях.

Конечная цель обработки экспериментальных результатов для естествоиспытателя чаще всего сводится к получению статистических данных об интересующих его теоретических величинах или к определению в рамках вероятностных заключений достоверности той или иной теоретической гипотезы. Этой проблеме в книге уделено основное внимание. Непосредственно в самих научных исследованиях редко возникает задача о принятии на основе полученных статистических данных решения, носящего необратимый характер и приводящего в случае его ошибочности к экономическим потерям. По этой причине авторы ограничились кратким изложением теории решений (гл. 6), составляющей важнейший для производства раздел прикладной математики.

Современная трактовка некоторых общих вопросов теории обработки экспериментальных данных на основе новейших методов математической статистики расходится с формулировками классической теории ошибок. В математической литературе до сих пор, к сожалению, соседствуют взаимоисключающие утверждения по этим вопросам. Это обстоятельство, по-видимому, и явилось главной причиной, побудившей авторов прибегнуть к параллельной трактовке основных вопросов теории оценок и теории решений на основе двух концепций: общего теоретико-информационного подхода и подхода, использующего теорему Бейеса. Несмотря на оговорки о небесспорности и неопределенности бейесовского подхода, авторы излагают весьма искусственный прием модернизации его, вводя так называемые субъективные вероятности.

В настоящем издании книги мы сочли необходимым сопроводить соответствующие места текста примечаниями, разъясняющими ограниченность области обоснованного использования подхода на основе теоремы Бейеса, а также осветить этот вопрос в первом разделе «Дополнения». Кроме того, в «Дополнение» включены еще два раздела, посвященные не рассмотренным в других монографиях вопросам применения принципа максимального правдоподобия.

Издание настоящей книги, безусловно, будет способствовать дальнейшему повышению уровня математической культуры обработки результатов экспериментальных исследований в физике и построению на их основе статистически обоснованных выводов.

А. А. Тяпкин

ПРЕДИСЛОВИЕ

Настоящий курс по статистике написан одним специалистом по статистике (В. Идье) и четырьмя физиками, работающими в области физики высоких энергий. Прежде всего он предназначен для физиков (и экспериментаторов в родственных науках), сталкивающихся с задачей извлечения информации из экспериментальных данных. Физикам часто не хватает элементарного знания статистики при решении проблем, требующих совершенных методов, если соответствующие методы вообще существуют. Для того чтобы по возможности охватить все вопросы, потребовалось бы написать очень объемный курс. Такие книги имеются и в настоящее время [1 — 3], но физики, как правило, не имеют времени для их изучения.

Мы попытались создать курс, который разумно краток и тем не менее достаточен для экспериментальной физики. Обычно это требует некоторого компромисса между теоретической строгостью и количеством описанных методов. Поэтому многие результаты приводятся без строгого доказательства (или вовсе без какого-либо доказательства). Тем не менее мы старались написать нечто большее, чем справочник рецептов и формул, не касаясь целого ряда вопросов, которые, по нашему мнению, несущественны для экспериментальной физики.

В то же время мы вводим много теоретических понятий, которые на первый взгляд кажутся ненужными экспериментатору. Нам представляется это необходимым по двум причинам. Во-первых, экспериментатор может захотеть изучить некоторую теорию или некоторые «общие методы» для того, чтобы разработать свои собственные методы — ведь экспериментальная физика всегда ставит новые вопросы. Вот почему мы останавливаемся на теории информации (гл. 5) и пытаемся в гл. 7 дать определение «общего» метода оценки. Мы надеемся, что разработанный читателем метод, если и не будет оптимальным, тем не менее он будет полезным. Во-вторых, экспериментатор должен знать о предположениях, лежащих в основе метода, будь он стандартным или его собственным. Именно по этой причине так много говорится о центральной предельной теореме, которая служит основой всей асимптотической статистики (гл. 3 и 7).

При цитировании теоремы мы также стараемся указать область ее применения для того, чтобы предостеречь читателя от неосторожного использования некоторых методов.

Среди главных предположений особенно важными являются предположения о распределениях генеральной совокупности данных, поскольку они определяют результаты. В гл. 4 приводится перечень полезных идеальных распределений; в реальной жизни они могут быть либо усечены (§ 4.3), либо преобразованы с учетом экспериментального разрешения (§ 4.3), либо может быть учтена эффективность регистрации (§ 8.5). Кроме того, истинное распределение может быть неизвестно; это приводит нас либо к эмпирическим распределениям (§ 4.3), либо к устойчивой оценке (§ 8.7) и критериям, не зависящим от распределений (гл. 11).

При использовании статистики в настоящее время молчаливо предполагается всеми, что объем экспериментального материала достаточно велик, чтобы считать выполненными условия асимптотики. В книге проводится четкая грань между асимптотическими свойствами (обычно простыми, если они известны) и свойствами выборки конечного размера (которые обычно неизвестны), часто даются асимптотические разложения, чтобы указать, как быстро асимптотические свойства становятся правильными.

В целом мы придаем особое значение различным понятиям оптимальности. Объяснение этого состоит не только в том, что для статистика, сторонника классической школы, это служит единственным способом выбора между различными методами, но также и в том, что физики-экспериментаторы обрабатывают все возрастающие объемы данных и поэтому нуждаются во все более оптимальных методах. Однако существует «оптимальная оптимальность», поскольку незначительное улучшение оптимальности часто достигается только большой ценой. Для этого нужно привлекать соображения экономии, что и пытаемся подчеркнуть во многих случаях.

В тех вопросах, по которым бейесовцы и антибейесовцы («классические» статистики) имеют разногласия, мы стоим на позициях классической школы (из-за профессиональных соображений), однако всюду, где это нужно, информируем читателя о подходе бейесовцев. Такое отношение мы объясняем следующим образом. В гл.6 показано, как решение, принятое на основе ограниченного количества информации, приводит к фундаментальной неопределенности: любое решение зависит от априорных предположений. Поскольку такие предположения по своей природе являются субъективными, можно считать, что функция экспериментаторов состоит не в том, чтобы принимать решения, а представлять результаты своего эксперимента таким образом, чтобы передавать максимум информации о предмете измерения. Тогда задача выбора решения в некотором смысле сводится к установлению единой точки зрения по тому или иному вопросу.

Этим объясняется, почему мы к вопросам оценки подходим с позиций теории информации. Логически теория проверок должна

быть тогда бейесовской (поскольку проверка в действительности и есть решение). Объяснение, почему в теории проверок (гл. 10 и 11) мы не стоим на позиции бейесовцев, состоит в том, что физики в практической деятельности рассматривают доверительный уровень как объективную меру «расстояния» эксперимента от проверяемой гипотезы.

В противоположность большинству (если не всем) книг по теории вероятностей и статистике мы не приводим примеров из азартных игр. Это объясняется нашими профессиональными интересами. Физики зачастую считают тщетными попытки использовать такие примеры в физике; поэтому наши примеры берутся из физики (главным образом из физики высоких энергий).

Наконец, отметим, что мы не обсуждаем вопросы оптимизации численных методов, очень важных, например, в методах максимального правдоподобия и наименьших квадратов. Причина этого состоит в том, что существуют отличные, по нашему мнению, курсы [4, 5] и экспериментатор может найти там все детали оптимизации алгоритмов поиска минимумов. Кроме того, большинство физиков имеют в своем распоряжении мощные программы оптимизации (например, в ЦЕРНе — библиотека программ на ЭВМ), что избавляет их от излишнего беспокойства.

ГЛАВА 1

ВВОДНАЯ ЧАСТЬ

§ 1.1. СОДЕРЖАНИЕ КНИГИ

По содержанию книгу можно разделить на две основные части: *теория вероятностей* (см. гл. 2 — 4) и *статистика* (см. гл. 5 — 11).

Теория вероятностей излагается лишь в той мере, в которой это необходимо для изучения статистики.

В гл. 5 и 6 описаны два общих подхода к выбору оценок: подход с позиций информации и подход теории решений. Первый подход состоит в том, чтобы при оценке сохранять максимальное количество информации, тогда как второй — в стремлении сделать минимальной потерю, связанную с выбором неверного решения о значении параметра. В пределе больших статистик эти два подхода эквивалентны, но мы будем отмечать различия там, где они имеют место.

Методы определения оценок параметров излагаются в трех главах: в гл. 7 и 8 — теория и практика оценки значения параметра, а в гл. 9 — оценка параметра интервалами. Критерии проверки гипотез подразделяются на общие критерии (см. гл. 10) и критерии согласия (см. гл. 11).

§ 1.2. ЯЗЫК КНИГИ

Статистика, подобно любой другой отрасли знаний, имеет свою собственную терминологию, к которой необходимо привыкнуть. Некоторые недоразумения могут возникнуть в тех случаях, когда один и тот же термин имеет разный смысл в статистике и в физике или когда одно и то же понятие имеет различные названия. В первом случае мы вкладываем в термин тот смысл, который он имеет в статистике (физик должен понять и изучить различие); во втором — понятие часто обозначаем термином, употребляемым в физике.

Пример первого случая:

физики говорят	статистики говорят
определить	оценить
оценить	угадать

Таким образом, слово «оценить» имеет различный смысл в физике и статистике. В данной книге оно применяется так, как это делают статистики. (Три главы книги посвящены объяснению смысла, вкладываемого статистиками.)

Во втором случае — это «демографический подход» к экспериментальной физике. Своим развитием статистика в значительной мере обязана изучению совокупностей (социология, медицина, сельское хозяйство) и производственной деятельности (контроль качества продукции). Поскольку трудно изучить всю совокупность, необходимо делать выборку. Сама же совокупность существует в действительности.

В экспериментальной физике набор всех изучаемых измерений соответствует выборке. Увеличивая число измерений, физик увеличивает «размер выборки», но он никогда не достигает всей «совокупности». Поэтому «совокупность» — это абстракция, не имеющая в действительности какого-либо аналога. Вследствие этого «демографические» термины до некоторой степени неприемлемы и необязательны и мы постараемся избегать некоторых из них.

Для «демографического» термина	мы используем физический термин
выборка	данные (набор)
сделать выборку	наблюдать, измерить
выборка размера N	N наблюдений
популяция	пространство наблюдений

Необходимо также делать различие между, скажем, средним значением полученных данных и средним, если бы объем данных был бесконечным. Когда это различие необходимо, мы используем термины «среднее выборки», «дисперсия выборки» и т. д. в противоположность среднему генеральной совокупности, дисперсии генеральной совокупности и т. д. или среднему и дисперсии распределения, соответствующего выборке. Таким образом, среднее генеральной совокупности = среднему распределения наблюдений = = среднему совокупности.

Мы избегаем физического термина «ошибка», который может ввести в заблуждение, и используем вместо него термины «дисперсия оценки», «доверительный интервал» или «оценка интервала». Мы также постараемся избегать термина «точность», поскольку он точно не определен. Во многих книгах по статистике можно найти целые главы, посвященные «переносу ошибок». Применение такого термина, по нашему мнению, может привести к недоразумениям; поэтому мы используем термин «замена переменных».

Может показаться, что некоторые вопросы вообще не затронуты в этой книге. Однако иногда они имеют другие названия.

Например, термин «регрессионный анализ» никогда не используется, но те же самые проблемы рассматриваются при подгонке линейных моделей методом наименьших квадратов.

§ 1.3. ДВЕ ФИЛОСОФИИ

К сожалению, специалисты по статистике не пришли к соглашению по основным принципам. Грубо говоря, их можно разделить на две школы: бейесовская (современная*) и антибейесовская (клас-

* Хотя в этой книге термин «современный» соответствует бейесовскому подходу, оригинальная работа Т. Бейеса появилась в 1763 г. [6].

сическая)*. Названия происходят из-за различного подхода к теореме Бейеса. Мы попытаемся представить основные результаты с точки зрения обоих подходов.

Бейесовский подход ближе к целям и складу мышления обычного физика. Физик хочет знать, что дает его эксперимент: может ли он подтвердить или опровергнуть какую-то теорию, является ли наблюдаемый пик резонансом или нет? Бейесовцы используют индуктивный метод при анализе своих данных, тогда как антибейесовцы не хотят делать так. Хотя бейесовский подход более завершен, он не бесспорен**. Вся трудность состоит в априорном знании. Бейесовец представляет все предыдущее знание об искомом параметре в виде некоторого вероятностного закона. Эксперимент модифицирует это знание, преобразуя априорный закон в апостериорный. *Решение* о параметре может быть сделано на основании этого *апостериорного* закона.

Предыдущее знание уже используется при планировании эксперимента. Если предыдущее знание велико, оно не может быть существенно изменено в результате небольшого эксперимента, и экспериментатор решает, делать эксперимент с существенно большим содержанием информации или не делать его совсем. Если предыдущее знание очень мало, то на апостериорный результат слабо влияет использованное априорное распределение.

В гл. 6 мы увидим, как решение, основанное на небольшом количестве информации, приводит к фундаментальной неопределенности: любое решение зависит от априорных предположений. Эти предположения, будучи, главным образом, субъективными по природе, могут привести к тому, что экспериментатор не сможет принять решение. В таком случае его работа сводилась бы к простому суммированию результатов эксперимента с тем, чтобы сохранить максимум информации об измеряемых неизвестных.

Однако экспериментатор-бейесовец, защищенный апостериорным знанием, может использовать теорию решений, изложенную в гл. 6, и делать решения. До момента, когда антибейесовец собирается оценивать параметры, большинство практических результатов совпадает в этих двух подходах, но их интерпретация совершенно различна. Следуя по существу антибейесовскому подходу, мы будем отмечать эти различия там, где это необходимо.

---

* Примененный авторами термин «современная» явно неудачен по отношению к подходу, возникшему на основе работы Томаса Бейеса 1763 г. В советской литературе современной называют формулировку теории оценок, использующую понятие правдоподобия; термин же «классическая» относят к формулировке теории ошибок, данной Лапласом на основе постулата Бейеса. — *Прим. ред.*

** Этот метод заведомо не обоснован в тех случаях, когда априорная вероятность неизвестна и теорема Бейеса дополняется постулатом Бейеса—Лапласа о равномерной плотности распределения априорной вероятности. Вопросы формулировки основной задачи теории оценок подробно обсуждаются в «Дополнении» к настоящему изданию книги. — *Прим. ред.*

§ 1.4. ОБОЗНАЧЕНИЯ

$b(\hat{\theta})$ — смещение оценки $\hat{\theta}$

$D(X, Y)$ — смешанный второй момент случайных переменных X и Y

D_N, D_{MN}, $D_N^{\pm}$ — статистика Колмогорова

$e(X, X')$ — эффективность регистрации

$E(X)$ — оператор математического ожидания

$f(X)$ — функция плотности вероятности (ф. п. в.)

$F(X)$ — функция распределения

$G(X)$ — производящая функция

H, H_i — гипотеза

H_0 — проверяемая гипотеза, «нулевая гипотеза»

$I_{\mathbf{X}}(\theta)$, $I_N(\theta)$ — информационная матрица

K — параметр нецентральности

$K(t)$ — производящая функция семиинвариантов

K_r — семиинвариант

l — отношение правдоподобий

L_p — норма степени p

$L(\mathbf{X}|\theta)$ — функция правдоподобия

n, N, r — число случайных переменных, событий (экспериментов)

$N(\mu, \sigma^2)$ — нормальное распределение со средним μ и дисперсией σ^2

$O(n^{-1})$ — член порядка n^{-1} или меньше

$p(\theta)$, $1-\beta$ — мощность критерия

$P(A)$ — вероятность того, что A верно

$P(A|B)$ — условная вероятность того, что при заданном B верно A

Q^2 — квадратичная форма, ковариационная форма

$r(X, X')$ — функция разрешения

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ — несмещенная оценка дисперсии

$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ — дисперсия выборки

S_n — сумма n случайных переменных, функция распределения для порядковой статистики

t, T — статистика (функция результатов наблюдений)

tr — след (матрицы)

$D(X)$, σ_X^2 — дисперсия

$D_{\mathbf{X}}$ — матрица вторых моментов

$D_{(rs)}$ — подматрица матрицы D, имеющая r строк и s столбцов

w_α — критическая область критерия с уровнем значимости α

w_i — вес

W — пространство проверочной статистики

X, X_i — случайные переменные

$X_n^2 = \sum_{i=1}^{n} X_i^2$ — сумма квадратов случайных переменных

Y, Z, Y_i, Z_i — случайные переменные

X_α, Y_α, Z_α — α-точка $f(X)$ и т. д. [определяется уравнением $F(X_\alpha) = \alpha$]

α — доверительный уровень, уровень значимости, потеря

β — примесь, доверительный уровень

γ_1 — асимметрия

γ_2 — эксцесс

$\delta(X)$ — δ-функция Дирака

θ, θ_i — теоретический параметр

θ_0 — истинное значение θ, значение θ, соответствующее нулевой гипотезе

λ — отношение максимумов правдоподобий

λ_α — α-точка нормального распределения [определяется уравнением $\Phi(\lambda_\alpha) = \alpha$]

μ, μ' — среднее значение

μ_n, μ'_n — центральный и алгебраический момент порядка n

ν_n, ν'_n — абсолютный момент порядка n (центральный и алгебраический)

$\pi(\theta)$ — априорная плотность

$\pi(\theta|\mathbf{X})$ — апостериорная плотность

ρ — коэффициент корреляций

σ^2 — дисперсия

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ — несмещенная оценка дисперсии при заданном среднем μ

$\chi^2(N)$ — χ^2-распределение с N степенями свободы

$\zeta(t)$ — характеристическая функция

$\Phi(X)$ — интеграл вероятности нормального распределения

Ω — пространство случайной переменной

$\wedge$ (например $\hat{\theta}$) — оценка

$\sim$(например $\underset{\sim}{D}$) — матрица

$\underset{\sim}{T}$(например $\underset{\sim}{A}^T$) — транспонирование матрицы $(\underset{\sim}{A})$

$[\theta_a, \theta_b]$ — интервал $\theta_a \leqslant \theta \leqslant \theta_b$

$\binom{p}{q} = \frac{p!}{q!\,(p-q)!}$ — биномиальный коэффициент

ГЛАВА 2

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

В этой главе введены понятия вероятности событий и определены основные правила работы с ними, затем эти правила обобщены и введены понятия случайных переменных и распределений вероятностей для непрерывных величин. Определяются важные характеристики распределений: ожидание, среднее, дисперсия, корреляция, смешанные вторые моменты, моменты. Наконец, вводится ряд вспомогательных понятий для работы с распределениями: характеристическая функция, производящая функция семиинвариантов и производящая функция вероятностей.

§ 2.1. ОПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ

До сих пор нет общепринятого определения вероятности. Ниже мы приводим два определения: определение вероятности как предела частоты и современное определение, основанное на теории множеств. Более подробно этот вопрос обсуждается в книгах по теории вероятностей (см. библиографию).

2.1.1. Определение вероятности как предела частоты. Рассмотрим эксперимент, в котором зарегистрирован ряд событий, и предположим, что некоторые из них принадлежат типу X. Пусть N — полное число событий, а n — число событий типа X. Тогда вероятность того, что какое-то событие будет событием типа X, может быть определена как предел отношения (частоты):

$$\frac{n}{N} \to P(X) \quad \text{при} \quad N \to \infty.$$

Даже если предельное значение неизвестно, физик склонен предположить, что оно существует.

Такое определение не очень нравится математикам, поскольку оно основано на экспериментировании и фактически подразумевает неосуществимые эксперименты ($N \to \infty$). Кроме того, такое определение ограничивает применимость вероятности только к наблюдаемым величинам.

2.1.2. Современное определение. Хотя строгое определение требует знания некоторых положений из теории меры и множеств, основные идеи просты и могут быть изложены без лишних деталей.

Пусть Ω — множество всех возможных *элементарных событий* X_i, которые *взаимно* исключают друг друга, т. е. появление одного из них означает, что другое событие не происходит.

Введем понятие *вероятности появления события* X_i, $P(X_i)$, имеющее следующие свойства:

$$\left.\begin{array}{ll} \text{а)} & P[X_i] \geqslant 0 \text{ для всех } i; \\ \text{б)} & P[X_i \text{ или } X_j] = P[X_i] + P[X_j]; \\ \text{в)} & \sum\limits_{\Omega} P[X_i] = 1^*. \end{array}\right\} \qquad (2.1)$$

Используя эти свойства, можно получить выражения для вероятностей более сложных событий, состоящих из множеств элементарных событий, а также для событий, представляющих собой взаимно перекрывающиеся множества элементарных событий.

§ 2.2. СВОЙСТВА ВЕРОЯТНОСТИ

Множество A элементарных событий X_i может быть опять рассмотрено как событие, хотя и *не элементарное*. Появление A определяется как появление по крайней мере одного события из множества A.

Обозначим $P(A)$ вероятность появления хотя бы одного X_i из множества A. Для таких множеств имеют место *закон сложения* и *закон умножения*, если множества *независимы*. Затем мы вводим понятие *условных вероятностей*, для которых выполняется *теорема Бейеса*. Наконец, мы переходим от события к случайным переменным.

2.2.1. Закон сложения для множеств элементарных событий. Рассмотрим два взаимно неисключающих множества A и B элементарных событий X_i, т. е. некоторые события X_i могут принадлежать и к A, и к B. В таком случае из определений (2.1) следует, что вероятность появления события, принадлежащего либо множеству A, либо множеству B, или обоим сразу, равна

$$P[A \text{ или } B] = P[A] + P[B] - P[A \text{ и } B], \qquad (2.2)$$

где A или B означает множество событий X_i, которые принадлежат либо множеству A, либо множеству B, либо обоим, а A и B — множество событий X_i, принадлежащих и A, и B.

---

* Следует дополнительно отметить, что в этом определении имеется в виду множество случайных событий, частота появления которых характеризуется некоторой вероятностной мерой, являющейся определенной функцией, заданной на всем множестве и удовлетворяющей приведенным выше общим свойствам. Иначе говоря, не частота определяет вероятность, а напротив, в наблюдаемой частоте появления определенных событий проявляется вероятность как мера, определяемая свойствами, присущими всему множеству событий. — *Прим. ред.*

Соотношение (2.2) может быть обобщено на случай нескольких множеств A_1, ..., A_N [7]. Обозначим

$$P_i = P(A_i);$$

$$P_{ij} = P(A_i \text{ и } A_j),\ i < j;$$

$$P_{ijk} = P(A_i \text{ и } A_j \text{ и } A_k),\ i < j < k \text{ и т. д.}$$

Пусть S_r обозначает суммы: $S_1 = \sum_i p_i$; $S_2 = \sum_{i<j} p_{ij}$; $S_3 = \sum_{i<j<k} p_{ijk}$ и т. д. Тогда вероятность появления события, принадлежащего по крайней мере одному из множеств A_i, равна

$$P[A_1 \text{ или } A_2, \text{ или } \ldots, \text{ или } A_N] = S_1 - S_2 + S_3 - \ldots - (-1)^N S_N.$$

2.2.2. Условная вероятность и независимость. Теперь можно ввести понятие *условной вероятности*, т. е. вероятности появления события A при заданном B, которую обозначим $P[A \mid B]$. Она означает вероятность того, что элементарное событие, принадлежащее множеству B, также принадлежит множеству A. Условная вероятность определяется соотношением

$$P[A \text{ и } B] = P[A \mid B]\, P[B] = P[B \mid A]\, P[A]. \qquad (2.3)$$

Говорят, что множества A и B *независимы*, если

$$P[A \mid B] = P[A]. \qquad (2.4)$$

Это означает, что (предшествующее) появление B совершенно не связано с появлением A. Из (2.3) получаем

$$P[A \text{ и } B] = P[A]\, P[B], \qquad (2.5)$$

если A и B независимы. Соотношение (2.5) есть необходимое и достаточное условие независимости A и B, которое оказывается очень полезным.

Между прочим отметим, что коэффициент корреляции, определяемый ниже, равен нулю, если A и B независимы, но равенство нулю коэффициента корреляции еще не означает независимости.

2.2.3. Пример закона сложения: эффективность просмотра. Предположим, что при отборе событий одного и того же типа фотографии с пузырьковой камеры просматриваются дважды. Пусть в первом просмотре найдено $(C + B_1)$ событий, во втором — $(C + B_2)$. Здесь C — количество событий, найденных в обоих просмотрах, тогда как B_1 и B_2 — число событий, найденных только в первом или только во втором просмотре соответственно. Оценим истинное число событий на фотографиях и общую эффективность просмотра в предположении, что вероятность обнаружения различных событий в одном и том же просмотре одинакова и что оба просмотра независимы.

Первое предположение позволяет работать с множествами событий 1 и 2, найденных в первом и втором просмотрах, а не с каждым событием в отдельности. Пусть $P(1)$ — вероятность найти данное событие при первом просмотре, а $P(2)$ — при втором.

Второе предположение позволяет записать [на основе соотношений (2.3) и (2.5)], что:

$$P\,(1\,|\,2) = P\,(1)$$

или

$$P\,(2|1) = P\,(2)$$

или

$$P\,(1 \text{ и } 2) = P\,(1)\,P\,(2).$$

Используя числа найденных событий, можно оценить эффективности просмотров в двух просмотрах посредством

$$\hat{P}\,(1\,|\,2) = \frac{C}{C+B_2} = \hat{P}\,(1)$$

и

$$\hat{P}\,(2\,|\,1) = \frac{C}{C+B_1} = \hat{P}\,(2)$$

(буква со знаком «$\wedge$» обозначает оценку той или иной величины).

Чтобы оценить эффективность просмотра, используем соотношение (2.2):

$$\hat{P}\,(1 \text{ или } 2) = \hat{P}\,(1) + \hat{P}\,(2) - \hat{P}\,(1 \text{ и } 2) = \hat{P}\,(1) + \hat{P}\,(2) - \hat{P}\,(1)\,P\,(2) =$$

$$= \frac{C}{C+B_2} + \frac{C}{C+B_1} - \frac{C^2}{(C+B_1)\,(C+B_2)} = \frac{C\,(C+B_1+B_2)}{(C+B_1)\,(C+B_2)}\,.$$

Для того чтобы оценить полное число N событий на пленке, можно использовать выражение

$$\hat{P}(1) = \frac{C+B_1}{\hat{N}}\,,$$

что дает

$$\hat{N} = \frac{(C+B_1)\,(C+B_2)}{C}\,.$$

Более подробное обсуждение этого вопроса можно найти в работах [8, 9]*

2.2.4. Теорема Бейеса для дискретных событий. Теорема Бейеса [6] связывает $P\,[A \mid B]$ с $P\,[B \mid A]$. Смысл теоремы состоит в том, что для множеств A и B (событий X_i) имеет место соотношение

$$P\,[A \mid B] = P\,[B \mid A]\,P\,[A]/P\,[B\,]. \qquad (2.6)$$

Очевидно, что оно следует из определения условной вероятности [см. соотношение (2.3)]. В общем случае, если $A_1, \ldots, A_N$ — взаимно исключающие и образующие полную систему множества (т. е. наблюдаемое элементарное событие должно принадлежать одному и

---

* Оценка дисперсии величины $\hat{N}$ равна $\hat{\sigma}^2 = \hat{N}\,\frac{B_1 B_2}{C^2}$. См. например, работу С. Н. Соколова и К. Д. Толстова. Контроль эффективности наблюдений и оценка истинного числа событий. — Препринт ОИЯИ, Р-1085, Дубна, 1962. — *Прим. ред.*

только одному из множеств A_i) и если B какое-нибудь событие, то теорема Бейеса может быть записана в виде

$$P[A_i|B] = \frac{P[B|A_i]\,P[A_i]}{\sum_i P[B|A_i]\,P[A_i]}. \qquad (2.7)$$

Пример. Планируется эксперимент для изучения лептонных распадов K^0-мезонов и для детектирования лептонных распадов используется черенковский счетчик. Желательно знать, достаточно ли одного счетчика для детектирования лептонных распадов при наличии небольшого фона от других реакций, которые также могут запускать счетчик? Введем такие обозначения:

$P(B)$ ≡ вероятность запуска черенковского счетчика;
$P(A)$ ≡ вероятность лептонного распада K^0-мезона;
$P(B|A)$ ≡ вероятность того, что лептонный распад вызовет срабатывание черенковского счетчика;
$P(A|B)$ ≡ вероятность того, что данное срабатывание черенковского счетчика вызвано лептонным распадом.

$P(B)$ может быть измерена по числу срабатываний черенковского счетчика при включенном пучке;
$P(A)$ предполагается известной из других экспериментов;
$P(B|A)$ может быть рассчитана, если измерена эффективность счетчика и известны его геометрические размеры.

Тогда нужное экспериментатору значение $P(A|B)$ может быть рассчитано с помощью теоремы Бейеса [см. соотношение (2.6)].

2.2.5. Бейесовский подход к теореме Бейеса. Когда A или B обозначает не множество *событий*, а *гипотезу*, смысл теоремы Бейеса менее очевиден. Именно в этом лежит основа возникновения двух школ статистиков. Бейесовец посредством $P(\theta)$ выражает *степень веры* в гипотезу θ и пытается выразить ее численно. Он считает теорему Бейеса справедливой также и для гипотез, тогда как антибейесовец не приемлет такую интерпретацию.

Посмотрим, как бейесовец трактует различные множители в соотношении

$$P(\theta_i|X) = P(X|\theta_i)\,P(\theta_i)/P(X). \qquad (2.8)$$

Для него θ_i — различные гипотезы, а X — экспериментальные значения; $P(\theta_i|X)$ соответствуют *апостериорному* знанию; $P(X|\theta_i)$ есть вероятность получить значения X, если справедлива гипотеза θ_i (это может быть функция правдоподобия); $P(\theta_i)$ соответствует *априорному* знанию или степени веры в различные гипотезы.

Критика антибейесовцев состоит в том, что все физики будут иметь различные степени веры и поэтому заключения [левая часть соотношения (2.8)] будут субъективными. На это бейесовцы отвечают, что в действительности $P(\theta_i)$ должно быть записано как $P(\theta_i|\theta)$, где θ обозначает множество всех гипотез и все предыдущее знание, и что если бы физики объединили все свое предыдущее знание, они смогли бы договориться относительно распределения $P(\theta_i)$.

Вероятность $P(X)$ получить результат X, если справедлива какая-нибудь гипотеза, может быть неизвестна. Если гипотезы

θ_i взаимно исключают друг друга и образуют полный набор гипотез, что, конечно, бывает очень редко, то

$$P(X) = \sum_i P(X \mid \theta_i) P(\theta_i).$$

В общем случае $P(X)$ неизвестна, и соотношение (2.8) приобретает более слабую форму:

$$P(\theta_i \mid X) \approx P(X \mid \theta_i) P(\theta_i). \qquad (2.9)$$

Хотя в таком случае нельзя рассчитать *апостериорные* вероятности, все-таки можно рассчитать реалистичность различных гипотез, знание которых также полезно. Шансы гипотезы θ_i по сравнению с гипотезой θ_j определяются с помощью отношения

$$P(\theta_i \mid X)/P(\theta_j \mid X). \qquad (2.10)$$

2.2.6. Случайная переменная. *Случайным* называют *событие*, которое имеет более чем один возможный исход. Каждому исходу может быть приписана вероятность. Результат случайного события нельзя предсказать, известны только вероятности возможных исходов. Напротив, событие только с одним исходом можно предсказать и вероятность исхода равна единице.

Случайному событию A может быть поставлена в соответствие *случайная переменная* X, которая принимает различные численные значения X_1, X_2, ..., соответствующие различным возможным исходам. Соответствующие вероятности $P(X_1)$, $P(X_2)$ ... образуют *распределение вероятностей*.

В случае более чем одной случайной переменной или нескольких наблюдений одной и той же случайной переменной должен быть рассмотрен вопрос о *независимости* различных наблюдений. Если они независимы, то на распределение каждой случайной переменной не влияет знание какого-нибудь другого наблюдения. Зависимость наблюдений означает, что распределение одной переменной изменяется, если результат другого наблюдения известен. Зависимые переменные тем не менее случайны и результат наблюдения предсказуем только в терминах вероятностей возможных значений в соответствии с видом распределения. Только в вырожденном случае полной зависимости, когда знание одного наблюдения точно определяет значение второй переменной, вторая переменная становится предсказуемой.

Если эксперимент состоит из n повторных наблюдений одной и той же случайной переменной X, то он иногда рассматривается как одно наблюдение n-мерного случайного вектора с компонентами X_1, ..., X_n.

В экспериментальной науке иногда принято говорить о некоторых переменных как о «более случайных», чем другие. Обычно это значит, что эти переменные распределены более равномерно либо их распределение шире или, наконец, их последовательные наблюдения менее коррелированы. Нам кажется неразумным использовать слово «случайный» для характеристики ре-

альных свойств распределений переменных, поэтому мы используем его только в том смысле, который вкладывался в это слово выше. Если переменную называют случайной, это означает, что она имеет распределение возможных значений и ничего не говорит ни о форме распределения, ни о какой-либо зависимости.

§ 2.3. НЕПРЕРЫВНЫЕ СЛУЧАЙНЫЕ ПЕРЕМЕННЫЕ

Обобщим понятие вероятностей событий и введем понятие распределения вероятностей случайных переменных. Для дискретных случайных переменных это обобщение очевидно. Для непрерывных случайных переменных (чьи возможные значения могут покрывать непрерывные интервалы) необходимо ввести *понятие функции плотности вероятности* (ф. п. в.) и интеграла от нее, т. е. *функции распределения*. Кроме того, введем понятие маргинальных и условных распределений и покажем, как они применяются в теореме Бейеса для непрерывных переменных.

2.3.1. Функция плотности вероятности. Рассмотрим для примера эксперимент с использованием счетчиков, в котором направление частиц пучка определяется двумя последовательными массивами счетчиков X и Y. Каждое зарегистрированное событие характеризуется двумя случайными дискретными параметрами — значениями i и j. Это значит, что частица пересекла счетчики X_i и Y_j в массивах X и Y соответственно. В этом случае можно ввести двумерное дискретное распределение вероятности P (X и Y).

Физик часто склонен думать, что реальное распределение пучка может быть описано непрерывным распределением вероятности $f(X, Y)$, и единственная причина для представления результатов в терминах дискретных переменных состоит в том, что счетчики имеют конечные размеры ΔX и ΔY.

Связь между P (X и Y) и $f(X, Y)$ может быть записана в виде

$$f(X,Y) = \lim_{\substack{\Delta X \to 0 \\ \Delta Y \to 0}} \frac{P(X \text{ и } Y)}{\Delta X \Delta Y} =$$

$$= \lim_{\substack{\Delta X \to 0 \\ \Delta Y \to 0}} \frac{P\left[\left(X - \frac{\Delta X}{2} \leqslant X < X + \frac{\Delta X}{2}\right) \text{ и } \left(Y - \frac{\Delta Y}{2} \leqslant Y < Y + \frac{\Delta Y}{2}\right)\right]}{\Delta X \Delta Y}. \quad (2.11)$$

Из соображений размерности понятно, что $f(X, Y)$ представляет *плотность* вероятности на единицу длины массивов X и Y. Поэтому $f(X, Y)$ называют *функцией* плотности вероятности (ф. п. в.) или совместной ф. п. в. (для функций плотности более чем одной переменной).

Ф. п. в. нормируется аналогично соотношению (2.1). В частности, совместная ф. п. в. в формуле (2.11) должна удовлетворять уравнению

$$\iint_{\Omega} f(X,Y)\, dX dY = 1, \quad (2.12)$$

где Ω — пространство всех возможных значений X и Y (полная длина обоих массивов счетчиков).

Ф. п. в. могут быть функциями произвольного числа непрерывных случайных переменных. Они могут быть также смешанными функциями непрерывных и дискретных случайных переменных. Например, функция вида

$$h(X) = p_0 \delta(X) + p_1 \delta(X-1) + (1-P_0-P_1)\frac{1}{\sqrt{2\pi}} \exp(-X^2/2)$$

является функцией плотности случайной переменной, принимающей значения

$X = 0$ с вероятностью p_0;

$X = 1$ с вероятностью p_1

и с вероятностью $(1 - p_0 - p_1)$, нормально распределенной со средним значением нуль. В таких случаях условие нормировки (2.12) имеет вид интеграла по всем непрерывным параметрам и суммы по всем дискретным переменным.

2.3.2. Замена переменных. Предположим, что $f(X)$ известна и нам нужно знать плотность распределения $g(Y)$, где

$$Y = h(X). \tag{2.13}$$

При таком преобразовании интервалу $(X, X + dX)$ соответствует интервал $(Y, Y + dY)$. Если преобразование (2.13) взаимно-однозначно, то

$$g(Y)\,dY = f(X)\,dX$$

и

$$g(Y) = \frac{f(X)}{|h'(X)|}. \tag{2.14}$$

Здесь $|h'(X)|$ — абсолютное значение производной функции $h(X)$. В многомерном случае, когда X и Y — векторы, $|h'|$ — якобиан преобразования.

Если преобразование (2.13) не взаимно-однозначно, то несколько сегментов $(X, X + dX)$ преобразуются в $(Y, Y + dY)$ и суммирование следует проводить по всем таким сегментам:

$$g(Y) = \sum \frac{f(X)}{|h'(X)|}. \tag{2.15}$$

Например, если преобразование (2.13) имеет вид $Y = X^2$, плотность $g(X^2)$ равна сумме двух членов, т. е.

$$g(X^2) = \frac{f(X) + f(-X)}{2|X|}.$$

Один важный случай замены переменных рассмотрен в п. 2.4.4 когда требуется найти распределение отношения двух случайных переменных.

2.3.3. Функции распределения, маргинальные и условные распределения. Обычно физик называет ф. п. в. распределением (например, распределением массы). Статистик использует слово «распределение» для проинтегрированных ф. п. в. Мы будем называть их *функциями распределений*. Случайная переменная X характеризуется либо своей ф. п. в. $f(X)$, либо своей функцией распределения $F(X)$:

$$F(X) = \int_{X_{\text{мин}}}^{X} f(X')\,dX'. \tag{2.16}$$

Из определения функции распределения следует, что

$$F(X_{\text{мин}}) = 0, \quad F(X_{\text{макс}}) = 1,$$

если область возможных значений $X_{\text{мин}} \leqslant X \leqslant X_{\text{макс}}$. $F(X)$ является монотонной функцией X, т. е. $F(X_1) > F(X)$ для всех $X_1 > X$. Очевидно, $F(X_1)$ соответствует вероятности того, что X окажется меньше X_1, т. е.

$$F(X_1) = P(X \leqslant X_1).$$

В общем случае вероятность найти значения X и Y в некоторой области $R(X, Y)$ равна

$$P(X \text{ и } Y \text{ в } R) = \iint_{R(X,Y)} dF(X,Y).$$

Проекции распределений называются *маргинальными распределениями*. Например, проекция двумерной плотности (поверхности) $f(X, Y)$ на ось X есть маргинальная функция плотности величины X:

$$g(X) = \int_{Y_{\text{мин}}}^{Y_{\text{макс}}} f(X,Y)\,dY \tag{2.17}$$

(см. рис. 2.1).

Пример. В реакции с тремя частицами в конечном состоянии ф. п. в. является функцией двух переменных:

$$X = m_{12}^2, \quad Y = m_{23}^2,$$

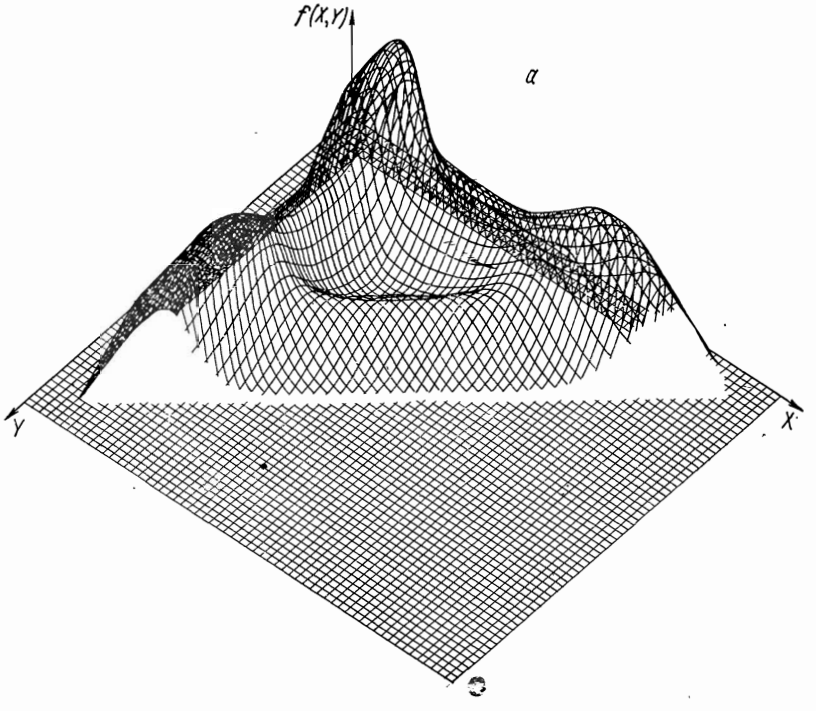
где m_{ij}^2 — квадраты инвариантных масс частиц i и j. Совместная плотность $f(X, Y)$, нормированная в соответствии с (2.12), соответствует плотности точек на графике Далица. Маргинальное распределение X соответствует проекции $f(X, Y)$ на ось X, т. е. «распределению» m_{12}^2.

Сечения распределений называют *условными распределениями*. Таким образом, нормированное сечение функции плотности $f(X, Y)$ при $X = X_0$ дает условную функцию плотности Y при заданном

$X = X_0$, которая обозначается $f(Y \mid X_0)$. По аналогии с дискретным случаем можно записать:

$$f(Y \mid X_0) = \frac{f(X_0, Y)}{\int f(X_0, Y)\, dY} = \frac{f(X_0, Y)}{g(X_0)}, \tag{2.18}$$

где $g(X)$ — маргинальное распределение переменной X.



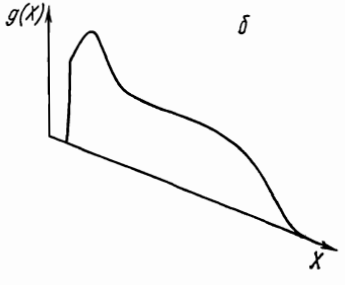


Рис. 2.1. Пример совместной ф.п.в. $f(X, Y)$ двух случайных переменных ($а$) и маргинальное распределение $g(X) = \int f(X, Y) dY$ ($б$)

В более общем случае условная плотность Y при заданном $X = h(Y)$ имеет вид

$$q[Y \mid X = h(Y)] = \frac{f[h(Y), Y]}{\int f[h(Y), Y]\, dY}. \tag{2.19}$$

2.3.4. Теорема Бейеса для непрерывных переменных. Рассмотрим совместную ф. п. в. $f(X, Y)$ для двух переменных X и Y с маргинальными функциями плотностей $g(X)$ и $h(Y)$ и условными функциями плотностей $P(X \mid Y)$ и $g(Y \mid X)$. Из определений п. 2.3.3 имеем:

$$f(X, Y) = p(X \mid Y)\, h(Y) = q(Y \mid X)\, g(X).$$

Тогда сразу же можно написать выражение для теоремы Бейеса в непрерывных переменных:

$$p(Y \mid X) = \frac{p(X \mid Y)\, h(Y)}{g(X)}. \tag{2.20}$$

2.3.5. Использование теоремы Бейеса для непрерывных переменных бейесовцами. В п. 2.2.6 мы познакомились с теоремой Бейеса для одной дискретной случайной переменной X. Приведем пример использования этой теоремы бейесовцами применительно к нескольким непрерывным переменным, например n независимым наблюдениям одной случайной переменной X.

Обозначим $f_i(X_i \mid \theta)$ ф. п. в. i-й случайной переменной, где действительное число θ — параметр, соответствующий семейству распределений f_i. Совместная плотность вероятности n случайных переменных дается выражением

$$p(\mathbf{X} \mid \theta) = \prod_{i=1}^{n} f_i(X_i \mid \theta). \tag{2.21}$$

Предполагается, что всем переменным соответствует одно и то же значение параметра. В частности, если имеются n наблюдений одной и той же переменной X, уравнение (2.21) упрощается до

$$p(\mathbf{X} \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta).$$

Заметим, что компоненты X_i все еще рассматриваются как различные переменные, хотя форма функции плотности одна и та же.

Возникает вопрос: можно ли сказать что-нибудь о значении θ, если проведено n наблюдений X_i^0, распределенных в соответствии с $f(X \mid \theta)$? Классическая школа формулирует условия так: параметр θ имеет истинное значение, которое фиксировано, но неизвестно. Можно только оценить параметр с помощью методов, которые будут обсуждаться в гл. 7 и 8. В частности, теорема Бейеса (2.20) там не используется.

Однако бейесовец не считает θ фиксированной величиной. Вместо этого он описывает свое знание о параметре посредством ф. п. в. $p(\theta)$, которая выражает *степень веры* в различные возможные значения θ. Для конструирования $p(\theta)$ он использует любое предварительное знание, которое он имеет: область возможных значений или веру в то, что некоторые значения более разумны, чем другие.

Тогда, используя теорему Бейеса (2.20), можно написать условное распределение по θ при заданном $\mathbf{X}$:

$$p(\theta \mid \mathbf{X}) = \frac{p(\mathbf{X} \mid \theta)\, p(\theta)}{\int p(\mathbf{X} \mid \theta)\, p(\theta)\, d\theta}.$$

Для определенного набора наблюдений $\mathbf{X}^0$ получается определенное условное распределение $p(\theta \mid \mathbf{X}^0)$ величины θ.

Различие между $p(\theta)$ и $p(\theta \mid \mathbf{X}^0)$ показывает, как какое-либо знание о параметре θ (степень веры) изменяется в результате наблюдения величин $\mathbf{X}^0$.

Распределение $p(\theta \mid \mathbf{X})$ суммирует все имеющееся знание о параметре θ и может быть впоследствии использовано так же.

В качестве единственного значения θ можно выбрать такое его значение, при котором $p(\theta \mid \mathbf{X})$ становится максимальным. В главе, посвященной теории решений (см. гл. 6), обсуждаются другие способы использования $p(\theta \mid \mathbf{X})$.

§ 2.4. СВОЙСТВА РАСПРЕДЕЛЕНИЙ

Дадим определение многим полезным числовым характеристикам распределения вероятностей: *оператору математического ожидания*, *среднему*, *дисперсии*, *асимметрии*, *эксцессу*, *семиинвариантам и другим моментам*. Познакомимся также со вспомогательными функциями, помогающими находить эти величины: *характеристической* функцией, *производящей функцией семиинвариантов и производящей функцией вероятности*.

2.4.1. Математическое ожидание, среднее и дисперсия. Функции плотности вероятности используются как весовые функции для получения информации о случайных переменных. Если $g(X)$ представляет собой некоторую функцию случайной переменной X с плотностью $f(X)$, то математическое ожидание $g(X)$ есть число

$$E(g) = \int_{\Omega} g(X) f(X)\, dX, \tag{2.22}$$

где Ω — все пространство переменных X.

Ожидание E есть *линейный оператор*:

$$E[ag(X) + bh(X)] = aE[g(X)] + bE(h(X)]. \tag{2.23}$$

Заметим, что $E[g(X)]$ не является функцией X (поскольку по X проведено интегрирование в заданных пределах).

Математическое ожидание самой случайной величины X называется *средним* по плотности $f(X)$ или *ожидаемым значением* X для плотности $f(X)$ и обозначается μ (иногда пишут X или $\langle X \rangle$):

$$\mu = \int X f(X)\, dX. \tag{2.24}$$

Математическое ожидание функции $(X-\mu)^2$ называется *дисперсией* $D(X)$ плотности $f(X)$:

$$D(X)=\sigma^2=E[(X-\mu)^2]=E[X^2-2\mu X+\mu^2]=E[X^2]-\mu^2=$$
$$=\int [X-\mu]^2 f(X)\,dX. \qquad (2.25)$$

$D(X)$ также является числом, а не функцией X. Величина σ называется *стандартным отклонением*. Это просто определение и никаких утверждений о связи между вероятностным содержанием и стандартным отклонением не может быть сделано. Для таких утверждений необходимо знать $f(X)$; в главе о доверительных интервалах (см. гл. 9) даны соответствующие примеры.

Ожидаемое значение $E(X)$ является мерой расположения распределения, тогда как дисперсия — мерой протяженности распределения на пространстве X. В п. 2.4.6 введены другие характеристики формы распределения.

Заметим, что среднее распределения не всегда существует. В качестве примера можно привести распределение Коши* с плотностью вероятности

$$f(X)=\frac{1}{\pi(1+X^2)},$$

поскольку интеграл

$$\int_{-\infty}^{\infty} X f(X)\,dX=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{X\,dX}{1+X^2}$$

не определен. Дисперсия переменной X, $D(X)$ бесконечна. Необходимо заметить, что $f(X)$ является формулой Брейта—Вигнера, часто встречающейся в физике.

Однако, если мы рассмотрим условное распределение X при условии, что $-A \leqslant X \leqslant A$, то плотность вероятности приобретает вид

$$g(X)=\frac{f(X)}{2\operatorname{arctg} A}.$$

Для этого распределения среднее и дисперсия X существуют:

$$E(X\,|-A \leqslant X \leqslant A)=0$$

и

$$D(X\,|-A<X<A)=\frac{A}{\operatorname{arctg} A}-1.$$

2.4.2. Смешанный второй момент и корреляция. Понятие математического ожидания (2.22) легко обобщить на многомерный случай. В частности, ожидание функции $g(X, Y)$ двух случайных переменных при заданной совместной плотности $f(X, Y)$

$$E[g(X, Y)]=\iint g[X, Y]\,dX dY,$$

---

* Свойства этого распределения см. в п. 4.2.11.

а среднее и дисперсии распределения $f(X, Y)$

$$\left.\begin{aligned}
&\mu_X = E(X) = \iint X f(X, Y)\, dX dY = \int X \int f(X, Y)\, dX;\\
&\mu_Y = E(Y) = \iint Y f(X, Y)\, dX dY;\\
&\sigma_X^2 = E[(X - \mu_X)^2];\\
&\sigma_Y^2 = E[(Y - \mu_Y)^2].
\end{aligned}\right\} \qquad (2.26)$$

Важными числовыми характеристиками совместной плотности являются *смешанный второй момент*, определяемый равенством

$$D(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y), \qquad (2.27)$$

и *коэффициент корреляции*

$$\rho(X, Y) = \frac{D(X, Y)}{\sigma_X \sigma_Y}. \qquad (2.28)$$

Коэффициент корреляции изменяется в интервале от -1 до $+1$. Чтобы доказать это, рассчитаем дисперсию линейной комбинации X и Y. Очевидно, что дисперсия является положительной величиной, поскольку она получается в результате интегрирования положительной функции. Поэтому

$$D(\alpha X + Y) = \alpha^2 D(X) + D(Y) + 2\alpha D(X, Y) \geqslant 0.$$

Неравенство выполняется для всех α; отсюда следует, что

$$[D(X, Y)]^2 - D(X)\, D(Y) \leqslant 0$$

или

$$-1 \leqslant \rho \leqslant 1.$$

(Фактически это доказательство леммы Шварца.)

Говорят, что случайные переменные $(X_1, \ldots, X_n)$ взаимно независимы, если и только если их совместная плотность полностью факторизуема, т. е.

$$f(\mathbf{X}) = f(X_1, \ldots, X_n) = f_1(X_1)\, f_2(X_2) \ldots f_n(X_n).$$

Мы уже использовали этот результат в п. 2.3.5 при написании совместной плотности множества наблюдаемых переменных. Ожидание произведения взаимно независимых переменных X и Y:

$$E(XY) = \int XY f(X, Y)\, dX dY = \int X f_1(X)\, dX \times$$

$$\times \int Y f_2(Y)\, dY = E(X)E(Y). \qquad (2.29)$$

Сравнивая этот результат с равенствами (2.27) и (2.28), видим, что для *независимых переменных смешанный второй момент* и *корреляция равны нулю*. Обратное утверждение не всегда справедливо.

О случайных переменных, для которых $\rho = 0$, говорят как о *некоррелированных* (но не обязательно независимых).

Например, предположим, что X симметрично распределено относительно нуля с функцией плотности $f(X)$. Пусть $Y = X^2$. Ясно, что X и Y действительно зависимы, поскольку они функционально связаны. Но корреляция между X и Y равна нулю. Именно $E(X) = 0$ и

$$E(Y) = \int X^2 f(X)\, dX = D(X) = \sigma^2.$$

Смешанный второй момент для X и Y равен

$$D(X, Y) = E[X(X^2 - \sigma^2)] = E(X^3) - \sigma^2 E(X) = 0,$$

так как $f(X)$ симметрична. Отсюда следует, что X и Y не коррелированы.

Обычно говорят «некоррелированные переменные», подразумевая «независимые переменные». В статистике понятие «некоррелированные» гораздо более слабое, чем «независимые», поскольку первое следует из последнего, но не наоборот. Если случайных переменных больше, чем две, смешанный второй момент и корреляция также могут быть определены для каждого маргинального двумерного совместного распределения (по X_i и X_j).

Матрица с элементами $D(X_i, X_j)$ называется *матрицей вторых моментов* или иногда *дисперсионной* матрицей или *матрицей ошибок*. Очевидно, диагональные элементы являются дисперсиями $\sigma^2_{X_i}$ (если для матрицы вторых моментов не существует обратной матрицы, то это означает, что имеется по крайней мере одна линейная комбинация между X_i).

Рассмотрим корреляции $\rho(X_k, Y)$ между переменной X_k и некоторой произвольной линейной комбинацией Y всех других переменных X_i, $i \neq k$. *Сводный коэффициент корреляции* ρ_k, определенный как максимально возможное значение $\rho(X_k, Y)$, служит мерой общей корреляции между X_k и всеми другими переменными. Если $\rho_k = 0$, то переменная X_k не коррелирована со всеми другими переменными. Если $\rho_k = 1$, X_k полностью коррелирована по крайней мере с одной линейной комбинацией других переменных. Сводный коэффициент корреляции выражается через диагональные элементы D_{kk} матрицы вторых моментов и диагональные элементы обратной ей матрицы $(D^{-1})_{kk}$:

$$\rho_k = \sqrt{1 - [D_{kk}(D^{-1})_{kk}]^{-1}}.$$

2.4.3. Линейные функции случайных переменных. Применим введенные ранее понятия к важному случаю линейных случайных переменных. Особый интерес представляет выражение для среднего нескольких наблюдений случайной величины.

Математическое ожидание *линейной функции* нескольких случайных переменных $X_1, \ldots, X_n$ равно:

$$E\left\{\sum_{i=1}^{n} a_i X_i\right\} = \sum_{i=1}^{n} a_i E\{X_i\} = \sum_{i=1}^{n} a_i \mu_{X_i}. \tag{2.30}$$

Это можно показать с помощью соотношений (2.22), (2.26) и на основе линейности оператора ожидания [см. соотношение (2.23)]. Дисперсия равна:

$$D\left\{\sum_{i=1}^{n} a_i X_i\right\} = \sum_{i=1}^{n} a_i^2 D(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j D(X_i, X_j). \tag{2.31}$$

Если X_i взаимно не коррелированы, соотношение (2.31) приобретает простую форму:

$$D\left\{\sum_{i=1}^{n} a_i X_i\right\} = \sum_{i=1}^{n} a_i^2 D(X_i). \tag{2.32}$$

Таким образом, дисперсия линейной функции является квадратичной функцией коэффициентов разложения.

Для того чтобы показать важность соотношений (2.30) — (2.32), рассмотрим случай, когда X_i представляет собой n различных испытаний в одном и том же эксперименте. В этом случае *среднее* (часто называемые *средним выборки*) равно среднему значению наблюдений X_i:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2.33}$$

Ожидание одного испытания предполагается равным

$$E(X_i) = \mu \text{ для всех } i$$

и дисперсия

$$D(X_i) = \sigma^2 \text{ для всех } i.$$

Заметим, что μ (часто называемое математическим ожиданием) не совпадает с $\overline{X}$. Тогда *математическое ожидание среднего n наблюдений* равно:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu, \tag{2.34}$$

а дисперсия среднего

$$D(\overline{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) +$$

$$+\frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D(X_i, X_j) = \frac{\sigma^2(X)}{n} + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D(X_i, X_j). \tag{2.35}$$

Если испытания *независимы*, $D(X_i, X_j) = 0$ для любой пары (i, j) и член с двойной суммой исчезает. В результате мы приходим к известной формуле

$$\sigma(\overline{X}) = \sqrt{D(\overline{X})} = \frac{\sigma(X)}{\sqrt{n}}, \tag{2.36}$$

из которой следует, что (при заданных условиях) стандартное отклонение среднего уменьшается с ростом n как $n^{-1/2}$.

Однако, если измерения коррелированы, стандартное отклонение среднего с ростом n уменьшается не как $n^{-1/2}$. Это зависит от коэффициента корреляции. Рассмотрим пример двух случайных переменных с одинаковыми средними и дисперсиями.

Если $\rho \neq 0$, получим в общем случае

$$\sigma^2(\overline{X}) = \frac{1}{2}\sigma^2(X) + \frac{1}{2}\sigma^2(X)\rho = \frac{1}{2}\sigma^2(X)(1+\rho).$$

Поскольку ρ изменяется от -1 до $+1$, дисперсия $\sigma^2(\overline{X})$ принимает любые значения от 0 до $\sigma^2(X)$. Рассмотрим предельные случаи.

Максимальная положительная корреляция, т. е. $\rho = +1$. В этом случае $\sigma^2(\overline{X}) = \sigma^2(X)$, среднее двух измерений равно среднему одного испытания. Очевидно, коэффициент корреляции $\rho = +1$ означает, что второе измерение идентично первому, т. е. $\overline{X} = X_1 = X_2$.

Максимальная отрицательная корреляция, т. е. $\rho = -1$. Тогда $\sigma^2(\overline{X}) = 0$; это означает, что $\overline{X}$ уже не случайная, а вполне определенная величина. Чтобы получить $\rho = -1$, необходимо иметь (в соответствии с определением ρ) $(X_2 - \mu) = -(X_1 - \mu)$. Поэтому $\overline{X} = \mu$.

Из соотношения (2.35) следует, что *дисперсия среднего может быть очень малой, если возможно производить измерения X_i таким образом, что наблюдаемые ошибки отрицательно коррелированы друг с другом* (см. пример в § 5.5).

2.4.4. Отношение случайных переменных. Пусть X и Y распределены независимо с функциями плотностей $f(X)$ и $g(Y)$ соответственно. Произведем следующую замену переменных:

$$U = X/Y; \quad V = Y,$$

Кроме того, предположим, что область изменения Y не содержит нуля. Тогда совместная функция плотности $h(U, V)$ удовлетворяет соотношению

$$h(U,V)\,dUdV = f(X)\,g(Y)\,dXdY = f(UV)\,g(V)\left\|\frac{\partial(X,Y)}{\partial(U,V)}\right\|dUdV = f(UV)\,g(V)\,V\,dU\,dV.$$

Функция плотности переменной U

$$p(U) = \int_0^\infty h(U,V)\,dV = \int_0^\infty f(UV)\,g(V)\,V\,dV.$$

Если Y может быть и положительным и отрицательным, совместная функция плотности становится равной

$$h(U, V) = f(UV)\,g(V)\,|V|, \quad V < 0;$$
$$h(U, V) = f(UV)\,g(V)\,V, \quad V \geqslant 0$$

и функция плотности отношения

$$p(U) = \int_0^\infty f(UV)\,g(V)\,V\,dV - \int_{-\infty}^0 f(UV)\,g(V)\,V\,dV. \qquad (2.37)$$

Применяя эту формулу к случаю, когда X и Y распределены по нормальному закону* со средним 0 и дисперсией 1, т. е.

$$f(X)=g(X)=(1/\sqrt{2\pi})\exp(-X^2/2),$$

получаем

$$p(U)=\frac{1}{\pi(1+U^2)}.$$

Таким образом, отношение двух нормально распределенных величин с нулевым средним следует закону распределения Коши (см. п. 4.2.11) с неопределенным средним и бесконечной дисперсией.

Точно так же, применяя (2.37) к общему случаю нормальных переменных $N(\mu_X, \sigma_X^2)$ и $N(\mu_Y, \sigma_Y^2)$, получим ф. п. в. для $Z = Y/X$

$$p(Z)=\frac{1}{\pi}\frac{\sigma_X/\sigma_Y}{\left(1+Z^2\frac{\sigma_X^2}{\sigma_Y^2}\right)}\exp\left[-\frac{1}{2}\left(\frac{\mu_X^2}{\sigma_X^2}+\frac{\mu_Y^2}{\sigma_Y^2}\right)\right]\times$$

$$\times\left[1+b\,e^{\frac{1}{2}b^2}\int_0^b e^{-\frac{1}{2}V^2}dV\right], \qquad (2.38)$$

где

$$b=\frac{(\mu_X/\sigma_X)+Z(\sigma_X/\sigma_Y)(\mu_Y/\sigma_Y)}{\sqrt{1+Z^2(\sigma_X/\sigma_Y)^2}}.$$

Таким образом, $p(Z)$ имеет форму распределения Коши, умноженного на $f(Z)$. Можно показать, что дисперсия этого распределения также бесконечна, поскольку $f(Z)$ ведет себя, как $1/Z$ или как константа при $Z \to \infty$, в зависимости от того, равно ли нулю μ_Y или нет.

Если μ_X/σ_X достаточно велико, так что можно брать практически только положительные значения X, тогда функция плотности $Z = Y/X$ равна

$$p(Z)=\frac{1}{\sqrt{2\pi}}\frac{\mu_X\sigma_Y^2+\mu_Y\sigma_X^2 Z}{(\sigma_Y^2+\sigma_X^2 Z^2)^{3/2}}\exp\left[-\frac{1}{2}\frac{(\mu_Y-\mu_X Z)^2}{\sigma_Y^2+\sigma_X^2 Z^2}\right]. \qquad (2.39)$$

Отсюда следует, что переменная

$$(\mu_Y-\mu_X Z)/\sqrt{(\sigma_Y^2+\sigma_X^2 Z^2)}$$

распределена нормально с нулевым средним и единичной дисперсией [1].

2.4.5. Приближенная формула для дисперсии. Выведем приближенную формулу для дисперсии функций случайных переменных. Пусть мы имеем функцию $f(X_1, \ldots, X_n)$, где $X_1, \ldots, X_n$ — случайные переменные со средним $\mu_1, \ldots, \mu_N$. Предположим также, что функция может быть разложена в ряд Тейлора вблизи $\mu_1, \ldots, \mu_n$. Тогда

$$f(X_1, \ldots, X_n)=f(\mu_1, \ldots, \mu_n)+\sum_{i=1}^{n}(X_i-\mu_i)\frac{df}{\partial\mu_i}+O[(X_i-\mu_i^2)],$$

где

$$\frac{\partial f}{\partial\mu_i}=\frac{\partial f_i}{\partial X_i}\Big|_{\mu_i}.$$

* Мы будем говорить о свойствах нормального распределения в п. 4.2.1.

Опуская члены порядка, большего чем 1 по $\Delta X_i = (X_i - \mu_i)$, получаем для ожидания

$$E(f) \cong f(\mu_1, \ldots, \mu_n).$$

Соответственно для дисперсии f:

$$D[f(X_1, \ldots, X_n)] = E[f - E(f)]^2 = E\left[\sum_{i=1}^{n}(X_i - \mu_i)\frac{\partial t}{\partial \mu_i}\right]^2 =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\partial f}{\partial \mu_i}\frac{\partial f}{\partial \mu_j}E(X_i - \mu_i)(X_j - \mu_j)\right] \simeq$$

$$\simeq \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial f}{\partial \mu_i}\frac{\partial f}{\partial \mu_j}E(X_i, X_j). \qquad (2.40)$$

Заметим, что в матричных обозначениях соотношение (2.40) приобретает простую форму ADA^T, где D — матрица вторых моментов. Если переменные X_i независимы, (2.40) сводится к

$$D[f(X_1, \ldots, X_n)] \simeq \sum_{i=1}^{n}\left(\frac{\partial f}{\partial \mu_i}\right)D(X_i). \qquad (2.41)$$

Точно так же смешанный второй момент двух функций $f(X_1, \ldots, X_n)$ и $g(X_1, \ldots, X_n)$ приближенно равен

$$D(f, g) \simeq \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial f}{\partial \mu_i}\frac{\partial g}{\partial \mu_j}D(X_i, X_j) \simeq$$

$$\simeq \sum_{i=1}^{n}\frac{\partial f}{\partial \mu_i}\frac{\partial g}{\partial \mu_i}D(X_i), \qquad (2.42)$$

если X_i независимы.

Применяя (2.40) для *отношения двух случайных переменных* X и Y, мы имеем приближенно

$$D\left(\frac{X}{Y}\right) \simeq \left(\frac{\mu_X}{\mu_Y}\right)^2\left[\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2} - 2\frac{\rho_{XY}\sigma_X\sigma_Y}{\mu_X\mu_Y}\right], \qquad (2.43)$$

где μ_X, μ_Y, σ_X^2, σ_Y^2 — средние и дисперсии X и Y, а ρ — коэффициент корреляции.

Отметим, что полученное приближение справедливо только при довольно серьезных предположениях, а именно

$$|\mu_X| \gg \sigma_X; \quad |\mu_Y| \gg \sigma_Y \quad \text{и } Y \neq 0.$$

В частности, если X и Y распределены нормально, $D(X \mid Y) = \infty$ (см. п. 2.4.4).

2.4.6. Моменты. Ожидание переменной X^n с данной плотностью $f(X)$ называется *n-м моментом* $f(X)$ (или моментом n-го порядка). В общем случае

$\mu_n' = E(X^n)$ — *n-й алгебраический момент*;

$\mu_n = E\{[X - E(X)]^n\}$ — *n-й центральный момент*;

$\nu'_n = E(|X|^n)$ — n-й *абсолютный момент*;

$\nu_n = E[|X - E(X)|^n]$ — n-й *абсолютный центральный момент* плотности $f(X)$.

В частности, среднее μ является первым алгебраическим моментом, а дисперсия — вторым центральным моментом $f(X)$.

Моменты многомерных распределений определяются аналогично. Например, алгебраический момент $f(X, Y)$ порядка m по переменной X и порядка n по переменной Y равен

$$\mu'_{mn} = E(X^m Y^n).$$

Распределение может быть несимметричным относительно своего среднего. Степень асимметрии распределения задается коэффициентом β_1, который определен как

$$\beta_1 = \mu_3^2/\mu_2^3.$$

Очевидно, у симметричного распределения μ_3 и, следовательно, β_1 равны нулю (обратное утверждение не всегда справедливо).

Другим полезным коэффициентом для описания формы распределения служит

$$\beta_2 = \mu_4/\mu_2^2.$$

Для многих целей более удобными коэффициентами являются:

$\gamma_1 = \sqrt{\beta_1} = \mu_3/\mu_2^{3/2}$ — *асимметрия*

и $\gamma_2 = \beta_2 - 3$ — *эксцесс*.

Коэффициент «эксцесс» определен в таком виде специально, чтобы он для нормального распределения был равен нулю (см. п. 4.2.1).

§ 2.5. ХАРАКТЕРИСТИЧЕСКАЯ ФУНКЦИЯ

Фурье-преобразование ф. п. в. называется *характеристической функцией* случайной переменной. Она имеет много полезных и важных свойств. Например, распределение сумм случайных переменных проще всего найти с помощью характеристических функций. Для многих распределений моменты также рассчитать проще, если использовать свойства этих функций.

Частными случаями характеристических функций являются производящие функции вероятности, моментов и семиинвариантов.

2.5.1. Определение и свойства. Для заданной случайной переменной X с плотностью $f(X)$ характеристическая функция определяется как

$$\left.\begin{aligned} &\zeta_X(t) = E[\exp(itX)] && (t\text{—реальное});\\ &\zeta_X = \int_{-\infty}^{\infty} \exp(itX) f(X)\, dX && (X\text{—непрерывна});\\ &\zeta_X = \sum_k p_k \exp(itX_k) && (X\text{—дискретна}). \end{aligned}\right\} \quad (2.44)$$

Характеристическая функция $\zeta_X(t)$ полностью определяет распределение вероятностей случайной переменной. В частности, если $F(X)$ непрерывна всюду и $dF(X) = f(X)\, dX$, то

$$f(X) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \zeta_X(t) \exp(-iXt)\, dt.$$

Функция $\zeta(t)$ обладает следующими свойствами:

$$\zeta(0) = 1; \qquad |\zeta(t)| \leqslant 1$$

и существует для всех t.

Если a и b — константы, то характеристическая функция переменной $aX + b$ равна:

$$\zeta_{aX+b}(t) = \exp(ibt)\, \zeta_X(at),$$

что следует из

$$E[\exp(it(aX+b))] = E[\exp(itb)\exp(iatX)] =$$
$$= \exp(itb)\, \zeta_X(at).$$

Если X и Y — *независимые* случайные переменные с характеристическими функциями $\zeta_X(t)$, $\zeta_Y(t)$, то характеристическая функция суммы $(X + Y)$

$$\zeta_{X+Y}(t) = \zeta_X(t)\, \zeta_Y(t),$$

что следует из

$$\zeta_{X+Y}(t) = E[\exp(it(X+Y))] = E(\exp(itX)\exp(itY)) =$$
$$= E(\exp(itX))\, E(\exp(itY)) = \zeta_X(t)\, \zeta_Y(t).$$

Это служит хорошим примером удобства использования характеристических функций: характеристическая функция суммы независимых переменных имеет гораздо более простой вид, чем соответствующая ф. п. в. суммы:

$$f(Z = X+Y) = \int_{-\infty}^{\infty} f_X(Z-t)\, f_Y(t)\, dt. \quad (2.45)$$

В общем случае характеристическая функция суммы n независимых случайных переменных равна:

$$\zeta_{X_1+\ldots+X_n}(t)=\prod_{i=1}^{n}\zeta_{X_i}(t). \tag{2.46}$$

Связь между характеристической функцией и моментами распределения может быть найдена с помощью формального разложения:

$$\zeta_X(t)=E[\exp(itX)]=E\left[\sum_{r=0}^{\infty}\frac{(itX)^r}{r!}\right]=\sum_{r=0}^{\infty}\frac{(it)^r}{r!}E(X^r)=$$

$$=\sum_{r=0}^{\infty}\frac{(it)^r}{r!}\mu_r', \tag{2.47}$$

т. е. моменты μ_r' являются коэффициентами при членах $(it)^r/r!$ в разложении $\zeta(t)$.

Для распределения со средним μ и дисперсией σ^2 можно получить приближенную характеристическую функцию, если использовать разложение в ряд Тейлора:

$$\zeta(t)=1+i\mu t+\frac{1}{2}(\sigma^2+\mu^2)(it)^2+0(t^2).$$

Если среднее разложения равно μ, то с помощью преобразования

$$\exp(-i\mu t)\,\zeta(t)=E[\exp(it(X-\mu))]=\zeta_{X-\mu}(t) \tag{2.48}$$

можно рассчитать центральные моменты относительно среднего:

$$\left.\begin{aligned}\mu_r'&=\frac{1}{i^r}\left(\frac{d}{dt}\right)^r\zeta(t)\,|_{t=0};\\ \mu_r&=\frac{1}{i^r}\left(\frac{d}{dt}\right)^r\exp(-i\mu t)\,\zeta(t)\,|_{t=0}.\end{aligned}\right\} \tag{2.49}$$

Приведем два примера использования характеристических функций.

1. Предположим, что X распределено нормально — $N(\mu, \sigma^2)$*. Характеристическая функция

$$\zeta(t)=\frac{1}{\sqrt{2\pi}\,\sigma}\int_{-\infty}^{\infty}\exp(iXt)\exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]dX=$$

$$=\exp\left(i\mu t-\frac{1}{2}t^2\sigma^2\right).$$

Для вычисления моментов используем соотношение (2.48):

$$\exp(-i\mu t)\,\zeta(t)=\exp\left(-\frac{1}{2}t^2\sigma^2\right)=\sum_{r=0}^{\infty}\frac{(it)^{2r}}{r!}\left(\frac{\sigma^2}{2}\right)^r.$$

* Свойства нормального распределения изучены в п. 4.2.1.

Сравнив с (2.47), получим:

$$\mu_{2r-1}=0; \qquad \mu_{2r}=\frac{(2r)!}{r!}\left(\frac{\sigma^2}{2}\right)^r.$$

2. Характеристическая функция распределения Коши*

$$\zeta(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\exp(itX)}{1+X^2}\,dX=\exp(-|t|)$$

показана на рис. 2.2. Поскольку $\zeta(t)$ не имеет производных при $t=0$, то из уравнения (2.49) следует, что это распределение не имеет моментов. В частности, среднего распределения не существует.

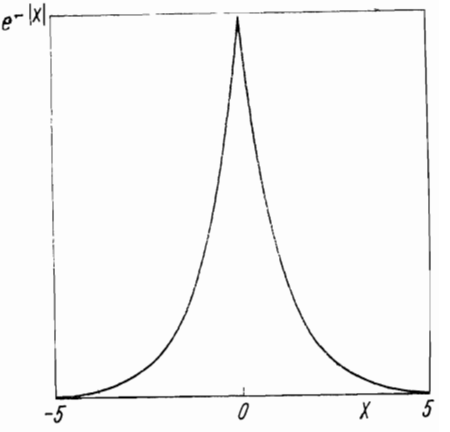


Рис. 2.2. Характеристическая функция распределения Коши $e^{-|X|}$ для $-5\leqslant X\leqslant 5$

2.5.2. Семиинварианты. Для случайной переменной X с характеристической функцией $\zeta(t)=E(\exp(itX))$ производящая функция семиинвариантов определяется как

$$K(t)=\ln\zeta(t). \tag{2.50}$$

Разлагая $K(t)$ в ряд по (it), получаем

$$K(t)=\sum_{r=0}^{\infty}K_r\frac{(it)^r}{r!}. \tag{2.51}$$

Коэффициенты K_r называются *семиинвариантами* или полуинвариантами распределения переменной X. Семиинварианты — это совокупность констант, служащих полезной мерой свойств распределения. В некоторых случаях знание семиинвариантов эквивалентно знанию распределения. Как следует из определения, они тесно связаны с моментами распределения. Для первых пяти се-

* О свойствах распределения Коши см. в п. 4.2.11.

мииинвариантов справедливы такие формулы [1]:

$$K_1 = \mu \equiv \text{среднее};$$
$$K_2 = \mu_2 \equiv \text{дисперсия};$$
$$K_3 = \mu_3;$$
$$K_4 = \mu_4 - 3\,\mu_2^2;$$
$$K_5 = \mu_5 - 10\,\mu_3\mu_2.$$

Для двух независимых случайных переменных X, Y с производящими функциями семиинвариантов $K_X(t)$ и $K_Y(t)$ производящая функция семиинвариантов суммы $X + Y$ равна $K_X(t) + K_Y(t)$, т. е. *семиинварианты суммы независимых случайных переменных равны суммам семиинвариантов.*

2.5.3. Производящая функция вероятности. Характеристическая функция может быть также определена для дискретной случайной переменной r. Пусть *функция вероятности* переменной r равна $P(r) = p_r$; тогда для характеристической функции получаем в соответствии с (2.44):

$$\zeta_r(t) = \sum_{r=0}^{\infty} p_r \exp(itr).$$

Однако удобнее переписать эту формулу с помощью замены $Z = \exp(it)$. Функция

$$G(Z) = E(Z^r) = \sum_{r=0}^{\infty} p_r Z^r \tag{2.52}$$

называется *производящей функцией вероятности* для r. $G(Z)$ регулярна по крайней мере для $|Z| < 1$, поскольку

$$p_r \geqslant 0 \quad \text{и} \sum_{r=0}^{\infty} p_r = 1.$$

Легко видеть, что значения производных $G(Z)$ в точке $Z = 1$ связаны с моментами. Например, среднее μ равно

$$G'(1) = G'(Z)|_{Z=1} = \sum_{r=0}^{\infty} rZ^{r-1} p_r|_{Z=1} = \sum_{r=0}^{\infty} rp_r = E(r) = \mu.$$

Используя вторую производную

$$G''(1) = G''(Z)|_{Z=1} = \sum_{r=0}^{\infty} r(r-1) Z^{r-2} p_r|_{Z=1} =$$

$$= \sum_{r=0}^{\infty} r(r-1) p_r = E[r(r-1)] = E(r^2) - E(r)$$

и определение дисперсии (2.25), получаем следующее выражение для дисперсии:

$$D(r) = G''(1) + G'(1) - [G'(1)]^2.$$

Подобная процедура — зачастую наиболее простой способ получения среднего и дисперсии дискретного распределения.

2.5.4. Суммы случайного числа случайных переменных. Рассмотрим сумму S_N независимых случайных переменных X

$$S_N = \sum_{i=1}^{N} X_i,$$

где само N — случайная переменная с производящей функцией вероятности $g(s) = \sum_{i=0}^{\infty} g_i s^i$. Предположим, что каждое X имеет производящую функцию вероятности в виде $f(s) = \sum_{j=0}^{\infty} f_j s^j$.

Тогда вероятность того, что S_N есть j, равна:

$$h_j = P(S_N = j) = \sum_{n=0}^{\infty} P(S_n = j \mid N = n)\, P(N = n). \qquad (2.53)$$

Для фиксированного n сумма S_n имеет такую производящую функцию вероятности:

$$\{f(s)\}^n = \sum_{j=0}^{\infty} f_j^{(n)} s^j,$$

т. е. $P(S_n = j \mid N = n) = f_j^{(n)}$. Следовательно, можно написать

$$h_j = \sum_{n=0}^{\infty} f_j^{(n)} g_n.$$

Тогда составная производящая функция вероятности для S_N

$$h(s) = \sum_{j=0}^{\infty} h_j s^j = \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} f_j^{(n)} g_n s^j = \sum_{n=0}^{\infty} g_n \{f(s)\}^n = g\{f(s)\}. \qquad (2.54)$$

Пример. Рассмотрим случай, когда N распределено по закону Пуассона (см. п. 4.1.3 и 4.1.4) со средним λ, а X_i распределены по тому же закону со средним μ.

Производящая функция для распределения Пуассона

$$g(s) = \exp(-\lambda + \lambda s).$$

В соответствии с (2.54) производящая функция S_N

$$h(s) = \exp[-\lambda + \lambda \exp(-\mu + \mu s)]. \qquad (2.55)$$

Это частный случай *составного распределения Пуассона* (см. п. 4.1.3 и 4.1.4), имеющего производящую функцию вероятности:

$$h(s, t) = \exp(-\lambda t + \lambda t f(s)).$$

Используя (2.55), легко показать в соответствии с результатами предыдущего параграфа, что

$$E(S_N) = \lambda\mu$$

и

$$D(S_N) = \lambda\mu(1 + \mu).$$

Для того чтобы найти вероятность (2.53), нужно посчитать

$$h_k = \frac{1}{k!} \frac{d^k h(s)}{ds^k}\Big|_{s=0}.$$

Можно написать, что

$$h_k = \left[\sum_{r=0}^{k} v_r^{(k)}\right] \exp(-\lambda + \lambda \exp(-\mu)),$$

где

$$v_r^{(k)} = \frac{r\mu}{k} v_r^{(k-1)} + \frac{\lambda\mu}{k} \exp(-\lambda) v_{r-1}^{(k-1)}, \quad r = 0, \ldots, k, \tag{2.56}$$

при этом $v_0^{(0)} = 1$ и $v_r^k = 0$, если $r < 0$ или $r > k$.

Точно так же относительно просто рассчитать распределение вероятности S_N для любого другого случая.

ГЛАВА 3

СХОДИМОСТЬ И ЗАКОН БОЛЬШИХ ЧИСЕЛ

В этой главе мы ознакомимся с фундаментальными теоремами, на основе которых ниже получены результаты, используемые на практике. Читатель, интересующийся только практическими применениями, может опустить § 3.1 и 3.2.

§ 3.1. ТЕОРЕМА ЧЕБЫШЕВА И ЕЕ СЛЕДСТВИЕ

Эта теорема со своим следствием, называемым неравенством Бьенеме—Чебышева, очень часто используется для доказательства теорем сходимости.

3.1.1. Теорема Чебышева. Пусть $h(X)$ — неотрицательная функция случайной переменной X. Тогда в соответствии с *теоремой Чебышева*, вероятность того, что функция $h(X)$ превысит некоторое значение K, меньше некоторого вполне определенного значения

$$P[h(X) \geqslant K] \leqslant \frac{E[h(X)]}{K} \tag{3.1}$$

для любого $K > 0$ и не зависит от формы $h(X)$, если только известно ее ожидание.

Доказательство очень просто. Действительно, ожидание

$$E[h(X)] = \int h(X) f(X)\, dX$$

больше или равно

$$K \int_R f(X)\, dX = KP[h(X) \geqslant K].$$

Здесь R — область значений X, на которой $h(X) \geqslant K$.

К несчастью, теорема слишком обща для того, чтобы быть полезной при практических вычислениях, поскольку верхний предел, который она устанавливает, обычно может быть уточнен только при использовании конкретного вида функции $h(X)$. Однако она полезна для иллюстрации теорем сходимости.

3.1.2. Неравенство Бьенеме—Чебышева. Произведем следующие замены в равенстве (3.1):

$$h(X) \longrightarrow [X - E(X)]^2;$$
$$K \longrightarrow (k\sigma)^2.$$

В результате получим

$$P(|X - E(X)| \geqslant k\sigma) \leqslant 1/k^2, \tag{3.2}$$

так называемое *неравенство* Бьенеме—Чебышева. Оно дает верхнюю границу для вероятности того, что будет наблюдаться превышение некоторого заданного числа стандартных отклонений независимо от формы функции при условии, что известна дисперсия. Подобно теореме Чебышева, оно имеет прежде всего чисто теоретическое значение. Однако в тех случаях, когда информация о функции распределения незначительна, оно находит и практическое применение.

Если имеется некая информация о ф. п. в., то могут быть получены более строгие ограничения. Например, если ф. п. в. непрерывна и имеет только один максимум в точке X_0, то неравенство таково:

$$P(|X - X_0| \geqslant k\tau) \leqslant 4/9\, k^2, \tag{3.3}$$

где

$$\tau^2 = \sigma^2 + (X_0 - \mu)^2.$$

Неравенство (3.3) может быть переписано в таком виде:

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{4}{9} \frac{1 + S^2}{(k - |S|)^2},$$

где

$$S = (\mu - X_0)/\sigma.$$

Подобно этому, если известен четный момент более высокого порядка, скажем μ_4, то можно показать, что предел равен [10]

$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{\gamma_2 + 2}{(k^2 - 1)^2 + \gamma_2 + 2}. \tag{3.4}$$

Это неравенство получается, если обозначения в (3.1) имеют такой смысл:

$$h(X) = 1 + \frac{\sigma^2 (k^2 - 1)[(X - \mu)^2 - k^2 \sigma^2]}{\mu_4 + k^4 \sigma^4 - 2k^2 \sigma^4};$$
$$K = 1;$$
$$\gamma_2 = (\mu_4/\sigma^4) - 3.$$

Если полученные выше формулы применить к нормальному распределению*, то получим результаты, приведенные в табл. 3.1.

---

* Свойства нормального распределения будут рассматриваться в п. 4.2.1.

Таблица 3.1

Границы для $P(|X-E(X)| \leqslant k\sigma)$ для нормального распределения

k	1	2	3	4
Неравенство (3.2)	1	1/4	1/9	1/16
Неравенство (3.3)	4/9	1/9	4/81	1/36
Более сильное неравенство (3.4)	1	2/11	2/66	2/277
Точное значение	0,317	0,0555	0,0027	0,000063

§ 3.2. СХОДИМОСТЬ

Обычно физики считают, что понятия «сходимость» и «предельные процессы» родственны соответствующим понятиям в обычных математических методах. В статистике понятие сходимости сложнее, поскольку дополнительно приходится иметь дело со случайными флуктуациями и вероятностями. Вследствие этого приходится вводить различные виды сходимости [11].

3.2.1. Сходимость по распределению. Это наиболее слабый вид сходимости. Рассмотрим последовательность $(X_1, \ldots, X_n)$ случайных переменных с функциями распределений $F_1(X), \ldots, F_n(X)$. Тогда говорят, что последовательность X_n *сходится по распределению* (при $n \to \infty$) к X с функцией распределения F, если для каждой точки, где F непрерывна,

$$\lim_{n\to\infty} F_n(X) = F(X). \tag{3.5}$$

Например, пусть X_n распределено по закону $N(0, 1/n)$; тогда X_n сходится по распределению к ф. п. в., имеющей в нуле вид δ-функции. Легко показать, что для $X \neq 0$ $F_n \to F$, где

$$F(X) = 0, \qquad X < 0;$$
$$F(X) = 1, \qquad X > 0.$$

В точке $X = 0$ уравнение (3.5) не выполняется, поскольку $F_n(0) = \frac{1}{2}$ и $F(0) = 0$. Но $F(X)$ не непрерывна в этой точке и наше определение, таким образом, справедливо.

3.2.2. Теорема Леви. Часто свойства сходимости легче всего проиллюстрировать, если использовать характеристическую функцию ζ случайной переменной. Теорема, позволяющая поступать таким образом, была сформулирована Паулем Леви.

Если $\{\zeta_k(t)\}$ — последовательность характеристических функций, соответствующих функциям распределения $F_k(X)$, если $\zeta_k(t)$ сходится к некоторой функции $\zeta(t)$ и если реальная часть $\zeta(t)$ непрерывна в $t = 0$ [вспомним, что $\zeta(t)$ комплексная функция], то

1) $\zeta(t)$ — характеристическая функция;

2) F_k сходится к F — функции распределения, соответствующей ζ.

3.2.3. Сходимость по вероятности. Говорят, что последовательность $(X_1, \ldots, X_k)$ *сходится по вероятности* к X, если для любого $\varepsilon > 0$ и любого $\eta > 0$ может быть найдено такое значение n, что

$$P(|X_k - X| > \varepsilon) < \eta$$

для всех $k > n$. Сходимость по распределению слабее этого вида сходимости, поскольку она ничего не говорит о «расстоянии» между X_k и X. Действительно, можно показать, что сходимость по вероятности означает сходимость по распределению, тогда как обратное в общем случае неверно.

Нужно отметить, что как X_k, так и X — случайные переменные. Понятно, что X_i обычно сильно коррелированы. Действительно, можно показать, что если X_i независимы (и следовательно, не коррелированы), то X *достоверная* переменная, (т. е. случайная переменная, у которой ф. п. в. имеет вид δ-функции) и сходимость по распределению эквивалентна сходимости по вероятности. В общем случае (X недостоверная переменная) сходимость по вероятности означает сходимость по распределению и соответствует сильной корреляции между X_i.

3.2.4. Более сильные типы сходимости. Наряду с такими типами сходимости в математической статистике используются понятия *почти достоверной сходимости и сходимости по квадратичному среднему*. Если имеет место любой из этих типов сходимости, то это означает и сходимость по вероятности. Хотя эти виды сходимости и имеют важное теоретическое значение, в данной работе они не представляют практического интереса и не обсуждаются. Соотношения между различными типами сходимости показаны на следующей диаграмме [11]:

§ 3.3. ЗАКОН БОЛЬШИХ ЧИСЕЛ

Наиболее важным применением теорем сходимости являются законы больших чисел. Эти законы объясняют сходимость среднего (выборочного среднего). Фактически существуют два закона больших чисел, известных как *слабый* и *сильный* законы, что соответствует различным типам сходимости. Различие между ними не представляет здесь интереса, и мы просто приводим общий результат.

Пусть $(X_1, \ldots, X_n)$ — последовательность независимых случайных переменных и каждое из них имеет одинаковое среднее и дисперсии σ_i^2. Вспомним определение среднего:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Если среднее μ существует, то *слабый закон* больших чисел утверждает, что при

$$\lim_{n \to \infty} \left[\frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \right] = 0$$

$\overline{X}$ сходится к μ по *квадратичному среднему*. *Сильный закон* означает, что если

$$\lim_{n \to \infty} \left[\sum_{i=1}^{n} \left(\frac{\sigma_i}{i} \right)^2 \right]$$

конечен, то $\overline{X}$ сходится почти достоверно к μ. В обоих случаях имеет место сходимость по вероятности [12]. Эти результаты легко обобщить на случай различных средних.

3.3.1. Интегрирование по методу Монте-Карло. Возможно, что наиболее общим применением использования закона больших чисел является вычисление интеграла по методу Монте-Карло. Из определения ожидания

$$E(g) = \int_a^b g(u) f(u)\, du. \tag{3.6}$$

Выберем для $f(u)$ нормированное равномерное распределение*

$$f(u) = 1/(b - a).$$

Из уравнения (3.2) следует, что если числа u_i выбраны случайным образом равномерно и независимо на интервале (a, b), то соответствующие значения функции g_i удовлетворяют равенству

$$\frac{1}{n} \sum_{i=1}^{n} g_i \to E(g) \text{ при } n \to \infty. \tag{3.7}$$

Из уравнений (3.6) и (3.7) легко получить обычную формулу для интеграла по методу Монте-Карло:

$$I = \frac{b-a}{n} \sum_{i=1}^{n} g(u_i) \to \int_a^b g(u)\, du,$$

где u_i выбираются равномерно от a до b.

* Свойства равномерного распределения будут рассмотрены в п. 4.2.7.

Отметим, что этот метод совершенно отличен от обычных численных методов, таких, как, например, метод трапеций. При некоторых общих условиях все эти методы обеспечивают сходимость к точному значению интеграла, но ошибка вычисления интеграла при конечных n зависит от метода и от функции g. Например, дисперсия оценки I методом Монте-Карло просто равна

$$D(I) = \frac{D(g)}{n}.$$

Несмотря на то что методы Монте-Карло тесно связаны со статистическими задачами в физике, их подробное рассмотрение выходит за рамки этой книги и мы отсылаем читателя к соответствующей работе [13].

3.3.2. Центральная предельная теорема. Эта теорема имеет важнейшее значение как для теоретических, так и для практических задач в статистике. В действительности мы уже ее использовали раньше. Теперь сформулируем ее полностью и обсудим несколько детальнее.

Сначала вспомним результат п. 2.4.3, а именно: если мы имеем последовательность независимых случайных переменных, для каждой из которых соответствующее распределение имеет среднее μ_i и дисперсию σ_i^2, то распределение суммы $S = \Sigma X_i$ имеет среднее $\Sigma\mu_i$ и дисперсию $\Sigma\sigma_i^2$. Этот результат справедлив для любых распределений при условии, что X_i независимы, а среднее и дисперсии каждой переменной существуют и возрастают не слишком быстро с i. Ничего не говорилось о распределении суммы, исключая его среднее и дисперсию.

При тех же условиях, которые были использованы выше, центральная предельная теорема говорит о том, как распределена сумма в пределе больших n (это распределение также не зависит от вида индивидуальных распределений). В частности, она утверждает, что при $n \to \infty$

$$\frac{S - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \to N(0, 1). \tag{3.8}$$

Теорема выполняется в указанном выше общем случае. Продемонстрируем ее для обычного случая, когда все $\mu_i = \mu$ и все $\sigma_i = \sigma$. Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные переменные с $E(X) = \mu$, $D(X) = \sigma^2 < \infty$ и с конечными моментами третьего порядка. Определим

$$Y_n = \frac{X_1 + \ldots + X_n}{n}$$

и переменную

$$Z_n = \frac{Y_n - \mu}{\sigma/\sqrt{n}}.$$

Тогда ожидание и дисперсия переменной Z_n равны:

$$E(Z_n) = 0; \quad D(Z_n) = 1.$$

Центральная предельная теорема утверждает, что Z_n сходится по распределению к стандартному нормальному распределению*

$$\lim_{n\to\infty} [P(Z_n < a)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-\frac{1}{2}X^2\right) dX.$$

Без потери общности предположим, что $\mu = 0$. Тогда характеристическую функцию переменной X можно представить в виде

$$\zeta(t) = E(\exp(itX)) = 1 + \sigma^2 \frac{(it)^2}{2} + O(t^3).$$

Отсюда следует, что характеристическая функция nY_n равна $[\zeta(t)]^n$, а характеристическая функция Z

$$\psi_{Z_n}(t) = \left[\zeta\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Логарифмируя обе части равенства, получаем

$$\ln \psi_{Z_n}(t) = n \ln \zeta\left(\frac{t}{\sigma\sqrt{n}}\right) = n \ln\left[1 + \frac{\sigma^2}{2}\frac{(it)^2}{n\sigma^2} + O\left(\frac{t^3}{n^{3/2}}\right)\right] =$$

$$= \frac{(it^2)^2}{2} + O\left(\frac{1}{\sqrt{n}}\right)$$

для любого фиксированного t. Таким образом, мы доказали, что в пределе

$$\psi_{Z_n}(t) \to \exp\left[-\frac{1}{2}(t^2)\right]$$

есть характеристическая функция стандартного нормального распределения. Такой же результат получается, если X_i имеют различные дисперсии σ_i^2 при условии, что все они конечны или по крайней мере не стремятся к бесконечности так же быстро, как i. Информация о более точных условиях приведена в работах [1, 14].

Асимметрия и эксцесс, определенные в п. 2.4.6, служат мерой того, как сильно отличается распределение от нормального. Для нормального распределения эти величины равны нулю.

Например, характеристическая функция χ^2-распределения** с n степенями свободы

$$\zeta(t) = (1 - 2it)^{n/2}.$$

---

* Свойства нормального стандартного распределения будут рассмотрены в п. 4.2.1.

** О свойствах χ^2-распределения см. п. 4.2.3.

Прологарифмировав это равенство, получим

$$\ln \zeta(t) = -\frac{n}{2} \ln(1-2it) = \frac{n}{2}\left[2it + 4\frac{(it)^2}{2} + 2^3\frac{(it)^3}{3} + 2^4\frac{(it)^4}{4} + \ldots\right].$$

Следовательно, моменты можно найти из равенств, что $\mu = n$; $\sigma^2 = 2n$; $\mu_3 = 8n$; $\mu_4 - 3\mu_2^2 = 48n$.

Асимметрия равна $\gamma_1 = 8n/(2n)^{3/2} = 2\sqrt{2/n}$,

а эксцесс — $\gamma_2 = 48n/4n^2 = 12/n$.

Обычно считается, что χ^2-распределение при $n > 30$ практически близко к нормальному (γ_1 и γ_2 меньше 0,5).

Отметим, однако, что малость этих коэффициентов еще недостаточна для того, чтобы распределение можно было считать нормальным. Иногда важны моменты более высокого порядка (см. рис. 4.6, *б*).

3.3.3. Пример: генератор случайных чисел гауссовского распределения. Для расчетов по методу Монте-Карло нужно иметь последовательность равномерно распределенных случайных чисел, поэтому на быстродействующих ЭВМ такой генератор случайных чисел оформлен в виде библиотечной подпрограммы. Однако часто полезно иметь генератор случайных чисел, распределенных не равномерно, а нормально, например, по закону $N(0, 1)$. С помощью центральной предельной теоремы можно легко сконструировать генератор гауссовского распределения, если уже имеется генератор равномерного распределения. Если u_i — i-е число, выбранное с помощью генератора равномерного распределения, то нужно вычислять величины

$$g = \frac{\sum_{i=1}^{n} u_i - n/2}{\sqrt{n/12}}. \tag{3.9}$$

Поскольку для равномерного распределения $\mu = 1/2$ и $\sigma^2 = 1/12$, то из условия (3.8) сразу же следует, что (3.9) распределено как $N(0, 1)$ при $n \to \infty$. На практике g уже близко к нормально распределенной переменной, когда $n = 10$. Наибольшие отличия наблюдаются на хвостах распределения переменной g, которое удовлетворяет условию

$$-\sqrt{3n} \leqslant g \leqslant \sqrt{3n}.$$

Удобно и надежно положить $n = 12$, при котором максимальные значения соответствуют шести стандартным отклонениям. При этом получается простая формула:

$$g = \sum_{j=1}^{12} u_i - 6.$$

Отметим, однако, что существуют точные методы, основанные на замене переменных, и они иногда могут быть более эффективными.

Точная ф. п. в. среднего $\bar{u}$ n-равномерно распределенных случайных чисел рассчитана в работе [15]

$$\left.\begin{aligned} p(\bar{u}) &= \frac{n^n}{(n-1)!}\sum_{i=0}^{k}(-1)^i\binom{n}{i}\left(\bar{u}-\frac{i}{n}\right)^{n-1}; \\ \frac{k}{n} &\leqslant \bar{u} \leqslant \frac{k+1}{n},\ k = 0, \ldots,\ n-1. \end{aligned}\right\} \tag{3.10}$$

Эта функция интересна тем, что она состоит из n дуг степени $n - 1$ по $\bar{u}$, соединенных в $n - 1$ точках с непрерывными первыми $n - 2$ производными. Функции вида (3.10) называют *кусочными*.

ГЛАВА 4

РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

Зачастую плотности вероятностей и функции распределения, встречающиеся в практической деятельности, с хорошей точностью могут быть приближены некоторыми математическими функциями. Поэтому в § 4.1 и 4.2 мы познакомимся с несколькими такими «идеальными» распределениями для дискретных и непрерывных случайных переменных соответственно. В § 4.3 рассмотрим распределения, встречающиеся на практике, и методы работы с ними. Более подробно этот вопрос рассмотрен в работах [16, 17].

§ 4.1. ДИСКРЕТНЫЕ РАСПРЕДЕЛЕНИЯ

4.1.1. Биномиальное распределение.

Переменная r, положительное целое $\leqslant N$.

Параметры N, положительное целое; p, $0 \leqslant p \leqslant 1$.

Функция вероятностей

$$P(r) = \binom{N}{r} p^r (1-p)^{N-r}, \; r = 0, \ldots, n. \tag{4.1}$$

Ожидаемое значение $E(r) = Np$. (4.2)

Дисперсия $D(r) = Np(1-p)$. (4.3)

Асимметрия $\gamma_1 = \dfrac{1-2p}{[Np(1-p)]^{1/2}}$.

Эксцесс $\gamma_2 = \dfrac{1-6p(1-p)}{Np(1-p)}$.

Производящая функция вероятностей $G(Z) = [p + (1-p)Z]^N$.

График рис. 4.1, *а*.

Биномиальное распределение дает вероятность получения r успешных испытаний, если общее число испытаний равно N и вероятность успеха в одном испытании равна p. Например, число событий в одной ячейке гистограммы распределено по биномиальному закону.

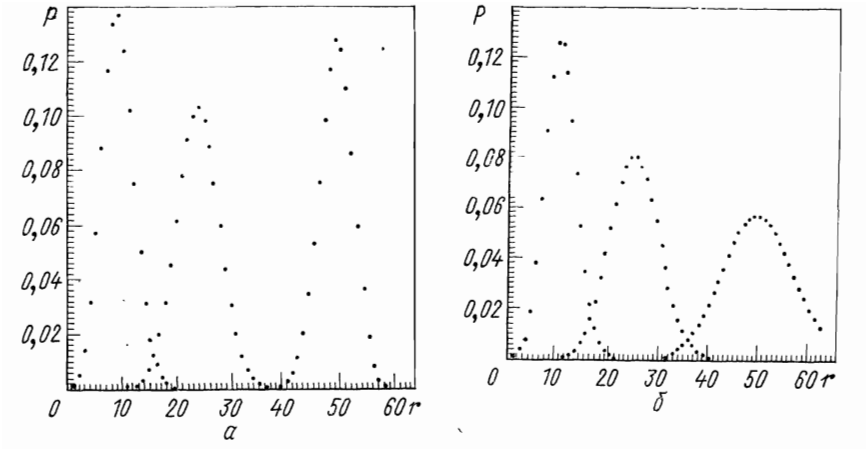


Рис. 4.1. Биномиальные распределения для $N=62$ с ожидаемыми средними $Np=10$, 25, 50 (*а*) и распределения Пуассона с ожидаемыми значениями $\mu=10$, 25, 50(*б*)

Если p неизвестно, то несмещенная оценка дисперсии

$$D(r)=\frac{N}{N-1}N\left(\frac{r}{N}\right)\left(1-\frac{r}{N}\right).$$

Примеры*: 1. Предположим, что результат A соответствует попаданию события в j-ю ячейку гистограммы, а $\overline{A}$ соответствует попаданию события в любую другую ячейку гистограммы. Вероятность p получить результат A обычно равна интегралу от ф. п. в. по ячейке (иногда его можно хорошо приблизить произведением ширины ячейки на значение ф. п. в. в середине ячейки). Вероятность попадания r событий в ячейку j и $N-r$ событий во все другие ячейки дается выражением (4.1). Ожидаемое число событий в ячейке j равно Np в соответствии с (4.2), а дисперсия — $Np(1-p)$ в соответствии с (4.3).

2. Предположим, что исследуется асимметрия вперед-назад и что эксперимент нужно закончить после набора N событий (N не случайная перемен-

---

* Для экспериментаторов, использующих электронные методы детектирования частиц, наиболее типичный пример применения биномиального распределения встречается при измерении эффективности счетчика, включенного в схему совпадения с другими счетчиками телескопа. Пропустив заданное число N_k частиц через телескоп из k счетчиков, регистрируют число соответствующих срабатываний схемы совпадений N_{k+1}, включающей исследуемый i-й счетчик. Тогда оценка эффективности этого счетчика, включая и эффективность электронной схемы совпадений, определяется отношением $\eta_i=\frac{N_{k+i}}{N_k}$. Оценка стандартного отклонения для измеренного значения будет равна

$$\hat{\sigma}_i=\sqrt{N_{k+i}(1-\hat{f}_i)/N_k}.$$

В практически важном случае, когда $f_i\sim 1$, корреляция в биномиальном распределении между случайной величиной N_{k+1} и заданной N_k приводит к значительному уменьшению погрешности измерений. — *Прим. ред.*

ная!). Пусть F и B — числа событий в передней и задней полусферах соответственно $N = F + B$. Тогда F распределено по биномиальному закону:

$$\binom{N}{F} p^F (1-p)^B$$

со средним Np и дисперсией $Np(1-p)$. При больших N дисперсия с хорошей точностью равна $F(1-p)$, а стандартное отклонение $\sqrt{F(1-p)}$ (не $\sqrt{F}$). Обычно асимметрия вперед-назад определяется как

$$r = \frac{F-B}{F+B} = \frac{2F}{N} - 1.$$

Тогда r распределено по закону

$$P(r) = \binom{N}{\frac{1}{2}(Nr+1)} p^{\frac{1}{2}(Nr+1)} (1-p)^{N-\frac{1}{2}(Nr+1)}$$

с дисперсией $4p(1-p)/N$. При больших N дисперсия равна приближенно

$$D(r) \simeq \frac{4FB}{N^3} = \frac{4FB}{(F+B)^3} \tag{4.4}$$

и, соответственно, стандартное отклонение

$$\sigma \approx \frac{2}{(F+B)} \sqrt{\frac{FB}{(F+B)}}.$$

3. Вероятность зарегистрировать нужное событие в одном экспериментальном испытании равна p. Предположим, что нужно узнать, как много испытаний следует проделать, чтобы вероятность регистрации по крайней мере одного события была α. Пусть X число нужных событий при N испытаниях, а N неизвестно. Задача состоит в том, чтобы найти такое N, для которого

$$P(X \geqslant 1) \geqslant \alpha$$

или

$$1 - P(X = 0) \geqslant \alpha,$$

или

$$P(X = 0) \leqslant 1 - \alpha.$$

Выражая левую часть неравенства с помощью (4.1) при $r = 0$, получаем

$$(1-p)^N \leqslant 1-\alpha.$$

Прологарифмировав неравенство с учетом того, что $(1-p) < 1$, получим

$$N \geqslant \ln(1-\alpha)/\ln(1-p).$$

4.1.2. Мультиномиальное распределение.

Переменные r_i, $i = 1, \ldots, k$, положительные целые $\leqslant N$.

Параметры N, положительное целое; k, положительное целое; $p_1 \geqslant 0$, $p_2 \geqslant 0, \ldots, p_k \geqslant 0$; с $\sum_{i=1}^{k} p_i = 1$.

Функция вероятности:

$$P(r_1, \ldots, r_k) = \frac{N!}{r_1! \ldots r_k!} p_1^{r_1} \ldots p_k^{r_k}. \tag{4.5}$$

Ожидаемые значения $E(r_i) = Np_i$.

Дисперсия $D(r_i) = Np_i(1 - p_i)$. (4.6)

Смешанные вторые моменты $D(r_i, r_j) = -Np_ip_j$, $i \neq j$. (4.7)

Производящая функция вероятностей

$$G(Z_2, \ldots, Z_k) = (p_1 + p_2 Z_2 + \ldots + p_k Z_k)^N.$$

Обобщением биномиального распределения на случай более чем двух возможных исходов эксперимента служит мультиномиальное распределение. Оно дает вероятность [см. уравнение (4.5)] получения r_i результатов типа i в N независимых испытаниях, где p_i соответствует вероятности i-го исхода в одном испытании ($i = 1, 2, \ldots, k$).

