ПРИКЛАДНОЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ

УЧЕБНОЕ ПОСОБИЕ

ДЛЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ

СПЕЦИАЛЬНОСТЬ



Прикладной статистический анализ

Е. И. Куликов

Горячая линия–Телеком

Е. И. Куликов

Прикладной статистический анализ

2-е издание, перераб. и доп.

Допущено УМО вузов по образованию в области информационной безопасности в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности 090106 – «Информационная безопасность телекоммуникационных систем», и аспирантов, обучающихся по специальности научных работников 05.13.19 – «Методы и системы защиты информации, информационная безопасность».

Москва
Горячая линия - Телеком
2008

УДК 519.2
ББК 22.172
К90

Рецензенты:

Институт криптографии, связи и информатики ФСБ России (нач. кафедры, доктор техн. наук *В. В. Сизых* и канд. физ.-мат. наук *В. А. Шапошников*);
кафедра «Автономные, информационные и управляющие системы» Московского государственного технического университета им. Н. Э. Баумана (зав. кафедрой доктор техн. наук, профессор *В. К. Хохлов*)

Куликов Е. И.

К90 Прикладной статистический анализ. Учебное пособие для вузов. – 2-е изд., перераб. и доп. – М.: Горячая линия–Телеком, 2008. – 464 с.: ил.
ISBN 978-5-9912-0021-9

Книга содержит обширный методический аппарат по прикладному статистическому анализу применительно к различным научно-техническим направлениям и состоит из двух частей. В первой части изложены основные положения теории вероятностей: законы распределения случайных величин, функциональные преобразования случайных величин, аппроксимация законов распределения, предельные соотношения и др. Во второй части освещены положения математической статистики: сущность выборочного метода, оценка параметров распределения, проверка гипотез, корреляционный, регрессионный и дисперсионный анализы, понятия современной теории эксперимента (планирования эксперимента). Изложение материала иллюстрируется численными примерами. Приложен набор таблиц, необходимых при статистических расчетах и выводах, а также задачи для самостоятельного решения.

Для студентов и аспирантов, изучающих дисциплины по информационной безопасности. Будет полезна как справочное пособие для инженеров и научных работников, интересующихся применением методов прикладной математической статистики.

ББК 22.172

Адрес издательства в Интернет www.techbook.ru

Учебное издание

Куликов Евгений Иванович
ПРИКЛАДНОЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ
Учебное пособие для вузов

Обложка художника В. Г. Ситникова
Компьютерная верстка Ю. Н. Чернышев

Подписано в печать 28.06.07. Формат 60 × 90/16. Печать офсетная. Уч.-изд. л. 29. Изд. № 721. Заказ № 8987.
ООО «Научно-техническое издательство «Горячая линия — Телеком»

Отпечатано с готовых диапозитивов в ОАО ордена «Знак Почета» «Смоленская областная типография им. В. И. Смирнова» 214000, г. Смоленск, проспект им. Ю. Гагарина, 2.

ISBN 978-5-9912-0021-9

© Е. И. Куликов, 2008
© Оформление издательства «Горячая линия–Телеком», 2008

Посвящается светлой памяти учителя и друга профессора Василия Ивановича Тихонова.

Предисловие

В последние десятилетия методы прикладного статистического анализа (прикладная статистика), основой которого являются теория вероятностей и математическая статистика, широко используются при изучении разнообразных вопросов науки и техники.

Следует отметить, что многие явления и системы, для исследования которых казалось достаточным применение классических методов, при более глубоком изучении потребовали отказа от «точного» описания этих явлений и перехода к их описанию методами теории вероятностей и математической статистики. Очень часто отказ от точного описания явления (системы) объясняется не только его сложностью, но и незнанием всех причин, связанных с возникновением изучаемого явления, а также тем, что не всегда имеется принципиальная возможность задания начального состояния явления. Методы теории вероятностей и математической статистики дают возможность обратиться непосредственно к законам, управляющим массовым явлением, минуя изучение отдельных его составляющих. Так возникли новые науки: статистическая физика, квантовая механика, статистическая радиотехника и другие науки.

Методы математической статистики не только позволяют обрабатывать данные исследований с оценкой достоверности получаемых статистических характеристик эксперимента, но также дают возможность наиболее целесообразным образом организовать (спланировать) проведение эксперимента с наименьшими стоимостными, временными и другими затратами. Это особенно важно при синтезе и анализе систем, эффективность работы которых зависит от большого числа причин (факторов). В подобной ситуации задачей исследования становится проблема выбора наиболее существенных факторов, конкретные значения которых в наибольшей степени оказывают влияние на функционирование сложных процессов и систем в различных областях научно-технической и общественно-политической деятельности человека.

Внедрение методов теории вероятностей и математической статистики в практику прикладных наук существенно возросло в результате «массового» применения современных вычислительных средств (персональных компьютеров), обладающих большим быстродействием выполнения арифметических и логических операций и большой долговременной и оперативной памятью. Появление современных персональных компьютеров позволило перейти от решения достаточно простых, как правило, одномерных и двумерных задач теории вероятностей и ма-

тематической статистики к значительно более сложным многомерным задачам. С этой целью создан ряд прикладных математических программ для решения как одномерных, так и многомерных задач. Однако продуктивное использование этих программ исследователями возможно лишь при наличии у них знаний основ прикладной теории вероятностей и математической статистики.

В течение длительного времени автор читает лекции для инженеров и аспирантов со средней математической подготовкой (в объеме втузовского курса). Эти лекции по прикладным методам статистического анализа и планирования эксперимента применительно к изучению и исследованию широкого спектра технических и естественных систем явились основой для написания данной книги. Вместе с тем по кругу рассматриваемых вопросов, их объему и прикладной направленности книга существенно выходит за рамки программы по курсу теории вероятностей и математической статистики для технических вузов.

Изложение материала подкрепляется решением задач, поясняющим сущность и способы практического использования описываемых методов аналитического и экспериментального исследований. Для проверки уяснения степени освоения излагаемого материала в конце каждой главы приведены контрольные вопросы и задачи. В конце книги на некоторые из задач приведены ответы.

В качестве учебного пособия книга предназначена для студентов и аспирантов. Как справочное пособие она может быть полезной для инженеров и научных работников, интересующихся применением методов прикладной математической статистики при проведении экспериментальных исследований. Для более глубокого изучения рассматриваемых вопросов предлагается использовать литературу, приведенную в конце книги.

Автор будет считать свою работу над книгой небесполезной, если она окажет помощь начинающим статистикам в изучении основ прикладной теории вероятностей и математической статистики для использования в своей практической деятельности.

В данном издании были учтены замеченные неточности и опечатки в ранее изданной книге автора. Кроме этого в книгу включен дополнительный материал, в основном связанный с многомерными задачами и непараметрическими методами анализа, а также увеличено число примеров и задач.

При изложении материала возможны неточности, описки и даже ошибки. Автор заранее благодарит всех за критические замечания как по существу излагаемого материала, так и за указание описок и ошибок.

Автор искренне благодарен начальнику ИКСИ Академии ФСБ России профессору А.П. Коваленко и его сотрудникам за помощь в издании книги.

Автор признателен профессору Ю.Н. Чернышову за редактирование и компьютерную верстку книги.

ЧАСТЬ I

Основные положения теории вероятностей

Порядок: Ни одна наука, созданная людьми, не смогла его соблюсти.

Паскаль Блез. «Мысли»

На правилах покоится игра.

Хессе Герман. «Игра в бисер»

ГЛАВА 1

События, случайные величины и их описание

1.1. События и вероятности событий

1.1.1. Определения

Научное познание происходящих в мире явлений имеет дело с закономерностями в повторяющихся наблюдениях. При этом из первоначальных понятий познания является абстрактное понятие *события*, которое определяется тем, что изучаемое явление при определенных условиях происходит или не происходит (наблюдается или не наблюдается). Абстрактное понятие события не имеет отношения к его природе, оно имеет отношение только к факту: произошло событие или не произошло.

В естествознании и других областях научного познания для описания закономерностей связи между теми или иными повторяющимися условиями S и событием A обычно используют два подхода: детерминистический и вероятностный (стохастический).

Детерминистический подход характеризуется тем, что при каждом осуществлении условий S наступает событие A. Иначе говоря, при детерминистическом подходе изучаемое явление (процесс) описывается

достаточно точно, в результате чего, если известно начальное состояние явления и известны внешние воздействия на это явление, то поведение изучаемого явления в будущем определяется однозначно. Подобная ситуация, например, наблюдается в классической механике, в которой утверждается, что при заданных силах, действующих на тело или систему тел, и заданных других начальных условиях движение тела или системы тел (событие) будет происходить однозначно определенным, заранее предсказуемым образом.

Сущность вероятностного подхода состоит в том, что при осуществлении заданных условий S в зависимости от привходящих обстоятельств событие A может произойти, а может и не произойти, т.е. является случайным.

В качестве иллюстрации эффективности вероятностного подхода можно привести пример, связанный с изучением свойств газа в целом (давление, температура и т.д.). Поскольку газ состоит из отдельных молекул, то теоретически для каждой молекулы можно составить дифференциальное уравнение ее движения в пространстве по законам механики. Однако для каждой из молекул не представляется возможным указать координаты и скорости в некоторый начальный момент времени; практически невозможно решить систему, состоящую из большого числа дифференциальных уравнений, и, тем более, ничего нельзя сказать о свойствах газа в целом. Вероятностный подход позволяет получить необходимый результат о свойствах газа, основанный на изучении некоторого усредненного эффекта случайно движущихся молекул.

При вероятностном подходе изучаются закономерности в случайных событиях и способы их количественного описания. Действительно, если наблюдать серии большого числа опытов, то обнаруживается, что, хотя результаты отдельных опытов могут сильно отличаться друг от друга, средние значения в каждой из серий остаются примерно равными, проявляя определенную устойчивость по мере увеличения числа опытов в серии. Эта *наблюдаемая статистическая закономерность, связанная с устойчивостью средних результатов, отражает вероятностную закономерность случайных событий.*

Первоначально идея о количественных закономерностях случайных событий возникла при карточных и других азартных играх. Как таковая теория вероятностного подхода начала развиваться с середины XVII века благодаря работам П. Ферма, Б. Паскаля и Я. Бернулли.

Математической базой современного вероятностного подхода является теория вероятностей. Над ее созданием работали многие великие философы, математики и физики. Как точная математическая наука она была обоснована А.Н. Колмогоровым.

Вероятность события — это числовая характеристика объективной возможности его появления. Теория вероятностей постулирует, что *вероятность события есть действительное число*, принимающее значение в диапазоне от нуля до единицы. К числу основных определений теории

вероятностей относятся понятия достоверного и невозможного событий, а также *аксиомы*, определяющие правила сложения и произведения вероятностей двух и более событий. Последние часто называют теоремами сложения и произведения (умножения) вероятностей событий.

Событие A, которое неизбежно происходит при каждом выполнении условий S, называется *достоверным*, и ему приписывается вероятность, равная единице. Событие A, которое не может произойти при выполнении условий S, называется *невозможным*, и ему приписывается вероятность, равная нулю. Событие A, которое при выполнении условий S может произойти, а может и не произойти, называется *случайным*. События называются *несовместными*, если никакие из них не могут произойти вместе при одном и том же опыте; в противоположной ситуации, т.е. когда события при одном и том же опыте могут произойти одновременно, такие события называются *совместными*. События называются *равновозможными*, если имеются основания полагать, что ни одно из них не является более возможным, чем любое другое. Несколько попарно несовместных событий образуют *полную группу*, если в результате опыта обязательно происходит хотя бы одно из них при каждом осуществлении условий S. Если два несовместных события образуют полную группу, то такие события называются *противоположными*, причем событие, противоположное событию A, принято обозначать $\bar{A}$.

Теория вероятностей обосновывает соответствующие правила, которые позволяют по вероятностям одних случайных событий находить вероятности других случайных событий, если последние каким-либо образом связаны с первыми случайными событиями. Особенностью случайных событий является то, что их закономерности проявляются лишь при многократном проведении опытов.

Для сравнения вероятностей появления тех или иных событий вводится некоторая численная мера вероятности события, Существует несколько определений этой меры.

Приведем классическую формулу вероятности события, когда имеется полная группа равновозможных и несовместных событий. При этом пусть событию A благоприятствуют m элементарных событий $A_1, \ldots, A_m$ из общего их числа n. Тогда отношение числа m благоприятствующих элементарных событий к общему их числу n есть *вероятность* события A:

$$\mathbf{P}\{A\} = p(A) \equiv p_A = \frac{m}{n}. \qquad (1.1.1)$$

Очевидно, что для вероятности события справедливо неравенство

$$0 \leqslant p(A) \leqslant 1. \qquad (1.1.2)$$

Применительно к комбинациям событий широкое применение находят понятия суммы и произведения событий.

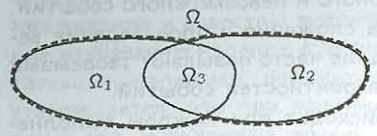


Рис. 1.1.1. Объединение и пересечение подобластей существования событий

Суммой (или объединением) двух событий A_1 и A_2 называется событие B, состоящее в появлении события A_1, или события A_2, или обоих этих событий одновременно. Иначе, суммой, или объединением, двух событий A_1 и A_2 называют событие B, состоящее в появлении хотя бы одного из этих двух событий. Аналогично суммой, или объединением, нескольких событий (m событий) $A_1, A_2, \ldots, A_m$ называют событие B, состоящее в появлении хотя бы одного из этих событий. Иначе говоря, событие B будет являться суммой событий $A_1, A_2, \ldots, A_m$, если произойдет либо одно из событий A_i, либо любые совместные события A_i, A_j или A_i, A_j, A_k и т.д., вплоть до совместного появления всех m событий $A_1, A_2, \ldots, A_m$. Сумму событий обычно обозначают как $B = A_1 + A_2 + \ldots + A_m$ или $B = \bigcup_{i=1}^{m} A_i$, где $\bigcup$ — знак объединения событий (по аналогии с объединением множеств). Сказанное условно можно выразить графически, как показано на рис. 1.1.1. Пусть Ω_1 — подобласть, соответствующая попаданию точек события A_1, а Ω_2 — подобласть, соответствующая попаданию точек события A_2. Тогда сумма событий $B = A_1 + A_2$ есть событие, соответствующее попаданию точек во всю выделенную пунктиром область Ω, которая является объединением подобластей Ω_1 и Ω_2.

Произведением нескольких событий (m событий) $A_1, A_2, \ldots, A_m$ называют событие C, состоящее в совместном появлении всех этих событий. Произведение событий обозначают как $C = A_1 A_2 \cdots A_m$, или $C = \bigcap_{i=1}^{m} A_i$, где $\bigcap$ — знак совмещения событий (пересечения множеств). Графически (см. рис. 1.1.1) произведение двух событий A_1 и A_2 представляет собой событие, состоящее в совместном попадании точек этих событий в область Ω_3, являющуюся пересечением подобластей Ω_1 и Ω_2.

События могут находиться между собой в определенной зависимости и взаимозависимости. Для таких событий в теории вероятностей введено понятие условной вероятности события. Если (применительно к двум событиям) событие A_2 зависит от того, произошло событие A_1 или не произошло, вероятность появления события A_2 называется *условной вероятностью* и записывается как $p(A_2 \mid A_1)$. Событие A_2 называется *независимым* от события A_1, если

$$p(A_2 \mid A_1) = p(A_2), \tag{1.1.3}$$

и, наоборот, событие A_2 называется *зависимым* от события A_1, если

$$p(A_2 \mid A_1) \neq p(A_2). \tag{1.1.4}$$

Следует иметь в виду, что понятие независимости событий относится как к одному опыту, так и к разным опытам, в то время как понятие совместности и несовместности событий применимо лишь к одному опыту. Так, например, при бросании игральной кости с цифрами от 1 до 6

выпадение одновременно простого (1, 3 или 5) и нечетных чисел являются событиями совместными, а одновременное выпадение этих же простых чисел и четного числа 4 или 6 — события несовместные.

1.1.2. Вероятность произведения событий (теорема умножения вероятностей)

Вероятность произведения двух совместных событий равна произведению вероятности одного события на условную вероятность другого события при условии, что произошло первое событие:

$$\mathbf{P}\{A_1 A_2\} = p(A_1)p(A_2 \mid A_1) = p(A_2)p(A_1 \mid A_2). \tag{1.1.5}$$

Это соотношение обычно называют *теоремой* или *формулой умножения вероятностей*.

Вероятность произведения нескольких событий $A_1, A_2, \ldots, A_m$ равна вероятности события A_1, умноженной на вероятность события A_2, при условии, что наступило событие A_1, и так далее, умноженной на вероятность события A_m, при условии, что наступили события $A_1, A_2, \ldots, A_{m-1}$:

$$\mathbf{P}\{A_1 A_2 \ldots A_m\} = p(A_1)p(A_2 \mid A_1)p(A_3 \mid A_1, A_2)\ldots \\ \ldots p(A_m \mid A_1, A_2, \ldots, A_{m-1}). \tag{1.1.6}$$

Если события взаимно независимы, то формула умножения событий упрощается и принимает вид

$$\mathbf{P}\{A_1 A_2 \ldots A_m\} = \prod_{i=1}^{m} p(A_i). \tag{1.1.7}$$

1.1.3. Вероятность суммы событий (теорема сложения вероятностей)

Вероятность суммы двух совместных событий равна сумме вероятностей этих событий за вычетом вероятности их произведения:

$$\mathbf{P}\{A_1 + A_2\} = p(A_1) + p(A_2) - p(A_1 A_2). \tag{1.1.8}$$

Это наглядно видно на рис. 1.1.1. Если события A_1 и A_2 независимы, то

$$\mathbf{P}\{A_1 + A_2\} = p(A_1) + p(A_2) - p(A_1)p(A_2). \tag{1.1.9}$$

Соотношение (1.1.8) часто называют *теоремой* или *формулой сложения вероятностей*.

Формула сложения двух событий обобщается на сумму более двух совместных событий. Применительно к трем событиям A_1, A_2 и A_3 имеем

$$\mathbf{P}\{A_1 + A_2 + A_3\} = p(A_1) + p(A_2) + p(A_3) - \\ - p(A_1 A_2) - p(A_1 A_3) - p(A_2 A_3) + p(A_1 A_2 A_3). \tag{1.1.10}$$

Применительно к сумме m несовместных событий $B = A_1 + A_2 + \ldots + A_m$

$$\mathbf{P}\{B\} = \sum_{i=1}^{m} p(A_i), \qquad (1.1.11)$$

т.е. вероятность суммы несовместных событий равна сумме вероятностей этих событий. Из формулы сложения вероятностей вытекают два очевидных следствия.