Примером мультиномиального распределения является гистограмма, содержащая N событий и состоящая из k ячеек; при этом в ячейке с номером i находится r_i событий. В этом примере случайная переменная имеет вид k-мерного вектора, но область значений этого вектора ограничена $(k - 1)$-мерным пространством, поскольку $\sum_{i=1}^{k} r_i = N$. Так как N фиксировано (см. ниже), то можно с помощью (4.7) показать, что коэффициент корреляции числа событий в двух ячейках i и j *отрицателен*:

$$\rho(r_i, r_j) = -\sqrt{p_i p_j/(1-p_i)(1-p_j)}. \quad (4.8)$$

Во многих экспериментальных ситуациях наблюдения r_i независимы, так что и общее число событий $N = \Sigma r_i$ может быть рассмотрено как случайная величина. Тогда данные можно рассматривать с двух точек зрения.

1. Ненормированный случай. Если удобно считать N случайной переменной, то рассматривают $(k + 1)$-переменную (r_i и N), из которых r_i взаимно не коррелированы, а N коррелированы с r_i. Соответствующая матрица вторых моментов имеет ранг k. Тогда распределение событий по ячейкам не является мультиномиальным (а пуассоновским, см. п. 4.1.3).

2. Нормированный случай. Если неудобно рассматривать N как случайную переменную (например, если не существует теории, которая предсказывает ожидаемое значение для N), то величины r_i должны быть рассмотрены как условные по отношению к фиксированному наблюдаемому значению N. В этом случае k значений r_i распределены по мультиномиальному закону, все они коррелированы друг с другом с корреляциями, вычисляемыми с помощью уравнения (4.8), а ранг матрицы вторых моментов равен $k - 1$. Когда $p_i \ll 1$ (много ячеек), уравнение (4.6) может быть приближенно записано в виде

$$D(r_i) \sim Np_i \sim r_i$$

и стандартное отклонение числа событий в ячейке становится равным:

$$\sigma_i \sim \sqrt{r_i}.$$

4.1.3. Распределение Пуассона.

Переменная r, положительное целое.

Параметр μ, положительное реальное число.

Функция вероятностей

$$P(r) = \frac{\mu^r \exp(-\mu)}{r!}. \tag{4.9}$$

Ожидаемое значение $E(r) = \mu$.

Дисперсия $D(r) = \mu$.

Асимметрия $\gamma_1 = 1/\sqrt{\mu}$.

Эксцесс $\gamma_2 = 1/\mu$.

Производящая функция вероятностей

$$G(Z) = \exp(\mu(Z - 1)).$$

График рис. 4.1, *б*, 4.2.

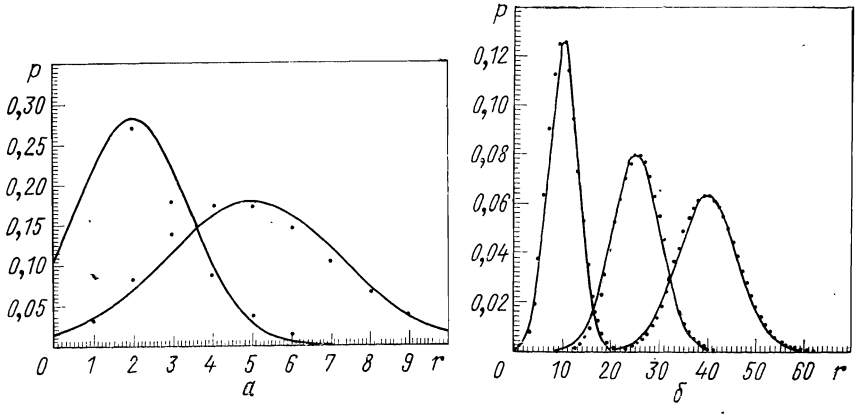


Рис. 4.2. Сравнение распределений Пуассона (точки) с нормальными распределениями (сплошные линии) с одинаковыми средними и дисперсиями:
а — распределения Пуассона с μ=2,5. Нормальные распределения N (2,2), N (5,5); *б* — с μ=10, 25, 40. Нормальные распределения N (10, 10), N (25, 25), N (40, 40)

Распределение Пуассона дает вероятность наблюдения r событий в заданный промежуток времени при условии, что события независимы и возникают с постоянной скоростью. Это распределение является предельным случаем биномиального распределения при $p \to 0$, но когда $Np = \mu$ — конечная константа.

Когда $\mu \to \infty$, распределение Пуассона приближается к нормальному распределению.

Существует важное соотношение между функцией распределения закона Пуассона и интегралом вероятности χ^2-распределения (см. п. 4.2.3), а именно:

$$P(r \leqslant N_0 \mid \mu) = 1 - P[\chi^2(2N_0 + 2) < 2\mu].$$

Примеры: 1. Предположим, что частицы испускаются радиоактивным источником со средней интенсивностью γ-частиц в единицу времени так, что вероятность испускания одной частицы за время δt равна $\nu\delta t$, а вероятность испускания более чем одной частицы за время δt есть величина второго порядка малости — $o(\delta t^2)$. Тогда распределение числа X частиц, испущенных в течение фиксированного времени t, — пуассоновское со средним νt:

$$P(X = r) = (\nu t)^r \exp(-\nu t)/r!$$

2. Распределение числа пузырьков на некоторой фиксированной длине трека частицы с постоянным импульсом также обычно считается пуассоновским. Этот вопрос подробно рассмотрен в примере следующего параграфа (составное пуассоновское распределение).

3. Рассмотрим отношение числа частиц, испущенных в переднем и заднем направлениях. Предположим, что числа F и B — независимые случайные переменные, распределенные согласно закону Пуассона.

Тогда F имеет среднее и дисперсию, равные одному и тому же числу, которое приближенно можно положить равным F. Асимметрия определяется соотношением

$$r = (F - B)/(F + B).$$

Предположив, что погрешности F и B малы (большое число событий), воспользуемся формулой дифференциала, чтобы получить приближенное значение для дисперсии r:

$$\frac{dr}{r} = \frac{d(F-B)}{F-B} - \frac{d(F+B)}{F+B} = 2\,\frac{BdF - FdB}{(F-B)(F+B)}.$$

Возведем обе части равенства в квадрат и вычислим ожидаемые значения

$$\frac{E(dr^2)}{r^2} \simeq 4\,\frac{B^2 E(dF^2) + F^2 E(dB)^2}{(F-B)^2(F+B)^2} \simeq \frac{4BF}{(F-B)^2(F+B)^2}.$$

В результате получим дисперсию

$$D(r) \simeq 4\,BF/(F + B)^3, \tag{4.10}$$

т. е. пришли к такому же результату, что и в примере 2 п. 4.1.1, в котором говорилось о биномиальном распределении [см. формулу (4.4)]. Отметим, однако, следующие отличия.

Уравнение (4.10) является приближенным (мы использовали дифференцирование), тогда как расчеты, приводящие к формуле (4.4), выполнены строго. В действительности, уравнение (4.10) отличалось бы от выражения (4.4), если бы мы не заменили μ_F на F, а μ_B на B.

В случае биномиального распределения число событий фиксировано *заранее*. Если N не фиксировано, то полное число событий подчинено закону Пуассона $e^{-\nu}\nu^N/N!$, а F и B — биномиальному закону, условному по отношению к N.

Таким образом, вероятность наблюдения N событий, из которых F вылетают вперед, а B назад, равна:

$$f(N, F, B) = e^{-\nu}\frac{\nu^N}{N!}\binom{N}{F} p^F (1-p)^B = e^{-\nu}\frac{(\nu p)^F [\nu(1-p)]^B}{F!\,B!}.$$

Это выражение имеет вид *произведения двух законов Пуассона для* F и B.

4.1.4. Составное распределение Пуассона.
Переменная r, положительное целое.
Параметры μ, λ, положительные реальные числа.
Функция вероятностей

$$P(r) = \sum_{N=0}^{n} \left[\frac{(N\mu)^r \exp(-N\mu)}{r!} \cdot \frac{\lambda^N \exp(-\lambda)}{N!} \right]. \qquad (4.11)$$

Ожидаемое значение $E(r) = \lambda\mu$.
Дисперсия $D(r) = \lambda\mu(1+\mu)$.
Производящая функция вероятностей

$$G(Z) = \exp[-\lambda + \lambda \exp(-\mu + \mu Z)].$$

Составное распределение Пуассона, иногда известное как распределение *ветвящегося процесса*, является распределением суммы N пуассоновских переменных n_i с одинаковым средним μ, а само N также является пуассоновской переменной со средним λ. Для того чтобы было понятно, как оно получается, введем переменную $r \equiv \sum_{i=1}^{N} n_i$. Каждое n_i распределено по закону

$$P(n_i) = \mu^{n_i} \exp(-\mu)/n_i!,$$

а N — по закону

$$P(N) = \lambda^N \exp(-\lambda)/N!$$

Используя фундаментальное соотношение

$$P(r) = \sum_{N} P(r/N)\, P(N)$$

и метод п. 2.5.4, а также тот факт, что сумма N пуассоновских переменных со средним μ сама является пуассоновской переменной со средним $N\mu$, легко получить функцию плотности вероятности (4.11).

С более общим случаем составного пуассоновского распределения приходится сталкиваться, когда переменная r является суммой N случайных наблюдений n_i, распределенных по некоторому закону, а N — пуассоновская переменная со средним λt. Тогда производящая функция вероятностей равна (см. п. 2.5.4)

$$G(Z, t) = \exp[-\lambda t + \lambda t f(Z)],$$

где $f(Z)$ — производящая функция вероятностей для n_i. Это распределение обладает замечательным свойством факторизуемости:

$$G(Z, t_1 + t_2) = G(Z, t_1)\, G(Z, t_2). \qquad (4.12)$$

Случайная переменная, связанная с $G(Z,t)$, может быть названа вкладом в промежуток t. Для составного процесса Пуассона результат (4.12) означает, что вклады неперекрывающихся периодов

являются независимыми случайными переменными также с составными пуассоновскими распределениями.

Это распределение имеет очень широкое применение в экологии, ядерных цепных реакциях, генетике и теории цепей. Мы покажем, как его можно использовать, на примере эксперимента по поиску кварков.

Пример. Эксперимент по поиску кварков. В качестве примера как простого, так и составного пуассоновского распределения рассмотрим процесс образования капелек заряженными частицами, проходящими через диффузионную камеру. Рассмотрим конкретный случай эксперимента по «поиску кварков», выполненный с помощью этой техники [18].

Когда быстрые (релятивистские) заряженные частицы проходят через диффузионную камеру, то ионизация молекул, вызванная этими частицами, приводит к образованию капелек, которые выглядят, как треки. Предполагается, что вероятность образования капли на единицу длины трека постоянна и пропорциональна квадрату заряда частицы. Поскольку почти все частицы, проходящие через камеру, обладают единичным зарядом, можно установить существование частиц с зарядом меньше единицы, если бы существовал трек, число капель которого на единицу длины было бы *значительно* меньше, чем у среднего трека. Цель эксперимента состоит в том, чтобы установить существование «кварка» — частицы с зарядом 2/3 единичного (протонного) заряда.

В соответствии с предположениями распределение числа капель на данной длине трека должно подчиняться закону Пуассона. Среднее этого распределения μ, определенное в результате подсчета числа капель на некоторой фиксированной длине обычных треков, оказалось равным 229. При просмотре был обнаружен трек только со 110 каплями. Возникает статистическая задача: рассчитать вероятность того, что частица с единичным зарядом образует только 110 (или меньше) капель, и сравнить ее со значением $1/55\,000 \sim 2\times 10^{-5}$, поскольку в эксперименте было проанализировано 55 000 треков. Эта вероятность равна

$$P(r \leqslant 110) = \sum_{i=0}^{110} \frac{229^i e^{-229}}{i!} \approx 1{,}6 \cdot 10^{-18},$$

отсюда следует, что результат значим (т. е. практически исключено, что частица с единичным зарядом даст такое число капель).

Однако в последующей работе [19] физические предположения, лежащие в основе анализа, были подвергнуты сомнению. В частности, вызывал сомнение механизм образования капель. Было указано, что закону Пуассона подчинено число элементарных актов рассеяния, в каждом из которых рождается около четырех капель. Предполагая, что в каждом таком акте рождается в точности четыре капли, можно пересчитать вероятность возникновения 110 или меньше капель (при этом числа 229 и 110 нужно округлить до ближайших чисел, кратных четырем):

$$P(r' \leqslant 28) = \sum_{i=0}^{28} \frac{57^i e^{-57}}{i!} \approx 6{,}7 \cdot 10^{-6}.$$

Полученное число существенно меньше и это показывает, насколько чувствителен результат к тому, сколь верны предположения.

Более точная модель образования капель должна была бы учесть, что в элементарных актах рассеяния рождается неодинаковое число капель, и что это число должно быть также распределено по закону Пуассона. Очевидно, что тогда число капелек будет подчинено закону составного рас-

пределения Пуассона. Полагая $\lambda\mu = 229$ и $\mu = 4$, получаем в соответствии с (2.56):

$$P(r \leqslant 110) = \sum_{i=0}^{110} \sum_{N=0}^{\infty} \left[\frac{(N\mu)^i e^{-N\mu}}{i!} \frac{\lambda^N e^{-\lambda}}{N!} \right] \sim 4{,}7 \cdot 10^{-5},$$

что еще значительнее уменьшает значимость наблюдения.

4.1.5. Геометрическое распределение.

Переменная r, положительное целое.

Параметры p, $0 \leqslant p \leqslant 1$.

Функция вероятностей

$$P(r) = p(1-p)^{r-1}. \tag{4.13}$$

Ожидаемое значение

$$E(r) = 1/p.$$

Дисперсия $D(r) = (1-p)/p^2$.

Асимметрия $\gamma_1 = \dfrac{(2-p)}{(1-p)^{1/2}}$.

Производящая функция вероятностей

$$G(Z) = \frac{pZ}{1-(1-p)Z}.$$

Эксцесс $\gamma_2 = \dfrac{p_2 - 6p + 6}{1-p}$.

График рис. 4.3.

Геометрическое распределение дает вероятность получить r безуспешных попыток, предшествующих первой успешной попытке, при условии, что вероятность успеха в одном испытании равна p.

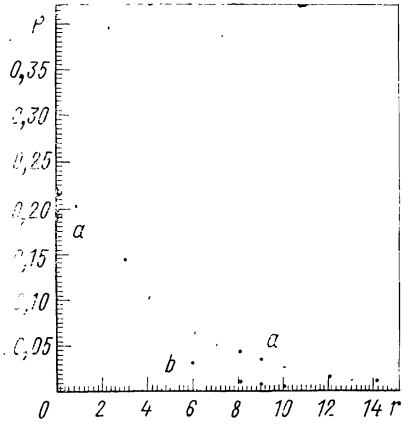


Рис. 4.3. Геометрические распределения:
$a - p = 0{,}2$; $b - p = 0{,}4$

4.1.6. Отрицательное биномиальное распределение.
Переменная r, положительное целое $\geqslant m$.
Параметры m, положительное целое; p, $0 \leqslant p \leqslant 1$.
Функция вероятностей

$$P(r) = \binom{r-1}{m-1} p^m (1-p)^{r-m}. \qquad (4.14)$$

Ожидаемое значение $E(r) = m/p$.
Дисперсия $D(r) = m(1-p)/p^2$.
Асимметрия $\gamma_1 = (2-p)/[m(1-p)]^{1/2}$.
Эксцесс $\gamma_2 = [p^2 - 6p + 6]/m(1-p)$.
Производящая функция вероятностей

$$G(Z) = \left(\frac{pZ}{1-(1-p)Z}\right)^m.$$

График рис. 4.4.
Отрицательное биномиальное распределение дает вероятность затраты r попыток до тех пор, пока число успешных попыток не станет равным m при условии, что вероятность успеха в одной попытке равна p.

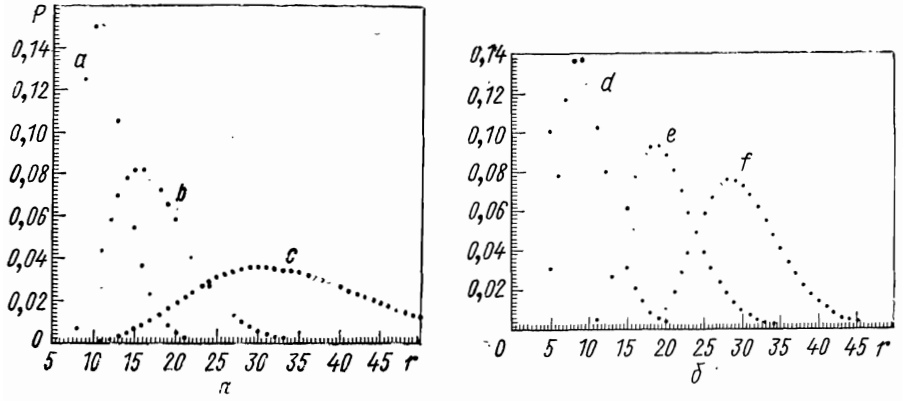


Рис. 4.4. Отрицательные биномиальные распределения:
$а$ — $m=7$ и $p=0{,}2$; $0{,}4$; $0{,}6$ для кривых a, b, c соответственно; $б$ — $p=0{,}5$ и $m=5$, 10, 15 для кривых d, e, f соответственно

В некоторых случаях нужно знать распределение числа безуспешных попыток в некоторой последовательности испытаний, заканчивающейся в момент получения m-го успеха. Соответствующая функция вероятностей равна

$$P(s) = \binom{s+m-1}{s} p^m (1-p)^s \quad \text{с} \quad E(s) = m(1-p)/p. \qquad (4.15)$$

Дисперсия и моменты более высокого порядка такие же. Распределение (4.15) можно также рассматривать как распределение числа успешных испытаний в единицу времени при условии, что скорость, с которой наблюдаются успешные испытания, является случайной переменной, имеющей гамма-распределение (а не константа как для распределения Пуассона).

4.2.1. Нормальное одномерное.

Параметры μ, реальное; σ, положительное реальное число.

Функции плотности вероятности

$$f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\,\frac{(X-\mu)^2}{\sigma^2}\right]. \qquad (4.16)$$

Функция распределения

$$F(X) = \Phi\left(\frac{X-\mu}{\sigma}\right), \text{ где } \Phi(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \exp\left(-\frac{1}{2}X^2\right) dX. \qquad (4.17)$$

Ожидаемое значение $E(X) = \mu$.

Дисперсия $D(X) = \sigma^2$.

Асимметрия $\gamma_1 = 0$.

Эксцесс $\gamma_2 = 0$.

Характеристическая функция

$$\zeta(t) = \exp\left[it\mu - \frac{1}{2}t^2\sigma^2\right].$$

Моменты

$$\mu_{2r} = \frac{(2r)!}{2^r(r)!}\sigma^{2r}; \quad \mu_{2r+1} = 0, \quad r \geqslant 1.$$

Графики рис. 4.2, 4.5, 4.6, *б*, 4.7, 4.10, 9.2.

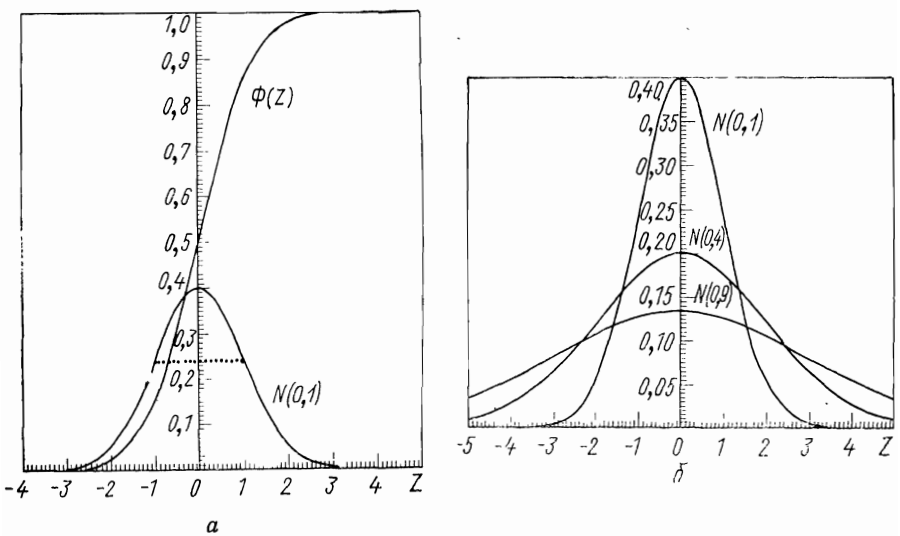


Рис. 4.5. Стандартное нормальное распределение $N(0,1)$ и его функция распределения $\Phi(Z)$. Линия из точек указывает стандартное отклонение σ (длина линии 2σ) (*а*) и нормальные распределения $N(0,1)$, $N(0,4)$ и $N(0,9)$ (*б*)

Ф. п. в. нормального или гауссовского распределения, сокращенно обозначаемого $N(\mu, \sigma^2)$, является наиболее важным теоретическим распределением в статистике. Его функция распределения [см. уравнение (4.17)] называется *интегралом вероятности нормального распределения или функцией ошибок*. Заметим, что в литературе существует много различных определений функции ошибок.

Отметим также, что стандартное отклонение σ не равно полуширине ф. п. в. на половине высоты. Полуширина на половине высоты равна 1,176 σ. Ниже даны вероятностные содержания различных интервалов:

$$P\left(-1{,}64 \leqslant \frac{X-\mu}{\sigma} \leqslant 1{,}64\right) = 0{,}90;$$

$$P\left(-1{,}96 \leqslant \frac{X-\mu}{\sigma} \leqslant 1{,}96\right) = 0{,}95;$$

$$P\left(-2{,}58 \leqslant \frac{X-\mu}{\sigma} \leqslant 2{,}58\right) = 0{,}99;$$

$$P\left(-3{,}29 \leqslant \frac{X-\mu}{\sigma} \leqslant 3{,}29\right) = 0{,}999.$$

Функция $N(0, 1)$ называется *стандартной** нормальной плотностью, а его функция распределения

$$\Phi(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} \exp\left(-\frac{1}{2} t^2\right) dt$$

— *стандартной нормальной функцией распределения.*

Можно получить ряд важных результатов для случайных переменных X_i, которые *независимы* и *нормальны*.

а. Любая линейная комбинация X_i также нормальна. Предположим, что средние и дисперсии переменных X_1 и X_2 соответственно равны μ_1, μ_2 и σ_1^2, σ_2^2. Тогда aX_1 и bX_2 являются независимыми нормальными переменными с характеристическими функциями:

$$\zeta_{aX_1}(t) = \exp\left(ita\mu_1 - \frac{1}{2} t^2 a^2 \sigma_1^2\right)$$

и

$$\zeta_{aX_2}(t) = \exp\left(itb\mu_2 - \frac{1}{2} t^2 b^2 \sigma_2^2\right).$$

---

* Случайная переменная называется *стандартизованной*, если ее ожидание равно нулю, а дисперсия единице. Поэтому, если X — случайная переменная со средним μ и дисперсией σ^2, тогда стандартизованная форма будет $(X - \mu)/\sigma$.

Характеристическая функция переменной $Z = aX_1 + bX_2$ равна

$$\zeta_Z(t) = \zeta_{aX_1}(t)\,\zeta_{bX_2}(t) = \exp\left[it(a\mu_1 + b\mu_2) - \frac{1}{2}t^2(a^2\sigma_1^2 + b^2\sigma_2^2)\right].$$

Таким образом, Z также подчиняется нормальному распределению со средним $a\mu_1 + b\mu_2$ и дисперсией $a^2\sigma_1^2 + b^2\sigma_2^2$. Так же можно показать, что любая линейная комбинация $Z = \sum_{i=1}^{n} a_i X_i$ имеет нормальное распределение.

б. *Среднее или выборочное среднее*

$$\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$$

и *выборочная дисперсия*

$$S^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad (4.18)$$

независимы, если и только если X_i имеют одинаковое нормальное распределение (с одинаковыми μ и σ). Это свойство присуще только нормальному распределению.

в. Если X_i — *стандартные* нормальные переменные, то плотность вероятности постоянна на сфере:

$$\sum_{i=1}^{n} X_i^2 = \text{const},$$

поскольку плотность — функция только $\sum X^2$. Это свойство радиальной симметрии также присуще только нормальному распределению [20].

г. Если $\mathbf{X}$ — вектор, компоненты которого — независимые случайные переменные, не обязательно одинаково распределенные, и если не существует нетривиального преобразования $\mathbf{X} = C\mathbf{Y}$, где $\mathbf{Y}$ — вектор независимых случайных переменных, тогда каждое X_i распределено нормально, преобразование ортогонально и каждое Y_i имеет нормальное распределение.

4.2.2. Нормальное многомерное.

Переменная $\mathbf{X}$, k-мерный вектор.

Параметры μ, k-мерный реальный вектор; $\underset{\sim}{D}$, $k \times k$ матрица, положительная полуопределенная.

Функция плотности вероятности

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2}\,|\underset{\sim}{D}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}-\mu)^T \underset{\sim}{D}^{-1}(\mathbf{X}-\mu)\right]. \qquad (4.19)$$

Ожидаемые значения

$$E(\mathbf{X}) = \mu.$$

Моменты второго порядка

$$D(\mathbf{X}) = \underset{\sim}{D};$$

$$D(X_i) = \underset{\sim}{D}_{ii};$$

$D(X_i, X_j) = (\underset{\sim}{D})_{ij}$ — ij-й элемент $\underset{\sim}{D}$.

Характеристическая функция

$$\zeta(\mathbf{t}) = \exp\left[i t\boldsymbol{\mu} - \frac{1}{2} t^T \underset{\sim}{D}\mathbf{t}\right].$$

При обобщении нормального распределения на многомерный случай естественно искать функцию плотности в виде экспоненты, показатель которой является квадратичной формой по компонентам векторов:

$$f(\mathbf{X}) \sim \exp\left[-\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\left(\frac{X_i-\mu_i}{\sigma_i}\right)\left(\frac{X_j-\mu_j}{\sigma_j}\right)\right].$$

Как и в одномерном случае, μ_i и σ_i — соответственно параметры расположения и масштабный параметр. Если использовать матричные обозначения, то функцию плотности можно записать в виде

$$f(\mathbf{X}) \sim \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \underset{\sim}{A}(\mathbf{X}-\boldsymbol{\mu})\right].$$

Поскольку $f(\mathbf{X})$ является функцией плотности, то $\int f(\mathbf{X})\, d\mathbf{X}$ должен сходиться и из этого следует, что матрица $\underset{\sim}{A}$ положительная полуопределенная. Определив константу пропорциональности из условия

$$\int f(\mathbf{X})\, d\mathbf{X} = 1,$$

придем к выражению (4.19) для функции плотности вероятности. Тогда *матрицу вторых моментов* $\mathbf{X}$ удобно определить в виде

$$\underset{\sim}{D} = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] = \underset{\sim}{A}^{-1} =$$

$$= \begin{bmatrix} \sigma_1^2 & & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1k}\sigma_1\sigma_k \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & & \cdots & \vdots \\ & & & & \\ & & & & \\ \vdots & & \vdots & & \vdots \\ \rho_{1k}\sigma_1\sigma_2 & \cdots & & & \sigma_k^2 \end{bmatrix},$$

где ρ_{ij} — коэффициент корреляции между компонентами X_i и X_j;

$$\rho_{ij} = E[(X_i-\mu_i)(X_j-\mu_j)]/\sigma_i\sigma_j.$$

Для случая двух переменных X и Y ф. п. в. (4.19) приобретает вид:

$$f(X,Y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(X-\mu_X)^2}{\sigma_X^2}-\right.\right.$$
$$\left.\left.-2\rho\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(Y-\mu_Y)^2}{\sigma_Y^2}\right\}\right]. \qquad (4.20)$$

Параметрами выражения (4.20) являются ожидания μ_X, μ_Y, стандартные отклонения σ_X, σ_Y и коэффициент корреляции ρ.

Условная плотность

$$f(X|Y)=\frac{1}{\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2}}\exp\left[-\frac{1}{2\sigma_X^2(1-\rho^2)}\times\right.$$
$$\left.\times\left\{X-\left(\mu_X+\frac{\rho\sigma_X}{\sigma_Y}(Y-\mu_Y)\right)\right\}^2\right]$$

также имеет вид нормального распределения со средним

$$\mu_X+\rho\frac{\sigma_X}{\sigma_Y}(Y-\mu_Y)$$ и дисперсией $\sigma_X^2(1-\rho^2)$.

Если $\mathbf{X}$ распределено по нормальному закону, а $\underset{\sim}{D}$ несингулярна, то величина $(\mathbf{X}-\boldsymbol{\mu})^T\underset{\sim}{D}^{-1}(\mathbf{X}-\boldsymbol{\mu})$ называется *ковариационной формой* $\mathbf{X}$. Она распределена по закону χ^2 с k степенями свободы.

Многомерное нормальное распределение интересно по многим причинам:

а) оно является функцией только средних дисперсий и корреляций двух переменных;

б) ф. п. в. в многомерном пространстве имеет «колоколообразную» форму;

в) уровни постоянной плотности вероятности соответствуют условию:

$$(\mathbf{X}-\boldsymbol{\mu})^T\underset{\sim}{D}^{-1}(\mathbf{X}-\boldsymbol{\mu})=\text{const};$$

г) любое сечение этого распределения, например, плоскостью $X_i=\text{const}$ опять-таки является нормальным распределением в $(k-1)$-мерном пространстве вида (4.19) с матрицей вторых моментов $\underset{\sim}{D}_{k-1}$, полученной в результате выбрасывания i-й строки и i-го столбца матрицы $\underset{\sim}{D}^{-1}$ и обращения результирующей подматрицы;

д) любая проекция на пространство меньшего числа измерений дает маргинальное распределение, которое также нормально и имеет вид (4.19) с матрицей вторых моментов, полученной путем зачеркивания соответствующих строк и столбцов D. В частности, маргинальное распределение X_i имеет вид $N(\mu_i, \sigma_i^2)$;

е) совокупность переменных, каждая из которых является линейной функцией переменных, распределенных нормально, также распределена по многомерному нормальному закону;

ж) средний вектор $\overline{\mathbf{X}}$ и *второй момент выборки* s_{ij} совокупности n независимых наблюдений определяются в виде:

$$\overline{X}_i = \frac{1}{n} \sum_{l=1}^{n} X_{il} \tag{4.21}$$

и

$$s_{ij} = \frac{1}{n-1} \sum_{l=1}^{n} (X_{il} - \overline{X}_i)(X_{jl} - \overline{X}_j),$$

где X_{il} — i-я компонента l-го наблюденного вектора. $\mathbf{X}$ и матрица (s_{ij}) распределены независимо, если и только если распределение генеральной совокупности $\mathbf{X}_k$ многомерное нормальное.

4.2.3. χ^2-Распределение.

Переменная X, положительное реальное число.

Параметр n, положительное целое (степень свободы).

Функция плотности вероятностей

$$f(X) = \frac{(1/2)\,(X/2)^{(n/2)-1} \exp(-X/2)}{\Gamma(n/2)}. \tag{4.22}$$

Ожидаемое значение $E(X) = n$.

Дисперсия $D(X) = 2n$.

Асимметрия $\gamma_1 = 2\sqrt{2/n}$.

Эксцесс $\gamma_2 = 12/n$.

Характеристическая функция

$$\zeta(t) = (1 - 2it)^{-n/2} \tag{4.23}$$

Графики рис. 4.6, 9.1.

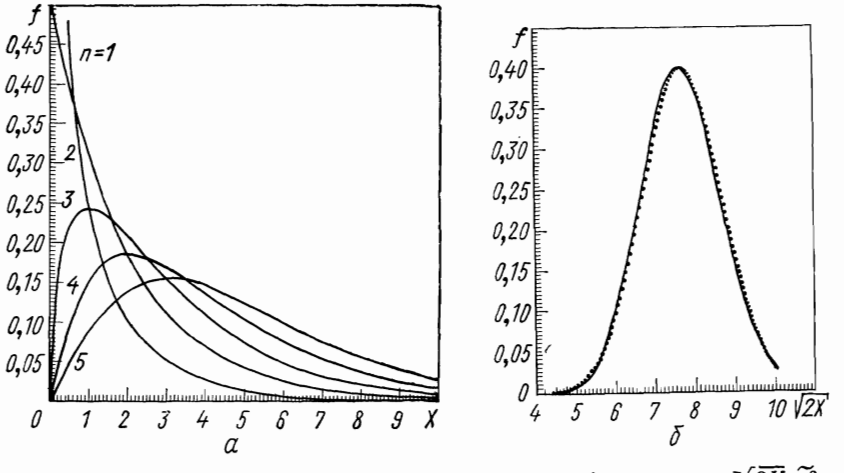


Рис. 4.6. χ^2-Распределения (*а*) и (*б* — сплошная линия) величины $\sqrt{2X}\,\tilde{\chi}^2_X$ (30) как функция $\sqrt{2X}$. Точки: $N(\sqrt{2 \cdot 30 - 1}, 1)$

Предположим, что $X_1, \ldots, X_n$ — независимые *стандартные нормальные* переменные каждая с распределением $N(0, 1)$. Тогда говорят, что сумма квадратов

$$X^2_{(n)} = \sum_{i=1}^{n} X_i^2 \tag{4.24}$$

имеет χ^2-*распределение* с n степенями свободы. Чтобы найти вид этого распределения, рассмотрим характеристическую функцию квадрата стандартной нормальной переменной X:

$$E[\exp(itX^2)] = \int_{-\infty}^{\infty} \exp(itX^2) \frac{1}{\sqrt{2\pi}} \times$$

$$\times \exp\left(-\frac{1}{2}X^2\right) dX = \frac{1}{\sqrt{(1-2it)}}. \tag{4.25}$$

Из уравнений (2.46), (4.24) и (4.25) следует, что характеристическая функция $X^2_{(n)}$ имеет вид (4.23), которая соответствует ф. п. в. (4.22).

Если $X^2_{(n)}$ и $X^2_{(m)}$ имеют независимые χ^2-распределения с n и m степенями свободы соответственно, то сумма $X^2_k = X^2_{(n)} + X^2_{(m)}$ имеет χ^2-распределение с $k = n + m$ степенями свободы. Это очевидно из определения (4.24).

Функция $R_n = \sqrt{X^2_{(n)}}$ имеет χ (n)-распределение с функцией плотности

$$P(R_n) = \frac{\left(\frac{1}{2}\right)^{n/2-1} R_n^{2n-1} \exp\left(-\frac{1}{2}R_n^2\right)}{\Gamma\left(\frac{1}{2}n\right)}.$$

В асимптотическом пределе χ^2 (n)- и χ (n)-распределения стремятся к нормальному, причем эти распределения с хорошей точностью можно считать нормальными при $n > 30$ (см. рис. 4.6, *б*).

Далее можно показать, что при больших n величины

$$Z_n = (X^2_{(n)} - n)/\sqrt{2n}$$

и

$$Z'_n = \sqrt{2X^2_{(n)}} - \sqrt{2n-1}$$

распределены по стандартному нормальному закону $N(0, 1)$.

Если независимые переменные X_i имеют нормальное распределение $N(\mu_i, 1)$ (не стандартные нормальные, как в предыдущем случае), то $X^2_{(n)}$ уравнения (4.24) имеет *нецентральное* χ^2-*распределение* с n степенями свободы, обозначаемое $\chi^2(n, \Delta)$, где $\Delta = \sum_{i=1}^{n} \mu_i^2$ называется *параметром нецентральности*.

Характеристическая функция $\chi^2(n, \Delta)$-распределения такова:

$$E[\exp(itX_{(n)})] = \frac{\exp[it\Delta/(1-2it)]}{(1-2it)^{n/2}}.$$

Отметим, что $\mu_1, \ldots, \mu_n$ входят в распределение только через Δ. Если $\Delta = 0$ (т. е. $\mu_i = 0$, $i = 1, \ldots, n$), то мы приходим к центральному χ^2-распределению.

Можно приближенно заменить $\chi^2(n, \Delta)$ переменной, пропорциональной центральному χ^2, например, $Y = \beta\chi^2(M)$, выбрав β и M так, что Y имеет такие же среднее и дисперсию, что и $\chi^2(n, \Delta)$. Это имеет место, когда

$$\beta = 1 + \Delta/(n + \Delta);$$

$$m = n + \frac{\Delta^2}{n+2\Delta} = \frac{(n+\Delta)^2}{n+2\Delta}.$$

4.2.4. *t*-Распределение Стьюдента.

Переменная t, реальное число.

Параметр n, положительное целое.

Функция плотности вероятностей

$$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\,\Gamma(n/2)} \frac{1}{(1+t^2/n)^{(n+1)/2}}. \tag{4.26}$$

Ожидаемое значение $E(t) = 0$.

Дисперсия $D(t) = \frac{n}{n-2}$, $n > 2$.

Асимметрия $\gamma_1 = 0$.

Эксцесс $\gamma_2 = \frac{6}{n-4}$, $n > 4$.

Моменты $\mu_{2r+1} = 0$, $2r < n$;

$$\mu_{2r} = \frac{n^r\,\Gamma(r+1/2)\,\Gamma(n/2-r)}{\Gamma(1/2)\,\Gamma(n/2)}, \quad \mu_{2r+1} = 0, \quad 2r < n.$$

Графики рис. 4.7, 9.2.

Моменты более высокого порядка, чем n, не существуют.

В п. 4.2.1 мы видели, что если измерения $X_1, X_2, \ldots, X_n$ нормальны каждая с распределением $N(\mu, \sigma^2)$, переменная $\sqrt{n}(\overline{X} - \mu)$ также нормальна с распределением $N(0, \sigma^2)$. Кроме того,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{4.27}$$

распределено независимо от $\overline{X}$. Можно показать, что $(n-1)\,s^2/\sigma^2$ распределено как χ^2-переменная с $(n-1)$ степенями свободы.

Довольно часто возникает потребность проверить, совместимо ли наблюдаемое среднее $\overline{X}$ с теоретическим значением μ. *Статистика**, которую нужно использовать для такой проверки, имеет вид

$$t = \sqrt{n}(\overline{X} - \mu)/s. \tag{4.28}$$

Эта величина подчиняется *t-распределению Стьюдента* с $(n-1)$ степенями свободы.

В общем случае переменная Y называется *оценкой величины* σ^2 *нормального закона* с k степенями свободы, если kY/σ^2 распределено как χ^2-переменная с k степенями свободы. Если X имеет распределение $N(0, \sigma^2)$ и независимо от Y, то отношение

$$t_k = X/Y$$

имеет t-распределение Стьюдента с k степенями свободы.

Распределение Стьюдента симметрично относительно $t = 0$. При $n = 1$ оно переходит в распределение Коши (см. п. 5.2.1), а при $n \to \infty$ оно приближается к стандартному нормальному распределению $N(0, 1)$. С помощью ф. п. в. можно сконструировать доверительные границы для μ или критерии проверки гипотезы, что $\mu = \mu_0$ (см. гл. 9 — 11).

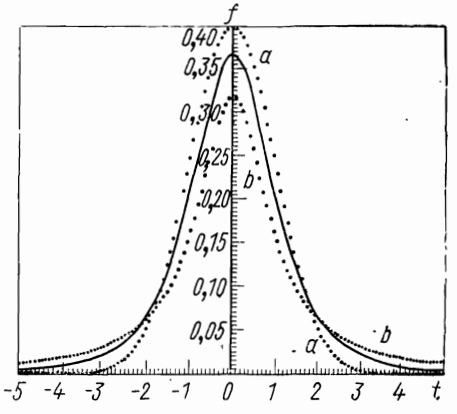


Рис. 4.7. Распределения Стьюдента для $n = 3$ (сплошная линия):
a — стандартное нормальное распределение $N(0,1)$ или распределение Стьюдента для $N = \infty$; b — распределение Коши или распределение Стьюдента для $N = 1$

Если $X_1, \ldots, X_m$ и $Y_1, \ldots, Y_n$ распределены по законам $N(\mu_X, \sigma^2)$ и $N(\mu_Y, \sigma^2)$ соответственно, то величина $(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)$ имеет распределение

$$N\left[0, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right].$$

Объединенная оценка σ^2

$$s^2 = \left[\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] \Big/ (n+m-2) \tag{4.29}$$

является оценкой дисперсии нормального закона с $m + n - 2$ степенями свободы. Здесь s^2 и $\overline{X} - \overline{Y}$ независимы [поскольку ни од-

* Определение проверочной статистики дано в гл. 10.

на из сумм в (4.29) не зависит ни от $\overline{X}$, ни от $\overline{Y}$]. Поэтому величина

$$t = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{1}{m}+\frac{1}{n}}\, s} \tag{4.30}$$

имеет t_{m+n-2}-распределение. Статистика (4.30) позволяет сформулировать критерии проверки того, является ли наблюдаемая разность средних $\overline{X} - \overline{Y}$ совместимой с теоретической разностью $\delta = \mu_X - \mu_Y$. В частности, если $\delta = 0$, то можно проверить справедливость равенства $\mu_X = \mu_Y$. Этот критерий существенно зависит от предположения о том, что два распределения имеют одинаковую дисперсию σ^2. Если это предположение неверно, то критерий становится более сложным.

Если X имеет распределение $N(\Delta\sigma, \sigma^2)$, а nY^2/σ^2 имеет $\chi^2(n)$-распределение (т. е. Y^2 является оценкой нормального закона для σ^2), то переменная $t' = X/Y$ имеет *нецентральное t-распределение* с n степенями свободы и параметром нецентральности Δ при условии, что X и Y независимы.

Очевидно,

$$P(t' = X/Y < k) = P(X - kY < 0). \tag{4.31}$$

Здесь X по определению является нормальной переменной, а Y практически нормально, если $n > 20$. К тому же ожидание и дисперсия разности $X - kY$ приблизительно равны

$$E(X - kY) \simeq \sigma(\Delta - k);$$

$$D(X - kY) \simeq \sigma^2(1 + k^2/n).$$

Следовательно, вероятность в (4.31) приблизительно равна

$$P\left(\frac{X}{Y} < k\right) \simeq \Phi\left(\frac{k-\Delta}{\sqrt{1+k^2/n}}\right),$$

где Φ — интеграл вероятности нормального распределения.

4.2.5. *F*- и *Z*-распределения Фишера—Снедекора.

Переменная F, положительное реальное число.

Параметры ν_1, ν_2, положительные целые (степени свободы).

Функция плотности вероятностей

$$f(F) = \frac{\nu_1^{\nu_1/2}\,\nu_2^{\nu_2/2}\,\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \frac{F^{\nu_1/2-1}}{(\nu_1+\nu_2 F)^{(\nu_1+\nu_2)/2}}. \tag{4.32}$$

Ожидаемое значение

$$E(F) = \frac{\nu_2}{\nu_2-2} \qquad \nu_2 > 2.$$

Дисперсия

$$D(F)=\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)},\quad \nu_2>4.$$

Пусть s_1^2, s_2^2 независимы и являются оценками нормального закона величин σ_1^2 и σ_2^2 с ν_1 и ν_2 степенями свободы соответственно, т. е. $\nu_1 s_1^2/\sigma_1^2$ распределено как $\chi^2(\nu_1)$ и $\nu_2 s_2^2/\sigma_2^2$ распределено как $\chi^2(\nu_2)$ и величины s_1^2 и s_2^2 определены аналогично s^2 в (4.27). Тогда говорят, что величина

$$F=\frac{s_1^2}{\sigma_1^2}\,\frac{\sigma_2^2}{s_2^2}$$

имеет *F-распределение* с (ν_1, ν_2) степенями свободы (или распределение отношения дисперсий). Распределение положительно асимметричное и приближается к нормальному при $\nu_1, \nu_2 \to \infty$, но довольно медленно ($\nu_1, \nu_2 > 50$).

У переменной $Z=\frac{1}{2}\ln F$ распределение гораздо ближе к нормальному. Его свойства таковы:

Переменная Z, реальное число.

Параметры ν_1, ν_2, положительные целые.

Функция плотности вероятностей

$$f(Z)=\frac{2\nu_1^{\nu_1/2}\,\nu_2^{\nu_2/2}}{B\left(\frac{1}{2}\nu_1,\frac{1}{2}\nu_2\right)}\,\frac{e^{\nu_1 Z}}{[\nu_1 e^{2Z}+\nu_2]^{(\nu_1+\nu_2)/2}}. \tag{4.33}$$

Ожидаемое значение

$$E(Z)=\frac{1}{2}\left[\frac{1}{\nu_2}-\frac{1}{\nu_1}\right]-\frac{1}{6}\left[\frac{1}{\nu_1^2}-\frac{1}{\nu_2^2}\right].$$

Дисперсия

$$D(Z)=\frac{1}{2}\left[\frac{1}{\nu_1}+\frac{1}{\nu_2}\right]+\frac{1}{2}\left[\frac{1}{\nu_1^2}+\frac{1}{\nu_2^2}\right]+\frac{1}{3}\left[\frac{1}{\nu_1^3}+\frac{1}{\nu_2^3}\right].$$

Характеристическая функция

$$\zeta(t)=\left(\frac{\nu_2}{\nu_1}\right)^{\frac{1}{2}it}\,\frac{\Gamma\left[\frac{1}{2}(\nu_2-it)\right]\Gamma\left[\frac{1}{2}(\nu_1+it)\right]}{\Gamma\left(\frac{1}{2}\nu_1\right)\Gamma\left(\frac{1}{2}\nu_2\right)}.$$

В настоящее время на практике предпочтением пользуется более простая статистика F, несмотря на то, что ее распределение значительно сильнее отличается от нормального.

При $\nu_1 = 1$ F-распределение сводится к $F = t^2_{\nu_2}$-распределению. При $\nu_2 \to \infty$ F приближается к $\frac{1}{\nu_1}\chi^2(\nu_1)$.

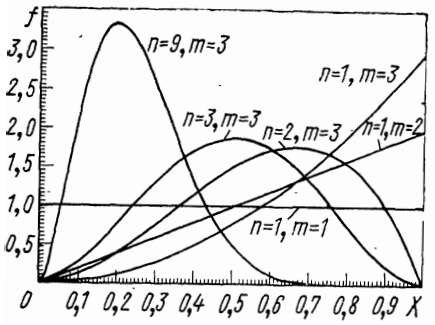


Рис. 4.8. Бета-распределение

Функция $U = \nu_1 F/(\nu_2 + \nu_1 F)$ эквивалентна F при проверке гипотез, монотонна в зависимости от F и имеет бета-распределение (см. п. 4.2.8) (рис. 4.8).

4.2.6. Равномерное распределение.

Переменная X, реальное число.

Параметры $a, b, a < b$.

Функции плотности вероятностей

$$f(X) = 1/(b-a), \quad a \leqslant X \leqslant b. \tag{4.34}$$

Ожидаемое значение $E(X) = (a+b)/2$.

Дисперсия $D(X) = (b-a)^2/12$.

Асимметрия $\gamma_1 = 0$.

Эксцесс $\gamma_2 = -1{,}2$.

Характеристическая функция

$$\zeta(t) = \frac{\sin h\left[\frac{1}{2} it(b-a)\right]}{it(b-a)} + \frac{1}{2} it(b+a).$$

Погрешности округления при арифметических вычислениях распределены *равномерно*.

4.2.7. Треугольное распределение.

Переменная X, реальное число.

Параметры $\Gamma > 0$, μ, реальные числа.

Функция плотности вероятностей

$$f(X) = -\frac{|X-\mu|}{\Gamma^2} + \frac{1}{\Gamma}, \quad \mu - \Gamma < X < \mu + \Gamma; \tag{4.35}$$

$f(X) = 0$ при всех других X.

Ожидаемое значение

$$E(X) = \mu.$$

Дисперсия

$$D(X) = \Gamma^2/6.$$

Асимметрия

$$\gamma_1 = 0.$$

Эксцесс

$$\gamma_2 = -0{,}6.$$

Примеры: 1. Сумма двух чисел, каждое из которых выбрано независимо в соответствии с равномерным распределением, имеет треугольное распределение со средним и дисперсией, равными удвоенным среднему и дисперсии равномерного распределения.

2. Импульсное распределение вторичных частиц синхротронного пучка зачастую очень близко к треугольному с центральным моментом μ и полной шириной Γ на полувысоте. Высота распределения равна $1/\Gamma$, и ее основание простирается от $\mu - \Gamma$ до $\mu + \Gamma$.

4.2.8. Бета-распределение.

Переменная X, реальное число.

Параметры n, m, положительные целые.

Функция плотности вероятностей

$$f(X) = \frac{\Gamma(n+m)}{\Gamma(n)\,\Gamma(m)} X^{m-1}(1-X)^{n-1}; \quad 0 \leqslant X \leqslant 1; \qquad (4.36)$$

$f(X) = 0$ при всех других X.

Ожидаемое значение

$$E(X) = \frac{m}{m+n}.$$

Дисперсия

$$D(X) = \frac{mn}{(m+n)^2(m+n+1)}.$$

Асимметрия

$$\gamma_1 = \frac{2(n-m)\sqrt{(m+n+1)}}{(m+n+2)\sqrt{mn}}.$$

Эксцесс

$$\gamma_2 = \frac{3(m+n+1)[2(m+n)^2+mn(m+n-6)]}{mn(m+n+2)(m+n+3)} - 3.$$

График рис. 4.8

Бета-распределение является основным распределением в статистике для переменных, ограниченных с обеих сторон, например, $0 \leqslant X \leqslant 1$. Например, оно используется как распределение доли совокупности, расположенной между наименьшим и наибольшим значениями в выборке, а также как распределение времени, затраченного на выполнение работы в анализах, которые проводятся по методике системы ПЕРТ. Частные случаи бета-распределений: равномерное, треугольное и параболическое.

4.2.9. Экспоненциальное распределение.

Переменная X, положительное реальное число.

Параметр λ, положительное реальное число.

Функция плотности вероятностей

$$f(X) = \lambda \exp(-\lambda X). \qquad (4.37)$$

Ожидаемое значение

$$E(X) = 1/\lambda.$$

Дисперсия

$$D(X) = 1/\lambda^2.$$

Асимметрия $\gamma_1 = 2$.

Эксцесс $\gamma_2 = 6$.

Характеристическая функция $\zeta(t) = (1 - it/\lambda)^{-1}$.

Предположим, что события распределены во времени случайно и среднее число событий в единицу времени равно λ. Согласно распределению Пуассона, вероятность наблюдения N событий в течение времени t:

$$P_N(t) = \frac{1}{N!}(\lambda t)^N \exp(-\lambda t).$$

Тогда вероятность того, что за время t не будет ни одного события, описывается *экспоненциальным распределением* $\exp(-\lambda t)$.

Рассмотрим интервал времени Z между двумя следующими друг за другом событиями. В течение этого интервала времени не наблюдается ни одного события. Таким образом, для фиксированного X

$$P(Z > X) = \exp(-\lambda X)$$

с функцией распределения экспоненциального закона

$$F(X) = P(Z \leqslant X) = 1 - \exp(-\lambda X).$$

Экспоненциальное распределение «не имеет памяти»: если до момента времени y не произойдет ни одного события, то вероятность того, что в течение последующего интервала времени x также не произойдет ни одного события, не зависит от y. Для фиксированного y

$$P(X > x + y \mid X > y) = \frac{\exp[-\lambda(x+y)]}{\exp(-\lambda y)} = \exp(-\lambda x) = P(\lambda > x).$$

В соответствии с экспоненциальным законом распределено время между моментами попадания частиц в счетчик.

С экспоненциальным распределением тесно связаны *гиперэкспоненциальное* распределение и эрланжиановское распределение. Гиперэкспоненциальное распределение описывает распределение времени между событиями в процессе, в котором с вероятностью p_1 события порождаются одним экспоненциальным законом со скоростью λ_1 и с вероятностью $(1-p_1)$ другим экспоненциальным законом со скоростью λ_2. Тогда функция плотности вероятности равна:

$$f(X) = p_1\lambda_1 \exp(-\lambda_1 X) + (1 - p_1)\lambda_2 \exp(-\lambda_2 X).$$

Таким образом, гиперэкспоненциальное распределение применимо там, где имеется смесь процессов, описываемых экспоненциальным распределением.

Эрланжиановское k-распределение описывает распределение времени между k-ми событиями экспоненциального процесса; оно является распределением суммы k случайных переменных, распределенных экспоненциально с частотой $k\mu$. Функция плотности вероятности равна

$$f(X) = \frac{k\mu}{(k-1)!} (k\mu X)^{k-1} \exp(-k\mu X).$$

Гиперэкспоненциальные процессы соответствуют экспоненциальным процессам, происходящим параллельно, тогда как эрланжиановские распределения — последовательным экспоненциальным процессам.

4.2.10. Гамма-распределение.

Переменная X, реальное положительное число.

Параметры a, b, реальные положительные числа.

Функция плотности вероятностей

$$f(X) = \frac{a(aX)^{b-1} \exp(-aX)}{\Gamma(b)}. \qquad (4.38)$$

Ожидаемое значение $E(X) = b/a$.

Дисперсия $D(X) = b/a^2$.

Асимметрия $\gamma_1 = 2/\sqrt{b}$.

Эксцесс $\gamma_2 = 6/b$.

Характеристическая функция

$$\zeta(t) = (1 - it/a)^{-b}.$$

График рис. 4.9.

Гамма-распределение является основным статистическим средством при описании переменных, ограниченных с одной стороны, например, $0 \leqslant X < \infty$. С помощью рис. 4.9 можно получить представление о разнообразии форм распределения, соответствующем различным значениям параметров. Заметим, что a является просто масштабным параметром.

Гамма-распределение применимо ко времени между повторными калибровками прибора, который нуждается в переградуировке после k употреблений, и ко времени отказа в работе для систем с запасными компонентами.

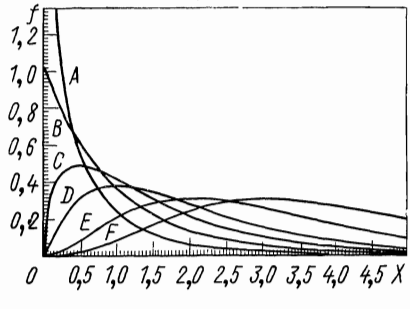


Рис. 4.9. Гамма-распределение:
$a = 1$ и $b = 0{,}5$: 1; 1,5; 2; 3; 4 для кривых $A - F$ соответственно

Экспоненциальное χ^2-распределение, а также эрланжиановское распределение являются частными случаями гамма-распределения.

4.2.11. Распределение Коши или распределение Брейта—Вигнера.

Переменная X, реальное число.

Параметры — нет, за исключением, может быть, параметров положения и масштаба.

Функция плотности вероятностей

$$F(X) = \frac{1}{\pi} \frac{1}{1+X^2}. \tag{4.39}$$

Ожидаемое значение $E(X)$ — не определено.

Дисперсия, асимметрия, эксцесс — расходящиеся.

Характеристическая функция $\zeta(t) = \exp(-|t|)$.

Графики рис. 2.2, 4.7, 4.10.

Распределение Коши — пример патологического распределения в статистике, поскольку, как мы уже отмечали в п. 2.4.1, его ожидание не определено, а все другие моменты расходящиеся. В физике оно соответствует важному случаю распределения Брейта — Вигнера, которое записывается в виде

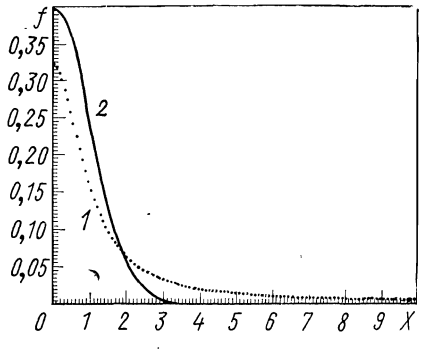


Рис. 4.10. Сравнение хвостов распределения Коши (*1*) и стандартного нормального распределения $N(0,1)$ (*2*)

$$f(X) = \frac{1}{\pi}\left(\frac{\Gamma}{\Gamma^2+(X-X_0)^2}\right). \tag{4.40}$$

Параметры X_0 и Γ являются параметрами расположения и масштаба или модой и полушириной на полувысоте соответственно. Тем не менее нужно отметить, что среднее и моменты распределения не определены.

В п. 5.3.3 мы увидим, как в результате обрезания распределение Коши приобретает свойства обычного распределения. С помощью рис. 4.10 можно сравнить ф.п.в. нормального и распределения Коши.

Отметим, что свойства распределения (4.40) могут быть совершенно другими, если Γ зависит от X.

4.2.12. Логарифмически-нормальное распределение.

Переменная X, положительное реальное число.

Параметры μ, реальное число; σ, положительное реальное число.

Функция плотности вероятностей

$$f(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \frac{1}{X} \exp\left[-\frac{1}{2\sigma^2}(\ln X - \mu)^2\right]. \tag{4.41}$$

Ожидаемое значение $E(X) = \exp\left(\mu - \frac{1}{2}\sigma^2\right)$.

Дисперсия $D(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$.

Асимметрия $\gamma_1 = \sqrt{(\exp(\sigma^2) - 1)}\ (\exp(\sigma^2) + 2)$.

Эксцесс $\gamma_2 = (\exp(\sigma^2) - 1)(\exp(3\sigma^2) + 3\exp(2\sigma^2) + 6\exp(\sigma^2) + 6)$.

График рис. 4.11.

Логарифмически-нормальное распределение соответствует случайной переменной, логарифм которой распределен по нормальному закону. Оно служит моделью для описания ошибки некоторого про-

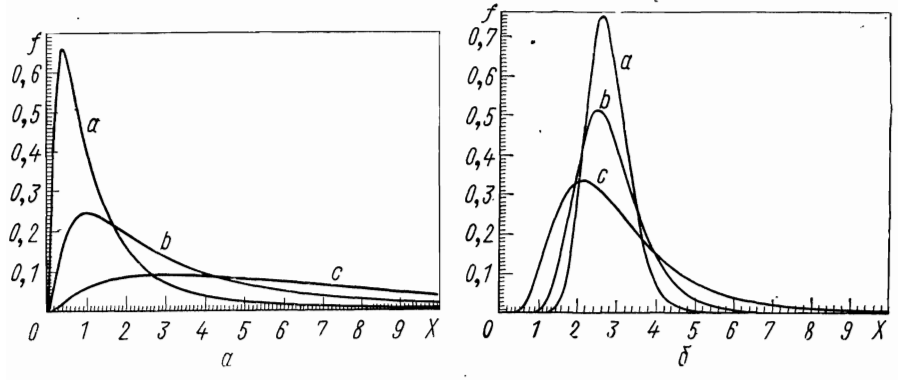


Рис. 4.11. Логарифмически-нормальное распределение:
а — $\sigma = 1$, $\mu = 0$, 1, 2 для кривых *a*, *b*, *c*; *б* — $\mu = 1$, $\sigma = 0{,}2$; 0,3; 0,5 для кривых *a*, *b*, *c*

цесса, включающего большое число малых мультипликативных ошибок (в соответствии с предельной центральной теоремой). Оно также применимо в тех случаях, когда наблюдаемая переменная является случайной долей предыдущего наблюдения.

4.2.13. Распределение экстремального значения.

Переменная X, реальное число.

Параметры μ, реальное число; σ, положительное реальное число.

Функция плотности вероятностей

$$f(X) = \frac{1}{\sigma}\exp\left[\pm\frac{\mu - X}{\sigma} - \exp\left[\pm(\mu - X)/\sigma\right]\right]. \qquad (4.42)$$

Функция распределения

$$F(X) = \exp\left[-\exp\left(\pm(\mu - X)/\sigma\right)\right].$$

Ожидаемое значение $E(X) = \mu \pm 0{,}5776\ \sigma$.

Дисперсия $D(X) = 1{,}645\sigma^2$.

Асимметрия $\gamma_1 = \pm 1{,}14$.

Эксцесс $\gamma_2 = 2{,}4$.

График рис. 4.12.

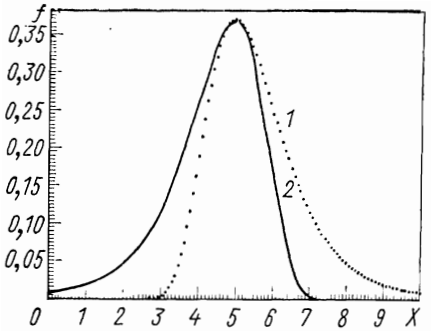


Рис. 4.12. Распределение экстремального значения наибольшего (*1*) и наименьшего (*2*) элемента для $\mu=5$, $\sigma=1$

Распределение экстремального значения дает предельное распределение для наибольшего (со знаком +) или наименьшего (со знаком —) элемента выборки независимых наблюдений из распределения экспоненциального типа (нормального, гамма, экспоненциального и т. д. [21]).

В качестве примера рассмотрим эксперимент, в результате которого фиксируется большое число n независимых гистограмм. Предположим, что необходимо описать отклонение каждой гистограммы от известной теории одним числом, например, среднеквадратичным отклонением в ячейках. Обозначим это число χ^2. Для того чтобы уметь оценивать значимость наблюдения одного особенно большого значения χ^2, нужно рассмотреть распределение вероятностей экстремального значения из n χ^2-распределений.

Предположим для простоты, что все гистограммы имеют одинаковое число ячеек и что число степеней свободы r всех n $\chi^2(r)$-распределений достаточно велико ($r \geqslant 30$), чтобы их можно было аппроксимировать нормальным законом:

$$X=\sqrt{2\chi^2}-\sqrt{2r-1}.$$

Случайная переменная X нормальна с нулевым средним, а наибольшее значение X описывается тогда распределением (4.42) [1] с

$$\mu=\sqrt{2\ln n};\ \sigma=\mu^{-1}.$$

Достоинством этого критерия по сравнению с другими более сложными критериями, описанными в гл. 11, является то, что он оперирует с одним числом и удобен при расчетах, выполняемых вручную.

4.2.14. Распределение Вейбула.

Переменная X, положительное реальное число.

Параметры η, σ, положительные реальные числа.

Функция плотности вероятностей

$$f(X)=\frac{\eta}{\sigma}\left(\frac{X}{\sigma}\right)^{\eta-1}\exp\left[-\left(\frac{X}{\sigma}\right)^{\eta}\right],\quad X\geqslant 0. \tag{4.43}$$

Ожидаемое значение $E(X)=\sigma\Gamma(1/\eta+1)$.

Дисперсия $D(X)=\sigma^2\left\{\Gamma\left(\frac{2}{\eta}+1\right)-\left[\Gamma\left(\frac{1}{\eta}+1\right)\right]^2\right\}$.

График рис. 4.13.

Распределение Вейбула описывает распределение ожидаемого времени отказа в работе большой разновидности сложных механизмов. Экспоненциальное распределение является частным случаем этого распределения, когда вероятность такого нарушения в момент времени t не зависит от t.

Соответствующий вид распределения Вейбула описывает также распределение минимума в выборке наблюдений из распределений переменных, ограниченных снизу.

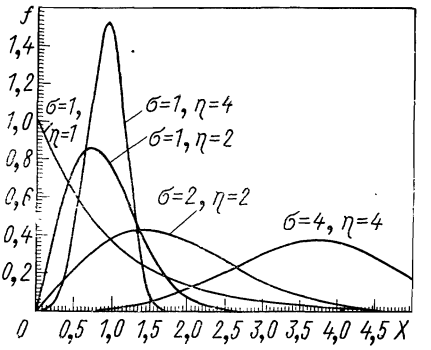


Рис. 4.13. Распределения Вейбула

4.2.15. Двойное экспоненциальное распределение

Переменная X, реальное число.

Параметры μ, реальное число; λ, положительное реальное число.

Функция плотности вероятностей

$$f(X) = \frac{\lambda}{2} \exp(-\lambda |X - \mu|). \qquad (4.44)$$

Ожидаемое значение $E(X) = \mu$.
Дисперсия $D(X) = 2/\lambda^2$.
Асимметрия $\gamma_1 = 0$.
Эксцесс $\gamma_2 = 3$.
Характеристическая функция

$$\zeta(t) = it\mu + \lambda^2/(\lambda^2 + t^2).$$

Двойное экспоненциальное распределение (иногда называют распределением Лапласа) является симметричной функцией, чьи хвосты

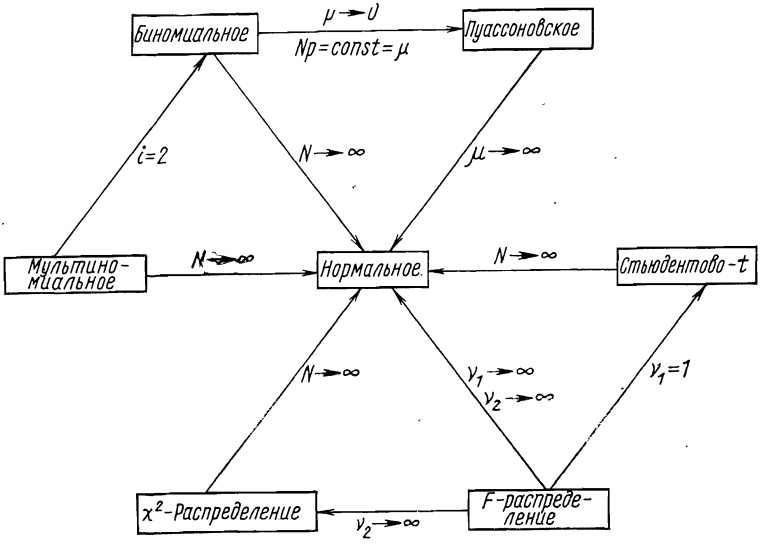


Рис. 4.14. Предельные условия для параметров распределений, при которых осуществляются условия сходимости, указанные стрелками

спадают менее резко, чем у гауссовского распределения, но быстрее, чем у распределения Коши. В отличие от последних двух распределений двойное экспоненциальное распределение имеет точку излома (первая производная имеет разрыв) при $X = \mu$.

Это распределение интересно с точки зрения оценки параметров, так как вследствие симметрии функции распределения наилучшей оценкой для среднего μ (по существу достаточной статистикой — см. § 5.3) служит медиана выборки.

4.2.16. Соотношения между распределениями в асимптотическом пределе. Некоторые из обсуждавшихся ранее распределений в асимптотическом пределе сходятся при различных дополнительных условиях к другим типам распределений. Не желая обсуждать детально каждый случай, мы иллюстрируем асимптотические свойства распределений с помощью рис. 4.14

§ 4.3. РАСПРЕДЕЛЕНИЯ, ВСТРЕЧАЮЩИЕСЯ НА ПРАКТИКЕ

В § 4.2 мы познакомились с наиболее важными идеальными распределениями, а также с их некоторыми полезными свойствами. В высшей степени наивно думать, что все физические распределения имеют некоторую идеализированную форму. Несмотря на то что при некоторых условиях идеальные распределения встречаются в физике, реальная жизнь, к несчастью, не столь проста, и часто приходится иметь дело с комбинацией этих идеальных распределений либо с распределениями, получающимися в результате искажений идеальных распределений. Покажем, как нужно работать с некоторыми типичными реальными распределениями, используя методику, разработанную ранее для идеальных распределений.

4.3.1. Применимость нормального распределения в общем случае. Возникает вопрос, естественно, с какой степенью веры последовательность измерений априори можно считать распределенной нормально. Например, часто предполагается, что человек, проводящий повторные измерения расстояния между двумя фиксированными точками, получит набор измерений, которые распределены нормально со средним, равным «истинному» расстоянию, и шириной, определяемой точностью используемого метода.

Математики иногда думают, что именно практика доказывает нормальность распределения измерительных ошибок. С другой стороны, экспериментаторы часто думают, что это обстоятельство доказано математически, и не проверяют выполнимость условий центральной предельной теоремы. Действительно, с обеих сторон (математической и экспериментальной) имеются подтверждения того, что нормальное распределение служит по меньшей мере хорошим приближением для реального распределения. Поэтому распределение считается нормальным потому, что с ним удобно работать (многие математически простые свойства присущи только нормальным распределениям), а также потому, что это обстоятельство до некоторой

степени подтверждается на практике. Ниже мы приводим два примера, в одном из которых предположение о нормальности распределения верно, в другом нет.

1. Сумма или среднее значение многих независимых наблюдений. В этом случае может быть применима центральная предельная теорема (см. п. 3.3.2), из которой следует, что сумма или среднее различных измерений будут (асимптотически) распределены по нормальному закону, даже если распределение генеральной совокупности не является нормальным (они даже не обязательно должны иметь одинаковые распределения генеральной совокупности).

Предположим, например, что при изучении π^--мезонов, вылетающих из мишени, расположенной в пучке высокоэнергетичных частиц, нужно измерить среднее значение импульса π^--мезонов. Известно, что импульсный спектр π^--мезонов, рождающихся в данной реакции, не является гауссовским и, кроме того, общий импульсный спектр π^--мезонов, рождающихся во всех реакциях, также не является нормальным. Однако, если рассматривать среднее значение импульса N следующих друг за другом π^--мезонов как случайную переменную, то повторные наблюдения такого типа будут распределены по нормальному закону при больших N.

2. Совокупность измерений, каждое из которых распределено по нормальному закону, имеет одинаковое среднее, но различные дисперсии. В этом случае, даже несмотря на то, что каждое отдельное измерение — нормально распределенная случайная переменная, распределение суммы нормальных распределений с различными ширинами никогда не является нормальным

В качестве примера рассмотрим массы узкого резонанса или стабильной частицы в эксперименте на пузырьковой камере. Если предполагается, что одно измерение массы распределено по нормальному закону со стандартной ошибкой ΔM, то результирующий спектр масс будет иметь вид нормального распределения, только если выбранные события имеют одинаковые ΔM. В более общем случае, когда имеется разброс значений ΔM (зачастую большой), результирующее распределение может быть очень далеким от нормального. Более детально этот важный случай описан в п. 4.3.4.

Подытоживая эти два примера, можно сказать, что при некоторых общих условиях распределение суммы независимых случайных переменных является нормальным, но сумма нормально распределенных переменных с различными дисперсиями (идеограмма) распределена не по нормальному закону.

4.3.2. Эмпирические распределения Джонсона. Иногда полезно представлять данные эмпирическим распределением. Это можно проделать вручную, но аппроксимация наблюденных данных некоторой математической формой обладает теми преимуществами, что она более объективна, выполняется автоматически и в результате получается математическое выражение, удобное для последующего использования.

Джонсоновские семейства распределений обладают довольно большим разнообразием форм. По своему назначению они служат преобразованиями нормального распределения. Пусть X—случайная

переменная, чье распределение должно быть получено эмпирически. Тогда в общем случае преобразование имеет вид:

$$Z = \gamma + \eta g\,(X;\ \varepsilon,\ \lambda).$$

Здесь Z — переменная стандартного нормального закона и $\eta,\ \lambda > 0$, $-\infty < \gamma,\ \varepsilon < \infty$; d — произвольная функция.

Джонсон предложил три различных типа функции g, соответствующие случаям, когда переменная X не ограничена, ограничена снизу и ограничена и сверху и снизу:

1) $g_1\,(X;\ \varepsilon,\ \lambda) = \operatorname{arcsin} h\,((X - \varepsilon)/\lambda)$, $-\infty < X < \infty$, называемый джонсоновским S_U семейством распределений;

2) $g_2\,(X;\ \varepsilon,\ \lambda) = \ln\,((X - \varepsilon)/\lambda)$, $X \geqslant \varepsilon$, называемый джонсоновским S_L семейством распределений;

3) $g_3\,(X;\ \varepsilon,\ \lambda) = \ln\left(\dfrac{X - \varepsilon}{\lambda - (X - \varepsilon)}\right)$, $\varepsilon \leqslant X \leqslant \lambda + \varepsilon$, называемый джонсоновским S_B семейством распределений.

Функции плотностей этих распределений имеют вид:

1) $$f_1(X) = \frac{\eta}{\sqrt{2\pi}}\,\frac{1}{\sqrt{(X-\varepsilon)^2 + \lambda^2}}\,\exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{(X-\varepsilon)}{\lambda} + \sqrt{\left(\frac{X-\varepsilon}{\lambda}\right)^2 + 1}\right)\right]^2\right\};$$

2) $$f_2(X) = \frac{\eta}{\sqrt{2\pi}\,(X-\varepsilon)}\,\exp\left\{-\frac{1}{2}\left[\gamma + \eta \ln\left(\frac{X-\varepsilon}{\lambda}\right)\right]^2\right\};$$

3) $$f_3(X) = \frac{\eta}{\sqrt{2\pi}}\,\frac{\lambda}{(X-\varepsilon)\,(\lambda - X + \varepsilon)}\,\exp\left\{-\frac{1}{2} \times \left[\gamma + \eta \ln\left(\frac{X-\varepsilon}{\lambda - X + \varepsilon}\right)\right]^2\right\}.$$

Со свойствами этих распределений более подробно можно познакомиться в работе [22].

4.3.3. Обрезание. Простейшая модификация идеальных распределений обусловлена тем фактом, что область возможных значений измерений никогда не является бесконечной. Обычно значения переменной X лежат внутри конечных пределов A и B, поэтому ф.п.в. приобретает такой вид:

$$f(X)\,dX \to \frac{f(X)\,dX}{\int\limits_A^B f(X)\,dX} = \frac{f(X)\,dX}{F(B) - F(A)}.$$

Хотя такая процедура обычно усложняет выражение, она иногда бывает полезной. Это имеет место для распределения Коши или Брейта—Вигнера. Выше мы видели, что это распределение не имеет моментов, если область определения не ограничена, но если она ограничена значениями A и B, то нормировка, среднее и дисперсия становятся конечными:

$$g(X) = \frac{f(X)}{\int\limits_{-A}^{A} \frac{1}{\pi} \frac{1}{1+X^2}\, dX} = \frac{1}{2 \operatorname{arctg} A} \frac{1}{1+X^2};$$

$$E(X) = \frac{1}{2 \operatorname{arctg} A} \int\limits_{-A}^{A} \frac{X}{1+X^2}\, dX = 0;$$

$$D(X) = \frac{1}{2 \operatorname{arctg} A} \int\limits_{-A}^{A} \frac{X^2}{1+X^2}\, dX = \frac{A}{\operatorname{arctg} A} - 1.$$

Если $A \to \infty$, то становится понятным, почему ожидание не существует:

$$\lim D(X) = \infty.$$

Хвосты распределения Коши падают настолько медленно, что дисперсия может быть произвольно большой, если A достаточно велико. (Это не случается с гауссовским распределением; см. рис. 4.10.)

Обрезание является всего лишь частным случаем более общего класса искажений, совокупность которых приводит к задаче определения *эффективности регистрации*. Если события регистрируются аппаратурой с различной вероятностью, то идеальная ф.п.в. $f(X)$ искажается и приобретает вид:

$$g(X) = \frac{\int\limits_{\Omega_Y} f(X)\, P(Y \mid X)\, e(X, Y)\, dY}{\int\limits_{\Omega_X} \int\limits_{\Omega_Y} f(X)\, P(Y \mid X)\, e(X, Y)\, dY\, dX},$$

где Y обозначает набор вспомогательных переменных, а $e(X, Y)$ имеет смысл плотности вероятности наблюдения события в точке (X, Y). Мы отложим детальное обсуждение этой проблемы до § 8.5.

4.3.4. Экспериментальное разрешение. Другим довольно частым источником искажений идеальных распределений служит экспериментальная неточность измерений, т. е. событие с истинным значением X будет измерено в некоторой другой точке X'. Распределение измеряемых значений X' относительно истинного X определяется плотностью вероятности, иногда называемой функцией разрешения: $r(X, X')$.

Тогда, если истинная плотность распределения X равна $f(X)$, то плотность распределения измеряемых величин будет иметь вид:

$$g(X') = \int\limits_{\Omega} r(X, X')\, f(X)\, dX. \qquad (4.45)$$

Это искажение принципиально отличается от всех тех, которые приводят к задаче определения эффективности регистрации, поскольку по «истинной» переменной было проведено интегрирование, и в результате осталась только измеряемая переменная X'. Это может привести, например, к тому, что измеряемая переменная принимает такие значения, для которых функция плотности идеального распределения равна нулю.

Как указывалось ранее в п. 4.3.1, часто предполагается, что функция разрешения $r(X, X')$ имеет вид нормального распределения

$$r(X, X') = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(X'-X)^2}{2\sigma^2}\right]. \qquad (4.46)$$

Легко получить два важных следствия, обусловленные применением этой идеальной функции разрешений. Если истинное распределение равно $N(\mu, \tau^2)$, а функция разрешения имеет вид (4.46), то результирующее измеряемое распределение таково:

$$g(X') = \frac{1}{\sqrt{2\pi}\,\sqrt{\sigma^2+\tau^2}} \exp\left[-\frac{(X'-\mu)^2}{2(\sigma^2+\tau^2)}\right] = N(\mu, \sigma^2+\tau^2),$$

т. е. дисперсия результирующего гауссовского распределения равна сумме дисперсий первоначального гауссовского распределения и гауссовской функции разрешения.

Часто встречается случай, когда первоначальное распределение $f(X)$ является распределением Коши или Брейта—Вигнера, а функция разрешения нормальна. Тогда интеграл (4.45) может быть выражен в виде *функции ошибок комплексного аргумента* [23].

Пусть $r(X, X')$ задана уравнением (4.46) и $f(X)$ — распределением Брейта—Вигнера:

$$f(X) = \frac{1}{\pi} \frac{\Gamma\sqrt{X_0}}{(X-X_0)^2+\Gamma^2 X_0}.$$

Сделаем подстановки:

$$x = (X'-X_0)/\sigma\sqrt{2};\quad y = \Gamma\sqrt{X_0}\,\sigma\sqrt{2};$$
$$z = x+iy;\quad t = (X'-X)/\sigma\sqrt{2},$$

тогда интеграл (4.45) может быть записан в виде:

$$g(X') = \frac{y}{\sqrt{\pi}} \int_{-\infty}^{\infty} dt\, \frac{y\exp(-t^2)}{(x-t)^2+y^2} = y\sqrt{\pi}\ \mathrm{Re}\, w(z),$$

где $w(z)$ — функция ошибок комплексного аргумента

$$w(z) = 2\exp(-z^2)\,\Phi(-i\sqrt{2}\,z).$$

В предельном случае, когда σ^2 мало по сравнению с Γ, форма результирующего распределения определяется только брейт-вигнеровским распределением. Когда верно обратное, то результирующее распределение почти $N(\mu, \sigma^2)$, где μ—ожидание обрезанного брейт-вигнеровского распределения.

4.3.5. Примеры переменного экспериментального разрешения. Если функция разрешения для данного измерения является гауссовской, но дисперсия σ^2 неодинакова для всех измерений, полученных в эксперименте, то общая функция разрешения эксперимента не гауссовская. Рассмотрим три примера *переменного разрешения*.

1. Если истинное распределение $f(X)$ равномерное или медленно изменяется, то наличие переменного разрешения приведет к возникновению пиков в тех точках, где разрешение наилучшее. Несмотря на то что этот факт очевиден, о нем часто забывают, поскольку обычно искажения делают измеренное распределение более широким или гладким по сравнению с истинным распределением.

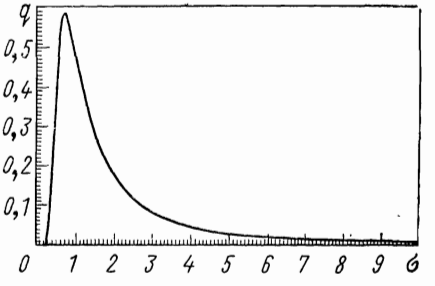


Рис. 4.15. Функция $q(\sigma)$ [см. уравнение (4.48)] с $\tau=1$

2. Предположим, что σ зависит от некоторых вспомогательных переменных (например, от расположения события в трековой камере) и ее распределение может быть описано функцией плотности $q(\sigma^2)$. Тогда функция разрешения будет равна

$$r(X, X')=\int_0^\infty q(\sigma^2)\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(X'-X)^2}{2\sigma^2}\right)d\sigma. \qquad (4.47)$$

В этом случае r будет гауссовской, только если $q(\sigma^2)$ является δ-функцией (т. е. σ^2 постоянно). Обычно измерение событий, регистрируемых с помощью соответствующей аппаратуры, осуществляется с различной точностью. Эта точность зависит, например, от такой величины, как длина трека, которая может быть почти независима от интересующей физика характеристики.

Рассмотрим сначала функциональную форму:

$$q(\sigma)=\frac{2}{\sqrt{2\pi}\,\tau\sigma^2}\exp\left(-\frac{1}{2\tau^2\sigma^2}\right). \qquad (4.48)$$

При больших значениях σ это распределение стремится к нулю как $1/\sigma^2$, а при малых значениях σ наблюдается довольно резкий обрыв ($q(\sigma)$ и все ее производные равны нулю при $\sigma=0$) (рис. 4.15). Функция распределения, соответствующая такому распределению ошибок, рассчитывается следующим образом:

$$r(X, X')=\int_0^\infty \frac{2}{2\pi\tau\sigma^3}\exp\left(-\frac{1}{2\tau^2\sigma^2}\right)\exp\left[-\frac{1}{2}\frac{(X-X')^2}{\sigma^2}\right]d\sigma=$$

$$=\frac{1}{\pi}\frac{\tau}{1+\tau^2(X-X')^2}.$$

В результате мы приходим к распределению Коши или Брейта—Вигнера, которое, как мы уже видели, не имеет среднего и дисперсии, если оно не обрезано. Поскольку на практике это распределение всегда обрезано, не следует беспокоиться о длинных хвостах, которые приводят к бесконечной дисперсии, а достаточно проверить, верно ли поведение $q(\sigma)$ при малых значениях σ. Если это поведение приблизительно верно, то функция распределения имеет форму Брейта—Вигнера, и это позволяет описывать как широкие, так и узкие резонансы, используя одну и ту же общую форму. Напротив, если при описании резонанса, ширина которого уже по сравнению с экспериментальным разрешением, используется формула Брейта—Вигнера, это эквивалентно предположению о том, что $q(\sigma)$ верна при малых σ.

3. Предположим, что $q(\sigma^2)$ в (4.47) соответствует равномерному распределению по переменной $1/\sigma$ в интервале $A \leqslant 1/\sigma \leqslant B$. В некоторых случаях такая формула является хорошим приближением к действительности, например, если точность зависит от длины трека в пузырьковой камере, а все длины треков равновероятно распределены между двумя конечными значениями.

Пусть $t = 1/\sigma$, тогда $q(t) = \dfrac{1}{B-A}\Big|_A^B$

и

$$r(X, X') = \frac{1}{(B-A)\sqrt{2\pi}} \int_A^B t \exp\left[-\frac{(X-X')^2 t^2}{2}\right] dt =$$

$$= \frac{1}{(B-A)\sqrt{2\pi}(X-X')^2} \{\exp[-(X-X')^2 A^2/2] -$$

$$-\exp[-(X-X')^2 B^2/2]\}.$$

Полагая $X' = 0$, можно найти ожидание и дисперсию $r(X, 0)$:

$$E(X) = 0; \; D(X) = 1/AB.$$

Таким образом, дисперсия этой функции разрешения может быть произвольно большой, если нижняя граница на точность становится произвольно малой.

Предположим, что нужно оценить $E(X)$ путем *усреднения* различных наблюдений, соответствующих этому распределению, с фиксированной максимальной точностью B. Очевидно, дисперсия будет меньше, если выбросить события с наихудшей точностью, но это также уменьшает объем выборки n. Выбрасывая все события с дисперсией между минимумом A и некоторым значением A', будем иметь такую дисперсию среднего:

$$D(\bar{X}) = \frac{1}{n} D(X) \sim \frac{1}{A' B (B-A')}.$$

Дисперсия минимальна, когда $A' = B/2$. Таким образом, при такой функции ошибок, если оценка среднего определяется как невзвешенное среднее, то минимум дисперсии получается в том случае, если не используются события с точностью $1/\sigma$, меньшей половины максимальной точности.

Кажущийся здесь парадокс легко объяснить: простое невзвешенное среднее не является наилучшей оценкой среднего. Позднее это будет обсуждаться в теории оценок.

ГЛАВА 5

ИНФОРМАЦИЯ

Понятие информации может быть определено по-разному. Для того чтобы как-то отличить друг от друга различные определения этого понятия, обычно используют фамилию человека, давшего соответствующее определение. Больше всего нам подходит определение информации, введенное Фишером. Требования, которым должно удовлетворять понятие информации, таковы.

1. Информация, содержащаяся в выборке наблюдений, должна *возрастать с увеличением числа наблюдений*. Удвоение числа событий должно удваивать количество доступной информации, если события независимы.

2. Информация должна быть *связана с тем, что мы изучаем* в эксперименте. Данные,которые не имеют отношения к проверяемой гипотезе (или измеряемым параметрам), не должны содержать какой-либо информации. Конечно, те же самые данные могут содержать информацию о других параметрах или гипотезах.

3. Информация *должна быть связана с точностью*: чем больше количество информации, тем выше точность эксперимента.

Очевидно, величина, обладающая такими свойствами, будет ценным инструментом при *обработке данных*. Сейчас трудно найти эксперимент, имеющий так мало первоначальных данных, что все они могут быть опубликованы или даже записаны. Поэтому нужен критерий для эффективного преобразования или обработки данных. При этом необходимо стремиться к максимальному сокращению первоначальных данных, совместимых с минимальной потерей информации. В частности, мы будем детально изучать вопрос, когда потеря информации может быть сделана равной нулю.

Примером других определений информации, которые здесь не будут рассматриваться, могут служить информация Кульбака, которая имеет чисто теоретический интерес, и информация Шеннона, используемая в теории связи [24—27].

§ 5.1. ОСНОВНЫЕ ПОНЯТИЯ

Мы будем рассматривать действительную случайную переменную (или совокупность n реальных случайных переменных) с ф.п.в. $f(\mathbf{X} \mid \theta)$, где θ является действительным параметром (или совокуп-

ностью k действительных параметров). Множество всех допустимых значений $\mathbf{X}$ (область $\mathbf{X}$) будет обозначаться Ω_θ, где индекс отражает возможную зависимость Ω от параметра $\boldsymbol{\theta}$.

5.1.1. Функция правдоподобия. Рассмотрим выборку из независимых наблюдений X, скажем, $X_1, \ldots, X_n$. Ими могут быть n событий, найденных в эксперименте, где каждому событию соответствует измерение p-величин. Совместная ф.п.в. всех X вследствие независимости равна:

$$L(\mathbf{X} \mid \boldsymbol{\theta}) = L(X_1, \ldots, X_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}). \tag{5.1}$$

Обычно принято называть $L(\mathbf{X} \mid \boldsymbol{\theta})$ функцией правдоподобия, если она рассматривается только как функция $\boldsymbol{\theta}$ при $\mathbf{X}$ фиксированных и равных тем, которые зарегистрированы в эксперименте. Мы будем использовать ее в этой форме в гл. 7—9. Сейчас же рассмотрим ее как функции и $\mathbf{X}$ и $\boldsymbol{\theta}$.

5.1.2. Функция результатов наблюдения (статистика). Введем новую случайную переменную в виде $Y = \Phi(X_1, \ldots, X_n)$. Любая такая функция Φ называется *функцией результатов наблюдений* (или статистикой). Например, среднее $\overline{X}$ [см. формулу (2.33)] есть статистика. Заметим, что, с другой стороны, «статистика» — это и название части прикладной математики, изучающей только что определенные статистики.

§ 5.2. ИНФОРМАЦИЯ ФИШЕРА

5.2.1. Определение информации. Объем информации о параметре θ, содержащейся в наблюдении X, определяется следующим выражением (при условии, что оно существует):

$$I_X(\theta) = E\left[\left(\frac{\partial \ln L(X \mid \theta)}{\partial \theta}\right)^2\right] = \int_{\Omega_\theta} \left(\frac{\partial \ln L(X \mid \theta)}{\partial \theta}\right)^2 L(X \mid \theta)\, dX. \tag{5.2}$$

Если $\boldsymbol{\theta}$ — k-мерное, формула приобретает вид

$$\left[\underset{\sim}{I}_X(\boldsymbol{\theta})\right]_{ij} = E\left[\frac{\partial \ln L(X \mid \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln L(X \mid \boldsymbol{\theta})}{\partial \theta_j}\right] =$$

$$= \int_{\Omega_\theta} \left[\frac{\partial \ln L(X \mid \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln L(X \mid \boldsymbol{\theta})}{\partial \theta_j}\right] L(X \mid \boldsymbol{\theta})\, dX. \tag{5.3}$$

Таким образом, в общем случае $I_X(\boldsymbol{\theta})$ есть матрица размерности $k \times k$.

Имеются два довода в пользу такого, на первый взгляд, произвольного определения: во-первых, оно работает, т. е. оно обладает требуемыми от информации свойствами; и, во-вторых, оно очень тесно связано с максимально доступной точностью оценки величины $\boldsymbol{\theta}$ при заданных X.

5.2.2. Свойства информации. Для того чтобы сделать информацию полезным понятием, определение (5.2) должно быть дополнено следующими двумя условиями:

1. Ω_θ не зависит от θ.

2. $L(\mathbf{X} \mid \theta)$ обладает необходимыми свойствами, позволяющими коммутировать операторам $\partial^2/\partial\theta_i\partial\theta_j$ и $\int dX$.

При выполнении первого условия выполняется, как правило, и второе условие для всех распределений, встречающихся в физике. Однако первое условие не всегда имеет место. Например, пусть нужно определить массу и направление движения K^0-мезона по его распаду на $\pi^+\pi^-$. Очевидно, что поперечный импульс каждого π-мезона изменяется в интервале от 0 до p^* импульса каждого π-мезона в системе центра инерции, который зависит от массы K^0. Поэтому Ω_θ (от 0 до p^*) зависит от θ (массы K^0).

Предполагая, что первое и второе условия выполнены, просто показать, что информация (5.3) может быть записана в виде

$$[I_X(\theta)]_{ij} = -E\left[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln L(X \mid \theta)\right].$$

Свойство аддитивности немедленно следует из второго условия. Информация об одном наборе параметров, полученная в разных экспериментах, может быть сложена. В частности, информация, содержащаяся в одном событии, связана с информацией, содержащейся в N событиях, соотношением

$$I_N(\theta) = NI_1(\theta). \tag{5.4}$$

Это замечательное свойство может быть очень полезным вследствие соотношения между информацией и точностью (см. § 7.4). Здесь оно может быть проиллюстрировано на простом примере.

Пусть переменная X распределена по нормальному закону с *известной* дисперсией σ^2 и неизвестным средним μ. Тогда информация, содержащаяся в одном наблюдении X, равна

$$I_1(\mu) = -\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\mu^2}\left(-\frac{(X-\mu)^2}{2\sigma^2} - \ln\sigma\sqrt{2\pi}\right)\frac{1}{\sqrt{2\pi}\,\sigma}\times$$

$$\times\exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right]dX = +\int_{-\infty}^{\infty}\frac{1}{\sigma^2}\,\frac{1}{\sqrt{2\pi}\,\sigma}\times$$

$$\times\exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right]dX = 1/\sigma^2.$$

Соответственно для N наблюдений X_i:

$$I_N(\mu) = -\int_{-\infty}^{\infty} \cdots \int_{\infty}^{-\infty} \frac{\partial^2}{\partial \mu^2} \left[\sum_{i=1}^{N} \left(-\frac{(X_i-\mu)^2}{\sigma^2} \right) - \ln \sigma \sqrt{2\pi} \right] \times$$

$$\times \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2} \frac{(X_i-\mu)^2}{\sigma^2} \right] dX_i = \frac{N}{\sigma^2} = NI_1(\mu). \quad (5.5)$$

Таким образом, $I_N(\mu)$ равно обратному значению дисперсии $\overline{X}$, а само $\overline{X}$ является обычной оценкой среднего μ генеральной совокупности. Для такого случая позже будет показано, что соотношение (5.5) дает обратное значение минимальной дисперсии, которая может быть получена при *любом* способе оценки μ.

Рассмотрим другое важное свойство информации, о котором говорилось во введении этой главы, а именно применимость этого понятия к задачам обработки данных. С этой целью введем понятие *достаточных статистик*.

§ 5.3. ДОСТАТОЧНЫЕ СТАТИСТИКИ

5.3.1. Достаточность. Говорят, что статистика $T = \Phi(X)$ *достаточна* для $\boldsymbol{\theta}$, если условная функция плотности X при данном T $f(X \mid T)$ независима от θ. T и θ могут быть многомерными и иметь *различные* размерности.

Если T является достаточной статистикой для θ, то любая строго монотонная функция T также будет достаточной статистикой для θ. Позже мы будем использовать это свойство. Важность понятия достаточности для обработки данных следует из определения: в T содержится столько же информации о θ, сколько и в первоначальных данных X. Иными словами, любая другая функция тех же самых данных не содержит какую-либо дополнительную информацию о θ (см. § 5.4).

Очевидно, что множество измерений $T = X$ достаточно для θ, поскольку оно содержит всю первоначальную информацию. Набор достаточных статистик будет полезен только в случае, когда он приводит к *сокращению объема* экспериментальных данных.

Из самого определения следует способ проверки достаточности, но предпочтительнее (из практических соображений) использовать следующий результат.

Необходимое и достаточное условие того, что $T(X)$ является достаточной статистикой для θ, состоит в том, что функция правдоподобия может быть представлена в виде

$$L(\mathbf{X} \mid \theta) = g(T, \theta) h(\mathbf{X}), \quad (5.6)$$

где: 1) $h(\mathbf{X})$ не зависит от θ и 2) $g(T, \theta)$ пропорционально $A(T \mid \theta)$ *условной плотности вероятности* распределения T при заданном θ.

Во многих случаях условие 2 выполняется. Например, если множество значений $\mathbf{X}$ не зависит от θ, можно записать:

$$A(T \mid \theta) = \int\limits_{\substack{\Omega_\theta \\ T=\text{const}}} L(\mathbf{X} \mid \theta)\, d\mathbf{X} = \int\limits_{\substack{\Omega_\theta \\ T=\text{const}}} g(T, \theta)\, h(\mathbf{X})\, d\mathbf{X} =$$

$$= g(T, \theta) \int\limits_{\substack{\Omega_\theta \\ T=\text{const}}} h(\mathbf{X})\, d\mathbf{X},$$

т. е. в этом случае условие 2 вытекает из условия 1. В общем же случае необходимо проверять выполнимость обоих условий, прежде чем использовать $T(\mathbf{X})$ в качестве достаточной статистики.

5.3.2. Примеры: 1. Предположим, что $\mathbf{X}$ распределено по нормальному закону $N(\mu, \sigma^2)$. Тогда функция правдоподобия равна:

$$L(\mathbf{X} \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right] = (2\pi\sigma^2)^{-n/2} \times$$

$$\times \exp\left[-\frac{n}{2\sigma^2}(\overline{X}-\mu)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\right], \qquad (5.7)$$

где среднее $\overline{X}$ определено уравнением(2.33).

Если σ^2 известно, то $\overline{X}$ распределено по нормальному закону $N(\mu, \sigma^2/n)$, а функция правдоподобия (5.7) может быть представлена в виде соотношения (5.6) с

$$g(\overline{X} \mid \mu) = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{n}{2\sigma^2}(\overline{X}-\mu)^2\right].$$

Очевидно, что условия 1 и 2 предыдущего параграфа оба выполняются. Поэтому $\overline{X}$ является достаточной статистикой для μ.

Когда, с другой стороны, μ известно, а σ^2 неизвестно, легко показать, что статистика

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

достаточна для σ^2. Условная ф.п.в. для $\hat{\sigma}^2$ при заданном σ^2 равна:

$$g(\hat{\sigma}^2 \mid \sigma^2) = \frac{1}{\Gamma\left(\frac{n}{2}\right)}\left(\frac{n}{2\sigma^2}\right)^{n/2} (\hat{\sigma}^2)^{(n/2-1)} \exp\left(-\frac{n}{2\sigma^2}\hat{\sigma}^2\right).$$

2. Приведем пример, когда невозможно найти достаточную статистику. Пусть функция плотности:

$$f(X \mid \theta) = \frac{\theta \exp(\theta X)}{\exp(\theta^2) - 1},$$

где $0 \leqslant X \leqslant \theta$.

Здесь Ω_θ зависит от θ. Для заданных наблюдений $X_1, \ldots, X_n$ функцию правдоподобия можно записать в виде

$$L(\mathbf{X} \mid \theta) = \left(\frac{\theta}{\exp(\theta^2) - 1)}\right)^n \exp(\theta T) \quad \text{с} \quad T = \sum_{i=1}^{n} X_i.$$

Очевидно, что функция правдоподобия может быть представлена в виде произведения двух членов с $h(\mathbf{X}) = 1$. Это приводит к тому, что $g(T \mid \theta) = L(X, \theta)$, которая не является условной ф.п.в. переменной T при заданном θ. Поэтому T не может быть достаточной статистикой для θ. Это не исключает возможность того, что некоторая другая статистика будет достаточной.

5.3.3. Минимальные достаточные статистики. Очевидно, что первоначальный набор данных $(X_1, \ldots, X_n)$ есть набор достаточных статистик. Можно показать, что если θ имеет размерность m, то можно найти S достаточных статистик для множества n измерений, причем очевидно $m \leqslant n$, $S \leqslant n$, а само S может быть меньше, равно или больше m. Набор достаточных статистик, имеющий наименьшее S, называется *минимально достаточным*. Он является функцией всех других наборов достаточных статистик. Очевидно, что такой набор является наиболее полезным, поскольку он дает максимальное сокращение объема данных.

Сокращение объема данных, производимое S достаточными статистиками, может быть таким, что S пропорционально n, т. е. число достаточных статистик зависит от числа наблюдений. Однако наличие достаточных статистик становится особенно полезным, когда S фиксировано при фиксированных m и не *зависит* от n. Распределения, допускающие такой набор, определяются теоремой Дармуа (см. п. 5.3.4.). В частности, ф.п.в. нормального закона $N(\mu, \sigma^2)$ имеет:

а) либо одну достаточную статистику для одного параметра μ или σ^2, если другой параметр фиксирован (это мы уже видели в п. 5.3.3.),

б) либо пару совместно достаточных статистик для двух параметров μ, σ^2, когда они оба неизвестны.

Можно показать, что для любой ф.п.в., для которой справедливо а), справедливо и б).

Например, возьмем:

$$f(X \mid \theta) = 1/\theta; \quad \theta - \frac{1}{2} \leqslant X \leqslant \theta + \frac{1}{2}.$$

Тогда можно показать, что $[\min X_i, \max X_i]$ образует пару достаточных статистик для θ, которая минимальна. Отметим, что в этом примере $S > m$ и не существует *единственной* достаточной статистики для θ. Или другой пример:

$$\prod_{i=1}^{n} f(X_i \mid \theta) = \frac{1}{\theta (2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\left[\frac{n^2}{\sigma^2}(\overline{X} - \mu)^2 + \sum_{i=2}^{n} X_i^2\right]\right\}.$$

Для этого случая среднее $\overline{X}$ — единственная достаточная (минимальная, конечно) статистика как для μ, так и для θ. Здесь $S < m$, и вся задача состоит в отыскании таких функций $\overline{X}$, которые являются наилучшими оценками μ и θ [2].

5.3.4. Теорема Дармуа. Эта теорема показывает, что только очень ограниченный класс ф.п.в. допускает существование некоторого числа достаточных статистик, не зависящего от количества наблюдений. Результаты теоремы для одномерного случая формулируются так:

1. Если при любом Ω_θ существует некоторое $n > 1$, такое, что множество $(X_1, \ldots, X_n)$ допускает достаточную статистику для θ, тогда ф. п. в. имеет «экспоненциальную форму»:

$$f(X \mid \theta) = \exp[\alpha(X)a(\theta) + \beta(X) + c(\theta)]. \tag{5.8}$$

2. Напротив, множество $(X_1, \ldots, X_n)$ имеет достаточную статистику для всех $n > 1$ (но только при условии, что Ω_θ не зависит от θ), если $f(X \mid \theta)$ имеет экспоненциальную форму и преобразование

$$(X_1, X_2, \ldots, X_n) \Rightarrow (R, X_2, \ldots, X_n)$$

взаимно однозначно и непрерывно дифференцируемо для всех X. Статистика R достаточна для параметра θ, и поэтому он является некоторой монотонной функцией R.

Для многомерного случая «экспоненциальная» форма имеет вид

$$f(\mathbf{X} \mid \theta) = \exp\left[\sum_{j=1}^{S} \alpha_j(\mathbf{X})\, a_j(\theta) + \beta(\mathbf{X}) + c(\theta)\right] \tag{5.9}$$

и

$$R_j = \sum_{i=1}^{n} \alpha_j(X_i), \quad j = 1, \ldots, S$$

служит одним возможным набором S совместно достаточных статистик для θ.

Примеры. Если использовать обозначения соотношения (5.8), ф.п.в. нормального закона $N(\mu, \sigma^2)$ может быть записана:

$$\ln f = X\frac{\mu}{\sigma^2} - \frac{X^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \ln(\sigma\sqrt{2\pi}).$$

1. Если предположить, что σ^2 известно, а μ неизвестно, то условия теоремы Дармуа выполняются при $a = \mu$; $\alpha = X/\sigma^2$; $c = -\mu^2/2\sigma^2$; $\beta = -X^2/2\sigma^2 - \ln(\sigma\sqrt{2\pi})$. Достаточная статистика для μ дается соотношением

$$A = \sum_{i=1}^{n} \alpha_i = \frac{n}{\sigma^2}\overline{X}.$$

В свою очередь, $\overline{X}$ тоже является достаточной статистикой.

2. Если известно μ, а σ^2 неизвестно, полагаем: $a=1/\sigma^2$; $\alpha = X\,(\mu - X/2)$; $c = -\mu^2/2\sigma^2 - \ln(\sigma\sqrt{2\pi})$; $\beta = 0$. Достаточная статистика имеет вид:

$$A = \sum_{i=1}^{n} \alpha_i = \mu \sum_{i=1}^{n} X_i - \frac{1}{2} \sum_{i=1}^{n} X_i^2.$$

Тогда статистика

$$A' \equiv N\hat{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 = 2A - n\mu^2$$

также достаточна для σ^2.

3. Если и μ и σ^2 оба неизвестны, полагаем: $a_1 = \mu/\sigma^2$; $\alpha_1 = X$; $c = -\mu^2/2\sigma^2 - \ln(\sigma\sqrt{2\pi})$; $a_2 = 1/\sigma^2$; $\alpha_2 = -X^2/2$; $\beta = 0$. Совместно достаточные статистики для μ и σ^2 имеют вид:

$$A_1 = \sum_{i=1}^{n} X_i \text{ и } A_2 = -\sum_{i=1}^{n} X_i^2/2.$$

4. Если μ и σ^2 по-прежнему оба неизвестны, но нам нужен только один параметр, можно выбрать $a = \mu$; $\alpha = X/\sigma^2$; $c = -\mu^2/2\sigma^2$; $\beta = -X^2/2\sigma^2 - \ln(\sigma\sqrt{2\pi})$, так что $R = \sum_{i=1}^{n} X_i/\sigma^2$ является достаточной статистикой для μ. Тогда независимо от σ^2 $\overline{X}$ также будут достаточной статистикой.

С другой стороны, невозможно найти статистику для σ^2, не зависящую от μ, поскольку единственно возможными выражениями для a, α, c, β являются выражения, приведенные в примере 2.

§ 5.4. ИНФОРМАЦИЯ И ДОСТАТОЧНОСТЬ

Мы уже указывали, что достаточная для θ статистика T содержит всю информацию о θ, содержащуюся в наблюдениях $X_1, .., X_n$. Таким образом, если существует достаточная статистика, то можно произвести уменьшение объема данных без потери информации. Если ф.п.в. не имеет экспоненциальной формы (5.9), то невозможно найти достаточную статистику. Тогда любое *уменьшение объема* данных с помощью фиксированного числа статистик, не зависящего от количества наблюдений, должно приводить к некоторой потере информации. Однако заметим, что можно уменьшить количество данных от n до αn, где $\alpha < 1$ без потери информации. Такой случай возможен, когда существуют $1/\alpha$ наблюдений (например, компоненты импульса) каждого из αn событий и в каждом событии нас интересует только одна величина (например, абсолютная величина импульса). Если *у данного события* для требуемой величины существует достаточная статистика, то некоторое сокращение объема данных возможно без потери информации.

Из нескольких недостаточных статистик следует выбирать такую статистику T, которая содержит максимум информации $I_T(\theta)$. Понятно, что $I_T(\theta)$ меньше $I_X(\theta)$ полной информации о θ, содержа-

щейся в измерениях X:

$$I_T(\theta) \leqslant I_X(\theta). \tag{5.10}$$

Доказательство (5.10) очень просто. Представим $L(X \mid \theta)$ в виде

$$L(X \mid \theta) = g(t \mid \theta) h(X, \theta), \tag{5.11}$$

что всегда возможно, поскольку h также является функцией θ, где $g(t, \theta)$ — условное распределение t при заданном θ. Подстановка (5.11) в (5.2) приводит к

$$I_X(\theta) = E\left\{\left[\frac{\partial}{\partial\theta} \ln g(t \mid \theta)\right]^2\right\} + E\left\{\left[\frac{\partial}{\partial\theta} \ln h(X, \theta)\right]^2\right\} =$$
$$= I_T(\theta) + E\left\{\left[\frac{\partial}{\partial\theta} \ln h(X, \theta)\right]^2\right\}.$$

Последний член всегда положителен, поэтому соотношение (5.10) доказано. Знак равенства в (5.10) обычно имеет место только тогда, когда h не зависит от θ. Но тогда T — достаточная статистика в соответствии с определением (5.6).