1. Сумма вероятностей несовместных событий $A_1, A_2, \ldots, A_m$, составляющих полную группу, равна единице:

$$\mathbf{P}\{A_1 + A_2 + \ldots + A_m\} = p(A_1) + p(A_2) + \ldots + p(A_m) = 1. \qquad (1.1.12)$$

2. Сумма вероятностей противоположных событий A и $\bar{A}$ равна единице:

$$\mathbf{P}\{A + \bar{A}\} = p(A) + p(\bar{A}) = 1. \qquad (1.1.13)$$

С учетом введения противоположного события и вероятности противоположного события формулу сложения вероятностей можно записать в другой форме. Действительно, вероятность наступления по крайней мере хотя бы одного из m событий $A_1, A_2, \ldots, A_m$ равна единице минус вероятность того, что не произойдет ни одного из этих событий (произойдет m противоположных событий):

$$\mathbf{P}\{A_1 + A_2 + \ldots + A_m\} = 1 - \mathbf{P}\{\bar{A}_1 \bar{A}_2 \ldots \bar{A}_m\}. \qquad (1.1.14)$$

Если события A_i независимы между собой, то независимы между собой и противоположные события $\bar{A}_i$. Применительно к независимым событиям

$$\mathbf{P}\{\bar{A}_1 \bar{A}_2 \ldots \bar{A}_m\} = \prod_{i=1}^{m} p(\bar{A}_i) = \prod_{i=1}^{m} [1 - p(A_i)]. \qquad (1.1.15)$$

Поэтому вероятность наступления хотя бы одного из m независимых событий A_i

$$p(A_1 + A_2 + \ldots + A_m) = 1 - \prod_{i=1}^{m} [1 - p(A_i)]. \qquad (1.1.16)$$

Для иллюстрации применения приведенных выше формул произведения и сложения вероятностей рассмотрим пример.

Пример 1.1.1. По цели производится три раздельных независимых выстрела. Обозначим события, соответствующие попаданию в цель при первом, втором и третьем выстрелах, как A_1, A_2 и A_3. Будем полагать, что вероятности этих событий заданы и соответственно равны: $p(A_1) = p_1 = 0{,}5$; $p(A_2) = p_2 = 0{,}6$; $p(A_3) = p_3 = 0{,}8$. Определим вероятность непопадания выстрелов в цель и вероятности попадания выстрелов в цель соответственно одного, двух и трех раз.

Обозначим события, соответствующие непопаданию и попаданию одного, двух и трех выстрелов в цель как H_0, H_1, H_2, H_3. Поскольку события A_i независимы и несовместны, то события H_i также независимы и несовместны. Отсюда вероятность того, что все три выстрела будут соответствовать трем попаданиям в цель, $p(H_3) = p_1 p_2 p_3 = 0{,}24$. Вероятность того, что ни одного попадания не произойдет, равна произведению вероятностей противоположных событий: $p(H_0) = (1 - p_1)(1 - p_2) \times (1 - p_3) = 0{,}04$. При попадании одного выстрела возможны три варианта события, т.е. событие состоит из суммы трех произведений несовместных событий:

$$H_1 = A_1 \bar{A}_2 \bar{A}_3 + \bar{A}_1 A_2 \bar{A}_3 + \bar{A}_1 \bar{A}_2 A_3.$$

Соответствующая вероятность того, что из трех выстрелов будет одно попадание,

$$p(H_1) = p_1(1 - p_2)(1 - p_3) + (1 - p_1)p_2(1 - p_3) + (1 - p_1)(1 - p_2)p_3 = 0{,}26.$$

Аналогично событие H_2, соответствующее двум попаданиям, также состоит из суммы трех произведений событий:

$$H_2 = A_1 A_2 \bar{A}_3 + \bar{A}_1 A_2 A_3 + A_1 \bar{A}_2 A_3.$$

Вероятность этого события

$$p(H_2) = p_1 p_2 (1 - p_3) + (1 - p_1)p_2 p_3 + p_1(1 - p_2)p_3 = 0{,}46.$$

Тогда вероятность того, что из трех выстрелов хотя бы один из них попадет в цель (событие H), равна сумме вероятностей:

$$p(H) = p(H_1) + p(H_2) + p(H_3) = 0{,}96.$$

Поскольку события H_0, H_1, H_2 и H_3 составляют полную группу событий, то убеждаемся в том, что вероятность этой группы событий равна единице.

1.1.4. Формулы полной и обратной вероятности (формула Байеса)

Следствием формул сложения и произведения является формула полной вероятности, а также формула обратной вероятности (формула Байеса).

Формула полной вероятности. Пусть некоторое событие A может появиться только в сочетании с одним из несовместных событий $H_1, H_2, \ldots, H_m$, которые составляют полную группу и называются *гипотезами событий*. Предположим, что известны вероятности этих событий $p(H_i)$, $i = 1, 2, \ldots, m$, а также условные вероятности $p(A \mid H_i)$ события A при условии, что произошло одно из несовместных событий H_i. Необходимо найти безусловную вероятность $p(A)$ события A. Так как событие A появляется тогда и только тогда, когда осуществляется одно из попарно несовместных событий $AH_1, AH_2, \ldots, AH_m$, то по формуле сложения вероятностей имеем

$$p(A) = \sum_{i=1}^{m} p(AH_i). \qquad (1.1.17)$$

Отдельные слагаемые в правой части находим по формуле умножения вероятностей: $p(AH_i) = p(H_i)p(A \mid H_i)$. Таким образом, формула полной вероятности имеет вид

$$p(A) = \sum_{i=1}^{m} p(H_i)p(A \mid H_i). \qquad (1.1.18)$$

Эту формулу называют *формулой полной вероятности событий*. Из неё видно, что безусловная вероятность события $p(A)$ есть не что иное, как некоторая усредненная вероятность появления события A.

Пример 1.1.2. Дополнительно к условиям предыдущего примера будем считать, что цель будет поражена: при одном попадании — с вероятностью 0,3; при двух попаданиях — с вероятностью 0,5; при трех попаданиях — с вероятностью 1,0. Вычислим вероятность того, что цель будет поражена.

Обозначим через A событие, соответствующее поражению цели. Тогда заданные условные вероятности поражения цели для различных событий (гипотез) H_i попадания в цель, будут $p(A \mid H_0) = 0$; $p(A \mid H_1) = 0{,}3$; $p(A \mid H_2) = 0{,}5$; $p(A \mid H_3) = 1$. Вероятности попадания в цель при трех независимых выстрелах вычислены в предыдущем примере и равны соответственно: $p(H_0) = 0{,}04$; $p(H_1) = 0{,}26$; $p(H_2) = 0{,}46$; $p(H_3) = 0{,}24$. Подставляя значения $p(A \mid H_i)$ и $p(H_i)$ в формулу полной вероятности (1.1.18), получим, что цель будет поражена с вероятностью $p(A) = 0{,}26 \cdot 0{,}3 + 0{,}46 \cdot 0{,}5 + 0{,}24 \cdot 1{,}0 = 0{,}548$.

Формула обратной вероятности (формула Байеса). Формулу полной вероятности группы событий можно рассматривать и иначе. Пусть известны вероятности гипотез $p(H_i)$ и условные вероятности $p(A \mid H_i)$ появления события A для каждой из гипотез. При этих условиях безусловная вероятность события A дается формулой полной вероятности (1.1.18).

Допустим, что произведено наблюдение, в результате которого появилось событие A. Выясним, с какой из гипотез H_i следует связывать событие A? Для решения вопроса необходимо вычислить условные вероятности гипотез $p(H_i \mid A)$ при условии, что произошло событие A. Затем эти условные вероятности надо сравнить между собой по заранее выбранному критерию (правилу). Наиболее часто в качестве такого правила используют критерий максимальной условной вероятности. Согласно этому критерию предпочтение отдается той из гипотез H_i, для которой условная вероятность $p(H_i \mid A)$ максимальна. На основании формулы умножения вероятностей имеем

$$p(AH_i) = p(A)p(H_i \mid A) = p(H_i)p(A \mid H_i).$$

Откуда получаем

$$p(H_i \mid A) = \frac{p(H_i)p(A \mid H_i)}{p(A)} = \frac{p(H_i)p(A \mid H_i)}{\sum_{i=1}^{m} p(H_i)p(A \mid H_i)}. \qquad (1.1.19)$$

Эту формулу называют *формулой обратной вероятности гипотез* или *формулой Байеса*. Формула Байеса позволяет оценить вероятности гипотез после того, как произошло событие A. Она играет фундаментальную роль во многих задачах статистических испытаний. При этом вероятности гипотез $p(H_i)$ обычно называют *априорными вероятностями* (вероятностями события до опыта), а условные вероятности гипотез $p(H_i \mid A)$ — *апостериорными вероятностями* (вероятностями события после опыта).

Пример 1.1.3. Применительно к условиям задач, рассмотренным в предыдущих двух примерах, будем полагать, что произошло поражение цели (событие A). Определим, какая из гипотез H_1, H_2 или H_3, соответствующая попаданию в цель одного, двух или трех выстрелов, наиболее вероятна.

Из предыдущих примеров имеем: $p(A) = 0{,}548$; $p(H_0) = 0{,}04$; $p(H_1) = 0{,}26$; $p(H_2) = 0{,}46$; $p(H_3) = 0{,}24$; $p(A \mid H_0) = 0{,}0$; $p(A \mid H_1) = 0{,}3$; $p(A \mid H_2) = 0{,}5$; $p(A \mid H_3) = 1{,}0$. Подставляя эти величины в формулу Байеса (1.1.19), получим значения апостериорных вероятностей трех гипотез: $p(H_1 \mid A) = 0{,}142$; $p(H_2 \mid A) = 0{,}420$; $p(H_3 \mid A) = 0{,}438$. Из сравнения апостериорных вероятностей этих гипотез видим, что наиболее вероятна гипотеза о попадании в цель трех выстрелов.

1.2. Случайные величины и их описание

Случайной величиной называется такая переменная величина, которая в результате однотипных опытов может принимать разные заранее неизвестные значения. В зависимости от возможных значений, принимаемых случайной величиной, случайные величины можно разделить на три группы: дискретные, непрерывные и непрерывно-дискретные (смешанные).

Как правило, случайные величины будем обозначать большими буквами латинского алфавита X, Y, Z и т.д., а их конкретные (наблюдаемые) значения будем обозначать малыми буквами этого же алфавита x, y, z и т.д.

1.2.1. Функция распределения, плотность вероятности и характеристическая функция

Для полного описания случайной величины необходимо указать возможные значения, принимаемые случайной величиной, и вероятности этих значений. Соотношение, устанавливающее в той или иной форме зависимость между возможными значениями случайной величины и их вероятностями, называется *законом распределения вероятностей* (в дальнейшем будем называть просто законом распределения).

Дискретная случайная величина может принимать конечное или счетное множество возможных значений. Закон распределения дискретной случайной величины X, принимающей значения $x_1, x_2, \ldots, x_n$, задается указанием вероятностей этих значений:

$$\mathbf{P}\{X = x_k\} = p(x_k) \equiv p_k, \quad k = 1, 2, \ldots, n. \qquad (1.2.1)$$

Зависимость между соответствующими значениями дискретной случайной величины и вероятностями может быть задана в виде графика (*многоугольника распределения*), а также в виде таблицы или аналитического выражения. Примерный вид многоугольника распределения вероятностей дискретной случайной величины изображен на рис. 1.2.1. Очевидно, что все вероятности должны удовлетворять условиям

$$p_k \geqslant 0, \quad \sum_{k=1}^{n} p_k = 1. \qquad (1.2.2)$$

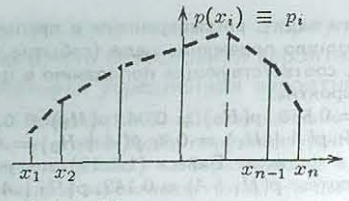


Рис. 1.2.1. Многоугольник распределения

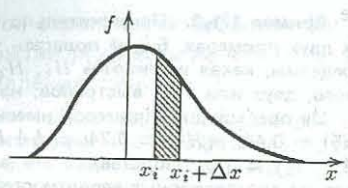


Рис. 1.2.2. Примерный вид плотности вероятности

Непрерывная случайная величина X принимает бесчисленное множество значений, причем вероятность попадания ее в любую бесконечно малую область бесконечно мала. Закон распределения вероятностей непрерывной случайной величины задается плотностью распределения вероятностей (в дальнейшем будем называть просто плотностью вероятности) $f(x)$. *Плотность вероятности* определяется как предел отношения вероятности попадания случайной величины X в малый интервал $[x, x+\Delta x]$ к длине этого интервала Δx при $\Delta x \to 0$:

$$f(x) = \lim_{\Delta x \to 0} \frac{\mathbf{P}\{x \leqslant X < x + \Delta x\}}{\Delta x}. \qquad (1.2.3)$$

Плотность вероятности $f(x)$ является размерной функцией; ее размерность обратна размерности рассматриваемой случайной величины. Плотность вероятности обладает следующими тремя основными свойствами:

1) она неотрицательна: $f(x) \geqslant 0$;

2) нормирована на единицу:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1; \qquad (1.2.4)$$

3) вероятность попадания непрерывной случайной величины в интервал $[x_1, x_2)$ равна интегралу от плотности вероятности в этих пределах:

$$\mathbf{P}\{x_1 \leqslant X < x_2\} = \mathbf{P}\{X < x_2\} - \mathbf{P}\{X < x_1\} = \int_{-\infty}^{x_2} f(x)\,dx - \int_{-\infty}^{x_1} f(x)\,dx = \int_{x_1}^{x_2} f(x)\,dx. \qquad (1.2.5)$$

Примерный вид плотности вероятности приведен на рис. 1.2.2.

Универсальным законом распределения, одинаково пригодным для описания как дискретных, так и непрерывных случайных величин, является функция распределения вероятностей (в дальнейшем будем называть просто функцией распределения) случайной величины. *Функцией распределения случайной величины* X называется функция $F(x)$, равная вероятности того, что случайная величина X примет значение, меньшее, чем x:

$$F(x) = \mathbf{P}\{X < x\}, \quad -\infty < x < \infty. \qquad (1.2.6)$$

В ряде пособий по теории вероятностей (особенно в зарубежной литературе) функция распределения определяется неравенством

$$\tilde{F}(x) = \mathbf{P}\{X \leqslant x\}, \quad -\infty < x < \infty. \qquad (1.2.7)$$

В противоположность (1.2.6) в выражении (1.2.7) точка x включается в интервал возможных значений случайной величины X. Это обстоятельство не играет роли при определении функции распределения непрерывной случайной величины, но его необходимо учитывать при вычислении функции распределения дискретной случайной величины. Функции распределения $F(x)$ и $\tilde{F}(x)$ связаны между собой равенствами

$$\tilde{F}(x) = F(x+0); \quad F(x) = \tilde{F}(x-0). \qquad (1.2.8)$$

Зная распределение вероятностей дискретной случайной величины, можно построить ее функцию распределения

$$F(x) = \mathbf{P}\{X < x\} = \sum_{x_k < x} p_k, \qquad (1.2.9)$$

где неравенство $x_k < x$ под знаком суммы указывает, что суммирование распространяется на все те значения x_k, которые меньше x. При этом, когда текущая переменная x проходит через какое-нибудь из возможных значений дискретной величины X, функция распределения меняется скачкообразно на величину, равную вероятности этого значения p_k.

Применительно к непрерывным случайным величинам функция распределения и плотность вероятности связаны соотношениями

$$F(x) = \int_{-\infty}^{x} f(y)\,dy; \quad f(x) = \frac{dF(x)}{dx}. \qquad (1.2.10)$$

Функция распределения обладает следующими очевидными свойствами:

1) $F(-\infty) = \lim\limits_{x \to -\infty} F(x) = 0$;

2) $F(+\infty) = \lim\limits_{x \to \infty} F(x) = 1$;

3) $F(x)$ — неубывающая функция, т.е. $F(x_2) \geqslant F(x_1)$ при $x_2 > x_1$;

4) $\mathbf{P}\{x_1 \leqslant X < x_2\} = F(x_2) - F(x_1)$.

Функция распределения дискретной случайной величины представляет собой ступенчатую функцию со скачками в точках x_1, x_2 и т.д. Примерные графики функций распределения дискретной и непрерывной случайных величин приведены на рис. 1.2.3,*а* и *б*.

Дискретная случайная величина X может быть представлена в виде плотности вероятности, если учесть понятие *дельта-функции*

$$\delta(z - z_0) = \begin{cases} \infty & \text{при } z = z_0, \\ 0 & \text{при } z \neq z_0, \end{cases} \qquad (1.2.11)$$

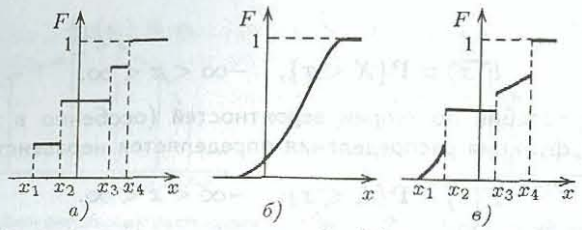


Рис. 1.2.3. Функции распределения дискретной (*а*), непрерывной (*б*) и смешаной величины (*в*)

причем $\delta(z) = \delta(-z)$.

Дельта-функция обладает следующими важными свойствами:

1) нормирована:

$$\int_{z_0-\alpha}^{z_0+\alpha} \delta(z-z_0)\,dz = 1, \quad \alpha > 0; \tag{1.2.12}$$

2) обладает «фильтрующим» свойством:

$$\int_{z_0-\alpha}^{z_0+\alpha} \varphi(z)\delta(z-z_0)\,dz = \varphi(z_0). \tag{1.2.13}$$

Из последнего соотношения следует, что спектр дельта-функции (преобразование Фурье) постоянен и равен единице:

$$\int_{-\infty}^{\infty} \delta(t)\mathrm{e}^{-\mathrm{i}\omega t}\,dt = 1.$$

где $\mathrm{i} = \sqrt{-1}$ – мнимая единица. Поэтому

$$\delta(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega t}\,d\omega. \tag{1.2.14}$$

С учетом понятия дельта-функции плотность вероятности дискретной случайной величины запишется в виде

$$f(x) = \sum_{k=1}^{n} p_k\delta(x-x_k), \tag{1.2.15}$$

где, как и выше, x_k — возможные значения случайной величины X; p_k — вероятности возможных значений x_k.