Обобщением соотношения (5.10) на случай, когда имеется k параметров θ, является соотношение

$$\widetilde{U I_T}\ U \leqslant \widetilde{U I_X}\ U,$$

где $\underset{\sim}{I_T}$ и $\underset{\sim}{I_X}$ — $k \times k$ матрицы, а U — произвольный k-мерный вектор.

Этот результат очень важен. Он показывает, что если, например, диагонализировать $\underset{\sim}{I_T}$ с помощью некоторого унитарного преобразования, то каждая диагональная компонента $\underset{\sim}{I_T}$ меньше, чем соответствующий диагональный элемент в недиагональной матрице $\underset{\sim}{I_X}$. В общем случае, если найден набор статистик T, который содержит величину $\underset{\sim}{I_T}(\boldsymbol{\theta})$ информации о k параметрах $\boldsymbol{\theta}$, то они содержат такое же количество информации о k линейных комбинациях параметров.

§ 5.5. ПРИМЕР ПЛАНИРОВАНИЯ ЭКСПЕРИМЕНТА

Как мы уже показали, задавая данные, можно сравнить количества информации, содержащейся в различных статистиках. Однако можно поступать по-другому. Пусть необходимо измерить некоторые параметры и существует несколько возможных способов сделать это. Тогда можно спланировать эксперимент, т. е. выбрать способ, дающий максимум информации о параметрах.

Например, нам нужно взвесить на весах четыре предмета. Самый простой подход — взвесить каждый предмет раздельно, в итоге мы проделаем четыре измерения. (Если точность весов не достаточно высока, можно повторить измерения, пока не будет достигнута требуемая точность.) Однако возникает вопрос, можно ли улучшить точность четырех измерений, если взвешивать четыре возможные комбинации предметов, а не каждый в отдельности [28, 29]?

Выберем четыре новые переменные X_i, которые являются линейными комбинациями весов θ_j $(j = 1, \ldots, 4)$ четырех предметов. Предположим, что X_i следует нормальному закону $N(\sum_j a_{ij}\theta_j, \sigma^2)$, где a_{ij} — коэффициенты линей-

ных комбинаций, а σ — неточность одного взвешивания. Мы предполагаем, что σ всегда одинаково независимо от линейной комбинации. Возможные комбинации определяются значениями $a_{ij} = -1, 0, +1$ (т. е. в i-м взвешивании объект с номером j либо находится в левой чаше весов, либо не принимает участия в измерении, либо находится в правой чаше весов соответственно). Функция правдоподобия измерений равна

$$L(\mathbf{X} \mid \boldsymbol{\theta}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^4 \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{4}\left(X_i - \sum_{j=1}^{4} a_{ij}\,\theta_j\right)^2\right].$$

С ее помощью можно рассчитать матрицу информации:

$$-\frac{\partial^2 \ln L(\mathbf{X} \mid \boldsymbol{\theta})}{\partial\theta_j\,\partial\theta_k} = \sum_{i=1}^{4}\frac{a_{ij}\,a_{ik}}{\sigma^2} = E\left(-\frac{\partial^2 \ln L}{\partial\theta_j\,\partial\theta_k}\right) = -[\underset{\sim}{I}(\boldsymbol{\theta})]_{jk}.$$

Сконструируем матрицу a с коэффициентами a_{ij}, так что

$$\underset{\sim}{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\,\underset{\sim}{a}^T\,\underset{\sim}{a}.$$

Сейчас мы можем сравнить два метода. В первом методе все четыре предмета взвешиваются независимо, поэтому $a_{ij} = \delta_{ij}$. Очевидно, в этом случае мы получаем

$$[\underset{\sim}{I}_1(\boldsymbol{\theta})]_{ij} = \delta_{ij}/\sigma^2.$$

Во втором методе производятся тоже четыре измерения, но все четыре предмета участвуют с коэффициентами a_{ij}, равными -1 или $+1$. Кроме того, выбранные комбинации должны быть линейно независимыми (чтобы можно было решить четыре уравнения относительно четырех неизвестных θ_j). Один возможный набор коэффициентов таков:

$$a = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Отсюда следует, что матрица информации

$$[\underset{\sim}{I}_2(\boldsymbol{\theta})]_{ij} = 4\delta_{ij}/\sigma^2 = 4\,[\underset{\sim}{I}_1(\boldsymbol{\theta})]_{ij}.$$

Очевидно, второй метод предпочтительнее, поскольку он дает вчетверо большую информацию. Это дает соответственно вдвое большую точность определения θ_j.

ГЛАВА 6

ТЕОРИЯ РЕШЕНИЙ

Классическая статистика помогает физику-экспериментатору обрабатывать результаты своего эксперимента с минимальной потерей информации. Однако цель эксперимента может состоять и в том, чтобы решить вопрос о правильности какой-либо гипотезы или о правильном значении фундаментальной константы, или просто о том, что делать дальше. Физику, таким образом, нужна *теория решений*, тесно связанная с мощными методами статистики*. В этой главе мы кратко опишем такую теорию. Основной подход теории решений может быть лучше всего проиллюстрирован простым примером.

Предположим, что работает детектор частиц, параметры которого в течение дня могут измениться по сравнению с оптимальными. Максимальное количество событий, регистрируемых этим детектором в течение дня, $\theta_{\text{макс}}$. Каждое утро необходимо решать вопрос, следует ли заново настраивать детектор, если результат измерения счетчика за предыдущий день равен t, а для настройки потребуется p-я часть дня?

Для того чтобы сделать рациональный выбор между двумя возможными решениями, необходимо ввести меру потерь, связанную с выбором каждого решения. Такая мера задается *функцией потерь* $L(\theta, d)$, определяющей величину потери при выборе решения d, если истинное число зарегистрированных событий θ. Такой функцией могла бы быть функция

$$L(\theta, d_1) = \theta_{\text{макс}} - \theta;$$

$$L(\theta, d_2) = p\theta_{\text{макс}}.$$

Здесь d_1 — решение «не настраивать» и d_2 — решение «настраивать».

---

* Теория решений вступает в действие в том случае, когда нельзя ограничиться даваемыми теорией статистических оценок предсказаниями вероятностей осуществления различных гипотез и требуется сделать окончательный выбор решения, учитывая реальные потери от возможной ошибки. Неодинаковость потерь для различных гипотез является причиной изменения выводов теории решений по сравнению с выводами теории оценок. Выбор менее вероятной гипотезы может оказаться более предпочтительным, если потери в случае ошибочности такого выбора окажутся меньше потерь, связанных с ошибочностью выбора более вероятной конкурирующей гипотезы. Типичный пример такой задачи представляет отбраковка изделий по параметру, нарушения в котором приводят к дорогостоящим авариям. В этом случае оказывается более выгодным нести потери на изготовлении изделий, попавших в брак вследствие переопределенных критериев отбраковки, чем допустить даже с малой вероятностью прием аварийноопасного изделия. — *Прим. ред.*

Если θ известно, то выбор решения очевиден. Детектор необходимо настраивать, если $\theta \leqslant (1-p)\theta_{\text{макс}}$. Однако θ неизвестно, известно только наблюдение величины, распределение которой зависит от значения θ. Назначение теории решений состоит в разработке техники выбора наилучшего решения на основе наблюдений. Основная цель этой техники — минимизировать потери, связанные с принятием неверного решения.

§ 6.1. ОСНОВНЫЕ ПОНЯТИЯ В ТЕОРИИ РЕШЕНИЙ

6.1.1. Субъективная вероятность, бейесовский подход*. В пп. 2.2.5 и 2.3.5 мы видели, как новые наблюдения модифицируют предварительное знание о некотором неизвестном параметре, если использовать теорему Бейеса. Для практического применения указанной там методики необходимо иметь количественное выражение предварительного знания в форме распределения вероятностей.

Предположим, мы хотим определить значение θ параметра распределения $f(X, \theta)$ наблюдаемой случайной переменной X. В рамках бейесовского подхода предполагается, что параметр θ имеет некоторое априорное распределение, которое отражает *степень веры* наблюдателя в различные возможные значения θ.

Сейчас мы покажем, как такую степень веры можно преобразовать в числа. Рассмотрим случай, когда параметр может иметь только два возможных значения θ_1 и θ_2. Тогда наблюдатель задает себе такой вопрос: «учитывая имеющуюся информацию, каковы минимальные шансы того, что θ равно θ_1, а не θ_2?»

Если ответ таков: «минимальные шансы равны $b = 4$», то это означает, что вера наблюдателя в значение θ_1 в четыре раза больше, чем в θ_2. Обозначим степень веры в θ_1 и θ_2 через $\pi(\theta_1)$ и $\pi(\theta_2)$. Поскольку известно только отношение $\pi(\theta_1)$ и $\pi(\theta_2)$

$$b = \pi(\theta_1)/\pi(\theta_2),$$

мы можем нормировать их на единицу:

$$\sum_{i=1}^{2} \pi(\theta_i) = 1,$$

---

* Собственно говоря, теорема Бейеса относится к решению весьма важной, но частной задачи математической статистики о вычислении функции распределения, учитывающей как проведенные измерения, так и предварительно имевшиеся сведения об определяемых величинах. Если отсутствуют сведения об априорной вероятности, то нет и самой задачи совместного учета предварительной информации и статистических данных, полученных в эксперименте.

В этом разделе авторами излагается точка зрения сторонников создания на основе теоремы Бейеса общего подхода в теории решений. По этой причине и введено понятие субъективной вероятности, выражающей степень веры в различные значения определяемой величины. В этом случае бейесовский подход формально становится общим. Но при отсутствии предварительных данных об определяемой величине такой подход дает распределение, отличающееся от полученных в измерениях статистических данных только искажениями, вносимыми учетом субъективной веры наблюдателя в те или иные значения определяемой величины. Некоторые критические замечания авторов о неопределенности такого обобщения бейесовского подхода приведены в конце параграфа. — *Прим. ред.*

поэтому

$$\pi(\theta_1) = b/(b+1); \quad \pi(\theta_2) = 1/(b+1).$$

Этот пример может быть обобщен на случай, когда необходимо получить распределение степеней веры при изменении θ в некотором непрерывном интервале значений.

Можно показать [29], что такие степени веры удовлетворяют аксиомам, справедливым для понятия вероятности, и с распределением $\pi(\theta)$ можно поэтому работать, как с распределением вероятности. Таким образом, распределение *субъективной вероятности* $\pi(\theta)$ обобщает знание наблюдателя о параметре θ.

Используя априорное распределение $\pi(\theta)$, наблюдения $\mathbf{X}$ с вероятностью $p(\mathbf{X} \mid \theta)$ и теорему Бейеса (2.20), получаем апостериорное распределение

$$p(\theta \mid \mathbf{X}) \sim p(\mathbf{X} \mid \theta)\pi(\theta).$$

Этот результат можно рассматривать как численное выражение знания о параметре θ при заданных наблюдениях $\mathbf{X}$.

В теории решений разработана методика для расчета оптимального решения при заданных $\mathbf{X}$. Апостериорное распределение является связующим звеном между наблюдениями $\mathbf{X}$ и параметром θ, относительно которого принимается решение.

6.1.2. Определения и терминология. Теория решений использует понятия трех различных пространств:

пространство наблюдений χ, в котором заключены все возможные наблюдения

$$\mathbf{X} = (X_1, \ldots, X_n);$$

пространство параметров Ω, которое содержит все возможные значения параметра θ или параметров $\Theta = (\theta_1, \ldots, \theta_p)$. Возможные значения θ часто называют *состояниями природы*;

пространство решений $\mathcal{D}$, которое содержит все возможные решения d.

Правило решения δ (или иногда *процедура решения*, *решающая функция*) указывает, какое решение d надо принять, если получены наблюдения $\mathbf{X}$, т. е.

$$d = \delta(\mathbf{X}).$$

Мы ограничимся рассмотрением *неслучайных правил решения* δ, которые полностью определяют d при заданных $\mathbf{X}$.

Для выбора между правилами решений необходима *функция потерь* $L(\theta, d)$, значение которой равно потере, связанной с выбором решения d, а θ предполагается в качестве истинного значения параметра.

Введение функции потерь составляет существенную часть теорий решений*. Любой рациональный выбор между решениями должен быть основан на расчете потери (например, *стоимости*) или выигрыша, связанного с принятием каждого решения. Рассмотрим функцию $L[\theta, \delta(\mathbf{X})]$, являющуюся случайной переменной. Она зависит от переменной $\mathbf{X}$. Функция риска для правила решения δ определяется как

$$R_\delta(\theta) \equiv E\{L[\theta, \delta(\mathbf{X})]\} = \int L[\theta, \delta(\mathbf{X})] f(\mathbf{X} \mid \theta)\, dX. \tag{6.1}$$

Таким образом, $R_\delta(\theta)$ служит средней потерей по всем возможным наблюдениям.

С точки зрения бейесовцев, параметр θ можно рассматривать как случайную переменную. Ожидаемый риск, усредненный по всем θ,

$$r_\pi(\delta) = \int R_\delta(\theta)\, \pi(\theta)\, d\theta \tag{6.2}$$

называется *апостериорным риском* использования правила решения δ при заданной априорной плотности $\pi(\theta)$. Используя (6.1), апостериорный риск (6.2) можно записать:

$$r_\pi(\delta) = E_\theta\{E_{\mathbf{X}}\{L[\theta, \delta(\mathbf{X})] \mid \theta\}\} = E_{\mathbf{X}}\{E_\theta\{L[\theta, \delta(\mathbf{X})] \mid \mathbf{X}\}\}.$$

Здесь индекс у оператора ожидания обозначает переменную, по которой производится усреднение. Величина

$$E_\theta\{L[\theta, \delta(\mathbf{X})] \mid \mathbf{X}\} \tag{6.3}$$

называется *апостериорной потерей* при заданных наблюдениях $\mathbf{X}$. Она соответствует среднему значению потерь, связанному с принятием решения $\delta(\mathbf{X})$. Усреднение производится по апостериорной плотности $p(\theta \mid \mathbf{X})$.

§ 6.2. ВЫБОР ПРАВИЛ РЕШЕНИЯ

6.2.1. Классический подход: правила предварительного упорядочения. Классический подход к выбору правила решения δ основан на функции риска $R_\delta(\theta)$ [см. соотношение (6.1)]. Наилучшим правилом решения является решение, которое дает наименьший риск. Таким образом, если δ и δ' — два возможных правила и если

$$R_\delta(\theta) < R_{\delta'}(\theta) \text{ для всех } \theta, \tag{6.4}$$

---

* Только учет потерь, связанных с выбором ошибочного решения, приводит к отличию теорий решений от теории статистических оценок. Однако определение самой функции потерь выходит за рамки собственно задач математической статистики. — *Прим. ред.*

тогда δ — лучшее правило решения по сравнению с δ'. Чтобы показать, что δ предпочтительнее, чем δ', мы будем использовать обозначение

$$\delta > \delta'.$$

Рассмотрим случай нескольких правил решения δ_i с функциями риска, изображенными на рис. 6.1. Поскольку $R_{\delta_1}(\theta) < R_{\delta_2}(\theta)$ для всех значений θ, то можно сказать, что правило решений δ_1 всегда лучше, чем правило решений δ_2. Однако правило решений δ_1 не лучше, чем все оставшиеся правила. Для некоторых значений θ лучшими являются правила решений δ_3 и δ_4.

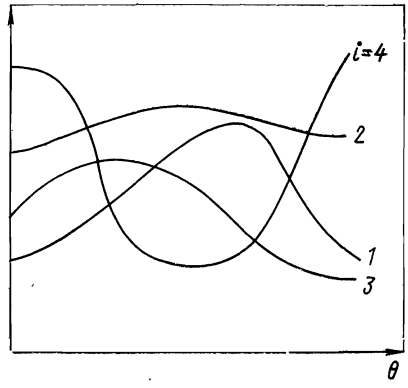


Рис. 6.1. Пример четырех функций риска R_{δ_i}

Таким образом, можно предварительно упорядочивать правила решения. В общем невозможно найти единственное наилучшее правило для всех θ, но для каждого значения θ будет оптимальное правило. Основная неопределенность в выборе связана с тем, что истинное значение θ неизвестно.

Правило решения δ' называется *допустимым*, если не существует правила δ такого, что неравенство (6.4) справедливо для всех значений θ, т. е. для некоторых значений θ справедливо неравенство $\delta' > \delta$.

Очевидно, что остается проблема выбора решения из класса допустимых решений. Таким образом, задача теории решений состоит, во-первых, в отыскании допустимых решений и, во-вторых, в выборе среди них одного решения. Полезным для этой цели является семейство бейесовских решений, обсуждаемых в следующем параграфе.

6.2.2. Бейесовский подход*. Бейесовский выбор правила решения δ для априорной плотности $\pi(\theta)$ основан на функции апостериорного риска $r_\pi(\delta)$ [см. формулу (6.2)]. Для бейесовцев наилучшим правилом решения δ, соответствующим априорной плотности $\pi(\theta)$, является то решение, которое дает наименьший апостериорный риск, т. е.

$$r_\pi(\delta) \leqslant r_\pi(\delta') \text{ для любого } \delta'. \qquad (6.5)$$

Таким образом, бейесовский подход состоит в том, что он переносит неопределенность в знании истинного значения θ в априорное распределение степеней веры $\pi(\theta)$. Основную неопределенность

* Этот подход применим в тех случаях, когда известна априорная вероятность. Он обеспечивает выбор наиболее предпочтительного правила, так как дополнительно учитывает сведения о θ, имевшиеся до измерения. — *Прим. ред.*

классического подхода можно обойти, если усреднить риск по апостериорной плотности.

Если существует решение δ, которое дает минимальную апостериорную потерю (6.3) для всех $\mathbf{X}$, то оно, очевидно, дает тоже минимальный апостериорный риск $r_\pi(\delta)$ и поэтому является бейесовским решением. Подобно этому, если существует решение, дающее минимальный риск $R_\delta(\theta)$ для всех θ, то оно является бейесовским решением.

При довольно общих предположениях (практически только предположение об ограниченности R_δ для всех δ) можно показать, что все *допустимые решения являются бейесовскими решениями*, т.е. если δ служит допустимым правилом решения, то тогда существует некоторая априорная плотность $\pi(\theta)$, такая, что δ служит бейесовским решением для π [12,30].

Напротив, если задано правило решения, то существует бейесовское правило, которое эквивалентно или является более предпочтительным. Класс бейесовских правил называется *полным* классом.

Если δ_B есть бейесовское правило, то не существует правила, которое лучше, чем δ_B для всех θ.

В то же время δ_B не всегда может быть допустимым, например, тогда, когда соответствующее априорное распределение $\pi_B(\theta)$ равно нулю для некоторых θ.

В частном случае, когда Ω есть конечное пространство дискретных значений $\theta_1, \ldots, \theta_r$, класс бейесовских решений для априорного распределения $\pi(\theta_i)$, $i = 1, 2, \ldots, r$ совпадает с классом допустимых решений при условии, что $\pi(\theta_i) \neq 0$, $i = 1, \ldots, r$.

6.2.3. Минимаксные решения. Среди различных способов, которые были предложены для выбора правил рещения, наиболее важным является метод *минимакса* Неймана. В соответствии с этим методом нужно выбирать такое правило решения, которое делает минимальным максимальный риск.

По определению правило минимакса, если оно существует, является допустимым. Можно показать, что при выполнении общих условий предыдущего параграфа оно существует. Отсюда следует, что минимаксное правило решения, являясь по существу понятием классической школы, должно быть бейесовским решением, соответствующим некоторому определенному априорному распределению $\pi_0(\theta)$.

Можно показать, что $\pi_0(\theta)$ — наиболее неблагоприятное априорное распределение, которое может быть выбрано в том смысле, что

$$\min [r_{\pi_0}(\delta)] \text{ для всех } \delta \geqslant \min [r_\pi(\delta)] \text{ для всех } \delta$$

для любого априорного распределения $\pi(\theta)$. Другими словами, бейесовское правило решения для π_0 имеет более высокий апостериорный риск, чем бейесовское правило для любого другого априорного распределения.

Поэтому минимаксное правило приводит к наиболее пессимистическому или консервативному решению.

§ 6.3. ОБЫЧНЫЕ ПРОБЛЕМЫ ВЫБОРА РЕШЕНИЙ В СТАТИСТИКЕ

Проблемы оценок и проверок являются основными в статистике, и все остальные главы посвящены им. В этом же разделе мы только покажем, как теория решений используется в задачах выбора решений при оценке значения параметра, оценке интервала значений и проверке гипотез.

6.3.1. Оценка значения параметра. Проблему решения при оценке значения параметра можно сформулировать так: какое значение $\hat{\theta}$ нужно выбрать для параметра θ, если имеются n наблюдений X_1, X_2, ..., X_n, распределенных в соответствии с плотностью $f(X, \theta)$. В этом случае пространство решений D находится во взаимно-однозначном соответствии с пространством параметров Ω: для каждого возможного значения θ существует решение d о том, чтобы приписать это значение параметру θ.

Пусть $\delta(\mathbf{X}) = \hat{\theta}$ — решающая функция, которая преобразует пространство наблюдений $(\chi)^n$ в $D = \Omega$.

Потеря соответственно будет функцией $L(\theta - \hat{\theta})$ расстояния между оценкой $\hat{\theta}$ и истинным значением θ. Функция потерь выбирается таким образом, чтобы она была минимальной при $\theta = \hat{\theta}$.

Поскольку регулярная функция может быть в окрестности минимума приближенно представлена квадратичной формой, то часто выбирают*

$$L(\theta - \hat{\theta}) = (\theta - \hat{\theta})^T W (\theta - \hat{\theta}) = \omega (\theta - \hat{\theta})^2 \qquad (6.6)$$

для рассматриваемого здесь одномерного случая. При такой функции потерь апостериорную потерю можно записать в виде

$$E[L(\theta - \hat{\theta}) \mid \mathbf{X}] = E[\omega (\theta - \hat{\theta})^2 \mid \mathbf{X}] = E\{\omega [\theta - E(\theta \mid \mathbf{X})]^2 \mid \mathbf{X}\} +$$
$$+ \omega [\hat{\theta} - E(\theta \mid \mathbf{X})]^2 = \omega \{D(\theta \mid \mathbf{X}) + [\hat{\theta} - E(\theta \mid \mathbf{X})]^2\}.$$

Очевидно, апостериорная потеря минимальна, когда правило решения таково, что $\hat{\theta} = E(\theta \mid \mathbf{X})$.

Для функции потерь в виде (6.6) оценка значения параметра для бейесовского подхода равна значению параметра θ, усредненному по апостериорной плотности распределения**. Обычно для различных функций потерь, соответствующих тому или иному случаю (например, соображение типа «лучше занизить точность, чем завысить»), необходимо отыскивать свое правило решения.

---

* Заметим, что если область значений θ не ограничена, то такая функция потерь тоже не ограничена и необходимо проявлять осторожность при использовании результатов § 6.2.

** Которая учитывает также известную априорную плотность распределения величины θ. — *Прим. ред.*

6.3.2. Оценка значений интервалом. Зачастую оценка только одного значения параметра $\hat{\theta}$ недостаточна, если она не связана с некоторым интервалом (как мы увидим в гл. 9). Всем хорошо известна запись вида $\hat{\theta} \pm \Delta\hat{\theta}$, где $\Delta\hat{\theta}$ соответствует стандартному отклонению оценки $\hat{\theta}$.

Так как апостериорная плотность $p(\theta \mid X)$ суммирует знание наблюдателя о параметре θ, то проблема решения состоит в выборе интервала (a, b) значений θ, который наилучшим образом описывает $p(\theta \mid X)$. В этом случае функция потерь является функцией $L(\theta; a, b)$ параметра θ и интервала. Возможной функцией потерь может быть

$$L(\theta; a, b) = \omega_1 (b - a)^2, \text{ если } \theta \in (a, b);$$
$$L(\theta; a, b) = \omega_2 (\theta - a)^2, \text{ если } \theta < a;$$
$$L(\theta; a, b) = \omega_3 (\theta - b), \text{ если } \theta > b.$$

Апостериорная потеря равна

$$E[L(\theta; a, b) \mid X] = \omega_1 \int_a^b (b-a)^2 p(\theta \mid X)\, d\theta +$$
$$+ \omega_2 \int_{-\infty}^{a} (\theta - a)^2 p(\theta \mid X)\, d\theta + \omega_3 \int_b^{\infty} (\theta - b)^2 p(\theta \mid X)\, d\theta.$$

Выбор решения соответствует выбору интервала (a, b), минимизирующего эту апостериорную потерю.

6.3.3. Проверка гипотез. Предположим, параметр θ принимает только два значения: 0 и 1. Соответственно существуют две гипотезы H_0 и H_1; гипотеза H_0 состоит в том, что θ равно нулю, H_1 — единице. Для выбора решения обычно определяют функцию потерь, которая равна 0 (нет потерь) для правильного решения. Такая функция потерь задается значением в табл. 6.1.

Таблица 6.1

Функция потерь

В действительности	Решение	
	Выбрать $\theta=0$	Выбрать $\theta=1$
H_0 верна	0	l_0
H_1 верна	l_1	0

В рамках бейесовского подхода каждой из гипотез H_0 и H_1 необходимо приписать априорные вероятности μ и $(1 - \mu)$ соответственно. Тогда, как это показано в табл. 6.2, для заданного правила решения δ существует конечная вероятность $\alpha(\delta)$ выбора H_1, когда верна H_0, и $\beta(\delta)$ выбора H_0, когда верна H_1.

Таблица 6.2

Вероятности решений $P(H_i \mid \theta)$

В действительности	Решение	
	Выбрать H_0	Выбрать H_1
H_0 верна	$1-\alpha(\delta)$	$\alpha(\delta)$
H_1 верна	$\beta(\delta)$	$1-\beta(\delta)$

В соответствии с бейесовским подходом необходимо выбрать такое правило решения δ, которое минимизирует

$$r_\mu(\delta) = \alpha(\delta) l_0 \mu + \beta(\delta) l_1 (1 - \mu).$$

Рассматриваемый случай можно отобразить графически, если нарисовать область допустимых точек $[\alpha(\delta), \beta(\delta)]$. Можно показать, что допустимая область выпукла и в общем случае имеет форму, показанную на рис. 6.2. Бейесовские правила решения соответствуют точкам на нижней границе этой области. Для $0 < \mu < 1$ семейство бейесовских решений идентично множеству допустимых правил решений. Очевидно, что минимальный риск будет получен для решения, соответствующего точке B, в которой прямая

$$r = l_0 \mu \alpha + l_1 (1 - \mu) \beta \quad (6.7)$$

касательна к области возможных точек. Это решение получается, если рассмотреть апостериорную потерю при заданных наблюдениях $\mathbf{X}$, т. е. ожидаемая потеря при выборе H_0 равна

$$l_1 P(H_1 \mid \mathbf{X}) = l_1 (1 - \mu) P(\mathbf{X} \mid H_1),$$

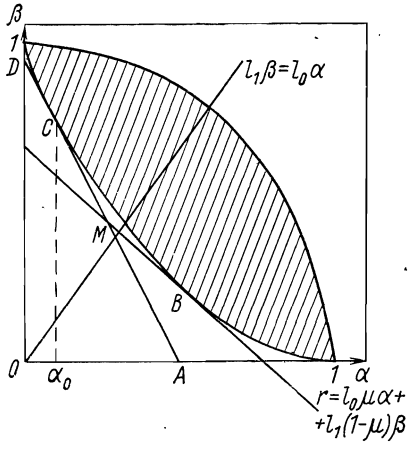


Рис. 6.2. Бейесовские правила решений

где $P(H_1 \mid \mathbf{X})$ — апостериорная вероятность того, что H_1 верна (т. е. $\theta = 1$).

С другой стороны, ожидаемая потеря при выборе H_1 равна

$$l_0 \mu P(\mathbf{X} \mid H_0).$$

Таким образом, правило решения с минимальным риском (точка B на рис. 6.2) таково: выбирается H_0, если

$$l_1 (1 - \mu) P(\mathbf{X} \mid H_1) < l_0 \mu P(\mathbf{X}_1 \mid H_0)$$

или

$$\frac{P(\mathbf{X} \mid H_1)}{P(\mathbf{X} \mid H_0)} < \frac{l_0 \mu}{l_1 (1-\mu)},$$

в противном случае выбирается H_1.

Сторонник классической школы не стал бы говорить о потерях l_0 и l_1, а выбрал бы произвольно определенный «уровень значимости»* α_0. Затем он выбрал бы такое решение, которое дает минимальное значение β. Этому решению соответствует точка C. Кажется очевидным, и это может быть показано строго [12], что такая процедура соответствует бейесовскому решению с определенным выбором отношения

$$l_0 \mu / l_1 (1 - \mu),$$

а именно равным отношению расстояний OD/OA на рис. 6.2. Далее, минимаксное правило решения соответствует точке M, точке пересечения с линией

$$l_1 \beta = l_0 \alpha.$$

Как подсказывает нам интуиция, малое значение α означает, что выбор H_1 происходит только в случае, когда наблюдения практически несовместимы с H_0. Такой способ рассуждения, не имеющий смысла при классическом подходе, подтверждается в бейесовской интерпретации тем фактом, что абсолютная величина наклона (6.7) в точке B есть уменьшающаяся функция α; это следует из выпуклости области. Тогда чем меньше α, тем больше должна быть априорная вероятность для H_0 (или потеря l_0), чтобы считать допустимыми маловероятные события и не принимать решения выбрать H_1 в качестве правильной гипотезы. Поэтому выбор малого значения α эквивалентен тому, что вы заранее отдаете предпочтение гипотезе H_0.

Малое значение доверительного уровня α можно интерпретировать как предположение о том, что H_0 с большей вероятностью верно и что наблюдатель не совершает ошибку при отбрасывании правильной теории более чем в $\alpha\%$ случаях своей профессиональной деятельности. С другой стороны, когда наблюдатель решает зафиксировать α, а не β, то это значит, что он считает маловероятным совершить ошибку, принимая H_0 за истинную, когда в действительности верна H_1 (и, может быть, менее опасным для своей профессиональной карьеры).

Очевидно, что это очень специальный способ формулирования проблемы решений и, конечно, он не универсален во всех случаях.

§ 6.4. ПРИМЕРЫ: НАСТРОЙКА АППАРАТУРЫ

Обратимся к менее классическим примерам, имея в виду, как теория решений может быть использована в «реальной жизни».

* Подробнее об уровнях значимости см. гл. 10, 11.

6.4.1. Настройка, вытекающая из оценки состояния аппаратуры. Вернемся к примеру, данному во введении к этой главе, когда было две возможности: d_1 не надо настраивать; d_2 надо настраивать; θ — среднее число событий, которое могла бы зарегистрировать аппаратура в течение дня. Потери, если бы в действительности средним было θ, были бы равны соответственно:

$$L(\theta, d_1) = (\theta_{\text{макс}} - \theta);$$
$$L(\theta, d_2) = \theta_{\text{макс}} p, \quad 0 < p < 1,$$

где $\theta_{\text{макс}}$ — максимальная скорость счета.

Предположим, что в течение предыдущего дня было зарегистрировано t событий. Тогда это число может быть использовано для оценки величины θ, характеризующей состояние аппаратуры.

Пусть распределение t относительно θ имеет вид

$$f(t, \theta) = \frac{\theta^t}{t!} \exp(-\theta)$$

(распределение Пуассона).

Предположим также, что априорное распределение θ выбрано в таком виде:

$$\pi(\theta) = \begin{cases} \dfrac{1}{\theta_{\text{макс}}}, & 0 \leqslant \theta \leqslant \theta_{\text{макс}}; \\ 0, & \theta < 0 \text{ или } \theta > \theta_{\text{макс}}, \end{cases}$$

т. е. θ распределено равномерно между нулем и $\theta_{\text{макс}}$. Такое распределение может быть выбрано, если из предыдущего опыта известно, что θ не концентрируется вокруг некоторой точки.

Тогда апостериорное распределение θ при заданном t будет таким:

$$P(\theta \mid t)\, d\theta \sim \frac{1}{\theta_{\text{макс}}} \frac{\theta^t}{t!} \exp(-\theta)\, d\theta.$$

Произведя замену переменных $u = 2\theta$, получим

$$P(u = 2\theta \mid t)\, du \sim \left(\frac{u}{2}\right)^t \exp\left(-\frac{u}{2}\right) du, \quad 0 < u < 2\theta_{\text{макс}}.$$

Это χ^2-распределение с $n = 2(1 + t)$ степенями свободы.

Если выбрано решение d_1, то апостериорная потеря равна

$$E_\theta\{L(d_1) \mid t\} = \int_0^{\theta_{\text{макс}}} (\theta_{\text{макс}} - \theta) P(\theta \mid t)\, d\theta = \theta_{\text{макс}} \int_0^{2\theta_{\text{макс}}} P(u \mid t)\, du -$$
$$- \frac{1}{2} \int_0^{2\theta_{\text{макс}}} u P(u \mid t)\, du. \qquad (6.8)$$

Здесь был использован тот факт, что $P(u)du = P(\theta)d\theta$. После некоторых преобразований соотношение (6.8) может быть записано:

$$E_\theta\{L(d_1) \mid t\} = \theta_{\text{макс}} P\{\chi^2(2+2t) < 2\theta_{\text{макс}}\} - (1+t) P\{\chi^2(3+2t) < 2\theta_{\text{макс}}\}.$$

Соответственно для решения d_2 апостериорная потеря равна

$$E_\theta\{L(d_2) \mid t\} = P\theta_{\text{макс}}.$$

Следовательно, бейесовское правило решения будет таким: выбрать d_1, если

$$E_\theta \{L(d_1) \mid t\} \leqslant E_\theta \{L(d_2) \mid t\},$$

в противном случае выбрать d_2.

Если $\theta_{\text{макс}}$ велико, полученные выражения упрощаются и решение о настройке аппаратуры принимается тогда, когда

$$t < (1-p)\,\theta_{\text{макс}} - 1.$$

6.4.2. Настройка, вытекающая из оценки оптимальной настройки. Предположим, что измеряется случайная переменная X, распределенная по нормальному закону $N(\theta, \tau^2)$, где θ неизвестно, а τ^2 известно. Параметр θ характеризует отклонение от некоторой реперной точки мишени, которой соответствует значение $X = 0$. Установка позволяет определенным образом смещать положение мишени. Решения таковы: $d(y)$ — произвести настройку с тем, чтобы уменьшить смещение на величину y; $d_2(y)$ — соответствующую настройку не производить.

В качестве функции потерь может быть выбрана любая из следующих двух функций:

$$1)\ L(d_y, \theta) = a(y-\theta)^2, \quad a > 0;$$

$$2)\ L(d_y, \theta) = \begin{cases} a(y-\theta)^2 + b, & y \neq 0; \\ a\theta^2, & y = 0. \end{cases}$$

В случае 2 как бы дается премия за решение не делать настройку. Предполагается, что априорная плотность распределения имеет вид $N(\mu, \sigma^2)$.

Тогда апостериорная плотность распределения θ пропорциональна

$$P(X \mid \theta)\, P(\theta) = \exp\left\{ -\frac{1}{2}\left[\frac{1}{\tau^2} + \frac{1}{\sigma^2}\right]\left[\theta - \frac{\dfrac{X}{\tau^2} + \dfrac{\mu}{\sigma^2}}{\dfrac{1}{\tau^2} + \dfrac{1}{\sigma^2}}\right]^2 \right\}.$$

Поэтому $P(\theta \mid X)$ равно

$$N\left(\frac{\dfrac{X}{\tau^2} + \dfrac{\mu}{\sigma^2}}{\dfrac{1}{\tau^2} + \dfrac{1}{\sigma^2}},\ \frac{1}{\dfrac{1}{\tau^2} + \dfrac{1}{\sigma^2}} \right).$$

Если $\sigma^2 \gg \tau^2$, апостериорная плотность имеет вид $N(X, \tau^2)$. Для функции потерь в случае 1 апостериорная потеря равна

$$E[(\theta - Y)^2 \mid X] = a[D[(\theta - Y) \mid X] + \{E[(\theta - Y) \mid X]\}^2] =$$
$$= a\{D(\theta - X) + [E(\theta \mid X) - Y]^2\}.$$

Таким образом, чтобы получить минимальную апостериорную потерю, нужно выбрать решение с

$$Y = E(\theta \mid X) = \frac{X/\tau^2 + \mu/\sigma^2}{1/\tau^2 + 1/\sigma^2}$$

с апостериорной потерей

$$a\left[\frac{1}{1/\tau^2 + 1/\sigma^2}\right].$$

В случае 2 можно показать подобным образом, что оптимально произвести настройку, если и только если

$$b < a\left(\frac{X/\tau^2+\mu/\sigma^2}{1/\tau^2+1/\sigma^2}\right),$$

и если настройка необходима, тогда нужно произвести настройку с

$$Y = E(\theta \mid X) = \frac{X/\tau^2+\mu/\sigma^2}{1/\tau^2+1/\sigma^2}.$$

§ 6.5. ЗАКЛЮЧЕНИЕ: НЕОПРЕДЕЛЕННОСТЬ В КЛАССИЧЕСКИХ И БЕЙЕСОВСКИХ РЕШЕНИЯХ

При классическом подходе необходимо выбирать среди различных возможных правил решения, каждое из которых оптимально для некоторого неизвестного значения параметров. При бейесовском подходе неопределенность переносится в субъективное априорное распределение по некоторой области возможных значений параметров, и после его выбора получается одно правило решения.

В классическом случае фундаментальная неопределенность лежит во многих возможных правилах решения, из которых должен быть сделан выбор. Для одного человека, являющегося сторонником бейесовского подхода, существует только одно правило решения, но в этом случае неопределенность лежит в возможных априорных распределениях, которые могут быть выбраны различными людьми по-разному. В асимптотике, когда число наблюдений становится большим, оба подхода приводят к совместимым результатам*.

Фактически это означает, что любое оптимальное классическое правило решения является также некоторым бейесовским правилом. Другими словами, даже если человек, принимающий решение, не является бейесовцем, то он поступает так, как если бы он был им!

Остается проблема философского плана, именно, как к этой основной неопределенности должен относиться ученый, который хочет быть объективным. Экспериментатор обычно чаще сталкивается с проблемой описания своих результатов таким образом, чтобы сохранить максимум информации о неизвестном параметре, чем с проблемой решения о величине параметра. Идя навстречу нуждам экспериментаторов, в гл. 7 и 8, посвященных проблеме оценок неизвестных параметров, мы будем использовать теоретико-информационный подход, а не методы теории решений. Следовательно, экспериментатор может не принимать решений, предоставляя это читателю, будь он физиком-теоретиком, специалистом по компиляции данных или же экспериментатором, желающим выяснить вопрос о постановке нового эксперимента.

---

* Следует отметить, что объясняется это, вопреки мнению авторов, вовсе не общностью бейесовского подхода, а уменьшением роли произвольно выбираемой априорной вероятности. — *Прим. ред.*

ГЛАВА 7

ТЕОРИЯ ОЦЕНОК

Поиск оценки может быть рассмотрен как измерение параметра (предполагается, что он имеет некоторое фиксированное, но неизвестное значение), основанное на ограниченном числе экспериментальных наблюдений. Тогда отыскание оценки состоит в определении либо единственного значения (оценка параметра точкой) или интервала значений (оценка параметра интервалом). В некотором смысле оценка параметра интервалом точнее, поскольку он либо ближе к истинному значению параметра, либо включает его*.

Существует тесная связь между отысканием оценки и информацией. В частности, очевидно, что оценка параметра есть *статистика* (т. е. она является функцией данных), и свойства статистик, обсужденные ранее, применимы также к оценкам.

§ 7.1. ОСНОВНЫЕ ПОНЯТИЯ

Чтобы найти оценку параметра, необходимо сначала выбрать функцию наблюдений (т. е. способ перехода от наблюдений к оценке); это и составляет метод оценки. Численное значение, получаемое в результате применения этого метода к определенному набору наблюдений, называется оценкой.

---

* При определении оценки параметра в виде фиксированного значения следует помнить, что найденное конкретное значение оценки параметра является величиной случайной, меняющейся в сериях повторных равноточных измерений в соответствии с определенным распределением $f(\hat{\theta}|\theta)$. Поэтому задача оценки параметра включает на самом деле не только определение по данным выборки конкретного значения оценки $\hat{\theta}$, но и определение ф. п. в. для этой случайной величины. При этом особенно важной задачей математической статистики является определение достаточной оценки $\hat{\theta}$, которая включает всю информацию о параметре, содержащуюся в выборке, и отвечает ф. п. в. $f(\hat{\theta}|\theta)$ с минимально возможной дисперсией.

Таким образом, оценка параметра фиксированным значением при условии определения ф. п. в., соответствующей достаточной оценке, содержит самую полную информацию об искомом параметре. Метод оценки параметра интервалом значений также содержит полную информацию, если определена функция распределения и для любого интервала значений может быть вычислена его достоверность. — *Прим. ред.*

Выбрав метод оценки, можно затем обсудить его качества в терминах четырех важных желаемых свойств: 1) состоятельность; 2) несмещенность; 3) информационная емкость или эффективность; 4) устойчивость.

Первые два свойства обсуждаются в этом разделе. Третье рассматривается на протяжении гл. 7 и 8. Четвертое свойство, несмотря на значительную практическую важность, не имеет отношения к основной теории отыскания оценок и будет обсуждаться только в § 8.7.

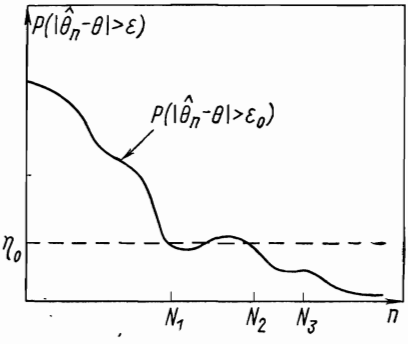


Рис. 7.1. Состоятельность по вероятности. Для любых $\eta>0$ и $\varepsilon>0$ может быть найдено такое N, что кривая расположится ниже значения η при всех $n>N$. При выбранных здесь значениях ε_0 и η_0 N_2 и N_3 удовлетворяют определению, а N_1 еще недостаточно велико

7.1.1. Состоятельность и сходимость. Метод оценки называется *состоятельным*, если оценки, полученные с его помощью, сходятся к истинному значению параметра с увеличением числа наблюдений. Могут быть определены различные виды состоятельности, если использовать различные виды сходимости (см. гл. 3). Дальше мы всегда будем подразумевать *состоятельность по вероятности*, определенную с использованием понятия сходимости по вероятности (см. п. 3.2.3). Пусть $\hat{\theta}_n$ — оценка параметра θ, полученная на основе n наблюдений. $\hat{\theta}_n$ является состоятельной оценкой θ, если для любых $\varepsilon > 0$ и $\eta > 0$ существует такое N, что

$$P(|\hat{\theta}_n - \theta| > \varepsilon) < \eta$$

для всех $n > N$. В этом случае говорят, что $\hat{\theta}_n$ сходится (по вероятности) к θ при возрастании n.

Закон больших чисел (см. § 3.3) эквивалентен утверждению, что среднее выборки* есть состоятельная оценка среднего совокупности (при условии, что дисперсия конечна).

Поскольку состоятельность в значительной мере является асимптотическим свойством, то отсюда не следует, что точность — монотонная функция n, т. е. даже если оценка состоятельна, добавление некоторого числа наблюдений отнюдь не всегда увеличивает точность (рис. 7.1).

---

* Напомним, что среднее выборки есть среднее имеющихся результатов наблюдения, тогда как среднее генеральной совокупности есть среднее распределения генеральной совокупности, из которого сделана выборка.

7.1.2. Смещение и состоятельность. Пусть $\hat{\theta}$ будет оценкой параметра θ, полученной на основе N наблюдений. Мы определяем смещение в оценке $\hat{\theta}$ как отклонение ожидания $\hat{\theta}$ от истинного значения θ_0:

$$b_N(\hat{\theta}) = E(\hat{\theta}) - \theta_0 = E(\hat{\theta} - \theta_0). \tag{7.1}$$

Поэтому оценка называется несмещенной, если для всех N и θ_0

$$b_N(\hat{\theta}) = 0 \quad \text{или} \quad E(\hat{\theta}) = \theta_0.$$

Смещение может быть определено по-разному, поскольку в качестве центра распределения могли бы быть выбраны другие характеристики. Обычно выбирается математическое ожидание (или арифметическое среднее). Это делается из соображений удобства, а также вследствие важности его свойств для нормального распределения. Заметим, что если оценка несмещена, то это вовсе не означает, что в среднем половина оценок будет лежать выше θ_0, а половина ниже. Это было бы правильно, если бы вместо арифметического среднего была бы выбрана медиана.

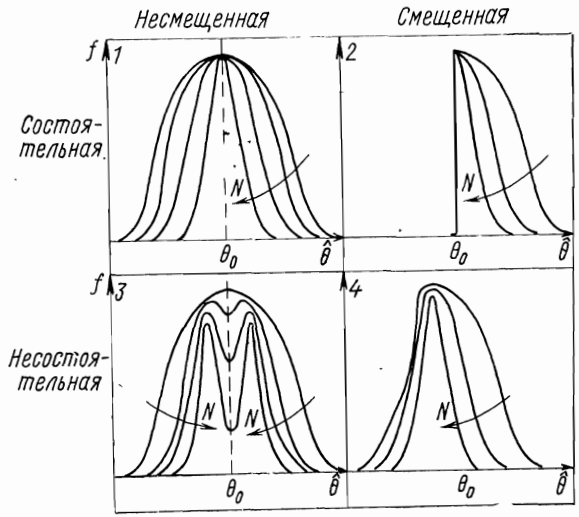


Рис. 7.2. Четыре различных вида ф.п.в. $f(\hat{\theta}, \theta_0)$, обладающей соответствующей комбинацией свойств состоятельности и смещенности (ненормированная ф.п.в.). Стрелки показывают движение кривых с увеличением N

Может показаться, что несмещенность и состоятельность связаны друг с другом, но можно легко показать, что из одного не следует другое. Предположим, что оценки $\hat{\theta}$, получаемые на основании выборок объемом в N наблюдений, распределены в соответствии с ф.п.в. $f(\hat{\theta}, \theta_0)$ относительно истинного значения θ_0. На рис. 7.2 приведены четыре различные функции $f(\hat{\theta}, \theta_0)$, использующие методы оценки с различной степенью смещения и состоятельности. В каждом квадрате рис. 7.2 семейство суживающихся кривых отражает поведение

$f(\hat{\theta})$ с ростом N. В квадрате *1* центр каждой $f(\hat{\theta})$ совпадает с θ_0, а пики с ростом N суживаются. Этот случай соответствует несмещенности и состоятельности. В квадрате *2* каждое распределение $f(\hat{\theta})$ смещено вправо от θ_0, но тем не менее такая оценка состоятельна, поскольку она сходится к истинному значению θ_0. В квадрате *3* каждое $f(\hat{\theta})$ не смещено, но нет сходимости к θ_0. Наконец, квадрат *4* соответствует случаю, когда оценка смещена и несостоятельна.

Если t — несмещенная оценка параметра θ, то можно было бы подумать, что t^2 — несмещенная оценка θ^2. К сожалению, это не так, как это легко может быть показано, если использовать свойство линейности оператора ожидания. Конечно, если t^2 — состоятельная оценка θ^2, то при больших объемах выборки оценка становится несмещенной (как показано в квадрате *2* рис. 7.2). По этой причине мы обычно считаем состоятельность более важным свойством, чем несмещенность.

§ 7.2. ОБЫЧНЫЕ СПОСОБЫ КОНСТРУИРОВАНИЯ СОСТОЯТЕЛЬНЫХ ОЦЕНОК

Предположим, что имеется n измерений случайной переменной X_i с ф.п.в. $f(X, \theta)$, где θ — неизвестный параметр. Если цель эксперимента состоит в извлечении информации об истинном значении θ_0 параметра θ, то необходимо иметь состоятельную оценку $\hat{\theta}$ значения θ_0, т. е. $\hat{\theta}$ должно сходиться к θ_0 с ростом n.

Легко сконструировать величину с подобными свойствами сходимости, если использовать *закон больших чисел* (см. § 3.3.). Из этого закона следует, что

$$n^{-1}\sum_{i=1}^{n} a(X_i) \xrightarrow[n\to\infty]{} E[a(X)] = \int a(X) f(X, \theta)\, dX, \qquad (7.2)$$

где $a(X)$ — некоторая функция X с конечной дисперсией $D[a(X)]$.

Наиболее распространены следующие три метода оценок, использующие закон больших чисел: *метод моментов*, *метод максимального правдоподобия* и *метод наименьших квадратов*. В этом разделе мы увидим, как с помощью этих методов получаются оценки, и покажем, что они состоятельны. Другие свойства этих методов будут обсуждаться ниже.

7.2.1. Метод моментов. Допустим, что можно найти такую функцию $a(X)$, что правая часть уравнения (7.2) становится известной функцией h от θ:

$$E[a(X)] = \int a(X) f(X, \theta)\, dX = h(\theta). \qquad (7.3)$$

Если это справедливо для истинного значения θ_0, то существует «обратная функция»* h^{-1} для h, определенная соотношением

$$h^{-1}[h(\theta_0)] \equiv \theta_0.$$

* Если существует много обратных функций, то только одна из них будет в общем случае давать состоятельную оценку.

Соотношение (7.3) можно «решить» относительно θ_0:

$$\theta_0 = h^{-1}\{E[a(X)]\} = h^{-1}\left[\int a(X) f(X, \theta_0)\, dX\right]. \qquad (7.4)$$

Интуитивно мы чувствуем, что для получения оценки $\hat{\theta}$ значения θ_0 необходимо заменить (7.2) на правую часть уравнения (7.4):

$$\hat{\theta} = h^{-1}\left[n^{-1}\sum_{i=1}^{n} a(X_i)\right]. \qquad (7.5)$$

Таким образом, оценка $\hat{\theta}$ становится функцией известных наблюдений X_i, а неизвестной ф.п.в. $f(X, \theta_0)$, и она состоятельна в соответствии с законом больших чисел [см. соотношение (7.2)].

Если использовать этот метод, то в одномерном случае можно выбрать

$$a(X) = X. \qquad (7.6)$$

Из (7.3) следует, что $h(\theta_0)$ тогда является просто средним $\mu(\theta_0)$ и состоятельная оценка $\hat{\theta}$ задается уравнением

$$\hat{\theta} = \mu^{-1}\left(n^{-1}\sum_{i=1}^{n} X_i\right).$$

Когда оцениваются значения нескольких параметров $\theta = (\theta_1, \ldots, \theta_r)$, то требуются различные функции $a_j(X)$. Соотношение (7.3) тогда приобретает такой вид:

$$E[a_j(X)] = h_j(\theta), \quad j = 1, \ldots, r, \qquad (7.7)$$

а оценками $\hat{\theta}_j$ являются решения системы уравнений, получаемых в результате замены

$$n^{-1}\sum_{i=1}^{n} a_j(X_i) \to E[a_j(X)]$$

в уравнениях (7.7). В частном случае, когда $a_j(X_i) = X^j$, функции $h_j(\theta)$ имеют вид моментов $\mu_j'(\theta)$ распределения $f(X, \theta)$. Отсюда и следует название *метод моментов*.

7.2.2. Неявно определенные оценки. Другая, более общая альтернатива — выбрать функцию $a(X)$ в уравнении (7.3) как функцию обеих переменных X и θ. Допустим, что можно найти такую функцию $a(X, \theta)$, что $h(\theta)$ равно нулю для истинного значения параметра $\theta = \theta_0$:

$$E[a(X, \theta_0)] = \int a(X, \theta_0) f(X, \theta_0)\, dX = h(\theta_0) = 0. \qquad (7.8)$$

Уравнение (7.8) и закон больших чисел (7.2) тогда приводят к неявному методу оценки. Давайте определим экспериментальную функцию ξ наблюдений X_i:

$$\xi(\theta) \equiv n^{-1}\sum_{i=1}^{n} a(X_i, \theta). \qquad (7.9)$$

В соответствии с законом больших чисел (7.2) $\xi(\theta)$ имеет в асимптотике те же самые корни, что и соотношение (7.8):

$$\xi(\theta_0) \xrightarrow[n\to\infty]{} E[a(X, \theta_0)] = 0.$$

В таком случае оценка $\hat{\theta}$ может быть определена неявно как корень уравнения

$$\xi(\theta) = 0. \qquad (7.10)$$

Очевидно, что в частном случае метода моментов в качестве $a(X_i, \theta)$ в (7.9) нужно взять $a(X_i) - E[a(X)]$. Тогда уравнение (7.10) приводит к такому же результату, что и (7.5).

Можно показать, что один из корней уравнения (7.8) дает состоятельную оценку θ_0 при двух условиях: 1) выполняются некоторые условия регулярности (в частности, ξ дифференцируема, среднее ξ и $\partial\xi/\partial\theta$ существуют) и 2) в пределе $n \to \infty$

$$E\left\{\left[\frac{\partial\xi(\theta)}{\partial\theta}\right]_{\theta=\theta_0}\right\} \neq 0. \qquad (7.11)$$

Условие (7.11) означает, что левая часть (7.9) имеет обратную функцию.

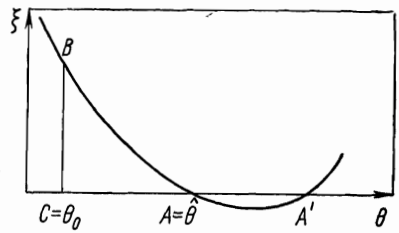


Рис. 7.3. Пример функции $\xi(\theta)$

Нетрудно доказать это утверждение для одномерного случая (рис. 7.3). Пусть A, ..., A' — корни уравнения (7.10); B — точка $(\theta_0, \xi(\theta_0))$, а C — точка $(\theta_0, 0)$. По закону больших чисел B сходится к C, а тангенс угла наклона $\xi(\theta)$ в точке B сходится к константе (7.11), которая по условию отлична от нуля.

Следовательно, по крайней мере одна из точек A, A', ... сходится к C Поэтому один из корней $\hat{\theta}$ является состоятельной оценкой θ [12].

Одна очевидная трудность при использовании такого неявного метода состоит в том, что уравнение (7.10) может иметь много решений и не существует способа для отличия их друг от друга. Конечно, метод хорош, если функция (7.8) имеет только один нуль.

На практике уравнения (7.10) конструируют путем наложения условия минимума или максимума на некоторую другую функцию $g(X, \theta)$, связанную с $a(X, \theta)$ соотношением

$$a(X_i, \theta) = \frac{\partial}{\partial\theta} g(X_i, \theta).$$

Типичными примерами являются хорошо известные методы максимального правдоподобия и минимума χ^2, к которым мы вскоре вернемся.

Если выполняются условия (7.8) и (7.11), то один из максимумов или минимумов экспериментальной величины $n^{-1}\sum_{i=1}^{n} g(X_i, \theta)$ дает

состоятельную оценку $\hat{\theta}$ значения θ_0 и это $\hat{\theta}$ является корнем уравнения

$$\xi(\theta) = \frac{1}{n} \frac{\partial}{\partial \theta} \sum_{i=1}^{n} g(X_i, \theta) = 0. \tag{7.12}$$

В соответствии с условием (7.8) функцию $a(X_i, \theta)$ нужно выбрать в виде

$$a(X_i, \theta_0) = \frac{\partial}{\partial \theta} \{g(X_i, \theta) - E[g(X_i, \theta)]\}_{\theta=\theta_0}$$

при условии, что операторы E и $\partial/\partial\theta$ коммутируют. В этом случае

$$E[a(X_i, \theta_0)] = E\left\{\frac{\partial g(X_i, \theta)}{\partial \theta} - E\left[\frac{\partial g(X_i, \theta)}{\partial \theta}\right]\right\}_{\theta=\theta_0} \equiv 0.$$

Практически $E(\partial g/\partial\theta)$ не должно зависеть от θ_0; если же оно зависит, то оценка бесполезна, поскольку истинное значение θ_0 неизвестно. Таким образом, соотношение (7.8) должно быть дополнено условием

$$\frac{\partial}{\partial \theta} E\left[\frac{\partial g(X_i, \theta)}{\partial \theta}\right]_{\theta=\theta_0} = 0.$$

Очевидно, что метод моментов соответствует такому выбору:

$$g(X_i, \theta) = a(X_i)\,\theta - \int_{\theta_0}^{\theta} \int a(X) f(X, \theta)\, dX\, d\theta. \tag{7.13}$$

7.2.3. Метод максимального правдоподобия. Важным случаем указанного выше подхода является *метод максимального правдоподобия*. В п. 5.1.1 мы определили правдоподобие набора из n независимых наблюдений X_i как

$$L(\mathbf{X} \mid \theta) = \prod_{i=1}^{n} f(X_i, \theta),$$

где $f(X_i, \theta)$ — ф.п.в.

Оценка максимального правдоподобия параметра θ есть такое значение $\hat{\theta}$, которое соответствует максимуму $L(\mathbf{X} \mid \theta)$ при заданных измерениях $\mathbf{X}$.

Метод максимального правдоподобия* — частный случай метода, описанного в п. 7.2.2, для которого

$$g(X_i, \theta) = \ln f(X_i, \theta). \tag{7.14}$$

Тогда (7.12) приобретает следующий вид:

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(X_i, \theta) = \frac{\partial}{\partial \theta} \ln L(\mathbf{X}, \theta) = 0. \tag{7.15}$$

* Отметим, что максимумы $\ln L$ и L совпадают.

Уравнение (7.15) называется *уравнением правдоподобия*. Оно отражает необходимое условие существования максимума $L(\mathbf{X}, \theta)$ при условии, что максимум не лежит на границе области изменения θ. Оценка максимального правдоподобия $\hat{\theta}$ является корнем уравнения (7.15).

Легко показать, что один из этих корней является *состоятельной оценкой* при условии, что f дважды дифференцируема; средние значения

$$\frac{\partial}{\partial\theta} \ln f(X_i, \theta)\,|_{\theta=\theta_0}$$

и

$$\frac{\partial^2}{\partial\theta^2} \ln f(X, \theta)\,|_{\theta=\theta_0}$$

существуют и операторы $\int dX$ и $\partial/\partial\theta$ коммутируют. Используя тот факт, что

$$\int f(X, \theta_0)\, dX = 1,$$

мы имеем

$$\int \frac{\partial}{\partial\theta} [f(X, \theta)]_{\theta=\theta_0}\, dX = 0$$

или

$$\int \frac{df}{\partial\theta}\,\frac{1}{f} f dX = 0. \tag{7.16}$$

С помощью оператора ожидания можно переписать это соотношение:

$$E\left\{\left[\frac{\partial}{\partial\theta} \ln f(X, \theta)\right]_{\theta=\theta_0}\right\} = 0.$$

Продифференцируем (7.16) еще раз по θ:

$$\int \frac{\partial^2 \ln f}{\partial\theta^2} f dX + \int \left(\frac{\partial \ln f}{\partial\theta}\right)^2 f dX = 0.$$

Отсюда следует, что экспериментальная величина

$$E\left[\frac{\partial\xi}{\partial\theta}\Big|_{\theta=\theta_0}\right] \equiv E\left[\frac{1}{n}\,\frac{\partial^2}{\partial\theta^2} \ln L(\mathbf{X}, \theta)\right]_{\theta=\theta_0}$$

в соответствии с законом больших чисел сходится к отрицательной константе

$$E\left(\frac{\partial\xi}{\partial\theta}\right)_{\theta=\theta_0} \xrightarrow{n\to\infty} E\left[\frac{\partial^2}{\partial\theta^2} \ln f(X, \theta)\right]_{\theta=\theta_0} =$$

$$= -E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial\theta}\right)^2\right]_{\theta=\theta_0} < 0. \tag{7.17}$$

Поэтому, как следует из п.7.2.2, корень $\hat{\theta}$ уравнения правдоподобия (7.15) действительно соответствует максимуму функции правдоподобия.

Из (7.16) также следует, что

$$D\left(\frac{\partial}{\partial\theta}\ln f(X,\theta)\right)=\int\left(\frac{\partial\ln f(X,\theta)}{\partial\theta^2}\right)^2 f(X,\theta)\,dX.$$

Таким образом,

$$\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta)=\frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}\ln f(X_i,\theta)\right]$$

распределена со средним нуль и дисперсией $nI(\theta)$ для всех значений n.

Более того, состоятельное решение в *асимптотике* дает абсолютный максимум правдоподобия. Чтобы доказать это, достаточно показать следующее:

1. В соответствии с законом больших чисел $\xi(\theta)$ сходится к постоянной функции

$$\int\frac{\partial f(X,\theta)}{\partial\theta}\,\frac{f(X,\theta_0)}{f(X,\theta)}\,dX$$

и поэтому корни уравнения $\xi(\theta)=0$ сходятся к константам θ_1, θ_2, ...

2. Поскольку логарифм — строго вогнутая функция, то

$$\int f(X,\theta_0)\ln\left[\frac{f(X,\theta)}{f(X,\theta_0)}\right]dX<\ln\int f(X,\theta_0)\,\frac{f(X,\theta)}{f(X,\theta_0)}\,dX=1$$

для всех $\theta\neq\theta_0$. Поэтому

$$\int f(X,\theta_0)\ln f(X,\theta)\,dX<\int f(X,\theta_0)\ln f(X,\theta_0)\,dX$$

для всех $\theta\neq\theta_0$.

Тогда, если $\hat{\theta}'_n$ — последовательность несостоятельных решений, а $\hat{\theta}_n$ — состоятельных, то

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln f(X,\hat{\theta}'_n)=E\left[\ln f(X,\lim_{n\to\infty}\hat{\theta}'_n)\right]<E\left[\ln f(X,\theta_0)\right]=$$

$$=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln f(X,\hat{\theta}_n).$$

Поскольку это доказательство справедливо для экспериментального правдоподобия только в пределе $n\to\infty$, нельзя быть уверенным, что максимум, найденный для конечного n, является абсолютным максимумом. Легко представить, что с ростом n относительные высоты различных максимумов $\ln L$ изменяются. Это отражает

фундаментальную неопределенность истинного значения θ_0 при наличии выборок *конечного размера*.

7.2.4. Методы наименьших квадратов. Рассмотрим множество наблюдений $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, распределенных случайно с ожиданиями

$$E(\mathbf{Y}) = \mathbf{M}(\theta) = \{m_i(\theta)\}, \quad i = 1, \ldots, n,$$

где $\mathbf{M}$ есть n-мерная функция r неизвестных параметров $\theta = \{\theta_1, \ldots, \theta_r\}$.

В соответствии с *методами наименьших квадратов* нужно отыскивать минимум квадратичной формы

$$g(\mathbf{Y}, \theta) = [\mathbf{Y} - \mathbf{M}(\theta)]^T \underset{\sim}{W} [\mathbf{Y} - \mathbf{M}(\theta)], \tag{7.18}$$

где $\underset{\sim}{W}$ — матрица весов, не зависящая от θ.

По-прежнему *оценка наименьших квадратов* является корнем уравнения (7.10), которое в этом случае приобретает форму

$$\xi(\theta) = \frac{\partial g}{\partial \theta} = -\frac{\partial \mathbf{M}(\theta)^T}{\partial \theta} \underset{\sim}{W} [\mathbf{Y} - \mathbf{M}(\theta)] = 0. \tag{7.19}$$

Если $\underset{\sim}{W}$ остается положительно определенной и если матрица вторых моментов случайных переменных Y_i остается ограниченной при стремлении n к бесконечности, то можно показать, что один из корней $\hat{\theta}$ уравнения (7.19) является *состоятельной оценкой*.

Для этого предположим, что $\underset{\sim}{W}$ диагональна (если это не так, то она может быть всегда диагонализирована, что не отразится на нашем доказательстве). Обозначим

$$\mathbf{X}_i(\theta) = \left\{ -\frac{\partial m_i(\theta)}{\partial \theta_k} w_{ii} [Y_i - m_i(\theta)] \right\}, \quad k = 1, \ldots, r.$$

Тогда вектор $\xi(\theta)$ с r компонентами в θ-пространстве

$$\xi_k(\theta) = -\sum_{i=1}^{n} \frac{\partial m_i(\theta)}{\partial \theta_k} w_{ii} [Y_i - m_i(\theta)].$$

может быть записан

$$\xi(\theta) = \sum_{i=1}^{n} \mathbf{X}_i(\theta). \tag{7.20}$$

По построению

$$E[\mathbf{X}_i(\theta_0)] = 0.$$

Сейчас непосредственно могут быть использованы выводы п. 7.2.2, поскольку уравнение (7.20) имеет форму (7.9). Поэтому, если уравнение (7.11) справедливо в пределе $n \to \infty$, то один из корней урав-

нения (7.19) сходится к θ_0 при условии, что $\partial\xi/\partial\theta$ сходится к ненулевой константе. Фактически

$$\frac{\partial\xi}{\partial\theta} = -\frac{\partial^2 \mathbf{M}^T}{\partial\theta^2} \underset{\sim}{W} [\mathbf{Y} - \mathbf{M}(\theta)] + \frac{\partial \mathbf{M}^T}{\partial\theta} \underset{\sim}{W} \frac{\partial \mathbf{M}}{\partial\theta}. \qquad (7.21)$$

В соответствии с законом больших чисел первый член правой части (7.21) сходится к нулю в пределе $n \to \infty$. Для $\theta = \theta_0$ второй член положительно определен. Поэтому корень уравнения (7.19), являющийся состоятельной оценкой θ_0, соответствует (в пределе $n \to \infty$) минимуму.

Важным частным случаем является *линейный метод наименьших квадратов*, когда $\mathbf{M}(\theta)$ — линейная функция θ, т. е.

$$\mathbf{M}(\theta) = E(\mathbf{Y}) = \underset{\sim}{A}\theta.$$

Матрица $\underset{\sim}{A}$ с постоянными элементами называется *матрицей планирования*.

Тогда оценка является единственным корнем $\hat{\theta}$ уравнения

$$\xi(\theta) = -2\underset{\sim}{A}^T \underset{\sim}{W} (\mathbf{Y} - \underset{\sim}{A}\theta) = 0. \qquad (7.22)$$

Этот корень равен

$$\hat{\theta} = (\underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{W} \mathbf{Y}. \qquad (7.23)$$

Если $\underset{\sim}{A}$ такова, что $\underset{\sim}{A}^T \underset{\sim}{W} \underset{\sim}{A}$ несингулярна, то $\hat{\theta}$ — состоятельная оценка.

§ 7.3. АСИМПТОТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ ОЦЕНОК

Асимптотическими мы всегда называем свойства оценок, когда число наблюдений n становится бесконечно большим.

7.3.1. Асимптотическая нормальность. Для того чтобы найти асимптотическое распределение состоятельных оценок, привлечем центральную предельную теорему (см. п.3.3.2). Из этой теоремы следует, что

$$n^{-1} \sum_{i=1}^{n} a(X_i) \qquad (7.24)$$

в асимптотическом пределе распределена по нормальному закону со средним $E[a(\mathbf{X})]$ и дисперсией $n^{-1}D[a(X)]$ при условии, что $D[a(X)]$ конечна.

Величину $\hat{\theta}$ в уравнении (7.5) можно разложить в ряд вблизи $E[a(X)]$:

$$\hat{\theta} = h^{-1}\left[n^{-1}\sum_{i=1}^{n} a(X_i)\right] = h^{-1}\{E[a(X)]\} + \frac{\partial h^{-1}}{\partial a}\{E[a(X)]\} \times$$
$$\times \left\{n^{-1}\sum_{i=1}^{n} a(X_i) - E[a(X)]\right\} + O(n^{-1}). \qquad (7.25)$$

Очевидно, что при такой степени точности $\hat{\theta}$ в асимптотическом пределе распределено по нормальному закону, поскольку выражение (7.24), а также все другие члены правой части (7.25) — константы.

Однако нормальность распределения оценки в виде (7.5) является в высшей степени асимптотическим свойством. Это следует из того факта, что случайная переменная (7.24) распределена по нормальному закону для конечного числа наблюдений, только если $A(X_i)$ распределена также по нормальному закону. Даже если бы это выполнялось, то функция h в общем случае нарушала бы свойство нормальности при конечном числе наблюдений.

Дисперсия случайной переменной (7.24) пропорциональна n^{-1}. Поэтому в приближении уравнения (7.25) дисперсия $\hat{\theta}$ пропорциональна n^{-1}. Нетрудно увидеть, что в асимптотике дисперсия

$$D(\hat{\theta}) = E(\hat{\theta} - \theta_0)^2 = \frac{1}{n}\left(\frac{\partial h^{-1}}{\partial a}\right)^2 D(a),$$

где $D(a)$ — дисперсия $a(X)$.

В случае многих параметров $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_r)$ функции h^{-1} и a имеют вид векторных функций $\mathbf{h}^{-1}$ и $\mathbf{a}$. Тогда матрица вторых моментов асимптотического распределения $\hat{\boldsymbol{\theta}}$ приобретает следующий вид:

$$\underset{\sim}{D}(\hat{\boldsymbol{\theta}}) = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T] = \frac{1}{n}\left[\frac{\partial \mathbf{h}^{-1}}{\partial \mathbf{a}}\right] \underset{\sim}{D}(\mathbf{a}) \left[\frac{\partial \mathbf{h}^{-1}}{\partial \mathbf{a}}\right]^T.$$

Здесь $\underset{\sim}{D}(\mathbf{a})$ — матрица вторых моментов $\mathbf{a}$; $\partial \mathbf{h}^{-1}/\partial \mathbf{a}$ — матрица с элементами $\partial h_i^{-1}/\partial a_j$ и $h_j^{-1}(\mathbf{a}) = \theta_{0i}$.

Смещенность оценки обусловлена членами порядка n^{-1}, которые были опущены в правой части уравнения (7.25). Из этого следует, что величина смещения по меньшей мере порядка n^{-1}. Аналогичные результаты получаются, если используются неявно определенные оценки. Например, свойство нормальности следует из того факта, что точка B на рис.7.3 распределена по нормальному закону вблизи C. Следовательно, в соответствии с центральной предельной теоремой $\xi(\theta_0)$ в (7.9) в окрестности нуля в асимптотике распределено по нормальному закону. Тогда в пределе, когда AC становится пропорциональной BC, разность $\hat{\theta} - \theta_0$ также распределена по нормальному закону. Матрица дисперсий для $\hat{\theta}$ равна:

$$\underset{\sim}{D}(\hat{\theta}) = \frac{1}{n}\left[E\left(\frac{\partial \xi(\theta)}{\partial \theta}\right)_{\theta=\theta_0}\right]^{-1} \underset{\sim}{D}[a(X, \theta_0)]\left[E\left(\frac{\partial \xi(\theta)}{\partial \theta}\right)_{\theta=\theta_0}\right]^{-1^T}, \tag{7.26}$$

поскольку $E[(\partial \xi(\theta)/\partial\theta)]_{\theta=\theta_0}$ в асимптотике равна тангенсу угла наклона BA, а $n^{-1}D[a(X, \theta_0)]$ — асимптотической дисперсии BC.

В качестве примера вычислим значение $D(\hat{\theta})$ в асимптотическом пределе для метода максимального правдоподобия, где

$$a(X, \theta) = \frac{\partial}{\partial \theta} \ln f(X, \theta).$$

Тогда получаем

$$D[a(X, \theta_0)] = E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial \theta}\right)^2\right]_{\theta=\theta_0} = -E\left(\frac{\partial^2 \ln f(X, \theta)}{\partial \theta^2}\right)_{\theta=\theta_0}$$

из уравнения (7.17) и равенства

$$E\left[\left(\frac{\partial \xi(\theta)}{\partial \theta}\right)_{\theta=\theta_0}\right] = E\left(\frac{\partial^2 \ln f(X, \theta)}{\partial \theta^2}\right)_{\theta=\theta_0}$$

в соответствии с законом больших чисел. Из уравнения (7.26) следует, что в асимптотике

$$D(\hat{\theta}) = \frac{1}{n}\left[E\left(\frac{\partial \ln f(X, \theta)}{\partial \theta}\right)^2\right]^{-1}_{\theta=\theta_0} =$$

$$= -\frac{1}{n}\left[E\left(\frac{\partial^2 \ln f(X, \theta)}{\partial \theta^2}\right)^2\right]^{-1}_{\theta=\theta_0} = [nI(\theta_0)]^{-1}, \qquad (7.27)$$

т. е. дисперсия равна обратному значению матрицы информации.

Поэтому для метода максимального правдоподобия величина $\sqrt{n}(\hat{\theta} - \theta_0)$ в асимптотике распределена по нормальному закону $N[0, I^{-1}(\theta_0)]$. Соответственно $\partial \ln L(\mathbf{X}, \theta)/\partial\theta$ асимптотически нормальна со средним значением нуль и матрицей дисперсий $nI(\theta)$.

7.3.2. Асимптотическое разложение моментов оценок. Покажем, как могут быть рассчитаны моменты оценок, используя разложение в ряд по степеням n^{-1}. Особый интерес представляют моменты $E(\hat{\theta} - \theta_0)$, $D(\hat{\theta})$, т. е. *смещение и дисперсия* θ соответственно.

Вспомним, что

$$E(\hat{\theta} - \theta_0) = E(\hat{\theta}) - \theta_0; \quad D(\hat{\theta} - \theta_0) = D(\hat{\theta});$$

$$E[(\hat{\theta} - \theta_0)^2] = D(\hat{\theta}) + [E(\hat{\theta} - \theta_0)]^2.$$

Пусть оценка $\hat{\theta}$ является корнем уравнения (7.12), причем $E[\xi(\theta_0)] = 0$ и соотношение (7.11) справедливо. Разложим $\xi(\hat{\theta})$ в ряд относительно истинного, но неизвестного значения θ_0, используя уравнение (7.12):

$$\xi(\hat{\theta}) = \xi(\theta_0) + (\hat{\theta} - \theta_0)\frac{\partial \xi}{\partial \theta}\Big|_{\theta=\theta_0} + \frac{1}{2}(\hat{\theta} - \theta_0)^2 \frac{\partial^2 \xi}{\partial \theta^2}\Big|_{\theta=\theta_0} + \ldots = 0.$$

Тогда

$$\xi(\theta_0) = -(\hat{\theta} - \theta_0)\frac{\partial \xi}{\partial \theta}\Big|_{\theta=\theta_0} - \frac{1}{2}(\hat{\theta} - \theta_0)^2 \frac{\partial^2 \xi}{\partial \theta^2}\Big|_{\theta=\theta_0} - \ldots \qquad (7.28)$$

Чтобы выразить $(\hat{\theta} - \theta_0)$ через $\xi(\theta_0)$, используем теорему Лагранжа об инверсии ряда, которая утверждает, что если

$$n = \sum_{i=1}^{\infty} a_i v^i, \qquad (7.29)$$

то

$$v = \sum_{j=1}^{\infty} \frac{u^j}{j!}\left[\frac{d^{j-1}}{\partial x^{j-1}}\left(\sum_{i=1}^{\infty} a_i X^{i-1}\right)^{-j}\right]_{X=0}$$

при условии, что $a_1 \neq 0$ и ряд (7.29) сходятся. Применяя эту теорему к ряду (7.28), получаем:

$$\hat{\theta}-\theta_0=\left[\frac{\xi(\theta_0)}{-\partial\xi/\partial\theta}+\frac{1}{2}\frac{\xi^2(\theta_0)\,\partial^2\xi/\partial\theta^2}{(-\partial\xi/\partial\theta)^3}+\ldots\right]_{\theta=\theta_0}. \qquad (7.30)$$

По закону больших чисел все производные $\left.\frac{\partial^k\xi(\theta)}{\partial\theta^k}\right|$ сходятся к фиксированным значениям для $n\to\infty$, в частности

$$\left.\frac{\partial\xi}{\partial\theta}\right|_{\theta=\theta_0}\xrightarrow[n\to\infty]{}E\left[\left(\frac{\partial\xi}{\partial\theta}\right)_{\theta=\theta_0}\right]\neq 0$$

в соответствии с предположением [см. уравнение (7.11)]. Таким образом, использование теоремы Лагранжа оправдано при достаточно больших n. Также по закону больших чисел $\xi(\theta_0)$ сходится к $E[\xi(\theta_0)]=0$. Введем новые случайные переменные X, Y, Z:

$$\xi(\theta_0)=E[\xi(\theta_0)]+X=X;$$

$$\left.\frac{\partial\xi}{\partial\theta}\right|_{\theta=\theta_0}=E\left[\left(\frac{\partial\xi}{\partial\theta}\right)_{\theta=\theta_0}\right]+Y\equiv -a+Y;$$

$$\left.\frac{\partial^2\xi}{\partial\theta^2}\right|_{\theta=\theta_0}=E\left[\left(\frac{\partial^2\xi}{\partial\theta^2}\right)_{\theta=\theta_0}\right]+Z=b+Z.$$

Здесь a и b введены с целью удобства обозначений. По построению

$$E(X)=E(Y)=E(Z)=0.$$

Тогда легко показать*, что моменты имеют следующую зависимость от n:

$$E(X^\alpha Y^\beta Z^\gamma)=n^{1-(\alpha+\beta+\gamma)}.$$

Переписывая (7.30) с использованием новых переменных и оставляя члены порядка n^{-1}, получаем

$$\hat{\theta}-\theta_0=\frac{X}{a}+\frac{1}{a^2}\left(XY+\frac{b}{2a}X^2\right)+O\left(\frac{1}{n^2}\right). \qquad (7.31)$$

Наконец, для средних и дисперсий получаем:

$$E(\hat{\theta}-\theta_0)=\frac{1}{a^2}\left[E(XY)+\frac{b}{2a}E(X^2)\right]+O\left(\frac{1}{n^2}\right);$$

$$D(\hat{\theta}-\theta_0)=\frac{1}{a^2}E(X^2)+O\left(\frac{1}{n^2}\right).$$

Возвращаясь обратно к функциям g (для которых отыскивается минимум или максимум), можно выразить смещение и дисперсию

---

* Для этого нужно использовать независимость членов суммы

$$X=\frac{1}{N}\sum_{i=1}^{N}a(X_i,\theta_0)$$

и подобных сумм для Y и Z. Утверждение справедливо, когда по крайней мере два из трех показателей степени α, β и γ $\leqslant 1$.

с точностью до членов порядка $1/n$:

$$b_n(\hat{\theta}) = E(\hat{\theta} - \theta_0) = n^{-1}\left[E\left(\frac{\partial^2 g}{\partial\theta^2}\right)\right]^{-2}_{\theta=\theta_0}\left\{D\left(\frac{\partial g}{\partial\theta}, \frac{\partial^2 g}{\partial\theta^2}\right) - \right.$$

$$\left. - \frac{E(\partial^3 g/\partial\theta^3)}{E(\partial^2 g/\partial\theta^2)} E\left[\left(\frac{\partial g}{\partial\theta}\right)^2\right]\right\}_{\theta=\theta_0} + O\left(\frac{1}{n^2}\right); \qquad (7.32)$$

$$D(\hat{\theta} - \theta) = \frac{1}{n}\left\{\frac{E[(\partial g/\partial\theta)^2]}{[E(\partial^2 g/\partial\theta^2)]^2}\right\}_{\theta=\theta_0} + O\left(\frac{1}{n^2}\right),$$

где

$$D\left(\frac{\partial g}{\partial\theta}, \frac{\partial^2 g}{\partial\theta^2}\right) = E\left(\frac{\partial g}{\partial\theta}\frac{\partial^2 g}{\partial\theta^2}\right) - E\left(\frac{\partial g}{\partial\theta}\right)E\left(\frac{\partial^2 g}{\partial\theta^2}\right).$$

Остается еще свобода выбора различных функций $g(X, \theta)$, и это обстоятельство может быть использовано для минимизации смещения или (и) дисперсии. Например, можно было бы рассчитать новую оценку $\hat{\theta}' = \hat{\theta} - E(\hat{\theta} - \theta_0)$, которая была бы несмещенной. Но обычно $E(\hat{\theta} - \theta_0)$ зависит от θ_0 и в этом случае можно оставить в $E(\hat{\theta} - \theta_0)$ главный член по n, таким образом, уменьшая смещение до членов более высокого порядка по n^{-1}. Однако такое уменьшение смещения достигается ценой возрастания дисперсии. В большинстве случаев

$$D(\hat{\theta}' - \theta_0) > D(\hat{\theta} - \theta_0).$$

7.3.3. Асимптотическое смещение и дисперсия обычных оценок. Рассчитаем точно смещение и дисперсию с помощью (7.32) для трех обычных оценок. Ограничимся случаем только одного параметра θ.

1. Метод моментов. В соответствии с (7.6) и (7.13) функция g для этого метода равна

$$g(X_i, \theta) = X_i\,\theta - \int_{\theta_0}^{\theta}\int Xf(X, \theta)\,dX d\theta.$$

Из этого следует, что $E[\xi(\theta_0)] = E\left(\frac{\partial g}{\partial\theta}\right)_{\theta=\theta_0} = E(X) - \mu_1'(\theta_0) = 0$;

$$E\left(\frac{\partial^2 g}{\partial\theta^2}\right) = -\frac{\partial\mu_1'(\theta)}{\partial\theta}; \quad E\left(\frac{\partial g}{\partial\theta}\frac{\partial^2 g}{\partial\theta^2}\right) = 0;$$

$$D\left(\frac{\partial g}{\partial\theta}\right) = E\left[\left(\frac{\partial g}{\partial\theta}\right)^2\right] = \mu_2(\theta_0);$$

$$E\left(\frac{\partial^3 g}{\partial\theta^3}\right) = -\frac{\partial^2\mu_1'(\theta_0)}{\partial\theta^2}.$$

Здесь для моментов используются те же обозначения, что и в 2.4.6. В соответствии с (7.32) для смещения и дисперсии $\hat{\theta} - \theta_0$ получаем

$$b_n(\hat{\theta}) = E(\hat{\theta} - \theta_0) = -\frac{1}{2n}\mu_2 \frac{\partial\mu_1'/\partial\theta}{(\partial\mu_1'/\partial\theta)^3}\Bigg|_{\theta=\theta_0};$$

$$D(\hat{\theta} - \theta_0) = \frac{1}{n}\frac{\mu_2}{(\partial\mu_1'/\partial\theta)^2}\Bigg|_{\theta=\theta_0}.$$

2. Метод наименьших квадратов в линейном случае. В этом случае g задается уравнением (7.18), а ξ — (7.22), $A = a$, $W = 1/\sigma^2$ для одномерного случая. Следовательно,

$$E[\xi(\theta_0)] = E\left(\frac{\partial g}{\partial\theta}\right)_{\theta=\theta_0} = -2E\left[\frac{a(Y - a\theta)}{\sigma^2}\right] = 0.$$

При этом

$$E\left(\frac{\partial^2 g}{\partial\theta^2}\right) = 2\left(\frac{a}{\sigma}\right)^2; \quad E\left(\frac{\partial g}{\partial\theta}\frac{\partial^2 g}{\partial\theta^2}\right) = 0;$$

$\partial^m g/\partial\theta^m = 0$ для $m \geqslant 3$.

Видно, что оценка (7.23) метода наименьших квадратов в линейном случае не смещена во всех порядках по n [члены, входящие в $O(n^{-2})$ в (7.32), пропорциональны производным более высокого порядка, которые равны нулю].

3. Метод максимального правдоподобия. В этом случае g задается уравнением (7.14), и мы уже видели в (7.16), что $E[\xi(\theta_0)] = 0$. Используя соотношение (7.27) между вторыми производными g и информацией I (на одно событие) и соотношение (7.17), получаем для смещения и дисперсии:

$$\left.\begin{aligned} b_n(\hat{\theta}) &= E(\hat{\theta} - \theta_0) = \frac{K + 2J}{nI^2}; \\ D(\hat{\theta} - \theta_0) &= 1/nI, \end{aligned}\right\} \tag{7.33}$$

где

$$K \equiv E\left(\frac{\partial^3 \ln f}{\partial\theta^3}\right)_{\theta=\theta_0};$$

$$J = E\left(\frac{\partial^2 \ln f}{\partial\theta^2}\frac{\partial \ln f}{\partial\theta}\right)_{\theta=\theta_0}.$$

В этих обозначениях уравнение (7.31) приобретает следующий вид:

$$\hat{\theta} - \theta_0 = \frac{X}{I\sqrt{n}}\left(1 + \frac{Y}{I\sqrt{n}} + \frac{1}{2}\frac{X^2 K}{nI^3}\right) + O\left(\frac{1}{n^2}\right).$$

Когда I, J и K вычисляются через производные экспериментальной функции правдоподобия при ее максимальном значении, учитываются случайные члены порядка $n^{-1/2}$. Таким образом, выражения (7.33) для смещения и дисперсии справедливы с точностью до членов порядка $n^{-3/2}$.

Мы вернемся к вопросу уменьшения смещения в § 8.6. Заметим, что для многомерного случая все эти выражения неверны. Однако соответствующие формулы для многомерного случая могут быть получены таким же методом.

§ 7.4. ИНФОРМАЦИЯ И ТОЧНОСТЬ ОЦЕНКИ

Во введении в эту главу указывалось, что оценка есть случайная переменная и поэтому она имеет распределение вероятности по θ, часто называемое *выборочным распределением*. В § 7.1 мы определили состоятельность и несмещенность таким образом, что выборочное распределение *состоятельной* оценки для достаточно большого числа наблюдений сконцентрировано сколь угодно близко к истинному значению θ_0, а выборочное распределение *несмещенной состоятельной* оценки всегда имеет в качестве центра θ_0.

Интуитивно мы понимаем, что точность оценки связана с шириной этого распределения, т. е. обратна дисперсии оценки. В этом параграфе мы рассмотрим связь между информацией оценки и дисперсией ее выборочного распределения.

7.4.1. Нижние границы для дисперсии — неравенство Крамера—Рао. Пусть $\mathbf{X}$ — наблюдения случайной величины, распределенной в соответствии с функцией плотности $f(\mathbf{X}\mid\theta)$, и пусть оценка $\hat{\theta}$ имеет выборочное распределение $q(\hat{\theta}\mid\theta)$. Обозначим функцию правдоподобия наблюдений $L_{\mathbf{X}}$, функцию правдоподобия оценки $L_{\hat{\theta}}$ и соответствующие информации через $I_{\mathbf{X}}$ и $I_{\hat{\theta}}$.

В соответствии с (7.1) смещение есть функция истинного значения

$$b = E(\hat{\theta}) - \theta_0 = \int \hat{\theta}(\mathbf{X}) f(\mathbf{X}\mid\theta_0)\, d\mathbf{X} - \theta_0.$$

Временно опустим индекс у θ. Дисперсия выборочного распределения

$$D(\hat{\theta}) = \int [\hat{\theta} - E(\hat{\theta})]^2 q(\hat{\theta}\mid\theta)\, d\hat{\theta} \tag{7.34}$$

связана с информацией *неравенством Крамера—Рао*

$$D(\hat{\theta}) \geqslant \frac{[1+(db/d\theta)]^2}{I_{\hat{\theta}}} \geqslant \frac{[1+(db/d\theta)]^2}{I_{\mathbf{X}}}. \tag{7.35}$$

Первая часть этого важного неравенства утверждает, что дисперсия несмещенной оценки ограничена снизу величиной, обратной информации, которую она содержит.

Вторая часть неравенства вытекает непосредственно из (5.10), если t заменить $\hat{\theta}$, и она означает, что дисперсия любой несмещенной оценки ограничена снизу величиной, обратной информации, содержащейся в наблюдениях. Заменяя $I_{\mathbf{X}}$ в соответствии с (5.2), получаем для любой оценки

$$D(\hat{\theta}) \geqslant \left\{ \left(\frac{d\tau(\theta)}{d\theta} \right)^2 \Big/ E\left[\left(\frac{\partial \ln L_{\mathbf{X}}}{\partial \theta} \right)^2 \right] \right\}, \tag{7.36}$$

где

$$\tau(\theta) \equiv E(\hat{\theta}) = \theta + b(\theta). \tag{7.37}$$

Неравенства (7.35) и (7.36) справедливы при условии, если: 1) область изменения переменных $\mathbf{X}$ не зависит от θ; 2) $L_{\mathbf{X}}$ должна быть достаточно регулярной, чтобы дифференцирование по θ и интегрирование по $\mathbf{X}$ коммутировали друг с другом.

Доказательство неравенства приведено ниже.

Запишем в соответствии с определением

$$\int L_{\hat{\theta}}\, d\hat{\theta} = \int q(\hat{\theta} \mid \theta)\, d\hat{\theta} = 1.$$

Таким образом, неважно, назовем ли мы его выборочным распределением $q(\hat{\theta} \mid \theta)$ или $L_{\hat{\theta}}$. Дифференцирование по θ дает

$$\int \frac{\partial q}{\partial \theta}\, d\hat{\theta} = \int \frac{\partial (\ln q)}{\partial \theta}\, q\, d\hat{\theta} = 0. \tag{7.38}$$

В результате дифференцирования (7.37) по θ получим

$$\frac{\partial}{\partial \theta} \int \hat{\theta} q(\hat{\theta} \mid \theta)\, d\hat{\theta} = \int \hat{\theta}\, \frac{\partial (\ln q)}{\partial \theta}\, q\, d\hat{\theta} = 1 + \frac{db}{d\theta}. \tag{7.39}$$

Поскольку $\theta + b(\theta)$ не зависит от переменной интегрирования, можно записать:

$$\int [\theta + b(\theta)]\, \frac{\partial (\ln q)}{\partial \theta}\, q\, d\hat{\theta} = 0.$$

Вычитая это выражение из (7.39), получаем

$$\int [\hat{\theta} - \theta - b(\theta)]\, \frac{\partial (\ln q)}{\partial \theta}\, q\, d\hat{\theta} = 1 + \frac{db}{d\theta}.$$

Использовав далее неравенство Шварца, получим

$$\int [\hat{\theta} - \theta - b(\theta)]^2 q(\hat{\theta} \mid \theta)\, d\hat{\theta} \int \left[\frac{\partial \ln q}{\partial \theta}\right]^2 q(\hat{\theta} \mid \theta)\, d\hat{\theta} \geqslant \left(1 + \frac{db}{d\theta}\right)^2. \tag{7.40}$$

Поскольку в соответствии с (7.34) и (7.37) первый интеграл является выражением для $D(\hat{\theta})$, неравенство (7.40) приобретает следующий вид:

$$D(\hat{\theta}) \geqslant \frac{\left(1 + \frac{db}{d\theta}\right)^2}{E\left[\left(\frac{\partial \ln q}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{db}{d\theta}\right)^2}{E\left[\left(\frac{\partial \ln L_{\hat{\theta}}}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{db}{d\theta}\right)^2}{I_{\hat{\theta}}},$$

что и доказывает неравенство Крамера — Рао (7.35).

7.4.2. Эффективность и минимальная дисперсия. Рассмотрим вначале условия, при которых неравенство (7.40) и первое неравенство в (7.35) становятся равенствами. Это происходит тогда, когда

$$\frac{\partial \ln L_{\hat{\theta}}}{\partial \theta} = A(\theta)[\hat{\theta} - \theta - b(\theta)]. \tag{7.41}$$

Из (7.41) следует, что *минимум дисперсии* для статистики $\hat{\theta}$

$$D(\hat{\theta}) = (I_{\hat{\theta}})^{-1}\left(1+\frac{db}{d\theta}\right)^2 \tag{7.42}$$

достигается, когда выборочное распределение $\hat{\theta}$ имеет экспоненциальную форму (5.8):

$$L_{\hat{\theta}} = q(\hat{\theta}\,|\,\theta) = \exp[a(\theta)\hat{\theta} + \beta(\hat{\theta}) + c(\theta)]. \tag{7.43}$$

Здесь $a(\theta)$ — интеграл от произвольной функции $A(\theta)$ в (7.41), $\beta(\hat{\theta})$ — константа интегрирования, а $c(\theta)$ — произвольная функция.

Когда также и второе неравенство в (7.35) превращается в равенство

$$I_{\hat{\theta}} = I_{\mathbf{X}}, \tag{7.44}$$

то дисперсия $D(\hat{\theta})$ достигает своей *границы минимальной дисперсии*. В этом случае говорят, что статистика $\hat{\theta}$ является *эффективной оценкой*. Важно отметить различие между:

1) *минимальной дисперсией*, когда данная оценка $\hat{\theta}$ имеет наименьшую дисперсию в семействе рассматриваемых оценок;

2) *границей минимальной дисперсии* для выборки данных $\mathbf{X}$ достигаемой, когда выполняются соотношения (7.42) и (7.44) (оценка $\hat{\theta}$ эффективна).

Важным частным случаем является случай, когда выборочное распределение $q(\hat{\theta}\,|\,\theta)$ нормально. В этом случае фразы о том, что оценка $\hat{\theta}$ достигает границы минимальной дисперсии и что извлекается максимум информации, полностью эквивалентны.

Уравнение (7.44) является необходимым и достаточным условием того, что $\hat{\theta}$ служит *достаточной* статистикой для θ.

Если существует достаточная статистика t, то в соответствии с теоремой Дармуа (см. п. 5.3.4) ф.п.в. $f(\mathbf{X}, \theta)$ имеет экспоненциальную форму (5.8). С другой стороны, если $f(\mathbf{X}, \theta)$ имеет экспоненциальную форму (5.8), то

$$t = n^{-1}\sum_{i=1}^{n}\alpha(X_i) \tag{7.45}$$

является достаточной статистикой из выборочного распределения экспоненциальной формы

$$q(t\,|\,\theta) = \exp[a(\theta)t + \beta_1(t) + c(\theta)].$$

Здесь β_1 можно выразить через β, если использовать уравнение (7.45). Можно показать, что $E(t) = -(dc/d\theta)/(da/d\theta) \equiv r(\theta)$. Поэтому статистика (7.45) служит несмещенной оценкой* для $r(\theta)$,

* Она является также оценкой максимального правдоподобия.

и поскольку $q(t|\theta)$ имеет экспоненциальную форму, справедливо соотношение

$$D(t) = I_t^{-1}[r(\theta)] = I_{\mathbf{X}}^{-1}[r(\theta)].$$

Наконец, несмещенная оценка t с дисперсией, равной границе минимальной дисперсии, существует тогда и только тогда, когда распределение $f(\mathbf{X}, \theta)$ имеет достаточные статистики. В этом случае функция $r(\theta)$, оцениваемая с помощью t, единственна с точностью до линейного преобразования [12].

Примеры. Для иллюстрации этих результатов рассмотрим примеры из п. 5.3.4.

В случае 1, когда μ неизвестно, а σ^2 известно, $r(\mu) = -\mu/\sigma^2$ и оценка μ эффективна.

В случае 2 (μ — известно, а σ^2 — неизвестно) $r(\sigma) = \mu^2/2 + \sigma^2/2\sqrt{2\pi}$ и эффективная несмещенная оценка существует только для функции σ^2.

Наконец, в случае 3, когда неизвестны и μ, и σ^2, мы имеем $r_1(\mu) = \mu$ и $r_2(\sigma) = \mu^2/2 + \sigma^2/2\sqrt{2\pi}$. Поэтому для μ и σ^2 существуют несмещенные эффективные оценки. Этот пример показывает, почему обычно отыскивают оценку σ^2, а не σ; σ^2 — единственная функция σ, имеющая эффективную и несмещенную оценку. В то же время можно сконструировать несмещенную оценку σ, которая, однако, не достигает границы минимальной дисперсии. Иными словами, она не содержит максимума информации о σ, т. е. менее эффективна. Такой оценкой является [12]

$$T = \sqrt{\frac{n}{2}}\,\frac{\Gamma(n/2)}{\Gamma((n+1)/2)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}\,. \qquad (7.46)$$

В сравнении со всеми несмещенными оценками T обладает минимальной дисперсией. Любая другая оценка с меньшей дисперсией не была бы несмещенной. Это приводит нас к важному результату большей общности, чем приведенный пример.

Если существует достаточная статистика t, то необходимое и достаточное условие того, что несмещенная оценка T имеет минимальную дисперсию, состоит в том, что она должна быть функцией только t:

$$T = T(t). \qquad (7.47)$$

В только что приведенном примере T была функцией достаточной статистики

$$t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2}.$$

Статистика T, как и в общем случае, неэффективна, т. е. она необязательно достигает границы минимальной дисперсии, поскольку существует только одна функция параметра, оценка которой может быть несмещенной и эффективной.

7.4.3. Неравенство Крамера—Рао для нескольких параметров. Полученные результаты очень просто обобщить на случай нескольких параметров:

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k).$$

Как показано в п.5.2.1, информация имеет вид матрицы $I_{\mathbf{X}}$, элементы которой вычисляются в соответствии с (5.3). Для несмещенных оценок θ неравенство Крамера—Рао приобретает такой вид:

$$D(\hat{\theta}_i) \geqslant [I_{\hat{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\theta})]_{ii} \geqslant [I_{\mathbf{X}}^{-1}(\boldsymbol{\theta})]_{ii}, \quad i = 1, \ldots, k, \tag{7.48}$$

где $[I^{-1}]_{ii}$ обозначает i-й диагональный элемент матрицы, обратной матрице информации.

Чтобы сформулировать этот результат для более общего случая, введем эллипсоид в $\boldsymbol{\theta}$-пространстве:

$$(\mathbf{u}-\boldsymbol{\theta})^T M (\mathbf{u}-\boldsymbol{\theta}) = \text{const},$$

где $\mathbf{u}$ — некоторая векторная переменная, а M — положительно определенная матрица. Тогда неравенство Крамера — Рао означает, что эллипсоиды

$$(\mathbf{u}-\boldsymbol{\theta})^T I_{\hat{\boldsymbol{\theta}}} (\mathbf{u}-\boldsymbol{\theta}) = c$$

лежат где-то внутри эллипсоида

$$(\mathbf{u}-\boldsymbol{\theta})^T D(\hat{\boldsymbol{\theta}}) (\mathbf{u}-\boldsymbol{\theta}) = c,$$

где $D(\hat{\boldsymbol{\theta}})$ — матрица вторых моментов оценок $\hat{\theta}$ с элементами

$$d_{ij} = E(\hat{\theta}_i \hat{\theta}_j) - E(\hat{\theta}_i) E(\hat{\theta}_j).$$

Знаки равенства в (7.48) имеют место тогда, когда оценки совместно достаточны для $\boldsymbol{\theta}$ [12].

7.4.4. Теорема Гаусса—Маркова. В тех случаях, когда достаточные статистики не существуют, наша цель должна быть более скромной. Рассмотрим класс *несмещенных оценок*, которые являются *линейными функциями* результатов наблюдений.

Пусть имеется выборка размером n наблюдений $Y_1, \ldots, Y_n$ с ф.п.в. $f(\mathbf{Y}|\boldsymbol{\theta})$, зависящей от k неизвестных параметров $\boldsymbol{\theta}$, и с матрицей вторых моментов D. Необходимое условие существования несмещенных линейных оценок $\hat{\boldsymbol{\theta}}$ состоит в том, что

$$E(\mathbf{Y}) = A\boldsymbol{\theta},$$

где A — $n \times k$-матрица констант. Тогда оценки задаются уравнением (7.23), с $W = D^{-1}$:

$$\hat{\boldsymbol{\theta}} = (A^T D^{-1} A)^{-1} A^T D^{-1} \mathbf{Y}. \tag{7.49}$$

Теорема Гаусса—Маркова утверждает, что среди оценок рассматриваемого класса оценки (7.49) метода наименьших квадратов для линейного случая имеют минимальную дисперсию.

Предположим, что $\mathbf{t} = \underset{\sim}{L}\mathbf{Y}$ — несмещенная оценка $\boldsymbol{\theta}$. Мы покажем, что $\underset{\sim}{D}(t_j)$, $j = 1, \ldots, k$ минимальна, когда

$$\underset{\sim}{L} = (\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1}. \tag{7.50}$$

Матрица вторых моментов для $\mathbf{t}$ равна:

$$\underset{\sim}{D}(\mathbf{t}) = E[(\underset{\sim}{L}\mathbf{Y} - \boldsymbol{\theta})(\underset{\sim}{L}\mathbf{Y} - \boldsymbol{\theta})^T] = E[\underset{\sim}{L}(\mathbf{Y} - \underset{\sim}{A}\boldsymbol{\theta})(\mathbf{Y} - \underset{\sim}{A}\boldsymbol{\theta})^T \underset{\sim}{L}^T] = \sigma^2 \underset{\sim}{L}\underset{\sim}{D}\underset{\sim}{L}^T. \tag{7.51}$$

Поскольку $\mathbf{t}$ — несмещенная оценка $\boldsymbol{\theta}$, то

$$E(\mathbf{t}) = E(\underset{\sim}{L}\mathbf{Y}) = \underset{\sim}{L}\underset{\sim}{A}\boldsymbol{\theta} = \boldsymbol{\theta}$$

для всех $\boldsymbol{\theta}$. Тогда (7.51) может быть записано:

$$\sigma^2 \underset{\sim}{L}\underset{\sim}{D}\underset{\sim}{L}^T = \sigma^2 \{[(\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1}]\underset{\sim}{D}[(\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1}]^T +$$
$$+[\underset{\sim}{L} - (\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1}]\underset{\sim}{D}[\underset{\sim}{L} - (\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1}]^T\}.$$

Это легко проверить, оценив оба члена правой части равенства. Каждый из этих членов имеет форму $\underset{\sim}{C}\underset{\sim}{C}^T$ и поэтому имеет неотрицательные диагональные элементы. Поскольку только второй член является функцией $\underset{\sim}{L}$, диагональные элементы $\underset{\sim}{L}\underset{\sim}{D}\underset{\sim}{L}^T$ минимальны, когда второй член равен нулю, т. е. когда выполняется равенство (7.50). Это доказывает теорему.

Важно отметить, что не делались никакие предположения о распределении наблюдений $\mathbf{Y}$, кроме того что их среднее значение есть линейная функция параметров. Оптимальные свойства оценок метода наименьших квадратов, а именно минимальная дисперсия и несмещенность, следуют прямо из линейности задачи. Кроме того, эти свойства имеют место при любом числе наблюдений.

7.4.5. Асимптотическая эффективность. Покажем, что оценка $\hat{\boldsymbol{\theta}}$ максимального правдоподобия асимптотически эффективна.

Поскольку $\hat{\boldsymbol{\theta}}$ есть состоятельная оценка $\boldsymbol{\theta}$ (см. п. 7.2.3), можем разложить ф.п.в. наблюдений $\mathbf{X}$ вблизи $\hat{\boldsymbol{\theta}}$:

$$f(\mathbf{X}, \boldsymbol{\theta}) = \exp\left[\ln f(\mathbf{X}, \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 \frac{\partial^2 \ln f(\mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2}\right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} + O(n^{-3}). \tag{7.52}$$

Заметим, что вторая производная сходится к константе:

$$\left.\frac{\partial^2 \ln f}{\partial \boldsymbol{\theta}^2}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \xrightarrow[n \to \infty]{} E\left(\frac{\partial^2 \ln f}{\partial \boldsymbol{\theta}^2}\right) = -I_{\mathbf{X}}.$$

Таким образом, ф.п.в. (7.52) имеет экспоненциальную форму (5.8) и из этого следует, что в асимптотике $\hat{\boldsymbol{\theta}}$ является достаточной статистикой для $\boldsymbol{\theta}$. В соответствии с результатами п. 7.4.2 оценка $\hat{\boldsymbol{\theta}}$ эффективна, а из (7.52) следует, что $\hat{\boldsymbol{\theta}}$ распределено по нормальному закону относительно $\boldsymbol{\theta}$ [12].

Соответственно из единственности эффективных несмещенных оценок* следует, что все асимптотически эффективные оценки кор-

* Поскольку рассматриваемые нами оценки состоятельны, то при неограниченном объеме выборки они ведут себя как несмещенные оценки. Доказательство свойства единственности может быть найдено в [12].

релированы, в частности, что все они являются функциями оценки максимального правдоподобия.

Отметим, что это свойство оценок максимального правдоподобия — опять-таки является асимптотическим свойством.

§ 7.5. БЕЙЕСОВСКИЙ ПОДХОД

7.5.1. Выбор априорной плотности. В гл. 6 мы уже обсуждали интерпретацию субъективных априорных плотностей как отражение знания наблюдателя о неизвестных параметрах. Когда наблюдатель выбирает априорную плотность, то сначала он выделяет некоторое семейство распределений с соответствующим диапазоном возможных форм, а затем из этого семейства отбирает такое распределение, которое наилучшим образом отражает его мнение о неизвестном параметре.

Например, если наблюдатель знает, что интервал возможных значений параметра (0,1), то, например, бета-распределения обладают широким диапазоном возможных форм (см. рис. 4.8). Для $a = b = 0$ наблюдатель получает однородное распределение, при $a = b = 10$ распределение имеет резкий пик вблизи $\theta \simeq 1/2$, тогда как для $a = b = -1/2$ плотность имеет большие значения вблизи $\theta = 0$ и $\theta = 1$, отражая ту точку зрения, что параметр может быть либо 0 либо 1, но маловероятно, что он равен 1/2. Когда интервал возможных значений — от нуля до бесконечности, удобным семейством являются гамма-распределения (4.38) (см. рис. 4.9).

Если же область возможных значений соответствует всей реальной оси, тогда нормальные распределения $N(\mu, \tau^2)$ дают соответствующий набор априорных плотностей. Наиболее вероятное значение параметра θ равно μ, а τ^{-2} отражает точность знания наблюдателя.

Возникает вопрос, как поступает наблюдатель при полном отсутствии знания о неизвестном параметре. В случае, когда априорные распределения соответствуют нормальному закону, то плотность вероятности стремится к $p(\theta)d\theta = d\theta$, $-\infty < \theta < \infty$ при $\tau^2 \to \infty$. Если параметром служит, например, дисперсия σ^2, то наблюдатель знает совершенно точно только одно, а именно, что $\sigma^2 > 0$. Однако, если он попытается отразить это знание посредством однородного распределения по σ^2 в некотором интервале значений, скажем, $(0, A^2)$, то антибейесовцами может быть высказана критика в том смысле, что другой физик, который тоже ничего не знает, кроме области значений, но который хочет оценить σ, выбрал бы в качестве своей априорной плотности однородное распределение по σ в интервале $(0, A)$. Эти два распределения несовместимы и приведут к различным апостериорным плотностям для σ^2, даже если будут они исходить из одних и тех же данных [31, 32].

Джеффрис предлагает выражать незнание о θ с некоторым ограниченным интервалом значений посредством априорной плотности, однородной по $\ln \theta$. Он доказывает, что такие плотности удовлет-

воряют различным условиям и, в частности, инвариантны к измерениям масштаба и степени оцениваемого параметра [33]. Линдлей указывает, что однородное распределение разумно использовать в качестве априорной плотности для параметров положения распределения, таких, как среднее μ. [Параметр θ называется *параметром положения* X, если ф.п.в. $f(X \mid \theta)$ является функцией только $(X - \theta)$.] Затем он доказывает, что когда параметром θ служит дисперсия, то $\varphi = \ln \theta$ является параметром положения для преобразованной переменной $Z = \ln (X - \mu)^2$, когда X распределено по нормальному закону. Поэтому в случае полного незнания кажется разумным использовать в качестве априорного распределения однородное распределение по $\ln \theta$ [34].

Желаемой характеристикой априорной плотности является то, что апостериорная плотность должна принадлежать тому же семейству. О такой априорной плотности говорят как о *замкнутой для* выборок различного объема. Предположим, имеются наблюдения X с ф.п.в. в $f(\mathbf{X} \mid \theta)$.

Пусть $\xi(\theta, \alpha)$ — априорная плотность θ, где α указывает на определенный член семейства ξ. Тогда плотность замкнута для выборок различного объема, если апостериорная плотность может быть записана как $\xi(\theta; \beta)$, где $\beta \equiv \beta(\alpha, X)$, т. е.

$$\xi(\theta; \beta) = \frac{\left[\prod_{i=1}^{n} f(\mathbf{X}_i \mid \theta)\right] \xi(\theta, \alpha)}{\int \left[\prod_{i=1}^{n} f(\mathbf{X}_i \mid \theta)\right] \xi(\theta, \alpha)\, d\theta}.$$

В этом случае изменение знания в результате наблюдений $\mathbf{X}$ сводится к замене индекса α на $\beta(\alpha, \mathbf{X})$.

Любое дискретное априорное распределение замкнуто для выборок различного объема. Можно показать, что для непрерывных распределений замкнутое семейство существует, если только ф.п.в. $f(X \mid \theta)$ имеет экспоненциальную форму. Тогда семейство распределений ξ имеет форму

$$\xi(\theta) \sim G(\theta)^{\alpha} \exp\{\beta A(\theta)\},$$

где α, β не зависят от θ, а A, G — произвольные функции θ.

Далее мы проиллюстрируем применение бейесовского подхода к оценке параметров нормального распределения, если имеется выборка наблюдений $X_1, \ldots, X_n$ из нормального распределения $N(\mu, \sigma^2)$.

7.5.2. Заключение о среднем, когда известна дисперсия. Пусть $\mathbf{X}$ распределены в соответствии с законом $N(\theta, \sigma^2)$, где θ неизвестно, а σ^2 известно.

Тогда апостериорная плотность θ выражается через наблюдения $\mathbf{X}$ и априорную плотность $\pi(\theta)$:

$$\pi(\theta \mid \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\sum_{i=1}^{n} \frac{(X_i-\theta)^2}{2\sigma^2}\right] \pi(\theta). \tag{7.53}$$

Нетрудно видеть, что если выбрать $\pi(\theta)$ в виде $N(\mu_0, \sigma_0^2)$, то $\pi(\theta \mid \mathbf{X})$ также нормально.

Опуская члены, не зависящие от θ, получаем вместо (7.53):

$$\pi(\theta \mid \mathbf{X}) \sim \exp\left[-\frac{1}{2}\frac{(\overline{X}-\theta)^2 n}{\sigma^2} - \frac{1}{2}\frac{(\theta-\mu_0)^2}{\sigma_0^2}\right] \sim$$

$$\sim \exp\left[-\frac{1}{2}\theta^2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)+\theta\left(\frac{n\overline{X}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right]. \tag{7.54}$$

Таким образом, апостериорная плотность θ нормальна, $N(\nu, \tau^2)$ со средним

$$\nu = \left(\frac{n\overline{X}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) \Big/ \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right) \tag{7.55}$$

и дисперсией

$$\tau^2 = (n/\sigma^2 + 1/\sigma_0^2)^{-1}.$$

Таким образом, среднее апостериорной плотности θ равно взвешенному среднему априорного среднего μ_0 и наблюдаемого среднего $\overline{X}$, причем веса пропорциональны обратным значениям дисперсий (точности). Точность τ^{-2} апостериорной плотности равна сумме точностей априорной плотности и наблюдаемого среднего.

Апостериорная плотность $\pi(\theta \mid \mathbf{X})$ суммирует все знание о θ, содержащееся в наблюдениях $\mathbf{X}$. Однако необходимо помнить, что бейесовская интерпретация ν и τ^2 отличается от классической интерпретации.

С точки зрения теории решений (см. гл. 6), при заданной параболической функции потерь оптимальным значением для θ является среднее апостериорной плотности (7.55):

$$\theta_{\text{опт}} = \nu.$$

Случаю полного отсутствия знания соответствует $\sigma_0 \to \infty$. Тогда, как и следовало ожидать, $\theta_{\text{опт}}$ становится равным $\overline{X}$.

Поскольку апостериорная плотность равна произведению функции правдоподобия и априорной плотности, вид априорной плотности в областях, далеких от $\overline{X}$, не имеет практического значения. Поэтому, когда наблюдатель располагает малым априорным знанием, любая разумно гладкая априорная плотность приведет к апостериорным плотностям, близким к плотности нормального распределения (7.54).

7.5.3. Заключение о дисперсии, когда известно среднее. Предположим, наблюдения $\mathbf{X}$ распределены по закону $N(\mu, \theta)$, для которого известно только μ. Обозначив $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$, можно записать правдоподобие для наблюдений:

$$L(\mathbf{X}, \theta) = (2\pi\theta)^{-n/2} \exp[-n\hat{\sigma}^2/2\theta]. \tag{7.56}$$

При выборе априорной плотности необходимо выбрать семейство той же формы, что и (7.56), например, параметризацию

$$\pi(\theta) \sim \theta^{-\frac{1}{2}\nu_0 - 1} \exp[-\nu_0 \sigma_0^2/\theta]. \tag{7.57}$$

Эта плотность имеет среднее $E(\theta) = \sigma_0^2\nu_0/(\nu_0 - 2)$ и наиболее вероятное значение $\sigma_0^2\nu_0/(\nu_0 + 2)$. Дисперсия плотности равна

$$D(\theta) = \frac{2\sigma_0^4\,\nu_0^2}{(\nu_0-2)^2(\nu_0-4)} \text{ для } \nu > 4.$$

Можно смещать распределение, варьируя σ_0^2 и уменьшать дисперсию, увеличивая ν_0, что обеспечивает определенный диапазон возможных форм для априорной плотности (рис. 7.4).

Плотность (7.57) имеет другое интересное свойство, а именно $\varphi = \nu_0\sigma_0^2/\theta$ распределено как $\chi^2(\nu_0)$.

Отсутствие знания соответствует случаю $\nu_0 \to 0$. Тогда можно показать, что уравнение (7.57) соответствует однородному распределению $1/\theta$ по всей реальной оси.

Рассмотрим апостериорную плотность

$$\pi(\theta \mid \mathbf{X}) \sim L(\mathbf{X}, \theta)\,\pi(\theta) \sim \theta^{-\frac{1}{2}(\nu_0+n)-1} \exp\left[-\frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2}{\theta}\right]. \quad (7.58)$$

Очевидно, уравнение (7.58) имеет ту же форму, что и априорная плотность (7.57), с заменой σ_0^2

$$\sigma_1^2 = \frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2}{\nu_0 + n}$$

—взвешенным средним априорной и выборочной дисперсий, и заменой ν_0: $\nu_1 = \nu_0 + n$. Переменная $\nu_1\sigma_1^2/\theta$ распределена как $\chi^2(\nu_0 + n)$. Апостериорное среднее, дающее оптимальное значение θ для квадратичной функции потерь, равно

$$E(\theta \mid \mathbf{X}) = \frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2}{\nu_0 + n - 2}.$$

Соответственно апостериорная дисперсия равна

$$D(\theta \mid \mathbf{X}) = \frac{2(\nu_0\sigma_0^2 + n\hat{\sigma}^2)^2}{(\nu_0+n-2)^2(\nu_0+n-4)}.$$

Коэффициент отклонения, определенный как отношение $\sqrt{\text{дисперсия}}$ к среднему, равен $\sqrt{2/(\nu_0+n-4)}$. Это показывает, что с ростом числа наблюдений стандартная погрешность становится малой по сравнению со средним. По мере роста n апостериорная плотность все более не зависит от априорной плотности.

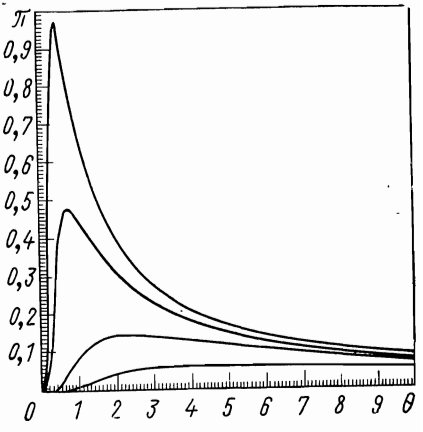


Рис. 7.4. Примеры априорных плотностей вида (7.57) (ненормированных)

В случае неопределенного априорного знания, представленного однородным распределением $(1/\theta)$, апостериорное распределение таково, что $n\hat{\sigma}^2/\theta$ распределено как $\chi^2(n)$.

Необходимо отметить, что хотя это последнее утверждение выглядит очень похожим на результат, полученный классическим способом, в этом случае переменной является величина $1/\theta$, тогда как в классическом случае σ^2.

7.5.4. Заключение о среднем и дисперсии. Рассмотрим случай, когда наблюдения $\mathbf{X}$ распределены по закону $N(\theta_1, \theta_2)$, где θ_1 и θ_2 оба неизвестны.

Необходимо ввести совместное априорное распределение для θ_1 и θ_2. Предположим, что θ_1 и θ_2 независимы, т. е. степень веры наблюдателя в параметр θ_1 не изменилась бы, если бы параметру θ_2 было сообщено некоторое значение. В случае неопределенного априорного знания об обоих параметрах мы считаем θ_1 и $\ln \theta_2$ однородно распределенными в интервале $(-\infty, \infty)$. Как указывалось выше, такое распределение по θ_2 разумным образом отражает полное незнание о параметре.

Тогда правдоподобие наблюдений равно:

$$L(\mathbf{X} \mid \theta_1 \theta_2) = (2\pi\theta_2)^{-n/2} \exp\left[-\sum_{i=1}^{n} (X_i - \theta_1)^2/2\theta_2\right],$$

а совместная апостериорная плотность θ_1 и θ_2 равна

$$\pi(\theta_1, \theta_2 \mid \mathbf{X}) \sim L(\mathbf{X} \mid \theta_1, \theta_2)\,\pi(\theta_1 \theta_2) \sim \theta_2^{-\frac{1}{2}(n+2)} \times$$

$$\times \exp\left[-\sum_{i=1}^{n} (X_i - \theta_1)^2/2\theta_2\right].$$

Мы можем написать:

$$\sum_{i=1}^{n} (X_i - \theta_1)^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + n(\overline{X} - \theta_1)^2 = (n-1)s^2 + n(\overline{X} - \theta_1)^2.$$

Поэтому апостериорная плотность может быть записана:

$$\pi(\theta_1, \theta_2 \mid \mathbf{X}) \sim \theta_2^{-\frac{1}{2}(n+2)} \exp\{-[(n-1)s^2 + n(\mathbf{X} - \theta_1)^2]/2\theta_2\}. \quad (7.59)$$

Интегрируя (7.59) по θ_2, получаем апостериорную плотность:

$$\pi_1(\theta_1 \mid \mathbf{X}) \sim \left\{1 + \frac{n(\overline{X} - \theta)^2}{(n-1)s^2}\right\}^{-n/2}.$$

Следовательно, $t = \sqrt{n}\,(\overline{X} - \theta_1)/s$ имеет t-распределение с $(n-1)$ степенями свободы*.

Подобно этому мы можем получить апостериорную плотность для θ_2

$$\pi_2(\theta_2 \mid \mathbf{X}) \sim \theta_2^{-\frac{1}{2}(n-1)-1} \exp\left[-\frac{(n-1)s^2}{2\theta_2}\right].$$

Таким образом, величина $[(n-1)s^2]/\theta_2$ имеет χ^2-распределение с $(n-1)$ степенями свободы.

7.5.5. Заключение. Таким образом, если дисперсия распределения известна, то величина $\sqrt{n}\,(\theta_1 - \overline{X})/\sigma$ распределена по закону $N(0, 1)$. Если дисперсия неизвестна, величина $\sqrt{n}\,(\theta_1 - \overline{X})/s$ имеет t-распределение с $(n-1)$ степенями свободы. Большая ширина t-распределения по сравнению с нормальным отражает разницу в знании, предшествующем получению наблюдений. Подобно этому, если сред-

* Свойства t-распределения изложены в п. 4.2.4.

нее известно, апостериорная плотность дисперсии θ_2 такова, что $\sum_{i=1}^{n} (X_i - \mu)^2/\theta_2$ распределена по закону χ^2 с n степенями свободы. Если среднее неизвестно, эта величина имеет χ^2-распределение с $(n-1)$ степенями свободы. Здесь уменьшение в априорном знании отражено в уменьшении числа степеней свободы от n до $n-1$.

Обратимся еще раз к совместной апостериорной плотности (7.59). Условное распределение θ_1 при фиксированном θ_2 имеет плотность

$$p(\theta_1|\,\theta_2, \mathbf{X}) \sim \theta_2^{-\frac{1}{2}} \exp\left[-\frac{n(\bar{X}-\theta_1)^2}{2\theta_2}\right],$$

т. е. $N[\bar{X}, (\theta_2/n)]$. Таким образом, параметры θ_1 и θ_2 не независимы, что противоречит априорному предположению.

Чем больше θ_2, тем больше ширина распределения по θ_1 и тем меньше точность нашего знания о θ_1. Каждое наблюдение X_i имеет дисперсию θ_2, и чем больше θ_2, тем больше рассеяние вблизи μ и тем меньше информации о μ содержит каждое наблюдение.

ГЛАВА 8

ОЦЕНКА ПАРАМЕТРА ФИКСИРОВАННЫМ ЗНАЧЕНИЕМ НА ПРАКТИКЕ

В § 8.1 этой главы мы обсудим различные требования, которым должна отвечать хорошая оценка. Многие из них уже обсуждались в гл. 7; другие — по своему характеру чисто «практические» требования — впервые будут рассмотрены здесь. В § 8.2—8.4 мы вновь возвратимся к обычным оценкам с целью более детального обсуждения. Поэтому некоторое повторение гл. 7 неизбежно. В то же время мы будем ссылаться без доказательства на свойства состоятельности, несмещенности и эффективности обычных оценок (несмотря на то, что читатель мог опустить эти части гл. 7).

В остальных параграфах рассматриваются различные стороны оценки параметров точкой в реальной жизни.

§ 8.1. ВЫБОР МЕТОДА ОЦЕНКИ

При выборе хорошего метода оценки встречаются противоречивые требования. Чтобы найти компромисс, необходимо вначале установить порядок важности этих требований. В п. 8.1.1 предлагается соответствующий перечень, составленный на основе важности того или иного требования с точки зрения его статистических свойств. Иногда приходится устанавливать другой порядок, взяв за основу, например, стоимость. В последующих параграфах мы обсудим, как и когда надо искать компромисс между различными требованиями.

8.1.1. Желаемые свойства оценок. 1. Состоятельность. Главное требование — это то, что оценка должна сходиться к истинному значению параметра с ростом числа наблюдений. Практически не бывает такой ситуации, когда экспериментатор выбрал бы несостоятельную оценку, заведомо зная об этом. Например, если было бы установлено, что оценка асимптотически смещена, экспериментатор захотел бы разработать некоторый метод, чтобы избавиться от смещения. Если все-таки он решил использовать несостоятельную оценку, то он наложил бы некоторое минимальное требование на верхнюю границу степени несостоятельности.

2. Минимальные потери информации. Когда оценка суммирует результаты эксперимента в виде одного числа, то любому, кто будет использовать ее, важно знать, что никакая

другая оценка не могла бы содержать больше информации об искомом параметре.

3. Минимальная дисперсия (эффективная оценка). Чем меньше дисперсия оценки, тем с большей определенностью сторонник классической статистики может сказать, что оценка находится вблизи истинного значения параметра (если пренебречь смещением).

4. Минимальные затраты (минимальная стоимость). С точки зрения теории решений это означает, что в качестве искомого значения параметра должно быть выбрано такое, которое приводит к минимальной средней потере (стоимости) в случае, если принято неверное решение. Таким образом, это требование имеет такую же относительную важность, как и требование минимальной дисперсии, только оно возникает с позиций совершенно другого подхода.

Все названные выше свойства предполагают знание распределения генеральной совокупности данных, когда такое знание отсутствует, или основано на ненадежных предположениях, интерес представляет следующее очень важное свойство.

5. Устойчивость. Оценка не должна зависеть от вида распределения или должна быть нечувствительной к отклонениям от предполагаемого распределения. В § 8.7 мы обсудим некоторые частные виды устойчивых оценок. Как правило, количество информации, содержащейся в таких оценках, уменьшается, поскольку при их получении игнорируется форма распределения.

6. Простота. Когда физик знакомится со статьей, в которой приводится величина оцениваемого параметра, он обычно предполагает, что оценка не смещена, распределена по нормальному закону и некоррелирована с другими оценками. Поэтому весьма желательно, чтобы оценки обладали этими простыми свойствами.

7. Требование минимальности времени на ЭВМ. Хотя это требование и не носит фундаментального характера, оно может привести к выбору менее эффективной оценки в связи с недостатком времени на ЭВМ либо по причинам экономии.

8. Минимальные затраты времени физика. На практике это требование зачастую рассматривается как наиболее важное. Цель этой книги, в частности, состоит в том, чтобы показать, что физику иногда полезно «потерять» некоторое количество времени на выбор хорошей оценки.

8.1.2. Компромисс между различными статистическими свойствами. Приведенный выше порядок желаемых свойств отражает ту нашу точку зрения, что статистические свойства 1—5 более важны, чем другие. Среди названных статистических свойств мы считаем требование минимальности потерь информации более важным, чем требование минимальной дисперсии. То обстоятельство, что может быть выбор между свойствами 2 и 3, следует из факта, что хотя дисперсия и ограничена снизу величиной, обратной информации, эта

нижняя граница не всегда достижима, т. е. могут быть такие две оценки t_1 и t_2 параметра θ с $I_{t_1}(\theta) < I_{t_2}(\theta)$, но с $D_{t_1} < D_{t_2}$.

В таком случае мы рекомендуем выбирать оценку t_2. Это отражает нашу надежду на то, что бóльшее количество информации, содержащейся в t_2, будет более полезным, когда впоследствии будет приниматься решение о параметре θ. Подобная информация следует из нашего знания распределения оценки t_2.

В то же время, если бы решение основывалось только на анализируемых экспериментальных данных, тогда бы оценка t_1 с минимальной дисперсией была более предпочтительной. Или если подходить с позиций теории решений, было бы выбрано такое значение, для которого функция потерь $L(t, \theta)$ в среднем была бы минимальна.

Интересно отметить, что если $L(t, \theta)$ квадратично по t с минимумом при $t = \theta$, разумный выбор t при фиксированном θ находится из минимума:

$$E[(t - \theta)^2] = D(t) - [E(t - \theta)]^2.$$

Это эквивалентно минимизации дисперсии только в случае несмещенных оценок. Ранее мы показали, что требование минимума потерь информации легко удовлетворяется только в двух случаях: 1) при использовании достаточных статистик, когда они существуют (см. п. 8.3.2); 2) в асимптотическом пределе, когда оценка максимального правдоподобия оптимальна.

Во всех других случаях не существует никаких общих методов нахождения оптимальной оценки. Тогда наиболее разумным шагом был бы отказ от стремления представить результат эксперимента в виде одного числа, т. е. нужно задавать форму функции правдоподобия по крайней мере в окрестности ее максимума (минимума).

8.1.3. Способы достижения простоты. Часто приходится жертвовать некоторым количеством информации, чтобы удовлетворить требованиям простоты (несмещенность, некоррелированность, нормальность).

Оценки нескольких параметров можно сделать некоррелированным путем диагонализации матрицы вторых моментов и нахождения соответствующих линейных комбинаций параметров. Но новые параметры часто не имеют физического смысла.

Оценки могут быть сделаны несмещенными различными методами. С одним из них мы встретимся ниже в примере 3, общее же обсуждение мы отложим до § 8.6.

Когда существуют достаточные статистики, то, очевидно, их следует использовать для оценки параметра, поскольку с их помощью оценка производится оптимально (см. п. 7.3.2.).

Большинство обычных оценок асимптотически несмещенны и распределены по нормальному закону. При этом возникает лишь вопрос, насколько хорошо асимптотическое приближение в некотором конкретном эксперименте. На такой вопрос не может быть дан ответ, если не найден вид точного распределения оценок, получаемых из выборки конечного объема. Однако некоторые указания на

справедливость асимптотического приближения могут быть найдены одним из следующих методов:

а) проверить, что логарифм функции правдоподобия имеет вид параболы;

б) если существуют две асимптотически эффективные оценки, проверить, что результаты статистически не отличаются. Например, такими двумя оценками могут быть оценки метода наименьших квадратов, применимого к двум различным разбиениям одних и тех же данных;

в) вычислить первые несколько моментов выборочного распределения с невысокой точностью по степеням n^{-1} и проверить, что смещение, асимметрия и эксцесс малы;

г) промоделировать поведение оценок методом Монте-Карло. Однако моделирование эксперимента при большом числе повторений требует больших затрат времени на ЭВМ.

Иногда оценка может быть сделана более простой или даже оптимальной заменой переменных. При этом надо различать:

а) замену оценки, т. е. замену переменной, зависящей от наблюдений, $\hat{\theta}_2 = g(\hat{\theta}_1, X)$;

б) от замены переменной (или параметра теории), не зависящей от наблюдений, $\theta_2 = g(\theta_1)$.

Замена а изменяет дисперсию и для нее не существует какого-либо общего метода конструирования нормально распределенной оценки.

В случае б можно, не меняя выбранного метода оценки (например, метод максимального правдоподобия), получить нормально распределенные оценки новых параметров (в терминах новых переменных). Однако обычно невозможно исключить смещение и сделать оценку распределенной по нормальному закону такой заменой. Ниже это показано в примере 3 [2].

Наилучший метод может быть таким: использовать первоначально выбранный метод оценки для нахождения оценки $\hat{\theta}$ параметра θ, затем оценить первые четыре момента $\hat{\theta}$ и, наконец, использовать общее семейство распределений (например, семейство Джонсона, см. п.4.3.2) с тем, чтобы получить лучшее по сравнению с асимптотически нормальным приближение к истинному распределению оценки $\hat{\theta}$.

Примеры: 1. Если неизвестным параметром является комплексное число z, то лучше параметризовать его в виде $\mathrm{Re}\, z$ и $\mathrm{Im} z$, а не через $\rho = |z|$ и $\varphi = \arg z$, поскольку φ становится плохо определяемой, когда ρ мало. Иными словами, матрица информации становится сингулярной для $\rho = 0$.

2. Пусть заданы две коррелированные оценки X и Y с коэффициентом корреляции ρ и дисперсиями σ_X^2 и σ_Y^2. Тогда оценки $u = X + \alpha Y$; $v = X - \alpha Y$ не коррелированы, если $\alpha = \sigma_X/\sigma_Y$. Действительно, смешанный второй момент оценок u и v равен нулю:

$$D(u, v) = \sigma_X^2 - \alpha^2 \sigma_Y^2 + (\alpha - \alpha)\rho\sigma_X \sigma_Y = 0.$$

3. Оценка параметра по гистограмме с помощью метода максимального правдоподобия. Предположим, что гистограмма содержит k ячеек, в i-ю ячейку попадает n_i число событий и суммарное число событий в гистограмме N. Вероятность попадания событий в i-ю ячейку $p_i(\theta)$, где θ — неизвестный параметр. Функция правдоподобия имеет следующий вид:

$$L(n \mid \theta) = N! \prod_{i=1}^{k} \left(p_i^{n_i}/n_i!\right).$$

Халдейн и Смит [35] показали, что смещение, дисперсия и асимметрия оценки максимального правдоподобия $\hat{\theta}$ соответственно равны:

$$b(\hat{\theta}) = -\frac{1}{2} N^{-1} A_1^{-2} B_1 + O(N^{-2});$$

$$D(\hat{\theta}) = N^{-1} A_1^{-1} + O(N^{-2});$$

$$\gamma_1(\hat{\theta}) = N^{-\frac{1}{2}} A_1^{-3/2} (A_2 - 3B_1) + O(N^{-2}),$$

при этом

$$A_1 = \sum_{i=1}^{k} \frac{1}{p_i} \left(\frac{\partial p_i}{\partial \theta}\right)^2;$$

$$A_2 = \sum_{i=1}^{k} \frac{1}{p_i^2} \left(\frac{\partial p_i}{\partial \theta}\right)^3;$$

$$B_1 = \sum_{i=1}^{k} \frac{1}{p_i} \left(\frac{\partial p_i}{\partial \theta}\right) \left(\frac{\partial^2 p_i}{\partial \theta^2}\right).$$

Отметим, что A_1 есть информация, определенная в гл. 5.

Чтобы избавиться от смещения, нужно найти такое преобразование $\tau = \tau(\theta)$, для которого $B_1(\tau) = 0$.

Учитывая, что

$$\frac{\partial p_i}{\partial \tau} = \frac{\partial p_i}{\partial \theta} \frac{\partial \theta}{\partial \tau};$$

$$\frac{\partial^2 p_i}{\partial \tau^2} = \frac{\partial^2 p_i}{\partial \theta^2} \left(\frac{\partial \theta}{\partial \tau}\right)^2 + \left(\frac{\partial p_i}{\partial \theta}\right) \left(\frac{\partial^2 \theta}{\partial \tau^2}\right),$$

получаем

$$B_1(\tau) = B_1 \left(\frac{\partial \theta}{\partial \tau}\right)^3 + A_1 \frac{\partial \theta}{\partial \tau} \frac{\partial^2 \theta}{\partial \tau^2} = 0.$$

Величины B_1 и A_1 в последнем уравнении рассчитываются при значении $\tau = \theta$, что соответствует $\partial\theta/\partial\tau = 1$ и $\partial^2\theta/\partial\tau^2 = 0$. Полученное дифференциальное уравнение может быть решено относительно параметра τ, оценка которого находится методом максимального правдоподобия и не имеет смещения. Однако, хотя смещение и равно нулю, асимметрия

$$\gamma_1(\hat{\tau}) = N^{-\frac{1}{2}} A_1^{-\frac{3}{2}} A_2,$$

причем A_2 в общем случае не равно нулю. Следовательно, оценка $\hat{\tau}$ не может быть сделана распределенной по нормальному закону для выборки данных конечного объема. Заметим, что

$$A_1(\tau)^{-3/2} A_2(\tau) = A_1^{-3/2} A_2.$$

8.1.4. Соображения экономии. Обычно под экономией понимается возможность отыскания оценки с минимальными затратами времени. Как правило, оптимальный метод оценки представляет собой некоторую итеративную процедуру, требующую больших затрат времени на ЭВМ. Поэтому за быстроту нахождения оценки приходится платить ценой уменьшения эффективности. Сейчас мы рассмотрим три подхода к проблеме оценки, представляющих собой некоторый компромисс между эффективностью и стоимостью.

А. Линейные методы. Линейные методы требуют минимальных затрат времени, так как их реализация обходится без итераций. Естественно, что эти методы могут быть использованы тогда, когда ожидаемые значения наблюдений являются линейными функциями параметров. Среди линейных методов нахождения несмещенных оценок наилучшим является метод наименьших квадратов (Теорема Гаусса—Маркова, см. п. 7.4.4).

Когда вид теоретического распределения неизвестен, с чем часто приходится сталкиваться в практической деятельности, то, если это возможно, нужно выбирать его из семейства экспоненциальных распределений (5.8). В этом случае нахождение оценок требует небольших вычислений, а сами оценки обладают оптимальными свойствами (см. п. 8.3.1 и 7.4.2).

Б. Оценка в два этапа. Определенная экономия времени вычислений может быть получена, если оценку проводить в два этапа:

1) оценить параметры простым, быстрым, неэффективным методом;

2) использовать эти оценки в качестве начального приближения для оптимального метода оценки.

Возможно, что этот метод потребует несколько больших затрат времени физика, ввиду того что нужно осмыслить результаты первого этапа, тем не менее он может привести к более полному пониманию ситуации (в дополнение к экономии времени на ЭВМ).

В. Оценка в три этапа. Предыдущие методы могут быть улучшены следующим образом.

1. Извлечь из данных некоторое число статистик, которые суммируют наблюдения в более компактном виде и по возможности имеют определенный физический смысл. Одним примером такого суммирования служит гистограмма, которая сводит всю совокупность наблюдений к некоторому числу ячеек в гистограмме. Другим примером может быть представление условного распределения в виде коэффициентов разложения этого распределения по полиномам Лежандра. Эти коэффициенты могут быть очень быстро найдены методом моментов (см. п. 8.2.1) и их физическая интерпретация очень наглядна.

2. Найти оценки параметров на основе такого суммирования данных. Если данные в таком «суммированном» виде имеют некий интуитивный физический смысл, то проблема оценки может быть существенно упрощена.

Таблица 8.1

Оценка для непрерывных распределений

Метод оценки	Вид распределения генеральной совокупности	Эффективность оценки	Свойство нормальности распределения оценки	Режим работы ЭВМ
Максимальное правдоподобие (см. п. 7.2.3; § 8.3)	Общий случай	Неэффективна	Распределение не по нормальному закону	Медленный, исключая случаи выбора хорошего начального приближения (см. 8.1.4)
	Достаточная статистика существует (см. п. 7.4.2) Например, распределение генеральной совокупности имеет вид нормального закона*	Можно найти эффективную и несмещенную оценку функции $g(\theta)$ параметров	В общем случае $g(\theta)$ распределена не по нормальному закону, но ее дисперсия известна	Быстрый: оценка $g(\theta)$ находится простыми способами. Обратное преобразование к θ может потребовать итераций
Наименьшие квадраты (см. п. 7.2.4; § 8.4)	Оценка для линейного случая (см. п. 8.4.1)	Неэффективна, но оптимальна среди несмещенных линейных оценок	В общем нет (но если ф. п. в. наблюдений имеет вид нормального закона, то распределение оценки классическое**)	Быстрый: линейные методы (см. п. 8.4.1; 8.4.2; 8.4.3)
	Оценка для нелинейного случая	Неоптимальна	Нет	Медленный
Моменты (см. п. 7.2.1; § 8.2)	Общий случай	Неоптимальна	Распределена не по нормальному закону, но не смещена	Быстрый, если функции ортогональны

* В случае нормальности распределения генеральной совокупности методы максимального правдоподобия и наименьших квадратов идентичны п. 8.4.4.

** Стьюдента, Фишера—Снедекора и т. д.

Таблица 8.2

Оценка для непрерывных распределений. Асимптотические пределы § 7.3*

Метод оценки	Вид распределения генеральной совокупности	Эффективность оценки	Режим работы ЭВМ
Максимальное правдоподобие (см. п. 7.2.3, § 8.3)	Общий случай	Оценки достигают границы минимальной дисперсии	Медленный**
	Достаточная статистика существует		Быстрый (см. табл. 8.1)
	Например, распределение генеральной совокупности имеет вид нормального закона		Быстрый (см. табл. 8.1)
Наименьшие квадраты (см. п. 7.2.4, § 8.4)	Линейная оценка	Оптимальна среди несмещенных линейных оценок	Быстрый — линейные методы
	Нелинейная оценка	Неоптимальна	Медленный**
Моменты (см. п. 7.2.1, § 8.2)	Общий случай	Неоптимальна	Быстрый, если функции ортогональны

* Оценка распределена по нормальному закону и не смещена в соответствии с центральной предельной теоремой.
** Все же быстрее, чем для выборки конечного объема, из-за параболической формы функции вблизи минимума.

3. Использовать полученные на втором этапе оценки в качестве начального приближения для оптимального метода оценки применительно к данным в их первоначальном виде. Как правило, этот этап считается «излишним» из-за недостатка времени и из-за непонимания того, что на первом этапе в процессе суммирования происходит некоторая потеря информации.

Когда информация, содержащаяся в первоначальных данных, не очень велика («малые статистики»), последний этап особенно важен. Кроме того, когда оценку проводят в три этапа, то метод

Таблица 8.3

Выборка конечного объема. Оценка по гистограмме п. 8.4.5*

Метод оценки	Эффективность оценки	Режим работы ЭВМ
Максимальное правдоподобие	Быстрее становится эффективной, чем другие	Зависит от задачи; обычно одинаков, если используется общая программа минимизации
Минимум χ^2 (теоретические ошибки)	—	То же, исключая случай линейной модели, когда решение быстро
Модифицированный минимум χ^2 (экспериментальные ошибки)	—	

* В общем случае о нормальности распределения оценки ничего нельзя сказать.

Таблица 8.4

Асимптотический предел. Оценка по гистограмме п. 8.4.5

Метод оценки	Эффективность оценки	Свойство нормальности распределения оценки	Режим работы ЭВМ
Максимальное правдоподобие	Граница минимальной дисперсии	Распределено по нормальному закону. Оценки полностью коррелированы с точностью до членов порядка $1/N^{3/2}$	Одинаков
Минимум χ^2 (теоретические ошибки)			
Модифицированный минимум χ^2 (экспериментальные ошибки)			

оценки, используемый на втором этапе, может быть упрощенным и приближенным.

8.1.5. Краткая сводка свойств оценок. В табл. 8.1 и 8.2 приведены свойства методов оценок применительно к непрерывным распределениям, а в табл. 8.3 и 8.4 — применительно к гистограммам. Отметим также, что гистограммирование данных приводит к потере информации по сравнению с первоначально негистограммированными данными.

§ 8.2. МЕТОД МОМЕНТОВ

Пусть задана выборка наблюдений $X_1, \ldots, X_n$, распределение генеральной совокупности которых $f(X, \theta)$. Один из методов оценки параметров θ состоит в вычислении *моментов выборки*

$$m_j' = \frac{1}{n} \sum_{i=1}^{n} X_i^j. \tag{8.1}$$

Вспомним определение моментов (см. п. 2.4.6)

$$\mu_j'(\theta) = \int X^j f(X, \theta)\, dX. \tag{8.2}$$

Можно использовать экспериментальные (выборочные) моменты (8.1) в качестве оценок теоретических (генеральной совокупности) моментов (8.2):

$$m_j' = \widehat{\mu_j'(\theta)}. \tag{8.3}$$

Решая эту систему уравнений относительно неизвестных параметров θ, мы получим оценки $\hat{\theta}$. Это есть обычный *метод моментов*, и он приводит к оценкам, которые состоятельны (см. п.7.2.1) и асимптотически не смещены (см. п. 7.3.3). Легко показать, что их дисперсия равна

$$D(m_j') = (\mu_{2j}' - \mu_j'^2)/n,$$

а смешанный второй момент [1]

$$D(m_i', m_j') = (\mu_{i+j}' - \mu_i' \mu_j')/n.$$

В некоторых случаях этот метод может быть очень прост для вычисления на ЭВМ, но в целом его оценки не столь эффективны, как оценки максимального правдоподобия, поскольку моменты выборки, как правило, не являются достаточными статистиками для моментов распределения генеральной совокупности. Однако, если число используемых моментов больше, чем число параметров, то потеря информации может быть уменьшена.

8.2.1. Ортогональные функции. Метод моментов представляет особый интерес, когда ф. п. в. наблюдений $\mathbf{X}$ может быть выражена через *ортогональные функции* $g_k(X)$:

$$f(X, \theta) = \alpha + \sum_{k=1}^{r} \theta_k g_k(X), \tag{8.4}$$

где α—нормировочная константа. Ортогональность выражается условиями:

$$\int g_j(X)\, g_k(X)\, dX = \delta_{jk}; \quad \int g_i(X)\, dX = 0.$$

Математические ожидания функций $g_k(X)$ равны:

$$E[g_k(X)] = \int g_k(X) f(X, \theta)\, dX = \int g_k(X)\, dX + \\ + \sum_{j=1}^{r} \theta_j \int g_j(X)\, g_k(X)\, dX = \theta_k.$$

В соответствии с результатами п. 7.2.1 это означает, что состоятельной и несмещенной оценкой θ_k является

$$\hat{\theta}_k = n^{-1} \sum_{i=1}^{n} g_k(X_i).$$

Согласно центральной предельной теореме (см. п. 3.3.2 и 7.3.1), $\hat{\theta}_k$ в асимптотике распределено по нормальному закону относительно θ_k с дисперсией, оценка которой может быть вычислена по формуле

$$S_k^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} [g_k(X_i) - \hat{\theta}_k]^2.$$

Можно приближенно вычислить доверительные области, если использовать тот факт, что отношение

$$\frac{\hat{\theta}_k - \theta_k}{S_k} = t_{n-1}$$

имеет (в асимптотике) распределение Стьюдента с $(n-1)$ степенями свободы.

Пример. При анализе экспериментов по измерению поляризации часто встречается форма (8.4). Например, при распаде Λ-гиперона, косинус угла вылета распадного пиона в с.ц.м. Λ-гиперона распределен в соответствии с функцией плотности

$$f(\cos\theta_\pi, P) = \frac{1}{2}(1 + \alpha P \cos\theta_\pi),$$

где α — параметр асимметрии, предполагаемый известным. Поскольку

$$\int_{-1}^{1} \cos\theta d(\cos) = 0$$

и

$$\int_{-1}^{1} \frac{1}{2}\cos^2\theta d(\cos\theta) = \frac{1}{3},$$

оценка αP равна $\alpha\hat{P} = (3/N)\Sigma\cos\theta_i$.

Дисперсия этой оценки может быть вычислена по формуле

$$D(\alpha\hat{P}) = \frac{9}{N}\left[\frac{1}{3} - \frac{(\alpha P)^2}{9}\right] = \frac{1}{N}[3-(\alpha P)^2]. \tag{8.5}$$

8.2.2. Сравнение метода максимального правдоподобия и метода моментов. Покажем, что метод моментов менее эффективен, чем метод максимального правдоподобия (м. п.).

Для ф. п. в., заданной (8.4), логарифм функции правдоподобия имеет вид

$$\ln L(\mathbf{X}\,|\,\theta) = \sum_{j=1}^{N} \ln\left[\alpha + \sum_{i=1}^{r} \theta_i g_i(X_j)\right].$$

Следовательно, оценки м. п. могут быть найдены из уравнений

$$\sum_{j=1}^{N} \frac{g_i(X_j)}{\left[\alpha + \sum\limits_{k=1}^{r} \theta_k g_k(X_j)\right]} = 0. \tag{8.6}$$

Можно сразу увидеть, что если $\sum\limits_{k=1}^{r} \theta_k g_k$ мало по сравнению с α (случай малых поляризаций), то оба метода эквивалентны, поскольку уравнение (8.6) можно записать в виде

$$\sum_{j=1}^{N} g_i(X_j)\left[\alpha - \sum_{k=1}^{r} \hat{\theta}_k g_k(X_j) + \ldots\right] = 0, \quad i = 1, \ldots, r.$$

Поэтому

$$\alpha \sum_{j=1}^{N} g_i(X_j) \simeq \sum_{k=1}^{r} \hat{\theta}_k \sum_{j=1}^{N} g_i(X_j)\, g_k(X_j) \simeq N\hat{\theta}_i.$$

Если $\sum\limits_{k}^{r}\theta_k g_k$ пренебречь нельзя, то метод моментов менее эффективен.

Рассмотрим пример, приведенный в п. 8.2.1, и положим $\alpha P = 0$. Тогда функция плотности будет равна

$$f(X, \theta) = \frac{1}{2}(1+\theta X); \quad -1 < X < 1.$$

Дисперсия оценки θ, полученной методом моментов, равна в соответствии с (8.5)

$$N^{-1}(3-\theta^2). \tag{8.7}$$

Асимптотическая дисперсия оценки м. п. равна:

$$\frac{1}{N}\left\{-E\left[\frac{\partial^2 \ln(1+\theta X)}{\partial\theta^2}\right]\right\}^{-1} = \frac{1}{N}\left\{\frac{1}{\theta^2}\left[-1+\frac{1}{2\theta}\ln\left(\frac{1+\theta}{1-\theta}\right)\right]\right\}^{-1} =$$

$$= \frac{\theta^2}{N}\left[-1+\frac{1}{2\theta}\ln\left(\frac{1+\theta}{1-\theta}\right)\right]^{-1} = \frac{1}{N}\left(3-\frac{9}{5}\theta^2+\ldots\right). \tag{8.8}$$

Сравнение уравнений (8.7) и (8.8) показывает, что дисперсия оценки м. п. меньше дисперсии оценки метода моментов.

§ 8.3. МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

В п. 7.2.3 мы назвали оценкой максимального правдоподобия $\hat{\theta}$ параметра θ такое его значение, при котором логарифм правдоподобия

$$\ln L(X \mid \theta) = \sum_{i=1}^{N} \ln f(X_i, \theta)$$

максимален для выборки объемом N независимых наблюдений $X_1, \ldots, X_N$, распределенных в соответствии с $f(\mathbf{X}, \theta)$. Если максимум лежит внутри области изменения θ, то необходимое условие его существования приводит к уравнению правдоподобия:

$$\frac{\partial \ln L(\mathbf{X} \mid \theta)}{\partial\theta} = 0. \tag{8.9}$$

Предполагается, что функция L нормирована в пространстве $\mathbf{X}$ так, что

$$\int L(\mathbf{X} \mid \theta)\, d\mathbf{X} = 1.$$

Важное свойство $L(\mathbf{X}, \theta)$ состоит в том, что этот интеграл должен не зависеть от θ.

Для r параметров θ_j получается система r уравнений правдоподобия

$$\partial \ln L(\mathbf{X} \mid \theta)/\partial\theta_j = 0;\ j = 1, \ldots, r.$$

8.3.1. Сводка свойств оценки максимального правдоподобия. Важно различать *асимптотические* свойства, которые имеют место для достаточно больших N, и *свойства выборки конечного объема.*

Оценка м. п. $\hat{\theta}$ асимптотически состоятельна: один из максимумов будет сколь угодно близок к истинному значению (см. п. 7.2.3 и работу [12]). Это означает асимптотическую несмещенность.

При очень общих условиях $\hat{\theta}$ в асимптотике распределена по нормальному закону, причем дисперсия распределения минимальна и равна граничному значению Крамера—Рао (см. п. 7.4.1)

$$D(\hat{\theta}) \xrightarrow[N\to\infty]{} \left\{ E\left[\left(\frac{\partial \ln L}{\partial \theta} \right)^2 \right] \right\}^{-1}. \tag{8.10}$$

Когда область изменения $\mathbf{X}$ не зависит от θ, ее оценку можно рассчитывать по формуле

$$\hat{D}(\hat{\theta}) = \left\{ \left(-\frac{\partial^2 \ln L}{\partial \theta^2} \right) \Big|_{\theta=\hat{\theta}} \right\}^{-1}.$$

Оценка $\sqrt{N}\ (\hat{\theta} - \theta)$ распределена по закону $N[0, I^{-1}\ (\theta)]$.

Во многих случаях асимптотический предел, когда реализуются эти оптимальные свойства, будет достигаться очень медленно. В частности, функция правдоподобия для конечного N может иметь более чем один максимум и нет способа решить, какой максимум соответствует истинному значению θ.

Для *конечного* N существует только один случай, когда оценка м. п. имеет оптимальные свойства. Это случай, когда распределение генеральной совокупности наблюдений $\mathbf{X}$ имеет экспоненциальную форму (5.8):

$$f(X \mid \theta) = \exp\ [\alpha(X)\ a(\theta) + \beta(X) + c(\theta)].$$

У таких распределений существуют достаточные статистики (см. п. 5.3.4). Ранее было показано (см. п. 7.4.2), что оценка величины $r(\theta) = -(dc/d\theta)/(da/d\theta)$ может быть найдена методом максимального правдоподобия (7.45), причем эта оценка эффективна и не смещена:

$$\hat{r} = r(\hat{\theta}) = N^{-1} \sum_{i=1}^{N} \alpha(X_i).$$

(См. также пример в п. 8.6.1.)

Важным свойством оценки м. п. является ее инвариантность: оценка м. п. $\hat{\tau}$ функции $\tau(\theta)$ равна

$$\hat{\tau} = \tau(\hat{\theta}). \tag{8.11}$$

Например, оценка м. п. θ^2 равна квадрату оценки θ.

Однако только одна функция $\tau(\theta)$ будет иметь несмещенную оценку для конечных N (см. п. 7.4.2) и поэтому некоторые функции (которые приближаются к асимптотическому пределу более медленно) могут иметь произвольно смещенные оценки.

Таким образом, вопреки сложившемуся у физиков мнению, метод м. п. необязательно дает лучшие оценки для выборок *конечного* объема. Несмотря на то что функция правдоподобия содержит всю информацию, заключенную в данных, это вовсе не означает, что метод м. п. использует эту информацию наилучшим образом.

В некоторых случаях можно показать, что оценки м. п. быстрее сходятся к асимптотическому пределу, чем некоторые другие классы асимптотически эффективных оценок. Позже мы встретимся с примером (см. п. 8.4.5), когда оценка м. п. (получаемая путем использования гистограммы) сходится быстрее, чем оценки асимптотически эффективных методов наименьших квадратов и модифицированных наименьших квадратов. Даже тогда нельзя сказать, что оценка м. п. будет лучше или даже более чувствительна для фиксированного, но малого числа наблюдений N.

В п. 7.5.2 мы ознакомились с методами расчета моментов распределения оценки м. п. при малых N, когда распределение еще ненормально. Используя оценку м. п., можно сконструировать нормально распределенную оценку, однако это достигается ценой увеличения дисперсии.

8.3.2. Пример: определение времени жизни странной частицы по ее распаду в некотором ограниченном объеме. Для определения времени жизни странной частицы, распадающейся со временем, используется N событий, характеризуемых длиной пролета l_i частиц от их точки рождения. К несчастью, из-за конечности объема, ограничивающего эксперимент, эта длина имеет верхнюю $l_{\text{верх},\,i}$ и нижнюю $l_{\text{нижн},\,i}$ границы.

В действительности измеряемой величиной является собственное время

$$t_i = l_i/\gamma_i \beta_i c,$$

где γ_i, β_i — обычные лоренцевские величины, а c — скорость света. Следовательно, границы изменения равны

$$t_{\text{верхн},\,i} = l_{\text{верхн},i}/\gamma_i \beta_i c$$

и

$$t_{\text{нижн},\,i} = l_{\text{нижн},\,i}/\gamma_i \beta_i c.$$

Функция правдоподобия для N наблюдений t_i равна:

$$L = \prod_{i=1}^{N} \frac{\tau^{-1} \exp(-t_i/\tau)}{\exp(-t_{\text{нижн},\,i}/\tau) - \exp(-t_{\text{верхн},\,i}/\tau)},$$

где τ — искомое время жизни. Тогда

$$\ln L = \sum_{i=1}^{N} \left\{ -\frac{t_i}{\tau} - \ln\left[\exp\left(-\frac{t_{\text{нижн},\,i}}{\tau}\right) - \exp\left(-\frac{t_{\text{верхн},\,i}}{\tau}\right)\right]\right\} - N \ln \tau.$$

Оценка максимального правдоподобия $\hat{\tau}$ получается в результате решения уравнения

$$\frac{\partial \ln L}{\partial \tau} = \frac{1}{\tau^2}\left\{\sum_{i=1}^{N} [t_i - g_i(\tau)] - N\tau\right\} = 0, \tag{8.12}$$

где

$$g_i(\tau) = \frac{t_{\text{нижн},\,i} \exp(-t_{\text{нижн},\,i}/\tau) - t_{\text{верхн},\,i} \exp(-t_{\text{верхн},\,i}/\tau)}{\exp(-t_{\text{нижн},\,i}/\tau) - \exp(-t_{\text{верхн},\,i}/\tau)}.$$

Уравнение (8.12) может быть решено методом Ньютона. Определим $F(\tau)$ через

$$F(\tau)=\sum_{i=1}^{N}[t_i-g_i(\tau)]-N\tau.$$

Выбирая в качестве начального приближения $\tau=\tau_0$ (например, наилучшее из ранее измеренных значений), получаем для следующего приближения

$$\tau_1=\tau_0-F(\tau_0)/F'(\tau_0).$$

Здесь $F'(\tau)=-\sum_{i=1}^{N} g_i'(\tau)-N$ и

$$g_i'(\tau)=-\frac{1}{\tau^2}\left[g_i(\tau)^2-\right.$$

$$\left.-\frac{t_{\text{нижн},\,i}^2\exp(-t_{\text{нижн},\,i}/\tau)-t_{\text{верхн},\,i}^2\exp(-t_{\text{верхн},\,i}/\tau)}{\exp(-t_{\text{нижн},\,i}/\tau)-\exp(-t_{\text{верхн},\,i}/\tau)}\right].$$

Соответственно получаем

$$\tau_1=\frac{N\tau_0+F(\tau_0)+\tau_0\sum_{i=1}^{N} g_i'(\tau_0)}{N+\sum_{i=1}^{N} g_i'(\tau_0)}.$$

Последующее приближение рассчитывается по формуле

$$\tau_{j+1}=\frac{\tau_j\,\overline{g'(\tau_j)}+\bar{t}-\overline{g(\tau_j)}}{\overline{g'(\tau_j)}+1},$$

где $\overline{g(\tau_j)}=\frac{1}{N}\sum_{i=1}^{N} g_i(\tau_j)$ и т. д.

Приближение τ_1 будет достаточно, если величина $F'(\tau_1)$ того же порядка, что и $\sqrt{F''(\tau_1)}$, или, что то же самое, величина $(\tau_0-\tau_1)$ порядка погрешности τ_1.]

Чтобы определить дисперсию, рассчитаем вторую производную логарифма функции правдоподобия:

$$\frac{\partial^2 \ln L}{\partial\tau^2}=-\frac{2}{\tau^3}F+\frac{1}{\tau^2}F'\simeq\frac{1}{\tau_0^2}\left(F'-\frac{2}{\tau_0}F\right).$$

Значение, обратное этому значению, приближенно равно асимптотической дисперсии оценки $\hat{\tau}$ параметра τ, т. е.

$$\sigma(\hat{\tau})\simeq\frac{\tau_0}{\sqrt{N\left[-1+\bar{g}'+\frac{2}{\tau_0}(\bar{t}-\bar{g})\right]}}.$$

8.3.3. Академический пример плохой оценки максимального правдоподобия. Все то, что говорилось ранее о методе м. п., исходило из предположения о независимости области наблюдений от параметра θ. Рассмотрим в качестве примера следующий экспери-

мент. Предположим, что в эксперименте регистрируются N событий X_i, распределенных однородно в интервале от 0 до θ, где θ — неизвестный параметр. Поскольку всегда $\theta \geqslant X_i$ для всех i, функция правдоподобия $L = \theta^{-N}$ имеет максимум при значении $\hat{\theta} = X_N$, где X_N — наибольшее значение, зарегистрированное в эксперименте. Меньшие значения θ недопустимы, а большие значения дали бы меньшее значение функции правдоподобия. На рис. 8.1 показаны распределения $\hat{\theta}$ для случаев $N = 1, 2, 3$. Видно, что эта оценка всегда дает результат, который слишком мал, и его можно было бы улучшить из простых общих соображений без использования какого-либо правдоподобия. Поскольку всегда $\theta_0 \geqslant X_N$, хорошей оценкой неизвестного параметра была бы функция

$$\hat{\theta}_{cs} = X_N + X_N/N.$$

Легко проверить, что эта оценка уже не смещена.

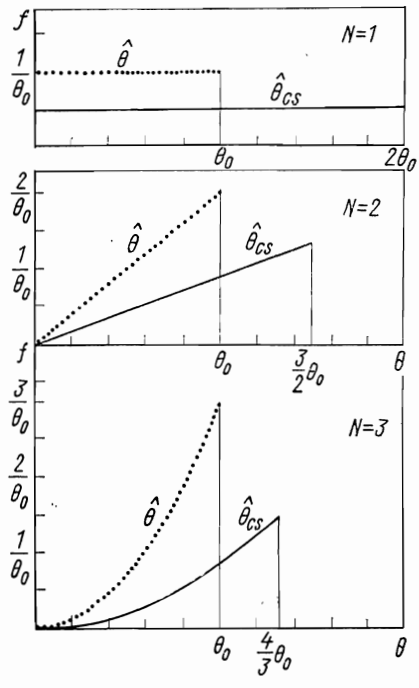


Рис. 8.1. Распределение $\hat{\theta}$ и $\hat{\theta}_{cs}$

На рис. 8.1 приведены кривые, соответствующие распределениям $\hat{\theta}_{cs}$. Среднее значение каждого из этих распределений равно θ_0. Но такое распределение все-таки нельзя считать хорошим, поскольку его пик всегда смещен вправо от истинного значения. В частности, резкий излом не исчезнет ни при каком N, так что ни $\hat{\theta}$, ни $\hat{\theta}_{cs}$ не будут асимптотически нормальными, так как область наблюдений зависит от значения θ.

Приведенный пример носит чисто академический характер, поскольку на практике всегда нужно включать аппаратурные функции разрешения в теоретическое распределение. Это приводит к размазыванию однородного распределения вблизи максимального значения и θ уже не является верхней границей распределения. В связи с этим область наблюдений не зависит от параметра. В дальнейшем этот пример рассматривается в п. 8.6.2.

8.3.4. Параметры при наличии связей. В практических задачах часто случается так, что между искомыми параметрами имеется связь, отражающая, например, некоторый физический закон. В этом параграфе мы обсуждаем связи, имеющие следующую форму:

$$\mathbf{g}(\hat{\theta}) = 0. \qquad (8.13)$$

Случаи же, для которых связи, накладываемые на функции параметров, имеют вид неравенств (например, положительность данного параметра), нами практически не рассматриваются.

Наложение связей всегда означает добавление некоторой информации и обычно ошибки параметров становятся меньше. Поэтому нужно быть осторожным, чтобы избежать добавления «ложной» информации, т. е. необходимо проверять совместимость данных со связями. Например, прежде чем положить параметр равным его теоретическому значению, необходимо проверить, нет ли зна́чимого отклонения от него. Это полезно делать даже и в том случае, если теория верна, поскольку в эксперименте могут быть погрешности и проверка выполнимости связей служит хорошим методом отыскания их. Способы таких проверок будут описаны в гл. 10.

При работе на ЭВМ самый эффективный способ использования связей состоит в такой замене переменных, чтобы уравнение (8.13) стало тривиальным. Например, если

$$g(\theta) = \theta_1 + \theta_2 - 1 = 0,$$

то следует θ_2 заменить $1 - \theta_1$ в функции правдоподобия и отыскивать ее максимум по θ_1. Таким образом, используется свойство инвариантности (8.11). Подобные методы могут применяться и в случае, когда θ ограничено простыми неравенствами $\theta' < \theta < \theta''$. После соответствующей замены переменных

$$\theta = \theta' + \frac{1}{2}(\sin\xi + 1)(\theta'' - \theta') \tag{8.14}$$

нужно искать максимум по ξ.

В случае, когда уравнения связи имеют вид

$$0 \leqslant \theta_i \leqslant 1, \quad i = 1, \ldots, n;$$

$$\sum_{i=1}^{n} \theta_i = 1,$$

существует хорошо известный рецепт для замены переменных. Для $n = 4$ он сводится к следующему:

$$\theta_1 = \xi_1;$$
$$\theta_2 = (1 - \xi_1)\,\xi_2;$$
$$\theta_3 = (1 - \xi_1)(1 - \xi_2)\,\xi_3;$$
$$\theta_4 = (1 - \xi_1)(1 - \xi_2)(1 - \xi_3),$$

где область изменения ξ_i между 0 и 1 задается уравнением (8.14).

При этом теряется первоначальная симметрия между параметрами, что иногда нежелательно.

Обычно подобные приемы трудноприменимы и, естественно, приходится пользоваться методом множителей Лагранжа. Если за-

дана функция правдоподобия $L(\mathbf{X} \mid \theta)$, то отыскивают экстремум функции

$$F = \ln L(\mathbf{X} \mid \theta) + \boldsymbol{\alpha}^T \mathbf{g}(\theta)$$

по переменным θ и $\boldsymbol{\alpha}$. Уравнение правдоподобия (8.9) заменяется системой уравнений

$$\left.\begin{aligned} \frac{\partial}{\partial \theta} \ln L(\mathbf{X}, \hat{\theta}) + \hat{\boldsymbol{\alpha}}^T \frac{\partial \mathbf{g}(\hat{\theta})}{\partial \theta} = 0; \\ \mathbf{g}(\hat{\theta}) = 0. \end{aligned}\right\} \tag{8.15}$$

Состоятельность и асимптотическая нормальность таких оценок могут быть доказаны методами, применявшимися в § 7.3.

Для того чтобы записать выражение для дисперсии $\hat{\theta}$ и $\hat{\boldsymbol{\alpha}}$, определим $\underset{\sim}{I}$, матрицу вторых производных F со знаком минус и ее подматрицы $\underset{\sim}{A}$ и $\underset{\sim}{B}$:

$$\underset{\sim}{I} \equiv \begin{pmatrix} \underset{\sim}{A} & \underset{\sim}{B} \\ \underset{\sim}{B}^T & 0 \end{pmatrix} = -E \begin{pmatrix} \frac{\partial^2 F}{\partial \theta\, \partial \theta'} & \frac{\partial^2 F}{\partial \theta\, \partial \alpha} \\ \left(\frac{\partial^2 F}{\partial \theta\, \partial \alpha}\right)^T & \frac{\partial^2 F}{\partial \alpha^2} \end{pmatrix} = -E \begin{pmatrix} \frac{\partial^2 \ln L}{\partial \theta\, \partial \theta'} & \frac{\partial \mathbf{g}}{\partial \theta} \\ \left(\frac{\partial \mathbf{g}}{\partial \theta}\right)^T & 0 \end{pmatrix}.$$

Обратное значение $\underset{\sim}{I}$ равно

$$\underset{\sim}{I}^{-1} = \begin{pmatrix} \underset{\sim}{A}^{-1} - \underset{\sim}{A}^{-1} \underset{\sim}{B} (\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1} \underset{\sim}{B}^T \underset{\sim}{A}^{-1} & \underset{\sim}{A}^{-1} \underset{\sim}{B} (\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1} \\ (\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1} \underset{\sim}{B}^T \underset{\sim}{A}^{-1} & -(\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1} \end{pmatrix}.$$

Разлагая в ряд Тейлора вектор

$$\begin{pmatrix} \partial F/\partial \theta \\ \partial F/\partial \alpha \end{pmatrix}_{\substack{\theta = \hat{\theta} \\ \alpha = \hat{\alpha}}} = \begin{pmatrix} \partial F/\partial \theta \\ \partial F/\partial \alpha \end{pmatrix}_{\substack{\theta = \theta_0 \\ \alpha = 0}} - \begin{pmatrix} \underset{\sim}{A} & \underset{\sim}{B}^T \\ \underset{\sim}{B} & 0 \end{pmatrix}_{\substack{\theta = \theta_0 \\ \alpha = 0}} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\alpha} \end{pmatrix} + \ldots$$

и используя тот факт, что этот вектор в соответствии с (8.15) равен нулю, можно записать выражение для матрицы вторых моментов:

$$E\left[\begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\alpha} \end{pmatrix} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\alpha} \end{pmatrix}^T\right] = \underset{\sim}{I}^{-1} \begin{pmatrix} \underset{\sim}{A} & 0 \\ 0 & 0 \end{pmatrix} \underset{\sim}{I}^{-1}.$$

После утомительных, но несложных вычислений получаем:

$$\underset{\sim}{D}(\hat{\theta} - \theta_0) = \underset{\sim}{A}^{-1} - \underset{\sim}{A}^{-1} \underset{\sim}{B} (\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1} \underset{\sim}{B}^T \underset{\sim}{A}^{-1} \tag{8.16}$$

и

$$\underset{\sim}{D}(\hat{\alpha}) = (\underset{\sim}{B}^T \underset{\sim}{A}^{-1} \underset{\sim}{B})^{-1}.$$

Первый член $\underset{\sim}{A}^{-1}$ в правой части (8.16) — обычная матрица дисперсий в отсутствие связей, тогда как второй член соответствует уменьшению дисперсии, полученной в результате добавления информации

в виде связей. Смешанный второй момент $\hat{\theta}$ и $\hat{\alpha}$ равен нулю: оценки параметров не коррелированы с $\hat{\alpha}$.

Отметим, что экстремум F является седловой точкой: F максимально по $\hat{\theta}$, но минимально по $\hat{\alpha}$. Это полностью может сделать неприменимыми обычные методы оптимизации, основанные на сравнении последовательных значений F (методы «подъема на холм»). Однако целью настоящей книги не является обсуждение достоинств различных методов оптимизации. Тех, кто интересуется этим, мы отсылаем к специальной литературе [4, 5].

Чтобы избежать этой трудности, можно, например, отыскивать максимум некоторой взвешенной суммы квадратов производных в уравнениях (8.15), а не максимум F.

Все сказанное выше относилось к случаю, когда I не сингулярна. В частности, $\partial^2 \ln L/\partial\theta\partial\theta'$ не должно быть сингулярно для $\theta = \theta_0$.

Это предположение неверно на практике, когда условие $g(\theta) = 0$ необходимо для однозначного определения параметров.

Например, если теоретическое распределение есть однородная функция θ, то нужно наложить некоторое условие типа

$$\sum_{i=1}^{n} \theta_i = 1.$$

Если из соображений симметрии нежелательно такую связь учитывать путем замены переменных, то метод лагранжевых множителей применим, если F заменить

$$F' = \ln L(\mathbf{X} \mid \theta) - g^2(\theta) + \alpha^T g(\theta).$$

Тогда уравнения (8.15) приобретают вид

$$\frac{\partial \ln L(\mathbf{X} \mid \theta)}{\partial \theta} - 2g(\theta)\frac{\partial g}{\partial \theta} + \hat{\alpha}^T \frac{\partial g}{\partial \theta}(\hat{\theta}) = 0;$$

$$g(\hat{\theta}) = 0$$

и $\underset{\sim}{I}$ заменяется

$$\underset{\sim}{I'} = -E\begin{pmatrix} \dfrac{\partial^2 \ln L}{\partial\theta\,\partial\theta'} + \dfrac{\partial g}{\partial\theta}\dfrac{\partial g}{\partial\theta'} & \dfrac{\partial g}{\partial\theta'} \\ \dfrac{\partial g^T}{\partial\theta} & 0 \end{pmatrix}_{\theta=\theta_0}$$

Величина

$$\frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'} + \frac{\partial g}{\partial\theta}\frac{\partial g}{\partial\theta'}$$

уже больше не сингулярна и все наши результаты (включая и матрицу вторых моментов с заменой F на F') справедливы.

§ 8.4. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (χ^2-МЕТОД)

Другим общим методом оценки является метод *наименьших квадратов*, часто называемый χ^2-методом. Путаница в терминологии возникает из-за того, что важный класс функций наименьших

квадратов ведет себя подобно классу функций χ^2, введенных в п. 4.2.3. Однако такое поведение характерно не для любой суммы квадратов разностей. (В п. 8.7.3 обсуждается качественно вопрос о том, почему используются наименьшие квадраты, а, например, не наименьшие кубы или наименьшие абсолютные значения.)

Рассмотрим выборку $Y_1, \ldots, Y_N$ измерений, причем $E(Y_i, \theta)$ и $\underset{\sim}{D}$ являются средними значениями и матрицей вторых моментов соответствующего распределения. Здесь θ — неизвестные параметры, а $E(Y, \theta)$ и $D_{ij}(\theta)$ — известные функции θ.

В методе наименьших квадратов оценками параметров θ_k служат такие значения $\hat{\theta}_k$, которые обращают в минимум ковариационную (квадратичную) форму (7.18) или

$$Q^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} [Y_i - E(Y_i, \theta)] (\underset{\sim}{D}^{-1})_{ij} [Y_j - E(Y_j, \theta)] = \\ = [\mathbf{Y} - E(\mathbf{Y}, \theta)]^T \underset{\sim}{D}^{-1} [\mathbf{Y} - E(\mathbf{Y}, \theta)]. \tag{8.17}$$

В п. 7.2.4 и 7.3.3 было показано, что оценки метода наименьших квадратов состоятельны, а для линейного случая они еще и не смещены.

В общем случае матрица вторых моментов $\underset{\sim}{D}$ не диагональна. Если наблюдения Y_i независимы и, следовательно, не коррелированы, матрица вторых моментов диагональна и ее элементы

$$D_{ii} = \sigma_i^2(\theta).$$

В этом случае квадратичная форма (8.17) сводится к хорошо известной сумме квадратов:

$$Q^2 = \sum_{i=1}^{N} \sigma_i^{-2}(\theta) [Y_i - E(Y_i \theta)]^2.$$

Наблюдениями могут быть, например, числа событий в ячейках гистограммы или координаты точек трека в пузырьковой камере.

8.4.1. Линейная модель. Метод наименьших квадратов (см. п. 7.2.4) становится линейным, когда дисперсии σ_i^2 не зависят от r параметров $\theta = (\theta_1, \ldots \theta_r)$, а ожидания $E(Y_i, \theta)$ являются линейными функциями параметров θ_j:

$$E(Y_i, \theta) = \sum_{j=1}^{r} a_{ij} \theta_j, \quad i = 1, \ldots, n,$$

или, если использовать матричные обозначения,

$$E(\mathbf{Y}, \theta) = \underset{\sim}{A}\theta. \tag{8.18}$$

Элементы a_{ij} *матрицы планирования* $\underset{\sim}{A}$ задаются моделью.

Рассмотрим следующий пример. Предположим, нужно аппроксимировать трек в пузырьковой камере полиномом третьей степени по X. Пусть различным фиксированным значениям точек X_i соответствуют измеренные Y координаты Y_i с дисперсией σ^2.

Уравнение кривой, аппроксимирующей трек,

$$Y_i = Y(X_i) = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \theta_3 X_i^3, \tag{8.19}$$

имеет вид (8.18). Чтобы задать матрицу $\underset{\sim}{A}$, нужно только вычислить $(j-1)$-ю степень X_i.

Решая *обычные уравнения* $\partial\theta^2/\partial\theta = 0$, получаем в соответствии с (7.23) для оценок $\hat{\theta}$ метода наименьших квадратов в линейном случае:

$$\hat{\theta} = (\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \underset{\sim}{D}^{-1} \mathbf{Y}. \tag{8.20}$$

Отметим, что такая минимизация осуществляется простым обращением матриц без каких-либо приближений, поисков или итеративных процедур. Она дает точное и единственное решение при условии, если $\underset{\sim}{A}^T\underset{\sim}{D}^{-1}\underset{\sim}{A}$ не вырождена. Метод наименьших квадратов без весов соответствует $\underset{\sim}{D} = \underset{\sim}{I}$, где $\underset{\sim}{I}$ — единичная матрица, и решение (8.20) переходит в

$$\hat{\theta} = (\underset{\sim}{A}^T \underset{\sim}{A})^{-1} \underset{\sim}{A}^T \mathbf{Y}. \tag{8.21}$$

Из уравнений (8.20) и (8.21) следует, что $\hat{\theta}$ являются *линейными оценками*.

В п. 7.3.3 было показано, что оценки (8.20) не смещены для всех порядков по N. Далее из теоремы Гаусса — Маркова (см. п. 7.4.4) следует, что среди всех линейных несмещенных оценок

$$\mathbf{t} = \underset{\sim}{L}\ \mathbf{Y}$$

оценки (8.20) имеют минимальную дисперсию. Эти дисперсии равны диагональным элементам матрицы:

$$D(\hat{\theta}) = E[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] = \sigma^2 [(\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A})^{-1}]. \tag{8.22}$$

Для нахождения $\hat{\theta}$ матрица $\underset{\sim}{D}$ должна быть известна с точностью до постоянного множителя.

Однако для определения *дисперсии* $\hat{\theta}$ должна быть известна и сама σ^2. Если она неизвестна, то ее можно оценить по минимальному значению Q^2:

$$Q^2_{\text{мин}} = (\mathbf{Y} - \underset{\sim}{A}\hat{\theta})^T \underset{\sim}{D}^{-1} (\mathbf{Y} - \underset{\sim}{A}\hat{\theta}), \tag{8.23}$$

называемому *результирующей* суммой *квадратов*. Используя матричные преобразования, можно получить следующее выражение для ожидания $Q^2_{\text{мин}}$ [2]:

$$E(Q^2_{\text{мин}}) = \sigma^2 (N - r).$$

Вследствие этого величина

$$\hat{\sigma}^2 = Q^2_{\text{мин}}/(N-r) \tag{8.24}$$

является несмещенной оценкой σ^2. Для метода наименьших квадратов без весов, когда $D = I$, уравнение (8.24) дает несмещенную оценку общей дисперсии наблюдений.

Отметим, что не было сделано никаких предположений о распределении генеральной совокупности данных **Y**. Оптимальные свойства оценок метода наименьших квадратов обусловлены только линейностью.

Когда данные распределены по Гауссу, то вследствие линейности оценки метода наименьших квадратов тоже распределены по Гауссу с матрицей вторых моментов, задаваемой уравнением (8.22). В этом случае метод наименьших квадратов эквивалентен методу максимального правдоподобия.

8.4.2. Полиномиальная модель. Если степень полинома в аппроксимирующем выражении (8.19) больше шести или семи, возникают серьезные трудности при обращении матрицы A^TA. Тогда следует преобразовать аппроксимирующее выражение, используя ортогональные полиномы $\xi_j(X)$ порядка $j = 1, \ldots, r$. Такое преобразование всегда возможно. Свойство ортогональности выражается следующим образом:

$$\sum_{i=1}^{N} \xi_j(X_i)\,\xi_k(X_i) = \delta_{jk}. \tag{8.25}$$

В этом случае элементы матрицы A равны: $A_{ij} = \xi_j(X_i)$ и $A^TA = 1$ есть единичная матрица. Таким образом избавляются от некоторой потери точности, вызываемой обращением матрицы. Тем не менее некоторые ограничения на получаемую точность возникают при выборе ортогональных полиномов $\xi_j(X)$.

Итак, модель имеет теперь такой вид: $Y(X) = \sum_j \psi_j \xi_j(X)$. Оценки параметров равны

$$\hat{\psi}_j = \sum_{i=1}^{N} \xi_j(X_i) Y(X_i)$$

с матрицей вторых моментов $\sigma^2 I$. Как видно, оценки распределены независимо. Соответственно *результирующая сумма квадратов* (8.23) приобретает вид

$$Q^2_{\text{мин}} = \sum_{i=1}^{N} Y^2(X_i) = \sum_{j=1}^{r} \hat{\psi}_j^2$$

и несмещенная оценка дисперсии σ^2 определяется уравнением (8.24). Важной особенностью оценок $\hat{\psi}_j$ является тот факт, что они распределены независимо.

В более общем случае, когда матрица вторых моментов наблюдений Y_i, $i = 1, \ldots, N$ имеет вид $\sigma^2 \underset{\sim}{D}$, полиномы $\xi_j(X)$ должны обладать свойством

$$\sum_{i=1}^{N} \xi_j(X_i)\,\xi_k(X_i)\,(\underset{\sim}{D}^{-1})_{jk} = \delta_{jk}$$

вместо свойства (8.25). Соответственно результирующая сумма квадратов (8.23) приобретает вид

$$Q^2_{\text{мин}} = \sum_{i=1}^{N}\sum_{j=1}^{N} Y_i Y_j (\underset{\sim}{D}^{-1})_{ij} - \sum_{j=1}^{r} \hat{\psi}_j^2.$$

В полиномиальных моделях ортогональные функции $\xi_j(X_i)$ могут быть выбраны методом Форсайта. Рекомендуемая техника позволяет достигать значительно более высоких числовых точностей. Если какая-то модель может быть выражена через ортогональные функции (например, полиномы Лежандра или присоединенные полиномы Лежандра), то указанные выше свойства выполняются и, в частности, независимость оценок параметров.

Важно отметить, что значимость k-й степени при аппроксимации измеряется параметром $\hat{\psi}_k$, а не коэффициентом $\hat{\theta}_k$, как это было в обычной полиномиальной модели. Подробнее этот вопрос будет рассматриваться в главе, посвященной проверке гипотез (см. гл. 10).

8.4.3. Связанные параметры в линейной модели. 1. Метод оценок. Пусть для N наблюдений Y_i и r параметров θ_j модель имеет вид $Y = A\theta + \varepsilon$. Тогда ожидаемое значение задается уравнением (8.18). Независимые переменные A_{ij}, образующие матрицу A, могут иметь любую форму. В частности, A_{ij} может быть равно X_i^{j-1}, что соответствует аппроксимации наблюдений Y_i, измеренных в точках X_i, полиномом r-й степени. Такая аппроксимация уже обсуждалась в двух предыдущих параграфах.

Предположим, что матрица вторых моментов величин ε_i равна $\sigma^2 \underset{\sim}{D}$ и параметры θ удовлетворяют m линейным уравнениям

$$\sum_{j=1}^{r} l_{kj}\theta_j = R_k, \quad k = 1, \ldots, m$$

или в матричных обозначениях

$$\underset{\sim}{L}\theta = \mathbf{R}.$$

Чтобы получить *оценки наименьших квадратов* для параметров θ, воспользуемся вектором множителей Лагранжа λ при минимизации функционала

$$Q^2 = (\mathbf{Y} - \underset{\sim}{A}\theta)^T \underset{\sim}{D}^{-1} (\mathbf{Y} - \underset{\sim}{A}\theta) + \lambda^T (\underset{\sim}{L}\theta - \mathbf{R}).$$

Дифференцируя его по θ и λ, получаем *нормальные уравнения*

$$\left.\begin{aligned} \underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A}\theta + \underset{\sim}{L}^T \lambda &= \underset{\sim}{A}^T \underset{\sim}{D}^{-1} \mathbf{Y}; \\ \underset{\sim}{L}\theta &= \mathbf{R}. \end{aligned}\right\} \tag{8.26}$$

Записав их в виде

$$\begin{bmatrix} \underset{\sim}{C} & \vdots & \underset{\sim}{L}^T \\ \cdots & \cdots & \cdots \\ \underset{\sim}{L} & \vdots & \underset{\sim}{\zeta} \end{bmatrix} = \begin{bmatrix} \theta \\ \cdots \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{S} \\ \cdots \\ \mathbf{R} \end{bmatrix},$$

где $\underset{\sim}{C} = \underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A}$, $\mathbf{S} = \underset{\sim}{A}^T \underset{\sim}{D}^{-1} \mathbf{Y}$, а $\underset{\sim}{\zeta}$ — нулевая матрица, получим для $\hat{\theta}$ и $\hat{\lambda}$:

$$\begin{bmatrix} \theta \\ \cdots \\ \lambda \end{bmatrix} = \begin{bmatrix} \underset{\sim}{F} & \vdots & \underset{\sim}{G}^T \\ \cdots & \cdots & \cdots \\ \underset{\sim}{G} & \vdots & \underset{\sim}{H} \end{bmatrix} \begin{bmatrix} \mathbf{S} \\ \cdots \\ \mathbf{R} \end{bmatrix}. \tag{8.27}$$

Несложные матричные преобразования приводят к таким выражениям:

$$\underset{\sim}{F} = \underset{\sim}{C}^{-1} - \underset{\sim}{C}^{-1} \underset{\sim}{L}^T (\underset{\sim}{L} \underset{\sim}{C}^{-1} \underset{\sim}{L}^T)^{-1} \underset{\sim}{L} \underset{\sim}{C}^{-1};$$
$$\underset{\sim}{G} = (\underset{\sim}{L} \underset{\sim}{C}^{-1} \underset{\sim}{L}^T)^{-1} \underset{\sim}{L} \underset{\sim}{C};$$
$$\underset{\sim}{H} = -(\underset{\sim}{L} \underset{\sim}{C}^{-1} \underset{\sim}{L}^T)^{-1}.$$

Используя уравнения (8.27) и соотношения между матрицами $\underset{\sim}{C}$, $\underset{\sim}{L}$, $\underset{\sim}{F}$, $\underset{\sim}{G}$ и $\underset{\sim}{H}$, можно изучить свойства оценок параметров. Соотношения для матриц $\underset{\sim}{C}$, $\underset{\sim}{L}$, $\underset{\sim}{F}$, $\underset{\sim}{G}$ и $\underset{\sim}{H}$ таковы:

$$\underset{\sim}{C}\underset{\sim}{F} + \underset{\sim}{L}^T \underset{\sim}{G} = \underset{\sim}{I}; \quad \underset{\sim}{L}\underset{\sim}{F} = \underset{\sim}{\zeta};$$
$$\underset{\sim}{C}\underset{\sim}{G}^T + \underset{\sim}{L}^T \underset{\sim}{H} = \underset{\sim}{\zeta}; \quad \underset{\sim}{L}\underset{\sim}{G}^T = \underset{\sim}{I}.$$

Для оценок получается выражение:

$$\hat{\theta} = \underset{\sim}{F}\mathbf{S} + \underset{\sim}{G}^T \mathbf{R} = \underset{\sim}{F}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \mathbf{Y} + \underset{\sim}{G}^T \mathbf{R}. \tag{8.28}$$

Так как мы предполагаем, что в среднем ошибки равны нулю, ожидаемые значения параметров приобретают вид

$$E(\hat{\theta}) = E\left[\underset{\sim}{F}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} (\underset{\sim}{A}\theta + \varepsilon) + \underset{\sim}{G}^T \underset{\sim}{L}\theta\right] =$$
$$= E(\underset{\sim}{F}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \varepsilon + \theta) = \theta. \tag{8.29}$$

Это показывает, что оценки не смещены. Соответственно для множителей Лагранжа $\hat{\lambda}$:

$$\hat{\lambda} = \underset{\sim}{G}\mathbf{S} + \underset{\sim}{H}\mathbf{R} = \underset{\sim}{G}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \underset{\sim}{A}\theta + \underset{\sim}{H}\underset{\sim}{L}\theta + \underset{\sim}{G}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \varepsilon = \underset{\sim}{G}\underset{\sim}{A}^T \underset{\sim}{D}^{-1} \varepsilon.$$

Следовательно, $E(\hat{\lambda}) = \zeta \equiv 0$.

Таким образом, среднее значение множителей Лагранжа, вычисляемое таким способом, равно нулю. Этот результат вполне разумен, поскольку уже предполагалось, что истинные значения параметров удовлетворяют уравнениям связи.

2. Матрица вторых моментов оценок. Матрицу вторых моментов оценок можно найти с помощью уравнения (8.29) и того факта, что дисперсия ошибок по определению равна $\underset{\sim}{D}(\varepsilon) = \sigma^2 \underset{\sim}{D}$

$$\underset{\sim}{D}(\hat{\theta}) = E\left[\underset{\sim}{F}\,\underset{\sim}{A}^T\,\underset{\sim}{D}^{-1}\,\varepsilon\,\varepsilon^T\,\underset{\sim}{D}^{-1}\,\underset{\sim}{A}\,\underset{\sim}{F}\right] = \sigma^2 \underset{\sim}{F}\,\underset{\sim}{C}\,\underset{\sim}{F} = \sigma^2 \underset{\sim}{F} =$$
$$= \sigma^2\left[\underset{\sim}{C}^{-1} - \underset{\sim}{C}^{-1}\,\underset{\sim}{L}^T\,(\underset{\sim}{L}\,\underset{\sim}{C}\,\underset{\sim}{L}^T)^{-1}\,\underset{\sim}{L}\,\underset{\sim}{C}^{-1}\right]. \qquad (8.30)$$

Если бы уравнения связи отсутствовали, то матрица вторых моментов оценок θ была бы равна $\sigma^2 C^{-1}$. Диагональные элементы второго члена служат мерой уменьшения дисперсий оценок, получаемой в результате добавления связей. Что касается смешанных вторых моментов оценок параметров, то они могут увеличиваться или уменьшаться в зависимости от конкретной задачи.

Матрица вторых моментов лагранжевых множителей равна:

$$\underset{\sim}{D}(\hat{\lambda}) = E\left[\underset{\sim}{G}\,\underset{\sim}{A}^T\,\underset{\sim}{D}^{-1}\,\varepsilon\varepsilon^T\,\underset{\sim}{D}^{-1}\,\underset{\sim}{A}\,\underset{\sim}{G}^T\right] = \sigma^2 \underset{\sim}{G}\,\underset{\sim}{C}\,\underset{\sim}{G}^T =$$
$$= -\sigma^2 \underset{\sim}{H} = \sigma^2 (\underset{\sim}{L}\,\underset{\sim}{C}^{-1}\,\underset{\sim}{L}^T)^{-1}.$$

Можно также показать, что смешанный второй момент для $\hat{\theta}$ и $\hat{\lambda}$ равен нулю.

Интересно отметить, что матрица вторых моментов параметров $\hat{\theta}$ является соответствующей подматрицей, обратной матрице «нормальных уравнений» (8.26), а матрица вторых моментов $\hat{\lambda}$ — соответствующей подматрицей со знаком минус.

3. Оценка σ^2. Результирующая сумма квадратов задается уравнением (8.23)

$$Q^2_{\text{мин}} = (\mathbf{Y} - \underset{\sim}{A}\,\hat{\theta})^T\,\underset{\sim}{D}^{-1}\,(\mathbf{Y} - \underset{\sim}{A}\,\hat{\theta}) = \mathbf{Y}^T\,\underset{\sim}{D}^{-1}\,\mathbf{Y}^T -$$
$$- \hat{\theta}^T\,\underset{\sim}{S} - \underset{\sim}{S}^T\,\hat{\theta} + \hat{\theta}^T\,\underset{\sim}{C}\,\hat{\theta} = \mathbf{Y}^T\,\underset{\sim}{D}^{-1}\,\mathbf{Y}^T - 2\hat{\theta}^T\,(\underset{\sim}{C}\,\hat{\theta} + \underset{\sim}{L}^T\,\hat{\lambda}) +$$
$$+ \hat{\theta}^T\,\underset{\sim}{C}\,\hat{\theta} = \mathbf{Y}^T\,\underset{\sim}{D}^{-1}\,\mathbf{Y}^T - (\hat{\theta}^T\,\underset{\sim}{C}\,\hat{\theta} + 2\hat{\lambda}^T\,\mathbf{R}),$$

или

результирующая сумма квадратов = полной сумме квадратов — — сумма квадратов, полученная после минимизации.

Для ожидаемых значений получаем

$$E\,(\mathbf{Y}^T\,\underset{\sim}{D}^{-1}\,\mathbf{Y}^T) = E\left[(\underset{\sim}{A}\theta + \varepsilon)^T\,\underset{\sim}{D}^{-1}\,(\underset{\sim}{A}\,\theta + \varepsilon)\right] =$$
$$= \theta^T\,\underset{\sim}{C}\,\theta + E\,(\varepsilon^T\,\underset{\sim}{D}^{-1}\,\varepsilon),$$

поскольку для любой матрицы $\underset{\sim}{B}$

$$E\,(\varepsilon^T\,\underset{\sim}{B}\,\varepsilon) = E\left[\sum_i \sum_j \varepsilon_i\,B_{ij}\,\varepsilon_j\right] = \sigma^2 \sum_i \sum_j B_{ij} D_{ij} = \operatorname{tr}(\underset{\sim}{B}\,\underset{\sim}{D}).$$

Отсюда следует, что

$$E\left(\mathbf{Y}^T \underset{\sim}{D}^{-1} \mathbf{Y}^T\right) = \theta^T \underset{\sim}{C}\, \theta + \sigma^2 \operatorname{tr}\left(\underset{\sim}{D}^{-1} \underset{\sim}{D}\right) = \theta^T \underset{\sim}{C}\, \theta + N\sigma^2.$$

Точно так же можно показать, что

$$E\left(\hat{\theta}\, \underset{\sim}{C}\, \hat{\theta} + 2\hat{\lambda}^T \mathbf{R}\right) = \hat{\theta}^T \underset{\sim}{C}\, \theta + (r-m)\,\sigma^2.$$

В итоге мы получаем следующий результат:

$$E\left(Q^2_{\text{мин}}\right) = N\sigma^2 - (r-m)\,\sigma^2,$$

и

$$\hat{\sigma}^2 = Q^2_{\text{мин}}/(N-r+m) \tag{8.31}$$

служит несмещенной оценкой σ^2.

Таким образом, оценки наименьших квадратов (8.28), (8.30) и (8.31) параметров, подчиняющихся уравнениям связи, имеют те же самые свойства, что и оценки (8.21), (8.22) и (8.24) для параметров без связей. В частности, эти результаты для вопроса оценок значений параметров точкой не зависят от распределения наблюдений. Однако при оценке параметра интервалом значений знание распределения генеральной совокупности необходимо. С этим мы встретимся в гл. 9.

8.4.4. Нелинейные модели в случае, когда данные распределены по нормальному закону. Предположим, что наблюдения Y_i распределены по нормальному закону со средним $E(Y_i)$, не коррелированы и имеют дисперсии σ_i^2.

Пусть параметры θ фигурируют в некоторой нелинейной модели, по которой

$$E(\mathbf{Y}_i) = f_i(\theta). \tag{8.32}$$

Тогда функция правдоподобия будет равна

$$L(\mathbf{Y}, \theta) = \prod_{i=1}^{N} \quad {}^{2})^{-\frac{1}{2}} \exp\left\{-\frac{[Y_i - f_i(\theta)]^2}{2\sigma_i^2}\right\},$$

соответственно

$$\ln L(\ , \theta) = -\frac{1}{2} \sum_{i=1}^{N} \ln(2\pi\sigma_i^2) - \sum_{i=1}^{N} \frac{[Y_i - f_i(\theta)]^2}{2\sigma_i^2}.$$

Поскольку первый член является константой, $\ln L$ максимален, когда

$$Q^2 = \sum_{i=1}^{N} \frac{[Y_i - f_i(\theta)]^2}{\sigma_i^2} \tag{8.33}$$

минимально, т. е. в этом случае метод наименьших квадратов эквивалентен методу м. п. Он оптимален для всех N, поскольку в этом случае существуют достаточные статистики (см. п. 7.4.2). Когда

параметры θ известны, Q^2 является суммой квадратов N стандартных переменных нормального закона и, следовательно, она распределена по закону χ^2 с N степенями свободы. Но если оценки θ должны быть найдены из условия минимума Q^2, закон распределения Q^2 уже другой. В этом случае минимальное значение Q^2 является суммой квадратов коррелированных, нецентральных (среднее значение не равно нулю) случайных переменных, которые, вообще говоря, распределены не по нормальному закону. Точное распределение этой величины очень трудно найти и, как описано в гл. 7, приходится изучать ее асимптотические свойства.

Когда отдельные наблюдения Y_i распределены *не* по нормальному закону, метод наименьших квадратов все еще используется для оценки параметров, но он уже не обладает такими общими оптимальными свойствами, которые так важны при малых N. Даже в асимптотическом пределе его оценки не всегда обладают минимальной дисперсией (в нелинейных моделях).

8.4.5. Оценка параметров по гистограммам: сравнение методов максимального правдоподобия и наименьших квадратов. Рассмотрим случай, когда наблюдениями Y_i являются количества событий в ячейках гистограмм. В этом случае Y_i будут иметь мультиномиальное распределение (см. п. 4.1.2). Как мы знаем, в пределе больших чисел событий это распределение становится асимптотически нормальным. Можно ожидать, что в этих условиях становится применимым метод наименьших квадратов. Выражение для ожидаемого числа событий в i-й ячейке как функцию параметра θ можно получить с помощью соответствующей модели (8.32).

Предполагая, что гистограмма имеет много ячеек (N велико), так что $p \ll 1$, мы приходим к примеру 2, рассмотренному в п. 4.1.2:

$$\sigma^2 (Y_i) \approx f_i (\theta). \tag{8.34}$$

Если пренебречь корреляциями между различными ячейками, то нужно минимизировать функционал (8.33):

$$Q^2 = \sum_{i=1}^{N} \frac{[Y_i - f_i(\theta)]^2}{f_i(\theta)}. \tag{8.35}$$

В результате мы приходим к обычному *методу минимума* χ^2 для случая оценки параметров по гистограммам. Причина для такого названия состоит в том, что Q^2 в асимптотике имеет распределение χ^2, если θ известно и в каждой ячейке содержится большое число событий (приближение нормального распределения). Отметим, что когда в ячейке гистограммы содержится только очень малое число событий, соотношение (8.34) все еще выполняется, поскольку дисперсия биномиального распределения равна $Np(1-p) \approx Np$ для всех N. Но при этом Y_i распределены уже не по нормальному закону, поэтому невозможно сделать каких-либо общих предсказаний относительно распределения Q^2.

Поскольку мы пренебрегли корреляцией и сделали приближения при конструировании функционала (8.35), нет оснований ожидать каких-либо интересных свойств у оценки, получаемой таким методом. Тем не менее оказывается, что такие оценки обладают оптимальными асимптотическими свойствами [2], они состоятельны, асимптотически нормально распределены и эффективны (имеют минимальную дисперсию). Это как раз свойства *наилучших асимптотически нормальных* оценок (н. а. н.).

При оценке параметров по гистограммам часто используется так называемый *модифицированный метод минимума* χ^2. Метод состоит в минимизации величины

$$Q^2 = \sum_{i=1}^{N} \frac{[Y_i - f_i(\theta)]^2}{Y_i}. \tag{8.36}$$

Это обусловлено практическими соображениями: более простой формой, экономией при расчетах и т. д.

Поскольку $E(Y_i) = f_i(\theta)$, то можно ожидать, что модифицированный метод будет разумным приближением к обычному. Действительно, можно показать, что в *асимптотике* при больших N эта оценка совпадает с оценкой метода минимума χ^2 и также является наилучшей асимптотически нормальной оценкой.

Наконец, третий способ — применить метод м. п. к мультиномиальному распределению гистограммированных событий, а не к первоначальным событиям, как это предлагалось в § 8.3. *Оценки максимального правдоподобия для мультиномиального распределения* находятся из условия обращения в максимум функционала:

$$\ln L = \sum_{i=1}^{N} Y_i \ln f_i(\theta). \tag{8.37}$$

Здесь N — число ячеек в гистограмме.

Используя общие свойства максимального правдоподобия (см. гл. 7), можно показать, что эти оценки асимптотически идентичны предыдущим двум оценкам и они также являются н. а. н. оценками.

Таким образом, эти три метода асимптотически нормальны. Тем не менее можно показать [2, 36], что в определенном смысле метод м. п. быстрее становится эффективным при выполнении условий регулярности и что модифицированный метод минимума χ^2 наихудший с этой точки зрения.

Таким образом, если нет особых причин выбрать один из этих методов (более быстрый расчет, например, вследствие линейности), *то рекомендуется использовать метод максимального правдоподобия* с использованием уравнения (8.37).

В частности, при использовании метода м. п. не возникает никаких трудностей, если в ячейку гистограммы попадет мало событий либо совсем не попадет, тогда как другие методы либо теряют эффективность из-за отсутствия нормальности распределения чис-

ла событий, либо вообще неприменимы из-за того, что некоторые члены в (8.36) становятся бесконечными.

Эти недостатки обычно обходят путем группировки ячеек, но при этом теряется информация. С другой стороны, метод м. п. для гистограммированных событий становится эквивалентным обычному методу м. п., когда ширина ячеек стремится к нулю.

Случай взвешенных событий будет рассмотрен в п. 8.5.4.

§ 8.5. ВЕСА И ЭФФЕКТИВНОСТЬ РЕГИСТРАЦИИ

Обычно физик-экспериментатор не имеет возможности непосредственно наблюдать изучаемое явление. Как правило, аппаратура вносит некоторые искажения или смещения и этот эффект необходимо принимать в расчет.

В физике высоких энергий источником такого смещения является недостаточно высокая *эффективность регистрации* детектора частиц. Хорошим примером может быть эксперимент по изучению нейтральных частиц в приборе, регистрирующем только заряженные частицы. Подобным прибором может быть пузырьковая камера. Нейтральная частица «наблюдается» только тогда, когда она распадается в пределах детектора на заряженные частицы. Это приводит к искажениям трех типов.

1. События не регистрируются, если нейтральная частица распадается на нейтральные частицы. Поскольку вероятность таких распадов не зависит от импульса, ее углов и т. д., это приводит к уменьшению общего числа событий.

2. События не регистрируются, если нейтральная частица распадается за пределами детектора. Такая потеря больше для быстрых частиц, чем для медленных, а также для частиц, проходящих более короткие пути внутри детектора.

3. События часто теряются, если нейтральная частица распадается вблизи точки рождения, что делает распад неотличным от реакций без нейтральной части. Это приводит не только к потере событий с нейтральными частицами, но также может служить источником дополнительного фона при изучении других реакций. Величина этого искажения зависит от импульса частицы.

Способы учета таких смещений могут быть различны в зависимости от их важности. Если эффект велик (эффективность регистрации очень мала для некоторых классов событий), то во избежание больших потерь информации необходимо точно решать задачу. Соответствующий метод описан в следующем параграфе. С другой стороны, если детектор вносит не слишком большие искажения (скажем $\sim 20\%$), то могут быть использованы некоторые приближения, требующие менее точного знания вида погрешности. Если говорить кратко, то точное решение состоит в модификации основной *физической* ф. п. в. В результате такой модификации получается искаженная *наблюдаемая* ф. п. в., которую затем можно сравнивать непосредственно с наблюдаемыми данными. Приближенные методы

сохраняют физическую ф. п. в. и вместо этого корректируют данные путем приписывания *весов* событиям в соответствии с вероятностью наблюдения события данного класса. В этом случае основная задача состоит в правильном расчете матрицы вторых моментов оценок.

8.5.1. Идеальный метод — максимальное правдоподобие. Предположим, что измеряемые переменные могут быть разделены на две группы: на переменные $\mathbf{X}$, ф. п. в. которых непосредственно зависит от оцениваемых параметров $\boldsymbol{\theta}$, и переменные $\mathbf{Y}$, зависящие от $\mathbf{X}$ и, следовательно, косвенно от $\boldsymbol{\theta}$. Соответственно

$$P(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}) = P(\mathbf{X} \mid \boldsymbol{\theta})\, Q(\mathbf{Y} \mid \mathbf{X})$$

и, как всегда,

$$\int P(\mathbf{X} \mid \boldsymbol{\theta})\, d\mathbf{X} = \int Q(\mathbf{Y} \mid \mathbf{X})\, d\mathbf{Y} = 1.$$

Пусть $e(\mathbf{X}, \mathbf{Y})$ — вероятность того, что регистрируется событие с величинами $(\mathbf{X}, \mathbf{Y})$. Тогда $e(\mathbf{X}, \mathbf{Y})$ является *эффективностью регистрации* прибора, $P(\mathbf{X} \mid \boldsymbol{\theta})$ — истинное распределение величин $\mathbf{X}$, а ф. п. в. реальных наблюдений может быть записана:

$$g(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta}) = \frac{P(\mathbf{X} \mid \boldsymbol{\theta})\, Q(\mathbf{Y} \mid \mathbf{X})\, e(\mathbf{X}, \mathbf{Y})}{\int P(\mathbf{X} \mid \boldsymbol{\theta})\, Q(\mathbf{Y} \mid \mathbf{X})\, e(\mathbf{X}, \mathbf{Y})\, d\mathbf{X}\, d\mathbf{Y}}.$$

Тогда функция правдоподобия параметров $\boldsymbol{\theta}$ при заданной выборке наблюдений $(\mathbf{X}_i, \mathbf{Y}_i)$, $i = 1, \ldots, N$ равна

$$L(\mathbf{X}_1, \ldots, \mathbf{X}_N, \mathbf{Y}, \ldots, \mathbf{Y}_N \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} g(\mathbf{X}_i, \mathbf{Y}_i \mid \boldsymbol{\theta}) =$$

$$= \prod_{i=1}^{N} \frac{P(\mathbf{X}_i \mid \boldsymbol{\theta})\, Q(\mathbf{Y}_i \mid \mathbf{X}_i)\, e(\mathbf{X}_i, \mathbf{Y}_i)}{\int P(\mathbf{X} \mid \boldsymbol{\theta})\, Q(\mathbf{Y} \mid \mathbf{X})\, e(\mathbf{X}, \mathbf{Y})\, d\mathbf{X}\, d\mathbf{Y}}. \tag{8.38}$$

Пример. Предположим, что с помощью пузырьковой камеры изучается механизм взаимодействий в реакции

$$K^+ p \to K^0 \pi^+ p.$$

Пусть $\mathbf{X}$ в этом случае обозначает все переменные, определяющие кинематику события (импульсы, направление всех треков), а $\mathbf{Y}$ — координаты точки взаимодействия в объеме камеры. Очевидно, что механизм взаимодействий не связан с $\mathbf{Y}$, но эффективность регистрации зависит от времени жизни K^0 (частично от $\mathbf{X}$) и от возможной длины трека ($\mathbf{Y}$). Таким образом, эффективность регистрации связывает $\mathbf{X}$ и $\mathbf{Y}$, так что зависимость от $\mathbf{Y}$ остается в функции правдоподобия.

Используя (8.38), можно записать

$$\ln L = W + \sum_{i=1}^{N} \ln(e_i Q_i), \tag{8.39}$$

где

$$W = \sum_{i=1}^{N} \ln P_i - N \ln \int PeQ. \tag{8.40}$$

Здесь использованы частично сокращенные обозначения. Для простоты рассмотрим только один параметр θ с оценкой θ, обращающей в максимум W.

В соответствии с результатами п. 7.3.2 дисперсию $\sigma^2_{\hat{\theta}}$ оценки максимального правдоподобия $\hat{\theta}$ можно определить из равенства

$$\frac{1}{\sigma^2_{\hat{\theta}}} = N\left[\frac{\int \frac{1}{P} P'^2 eQ}{\int PeQ} - \left(\frac{\int P' eQ}{\int PeQ}\right)^2\right]. \tag{8.41}$$

Здесь $P' = \partial P(\mathbf{X} \mid \theta)/\partial\theta$, т. е. оценка

$$\hat{\sigma}^{-2}_{\hat{\theta}} = -\frac{\partial^2 W}{\partial \theta^2}\Big|_{\theta=\hat{\theta}}. \tag{8.42}$$

Нетрудно показать, что ожидаемое значение $\sigma^{-2}_{\hat{\theta}}$ равно $\sigma^{-2}_{\hat{\theta}}$.

Если аналитическое выражение для нормировки PeQ не существует, то для ее нахождения приходится применять различные численные методы. Одним из таких методов может быть метод Монте-Карло, требующий больших затрат времени. Кроме того, результаты зависят от формы Q, которая зачастую плохо определена. По этим соображениям приходится отыскивать методы, исключающие Q из выражений для $\hat{\theta}$ и $\sigma^2_{\hat{\theta}}$: очевидно, подобные методы приводят к уменьшению информации и ожидаемой точности оценок, но оценка становится менее зависимой* от формы Q.

8.5.2. Приближенный метод. Заменим в уравнении (8.39) W:

$$W' = \sum_i \left(\frac{1}{e_i} \ln P_i\right). \tag{8.43}$$

Интуитивным доводом в пользу приближения (8.43) служит тот факт, что каждое событие, наблюдаемое с вероятностью регистрации e_i, соответствует в некотором смысле числу $w_i = \frac{1}{e_i}$ реально происшедших событий. Тогда «правдоподобие» таких событий должно было бы иметь вид $P_i^{w_i}$ вместо P_i. Такое введение величин e_i соответствует искажению наблюдений функцией $1/e\ (X, Y)$ и при оценке параметров они должны фигурировать с теоретическим распределением $P\ (X \mid \theta)$. Величина $Q\ (\mathbf{Y})$ уже не будет фигурировать в «правдоподобии».

Несмотря на качественный характер таких аргументов, оценки $\hat{\theta}$, полученные из условия максимума W', также распределены

* См. § 8.7.

асимптотически нормально относительно истинного значения θ_0. Однако надо быть осторожным при вычислении дисперсии этой оценки. Нельзя использовать для оценки информации вторую производную W', поскольку

$$\sum_{i=1}^{N} w_i > N,$$

где N — число зарегистрированных событий. Чтобы обойти эту трудность, иногда производят перенормировку $w_i' = Nw_i/\Sigma w_i$, но это неудовлетворительный метод, если не все веса w_i близки к единице (что типично для пузырьковых камер).

Использовав общие методы п. 7.3.2, получим точное выражение для дисперсии оценки $\hat{\theta}'$, найденной из условия максимума W'. Наконец, покажем, что дисперсия оценки может быть уменьшена в результате выбрасывания событий. Этот парадоксальный результат следует из того, что с точки зрения использования информации этот метод не оптимален.

Имея в виду обозначения п. 7.3.2, находим

$$\xi(\theta) = \frac{1}{N}\frac{\partial W'}{\partial \theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \theta}\left(\frac{1}{e_i}\ln P_i\right). \tag{8.44}$$

Можно легко проверить, что $E[\xi(\theta_0)] = 0$, т. е. выполняются требования п. 7.3.2.

Отсюда следует, что дисперсия оценки $\hat{\theta}'$, удовлетворяющей равенству $\xi(\hat{\theta}') = 0$ (т. е. обращающей в максимум W'), равна:

$$\sigma^2_{\hat{\theta}'} = D_2'/ND_1'^2,$$

где

$$D_2' = \frac{\int \frac{P'^2}{P}\frac{Q}{e}}{\int PeQ} \quad \text{и} \quad D_1' = \frac{\int \frac{P'^2}{P}Q}{\int PeQ}.$$

Таким образом, величина, соответствующая уравнению (8.41), приобретает вид

$$\frac{1}{\sigma^2_{\hat{\theta}'}} = \frac{N\left(\int \frac{P'^2}{P}Q\right)^2}{\left(\int \frac{P'^2}{P}\frac{Q}{e}\right)\left(\int PeQ\right)}. \tag{8.45}$$

Хотя Q и не появляется в (8.44), оно еще фигурирует в выражении для дисперсий (8.45). Однако укажем способ оценки $1/\sigma^2_{\hat{\theta}}$ без использования Q. Очевидно, D_1' и D_2' могут быть выражены в виде ожиданий

$$D_1' = E\left[\frac{1}{e}\left(\frac{P'}{P}\right)^2\right]; \qquad D_2' = E\left[\left(\frac{P'}{Pe}\right)^2\right].$$

Далее,

$$\frac{\partial^2 W'}{\partial\theta^2} = \sum_{i=1}^{N} \frac{1}{e_i P_i} \frac{\partial^2 P_i}{\partial\theta^2} - \sum_{i=1}^{N} \frac{1}{e_i P_i^2} \left(\frac{\partial P_i}{\partial\theta} \right)^2$$

и

$$E\left(\frac{1}{eP} \frac{\partial^2 P}{\partial\theta^2} \right) = \frac{\int \frac{P'' PeQ}{eP}}{\int PeQ} = \frac{\int P'' Q}{\int PeQ} = 0,$$

поскольку $\int PQ = 1$. Таким образом, в качестве оценок для D_1' и D_2' получаем:

$$\hat{D}_1' = -\frac{\partial^2}{\partial\theta^2} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\ln P_i}{e_i} \right];$$

$$\hat{D}_2' = -\frac{\partial^2}{\partial\theta^2} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\ln P_i}{e_i^2} \right].$$

В многомерном случае получается иной результат [37]. Определим матричные элементы

$$H_{kl} = E\left[\frac{1}{e} \left(\frac{\partial \ln P}{\partial\theta_k} \right) \left(\frac{\partial \ln P}{\partial\theta_l} \right) \right]$$

и

$$H_{kl}' = E\left[\frac{1}{e^2} \left(\frac{\partial \ln P}{\partial\theta_k} \right) \left(\frac{\partial \ln P}{D\theta_l} \right) \right].$$

Оценки таких матричных элементов могут быть рассчитаны по формулам:

$$\left.\begin{aligned} H_{kl} &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{eP^2} \left(\frac{\partial P}{\partial\theta_k} \right) \left(\frac{\partial P}{\partial\theta_l} \right) \right\}_i; \\ H_{kl}' &= \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{1}{e^2 P^2} \left(\frac{\partial P}{\partial\theta_k} \right) \left(\frac{\partial P}{\partial\theta_l} \right) \right\}_i. \end{aligned}\right\} \qquad (8.46)$$

В то же время, если аналитический вид первых производных неизвестен, то можно численно рассчитать матричные элементы по формулам:

$$H_{kl} = -\frac{\partial^2 W'}{\partial\theta_k \partial\theta_l} = -\frac{\partial^2}{\partial\theta_k \partial\theta_l} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\ln P_i}{e_i} \right] \qquad (8.47)$$

и

$$H_{kl}' = -\frac{\partial^2}{\partial\theta_k \partial\theta_l} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\ln P_i}{e_i^2} \right].$$

Тогда матрица вторых моментов $\hat{\theta}'$ равна

$$D(\hat{\theta}') = \underset{\sim}{H}^{-1} \underset{\sim}{H}' \underset{\sim}{H}^{-1}, \tag{8.48}$$

Очевидно, если $e = 1$, то $\underset{\sim}{H}' = \underset{\sim}{H}$, и формула (8.48) приводит к результату обычного метода м. п. Итак, приближенный метод для многомерного случая таков. Отыскиваются оценки $\hat{\theta}'$ из условия максимума (8.43). Если это возможно, по уравнениям (8.46) рассчитываются матрицы H и H'. Если аналитический вид первых производных в уравнениях (8.46) неизвестен, используются уравнения (8.47). Затем по уравнению (8.48) вычисляется матрица вторых моментов.