Случайная величина является *непрерывно-дискретной* (*смешанной*), если она в фиксированном числе точек $x_1, x_2, \ldots, x_n$ имеет конечные значения вероятностей $p_1, p_2, \ldots, p_n$, а на остальном интервале описывается некоторой функцией $f_1(x)$. Плотность вероятности такой случайной величины X определяется формулой

$$f(x) = f_1(x) + \sum_{k=1}^{n} p_k\delta(x-x_k). \tag{1.2.16}$$

Характер смешанной функции распределения виден из рис. 1.2.3,*в*.

Вместо плотности вероятности или функции распределения для описания вероятностных свойств случайной величины можно использовать некоторые функции, однозначно связанные с плотностью вероятности или функцией распределения. Наиболее продуктивной функцией является *характеристическая функция* $\theta(\vartheta)$, являющаяся (с точностью до знака над экспонентой) преобразованием Фурье от плотности вероятности $f(x)$

$$\theta(\vartheta) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}\vartheta x}\,dx = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\vartheta x}\,dF(x). \tag{1.2.17}$$

На основании обратного преобразования Фурье можно записать

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \theta(\vartheta)\mathrm{e}^{-\mathrm{i}\vartheta x}\,d\vartheta. \tag{1.2.18}$$

Применительно к дискретной случайной величине характеристическая функция имеет вид

$$\theta(\vartheta) = \int_{-\infty}^{\infty}\sum_{k=-\infty}^{\infty} p_k\delta(x-x_k)\mathrm{e}^{\mathrm{i}\vartheta x}dx = \sum_{k=-\infty}^{\infty} p_k\mathrm{e}^{\mathrm{i}\vartheta x_k}. \tag{1.2.19}$$

В теории вероятностей широко применяется понятие математического ожидания. *Математическим ожиданием* функции $\varphi(X)$ случайной величины X с плотностью вероятности $f(x)$ называется интеграл вида

$$\mathrm{M}\{\varphi(X)\} = \int_{-\infty}^{\infty} \varphi(x)f(x)\,dx. \tag{1.2.20}$$

Здесь и в дальнейшем символ $\mathrm{M}\{*\}$ означает операцию математического ожидания над величиной, указанной в фигурных скобках. Операция математического ожидания, по существу, есть осреднение рассматриваемой функции от случайной величины $\varphi(X)$ с весами, соответствующими плотности вероятности $f(x)$ этой случайной величины.

Характеристическую функцию (1.2.17) можно рассматривать как математическое ожидание случайной функции $\exp(\mathrm{i}\vartheta X)$, т.е.

$$\theta(\vartheta) = \mathrm{M}\{\exp(\mathrm{i}\vartheta X)\} = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}\vartheta x}\,dx.$$

Из определения характеристической функции следуют условия

$$|\theta(\vartheta)| \leqslant \int_{-\infty}^{\infty} f(x)\,dx = 1, \quad \theta(\vartheta)|_{\vartheta=0} = 1.$$

Если плотность вероятности — четная функция $f(x) = f(-x)$, то $\theta(\vartheta) = \theta(-\vartheta)$. Кроме того, для линейного преобразования вида $Y =$

$= a + bX$ характеристическая функция определяется как

$$\theta_Y(\vartheta) = \mathrm{M}\{\mathrm{e}^{\mathrm{i}\vartheta Y}\} = \mathrm{M}\{\mathrm{e}^{\mathrm{i}\vartheta(a+bX)}\} = \mathrm{e}^{\mathrm{i}a\vartheta}\theta_X(b\vartheta), \tag{1.2.21}$$

где $\theta_X(\vartheta)$– характеристическая функция случайной величины X.

В ряде задач теории вероятностей приходится находить плотность вероятности и характеристическую функцию случайной величины X, являющейся суммой n независимых случайных величин X_k:

$$X = X_1 + X_2 + \ldots + X_n. \tag{1.2.22}$$

Обозначая через $f_{X_k}(x)$, $\theta_{X_k}(\vartheta)$ плотность вероятности и характеристическую функцию слагаемого X_k, для характеристической функции случайной величины X можно написать

$$\theta_X(\vartheta) = \mathrm{M}\left\{\exp\left(\mathrm{i}\vartheta\sum_{k=1}^{n} X_k\right)\right\} = \mathrm{M}\left\{\prod_{k=1}^{n}\exp(\mathrm{i}\vartheta X_k)\right\}.$$

Для независимых случайных величин X_k математическое ожидание произведения случайных функций $\varphi(X_k)$ равно произведению математических ожиданий этих функций. Поэтому характеристическая функция суммы независимых случайных величин (1.2.22) определяется через произведение характеристических функций слагаемых:

$$\theta_X(\vartheta) = \prod_{k=1}^{n}\theta_{X_k}(\vartheta). \tag{1.2.23}$$

Если слагаемые X_k одинаково распределены, т.е. $f_{X_k}(x) \equiv f_{X_0}(x)$; $\theta_{X_k}(\vartheta) \equiv \theta_{X_0}(\vartheta)$, то характеристическая функция случайной величины X определяется формулой

$$\theta_X(\vartheta) = [\theta_{X_0}(\mathrm{i}\vartheta)]^n. \tag{1.2.24}$$

При этом плотность вероятности случайной величины X определяется через характеристическую функцию $\theta_X(\vartheta)$ согласно (1.2.18).

Вычисление плотности вероятности суммы независимых случайных величин по известным плотностям вероятности слагаемых называется *композицией законов распределения*.

В приложении 1 приведена таблица с распределениями ряда дискретных и непрерывных случайных величин [15, 36, 59, 61].

1.2.2. Числовые характеристики случайных величин

Во многих прикладных задачах трудно или даже невозможно определить функцию распределения или плотность вероятности случайной величины. Иногда в этом нет необходимости. В подобных ситуациях описание случайной величины при помощи законов распределения может быть заменено указанием отдельных параметров этих законов распределения. К подобным параметрам законов распределения относятся

числовые характеристики случайных величин. Приведем определения основных числовых характеристик случайной величины.

Полагая в (1.2.20) $\varphi(X) = X$, получим математическое ожидание случайной величины

$$m = \mathrm{M}\{X\} = \int_{-\infty}^{\infty} x f(x)\,dx. \tag{1.2.25}$$

Математическое ожидание случайной величины обычно называют *средним значением* случайной величины. Если интерпретировать $f(x)$ как плотность массы, распределенной вдоль оси x, то среднее значение случайной величины есть абсцисса центра тяжести этой массы. Иначе говоря, среднее значение случайной величины характеризует положение «центра тяжести» плотности вероятности $f(x)$. Для дискретной случайной величины, принимающей n возможных значений, среднее значение определяется соответственно как

$$m = \sum_{i=1}^{n} p_i x_i. \tag{1.2.26}$$

Кроме среднего значения, в качестве характеристики положения случайной величины X на оси абсцисс x применяются понятия медианы x_{med}, квантили x_p и моды x_{mod}.

Медианой (срединным значением) называется такое значение x_{med} случайной величины X, которое является корнем уравнения $F(x_{\mathrm{med}}) = 0{,}5$, т.е. медиана находится из уравнения

$$\mathbf{P}\{X \leqslant x_{\mathrm{med}}\} = \mathbf{P}\{X \geqslant x_{\mathrm{med}}\} = 0{,}5. \tag{1.2.27}$$

Иначе говоря, медианой является такое значение случайной величины, при котором для непрерывных случайных величин площадь под плотностью вероятности делится пополам. Если существует интервал (a, b), на котором $F(x) = 0{,}5$, то любая точка этого интервала является медианой.

Квантилью, соответствующей заданному уровню вероятности p, называется такое значение x_p, при котором функция распределения $F(x)$ принимает значение, равное p, т.е. квантиль порядка p есть корень уравнения

$$F(x_p) = \mathbf{P}\{X < x_p\} = p, \tag{1.2.28}$$

где p — некоторое заданное число ($0 \leqslant p \leqslant 1$). Нетрудно видеть, что медиана есть квантиль порядка $1/2$: $x_{\mathrm{med}} = x_{1/2}$. Часто порядок квантили выражают в виде процентов ($0\ \% \leqslant p \leqslant 100\ \%$). Так, медианой является 50%-ная квантиль.

Модой x_{mod} непрерывной случайной величины называется любая точка максимума плотности вероятности. Если плотность вероятности имеет один максимум, то она называется называется *унимодальной*, а если несколько, то *полимодальной*. Для дискретных «одновершинных» распределений в качестве моды обычно принимается то

значение $X = x_{\text{mod}}$, при котором вероятность $\mathrm{P}\{X = x_{\text{mod}}\}$ достигает своего максимума.

Для симметричных одномодальных плотностей вероятностей среднее значение, мода и медиана совпадают.

Более общими числовыми характеристиками случайной величины являются моменты. *Моментом* k-го порядка (в дальнейшем для сокращения иногда будем называть k-м моментом) случайной величины X относительно произвольного значения a называется математическое ожидание величины $(X-a)^k$:

$$m_k(a) = \mathrm{M}\{(X-a)^k\} = \int_{-\infty}^{\infty} (x-a)^k f(x)\,dx.$$

Момент k-го порядка, рассматриваемый относительно начала координат ($a = 0$), называется *начальным моментом* k-го порядка:

$$m_k = \int_{-\infty}^{\infty} x^k f(x)\,dx. \qquad (1.2.29)$$

Момент k-го порядка, рассматриваемый относительно среднего значения ($m_1 \equiv m$), называется *центральным моментом* этого же порядка:

$$\mu_k = \mathrm{M}\{(X-m)^k\} = \int_{-\infty}^{\infty} (x-m)^k f(x)\,dx. \qquad (1.2.30)$$

Центральные моменты можно выразить через начальные моменты, используя формулу бинома Ньютона для разложения $(X-m)^k$:

$$\mu_k = \mathrm{M}\left\{\sum_{r=0}^{k}(-1)^r C_k^r X^r m^{k-r}\right\} = \sum_{r=0}^{k}(-1)^r C_k^r m_r m^{k-r}. \qquad (1.2.31)$$

Здесь C_k^r — число сочетаний из k элементов по r элементов:

$$C_k^r = \frac{k!}{r!(k-r)!} \equiv \binom{k}{r}, \qquad (1.2.32)$$

где $k! = 1\cdot 2\cdot 3\cdots k$, причем $0! = 1$, а для $k \gg 1$ справедлива приближенная формула (*формула Стирлинга*)

$$k! \approx \sqrt{2\pi k}\,k^k \mathrm{e}^{-k}. \qquad (1.2.33)$$

Формула Стирлинга имеет погрешность менее $(25/3k)$ %.

Из формулы (1.2.31), в частности, получаем выражения для первых четырех центральных моментов:

$$\mu_0 = 1;\quad \mu_1 = 0;\quad \mu_2 = m_2 - m^2;\quad \mu_3 = m_3 - 3mm_2 + 2m^3;$$
$$\mu_4 = m_4 - 4mm_3 + 6m^2 m_2 - 3m^4. \qquad (1.2.34)$$

Из моментов высшего порядка наиболее часто применяются второй, третий и четвертый центральные моменты, причем наиболее широко используется второй центральный момент

$$\mu_2 = \mathrm{M}\{(X-m)^2\} \equiv D = \int_{-\infty}^{\infty} (x-m)^2 f(x)\,dx, \qquad (1.2.35)$$

называемый *дисперсией* случайной величины X. Дисперсия определяет рассеяние (разброс) значений случайной величины относительно ее среднего значения. Усредненное значение этого рассеяния случайной величины характеризуют *средним квадратическим отклонением*, численно равным положительному значению квадратного корня от дисперсии:

$$\sigma = \sqrt{D}. \qquad (1.2.36)$$

С центральным моментом третьего порядка μ_3 связан *коэффициент асимметрии*

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} (x-m)^3 f(x)\,dx. \qquad (1.2.37)$$

Коэффициент асимметрии является количественной мерой асимметрии плотности вероятности (распределения вероятностей в дискретном случае) относительно среднего значения.

На рис. 1.2.5 приведены две кривые плотности вероятностей, одна из которых (*а*) имеет более пологий спад справа от среднего значения, а вторая (*б*) — более пологий спад слева от среднего значения. Для первой кривой в формуле (1.2.37) кубы положительных отклонений превысят кубы отрицательных отклонений, так что коэффициент γ_1 будет положителен. В подобной ситуации говорят, что плотность вероятности обладает положительной асимметрией. Если коэффициент асимметрии принимает отрицательное значение, то асимметрия отрицательная.

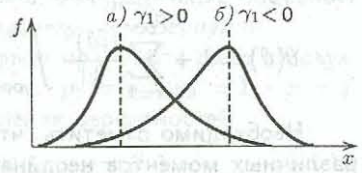


Рис. 1.2.5. Две асимметричные плотности вероятности

Для унимодальных распределений в качестве величины, характеризующей вид плотности вероятности около максимального значения, используют величину

$$\gamma_2 = (\mu_4/\mu_2^2) - 3, \qquad (1.2.38)$$

которую называют *коэффициентом эксцесса*. Чем острее вершина плотности вероятности, тем больше коэффициент эксцесса. Как будет показано ниже, нулевое значение коэффициента эксцесса соответствует нормальному (гауссовскому) распределению случайной величины.

Начальные моменты относительно несложно вычисляются, если известна характеристическая функция. Найдем эту связь, для чего продифференцируем характеристическую функцию (1.2.17) k раз по пере-

менной ϑ:

$$\frac{d^k\theta(\vartheta)}{d\vartheta^k} = \int_{-\infty}^{\infty} (\mathrm{i}x)^k f(x)\exp(\mathrm{i}\vartheta x)\,dx,$$

откуда

$$\int_{-\infty}^{\infty} x^k f(x)\exp(\mathrm{i}x)\,dx = \mathrm{i}^{-k}\frac{d^k\theta(\vartheta)}{d\vartheta^k}.$$

Отсюда нетрудно видеть, что k-я производная от характеристической функции при $\vartheta = 0$ представляет k-й начальный момент

$$m_k = \int_{-\infty}^{\infty} x^k f(x)\,dx = \mathrm{i}^{-k}\left.\frac{d^k\theta(\vartheta)}{d\vartheta^k}\right|_{\vartheta=0}. \qquad (1.2.39)$$

В частном случае среднее значение

$$m = \frac{1}{\mathrm{i}}\left.\frac{d\theta(\vartheta)}{d\vartheta}\right|_{\vartheta=0}. \qquad (1.2.40)$$

Знание начальных моментов позволяет восстановить характеристическую функцию, а следовательно, и плотность вероятности случайной величины. Действительно, разлагая экспоненту под знаком интеграла (1.2.17) в степенной ряд, имеем

$$\theta(\vartheta) = 1 + \sum_{k=1}^{\infty}\frac{(\mathrm{i}\vartheta)^k}{k!}\int_{-\infty}^{\infty} x^k f(x)\,dx = 1 + \sum_{k=1}^{\infty} m_k\frac{(\mathrm{i}\vartheta)^k}{k!}. \qquad (1.2.41)$$

Необходимо отметить, что в описание случайной величины вклад различных моментов неодинаков. Он тем выше, чем меньше порядок момента. Это наглядно можно показать на примере вычисления математического ожидания произвольной, но достаточно «плавной» функции $\varphi(x)$, полагая, что плотность вероятности $f(x)$ имеет наиболее существенное значение на интервале порядка $\sigma = \sqrt{D_X}$ около среднего значения m_X. Функцию $\varphi(x)$ в рассматриваемом интервале представим в виде степенного ряда в окрестности точки m_X:

$$\varphi(X) = \varphi(m_X) + \varphi'(m_X)(X-m) + \frac{1}{2}\varphi''(m_X)(X-m_X)^2 + \ldots.$$

Отсюда

$$\mathrm{M}\{\varphi(X)\} \approx \varphi(m_X) + \frac{1}{2}\varphi''(m_X)D_X.$$

Таким образом, видно, что среднее значение рассматриваемой функции от случайной величины приближенно определяется средним значением и дисперсией исходной случайной величины X.

Кроме моментов, характеризующих распределение случайных величин, часто применяется другая совокупность постоянных, называемых *кумулянтами* или *семиинвариантами* [15, 36, 37, 59, 60 и др.],

$$\kappa_k = \mathrm{i}^{-k}\left[\frac{d^k}{d\vartheta^k}\ln\theta(\vartheta)\right]_{\vartheta=0}, \qquad (1.2.42)$$

которые взаимооднозначно связаны с моментами распределения. В указанной литературе приведены соотношения, связывающие кумулянты с моментами и наоборот.

1.3. Законы распределения дискретных случайных величин

Из большого разнообразия законов распределения вероятностей рассматриваются наиболее часто встречающиеся в прикладных задачах теории вероятностей и математической статистики.

1.3.1. Закон распределения Бернулли

Простейшей дискретной случайной величиной X является случайная величина, принимающая два значения: x_1 с вероятностью p_1 и x_2 с вероятностью $p_2 = 1 - p_1$. Подобная ситуация характерна для систем передачи кодированных сообщений (бинарных сообщений), когда элементарными сигналами являются $x_1 = 1$ и $x_0 = 0$. Распределение подобной случайной величины часто называют *распределением Бернулли*.

Итак, пусть случайная величина X может принимать одно из двух возможных значений 1 или 0 с вероятностями $p_1 = p$ и $p_0 = 1 - p = q$ соответственно. Выражение для распределения вероятностей

$$p(x) \equiv p_x = p^x(1-p)^{1-x}, \quad 0 \leqslant p \leqslant 1; \quad x = 0; 1. \qquad (1.3.1)$$

Функция распределения определяется как

$$F(x) = \begin{cases} \mathbf{P}\{X < 0\} = 0 & \text{при } x < 0; \\ \mathbf{P}\{0 \leqslant X < 1\} = 1 - p & \text{при } 0 \leqslant x < 1; \\ \mathbf{P}\{X \geqslant 1\} & \text{при } x \geqslant 1. \end{cases} \qquad (1.3.2)$$

На рис. 1.3.1 приведены графики распределения вероятностей и функции распределения бернуллиевской случайной величины с $p = 0{,}75$.