8.5.3. Выбрасывание событий с большим весом. Появление одного события с очень большим весом обычно приводит к несрабатыванию приближенного метода в отличие от идеального метода. Уменьшение числа событий, конечно, уменьшает точность, но исключение событий с очень большими весами увеличило бы точность. Таким образом, должно быть оптимальное число событий.

Введем ступенчатую функцию $H(e - e_0)$, которая исключает события, эффективность регистрации которых меньше e_0. Тогда в соответствии с (8.45) дисперсия равна

$$\sigma^2(e_0) = \left(\frac{1}{N}\int PeQ\right) \frac{\int \frac{P'^2}{P}\frac{Q}{e} H}{\left(\int \frac{P'^2}{P} QH\right)^2}.$$

Вычислим производную по e_0:

$$\frac{d\sigma^2(e_0)}{de_0} = \left(\frac{1}{N}\int PeQ\right) \frac{\int \frac{P'^2}{P} Q\delta}{\int \frac{P'^2}{P} QH} \left[2\frac{\int \frac{P'^2}{P}\frac{QH}{e}}{\int \frac{P'^2}{P} QH} - \frac{1}{e_0}\right],$$

где $\delta \equiv \delta(e - e_0)$. Решая уравнение $d\sigma^2/de_0 = 0$, найдем

$$\sigma^2_{\text{мин}} = \left(\frac{1}{N}\int PeQ\right)\left[2e_{\text{мин}}\int \frac{P'^2}{P} QH(e - e_{\text{мин}})\right]^{-1}.$$

Пример. Выбор функций

$$p = \frac{1}{2}(1+\alpha X),\ Q = \frac{1}{2}(1+\beta Y),\ e = |(Y-X)/2|^{\mu},$$

$$-1 \leqslant X,\ Y \leqslant 1, \quad -1 \leqslant \alpha,\ \beta \leqslant 1$$

позволяет провести аналитическое сравнение идеального и приближенного методов. Рассмотрим вместо дисперсии σ^2 величину $N\sigma^2$ (дисперсия имеет обычную N^{-1} зависимость).

При заданных α, β и μ идеальный метод дает $N\sigma^2$, тогда как приближенный приводит к

$$N\sigma'^2 = f(e_0) \tag{8.49}$$

— функции пороговой эффективности. Определим величины:

$N\sigma'^2_{\text{мин}}$ = минимум $f(e_0)$, соответствующий порогу $e_{\text{мин}}$ и доле $r_{\text{мин}}$ % выброшенных событий;

$$N\sigma'^2_1 = \text{значению при } e_0 = 0{,}1;$$

$$N\sigma'^2_2 = \text{значению при } r = 1\%.$$

Отношения

$$R_0 \equiv \frac{\sigma'^2_{\text{мин}}}{\sigma^2}, \quad R_1 \equiv \frac{\sigma'^2_1}{\sigma'^2_{\text{мин}}}, \quad R_2 \equiv \frac{\sigma'^2_2}{\sigma'^2_{\text{мин}}},$$

оказывается, почти целиком зависят только от μ, а не от α или β. Кроме того, $e_{\text{мин}}$ и $r_{\text{мин}}$ являются функциями практически только μ. На рис. 8.2 показан график функции (8.49) для $\alpha = \beta = 0{,}3$ и $\mu = 0{,}99$. Он отражает существование оптимума при

$$e_{\text{мин}} = 0{,}25, \quad r_{\text{мин}} = 15\%;$$

$$N\sigma'^2_{\text{мин}} = 3{,}18, \quad N\sigma^2 = 2{,}50.$$

(Кривая на рис. 8.2 остается почти неизменной с точностью до некоторого множителя при различных значениях α и β. На выбор величины μ влияет то обстоятельство, что $N\sigma'^2$ расходится при $e_0 = 0$ и $\mu \geqslant 1$.)

Таким образом, $e_{\text{мин}}$, $r_{\text{мин}}$ приводят к $R_0 = 1{,}25$, $e = 0{,}1$ — к $R_\perp = 1{,}15$ и $r = 1\%$ — к $R_2 = 1{,}38$.

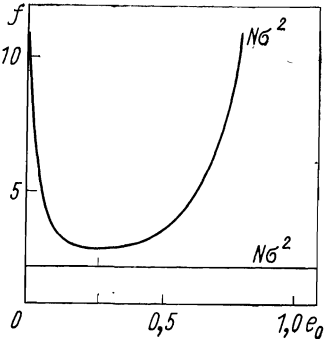


Рис. 8.2. Зависимость дисперсии (8.49) в приближенном методе от пороговой эффективности e_0 ($\alpha = \beta = 0{,}3$; $\mu = 0{,}99$)

8.5.4. Метод наименьших квадратов. Рассмотрим гистограмму с числом событий k_i в i-й ячейке, $i = 1, \ldots, N$. Предположим, модель дает выражение для нормировки и распределения. Пусть ожидаемое число событий в i-й ячейке равно:

$$a_i(\theta) = N(\theta) \frac{\int_i PeQ}{\int PeQ}, \tag{8.50}$$

где $N(\theta)$ $\left(\neq \sum_{i=1}^{N} k_i\right)$ — полное число предсказываемых событий, а $\int_i$ означает интегрирование по ячейке i. В нашем распоряжении

имеются два метода (8.36) и (8.35), для которых

$$Q_1^2 = \sum_{i=1}^{N} \frac{(k_i - a_i)^2}{k_i}, \quad \frac{\partial Q_1^2}{\partial \theta} = -2 \sum_{i=1}^{N} \left(1 - \frac{a_i}{k_i}\right) \frac{\partial a_i}{\partial \theta};$$

$$Q_2^2 = \sum_{i=1}^{N} \frac{(k_i - a_i)^2}{a_i}, \quad \frac{\partial \theta_2^2}{\partial \theta} = -\sum_{i=1}^{N} \left[\left(\frac{k_i}{a_i}\right)^2 - 1\right] \frac{\partial a_i}{\partial \theta}.$$

Используя метод и обозначения п. 7.3.3 (2), можно показать, что как для Q_1^2, так и для Q_2^2 асимптотически $E[\xi(\theta_0)] = 0$, где $\xi = N^{-1}\, \partial Q^2/\partial\theta$, поскольку

$$E\left(\frac{1}{k_i}\right) \to \frac{1}{a_i}, \quad E\left(\frac{1}{k_i^2}\right) \to \frac{1}{a_i^2}.$$

Введем сейчас приближенный метод, который обходится без Q в интегралах. Пусть b_i — предсказываемое число событий в i-й ячейке, когда $e = 1$ (a_i — предсказываемое число событий, когда $e < 1$). Скорректировав числа с помощью известной экспериментальной эффективности, получим числа c_i, такие, что

$$E\,(c_i) = a_i. \tag{8.51}$$

Тогда

$$b_i = N'(\theta)\, \frac{\int_i PQ}{\int PQ} = N(\theta)\, \frac{\int_i PQ}{\int PeQ}. \tag{8.52}$$

Здесь N' — полное число событий, предсказываемое для случая $e = 1$. Из соотношений (8.50) и (8.52) следует

$$a_i = b_i\, \frac{\int_i PeQ}{\int_i PQ}.$$

Это отношение интегралов, которое должно быть рассчитано, является как бы ожиданием $E_i(e)$ эффективности e для i-й ячейки или средним $[E_i\ (1/e)]^{-1}$ веса $w = 1/e$. Для того чтобы оценить его, возьмем арифметическое среднее событий, которые *на эксперименте* попали в i-ю ячейку. Другими словами, нормировкой является $\int_i PeQ$, т. е.

$$\frac{\int_i PQ}{\int_i PeQ} = E_i\left(\frac{1}{e}\right) \approx \frac{\sum_{j=1}^{k_i} w_{ij}}{k_i},$$

где $w_{ij} = 1/e_i$ для j-го события в i-й ячейке (суммирование проводится по всем событиям этой ячейки).

Очевидно, задача несимметрична по отношению к оценкам $E_i(e)$ и $[E_i(1/e)]^{-1}$, поскольку величина $E_i(e)$ может быть исполь-

зована только для выборки событий с эффективностью регистрации $e = 1$. Наконец, мы получаем

$$c_i = \frac{b_i k_i}{\sum_{j=1}^{k_i} w_{ij}},$$

где выражение для b_i точно известно. Полученное соотношение удовлетворяет уравнению (8.51). Тогда минимизируемая функция приобретает следующий вид:

$$Q^2 = \sum_{i=1}^{N} \frac{1}{\sigma_1^2}\left(k_i - b_i \frac{k_i}{\sum_j w_{ij}}\right)^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i'^2}\left(\sum_{j=1}^{k_i} w_{ij} - b_i\right)^2 \qquad (8.53)$$

и

$$\frac{1}{\sigma_i'^2} = \frac{1}{\sigma_i^2}\left(\frac{k_i}{\sum_j w_{ij}}\right)^2.$$

Используем второе выражение в правой части (8.53). Для $\sigma_i'^2$ имеем

$$\sigma_i'^2 = E\left[\left(\sum_{j=1}^{k_i} w_{ij} - b_i\right)^2\right] = E\left[\left(\sum_{j=1}^{k_i} w_{ij}\right)^2\right] - b_i^2,$$

поскольку $E\left[\sum_j w_{ij}\right] = b_i$.

Используя тот факт, что k_i распределено по закону Пуассона, можно показать

$$E\left[\left(\sum_{j=1}^{k_i} w_{ij}\right)^2\right] = E(k_i)\, E(w_i^2) + b_i^2.$$

Оценка $E(w_i^2)$ может быть рассчитана как арифметическое среднее:

$$E(w_i^2) \simeq \frac{1}{k_i} \sum_{j=1}^{k_i} w_{ij}^2.$$

Что касается оценки $E(k_i)$, то она может быть найдена двумя методами, соответственно двум различным способам (8.36) и (8.35):

$$E(k_i) = k_i,$$

это дает

$$Q_1'^2 = \sum_{i=1}^{N}\left[\left(\sum_{j=1}^{k_i} w_{ij} - b_i\right)^2 \Big/ \sum_{j=1}^{k_i} w_{ij}^2\right]$$

или $E(k_i) = a_i$ соответственно $Q_2'^2 = \sum_{i=1}^{N}$. Очевидно, Q' стремится к Q при $w \to 1$.

Как и в обычном (без весов) случае, использование $Q_2'^2$ лучше обосновано, но по практическим соображениям можно предпочесть $Q_1'^2$ (процесс минимизации требует меньших затрат времени на ЭВМ).

Этот метод дает особый выигрыш, если b_i является линейной функцией параметров, поскольку решение системы может быть получено (записано) точно. Во всех других случаях приходится прибегать к итерационным процедурам, одинаково медленным для $Q_1'^2$ и для $Q_2'^2$.

Применение общего результата п. 7.3.2 приводит к довольно сложной формуле для дисперсии оцениваемых параметров.

§ 8.6. УМЕНЬШЕНИЕ СМЕЩЕНИЯ

В п. 7.4.2 уже обсуждалась связь между смещением и минимальной дисперсией (максимальной информацией). В приведенном там примере уменьшение смещения достигалось заменой переменных. Мы пришли к заключению, что в общем случае уменьшение смещения может быть достигнуто только ценой некоторой потери информации. Ниже обсудим два общих метода уменьшения смещения.

8.6.1. Точное распределение оценки известно. Если ф. п. в. оценки полностью известно, то смещение b также известно и ее можно вычислить по (7.1). Тогда вместо смещенной оценки t можно использовать несмещенную оценку:

$$t_1 = t - b. \tag{8.54}$$

Дисперсия не изменяется:

$$D\ (t_1) = D\ (t - b) = D\ (t).$$

Как правило, b зависит от неизвестных параметров ф. п. в. экспериментальных наблюдений и, следовательно, также должно быть найдено. Несмещенная оценка параметра $\hat{b}$ в (8.54) увеличивает дисперсию окончательной оценки. Но такая потеря точности может быть вполне допустимой.

Пример. Пусть имеется выборка из N независимых наблюдений $X_1, \ldots, \ldots, X_N$, распределенных по нормальному закону с неизвестными средним и дисперсией. Найдем оценку дисперсии этого нормального распределения методом максимального распределения.

Ф.п.в. имеет вид

$$f(X, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[- \frac{(X-\mu)^2}{2\sigma^2} \right].$$

Функция правдоподобия равна

$$L(X, \sigma^2) = \frac{1}{(2\pi)^{N/2}\,\sigma^N} \exp\left[- \frac{1}{2\sigma^2} \sum_{i=1}^{N} (X_i - \mu)^2 \right].$$

Нетрудно видеть, что оценка максимального правдоподобия равна

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{N}X_i^2-\left(\frac{\Sigma X_i}{N}\right)^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2.$$

(В (4.18) эта оценка была обозначена как S^2). Смещение находится из

$$E(\hat{\sigma^2})=\frac{1}{N}NE(X_i-\overline{X})^2=E[(X_i-\mu)^2+(\overline{X}-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)]=$$

$$=\sigma^2+\frac{\sigma^2}{N}-\frac{2\sigma^2}{N}=\sigma^2-\frac{\sigma^2}{N}.$$

Таким образом, смещение оценки максимального правдоподобия равно $-\sigma^2/N$. Тогда несмещенной оценкой для σ^2 (для любого распределения) является [см. уравнение (4.27)]:

$$s^2=\frac{N}{N-1}\hat{\sigma}^2=\frac{1}{N-1}\sum_{i=1}^{N}(X-\overline{X})^2=\hat{\sigma}^2+\frac{\hat{\sigma}^2}{N-1}.$$

Теперь можно сравнить дисперсии этих оценок. Для нормального распределения можно рассчитать дисперсию оценки $\hat{\sigma}^2$ [2]:

$$D(\hat{\sigma}^2)=2\sigma^4(N-1)/N^2.$$

Соответственно дисперсия равна

$$D(s^2)=\left(\frac{N}{N-1}\right)^2D(\hat{\sigma}^2)=\frac{2\sigma^4}{(N-1)}.$$

Таким образом, избавляясь от смещения, мы увеличили дисперсию оценки на величину порядка $1/N^2$, т. е. на

$$\frac{2\sigma^2}{N^2}\left(1+\frac{1}{N-1}\right).$$

Соответственно стандартное отклонение при этом возросло на

$$\sqrt{D(s^2)}-\sqrt{D(\hat{\sigma}^2)}=\sqrt{D(\hat{\sigma}^2)}\left[\frac{N}{N-1}-1\right]=\frac{\sigma^2}{N}\sqrt{\frac{2}{N-1}}.$$

При больших N такая потеря в точности гораздо меньше смещения.

8.6.2. Точное распределение оценки неизвестно. Если детальный вид ф. п. в. неизвестен, то существуют довольно мощные методы уменьшения смещения с точностью до $1/N^2$ [38]. Приведем здесь наиболее простой из них.

Предположим, что t_N является оценкой параметра θ, обладающей смещением, и что мы выразили ее в виде ряда по степеням $1/N$ для больших N. Тогда главным членом будет θ (независимо от N), а следующим — член, соответствующий дисперсии и уменьшающийся, как $N^{-1/2}$. Среднее значение этого члена равно нулю. Главный член смещения (среднее значение не равно нулю для конечных N) обычно ведет себя, как N^{-1}, и именно этот член можно исключить при более тщательном анализе данных. В разложении, о котором

говорилось выше, рассмотрим главные члены с ненулевыми средними для конечных N:

$$E(t_N) = \theta + \frac{1}{N}\beta + O\left(\frac{1}{N^2}\right);$$

$$E(t_{2N}) = \theta + \frac{1}{2N}\beta + O\left(\frac{1}{N^2}\right).$$

Здесь смещение равно $b = \beta/N$. Комбинируя эти два значения, получаем

$$E(2t_{2N} - t_N) = \theta + O(1/N^2).$$

Таким образом, используя соответствующим образом t_{2N} (оценку по всем данным) и t_N (оценку по половине данных), мы исключили главный член смещения. Можно распорядиться данными еще лучше, если разделить их на две равные половины, получить оценки t_N и t'_N, вычислить оценку по формуле

$$2t_{2N} - \frac{1}{2}(t_N + t'_N).$$

Как и в предыдущем случае, дисперсия обычно возрастает на величину порядка $1/N$. Существует метод, при котором дисперсия возрастает только на величину порядка $1/N^2$ [2]. К несчастью, этот метод не очень полезен, поскольку он требует N различных оценок.

Пример. Применим этот метод к примеру в п. 8.3.3 (см. рис. 8.1). Перенумеруем все события выборки в порядке возрастания их величин и разделим всю выборку случайным образом на две половины. В одну из этих половин попадает X_N и для нее

$$t_M = t_{2M} = X_N.$$

Среднее значение оценки для другой половины:

$$t'_M = \frac{1}{2}X_{N-1} + \frac{1}{4}X_{N-2} + \frac{1}{8}X_{N-3} + \cdots$$

Здесь 1/2 в точности обозначает вероятность, что X_{N-1} является наибольшим значением X во второй выборке, 1/4 — то же самое для X_{N-2} и т. д. В среднем расстояние между соседними значениями X равно X_N/N. Тогда для больших N мы получаем

$$\hat{\theta}_{corr} \simeq 2X_N - \frac{X_N + \left(X_N - \dfrac{2X_N}{N}\right)}{2} \simeq X_N + \frac{X_N}{N}.$$

Такая оценка уже не смещена.

§ 8.7. УСТОЙЧИВЫЕ (НЕ ЗАВИСЯЩИЕ ОТ РАСПРЕДЕЛЕНИЯ) ОЦЕНКИ

В предыдущих параграфах мы рассмотрели задачу оценки параметров, предполагая известным вид ф. п. в. Если вид ф. п. в. точно неизвестен, то возникают следующие вопросы:

1. Оценки каких параметров могут быть найдены без каких-либо предположений о форме ф. п. в.?

2. Насколько надежны оценки параметров, если выбранный вид ф. п. в. не совсем правилен, например, если данные включают неверно идентифицированные события.

По традиции оценки типа 1 называют *не зависящими от распределения*. Однако этот термин может ввести в заблуждение [39, 40], поскольку свойства таких оценок (особенно дисперсия) могут сильно зависеть от распределения. Поэтому более правильным будет термин *устойчивая оценка*. Обычно под *устойчивостью* принято понимать нечувствительность к малым отклонениям от точного распределения.

Об устойчивых оценках [41] известно относительно мало, несмотря на их практическую важность для анализа данных. Поэтому мы ограничимся описанием нескольких методов, которые дают более устойчивые оценки, чем обычно используемые. Если виды точных распределений и рассматриваемых оценок могут быть ограничены некоторыми общими классами, то иногда становится возможным определить «оптимально устойчивые» оценки.

Существует тесная связь между методами нахождения устойчивых оценок, которые излагаются здесь, и проверками гипотез, не зависящими от вида распределения. Этот вопрос обсуждается в гл. 11. Хотя проблемы нахождения оценок и проверки гипотез могут быть сформулированы так, что они не зависят от вида распределения, тем не менее дисперсии таких оценок и мощность таких критериев проверки гипотез сильно зависят от вида распределения.

8.7.1. Устойчивая оценка центра распределения. Единственный вопрос проблемы устойчивых оценок, который интенсивно обсуждался в литературе, — это оценка центра неизвестного симметричного распределения. Центром распределения может быть определен любой из *параметров расположения*, например, *среднее*, медиана, мода, полусумма крайних значений*. Не у каждого распределения существуют все эти величины, например, распределение Коши не имеет среднего (см. п. 4.2.11). Среднее выборки является наиболее обычной и наиболее часто используемой оценкой положения, так как:

1) она состоятельна всегда, когда дисперсия соответствующего распределения конечна (закон больших чисел § 3.3);

---

* Медиана — это такое значение X, для которого функция распределения $F(X) = 0{,}5$. Мода — такое значение X, при котором ф. п. в. максимальна. Если область возможных значений X ограничена значениями $[X_{\text{мин}}, X_{\text{макс}}]$, то *полусумма* крайних значений равна $(X_{\text{мин}} + X_{\text{макс}})/2$.

2) она оптимальна (не смещена и обладает минимальной дисперсией), если соответствующее распределение нормально. Однако, если соответствующее распределение не нормально, то среднее выборки не является наилучшей оценкой. Ниже мы перечисляем наилучшие оценки параметра расположения для некоторых важных распределений:

Распределение	Оценка расположения с минимальной дисперсией
Нормальное	Среднее выборки
Равномерное	Полусумма крайних значений
Коши	Оценка максимального правдоподобия
Двойное экспоненциальное	Медиана (среднее значение)

Конечно, все эти оценки являются оценками максимального правдоподобия, но в случае распределения Коши оценка не может быть выражена более простым (не зависящим от распределения) образом. Как и в п. 7.4.2, определим *эффективность* оценки, как отношение границы минимальной дисперсии к оценке дисперсии. Тогда эффективность (асимптотическая) названных выше оценок равна 1 для соответствующего распределения. Сравнивая эффективности, можно получить устойчивость этих оценок применительно к другим распределениям. (В табл. 8.5 даны асимптотические значения эффективностей. Задача также была решена и для малых N [42]. Асимптотические значения дисперсий приведены в табл. 8.6.)

Таблица 8.5

Асимптотические эффективности оценок параметра положения

Распределение	Медиана	Среднее	Полусумма крайних значений
Нормальное	0,64	1,00	0,00
Равномерное	0,00	0,00	1,00
Коши	0,82	0,00	0,00
Двойное экспоненциальное	1,00	0,50	0,00

Ни одна из трех рассмотренных здесь оценок не имеет отличную от нуля асимптотическую эффективность для всех четырех распределений. Причина состоит в том, что среднее симметричных распределений, определенных на ограниченной области значений (равномерное, треугольное, β-распределение и т. д.), определяется указанием концевых точек, поэтому большая часть информации содержится в хвостах. В этом случае отличной оценкой параметра расположения является полусумма крайних значений. С другой стороны, для распределений, определенных на неограниченной области значений, полусумма крайних значений всегда имеет нулевую асимптотическую эффективность.

Так как обычно всегда известно, ограничена или не ограничена область значений, нет серьезных ограничений рассматривать только распределения с неограниченной областью значений. Предположим, что соответствующее распределение нормально, с неизвестной (но ненулевой) примесью событий, распределенных в соответствии с распределением Коши. Тогда среднее выборки имеет бесконечную дисперсию, но медиана (в соответствии с табл. 8.5) имеет по меньшей мере 64%-ную эффективность независимо от величины примеси. В общем случае для распределений с неограниченной областью значений медиана выборки служит более устойчивой оценкой расположения, чем среднее выборки. Ниже мы обсуждаем некоторые другие оценки, которые даже более устойчивы для этого класса распределений.

8.7.2. Выравнивание и винсоризация. Введем две новые оценки:

1. *Выровненным* средним выборки в N наблюдений назовем среднее выборки объемом $N - m$ наблюдений, оставшихся после выбрасывания $m/2$ наибольших и $m/2$ наименьших значений.

2. *Винсоризованным* средним выборки в N наблюдений называется среднее выборки объемом N наблюдений, полученной из предыдущей выборки путем замены наибольших $m/2$ и наименьших значений наибольшим и наименьшим значением в соответственно оставшейся части выборки объемом $N - m$.

Для обеих оценок имеется один свободный параметр, в качестве которого обычно выбирают половину доли оставшихся (неизменяющихся или неотбрасываемых) значений:

$$r = (N - m)/2N.$$

Отметим, что при $r = 0{,}5$ обе оценки фактически являются средними выборки и при $r \to 0$ обе оценки эквивалентны медиане выборки. На рис. 8.3 проводится сравнение асимптотических эффективностей выровненных и винсоризованных средних как функций r. Выбор наилучшего значения r аналогичен проблеме выбора правил решения (см. § 6.2), поскольку *априори* неизвестен вид распределения. Для оптимальной устойчивости должно быть использовано правило минимакса: сделать максимальной минимально возможную эффективность*. В соответствии с рис. 8.3 такой подход дает $r = 0{,}23$ для выровненного среднего; при этом значении по меньшей мере гарантирована эффективность 82% независимо от того, какое из трех распределений имеет место.

Видно, что винсоризованное среднее менее устойчиво, поскольку точка минимакса гарантирует эффективность только 71%.

8.7.3. Обобщенные нормы p-й степени. В § 8.4 и в гл. 7 уже обсуждались детально свойства оценок метода наименьших квадратов. В частности, мы знаем, они оптимальны при оценке среднего

---

* При использовании правила минимакса выбирается наиболее пессимистическая точка зрения по поводу возможного распределения. Такой подход совпадает с требованиями, предъявленными к устойчивой оценке.

(или при аппроксимации измеренных точек некоторой линией), если соответствующее распределение (или измерительные ошибки) нормально (см. п. 8.4.4). В этом параграфе мы поговорим об этих оценках в более широком плане, рассмотрев два связанных вопроса.

а. Каковы свойства оценок наименьших квадратов, если соответствующие распределения не нормальны?

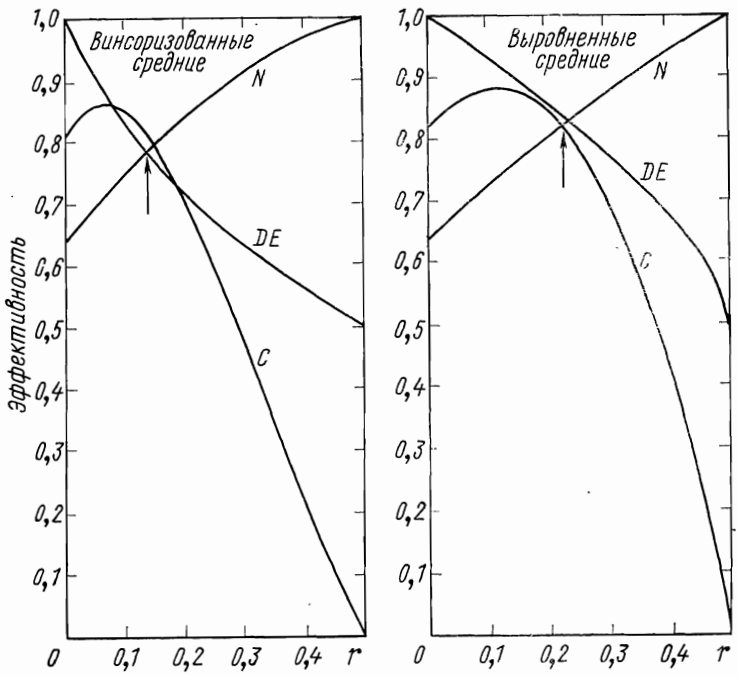


Рис. 8.3. Асимптотические эффективности выровненного и винсоризованного средних для нормального (N), двойного экспоненциального (DE) и Коши (C) распределений

б. Каковы свойства других оценок, которые находятся из условия минимума не квадратов отклонения, а некоторых других степеней? Следует ли иногда использовать наименьшие кубы, наименьшие абсолютные значения и т. д.?

Опять-таки исследуем проблему оценки параметра положения; применение рассмотренных ниже методов к другим оценкам, обычно находимых методом наименьших квадратов, очевидно. Назовем величину a_p L_p-оценкой расположения, если a_p минимизирует величину:

$$L_p = \sum_{i=1}^{N} |X_i - a|^p.$$

Для всех $p \geqslant 1$ L_p-оценка существует, и свойства этих оценок уже изучались [43]. Когда $p = 2$, L_p-я оценка является средним выборки, а когда $p = 1$, она — медиана выборки. Если $p = \infty$, в этом случае L_p иногда называется нормой Чебышева, оценка эквивалентна полусумме крайних значений, поскольку только экстремальные значения дают вклад. В табл. 8.6 мы приводим дисперсии этих оценок для различных распределений (дисперсии, обведенные квадратиками, — минимально возможные).

Таблица 8.6

Асимптотические дисперсии L_p оценки положения

Распределение	L_1 (медиана)	L_2 (среднее)	L_∞ (полусумма крайних значений)
Равномерное	$1/4N$	$1/12N$	$\boxed{1/2N^2}$
Треугольное	—	$1/6N$	$4-\pi/4N$
Нормальное	$\pi/2N$	$\boxed{1/N}$	$\pi^2/12 \ln N$
Двойное экспоненциальное	$\boxed{1/2N}$	$2/N$	$\pi^2/12$
Коши	$\pi^2/4N$	∞	∞

Часто уже отмечалось, что поскольку эффективность различных методов оценки определена в терминах дисперсии их оценок, то мы выбираем L_2-критерий для сравнения методов оценки. Эта кажущаяся произвольной процедура в большинстве случаев оправдана ввиду асимптотической нормальности оценок.

Как видно из табл. 8.6, популярность L_2-оценки частично обусловлена важностью нормального распределения (см. п. 4.3.1). Кроме того, среди всех оценок, для которых ожидаемые величины являются линейными функциями наблюдений, L_2 оценка оптимальна для любого распределения с конечной дисперсией. Для распределений с более длинными «хвостами», чем у нормального распределения, оценка с $p < 2$ обычно более эффективна, а для распределений с более короткими «хвостами» более эффективна $p > 2$.

Отметим также своеобразную зависимость от N оценки L_∞. Поскольку L_p оценки в области $1 \leqslant p \leqslant 2$ обладают довольно интересными свойствами, казалось бы, следует расширить определение L_p на область $p < 1$ [44]. Такие оценки неудобны с точки зрения вычислений и обычно они не представляют большого интереса, за исключением важного класса проблем, рассмотренных в следующем параграфе.

8.7.4. Оценки положения для асимметричных распределений. Рассмотрим задачу оценки центра узкого «сигнала», на который накладывается неизвестное, но более широкое асимметрическое «фо-

новое» распределение. Ввиду асимметрии фона трудно использовать выравнивание или винсоризацию.

Общий метод состоит в параметризации сигнала и фона некоторым произвольным выражением. Затем проводится минимизационная процедура метода наименьших квадратов или метода максимального правдоподобия с целью получить оптимальные значения параметров, включая параметр положения. Такой метод не является устойчивым, поскольку оценка положения зависит от параметризации фона и от корреляций.

Устойчивым для такой задачи является метод, при котором оценивается *мода* наблюдаемого распределения, так как она служит хорошо определяемым параметром также и для асимметричного распределения и почти инвариантна при изменении гладкого фона [45, 46].

Устойчивая техника для оценки моды состоит в использовании $a_{-\infty}$ L_p-оценки для $p = -\infty$. Ранее мы уже видели, что с уменьшением p оценка a_p игнорирует информацию в хвостах распределения и приписывает бóльший вес наблюдениям в области с наибольшей плотностью. $L_{-\infty}$-оценка на практике осуществляется следующим образом: находятся два наблюдения, располагающиеся на наименьшем расстоянии друг от друга и выбирается такое из них, для которого среди оставшихся наблюдений находится ближайшая соседняя точка. Положение этого наблюдения затем выбирается в качестве оценки моды.

Асимптотическая эффективность $L_{-\infty}$-оценки почти в шесть раз меньше, чем эффективность L_1-оценки применительно к нормальному распределению, и почти в три раза меньше, чем эффективность L_1-оценки для распределения Коши. С другой стороны, $a_{-\infty}$ совершенно нечувствительна к асимметрии распределения. В качестве примера применим этот метод к χ^2 (6)-распределению (см. п. 4.2.3), для которого мода равна 4, медиана равна 5,348, среднее равно 6. Для $N = 80$ наблюдениям ожидания и дисперсии L_p оценок равны:

$$E(a_{-\infty}) \approx 4{,}1;\ D(a_{-\infty}) \approx 1{,}3;$$
$$E(a_1) \approx 5{,}0;\ D(a_1) \approx 1{,}3;$$
$$E(a_2) \approx 5{,}9;\ D(a_2) \approx 1{,}9.$$

Однако точное распределение $a_{-\infty}$ неизвестно даже для простых распределений и она может плохо вести себя при малом числе наблюдений.

ГЛАВА 9

ОЦЕНКА ПАРАМЕТРОВ ИНТЕРВАЛОМ ЗНАЧЕНИЙ

В гл. 7 и 8, посвященных оценке параметров, мы уже рассмотрели методы оценки неизвестных параметров одним значением. При оценке параметров интервалом значений мы хотим найти область $\theta_a \leqslant \theta \leqslant \theta_b$, которая с вероятностью β содержит истинное значение θ_0. Задача состоит в том, чтобы среди всех таких областей $[\theta_a, \theta_b]$ найти, в некотором смысле, наилучшую при заданном β. Например, возможным критерием для выбора такого интервала может быть требование минимальности его длины среди всех интервалов с вероятностным содержанием β. Такие интервалы называются *доверительными для* θ *с вероятностным содержанием* β.

На языке экспериментальной физики «доверительный интервал для θ» соответствует «ошибке параметра θ».

Экспериментатор всегда хочет быть уверен, что выбранный интервал действительно содержит истинное значение θ_0. С этой целью он выбирает β разумно большим, например, 95 или 99%. В экспериментальной физике общепринято выбирать $\beta = 68{,}3$ или 95,5%, а соответствующие «ошибки» (доверительные интервалы) называются ошибками с одним стандартным отклонением или ошибками с двумя стандартными отклонениями. Такая терминология может ввести в заблуждение, поскольку обычно она верна только для нормального распределения.

Если переменная X распределена в соответствии с ф. п. в. $f(X \mid \theta)$, то вероятностное содержание β области значений $[a, b]$ в пространстве X равно:

$$\beta = p\ (a \leqslant X \leqslant b) = \int_a^b f(X \mid \theta)\, dX. \tag{9.1}$$

Если известны плотность f и параметр θ, то всегда можно рассчитать β для заданных a и b.

Если параметр θ неизвестен, то необходимо найти другую переменную

$$Z = Z(X, \theta),$$

являющуюся функцией наблюдения X и параметра θ, ф. п. в. которой не зависит от неизвестного θ. Если такая переменная может быть найдена, то тогда можно преобразовать (9.1) к виду, соответствующему

задаче оценки интервала, а именно: для заданного β найти оптимальную область $[\theta_a, \theta_b]$ в пространстве θ, для которого

$$p(\theta_a \leqslant \theta_0 \leqslant \theta_b) = \beta. \tag{9.2}$$

Наш подход к проблеме нахождения доверительных интервалов по существу классический: θ_0 — неизвестная константа, а θ_a и θ_b — функции случайной переменной X. Однако в последнем параграфе мы также кратко изложим бейесовскую интерпретацию проблемы доверительных интервалов.

§ 9.1. ДАННЫЕ РАСПРЕДЕЛЕНЫ ПО НОРМАЛЬНОМУ ЗАКОНУ

9.1.1. Доверительные интервалы для среднего. Пусть $f(X \mid \theta)$ в уравнении (9.1) есть ф. п. в. нормального распределения $N(\mu, \sigma^2)$ [см. уравнение (4.16)]. Если μ и σ^2 известны, то уравнение (9.1) может быть записано:

$$\beta = p(a \leqslant X \leqslant b) = \int_a^b N(\mu,\sigma^2)\, dX = \Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma).$$

Здесь Φ — интеграл вероятности нормального распределения [см. уравнение (4.17)].

Если μ неизвестно, то уже нельзя рассчитать вероятностное содержание интервала $[a, b]$. Вместо этого можно вычислить вероятность β того, что X лежит в некотором интервале относительно своего неизвестного среднего, скажем $[\mu + c, \mu + d]$:

$$\beta = p(\mu + c \leqslant X \leqslant \mu + d) = \int_{\mu+c}^{\mu+d} N(\mu, \sigma^2)\, dX' =$$

$$= \int_{c/\sigma}^{d/\sigma} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} Y^2\right] dY = \Phi(d/\sigma) - \Phi(c/\sigma). \tag{9.3}$$

Можно преобразовать вероятностное утверждение в (9.3) с тем, чтобы оно приняло форму утверждения (9.2):

$$\beta = p(X - d \leqslant \mu \leqslant X - c). \tag{9.4}$$

Отметим, что это выражение является вероятностным утверждением о случайной переменной X, хотя оно выглядит так, как будто бы оно задает границы на интервале значений μ.

Когда μ и σ неизвестны, используют *стандартизованную переменную* $Z = (X - \mu)/\sigma$. Известно, что эта переменная распределена по закону *стандартного нормального распределения* $N(0, 1)$, если X распределено нормально. Вероятностное утверждение относительно Z приобретает вид:

$$\beta = p(c \leqslant Z \leqslant d) = \int_c^d \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} Z^2\right] dZ = \Phi(d) - \Phi(c). \tag{9.5}$$

Тогда доверительный интервал (9.4) можно получить преобразованием неравенства для Z в неравенство для μ. Для нормального распределения такое преобразование сделать алгебраически просто вследствие симметрии функции относительно X и μ. Но в общем случае оно не столь просто и даже не всегда алгебраически возможно. В § 9.2 излагаются методы решения подобной задачи в общем случае. А сейчас введем одно полезное понятие: для случайной переменной X с ф. п. в. $f(X)$ и функцией распределения $F(X)$ введем понятие *α-точки*:

$$\int_{-\infty}^{X_\alpha} f(X)\,dX = F(X_\alpha) = \alpha.$$

Очевидно, что интервал $[c, d]$ в соотношении (9.5) тогда может быть записан как

$$[Z_\alpha, Z_{\alpha+\beta}]. \qquad (9.6)$$

Для стандартного нормального распределения этот случай отражен на рис. 9.1. Однако, поскольку интервал (9.6) имеет вероятностное содержание β для любого α, обычно существует бесконечно много доверительных интервалов, удовлетворяющих уравнению (9.5).

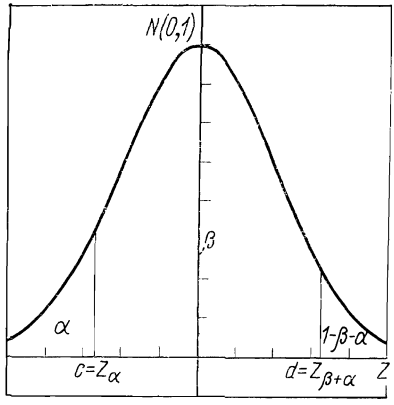


Рис. 9.1. Области с вероятностным содержанием α, β и $1-\beta-\alpha$ в распределении $N(0,1)$: c — α-точка, а d — $(\alpha+\beta)$-точка

Обычно полагают $\alpha = (1 - \beta)/2$, что дает *центральный* (и также минимальный по длине) интервал, симметричный относительно нуля.

Пример. Для $\beta = 0{,}95$, $\alpha = 0{,}025$ имеем $Z_\alpha = -1{,}96$, $Z_{\alpha+\beta} = 1{,}96$, что дает 95%-ный доверительный интервал для μ:

$$[X - 1{,}96\sigma,\ X + 1{,}96\sigma].$$

Для $\beta = 0{,}9545$, $\alpha = 0{,}02275$ получается $Z_\alpha = -2{,}0$, $Z_{\alpha+\beta} = 2{,}0$. Это дает «погрешность» в два стандартных отклонения или 95,45%-ный доверительный интервал:

$$[X - 2\sigma,\ X + 2\sigma].$$

9.1.2. Доверительные интервалы для нескольких параметров. Результаты предыдущих глав свидетельствуют о практической важности нормального распределения. В асимптотике все обычные методы оценок дают нормально распределенные оценки параметров. Поэтому было бы полезным ввести понятие *доверительных областей* с данным вероятностным содержанием β для случая многих параметров.

Предположим, что существует оценка $\mathbf{t}$ параметров $\boldsymbol{\theta}$ и эта оценка распределена по нормальному многомерному закону со средним $\boldsymbol{\theta}$ и матрицей вторых моментов $\underset{\sim}{D}$. Тогда в соответствии с (4.19) ф. п. в. равна:

$$f(\mathbf{t}\,|\,\boldsymbol{\theta}) = \frac{1}{(2\pi)^{N/2}\,|\underset{\sim}{D}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{t}-\boldsymbol{\theta})^T \underset{\sim}{D}^{-1}(\mathbf{t}-\boldsymbol{\theta})\right]. \quad (9.7)$$

Из нормальности распределения оценок $\mathbf{t}$ следует, что *квадратичная форма*

$$Q(\mathbf{t}, \boldsymbol{\theta}) = (\mathbf{t}-\boldsymbol{\theta})^T \underset{\sim}{D}^{-1}(\mathbf{t}-\boldsymbol{\theta})$$

распределена по закону χ^2 (N) (см. п. 4.2.2). Это значит, что распределение Q не зависит от $\boldsymbol{\theta}$. Поэтому можно написать вероятностное утверждение, подобное (9.5), а именно:

$$P[Q(\mathbf{t}, \boldsymbol{\theta}) \leqslant K_\beta^2] = \beta, \quad (9.8)$$

где K_β — β-точка для χ^2 (N)-распределения.

Область в пространстве $\mathbf{t}$, задаваемая уравнением

$$Q(\mathbf{t}, \boldsymbol{\theta}) = K_\beta^2, \quad (9.9)$$

имеет вид гиперэллипсоида, поверхность которого соответствует постоянному значению плотности вероятности для функции (9.7). В соответствии с (9.8) вероятностное содержание этой области равно β. Поскольку Q $(\mathbf{t}, \boldsymbol{\theta})$ симметрично по переменным $\mathbf{t}$ и $\boldsymbol{\theta}$, то вполне понятно, как из вероятностного утверждения (9.8) о $\mathbf{t}$ получить

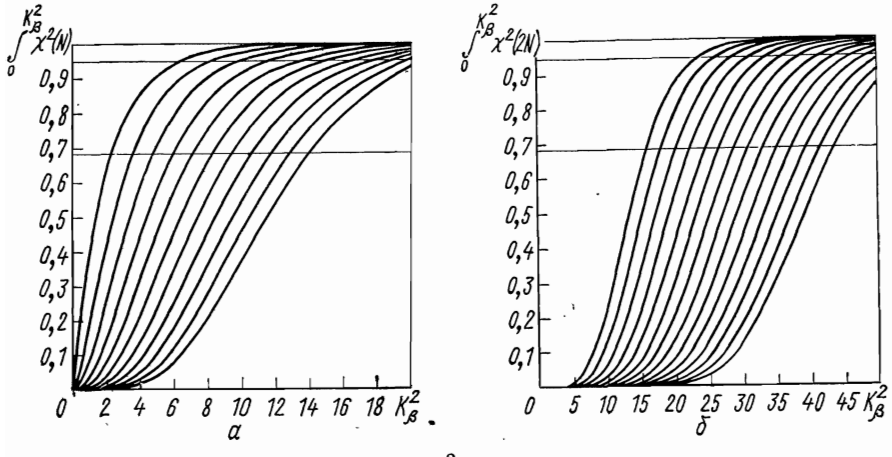


Рис. 9.2. Функции χ^2-распределений $\int_0^{K_\beta^2} \chi^2(N)$ для $N=2, \ldots, 12$ $(а)$ и χ^2-распределений $\int_0^{K_\beta^2} \chi^2(2N)$ для $N=7, \ldots, 20$ $(б)$

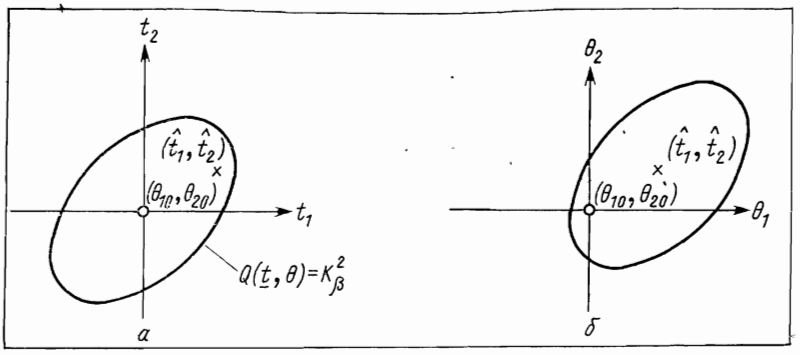


Рис. 9.3. Доверительные зоны с вероятностным содержанием β в **t**-пространстве (*а*) и в $\boldsymbol{\theta}$-пространстве (*б*). Истинные значения — $(\theta_{10}, \theta_{20})$, оценки — $(\hat{t}_1, \hat{t}_2)$

доверительную область для $\boldsymbol{\theta}$. Доверительная область в пространстве переменных $\boldsymbol{\theta}$ с вероятностным содержанием β имеет вид гиперэллипсоида с центром в точке **t** и уравнением (9.9). Приведенные на рис. 9.2 зависимости K_β^2 от N и β позволяют находить такие области.

Двумерный случай достаточно общ, чтобы можно было понять свойства доверительных областей, когда пространство параметров многомерно. Уравнение (9.8) имеет смысл вероятностного утверждения об *одновременных* значениях всех параметров θ и в пространстве двух параметров оно определяет эллиптическую область, показанную на рис. 9.3, *а*. Здесь $\hat{t}_1$, $\hat{t}_2$ — оценки параметров, а θ_{10}, θ_{20} — истинные значения неизвестных параметров θ_1, θ_2. На рис. 9.3, *б* показана соответствующая область в θ-пространстве, полученная в результате преобразования вероятностного утверждения (9.8).

На рис. 9.4 приведены эллиптические области (9.9) в **t**-пространстве для заданного значения β и при различных значениях коэффициента корреляции в предположении, что матрица вторых моментов имеет вид:

$$\mathscr{D} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

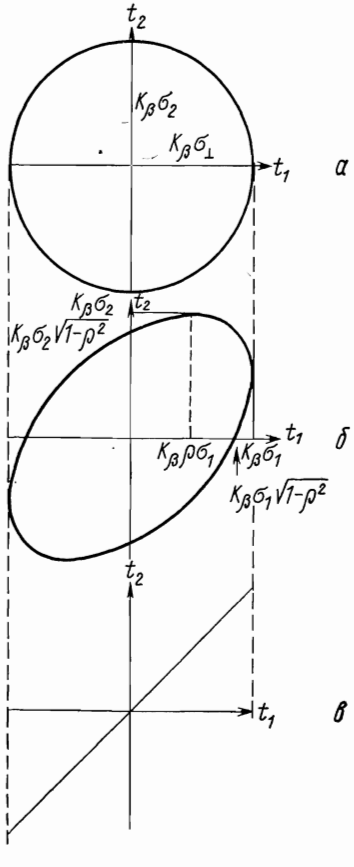


Рис. 9.4. Доверительные области с вероятностным содержанием β для оценок t_1, t_2 (*а*) $\rho=0$, оценки не коррелированы и независимы; корреляция $0<\rho<1$ (*б*) и $\rho=1$ полностью коррелированы; матрица вторых моментов сингулярна (*в*)

В то же время вероятностные утверждения можно было бы делать о различных параметрах независимо. На рис. 9.4 это соответствовало бы прямоугольным областям

$$\theta_1 - K_\beta \sigma_1 \leqslant t_1 \leqslant \theta_1 + K_\beta \sigma_1;$$
$$\theta_2 - K_\beta \sigma_2 \leqslant t_2 \leqslant \theta_2 + K_\beta \sigma_2,$$

включающим в себя эллипсы (отметим, что рис. 9.4, *а* соответствует случаю *независимых переменных*; вероятностное утверждение для одновременных значений θ_1 и θ_2 отражено кругом).

Однако вероятностное содержание описанного прямоугольника очевидно больше, чем эллипса. Кроме того, оно зависит от коэффициента корреляции ρ. На рис. 9.5, *а* нами показана связь между K_β прямоугольных областей с β и ρ вписанного эллипса для нормального распределения.

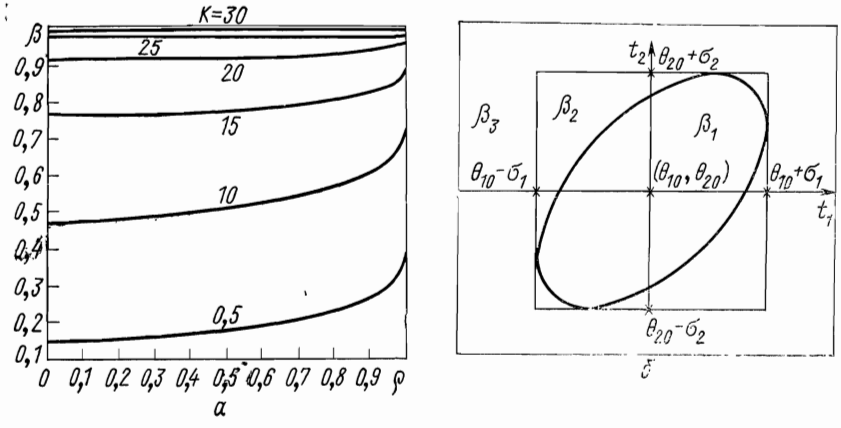


Рис. 9.5. Связь между K_β описанного прямоугольника и вероятностным содержанием β вписанного эллипса с коэффициентом корреляции ρ (*а*) и доверительные области для нормальных оценок t_1, t_2 с $K_{\beta_1}=1$, $\rho=0{,}5$ (*б*).

Вероятностное содержание равно β_1 для эллиптической области [см. уравнение (9.9)], β_2 — для описанного прямоугольника [см. уравнение (9.10)] и β_3 — для горизонтальной полосы [см. уравнение (9.11)]

Рис. 9.5, *б* изображает три различные доверительные области для частного случая $K_{\beta_1} = 1$, $\rho = 0{,}5$. Эллипс соответствует доверительной области (9.9). Описанный вокруг него прямоугольник соответствует доверительной области

$$p\,[(\theta_1 - \sigma_1 \leqslant t_1 \leqslant \theta_1 + \sigma_1) \text{ и } (\theta_2 - \sigma_2 \leqslant t_2 \leqslant \theta_2 + \sigma_2)] = \beta_2 \qquad (9.10)$$

Длинная горизонтальная полоса, ограниченная прямыми $t_2 = \theta_{20} \pm \sigma_2$, задает однопараметрический доверительный интервал $[t_2 - \sigma_2,\ t_2 + \sigma_2]$, соответствующий утверждению, что при любом значении

$$p\,[\theta_2 - \sigma_2 \leqslant t_2 \leqslant \theta_2 + \sigma_2] = \beta_3. \qquad (9.11)$$

По графикам на рис. 9.2 и 9.5, *а* можно найти различные вероятностные содержания: $\beta_1 = 39{,}3\%$; $\beta_2 = 49{,}8\%$; $\beta_3 = 68{,}3\%$.

Таким образом, в многомерном случае $\pm\sigma_i$ стандартное отклонение для одного параметра имеет вероятностное содержание 68,3% только тогда, когда не оцениваются доверительные интервалы *всех*

других параметров. Если доверительные интервалы для двумерного случая выражаются в виде

$$t_1 \pm k\sigma_1,\ t_2 \pm k\sigma_2,$$

то значение k, соответствующее данному вероятностному содержанию β, зависит от коэффициента корреляции между оценками. Даже если коэффициент корреляции равен нулю, то значение k, дающее 68,3%-ный совместный доверительный интервал, равно $k = Z_{0,913} = 1{,}36$ (см. рис. 9.5, *а*). Этот эффект возрастает с ростом числа параметров. Поэтому очень важно понимать, какие параметры оцениваются совместно и какие оценки коррелированы.

Пример. По наблюдаемому спектру масс оценивают массу M и ширину Γ резонанса. Предположим, что получены нормально распределенные оценки с известной матрицей вторых моментов. Если нужно привести доверительные интервалы для M и Γ, то можно поступать по-разному.

Например, если хотят указать вероятность того, что точка (M, Γ) лежит в некоторой области, то можно использовать уравнение (9.9) для эллиптической или уравнение (9.10) для прямоугольных областей. Напротив, если нужно привести доверительный интервал только для M, то можно использовать (9.11), которое при заданном β позволяет установить более компактные границы на M вследствие того, что ничего не говорится относительно Γ. (Это не исключает того, что другой сотрудник на основе тех же самых данных получит границы только для Γ. В результате будет получено два несовместимых результата). Если вы поступаете именно так, что 68%-ный доверительный интервал для M и 68%-ный доверительный интервал для Γ соответствуют только 46% вероятности, что обе области содержат истинное значение одновременно (предполагая, что оценки не коррелированы).

Наконец, третий способ — дать доверительный интервал для M при условии, что известно значение Γ. Предположим, что матрица вторых моментов оценок M и Γ имеет вид:

$$\underset{\sim}{D} = \begin{pmatrix} \sigma_M^2 & \rho\sigma_M\sigma_\Gamma \\ \rho\sigma_M\sigma_\Gamma & \sigma_\Gamma^2 \end{pmatrix}.$$

Тогда условное по отношению к параметру Γ распределение M нормально со средним $M_0 + (\sigma_M/\sigma_\Gamma)\rho\,(\Gamma - \Gamma_0)$ и дисперсией $\sigma_M^2\,(1 - \rho^2)$, где M_0 — истинное значение M. Следовательно, стандартизованная переменная имеет вид:

$$Z = \frac{M - \left[M_0 + \dfrac{\rho\sigma_M}{\sigma_\Gamma}(\Gamma - \Gamma_0)\right]}{\sigma_M\sqrt{1-\rho^2}},$$

и можно записать вероятностное утверждение:

$$p\,(Z_\alpha \leqslant Z \leqslant Z_{\beta+\alpha}\,|\,\Gamma) = \beta.$$

Тогда симметричный доверительный интервал для M_0 с вероятностным содержанием β [или $(1/2)Z_{(1+\beta)/2}$ стандартных отклонений], условный по отношению к $\Gamma = \hat{\Gamma}$, равен

$$\hat{M} - \frac{\rho\sigma_M}{\sigma\Gamma}(\hat{\Gamma} - \Gamma_0) - Z_{\frac{1+\beta}{2}}\sigma_M\sqrt{1-\rho^2} \leqslant M_0 \leqslant \hat{M} -$$

$$- \frac{\rho\sigma_M}{\sigma_\Gamma}(\hat{\Gamma} - \Gamma_0) - Z_{\frac{1-\beta}{2}}\sigma_M\sqrt{1-\rho^2}.$$

Этот интервал является функцией истинного значения Γ_0. Смысл конструирования таких интервалов сейчас станет ясным. Если из другого источника (имеются в виду другие данные) известно значение Γ^* для Γ_0, то информация о $\hat{\Gamma}$ уже не требуется для конструирования доверительных интервалов для Γ_0. Вместо этого полученную информацию можно использовать для оценки M_0 путем конструирования условного доверительного интервала

$$\hat{M}-\frac{\rho\sigma_M}{\sigma_\Gamma}(\hat{\Gamma}-\Gamma^*)-z_{\frac{1-\beta}{2}}\sigma_M\sqrt{1-\rho^2}\leqslant M_0\leqslant\hat{M}-$$

$$-\frac{\rho\sigma_M}{\sigma_\Gamma}(\hat{\Gamma}-\Gamma^*)-z_{\frac{1+\beta}{2}}\sigma_M\sqrt{1-\rho^2}.$$

Отметим, что было бы неправильным конструировать условный интервал для M, полагая $\Gamma_0=\hat{\Gamma}$. Условное распределение можно использовать только тогда, когда добавляется некоторая дополнительная информация, в данном случае известное значение Γ^*.

9.1.3. Интерпретация матрицы вторых моментов. В общем многомерном случае задача нахождения доверительной области для подмножества параметров $\boldsymbol{\theta}_{(r)}=(\theta_1,\ \ldots,\ \theta_r)$ может формулироваться так:

1) указать доверительную область независимо от оценок $\mathbf{t}_{(s)}$ других $s=(N-r)$-параметров;

2) указать доверительную область, условную по отношению к $t_{(s)}$-оценкам параметров $\boldsymbol{\theta}_{(s)}=(\theta_{r+1},\ \ldots,\ \theta_N)$, в предположении, что они имеют известные значения $\boldsymbol{\theta}^*_{(s)}$.

Оценки $\mathbf{t}_{(N)}$ распределены по многомерному нормальному закону со средним $\boldsymbol{\theta}_{(N)}$ и матрицей вторых моментов:

$$\underset{\sim}{D}=\begin{pmatrix}\sigma_1^2 & \ldots\ldots & \rho_{1N}\sigma_1\sigma_N\\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \rho_{1N}\sigma_1\sigma_N & \ldots\ldots & \sigma_N^2\end{pmatrix}.$$

Она может быть разбита на подматрицы:

$$\underset{\sim}{D}=\begin{pmatrix}\underset{\sim}{D}_{(rr)} & \vdots & \underset{\sim}{D}_{(rs)}\\ \ldots & & \ldots\\ \underset{\sim}{D}^T_{(rs)} & \vdots & \underset{\sim}{D}_{(ss)}\end{pmatrix}.$$

В первом случае используется тот факт, что распределение оценок $\mathbf{t}_{(r)}=(t_1,\ \ldots,\ t_r)$ нормальное, r-мерное со средним $\boldsymbol{\theta}_{(r)}$ и матрицей вторых моментов $\underset{\sim}{D}_{(rr)}$, образованной r-строками $(1,\ \ldots,\ r)$ и r-столбцами $(1,\ \ldots,\ r)$ матрицы $\underset{\sim}{D}$.

Во втором случае распределение оценок $\mathbf{t}_{(r)}$, условное по отношению к наблюдаемым оценкам $\mathbf{t}_{(s)}$ — нормальное r-мерное со средним

$$E(\mathbf{t}_{(r)}\,|\,\mathbf{t}_{(s)})=\boldsymbol{\theta}_{(r)}+\underset{\sim}{D}_{(rs)}\underset{\sim}{D}^{-1}_{(ss)}(\mathbf{t}_{(s)}-\boldsymbol{\theta}_{(s)})$$

и матрицей вторых моментов $\underset{\sim}{D}_{(rr\,|\,s)}$, полученной из обратной матрицы $\underset{\sim}{D}^{-1}$ вычеркиванием строк $(r+1,\ldots, N)$ и столбцов $(r+1, \ldots, N)$ и последующим обращением оставшейся матрицы. Тогда условные доверительные области задаются соотношением

$$Q(\mathbf{t}_{(r)}, \boldsymbol{\theta}_{(r)} \,|\, \mathbf{t}_{(s)}, \boldsymbol{\theta}^*_{(s)}) \leqslant K^2_\beta,$$

где K^2_β — значение, соответствующее уровню значимости β для $\chi^2(r)$-распределения, и

$$Q(\mathbf{t}_{(r)}, \boldsymbol{\theta}_{(r)} \,|\, \mathbf{t}_{(s)}, \boldsymbol{\theta}^*_{(s)}) = [\mathbf{t}_{(r)} - \{\boldsymbol{\theta}_{(r)} + \underset{\sim}{D}_{(rs)}\, \underset{\sim}{D}^{-1}_{(ss)}\, (\mathbf{t}_{(s)} - \boldsymbol{\theta}^*_{(s)})\}]^T\, \underset{\sim}{D}^{-1}_{(rr|s)} \times$$
$$\times [\mathbf{t}_{(r)} - \{\boldsymbol{\theta}_{(r)} + \underset{\sim}{D}_{(rs)}\, \underset{\sim}{D}^{-1}_{(ss)}\, (\mathbf{t}_{(s)} - \boldsymbol{\theta}^*_{(s)})\}].$$

Опять-таки необходимо отметить, что эта область зависит как от наблюдаемых оценок $\mathbf{t}_{(s)}$, так и от известных значений параметров $\boldsymbol{\theta}^*_{(s)}$.

В качестве примера рассмотрим случай $N = 3$. Условное распределение t_1 при заданных значениях t_2, t_3 нормально со средним

$$E(t_1 \,|\, t_2, t_3, \theta_2, \theta_3) = \theta_1 - \frac{\sigma_1}{\sigma_2}\left(\frac{\rho_{13}\,\rho_{23} - \rho_{12}}{1-\rho^2_{23}}\right)(t_1 - \theta_2) -$$
$$- \frac{\sigma_1}{\sigma_3}\left(\frac{\rho_{12}\,\rho_{23} - \rho_{13}}{1-\rho^2_{23}}\right)(t_1 - \theta_3)$$

и дисперсией

$$D(t_1) \equiv D(t_1 \,|\, t_2, t_3) = \sigma_1^2\left(1 - \frac{\rho^2_{12} + \rho^2_{13} - 2\rho_{13}\,\rho_{12}\,\rho_{23}}{1-\rho^2_{23}}\right).$$

Доверительная область с вероятностным содержанием β, условная по отношению к наблюдениям t_2, t_3, при заданных значениях θ^*_2 и θ^*_3 имеет вид:

$$E(t_1 \,|\, t_2, t_3, \theta^*_2, \theta^*_3) - Z_{\beta+\alpha}\sqrt{D(t_1)} \leqslant \theta_1 \leqslant E(t_1 \,|\, t_2, t_3, \theta^*_2, \theta^*_3) -$$
$$- Z_\alpha \sqrt{D(t_1)}.$$

В случае 2 параметры θ_s как бы «заморожены» при известных значениях θ^*_s. Тогда размер доверительной области «погрешности» всегда меньше, чем в случае 1, поскольку, фиксируя $\boldsymbol{\theta}_{(s)}$, мы увеличиваем информацию о $\boldsymbol{\theta}_{(r)}$. Эта информация передается посредством корреляций между $\mathbf{t}_{(r)}$ и $\mathbf{t}_{(s)}$.

Читателям, желающим более детально ознакомиться с условными распределениями, мы рекомендуем использовать книги Кендалла [1—3].

§ 9.2. ОБЩИЙ ОДНОМЕРНЫЙ СЛУЧАЙ

9.2.1. Доверительные интервалы и зоны. Изложим классический метод «обращения» вероятностного утверждения (9.1). Предположим, что методу оценки t $(\mathbf{X})$ соответствует ф. п. в. $f(t \,|\, \theta)$. Тогда

уравнение (9.1) может быть записано в виде:

$$\beta = P(t_1 \leqslant t \leqslant t_2 \mid \theta) = P[t_1(\theta) \leqslant t \leqslant t_2(\theta)] = \int_{t_1}^{t_2} f(t \mid \theta)\, dt. \quad (9.12)$$

Решение уравнения (9.12) имеет вид интервала в t-пространстве $[t_1(\theta), t_2(\theta)]$, *заранее* известный.

В частности, центральный интервал определяется уравнением

$$\int_{-\infty}^{t_1} f(t \mid \theta)\, dt = \frac{1-\beta}{2} = \int_{t_2}^{\infty} f(t \mid \theta)\, dt.$$

Обратное по отношению к (9.12) вероятностное утверждение имеет вид интервала в θ-пространстве $[\theta_A, \theta_B]$. На рис. 9.6 показан соответствующий пример. Область между $t_1(\theta)$ и $t_2(\theta)$ называется *доверительной зоной*.

Отметим, что доверительная зона *строится горизонтально* [если для всех θ известны t-интервалы, являющиеся решениями уравнения (9.12)], а *читается вертикально*.

По построению вероятностное содержание зоны вдоль любого горизонтального направления (при данном фиксированном θ) равно β. Если из эксперимента получена оценка $\hat{t}$, то требуемый доверительный интервал для θ задается точками пересечения вертикальной линии $t = \hat{t}$ с кривыми $t_1(\theta)$ и $t_2(\theta)$. Тот факт, что интервал $[\theta_A, \theta_B]$ служит доверительным, следует из таких рассуждений.

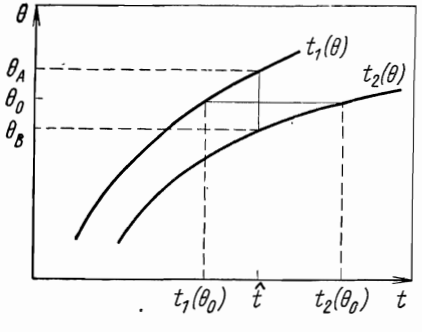


Рис. 9.6. Доверительная зона для параметра θ, полученная с помощью оценки $\hat{t}$

Предположим, что θ_0 — истинное, но неизвестное значение θ. Тогда при большом числе испытаний доля значений $t(X)$, попадающих в интервал $t_1(\theta_0) \leqslant t \leqslant t_2(\theta_0)$, равна β по определению. Но из рис. 9.6 ясно, что интервалы $[\theta_A, \theta_B]$, соответствующие значениям t в $[t_1(\theta_0), t_2(\theta_0)]$ и содержащие истинное значение θ_0, составляют долю β общего числа интервалов. Другими словами, мы можем утверждать, что

$$P(\theta_B \leqslant \theta_0 \leqslant \theta_A) = \beta.$$

В патологических случаях доверительная зона может иметь изгибы, так что результирующий доверительный интервал состоит из нескольких частей. Несмотря на то что такой результат математически правилен, к нему нужно относиться осторожно с тем, чтобы он имел разумную интерпретацию.

Рассмотрим другой подход к этой задаче. На рис. 9.7 показана функция распределения $F(t|\theta)$ переменной t при различных θ. Если в результате эксперимента зафиксировано наблюдение $\hat{t}$, то через точку $\hat{t}$ на оси t проводят вертикальную линию и выделяют интервал длиной AB. Затем выделяют такие функции распределения $F(t|\theta_A)$ и $F(t|\theta_B)$, графики которых проходят через точки A и B. Они и дают доверительный интервал

$$P(\theta_B \leqslant \theta_1 \leqslant \theta_A) = \beta.$$

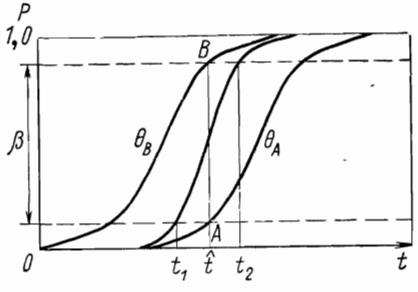


Рис. 9.7. Доверительная зона рис. 9.6 на языке функции распределения $F(t|\theta)$

9.2.2. Доверительные границы (верхний и нижний пределы). Как уже отмечалось, для каждой задачи существует бесконечно много доверительных интервалов и с точки зрения статистики центральный интервал не обладает какими-то особыми достоинствами. Для некоторых задач даже более предпочтительны (с точки зрения физики) предельно нецентральные интервалы, для которых верхняя или нижняя доверительная *граница* равна $+\infty$ или $-\infty$, например: $P(\theta \geqslant t_1) = \beta$. Такая граница может быть полезна тогда, когда хотят узнать, насколько мал может быть какой-то параметр.

Обычно нецентральные доверительные интервалы больше полезны, когда область значений параметров ограничена с той или иной стороны.

Пример. Рассмотрим задачу нахождения нижней доверительной границы для среднего λ пуассоновского распределения. По такому закону распределено число событий N в ячейке гистограммы, когда вероятность попадания в такую ячейку мала (см. п. 4.1.1, 4.1.3):

$$P(N = r|\lambda) = e^{-\lambda}\lambda^r/r!$$

Обозначим $\lambda_.$ и $\lambda^.$ соответственно нижнюю и верхнюю границы.

Нижняя граница для λ является решением уравнения

$$\beta = P(N \geqslant N_0|\lambda_.) = \sum_{r=N_0}^{\infty} e^{\frac{-\lambda_. \lambda_.^r}{r!}} \qquad (9.13)$$

для заданного β. Такое уравнение имеет единственное решение, поскольку правая часть уравнения (9.13) является функцией распределения $\chi^2_\beta(2N_0)$. Поэтому обратное по отношению к (9.13) вероятностное утверждение имеет вид

$$\beta = P[\chi^2(2N_0) \leqslant 2\lambda_.]$$

или

$$\lambda_. = \frac{1}{2}\chi^2_\beta(2N_0),$$

где $\chi^2_\beta (2N_0)$ — β-точка χ^2-распределения с $2N_0$ степенями свободы. По таблицам $\chi^2 (N)$ или по рис. 9.2 можно найти 95%-ные нижние границы $\lambda_.$ для заданного N_0. В табл. 9.1 приведено несколько таких значений.

Таблица 9.1

Зависимость нижней доверительной границы от нижней границы числа событий

N_0	1	2	3	4	5	10	20	50
$\lambda_.$	0,051	0,355	0,82	1,37	1,97	5,42	13,25	38,96

Тогда, например, при $N_0 = 2$ можно записать

$$P\,(N < 2 \mid \lambda = 0{,}355) = 0{,}95$$

и для всех $\lambda > 0{,}355$ $P\,(N < 2 \mid \lambda) > 0{,}95$.

Важно отметить, что доверительное утверждение

$$P\,(\lambda > \lambda_. \mid N_0) = 0{,}95$$

является вероятностным утверждением относительно $\lambda_.$, которое служит функцией наблюдения N_0.

§ 9.3. ИСПОЛЬЗОВАНИЕ ФУНКЦИИ ПРАВДОПОДОБИЯ

9.3.1. Параболический логарифм функции правдоподобия. Рассмотрим случай нормального распределения с единичной дисперсией и неизвестным средним μ. Тогда правдоподобие для совокупности N наблюдений дается уравнением (5.7) с $\sigma = 1$.

Как функция μ, $L\,(X \mid \mu)$ имеет такую же хорошую форму, что и функция плотности нормального распределения. График логарифма функции правдоподобия имеет вид параболы в зависимости от μ:

$$\ln L\,(X \mid \mu) = \ln c - \frac{1}{2}\sum_{i=1}^{N}(X_i - \overline{X})^2 - \frac{N}{2}(\mu - \overline{X})^2$$

с максимумом при $\mu = \overline{X}$. На рис. 9.8 показан случай $N = 1$; при этом произведена замена переменных с тем, чтобы $\ln L = 0$ в максимуме. Горизонтальная линия, проведенная при значении $\ln L = -0{,}5$, задает интервал $\overline{X} - 1 \leqslant \mu \leqslant \overline{X} + 1$ [из условия $(\mu - \overline{X})^2/2 = {}^1/_2$].

Из свойств нормального распределения следует, что

$$P\,[(\overline{X} - \mu)^2 \leqslant 1] = 68{,}3\%$$

или

$$P\,[-1 \leqslant \overline{X} - \mu \leqslant 1] = 68{,}3\%,$$

или

$$P\,[\overline{X} - 1 \leqslant \mu \leqslant \overline{X} + 1] = 68{,}3\%.$$

Соответственно прямая, проведенная при значении $\ln L = -2$, соответствует 95,5%-ному доверительному интервалу для μ. Как и прежде, вероятностные утверждения основаны на свойствах случайной переменной $\overline{X}$, а не на каких-либо свойствах параметра μ. В частности, интервал имеет одни и те же свойства, независимо от

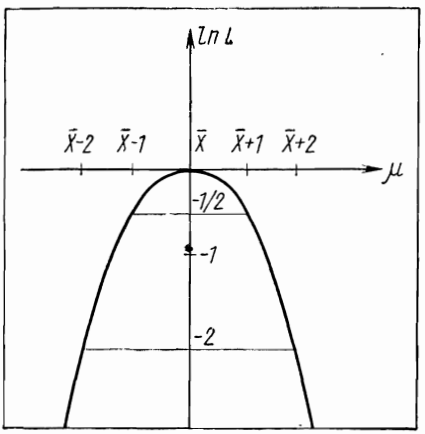


Рис. 9.8. Случай, когда логарифм функции правдоподобия параболичен

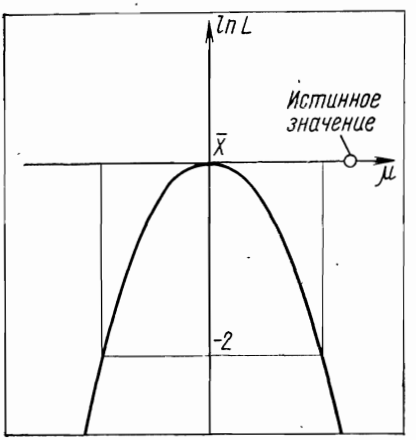


Рис. 9.9. Пример 95,5%-го доверительного интервала, который не содержит истинного значения

числа испытаний. Таким образом, в 4,5% испытаний интервал, полученный в результате анализа функции правдоподобия, *не будет включать истинное значение* (рис. 9.9).

9.3.2. Непараболические функции правдоподобия. Предположим, что в результате двух экспериментов получены две группы данных $\mathbf{X}_1$ и $\mathbf{X}_2$ и цель каждого эксперимента состояла в определении одного неизвестного параметра θ_1 и θ_2 соответственно.

Если функции правдоподобия $L_1(\mathbf{X}_1 \mid \theta_1)$ и $L_2(\mathbf{X}_2 \mid \theta_2)$ равны, тогда мы должны делать одни и те же выводы о θ_2, как и о θ_1. Так же как и оценки, доверительные интервалы и т. д. должны согласовываться. То же самое происходит, если две функции пропорциональны:

$$L_1(\mathbf{X}_1 \mid \theta_1) = kL_2(\mathbf{X}_2 \mid \theta_2),$$

где k не зависит от θ_1 и θ_2. Это свойство было названо в п. 8.3.1 *инвариантностью* оценок максимального правдоподобия.

Рассмотрим сейчас более общий случай, чем в п. 8.3.1, и предположим, что $\ln L(\mathbf{X} \mid \theta)$ непрерывная функция θ и имеет только один максимум в области возможных значений. Тогда существуют

преобразования g (θ, $\mathbf{X}$), которые преобразуют кривую $\ln L$ ($\mathbf{X} \mid \theta$) в параболу вблизи функции G наблюдений:

$$\ln L_g(\mathbf{X} \mid g) = \frac{1}{2}[g - G(\mathbf{X})]^2.$$

В асимптотическом случае доказывается, что g можно выбрать независимо от $\mathbf{X}$. В таком случае G служит оценкой максимального правдоподобия для g.

Используя свойство инвариантности, можно сделать такие же выводы о параметре g, какие делались о параметре нормального рас-

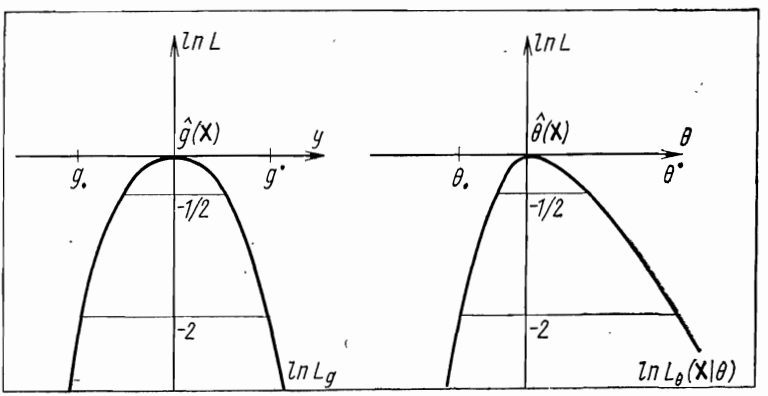


Рис. 9.10. Преобразование непараболической функции правдоподобия в параболическую

пределения μ. В частности, 95,5%-ный доверительный интервал, скажем $g_* \leqslant g \leqslant g^*$, равен:

$$G(\mathbf{X}) - 2 \leqslant g \leqslant G(\mathbf{X}) + 2,$$

где $G(\mathbf{X}) = \hat{g}$.

На рис. 9.10 интервал $[g_*, g^*]$ соответствует пересечению графика функции $\ln L_g$ с прямой $\ln L = -2$. Преобразуем этот доверительный интервал для g в соответствующий интервал для θ. Поскольку

$$\ln L_\theta(\mathbf{X} \mid \theta) = \ln L_g[\mathbf{X} \mid g(\theta)],$$

то значения θ, соответствующие интервалу $[g_*, g^*]$, являются точками пересечения $\ln L = -2$ с графиком функции $\ln L_\theta(\mathbf{X} \mid \theta)$ (см. рис. 9.10).

Таким образом, используя свойства функции правдоподобия и оценок м. п., можно делать выводы о параметре непараболической функции правдоподобия, *не находя соответствующего преобразования* к параболической функции правдоподобия. Такая процедура исключительно проста и изящна, но она все-таки имеет свои ограничения.

1. Способ преобразования правилен с точностью *только до порядка* $1/N$. Он устраняет первый член смещения в разложении функции правдоподобия, но оставляет другие члены более высокого порядка по $1/N$. Главная особенность такого преобразования состоит в том, что мы преобразуем экспериментальную функцию правдоподобия в функцию правдоподобия нормального распределения по параметру, тогда как нам нужно было бы сделать нормальным *теоретическое* распределение. С точностью до порядка $1/N$ это одно и то же для выборок больших объемов, но для малых выборок это не так.

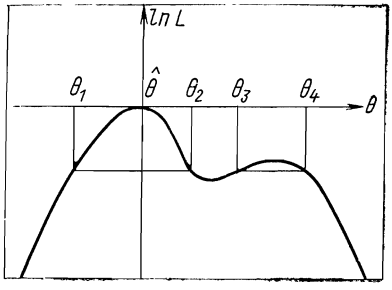


Рис. 9.11. «Патологическая» функция правдоподобия

2. Полученный интервал является центральным для преобразованной переменной. Он может дать очень широкие и не центральные границы для первоначальной переменной.

3. Если функция правдоподобия имеет несколько максимумов («патологический» случай), такая процедура будет давать не один непрерывный интервал, а в общем случае несколько. Тогда можно делать такие доверительные утверждения:

$$p\ (\theta_1 \leqslant \theta \leqslant \theta_2 \text{ или } \theta_3 \leqslant \theta \leqslant \theta_4) = 1 - \beta,$$

т. е. вероятность, что либо

$$[\theta_1\ (X) \leqslant \theta \text{ и } \theta_2\ (X) \geqslant \theta],$$

либо

$$[\theta_3\ (X) \leqslant \theta \text{ и } \theta_4\ (X) \geqslant \theta]$$

равна $1-\beta$. Соответствующий пример показан на рис. 9.11.

Интерпретация такого утверждения в качестве доверительного интервала для θ до некоторой степени сомнительна. Можно было бы найти один непрерывный интервал с тем же самым вероятностным содержанием; такой интервал мог бы быть более значимым. В таких случаях, однако, доверительный интервал дает очень неполное представление о результатах.

4. Когда предложенный выше метод дает неопределенные (или бесконечные) доверительные интервалы, то вероятнее всего используемое преобразование недостаточно хорошо по своей природе. Возможно, что был неправильно поставлен вопрос об исследуемых параметрах или была потеряна часть информации о модели или данных.

В таких случаях можно было б отыскать преобразование, дающее некоторое другое распределение, отличное от нормального. Возможно также, что для этой задачи требуется более сложная ин-

терпретация, чем интерпретация ее в виде единственной оценки м. п. и доверительного интервала.

Ниже (в п. 9.4.3) мы вернемся к обсуждению многомерного случая (случай нескольких параметров).

§ 9.4. ИСПОЛЬЗОВАНИЕ АСИМПТОТИЧЕСКИХ ПРИБЛИЖЕНИЙ

9.4.1. Асимптотическая нормальность оценки максимального правдоподобия. В п. п. 7.3.1 и 7.3.3 мы уже видели, что оценка м. п. $\hat{\theta}$ в асимптотическом пределе распределена по закону $N(\theta, 1/NI)$, где N — число наблюдений, а I — информация о параметре θ, содержащаяся в одном наблюдении. Отсюда мы сразу же получаем доверительный интервал, если использовать результаты п. 9.3.1 для случая нормального распределения, а именно:

$$\theta_{\cdot}^{\cdot} = \hat{\theta} \pm \lambda_{\beta/2} / \sqrt{NI(\hat{\theta})}, \tag{9.14}$$

где $\lambda_{\beta/2}$ — $(\beta/2)$-точка стандартного нормального распределения.

Однако вследствие смещения оценки $\hat{\theta}$ и из-за того, что асимптотическое поведение характерно для больших N, этот метод дает только грубый доверительный интервал.

9.4.2. Асимптотическая нормальность распределения $\partial \ln L/\partial\theta$. Более точный доверительный интервал (фактически наиболее точный интервал вблизи истинного значения θ_0) можно получить, если использовать то обстоятельство, что $\partial \ln L/\partial\theta$ распределено со средним 0 и дисперсией $NI(\theta)$ для всех N. При этом используется приближение, что переменная

$$\frac{1}{\sqrt{NI(\hat{\theta})}} \frac{\partial \ln L}{\partial \theta}$$

имеет стандартное нормальное распределение. Тогда доверительный интервал получается из условия

$$\left| \frac{1}{\sqrt{NI(\hat{\theta})}} \frac{\partial \ln L}{\partial \theta} \right| < \lambda_{\beta/2}. \tag{9.15}$$

Пример. Предположим, что в эксперименте зарегистрированы наблюдения $\mathbf{X}$, распределенные по закону отрицательного экспоненциального распределения (см. п. 4.2.9) со средним μ:

$$f(X, \mu) = \frac{1}{\mu} \exp(-\chi/\mu).$$

Логарифм правдоподобия равен:

$$\ln L = -N \ln \mu - \sum_{i=1}^{N} \frac{X_i}{\mu} = -N(\ln \mu + \overline{X}/\mu)$$

и ее производные

$$\frac{\partial \ln L}{\partial \mu} = -\frac{N}{\mu} + \frac{N\overline{X}}{\mu^2}; \quad \frac{\partial^2 \ln L}{\partial \mu^2} = \frac{N}{\mu^2} - \frac{2\sum_{i=1}^{N} X_i}{\mu^3}.$$

Информация $I(\hat{\mu}$ для $\hat{\mu} = \overline{X}$ находится из соответствующего выражения для математического ожидания

$$E\left(-\frac{\partial^2 \ln L}{\partial \mu^2}\right) = \frac{N}{\mu^2} = NI(\mu).$$

Поэтому

$$NI(\hat{\mu}) = \frac{N}{\overline{X}^2}.$$

Тогда грубый доверительный интервал [см. уравнение (9.14)] равен:

$$\overline{X}\left(1 - \frac{\lambda_{\beta/2}}{\sqrt{N}}\right) \leqslant \mu \leqslant \overline{X}\left(1 + \frac{\lambda_{\beta/2}}{\sqrt{N}}\right).$$

Асимптотически более точный интервал [см. уравнение (9.15)] получается из условия

$$\frac{\left|\frac{N\overline{X}}{\mu^2} - \frac{N}{\mu}\right|}{\sqrt{N}/\mu} \leqslant \lambda_{\beta/2}$$

или если преобразовать его в доверительный интервал для μ:

$$\frac{\overline{X}}{1 + \frac{\lambda_{\beta/2}}{\sqrt{N}}} \leqslant \mu \leqslant \frac{\overline{X}}{1 - \frac{\lambda_{\beta/2}}{\sqrt{N}}}.$$

Если бы мы отыскивали доверительный интервал для $1/\mu$, то уравнения (9.14) и (9.15) дали бы один и тот же результат.

9.4.3. Многомерные доверительные области. Доверительная область для k параметров получается в результате обобщения уравнения (9.15):

$$\frac{1}{N}\left(\frac{\partial L(\mathbf{X} \mid \theta)}{\partial \theta}\right)^T I^{-1}(\theta)\left(\frac{\partial L(\mathbf{X} \mid \theta)}{\partial \theta}\right) \leqslant \chi_\beta^2(k), \qquad (9.16)$$

где $I(\theta)$ — матрица информации, а $\chi_\beta^2(k)$ — β-точка χ^2-распределения с k степенями свободы. Если использовать приближение

$$I(\theta) = \left(-\frac{\partial^2 \ln L}{\partial \theta\, \partial \theta'}\right)_{\theta = \hat{\theta}},$$

то (9.16) задает гиперэллипсоидальную область для $\partial L/\partial\theta$. Соответствующая область в θ-пространстве обычно не эллипсоидальна.

Для того чтобы найти многомерные доверительные области для произвольной функции правдоподобия, нужно обобщить метод п. 9.3.2 на k переменных, т. е. численно найти такие значения $\theta_\cdot$ и $\theta^\cdot$, при которых гиперповерхность

$$\ln L_\theta(\mathbf{X} \mid \theta)$$

пересекается с гиперплоскостью

$$\ln L = \ln L_{\text{макс}} - \frac{1}{2}\chi^2_\alpha(k).$$

Тогда область $[\theta_\cdot, \theta^\cdot]$ называется доверительной с вероятностным содержанием α.

Такой подход может быть обоснован в асимптотическом пределе, если использовать свойства отношения правдоподобия

$$l = \frac{L(\mathbf{X} \mid \theta)}{L(\mathbf{X} \mid \hat{\theta})}.$$

В § 10.5 показано, что переменная $-2\ln l(\theta)$ распределена асимптотически, как $\chi^2(k)$. Это непосредственно приводит к вероятностному утверждению

$$\beta = p[-2\ln l(\theta) \leqslant \chi^2_\beta(k)],$$

которое в принципе может быть преобразовано в доверительную область для θ.

§ 9.5. СВОЙСТВА ТРЕХ ОБЩИХ МЕТОДОВ ОЦЕНКИ ИНТЕРВАЛА ДЛЯ ВЫБОРКИ КОНЕЧНОГО ОБЪЕМА

Три метода оценки параметра интервалом в общем случае (когда данные распределены не по нормальному закону, а θ — p-мерно) основаны на следующих трех асимптотических свойствах функции правдоподобия.

1. Переменная $\sqrt{N}(\hat{\theta} - \theta)$ асимптотически распределена по нормальному закону со средним 0 и матрицей вторых моментов I^{-1} (см. п. п. 7.2.3, 7.3.1 и 7.3.3 (3)]; $\underset{\sim}{I}$ — информационная матрица с элементами:

$$(\underset{\sim}{I})_{ij} = E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial \theta_i}\right)\left(\frac{\partial \ln f(X, \theta)}{\partial \theta_j}\right)\right] = -E\left[\frac{\partial^2 \ln f(X, \theta)}{\partial \theta_i \partial \theta_j}\right].$$

2. $\partial \ln f(X, \theta)/\partial\theta$ распределено со средним 0 и матрицей вторых моментов $\underset{\sim}{I}(\theta)$ и асимптотически нормально (см. п. 7.2.3).

3. $$2\sum_k \left[\ln f(X_k, \hat{\theta}) - \ln f(X_k, \theta)\right]$$

асимптотически ведет себя как χ^2-распределение с p степенями свободы.

Таблица 9.2

Вид функций для оценки интервала в общем случае

Метод		Функция $Q^2(\theta)$	Асимптотическое разложение	Комментарии
Используется $(\hat{\theta}-\theta)$	1а. Информация рассчитывается аналитически	$N(\hat{\theta}-\theta)\,I(\theta)\,(\hat{\theta}-\theta)$	$\frac{X^2}{I}+\frac{2X^2Y}{I^2\sqrt{N}}+\frac{kX^3}{I^3\sqrt{N}}$ Среднее равно $1+\frac{1}{N}(2a+b)$	Смещена. При замене параметров можно исключить смещение среднего
	1б. Информация оценивается по данным	$-N(\hat{\theta}-\theta)\,I_N(\theta)\,(\hat{\theta}-\theta)$	$\frac{X^2}{I}+\frac{X^2Y}{I^2\sqrt{N}}+\frac{kX^3}{I^3\sqrt{N}}$ Среднее равно $1+\frac{1}{N}(a+b)$	Смещена. Смещение можно исключить
	1в. Информация оценивается по данным при $\theta=\hat{\theta}$	$-N(\hat{\theta}-\theta)\,I_N(\hat{\theta})\,(\hat{\theta}-\theta)$	$\frac{X^2}{I}+\frac{X^2Y}{I^2\sqrt{N}}$ Среднее равно $1+\frac{1}{N}a$	То же
Используется $\sum_k \frac{\partial \ln f(X_k,\theta)}{\partial\theta}$	2а. Информация рассчитывается аналитически	$\frac{1}{N}\left(\sum_k \frac{\partial \ln f_i}{\partial\theta}\right)\times$ $\times I^{-1}(\theta)\left(\sum_k \frac{\partial \ln f_i}{\partial\theta}\right)$	$\frac{X^2}{I}$ Среднее равно 1	Не смещена

Продолжение табл. 9.2

Метод		Функция $Q^2(\theta)$	Асимптотическое разложение	Комментарии
	2б. Информация оценивается по данным	$-\frac{1}{N}\left(\sum_k \frac{\partial \ln f}{\partial \theta}\right)\times$ $\times I_N^{-1}(\theta)\left(\sum_k \frac{\partial \ln f}{\partial \theta}\right)$	$\frac{X^2}{I}+\frac{X^2 Y}{I^2\sqrt{N}}$ Среднее равно $1+\frac{1}{N}a$	Смещена. Смещение можно исключить
	2в. Информация оценивается по данным при $\hat{\theta}$	$-\frac{1}{N}\left(\sum_k \frac{\partial \ln f}{\partial \theta}\right)\times$ $\times I_N^{-1}(\hat{\theta})\left(\sum_k \frac{\partial \ln f}{\partial \theta}\right)$	$\frac{X^2}{I}+\frac{X^2 Y}{I^2\sqrt{N}}+\frac{kX^3}{I^3\sqrt{N}}$ Среднее равно $1+\frac{1}{N}(a+b)$	То же
Отношение правдоподобий	3	$2\left(\sum_k \ln f_i(X_k, \hat{\theta})-\right.$ $\left.-\sum_k \ln f(X_k, \theta)\right)$	$\frac{X^2}{I}+\frac{X^2 Yi}{I^2\sqrt{N}}+\frac{kX^3}{3I^3\sqrt{N}}$ Среднее равно $1+\frac{1}{N}\left(a+\frac{b}{3}\right)$	Смещена. Смещение инвариантно по отношению к замене переменных. Может быть найдено более точное приближение

Любое из этих свойств может быть использовано для того, чтобы ввести соответствующую функцию Q^2 (**X**, θ), распределенную независимо от θ, т. е. любая из следующих функций

$$Q_1^2 = N(\hat{\theta}-\theta)^T \underset{\sim}{I}(\theta)(\hat{\theta}-\theta); \tag{9.17}$$

$$Q_2^2 = \left[\frac{1}{N}\sum_{k=1}^{N}\frac{\partial \ln f_i(X_k, \theta)}{\partial\theta}\right]^T N\underset{\sim}{I}^{-1}(\theta)\left[\frac{1}{N}\sum_{k=1}^{N}\frac{\partial \ln f_i(X_k\theta)}{\partial\theta}\right]; \tag{9.18}$$

$$Q_3^2 = 2\sum_{k=1}^{N}[\ln f(X_k, \hat{\theta}) - \ln f(X_k, \theta)] \tag{9.19}$$

асимптотически распределена по χ^2 (p)-закону.

На практике не всегда можно вычислить матрицу информации $\underset{\sim}{I}(\theta)$ аналитически. Тогда в уравнениях (9.17) и (9.18) она может быть заменена на соответствующую свою оценку, полученную на основе данных. Это можно сделать двумя способами. В первом случае оценки матричных элементов $(\underset{\sim}{I})_{ij}$ полагаются равными выборочному среднему вторых производных:

$$\left[\underset{\sim}{I}_N(\theta)\right]_{ij} = -\frac{1}{N}\sum_{k=1}^{N}\frac{\partial^2 \ln f_i(X_k, \theta)}{\partial\theta_i\,\partial\theta_j}. \tag{9.20}$$

Второй способ аналогичен первому, только θ в (9.20) фиксировано и равно $\hat{\theta}$. При этом исходят из того, что оценка θ находится вблизи истинного значения θ:

$$\left[\underset{\sim}{I}_N(\hat{\theta})\right]_{ij} = -\left\{\frac{1}{N}\sum_{k=1}^{N}\left[\frac{\partial^2 \ln f_i(X_k, \theta)}{\partial\theta_i\,\partial\theta_j}\right]\right\}_{\theta=\hat{\theta}}. \tag{9.21}$$

Таким образом, мы имеем семь возможных функций Q^2 (**X**, θ): уравнения (9.17) — (9.19) или уравнения (9.17), (9.18) с двумя различными выражениями для матрицы информации (9.20) или (9.21). В табл. 9.2 приведены точные виды всех этих функций.

Тогда вероятностные утверждения приобретают такой вид:

$$P\left[\chi^2_{\frac{1-\beta}{2}}(p) \leqslant Q^2(\theta) \leqslant \chi^2_{\frac{1+\beta}{2}}(p)\right] = \beta. \tag{9.22}$$

Соответственно доверительные области для θ можно найти из уравнения (9.22). Интересно отметить, что доверительная область имеет вид эллипсоида в θ-пространстве только тогда, когда используются Q_{1c}^2 [см. уравнение (9.17), выражение (9.21) для матрицы информации и табл. 9.2].

Рассмотрим распределения Q^2 (**X**, θ) для выборок конечного объема, когда они не имеют вид χ^2 (p)-распределений и уравнение (9.22) еще не верно. В частности, рассмотрим смещение распределений Q^2 (**X**, θ) по отношению к χ^2 (p) (рис. 9.12). Очевидно, если

$P[Q^2(\theta)]$ смещено относительно $P[\chi^2(p)]$, то вероятностное содержание по любую сторону интервала в уравнении (9.22)

$$\left[\chi^2_{\frac{1-\beta}{2}}(p),\quad \chi^2_{\frac{1+\beta}{2}}(p)\right]$$

будет неодинаково для распределения $Q^2(\theta)$.

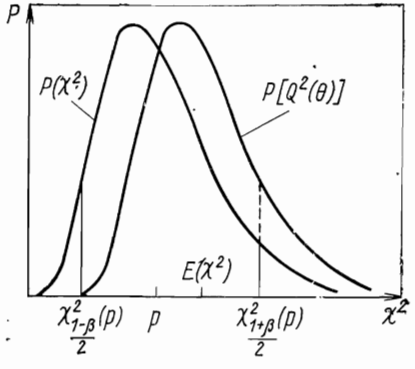


Рис. 9.12. $\chi^2(p)$-Распределение и смещенное приближенное распределение, соответствующее $Q^2(\mathbf{X}, \theta)$ для выборки конечного объема

Разложим каждую из функций $Q^2(\mathbf{X}, \theta)$ в ряд по степеням $1/N$. Если ограничиться простейшим случаем только одного параметра ($p = 1$), то можно написать:

$$\frac{1}{N}\sum_k \frac{\partial \ln f_k(X_k, \theta)}{\partial\theta} = E\left(\frac{\partial \ln f(X, \theta)}{\partial\theta}\right) + \frac{X}{\sqrt{N}} = \frac{X}{\sqrt{N}};$$

$$\frac{1}{N}\sum_k \frac{\partial^2 \ln f_k(X_k, \theta)}{\partial\theta^2} = E\left(\frac{\partial^2 \ln f(X, \theta)}{\partial\theta^2}\right) + \frac{Y}{\sqrt{N}} = -I + \frac{Y}{\sqrt{N}},$$

$$\frac{1}{N}\sum_k \frac{\partial^3 \ln f(X_k, \theta)}{\partial\theta^3} = E\left(\frac{\partial^3 \ln f(X, \theta)}{\partial\theta^3}\right) + \frac{Z}{\sqrt{N}} = k + \frac{Z}{\sqrt{N}}.$$

В соответствии с предельной центральной теоремой X, Y и Z асимптотически распределены по нормальному закону со средним нуль и конечной дисперсией. В п. 7.3.3 (условие (3)) было показано, что

$$(\hat{\theta}-\theta) = \frac{X}{I\sqrt{N}}\left[1 + \frac{Y}{I\sqrt{N}} + \frac{1}{2}\frac{kX}{I^2\sqrt{N}}\right] + \ldots,$$

отсюда следует, что

$$-I_N(\hat{\theta}) = -I_N(\theta) + (\hat{\theta}-\theta)\frac{1}{N}\sum_{k=1}^{N}\frac{\partial^3 \ln f(X_k, \theta)}{\partial\theta^3} + \ldots = -I(\theta) +$$

$$+ \frac{Y}{\sqrt{N}} + \frac{kX}{I\sqrt{N}} + \ldots = -I(\theta)\left[1 - \frac{Y}{I\sqrt{N}} - \frac{kX}{I^2\sqrt{N}}\right] + \ldots$$

Точно так же можно показать, что

$$2\sum_k [\ln f(X_k, \hat{\theta}) - \ln f(X_k, \theta)] = 2N\left[\frac{1}{N}\sum_k \frac{\partial \ln f(X_k, \theta)}{\partial\theta}(\hat{\theta}-\theta) + \right.$$

$$\left. + \frac{1}{2N}\sum_k \frac{\partial^2 \ln f(X_k, \theta)}{\partial\theta^2}(\hat{\theta}-\theta)^2 + \frac{1}{6N}\sum_k \frac{\partial^3 \ln f(X_k, \theta)}{\partial\theta^3}(\hat{\theta}-\theta)^3\right] + \cdots =$$

$$= \frac{2X^2}{I}\left(1 + \frac{X}{I\sqrt{N}} + \frac{1}{2}\frac{kX}{I^2\sqrt{N}}\right) + \frac{X^2}{2I}\left(-1 + \frac{Y}{I\sqrt{N}}\right) \times$$

$$\times \left(1 + \frac{2Y}{I\sqrt{N}} + \frac{kX}{I^2\sqrt{N}}\right) + \frac{kX^3}{6I^3\sqrt{N}} + O\left(\frac{1}{N}\right) = \frac{X^2}{2I} +$$

$$+ \frac{X^2 Y}{I^2\sqrt{N}} + \frac{kX^3}{3I^3\sqrt{N}} + O\left(\frac{1}{N}\right).$$

Можно написать разложения для каждой функции; они приведены в табл. 9.2. Средние значения выражаются через две константы a и b, где

$$a = E(\sqrt{N}X^2 Y)/I^2; \quad b = E(\sqrt{N}\,\tilde{k}X^3)/I^3.$$

Для того чтобы показать, что a и b конечны, запишем

$$X^2 Y = \frac{1}{N}\left(\sum_i g_i\right)^2\left[\sqrt{N}\,I_0 + \frac{1}{\sqrt{N}}\sum_i h_i\right],$$

где

$$g_i = \frac{\partial \ln f(X_i, \theta)}{\partial\theta}$$

и

$$h_i = \frac{\partial^2 \ln f(X_i, \theta)}{\partial\theta^2}.$$

Тогда

$$\frac{E(\sqrt{N}X^2 Y)}{I^2} = \frac{1}{I^2}E\left[I\left(\sum_i g_i\right)^2\right] + \frac{1}{I^2}E\left[\frac{1}{N}\left(\sum_i g_i\right)^2\sum_i h_i\right] =$$

$$= \frac{1}{I}E\left(\sum_i g_i^2\right) + \frac{1}{NI^2}E\left[\sum_i g_i^2 h_i + \sum_{i\neq j} g_i^2 h_j\right] =$$

$$= N + \frac{E(g^2 h)}{I^2} + \frac{N(N-1)}{NI^2}I(-I) = \frac{E(g^2 h)}{I^2} + 1.$$

Поэтому

$$a = 1 + \frac{1}{I^2}E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial\theta}\right)^2\frac{\partial^2 \ln f(X, \theta)}{\partial\theta^2}\right]$$

и соответственно

$$b = \frac{1}{I^3}E\left(\frac{\partial^3 \ln f(X, \theta)}{\partial\theta^3}\right)E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial\theta}\right)^3\right].$$

Как видно из табл. 9.2, все функции $Q^2(\theta)$, за исключением $Q^2_{2a}(\theta)$, смещены по отношению к χ^2. Это смещение пропорционально $1/N$ и существует, даже если логарифм функции правдоподобия $\ln f(\mathbf{X}, \theta)$ симметричен по θ. В этом отношении оценка интервала отличается от оценки параметра значением, для которой достаточным условием отсутствия смещения была симметрия логарифма функции правдоподобия.

Смещение можно исключить в результате замены переменных, за исключением последнего метода (метода отношений правдоподобия), который инвариантен по отношению к таким преобразованиям.

Функция $Q^2_{2a}(\theta)$ не смещена по отношению к χ^2. Тем не менее величина χ^2/I распределена не по χ^2-закону. Действительно, дисперсия X^2/I равна:

$$D\left(\frac{X^2}{I}\right) = E\left(\frac{X^4}{I^2}\right) - \left[E\left(\frac{X^2}{I}\right)\right]^2 =$$

$$= 2 + \frac{1}{N}\left\{\frac{1}{I^2} E\left[\left(\frac{\partial \ln f(X, \theta)}{\partial \theta}\right)^4\right] + 1\right\}.$$

Член с множителем $1/N$ дает смещение по сравнению с дисперсией $\chi^2(1)$-распределения. О функциях, дающих более близкое приближение к χ^2, читатель может узнать в работе [2].

В других методах дополнительные члены будут точно так же давать вклад в дисперсию и более высокие моменты распределения. Однако для метода отношения правдоподобий может быть показано, что функция

$$\chi^2_4 = \frac{2\left[\sum_k \ln f(X_k, \hat{\theta}) - \sum_k \ln f(X_k, \theta)\right]}{1 + (1/N)\left[a(\theta) + (1/3)\, b(\theta)\right]}$$

распределена по χ^2, включая и члены порядка $1/N$ [2,47].

Члены более высоких порядков обычно не вычисляют.

Если можно аналитически вычислить $I(\theta)$, то для оценки интервалов лучше всего использовать функцию Q^2_{2a}. В этом случае отсутствует смещение по отношению к χ^2-распределению. Во всех других случаях наилучшим представляется метод отношения правдоподобий (см. п. 9.4.3. и § 10.5), поскольку он обладает полезным свойством инвариантности при замене оцениваемых переменных.

§ 9.6. БЕЙЕСОВСКИЙ ПОДХОД

Бейесовский подход к задаче оценки параметра интервалом внешне подобен использованию метода отношения правдоподобий, но интерпретация результирующих интервалов совершенно другая. В классическом подходе неизвестный параметр имеет постоянное значение θ_0, тогда как в бейесовском подходе существует распределение возможных истинных значений $\pi(\theta \mid \mathbf{X})$.

Как следует из п. 2.3.5, это апостериорное распределение отражает распределение степени доверия при заданных данных $\mathbf{X}$ и заданном *априорном* знании $\pi(\theta)$:

$$\pi(\theta \mid \mathbf{X}) = \frac{\prod_{i=1}^{N} f(X_i, \theta)\,\pi(\theta)}{\int \prod_{i=1}^{N} f(X_i, \theta)\,\pi(\theta)\,d\theta}. \tag{9.23}$$

Распределение (9.23) суммирует полное знание наблюдателя о параметре θ. Используя этот факт, бейесовский доверительный интервал $[\theta_\cdot, \theta^\cdot]$ с вероятностным содержанием β определяется как

$$\int_{\theta_\cdot}^{\theta^\cdot} \pi(\theta \mid \mathbf{X})\,d\theta = \beta. \tag{9.24}$$

В соответствии с этим определением интервал $[\theta_\cdot, \theta^\cdot]$ соответствует доле β полной веры наблюдателя в параметр θ, т. е. наблюдатель готов заключить пари, что истинное значение лежит в этом

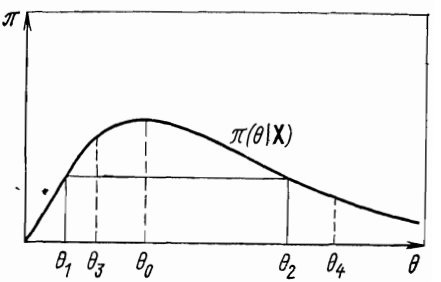


Рис. 9.13. Плотность распределения апостериорной веры $\pi(\theta \mid \mathbf{X})$

интервале с шансами на выигрыш $\beta/(1-\beta)$. Это означает, что полное апостериорное распределение (9.23) как бы заменяется дискретным двухточечным распределением:

$$P(\theta_\cdot < \theta < \theta^\cdot) = \beta;$$

$$P\{\theta \notin [\theta_\cdot, \theta^\cdot]\} = 1 - \beta.$$

Как и в классическом подходе, интервал $[\theta_\cdot, \theta^\cdot]$ не единственный. Однако, если добавить условие, что любая точка внутри интервала должна быть более вероятна, чем любая точка вне его, можно получить единственный интервал. Интуитивно мы понимаем, что такой интервал будет самым коротким из всех возможных интервалов с вероятностным содержанием β. На рис. 9.13 показан такой интервал $[\theta_1, \theta_2]$, определяемый условиями:

$$\int_{\theta_1}^{\theta_2} \pi(\theta \mid \mathbf{X})\,d\theta = \beta \quad \text{и} \quad \pi(\theta_1 \mid \mathbf{X}) = \pi(\theta_2 \mid \mathbf{X}).$$

Другим доверительным интервалом с тем же самым вероятностным содержанием мог бы быть $[\theta_3, \theta_4]$. Тогда вероятностные содержания интервалов $[\theta_1, \theta_3]$ и $[\theta_2, \theta_4]$ равны. Но поскольку по определению $\pi(\theta, \mathbf{X})$ больше в первом интервале, чем во втором, то из этого следует, что

$$\theta_3 - \theta_1 < \theta_4 - \theta_2,$$

т. е.

$$\theta_2 - \theta_1 < \theta_4 - \theta_3.$$

Сопоставим такую интерпретацию с классической. На рис. 9.14 нами показаны три функции плотности для наблюдаемой статистики $t(\mathbf{X})$. Для простоты предположим, что θ является средним значением $f(t|\theta)$. Истинное значение θ_0 и плотность $f(t|\theta_0)$ неизвестны. При заданном наблюдении $t(\mathbf{X})$ устанавливается нижняя граница $\theta_\cdot$, так что

$$\int_{t(\mathbf{X})}^{\infty} f(t|\theta_\cdot) = (1-\beta)/2$$

(на рис. 9.14 эта область заштрихована реже) и значение $\theta^\cdot$ таково

$$\int_{-\infty}^{t(\mathbf{X})} f(t|\theta^\cdot) = (1-\beta)/2$$

(более плотная штриховка на рис. 9.14).

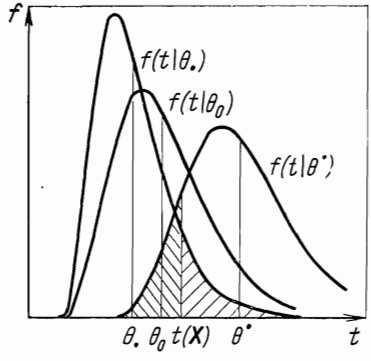


Рис. 9.14. Плотности вероятностей для классических доверительных границ $\theta_\cdot$ и $\theta^\cdot$ и для классического истинного значения θ_0

Тогда $[\theta_\cdot, \theta^\cdot]$ будет центральным интервалом с вероятностным содержанием β.

Вероятность того, что классический интервал, сконструированный таким образом, не содержит истинное значение θ_0, равна $1-\beta$. В бейесовском подходе доля веры в то, что θ_0 не внутри бейесовского интервала, равна $1-\beta$.

Отметим, что существует и другой случай, когда бейесовский интервал может не включать θ_0, если наше первоначальное знание $\pi(\theta)$ неверно. Тогда $\pi(\theta|\mathbf{X})$ будет концентрироваться вблизи θ_0 только в асимптотическом пределе.

В обоих подходах недостаточно приводить только оценку интервала, поскольку это не дает каких-либо указаний на то, какие значения внутри интервала рассматриваются как наиболее вероятные. Поэтому лучше в смысле представления результатов привести оценку значения θ вместе с оценкой интервала.

ГЛАВА 10

ПРОВЕРКА ГИПОТЕЗ

В предыдущих трех главах мы познакомились с тем, как экспериментальная информация может быть использована для оценки параметров. В этой и последующих главах будет показано, как экспериментальные данные могут быть использованы для *проверки*: для того чтобы подтвердить или опровергнуть теорию или гипотезу или помочь в выборе одной из альтернативных гипотез.

Если *проверяемой гипотезе* соответствует определенное значение некоторого параметра, то задачи *оценки параметра* и *проверки гипотезы* взаимосвязаны; например, хорошие методы оценок часто приводят к аналогичным процедурам проверки. Однако выводы, получающиеся в этих двух случаях, совершенно различны и не надо их путать (на практике это часто случается!).

Допустим, что априори ничего не известно о некотором параметре. Естественно попытаться использовать данные для того, чтобы оценить значение этого параметра с помощью методов, описанных в гл. 7—9. С другой стороны, если имеется некое теоретическое предсказание о величине параметра, то задача может быть сформулирована как проверка совместимости данных с этим значением. В любом случае природа задачи (оценка или проверка) должна быть ясной с самого начала и совместимой с конечным результатом.

Гипотеза, которая может быть сформулирована без каких-либо предположений, называется *простой гипотезой*. Совокупность гипотез, состоящая из более чем одной простой гипотезы, называется *сложной гипотезой*. Типичным примером сложной гипотезы является гипотеза, содержащая свободный параметр. Для того чтобы познакомиться с основными понятиями, используемыми при проверке гипотез, достаточно рассмотреть простые гипотезы, что мы и будем делать в основной части этой главы.

Наиболее важная задача состоит в проверке заданной гипотезы (обозначим ее H_0) по отношению ко всем другим возможным гипотезам (не H_0). Гл. 11 посвящена таким проверкам; соответствующие критерии называются *критериями согласия*.

§ 10.1. ФОРМУЛИРОВКА КРИТЕРИЯ

10.1.1. Основные понятия при проверке гипотез. Предположим, что нужно проверить гипотезу H_0 (называемую нулевой гипотезой) по отношению к противоположной гипотезе H_1 на основе некоторых экспериментальных наблюдений. Пусть X будет некоторой функцией наблюдений, называемой *проверочной статистикой*, а W — пространством всех возможных значений X. Разделим пространство W на *критическую* w и *допустимую* $W - w$ *области*, такие, что если функция наблюдений X попадает в область w, то это означает, что нулевая гипотеза неверна. Таким образом, выбор критерия проверки H_0 сводится к выбору проверочной статистики X и критической области w.

Обычно размер критической области выбирается таким, чтобы получить желаемый *уровень значимости* α (величина критерия), определенный как вероятность попадания X в w, когда гипотеза H_0 верна:

$$P\ (X \in w \mid H_0) = \alpha. \tag{10.1}$$

Таким образом, α — вероятность того, что H_0 будет отброшена даже если она в действительности верна.

Полезность критерия проверки зависит от его способности отделить данную гипотезу от противоположной гипотезы H_1. Мерой этой полезности служит *мощность критерия*, определенная как вероятность $1-\beta$ попадания X в критическую область, если H_1 верна:

$$P\ (X \in W - w \mid H_1) = 1-\beta. \tag{10.2}$$

Иными словами, β — вероятность попадания X в допустимую область, если верна противоположная гипотеза:

$$P\ (X \in W - w \mid H_1) = \beta.$$

При проверке гипотез обычно разделяют два вида неверных заключений:

а) *ошибка первого рода* или *потеря*: отбрасывание нулевой гипотезы, когда она верна. Вероятность такой ошибки равна α;

б) *ошибка второго рода* или *примесь*: принятие нулевой гипотезы, когда она неверна. Вероятность такой ошибки равна β*.

В следующих параграфах мы на конкретных примерах познакомимся с этими двумя видами ошибок.

10.1.2. Пример: разделение двух классов событий. Предположим, что нужно отделить события с упругим рассеянием протона $pp \to pp$ (проверяемая гипотеза H_0) от событий неупругого процесса $pp \to pp\pi^0$ (противоположная гипотеза H_1) в экспериментальной установке, в которой измеряются траектории протонов, а π^0-мезоны непосредственно не регистрируются. Тогда

* Символы α и β обычно используются большинством авторов для измерения риска первого и второго рода, но некоторые авторы используют $1-\beta$ там, где мы пишем β.

разделение событий должно быть основано на существующей кинематической информации (измеренных углах и импульсах протонов). Эта информация в свою очередь определяет *недостающую* массу (полную энергию в системе центра всех невидимых частиц) события. Недостающая масса M может быть выбрана в качестве проверочной статистики.

В случае, если верна гипотеза H_0, истинное значение $M = 0$, если же справедлива H_1, то $M = M_{\pi^0} = 135$ *Мэв/с²*. Поэтому допустимая область должна быть выбрана вблизи $M = 0$, а критическая область — вблизи $M = 135$ *Мэв/с²*. Если недостающая масса у истинных событий с упругим рассеянием протона близка к 135 *Мэв/с²* вместо того, чтобы быть близкой к нулю (из-за измерительных погрешностей), они будут приняты на неупругое и образуют, таким образом, потерю. Истинно неупругие события с M, близким к нулю, образуют примесь к группе упругих событий.

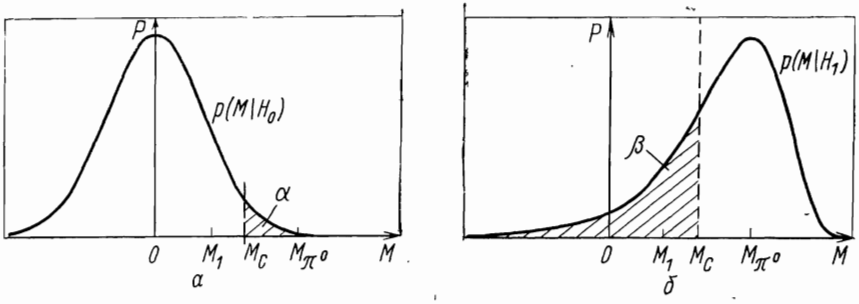


Рис. 10.1. Функции размещения для недостающей массы M для гипотез H_0 и H_1, описанных в тексте

Точный выбор критической области зависит от того, какие величины потери и примеси мы сочтем допустимыми. Очевидно, что экспериментатор всегда стремится уменьшить и потерю α, и примесь β.

Предположим, что известны ф. п. в. случайной переменной M для каждой из этих двух гипотез — функции $p(M \mid H_0)$ и $p(M \mid H_1)$ — пусть они имеют вид, изображенный на рис. 10.1. Вид этих функций распределений можно предсказать, если экспериментатор правильно понимает работу измерительной аппаратуры, знает источники и величину всех погрешностей и т. д. Такие распределения называются *функциями разрешения*.

Рассмотрим случай, когда измерено только одно событие. Исходя из формы кривой на рис. 10.1, *а* разумно выбрать критическую область в виде

$$M > M_c$$

так, что

$$P(M > M_c) = \int_{M_c}^{\infty} p(M \mid H_0)\, dM = \alpha.$$

Здесь α — желаемый уровень значимости. Тогда, если недостающая масса события удовлетворяет неравенству $M \leqslant M_c$, верной будет считаться гипотеза H_0.

Если справедлива гипотеза H_1, то в соответствии с рис. 10.1 вероятность попадания наблюдения в допустимую область $M \leqslant M_c$ равна:

$$P(M \leqslant M_c \mid H_1) = \int_{-\infty}^{M_c} p(M \mid H_1)\, dM = \beta.$$

Таким образом, можно вычислить мощность критерия (1 — β). Обычно α выбирается малым, например, 0,05 или 0,01. Тогда β определяется видом выбранного критерия. Очевидно, что для данного примера (см. рис. 10.1) $\beta \gg \alpha$. Таким образом, вероятность классификации события как упругого, если оно в действительности неупругое, велика. Например, событие со значением недостающей массы, равной M_1 (см. рис. 10.1), будет названо упругим, хотя неупругое событие имеет большую вероятность иметь недостающую массу M_1.

Отметим, что реальная примесь (т. е. число событий с π^0-мезоном, отнесенных к классу упругих событий) равна βR, где R — априорное число всех событий с рождением π^0-мезона. Если R мало, то β можно выбрать довольно большим.

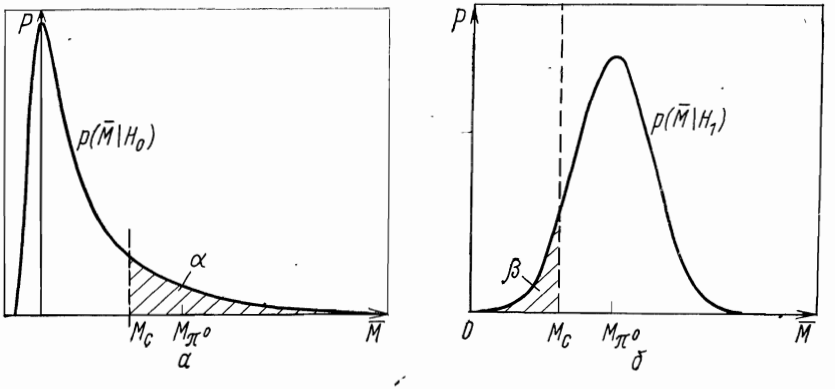


Рис. 10.2. Функции размещения для среднего $\bar{M}$ нескольких наблюдений недостающей массы M для гипотез H_0 и H_1, описанных в тексте

На основе подобного анализа можно также изучить вопрос о необходимости усовершенствования аппаратуры с тем, чтобы улучшить точность измерения импульса и тем самым получить более узкую функцию разрешения. Тогда положение может улучшиться и, как видно из рис. 10.2, распределения становятся более отчетливыми, а вероятность β примеси гораздо меньше.

В общем случае задача состоит не только в том, чтобы сконструировать точный критерий (когда вероятности α и β известны), но и наилучший критерий. Например, вместо того чтобы использовать в качестве проверочной статистики только недостающую массу, можно было бы использовать недостающий импульс или недостающую энергию, чтобы прийти к лучшему разделению между двумя гипотезами. В последующих параграфах мы вернемся к этой задаче.

§ 10.2. СРАВНЕНИЕ КРИТЕРИЕВ

10.2.1. Мощность. В п. 10.1.1 мы определили мощность $p(\theta_1)$ критерия для выделения гипотезы

$$H_0 : \theta = \theta_0$$

в сравнении с противоположной

$$H_1 : \theta = \theta_1$$

как

$$p\ (\theta_1) = 1 - \beta.$$

В общем случае (сложных гипотез) можно считать мощность функцией

$$p\ (\theta) = 1 - \beta\ (\theta)$$

значения θ для произвольной противоположной гипотезы. Тогда по построению

$$p\ (\theta_0) = 1 - \beta\ (\theta_0) = \alpha.$$

На рис. 10.3 даны примеры трех возможных *функций мощности.*

Критерии можно сравнить на основании их функций мощности: если *противоположная гипотеза* H_1 *простая,* то *наилучшим критерием для* H_0 *по сравнению с* H_1 *с данным уровнем значимости* α *служит критерий с максимальной мощностью при* $\theta = \theta_1$. В примере, показанном на рис. 10.3, критерий B имеет максимальную мощность (или минимальную примесь β) для всех $\theta > \theta'$ и, в частности, при $\theta = \theta_1$. В свою очередь критерий C — наилучший при проверке гипотезы H_0 по отношению к противоположной гипотезе, которая предполагает, что истинное значение θ лежит в интервале (θ_0, θ').

Если при данном значении θ некоторый критерий обладает по

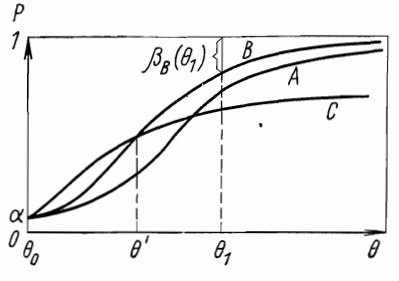


Рис. 10.3. Функции мощности критериев A, B и C с доверительным уровнем α

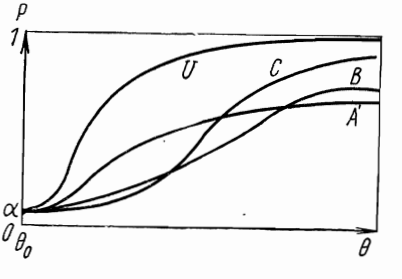


Рис. 10.4. Пример функций мощности четырех критериев, один из которых (U) является р.н.м.-критерием

меньшей мере такой же мощностью, как и любой другой возможный критерий, то он называется *наиболее мощным критерием* для такого значения θ. Критерий, который является наиболее мощным для всех рассматриваемых значений θ, называется *равномерно наиболее мощным* (р. н. м.). Позже мы рассмотрим условия, при которых существует р. н. м.-критерий.

В примере, показанном на рис. 10.4, критерий U является р. н. м.-критерием. Если мы сконструировали наиболее мощный критерий для определенного значения θ, например θ_1, и он не зависит от θ_1, то такой критерий является р.н.м. Иногда из соображения простоты или *устойчивости*, т. е. малой чувствительности к возможным изменениям нулевой гипотезы, предпочтение может быть отдано не р. н. м.-критерию, а другому критерию.

10.2.2. Состоятельность. Очень важным свойством критерия является свойство состоятельности, т. е. свойство критерия с увеличением числа наблюдений все лучше разделять гипотезы. Говорят, что критерий состоятелен, если его мощность стремится к единице при стремлении числа наблюдений к бесконечности. На математическом языке состоятельность выражается в виде

$$\lim_{N\to\infty} P(\mathbf{X} \in w_d \mid H_1) = 1,$$

где $\mathbf{X}$ — выборка из N наблюдений; w_α — критическая область с размером α при условии, что справедлива гипотеза H_0; у состоятельного критерия функция мощности стремится к ступенчатой функции при $N \to \infty$ (рис. 10.5).

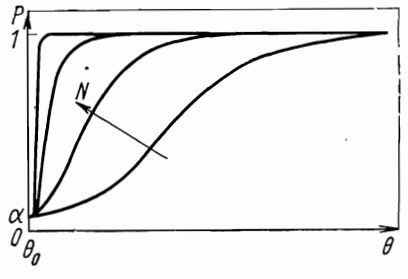


Рис. 10.5. Функция мощности для состоятельного критерия как функция N. С ростом N она стремится к ступенчатой функции

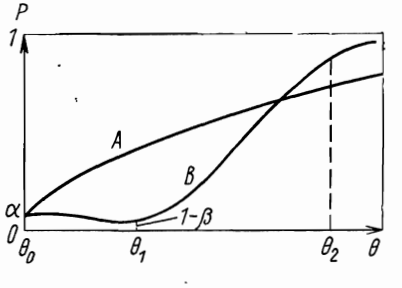


Рис. 10.6. Функция мощности для смещенного (B) и несмещенного (A) критериев

10.2.3. Смещение. Рассмотрим график функции мощности, которая обращается в минимум не при значении $\theta = \theta_0$ (кривая B на рис. 10.6). В этом случае вероятность выбора гипотезы H_0, т. е. $\theta = \theta_0$, больше, когда $\theta = \theta_1$, чем когда $\theta = \theta_0$:

$$1 - \beta < \alpha,$$

т. е. чаще будет выбираться нулевая гипотеза, когда она неверна, чем когда она верна. Такой критерий называется *смещенным* и понятно, что такое свойство в общем случае нежелательно.

Тем не менее возможны частые случаи, когда смещенный критерий предпочтителен из-за его большой мощности при определенных значениях θ. Например, как видно из рис. 10.6, критерий B (смещенный) обладает бóльшей мощностью по сравнению с критерием A (несмещенным) при значениях θ, близких к θ_2. Поэтому, если особенно важно разделить гипотезы при значении $\theta = \theta_2$, целесообразнее выбрать именно этот критерий. Однако, пользуясь критерием B, необходимо помнить, что он не позволит отличить значение θ_0 от значения θ_1.

Переходя к случаю, когда нулевая гипотеза H_0 включает некоторую область θ_0 значений θ, несмещенный критерий можно определить следующим образом:

$$P(X \in w_\alpha | \theta) \leqslant \alpha \text{ для всех } \theta \in \theta_0$$

и

$$P(X \in w_\alpha | \theta) \geqslant \alpha \text{ для всех } \theta \in \theta - \theta_0.$$

Из определений наиболее мощного и несмещенного критериев сразу же вытекает такое следствие: если существует р. н. м.-критерий и по крайней мере один несмещенный критерий, то р. н. м.-критерий — несмещенный.

Отметим, что критерии, используемые в повседневной деятельности, обычно обладают свойством несмещенности, но не являются р. н. м.

10.2.4. Устойчивые критерии. Основная масса критериев, применяемых на практике, является стандартными методами, разработанными не для какой-то одной специальной задачи. Такой критерий обязательно должен обладать следующим свойством: распределение проверочной статистики (и, следовательно, размер критической области) должно быть известно или может быть рассчитано *независимо от вида распределения*, соответствующего проверяемой гипотезе. Подобные критерии называются устойчивыми и это означает, что распределение проверочной статистики зависит только от правильности нулевой гипотезы и обычно от объема доступных данных, но не от самих гипотез.

Так, хорошо известный χ^2-критерий, как мы увидим в дальнейшем, устойчив, поскольку в простейшем случае уровень значимости α зависит только от числа ячеек гистограммы, а не от вида распределений.

Необходимо подчеркнуть, что термин устойчивый относится к уровню значимости, но не к другим свойствам критерия. В частности, мощность критерия обычно очень сильно зависит от вида распределений, даже для устойчивого критерия.

10.2.5. Выбор критерия. По традиции критерий выбирается из сравнения мощностей различных критериев. Этот вопрос уже затрагивался в предыдущих параграфах. Подобная процедура выглядит так: предполагается, что ошибка первого рода (или потеря) равна заданной константе α, и затем выбирается такой критерий, который дает минимальную ошибку второго рода (или примесь) β. Понятно, что такой метод нарушает симметрию между α и β. Более того, он часто нереален в практической деятельности, поскольку при заданном α минимальное значение β может быть еще неприемлемым.

Поэтому при сравнении различных критериев для разделения двух гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1$ будем считать α и β переменными. Как и в п. 10.2.1, для данного критерия и данного значения $\alpha = p(\theta_0)$ можно вычислить $\beta = p(\theta_1)$. Тогда каждому критерию

может быть поставлена в соответствие кривая зависимости β и α подобно тому, как это показано на рис. 10.7.

Пунктирная линия на рис. 10.7 соответствует условию $1 - \beta = \alpha$, поэтому кривые всех несмещенных критериев будут лежать ниже этой линии и проходить через две точки (1,0) и (0,1). Поскольку задача экспериментатора состоит в выборе критерия с минимальными α и β, критерий C будет наихудшим для всех α и β по сравнению со всеми другими критериями, отраженными на этом рисунке. Если гипотезы простые (полностью определены), то тогда существует критерий (Неймана — Пирсона), который при всех α и β по меньшей мере столь же хорош, как и любой другой критерий (см. п. 10.3.1). Если этот критерий слишком сложен или не существует (в случае сложных гипотез), то может возникнуть проблема выбора между различными критериями. Такими критериями могут быть критерии A и B, показанные на рисунке. Очевидно, что критерий A должен быть выбран для $\alpha < \alpha_1$, а B — для $\alpha > \alpha_1$.

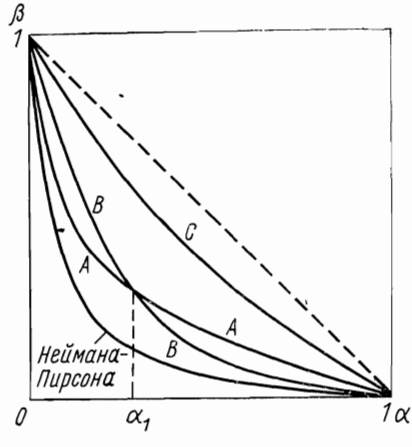


Рис. 10.7. Сравнение критериев на языке α-, β-значений

Если исследуемые гипотезы не простые либо θ может принимать более чем два значения, то рис. 10.7 превращается в многомерную диаграмму с новыми осями, соответствующими θ или другим неопределенным параметрам задачи, либо каждый критерий может быть представлен семейством кривых в α — β-плоскости.

Когда анализируемые гипотезы включают дискретные распределения, то кривые в α — β-плоскости перестают быть непрерывными в связи с тем, что возможен только дискретный набор α. Это приводит к дополнительным трудностям, однако они носят второстепенный характер.

Изложенный выше подход позволяет выбрать наилучший критерий и рассчитать β, если дано α (и все другие возможные параметры определены), или вычислить α, если известно β. Следующий шаг состоит в выборе определенных значений α и β. Для этого нужно знать потери, связанные с принятием ошибочного решения (см. § 10.6 и гл. 6).

§ 10.3. КРИТЕРИИ ДЛЯ ПРОСТЫХ ГИПОТЕЗ

10.3.1. Критерий Неймана — Пирсона. Проблема нахождения *наиболее мощного критерия* проверки гипотезы H_0 в сравнении с гипотезой H_1 может быть сформулирована как задача нахождения

наилучшей критической области (н. к. о.) в X-пространстве. Предположим, что случайная переменная $\mathbf{X} = (X_1, \ldots, X_N)$ имеет ф. п. в. $f_N(\mathbf{X} \mid \theta)$, а пространство параметров состоит только из двух точек θ_0 и θ_1. Предположим также, что отношение функций $f_N(\mathbf{X} \mid \theta_0)$ и $f_N(\mathbf{X} \mid \theta_1)$ всюду непрерывно.

Из определения (10.2) имеем для мощности $(1 - \beta)$-критерия с уровнем значимости α

$$\int_{w_\alpha} f_N(\mathbf{X} \mid \theta_0)\, dx = \alpha; \tag{10.3}$$

$$1 - \beta = \int_{w_\alpha} f_N(\mathbf{X} \mid \theta_1)\, d\mathbf{X}. \tag{10.4}$$

Наша задача — найти для заданного α такую область w_α, которая дает максимальное значение $1 - \beta$. Уравнение (10.4) может быть переписано:

$$1 - \beta = \int_{w_\alpha} \frac{f_N(\mathbf{X} \mid \theta_1)}{f_N(\mathbf{X} \mid \theta_0)} f_N(\mathbf{X} \mid \theta_0)\, d\mathbf{X} = E_{w_\alpha}\left(\frac{f_N(\mathbf{X} \mid \theta_1)}{f_N(\mathbf{X} \mid \theta_0)} \Bigg|_{\theta = \theta_0} \right).$$

Очевидно, что эта величина максимальна, если и только если w_α — такая часть α пространства $\mathbf{X}$, которая содержит наибольшие значения $f_N(\mathbf{X} \mid \theta_1)/f_N(\mathbf{X} \mid \theta_0)$. Таким образом, н. к. о. w_α состоит из точек, удовлетворяющих неравенству

$$l_N(\mathbf{X}, \theta_0, \theta_1) \equiv \frac{f_N(\mathbf{X} \mid \theta_1)}{f_N(\mathbf{X} \mid \theta_0)} \geqslant c_\alpha, \tag{10.5}$$

при этом c_α выбрано таким, чтобы выполнялось условие (10.3).

В результате получается такой критерий:

если $l_N(\mathbf{X}, \theta_0, \theta_1) > c_\alpha$, то принимается $H_1 : f_N(\mathbf{X} \mid \theta_1)$;

если $l_N(\mathbf{X}, \theta_0, \theta_1) \leqslant c_\alpha$, то принимается $H_0 : f_N(\mathbf{X} \mid \theta_0)$.

Такой критерий называется *критерием Неймана — Пирсона*. Проверочная статистика l_N по существу является отношением правдоподобий для двух гипотез и это отношение должно быть известно для всех точек $\mathbf{X}$-пространства наблюдений. Поэтому обе гипотезы H_0 и H_1 должны быть полностью определенными простыми гипотезами и в этих условиях критерий Неймана — Пирсона — *наилучший*.

Соответствующий критерий для непростых гипотез не обязательно оптимален. Эти вопросы обсуждаются в § 10.4 и 10.5.

10.3.2. Пример: критерий знаков в сравнении с критерием для нормального распределения. Предположим, что измеряется некоторая величина, например, интенсивность пучка X. Необходимо знать, увеличилась ли интенсивность пучка после внесения в аппаратуру некоторых модификаций. Чтобы

ответить на этот вопрос, измерим N раз интенсивность пучка и вычислим среднее X. Предположим, что измеренные погрешности распределены по нормальному закону с дисперсией σ^2. Тогда критерий нормального распределения может быть сформулирован следующим образом:

нулевая гипотеза, $H_0 : \mu = \mu_0$, распределение $N(\mu_0, \sigma^2)$,

альтернативная гипотеза, $H_1 : \mu = \mu_1$, распределение $N(\mu_1, \sigma^2)$.

Тогда наиболее мощный критерий основан на проверочной статистике

$$t = \frac{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i-\mu_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(X_i-\mu_0)^2\right]}$$

или

$$\sigma^2 \ln t = \frac{1}{2} N(\mu_0^2-\mu_1^2)+(\mu_1-\mu_0)\sum_{i=1}^{N} X_i.$$

Видно, что t является монотонной функцией переменной

$$z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{N}},$$

которая распределена по нормальному закону $N(\mu_1-\mu_0, 1)$, если верна H_1, и закону $N(0, 1)$, если верна H_0.

Критерий Неймана — Пирсона, таким образом, эквивалентен введению некоторого обрезания на z: если α есть уровень значимости, то критической областью является часть X-пространства, где $z > \lambda_\alpha$,

$$\{X;\ z > \lambda_\alpha\}.$$

λ_α есть α-точка стандартного нормального распределения. Мощность критерия $1-\beta = P(z > \lambda_\alpha)$ при условии, что X_i распределены по $N(\mu, \sigma^2)$ и z по $N(d, 1)$ равна

$$1-\beta = \int_{\lambda_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-(z-d)^2/2]\, dz = \Phi(d-\lambda_\alpha),$$

где $d = \frac{\sqrt{N}}{\sigma}(\mu_1-\mu_0)$.

Мощность $1-\beta$ является возрастающей функцией $(\mu_1-\mu_0)$ и N. Очевидно, $P(\mu_1 = \mu_0) = \alpha$, как и следовало ожидать.

Наряду с таким критерием может быть сконструирован *критерий знаков*. Действительно, пусть:

гипотеза H_0 : N наблюдений имеют в качестве среднего μ_0 и распределены симметрично;

гипотеза H_1 : N наблюдений имеют некоторое другое среднее μ_1.

Пусть далее N_+ будет числом наблюдений, для которых $(X_i - \mu_0)$ положительно, соответственно $N_- = N - N_+$. Если верна гипотеза H_0, то проверочная статистика N_+ распределена по закону:

$$P(N_+ = r) = \binom{N}{r}\left(\frac{1}{2}\right)^N.$$

Предположим, например, что $N = 10$ и имеются два наблюдения, для которых $(X_i - \mu_0)$ отрицательно, $N_- \leqslant 2$. Тогда уровень значимости α равен:

$$\alpha = \frac{1}{2^{10}}\left[\binom{10}{0} + \binom{10}{1} + \binom{10}{2}\right] = \frac{7}{128} = 0{,}0547.$$

Пусть q_+ и q_- соответственно вероятности получить положительный или отрицательный знак, так что $q_+ + q_- = 1$, и пусть $\mu = \mu_1 - \mu_0$.

Тогда $q_- = \Phi(\mu/\sigma)$ и мощность равна

$$p(\mu) = q_+^{10} + 10q_+^9 q_- + 45q_+^8 q_-^2.$$

В табл. 10.1 приведены значения мощностей для $\sigma = 1$.

Таблица 10.1

Значения мощностей для $\sigma = 1$

μ	0	0,5	1
q_-	0,5	0,6915	0,8413
$p(\mu)$	0,055	0,360	0,797

Можно сравнить эти два критерия и показать, что при одинаковых α и N мощность критерия знаков всегда меньше, чем мощность критерия нормального распределения. Поскольку последний критерий является критерием

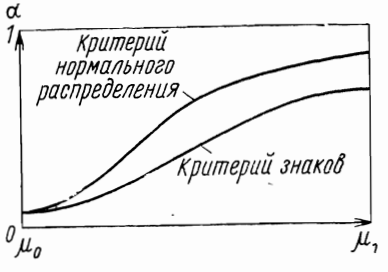


Рис. 10.8. Функции мощности критерия нормального распределения и критерия знаков

Неймана—Пирсона, он принадлежит классу р. н. м.-критериев и его мощность должна быть по меньшей мере не меньше, чем мощность критерия знаков для всех значений μ (рис. 10.8).

§ 10.4. КРИТЕРИИ ДЛЯ СЛОЖНЫХ ГИПОТЕЗ

В предыдущих двух параграфах мы узнали, как можно сконструировать наилучший критерий для разделения двух простых гипотез. Однако не существует общего оптимального метода, когда гипотеза H_0 или H_1 (или обе) не является полностью определенной.

Фактически случай сложных гипотез является наиболее общим, поскольку он возникает всякий раз, когда проверяемая теория со-

держит один или несколько свободных параметров. То же самое может произойти, если делается проверка, которая основана на экспериментальных результатах, включающих ранее неизмерявшиеся эффекты.

В дальнейшем будем различать случаи, когда гипотезы принадлежат *одному непрерывному семейству* и когда они относятся к *различным семействам* гипотез. В первом случае ф. п. в. имеет одну и ту же аналитическую форму; различие между гипотезами обусловлено разными значениями параметров. Поэтому целое параметрическое семейство может быть получено в результате непрерывного изменения параметров. Например, гипотезы могут иметь такой вид:

H_0 : справедливо $f(\mathbf{X}|\theta)$ со значениями $\theta < \theta_0$;

H_1: справедливо то же $f(\mathbf{X}|\theta)$, но со значениями $\theta > \theta_1$.

Гипотезы, принадлежащие различным семействам, не могут быть получены одно из другого непрерывным изменением параметров. Мы встретимся с таким примером в п. 10.5.6. Очевидно, из двух различных семейств $f(\mathbf{X}|\psi)$ и нового параметра λ можно сконструировать более общее параметрическое семейство:

$$\lambda f(\mathbf{X}|\varphi) + (1-\lambda)\, g(\mathbf{X}|\psi).$$

Таким образом, меняя λ от 0 до 1, можно от g н епрерывно перейти к f. Мы обсудим этот случай в п. 10.5.7.

В последующих параграфах мы укажем, при каких условиях для сложных гипотез существует наилучший р. н. м.-критерий и как сконструировать разумные критерии, когда р. н. м.-критерий не существует.

Ввиду важности сложных гипотез основное внимание будет уделено свойствам критерия максимума отношения правдоподобий (см. § 10.5), который обычно используется в таких случаях.

10.4.1. Существование равномерного наиболее мощного критерия для экспоненциального семейства. Так же как и в теореме Дармуа о достаточных статистиках (см. п. 5.3.4), рассмотрим *экспоненциальное семейство* функций, для которых существует р. н. м.-критерий. Следующая теорема формулирует условия, при которых существует р. н. м.-критерий.

Если $X_1, \ldots, X_N$ — независимые, одинаково распределенные случайные переменные с ф. п. в. вида

$$F(X)\, G(\theta) \exp[A(X)\, B(\theta)], \tag{10.6}$$

где $B(\theta)$ — строго монотонная функция, то тогда существует р. н. м.-критерий для гипотезы $H_0 : \theta = \theta_0$ в сравнении с гипотезой $H_1 : \theta > \theta_0$ (отметим, что этот критерий односторонний (см. п. 10.4.2 и работу [48]).

Если использовать критерий Неймана — Пирсона, то наиболее мощный критерий для разделения гипотезы $H_0 : \theta = \theta_0$ против

$H_1 : \theta = \theta_1$ приобретает форму:

$$\frac{G^N(\theta_1)}{G^N(\theta_0)} \exp\left\{\left[\sum_{i=1}^{N} A(X_i)\right][B(\theta_1) - B(\theta_0)]\right\} \gtrless c_\alpha$$

или

$$\sum_{i=1}^{N} A(X_i) \gtrless k_\alpha.$$

Здесь k выбрано таким, чтобы давать требуемый уровень значимости, когда $\theta = \theta_0$. Этот критерий не зависит от θ_1 и, следовательно, является р. н. м.

10.4.2. Односторонние и двусторонние критерии. Если проверяемая гипотеза включает в себя только один параметр θ, то обычно разделяют критерии на односторонние

$$H_0 : \theta = \theta_0;\ H_1 : \theta > \theta_0$$

и двусторонние

$$H_0 : \theta = \theta_0;\ H_1 : \theta \neq \theta_0.$$

Мы уже узнали, что для экспоненциального семейства односторонний р. н. м.-критерий существует (см. п. 10.4.1). Однако двусторонних р. н. м.-критериев обычно не существует. Это можно понять из рис. 10.9. На этом рисунке кривая 1^+ — есть кривая мощности для критерия, который является р. н. м.-критерием для $\theta > \theta_0$. Кривая 1^- — кривая мощности р. н. м.-критерия для

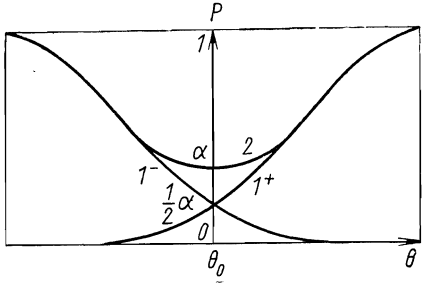


Рис. 10.9. Двусторонний критерий 2 является суммой односторонних критериев 1^+ и 1^-

$\theta < \theta_0$, а кривая 2 — двусторонний критерий, соответствующий сумме критериев 1^+ и 1^-. Если односторонние критерии имеют уровень значимости $\alpha/2$, то двусторонний критерий имеет уровень α, и если размер двустороннего критерия также уменьшить до $\alpha/2$, то понятно, что он менее мощен (по каждую сторону), чем соответствующий односторонний критерий.

В случае экспоненциального семейства р. н. м.-критерий существует для отделения гипотезы $H_0 : \theta \leqslant \theta_1$ или $\theta \geqslant \theta_2$ против гипотезы $H_1 : \theta_1 < \theta < \theta_2$ [48]. Однако не существует р. н. м.-критерия для отделения H_1 против H_0.

Когда нужно построить двусторонний критерий, то его обычно получают, объединяя два односторонних р. н. м.-критерия. Напоминаем, что двустороннего р. н. м.-критерия не существует.

10.4.3. Получение максимальной локальной мощности. Если р. н. м.-критерий не существует, то важное значение приобретают критерии, которые являются наиболее мощными в окрестности нулевой гипотезы. Тогда мы имеем

$$H_0 : \theta = \theta_0; \; H_1 : \theta = \theta_0 + \Delta,$$

где Δ мало.

Разложим логарифм функции правдоподобия в ряд:

$$\ln L(\mathbf{X}, \theta_1) = \ln L(\mathbf{X}, \theta_0) + \Delta \frac{\partial \ln L}{\partial \theta}\Big|_{\theta=\theta_0} + \ldots$$

Если мы опять применим лемму Неймана — Пирсона к H_0 и H_1, то критерий приобретает следующий вид:

$$\ln L(\mathbf{X}, \theta_1) - \ln L(\mathbf{X}, \theta_0) \gtrless c_\alpha$$

или

$$\frac{\partial \ln L}{\partial \theta}\Big|_{\theta=\theta_0} \gtrless k_\alpha.$$

Если наблюдения независимы и распределены одинаково, то

$$E\left(\frac{\partial \ln L}{\partial \theta}\Big|_{\theta=\theta_0}\right) = 0;$$

$$E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right] = +NI.$$

Здесь N — число наблюдений, а I матрица информации (см. п. 5.2.1). При соответствующих условиях (см. п. 7.3.1) $\partial \ln L/\partial\theta$ — приблизительно нормально. Следовательно, локально наиболее мощный критерий приблизительно задается соотношением

$$\frac{\partial \ln L}{\partial \theta}\Big|_{\theta=\theta_0} \gtrless \lambda_\alpha \sqrt{NI}. \tag{10.7}$$

§ 10.5. КРИТЕРИЙ ОТНОШЕНИЯ ПРАВДОПОДОБИЙ

Пусть наблюдения $\mathbf{X}$ распределены по закону $f(\mathbf{X} \mid \boldsymbol{\theta})$, зависящему от двух параметров $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Тогда функция правдоподобия равна

$$L(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta}). \tag{10.8}$$

В общем случае, если Θ — все пространство параметров θ, а ν — некоторое подпространство Θ, то любой критерий параметрических гипотез (принадлежащих одному семейству) может быть сформулирован так:

$$\left.\begin{aligned} H_0 &: \boldsymbol{\theta} \in \nu; \\ H_1 &: \boldsymbol{\theta} \in \Theta - \nu. \end{aligned}\right\} \tag{10.9}$$

Примеры гипотез, относящихся к одному параметрическому семейству:

1) $H_0 : \theta_1 = a, \theta_2 = b;$
   $H_1 : \theta_1 \neq a, \theta_2 \neq b;$
2) $H_0' : \theta_1 = c, \theta_2$ — не определен;
   $H_1' : \theta_1 \neq c, \theta_2$ — не определен;
3) $H_0'' : \theta_1 + \theta_2 = d;$
   $H_0' : \theta_1 + \theta_2 \neq d.$

10.5.1. Проверочная статистика. Используя функцию правдоподобия (10.8) и обозначения (10.9), можно определить *отношение максимумов правдоподобия*:

$$\lambda = \frac{\max\limits_{\boldsymbol{\theta} \in \nu} L(\mathbf{X} \mid \boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta} \in \Theta} L(\mathbf{X} \mid \boldsymbol{\theta})}. \tag{10.10}$$

Очевидно, $0 \leqslant \lambda \leqslant 1$. Конечно, λ служит разумной проверочной статистикой для H_0. Для случая простой H_0 по сравнению с простой H_1 она обычно соответствует наиболее мощному критерию (см. п. 10.3.1). В других более сложных случаях полезность λ как проверочной статистики обусловлена тем фактом, что она всегда является функцией достаточной статистики задачи. Основное достоинство этого метода — в выработке рабочих критериев с хорошими свойствами по крайней мере при больших объемах наблюдений.

Гипотезы (10.9) обычно имеют такую форму:

$$H_0 : \theta_i = \theta_{i0},\ i = 1, 2, \ldots, r \text{ [скажем, } \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}];$$

$$\theta_j \text{ — не определены [скажем, } \boldsymbol{\theta}_s];$$

$$H_1 : \theta_i \neq \theta_{i0},\ i = 1, 2, \ldots, r;$$

$$\theta_j \text{ — не определены; } j = 1, \ldots, s.$$

Здесь нулевая гипотеза приписывает определенные значения подмножеству параметров, тогда как альтернативная гипотеза H_1 приписывает параметрам любые значения, отличные от указанных гипотезой H_0.

Тогда *отношение максимумов правдоподобия* (10.10) есть отношение максимального значения $L(\mathbf{X} \mid \boldsymbol{\theta})$ на подмножестве θ_j, $j = 1, \ldots, s$ с фиксированными $\theta_i \neq \theta_{i0}$, $i = 1, \ldots, r$ к максимуму $L(\mathbf{X} \mid \boldsymbol{\theta})$ на всем пространстве параметров.

В этих обозначениях статистика (10.10) приобретает вид:

$$\lambda = \frac{\max\limits_{\theta_s} L(\mathbf{X} \mid \theta_{r_0}, \theta_s)}{\max\limits_{\theta_r,\ \theta_s}{}' L(\mathbf{X} \mid \theta_r, \theta_s)} \tag{10.11}$$

или

$$\lambda = \frac{L(\mathbf{X} \mid \theta_{r0}, \theta_s'')}{L(\mathbf{X} \mid \theta_r', \theta_s')}.$$

Здесь θ_s'' — значение θ_s в максимуме в ограниченной θ-области, а θ_r', θ_s' — значения θ_r, θ_s в максимуме во всей θ-области.

10.5.2. Асимптотическое распределение для непрерывных семейств гипотез. Мы уже знаем, что отношение правдоподобий является разумной проверочной статистикой. Теперь нужно по распределению этой статистики для нулевой гипотезы определить критическую область. Однако часто эта процедура довольно затруднительна, поскольку распределение может быть неизвестно либо слишком громоздко. Иногда можно прибегать к помощи метода Монте-Карло, но это не всегда устраивает физика. Обычно поступают так: отыскивают *асимптотическое распределение* отношения правдоподобий и используют его как приближение к истинному распределению.

Асимптотически дисперсия оценки метода м. п. приближается к границе минимальной дисперсии (см. п. 7.4.1). Из этого следует (см. п. 7.3.2), что функция правдоподобия $L(\mathbf{X} \mid \theta)$ должна иметь форму

$$\frac{\partial \ln L}{\partial \theta} = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right](\hat{\theta} - \theta)$$

или

$$L(\mathbf{X} \mid \theta) \sim \exp\left[\frac{1}{2} E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)(\hat{\theta} - \theta)^2\right]. \tag{10.12}$$

В случае многомерного θ отношение (10.12) приобретает вид

$$L(\mathbf{X} \mid \theta) = L(\mathbf{X} \mid \theta_r, \theta_s) \sim \exp\left[-\frac{1}{2}(\hat{\theta} - \theta)^T \underset{\sim}{I}(\hat{\theta} - \theta)\right], \tag{10.13}$$

где $\underset{\sim}{I}$ — матрица информации для θ:

$$\underset{\sim}{I} = \begin{bmatrix} \underset{\sim}{I}_r & \vdots & \underset{\sim}{I}_{rs} \\ \cdots & \cdots & \cdots \\ \underset{\sim}{I}_{rs}^T & \vdots & \underset{\sim}{I}_s \end{bmatrix}.$$

Поэтому уравнение (10.13) может быть переписано:

$$L(\mathbf{X}\mid\theta_r,\theta_s)\sim\exp\left\{-\frac{1}{2}\left[(\hat{\theta}_r-\theta_r)^T\underset{\sim}{I}_r(\hat{\theta}_r-\theta_r)+\right.\right.$$
$$\left.\left.+2(\hat{\theta}_r-\theta_r)^T\underset{\sim}{I}_{rs}(\hat{\theta}_s-\theta_s)+(\hat{\theta}_s-\theta_s)^T\underset{\sim}{I}_s(\hat{\theta}_s-\theta_s)\right]\right\}. \quad (10.14)$$

Нам также известно из п. 7.3.1, что асимптотическое распределение $\hat{\theta}=(\hat{\theta}_r,\hat{\theta}_s)$ — $(r+s)$-мерное нормальное с матрицей вторых моментов $\underset{\sim}{I}^{-1}$. Таким образом, видно, что функция правдоподобия асимптотически стремится к ф. п. в. $\hat{\theta}$.

Рассмотрим сейчас значение выражения (10.13), когда определяется максимум $L(\mathbf{X}\mid\theta)$ по всем θ:

$$L(\mathbf{X}\mid\theta_r',\theta_s').$$

Из того факта, что L может быть записано в виде (10.13), следует, что

$$\theta_r'=\hat{\theta}_r \text{ и } \theta_s'=\hat{\theta}_s$$

и

$$L(\mathbf{X}\mid\theta_r',\theta_s')\sim 1. \quad (10.15)$$

Пусть опять-таки ограниченная область задается условием $\theta_r=\theta_{r0}$. Значение $L(\mathbf{X}\mid\theta)$ в максимуме на этой области равно $L(\mathbf{X}\mid\theta_{r0},(\theta_s'')$. Аналогично из того факта, что L может быть записана в виде (10.13), следует, что $\theta_s''=\hat{\theta}_s$. Тогда уравнение (10.14) приобретает вид

$$L(\mathbf{X}\mid\theta_{r0},\theta_s'')\sim\exp\left[-\frac{1}{2}(\hat{\theta}_r-\theta_{r0})^T\underset{\sim}{I}_r(\hat{\theta}_r-\theta_{r0})\right]. \quad (10.16)$$

Используя в уравнении (10.11) соотношения (10.15) и (10.16), получаем для отношения правдоподобий

$$\lambda=\frac{L(\mathbf{X}\mid\theta_r,\theta_s'')}{L(\mathbf{X}\mid\theta_r',\theta_s')}=\exp\left[-\frac{1}{2}(\hat{\theta}_r-\theta_{r0})^T\underset{\sim}{I}_r(\hat{\theta}_r-\theta_{r0})\right].$$

Из асимптотических свойств $\hat{\theta}_r$ следует, что асимптотическое распределение величин

$$-2\ln\lambda=\left[(\hat{\theta}_r-\theta_{r0})^T\underset{\sim}{I}_r(\hat{\theta}_r-\theta_{r0})\right]$$

имеет вид центрального χ^2 (r)-распределения, если верна H_0, и нецентрального χ^2 (r)-распределения с параметром нецентральности

$$K_1=(\theta_r-\theta_{r0})^T\underset{\sim}{I}_r(\theta_r-\theta_{r0}),$$

если верна H_1. Иными словами, если H_0 накладывает r связей на $(s+r)$ параметров в гипотезах H_0 и H_1, то величина $-2\ln\lambda$ распределена, как χ^2 (r), если верна H_0.

10.5.3. Асимптотическая мощность для непрерывных семейств гипотез. Используя результаты п. 10.5.2, можно рассчитать мощность критерия. Критическая область задается значениями $-2\ln\lambda$, большими, чем α-точка $\chi^2(r)$-распределения, обозначим ее $\chi^2_\alpha(r)$. Параметр нецентральности K_1 является мерой «расстояния» альтернативной гипотезы от H_0; матрица I_r, входящая в $-2\ln\lambda$, может быть рассчитана. Тогда мощность критерия задается соотношением:

$$p(K)=\int\limits_{\chi^2_\alpha(r)}^{\infty} dF_1[\chi^2(2, K_1)], \tag{10.17}$$

где $F_1[\chi^2(r, K_1)]$ — функция нецентрального χ^2-распределения. Используя приближение для $\chi^2(N, \Delta)$, данное в конце п. 4.2.3, получаем

$$p(K)=\int\limits_{\left(\frac{r+K_1}{r+2K_1}\right)\chi^2_\alpha(r)}^{\infty} dF_2\left[\chi^2\left(r+\frac{K_1^2}{r+2K_1}\right)\right],$$

где F_2 — функция центрального χ^2-распределения.

Можно показать, что критерий отношения правдоподобий *состоятелен*. С ростом числа наблюдений N мощность критерия стремится к 1 для всех альтернативных гипотез H_1, т. е.

$$\lim_{N\to\infty} p(\mathbf{X}\in w_\alpha \mid H_1)=1,$$

где w_α — критическая область, заданная критерием.

Из состоятельности критерия следует, что он *асимптотически не смещен*. Однако несмещенность — довольно важное свойство и его желательно иметь также для выборок малого объема.

При проверке гипотез о параметрах расположения и масштаба с использованием критерия отношения правдоподобий имеет место такой факт: если обеспечена несмещенность оценок максимального правдоподобия, то результирующий критерий также не смещен. Это служит хорошим поводом для устранения смещения оценок.

Отметим, что во всех этих рассуждениях предполагались «обычные» условия регулярности и независимости области измерений от параметров. Если эти предположения не выполняются, то при некоторых условиях [2, 49] величина $-2\ln\lambda$ распределена точно, как $\chi^2(2r)$.

10.5.4. Примеры. 1. Проверка справедливости $\Delta S/\Delta Q$ правила в лептонных распадах K^0-мезонов.

Предположим, что цель данного эксперимента состоит в определении отношения X двух распадных амплитуд, которые являются комплексными числами

$$X=\frac{A(\Delta S=-\Delta Q)}{A(\Delta S=\Delta Q)}.$$

В общем случае X может быть произвольным комплексным числом, но существуют три важные физические гипотезы, которые дают такие предсказания о X [рис. 10.10 и уравнение (10.18), формулирующие гипотезы]:

гипотеза A : если справедливо правило $\Delta S = \Delta Q$, то $X = 0$;

гипотеза B : если в распаде CP-четность сохраняется, то $\mathrm{Im}\ (X) = 0$;

гипотеза C : если справедлива теория «максимального CP-нарушения», то число X чисто мнимое и не равно нулю.

Современное состояние теории слабых взаимодействий таково, что значение X нужно только для того, чтобы подтвердить одну из гипотез A, B, C либо исключить все эти гипотезы. Поэтому в данном случае мы имеем дело не столько с оценкой, сколько с проверкой гипотез.

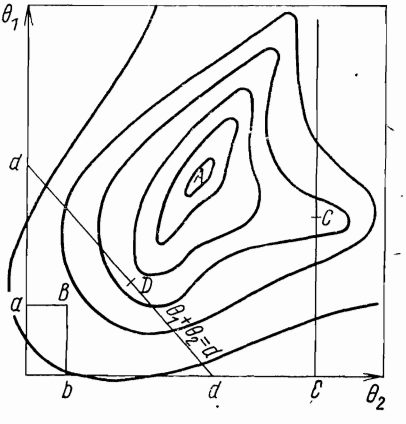


Рис. 10.10. Линии уровня правдоподобия и критерии гипотез для примера о проверке правила $\Delta S/\Delta Q$

В этом эксперименте гипотеза A простая, а гипотеза B — сложная. Гипотеза A является частным случаем гипотезы B. Гипотеза C также сложная и отличается от гипотез B и A. Альтернативой по отношению ко всем этим гипотезам является предположение о том, что $\mathrm{Re}\ (X)$ и $\mathrm{Im}\ (X)$ обе не равны нулю.

Если из эксперимента известна функция правдоподобия, то можно по крайней мере в принципе составить диаграмму линий уровня правдоподобия $L(X, \boldsymbol{\theta})$. (Здесь X имеет смысл вектора наблюдений!)

В примере на рис. 10.10 точка d соответствует максимуму L. Точки b и c соответствуют максимуму функций правдоподобия, когда верна гипотеза B или C соответственно.

Отношение правдоподобий для гипотезы A равно

$$\lambda_a = L\,(0)/L\,(d).$$

Если гипотеза A верна, то величина $-2 \ln \lambda_a$ распределена асимптотически, как χ^2 (2), и это соответствует обычному критерию правила $\Delta S = \Delta Q$.

Для проверки гипотезы B нужно составить такое отношение максимумов правдоподобия:

$$\lambda_b = L\,(b)/L\,(d).$$

Если верна гипотеза B, то $-2 \ln \lambda_b$ распределена асимптотически, как χ^2 (1).

Наконец, если предположить, что CP сохраняется, то можно использовать такое отношение максимумов правдоподобий:

$$\lambda_{ab} = L\,(0)/L\,(b).$$

Если A (и, следовательно, B) справедливо, то $-2 \ln \lambda_{ab}$ распределено асимптотически, как χ^2 (1).

2. Проверка равенства нулю среднего данных, распределенных по нормальному закону с неизвестной дисперсией.

Предположим, что данные $X_1, \ldots, X_N$ распределены по закону $N(\mu, \sigma^2)$; $H_0 : \mu = 0^2$, σ^2 — неизвестно, $H_1 : \mu \neq 0$, σ^2 неизвестно. Функция правдоподобия равна

$$L(\mu, \sigma^2) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\frac{(X_i-\mu)^2}{\sigma^2}\right].$$

Если справедлива H_0, то оценка σ^2, найденная методом м. п., равна:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2.$$

Поэтому

$$\max_{\sigma^2} L(0, \sigma^2) = \exp(-N/2)/(2\pi)^{N/2}\left(\frac{\Sigma X_i^2}{N}\right)^{N/2}.$$

В рамках общей гипотезы оценки м. п. равны:

$$\hat{\mu} = \overline{X};\ s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2$$

и

$$\max_{\mu,\,\sigma^2} L(\mu, \sigma^2) = \frac{\exp(-N/2)}{(2\pi)^{N/2}\left[\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2\right]^{N/2}}.$$

Составляя отношение, приходим к такому выражению:

$$\lambda = \left[\frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{\sum_{i=1}^{N}X_i^2}\right]^{N/2}.$$

После некоторых преобразований получим

$$\lambda^{2/N} = \frac{1}{(1+t^2/(N-1))},$$

где

$$t = \frac{\sqrt{N}\,\overline{X}}{\sqrt{\frac{1}{N-1}\Sigma(X_i-\overline{X})^2}} = \sqrt{N}\,\overline{X}/S$$

является переменной распределения Стьюдента (см. п. 4.2.4). Таким образом, критерий, основанный на λ, эквивалентен критерию, основанному на двустороннем t-критерии Стьюдента. Критическая область соответствует малым значениям λ.

3. Проверка одинаковости среднего у данных, распределенных по закону Пуассона.

Допустим, что $X_1, \ldots, X_N$ — независимые наблюдения, распределенные по закону Пуассона со средними μ_i. Нулевая гипотеза такова:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu.$$

Ее нужно сравнить с противоположной гипотезой: $H_1 : \mu_1$ произвольные, но не все равны друг другу. Функция правдоподобия равна

$$L(\mu_1 \ldots, \mu_N) = \prod_{i=1}^{N} \exp(-\mu_i)\, \mu_i^{X_i}/X_i!.$$

Если верна H_0, то оценка максимального правдоподобия такова: $\hat{\mu} = \overline{X}$; если же верна H_1, то $\hat{\mu}_i = X_i$. Тогда

$$\max_{\mu} L(\mu) = \frac{\exp(-N\overline{X})\, \overline{X}^{N\overline{X}}}{\prod_{i=1}^{N} X_i!}$$

и

$$\max_{\mu_i} L(\mu_1, \ldots, \mu_N) = \exp(-N\overline{X}) \prod_{i=1}^{N} X_i^{X_i}/X_i!$$

Следовательно, отношение правдоподобий равно:

$$\lambda = \overline{X}^{N\overline{X}} \Big/ \prod_{i=1}^{N} X_i^{X_i}$$

или

$$\ln \lambda = N\overline{X} \ln \overline{X} - \sum_{i=1}^{N} \overline{X}_i \ln \overline{X}_i.$$

Это выражение пока еще не похоже на обычные критерии, однако в асимптотическом пределе оно эквивалентно такому выражению:

$$\sum_{i=1}^{N} (X_i - \overline{X})^2/\overline{X},$$

которое в том же пределе распределено как χ^2 с $(N-1)$ степенями свободы. Докажем это утверждение. Положим $X_i = \overline{X} + \delta_i$ и предположим, что δ_i мало. Тогда выражение $\ln \lambda$ можно преобразовать:

$$\ln \lambda = N\overline{X} \ln \overline{X} - \sum_{i=1}^{N} (\overline{X}+\delta_i) \ln (\overline{X}+\delta_i) = N\overline{X} \ln \overline{X} -$$

$$- \sum_{i=1}^{N} (\overline{X}+\delta_i) \ln \overline{X} - \sum_{i=1}^{N} (\overline{X}+\delta_i) \left(\frac{\delta_i}{\overline{X}} - \frac{\delta_i^2}{2\overline{X}^2} + \ldots \right) =$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \frac{\delta_i^2}{\overline{X}} + O(\delta_i^3).$$

Сумма членов порядка δ_i равна нулю по построению. Отметим, что $\sum_{i=1}^{N} \delta_i^2/\overline{X}$ есть уже χ^2-статистика. Таким образом, мы приходим к обычному резуль-

тату, т. е. в окрестности нулевой гипотезы — $2 \ln \lambda$ асимптотически эквивалентно χ^2 $(N - 1)$.

Для выборок сравнительно небольших объемов неясно, что лучше: χ^2-критерий или критерий отношения правдоподобий. Скорее всего это зависит от альтернативных гипотез.

10.5.5. Случай выборок небольших размеров. Хотя асимптотические свойства критерия отношений правдоподобия для непрерывных семейств гипотез просты, его свойства для малых выборок не совсем ясны.

Если справедлива гипотеза H_0, то можно найти более точное приближение для распределения $-2 \ln \lambda$, чем то, которое получается в асимптотическом пределе. В § 9.5 было показано (см. табл. 9.2), что

$$E[-2 \ln \lambda] = r\left(1 + \frac{a}{N} + \ldots\right),$$

где в случае одного параметра

$$a = \frac{1}{I}\left\{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2 \frac{\partial^2 \ln L}{\partial \theta^2}\right] - \right.$$
$$\left. - \frac{1}{3} \frac{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^3\right] E\left[\frac{\partial^3 \ln L}{\partial \theta^3}\right]}{I}\right\}.$$

Тогда оказывается, что величина $-2 \ln \lambda/(1 + a/N)$ ведет себя, как χ^2 (r), с точностью до членов порядка $1/N^2$[2]. Соответствующий масштабный множитель, несомненно, улучшает приближение и его нужно использовать, когда $N < 100$.

Особого внимания заслуживает частный случай, когда приходится сталкиваться с *линейной моделью* (см. п. 8.4.1 и 8.4.3). Для такой модели распределение известно также для конечного N. Предположим, что некоторые наблюдения Y_i выражаются через другие наблюдения X_i и случайную ошибку ε_i:

$$Y_i = \sum_k \theta_k f_k(X_i) + \varepsilon_i,$$

где ε — распределено по нормальному закону со средним 0 и постоянной дисперсией.

Предположим, что необходимо проверить, равны ли θ_k некоторым постоянным значениям θ_{0k} или, если говорить вообще, удовлетворяют ли компоненты вектора $\boldsymbol{\theta}$ некоторой системе r линейных уравнений:

$$\underset{\sim}{A}\boldsymbol{\theta} = \mathbf{b}, \tag{10.18}$$

где $\underset{\sim}{A}$ и $\mathbf{b}$ известны. Частным случаем подобной задачи может быть задача о влиянии последнего члена разложения некоторой функции в ряд.

Правдоподобие для обеих гипотез [H_0: вектор θ удовлетворяет уравнению (10.18), H_1: уравнение связи не имеет места]:

$$L(\mathbf{X} \mid \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left[Y_i - \sum_{k=1}^{r} \theta_k \rho_k(\mathbf{X}_i)\right]^2\right\}. \tag{10.19}$$

Будем считать вначале, что дисперсия σ^2 *известна*. Если верна гипотеза H_0, то оценки $\hat{\theta}_{0k}$ параметров нужно отыскивать из условия максимума (10.19) при одновременном выполнении уравнений связи (10.18). Если же верна H_1, то оценки $\hat{\theta}_{1k}$ параметров находятся из абсолютного максимума (10.19). Тогда отношение максимумов правдоподобий равно

$$-2\ln\lambda = \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[Y_i - \sum_{k=1}^{r} \hat{\theta}_{0k} \rho_k(X_i)\right]^2 - \frac{1}{\sigma^2} \sum_{i=1}^{N} \left[Y_i - \sum_{k=1}^{r} \hat{\theta}_{1k} \rho_k(X_i)\right]^2.$$

Было показано, что второй член можно представить в виде суммы первого члена и некоторой квадратичной формы вектора ошибок ε. Вывод о том, что величина $-2\ln\lambda$ ведет себя, как $\chi^2(r)$, доказанный в п. 10.5.2 для асимптотического предела, здесь выполняется точно [50]. Такой же результат справедлив и для случая, когда ошибки ε не независимы, а имеют некоторую известную матрицу вторых моментов. При этом нужно использовать оценки, полученные из условия минимума взвешенных наименьших квадратов (см. § 8.4).

Если дисперсия σ^2 *неизвестна* и ее нужно оценить по данным, то поступают несколько иначе. Функция правдоподобия $L(\mathbf{X} \mid \theta, \sigma^2)$ по-прежнему задается правой частью (10.19). Результирующая сумма квадратов для гипотезы H_0 равна

$$\sum_{i=1}^{N} \left[Y_i - \sum_{k=1}^{r} \hat{\theta}_{0k} \rho_k(X_i)\right]^2.$$

Это приводит к оценке

$$S_\omega^2 = \frac{1}{N} \sum_{i=1}^{N} \left[Y_i - \sum_{k=1}^{r} \hat{\theta}_{0k} \rho_k(X_i)\right]^2, \tag{10.20}$$

где ω — подпространство пространства θ, задаваемое уравнениями связи. Тогда максимальное значение правдоподобия для гипотезы H_0 равно

$$\max_{\omega} L(X \mid \theta, \sigma^2) = \frac{1}{(2\pi)^{N/2} \left(S_\omega^2\right)^{N/2}} \exp(-N/2).$$

Подобно этому, если NS_{Ω}^2 — результирующая сумма квадратов для гипотезы H_1 и если Ω обозначает все θ-пространство, максимальное значение правдоподобия равно

$$\max_{\Omega} L(\mathbf{X} \mid \theta, \sigma^2) = \frac{1}{(2\pi)^{N/2} \left(S_{\Omega}^2\right)^{N/2}} \exp(-N/2).$$

Тогда для отношения правдоподобий получаем:

$$\lambda = (S_{\Omega}^2/S_{\omega}^2)^{N/2}$$

или

$$\lambda^{-2/N} = 1 + (S_{\omega}^2 - S_{\Omega}^2)/S_{\Omega}^2.$$

Можно показать [50], что $(S_{\omega}^2 - S_{\Omega}^2)/\sigma^2$ и S_{Ω}^2/σ^2 распределены независимо как χ^2 с r и $(N - S)$-степенями свободы соответственно. Поэтому переменная

$$F = (\lambda^{-2/N} - 1)(N - S)/r$$

распределена, как переменная Фишера — Снедекора $F(r, N - s)$, если уравнения связи выполняются.

В случае только одного уравнения связи мы приходим к критерию Стьюдента, где $t^2 = F(1, N - 1)$. Когда уравнения связи не выполняются, функция (10.20) следует нецентральному закону Фишера — Снедекора, что позволяет вычислить мощность критерия.

Пример. Аппроксимация кривой полиномом. Как указывалось в п. 8.4.2, лучше всего при оценке использовать полиномы $\xi_j(X)$, ортогональные на пространстве наблюдений X. Тогда аппроксимирующее выражение может быть записано как

$$Y_i = \sum_j \psi_j \xi_j(X_i) + \varepsilon_i.$$

Оценки параметров

$$\hat{\psi}_j = \sum_{i=1}^{N} \xi_j(X_i) Y(X_i).$$

Важным свойством оценок $\hat{\psi}_j$ является их независимость и, в частности, оценка $\hat{\psi}_j$ не зависит от максимальной степени используемых полиномов.

Какова должна быть максимальная степень полинома?

Пусть S_j — результирующая сумма квадратов, которая включает все степени полиномов по j включительно. Из предыдущих рассуждений следует, что

$$F = \frac{S_{j-1} - S_j}{S_j}(N - j - 1)$$

распределена, как переменная Фишера — Снедекора $F(1, N - j - 1)$, если введение полинома j-й степени не оправдано. Тогда, используя табличные значения F-распределения, можно выработать соответствующие рекомендации. Они приведены в табл. 10.2.

Подобные методы могут быть выработаны для проверки дисперсии ε. В табл. 10.3 дана сводка критериев для среднего и дисперсии нормального распределения (в одномерном случае, рис. 10.11).

Таблица 10.2

Рекомендации по выбору максимальной степени полинома

$N-j-1$	2	3	4	6	8	12	20	60	120
Отбросить полином j-го порядка с 95%-ным доверительным уравнением, если F меньше	18,5	10,1	7,7	6	5,3	4,7	4,3	4	3,9

10.5.6. Пример различных семейств гипотез. Перейдем теперь от непрерывных семейств гипотез к различным параметрическим семействам. Возьмем для примера два семейства

$$f(X, \varphi) = \varphi \exp(-\varphi X)$$

или экспоненциальное распределение (4.37) и

$$g(X, \psi) = \psi^2 X \exp(-\psi X).$$

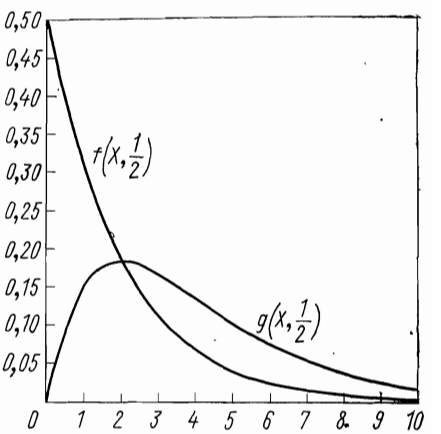


Рис 10.11. Для $\varphi = \psi = 1/2$, $f(X, \varphi)$ и $g(X, \psi)$ распределены, как $\chi^2(2)$- и $\chi^2(4)$-распределения соответственно

Параметры φ и ψ неизвестны, но, используя метод максимального правдоподобия, получаем для их оценок:

$$\hat{\varphi} = 1/\overline{X}, \quad \hat{\psi} = 2/\overline{X}.$$

Здесь $\overline{X}$ — среднее N наблюдений. Отношение максимальных значений функций правдоподобия равно

$$-2 \ln \lambda = -2 \ln \frac{\prod_{i=1}^{N} f(X_i, \hat{\varphi})}{\prod_{i=1}^{N} g(X_i, \hat{\psi})} = -2 \left[N \ln \hat{\varphi} - \hat{\varphi} N \overline{X} - 2N \ln \hat{\psi} - \right.$$

$$\left. - \sum_{i=1}^{N} \ln X_i + \hat{\psi} N \overline{X} \right] \qquad (10.21)$$

Таблица 10.3

Критерии для нормального распределения, одномерный случай

	Критерий для среднего	
	Дисперсия известна	Дисперсия неизвестна
Асимптотические свойства	$-2\ln\lambda=\frac{N}{\sigma^2}(\overline{X}-m_0)^2\sim\chi^2(1)$, если m_0 — истинное среднее	$-2\ln\lambda=N\ln\left[\frac{\frac{1}{N}\sum_i(X_i-\overline{X})^2}{\frac{1}{N}\sum_i(X_i-m_0)^2}\right]=N\ln\left(\frac{S^2}{\hat{\sigma}^2}\right)\sim\chi^2(1)$, если m_0 — истинное среднее
Выборка конечного размера	$-2\ln\lambda\sim\chi^2(1)$, если m_0 — истинное среднее	$\lambda^{2/N}=\frac{\frac{1}{N}\sum_i(X_i-m_0)^2}{\frac{1}{N}\sum_i(X_i-\overline{X})}=\frac{\hat{\sigma}^2}{S^2}=\frac{1}{1+\frac{1}{N-1}t^2}$; $t=\frac{\sqrt{N}(\overline{X}-m_0)}{\frac{1}{N-1}\sum_i(\overline{X}_i-\overline{X})^2}\sim$ закону Стьюдента с $N-1$ степенями свободы, если m_0 — истинное среднее

	Критерий для дисперсии	
	Среднее известно	Среднее неизвестно
Асимптотические свойства	$-2\ln\lambda=\frac{N}{2}\left[\frac{\hat{\sigma}^2}{\sigma_0^2}+\ln\frac{\hat{\sigma}^2}{\sigma_0^2}-1\right]\sim\chi^2(1)$, если σ_0^2 — истинная дисперсия $\hat{\sigma}^2=\frac{1}{N}\sum_i(X_i-m_0)^2$	$-2\ln\lambda=\frac{N}{2}\ln\frac{S^2}{\sigma_0^2}\sim\chi^2(1)$, если σ_0^2 — истинная дисперсия $S^2=\frac{1}{N}\sum_i(X_i-\overline{X})^2$
Свойства выборки конечного размера	$-2\ln\lambda$ — монотонная функция $\hat{\sigma}^2/\sigma_0^2$ $N\frac{\hat{\sigma}^2}{\sigma_0^2}\sim\chi^2(N)$, если σ_0^2 — истинная дисперсия	$-2\ln\lambda$ — монотонная функция S^2/σ_0^2 $(N-1)\frac{S^2}{\sigma_0^2}\sim\chi^2(N-1)$, если σ_0^2 — истинная дисперсия

или

$$-2N^{-1}\ln\lambda = -2\left[\ln\overline{X} - N^{-1}\sum_{i=1}^{N}\ln X_i + 1 - 2\ln 2\right]. \qquad (10.22)$$

К каждой из сумм такого выражения мы уже можем применить центральную предельную теорему. В пределе больших N получим

$$-2N^{-1}\ln\lambda \to -2[\ln E(X) - E(\ln X) + 1 - 2\ln 2]. \qquad (10.23)$$

Проверочная статистика (10.22) распределена вблизи значения (10.23) по нормальному закону с дисперсией

$$\sigma^2_{-2N^{-1}\ln\lambda} = \frac{4}{N}\left\{\frac{\sigma_X^2}{[E(X)]^2} + E[(\ln^2(X))] - [E^2(\ln(X))] - 2\times \right.$$
$$\left. \times\frac{E(\overline{X}\ln X)}{E(X)} + 2E(\ln X)\right\}. \qquad (10.24)$$

Здесь мы использовали асимптотическое разложение

$$\ln\overline{X} = \ln[E(X)] + \frac{1}{E(X)}[\overline{X} - E(X)] + \ldots$$

Очевидно, величина ожиданий, входящих в предыдущие выражения, зависит от того, какая гипотеза верна. Все ожидания могут быть вычислены аналитически (представляем читателю проделать это в качестве упражнения):

$$\sigma^2_{-2N^{-1}\ln\lambda} = \begin{cases} \dfrac{4}{N}\left(\dfrac{\pi^2}{6} - 1\right), & \text{если } f \text{ верна;} \\ \dfrac{4}{N}\left(\dfrac{\pi^2}{6} - \dfrac{1}{2}\right), & \text{если } g \text{ верна.} \end{cases} \qquad (10.25)$$

10.5.7. Общие методы проверки гипотез, принадлежащих разным семействам. 1. В п. 10.5.6 мы видели, что распределение статистики (10.21) в общем случае зависит от N (и от истинного значения параметров [51]) и не является χ^2-распределением, когда f верна. Однако это не означает, что отношение правдоподобий не дает никакой пользы при сравнении гипотез, принадлежащих различным семействам. Если распределение отношения учесть должным образом, то его можно использовать при конструировании других проверочных статистик. Например, функция [51]

$$\frac{-2N^{-1}\ln\lambda - E(-2N^{-1}\ln\lambda)}{\sigma_{-2N^{-1}\ln\lambda}} \qquad (10.26)$$

асимптотически распределена по нормальному закону, если ожидания вычислены по отношению к правильной гипотезе. Для распределений f и g предыдущего параграфа ошибка первого рода α соответствует такой λ_α-точке функции (10.26):

$$\lambda_\alpha = \frac{1 + c - 3\ln 2}{\frac{1}{N}\left(\frac{\pi^2}{6} - 1\right)},$$

где $c \sim 0{,}577$ — константа Эйлера. λ_β-Точка, соответствующая ошибке второго рода β, равна:

$$\lambda_\beta = \frac{\frac{2\lambda_\alpha}{N}\left(\frac{\pi^2}{6} - 1\right) + 1 + \ln 2}{\frac{2}{N}\left(\frac{\pi^2}{6} - \frac{1}{2}\right)}.$$

Если моменты $-2\ln\lambda$ зависят от истинного значения параметров, то можно построить подобные критерии, заменяя неизвестные значения их оценками (и соответственно модифицируя дисперсию).

2. Есть другой подход, который обходится без вычисления этих моментов. Он состоит в том, что конструируются *общие* параметрические гипотезы. Для проверки $f(\mathbf{X}, \varphi)$ против $g(\mathbf{X}, \psi)$ общее семейство имеет вид:

$$h(\mathbf{X}, \theta, \varphi, \psi) = (1 - \theta) f(X, \varphi) + \theta g(\mathbf{X}, \psi).$$

Теперь можно использовать критерий отношения максимумов правдоподобия для сравнения гипотезы:

$$H_0 : \theta = 0, \ \varphi, \ \psi \text{ неизвестны}$$

с гипотезой: $H_1 : \theta \neq 0, \ \varphi, \ \psi$ неизвестны.

Проверочная статистика равна:

$$-2\ln\lambda = -2\ln f(\mathbf{X}, \hat{\varphi}') + 2\ln\left[(1-\hat{\theta}'')\, f(\mathbf{X}, \hat{\varphi}'') + \hat{\theta}_g''(\mathbf{X}, \hat{\psi}'')\right], \tag{10.27}$$

где $\hat{\varphi}'$ — оценка максимального правдоподобия параметра φ на ограниченной области $\theta = 0$, тогда как $\hat{\theta}'', \hat{\varphi}'', \hat{\psi}''$ — оценки максимального правдоподобия на всем пространстве параметров. Можно показать, что в рамках гипотезы H_0 величина $-2\ln\lambda$ распределена асимптотически, как χ^2 (1), поскольку одна связь, а именно $\theta = 0$, была наложена на параметрическое пространство.

Мощность такого критерия может быть рассчитана на основе того факта, что в случае справедливости противоположной гипотезы величина $-2\ln\lambda$ в асимптотике распределена по закону нецентрального χ^2 (1) с параметром нецентральности θ^2/I, где

$$I = E\left\{\frac{[f(\mathbf{X}, \varphi) - g(\mathbf{X}, \psi)]^2}{[(1-\theta) f(\mathbf{X}, \varphi) + \theta g(\mathbf{X}, \psi)]^2}\right\}.$$

В частности, для случая противоположной гипотезы $\theta = 1$

$$I = E\left\{\left[\frac{f(\mathbf{X}, \varphi)}{g(\mathbf{X}, \psi)} - 1\right]^2\right\}.$$

Поскольку этот критерий сравнивает $f(\mathbf{X}, \varphi)$ с альтернативной гипотезой, являющейся смесью различных семейств, его мощность не очень велика. Иначе говоря, в результате образования общего семейства происходит потеря точности.

3. Рассмотрим непосредственно отношение правдоподобий

$$l = \frac{f(\mathbf{X}, \hat{\varphi})}{g(\mathbf{X}, \hat{\psi})}. \tag{10.28}$$

Здесь $\hat{\varphi}$, $\hat{\psi}$ — оценки максимального правдоподобия, полученные для каждого семейства в отдельности. Поскольку $g(\mathbf{X}, \hat{\psi})$ соответствует максимуму $h(\mathbf{X}, \theta, \varphi, \psi)$ на ограниченной области $\theta = 1$, то

$$g(\mathbf{X}, \hat{\psi}) \leqslant h(\mathbf{X}, \hat{\theta}'', \hat{\varphi}'', \hat{\psi}''), \tag{10.29}$$

если использовать те же обозначения, что и в уравнении (10.27). Используя уравнения (10.28) и (10.29), получаем

$$-2 \ln l = 2 \ln g(\mathbf{X}, \hat{\psi}) - 2 \ln f(\mathbf{X}, \hat{\varphi}) \leqslant$$

$$\leqslant 2 \ln h(\mathbf{X}, \theta'', \varphi'', \psi'') - 2 \ln f(\mathbf{X}, \hat{\varphi})$$

или

$$-2 \ln l \leqslant -2 \ln \lambda.$$

Тогда, если $-2 \ln l \geqslant k_\alpha$, то и $-2 \ln \lambda > k_\alpha$. Следовательно, при *достаточно больших значениях* $-2 \ln l$ *гипотеза* $f(\mathbf{X}, \varphi)$ *с гарантией может быть отвергнута на основе критерия, построенного для общего семейства гипотез.* Например, пусть при уровне значимости 0,2% $k_\alpha = 9$. Тогда, если $-\ln l \geqslant 4,5$, то можно с уверенностью отвергнуть гипотезу о том, что истинное распределение имеет вид $f(\mathbf{X}, \varphi)$. Такой критерий может быть полезен на практике, особенно для большого числа наблюдений. В примере предыдущего параграфа этот критерий выполняется при больших N и поэтому он состоятелен.

§ 10.6. ПРОВЕРКА ГИПОТЕЗ И ТЕОРИЯ РЕШЕНИЙ

Продолжим обсуждение, начатое в п. 6.3.3, в таком плане: какое решение должно быть принято, если имеется некоторый критерий для разделения гипотез H_0 и H_1. В п. 6.3.3 мы пришли к выводу, что для простых гипотез:

1) критерий Неймана — Пирсона можно рассматривать как бейесовское правило решения;

2) выбор уровня значимости α эквивалентен *априори* предпочтению гипотезы H_0;

3) классическая теория проверки гипотез является частной формулировкой задачи о решениях и целиком ответственна за несимметрию H_0 и H_1;

4) практические достоинства теории решений заключены в ее относительно бо́льшей гибкости в сравнении с классической теорией и в отчетливом представлении предположений, лежащих в основе того или иного метода.

В последующих параграфах мы постараемся продемонстрировать силу теории решений.

10.6.1. Бейесовский подход к выбору семейств распределений. Допустим, что нужно решить, какое из двух семейств распределений $f(\mathbf{X}\,|\,\varphi)$ и $g(\mathbf{X}\,|\,\psi)$ лучше описывает данные $\mathbf{X}$. Сконструируем новое семейство распределений, достаточно общее, чтобы содержать оба семейства f и g:

$$h(\mathbf{X}\,|\,\theta, \varphi, \psi) = \theta f(\mathbf{X}\,|\,\varphi) + (1 - \theta)\, g(\mathbf{X}, \psi),$$

где $\theta = 0$ или 1 (но не промежуточное).

Решение должно ответить на вопрос: соответствуют ли данные некоторому семейству гипотез и, если принадлежат, то какому члену семейства?

Сгруппируем возможные решения:

d_0: выбор невозможен; результат неоднозначен;

d_1, φ^*: искомым семейством является $f(\mathbf{X}\,|\,\varphi)$ с $\varphi = \varphi^*$;

d_2, ψ^*: искомым семейством является $g(\mathbf{X}\,|\,\psi)$ с $\psi = \psi^*$.

Отметим, что каждое из двух последних решений соответствует непрерывному множеству решений.

Для выбора какого-то решения нужен критерий. Таким критерием может быть критерий минимальности цены, т. е. минимальность величины функции стоимости* $L(d, \theta, \varphi, \psi)$, определенной в табл. 10.4.

Таблица 10.4

Функция стоимости

Решения	Что верно в действительности	
	$\theta = \theta_1 = 1, \varphi$	$\theta = \theta_2 = 0, \psi$
d_0	β_1	β_2
d_1, φ^*	$\alpha_1(\varphi^* - \varphi)^2$	γ_1
d_2, ψ^*	γ_2	$\alpha_2(\varphi^* - \psi)^2$

Смысл функции стоимости таков: если в действительности верны значения (θ_1, φ), то принятие решения «результат неоднозначен» имеет стоимость β_1. Выбор правильного семейства, но неверного значения параметра φ^* стоит $\alpha_1(\varphi^* - \varphi)^2$ и т. д.

Нужно также ввести степень веры наблюдателя в параметры θ, φ, ψ в виде априорного распределения. Применительно к данному

* Отметим, что *функция стоимости* была названа *функцией потерь* в гл. 6. Мы хотим избежать двойного использования слова «потеря» в этой главе. Поэтому *потеря* будет иметь значение ошибки первого рода.

примеру можно предположить, что вероятность реализации семейства f и g равна μ и $1-\mu$, тогда как априорные распределения φ и ψ независимы и равны $\pi(\varphi)$ и $\pi(\psi)$. Совместная априорная плотность для θ, φ, ψ может быть записана в виде

$$\pi(\theta, \varphi, \psi) = \mu\pi(\varphi)\pi(\psi) \times \times \delta(\theta) + (1-\mu)\pi(\varphi)\pi(\psi)\delta(1-\theta). \quad (10.30)$$

В соответствии с теоремой Бейеса апостериорная плотность для θ, φ и ψ равна:

$$\pi(\theta, \varphi, \psi \mid \mathbf{X}) = k[\mu f(\mathbf{X} \mid \varphi)\pi(\psi)\pi(\varphi)\delta(\theta) + (1 - -\mu) g(\mathbf{X} \mid \psi)\pi(\varphi)\pi(\psi)\delta(1-\theta)]. \quad (10.31)$$

Нормировочная константа k соответственно равна:

$$\frac{1}{k} = \mu F(\mathbf{X}) + (1-\mu) G(\mathbf{X}),$$

где

$$F(\mathbf{X}) = \int f(\mathbf{X} \mid \varphi)\pi(\varphi)\, d\varphi \quad (10.32)$$

и

$$G(\mathbf{X}) = \int g(\mathbf{X} \mid \psi)\pi(\psi)\, d\psi. \quad (10.33)$$

Сейчас можно вычислить апостериорную цену каждого возможного рашения, если усреднить (θ, φ, ψ) по апостериорному распределению. Например, апостериорная стоимость для решения d_0

$$E[L(\theta, \varphi, \psi, d_0 \mid \mathbf{X})] = \int L(d_0, \theta, \varphi, \psi)\pi(\theta, \varphi, \psi \mid \mathbf{X})\, d\theta d\varphi d\psi = = \frac{\beta_1 \mu F(\mathbf{X}) + \beta_2 (1-\mu) G(\mathbf{X})}{\mu F(\mathbf{X}) + (1-\mu) G(\mathbf{X})}. \quad (10.34)$$

Соответственно апостериорные цены решений (d_1, φ^*) и (d_2, ψ^*) таковы:

$$E[L(\theta, \varphi^*, \psi, d_1 \mid \mathbf{X})] = k[\int \alpha_1 (\varphi^* - \varphi)^2 \mu f(\mathbf{X} \mid \varphi)\pi(\varphi)\, d\varphi + + \gamma_1 (1-\mu) G(\mathbf{X})]; \quad (10.35)$$

$$E[L(\theta, \varphi, \psi^*, d_2 \mid \mathbf{X})] = k[\gamma_2 \mu F(\mathbf{X}) + \int \alpha_2 (\psi^* - \psi)^2 (1-\mu) \times \times g(\mathbf{X} \mid \psi)\pi(\psi)\, d\psi].$$

Для любого $\mathbf{X}$ сейчас можно выбрать решение с минимальной апостериорной ценой. Очевидно, что в качестве φ^* должно быть выбрано такое значение, которое обращает в минимум правую часть первого равенства (10.35). Таким значением является

$$\varphi^* = \frac{\int \varphi f(\mathbf{X} \mid \varphi)\pi(\varphi)\, d\varphi}{\int f(\mathbf{X} \mid \varphi)\pi(\varphi)\, d\varphi} = E(\varphi \mid \mathbf{X}), \quad (10.36)$$

т. е. среднее для φ по апостериорному распределению. Точно так же

$$\psi^* = E(\psi \mid \mathbf{X}). \quad (10.37)$$

(Этот результат справедлив всегда, когда цена параболична, т. е. пропорциональна квадрату отклонения.) При таких значениях φ^*, ψ^* апостериорные цены равны:

$$E[L(\theta, \varphi^*, \psi, d_1 \mid \mathbf{X})] = \frac{\alpha_1 D(\varphi \mid \mathbf{X}) \mu F(\mathbf{X}) + \gamma_1 (1-\mu) G(\mathbf{X})}{\mu F(\mathbf{X}) + (1-\mu) G(\mathbf{X})} \quad (10.38)$$

и

$$E[L(\theta, \varphi, \psi^*, d_2 \mid \mathbf{X})] = \frac{\gamma_2 \mu F(\mathbf{X}) + \alpha_2 D(\psi \mid \mathbf{X}) (1-\mu) G(\mathbf{X})}{\mu F(\mathbf{X}) + (1-\mu) G(\mathbf{X})}, \quad (10.39)$$

где $D(\varphi \mid \mathbf{X})$ и $D(\psi \mid \mathbf{X})$ — дисперсии апостериорных распределений φ и ψ соответственно.

Теперь понятно, какое правило решения нужно выбрать. Очевидно, искомым правилом решения будет то, которому соответствует минимальное из величин (10.34), (10.38) и (10.39). Если выбирается решение с неопределенным результатом, то значение параметра определяется соотношением (10.36) или (10.37). Это правило решения проиллюстрировано ниже.

Пример. Допустим, что два семейства f и g имеют вид:

$$f(X \mid \varphi) = \varphi \exp(-\varphi X);$$

$$g(X \mid \psi) = \psi^2 X \exp(-\psi X),$$

а априорные распределения для φ и ψ таковы:

$$\pi(\varphi) = a^2 \varphi \exp(-a\varphi); \quad \pi(\psi) = b^2 \psi \exp(-b\psi). \quad (10.40)$$

Такой выбор априорных распределений продиктован следующими соображениями: известно, что φ и ψ не равны нулю и по порядку величины равны

$$E(\varphi) = 2/a; \; E(\psi) = 2/b.$$

Легко показать, что величина (10.32) при таких условиях равна

$$F(\mathbf{X}) = a^2 \Gamma(N+2) (a + N\overline{X})^{-N-2},$$

где $\overline{X}$ — среднее наблюдений X_i. Использовав формулу (10.36), получим

$$\varphi^* = (N+2)/(a + N\overline{X}).$$

Поэтому дисперсия апостериорного распределения φ равна:

$$D(\varphi \mid \mathbf{X}) = E[(\varphi - \varphi^*)^2] = (N+2) (a + N\overline{X})^{-2}.$$

Точно так же можно вычислить величины (10.33) и (10.37):

$$G(\mathbf{X}) = b^2 (b + N\overline{X})^{-2N-2} \prod_{i=1}^{N} X_i \, \Gamma(2N+2);$$

$$\psi^* = (2N+2)/(b + N\overline{X}),$$

что, в свою очередь, позволяет рассчитать дисперсию апостериорного распределения для ψ:

$$D(\psi \mid \mathbf{X}) = E[(\psi - \psi^*)^2] = (2N+2) (b + N\overline{X})^{-2},$$

наконец, апостериорные цены (10.34), (10.38) и (10.39) равны:

$$E[L(d_0 \mid \mathbf{X})] = \frac{1}{D}\left[\frac{\beta_1 \mu a^2 \Gamma(N+2)}{(a+N\bar{X})^{N+2}} + \frac{\beta_2 (1-\mu) b^2 \Gamma(2N+2)}{(b+N\bar{X})^{2N+2}} \prod_{i=1}^{N} X_i\right]; \tag{10.41}$$

$$E[L(d_1 \mid \mathbf{X})] = \frac{1}{D}\left[\frac{\alpha_1 \mu a^2 \Gamma(N+3)}{(a+N\bar{X})^{N+4}} + \frac{\gamma_1 (1-\mu) b^2 \Gamma(2N+2)}{(b+N\bar{X})^{2N+2}} \prod_{i=1}^{N} X_i\right]; \tag{10.42}$$

$$E[L(d_2 \mid \mathbf{X})] = \frac{1}{D}\left[\frac{\gamma_1 \mu a^2 \Gamma(N+2)}{(a+N\bar{X})^{N+2}} + \frac{\gamma_1 (1-\mu) b^2 \Gamma(2N+3)}{(b+N\bar{X})^{2N+3}} \prod_{i=1}^{N} X_i\right] \tag{10.43}$$

с общим знаменателем

$$D = \frac{\mu a^2 \Gamma(N+2)}{(a+N\bar{X})^{2N+2}} + \frac{(1-\mu) b^2 \Gamma(2N+2)}{(b+N\bar{X})^{2N+2}}.$$

Для того чтобы получить некоторое представление о задаче, предположим, что $a \approx 2b$. При больших N для $\Gamma(N)$ можно использовать формулу Стирлинга, а различные степени $(a + N\bar{X})$ и $(b + N\bar{X})$ можно приближенно выразить через соответствующие экспоненты. Тогда выражения (10.41), (10.43) сводятся к

$$E[L(d_0 \mid \lambda)] \approx \frac{1}{D}[\beta_1 \mu + \beta_2 (1-\mu) \exp(-\ln \lambda)]; \tag{10.44}$$

$$E[L(d_1 \mid \lambda)] \approx \frac{1}{D}\left[\frac{\alpha_1 \mu}{N\bar{X}^2} + \gamma_1 (1-\mu) \exp(-\ln \lambda)\right]; \tag{10.45}$$

$$E[L(d_2 \mid \lambda)] \approx \frac{1}{D}\left[\gamma_2 \mu + \alpha_2 (1-\mu) \frac{2}{N\bar{X}^2} \exp(-\ln \lambda)\right], \tag{10.46}$$

где

$$D \approx \mu + (1-\mu) \exp(-\ln \lambda)$$

и

$$\ln \lambda = N\left(\ln \bar{X} - N^{-1} \sum_{i=1}^{N} \ln X_i + 1 - 2 \ln 2\right).$$

Таким образом, получается, что в случае $\alpha_1 = \alpha_2 = 0$ (при оценке не происходит никаких потерь) такой бейесовский критерий эквивалентен критерию отношений правдоподобия, поскольку λ есть отношение правдоподобий:

$$\lambda = f(\mathbf{X}, \hat{\varphi})/g(\mathbf{X}, \hat{\psi}).$$

На рис. 10.12 показана зависимость выражений в квадратных скобках формул (10.44), (10.45) и (10.46) для такого случая. Неудивительно, что когда цена выбора решения d_0 довольно велика, следует выбирать решение d_1 или d_2.

Если α_1 и α_2 не равны нулю, то это приведет к росту соответствующих цен, что связано с флуктуациями оценок.

Отметим, что при N, стремящемся к бесконечности, φ^* и ψ^* приближаются к соответствующим оценкам максимального правдоподобия $\hat{\varphi}$ и $\hat{\psi}$. Это связано с уменьшением влияния априорного знания на выбор оценки.

10.6.2. Последовательные критерии для оптимального числа наблюдений. При конструировании функции стоимости для некоторой группы правил решения естественно учитывать также и цену, связанную с получением данных. Тогда оптимальное решение (в бейесовском смысле) становится функцией числа наблюдений. Классические правила решения (с фиксированным числом наблюдений) не обязательно оптимальны, поскольку они не допускают возможности совершенствования процесса эксперимента в зависимости от предыдущих наблюдений. Критерии, которые можно модифицировать, называются *последовательными*.

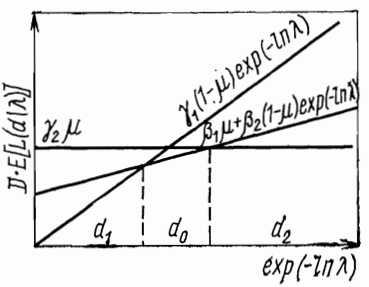


Рис. 10.12. Зависимость апостериорной цены от отношения правдоподобий для случая, описанного в тексте

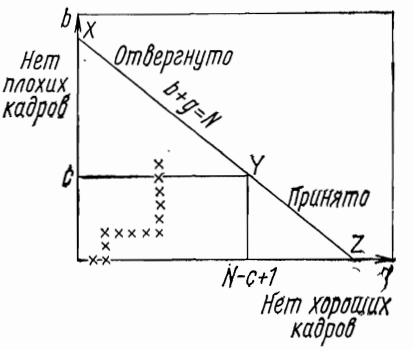


Рис. 10.13. План просмотра в случае обычного последовательного критерия для примера с выбраковкой рулона пленки

Пример. Допустим, что нужно решить вопрос, измерять или не измерять рулон пленки, исходя из того, как много хороших или плохих кадров она содержит.

Обычно поступают так: решают, при какой доле хороших кадров рулон следует считать приемлемым. Затем на некотором рулоне отбирается N кадров для измерения. Исходя из количества хороших измерений рулон либо принимается, либо отвергается.

Примерный план действия таков.

Измерить N кадров. Отвергнуть рулон, если число плохих кадров равно C или больше, в противном случае считать рулон приемлемым. Соответствующие правила решения отражены на рис. 10.13.

Выбор правила решения соответствует разделению линии XZ [$b + g = N$] на две части, XY и YZ. Если наблюдение (b, g) лежит на XY, то нужно отвергнуть рулон; если (b, g) на YZ, то его следует принять.

Очевидно, следующий план действий является альтернативным предыдущему. Один за другим измеряются кадры на рисунке. Пусть число хороших событий равно g, число плохих b. Если b становится равным c одновременно или позже, чем g равным $N - c + 1$, то рулон принимается; в противном случае рулон считается непригодным.

На диаграмме это соответствовало бы пересечению линий $b = c$ и $g = N - c + 1$. Очевидно, что при таком плане действий количество просмотренных кадров меньше, чем при первом. В данном случае окончательные решения будут идентичны и, очевидно, на практике нужно использовать второе правило. Таким образом, ясно, что оптимальное решение следует искать в классе последовательных критериев.

Сконструируем последовательный критерий для разделения *двух простых гипотез*: $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1$. Полная функция стоимости будет складываться из

l_0 : цены, обусловленной ошибкой первого рода* (неверно отвергнута H_0);
l_1 : цены, обусловленной ошибкой второго рода (неверно отвергнута H_1);
c : стоимость измерения событий.

Полная функция стоимости для N событий приведена в табл. 10.5.

Таблица 10.5

Полная функция стоимости

Решение	Что верно в действительности	
	H_0	H_1
H_0 верна	Nc	l_1+Nc
H_1 верна	l_0+Nc	Nc

Пусть вероятности ошибок первого и второго рода равны α и β соответственно и пусть ожидаемое число событий, необходимое для принятия решения, равно $E_i(\delta, N)$, если верна H_1. Тогда, если априорная вероятность H_0 равна μ, *апостериорный риск* для правила решения δ равен:

$$r_\mu(\delta) = \mu[\alpha(\delta, N)\, l_0 + cE_0(\delta, N)] + (1-\mu)\times \\ \times[\beta(\delta, N)\, l_1 + cE_1(\delta, N)]. \qquad (10.47)$$

Наконец, оптимальное правило решения соответствует минимуму (10.47) по переменным α, β или N. Среди непоследовательных критериев для фиксированного N оптимальным (в классическом смысле) в соответствии с леммой Неймана — Пирсона является критерий, основанный на отношении правдоподобий:

$$\lambda_N = \frac{L(\mathbf{X} \mid \theta_1)}{L(\mathbf{X} \mid \theta_0)} = \prod_{i=1}^{N} \frac{f(\mathbf{X}_i \mid \theta_1)}{f(\mathbf{X}_i \mid \theta_0)}. \qquad (10.48)$$

Тогда если $\lambda_N < c_\alpha$, нужно выбрать H_0, в противном случае H_1, c_α определено в п. 10.3.1. Вальд [52, 53] доказал следующее обобщение: можно выбрать такие две константы A и B, что эксперимент следует продолжать, пока выполняется неравенство

$$B < \lambda_N < A, \qquad (10.49)$$

и прекратить, когда неравенства нарушаются.

Далее,

если $\lambda_N < B$, нужно выбрать $H_0 : \theta = \theta_0$,

если же $\lambda_N > A$, нужно выбрать $H_1 : \theta = \theta_1$.

Были доказаны такие теоремы [48].

* См. сноску в п. 10.6.1.

1. Подход с позиций теории решений.

Последовательный критерий отношения вероятностей дает минимум апостериорного риска (10.47) при значениях:

$$A = \frac{\mu}{1-\mu} \frac{1-\mu_1}{\mu_1};$$

$$B = \frac{\mu}{1-\mu} \frac{1-\mu_0}{\mu_0},$$

где μ_i — априорные вероятности, соответствующие случаю, когда оптимальным решением было бы отвергнуть гипотезу H_i без проведения каких-либо наблюдений. (Они всегда существуют и $\mu_0 \leqslant \mu_1$.)

2. Подход с позиций классической теории.

Среди всех критериев (последовательных или нет), для которых

$$P \text{ (отбросить } H_0 \mid H_0) \leqslant \alpha \text{ (потеря)};$$

$$P \text{ (принять } H_0 \mid H_1) \leqslant \beta \text{ (примесь)}$$

и для которых $E_0(N)$ и $E_1(N)$ конечны, последовательный критерий отношения вероятностей минимизирует одновременно $E_0(N)$ и $E_1(N)$.

Кроме того, приближенно справедливы равенства:

$$A \simeq (1-\beta)/\alpha; \quad B \simeq \beta/(1-\alpha). \tag{10.50}$$

Можно сравнить оба подхода [48] подобно тому, как это делалось в п. 6.3.3.

Если гипотезы не простые, то соответствующий последовательный критерий не обладает такими оптимальными свойствами, как критерий Неймана — Пирсона. Это особенно неприятно, поскольку наибольший практический интерес представляют непрерывные параметрические семейства гипотез. Тогда остается только сформулировать задачу так, чтобы можно было использовать метод «простой гипотезы», хотя он уже не обязательно оптимален. Например, часто встречается проблема выбора между гипотезой

$$H_0 : \theta \geqslant \theta_0$$

и

$$H_1 : \theta < \theta_0.$$

Эту задачу можно переформулировать, как выбор между

$$H_0 : \theta > \theta_0; \quad H_1 : \theta < \theta_1,$$

где $\theta_1 < \theta_0$.

Пример. Вернемся к примеру с пленкой и используем подобный прием. Для этого случая θ может быть долей плохих кадров. Рулон пленки должен быть либо принят, либо отвергнут с уровнем зависимости α и мощностью $1-\beta$. Уравнение (10.48) приобретает вид:

$$\lambda_N = \frac{\binom{N}{b}\theta_1^b(1-\theta_1)^{N-b}}{\binom{N}{b}\theta_0^b(1-\theta_0)^{N-b}}.$$

Последовательный критерий отношения вероятностей задается уравнением 10.49):

$$B < \left(\frac{\theta_1}{\theta_0}\right)^b \left(\frac{1-\theta_1}{1-\theta_0}\right)^{N-b} < A.$$

Тогда

$$\ln B < b \ln\frac{\theta_1}{\theta_0} + (N-b)\ln\frac{1-\theta_1}{1-\theta_0} < \ln A,$$

где B и A определены уравнениями (10.50). На рис. 10.14 показаны соответствующие области этого критерия.

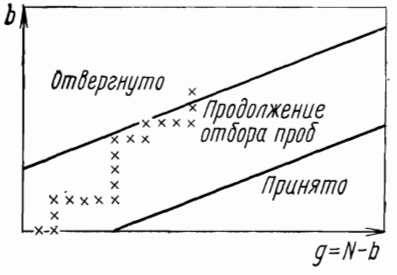


Рис. 10.14. План просмотра пленки для задачи, переформулированной на языке критерия Неймана — Пирсона для простых гипотез

10.6.3. Последовательный критерий отношения вероятностей для непрерывного семейства гипотез. Предположим, что для двух значений θ_0 и θ_1 параметров разработан последовательный критерий отношения правдоподобий. Тогда можно показать [48], что функция мощности (также называемая операционной характеристикой) критерия для дискриминации $f(\mathbf{X}\mid\theta_0)$ против $f(\mathbf{X}\mid\theta)$ равна

$$B(\theta) \simeq \frac{A^{h(\theta)}-1}{A^{h(\theta)}-B^{h(\theta)}},$$

где $h(\theta)$ — ненулевое решение уравнения

$$\int\left[\frac{f(X\mid\theta_1)}{f(X\mid\theta_0)}\right]^{h(\theta)} f(X\mid\theta)\,dX = 1.$$

Средний объем выборки как функция θ равен

$$E(N \mid \theta) \simeq \frac{\beta(\theta) \ln B - [1 - \beta(\theta)] \ln A}{E\left\{ \ln \left[\frac{f(X \mid \theta_1)}{f(X \mid \theta_0)} \right] \middle| \theta \right\}}$$

при условии, что знаменатель не равен нулю.

Последовательный критерий отношения вероятностей оптимален для двух фиксированных точек θ_0 и θ_1, но часто размер выборки пре-

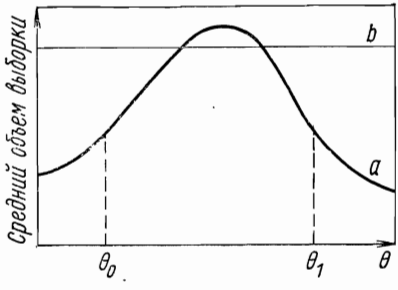


Рис. 10.15. Пример последовательного критерия отношения правдоподобий *a* и критерия Неймана—Пирсона *b*, обсуждаемого в тексте

вышает размер эквивалентной единичной выборки (рис. 10.15). Это означает, что последовательный критерий отношения вероятностей не является наилучшим последовательным критерием в случае сложных гипотез.

§ 10.7. СВОДКА ОПТИМАЛЬНЫХ КРИТЕРИЕВ

В табл. 10.6 даны рекомендации: какой критерий выбрать в зависимости от того, простые или сложные гипотезы.

В табл. 10.6 не отражен случай, когда нужно сравнить данную гипотезу со всеми другими возможными гипотезами. Такие непараметрические критерии рассмотрены в гл. 11.

Таблица 10.6

Сводка оптимальных критериев

Тип гипотезы	Оптимальные критерии
Простая гипотеза против простой гипотезы	Существует один оптимальный критерий: критерий Неймана—Пирсона (см. п. 10.3.1). Ожидаемое число наблюдений может быть уменьшено, если использовать последовательный критерий отношения правдоподобий (см. п. 10. 6.2)

Продолжение табл. 10.6

Тип гипотезы	Оптимальные критерии
Сложные гипотезы	В общем случае равномерно наиболее мощный критерий не существует. Одно исключение: односторонний критерий для экспоненциального семейства (см. п. 10. 4. 2). Критерий должен обладать тем не менее разумными свойствами: несмещенностью (см. п. 10. 2. 3), быть наиболее мощным в интересующей области (см. п. 10. 2. 5), иметь максимальную локальную мощность (см. п. 10. 4. 3). Непрерывное семейство гипотез (уравнения связи на параметры, проверка значимости дополнительных параметров). Общий метод: критерий отношения правдоподобий (см. § 10. 5). В асимптотике распределения просты (χ^2) (см. п. 10. 5. 3). Для выборок конечного объема распределения относятся к классическим только в случае нормального распределения (см. п. 10. 5. 5, табл. 10. 2). Различные семейства гипотез (никакой связи между гипотезами). Общий метод (см. п. 10. 5. 7 (1)). Общая модель и использование критерия отношения правдоподобий (см. п. 10. 5. 7 (2)). Важность отношения правдоподобия в общем случае (см. п. 10. 5. 7 (3)).