Используя понятия дельта-функции (1.2.11), можно записать аналитические выражения для плотности вероятности подобной дискретной случайной величины:

$$f(x) = (1-p)\delta(x) + p\delta(x-1).$$

Согласно (1.2.19) характеристическая функция рассматриваемой случайной величины имеет вид

$$\theta(\vartheta) = \sum_{x=0}^{1} p_x \mathrm{e}^{\mathrm{i}\vartheta x} = q + p\mathrm{e}^{\mathrm{i}\vartheta} = 1 + p(\mathrm{e}^{\mathrm{i}\vartheta} - 1). \qquad (1.3.3)$$

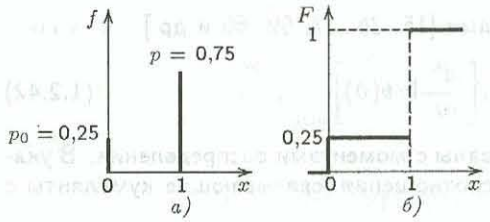


Рис. 1.3.1. Графики распределения вероятностей (*а*) и функции распределения (*б*) бернуллиевской случайной величины

Среднее значение и дисперсия этой случайной величины равны

$$m = \sum_{x=0}^{1} x p_x = 0 \cdot q + 1 \cdot p = p; \tag{1.3.4}$$

$$D = \mathbf{M}\{(X-p)^2\} = \sum_{x=0}^{1} p_x (x-m)^2 =$$

$$= (1-p)(0-p)^2 + p(1-p)^2 = p(1-p). \tag{1.3.5}$$

1.3.2. Биномиальный закон распределения

Пусть производится серия из n одинаковых независимых опытов, в каждом из которых событие A может произойти с вероятностью p и не произойти с вероятностью $q = 1 - p$. В дальнейшем удобнее событию A приписать величину, равную единице, а противоположному событию $\bar{A}$ приписать нуль, т.е. $\mathbf{P}\{A = 1\} = p(A) = p$, $\mathbf{P}\{\bar{A} = 0\} = p(\bar{A}) = q = 1 - p$. Необходимо определить вероятность того, что из n опытов в k опытах произойдет событие A.

Припишем событию A_i в i-м опыте случайную величину X_i, принимающую значения 1 или 0 с вероятностями $\mathbf{P}\{X_i = 1\} = p$, $\mathbf{P}\{X_i = 0\} = q = 1 - p$. Тогда случайная величина

$$K = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i \tag{1.3.6}$$

определяет суммарное число появления событий A в серии из n независимых опытов и является суммой n независимых бернуллиевских случайных величин. Случайная величина K может принимать значение $0, 1, 2, \ldots, n$

Пусть случайная величина K принимает конкретное значение $k \neq 0$ из n возможных. Отношение k/n числа положительных исходов k в данном опыте к общему числу опытов n называют *относительной частотой* (*частостью*) *события* (в дальнейшем будем говорить просто частота).

Определим вероятность $\mathbf{P}\{K = k\} = p_n(k)$ того, что случайная величина K примет конкретное значение $k = 0, 1, 2, \ldots, n$ из $n + 1$ возможных. Для этого используем характеристическую функцию для распределения Бернулли (1.3.3). Согласно (1.2.24) характеристическая

функция для случайной величины K

$$\theta(\vartheta) = (p e^{i\vartheta} + q)^n = \sum_{k=0}^{n} C_n^k p^k q^{(n-k)} e^{i\vartheta k}. \tag{1.3.7}$$

В то же время эта характеристическая функция связана с распределением вероятностей $p_n(k)$ формулой

$$\theta(\vartheta) = \mathbf{M}\{e^{i\vartheta K}\} = \sum_{k=0}^{n} p_n(k) e^{i\vartheta k}.$$

Сравнивая последнее соотношение с (1.3.7), имеем

$$p_n(k) = C_n^k p^k q^{(n-k)}, \quad k = 0, 1, 2, \ldots \tag{1.3.8}$$

Это распределение называют *биномиальным*, поскольку вероятности $p_n(k)$ можно рассматривать как коэффициенты при k-й степени разложения бинома $(q + pz)^n$ в ряд по степеням величины z:

$$(q + pz)^n = \sum_{k=0}^{n} C_n^k p^k q^{(n-k)} z^k.$$

Для изучения характера изменения распределения $p_n(k)$ от числа положительных исходов k рассмотрим отношение

$$\frac{p_n(k)}{p_n(k+1)} = \frac{C_n^k p^k q^{(n-k)}}{C_n^{(k+1)} p^{(k+1)} q^{(n-k-1)}} = \frac{(k+1)}{(n-k)} \frac{q}{p}. \tag{1.3.9}$$

Отсюда вытекает, что

$$p_n(k+1) < p_n(k), \quad \text{если} \quad k > np - q;$$

$$p_n(k+1) = p_n(k), \quad \text{если} \quad k = np - q;$$

$$p_n(k+1) > p_n(k), \quad \text{если} \quad k < np - q.$$

Если $np - q$ является целым числом, то максимальное значение вероятность $p_n(k)$ принимает для двух значений переменной k, а именно для $k_1 = np - q$ и $k_2 = np - q + 1$. Если $np - q$ не является целым числом, то максимальное значение $p_n(k)$ достигается при k_3, равном наименьшему целому числу, но большему, чем k_1. Для большого числа возможных исходов $n \gg 1$ максимальное значение вероятность $p_n(k)$ достигает при $k_4 \approx np$, которое, как будет показано ниже, соответствует среднему значению числа положительных исходов m. Для $(np - q) < 0$ справедливо неравенство $p_n(0) > p_n(1) > \ldots > p_n(n)$. Если $(np - q) = 0$, то $p_n(0) = p_n(1) > p_n(2) > \ldots > p_n(n)$.

Согласно (1.3.9) биномиальное распределение легко табулируется, если известна одна из вероятностей $p_n(k)$:

$$p_n(k+1) = p_n(k) \frac{n-k}{k+1} \frac{p}{q}; \quad p_n(k-1) = p_n(k) \frac{k}{n-k+1} \frac{q}{p}. \tag{1.3.10}$$

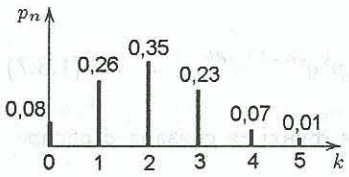


Рис. 1.3.2. Биноминальный закон распределения для $n = 5$, $p = 0{,}4$

Пример 1.3.1. Используя первое выражение (1.3.10), вычислим вероятности биномиального распределения для $p = 0{,}4$ и $n = 5$.

В качестве исходной вероятности возьмем $p_5(0) = C_5^0 \cdot 0{,}4^0 \cdot 0{,}6^5 = 0{,}07776$. Тогда в соответствии с первым выражением (1.3.10) получим $p_5(1) = 0{,}25920$; $p_5(2) = 0{,}34560$; $p_5(3) = 0{,}23040$; $p_5(4) = 0{,}07680$; $p_5(5) = 0{,}01024$; На рис. 1.3.2 показан график вероятностей полученного биномиального распределения.

В [8, 51 и др.] приведены таблицы значений вероятностей (1.3.8) для ряда параметров p и n. Там же имеются таблицы вероятностей

$$\mathbf{P}\{K \leqslant k\} = \sum_{j=0}^{k} p_n(j) = \sum_{j=0}^{k} C_n^j p^j (1-p)^{(n-j)}. \qquad (1.3.11)$$

Для определения среднего значения и дисперсии биномиальной случайной величины воспользуемся соотношениями (1.2.39), связывающими начальные моменты с производными от характеристической функции. Среднее значение

$$m = \frac{1}{\mathrm{i}} \frac{d\theta(\vartheta)}{d\vartheta}\bigg|_{\vartheta=0} = \left[\frac{1}{\mathrm{i}} n(q + p\mathrm{e}^{\mathrm{i}\vartheta})^{(n-1)} \mathrm{i} p \mathrm{e}^{\mathrm{i}\vartheta}\right]_{\vartheta=0} = np. \qquad (1.3.12)$$

Таким образом, если серия из n опытов повторяется много раз, то среднее число успехов (среднее число появления единиц) будет равно вероятности успеха в одном испытании p, умноженной на число опытов n. Аналогично вычисляется дисперсия биномиальной случайной величины $D = m_2 - m^2$. Здесь

$$m_2 = \frac{1}{\mathrm{i}^2} \frac{d^2\theta(\vartheta)}{d\vartheta^2}\bigg|_{\vartheta=0} = n(n-1)p^2 + np.$$

Поэтому

$$D = n(n-1)p^2 + np - n^2p^2 = np(1-p) = npq. \qquad (1.3.13)$$

Из отношения среднего квадратического отклонения к среднему числу успехов (положительных исходов) $\sqrt{D}/m = \sqrt{q/np}$ видно, что с увеличением числа опытов n относительное рассеяние величины положительных исходов k уменьшается, в то время как абсолютное рассеяние увеличивается.

Аналогичным образом можно вычислить и другие центральные моменты. В частности, третий и четвертый центральные моменты определяются формулами

$$\mu_3 = npq(q-p); \quad \mu_4 = 3(npq)^2 + npq(1-6pq). \qquad (1.3.14)$$

Соответственно коэффициенты асимметрии и эксцесса равны

$$\gamma_1 = \frac{\mu_3}{D^{3/2}} = \frac{q-p}{\sqrt{npq}}; \quad \gamma_2 = \frac{\mu_4}{D^2} - 3 = \frac{1-6pq}{npq}. \qquad (1.3.15)$$

Из последних формул видно, что если $p < 1/2$, то асимметрия положительная; если $p > 1/2$, асимметрия отрицательная; при $p = 1/2$ асимметрия равна нулю. Оба коэффициента (γ_1, γ_2) стремятся к нулю, если $n \gg 1$.

Важным свойством биномиального распределения является его воспроизводимость при композиции r независимых биномиально распределенных случайных величин X_l, $l = 1, 2, \ldots, r$, с одинаковыми элементарными вероятностями $p_l = p$ и разными значениями числа опытов n_l $(n_1, n_2, \ldots, n_r)$. При этих условиях характеристическая функция (1.2.23) суммы случайных величин $Y = \sum X_l$ конкретизируется и примет вид

$$\theta_Y(\vartheta) = \prod_{l=1}^{r} \theta_{X_l}(\vartheta) = (q + p\mathrm{e}^{\mathrm{i}\vartheta})^{n_1}(q + p\mathrm{e}^{\mathrm{i}\vartheta})^{n_2} \ldots (q + p\mathrm{e}^{\mathrm{i}\vartheta})^{n_r} =$$
$$= (q + p\mathrm{e}^{\mathrm{i}\vartheta})^{n=n_1+n_2+\ldots+n_r}.$$

Это свидетельствует о том, что результирующий закон распределения будет биномиальным с параметрами p и $n = n_1 + n_2 + \ldots + n_r$.

Определение вероятностей $p_n(k)$ при $n \gg 1$ представляет значительные вычислительные трудности. Для этого случая имеется асимптотическая формула биномиального распределения, получающаяся в результате предельного перехода исходного выражения (1.3.8). Приведем вывод асимптотической формулы [52 и др.], для чего в (1.3.8) обозначим $l = n - k$ и применим формулу Стирлинга (1.2.33), полагая $l \gg 1$. Получим

$$p(k) \approx \frac{1}{\sqrt{2\pi}} \left(\frac{n}{kl}\right)^{1/2} \left(\frac{np}{k}\right)^k \left(\frac{nq}{l}\right)^l. \qquad (1.3.16)$$

Введем новую переменную $x_{nk} = (k - np)/\sqrt{D} = (nq - l)/\sqrt{D}$. В дальнейшем полагается, что $n \gg 1$, а величина x_{nk} ограничена, т.е. $|x_{nk}| < c$, где $c > 0$. Тогда для $n \to \infty$ справедливы предельные значения: $k = np + x_{nk}\sqrt{D} \to \infty$, $l = nq - x_{nk}\sqrt{D} \to \infty$. При этом

$$\frac{kl}{n} = n\left(p + \frac{x_{nk}\sqrt{D}}{n}\right)\left(q - \frac{x_{nk}\sqrt{D}}{n}\right) \to npq = D.$$

Для получения предельного значения произведения последних двух сомножителей (1.3.16) прологарифмируем это произведение и учтем разложение логарифма в степенной ряд

$$\ln(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \frac{1}{4}z^4 + \ldots, \quad -1 < z \leqslant 1. \qquad (1.3.17)$$

Ограничиваясь первыми двумя членами разложения, получим

$$\ln\left[\left(\frac{np}{k}\right)^k\left(\frac{nq}{l}\right)^l\right] = -(np+x_{nk}\sqrt{D})\ln\left(1+x_{nk}\frac{\sqrt{D}}{np}\right) -$$

$$-(nq-x_{nk}\sqrt{D})\ln\left(1-x_{nk}\frac{\sqrt{D}}{nq}\right) \approx -\frac{x_{nk}^2}{2}.$$

В результате выражение для вероятностей $p_n(k)$ примет вид

$$p_n(k) \approx \frac{1}{\sqrt{npq}} f\left(\frac{k-np}{\sqrt{npq}}\right) = \frac{1}{\sqrt{2\pi npq}}\exp\left(-\frac{(k-np)^2}{2npq}\right), \quad (1.3.18)$$

где

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right). \quad (1.3.19)$$

Эту формулу называют *локальной формулой Муавра–Лапласа*. Непрерывную кривую, описываемую формулой (1.3.19), обычно называют *гауссовой* или *нормальной кривой* (кривой нормального распределения). Заметим, что разность $(k-np)$ представляет собой отклонение числа положительных исходов от его наиболее вероятного значения.

Для вычисления вероятностей типа

$$\mathbf{P}\{k_1 \leqslant K \leqslant k_2\} = \sum_{k=k_1}^{k_2} p_n(k) \quad (1.3.20)$$

при $n \gg 1$ и сделанных выше ограничениях на величину $(k-np)/\sqrt{D}$ вместо $p_n(k)$ подставляется ее приближенное выражение, а сумма приближается с помощью вычисления площади, ограниченной непрерывной кривой (1.3.18) в пределах от $k_1 - 1/2$ до $k_2 + 1/2$ (рис. 1.3.3). Величины $1/2$ являются поправками на непрерывность при суммировании площадей прямоугольников, заменяющих дискретное распределение вероятностей их кусочно-непрерывным аналогом. В результате такого приближения имеем

$$\mathbf{P}\{k_1 \leqslant K \leqslant k_2\} \approx \frac{1}{\sqrt{2\pi}\sqrt{npq}}\int_{k_1-1/2}^{k_2+1/2}\exp\left(-\frac{(k-np)^2}{2npq}\right)dk.$$

Осуществляя замену переменных $[(k-np)/\sqrt{npq}] = t$, получим

$$\mathbf{P}\{k_1 \leqslant K \leqslant k_2\} \approx \Phi\left(\frac{k_2+1/2-np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1-1/2-np}{\sqrt{npq}}\right). \quad (1.3.21)$$

Здесь $\Phi(x)$ — табулированный интеграл [8, 51, 72 и др.]:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp(-z^2/2)\,dz; \quad \Phi(-x) = 1-\Phi(x). \quad (1.3.22)$$

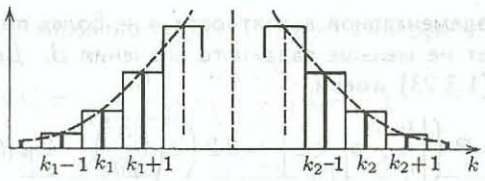


Рис. 1.3.3. Аппроксимирующая кривая

Выражение (1.3.21) часто называют *интегральной формулой Муавра–Лапласа*.

Для конечных значений n формулы (1.3.18) и (1.3.21) являются приближенными, и поэтому при вычислениях по этим формулам надо относиться с осторожностью. Более точные приближения приведены в [8]. При этом для фиксированных значений $n \gg 1$ наименьшая точность будет при $p = q = 1/2$. Для некоторых прикладных задач удовлетворительная точность вычислений искомых вероятностей в диапазоне $(np - 3\sqrt{npq} \leqslant k \leqslant (np + 3\sqrt{npq})$ по формулам (1.3.18) и (1.3.21) обеспечивается уже при $npq > 9$. Применение приведенных приближенных формул недопустимо в случае, когда асимметрия биномиального распределения очень велика. В частности, применение приближенных формул недопустимо, когда $p_n(k)$ монотонно убывает или монотонно возрастает. Эта ситуация соответствует $(np - q) < 0$ и $(np - q) > (n-1)$), что, в свою очередь, накладывает ограничения на элементарную вероятность p: $p < 1/(n+1)$ и $p > n/(n+1)$.

Интегральная формула Муавра–Лапласа позволяет оценить вероятность модуля отклонения частоты события k/n от его вероятности p на величину, меньшую ε, где ε — некоторое заранее заданное постоянное число. Учитывая (1.3.18), (1.3.20) и (1.3.21), имеем

$$\mathbf{P}\left\{\left|\frac{k}{n}-p\right| < \varepsilon\right\} = \mathbf{P}\left\{-\varepsilon\sqrt{\frac{n}{pq}} < x = \frac{k-np}{\sqrt{npq}} < \varepsilon\sqrt{\frac{n}{pq}}\right\} \approx$$

$$\approx \frac{1}{\sqrt{2\pi}}\int_{-\varepsilon\sqrt{n/pq}}^{\varepsilon\sqrt{n/pq}}\exp(-x^2/2)\,dx = 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right) - 1. \quad (1.3.23)$$

Если $n \to \infty$, то $\Phi\left(\varepsilon\sqrt{n/pq}\right) \to 1$ и, следовательно

$$\lim_{n\to\infty}\mathbf{P}\left\{\left|\frac{k}{n}-p\right| < \varepsilon\right\} = 1. \quad (1.3.24)$$

Предельное соотношение (1.3.24), являясь частным случаем закона больших чисел, впервые было получено Я. Бернулли. Оно называется *теоремой (формулой) Бернулли*. Ниже в п. 2.4 будет дано более общее доказательство этой формулы.