ГЛАВА 11

КРИТЕРИИ СОГЛАСИЯ

Рассмотрим вопрос проверки нулевой гипотезы H_0 с помощью проверочной статистики T, обладающей некоторой критической областью ω с уровнем значимости α. Однако в отличие от предыдущего альтернативная гипотеза H_1 здесь означает множество всех возможных альтернатив для H_0. Поэтому H_1 не может быть сформулирована, а ошибка второго рода β неизвестна.

Критерии согласия сравнивают экспериментальные данные с ф. п. в., в соответствии с которым согласно гипотезе H_0 распределены эти данные. Результатом такого сравнения должно быть заключение: если бы H_0 была верна и если бы эксперимент был повторен много раз, то как часто (редко) наблюдались бы результаты, получаемые в эксперименте. Сравнение между H_0 и H_1 в том плане, как это делалось в предыдущей главе, здесь уже теряет смысл, поскольку H_1 носит настолько общий характер, что она опишет данные с любой точностью. Поэтому заранее известно, что с чисто статистической точки зрения гипотеза H_1 лучше, чем H_0.

Из практических соображений рассмотрим только критерии, не зависящие от вида распределения: чтобы применить такой критерий, нужно знать только его ф. п. в., а для многих хорошо известных, не зависящих от распределений критериев ф. п. в. приведены в таблицах. Конечно, для какой-то специальной задачи может существовать свой более совершенный критерий, но часто физик не желает тратить время на его разработку и предпочитает использовать некий общий критерий.

Мы уже указывали, что при группировке событий в гистограмму* происходит некоторая потеря информации в ячейке. Следовательно, нужно ожидать, что критерии для данных, сгруппированных в гистограмму, хуже, чем когда каждое событие рассматривается индивидуально. Однако требование независимости критерия от вида распределений ограничивает выбор по существу только одним классом критериев для данных, не сгруппированных в гистограмму (см. § 11.4). Кроме того, этот класс критериев применим к данным, зависящим только от одной случайной переменной, а гипотезы H_0

* Ячейки гистограммы часто называют *классами данных* в литературе по статистике.

не должны содержать параметры θ, оценки которых нужно находить на основе экспериментальных данных.

Когда данные сгруппированы в гистограмму, важно, чтобы число событий в ячейке было достаточно большим. Это требование необходимо для того, чтобы сработал тот или иной метод, поскольку большинство важных свойств критериев (например, независимость от распределения) справедливо только в асимптотическом пределе. Такое требование может серьезно ограничить область применения критериев значимости для данных, зависящих от нескольких случайных переменных.

Конечно, при данном разбиении существует много гипотез, которые предсказывают полученное на опыте распределение событий. Поскольку для проверки нулевой гипотезы используется только это распределение, критерий, собственно говоря, проверяет не только H_0, но и целый класс гипотез, неотличимых друг от друга при данном разбиении.

§ 11.1. КРИТЕРИЙ ОТНОШЕНИЯ ПРАВДОПОДОБИЙ

Предположим, что N наблюдений сгруппированы в гистограмму, содержащую k ячеек, и в ячейку с номером i попадает n_i событий (что не означает, что наблюдаемая случайная переменная одномерна). Пусть p_i означает вероятность попадания события в ячейку с номером i, предсказываемую гипотезой H_0, а q_i — вероятность попадания в ту же ячейку, соответствующая истинной неизвестной гипотезе. Правдоподобия данных в рамках этих двух гипотез соответственно равны:

$$L_0(\mathbf{n}\,|\,\mathbf{p}) = N! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!};$$

$$L(\mathbf{n}\,|\,\mathbf{q}) = N! \prod_{i=1}^{k} \frac{q_i^{n_i}}{n_i!}.$$

Оценка $\hat{q}_i$ величины q_i может быть найдена из условия максимума $\ln L\ (\mathbf{n}\,|\,\mathbf{q})$ относительно величин q, удовлетворяющих условию $\sum_{i=1}^{k} q_i = 1$;

$$\frac{\partial \ln L(\mathbf{n}\,|\,\mathbf{q})}{\partial q_i} = \frac{n_i}{q_i} - \frac{n_k}{q_k} = 0. \tag{11.1}$$

Суммируя правую часть по i, имеем:

$$\sum_{i=1}^{k-1} q_k n_i = \sum_{i=1}^{k-1} n_k q_i.$$

Используя очевидные соотношения $\sum_{i=1}^{k} n_i = N$ и $\sum_{i=1}^{k} q_i = 1$, получаем

$$q_k = n_k/N.$$

Наконец, подставляя это выражение в (11.1), приходим к такому результату:

$$\hat{q}_i = n_i/N. \tag{11.2}$$

Проверочная статистика имеет вид отношения правдоподобий:

$$\lambda = \frac{L_0(\mathbf{n} \mid \mathbf{p})}{L(\mathbf{n} \mid \mathbf{q})} = N^N \prod_{i=1}^{k} \left(\frac{p_i}{n_i}\right)^{n_i}.$$

Как мы уже видели в § 10.5, ф. п. в. этой проверочной статистики для конечного N все еще зависит от вида распределения, но $(-2 \ln \lambda)$ асимптотически ведет себя, как χ^2 $(k-1)$. Число степеней свободы $(k-1)$ обусловлено тем, что k членов q_i удовлетворяют одному уравнению связи

$$\sum_{i=1}^{k} q_i = 1.$$

§ 11.2. χ^2-КРИТЕРИЙ ПИРСОНА

Как известно, ф. п. в. мультиномиального распределения в асимптотическом пределе сходится к ф. п. в. нормального распределения. Имея в виду это обстоятельство, Пирсон предложил использовать вместо отношения правдоподобий статистику вида (8.17):

$$(\mathbf{n} - N\mathbf{p})^T \underset{\sim}{D}^{-1} (\mathbf{n} - N\mathbf{p}),$$

где $\underset{\sim}{D}$ — матрица вторых моментов наблюдений $\mathbf{n}$. Поскольку из-за соотношения $\sum_{i}^{k} n_i = N$ эта матрица имеет ранг $(k-1)$, мы будем использовать $(k-1)$ членов.

Для определенности пусть ими будут первые $(k-1)$ члены (позже мы восстановим симметрию между всеми k членами). Пусть $\underset{\sim}{W}$ — матрица порядка $(k-1) \times (k-1)$, полученная из $\underset{\sim}{D}$ вычеркиванием k-х столбца и строки. Можно легко показать, что

$$N(\underset{\sim}{W}^{-1})_{ij} = \frac{1}{p_i}\delta_{ij} + \frac{1}{p_k}.$$

Теперь рассмотрим статистику

$$T = (\mathbf{n} - N\mathbf{p})^T \underset{\sim}{W}^{-1} (\mathbf{n} - N\mathbf{p}), \tag{11.3}$$

в которой фигурируют только первые $(k-1)$ компоненты $\mathbf{n}$ и $\mathbf{p}$. Случайная переменная $(\mathbf{n} - N\mathbf{p})$ асимптотически распределена по многомерному нормальному закону с матрицей вторых моментов $\underset{\sim}{W}$. Таким образом, в том же пределе T распределена по χ^2 $(k-1)$-закону.

До сих пор $(k-1)$ членов в (11.3) выбирались произвольно. Найдем выражение, эквивалентное T, в котором все ячейки представлены одинаково:

$$T=\frac{1}{N}\left\{\sum_{i=1}^{k-1}\frac{(n_i-Np_i)^2}{p_i}+\frac{1}{p_k}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}(n_i-Np_i)(n_j-Np_j)\right\}=$$

$$=\frac{1}{N}\left\{\sum_{i=1}^{k-1}\frac{(n_i-Np_i)^2}{p_i}+\frac{1}{p_k}\left[\sum_{i=1}^{k-1}(n_i-Np_i)\right]^2\right\}$$

или окончательно

$$T=\frac{1}{N}\sum_{i=1}^{k}\frac{(n_i-Np_i)^2}{p_i}=\frac{1}{N}\sum_{i=1}^{k}\frac{n_i^2}{p_i}-N. \qquad (11.4)$$

Это обычный критерий согласия χ^2. При практических расчетах обычно считают, что асимптотические свойства статистики (11.4) применимы тогда, когда ожидаемое число события на одну ячейку (Np_i) больше, чем 5. Кочран [53, 54] утверждает, что такое приближение допустимо и тогда, когда не более чем в 20% ячеек ожидаемые числа событий заключены между 1 и 5.

Проведенное выше рассуждение может показаться не очень убедительным, однако окончательный результат (11.4) может быть доказан строго [12].

Какой из двух приведенных критериев следует считать предпочтительным? Мы только что видели, что они совпадают в асимптотическом пределе. Для выборок конечных размеров они отличаются друг от друга и невозможно что-нибудь сказать об этом в общем случае. В следующем параграфе мы вкратце рассмотрим моменты статистики (11.4).

11.2.1. Моменты статистики Пирсона. Пусть H_T — истинная гипотеза, а p_i — вероятность попадания события в i-ю ячейку гистограммы, предсказываемая этой гипотезой. Тогда

$$E(T\mid H_T)=\sum_i\frac{q_i(1-q_i)}{p_i}+N\left(\sum_i\frac{q_i^2}{p_i}-1\right), \qquad (11.5)$$

и когда $q_i=p_i$

$$E(T\mid H_0)=k-1.$$

Можно показать, что $E(T)$ минимально при $q_i=p_i$, минимизируя $E(T\mid H_T)$ по q_i при условии, что $\sum_{i=1}^{k}q_i=1$. Следовательно,

$$E(T\mid H_0)=k-1;$$
$$E(T\mid H_T)\geqslant k-1,$$

каково бы ни было N (конечные статистики).

Если гистограмма сконструирована так, что все ячейки имеют одинаковое вероятностное содержание $p_i = 1/k$ для гипотезы H_0, то легко можно вычислить следующие моменты [2]:

$$D(T \mid H_0) = 2(k-1);$$
$$D(T \mid H_1) = 4(N-1)k^2\left\{\sum_i q_i^3 - \left(\sum_i q_i^2\right)^2\right\} + 2k^2\left\{\sum_i q_i^2 - \left(\sum_i q_i^2\right)^2\right\};$$
$$E(T \mid H_1) = (k-1) + (N-1)\left\{k\sum_i q_i^2 - 1\right\}.$$

11.2.2. χ^2-Критерий при оценке параметров. Если распределение генеральной совокупности зависит от параметра θ, который нужно оценить по данным, то статистика T не распределена по закону χ^2 $(k-1)$. Рассмотрим два метода оценки θ: оценки максимального правдоподобия, полученные по гистограмме и по данным, когда каждое событие рассматривается индивидуально.

В первом случае (названном в п. 8.4.5 оценкой «мультиномиального» распределения максимального правдоподобия) нужно отыскивать максимум

$$L(\mathbf{n} \mid \theta) = N! \prod_{i=1}^{k} \frac{[p_i(\theta)]^{n_i}}{n_i!}$$

или логарифм относительно параметров θ_j, $j = 1, \ldots, r$, где θ имеет r компонент. Тогда можно показать, что статистика T, использующая такие оценки, ведет себя асимптотически, как χ^2 $(k-r-1)$-распределение. Это как бы означает, что при оценке θ происходит потеря r степеней свободы.

Во втором случае при оценке θ используется вся доступная информация (следовательно, этот метод лучше), а при построении T используется распределение событий в гистограмме. Можно показать [2], что функция распределения T является неким промежуточным распределением между χ^2 $(k-1)$ (которое имеет место, когда θ фиксировано) и χ^2 $(k-r-1)$ (которое имеет место при оценке параметров первым методом). Критерий в этом случае уже зависит от распределения, но если k велико, а r мало, эти два распределения χ^2 $(k-1)$ и χ^2 $(k-r-1)$ достаточно близки друг к другу, чтобы сделать критерий практически не зависящим от распределения (см. рис. 9.2 функций распределений χ^2).

11.2.3. Выбор оптимального размера ячейки. На практике приходится решать, каким должен быть размер ячейки гистограммы. При слишком малом числе ячеек теряется значительная часть информации, а при большом их числе в ячейки гистограммы будет попадать слишком малое число событий. Большинство результатов, которые мы приводим, например, нормальность мультиномиального распределения, справедливы только в асимптотическом пределе. Одно из основных правил при выборе размера ячеек состоит в том, что при числе ячеек k они должны соответствовать одинаковому вероятностному содержанию в рамках гипотезы H_0.

Критерий T состоятелен (и, следовательно, асимптотически не смещен) независимо от способа разбиения. Однако для конечных статистик этот критерий, вообще говоря, не является несмещенным. Одной из причин, заставляющих проводить разбиение на равновероятностные ячейки, состоит в том, что этот критерий локально не смещен. Мы не приводим здесь подробно доказательства [2], а изложим вкратце его идею.

Обозначим $\varepsilon_i = p_i - (1/k)$ и рассмотрим мощность критерия как функцию $\boldsymbol{\varepsilon}$. Случай равновероятностных ячеек соответствует $\boldsymbol{\varepsilon} = 0$, а локальная несмещенность означает, что $P(\boldsymbol{\varepsilon}) \geqslant P(\mathbf{0})$ для малого $\boldsymbol{\varepsilon}$. Можно разложить $P(\boldsymbol{\varepsilon})$ в ряд Тейлора:

$$P(\boldsymbol{\varepsilon}) = P(\mathbf{0}) + A \sum_i \varepsilon_i^2 + \ldots$$

Далее можно показать, что в этом разложении $A > 0$.

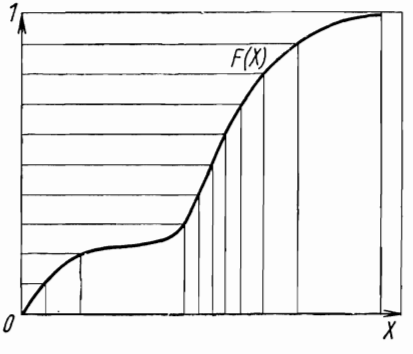


Рис. 11.1. Способ нахождения равновероятностных ячеек

При описании свойств распределения T в предыдущих параграфах предполагалось, что ячейки не зависят от наблюдений. Однако, если положение ячеек оценивается по данным, то это не влияет на асимптотические свойства распределения T. В частности:

1) можно использовать оценки параметров при построении гистограммы с равновероятностным содержанием в ячейках;

2) можно начинать с большого числа ячеек равных размеров, затем группировать соседние ячейки, пока не будет получено требуемое вероятностное содержание в каждой ячейке;

3) очевидно, нельзя выбрать такое, которое дает наименьшее T. Такая проверочная статистика не будет распределена по закону χ^2.

При построении схемы разбиения, основанной на некотором ожидаемом распределении в рамках гипотезы H_0, можно использовать таблицы, если распределение достаточно классическое. В большинстве практических случаев приходится находить функцию распределения $F(X)$ интегрированием, затем делить интервал изменения $F(X)$, $[0, 1]$ на k равных частей, чтобы найти соответствующие ячейки по переменной X (рис. 11.1). Если переменная X одномерна, то подобная процедура довольно проста.

Когда же $\mathbf{X} = X_1, \ldots, X_n$ есть n-мерная величина, то можно было бы вычислить функцию распределения для одной размерности и подобным же образом выбрать ячейки, но при этом теряется вся информация о других компонентах. Чтобы избежать этого, можно для каждой ячейки по X_i вычислить $F(X_2 | X_1)$, затем использовать такой же метод и, наконец, получить K^2 ячеек; при этом, конечно,

теряется информация об $X_3, \ldots, X_n$. Этот метод может быть повторен n раз, тогда

$$\left.\begin{aligned} Z_1 &= F_1(X_1); \\ Z_2 &= F_2(X_2 \mid X_1); \\ Z_i &= F_i(X_i \mid X_1, \ldots, X_{i-1}), \quad i = 1, n. \end{aligned}\right\} \tag{11.6}$$

Таким образом, получается преобразование $\mathbf{Z} = R(\mathbf{X})$ и по построению Z распределена равномерно внутри n-мерного гиперкуба длиной $[0, 1]$ по каждому измерению. Теперь уже можно разделить этот гиперкуб на равные ячейки

$$c_{j_1 j_2 \ldots j_N} = \left\{\mathbf{Z};\ \frac{j_i - 1}{m_i} \leqslant Z_i \leqslant \frac{j_i}{m_i},\ i = 1, \ldots, n\right\}, \tag{11.7}$$

где $j_i = 1, \ldots, m_i$, если выбрано разбиение на m_i частей по каждому измерению. (Если быть строгим, то нужно использовать более точные обозначения, поскольку при таких обозначениях границы соседних ячеек перекрываются.) Тогда статистика T приобретает вид:

$$T = \frac{k}{N} \sum_{i=1}^{k} n_i^2 - N \quad \text{и} \quad k = \prod_{i=1}^{n} m_i.$$

При таком способе полное число ячеек не может быть выбрано равным некоторому заранее выбранному числу (например, нельзя использовать первый пример). Однако обычно не очень важно точно выбирать k. Мы увидим, как можно оптимизировать это число.

Оставим открытой возможность различных разбиений, т. е. m_i не обязательно равны. Действительно, могут существовать физические причины для этого. Например, одна компонента может быть не очень чувствительна к проверяемой гипотезе, тогда как другая, напротив, очень чувствительна и есть основания оставить неизмененным полное число ячеек. Или когда в распределение, предсказываемое гипотезой H_0, не включена функция разрешения, то не имеет смысла использовать ячейки меньшего размера, чем разрешение. Таким образом, схема разбиения (11.7), будучи разумно симметричной по всем компонентам, может быть не лучшей для некоторой частной задачи. Например, можно делить n-мерное пространство $\mathbf{X}$ на гиперобъемы равновероятностного содержания, которые могут иметь определенный физический смысл для изучаемой задачи. Очень часто это сводится к замене переменных, когда некоторые из координат не несут информации и по ним может быть проведено интегрирование.

Рассмотрим вопрос оптимизации числа ячеек k. Для этого найдем предельную мощность при N, стремящемся к бесконечности. Поскольку критерий состоятелен, то его мощность по сравнению с лю-

бой другой фиксированной гипотезой асимптотически становится равной единице.

Поэтому, чтобы найти предельную функцию мощности, рассмотрим последовательность альтернативных гипотез H_T, где $q_i = p_i [1 + (\Delta_i/\sqrt{N})]$, которая сходится к H_0 с ростом N (Δ_i — фиксированы). Тогда в рамках гипотезы H_T статистика T асимптотически распределена, как нецентральный χ^2 с параметром нецентральности $K = \sum_i p_i \Delta_i^2$ и $(k-r-1)$-степенями свободы (оценка r параметров находится методом максимального правдоподобия по данным, сгруппированным в гистограмму).

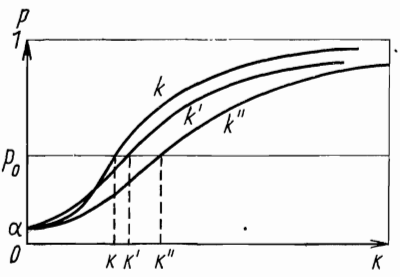


Рис. 11.2. Пример функций мощности для различного числа k ячеек в гистограмме

Найдем в классе критериев с равновероятностными ячейками оптимальный критерий. Поскольку в этом классе нет р. н. м.-критерия, нужно привлечь дополнительные соображения. Будем считать, например, p_0 допустимой мощностью (т. е. примесь $1 - p_0$). Точки пересечения прямой $p = p_0$ с функцией мощности $p(k, K)$ определяют связь $K_0(k)$ (рис. 11.2). Из всех критериев выберем тот, который соответствует наименьшему значению k. Величина K служит мерой близости альтернативной гипотезы к H_0.

Можно написать выражение для мощности критерия $p(k, K)$, если для нецентрального χ^2-распределения гипотезы H_T воспользоваться приближением нормального закона

$$p = \Phi\left[\sqrt{2}(N-1)\left(\frac{b}{k}\right)^{5/2} - \lambda_\alpha\right].$$

Решая это уравнение относительно k, получаем для искомого оптимума:

$$k = b\left[\frac{\sqrt{2}(N-1)}{\lambda_\alpha + \lambda_{1-p_0}}\right]^{2/5}. \qquad (11.8)$$

Здесь λ_α есть α-точка стандартного нормального распределения.

Для случая простых гипотез (нет свободных параметров) и $p_0 = 0{,}5$ было получено [55] значение $b = 4$. В общем случае для b нужно выбирать значение между 2 и 4. Из (11.8) видно, что k ведет себя, как $N^{2/5}$. Кроме того, ожидаемое число событий в ячейках гистограммы $N/k \sim N^{3/5}$ должно быть не слишком малым, например, $\geqslant 5$. В табл. 11.1 мы приводим ряд чисел, которые следует использовать при практических расчетах.

Таблица 11.1

Значения $k(N/k)$, рассчитанные с помощью (11.8) для значений $b = 4$, $\lambda_{0,5}=0$, $\lambda_{0,2}=0{,}84$, $\lambda_{0,05}=1{,}64$, $\lambda_{0,01}=2{,}33$. Вероятность принять H_0, когда она неверна, равна $1-p_0$

N	α	P_0	
		0,5	0,8
200	0,01	27 (7,4)	24 (8,3)
	0,05	31 (6,5)	27 (7,4)
500	0,01	39 (13)	35 (14)
	0,05	45 (11)	39 (13)

В итоге приходим к таким рекомендациям.

1. Найти число ячеек, используя (11.8), для $b \sim$ от 2 до 4.

2. Если окажется, что N/k мало, уменьшить k, чтобы выполнялось неравенство $N/k \geqslant 5$.

3. Сформировать k равновероятностных ячеек гистограммы либо на основе теоретического распределения, предсказываемого H_0, либо на основе данных. Отметим, что если измеряемая случайная величина многомерна, то существуют различные способы формирования ячеек с одинаковым вероятностным содержанием в k ячейках.

4. Если нужно оценить параметры, то следует использовать метод максимального правдоподобия с использованием индивидуальных измерений, но при этом нужно помнить, что проверочная статистика не полностью независима от вида распределения (см. п. 11.2.2).

§ 11.3. ДРУГИЕ КРИТЕРИИ ДЛЯ ДАННЫХ, СГРУППИРОВАННЫХ В ГИСТОГРАММУ

11.3.1. Критерий серии. Недостатком T-статистики является необходимость группировки данных в гистограмму. Кроме того, эта статистика не использует информацию о знаках разности $(n_i - Np_i)$. Критерий *серий* как раз использует эту информацию. Основное достоинство этого критерия состоит в том, что для *простых гипотез* он не зависит от χ^2-критерия для той же самой гистограммы и поэтому несет дополнительную информацию.

Если верна гипотеза H_0, то оба вида знаков равновероятны. Этот простой факт позволяет получить такие результаты [56]. Пусть M будет числом положительных отклонений, N — числом отрицательных отклонений, а R — полное число *серий*. Серией называется последовательность отклонений одинакового знака, которой предшествует и за которой следует отклонение противоположного знака

(если она не содержит граничные ячейки). Тогда

$$P(R=2s)=\frac{2\binom{M-1}{s-1}\binom{N-1}{s-1}}{\binom{M+N}{M}};$$

$$P(R=2s-1)=\frac{\binom{M-1}{s-2}\binom{N-1}{s-1}+\binom{M-1}{s-1}\binom{N-1}{s-2}}{\binom{M+N}{M}}.$$

Критическая область такого критерия соответствует малым, практически неосуществимым значениям $R: R \leqslant R_{\text{мин}}$. Задав вероятность R, можно вычислить $R_{\text{мин}}$, соответствующее требуемому уровню значимости. Ожидание и дисперсия R равны:

$$E(R)=1+2MN/(M+N);$$

$$D(R)=\frac{2MN(2MN-M-N)}{(M+N)^2(M+N-1)}.$$

Хотя мощность этого критерия обычно меньше мощности χ^2-критерия Пирсона, он не зависит от него и оба эти критерия можно использовать совместно. В результате получается очень важный критерий, свойства которого рассматриваются в § 11.6.

11.3.2. Критерий числа пустых ячеек, порядковые статистики. Предположим, что согласно нулевой гипотезе наблюдения $\mathbf{X}$ имеют функцию распределения $F_0(\mathbf{X})$. Разделим пространство X на M ячеек одинакового вероятностного содержания, как это делалось в п. 11.2.3, и пусть измерения распределены по ячейкам так, что S_0 ячеек не содержит ни одного наблюдения. Назовем S_0 *числом пустых ячеек*. Тогда функция вероятности S_0 для нулевой гипотезы имеет вид:

$$P(S_0=s)=\frac{1}{M^N}\binom{M}{s}\sum_{i=0}^{M-s}\binom{M-s}{i}(-1)^i(M-s-1)^N;$$

s принимает значения $(K, K+1, \ldots, M-1)$, где $K=\max(0, M-N)$.

Имеются таблицы этого распределения. Ожидание и дисперсия S_0 равны:

$$E(S_0)=M(1-1/M)^N;$$

$$D(S_0)=M(M-1)(1-2/M)^N+M(1-1/M)^N-M^2(1-1/M)^{2N}.$$

Если M и N стремятся к бесконечности так, что $\rho=N/M>0$, S_0 асимптотически распределено по нормальному закону $N\{M\exp(-\rho), M[\exp(-\rho)-\exp(-2\rho)(1+\rho)]\}$, что позволяет легко формулировать приближенные критерии. В литературе рекомен-

дуется использовать значение $\rho = 1{,}255$, поскольку при этом распределение быстрее всего сходится к нормальному. Сейчас уместно ввести понятие *порядковой статистики* X_i. Расположим N независимых наблюдений $X_1, \ldots, X_N$ случайной переменной X в порядке их возрастания: $X_{(1)} \leqslant X_{(2)} \leqslant , \ldots, X_{(N)}$ (это всегда возможно, поскольку наблюдения независимы). Упорядоченные наблюдения $X_{(i)}$ называются порядковой статистикой. Их функция распределения (рис. 11.3) имеет вид:

$$S_N(X) = \begin{cases} 0 & X < X_{(1)}; \\ i/N & \text{для } X_{(i)} \leqslant X \leqslant X_{(i+1)}, \\ i = 1, \ldots, N-1; \\ 1 & X_{(N)} \leqslant X. \end{cases} \tag{11.9}$$

Отметим, что $S_N(X)$ всегда возрастает ступеньками равной высоты N^{-1}. В соответствии с определением функции распределения $F(X)$ [см. уравнение (2.16)] и законом больших чисел (см. § 3.3)

$$\lim_{N \to \infty} \{P[S_N(X) = F(X)]\} = 1.$$

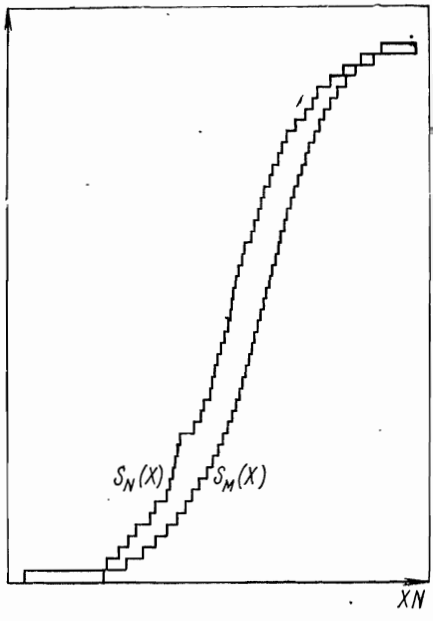


Рис. 11.3. Пример функций распределения $S_N(X)$ и $S_M(X)$ типа (11.9)

В § 11.2 в качестве проверочных статистик будут рассмотрены различные нормы разности $S_N(X) - F(X)$.

Критерий числа пустых ячеек можно использовать при сравнении двух экспериментальных распределений. Пусть первая группа экспериментальных данных содержит N наблюдений X_i с порядковой статистикой $X_{(r)}$, $r = 1, \ldots, N$. Определим совокупность $N + 1$ ячеек I, которая содержит порядковую статистику.

Ячейка I_1 простирается от $-\infty$ до X_1, а ячейка $I_{(N+1)}$ — от $X_{(N)}$ до ∞.

Пусть, далее, второй набор данных содержит M наблюдений Y_i и в ячейку I_i попадает r_i наблюдений этого набора.

Обозначим S_0' число ячеек I_i, в каждую из которых не попадает ни одного наблюдения Y ($r_i = 0$). Функция вероятности S_0' равна

$$P(S_0' = s) = \frac{\binom{N+1}{s}\binom{M-1}{N-s}}{\binom{N+M}{N}}.$$

Параметр s может принимать значения $(K, K+1, \ldots, N)$, где $K = \max[0, N+1-M]$. В асимптотическом пределе S_0' распределено по нормальному закону со средним $(N+1)/(1+l)$ и дисперсией $(N+1)\,l^2/(1+l)^3$, где $l = M/N$.

11.3.3. Критерий «гладкости» Неймана—Бартона. Чтобы критерии были независимы от функции распределения, предсказываемой нулевой гипотезой H_0, нужно представить данные в каком-то стандартном виде. Нейман и Бартон использовали преобразование (11.6). Они рассмотрели одномерный случай и вместо переменных X_i, $i = 1, \ldots, N$ использовали переменные $Y_i = F_0(X_i)$. Если верна гипотеза H_0, то Y_i распределены равномерно между 0 и 1. Следовательно, нужно построить критерии, «чувствующие» отклонения от однородности. Альтернативная гипотеза H_s имеет вид: $H_s : Y_i$ распределены с функцией плотности $f(Y \mid H_s)$, где

$$f(Y \mid H_s) = c(\theta_1, \ldots, \theta_s) \exp\left[1 + \sum_{r=1}^{s} \theta_r l_r(Y)\right].$$

Здесь функции $l_r(Y)$ — полиномы Лежандра порядка r, ортогональные на отрезке $(0,1)$. Тогда нужно сравнить гипотезу

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_s = 0$$

или

$$H_0 : \sum_{r=1}^{s} \theta_r^2 = 0 \tag{11.10}$$

с любой другой, когда эти равенства не выполняются.

Напишем функцию правдоподобия для наблюдений Y_i:

$$L(\mathbf{Y} \mid \theta) = [c(\theta)]^N \exp\left[\sum_{r=0}^{s} \theta_r \sum_{i=1}^{N} l_r(Y_i)\right].$$

Статистика $t_r = \sum_{i=1}^{N} l_r(Y_i)$ является достаточной для θ_r. Поскольку H_0 было записано в виде (11.10), то представляется разумным использовать в качестве проверочной статистики некоторую функцию t_r^2, например,

$$p_s^2 = \frac{1}{N} \sum_{r=1}^{s} t_r^2 = \frac{1}{N} \sum_{r=1}^{s} \left[\sum_{i=1}^{N} l_r(Y_i)\right]^2.$$

Если верна гипотеза H_s, то величина p_s^2 асимптотически распределена по закону нецентрального $\chi^2(s, K)$ с параметром нецентральности

$$K = N \sum_{r=1}^{s} \theta_r^2.$$

Соответственно доверительные уровни для нулевой гипотезы получаются из центрального χ^2 (s)-распределения. Было показано, что распределения величин p_1^2 и p_2^2 достаточно близки к χ^2 (1) и χ^2 (2) соответственно, если $N \geqslant 20$.

Предположим, что наблюдения X_i, $i = 1, \ldots, N$ группируются в гистограмму, содержащую k ячеек с вероятностями:

$$p_i = p_{0i}, \text{ если верна } H_0, \text{ и}$$

$$p_i = p_{si}, \text{ если верна альтернативная гипотеза } H_s.$$

Тогда преобразование к равномерному распределению имеет вид:

$$Y_i' = \sum_{j=1}^{i-1} p_{0j} + \frac{1}{2} p_{0i}.$$

Остается выбрать набор полиномов $P_r(Y')$, удовлетворяющих равенствам:

$$\sum_{i=1}^{k} p_{0i} p_r(Y_i') P_t(Y_i') = \delta_{rt}. \tag{11.11}$$

Проверочная статистика приобретает вид:

$$p_{k-1}^2 = \sum_{r=1}^{k-1} u_r^2 = \frac{1}{N} \sum_{r=1}^{k-1} \left[\sum_{i=1}^{k} n_i P_r(Y_i') \right]^2,$$

где n_i — число наблюдений в i-й ячейке. С помощью (11.11) это выражение точно приводится к уравнению (11.4)

$$p_{k-1}^2 = \sum_{i=1}^{k} \frac{n_i}{N p_{0i}} - N = T.$$

Таким образом, p_{k-1}^2-критерий идентичен χ^2-критерию (см. п. 11.1.2). Статистики p_r^2 порядка $r < k - 1$ могут быть названы компонентами или разбиениями T-статистики.

Несомненным достоинством P_r^2-статистик в случае, когда данные сгруппированы в гистограмму, является тот факт, что они позволяют выделить соответствующие функции Y_i для проверки с тем, чтобы получить максимальную мощность; иными словами, они выбирают важные компоненты T-статистики.

χ^2-Критерий для важных гипотез (когда параметры оцениваются по данным) эквивалентен p_k^2-критериям; при этом число степеней свободы уменьшается на единицу для каждого параметра, оцениваемого методом максимума правдоподобия мультиномиального распределения.

§ 11.4. КРИТЕРИИ, НЕ СВЯЗАННЫЕ С ГРУППИРОВКОЙ ДАННЫХ В ГИСТОГРАММУ

Как уже отмечалось, критерии для данных, не сгруппированных в гистограмму, в принципе должны быть лучше, чем критерии для данных, представленных в виде гистограммы. На первый взгляд может показаться, что хорошим кандидатом для этого может служить правдоподобие данных. К несчастью, оно содержит мало информации (см. п. 11.4.3) как проверочная статистика.

Наиболее удачные критерии основаны на сравнении функции распределения $F(X)$, предсказываемой гипотезой H_0, с эквивалентным распределением данных [см. уравнение (11.9)]. Проверочная статистика является некоторой мерой «расстояния» между экспериментальной и гипотетической функциями распределений. Для критерия Смирнова — Крамера — Мизеса (см. п. 11.4.1) такой мерой является среднеквадратичная разность между двумя функциями, для критерия Колмогорова (см. п. 11.4.2) — минимальная или максимальная разность.

11.4.1. Критерий Смирнова — Крамера — Мизеса. Рассмотрим статистику

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - F(X)]^2 f(X)\, dX, \tag{11.12}$$

где $f(X)$ — ф. п. в., соответствующая гипотезе H_0, $F(X)$ — ее функция распределения, а $S_N(X)$ — определено уравнениями (11.9). Подставляя (11.9) в (11.12), получаем:

$$W^2 = \int_{-\infty}^{X_1} F^2(X)\, dF(X) + \sum_{i=1}^{N-1} \int_{X_i}^{X_{i+1}} \left[\frac{i}{N} - F(X)\right]^2 dF(X) +$$

$$+ \int_{X_N}^{\infty} [1 - F(X)]^2\, dF(X) =$$

$$= \frac{1}{N}\left\{\frac{1}{12N} + \sum_{i=1}^{N}\left[F(X_i) - \frac{2i-1}{2N}\right]^2\right\}. \tag{11.13}$$

При этом использовались такие свойства функции распределения:

$$F(-\infty) \equiv 0 \text{ и } F(+\infty) = 1.$$

Для фиксированного значения X, $S_N(X)$ распределено по биномиальному закону с моментами:

$$E[S_N(X)] = F(X);$$

$$E\{[S_N(X) - F(X)]^2\} = \frac{1}{N} F(X)[1 - F(X)].$$

Соответственно среднее и дисперсия статистики (11.3) равны:

$$E(W^2) = \frac{1}{N}\int_0^1 F(1-F)\,dF = \frac{1}{6N};$$

$$D(W^2) = E(W^4) - E(W^2) = (4N-3)/180\,N^3.$$

Распределение W^2 совершенно не зависит от распределения X даже для конечных N. Это нетрудно увидеть, если произвести замену переменных $Y = F(X)$. Тогда из определения W^2 следует:

$$W^2 = \int_0^1 [S_N(Y) - Y]^2\,dY.$$

В таком виде W^2 уже не зависит от $f(X)$.

Асимптотическая характеристическая функция NW^2 равна [57]:

$$\lim_{N\to\infty} E[\exp(i\,tNW^2)] = \sqrt{\frac{\sqrt{2it}}{\sin\sqrt{2it}}}. \qquad (11.14)$$

Это соотношение можно обратить, что позволяет вычислить критические значения [58]:

Уровень значимости α	Критическое значение NW^2
0,10	0,347
0,05	0,461
0,01	0,743
0,001	1,168

Было показано [59], что с точностью, с которой вычислены критические значения, асимптотический предел достигается, когда $N \geqslant 3$.

Если гипотеза H_0 — сложная, то W^2 в общем случае зависит от вида распределения. Кроме того, если переменная X многомерна, то критерий теряет свои свойства, если компоненты зависимы. Однако можно образовать критерий для сравнения двух распределений $F(X)$ и $G(X)$. Пусть число наблюдений равно N и M соответственно и пусть гипотеза H_0 такова: $F(X) = G(X)$. Тогда проверочная статистика имеет вид

$$W^2 = \int_{-\infty}^{\infty} [S_N(X) - S_M(X)]^2\,d\left[\frac{NF(X)+MG(X)}{N+M}\right]. \qquad (11.15)$$

Тогда уже величина $\frac{MN}{M+N}\,W^2$, где W^2 задается уравнением (11.15), будет иметь асимптотическую характеристическую функцию (11.14).

11.4.2. Критерий Колмогорова. Проверочной статистикой для этого критерия является максимум отклонения наблюдаемого рас-

пределения $S_N(X)$ [см. уравнение (11.9)] от распределения $F(X)$, предсказываемого гипотезой H_0. Она определяется либо как

$$D_N = \max |S_N(X) - F(X)| \quad \text{для всех } X,$$

либо как

$$D_N^{\pm} = \max \{\pm [S_N(X) - F(X)]\} \quad \text{для всех } X,$$

если рассматриваются только односторонние критерии.

Можно показать, что предельными распределениями являются:

для $\sqrt{N}\, D_N$

$$\lim_{N\to\infty} p(\sqrt{N}\, D_N > z) = 2 \sum_{r=1}^{\infty} (-1)^{r-1} \exp(-2r^2 z^2) \qquad (11.16)$$

и для $\sqrt{N}\, D_N^{\pm}$

$$\lim_{N\to\infty} p\ (\sqrt{N}\, D_N^{\pm} > z) = \exp(-2z^2). \qquad (11.17)$$

С другой стороны, вероятностное утверждение (11.17) может быть сформулировано как

$$\lim_{N\to\infty} p[2N (D_N^{\pm})^2 \leqslant 2z] = 1 - \exp(-2z^2). \qquad (11.18)$$

Таким образом, величина $4\ N (D_N^{\pm})^2$ распределена по χ^2 (2)-закону. Считается, что предельные распределения (11.16), (11.17) и (11.18) справедливы при $N \geqslant 80$. Ниже приведены некоторые критические значения $\sqrt{N} D_N$ для распределения (11.16).

Уровень значимости α	Критическое значение $\sqrt{N}\, D_N$
0,01	1,63
0,05	1,36
0,10	1,22
0,20	1,07

Для сравнения двух распределений $S_N(X)$ и $S_M(X)$ (см. рис. 11.3) эквивалентная статистика имеет вид:

$$D_{MN} = \max |S_N(X) - S_M(X)| \quad \text{для всех } X$$

или для односторонних критериев

$$D_{MN}^{\pm} = \max \{\pm [S_N(X) - S_M(X)]\} \quad \text{для всех } X.$$

Соответственно величины $\sqrt{MN/(M+N)}\, D_{MN}$ и $\sqrt{MN/(M+N)} D_{MN}^{\pm}$ имеют предельные распределения (11.16) и (11.17).

Наконец, можно обратить вероятностное утверждение о D_N с тем, чтобы получить *доверительную область* для $F(X)$. Утверждение

$$P\{D_N = \max |S_N(X) - F(X)| > d_{\alpha}\} = \alpha$$

определяет d_α как α-точку D_N. Отсюда следует, что

$$P\{S_N(X) - d_\alpha \leqslant F(X) \leqslant S_N(X) + d_\alpha\} = 1 - \alpha.$$

Поэтому, если граница доверительной области выбирается на расстоянии $\pm d_\alpha$ от $S_N(X)$, то это означает, что с вероятностью $(1 - \alpha)$ $F(X)$ лежит целиком внутри доверительной области (подобно этому может быть использовано $d_\alpha^{\pm}$ для образования односторонних границ). Таким образом, можно рассчитать число наблюдений, необходимое для получения $F(X)$, с некоторой определенной точностью. Предположим, что нужно получить $F(X)$ с погрешностью 0,05 с вероятностью 0,99. Тогда для этого требуется $N = (1{,}628/0{,}05)^2 \sim 1000$ наблюдений.

11.4.3. Использование функции правдоподобия. Предположим, что N наблюдений $\mathbf{X}$ имеют ф. п. в. $f(\mathbf{X})$. Логарифм функции правдоподобия равен:

$$\lambda = \sum_{i=1}^{N} \ln f(X_i).$$

Если по данным не требуется найти оценки неизвестных параметров, то можно в принципе вычислить среднее и дисперсию λ:

$$E_{\mathbf{X}}(\lambda) = N \int \ln f(\mathbf{X})\, f(\mathbf{X})\, d\mathbf{X};$$

$$D_{\mathbf{X}}(\lambda) = N \int [\ln f(\mathbf{X}) - N^{-1} E_{\mathbf{X}}(\lambda)]^2 f(\mathbf{X})\, d\mathbf{X}$$

и даже более высокие моменты, если есть основания думать, что *приближение* нормальным *законом* не совсем верно.

Если же по данным нужно найти оценки неизвестных параметров методом максимального правдоподобия, то расчеты становятся гораздо сложнее. Простое, но очень дорогостоящее решение состоит в том, чтобы генерировать искусственные события методами Монте-Карло. Применяя затем метод максимального правдоподобия, скажем, к каждому из 50 смоделированных экспериментов, можно получить λ для каждого и, таким образом, грубо определить распределение λ. Параметры $\boldsymbol{\theta}$ распределения генеральной совокупности $f(\mathbf{X}, \boldsymbol{\theta})$ при этом полагаются равными оценке $\hat{\boldsymbol{\theta}}$, полученной из реального эксперимента.

Даже если такой метод приемлем с точки зрения затрат (времени на ЭВМ), могут встретиться другие трудности. Например, заранее неизвестно, мощный этот критерий или нет.

Предположим, что данные X имеют ф. п. в. на области конечного размера, и произведем такую замену переменных, при которой новая переменная $X'(X)$ распределена равномерно:

$$g(X') = \frac{f(X)}{|dX'/dX|} = \text{const}.$$

Тогда функция правдоподобия в новых переменных X' также константа, равная Δ^{-N} для N наблюдений, если Δ — область изменения X'. Можно всегда найти такую замену переменных $X'(X)$, которая дает определенный результат, не зависящий от данных. (Отметим, однако, что такая замена переменных не изменяет T-статистику χ^2-критерия.)

§ 11.5. ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ

11.5.1. Наблюдение тонкой структуры. В экспериментальной физике часто возникают случаи, когда некоторое важное явление (резонанс, спектральная линия) проявляется в виде относительно узкого сигнала (пика или провала), наложенного на некоторый гладкий фон (рис. 11.4). Первым, естественно, возникает вопрос: означают ли наблюдения наличие тонкой структуры в области

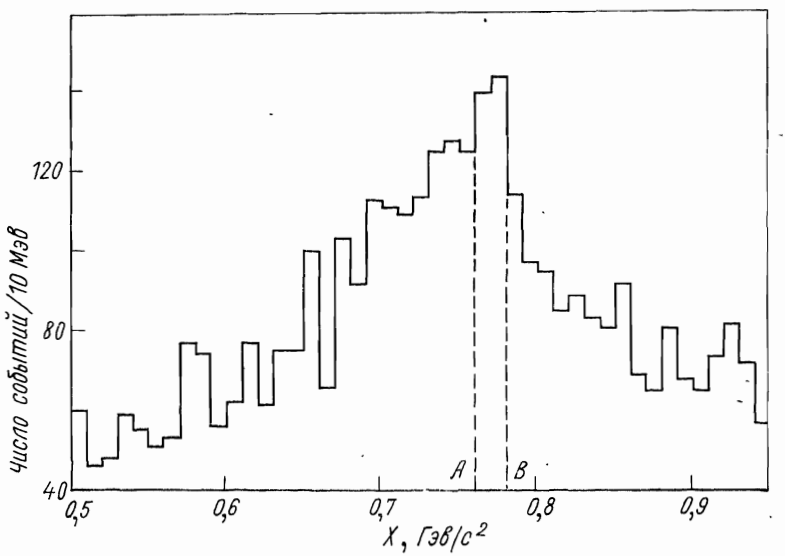


Рис. 11.4. Экспериментальная гистограмма с тонкой структурой в области AB

AB? Если да, то следующая задача состоит в оценке параметров этой структуры, таких, как размер и положение сигнала.

С другой стороны, можно попытаться определить вероятность наблюдения статистической флуктуации данного размера в данной области AB или в произвольной области такой же ширины. Понятно, что нулевая гипотеза (H_0: в области AB нет реального физического эффекта) требует некоторого знания формы фона или шума. Иногда она известна с большой точностью, но чаще она неизвестна и единственно разумным предположением является предположе-

ние о его «гладкости». Это предположение обычно соответствует представлению фона в виде полинома невысокого порядка.

Отметим, что если вся область изменения случайной переменной гораздо больше, чем AB, то проверочная статистика не должна применяться ко всей области, поскольку это легко может привести к принятию H_0, хотя она и неверна (ошибка второго рода). Это действительно может происходить, поскольку чем больше рассматриваемая область, тем больше вероятность того, что где-либо произойдет одна статистическая флуктуация шириной AB. Поэтому единственно, что требуется от фона, — это предсказание его величины и формы в области AB.

Будем описывать фон функцией $b(\mathbf{X}, \theta)$ наблюдений $\mathbf{X}$ и неизвестных параметров θ. Оценки $\hat{\theta}$ и матрица вторых моментов $\underset{\sim}{D}$ могут быть получены методами, описанными в гл. 7 и 8 (исключая область AB), что приводит к

$$\hat{b}_{AB} = \int_A^B b(\mathbf{X}, \hat{\theta})\, d\mathbf{X}.$$

Поскольку $\hat{b}_{AB}$ — функция $\hat{\theta}$, то можно вычислить ее дисперсию с помощью обычных методов замены переменных (см. п. 2.3.2):

$$\sigma_{AB}^2 = \mathbf{D}^T \underset{\sim}{D} \mathbf{D},$$

где $\mathbf{D}$ — вектор производных

$$D_i = \left.\frac{\partial \hat{b}_{AB}}{\partial \theta_i}\right|_{\hat{\theta}_i} = \int_A^B \frac{\partial}{\partial \theta_i} b(\mathbf{X}, \hat{\theta})\, d\mathbf{X}.$$

Пусть число наблюдений в AB равно n_{AB}. Естественной проверочной статистикой для определения того, отличается ли n_{AB} значительно от $\hat{b}_{AB}$, является

$$T = \frac{(n_{AB} - \hat{b}_{AB})^2}{D(n_{AB} - \hat{b}_{AB})},$$

где $D(n_{AB} - \hat{b}_{AB})$ — дисперсия разности. Если верна гипотеза H_0 (см. п. 4.1.2),

$$E(n_{AB}) = b_{AB}; \quad D(n_{AB}) = b_{AB}$$

и оценка b_{AB} равна $\hat{b}_{AB}$. Поэтому

$$D(n_{AB} - \hat{b}_{AB}) \approx \hat{b}_{AB} + \hat{\sigma}_{AB}^2 - 2D(n_{AB}, \hat{b}_{AB}).$$

Поскольку при оценке θ мы исключили область AB, θ, n_{AB} и $\hat{b}_{AB}$ не коррелированы:

$$D(n_{AB} - \hat{b}_{AB}) \approx \hat{b}_{AB} + \hat{\sigma}^2_{AB} \tag{11.19}$$

и

$$T \approx \frac{(n_{AB} - \hat{b}_{AB})^2}{\hat{b}_{AB} + \hat{b}^2_{AB}}. \tag{11.20}$$

Если n_{AB} достаточно велико, то оно распределено по нормальному закону относительно $\hat{b}_{AB}$ и T ведет себя, как χ^2 (1). Часто статистику (11.20) выражают в терминах *стандартных отклонений* (это эквивалентно предыдущему):

$$d = \sqrt{T}. \tag{11.21}$$

Тогда можно говорить о *шансах* того, что не будет наблюдаться превышение в данное число стандартных отклонений (табл. 11.2). (Предостережение: поскольку приближение о том, что n_{AB} распределено по нормальному закону, становится плохим по мере возрастания d, то шансы в этой таблице для больших d могут не соответствовать действительности.)

Таблица 11.2

Шансы того, что не будет превышено d стандартных отклонений для нормального распределения

d	Шансы	d	Шансы
1	2,15 : 1	4	16000,0 : 1
2	21,0 : 1	5	$1{,}7 \cdot 10^6$: 1
3	370,0 : 1		

До сих пор мы неявно предполагали, что область AB *выбрана независимо* от *наблюдений*. Например, тонкая структура уже могла наблюдаться в предыдущих экспериментах и нужно точно установить ее существование и ее форму. Однако, если выбор области AB основан на данных, то числа в табл. 11.2 уже не соответствуют действительности, поскольку они не учитывают вероятность появления сигнала в произвольном месте области изменения измеряемой переменной.

Для иллюстрации этого рассмотрим сигналы, которые целиком умещаются в одной ячейке. Пусть p будет вероятностью превысить d стандартных отклонений в данной ячейке. Если заранее номер ячейки неизвестен, то вероятность флуктуации размером более чем d стандартных отклонений по крайней мере в одной из ячеек k равна:

$$q = 1 - (1-p)^k \approx kp.$$

Например, если гистограмма содержит 40 ячеек, то флуктуация в 3 стандартных отклонения в данной ячейке имеет ту же вероятность, что и флуктуация в 4 стандартных отклонения в какой-нибудь одной (произвольной) ячейке. В общем случае, если для данной ячейки j вероятность, что d_j превысит α-точку λ_α, равна

$$P(d_j > \lambda_\alpha) = \alpha,$$

тогда

$$P \text{ (у указанных } l \text{ ячеек } d_j > \lambda_\alpha) = \binom{k}{l} \alpha^l (1 - \alpha)^{k-l}.$$

и

$$P \text{ (у любых } l \text{ ячеек } d_j > \lambda_\alpha) = \sum_{m \geqslant l}^{k} \binom{k}{m} \alpha^m (1 - \alpha)^{k-m}.$$

Что случится, если ширина сигнала равна двум ячейкам, трем ячейкам и т. д.? Если ширина сигнала известна до эксперимента, то можно выбрать длину ячеек такой, чтобы сигнал умещался в одной ячейке. Если, с другой стороны, соответствующее решение принимается на основе данных, то задача существенно усложняется. По существу эта задача подобна задаче о функции разрешения, с которой мы встречались раньше (см. п. 4.3.4).

Если ожидаемая ширина тонкой структуры мала по сравнению с шириной экспериментальной функции разрешения, то ее трудно наблюдать: узкий пик над фоном станет ниже и шире, а узкий провал будет выше и шире. Таким образом, при оценке эффекта важно учитывать функцию разрешения.

Предположим, что на основе статистики (11.20) или (11.21) было сделано заключение, что наблюдаемый сигнал значим, т. е. отвергается нулевая гипотеза

H_0: в области AB нет реального физического сигнала.

Следующий шаг состоит в том, чтобы оценить размер сигнала. Для этого можно использовать величину $s = (n_{AB} - \hat{b}_{AB})$, но для дисперсии оценки уже нельзя использовать формулу (11.19). Теперь нужно проверять не согласие гипотезы с экспериментальными данными, а гипотезу H_1 (предполагая ее верной) против H_0, где

H_1 : в области AB реальный физический сигнал имеет величину s, а фон — величину b_{AB}.

Тогда дисперсия (11.19) приобретает вид:

$$D(n_{AB} - \hat{b}_{AB}) \approx n_{AB} + \hat{\sigma}^2_{AB}. \qquad (11.22)$$

По той же причине, что и раньше, в этой формуле отсутствуют корреляционные члены.

Пусть ошибка второго рода равна:

$$\beta = P\,(d \leqslant \lambda_\alpha \mid H_1),$$

где d определяется уравнением (11.21). Если верна H_1, то ожидание и дисперсия n_{AB} равны:

$$E(n_{AB}) = D(n_{AB}) = b_{AB} + s$$

и d — приблизительно распределено, как $N(\mu, \sigma^2)$, с

$$\mu = \frac{s}{\sqrt{\hat{b}_{AB} + \hat{\sigma}^2_{AB}}}, \quad \sigma^2 = \frac{\hat{b}_{AB} + s + \hat{\sigma}^2_{AB}}{\hat{b}_{AB} + \hat{\sigma}^2_{AB}}.$$

Отсюда следует, что

$$\beta = \Phi\left(\frac{\lambda_\alpha \sqrt{\hat{b}_{AB} + \hat{\sigma}^2_{AB}} - s}{\sqrt{\hat{b}_{AB} + s + \hat{\sigma}^2_{AB}}}\right). \tag{11.23}$$

11.5.2. Комбинирование независимых оценок. Предположим, что нужно получить более точную оценку набора r параметров θ, имея комбинации независимых оценок $\hat{\theta}$ из N различных экспериментов, и предположим, что матрица вторых моментов оценок в эксперименте J равна D_J. Наиболее точная процедура состояла бы в суммировании информации о θ с помощью правдоподобия эксперимента или, может быть, с помощью достаточной статистики, но обычно ни то, ни другое неизвестно (а последнее редко, когда существует).

Если использовать только оценки $\hat{\theta}_J$ и матрицу вторых моментов D_J, то это эквивалентно предположению о нормальном распределении оценок, что в свою очередь справедливо при достаточно большом числе наблюдений. К несчастью, часто делают даже более грубые приближения. Это случается, когда (как правило) экспериментатор приводит оценку $\hat{\theta}_i$ и 68,3%-ный доверительный интервал $\pm\, \varepsilon_i$ для каждого параметра i, как если бы они были не коррелированы (хотя, возможно, что при определении ε он использовал полную матрицу вторых моментов). В этом случае приходится приближать D диагональной матрицей с элементами ε_i^2.

Будем предполагать, что оценки распределены по нормальному закону и что их матрицы вторых моментов известны. Предположим далее, что решено исключить $N - M$ экспериментов, как содержащие слишком мало информации (малая точность) по сравнению с другими экспериментами. Если предположение о нормальности действительно верно для исключенных экспериментов, то теряется малая доля информации. Если такое приближение не оправдано, то учет соответствующих экспериментов может сильно загрубить общий результат.

Очевидно, здесь мы сталкиваемся с задачей определения потери (оценок, распределенных по нормальному закону) и примеси (оце-

нок, ошибочно считающихся распределенными по нормальному закону), для которой не существует общего решения.

Предположим, что оставшимся M экспериментальным результатам можно без сомнения доверять. Тогда можно было бы, используя предположение о нормальности, сконструировать достаточную статистику, что позволило бы получить суммарную оценку. Однако может возникнуть вопрос, а не смещены ли некоторые оценки, либо, может быть, дисперсии некоторых оценок занижены или завышены?

Поэтому вначале следует провести *проверку совместимости* этих результатов, а затем, может быть, отбросить некоторые из них.

Пусть нулевой гипотезой будет:

H_0 : все эксперименты — это выборка наблюдений из распределений с одинаковыми средними и с правильно выбранными распределениями.

Чтобы найти проверочную статистику, мы воспользуемся методом наименьших квадратов. Пусть $\hat{\theta}$ будет оценкой, для которой *результирующая сумма квадратов*

$$Q^2(\theta) = \sum_{J=1}^{M} (\theta - \hat{\theta}_J)^T \underset{\sim}{D_J^{-1}} (\theta - \hat{\theta}_J)$$

минимальна: $Q^2_{\text{мин}} = Q^2(\hat{\hat{\theta}})$.

Решив уравнения

$$\left.\frac{\partial Q^2(\theta)}{\partial \theta}\right|_{\theta = \hat{\hat{\theta}}} = 0,$$

получим

$$\hat{\hat{\theta}} = \left(\sum_{J=1}^{M} \underset{\sim}{D_J^{-1}}\right)^{-1} \left(\sum_{J=1}^{M} \underset{\sim}{D_J^{-1}}\, \hat{\theta}_J\right). \tag{11.24}$$

Дисперсия оценки (11.24) равна:

$$\underset{\sim}{D} = \left(\sum_{J=1}^{M} \underset{\sim}{D_J^{-1}}\right)^{-1}.$$

Хотя здесь и предполагается, что различные эксперименты не коррелированы, однако можно легко обобщить этот метод на случай когда они коррелированы, например, когда они используют одну и ту же константу, известную с не слишком хорошей точностью.

Критерий использует статистику $Q^2_{\text{мин}}$, о которой известно (в соответствии с предположением о нормальности), что она ведет себя, как $\chi^2\,[r\,(M-1)]$, если использовано M экспериментов, а θ — r-мерно. Эксперименты совместимы с доверительным уровнем $(1-\alpha)$, если

$$Q^2_{\text{мин}} \leqslant \lambda_\alpha, \tag{11.25}$$

где λ_α — α-точка χ^2 $[r\,(M-1)]$-распределения. Если $r=1$, то эта процедура сводится к хорошо известной *процедуре определения взвешенного среднего*:

$$\hat{\hat{\theta}}=\frac{\sum_{J=1}^{M}\sigma_J^{-2}\hat{\theta}_J}{\sum_{J=1}^{M}\sigma_J^{-2}}, \tag{11.26}$$

где σ_J^2 — дисперсия $D\,(\hat{\theta}_J)$ оценки θ в эксперименте J.

Если весь набор M экспериментов удовлетворяет критерию (11.25), то они совместимы и нет сомнения, что наилучшей будет оценка (11.24) или (11.26). Теперь давайте рассмотрим случай, когда эксперименты в соответствии с критерием (11.25) несовместимы. Как выделить плохие эксперименты, ответственные за невыполнение критерия?

Сначала нужно обратиться к физическим аргументам, а не к статистическим, поскольку, даже когда очевидно, что несовместимость обусловлена конкретным экспериментом I, статистика не может сказать, что эксперимент I неверен, а остальные $M-1$ экспериментов верны и наоборот. Иначе говоря, эксперименты могут быть с надежностью отвергнуты только на основе физических аргументов (если они существуют).

Сейчас мы рассмотрим два частных случая подхода к проблеме несовместимости с позиций статистики.

1. Выделение одного эксперимента с номером I.

Можно проверять каждый эксперимент, используя при конструировании проверочной статистики величину $\hat{\hat{\theta}}$. Проверочная статистика является функцией разности между $\hat{\hat{\theta}}$ и $\hat{\theta}_I$:

$$T_I=(\hat{\hat{\theta}}-\hat{\theta}_I)^T\,\underset{\sim}{W_I}^{-1}\,(\hat{\hat{\theta}}-\hat{\theta}_I),$$

где $\underset{\sim}{W_I}$ — дисперсия этой разности. Матрицу $\underset{\sim}{W_I}$ можно вычислить следующим образом:

$$\underset{\sim}{W_I}=E\,[(\hat{\hat{\theta}}-\hat{\theta}_I)\;(\hat{\hat{\theta}}-\hat{\theta}_I)^T]=E\,\{[(\hat{\hat{\theta}}-\theta)-(\hat{\theta}_I-\theta)]\,[(\hat{\hat{\theta}}-\theta)-$$

$$-(\hat{\theta}_I-\theta)]^T\}=E\,[(\hat{\hat{\theta}}-\theta)\,(\hat{\hat{\theta}}-\theta)^T]-2E\,[(\hat{\hat{\theta}}-\theta)\,(\hat{\theta}_I-\theta)^T]+$$

$$+E\,[(\hat{\theta}_I-\theta)\,(\hat{\theta}_I-\theta)^T]=\underset{\sim}{D}-2E\,[(\hat{\hat{\theta}}-\theta)\,(\hat{\theta}_I-\theta)^T]+\underset{\sim}{D_I}.$$

Перекрестный член может быть найден, если для $\hat{\hat{\theta}}$ использовать выражение (11.24):

$$E\,[(\hat{\hat{\theta}}-\theta)\,(\hat{\theta}_I-\theta)^T]=\underset{\sim}{D}\sum_J\underset{\sim}{D_J}^{-1}\,E\,[(\hat{\theta}_J-\theta)\,(\hat{\theta}_I-\theta)^T].$$

Поскольку мы предполагаем, что эксперименты не коррелированы, остаются только члены с $J = I$. В результате перекрестный член равен просто $\underset{\sim}{D}$ и соответственно

$$\underset{\sim}{W}_I = \underset{\sim}{D}_I - \underset{\sim}{D}$$

или

$$T_I = (\hat{\hat{\theta}} - \hat{\theta}_I)^T (\underset{\sim}{D}_I - \underset{\sim}{D})^{-1} (\hat{\hat{\theta}} - \hat{\theta}_I). \qquad (11.27)$$

Если эксперименты коррелированы, то ход рассуждений такой же, только нужно учитывать вклад недиагональных членов. Если θ одномерно, уравнение (11.27) сводится к

$$T_I = \frac{(\hat{\hat{\theta}} - \hat{\theta}_I)^2}{\sigma_I^2 - \sigma^2} \qquad (11.28)$$

и влияние эксперимента I на оценку $\hat{\hat{\theta}}$ равно:

$$\frac{\hat{\hat{\theta}} - \hat{\theta}_I}{\sqrt{\sigma_I^2 - \sigma^2}}. \qquad (11.29)$$

Хотя выражение (11.29) общеизвестно, тот факт, что под корнем стоит знак минус, представляется удивительным для многих пользователей, что обусловлено сильной корреляцией ($\rho = \sigma/\sigma_I$) между $\hat{\hat{\theta}}$ и $\hat{\theta}_I$. Статистика (11.29) имеет стандартное нормальное распределение $N(0, 1)$, нормальное по предположению и стандартное по построению.

2. *Случай, когда для всех M экспериментов необходимо введение масштабного множителя.* Если измеренные дисперсии σ_j^2 одного параметра θ известны с точностью до некоторого общего множителя k^2, то часть информации, содержащейся в данных, должна быть использована для оценки масштабного множителя k. Понятно, что при этом любые критерии несовместимости данных будут менее чувствительными.

Для того чтобы проверить, не противоречит ли наблюдение другим наблюдениям, можно использовать проверочную статистику

$$T_I' = \frac{(\hat{\hat{\theta}} - \hat{\theta}_I)^2}{\hat{k}^2 (\sigma_I^2 - \sigma^2)}. \qquad (11.30)$$

В соответствии с (8.24) оценка масштабного множителя равна:

$$\hat{k} = \sqrt{Q^2_{\text{мин}}/(M - 1)},$$

статистика

$$\tau_1 = \sqrt{T_I'} = (\hat{\hat{\theta}} - \hat{\theta}_I)/\hat{k}\sqrt{\sigma_I^2 - \sigma^2}$$

имеет распределение с ф. п. в.*:

$$p(\tau_1) = \frac{1}{\sqrt{\pi \cdot (M-1)}} \frac{\Gamma((M-1)/2)}{\Gamma((M-2)/2)} (1-\tau_1^2/(M-1))^{(M-4)/2};$$

$$M \geqslant 3.$$

Отметим, что область изменения τ_1 конечна: $-\sqrt{M-1} \leqslant r_1 \leqslant \sqrt{M-1}$ и при $M = 4$ ее ф. п. в. просто константа. При больших M оно близко к стандартному нормальному распределению (табл. 11.3).

Таблица 11.3

Зависимость критических значений статистики $|\tau|$ от числа экспериментов M при различных уровнях значимости

Число экспериментов M	Двусторонние уровни значимости для τ_1				
	10%	5%	1%	0,1%	Граничное значение
4	1,56	1,64	1,72	1,73	1,732
6	1,63	1,81	2,05	2,18	2,236
8	1,64	1,87	2,21	2,45	2,646
10	1,65	1,90	2,29	2,62	3,000
∞ (случай нормального распределения)	1,64	1,96	2,58	3,29	∞

Другой, более простой метод состоит в том, чтобы рассчитать $\hat{\hat{\theta}}$, $\hat{k}$ и σ, не учитывая эксперимент I. Тогда проверочная статистика приобретает такой вид:

$$\tau_2 = \frac{\hat{\hat{\theta}} - \hat{\theta}_I}{\hat{k}\sqrt{\sigma^2 + \sigma_I^2}}.$$

Величина τ_2 имеет t-распределение Стьюдента с $M - 2$ степенями свободы.

11.5.3. Сравнение распределений. Аналогично тому, как статистика Колмогорова позволяет сравнивать два одномерных распределения данных, не сгруппированных в гистограммы, хотелось бы иметь метод для сравнения различных гистограмм. С этой целью мы обратимся к χ^2-статистике (см. § 11.2).

---

* Мы благодарны доктору В. И. Лейху за ознакомление с этим результатом.

Пусть число гистограмм равно r и каждая имеет k ячеек. Пусть i-я ячейка j-й гистограммы содержит n_{ij} наблюдений, так что

$$\sum_{i=1}^{k} n_{ij} = N_j, \quad j = 1, \ldots, r;$$

$$\sum_{j=1}^{r} n_{ij} = m_i, \quad i = 1, \ldots, k.$$

Нулевая гипотеза задается ф. п. в. распределения генеральной совокупности с вероятностным содержанием p_i в i-й ячейке, одинаковым для всех r гистограмм. Величины p_i должны быть оценены прежде, чем может быть сформулирован критерий. Пусть μ_j будет ожидаемым значением для N_j, тогда правдоподобие наблюдений равно:

$$L(\mathbf{n} \mid \mathbf{p}) = \prod_{j=1}^{r} \left\{ \left(\frac{\mu_j^{N_j} \exp(-\mu_j)}{N_j!} \right) (N_j!) \left(\prod_{i=1}^{k} \frac{p_{ij}^{n_{ij}}}{n_{ij}!} \right) \right\} =$$

$$= \left[\left(\sum_i m_i \right)! \prod_{i=1}^{k} \left(\frac{p_i^{m_i}}{m_i!} \right) \right] \left[\frac{\prod_{j=1}^{r} N_j!}{\left(\sum_{j=1}^{r} N_j \right)!} \right] \left[\prod_{i=1}^{k} \frac{m_i!}{\sum_{j=1}^{r} (n_{ij}!)} \right] \times$$

$$\times \left(\prod_{j=1}^{r} \frac{\mu_j^{N_j} \exp(-\mu_j)}{N_j!} \right). \tag{11.31}$$

При вычислении $\partial \ln L / \partial p_i$ отметим, что только первая скобка в (11.31) зависит от p_i. Соответственно производная по p_i идентична производной по $\hat{q}_i$ в § 11.1 и

$$\hat{p}_i = m_i / \sum_{i=1}^{k} m_i.$$

Выберем в качестве проверочной статистики

$$T = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{[n_{ij} - E(n_{ij})]^2}{E(n_{ij})}, \tag{11.32}$$

где $E(n_{ij}) = p_i E(N_j)$. Заменив p_i и $E(N_j)$ на их оценки $\hat{p}_i$ и N_j соответственно, получим

$$E(n_{ij}) = \frac{m_i}{\sum_i m_i} N_j = \frac{\sum_{l=1}^{r} n_{il}}{\sum_{l=1}^{r} \sum_{q=1}^{k} n_{ql}} N_j.$$

В асимптотическом пределе, когда число событий достаточно велико, для того чтобы можно было считать n_{ij} распределенной по нормальному закону $N\{E(n_{ij}), E(n_{ij})\}$, величина T ведет себя, как χ^2. Число степеней свободы может быть рассчитано по числу наблюдений (rk) и числу оцениваемых параметров: r значений N_j и $(k-1)$ значений p_i (поскольку их сумма равна 1 : $\sum_i p_i = 1$). Таким образом, T имеет χ^2 $[(k-1)(r-1)]$-распределение.

Статистика (11.32) может быть записана по-разному:

$$T = \left(\sum_{j=1}^{r} N_j\right)\left[\sum_{j=1}^{r}\left(\frac{1}{N_j}\sum_{i=1}^{k}\frac{n_{ij}^2}{m_i}\right)-1\right] =$$

$$= \frac{\left(\sum_{j=1}^{r} N_j\right)}{\left(\prod_{j=1}^{r} N_j\right)}\left\{\sum_{i=1}^{k}\left[\frac{\sum_{j=1}^{r}\left(n_{ij}^2\prod_{l\neq j}^{r} N_l\right)}{m_i}\right]-\prod_{j=1}^{r} N_j\right\} =$$

$$= \left(\sum_{j=1}^{r} N_j\right)\left\{(r-1)-\sum_{j=1}^{r}\sum_{l>j}^{r}\left[\frac{N_j+N_l}{N_j N_l}\sum_{i=1}^{k}\left(\frac{n_{ij}\,n_{il}}{m_i}\right)\right]\right\}. \qquad (11.33)$$

Для простого случая $r = 2$, соотношение (11.33) может быть переписано в виде:

$$T = \left(\frac{N_1+N_2}{N_1 N_2}\right)\left[(N_1+N_2)\sum_{i=1}^{k}\left(\frac{n_{i1}^2}{m_i}\right)-N_1^2\right] =$$

$$= \left(\frac{N_1+N_2}{N_1 N_2}\right)\left[(N_1+N_2)\sum_{i=1}^{k}\left(\frac{n_{i2}^2}{m_i}\right)-N_2^2\right].$$

Если независимо от наблюдений можно разделить ячейки на две или несколько подгрупп, то внутри каждой подгруппы можно проверять сходство распределений. Это приведет к *расщеплению* полного χ^2 на компоненты, каждая из которых распределена независимо от других по χ^2-закону. Такое разделение часто может улучшать точность χ^2-критерия. Конечно, это зависит от возможности *априори* разделить ячейки на подгруппы, между которыми или внутри которых предполагается найти отклонения.

Мы проиллюстрируем этот метод на примере. Допустим, что можно разделить совокупность k ячеек на две группы A и B и внутри нужно проверить, нет ли отклонений. Разделение на две группы A и B сделано *априори без какой-либо информации о наблюдениях*. Тогда полная статистика T может быть разделена на три статистики:

$$T = T_A + T_B + T_{A\,|\,B},$$

каждая из которых имеет χ^2-распределение. Эти статистики имеют вид:

$$T_A = \frac{(N_1+N_2)^2}{N_1 N_2}\left[\sum_{i\in A}\left(\frac{n_{i1}^2}{m_i}\right)-\left(\frac{n_{A1}^2}{m_A}\right)\right];$$

$$T_B = \frac{(N_1+N_2)^2}{N_1 N_2}\left[\sum_{i\in B}\left(\frac{n_{i1}^2}{m_i}\right)-\left(\frac{n_{B1}^2}{m_B}\right)\right];$$

$$T_{A\mid B} = \frac{(N_1+N_2)^2}{N_1 N_2}\left[\frac{n_{A1}^2}{m_A}+\frac{n_{B1}^2}{m_B}-\frac{N_1^2}{N_1+N_2}\right],$$

где мы использовали обозначения $n_{Aj} = \sum\limits_{i\in A} n_{ij}$, $m_A = \sum\limits_{i\in A} m_i$ и те же самые обозначения для B. По построению эти статистики имеют распределения χ^2 $(k_A - 1)$, χ^2 $(k_B - 1)$ и χ^2 (1) соответственно, а k_A, k_B — числа ячеек во множествах A и B.

Эти результаты легко обобщить на случай, когда подгрупп больше чем две. Однако всегда нужно помнить, что такое разделение должно быть сделано без *использования наблюдений.*

§ 11.6. КОМБИНИРОВАНИЕ НЕЗАВИСИМЫХ КРИТЕРИЕВ

Может случиться, что два или больше двух критериев могут быть применены к одним и тем же данным или один и тот же критерий применен к различным группам данных таким образом, что хотя ни один критерий сам по себе не зна́чим, комбинация критериев становится зна́чимой. Конечно, при использовании комбинированного критерия нужно знать свойства каждого критерия и дополнительно к этому возникают две новые проблемы:

а) установление, что индивидуальные критерии *независимы*;

б) нахождение *уровня значимости* комбинированного критерия.

11.6.1. Независмоисть критериев. Каждый из комбинируемых критериев задается своей проверочной статистикой. Говорят, что критерии *независимы*, если проверочные статистики являются случайными переменными в том смысле, как это определено в п. 2.2.2. Это означает, что выражение для вероятности (при заданной нулевой гипотезе) получения данных значений проверочных статистик t_i имеет вид произведения различных членов, каждый из которых зависит только от одной статистики:

$$P(t_1, t_2 \mid H_0) = P(t_1 \mid H_0)\, P(t_2 \mid H_0). \qquad (11.34)$$

На практике иногда бывает трудно записать вероятность (11.34), хотя часто по способу построения различных критериев можно понять, возможно ли такое представление в принципе. Из критериев, описанных в этой главе, только χ^2-критерий Пирсона и критерий серий (асимптотически) независимы [2]. Интуитивно это понятно, поскольку критерий Пирсона не зависит от упорядочения ячеек или от знаков отклонений в каждой ячейке, тогда как именно это являет-

ся единственной информацией для критерия серий. Обычно критерий Пирсона, хоть он чаще всего используется при анализе данных, критикуют за недостаток мощности, обусловленный тем, что эта информация не учитывается. Если комбинировать эти два критерия, то уже не будет основания для подобной критики.

Если H_0 не является действительно простой гипотезой (т. е. когда по данным оценивается один или несколько параметров), критерий Пирсона все еще асимптотически независим от критерия серий, но в этом случае распределение статистики Пирсона известно лишь приближенно (см. п. 11.2.2), а распределение серий обычно вовсе не известно, так что комбинирование критериев не имеет практического смысла.

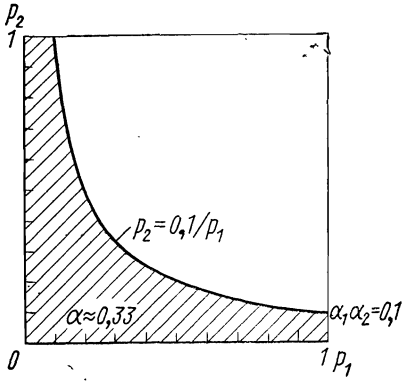


Рис. 11.5. Комбинирование двух непрерывных критериев

11.6.2. Уровень значимости комбинированного критерия. Предположим, что к одному и тому же набору данных были применены два независимых критерия с уровнями значимости $p_1 = \alpha_1$ и $p_2 = \alpha_2$. Тогда можно было бы предположить, что общий уровень значимости α (т. е. вероятность, что p_1 и p_2 таковы, что $p_1p_2 < \alpha_1\alpha_2$) равен $\alpha = \alpha_1\alpha_2$. При этом рассуждения были бы такими: предположим, что H_0 верна; тогда вероятность, что $p_1 \leqslant \alpha_1$ и одновременно $p_2 \leqslant \alpha_2$, конечно, равна $\alpha_1\alpha_2$. Подобно многим утверждениям о вероятностях, это утверждение верно, но только его нельзя использовать для установления общего уровня значимости α [60]. Трудность состоит в том, что $p_1 \leqslant \alpha_1$ и $p_2 \leqslant \alpha_2$ является достаточным условием для $(p_1p_2 \leqslant \alpha_1\alpha_2)$, но не необходимым, т. е. что вероятность произведения не равна произведению вероятностей.

Действительно, например, когда соответствующие переменные непрерывны и распределения их индивидуальных доверительных уровней p_1 и p_2 равномерны (если верна H_0) между нулем и единицей, задача очевидна. Вероятность, что $p_1p_2 \leqslant \alpha_1\alpha_2$, получается в результате интегрирования функции, изображенной на рис. 11.5:

$$\alpha = \alpha_1\alpha_2 [1 - \ln (\alpha_1\alpha_2)]. \qquad (11.35)$$

Это больше, чем $\alpha_1\alpha_2$.

Если имеется больше чем два критерия, аналитические выражения усложняются, но в результате простого преобразования все равно можно получить χ^2-распределение. Пусть имеется N независимых

критериев с (непрерывными) доверительными уровнями α_i. Положим

$$\alpha' = -2 \ln \prod_{i=1}^{N} \alpha_i.$$

Тогда α' распределено (если верна H_0), как $\chi^2(2N)$.

Если одна или несколько проверочных статистик дискретны (например, могут принимать только целые значения), проведенный выше анализ уже не верен [59]*, хотя он может служить хорошим приближением, если дискретная статистика «почти» непрерывна (может принимать много значений). Чтобы увидеть, как должны быть модифицированы расчеты, рассмотрим комбинирование одного непрерывного критерия и одного дискретного критерия (это соответствует случаю комбинирования критерия Пирсона χ^2 и критерия серий). При этом набор возможных значений α_1 и α_2 не заселяет равномерно единичный квадрат, вместо этого он ограничен некоторыми значениями, соответствующими, например, вертикальным линиям на рис. 11.6. Как и прежде, мы проводим кривую $p_2 = \alpha_1\alpha_2/p_1$ так, что для всех точек ниже кривой произведение меньше, чем $\alpha_1\alpha_2$. Для оценки α нужно просто вычислить вероятность попадания разрешенной точки под эту кривую. Соответствующая величина равна не полной области под кривой, а области, заштрихованной на рис. 11.6. Поскольку эта область всегда меньше, чем полная область под кривой, истинный доверительный уровень *меньше*, чем тот, который дается «непрерывным» приближением [см. уравнение (11.35)].

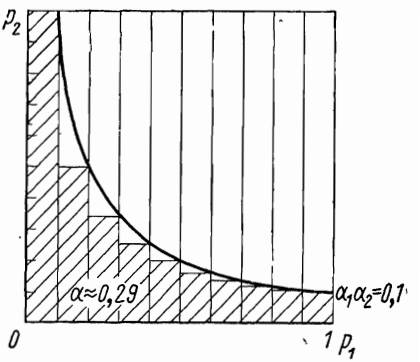


Рис. 11.6. Комбинирование одного непрерывного и одного дискретного критериев

§ 11.7. ОРТОГОНАЛЬНЫЕ ПОЛИНОМЫ (ПРИЛОЖЕНИЕ)

В этой книге мы часто использовали ортогональные полиномы (см. пп. 8.1.5, 8.2.1, 8.4.2 и 10.5.5). В этом приложении мы сделаем несколько замечаний по поводу их использования в различных статистических методах.

11.7.1. Определения. Для того чтобы не дублировать рассуждения для непрерывных и дискретных переменных, мы вначале введем

---

* Мы благодарны мистеру Х. де Стеблеру за ознакомление нас с этой работой.

понятие скалярного произведения двух функций $g(X)$ и $h(X)$, $(g, h) = \int w(X)\, g(X)\, h(X)\, dX$ или

$$(g, h) = \sum_{i=1}^{N} w^*(X_i)\, g(X_i)\, h(X_i), \tag{11.36}$$

где функция $w(X) \geqslant 0$ [или $w^*(X_i) \geqslant 0$] — весовая *функция*.

Определим свойство ортогональности через соотношение

$$(f_i\, f_j) = \beta_i\, \delta_{ij},$$

где β — нормировочная константа. Имея в своем распоряжении данный набор функций $g_i(X)$, построим новый набор функций $f_i(X)$, используя метод Шмидта. Запишем g_i через f_j, $j = 1, \ldots, i$:

$$g_i = \sum_{j=1}^{i} a_{ij} f_j.$$

Из ортогональности функций f следует:

$$a_{ii} f_i = g_i - \sum_{j=1}^{i-1} \left[\frac{(g_i, f_j)}{\beta_j}\right] f_i. \tag{11.37}$$

Соответственно из нормировки f_i следует:

$$a_{ii}^2 = \frac{1}{\beta_i}\left\{(g_i, g_i) - \sum_{j=1}^{i-1} \frac{1}{\beta_j} (g_i, f_j)^2\right\}. \tag{11.38}$$

Знаки коэффициентов a_{ii} несущественны: произвольно они могут быть выбраны положительными.

Конкретизируем семейство полиномов. Положив $g_i = X^i$, $i = 0, \infty$ и взяв $f_0 = \sqrt{\beta_0}$, получим в соответствии с (11.37) и (11.38):

$$f_n = \left\{\frac{\beta_n}{(X^n, X^n) - \sum_{j=0}^{n-1} \frac{(X_i^n f_j)^2}{\beta_j}}\right\}^{1/2} \left\{X^n - \sum_{j=0}^{n-1} \frac{(X^n, f_j)}{\beta_j} f_j\right\}. \tag{11.39}$$

Если весовая функция и нормировочные константы β_i выбраны, то функции f_i полностью определены. Если весовая функция достаточно проста (и, кроме того, для дискретного случая, если расположение точек X_i достаточно регулярно), то можно получить более простые формулы для нахождения f_i. В литературе, например, в [23] можно найти такие формулы вместе с таблицами для полиномов Якоби, Лежандра, Чебышева и т. д. Однако, если нужно сконструировать ортогональные полиномы на произвольном наборе дискретных точек, то нужно использовать уравнение (11.39).

11.7.2. Использование ортогональных функций. Вспомним два случая, когда были использованы ортогональные функции. Мы видели, что в случае линейной модели метода наименьших квадратов (см. пп. 8.4.1 и 8.4.2) ортогональные функции (фактически полино-

мы) приводят к более простым вычислениям, поскольку инвертируемая матрица диагональна. Эта матрица пропорциональна матрице дисперсий оценок, откуда следует, что эти оценки не коррелированы. Соответственно это упрощает проверку необходимости введения в разложение полиномов более высокого порядка для описания данных (см. п. 10.5.5).

Во втором случае (см. п. 8.2.1) ф. п. в. наблюдений была представлена в виде линейной по параметрам функций, а коэффициенты разложения были ортогональными функциями.

Отметим, что в обоих случаях процедура ортогонализации одинакова. В первом случае ортогональность имеет место на дискретном множестве значений Y_i, наблюденных в точках X_i, тогда как во втором случае доступной информацией является набор точек X_i, так что ортогональность определена независимо от наблюдений на непрерывной переменной X. Рассмотрим первый случай. Предполагая, что матрица дисперсий равна $\underset{\sim}{D}_Y$, и используя линейную модель

$$Y_i = \sum_{k=1}^{n} \theta_k f_k(X_i),$$

получаем для оценок и параметров при N наблюдениях:

$$\hat{\theta}_k = \frac{1}{N} \sum_{i=1}^{N} Y_i f_k(X_i).$$

После несложных вычислений получим:

$$(\hat{\theta}_k - \theta_k) = \frac{1}{N} \sum_i \varepsilon_i f_k(X_i) + \frac{1}{N} \sum_{j \neq k} \theta_j \sum_i f_j(X_i) f_k(X_i) +$$

$$+ \alpha_k \left[\frac{1}{N} \sum_i f_k^2(X_i) - 1 \right],$$

где $Y_i = Y_i^0 + \varepsilon_i$; Y_i^0 — истинное неизвестное значение функции Y в точке X и ε^2 соответствует дисперсии наблюдения.

Таким образом, при вычислении $E(\hat{\theta}_k - \theta_k)$ первый член по ε исчезает, поскольку $E(\varepsilon_i) = 0$, но остаются два других члена порядка $1/N$. Они исчезают, когда функции f ортогональны на дискретном множестве точек X_i.

Если же точки X_i распределены не регулярным образом, обычно не думают о том, чтобы использовать свойство ортогональности для непрерывной переменной. Однако если анализируется гистограмма с равными одинаковыми соседними ячейками, то иногда пытаются их использовать. Например, при изучении углового распределения, когда косинус полярного угла меняется от -1 до $+1$, используются полиномы Лежандра. Тогда величина смещения оценок $\hat{\theta}$ порядка $1/N$ может быть того же порядка величины, что и неопределенность оценок.

СПИСОК ЛИТЕРАТУРЫ

1. **Kendall M. G., Stuart A.** The advanced theory of statistics. V. 1. Charles Griffin and Co. Ltd, Lond., 1963. (См. также русский перевод: **Кендалл М., Стьюарт А.** Теория распределений. М., «Наука», 1966).
2. **Kendall M. G., Stuart A.** The advanced theory of statistics. V. 2. Charles Griffin and Co. Ltd, Lond., 1967. (См. также русский перевод: **Кендалл М., Стьюарт А.** Статистические выводы и связи. М., «Наука», 1973).
3. **Kendall M. C., Stuart A.** The advanced theory of statistics. V. 3 Charles Griffin and Co. Ltd, Lond., 1968.
4. **Kowalik J., Osborne M. R.** Methods for unconstrained optimization problems. American Elsevier, N. Y., 1968.
5. **Fletcher R.** Function minimization without evaluating derivatives, a review. — «Computer J.», 1967, v. 8, p. 33.
6. **Bayes T.** An essay towards solving a problem in the doctrine of chances. — «Biometrika», 1958, v. 45, p. 293 (Preprint of 1763).
7. **Feller W.** An introduction to probability theory and its applications. V. 1, N. Y., John Wiley and Sons, 1968. (См. также русский перевод 2 изд. 1957 г.: **Феллер В.** Введение в теорию вероятностей и ее приложения. М., «Мир», 1964).
8. **Evans D. A., Barkas W. H.** Exact treatment of search statistics. — «Nucl. Instrum. and Methods», 1967, v. 56, p. 289.
9. **Knop R. E.** Error in estimation of total events. — «Rev. Scient. Instrum.», 1970, v. 41, p. 1518.
10. **Крамер Г.** Математические методы статистики. Пер. с англ. М., Изд-во иностр. лит., 1948.
11. **Tortrat A.** Principles de statistique mathematique. Dunod, Paris, 1961.
12. **Fourgeaud C., Fuchs A.** Statistique. Dunod, Paris, 1967.
13. **Hammersley J. M., Handsomb D. C.** Monte Carlo Methods. Methuen's statistical monographs, Lond,. 1964.
14. **Feller W.** An introduction to probability theory and its applications. V. 2, N. Y., John Wiley and Sons, 1966.
15. **Hall P.** The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. — «Biometrika», 1927, v. 19, p. 240.
16. **Johnson I., Kotz S.** Discrete distributions. Houghton Mifflin Company, Boston, 1969.
17. **Beyer W. H.** Handbook of tables for probability and statistics. CRC Publications, Cleveland, 1966.
18. **McCusker C., Cairns I.** Evidence of quarks in air-shower cores. — «Phys. Rev., Lett.», 1969, v. 23, p. 658.
19. **Adair R., Kasha H.** Analysis of some results of quarks searches. — «Phys. Rev. Lett»., 1969, v. 23, p. 1355.
20. **Bartlett M. S.** The vector representation of a sample. — «Proc. Cambr. Phil. Soc., 1934, v. 30, p. 327.
21. **NBS,** Probability tables for the analysis of extreme value data. National Bureau of Standarts, Applied Mathematics Series, 22, Washington, 1953.
22. **Hahn G. J., Shapiro S. S.** Statistical models in engineering. N. Y., John Wiley and Sons, 1967.
23. **Abramovitz M., Stegun I.** Handbook of mathematical functions. Dover Publications Inc., N. Y., Reprinted from National Bureau of Standards, Applied Mathematics, Series, 55, Washington, 1964.
24. **Kullback.** Information theory and statistics. Dover Publications Inc., N. Y., 1968. (См. также русский перевод изд. 1959 г.: **Кульбак С.** Теория информации и статистика. М., «Наука», 1967).
25. **Shannon C. E.** A mathematical theory of communication. — «Bell System Techn. J.», 1948, v. 27, p. 379—423, 623—656.
26. **Shannon C. E., Weaver W.** The mathematical theory of communication. University of Illinois Press, Urbana, 1949.

27. **Wiener N.** Cybernetics. N. Y., John Wiley and Sons, 1948; **Винер Н.** Кибернетики Пер. с англ. М., «Советское радио», 1958.
28. **Lindley D. V.** Introduction to probability and statistics. V. 1. Cambridge University Press, Lond., 1956.
29. **Lindley D. V.** Introduction to probability and statistics. V. 2. Cambridge University Press, Lond., 1965.
30. **Wald A.** Statistical decision function. N. Y., John Wiley and Sons, 1950. (См. также русский перевод: **Вальд А.** Статистические решающие функции. В сб.: Позиционные игры. М., «Наука», 1967, с. 300.)
31. **Hudson D. J.** Lectures on elementary statistics and probability. CERN 63 — 29, 1963.
32. **Hudson D. J.** Statistics lectures II: Maximum likelihood and least squares theory. CERN 64—18, 1964. (См. также русский перевод: **Худсон Д.** Статистика для физиков. М., «Мир», 1967.)
33. **Jeffreys H.** Theory of probability. Oxford University Press, Oxford, 1948.
34. **Lindley D. V.** The use of prior probability in statistical inference and decisions. Proc. 4 th Berkeley Symp. on Mathematical statistics and probability. V. 1. Univ. of California Press, Berkeley and Los Angeles, 1961, p. 453.
35. **Haldane B. S., Smith S. M.** The sampling distribution of a maximum likelihood estimate. — «Biometrika», 1956, v. 43, p. 96.
36. **Rao C. R.** Efficient estimates and optimum inference procedures in large samples. — «J. Roy. Statist. Soc.», 1962, v. B24, p. 46.
37. **Solmitz F.** Analysis of experiments in particle physics. Report UCRL 11413, 1964.
38. **Quenouille M. H.** Notes on bias in estimation. — «Biometrika», 1956, v. 43, p. 353.
39. **Huber P. J.** Robust estimation of a location parameter. — «Ann. Math. Statist.», 1964, v. 35, p. 73.
40. **Huber P. J.** A robust version of the probability ratio test. — «Ann. Math. Statist.», 1965, v. 36, p. 1753.
41. **Tukey J. M.** The future of data analysis. — «Ann. Math. Statist.», 1962, v. 33, p. 1.
42. **Crow E. L., Siddiqui M. M.** Robust estimation of location. — «J. Amer. Stat. Ass.», 1967, v. 62, p. 353.
43. **Rice J. R., White J. S.** Norms for smoothing and estimations. — «Siam Rev.», 1964, v. 6, p. 243.
44. **Ekblom H., Enriksson S.** L_p criteria for the estimation of location parameters. — «Siam J. Appl. Math.», 1969, v. 17, p. 1130.
45. **Dalenius T.** The mode, a neglected parameter. — «J. Roy. Stat. Soc.», 1965, v. A.128, p. 110.
46. **Grenander U.** Some direct estimates of the mode. — «Ann. Math. Statist.», 1965, v. 36, p. 131.
47. **Lawley D. N.** A general method for approximating to the distribution of likelihood ratio criteria. — «Biometrika», 1956, v. 43, p. 295.
48. **Lehmann E. L.** Testing statistical hypothesis. N. Y., John Wiley and Sons, 1959. (См. также русский перевод: **Леман Э.** Проверка статистических гипотез. М., «Наука», 1966.)
49. **Hogg R. V., Craig A. T.** Sufficient statistics in elementary distribution theory. — «Sankhya», 1956, v. 17, p. 209.
50. **Cochran W. G.** The distribution of quadratic forms in a normal system with application to the analysis of covariance. — «Proc. Cambridge Philos. Soc.», 1934, v. 30, p. 178.
51. **Cox D. R.** Tests of separate families of hypotheses, Proc. 4th Berkeley Symp. on mathematical statistics and probability. V. 1. University of California Press, Berkaley and Los Angeles, 1961, p. 105.
52. **Wald A.** Theory and application of sequental probability ratio tests. N. Y., John Wiley and Sons. 1947.
53. **Cochran W. G.** The chi-squared test of goodness-of- fit. — «Ann. Math. Statist.», 1952, v. 23, p. 315.

54. **Cochran W. G.** Some methods for strengthening the common chi-squared tests. — «Biometrika», 1954, v. 10, p. 417.
55. **Mann H. B., Wald A.** On the choice of number of intervals in the application of the chi-squared test. — «Ann. Math. Statist.», 1942, v. 18, p. 50.
56. **Wilks S. S.** Mathematical Statistics. N. Y., John Wiley and Sons, 1962. (См. также русский перевод: **Уилкс С.** Математическая статистика. М., «Наука», 1967.)
57. **Smirnov N. V.** Sur la distribution de W^2. — «Compt. rend. Acad. Sci.», 1936, v. 202, p. 449.
58. **Anderson T. W., Darling D. A.** Asymptotic theory of certain goodness — of — fit criteria based on stochastic processes. — «Ann. Math. Statist.», 1952, v. 23, p. 193.
59. **Marchall A. W.** The small sample distribution of nW_n^2. — «Ann. Math. Statist.», 1958, v. 29, p. 307.
60. **Wallis W. A.** Compounding probabilities from independent significance tests. — «Econometrica», 1942, v. 10, p. 229.
61. **Fisher R. A.** Statistical methods for research workers. Oliver and Boyd, Edinburgh — Lond., 1958. (См. также русский перевод: **Фишер Р. А.** Статистические методы для исследователей. М., Госстатиздат, 1958.)

ДОПОЛНЕНИЕ

I. ПО ПОВОДУ ТРАКТОВКИ ОСНОВНЫХ ПРОБЛЕМ ТЕОРИИ ОЦЕНОК

А. А. Тяпкин

ПРЕДВАРИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Разнообразие задач, решаемых в научных исследованиях с привлечением методов математической статистики, делает необходимым более детальное рассмотрение принципиальных вопросов построения вероятностных заключений в теории оценок. Для правильного выбора статистических критериев и самостоятельного формулирования статистически обоснованных выводов в различных ситуациях, возникающих при анализе и планировании эксперимента, специалистам требуется достаточно глубокое понимание самой сути проблемы получения ответов на поставленные вопросы в виде вероятностных суждений.

Однако читатель, пожелавший изучить эту проблему, неминуемо столкнется с большими трудностями при разборе имеющихся в современной специальной литературе плохо согласующихся между собой высказываний по одним и тем же вопросам. Во введении авторы настоящей книги обращают внимание на отсутствие среди специалистов единой точки зрения на формулировку основных задач теории оценок и теории решений. По этой причине они предпочли параллельное изложение двух интерпретаций постановки одних и тех же задач математической статистики. Оба подхода в их изложении, к сожалению, противопоставлены друг другу в качестве одинаково общих и независимых трактовок любых задач теории ошибок. Проблему же существования двух несовместимых концепций авторы отнесли к области философии.

На самом же деле в той области задач, где действительно могут быть обоснованны оба подхода, эти концепции вполне совместимы в рамках строгих математических понятий теории вероятностей. Противоречивость высказываний по этим вопросам авторитетных специалистов доказывает лишь факт существования в литературе определенной неувязки в освещении этого круга вопросов. Эти разногласия специалистов вовсе не привели к постановке проблемы существования двух несовместимых интерпретаций задач теории оценок, для разрешения которой требовалось бы дальнейшее развитие математической статистики или даже выход за пределы теории вероятности в область философии. Такой проблемы, не разрешимой методами современной математической статистики, просто не существует. Что же касается некоторого разногласия специалистов

по общим вопросам формулировки задач теории оценок, то его возникновение обусловлено рядом причин, и в том числе исторически возникшим недоразумением с формулировкой основной задачи теории ошибок.

Далее мы покажем, что как недоразумения с разногласиями ложного характера, так и прямые ошибки прикладного значения чаще всего возникали от использования понятия вероятности для определенного типа событий без четкого определения рассматриваемого статистического коллектива. Приведем простейший пример возникновения ложного разногласия по этой причине. Пусть речь идет о надежности срабатывания какого-либо устройства однократного использования, т. е. о вероятностной характеристике срабатывания устройства, на данном экземпляре которого невозможно провести предварительные испытания. В этом случае можно представить себе ситуацию, когда один из сотрудников ОТК к данному изделию относит надежность α_1, определенную по серии контрольных испытаний для продукции соответствующего года, а другой характеризует надежность срабатывания того же устройства величиной $\alpha_2 < \alpha_1$, учитывая, что данное изделие изготовлено в последнем месяце года, продукция которого также на основании выборочных испытаний отличается меньшей величиной надежности. Ясно, что здесь нет какого-либо противоречия. Использование вероятностных характеристик для данного единичного изделия всегда подразумевает присоединение его к определенным статистическим коллективам, которые, вообще говоря, могут выбираться различным образом. Знаменатель дроби, которая асимптотически выражает вероятность события данного типа, равен полному числу испытаний, составляющих определенный статистический коллектив. Такая же ситуация возникает и в теории ошибок, когда на основании одного и того же результата измерений строятся оценки определяемой величины, относящиеся к различным статистическим коллективам повторных измерений. Недоразумения же на этой основе возникают только в тех случаях, когда, говоря о вероятности данного события, забывают строго определить статистический коллектив, в рамках которого ведется рассмотрение случайного процесса.

Для уяснения сущности статистического подхода теории оценок разберем вначале классическую формулировку задачи теории ошибок, учитывая обязательное требование конкретизации статистического коллектива (ансамбля), а затем, выделив из нее часть, допускающую строгое обоснование, покажем ее совместимость с современной трактовкой проблемы.

Прямые задачи, решаемые в теории вероятностей, состоят в предсказании вероятности реализации конкретных событий в определенном статистическом коллективе повторных испытаний. Например, при известном соотношении числа черных и белых шаров в урне предсказывается вероятность вытащить из урны определенное число

черных шаров при заданном числе испытаний*. Но в теории вероятностей может быть поставлена и обратная задача оценки соотношения числа черных и белых шаров в урне по известному случайному результату отдельного испытания. Такого рода обратные задачи и составляют особый раздел теории вероятностей и математической статистики — теории оценок. Исходную, основную задачу этой теории представляет тот случай, когда необходимо получение вероятностных характеристик о неслучайной величине. Впервые такая задача была поставлена в середине XVIII в. Даниилом Бернулли. Однако строгая и общая формулировка решения этой задачи была дана только в 30-х годах нашего столетия английским статистиком Рональдом Фишером [1] уже после того, как им в 1912 г. был предложен принцип максимального правдоподобия [2] и затем в 1925 г. создан на его основе общий метод получения статистических оценок параметров [3].

Основное затруднение вызывала принципиальная сторона проблемы получения вероятностных характеристик о величине, не являющейся случайной. В рамках теории вероятности легко формулируется задача статистического описания случайных результатов X измерений некоторой неизвестной постоянной величины θ_0. Но в обращенной задаче приходится, с одной стороны, исходить из определенного результата измерений X_i, что заставляет отвлечься от статистического коллектива истинно случайных величин X, с другой стороны, находить некоторую функцию плотности вероятности для характеристики неизвестной величины, которая в исходной задаче теории ошибок не является случайной величиной.

Конечно, указанная трудность вовсе не такого масштаба, чтобы она оказалась неразрешимой для великих математиков прошлого. Эта трудность послужила только причиной уклонения от поиска строгой формулировки исходной задачи теории ошибок в сторону решения несколько иной задачи, в которой измеряемая величина является случайной. По этой причине вопрос о формулировке основной задачи теории ошибок не числился среди нерешенных проблем теории вероятностей, а найденная недостаточно обоснованная формулировка к тому же не вступила в явное противоречие с практикой ее применения в теории ошибок.

КЛАССИЧЕСКАЯ ФОРМУЛИРОВКА ОСНОВНОЙ ЗАДАЧИ ТЕОРИИ ОШИБОК

Прямая задача в теории вероятностей может быть поставлена для более сложного варианта двух или нескольких последовательных случайных процессов. Пусть в прежнем примере с урной само соотношение числа черных и белых шаров является случайной ве-

* Заметим, что для рассматриваемой задачи статистический коллектив состоит из повторных испытаний, проводимых с одной и той же урной после того, как в ней восстановлено начальное число белых и черных шаров.

личиной. Это может быть связано с каким-либо процессом изменения числа шаров в урне по известному случайному закону или в более простом случае, из которого мы в дальнейшем будем исходить, обусловлено выбором самой урны из начальной генеральной совокупности, составленной из урн с различным соотношением числа черных и белых шаров. Предсказание вероятности получения определенного числа черных шаров в этой более сложной задаче будет относиться к статистическому коллективу повторных испытаний, проводимых каждый раз с новой урной, взятой из начальной совокупности с известным распределением соотношения числа черных и белых шаров.

Обращение этой более сложной задачи не встречает, однако, отмеченной выше трудности построения вероятностных суждений о неслучайной величине. Применительно к случаю обработки конкретного результата измерения X_i эта задача также легко формулируется для статистического подансамбля повторных измерений, давших в пределах некоторого интервала dX один и тот же результат X_i. При этом обязательно имеется в виду, что повторные измерения* каждый раз проводятся с новым объектом, взятым из генеральной совокупности с известным распределением случайной величины θ.

Вся эта задача может быть строго сформулирована как задача объединения имеющихся предварительных сведений о случайной величине θ в виде априорной плотности вероятности $p(\theta)$ с новыми статистическими сведениями, полученными в результате измерения, проведенного с отдельным объектом из генеральной совокупности. Решение этой задачи было дано английским математиком XVIII в. Бейесом в виде теоремы, согласно которой для случая непрерывного распределения (сложная гипотеза) функция плотности апостериорной вероятности равна

$$p(\theta \mid X_i) = \frac{p(\theta)\, p(X_i \mid \theta)}{\int p(\theta)\, p(X_i \mid \theta)\, d\theta}, \qquad (I.1)$$

где $p(X_i \mid \theta)$ — функция плотности условной вероятности получения результата X_i в измерениях с объектами, обладающими одним и тем же значением θ измеряемой случайной величины.

Из полного статистического ансамбля повторных измерений, проводимых каждый раз с новым объектом из генеральной совокупности, мы должны выделить статистический подансамбль Бейеса, включающий только те измерения, которые в пределах некоторого интервала dX дали один и тот же результат X_i. Лишь в этом специально отобранном статистическом ансамбле реализуется полученная апостериорная плотность вероятности $p(\theta \mid X_i)$, которая отличается от заданной априорной плотности вероятности $p(\theta)$ меньшей величиной дисперсии той же случайной θ. Это преобразование функции плотности вероятности случайной θ связано с отбором измерений, давших результат X_i, и с учетом известной функции

* Это могут быть отдельные серии из определенного числа измерений, которые мы здесь рассматриваем как одно измерение.

плотности вероятности $p(X/\theta)$, описывающей случайные флуктуации процесса измерения заданной величины θ.

Смысл полученной плотности апостериорной вероятности состоит непосредственно в определении вклада в статистический подансамбль с фиксированным результатом измерений X_i случаев, когда из начальной совокупности были выбраны объекты с истинными значениями измеряемой величины в пределах от $\theta - d\theta/2$ до $\theta + d\theta/2$. Полученное решение имеет прямое отношение и к постановке задачи теории ошибок. Конкретный результат X_i был получен в измерении, проведенном с объектом, обладающим вполне определенным, но не известным нам значением θ_j измеряемой величины. В получении оценки именно этого значения θ_j и состоит задача теории ошибок. Подлежащее статистической оценке значение θ_j принадлежит начальной совокупности значений с известной плотностью вероятности $p(\theta)$. Это обстоятельство означает, что еще до учета результата измерения мы располагаем знанием некоторой априорной оценки величины θ_j. С другой стороны, статистический подансамбль Бейеса составлен из совершенно равноправных с θ_j значений случайной величины θ, при измерении которых был получен тот же результат X_i. Следовательно, для оценки значения θ_j мы можем использовать более узкое распределение апостериорной вероятности в подансамбле Бейеса. Получаемая при этом апостериорная оценка

$$\hat{\theta} = \bar{\theta} = \int \theta p(\theta \mid X_i)\, d\theta$$

с дисперсией $D(\hat{\theta}) = D(\theta \mid X_i)$ объединяет в себе априорные статистические данные с экспериментальными данными.

Эта задача об объединении статистических сведений об измеряемой величине θ_j решается на основе теоремы Бейеса вполне строго, и современная математическая статистика ничего существенного к этому решению добавить не может [4]. Но непременным условием простейшего обоснования применения теоремы Бейеса является возможность случайных изменений измеряемой величины θ.

Бейесовский подход наталкивается на определенные затруднения лишь при попытках распространить его за пределы этого условия, когда либо ничего не известно об априорном распределении измеряемой случайной величины, либо, напротив, известно, что измеряемая величина заведомо не является случайной. В первом из этих случаев, когда отсутствуют какие-либо предварительные сведения об априорном распределении измеряемой величины, вообще не должна вставать решаемая теоремой Бейеса задача совместного учета априорных знаний и статистических данных, полученных в эксперименте. Здесь следовало бы ограничиться решением первой задачи теории ошибок, состоящей в построении вероятностных суждений об измеряемой величине только на основе анализа проведенного эксперимента. Но именно трудности формулировки этой начальной задачи и вынудили одних авторов прийти к ошибочному заключению о неразрешимости задачи без знания априорного распределения,

а других авторов вернуться к бейесовской постановке вопроса путем весьма искусственного введения некоторой произвольной функции априорного распределения величины $p(\theta)$. Ходжес и Леман [5] при выборе априорного распределения предложили использовать субъективные вероятности, отражающие степень веры экспериментатора в те или иные статистические предположения об измеряемой величине. Эта точка зрения на введение субъективных вероятностей отражена Идье и др. в настоящей книге (см. стр. 94). Видимо, это сравнительно недавно появившееся обобщение теоремы Бейеса на случай субъективных априорных вероятностей имели в виду авторы, назвав современным бейесовский подход, насчитывающий уже более 200 лет.

Отсылая читателей к приведенной в известном «Курсе теории вероятностей» Б. В. Гнеденко [6] справедливой критике попыток использовать аппарат теории вероятностей для субъективных категорий, отметим, что в данном случае теорема Бейеса используется с целью обойти задачу непосредственного выявления объективных статистических данных, следующих из эксперимента, и представить их с учетом влияния произвольных факторов субъективного характера*.

Гораздо более серьезного рассмотрения заслуживает другой тип обобщения теоремы Бейеса, положенный в основу классической формулировки исходной задачи теории ошибок. Речь идет об использовании теоремы Бейеса для построения вероятностных суждений об определяемой в эксперименте неслучайной величине. Но прежде чем обсуждать правомерность такой постановки вопроса, рассмотрим более простой пример. В рамках обычного применения теоремы Бейеса возможен предельный случай, когда дисперсия $D(\theta)$ случайной величины θ по априорному распределению $p(\theta)$ становится много больше дисперсии $D(\theta \mid X_i)$ той же величины по апостериорному распределению $p(\theta/X_i)$. Ясно, что в этом случае априорное распределение уже не вносит в окончательное распределение каких-либо уточняющих сведений о величине θ_j. Процедуру построения бейесовского подансамбля уже нельзя считать объединением двух типов статистических сведений о θ_j. Ограниченность области значений случайной величины θ в этом подансамбле целиком обусловлена отбором из равномерного распределения случаев измерения, давших один и тот же результат X_i.

Следовательно, при достаточно широком распределении случайной величины θ в исходном ансамбле, применив к его отдельным элементам поочередно процедуру многократных измерений, можно непосредственно выявить статистические характеристики погрешностей измерения, если отобрать затем только те случаи измерения величин θ, лежащих в интервале $d\theta$, которые дали результат X_i

* К этим явно несостоятельным попыткам обойти трудности формулировки исходной задачи теории ошибок за счет введения субъективных вероятностей не следует причислять развитый Кудо [7] метод учета частичной априорной информации.

в пределах интервала dX. При этих условиях ($D(\theta) \gg D(\theta \mid X_i)$ априорное распределение $p(\theta)$ уже не оказывает влияния на окончательное распределение, и исходный ансамбль с равномерным распределением случайной величины требуется только для того, чтобы на его материале проявились статистические свойства процесса измерения, непосредственным выражением которых и является функция плотности условной вероятности $p(\theta \mid X_i)$, реализующаяся в отобранном бейесовском подансамбле. В соотношении (1.1) происходит сокращение плотности априорной вероятности, и теорема Бейеса дает не зависящее от величины этой плотности преобразование заданного распределения условной вероятности $p(X \mid \theta)$ случайной величины X в функции плотности условной вероятности $p(\theta \mid X_i)$ уже для случайной θ. Такое преобразование условных вероятностей называют обращением распределения или преобразованием в обратную вероятность.

В правомерности понятия обратной вероятности неоднократно возникали сомнения (см., например, монографию Яноши [8]). Но аргументы о противоречивости этого понятия не относятся к рассмотренному случаю реализации обратной вероятности в бейесовском подансамбле, полученном при постоянной плотности априорной вероятности в исходном статистическом коллективе. Сомнения возникали по поводу обоснования распространения этого подхода на случай, когда определяемая величина θ_0 не является случайной. Это обобщение перехода к обратным вероятностям строилось на основании постулата Бейеса, согласно которому во всех случаях, когда не известно априорное распределение измеряемой величины θ (или в общем случае определяемого параметра), следует все значения θ считать априори равновероятными. Даже если измеряемая величина — не известная нам постоянная, согласно постулату Бейеса следует принимать $p(\theta) = \text{const}$. Этот постулат вместе со знаменитой теоремой Бейеса появился в его работе, опубликованной в 1763 г. вскоре после смерти автора. Лаплас придал большое значение постулату Бейеса, положив его в основу процедуры обращения вероятности и формулировки исходной задачи теории ошибок.

Как отмечает Худсон [9], впервые этот постулат подвергся серьезной критике в 1854 г. в работе Буля. Высказывалось также предположение, что сам Бейес не публиковал свою работу из-за сомнений в справедливости постулата [9].

В настоящее время многие считают логически несостоятельными как сам постулат Бейеса, согласно которому априорное распределение в виде $\delta(\theta - \theta_0)$-функции при неизвестном θ_0 заменяется $p(\theta) = \text{const}$, так и принятую Лапласом классическую формулировку основной задачи теории ошибок на основе перехода к обратным вероятностям, когда об измеряемой величине говорится как о случайной, с некоторой вероятностью попадающей в определенный интервал значений θ^*.

---

* См., например, книги В. Н. Романовского [10] и Худсона [9].

Все эти сомнения и критические замечания в адрес обоснования классической формулировки основной задачи теории ошибок действительно обоснованны. Применение обратных вероятностей на основе формального предположения о постоянстве априорного распределения без рассмотрения соответствующих статистических коллективов создает впечатление о крайней несостоятельности всего подхода к проблеме формулировки вероятностных суждений о величинах, определяемых в эксперименте. Однако эти недостатки обоснования перехода к обратным вероятностям в задаче об измерении постоянной величины θ_0 вполне можно избежать, если каждый этап введения вероятностного описания прослеживать на конкретном примере реализации в соответствующем статистическом ансамбле.

УТОЧНЕНИЕ ОБОСНОВАНИЯ КЛАССИЧЕСКОЙ ФОРМУЛИРОВКИ

Пусть при измерении неизвестной нам постоянной величины θ_0 получен результат X_i. Требуется найти статистическую оценку измеряемой величины θ_0, используя полученный результат и известную функцию плотности условной вероятности $p(X \mid \theta)$, характеризующую случайные погрешности процесса измерения при различных заданных значениях θ измеряемой величины.

Прежде всего уточним статистический коллектив, в рамках которого осуществляется вероятностное описание процесса измерения, заданное распределением $p(X \mid \theta)$. Этот этап рассмотрения относится к области прямых задач теории вероятности. Каждому значению параметра θ распределения условной вероятности соответствует своя совокупность повторных измерений величины θ, результаты которых распределены по закону $p(X \mid \theta)$. Взятые вместе эти коллективы повторных измерений различных величин составляют общий статистический ансамбль, в котором каждое значение случайной X_i представлено многими случаями, полученными уже при измерении различных величин θ. Можно представить себе, что заданные функции распределения $p(X \mid \theta)$ при различных значениях параметра θ были определены в таких калибровочных измерениях при известных значениях θ. Во всяком случае, задать функции распределения $p(X \mid \theta)$ для различных θ и означает представить рассмотренные выше генеральные совокупности результатов калибровочных измерений, проведенных с различными истинными значениями измеряемой величины.

Однако содержащуюся в этом полном ансамбле калибровочных измерений информацию о случайных флуктуациях процесса измерения можно выразить по-разному. Можно сказать, что в силу случайных отклонений при измерении заданной величины θ мы получаем совокупность в общем случае отличных от θ значений случайной величины X, определенным образом группирующихся около истинного значения измеряемой величины. Описание статистического распределения значений X в этой совокупности результатов калибро-

вочных измерений дается функцией распределения условий вероятности $p(X \mid \theta)$.

С другой стороны, можно сказать, что в силу случайных отклонений один и тот же результат X_i получается при калибровочных измерениях отличающихся истинных значений величины θ. Сама по себе величина θ может и не быть случайной, но полный ансамбль калибровочных измерений обязательно включает измерения с различными истинными значениями этой величины*. Количественные характеристики случайных отклонений, происходящих в самом процессе измерения, при такой формулировке могут быть специально выделены из массива полной генеральной совокупности результатов калибровочных измерений путем отбора только тех случаев, которые дали заданный результат X_i в пределах некоторого интервала dX. Но, как следует из рассмотренного в предыдущем разделе примера, частота событий получения данного результата X_i только тогда является независимой количественной характеристикой случайных отклонений собственно процесса измерений, когда в полный ансамбль измерений в равном количестве включены измерения с различными значениями величины θ. В противном случае количественный состав отобранного подансамбля значений X_i будет искажен выбором неравного объема серий измерений при различных значениях θ.

Таким образом, знание функции распределения $p(X \mid \theta)$ при различных значениях параметра θ тождественно заданию полного ансамбля результатов калибровочных измерений, из массива которого может всегда быть выделен подансамбль Бейеса, распределение величины θ в котором описывается обращенной функцией распределения:

$$p(\theta \mid X_i) = \frac{p(X_i \mid \theta)}{\int p(X_i \mid \theta)\, d\theta}. \qquad (1.2)$$

Следовательно, можно себе представить, что операция обращения распределения $p(X_i \mid \theta)$ или переход к так называемым обратным вероятностям происходит в рамках полного ансамбля калибровочных измерений и сводится к простой переформулировке статистических сведений о случайных погрешностях процесса измерения. В этом отношении формальный переход к распределению условной вероятности $p(\theta \mid X_i)$ пока никак не связан с какими-либо предположениями об истинном распределении величин θ в предстоящих измерениях, и постулат Бейеса может трактоваться как обязательное условие формирования полного ансамбля калибровочных измерений. Специальный характер производимой переформулировки сведений о случайных погрешностях процесса измерения состоит лишь в том, что эти данные представлены в виде условной вероятности применительно к получению конкретного результата X_i,

* Последнее утверждение эквивалентно заданию функции распределения $p(X \mid \theta)$ для различных значений параметра θ.

но в этом, конечно, не больше произвола, чем в записи условного распределения $p(X \mid \theta)$. Для уяснения смысла понятий обращенных вероятностей нам понадобилось выделить как отдельный этап калибровочные измерения.

Итак, мы можем себе представить, что на предварительном этапе калибровочных измерений выяснилось, что результат X_i получается с определенной частотой $\sim p(\theta \mid X_i)$, зависящей от истинного значения θ измеряемой величины. Теперь нам остается выяснить, в какой мере эти статистические данные, полученные при калибровочных измерениях известных величин θ, можно использовать для статистической оценки в реальном измерении неизвестной величины θ_0. Классическая формулировка задачи может быть представлена как утверждение, что с вероятностью $p(\theta \mid X_i)$ использованное в калибровочных измерениях конкретное значение θ совпадает с неизвестной постоянной θ_0. Тогда среднее значение $\overline{\theta} = \int \theta p(\theta \mid X_i)\, d\theta$ с дисперсией $D(\theta \mid X_i)$ может быть взято в качестве оценки неизвестной величины θ_0.

Означает ли это, что измеряемая величина θ_0 превращена в случайную? Здесь правильнее было бы сказать, что к данному случаю получения результата X_i при измерении неизвестной постоянной величины θ_0 присоединяется бейесовский подансамбль случаев получения того же результата X_i в измерениях различных истинных значений θ, взятых равновероятными. Так как для такого подансамбля имеется полное статистическое описание, то оно распространяется и на случай измерения θ_0, рассматриваемый теперь уже как случайная выборка из этой генеральной совокупности, сформированной бейесовской процедурой отбора событий.

Для понимания этого важнейшего этапа классического подхода необходимо будет разобраться в следующих двух вопросах:

во-первых, следует выяснить общий вопрос о правомерности включения в какой-либо статистический коллектив конкретного случая измерения неизвестной постоянной величины;

во-вторых, необходимо выяснить обоснованность использования для этой цели подансамбля, отобранного описанной выше процедурой Бейеса.

Против положительного ответа на первый вопрос обычно приводится следующий аргумент: неизвестная постоянная величина θ_0 не может с какой-либо вероятностью, отличной от 0 или 1, совпадать с конкретным значением переменной θ или находиться внутри конкретного интервала этой переменной; величина θ_0 может либо совпадать с конкретным значением θ, и тогда вероятность этого события равна 1, либо не совпадать с ним, и тогда вероятность совпадения равна 0. Но такая постановка вопроса требует ответа, выходящего за пределы теории вероятностей. Статистический же ответ на поставленный вопрос всегда подразумевает присоединение данного конкретного случая к какому-либо статистическому коллективу. Ведь и в примере о надежности срабатывания конкретного устройст-

ва, запускающего ракету, можно сказать, что данное устройство при включении либо срабатывает, и тогда вероятность взлета ракеты равна 1, либо устройство не сработает, и тогда вероятность того же события следует принять равной 0. Как мы уже объяснили выше, статистический ответ на этот вопрос вовсе не означает превращение срабатывания данного запускающего устройства в случайное событие. Такой ответ лишь предполагает рассмотрение данного конкретного устройства в качестве равноправного элемента некоторого множества, для которого могут быть установлены определенные статистические свойства. Соответствующие вероятностные свойства могут быть экспериментально подтверждены в рамках выбранного статистического коллектива. Как уже отмечалось, могут быть основания для вероятностных предсказаний в рамках различных статистических коллективов, и тогда встает вопрос о сравнении различных статистических оценок рассматриваемого события. Таким образом, в классической формулировке основной задачи следует обсуждать лишь обоснованность подключения случая измерения неизвестной постоянной величины к статистическому подансамблю измерений, отобранных в соответствии с процедурой Бейеса. Мы уже видели, что бейесовский подансамбль, построенный на массиве калибровочных измерений различных известных величин θ, есть формальный прием представления случайных погрешностей процесса измерения.

Непосредственной областью приложения представленной таким образом информации о свойствах измерительного процесса, конечно, является статистический ансамбль измерений, давших результат X_i в пределах некоторого интервала, при условии, что измерениям подвергались неизвестные истинные значения θ, представленные равновероятно в начальном континууме. В отличие от бейесовского подансамбля, составленного на массиве полного калибровочного опыта, мы теперь не знаем, какие именно случаи, давшие результат X_i, были получены при измерении заданного истинного значения θ в интервале $d\theta$. Но знание обращенной функции распределения позволяет нам предсказать, что их доля в общем количестве случаев получения результата X_i пропорциональна величине $p(\theta \mid X_i)\, d\theta$.

Теперь нетрудно увидеть и определенные основания для, казалось бы, искусственного включения в этот статистически описываемый подансамбль Бейеса и случая измерения неизвестной постоянной величины θ_0, давшей тот же результат X_i. Ведь любое другое измерение, входящее в этот подансамбль, относится также к какому-то неизвестному истинному значению θ_j измеряемой величины, и каждый экспериментатор, получивший результат X_i при измерении конкретного, но неизвестного ему истинного значения θ_j, решает задачу оценки именно этой величины. Равноправность всех этих случаев получения результата X_i дает основу для их объединения в такой общий ансамбль, статистические свойства которого зависели бы только от случайных погрешностей процесса измерения. Но этому последнему требованию, как мы выяснили раньше, удовлетворяет бейесовский подансамбль только в том случае, если он

отобран из измерений неизвестных, но равновероятных представленных значений θ.

Таким образом, как формальное преобразование (2.2) функцией $p(X|\theta)$ в обращенное распределение $p(\theta \mid X)$ интерпретируется на основе полного ансамбля калибровочных измерений при обязательном условии равновероятности представленных в нем случаев измерения различных значений θ, так и статистический коллектив практического приложения обращенного распределения $p(\theta \mid X)$ должен формироваться из измерений равновероятно представленных неизвестных значений θ.

Итак, классическая формулировка задачи, утверждая, что при фиксированном значении полученного результата измерений X_i истинное значение измеряемой величины θ_0 с вероятностью $p(\theta|X_i)\,d\theta$ совпадает с некоторым значением θ в пределах интервала $d\theta$, подразумевает реализацию указанной вероятности в специально составленном ансамбле повторных измерений равновероятно представленных величин θ. Само по себе это специальное формирование коллектива, в рамках которого становится возможно статистическое описание, не должно вызывать возражений. Если, например, из урны с известным соотношением числа белых и черных шаров извлечен черный шар, но у вас завязаны глаза и вы не можете с достоверностью установить его цвет, то вам придется мысленно вернуть шар в урну для того, чтобы иметь хотя бы статистический ответ, который оправдается только в серии повторных испытаний. Подобно этому и в классической постановке задачи неизвестную измеряемую величину приходится погружать в коллектив равновероятных значений, с тем чтобы получить возможность статистически интерпретировать результат измерений. Ничто не может помешать представить, что измеренный объект предварительно был извлечен из генеральной совокупности, в которой равновероятно представлены различные значения измеряемой величины. Не может вызвать сомнений и строгость даваемого в классической теории ошибок статистического решения для этого условного статистического коллектива. Весь вопрос только в том, насколько удобно такое решение. Ведь для проверки предсказанного распределения $p(\theta \mid X_i)$ потребуется практическое осуществление такого ансамбля, так как повторные измерения должны производиться с разными объектами этой совокупности.

Наиболее просто классический подход обосновывается в рассматриваемом случае, когда статистически оцениваемая величина совпадает с измеряемой или связана с ней линейным соотношением. Рассмотрение принципиальной стороны вопроса уже на примере этой простейшей задачи показывает возможность более строгого обоснования классической постановки вопроса и позволяет доказать отсутствие какой-либо логической несостоятельности в самом построении вероятностных суждений, принятом в классической теории ошибок. Поэтому сделанные в адрес классической постановки вопроса критические замечания многих авторов следует относить прежде

всего к неточностям объяснения принятой формулировки, к ошибочной трактовке постулата Бейеса. Что же касается общности классического подхода, то эта проблема требует особого рассмотрения.

Даже решение рассмотренной выше простейшей задачи оказывается ограниченным в том смысле, что остается применимым лишь к условному статистическому ансамблю, который может быть невыполним на практике. Если, например, практическая реализация оказывается возможной только в ансамблях, не удовлетворяющих этому требованию, то необходимо искать другое решение задачи. Однако полное решение задачи на основе теоремы Бейеса оказывается невозможным в тех случаях, когда неизвестна в действительности реализующаяся плотность априорной вероятности в начальном статистическом ансамбле. В этом случае можно было бы искать решение задачи в рамках статистического ансамбля повторных измерений, проведенных с одним и тем же объектом. Но, как мы отмечали выше, классическая теория ошибок уклонилась от поиска такой формулировки задачи.

С неоднозначностью возможного построения подансамбля Бейеса приходится встречаться в более сложных задачах, когда оцениваемая величина связана нелинейным соотношением с измеряемыми величинами. Худсон [9] и авторы настоящей книги (см. стр. 128) разбирают в связи с этим пример оценки дисперсии нормального распределения. Однако эта проблема устраняется весьма просто. Нужно лишь помнить, что если условное распределение $p_1(\theta \mid X_i)$ получено при $p(\theta) = \text{const}$, то оно применимо только к ансамблям с равновероятным представлением величины θ. Иначе говоря, простая замена переменных, например на $v = \theta^2$, в этом распределении $p_1(\sqrt{v}/X_i)\frac{1}{2\sqrt{v}}$ вовсе не делает его применимым к ансамблю, в котором $p(v) = \text{const}$. Приведенное авторами второе решение $p_2(v/X_i)$ действительно отличается от первого $p_1(\sqrt{v}/X_i)\frac{1}{2\sqrt{v}}$, но оно имеет свою область применения — статистический ансамбль с $p(v) = \text{const}$. Так что это типичный пример ложного расхождения. Важно, что функции $p_1(\theta/X_i)$ и $p_2(v/X_i)$ применительно к собственным ансамблям с $p(\theta) = \text{const}$ и $p(v) = \text{const}$ дают совпадающие статистические результаты.

Иногда неосуществимость ансамбля с $p(\theta) = \text{const}$ видят в невозможности нормировки априорного распределения. Однако это соображение не существенно. Дело в том, что на самом деле постоянство распределения достаточно иметь в ограниченной области значимых величин вероятности $p(x_i \mid \theta)$, характеризующей случайные погрешности измерений. Ограниченность сформулированной в классической теории ошибок задачи может заключаться в неосуществимости требуемого статистического ансамбля по условиям самого опыта. Нередко измеряемая величина является случайной с неизвестным априорным распределением. В этом случае задача статистического анализа для полной системы с учетом двух последовательных

случайных процессов, конечно, не решается. Однако совершенно неправы те, кто в связи с этим считает невозможным постановку вопроса об учете только случайных погрешностей, вносимых непосредственно процессом измерений*. Частичное решение этой задачи, конечно, сохраняет неопределенность, связанную с предшествующим случайным выбором измеряемой величины. Поэтому даваемое классической теорией решение приложимо только к рассмотренному выше условному статистическому ансамблю, в котором проявляются лишь случайные погрешности самого процесса измерения. В современной же постановке задачи решение оказывается приложимым и к системе с неизвестной априорной вероятностью.

Ограниченность классического решения задачи в полной мере выявляется в сравнении с постановкой задачи в современной теории оценок, отличающейся не только большой простотой и ясностью, но и универсальностью.

СОВРЕМЕННАЯ ФОРМУЛИРОВКА ОСНОВНОЙ ЗАДАЧИ ТЕОРИИ СТАТИСТИЧЕСКИХ ОЦЕНОК

Различные задачи построения вероятностных суждений на основании случайных выборок трудно уместить в рамках собственно теории ошибок. Сейчас они составляют большой раздел математической статистики — теорию статистических оценок. Современный метод формулировки таких задач использует понятие правдоподобия предположений, сделанных на основании случайных выборок. Правдоподобие есть вероятность оказаться справедливым предположительному высказыванию (гипотезе) об оцениваемой величине в тех случаях, когда количественные характеристики предположения находятся по определенному способу на основе получаемых значений случайных величин. При этом от выборки к выборке сохраняется лишь способ построения данного типа заключения, содержащиеся же в нем конкретные количественные характеристики оцениваемой величины подвержены, естественно, случайным изменениям. Именно по этой причине данного типа заключение оказывается то справедливым, то ошибочным даже в отношении величин, не являющихся случайными.

Соответственно этому при подсчете величины правдоподобия предположения все события должны разбиваться только на два класса — на случаи подтверждающие и случаи, опровергающие данное предположение.

---

* Такую точку зрения отстаивал академик С. Н. Бернштейн [11]. С принципиальным возражением против этого мнения выступил в 1942 г. академик А. Н. Колмогоров [12]. Позднее, однако, авторы [13] высказали ту же (см. [11]) ошибочную точку зрения при обсуждении задачи выбора из двух гипотез о возможных радиоактивных источниках известной активности на основании единичного измерения интенсивности. Ниже мы приводим пример решения аналогичной задачи на основе предсказания отношения доверительных вероятностей.

Формулировки статистических задач на основе понятия правдоподобия обладают большой простотой и общностью. Они применимы для различных задач оценок как случайных, так и не случайных величин*.

Переформулировка основной задачи теории ошибок в неявном виде содержалась уже в первых работах английских статистиков Стьюдента [14] (1908 г.) и Фишера [2] (1912 г.) по созданию строго обоснованных методов статистического анализа случайных выборок малого объема. В явном же виде новая трактовка проблемы была предложена Фишером значительно позднее, уже после того, как им было завершено обоснование метода максимального правдоподобия и подробно исследованы свойства оценок, даваемых этим методом, хотя по логике вещей новая постановка всей проблемы должна бы предшествовать соответствующему новому методу решения части этой проблемы.

Введенное Фишером понятие доверительной вероятности для предположительных высказываний (гипотез) и предложенное им одно из определений количества информации легли в основу общего теоретико-информационного подхода к проблемам математической статистики и в определенной мере подготовили почву для последующего развития термодинамического подхода к информационным процессам.

Сейчас эта новая трактовка задач теории ошибок широко используется в прикладной математике, большое место занимает она и в педагогической литературе по математической статистике. Однако признание новой точки зрения пришло далеко не сразу. Как отмечает сам Фишер, говоря о первом издании своей книги «Статистические методы для исследователей» (1925 г.), по-новому представленный вопрос «не обратил на себя внимание преподавательских кругов и не заставил их сменить порочную практику обучения студентов неправильным методам статистического анализа» [15]. Особенно задержались отклики советских специалистов на новое направление математической статистики. Не обошлось здесь и без досадных недоразумений. Известный авторитет в области теории вероятностей академик С. Н. Бернштейн выступил в 1941 г. против доверительных вероятностей Фишера и основанной на них новой трактовки задач теории ошибок [11]. Эта ошибочная оценка работ Фишера нашла, к сожалению, некоторое отражение и в более позднем издании его фундаментального курса «Теория вероятностей», в котором выборочный метод в статистике излагался без упоминания о методе максимального правдоподобия, а по поводу формулировки задачи без обращения к теореме Бейеса было сделано лишь следующее краткое замечание: «Это допущение, к сожалению, противоречит принципам теории вероятностей ... должны существовать и априорные законы

* Следует подчеркнуть, что речь идет об общности метода формулировки теории оценок, а не метода решения этих задач.

вероятностей для этих величин, и, вопреки желанию Фишера, необходимо учитывать формулу Бейеса...» [16].

Решающее значение для возникновения интереса к новейшим методам математической статистики имела статья академика А. Н. Колмогорова [12]. Широкому распространению новых методов среди советских специалистов способствовали выход в 1947 г. книги академика В. И. Романовского [10] и издание в 1948 г. под редакцией академика А. Н. Колмогорова фундаментальной монографии стокгольмского профессора Крамера [17].

Сейчас развитые английскими статистиками Стьюдентом и Фишером методы точного решения задач при выборках малого объема широко освещаются во всех монографиях и руководствах по статистике. В 1964 г. вышла книга Н. П. Клепикова и С. Н. Соколова, целиком посвященная методу максимального правдоподобия [18]. К сожалению, авторы этой книги обошли молчанием вопрос о современной формулировке задачи определения степени достоверности оценки гипотез на основе функции правдоподобия.

В настоящей книге Идье и др. подробно освещен теоретико-информационный подход. Поэтому мы ограничимся лишь кратким изложением основных особенностей современной постановки проблемы и некоторыми замечаниями, касающимися выбора статистических критериев.

В новой постановке задача оценки истинного значения θ_0 измеряемой величины формулируется в рамках статистического коллектива повторных измерений неизменной величины θ_0. В функции $p(X \mid \theta)$, описывающей распределение результатов измерения X при фиксированном значении измеряемой величины, параметр θ трактуется как предположительное значение оцениваемой величины θ_0. Условная вероятность $p(X_i \mid \theta)$ получения заданного результата X_i в серии повторных опытов, если бы измеряемая величина действительно равнялась θ, принимается за вероятность оказаться справедливым предположению, что принятое значение $\theta = \theta(X)$ в пределах некоторого интервала $d\theta$ совпадает с истинным значением измеряемой величины. Эта последняя величина вероятности рассматривается в том же статистическом ансамбле повторных измерений некоторой неизвестной величины θ_0 для предполагаемых значений $\theta(X)$, фиксированным образом связанных с получаемыми результатами X. Например, если выбрано значение $\theta_r = X + r$, то $p(X \mid \theta_r)\, dX$ будет вероятностью совпадения с θ_0 предполагаемого значения $\theta_r = X + r$ (при фиксированном значении r) в некотором интервале $d\theta = \varphi(dX)$, зависящем от выбранного интервала dX, т. е. проверку вероятности $p(X \mid \theta_r)\, dX$ предсказания совпадения с θ_0 при каждом измерении надо проводить для $\theta_r = X_i + r$, зависящего от полученного значения X_i случайной величины X. Вероятность получения такого X_r, которое дает совпадение $X_r + r$ с истинным значением θ_0, в действительности равна $p(X_r \mid \theta_0)\, dX$, что совпадает с принятой для оценки $\theta_r = X + r$ вероятностью $p(X \mid \theta_r)\, dX$.

Функцию распределения плотности рассматриваемой вероятности $p(X \mid \theta)\, dX$ принято называть функцией правдоподобия. $L(X \mid \theta) = p(X \mid \theta)$. Величина θ рассматривается как аргумент функции правдоподобия, которому в пространстве параметров соответствует переменная предположительных значений истинной величины θ_0. Поскольку функция правдоподобия пропорциональна доверительной вероятности совпасть выбранной величине θ с истинным значением θ_0, то наибольших совпадений обеспечивает оценка $\hat{\theta}$, соответствующая максимуму функции правдоподобия. Приняв для точечной оценки принцип максимума правдоподобия, можно, конечно, забыть о связи каждой точки функции правдоподобия с доверительной вероятностью гипотезы о совпадении соответствующего значения аргумента с истинной величиной θ_0 и интересоваться в дальнейшем только статистическим распределением оценки $\hat{\theta}$, которая при каждой выборке случайным образом меняет свою величину. Но статистическую интерпретацию допускает не только точка максимума, но и вся функция правдоподобия*. На практике по ней определяют отношение доверительных вероятностей (отношение правдоподобия) для гипотез с фиксированными значениями параметра θ и устанавливают величину доверительного интервала.

Конечно, в каждом отдельном случае при анализе выборки в пространстве параметров мы получаем определенную функцию правдоподобия. И любое конкретное высказывание об истинном значении параметра может быть либо только ложным, либо справедливым. Статистические предсказания правдоподобия гипотез строго выполняются в ансамбле повторных измерений. Поэтому статистические выводы, полученные при анализе конкретной выборки, надо приложить к мысленно представленной себе статистической совокупности таких же анализов других случайных выборок из общей генеральной совокупности. Следовательно, надо представить себе совокупность различных кривых правдоподобия, максимумы и другие параметры которых смещаются в соответствии с полученными в измерениях результатами. Случайные смещения будут происходить и для доверительного интервала, построенного по кривой правдоподобия. Поэтому такой интервал будет перекрывать истинное значение измеряемой величины в количестве случаев, соответствующем его достоверности.

Так просто решается проблема о построении вероятностных выводов о неслучайной величине θ_0, что вызывает естественное недоумение долгий путь к уяснению очевидной истины: статистические выводы, построенные на основе анализа случайной величины, случайны по своей природе и не требуют обязательного внесения элемента случайности в сами оцениваемые величины. К приведенной нами трактовке бейесовской условной вероятности $p(\theta \mid X_i)$, как ве-

* Это утверждение, однако, не относится к случаю дискретного распределения измеряемой величины и непрерывных возможных значений оцениваемого параметра.

роятности совпадения некоторого калибровочного значения θ с истинным значением θ_0 измеряемой величины, необходимо лишь добавить понимание того, что в этом предсказании проверке может быть подвергнута какая-либо определенная величина рассматриваемой вероятности $p(\theta \mid X_i)$, которая в силу случайности используемого результата измерения предполагает каждый раз выбор нового соответствующего калибровочного значения $\theta(X)$. Это и означало бы установить еще одну область статистического приложения той же обращенной функции распределения.

То, что новая постановка проблемы использует коллектив повторных измерений и соответствующих повторных анализов, обеспечивает ей большую универсальность. Дело не только в том, что в это рассмотрение стало возможным включить случай повторных измерений одной и той же постоянной величины. Это решение основной задачи теории ошибок оказывается все же условным в тех случаях, когда в силу физических причин возможны лишь однократные испытания. Универсальность формулировки на основе доверительных вероятностей состоит в том, что эти вероятности осуществляются в любых повторных измерениях, в том числе и в измерениях случайных величин, меняющихся по неизвестному закону. Высказывания о доверительном интервале просто не зависят от априорного распределения. Нетрудно рассчитать мыслимый проверочный опыт и убедиться, что интервал заданной достоверности α, построенный по функциям правдоподобия различных экспериментальных выборок X_j именно с вероятностью α, включает соответствующие истинные значения измеряемых величин θ_j вне зависимости от закона распределения этих величин.

Высказывания об отношении правдоподобия двух конкурирующих гипотез также могут быть сформулированы независимо от априорных вероятностей данных гипотез. Представим себе, что по результату отдельного выстрела (или серии выстрелов) необходимо дать статистический ответ на вопрос, в какую из двух близко расположенных мишеней целился стрелок*. В классической постановке на этот вопрос дается только условный ответ для случая равновероятного выбора мишени. В современной постановке задачи по данному результату X_i может быть определено отношение правдоподобия $l_X = L(X_i \mid \theta_1)/L(X_i \mid \theta_2)$ и указан подколлектив результатов X, для которого разделение гипотез с тем же коэффициентом отбора l_X и с той же примесью $1/(l_X + 1)$ конкурирующей гипотезы выполняется вне зависимости от соотношения вероятностей участия гипотез в испытаниях. В этот подансамбль результатов равного отношения правдоподобия кроме результатов X_i в пределах интервала dX войдут симметричные им результаты X_j, для которых отношение

* Эта задача совершенно аналогична задаче о выборе одного из известных радиоактивных источников по измерению интенсивности излучения. Авторы книги [13] сочли эту задачу не разрешимой методами математической статистики в случае неизвестной априорной вероятности участия в измерениях каждого источника.

правдоподобия дает тот же коэффициент отбора $l_X = L(X_j | \theta_2) \times$ $\times L(X_j/\theta_1)^{-1}$ в пользу другой гипотезы. Характерная особенность нового подхода, на которую мы уже обращали внимание, состоит в статистических предсказаниях для величин $\theta(X)$, связанных с результатом случайной выборки и не зафиксированных в абсолютных значениях пространства параметров θ. При обработке конкретного результата рассматривается вопрос о выборе, конечно, вполне определенной гипотезы, например θ_1. Но даваемая методом достоверность выбора относится не вообще к первой гипотезе, а к достоверности ее выбора при заданном результате измерения. При анализе следующего результата мы можем с той же высокой надежностью отдавать предпочтение другой альтернативе, утверждая, что стрелок на этот раз целился уже во вторую мишень. Важно, что относительное число угадываний меняющихся истинных положений мишени действительно оказывается равной указанной достоверности.

Обсуждая проблему решения так называемой простой гипотезы, когда истинные значения определяемого параметра могут принимать лишь определенные дискретные значения, следует отметить, что в научных исследованиях эта задача ставится несколько иначе, чем в области массового производства. Дело не только в том, что в науке задача обычно кончается определением доверительных вероятностей соответствующих гипотез и возможной ошибки сделанного выбора. Значительно реже ставится задача об окончательном выборе решения с учетом потерь от возможных ошибок, но есть и другое принципиальное отличие в постановке задачи. Имея определенную выборку, ученый должен по случайному результату именно этой выборки сделать предсказание доверительной вероятности гипотез. Ясно, что при повторных выборках будут получаться результаты, дающие различные возможности для разделения двух гипотез. Например, могут быть попадания точно в середину между двумя мишенями и предсказанные по таким выборкам доверительные вероятности гипотез окажутся одинаковыми. Будут, естественно, и другие группы выборок, в том числе и дающие для гипотез значительно отличающиеся вероятности. При определении статистического коллектива для применения доверительных вероятностей, предсказанных на основании результата своей единичной выборки, экспериментатор должен иметь в виду только ту группу результатов из общего статистического коллектива, которые дают одинаковое отношение правдоподобия. Именно к такому отобранному коллективу значений X_i и им симметричных X_j применима вычисленная достоверность гипотез.

В статистических задачах массового производства, напротив, интерес представляет не вопрос о разделении гипотез по какой-то отдельной группе результатов, обладающих одинаковой мощностью выделения правильной гипотезы, а средние данные о разделении гипотез по всем возможным выборкам. Поэтому вся область значений результатов измерений разбивается на две части, соответствующие

выбору различных гипотез. Возможность совершенно различных материальных потерь при ошибках (например, попадание брака в принятые изделия и попадание хороших изделий в отбракованные) заставляет разделять их на ошибки первого и второго рода. Но так как соответствующие разделы математической статистики развивались с учетом приложения к задачам массового производства, то вероятности таких ошибок были рассчитаны для попадания результатов в широкие области отбора гипотез. Это обстоятельство следует иметь в виду научным сотрудникам при решении своих задач. Только при обработке большого экспериментального материала надо пользоваться статистическими критериями, установленными для целых областей критических и допустимых значений проверочных статистик. При обсуждении гипотез, касающихся природы зарегистрированного в эксперименте редкого события, следует определять достоверность и возможную ошибку выбора данной гипотезы, исходя из доверительных вероятностей, определенных именно по экспериментальным данным рассматриваемого единичного события.

Такой подход вовсе не означает построения вероятностных суждений вне статистического коллектива. Найденная вероятность ошибочного выбора гипотезы реализуется в конкретном статистическом подансамбле аналогичных событий, занимающих одинаковое положение по отношению конкурирующих гипотез.

Другая важная особенность современного подхода, использующего понятие доверительных вероятностей, состоит в том, что в нем самые сложные задачи формулируются одинаково просто. Метод применяют к анализу измерений многих величин для оценки нескольких теоретических параметров, связанных в общем случае сложными нелинейными соотношениями с измеряемыми величинами. Нелинейность этих уравнений приводит к многозначности оценок, зависящей от экспериментальных погрешностей. В связи с этим встает вопрос о доверительных вероятностях соответствующих решений, о возможности отбрасывания некоторых из них. Так как при нормальном законе погрешностей уравнение правдоподобия приводит к условию минимизации χ^2-функционала, то различным решениям соответствуют определенные значения этого функционала. Поскольку эти величины подчиняются χ^2-распределению, то допустимость решения рассматривается по χ^2-критерию. Обычно число измеряемых величин значительно превосходит число определяемых параметров, что приводит к большому числу степеней свободы рассматриваемого χ^2-распределения. Действительно, малая вероятность по χ^2-распределению говорит о необъяснимом погрешностями измерений большом расхождении экспериментальных данных с теоретическими значениями для рассматриваемого решения. Однако в практике статистического анализа не только использовался χ^2-критерий для отбрасывания заведомо ложных решений, но и неоднократно само χ^2-распределение необоснованно использовалось для определения доверительной вероятности оставленных решений [19]. На самом же деле последняя задача должна решаться на основе функции правдо-

подобия [19, 20]. Ошибка здесь той же самой природы, как и в том случае, когда вместо требуемого в задаче определения вероятности ухода броуновской частицы в конкретную точку, отстоящую от начальной на расстоянии R, вычисляют вероятность ухода ее в шаровой слой данного радиуса. Только в случае статистического анализа экспериментальных данных при большом числе степеней свободы χ^2-распределения в этой модели речь должна идти о шаровом слое в многомерном пространстве, что лишь усугубляет ошибку.

Другой пример неправомерного использования χ^2-распределения относится к определению допустимых областей оценок параметров [21]. Применив этот подход, авторы работы [22, 23] получили в несколько раз более широкие области допустимых значений определяемых фаз протон-протонного рассеяния, чем дает метод по вторым производным от логарифма функции правдоподобия.

Эти примеры показывают, как важно правильно выбрать статистический критерий, соответствующий поставленному вопросу.

СПИСОК ЛИТЕРАТУРЫ

1. **Fischer R. A.** Inverse probability. — «Proc. Cambr. Phil. Soc.,» 1930, v. 26, p. 528; The concepts of inverse probability and fiducial probability referring to unknown parameters. — «Proc. Roy. Soc.», 1933, v. 139, p. 343; The Design of Experiment, Edinburgh, Oliver and Boyd, 1935.
2. **Fisher R. A.** On an absolute criterion for fitting frequency curves. — «Messeger of Math.», 1912, v. 41, p. 155.
3. **Fisher R. A.** Theory of statistical estimation. — «Proc. Cambr. Phil. Soc.», 1925, v. 22, p. 700.
4. **Закс Ш.** Теория статистических выводов. Пер. с англ. М., «Мир», 1975, с. 28.
5. **Hodges J. L., Lehmann E. L.** The use of previous experience in reaching statistical decisions. — «Ann. Math. Statistics», 1952, v. 23, p. 396.
6. **Гнеденко Б. В.** Курс теории вероятностей. М., ГИТТЛ, 1954, с. 19.
7. **Kudo H.** On partial prior information and the property of parametric sufficiency. — In. Proc. Fifth Berkeley Symp. Math. Statist. Prob. V. 1, 1967, p. 251.
8. **Яноши Л.** Теория и практика обработки результатов измерений. Пер. с англ. М., «Мир», 1965, с. 185.
9. **Худсон Д.** Статистика для физиков. Пер. с англ. М., «Мир», 1967, с. 143.
10. **Романовский В. И.** Основные задачи теории ошибок. М. — Л., ГИТТЛ, 1947, § 2, с. 12.
11. **Бернштейн С. Н.** О «достоверных» вероятностях Фишера. — «Изв. АН СССР. Сер. матем.», 1941, т. 5, с. 85.
12. **Колмогоров А. Н.** Определение центра рассеивания и меры точности по ограниченному числу наблюдений. — «Изв. АН СССР. Сер. матем.», 1942, т. 6, с. 3.
13. **Гольданский В. И., Куценко А. В., Подгорецкий М. И.** Статистика отсчетов при регистрации ядерных частиц. М., Физматгиз, 1959, с. 65.
14. **Student.** The Probable Error of a Mean. — «Biometrika», 1908, v. 6, p. 1.
15. **Фишер Р. А.** Статистические методы для исследователей. Пер. с англ. М., Госстатиздат, 1958, с. 8.
16. **Бернштейн С. Н.** Теория вероятностей. М. — Л., Гостехиздат, 1946, с. 306.
17. **Крамер Г.** Математические методы статистики. Пер. с англ. М., Изд-во иностр. лит., 1948.

18. **Клепиков Н. П., Соколов С. Н.** Анализ и планирование экспериментов методом максимума правдоподобия. М., «Наука», 1964.
19. **Тяпкин А. А.** Об ошибочности использования χ^2-критерия в фазовом анализе для отбора статистически допустимых многозначных решений. Препринт ОИЯИ Е-2353, Дубна, 1965.
20. **Определение** вероятности канала ядерной реакции. —«Ядерная физика», 1967, т. 6, с. 90. Авт.: Мороз В. И., Никитин А. В., Троян Ю. А., Шахбазян Б. А.
21. **Тяпкин А. А.** К вопросу об определении допустимых областей фаз в фазовом анализе по методу «оврагов». Препринт ОИЯИ Д-642, Дубна, 1965.
22. **Фазовый** анализ *pp*-рассеяния при 95, 150 и 310 *Мэв*. Препринт ОИЯИ Д-598, Дубна, 1960. Авт.: Гельфанд И. М., Грашин А. Ф., Иванова Л. Н., Померанчук И. Я., Смородинский Я. А.
3. **Фазовый** анализ *pp*-рассеяния, 95 *Мэв*. —«Журн. эксперим. и теор. физ.», 1961, т. 40, с. 1106. Авт.: Боровиков В. А., Гельфанд И. М., Грашин А. Ф., Померанчук И. Я.

II. ВЫЧИТАНИЕ ФОНА ПРИ ОЦЕНКЕ ПАРАМЕТРОВ МЕТОДОМ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

В. С. Курбатов, А. А. Тяпкин

ДВА ВАРИАНТА ОПРЕДЕЛЕНИЯ ОЦЕНОК ПАРАМЕТРОВ МЕТОДАМИ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Наиболее общим методом оценки параметров теоретических функций по экспериментальным данным является метод максимального правдоподобия (м. п.), предложенный Фишером [1, 2]. Этот метод, как известно, обеспечивает получение максимально возможной точности определения параметров [3]. Типичным для ядерной физики и физики высоких энергий является случай, когда экспериментальные данные в виде определенных зарегистрированных событий, представляющих собой случайную выборку из некоторой генеральной совокупности, используются для определения оценок параметров теоретических выражений дифференциальных сечений наблюдаемых процессов.

Пусть N — объем выборки (число зарегистрированных событий), а $\mathbf{X}$ — совокупность измеренных для каждого события координат $X_1, X_2, \ldots, X_N$. Будем считать, что существует правильная теория, которая описывает плотность вероятности распределения событий по измеренным координатам в виде функций $f(\mathbf{X}, \theta)$, где θ — совокупность определяемых параметров $\theta_1, \theta_2, \ldots, \theta_m$. С дифференциальным сечением соответствующего процесса эта функция связана соотношением

$$f(\mathbf{X}; \theta) = \frac{d\sigma/d\mathbf{X}}{\int\limits_D (d\sigma/d\mathbf{X})\, d\mathbf{X}},$$

где D — область изменений измеряемых координат.

Тогда, согласно принципу максимального правдоподобия, оценками параметров θ служат такие значения $\hat{\theta}$, которые обращают в минимум следующий функционал:

$$\Phi = -\sum_{i=1}^{N} \ln f(\mathbf{X}_i; \theta), \qquad (II.1)$$

где $\mathbf{X}_i$ — значения координат $\mathbf{X}$ у i-го события.

К проблеме оценки параметров можно подойти несколько иначе, анализируя гистограмму распределения наблюдаемых событий. Для удобства предположим, что каждое событие характеризуется лишь одной измеряемой координатой, которая может изменяться

в некоторых пределах $a \leqslant \mathbf{X} \leqslant b$. Построим гистограмму распределения наблюдаемых событий по координате X с шагом h:

$$h = (b - a)/k,$$

где k — количество ячеек гистограммы. Для каждой ячейки гистограммы может быть рассчитана ожидаемая вероятность попадания координаты $\mathbf{X}$ в ячейку с номером j:

$$p_j(\theta) = \int_{\mathbf{x}_j - h/2}^{\mathbf{x}_j + h/2} f(\mathbf{X};\, \theta)\, d\mathbf{X}. \qquad \text{(II.2)}$$

Если в результате эксперимента в 1-ю ячейку гистограммы попало n_1 событий, во вторую n_2, в j-ю n_j ($\sum_{j=1}^{k} n_j = N$), то вероятность получить именно такое распределение по числу событий гистограммы следует закону мультиномиального распределения

$$p(n_1, n_2, \ldots, n_k) = \frac{N!}{n_1! \ldots n_k!} \prod_{j=1}^{k} [p_j(\theta)]^{n_j}. \qquad \text{(II.3)}$$

Дальше можно применять стандартную технику принципа максимального правдоподобия, т. е. отыскивать минимум такого функционала:

$$\Phi(\theta) = -\sum_{j=1}^{k} n_j \ln p_j(\theta). \qquad \text{(II.4)}$$

В предположении о нормальности распределения каждой из n_j (это справедливо при больших n_j) метод м. п. приводит к обычному методу наименьших квадратов, т. е. к минимизации функционала:

$$\Phi(\theta) = \sum_{j=1}^{k} \frac{[n_j - N p_j(\theta)]^2}{n_j}.$$

Кроме того, параметры можно оценивать методом минимума χ^2-функционала:

$$\chi^2(\theta) = \sum_{j=1}^{k} \frac{[n_j - N p_j(\theta)]^2}{N p_j(\theta)}.$$

Совершенно очевидно, что при ширине интервала h, значительно превышающей точность измерения координат событий, гистограммный метод представления экспериментальных данных приводит к потере полученной в эксперименте информации о координатах событий, которая соответственно сказывается и на результатах найденных оценок параметров θ. С другой стороны, при уменьшении h должны, естественно, уменьшаться эти потери первоначально имею-

щейся информации о координатах событий и метод определения оценок параметров из анализа гистограммного представления экспериментальных данных должен приближаться к рассмотренному выше методу непосредственного анализа полученной совокупности событий*.

Полностью тождественные результаты эти методы определения оценок параметров дадут не только при ширине h интервала ячейки, равной точности измерения координат σ_X, но и при $h > \sigma_X$, если в силу ограниченности объема выборки N разбиение событий по ячейкам гистограммы приведет к тому, что в каждой ячейке окажется не больше одного события. Рассмотрим для этого случая приведенный функционал (II.4), при минимизации которого находятся оценки искомых параметров. Так как

$$p_j(\theta) \cong f(X_j;\ \theta)\, h$$

и из суммы $\sum_{l}^{k}$ выпадают члены, для которых $n_j = 0$, то функционал (II.4) принимает вид:

$$\Phi(\theta) = -\sum_{i=1}^{N} \ln f(X_i;\ \theta) - N \ln h, \qquad \text{(II.5)}$$

где индекс i означает порядковый номер только тех ячеек гистограммы, в которые попало по одному зарегистрированному событию. Полученный функционал отличается от функционала (II.1) лишь последним членом, независящим от параметров θ, и тем, что в нем величины X_i обозначают не координаты зарегистрированных событий, а центры ячеек гистограммы, в которые попали эти события. Но это различие совершенно несущественно, так как оно не отражается на точности оценок, а также и на получаемых конкретных значениях оценок параметров, если можно пренебречь изменением функции $f(X;\ \theta)$ в пределах ширины интервала отдельной ячейки. Это значит, что дальнейшее уменьшение ширины интервала ячеек гистограммы уже не приведет к повышению точности определения оценок параметров, хотя гистограммное представление данных по-прежнему еще не соответствует точности измерения координат зарегистрированных событий. Совершенно очевидно, что высокая точность измерений координат случайных событий может сказаться на точности определения оценок теоретических параметров только при достаточно большом объеме статистической выборки событий.

---

* В литературе можно встретить постановку вопроса об оптимальной ширине ячейки гистограммы. Так, например, в книге Яноши [4, с. 279] рассмотрен вопрос об определении оптимальной ширины ячейки гистограммы с точки зрения χ^2-критерия. Однако, как отмечено в той же работе (с. 284 — 285), величина интервала гистограммы, определенная из условия получения достаточной статистической точности в каждом интервале, оптимальна лишь в рамках выбранного χ^2-критерия, а не общего принципа максимального правдоподобия.

Совпадение функционалов (II.5) и (II.1) означает, что рассмотренный метод определения оценок искомых параметров из анализа гистограммы распределения зарегистрированных событий в предельном случае переходит в оптимальный вариант оценки параметров по методу м. п., когда экспериментальный материал анализируется непосредственно без потери полученной в эксперименте информации.

При сопоставлении этих двух методов анализа экспериментальных данных мы не учитывали погрешности измерения самих координат событий, считая, что основную неопределенность вносят флуктуации в числе зарегистрированных событий. Конечно, при достаточно большой статистике метод гистограмм еще до выполнения условия попадания в каждый интервал ячейки не больше одного события достигнет предельно возможной точности, если ширина интервала h станет равной погрешности измерения соответствующей координаты.

Анализ экспериментальных данных, представленных в виде гистограмм, в связи с практическими удобствами использования метода наименьших квадратов широко применяется и в тех случаях, когда такое представление данных приводит к определенной потере полученной в эксперименте информации.

Однако при современных средствах вычислительной техники имеются все основания применять оптимальный метод оценки теоретических параметров из непосредственного анализа экспериментального материала без внесения огрублений при разбиении его по отдельным ячейкам гистограммы. Для практической реализации этого метода необходимо, однако, решить задачу учета информации, получаемой в фоновых измерениях.

УЧЕТ ФОНОВЫХ СОБЫТИЙ ПРИ ОЦЕНКЕ ПАРАМЕТРОВ

На практике анализировать в действительности приходится экспериментальные данные в виде зарегистрированных событий, обусловленных исследуемым эффектом и фоновыми причинами, а также экспериментальные данные, полученные отдельно в фоновом эксперименте. В гистограммном методе представления экспериментальных данных учет фоновых измерений сводится, естественно, к простому вычитанию числа фоновых событий*, зарегистрированных в соответствующих интервалах переменных, из числа событий, полученных в основном эксперименте при исследовании эффекта в присутствии фона. Найденная таким образом разностная гистограмма распределения числа событий используется затем для оценки параметров, определяющих теоретическое описание исследуемого эффекта. И лишь при определении погрешностей полученных оценок параметров учитывается, что дисперсия использованных при анализе

* При этом имеется в виду, что результаты полученные непосредственно в фоновом эксперименте по продолжительности эксперимента и интенсивности первичного пучка частиц, приведены в результате соответствующей перенормировки к условиям основного эксперимента.

разностей $n_j^{э+ф} - n_j^{ф}$ равна сумме числа событий, зарегистрированных в основном и фоновом экспериментах, $n_j^{э+ф} + n_j$.

Рассмотрим теперь задачу учета фоновых измерений в случае определения параметров методом м. п. из анализа, в котором используется вся полученная в эксперименте информация о координатах зарегистрированных событий.

События, зарегистрированные в основном эксперименте, представляют собой случайную выборку $M_{э+ф}$ событий с координатами $X_1, X_2, X_j, \ldots, X_{N_{э+ф}}$ из генеральной совокупности событий, характеризуемой некоторой функцией плотности вероятности распределения событий по переменным X:

$$f_{э+ф}(X) = (1-\beta) f_э(\mathbf{X}; \theta) + \beta f_ф(\mathbf{X}), \qquad (II.6)$$

где $f_э(\mathbf{X}; \theta)$ и $f_ф(\mathbf{X})$ — плотности вероятности распределения событий соответственно исследуемого эффекта и фонового происхождения, а β — истинное значение относительной доли фоновых событий.

Аналогично этому события, зарегистрированные в фоновом эксперименте, представляют собой случайную выборку $M_ф$ событий с координатами $X_1, X_2, \ldots, X_{N_ф}$ из генеральной совокупности, характеризуемой плотностью распределения фоновых событий $f_ф(\mathbf{X})$.

Анализируя методом м. п. данные только выборки $M_{э+ф}$ без учета результатов фонового эксперимента, получим кривую регрессии $\hat{f}_{э+ф}(\mathbf{X})$, являющуюся наилучшей оценкой ф. п. в. распределения событий генеральной совокупности $f_{э+ф}(\mathbf{X})$. В свою очередь, анализ данных фонового распределения позволяет определить некоторую кривую регрессии $\hat{f}_ф(\mathbf{X})$, являющейся оценкой истинной функции плотности распределения фоновых событий. В силу сходимости оценок, даваемых методом м. п., получаемые линии регрессии при неограниченном возрастании объемов выборок стремятся к истинным функциям распределения

$$\hat{f}_{э+ф}(\mathbf{X}) \to f_{э+ф}(\mathbf{X}) \quad \text{и} \quad \hat{f}_ф(\mathbf{X}) \to f_ф(\mathbf{X})$$
$$N_{э+ф} \to \infty \qquad N_ф \to \infty$$

При этом $N_ф/N_{э+ф} \to \beta$, а получаемые в основном и фоновом эксперименте плотности распределения событий стремятся соответственно к истинным плотностям распределения событий $N_{э+ф} f_{э+ф}(\mathbf{X})$ и $N_ф f_ф(\mathbf{X})$. Это значит, что разность экспериментально наблюдаемых плотностей распределения событий $N_{э+ф}\hat{f}_{э+ф}(\mathbf{X}) - N_э\hat{f}_э(\mathbf{X})$ стремится в предельном случае $N_{э+ф} \to \infty$ и $N_ф \to \infty$ в силу соотношения (II.6) к истинному распределению плотности событий исследуемого эффекта $N_э f_э(\mathbf{X}, \theta)$, эффективную оценку которого нам необходимо найти для реальных выборок ограниченного объема. Конечно, оценкой искомого распределения при ограниченном объеме случайных выборок $M_{э+ф}$ и $M_ф$ может быть функция распре-

деления $(N_{э+ф} - N_ф) f_э(\mathbf{X}, \hat{\theta})$, найденная из анализа разности $N_{э+ф} \hat{f}_{э+ф}(\mathbf{X}) - N_ф \hat{f}_ф(\mathbf{X})$. Однако предварительное получение кривых регрессий $\hat{f}_{э+ф}(\mathbf{X})$ и $\hat{f}_э(\mathbf{X})$ из отдельных анализов выборок $M_{э+ф}$ и $M_ф$ также связано с некоторой потерей первоначально имеющейся экспериментальной информации из-за огрублений, вносимых при выборе теоретической модели для описания этих функций. Кроме того, такой вариант анализа излишне осложнен как необходимостью выбора гипотезы теоретического описания фона, так и самим проведением предварительного анализа в целях определения вспомогательных кривых регрессии. Поэтому данный метод учета результатов фоновых измерений следует применять только в том случае, когда имеются надежные дополнительные сведения о функции распределения фоновых событий. Так, например, иногда имеются теоретические основания принять наиболее простую гипотезу о равномерности распределения фоновых событий. Такой пример учета фоновых событий при использовании метода м. п. рассмотрен в работе [5].

В настоящей работе предлагается наиболее общий метод вычитания фона при определении оценок параметров методом м. п. с учетом индивидуальных координат зарегистрированных событий. Этот метод не связан с принятием каких-либо гипотез о функции распределения фоновых событий и подобно обычному методу вычитания фона при гистограммном представлении экспериментальных данных учитывает только информацию, полученную об этой функции распределения в специальном фоновом эксперименте.

Представим себе, что у нас появилась возможность установить, какие именно события в выборке $M_{э+ф}$ являются фоновыми. Тогда, отбросив эти ложные события, мы могли бы представить вероятность получения событий, относящихся только к исследуемому эффекту, в следующем виде:

$$p_э = W(N_{э+ф} - N'_ф) \prod_{j=1}^{N_{э+ф}-N'_ф} f(\mathbf{X}_j;\ \theta)\, d\mathbf{X}_j, \qquad (II.7)$$

где $N'_ф$ — общее число фоновых событий, зарегистрированных в основном эксперименте; $W(N_{э+ф} - N'_ф)$ — вероятность получения $(N_{э+ф} - N'_ф)$ событий эффекта.

Требование определения оценок параметров θ из условия получения максимума вероятности $p_э$ и приводит к минимизации функционала (II.1) в случае отсутствия фона в измерениях*.

Но вместо простого отбрасывания известных фоновых событий можно было бы осуществить своеобразное «вычитание» их из вы-

---

* Заметим, что взятый с обратным знаком $\ln p_э$ отличается от приведенного функционала (II.1) на $-\ln W(N_{э+ф} - N'_ф)$, не зависящую от параметров $\boldsymbol{\theta}$, определяющих нормированную плотность распределения событий по координатам $f(\mathbf{X}; \boldsymbol{\theta})$. Этот член функционала правдоподобия необходимо учитывать при определении полного сечения исследуемого процесса.

ражения вероятности получения всей выборки $M_{э+ф}$

$$p'_{э+ф} = W(N_{э+ф}) \prod_{j=1}^{N_{э+ф}} f(\mathbf{X}_j;\ \theta)\, d\mathbf{X}_j, \qquad \text{(II.8)}$$

при составлении которого фоновые события были ошибочно отнесены к исследуемому эффекту. Действительно, мы придем к точному выражению (II.7), если ложно составленную вероятность $p'_{э+ф}$ умножить на

$$\gamma_0 = \frac{W(N_{э+ф} - N'_ф)}{W(N_{э+ф})} \frac{1}{\prod_{k}^{N'_ф} f_k(\mathbf{X}_k;\ \theta)\, d\mathbf{X}_k},$$

включив в произведение $\prod_{k=1}^{N'_ф} f(\mathbf{X}_k;\ \theta)\, d\mathbf{X}_k$ вероятность получения за счет исследуемого эффекта только событий, совпадающих с известными фоновыми событиями выборки $M_{э+ф}$.

Ту же процедуру «вычитания» следует применить и в том случае, когда ложные фоновые события в выборке $M_{э+ф}$ не известны, а статистическая оценка γ множителя γ_0 получена на основе специально поставленного фонового эксперимента, т. е. вместо множителя γ_0 мы воспользуемся множителем

$$\gamma = \frac{W(N_{э+ф} - N_ф)}{W(N_{э+ф})} \frac{1}{\prod_{i}^{N_ф} f_i(\mathbf{X}_i;\ \theta)\, d\mathbf{X}_i},$$

отношение которого к γ_0 в среднем равно единице.

После умножения $p'_{э+ф}$ на γ вместо $p_э$ получим оценку этой величины:

$$\hat{p}_э = W(N_{э+ф} - N_ф) \frac{\prod_{j=1}^{N_{э+ф}} f_j(\mathbf{X}_j;\ \theta)\, d\mathbf{X}_j}{\prod_{i}^{N_ф} f_i(\mathbf{X}_i;\ \theta)\, d\mathbf{X}_i}. \qquad \text{(II.9)}$$

Соответствующий этой оценке вероятности получения событий эффекта в выборке $M_{э+ф}$ функционал правдоподобия с точностью до члена $-\ln(N_{э+ф} - N_э)$ имеет вид:

$$\Phi = -\left[\sum_{j=1}^{N_{э+ф}} \ln f(\mathbf{X}_j;\ \theta) - \sum_{i=1}^{N_ф} \ln f(\mathbf{X}_i;\ \theta)\right]. \qquad \text{(II.10)}$$

Из условия минимизации этого функционала и должны находиться оценки параметров θ в случае, когда единственные сведения о вероятности присутствия ложных событий в результате основного

эксперимента получены в специально поставленном фоновом эксперименте, т. е. соответствующие оценки $\hat{\theta}$ ($\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_m$) параметров должны определяться из уравнений

$$\frac{\partial \Phi}{\partial \theta} = -\left[\sum_{j=1}^{N_{э+ф}} \frac{\partial \ln f(\mathbf{X}_j;\ \theta)}{\partial \theta} - \sum_{i=1}^{N_ф} \frac{\partial \ln f(\mathbf{X}_i;\ \theta)}{\partial \theta}\right] = 0. \qquad \text{(II.11)}$$

Предложенный метод учета результатов фонового эксперимента совершенно аналогичен вычитанию фоновых случаев в гистограммном представлении экспериментальных данных. В частности, исходный функционал правдоподобия, используемый при гистограммном представлении данных, в предельном случае, когда в каждую ячейку гистограммы попадает не больше одного события, переходит в полученный нами функционал (II.10). Действительно, разностная гистограмма представлена в каждом интервале переменных $\mathbf{X}$ числом событий, зарегистрированных в основном эксперименте $(n_{э+ф})_j$, за вычетом числа событий $(n_ф)_j$, зарегистрированных в соответствующем интервале в фоновом эксперименте. Следовательно, функционал (II.4), при минимизации которого определяются оценки параметров, в случае разностной гистограммы принимает вид:

$$\Phi(\theta) = -\sum_{j=1}^{k} [(n_{э+ф})_j - (n_ф)_j] \ln p_j(\theta) =$$

$$= \left[\sum_{j=1}^{k} (n_{э+ф})_j \ln p_j(\theta) - \sum_{j=1}^{k} (n_ф)_j \ln p_j(\theta)\right], \qquad \text{(II.12)}$$

где

$$p_j(\theta) = \int_{X_{j-h/2}}^{X_{j+h/2}} f_э(\mathbf{X};\ \theta)\, d\mathbf{X}. \qquad \text{(II.13)}$$

При достаточно малой величине интервала ячейки соотношение (II.13) может быть заменено

$$p_j(\theta) = f_э(\mathbf{X}_j;\ \theta)\, h. \qquad \text{(II.14)}$$

Рассмотрим функционал (II.12) в случае выбора настолько малого интервала h, что при заданной статистике все $(n_{э+ф})_j$ и $(n_ф)_j$ становятся равными либо 0, либо 1. Тогда

$$\sum_{j=1}^{k} (n_{э+ф})_j \ln h f_э(\mathbf{X}_j;\ \theta) = \sum_{i=1}^{N_{э+ф}} \ln h f_э(\mathbf{X}_i;\ \theta) =$$

$$= N_{э+ф} \ln h + \sum_{i=1}^{N_{э+ф}} \ln f_э(\mathbf{X}_i;\ \theta)$$

где суммирование идет только по тем ячейкам гистограммы, для которых $(n_{э+ф}) = 1$. Соответственно,

$$\sum_{j=1}^{k} (n_ф)_j \ln h f_э (X_j; \theta) = N_ф \ln h + \sum_{i=1}^{N_ф} \ln f (X_i; \theta).$$

Следовательно, функционал (II.12) принимает вид

$$\Phi(\theta) = -\left[\sum_{i=1}^{N_{э+ф}} \ln f_э (\mathbf{X}_i; \theta) - \sum_{i=1}^{N_ф} \ln f_э (\mathbf{X}_i; \theta)\right] - (N_{э+ф} - N_ф) \ln h, \quad \text{(II.15)}$$

который с точностью до члена $(N_{э+ф} - N_ф) \ln h$, независящего от параметра θ, совпадает с функционалом (II.10) для алгебраической совокупности событий выборки $M_{э+ф}$ и $M_ф$. А это и значит, что непосредственный анализ методом м. п. составленной таким образом алгебраической совокупности событий позволяет определить с учетом результатов фонового эксперимента оценки искомых параметров теоретического описания исследуемого физического процесса, полностью используя полученную в эксперименте информацию о координатах X зарегистрированных событий.

Следует отметить, что предложенный метод учета результатов фонового эксперимента отличается большой простотой, так как он сводится практически к обычному непосредственному анализу методом максимального правдоподобия совокупности событий, полученных в основном и фоновом экспериментах. Конкретно рассмотренный вариант данного метода учета фона исходил из полной тождественности условий фонового и основного экспериментов. В случае различия длительностей $T_{э+ф}$ и $T_ф$ этих экспериментов соответствующая перенормировка приводит к замене последнего члена функционала (II.10) на

$$\frac{T_{э+ф}}{T_ф} \sum_{i=1}^{N_ф} \ln f_э (\mathbf{X}_i; \theta).$$

ОПРЕДЕЛЕНИЕ ПОГРЕШНОСТЕЙ ОЦЕНОК ПАРАМЕТРОВ

Определим погрешности оценок параметров, найденных из уравнений (II.11). Для этой цели в уравнениях (II.11) первую сумму условно представим в виде отдельных сумм по событиям эффекта и фоновым событиям:

$$-\sum_{j=1}^{N_э} \frac{\partial \ln f (\mathbf{X}_j; \hat{\theta})}{\partial \theta} - \left[\sum_{k=1}^{N'_ф} \frac{\partial \ln f (\mathbf{X}_k; \hat{\theta})}{\partial \theta} - \sum_{i=1}^{N_ф} \frac{\partial \ln f_i (\mathbf{X}_i; \hat{\theta})}{\partial \theta}\right] = 0. \quad \text{(II.16)}$$

Равенство нулю первой суммы определяет, согласно (II.1), оценки параметров $\hat{\theta}_0$ в случае отсутствия фона в измерениях. Из-за несовпадения сумм, относящихся к фоновым событиям, зарегистрированных в двух отдельных экспозициях, получаемые из уравнений (II.16) оценки $\hat{\theta}$ параметров отличаются от оценок параметров $\hat{\theta}_0$, соответствующих бесфоновому эксперименту. Обозначив $\Delta\hat{\theta}$ разность $\hat{\theta} - \hat{\theta}_0$, произведем замену в первом члене уравнения (II.16) величины $\hat{\theta}$ на $\hat{\theta}_0 + \Delta\hat{\theta}$. Разлагая в ряд Тейлора и пренебрегая членами второго и более высокого порядка относительно разности $\Delta\hat{\theta}$, находим

$$\Delta\hat{\theta} \sum_{j=1}^{N_э} \frac{\partial^2 \ln f(\mathbf{X}_j; \hat{\theta}_0)}{\partial\theta^2} = -\sum_{k=1}^{N'_ф} \frac{\partial \ln f(\mathbf{X}_k; \hat{\theta})}{\partial\theta} + \sum_{i=1}^{N_ф} \frac{\partial \ln f(\mathbf{X}_i; \hat{\theta})}{\partial\theta}. \quad \text{(II.17)}$$

Так как θ_0 являются решениями уравнений $\sum_{j}^{N_э} \frac{\partial \ln f(\mathbf{X}_j; \theta_0)}{\partial\theta} = 0$, соответствующими функционалу правдоподобия при бесфоновых измерениях $\Phi_0 = -\sum_{i=1}^{N_э} \ln f(X_j; \theta)$, то взятые с обратным знаком суммы вторых производных от $\ln f(X_j, \theta_0)$, равные $\frac{\partial^2\Phi_0}{\partial\theta^2}$, представляют собой обратные величины дисперсий $D_0^{-1}(\theta_0)$ оценок $\hat{\theta}_0$ параметров, получаемых при отсутствии фона в измерениях.

Следовательно, отклонения оценок $\hat{\theta}$ параметров от $\hat{\theta}_0$ равны

$$\Delta\hat{\theta} = D_0(\hat{\theta}_0) \left[\sum_{k=1}^{N'_ф} \frac{\partial \ln f(\mathbf{X}_k; \hat{\theta})}{\partial\theta} - \sum_{i=1}^{N_ф} \frac{\partial \ln f(\mathbf{X}_i; \hat{\theta})}{\partial\theta} \right]. \quad \text{(II.18)}$$

Представим себе, что из неограниченного числа пар повторных независимых выборок $M_{э+ф}$ и $M_ф$ отобрана такая статистическая совокупность выборок, которая характеризуется совпадением в пределах погрешностей измерений событий, относящихся к исследуемому эффекту. Усредним в таком статистическом коллективе случайную величину разности:

$$E(\Delta\hat{\theta}) = D_0(\hat{\theta}_0)(\bar{N}'_ф - \bar{N}_ф) E\left(\frac{\partial \ln f(\mathbf{X}; \hat{\theta})}{\partial\theta}\right). \quad \text{(II.19)}$$

Так как $\bar{N}'_ф = N_ф$, то математическое ожидание $E(\Delta\hat{\theta}) = 0$. Следовательно, определяемые из уравнения (II.11) компоненты $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \ldots, \hat{\theta}_m)$ случайного вектора $\hat{\theta}$ являются несмещенными оценками величин компонент $(\hat{\theta}_{01}, \hat{\theta}_{02}, \ldots, \hat{\theta}_{0m})$ вектора θ_0, представляющими собой решения неизвестного нам уравнения

$$\sum_{j}^{N_э} \frac{\partial \ln f(\mathbf{X}_j; \hat{\theta}_0)}{\partial\theta} = 0, \quad \text{(II.20)}$$

в котором из выборки $M_{э+ф}$ учитываются только события исследуемого эффекта. Это также значит, что использованный нами множитель γ преобразования соотношения (II.8) в соотношении (II.9) в среднем равен истинному значению коэффициента γ_0, обеспечивающему преобразование соотношения (II.8) в соотношение (II.7).

Возведя в квадрат разность $\Delta\hat{\theta}$, определяемую соотношением (II.18), и усредняя по тому же статистическому коллективу, находим дисперсию:

$$D(\Delta\hat{\theta}_0) = E(\hat{\theta}-\hat{\theta}_0)^2 = D^2(\hat{\theta}_0) E\left[\left(\sum_{k=1}^{N'_\Phi} \frac{\partial \ln f(\mathbf{X}_k;\ \hat{\theta})}{\partial\theta}\right)^2 + \right.$$

$$\left. + \left(\sum_{i=1}^{N_\Phi} \frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right)^2 - 2\sum_{k=1}^{N'_\Phi} \frac{\partial \ln f(\mathbf{X}_k;\ \hat{\theta})}{\partial\theta} \sum_{i=1}^{N_\Phi} \frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right] =$$

$$= E\left[\sum_{k=1}^{N'_\Phi}\sum_{k'=1}^{N'_\Phi} \frac{\partial \ln f(\mathbf{X}_k;\ \hat{\theta})}{\partial\theta} \frac{\partial \ln f(\mathbf{X}_{k'};\ \hat{\theta})}{\partial\theta} + \sum_{i=1}^{N_\Phi}\sum_{i'=1}^{N_\Phi} \frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta} \times \right.$$

$$\left. \times \frac{\partial \ln f(\mathbf{X}_{i'};\ \hat{\theta})}{\partial\theta} - 2\sum_{k=1}^{N'_\Phi}\sum_{i=1}^{N_\Phi} \frac{\partial \ln f(\mathbf{X}_k;\ \hat{\theta})}{\partial\theta} \frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right] D^2(\hat{\theta}_0) =$$

$$= 2D^2(\hat{\theta}_0)\left\{\overline{N}_\Phi E\left[\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)^2\right] + \left[(\overline{N}_\Phi)^2 - \overline{N}_\Phi\right] \times \right.$$

$$\left. \times \left[E\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)\right]^2 - \overline{N}_\Phi^2\left[E\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)\right]^2\right\} =$$

$$= 2D^2(\hat{\theta}_0)\,\overline{N}_\Phi\left\{E\left[\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)^2\right] - \left[E\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)\right]^2\right\}. \quad \text{(II.21)}$$

Здесь

$$D(\hat{\theta}_0) = E(\hat{\theta}_0-\theta)^2 = \frac{1}{E\left(\frac{\partial^2 \Phi_0}{\partial\theta^2}\right)}.$$

Но нам не известны математические ожидания величин в этом соотношении, определяющем искомую дисперсию $D(\Delta\hat{\theta})$ случайной величины $\Delta\hat{\theta}$. Мы можем определить лишь оценки этих точных сред-

них значений на основе имеющихся результатов фонового эксперимента.

Так $\hat{E}(N_\Phi) = N_\Phi$:

$$\hat{E}\left[\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)^2\right] = \frac{1}{N_\Phi}\sum_{i=1}^{N_\Phi}\left(\frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right)^2;$$

$$\hat{E}\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta} = \frac{1}{N_\Phi}\sum_{i=1}^{N_\Phi}\frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta} \qquad \text{(II.22)}$$

и

$$\hat{D}(\hat{\theta}_0) = \frac{1}{\hat{E}\left(\frac{\partial^2 \Phi_0}{\partial\theta^2}\right)} = \left[\sum_{j=1}^{N_{\text{э}+\Phi}}\frac{\partial^2 \ln f(\mathbf{X}_j;\ \hat{\theta})}{\partial\theta^2} - \sum_{i=1}^{N_\Phi}\frac{\partial^2 \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta^2}\right]^{-1}.$$

Заменяя в соотношении (II.21) точные средние значения величин на их оценки (II.22), получим окончательное выражение для оценки дисперсии $D(\Delta\hat{\theta})$, характеризующей среднеквадратичное отклонение случайной величины $\hat{\theta}$ от $\hat{\theta}_0$:

$$\hat{D}(\Delta\hat{\theta}) = 2\hat{D}^2(\hat{\theta}_0)\left[\sum_{i=1}^{N_\Phi}\left(\frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right)^2 - \frac{1}{N_\Phi}\left(\sum_{i=1}^{N_\Phi}\frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right)^2\right]. \qquad \text{(II.23)}$$

Оценка полной дисперсии $D(\hat{\theta})$, характеризующей среднеквадратичное отклонение случайной величины $\hat{\theta}$ от истинного значения параметра $\hat{\theta}$, будет выражаться суммой $\hat{D}(\hat{\theta}_0) + \hat{D}(\Delta\hat{\theta})$:

$$D(\hat{\theta}) = \hat{E}(\hat{\theta}-\theta)^2 = \hat{D}(\hat{\theta}_0) + 2\hat{D}^2(\hat{\theta}_0)\left[\sum_{i=1}^{N_\Phi}\left(\frac{\partial \ln f(\mathbf{X};\ \hat{\theta})}{\partial\theta}\right)^2 - \frac{1}{N_\Phi}\left(\sum_{i=1}^{N_\Phi}\frac{\partial \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta}\right)^2\right], \qquad \text{(II.24)}$$

где оценка дисперсии параметра, определяемого в бесфоновом эксперименте, согласно (II.22), равна

$$\hat{D}(\hat{\theta}_0) = -\sum_{j=1}^{N_{\text{э}+\Phi}}\left[\frac{\partial^2 \ln f(\mathbf{X}_j;\ \hat{\theta})}{\partial\theta} - \sum_{i=1}^{N_\Phi}\frac{\partial^2 \ln f(\mathbf{X}_i;\ \hat{\theta})}{\partial\theta^2}\right]^{-1}.$$

Полученное соотношение (II.24) представляет собой условное выражение оценок диагональных членов дисперсионной матрицы $D_{\hat{\theta}_k,\,\hat{\theta}_l}$ при статистической независимости определяемых оценок.

Для получения членов дисперсионной матрицы в общем случае необходимо в уравнениях (II.11) и (II.16) провести дифференцирование по одному из параметров θ_k, а в отношении (II.17) ввести разность $\Delta\hat{\theta}_l$, относящуюся к другому параметру θ_l. Проводя расчеты, аналогичные (II.21), и переходя от точных значений средних к их оценкам (II.22), получаем общее выражение оценок членов дисперсионной матрицы для предложенного метода учета фоновых измерений:

$$\hat{D}_{\hat{\theta_k},\,\hat{\theta_l}} = E\left[(\hat{\theta}_k-\theta_k)(\hat{\theta}_l-\theta_l)\right] = [W^{-1}]_{kl} + 2\,[W^{-1}HW^{-1}]_{kl}.$$

Здесь

$$[W]_{kl} = -\left[\sum_{j=1}^{N_{э+ф}} \frac{\partial^2 \ln(\mathbf{X}_j;\,\hat{\theta})}{\partial\theta_k\,\partial\theta_l} - \sum_{i=1}^{N_ф} \frac{\partial^2 \ln(\mathbf{X}_i;\,\hat{\theta})}{\partial\theta_k\,\partial\theta_l}\right];$$

$$[H]_{kl} = \sum_{i=1}^{N_ф} \frac{\partial \ln f(\mathbf{X}_i;\,\hat{\theta})}{\partial\theta_k}\,\frac{\partial \ln f(\mathbf{X}_i;\,\hat{\theta})}{\partial\theta_l} -$$

$$-\frac{1}{N_ф}\left(\sum_{i=1}^{N_ф}\frac{\partial \ln f(\mathbf{X}_i;\,\hat{\theta})}{\partial\theta_k}\right)\left(\sum_{j=1}^{N_ф}\frac{\partial \ln f(\mathbf{X}_j;\,\hat{\theta})}{\partial\theta_l}\right). \qquad \text{(II.25)}$$

В случае вычитания фона при представлении данных в виде гистограмм оценка матрицы ошибок имеет подобную форму:

$$\hat{D}_{\hat{\theta}_k,\,\hat{\theta}_l} = \hat{E}\left[(\hat{\theta_k}-\theta_k)(\hat{\theta}_l-\theta_l)\right] = \{[W^{-1}]_{kl} + 2\,[W^{-1}HW^{-1}]_{kl}\},$$

$$[W]_{kl} = -\sum_{j=1}^{k}\left(n_j^{э+ф}-n_j^{ф'}\right)\frac{\partial^2 \ln p_i(\hat{\theta})}{\partial\theta_k\,\partial\theta_l};$$

$$[H]_{kl} = \sum_{i=1}^{k} n_i^{ф'}\,\frac{\partial \ln p_i(\hat{\theta})}{\partial\theta_k}\,\frac{\partial \ln p_i(\hat{\theta})}{\partial\theta_l} -$$

$$-\frac{1}{N_{ф'}}\sum_{i=1}^{k} n_i^{ф'}\,\frac{\partial \ln p_i(\hat{\theta})}{\partial\theta_k}\sum_{j=1}^{k} n_j^{ф'}\,\frac{\partial \ln p_j(\hat{\theta})}{\partial\theta_l},$$

где k — число ячеек гистограммы; $n_j^{э+ф}$ — число случаев, зарегистрированных в j-й ячейке гистограммы в эксперименте эффект + +фон; $n_j^{ф'}$ — то же самое, что и $n_j^{э+ф}$, только в фоновом эксперименте; $N_{ф'}$ — общее число событий в фоновом эксперименте; $p_i(\theta)$ — вероятность попадания события эффекта в i-ю ячейку гистограммы.

СПИСОК ЛИТЕРАТУРЫ

1. **Fisher R. A.** On an absolute criterion for fitting frequency curves. — «Messenger of Math.», 1912, v. 41, p. 155.
2. **Фишер Р. А.** Статистические методы для исследователей. Пер. с англ. М., Госстатиздат. 1958.
3. **Крамер Г.** Математические методы статистики. Пер. с англ. М., Изд-во иностр. лит. 1948.
4. **Яноши Л.** Теория и практика обработки результатов измерений. Пер. с англ. М., «Мир», 1965.
5. **Соколов С. Н., Уточкин Б. А.** Критерий согласия и требования к прибору в методике наложения проекции траектории на трек. Препринт ИФВЭ, СВМ/СЭФ 69-27. Серпухов, 1969.

III. ПОИСК МАКСИМУМА ФУНКЦИИ ПРАВДОПОДОБИЯ МЕТОДОМ ЛИНЕАРИЗАЦИИ

И. Н. Силин

При обработке экспериментальных данных часто требуется искать максимум функции правдоподобия. Правда, технически обычно этого добиваются минимизацией отрицательного логарифма функции правдоподобия, и мы будем поступать так же.

Для эффективной минимизации функции $s(\theta)$ со сложным рельефом нужно, чтобы алгоритм минимизации использовал информацию о профиле ее поверхности. Многие алгоритмы накапливают такую информацию в процессе поиска минимума [1, 2]. Для гладких функций большую информацию дают вторые производные $s(\theta)$ по варьируемым параметрам $\theta\ \{\theta_1, \ldots, \theta_m\}$. Но вычисление всех вторых производных аналитически громоздко, а численное занимает много машинного времени. Полноценно использовать вторые производные к тому же непросто, если мы не находимся в непосредственной окрестности минимума [3].

При обработке экспериментальных данных по методу максимума правдоподобия и его частотному случаю — методу наименьших квадратов минимизируемые выражения имеют специфический вид:

$$s = -\sum_{j=1}^{n} \ln p(\mathbf{X}, \theta) \qquad \text{(III.1)}$$

и

$$s = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{f_j(\mathbf{X}_j, \theta) - F_j}{\sigma_j}\right)^2, \qquad \text{(III.2)}$$

где $p(\mathbf{X}, \theta)$, подгоняемое под набор измерений $X_1, \ldots, X_n$ распределение вероятности, и $f_j(\mathbf{X}_j, \theta)$, подгоняемые под измерения $F_1 \pm \sigma_1, \ldots, F_n \pm \sigma_n$ в точках $X_1, \ldots, X_n$, — функции с искомыми параметрами* θ.

Пренебрегая нелинейностью подгоняемых функций $p(\theta)$ и $f_j(\theta)$, можно получить приближенные вторые производные s и использо-

* Более сложный случай с учетом корреляции F_j дает

$$s = \frac{1}{2}\sum_{j,\,j'=1}^{n}(f_j(X_j, \theta) - F_j)\,W_{jj'}\,(f_{j'}(X_{j'}, \theta) - F_{j'}).$$

вать их для итерационного поиска минимума совместно с первыми производными s.

Действительно:

$$Z_{ik}=\frac{\partial^2 s}{\partial\theta_i\,\partial\theta_k}=\sum_{j=1}^{n}\frac{1}{p_j^2}\frac{\partial p_j}{\partial\theta_i}\frac{\partial p_j}{\partial\theta_k}-\sum_{j=1}^{n}\frac{\partial^2 p_j}{\partial\theta_i\,\partial\theta_k}\frac{1}{p_j} \qquad \text{(III.3)}$$

и

$$Z_{ik}=\frac{\partial^2 s}{\partial\theta_i\,\partial\theta_k}=\sum_{j=1}^{n}\frac{1}{\sigma_j^2}\frac{\partial f_j}{\partial\theta_i}\frac{\partial f_j}{\partial\theta_k}+\sum_{j=1}^{n}\frac{(f_j-F_j)}{\sigma_j^2}\frac{\partial^2 f_j}{\partial\theta_i\,\partial\theta_k}. \qquad \text{(III.4)}$$

Пренебрегая вторыми членами в правых частях (III.3) и (III.4), получаем выражения, для вычисления которых требуются лишь первые производные f и p, что значительно проще как для аналитического, так и для численного дифференцирования.

Для случая метода наименьших квадратов [формулы (III.2) и (III.4)] этот прием используется во многих алгоритмах [4—6]: в методе линеаризации [7—11], также и для случая метода максимума правдоподобия [формулы (III.1) и (III.3)].

Метод линеаризации был разработан в Дубне С. Н. Соколовым и И. Н. Силиным и реализован И. Н. Силиным вначале на машинах «Стрела» и М-20 в виде стандартных подпрограмм в машинных кодах, а затем и на языках АЛГОЛ и ФОРТРАН (программа FUMILI). В дальнейшем он был значительно усовершенствован. В настоящее время метод линеаризации широко используется при обработке экспериментальных данных.

При разработке алгоритма ставилась цель по возможности ускорить обработку экспериментальных данных для сложных $f_j(\mathbf{X}, \boldsymbol{\theta})$ и $p_j(\mathbf{X}, \boldsymbol{\theta})$, требующих большого машинного времени на вычисление. От других алгоритмов подобного рода алгоритм FUMILI отличается прежде всего способом ограничения величины шага в процессе выполнения итераций.

Рассмотрим для ясности общий путь вычислений в FUMILI на примере метода наименьших квадратов. Пусть мы имеем n экспериментальных значений $F_1 \pm \sigma_1, \ldots, F_n \pm \sigma_n$, под которые мы хотим подогнать по методу наименьших квадратов гладкие теоретические функции $f_1(X_1, \boldsymbol{\theta}), \ldots, f_n(X_n, \boldsymbol{\theta})$, где $\boldsymbol{\theta}\{\theta_1, \ldots, \theta_m\}$ — подгоняемые параметры, $X_j\{X_{j,1}, \ldots, X_{j,r}\}$ — координаты экспериментальных точек. Аналитические выражения $f_j(X_j, \theta)$ в частном (довольно распространенном) случае могут быть одинаковыми при всех j, но могут быть и разными.

Нам нужно найти минимум выражения (III.2). В принципе функция $s(\boldsymbol{\theta})$ может иметь много локальных минимумов. Лишь в редких случаях (в частности, при линейных $f_j(\boldsymbol{\theta})$) может быть доказано, что минимум единственный. Мы будем искать какой-нибудь локальный минимум.

Прежде всего необходимо выбрать начальное приближение $\boldsymbol{\theta}^{(0)}\{\theta_1^{(0)}, \ldots, \theta_m^{(0)}\}$. Если решение нам примерно известно, то можно его

использовать в качестве $\theta^{(0)}$. В противном случае нужно взять какое-либо разумное начальное приближение, либо выбрать его случайным образом, используя датчик случайных чисел в области не бессмысленных значений θ.

Вспомним, что в минимуме

$$\partial s/\partial\theta_i = 0. \tag{III.5}$$

Вычисляем при $\theta = \theta^{(0)}$

$$s = \frac{1}{2}\sum_{j=1}^{n}\left(\frac{f_j(\mathbf{X}_j,\ \theta) - F_j}{\sigma_j}\right)^2; \tag{III.6}$$

$$\frac{\partial s}{\partial\theta_i} = \sum_{j=1}^{n}\frac{1}{\sigma_j^2}\frac{\partial f_j}{\partial\theta_i}(f_j(\mathbf{X}_j,\ \theta) - F_j) \tag{III.7}$$

и приближенные вторые производные s

$$Z_{ik} = \sum_{j=1}^{n}\frac{1}{\sigma_j^2}\frac{\partial f_j}{\partial\theta_i}\frac{\partial f_j}{\partial\theta_k}. \tag{III.8}$$

Будем искать приближенный минимум s так, как будто f_j линейно зависят от искомых параметров θ. Тогда из (III.5) следует, что искомые θ удовлетворяют системе уравнений:

$$\frac{\partial s}{\partial\theta_i}(\theta^{(0)}) + \sum_k Z_{ik}(\theta^{(0)})(\theta_k - \theta_k^{(0)}) = 0^*.$$

Обозначив $\theta_k - \theta_k^{(0)}$ через $\Delta\theta_k$ и $\partial s/\partial\theta_i$ через g_i, имеем:

$$\Delta\theta_k = -\sum_{i=1}^{m} g_i Z_{ik}^{-1}, \tag{III.9}$$

где Z_{ik}^{-1} — элементы матрицы, обратной к Z.

Если функции $f_j(\theta)$ в (III.6) действительно линейны, то $\theta + \Delta\theta$ указывает координаты точного минимума s.

Если это не так, то вектор $\Delta\theta$ лишь указывает одно из направлений, в котором s уменьшается. Тем не менее, до тех пор, пока мы будем двигаться вдоль $\Delta\theta$ в области, в которой $f_j(\theta + \Delta\theta) \simeq \sum_k \frac{\partial f_j(\theta)}{\partial\theta_k} \times \Delta\theta_k$ (и пока не придем в область минимума), s будет уменьшаться. Таким образом, $\Delta\theta$ отличается от других направлений (например, антиградиента s) тем, что эффективность движения вдоль него

---

* Мы воспользовались тем, что при линейных $f_j(\theta)$ $\frac{\partial s}{d\theta_i}$ в (III.7) так же линейны по θ, и следовательно:

$$\Delta\frac{\partial s}{\partial\theta_i} = \sum_k \frac{\partial^2 s}{\partial\theta_i\,\partial\theta_k}\Delta\theta_k.$$

практически ограничивается областью квазилинейности $f_j(\theta)$ (т. е. нелинейными эффектами), а не другими особенностями рельефа*.

В связи с этим в FUMILI был использован специфический аппарат ограничения шага. Построим параллелепипед W_0 с центром в точке $\theta^{(0)}$ с осями, параллельными осям координат θ_i, и с длинами осей, равными $2b_i$. $b_i^{(0)}$ выберем из соображения, чтобы при изменении θ_i в пределах $\pm b_i$ функции $f_j(\theta)$ были в среднем грубо линейными (нелинейная часть их приращения была по модулю меньше линейной). Если это затруднительно из-за сложности выражений $f_j(\theta)$, то можно вначале выбрать b_i из соображений разумного масштаба или случайно (в дальнейшем они будут уточнены).

Если шаг $\Delta\theta$ выводит за пределы W_0, то найдем точку $\theta^{(1)}$ пересечения вектора $\Delta\theta$ с поверхностью W_0 и попытаемся использовать ее в качестве очередного приближения.

Если $\theta + \Delta\theta$ лежит внутри W_0, то $\theta^{(1)} = \theta^{(0)} + \Delta\theta$.

Вычислим $s_1 = s(\theta^{(1)})$. Если выяснится, что s_1 больше s_0, это значит, нелинейные эффекты не позволяют двигаться столь большим шагом и, следовательно, b_i слишком велики (мы плохо оценили область квазилинейности $f_j(\theta)$). Нам следует уменьшить b_i и передвинуть $\theta^{(1)}$ ближе к $\theta^{(0)}$.

В действительности нам нужно еще позаботиться о подавлении возможных, плохо затухающих колебаний вокруг минимума (при сильной нелинейности). Поэтому шаг следует уменьшить также и в том случае, если s_1 мало уменьшилось по сравнению с s_0. Сделаем это так. Квадратично проинтерполируем s на прямой, соединяющей точки $\theta^{(0)}$ и $\theta^{(1)}$ функцией $\tilde{s}(t)$, используя значение $s(\theta^{(0)})$, $s(\theta^{(1)})$ и $\frac{\partial s(\theta^{(0)} + t(\theta^{(1)} - \theta^{(0)}))}{\partial t}\Big|_{t=0} = g_0(\theta^{(1)} - \theta^{(0)})$. Найдем $s_{\text{мин}} = s(t_{\text{мин}})$ при условии $0 \leqslant t_{\text{мин}} \leqslant 1$. Если

$$s_0 - s_1 < \frac{1}{2}(s_0 - \tilde{s}_{\text{мин}}) \qquad \text{(III.10)}$$

(коэффициент 1/2, как и некоторые последующие коэффициенты, выбран на основе интуиции и численных экспериментов), то шаг следует уменьшить в $t_{\text{мин}}$ раз. Тем не менее, если $t_{\text{мин}} < 1/4$, то возьмем $t_{\text{мин}} = 1/4$, так как при наличии криволинейных оврагов малые $t_{\text{мин}}$ часто оказываются (за счет квадратичной интерполяции) сильно заниженными. Все b_i также уменьшим в $t_{\text{мин}}$ раз. Заменим точкой $\theta^{(0)} + t_{\text{мин}}(\theta^{(1)} - \theta^{(0)})$ точку $\theta^{(1)}$ и повторим предыдущую процедуру проверки сходимости.

Если потребуется, проделаем повторное уменьшение шага. Однако по многим соображениям повторять процедуру уменьшения шага не следует слишком много раз. В частности, при конкретных неудачных $\theta^{(0)}$ могут возникнуть большие потери точности в вычислении s_0 или $\Delta\theta$, из-за которых s_1 никогда не будет меньше s_0. По-

* Например, если рельеф s представляет собой овраг, то эффективность движения вдоль $\Delta\theta$ определяется не наличием оврага, а его искривленностью.

этому ограничимся максимум двумя дроблениями шага (в FUMILI есть параметр, задающий максимальное число дроблений) и перейдем к следующей итерации.

Если уменьшение шага не потребовалось (условие (III.10) не выполнилось), так же перейдем к следующей итерации.

Вторая итерация получается из первой заменой в предыдущих манипуляциях итерационных индексов 0 на 1 и 1 на 2. Размеры области квазилинейности b_i следует иногда увеличивать (для ускорения сходимости). Поэтому, если в предыдущей итерации (или вообще говоря, в нескольких предыдущих итерациях) дробление шага не потребовалось, увеличим в четыре раза те b_i, которые сдерживают движения (т. е. только те, без увеличения которых длина шага $\theta^{(1)} - \theta^{(0)}$ не могла бы быть учетверена).

Избранный нами способ ограничения шага, как правило, требует меньше затрат на перевычисления s, чем в других алгоритмах минимизации, так как обычно не требуется слишком сильных изменений b_i от итерации к итерации.

Дополнительное важное достоинство этого способа то, что b_i являются мерой нелинейности $f_j(\theta)$ и поэтому содержат информацию, необходимую для численного дифференцирования $f_j(\theta)$.

На практике хорошие результаты дает выбор в качестве шага численного дифференцирования по параметру θ_i $0{,}01 b_i$.

Нам следует еще выработать критерий прекращения итерационного процесса. Оказывается, что матрица Z^{-1} является оценкой матрицы статистических ошибок искомых параметров (связанных с погрешностями эксперимента). Естественно, что искать оптимальные параметры с точностью, много превышающей статистическую, нет необходимости. Поэтому в качестве критерия прекращения итераций можно выбрать следующий:

$$| \Delta\theta_i | \leqslant \sqrt{Z^{-1}_{ii}} \text{ для всех } i,$$

где $\varepsilon \sim 0{,}01$.

Если желательно определить точность, с которой получено $s_{\text{мин}} = s(\theta_{\text{мин}})$, то ее можно оценить как $g\Delta\theta/2$ [оценка получается за счет линейной экстраполяции $f_j(\theta)$ и вычисления $s_{\text{мин}}$ как функции экстраполированных $f_j(\theta)$]. На практике эта оценка оказывается весьма близкой к истине.

Из матрицы Z можно извлечь еще очень полезную информацию — факторы корреляции R_i:

$$R_i = Z_{ii} Z^{-1}_{ii} \; (R_i \geqslant 1), \qquad \text{(III.11)}$$

R_i является хорошей мерой обусловленности Z. Во-первых, точность обратной матрицы Z^{-1} по крайней мере в $R_{i\,\text{макс}}$ раз хуже, чем точность матрицы Z. Во-вторых, $\rho_i = \sqrt{1 - 1/R_i}$ является сводным коэффициентом корреляции параметра θ_i с остальными параметрами (т. е. максимальным возможным коэффициентом корреляции параметров θ_i с произвольной линейной комбинацией остальных параметров).

То, что в формуле шага нами используются приближенные вторые производные s вместо точных, вообще говоря, дает большое преимущество. Кроме того, что их легче вычислить, чем точные, они еще и более устойчивы по отношению к смещениям относительно минимума, и в обычных применениях матрица Z не отрицательно определенная. В то же время точные вторые производные включают в себя член $\sum_{j=1}^{n} \frac{F_j - f_j(\theta)}{\sigma_j^2} \frac{\partial^2 f_j}{\partial\theta_i \partial\theta_k}$, содержащий расстояние экспериментальных точек от кривой и поэтому сильно неустойчивый по отношению к смещениям параметров.

Если эксперимент неполный, то, как правило, матрица Z имеет нулевой определитель (и хотя бы одну оценку ошибки θ_i — бесконечную) уже при начальных значениях параметров. Поэтому нет необходимости проводить итерации, чтобы убедиться, что это бессмысленно. С точными вторыми производными это не так.

Тем не менее, естественно, существуют ситуации, в которых приближенность вторых производных сказывается отрицательно.

Прежде всего может оказаться медленной сходимость в окрестности минимума. Однако самым неприятным является тот случай вырождения, когда определитель матрицы Z оказывается нулевым за счет обращения в нуль первой производной $f_j(\theta)$ при всех j по какому-либо параметру (либо линейной комбинации первых производных по нескольким параметрам). В этом случае вклад вторых производных всегда существеннее. В окрестности поверхности, на которой матрица Z вырождена, сходимость к минимуму полностью нарушается. А за счет дробления b_i мы можем сойтись к любой точке поверхности вырождения (если производные имеют соответствующий знак). Признаком такого нарушения сходимости является резкая неустойчивость факторов корреляции и оценок ошибок в процессе итераций и возрастание хотя бы некоторых из них до бесконечности по мере уменьшения s.

В связи с этим нужно обратить внимание на следующее. В методах минимизации произвольных функций часто используют для ограничения области изменения параметров замену переменных. Например, если хотят, чтобы минимум был найден при неотрицательных значениях параметра θ, делают замену $\theta = t^2$ и в качестве параметра используют t. Однако в случае метода линеаризации этого делать нельзя, так как $\frac{\partial f_j(\theta)}{\partial t} = \frac{\partial f_j}{\partial\theta}\frac{\partial\theta}{\partial t} = 2\frac{\partial f_j}{\partial\theta}t$ и при $t=0$ $\frac{\partial f_j}{\partial t} = 0$ для всех j.

Поэтому матрица Z оказывается вырожденной при $t = 0$ и сходимость алгоритма вблизи $t = 0$ полностью разрушается.

Ввиду этого в FUMILI введен специальный аппарат простых ограничений параметров типа $\theta_{i\,\text{мин}} \leqslant \theta_i \leqslant \theta_{i\,\text{макс}}$, где $\theta_{i\,\text{мин}}$ и $\theta_{i\,\text{макс}}$ константы. Хотя это и частный случай, но он сравнительно эффективно реализуется и часто достаточен для практических применений.

Реализуется он следующим образом. При построении параллелепипеда W учитываются не только b_i, но и границы $\theta_{i\ \text{мин}}$ и $\theta_{i\ \text{макс}}$. Если $a_i + b_i > \theta_{i\ \text{макс}}$, то соответствующей границей W делается $\theta_{i\ \text{макс}}$. Аналогично с $\theta_{i\ \text{мин}}$.

Далее, если оказывается, что очередное приближение лежит на границе W (за счет $\theta_{\text{мин}}$ или $\theta_{\text{макс}}$), то все параметры, которые лежат на границе, и составляющие градиента, s по которым таковы, что s уменьшается за пределы W, фиксируются.

Затем по остальным параметрам вычисляется шаг, и если оказывается, что какие-нибудь из незафиксированных параметров, лежащих на границе W, пытаются выйти за пределы W, то один из них также фиксируется, снова вычисляется $\Delta\theta$, делается очередная проверка незафиксированных граничных параметров, и так до тех пор, пока никакой из граничных параметров не будет пытаться выйти за границу W. По оставшимся незафиксированным параметрам вычисляется $\Delta\theta$ и выполняется шаг, ограниченный размерами W.

Критерий окончания итерационного процесса также модифицируется. Итерационный процесс считается законченным, если на границах параметры зафиксированы только из-за знака соответствующей составляющей градиента и одновременно выполняется критерий точности $|\Delta\theta_i| < \varepsilon\sigma_i$ по остальным параметрам.

Ограничения $\theta_{i\ \text{макс}}$ и $\theta_{i\ \text{мин}}$ могут применяться также для того, чтобы ограничить область изменения параметров разумными пределами. Иначе при сложной форме рельефа иногда итерационный процесс может уводить далеко в сторону даже от сравнительно хорошего начального приближения.

Как мы уже говорили, матрица Z^{-1} является оценкой матрицы ошибок σ^2_{ik} искомых параметров.

Если зависимость $f_j(\theta)$ от параметров линейная и физические ограничения на область изменения параметров отсутствуют, то Z^{-1} является точной матрицей ошибок оценки параметров, полученной в результате минимизации (если, конечно, величины экспериментальных погрешностей точек можно считать точными).

В случае нелинейной зависимости $f_j(\theta)$ от искомых параметров в большинстве случаев не существует доступных для современных вычислительных машин точных способов оценки погрешностей параметров.

Однако матрица Z^{-1} является хорошей оценкой погрешностей параметров, если в окрестности минимума в задаваемом ею эллипсоиде рассеяния, соответствующем отклонению на несколько ошибок от минимума, Z^{-1} меняется слабо.

В частности, можно смотреть, как меняются оценки ошибок параметров при приближении к минимуму. И если на расстоянии нескольких ошибок от минимума они при дальнейших итерациях меняются слабо, это может служить указанием на то, что погрешности определены хорошо.

На практике широко применяется другой способ оценки погрешностей, когда оцениваемый параметр меняют в окрестностях минимума и фиксируют, а по остальным параметрам ищут минимум. Профиль функции $s(\theta_k)$, полученной таким образом, используют для оценки погрешности θ_k, считая, что увеличение $2s(\theta_k)$ по отношению к $2s_{\text{мин}}$ на 1 соответствует отклонению θ_k на одну стандартную ошибку. Увеличение $2s(\theta_k)$ на 4 соответствует отклонению на две стандартные ошибки, и т. д. Этот способ дает точные результаты также лишь в случаях линейной зависимости f_j от искомых параметров (либо в случаях, сводящихся к ним заменой переменных).

В целом этот способ дает, видимо, более близкие к истине оценки погрешностей, чем Z^{-1}, но, к сожалению, имеет тенденцию занижать погрешности.

Если минимум $s(\theta)$ не единственный, то возникают дополнительные проблемы, связанные с поиском всех локальных минимумов $s(\theta)$. Часто нет ничего лучшего, кроме поиска, исходя из различных случайных начальных приближений (хотя могут быть приемы, повышающие вероятность поиска минимумов с хорошими χ^2). При этом возникают проблемы отбрасывания статистически необоснованных решений либо по χ^2-критерию, если разница в χ^2 велика, либо по более тонким критериям, как, например, τ-критерий [12, 13].

СПИСОК ЛИТЕРАТУРЫ

1. **Fletcher R., Reeves C. M.** Function minimization by conjugate gradients — «Comput. J.», 1964, v. 7, p. 149.
2. **Fletcher R., Powell M. J. D.** A rapidly convergent descent method for minimization. — «Comput. J.», 1963, v. 6, p. 163.
3. **Силин И. Н., Белявская Л. В.** Минимизации функций многих переменных с использованием вторых производных. — Препринт 2674, Дубна, 1966.
4. **Levenberg K.** A method for the solution certain non-linear problems in least squares. — «Quart. Appl. Math»., 1944, v. 2, p. 164.
5. **Marquardt D. W.** An algorithm for least-squares estimation of non-linear parameters. — «J. Soc. Industr. and Appl. Math.», 1963, v. 11, N 2, p. 431.
6. **Powell M. J. D.** A method for minimizing a sum of squares of non-linear functions without calculating derivatives. — «Comput. J.». 1965, v. 7, N 4, p. 303.
7. **Соколов С. Н., Силин И. Н.** Нахождение минимумов функционалов методом линеаризации. — Препринт ОИЯИ Д-810, Дубна, 1961.
8. **Казаринов Ю. М., Силин И. Н.** Фазовый анализ нуклон-нуклонного рассеяния при энергии 240 *Мэв*. — «Журн. эксперим. и теор. физ.», 1962, т. 43, с. 692.
9. **Библиотека** программ на фортране D-520 (И. Н. Силин), т. I, Б1-11-5144, Дубна; 1970, с. 180. Депонированная публикация ОИЯИ.
10. **Родионов А. И., Силин И. Н.** Алгольный вариант программы FUMILI — минимизация функционалов методом линеаризации. — «Совместный научный сборник ОИЯИ (Дубна, СССР) и ЦИФИ (Будапешт, Венгрия). Алгоритмы и программы для решения некоторых задач физики». Будапешт, 1974, KFKI-74-34, с. 113.
11. **Силин И. Н.** Программа сравнения гипотез по τ и r^2-критериям. — Там же, с. 161.
12. **Pazman A.** A method for testing complex linear h ypothesis and its use in the phase shift analysis. Пре принт ОИЯИ Е-5-3775, 1968.
13. **Пазман А., Силин И. Н.** Модификация τ-крите рия и r^2-критерий сравнения гипотез. — Сообщение ОИЯИ Р5-7174, Дубна, 1973.

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

(Числа обозначают раздел)

ОГЛАВЛЕНИЕ

Идье В., Драйард Д., Джеймс Ф., Рус М., Садуле Б.

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЭКСПЕРИМЕНТАЛЬНОЙ ФИЗИКЕ

Редактор *В. Н. Безрукова*
Художественный редактор *А. Т. Кирьянов*
Переплет художника *С. Н. Томилина*
Технический редактор *С. В. Долгополова*
Корректоры *М. А. Жарикова, Н. А. Смирнова*

Сдано в набор 29/VII 1975 г. Подписано к печати 2/III 1976 г.
Формат $60\times90^1/_{16}$ Бумага типографская № 2 Усл. печ. л. 21,0
Уч.-изд. л. 21,91. Тираж 6000 экз. Цена 2 р. 35 к. Зак. изд. 72290
Зак тип. 401
Атомиздат 103031 Москва, К-31, ул. Жданова, 5

Московская типография № 4 Союзполиграфпрома
при Государственном комитете Совета Министров СССР
по делам издательств, полиграфии и книжной торговли,
г. Москва, И-41. Б. Переяславская ул., дом 46