Приближенное соотношение (1.3.23) позволяет найти наименьшее число испытаний $n_{\min}$, при котором вероятность модуля отклонения ча-

стоты k/n от элементарной вероятности p не более положительной величины ε будет не меньше заданного значения β. Действительно, из соотношения (1.3.23) имеем

$$\mathbf{P}\left\{\left|\frac{k}{n}-p\right|<\varepsilon\right\}\approx 2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)-1\geqslant\beta. \qquad (1.3.25)$$

Для заданных численных значений p,q,ε и β с использованием таблиц интеграла (1.3.22) находится искомое значение $n_{\min}$. Если величину ε задать в долях $g>0$ от среднего квадратического отклонения ($\varepsilon=g\sqrt{npq}$), последнее соотношение для вычисления $n_{\min}$ можно представит в виде

$$\Phi(gn)\geqslant(1+\beta)/2. \qquad (1.3.26)$$

Например, пусть $\beta=0{,}99$; $g=0{,}1$. Так как $(1+\beta)/2=0{,}995$, из табл. 1б Приложения 2 для $\Phi(gn)=0{,}995$ находим значение $gn=2{,}57$. Отсюда получаем $n_{\min}\geqslant 26$.

Более точное выражение для $p_n(k)$ может быть получено при использовании дальнейших приближений в разложении (1.3.17). В частности, как будет показано ниже (п. 2.3.1), если учитывать только коэффициент асимметрии, то приближенно можно записать

$$p_n(k)\approx\frac{1}{\sqrt{2\pi D}}\exp\left(-\frac{z^2}{2D}\right)\left[1+\frac{1}{2}(p-q)\frac{z}{D}-\frac{1}{6}(p-q)\frac{z^3}{D^2}\right], \qquad (1.3.27)$$

где $z=k-np$, $D=npq$. При этом следует иметь в виду, что при такой аппроксимации нарушается свойство положительной определенности, так как при больших значениях z аппроксимирующая кривая может принимать отрицательные значения, недопустимые для распределения вероятностей.

Нормированные функции $f(x)=(2\pi)^{-1/2}\exp(-x^2/2)$ и $\Phi(x)$ широко используются в прикладных задачах теории вероятностей и математической статистики; таблицы этих функций приведены практически в большинстве пособий по теории вероятностей и математической статистике. Сокращенные варианты таблиц функции $f(x)$ и интеграла $\Phi(x)$ даны в табл. 1а и 1б Приложения 2.

1.3.3. Распределение Пуассона

Дискретное биномиальное распределение при $n\gg 1$ может быть преобразовано с помощью предельного перехода в новое распределение, когда вероятность появления события p очень мала и сравнима с величиной $\approx 1/n$, т.е. $p=\lambda/n$, где λ — некоторая положительная константа, а n — количество испытаний. Найдем предельное распределение $p_n(k)$ при $n\to\infty$, т.е. в случае, когда элементарная вероятность p стремится

к нулю. Исходную вероятность $p_n(k)\equiv p(k)$ запишем в виде

$$p(k)=\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}=$$

$$=\frac{(n-k+1)(n-k+2)\cdots(n-1)n}{k!}\left(1-\frac{\lambda}{n}\right)^n\frac{\lambda^k/n^k}{(1-\lambda/n)^k}=$$

$$=\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\frac{\left(1-\frac{k-1}{n}\right)\left(1-\frac{k-2}{n}\right)\cdots\left(1-\frac{1}{n}\right)}{(1-\lambda/n)^k}. \qquad (1.3.28)$$

При $n\to\infty$ ($n\gg 1$)

$$\lim_{n\to\infty}(1-\lambda/n)^n=\mathrm{e}^{-\lambda};\quad\lim_{n\to\infty}(1-\lambda/n)^k=1;$$

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)=1.$$

В результате предельного перехода имеем

$$p(k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda};\quad k=0,1,\ldots \qquad (1.3.29)$$

Анализ этого распределения показывает, что при $\lambda\leqslant 1$ с увеличением k величина $\lambda^k/k!$ уменьшается, а следовательно, с увеличением k от 0 до ∞ вероятность $p(k)$ убывает. Если $\lambda>1$, то с увеличением k множитель $\lambda^k/k!$ сначала возрастает, а затем убывает; поэтому и распределение при $\lambda>1$ сначала возрастает, а затем убывает. Это распределение называется *распределением Пуассона*. Оно дискретно и зависит от одного параметра λ, который всегда положителен. При этом нетрудно убедиться, что

$$\sum_{k=0}^{\infty}p(k)=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=1.$$

В [8, 51, 72 и др.] для ряда значений параметра λ приведены таблицы распределения вероятностей и функции распределения Пуассона, а также вероятности

$$\mathbf{P}\{K\leqslant k\}=\sum_{j=0}^{k}p(j)=\sum_{j=0}^{k}\exp(-\lambda)\frac{\lambda^j}{j!}. \qquad (1.3.30)$$

На рис. 1.3.4 приведены многоугольники распределения вероятностей пуассоновской случайной величины для значений $\lambda=0{,}5$; 1; 2 и 5.

Для большого, но конечного объема выборки n полученное распределение Пуассона является приближенным. Однако оно вполне приемлемо для вычисления вероятностей и параметров (например, среднего значения и дисперсии) редких событий, происходящих во времени или в

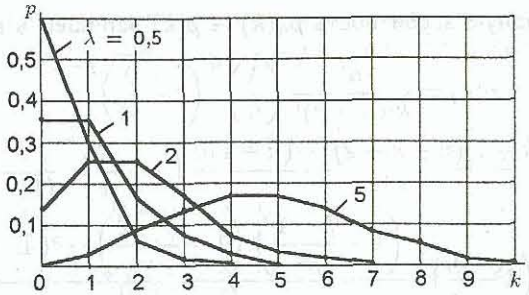


Рис. 1.3.4. Пуассоновский закон распределения

пространстве случайным образом и независимо друг от друга. При этом считается, что вероятность попадания точки в некоторую область зависит только от размеров области, а не от ее положения. Подобные задачи возникают при оценках надежности работы систем или их элементов, явлений радиоактивного распада, систем массового обслуживания и т.п.

Распределение Пуассона также часто используется для приближенного вычисления вероятностей биномиального распределения (1.3.8), когда n велико ($n \geqslant 50$), а параметр p мал. Возникающая при этом приближении ошибка имеет порядок λ^2/n и при постоянном λ сходится к нулю с увеличением n [72].

Найдем характеристическую функцию пуассоновской случайной величины

$$\theta(\vartheta) = \sum_{k=0}^{\infty} p(k)\mathrm{e}^{-\mathrm{i}\vartheta k} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\mathrm{e}^{\mathrm{i}\vartheta k} =$$

$$= \mathrm{e}^{-\lambda} \sum_{k=0}^{\infty} \frac{[\lambda \exp(\mathrm{i}\vartheta)]^k}{k!} = \exp[\lambda(\mathrm{e}^{\mathrm{i}\vartheta} - 1)]. \qquad (1.3.31)$$

Для вычисления среднего значения и дисперсии воспользуемся соотношениями (1.2.39), связывающими начальные моменты случайной величины с характеристической функцией. Среднее значение

$$m = \frac{1}{\mathrm{i}} \frac{d\theta(\vartheta)}{d\vartheta}\bigg|_{\vartheta=0} = \lambda \mathrm{e}^{\mathrm{i}\vartheta} \exp[\lambda(\mathrm{e}^{\mathrm{i}\vartheta} - 1)]|_{\vartheta=0} = \lambda. \qquad (1.3.32)$$

Второй начальный момент определяется формулой

$$m_2 = \frac{1}{\mathrm{i}^2} \frac{d^2(\vartheta)}{d\vartheta^2}\bigg|_{\vartheta=0} = \lambda \mathrm{e}^{\mathrm{i}\vartheta}\{[1 + \lambda \mathrm{e}^{i\vartheta}] \exp[\lambda(\mathrm{e}^{\mathrm{i}\vartheta} - 1)]\}|_{\vartheta=0} = \lambda + \lambda^2.$$

Из последнего выражения находим дисперсию пуассоновской случайной величины

$$D = m_2 - m^2 = \lambda + \lambda^2 - \lambda^2 = \lambda. \qquad (1.3.33)$$

Аналогично можно вычислить третий и четвертый центральные моменты:

$$\mu_3 = \lambda; \quad \mu_4 = \lambda + 3\lambda^2. \qquad (1.3.34)$$

При этом коэффициенты асимметрии и эксцесса равны

$$\gamma_1 = \mu_3/D^{3/2} = \lambda^{-1/2}; \quad \gamma_2 = \mu_4/D^2 - 3 = 1/\lambda. \qquad (1.3.35)$$

Таблица 1.3.1

k	3	4	6	8	10	12	14	16	17
$10^{-4}p(k)$	76	189	631	1126	1251	948	521	217	127
$10^{-4}p_N(k)$	109	207	570	1034	1262	1034	570	207	109
$\delta\%$	43,4	9,5	9,7	8,7	0,9	9,1	9,4	4,6	14,2

Из приведенных формул для моментов пуассоновского распределения видно, что дисперсия, среднее значение и третий центральный момент равны между собой, а коэффициенты асимметрии и эксцесса всегда положительны. При этом, если параметр распределения $\lambda \gg 1$ ($\lambda \to \infty$), то оба коэффициента стремятся к нулю. Последнее означает, что при $\lambda \gg 1$ график пуассоновского распределения симметричен относительно среднего значения $m = \lambda$ и в окрестности максимального значения (например, для $|k - \lambda| \leqslant \sqrt{\lambda}$) может быть аппроксимирован графиком нормальной кривой (1.3.18) со средним значением и дисперсией, равными λ.

В табл. 1.3.1 для $\lambda = 10$ приведены точные значения вероятностей $p(k)$, вычисленные по таблицам пуассоновского распределения [8], и вероятности $p_N(k)$, полученные на основе нормального приближения. Там же даны относительные погрешности (в процентах) $\delta[p(k)]\ \% = 100[p(k) - p_N(k)]/p(k)$.

Из приведенных данных видно, что при $\lambda = 10$ приемлемая погрешность не более 10% сохраняется только для $4 \leqslant k \leqslant 16$.

Для пуассоновского распределения справедливо утверждение: если независимые случайные величины X_l, $l = 1, 2, \ldots, r$, распределены по законам Пуассона с параметрами λ_l соответственно, то их сумма также будет распределена по закону Пуассона с параметром $\lambda = \sum_{l=1}^{r} \lambda_l$.

Данное утверждение доказывается аналогично изложенному для биномиального распределения, а именно характеристическая функция суммы независимых пуассоновских случайных величин согласно (1.2.23)

$$\theta_Y(\vartheta) = \mathrm{M}\left\{\exp\left(\mathrm{i}\vartheta \sum_{l=1}^{r} X_l\right)\right\} = \prod_{l=1}^{r} \theta_{X_l}(\vartheta) =$$

$$= \prod_{l=1}^{r} \exp[\lambda_l(\mathrm{e}^{\mathrm{i}\vartheta} - 1)] = \exp\left[(\mathrm{e}^{\mathrm{i}\vartheta} - 1)\sum_{l=1}^{r} \lambda_i\right] = \exp[\lambda(\mathrm{e}^{\mathrm{i}\vartheta} - 1)].$$

Это свидетельствует о том, что композиция l пуассоновских случайных величин также будет распределена по пуассоновскому закону с параметром $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_r$.

Как показано ниже (п. 2.3.1), для больших значений $\lambda \gg 1$ справедливо приближенное представление пуассоновского распределения выражением (с учетом асимметрии $\gamma_1 = \lambda^{-1/2}$)

$$p(k) \approx \frac{\exp[-(k-\lambda)^2/2\lambda]}{\sqrt{2\pi\lambda}} \left[1 - \frac{k-\lambda}{2\lambda} + \frac{(k-\lambda)^3}{6\lambda^2}\right]. \qquad (1.3.36)$$

Расчеты показывают, что приведенное приближение оказывается приемлемым для прикладных задач уже при $\lambda \geqslant 5$ [13]. Однако данному приближению свойственны недостатки, указанные для подобного приближения биномиального распределения (1.3.27).

Пример 1.3.2. На пункт сбора информации приходят сообщения нормализованной длительности $\tau = 10$ с. При этом предполагается, что сообщения по времени не перекрываются. В течение пяти часов пришло 36 сообщений. Определить вероятность того, что в течение одного часа будет не менее трех сообщений.

Число независимых отрезков времени ν, соответствующее пяти часам приема сообщений, равно $\nu = 5 \cdot 3600 : 10 = 1800$. При этом $p = 36 : 1800 = 0{,}02$ — вероятность появления одного сообщения в течение одного отрезка времени. Одному часу приема соответствует $n = 360$ независимых элементов (отрезков времени). Число сообщений k из $n = 360$ возможных распределено по биномиальному закону. Поэтому вероятность того, что в течение одного часа будет не менее трех сообщений, равна $\mathbf{P}\{K \geqslant 3\} = 1 - \sum_{k=0}^{2} p_n(k)$. Однако при $n \gg 1$ применение биномиального распределения для подобной задачи весьма затруднительно. В связи с этим для определения вероятности $\mathbf{P}\{k \geqslant 3\}$ можно воспользоваться распределением Пуассона, применительно к которому для $p = 0{,}02$ и $n = 360$ параметр $\lambda = 7{,}2$. Тогда вероятность того, что в течение одного часа будет не меньше трех сообщений

$$\mathbf{P}\{K \geqslant 3\} = 1 - [p(0) + p(1) + p(2)] = 1 - [1 + 7{,}2 + 25{,}92]\exp(-7{,}2) =$$
$$= 1 - 34{,}12 \cdot 0{,}000747 \approx 1 - 0{,}0255 \approx 0{,}975.$$

1.3.4. Равномерный закон распределения

Закон распределения дискретной случайной величины X, принимающей n конкретных значений x_k, $k = 1, 2, \ldots, n$, с одинаковыми вероятностями

$$p_i = 1/n, \quad k = 1, 2, \ldots, n, \qquad (1.3.37)$$

называется *равномерным законом распределения*. Среди подобных случайных величин наибольшее применение получили случайные величины, принимающие значения $x_k = k$, $k = 1, 2, \ldots, n$. С использованием дельта-функции (1.2.11) плотность вероятности записываются в виде

$$f(x) = \frac{1}{n}\sum_{k=1}^{n} \delta(x - k). \qquad (1.3.38)$$

Функция распределения рассматриваемой дискретной случайной величины является ступенчатой функцией со скачками в точках $1, 2, \ldots, n$. Для вычисления среднего значения, дисперсии и характеристической функции подобной дискретной случайной величины используем известные из алгебры выражения для сумм:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}; \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}; \qquad (1.3.39)$$

$$\sum_{k=1}^{n} aq^{k-1} = \frac{a(1-q^n)}{1-q}. \qquad (1.3.40)$$

С учетом (1.3.39) среднее значение и дисперсия рассматриваемой дискретной случайной величины равны соответственно

$$m = \frac{1}{n}\sum_{k=1}^{n} k = \frac{n+1}{2}; \qquad (1.3.41)$$

$$D = \sum_{k=1}^{n} p_k x_i^2 - \left[\frac{n+1}{2}\right]^2 = \frac{1}{n}\sum_{k=1}^{n} i^2 - \left[\frac{n+1}{2}\right]^2 = \frac{(n^2-1)}{12}. \qquad (1.3.42)$$

Принимая во внимание (1.3.40), общая формула для характеристической функции (1.2.19) конкретизируется и принимает вид

$$\theta(\vartheta) = \sum_{k=1}^{n} p_k e^{i\vartheta k} = \frac{\exp(i\vartheta)[1 - \exp(i\vartheta n)]}{n[1 - \exp(i\vartheta)]}. \qquad (1.3.43)$$

1.4. Распределение непрерывных случайных величин

1.4.1. Нормальный закон распределения

Непрерывная случайная величина X имеет *нормальное распределение* с параметрами m и σ, если плотность вероятности ее дается формулой

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]. \qquad (1.4.1)$$

Нормальная плотность вероятности $f(x)$ — четная функция относительно величины m. Она стремится к нулю при $x \to \pm\infty$, а ее максимальное значение (мода), равное $(\sqrt{2\pi}\sigma)^{-1}$, соответствует $x = m_{\text{mod}}$. Случайную величину, имеющую нормальную плотность вероятности, часто называют *гауссовской случайной величиной*.

Гауссовские случайные величины наиболее часто встречаются в практике вероятностно-статистических исследований. В теории вероятностей доказывается (п. 2.4), что к нормальному распределению сходится распределение суммы независимых или слабо зависимых случайных величин с конечными моментами при неограниченном увеличении их числа.

Функция распределения гауссовской случайной величины имеет вид

$$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\exp\left[-\frac{(y-m)^2}{2\sigma^2}\right]dy=$$
$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(x-m)/\sigma}\exp\left(-\frac{z^2}{2}\right)dz=\Phi\left(\frac{x-m}{\sigma}\right). \qquad (1.4.2)$$

Здесь, как и выше, $\Phi(z)$ — интеграл (1.3.22). Гауссовскую случайную величину с параметрами $m=0$ и $\sigma^2=1$ обычно называют *нормированной* или *стандартной* гауссовской случайной величиной. Для такой случайной величины имеются таблицы нормированных плотностей вероятности $f(x)$ и функций распределения $F(x)$ (сокращенные варианты этих таблиц приведены в табл. 1а и 1б Приложения 2).

Пользуясь таблицами интеграла (1.3.22), легко вычисляется вероятность попадания гауссовской случайной величины в заданный интервал $[a,b]$:

$$\mathbf{P}\{a\leqslant X\leqslant b\}=\frac{1}{\sqrt{2\pi}\sigma}\int_{a}^{b}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]dx=$$
$$=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right). \qquad (1.4.3)$$

В частности, можно установить, что вероятность попадания гауссовской случайной величины в интервал $m\pm\sigma$ равна 0,68; в интервал $m\pm2\sigma$ равна 0,96 и в интервал $m\pm3\sigma$ равна 0,997.

В прикладных задачах математической статистики часто используются таблицы, позволяющие по заданному значению вероятности β определить значение переменной $X=x_\beta$, удовлетворяющее уравнению $\mathbf{P}\{X>x_\beta\}=1-F(x_\beta)=\beta$. Значения x_β называют *критическими значениями* или процентными точками, если β выражено в процентах. Применительно к нормированной случайной величине $z=(x-m)/\sigma$ по заданному значению вероятности β решается уравнение

$$\mathbf{P}\{Z>z_\beta\}=1-\Phi(z_\beta)=\frac{1}{\sqrt{2\pi}}\int_{z_\beta}^{\infty}\exp(-z^2/2)\,dz=\beta \qquad (1.4.4)$$

относительно переменной $z_\beta=(x_\beta-m)/\sigma$. В табл. 1в Приложения 2 приведены критические значения для нормального распределения.

Характеристическая функция нормального распределения определяется соотношением

$$\theta(\vartheta)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\exp\left[-\frac{(x-m)^2}{2\sigma^2}+\mathrm{i}\vartheta x\right]dx. \qquad (1.4.5)$$

Делая замену переменной

$$z=\frac{x-m}{\sigma\sqrt{2}}-\mathrm{i}\frac{\sigma\vartheta}{\sqrt{2}}$$

и учитывая табличный интеграл вида [21]

$$\int_{-\infty}^{\infty}\exp(-z^2)\,dz=\sqrt{\pi}, \qquad (1.4.6)$$

после несложных преобразований выражение для характеристической функции примет вид

$$\theta(\vartheta)=\exp(\mathrm{i}\vartheta m-\sigma^2\vartheta^2/2). \qquad (1.4.7)$$

Выясним смысл параметров m и σ, определяющих поведение нормального закона распределения. Путем дифференцирования характеристической функции получаем среднее значение

$$\mathrm{M}\{X\}=\frac{1}{\mathrm{i}}\left.\frac{d\theta(\vartheta)}{d\vartheta}\right|_{\vartheta=0}=\left.\frac{\mathrm{i}m-\sigma^2\vartheta}{\mathrm{i}}\exp\left(\mathrm{i}\vartheta m-\frac{\sigma^2\vartheta^2}{2}\right)\right|_{\vartheta=0}=m,$$

т.е. параметр m является средним значением гауссовской случайной величины. Аналогично вычисляется дисперсия. Действительно,

$$m_2=\frac{1}{\mathrm{i}^2}\left.\frac{d^2\theta(\vartheta)}{d\vartheta^2}\right|_{\vartheta=0}=\left.\frac{(\mathrm{i}m-\sigma^2\vartheta)^2-\sigma^2}{\mathrm{i}^2}\exp\left(\mathrm{i}m\vartheta-\frac{\sigma^2\vartheta^2}{2}\right)\right|_{\vartheta=0}=$$
$$=m^2+\sigma^2.$$

Отсюда получаем, что параметр σ^2 является дисперсией нормально распределенной случайной величины: $D=m_2-m^2=\sigma^2$. Так как для рассматриваемого распределения мода нормального распределения $\mathbf{P}\{X<m\}=\mathbf{P}\{X>m\}=1/2$, то среднее значение одновременно является и медианой распределения. В связи с этим для нормального распределения $m=m_{\mathrm{mod}}=m_{\mathrm{med}}$.

Для определения центрального момента k-го порядка в исходной формуле

$$\mu_k=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}(x-m)^k\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]dx$$

осуществим замену переменной $z=(x-m)/(\sigma\sqrt{2})$. Тогда

$$\mu_k=\frac{(\sigma\sqrt{2})^k}{\sqrt{\pi}}\int_{-\infty}^{\infty}z^k\exp(-z^2)\,dz. \qquad (1.4.8)$$

Отсюда видно, что нулевой центральный момент ($k=0$) равен единице ($\mu_0=1$), а все нечетные моменты равны нулю. Интегрируя выражение (1.4.8) по частям, имеем

$$\mu_k=\frac{(\sigma\sqrt{2})^k}{\sqrt{\pi}}\int_{-\infty}^{\infty}z^{k-1}z\mathrm{e}^{-z^2}\,dz=\frac{(k-1)(\sigma\sqrt{2})^k}{2\sqrt{\pi}}\int_{-\infty}^{\infty}z^{k-2}\mathrm{e}^{-z^2}\,dz.$$

Поскольку выражение (1.4.8) справедливо для произвольного k, в

том числе и для $(k-2)$, можно записать

$$\mu_{(k-2)} = \frac{(\sigma\sqrt{2})^{(k-2)}}{\sqrt{\pi}} \int_{-\infty}^{\infty} z^{(k-2)} \exp(-z^2)\, dz.$$

Сравнивая между собой два последних выражения, получаем рекуррентное соотношение

$$\mu_k = (k-1)\sigma^2 \mu_{(k-2)}, \tag{1.4.9}$$

позволяющее выразить моменты высших порядков через моменты низших порядков. Принимая во внимание, что второй центральный момент является дисперсией μ_2, общая формула для четного центрального момента k-го порядка гауссовской случайной величины может быть записана в виде

$$\mu_k = 1 \cdot 3 \cdot 5 \cdots (k-1)\sigma^k \quad \text{при} \quad k \quad \text{четном.} \tag{1.4.10}$$

В частности, для $k = 4$ четвертый центральный момент равен $\mu_4 = 3\sigma^4$. Следовательно, для нормального распределения коэффициенты асимметрии и эксцесса, вычисленные по (1.2.37) и (1.2.38), равны нулю.

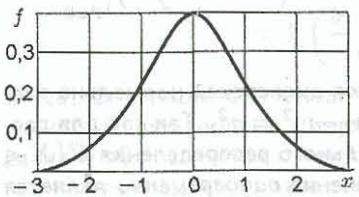


Рис. 1.4.1. График плотности вероятности гауссовской случайной величины

На рис. 1.4.1 приведен график плотности вероятности гауссовской случайной величины с нулевым средним значением $(m = 0)$ и единичной дисперсией $(\sigma^2 = 1)$.

Нормальное распределение воспроизводит себя при композиции. Иначе говоря, сумма n независимых гауссовских случайных величин

$$X = \sum_{k=1}^{n} X_k \tag{1.4.11}$$

со средними значениями m_k и дисперсиями σ_k^2 также будет распределена по нормальному закону со средним значением и дисперсией, определяемыми выражениями

$$m = \sum_{k=1}^{n} m_k; \quad \sigma^2 = \sum_{k=1}^{n} \sigma_k^2. \tag{1.4.12}$$

Действительно, характеристическая функция суммы независимых гауссовских случайных величин согласно (1.2.23) имеет вид

$$\theta(\vartheta) = \prod_{k=1}^{n} \exp\left(i\vartheta m_k - \frac{\sigma_k^2 \vartheta^2}{2}\right) = \exp\left(i\vartheta \sum_{k=1}^{n} m_k - \frac{\vartheta^2}{2} \sum_{k=1}^{n} \sigma_k^2\right).$$

Эта функция соответствует характеристической функции гауссовской случайной величины с параметрами (1.4.12).

1.4.2. Гамма-распределение

Непрерывная случайная величина X имеет *гамма-распределение* (Г-распределение) с параметрами α и β, если ее плотность вероятности задается формулой

$$f(x) = \frac{x^{\alpha}}{\beta^{\alpha+1}\Gamma(\alpha+1)} e^{-x/\beta}, \quad x \geqslant 0. \tag{1.4.13}$$

Здесь

$$\Gamma(\alpha+1) = \int_0^{\infty} t^{\alpha} e^{-t}\, dt, \quad \alpha+1 > 0, \tag{1.4.14}$$

— *гамма-функция*, а $\alpha > -1$ и $\beta > 0$ — параметры распределения, причем параметр α гамма-распределения характеризует форму распределения, а параметр β — масштаб распределения.

Функция распределения гамма-распределения определяется выражением

$$F(x) = \frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)} \int_0^{x} t^{\alpha} e^{-t/\beta}\, dt = \frac{\Gamma(\alpha+1, x/\beta)}{\Gamma(\alpha+1)}, \quad x \geqslant 0. \tag{1.4.15}$$

где

$$\Gamma(\alpha+1, x) = \int_0^{x} t^{\alpha} e^{-t}\, dt, \quad \alpha+1 > 0, \tag{1.4.16}$$

неполная гамма-функция. Для вычисления вероятностей гамма-распределения для ряда значений параметров α и λ в [8, 73 и др] приведены соответствующие таблицы.

Гамма-функция $\Gamma(z)$ и ее производные для всех действительных $\alpha > 0$ непрерывны; гамма-функция имеет единственный минимум на интервале $(0, \infty)$. Интегрированием (1.4.14) по частям можно показать, что справедливо рекуррентное соотношение

$$\Gamma(z+1) = z\Gamma(z), \quad z > 0. \tag{1.4.17}$$

Из (1.4.14) нетрудно вычислить, что $\Gamma(1) = 1$. Поэтому для целых чисел z справедливы соотношения

$$\Gamma(z+1) = 1 \cdot 2 \cdots z = z!; \quad \Gamma(1) = 0! = 1. \tag{1.4.18}$$

Гамма-функции $\Gamma(z)$ и $\Gamma(z, x)$ широко используются в теории вероятностей и математической статистике, они связаны с другими специальными функциями. Для дробных величин z таблицы значений гамма-функции приведены в справочниках по высшей математике; подобные таблицы помещены и в [8].

Если в формуле (1.4.13) положить $\beta = 1/\lambda$, а $\alpha + 1 = \nu$, где ν — целое положительное число, *получим плотность вероятности Эрланга*:

$$f(x) = \frac{\lambda^{\nu} x^{\nu-1}}{\nu!} \exp(-\lambda x), \quad x \geqslant 0. \tag{1.4.19}$$

При $\alpha = 0$ гамма-распределение переходит в *экспоненциальное распределение*, плотность вероятности которого с учетом принятого выше обозначения имеет вид

$$f(x) = (1/\lambda)\mathrm{e}^{-\lambda x}, \quad x \geqslant 0. \tag{1.4.20}$$

Нетрудно видеть, что экспоненциальное распределение является частным случаем более общего распределения Эрланга для $\nu = 1$. Экспоненциальное распределение и распределение Эрланга широко применяются в теории надежности и массового обслуживания.

Характеристическая функция гамма-распределения определяется выражением

$$\theta(\vartheta) = \frac{1}{\beta\Gamma(\alpha+1)}\int_0^\infty \left(\frac{x}{\beta}\right)^\alpha \exp\left[-\left(\frac{1}{\beta} - \mathrm{i}\vartheta\right)x\right] dx = (1-\mathrm{i}\beta\vartheta)^{-(\alpha+1)}. \tag{1.4.21}$$

При написании этой формулы учтено соотношение (1.4.14). Начальный момент k-го порядка гамма-распределения

$$m_k = \frac{1}{\beta\Gamma(\alpha+1)}\int_0^\infty x^k \left(\frac{x}{\beta}\right)^\alpha \mathrm{e}^{-x/\beta}\, dx = \frac{\Gamma(\alpha+k+1)}{\Gamma(\alpha+1)}\beta^k = \\ = \beta^k(\alpha+1)(\alpha+2)\cdots(\alpha+k). \tag{1.4.22}$$

Отсюда получаем формулы для среднего значения и дисперсии:

$$m = \beta(\alpha+1); \quad D = m_2 - m^2 = \beta^2(\alpha+1). \tag{1.4.23}$$

Аналогично вычисляются третий и четвертый центральные моменты, а также коэффициенты асимметрии и эксцесса

$$\mu_3 = 2\beta^3(\alpha+1); \quad \mu_4 = 3\beta^4(\alpha+1)(\alpha+3); \\ \gamma_1 = \mu_3/D^{3/2} = 2/\sqrt{\alpha+1}; \quad \gamma_2 = \mu_4/D^2 - 3 = 6/(\alpha+1). \tag{1.4.24}$$

Решая уравнения (1.4.23) относительно параметров α и β, имеем

$$\alpha = (m^2/D) - 1; \quad \beta = D/m. \tag{1.4.25}$$

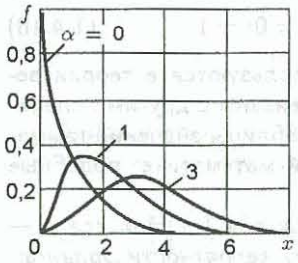


Рис. 1.4.2. График плотности вероятности гамма-распределения

На рис. 1.4.2 приведены графики плотности вероятности (1.4.13) для $\beta = 1$ и трех значений параметра $\alpha = 0;\ 1;\ 3$.

Если α — целое число, то гамма-распределение можно рассматривать как сумму $\nu = \alpha + 1$ независимых случайных величин X_i, каждая из которых имеет одинаковую экспоненциальную плотность вероятности (1.4.20). Действительно, согласно (1.2.23) характеристическая функция суммы ν независимых случайных слагаемых равна произведению характеристических функций слагаемых. Учитывая, что характеристическая функция экспоненциально распределенной величины

$$\theta(\vartheta) = (1 - \mathrm{i}\beta\vartheta)^{-1}, \tag{1.4.26}$$

их произведение совпадает с (1.4.21).

При больших значениях параметра α ($\alpha \gg 1$) коэффициенты асимметрии и эксцесса гамма-распределения стремятся к нулю. Следовательно, при больших значениях α гамма-распределение переходит в нормальное распределение с параметрами (1.4.23). При композиции g гамма-распределенных случайных величин с параметрами $\beta_j = \beta = \mathrm{const}$ и α_j, $j = 1, 2, \ldots, g$, можно утверждать, что результирующий закон будет гамма-распределением с параметрами β и $1 + \alpha = \sum_{j=1}^{g}(1+\alpha_j)$. В этом нетрудно убедиться, применяя формулу (1.2.23).

Пример 1.4.1. Время безотказной работы системы подчиняется экспоненциальному закону. Вычислить вероятность того, что система откажет (потребует ремонта) после двадцати часов работы, если известно, что в течение 1000 часов работы системы приходится ремонтировать ее в среднем десять раз.

Плотность вероятности времени t безотказной работы системы подчиняется экспоненциальному закону $f(t) = \lambda\exp(-\lambda t)$, где $1/\lambda = 100$ часов — среднее время безотказной работы системы. Поэтому вероятность того, что система откажет после двадцати часов работы

$$\mathbf{P}\{T < t = 20\} = \int_0^{t=20} \lambda \mathrm{e}^{-\lambda z} dz = 1 - \mathrm{e}^{-0{,}02} \approx 0{,}18.$$

1.4.3. Прямоугольное (равномерное) распределение

Плотность вероятности, описывающая непрерывные случайные величины с равновероятными исходами, называется *прямоугольной* или *равномерной плотностью вероятности* и имеет вид

$$f(x) = \begin{cases} 1/(b-a), & a \leqslant x \leqslant b; \\ 0, & x < a,\ x > b. \end{cases} \tag{1.4.27}$$

Соответственно функция распределения прямоугольно распределенной случайной величины определяется выражением

$$F(x) = \begin{cases} 0, & x < a; \\ (x-a)/(b-a), & a \leqslant x \leqslant b; \\ 1, & x > b. \end{cases} \tag{1.4.28}$$

На рис. 1.4.3 приведены графики равномерной плотности вероятности и соответствующей функции распределения.

Прямоугольное распределение встречается в задачах, связанных с округлением до целых чисел численных расчетов.

Начальный момент k-го порядка равномерно распределенной случайной величины определяется выражением

$$m_k = \frac{1}{b-a}\int_a^b x^k\, dx = \frac{b^{k+1} - a^{k+1}}{(b-a)(k+1)}. \tag{1.4.29}$$

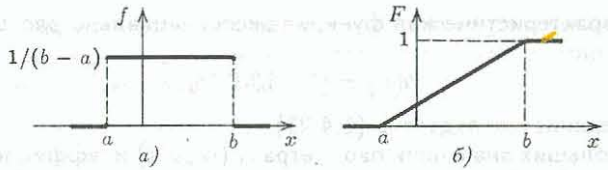


Рис. 1.4.3. Плотность вероятности (*а*) и функция распределения (*б*) равномерно распределенной случайной величины

При этом среднее значение

$$m = (a+b)/2. \tag{1.4.30}$$

Центральный момент k-го порядка находится из соотношения

$$\mu_k = \frac{1}{b-a}\int_a^b \left[x - \frac{a+b}{2}\right]^k dx. \tag{1.4.31}$$

Осуществляя замену переменной $y = x - (a+b)/2$, имеем

$$\mu_k = \frac{1}{b-a}\int_{-(b-a)/2}^{(b-a)/2} y^k\, dy = \frac{(b-a)^k[1-(-1)^{k+1}]}{2^{k+1}(k+1)}. \tag{1.4.32}$$

Из этой формулы видно, что если k — нечетное число, то, как и следовало ожидать, центральные моменты нечетного порядка равны нулю. Центральный момент четного порядка определяется выражением

$$\mu_k = \frac{(b-a)^k}{2^k(k+1)}, \quad k \text{ — четное число.} \tag{1.4.33}$$

Отсюда формулы для дисперсии и четвертого центрального момента принимают вид

$$D = (b-a)^2/12; \quad \mu_4 = (b-a)^4/80. \tag{1.4.34}$$

При этом коэффициенты асимметрии и эксцесса равны соответственно $\gamma_1 = 0$, $\gamma_2 = -1{,}2$.

Характеристическая функция прямоугольного распределения имеет вид

$$\theta(\vartheta) = \frac{1}{b-a}\frac{\mathrm{e}^{\mathrm{i}\vartheta b} - \mathrm{e}^{\mathrm{i}\vartheta a}}{\mathrm{i}\vartheta}. \tag{1.4.35}$$

С помощью линейного преобразования интервал, на котором случайная величина распределена равномерно, может быть переведен в любой другой заданный интервал. Так, например, с помощью преобразования переменных

$$y = \frac{x - (a+b)/2}{b-a}$$

случайная величина Y будет равномерно распределена по интервалу $(-1/2, 1/2)$. Поэтому в качестве нормированной («стандартной») равномерно распределенной случайной величины рассматривают случайную величину с плотностью вероятности вида

$$f(x) = \begin{cases} 1, & -1/2 \leqslant x \leqslant 1/2; \\ 0, & x > 1/2,\ x < -1/2. \end{cases} \tag{1.4.36}$$

При этом среднее значение равно нулю, дисперсия равна $D = 1/12$, а характеристическая функция определяется выражением

$$\theta(\vartheta) = \frac{1}{\mathrm{i}\vartheta}(\mathrm{e}^{\mathrm{i}\vartheta/2} - \mathrm{e}^{-\mathrm{i}\vartheta/2}) = \frac{2}{\vartheta}\sin(\vartheta/2). \tag{1.4.37}$$

Рассмотрим сумму n случайных величин $X = \sum X_k$ с одинаковыми плотностями вероятности (1.4.36) и, естественно, с одинаковыми характеристическими функциями (1.4.37). Согласно (1.2.24) характеристическая функция суммы

$$\theta_X(\vartheta) = \frac{2^n \sin^n(\vartheta/2)}{\vartheta^n}. \tag{1.4.38}$$

Отсюда видно, что композиция равномерно распределенных случайных величин не воспроизводит равномерное распределение. В то же время можно указать интервал $(-n/2, n/2)$, на котором плотность вероятности рассматриваемой суммы не равна нулю, так как в этот интервал попадает значение суммы $X = X_1 + X_2 + \ldots + X_n$. При этом среднее значение равно нулю, а дисперсия равна $D = n/12$.

Плотность вероятности для произвольного числа слагаемых определяется формулой [37 и др.]

$$f(x) = \frac{1}{(n-1)!}\left[\left(x + \frac{n}{2}\right)^{n-1} - C_n^1\left(x + \frac{n}{2} - 1\right)^{n-1} + C_n^2\left(x + \frac{n}{2} - 2\right)^{n-1} + \ldots\right], \tag{1.4.39}$$

где C_n^k — число сочетаний из n элементов по k элементов (1.2.32), причем в сумму (1.4.39) входят все слагаемые $(x+n/2)$, $(x+n/2-1)$,..., которые при данном x неотрицательны. Например, если $n = 2$, то

$$f(x) = \begin{cases} x+1, & -1 < x \leqslant 0; \\ 1-x, & 0 \leqslant x < 1. \end{cases} \tag{1.4.40}$$

График плотности вероятности двух слагаемых представляет собой равнобедренный треугольник с вершиной, равной единице в точке $x = 0$, а основанием является отрезок $(-1, +1)$. При $n = 3$ график плотности вероятности состоит из трех отрезков квадратичных парабол.

На рис. 1.4.4 приведены графики плотности вероятности для суммы двух и трех равномерно распределенных случайных величин. Там же показана для сравнения плотность вероятности одного слагаемого и нормальная плотность вероятности (*а*) с дисперсией $D = 3/12$, соот-

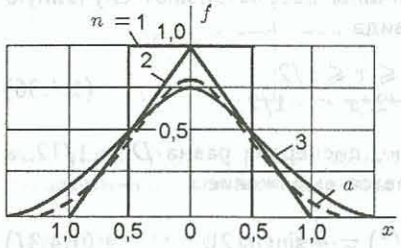


Рис. 1.4.4. Графики прямоугольного и «родственных» ему распределений

ветствующей дисперсии суммы трех случайных величин. Уже при трех слагаемых результирующих график функции распределения очень похож на кривую нормальной плотности вероятности. При возрастании числа слагаемых n плотность вероятности суммы независимых одинаково равномерно распределенных случайных величин быстро приближается к нормальной плотности вероятности.

Применительно к нормированным слагаемым с плотностью вероятности (1.4.36) при $n \gg 1$ такое распределение будет иметь нулевое среднее значение и дисперсию $D = n/12$. Практически подобное приближение для прикладных задач справедливо уже при $n \geqslant 4$.

Прямоугольное распределение часто используется при анализе ошибок от округления числа по интервалу $(-1/2, +1/2)$ в единицах последнего знака, а задача построения композиции равномерно распределенных случайных величин возникает в метрологии при суммировании элементарных систематических погрешностей, распределения которых обычно аппроксимируют прямоугольным распределением.

1.4.4. Логарифмически нормальное (логнормальное) распределение

Непрерывная положительная случайная величина X имеет *логарифмически нормальное* (*логнормальное*) распределение с параметрами a и σ^2, если ее плотность вероятности описывается выражением

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{(\ln x - a)^2}{2\sigma^2}\right], \quad x > 0. \tag{1.4.41}$$

Начальные моменты $m_k = \mathrm{M}\{X^k\}$ этой случайной величины связаны с параметрами a и σ^2 соотношением [2, 15, 36, 42 и др].

$$m_k = \exp\left(\frac{1}{2}k^2\sigma^2 + ka\right). \tag{1.4.42}$$

Отсюда нетрудно получить среднее значение m_X и ее дисперсию σ_X^2, а также коэффициенты асимметрии γ_1 и эксцесса γ_2:

$$m_X = \exp\left(\frac{1}{2}\sigma^2 + a\right); \tag{1.4.43}$$

$$\sigma_X^2 = \left[\exp(\sigma^2) - 1\right]\exp(\sigma^2 + 2a^2); \tag{1.4.44}$$

$$\gamma_1 = \sqrt{\exp(\sigma^2) - 1}\left[\exp(\sigma^2) + 2\right]; \tag{1.4.45}$$

$$\gamma_2 = \left[\exp(\sigma^2) - 1\right]\left[\exp(3\sigma^2) + 3\exp(2\sigma^2) + 6\exp(\sigma^2) + 6\right]. \tag{1.4.46}$$

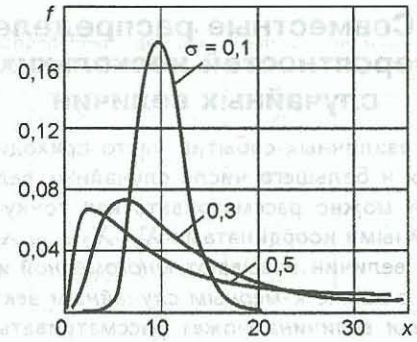


Рис. 1.4.5. Плотности вероятностей логарифмически нормального распределения

При этом мода и медиана логнормального распределения равны

$$x_{\mathrm{mod}} = \exp(-\sigma^2 + a); \quad x_{\mathrm{med}} = \mathrm{e}^a. \tag{1.4.47}$$

Логарифмически нормальное распределение случайной величины X появляется в результате суммирования независимых положительных случайных величин y_i, причем прирост суммарного эффекта случайных величин X, вызванный случайной величиной y_{i+1}, пропорционален произведению y_{i+1} и x_i: $x_{i-1} - x_i = y_{i+1}x_i$. Поэтому будем иметь

$$y_1 + y_2 + \cdots + y_n = \sum_{i=0}^{n-1} \frac{x_{i+1} - x_i}{x_i}.$$

Полагая прирост суммарного эффекта достаточно малым, сумму в правой части последнего выражения приближенно можно заменить интегралом:

$$y_1 + y_2 + \cdots + y_n = \int_{x_0}^{x} \frac{dt}{t} = \ln x - \ln x_0.$$

Здесь $x = x_n$ обозначает окончательный эффект суммирования, а x_0 — начальное значение. Как отмечалось выше и более строго показано в п. 2.4, при достаточно большом числе n независимых слагаемых y_i сумма в предыдущем выражении будет распределена нормально и, следовательно, случайная величина $\ln x$ также будет распределена нормально со средним значением $\ln x_0$.

На рис. 1.4.5 приведены графики плотности вероятностей логарифмически нормального распределения при $a = 1$ и $\sigma = 0{,}1$; 0,3; 0,5.

Логарифмически нормальное распределение достаточно удовлетворительно описывает многие физические и социально-экономические модели [2] (например, долговечность изделия, работающего в режиме износа и старения; размеры космических образований; увеличение заработной оплаты работника; приращение богатства при низких значениях дохода и др.).

1.5. Совместные распределения вероятностей нескольких случайных величин

При изучении различных событий часто приходится иметь дело с совокупностью двух и большего числа случайных величин. Систему k случайных величин можно рассматривать как точку в k-мерном пространстве со случайными координатами $X_1, X_2, \ldots, X_k$. Поэтому такую систему случайных величин называют *многомерной* или *k-мерной случайной величиной*, а также *k-мерным случайным вектором*. При $k = 2$ *двумерная случайная величина* может рассматриваться как случайная точка на плоскости, а при $k = 3$ — как случайная точка в пространстве. Подобная трактовка совокупности двух или трех случайных величин дает возможность пользоваться наглядными геометрическими представлениями. Системы случайных величин аналогично одномерным случайным величинам могут быть дискретными, непрерывными или смешанными в зависимости от типа случайных величин, образующих эту систему.

Полной вероятностной характеристикой многомерной случайной величины является *многомерный закон распределения*, устанавливающий связь между областями возможных значений многомерной случайной величины и вероятностями ее появления в этих областях. В связи с определенной сложностью описания многомерных случайных величин произвольного порядка, ниже более подробно рассматриваются двумерные случайные величины.

1.5.1. Двумерные случайные величины

Закон распределения двух случайных величин X и Y (случайного двумерного вектора с координатами X и Y) может быть задан в различных формах. Если эти случайные величины являются дискретными, то закон *распределения двумерной случайной величины* (X, Y), как правило, задается в табличной форме (табл. 1.5.1). В таблице обозначено

$$p_{ij} = \mathbf{P}\{X = x_i, Y = y_j\} = p_2(x_i, y_j); \tag{1.5.1}$$

$$\mathbf{P}\{X = x_i\} = p(x_i) = \sum_{j=1}^{q} p_{ij}; \quad \mathbf{P}\{Y = y_j\} = p(y_j) = \sum_{i=1}^{n} p_{ij};$$

$$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, q. \tag{1.5.2}$$

При этом, естественно,

$$\sum_{i=1}^{n} \sum_{j=1}^{q} p_{ij} = 1.$$

Если X и Y независимы, то $p_{ij} = p_i p_j$.

Универсальной характеристикой многомерных случайных величин, в том числе и двумерных, является функция распределения. Применительно к двумерной случайной величине *двумерной* (*совместной*) *функцией распределения* $F_{XY} \equiv F_2(x, y)$ называется вероятность совместного выполнения неравенств $X < x$ и $Y < y$:

$$F_2(x, y) = \mathbf{P}\{X < x, Y < y\}. \tag{1.5.3}$$

В геометрической интерпретации (рис. 1.5.1) функцию распределения $F_2(x, y)$ можно рассматривать как вероятность попадания случайной точки внутрь бесконечного левого нижнего квадранта с координатами вершины (x, y).

Если двумерные случайные величины X и Y непрерывны, то их распределение характеризуют *двумерной* (*совместной*) *плотностью вероятности*, которая определяется выражением

$$f_{XY}(x, y) \equiv f_2(x, y) = \frac{\partial^2 F_2(x, y)}{\partial x \partial y}. \tag{1.5.4}$$

При этом предполагается, что двумерная функция распределения непрерывна и имеет непрерывную смешанную производную второго порядка.

Функция распределения $F_2(x, y)$ и плотность вероятности $f_2(x, y)$ обладают следующими основными свойствами:

1) $F_2(x, y)$ — неубывающая функция своих аргументов, т.е.
$F_2(x_2, y) \geqslant F_2(x_1, y)$ при $x_2 > x_1$,
$F_2(x, y_2) \geqslant F_2(x, y_1)$ при $y_2 > y_1$;

2) $F_2(x, -\infty) = F_2(-\infty, y) = F_2(-\infty, -\infty) = 0$;

3) $F_2(\infty, \infty) = \int_{-\infty}^{\infty} \int_{\infty}^{\infty} f_2(x, y)\, dxdy = 1$ — условие нормировки;

4) $F_2(x, \infty) = F_X(x)$, $F_2(\infty, y) = F_Y(y)$, где, как и выше, $F_X(x)$ и $F_Y(y)$ — одномерные функции распределения случайных величин X и Y;

5) вероятность попадания случайной точки (X, Y) в произвольную область Q определяется выражением

$$p(Q) = \iint_Q f_2(x, y)\, dxdy. \tag{1.5.5}$$

Таблица 1.5.1

Распределение двумерной случайной величины

x_i	y_j						$\sum_{j=1}^{q} p_{ij}$
	y_1	y_2	$\ldots$	y_r	$\ldots$	y_q	
x_1	p_{11}	p_{12}	$\ldots$	p_{1r}	$\ldots$	p_{1q}	$p(x_1)$
x_2	p_{21}	p_{22}	$\ldots$	p_{2r}	$\ldots$	p_{2q}	$p(x_2)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_n	p_{n1}	p_{n2}	$\ldots$	p_{nr}	$\ldots$	p_{nq}	$p(x_n)$
$\sum_{i=1}^{n} p_{ij}$	$p(y_1)$	$p(y_2)$	$\ldots$	$p(y_r)$	$\ldots$	$p(y_q)$	

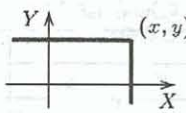


Рис. 1.5.1. К определению двумерной функции распределения

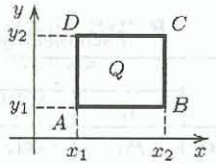


Рис. 1.5.2. Прямоугольная область

Пусть областью попадания случайной точки является прямоугольник, показанный на рис. 1.5.2, со сторонами, параллельными осям абсцисс и ординат, и координатами $A(x_1, y_1)$, $B(x_2, y_1)$, $C(x_2, y_2)$ и $D(x_1, y_2)$. Тогда вероятность попадания случайной точки в этот прямоугольник

$$p(Q) = \mathbf{P}\{x_1 \leqslant X < x_2, y_1 \leqslant Y < y_2\} = \\ = F_2(x_2, y_2) - F_2(x_1, y_2) - F_2(x_2, y_1) + F_2(x_1, y_1). \quad (1.5.6)$$

Одномерные функции распределения и плотности вероятности выражаются через двумерные функции распределения и плотности вероятности с помощью следующих соотношений:

$$F_X(x) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_2(x, y)\,dxdy; \quad (1.5.7)$$

$$f_X(x) = \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_2(x, y)\,dy; \quad (1.5.8)$$

$$F_Y(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f_2(x, y)\,dxdy; \quad (1.5.9)$$

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f_2(x, y)\,dx. \quad (1.5.10)$$

Закон распределения двух случайных величин X и Y определяется распределением каждой из них и зависимостью между ними. Степень зависимости случайных величин X и Y характеризуется условным законом распределения, под которым понимается закон распределения одной из случайных величин, найденный при условии, что другая случайная величина приняла определенное значение (значения).

Рассмотрим непрерывные случайные величины X и Y. Запишем условную вероятность $X < x$ при условии, что вторая случайная величина заключена в интервале $y \leqslant Y < y + \Delta y$,

$$\mathbf{P}\{X < x \mid y \leqslant Y < y + \Delta y\} = \frac{\mathbf{P}\{X < x; y \leqslant Y < y + \Delta y\}}{\mathbf{P}\{y \leqslant Y < y + \Delta y\}}. \quad (1.5.11)$$

Пусть $\Delta y \to 0$, тогда из последнего выражения следует

$$F_X(x \mid y) = \lim_{\Delta y \to 0} \frac{\mathbf{P}\{X < x; y \leqslant Y < y + \Delta y\}}{\mathbf{P}\{y \leqslant Y < y + \Delta y\}} =$$

$$= \lim_{\Delta y \to 0} \frac{\int_{-\infty}^{x}\int_{y}^{y+\Delta y} f_2(x, y)\,dxdy}{\int_{-\infty}^{\infty}\int_{y}^{y+\Delta y} f_2(x, y)\,dxdy} = \lim_{\Delta y \to 0} \frac{\Delta y \int_{-\infty}^{x} f_2(x, y)\,dx}{\Delta y \int_{-\infty}^{\infty} f_2(x, y)\,dx} = \\ = \frac{\int_{-\infty}^{x} f_2(x, y)\,dx}{\int_{-\infty}^{\infty} f_2(x, y)\,dx} = \frac{\int_{-\infty}^{x} f_2(x, y)\,dx}{f_Y(y)}. \quad (1.5.12)$$

При фиксированном y этот предел как функция от x представляет собой *условную функцию распределения* величины X при условии $Y = y$.

Дифференцируя последнее соотношение по переменной x, получаем выражение для *условной плотности вероятности* случайной величины X:

$$f_X(x \mid y) = \frac{dF_X(x \mid y)}{dx} = \frac{f_2(x, y)}{f_Y(y)}. \quad (1.5.13)$$

Аналогично из условной функции распределения $F_Y(y \mid X = x)$ имеем

$$f_Y(y \mid x) = \frac{dF_Y(y \mid x)}{dy} = \frac{f_2(x, y)}{f_X(x)}. \quad (1.5.14)$$

Поэтому для непрерывных случайных величин справедлива формула

$$f_2(x, y) = f_X(x) f_Y(y \mid x) = f_Y(y) f_X(x \mid y). \quad (1.5.15)$$

Применительно к дискретным случайным величинам в соответствии с формулой умножения вероятностей можно записать

$$p_2(x_i, y_j) = p(x_i)p(y_j \mid x_i) = p(y_j)p(x_i \mid y_j), \\ i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, q. \quad (1.5.16)$$

где $p(y_j \mid x_i)$, $p(x_i \mid y_j)$ — условные вероятности соответствующих дискретных случайных величин. Если дискретные случайные величины независимы, то

$$p(x_i, y_j) = p(x_i)p(y_j). \quad (1.5.17)$$

Для независимых непрерывных случайных величин аналогично имеем

$$f_2(x, y) = f_X(x) f_Y(y). \quad (1.5.18)$$

Условия (1.5.17) и (1.5.18) являются *необходимыми и достаточными условиями независимости случайных величин*. Согласно формуле Байеса (1.1.19) условную плотность вероятности можно трактовать как апостериорную плотность вероятности одной случайной величины, если вторая случайная величина приняла определенное конкретное значение.

Естественно, аналогичная трактовка справедлива и для условных вероятностей дискретных случайных величин.

Хотя законы распределения системы случайных величин являются исчерпывающими вероятностными характеристиками, однако их определение как теоретически, так и экспериментально, встречает большие

трудности. Поэтому при исследовании систем случайных величин широкое применение нашли их числовые характеристики, которые в определенной степени могут дать представление о характере закона распределения. В основу получения таких числовых характеристик положено понятие многомерных моментов (начальных и центральных), обобщающих понятия одномерных моментов (моментов одномерных случайных величин). Применительно к двумерной случайной величине рассматриваются следующие моменты.

1. *Начальный момент* m_{k_1,k_2} порядка $k_1 + k_2$

$$m_{k_1,k_2} = \mathrm{M}\{X^{k_1}Y^{k_2}\} = \sum_{i=1}^{n}\sum_{j=1}^{q} x_i^{k_1} y_j^{k_2} p_{ij},$$
$$i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, q, \tag{1.5.19}$$

для дискретных случайных величин;

$$m_{k_1,k_2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^{k_1} y^{k_2} f_2(x,y)\,dxdy \tag{1.5.20}$$

для непрерывных случайных величин.

2. *Центральный момент* μ_{k_1,k_2} порядка $k_1 + k_2$

$$\mu_{k_1,k_2} = \sum_{i=1}^{n}\sum_{j=1}^{q}(x_i - m_X)^{k_1}(y_j - m_Y)^{k_2} p_{ij} \tag{1.5.21}$$

для дискретных случайных величин;

$$\mu_{k_1,k_2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_X)^{k_1}(y - m_Y)^{k_2} f_2(x,y)\,dxdy \tag{1.5.22}$$

для непрерывных случайных величин.

Здесь m_X и m_Y — средние значения случайных величин X и Y, определяемые формулами

$$m_X = m_{10} = \mathrm{M}\{X^1Y^0\} = \mathrm{M}\{X\}; \quad m_Y = m_{01} = \mathrm{M}\{X^0Y^1\} = \mathrm{M}\{Y\}. \tag{1.5.23}$$

Из центральных моментов наибольшее применение находят моменты второго порядка. Два из них представляют собой дисперсии величин X и Y:

$$D_X = \mu_{20} = \mathrm{M}\{(X - m_X)^2\}; \quad D_Y = \mu_{02} = \mathrm{M}\{(Y - m_Y)^2\}. \tag{1.5.24}$$

Дисперсии D_X и D_Y характеризуют степень разброса возможных значений координат случайной точки в направлении осей абсцисс и ординат относительно средних значений m_X и m_Y соответственно. Среди смешанных моментов особую роль играет центральный смешанный момент второго порядка $\mu_{11} \equiv K_{XY}$, называемый *корреляционным моментом* или *ковариацией* случайных величин X и Y,

$$\mu_{11} = K_{XY} = K_{YX} = \mathrm{M}\{(X - m_X)(Y - m_Y)\}. \tag{1.5.25}$$

Корреляционный момент μ_{11} связан со смешанным вторым начальным моментом m_{11} очевидным соотношением

$$m_{11} = \mathrm{M}\{XY\} = \mu_{11} + m_X m_Y. \tag{1.5.26}$$

Две случайные величины X и Y называются *некоррелированными*, если для них $\mu_{11} = 0$, т.е. справедлива запись

$$m_{11} = \mathrm{M}\{XY\} = \mathrm{M}\{X\}\mathrm{M}\{Y\} = m_X m_Y. \tag{1.5.27}$$

Случайные величины X и Y называются *ортогональными*, если второй смешанный начальный момент равен нулю:

$$m_{11} = \mathrm{M}\{XY\} = 0. \tag{1.5.28}$$

Конкретизируем выражения для корреляционного момента применительно к дискретным и непрерывным случайным величинам. Для дискретных случайных величин корреляционный момент

$$K_{XY} = \sum_{i=1}^{n}\sum_{j=1}^{q}(x_i - m_X)(y_j - m_Y)p_{ij}. \tag{1.5.29}$$

Соответственно для непрерывных случайных величин корреляционный момент определяется из выражения

$$K_{XY} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - m_X)(y - m_Y)f_2(x,y)\,dxdy. \tag{1.5.30}$$

Корреляционный момент K_{XY} описывает как степень разброса возможных значений случайных величин X и Y относительно своих средних значений, так и статистическую связь между ними. Для описания лишь степени статистической связи между случайными величинами применяют безразмерную величину

$$\rho_{XY} = K_{XY}/\sqrt{D_X D_Y} = K_{XY}/(\sigma_X \sigma_Y), \tag{1.5.31}$$

называемую *коэффициентом корреляции* или *парным коэффициентом корреляции*. Коэффициент корреляции ρ_{XY} удовлетворяет условию

$$-1 \leqslant \rho_{XY} \leqslant 1. \tag{1.5.32}$$

Это условие вытекает из *неравенства Буняковского–Шварца*, которое утверждает: если две случайные величины Z и V имеют конечные значения первых двух моментов, то для квадрата математического ожидания их произведения справедливо соотношение [8, 15 и др.]

$$[\mathrm{M}\{ZV\}]^2 \leqslant \mathrm{M}\{Z^2\}\mathrm{M}\{V^2\}. \tag{1.5.33}$$

При этом неравенство переходит в равенство тогда и только тогда, когда

случайные величины связаны между собой детерминированной линейной зависимостью $Z = \text{const}\, V$.

Коэффициент корреляции определяет линейную вероятностную зависимость между случайными величинами X и Y. Пусть случайная величина Y линейно зависит от случайной величины X с параметрами $m_X = 0$ и D_X:

$$Y = a \pm bX,$$

где a и b — некоторые постоянные. При этом среднее значение и дисперсия Y соответственно равны $\mathrm{M}\{Y\} = a$, $D_Y = \mathrm{M}\{(Y-a)^2\} = b^2 D_X$. Поэтому

$$\rho_{XY} = \frac{\mathrm{M}\{X(Y-a)\}}{\sqrt{D_X D_Y}} = \frac{\pm b\mathrm{M}\{X^2\}}{bD_X} = \pm 1,$$

т.е. если случайные величины связаны между собой линейной зависимостью, то коэффициент корреляции равен ± 1. В общем случае произвольной статистической зависимости (не детерминированной) коэффициент корреляции принимает значения в пределах неравенства (1.5.32), причем знак плюс соответствует тому, что с увеличением одной случайной величины в среднем увеличивается и вторая случайная величина; знак минус характеризует противоположную ситуацию. Если случайные величины X и Y независимы [$f_2(x,y) = f_X(x) f_Y(y)$], то

$$K_{XY} = \int_{-\infty}^{\infty} (x - m_X) f_X(x)\, dx \int_{-\infty}^{\infty} (y - m_Y) f_Y(y)\, dy = 0$$

и, следовательно, для независимых случайных величин $\rho_{XY} = 0$. Две случайные величины, для которых коэффициент корреляции равен нулю, называются *некоррелированными*. Независимые случайные величины всегда некоррелированы. Зависимые же случайные величины могут быть как коррелированными, так и некоррелированными. В частности, как будет показано ниже, для гауссовских случайных величин некоррелированность означает также и их независимость.

Пример 1.5.1. Пусть двумерная плотность вероятности случайных величин X и Y равномерна внутри круга радиуса a:

$$f_2(x,y) = \begin{cases} 1/\pi a^2, & x^2 + y^2 \leqslant a^2, \\ 0, & x^2 + y^2 > a^2. \end{cases}$$

Найдем корреляционный момент согласно (1.5.30), учитывая, что $m_X = m_Y = 0$. Так как рассматриваемая двумерная плотность вероятности отлична от нуля только внутри круга радиуса a, для вычисления интеграла (1.5.30) целесообразно осуществить замену переменных $x = r\cos\varphi$, $y = r\sin\varphi$. В результате получим

$$K_{XY} = \frac{1}{\pi a^2} \int_0^{2\pi} \int_0^a r^3 \cos\varphi \sin\varphi\, dr\, d\varphi = 0.$$

Таким образом, рассматриваемые случайные величины X и Y являются некоррелированными, хотя и зависимы, так как координаты любой случайной точки внутри круга радиуса a между собой связаны уравнением окружности $x^2 + y^2 = r^2$.

Пример 1.5.2. Вычислим корреляционный момент случайных величин X и Y, связанных между собой детерминированной зависимостью $Y = X^2$, полагая при этом, что $m_X = 0$, а $f_X(x) = f_X(-x)$. В силу четности плотности вероятности случайной величины X все ее нечетные центральные моменты равны нулю. Поэтому

$$K_{XY} = \mathrm{M}\{(Y - m_Y)X\} = \mu_3 - m_Y m_X = 0.$$

Таким образом, рассматриваемые случайные величины X и Y являются некоррелированными, хотя и связаны между собой детерминированной («строгой») зависимостью.

1.5.2. Условные моменты. Регрессия первого рода

Для совокупности двух случайных величин часто используются понятия *условных моментов* одной случайной величины при фиксированном значении другой случайной величины. Например, *условное среднее значение* и *условная дисперсия* случайной величины Y относительно конкретного значения $X = x$ определяются формулами

$$m_Y(x) = m_{Y|x} = \int_{-\infty}^{\infty} y f_Y(y \mid x)\, dy; \qquad (1.5.34)$$

$$D_Y(x) = D_{Y|x} = \int_{-\infty}^{\infty} [y - m_Y(x)]^2 f_Y(y \mid x)\, dy. \qquad (1.5.35)$$

Аналогичным образом определяются и другие условные моменты Y относительно $X = x$ и условные моменты X относительно $Y = y$. Так как условная плотность вероятности случайной величины Y (1.5.14) зависит от конкретного значения $X = x$, последнюю можно рассматривать как независимую переменную. Поэтому любая числовая характеристика условной плотности вероятности $f_Y(y \mid x)$ также будет функцией от x. Наиболее часто в качестве условной характеристики берут условное среднее значение, которое используют в качестве ее прогнозируемого значения $y^* = m_Y(x)$. Эту условную характеристику называют *регрессией первого рода*. Для определения регрессий первого рода принципиально необходимо знать по крайней мере совместную двумерную плотность вероятности случайных величин X и Y. Однако за редким исключением, в основном связанным с гауссовскими случайными величинами и их функциональными преобразованиями, описание двумерных плотностей вероятности в удобном виде для дальнейших вычислений регрессий первого рода не представляется возможным. Задача еще более усложняется, когда необходимо получить статистическую зависимость случайной величины Y от нескольких переменных — случайных величин $X_1, X_2, \ldots, X_p$. В связи с этим изучаются *регрессии второго рода*.

При втором подходе (регрессия второго рода) класс возможных функций статистической взаимосвязи между случайными величинами X и Y ограничивают в суженном классе функций с точностью до некоторых неизвестных параметров (коэффициентов). Эти коэффициенты и их статистические характеристики определяют на основе исходной априорной информации о моментах анализируемых случайных величин. Если

априорная информация о моментах отсутствует, то оценки этих моментов получают на основе проведения соответствующих опытов. В этом случае регрессию называют *выборочной регрессией*.

Изучим свойства регрессии первого рода (условного среднего значения) применительно к случайной величине Y; она определяется выражением (1.5.34) и является регрессией Y на X. Если случайные величины X и Y независимы, то $f_Y(y \mid x) \equiv f_Y(y)$ и из (1.5.34) получаем $m_Y(x) \equiv m_Y$, т.е. в качестве прогнозируемого значения случайной величины Y следует взять ее априорное среднее значение $y^* = m_Y$. В этом случае условная дисперсия будет равна безусловной (априорной) дисперсии $D_{Y|x} \equiv D_Y$.

Покажем, что функция $m_Y(x)$ дает наилучшее описание зависимости Y от x в смысле минимума среднего квадрата отклонения величины Y от $m_Y(x)$. Для этого найдем такую функцию $h(x)$, чтобы обеспечивался минимум выражения

$$\mathrm{M}\{[Y-h(x)]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[y-h(x)]^2 f_2(x,y)\,dxdy = \\ = \int_{-\infty}^{\infty} f_X(x)\left[\int_{-\infty}^{\infty}[y-h(x)]^2 f_Y(y \mid x)\,dy\right]dx = \min.$$

Это выражение принимает минимальное значение, если будет минимальным значение внутреннего интеграла по y при любом x. Перепишем внутренний интеграл в виде

$$\int_{-\infty}^{\infty}\{[y-m_Y(x)]-[h(x)-m_Y(x)]\}^2 f_Y(y \mid x)\,dy = \\ = \int_{-\infty}^{\infty}[y-m_Y(x)]^2 f_Y(y \mid x)\,dy - \\ - 2[h(x)-m_Y(x)]\int_{-\infty}^{\infty}[y-m_Y(x)]f_Y(y \mid x)\,dy + \\ + [h(x)-m_Y(x)]^2\int_{-\infty}^{\infty} f_Y(y \mid x)\,dy = D_{Y|x} + [h(x)-m_Y(x)]^2.$$

Отсюда видим, что выражение будет принимать минимальное значение, равное условной дисперсии $D_{Y|x}$, если $h(x) = m_Y(x)$. Иначе говоря, среди всех возможных функций $h(x)$ минимум среднего квадрата отклонения прогнозируемого значения y^* относительно произвольной функции $h(x)$ достигается тогда, когда эта функция совпадает с условным средним значением: $h(x) = m_Y(x)$.

Рассмотрим некоторые статистические характеристики рассматриваемой регрессии $y^* = m_Y(x)$. Математическое ожидание (среднее значение) условного среднего значения m_Y

$$\mathrm{M}\{y^*\} = \mathrm{M}\{m_Y(x)\} = \int_{-\infty}^{\infty} m_Y(x) f_X(x)\,dx =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_Y(y \mid x) f_X(x)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_2(x,y)\,dxdy = \\ = \int_{-\infty}^{\infty} y f_Y(y)\,dy = m_Y, \qquad (1.5.36)$$

т.е. прогноз значения случайной величины Y по условному среднему значению $m_Y(x)$ в среднем совпадает с безусловной величиной m_Y. Вычислим корреляционные моменты между случайными величинами $Y - m_Y(x)$ и $m_Y(x) - m_Y$, с одной стороны, и случайными величинами $Y - m_Y$ и $m_Y(x) - m_Y$ — с другой:

$$\mathrm{M}\{[Y-m_Y(x)][m_Y(x)-m_Y]\} = \\ = \mathrm{M}\{Y m_Y(x)\} - m_Y\mathrm{M}\{Y\} - \mathrm{M}\{m_Y^2(x)\} + m_Y^2 = \\ = \mathrm{M}\{Y m_Y(x)\} - \mathrm{M}\{m_Y^2(x)\}.$$

Учитывая соотношение

$$\mathrm{M}\{Y m_Y(x)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y m_Y(x) f_2(x,y)\,dxdy = \\ = \int_{-\infty}^{\infty} m_Y(x) f_X(x)\left[\int_{-\infty}^{\infty} y f_Y(y \mid x)\,dy\right]dx = \int_{-\infty}^{\infty} m_Y^2(x) f_X(x)\,dx = \\ = \mathrm{M}\{m_Y^2(x)\},$$

из предыдущего выражения получаем

$$\mathrm{M}\{[Y-m_Y(x)][m_Y(x)-m_Y]\} = 0, \qquad (1.5.37)$$

т.е. рассматриваемые случайные величины $Y - m_Y(x)$ и $m_Y(x) - m_Y$ некоррелированы. В противоположность этому корреляционный момент между случайными величинами $Y - m_Y$ и $m_Y(x) - m_Y$ не равен нулю. Действительно,

$$K(Y, m_Y(x)) = \mathrm{M}\{(Y-m_Y)[m_Y(x)-m_Y]\} = \\ = \mathrm{M}\{Y m_Y(x)\} - m_Y\mathrm{M}\{Y\} - m_Y\mathrm{M}\{m_Y(x)\} + m_Y^2 = \\ = \mathrm{M}\{[m_Y(x)-m_Y]^2\} = D_m,$$

где

$$D_m = \mathrm{M}\{[m_Y(x)-m_Y]^2\} = \int_{-\infty}^{\infty}[m_Y(x)-m_Y]^2 f_X(x)\,dx \qquad (1.5.38)$$

— средний квадрат (дисперсия) отклонения условного среднего $m_Y(x)$ относительно безусловного среднего m_Y. При этом коэффициент корреляции между случайной величиной Y и ее прогнозируемым значением $m_Y(x)$

$$\rho(Y, m_Y(x)) = \frac{K(Y, m_Y(x))}{\sqrt{D_Y D_m}} = \sqrt{\frac{D_m}{D_Y}}. \qquad (1.5.39)$$

Аналогично вычисляются смешанные моменты между случайными величинами $Y - m_Y(x)$ и X и случайными величинами $Y - m_Y$ и X:

$$\mathrm{M}\{[Y - m_Y(x)]X\} =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yx f_2(x,y)\,dxdy - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_Y(y \mid x) f_X(x)\,dxdy = 0;$$

$$\mathrm{M}\{(Y - m_Y)X\} = \mathrm{M}\{(Y - m_Y)[(X - m_X) + m_X]\} = K_{XY}.$$

Выразим безусловную дисперсию $D_Y = \mathrm{M}\{(Y - m_Y)^2\}$ случайной величины Y через среднее значение условной дисперсии $\bar{D}_{Y|x}$:

$$\bar{D}_{Y|x} = \mathrm{M}\{[Y - m_Y(x)]^2\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - m_Y(x)]^2 f_2(x,y)\,dxdy, \tag{1.5.40}$$

и дисперсию D_m, определяемую (1.5.38). В результате получим

$$D_Y = \mathrm{M}\{(Y - m_Y)^2\} = \mathrm{M}\{([Y - m_Y(x)] + [m_Y(x) - m_Y])^2\} =$$
$$= \mathrm{M}\{[Y - m_Y(x)]^2\} + \mathrm{M}\{[m_Y(x) - m_Y]^2\} = \bar{D}_{Y|x} + D_m. \tag{1.5.41}$$

Отношение среднего значения условной дисперсии $\bar{D}_{Y|x}$ прогнозируемого значения случайной величины к ее безусловной (априорной) дисперсии D_Y

$$\zeta = \frac{\bar{D}_{Y|x}}{D_Y} = 1 - \frac{D_m}{D_Y} \tag{1.5.42}$$

может служить мерой точности прогнозирования $y^* = m_Y(x)$.

Величину, численно равную квадрату коэффициента корреляции (1.5.39),

$$\eta^2_{Y|x} = \frac{D_m}{D_Y} = 1 - \frac{\bar{D}_{Y|x}}{D_Y} = \frac{1}{D_Y}\int_{-\infty}^{\infty} [m_Y(x) - m_Y]^2 f_X(x)\,dx \tag{1.5.43}$$

называют *корреляционным отношением* случайной величины Y по случайной величине X. Если случайные величины X и Y независимы, то $m_Y(x) = m_Y$ при любом $X = x$ и, следовательно, $\eta^2_{Y|x} = 0$.

В случае линейной зависимости Y от X для выполнения условия $m_Y(x) = m_Y$ достаточно, чтобы случайные величины X и Y были некоррелированы. При этом регрессия Y на X представляет собой прямую, параллельную оси абсцисс. Ниже при изучении линейной регрессии второго рода это утверждение вытекает из (1.5.51). Вместе с этим необходимо подчеркнуть, что равенство нулю корреляционного отношения и коэффициента корреляции не означает независимости случайных величин. Это обусловлено тем, что при $\rho_{XY} = 0$ для нелинейной зависимости Y от X справедливо неравенство $m_Y(x) \neq m_Y$. Последнее наглядно иллюстрируется примерами 1.5.1 и 1.5.2.

Действительно, применительно к данным примера 1.5.2 плотность вероятности случайной величины Y можно записать, используя δ-функцию (1.2.11): $f_Y(y) = \delta(y - x^2)$. В результате получим

$$m_Y(x) = \int_{-\infty}^{\infty} y\delta(y - x^2)\,dy = x^2; \quad m_Y = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = D_X.$$

т.е. $m_Y(x) \neq m_Y$.

При независимости случайной величины Y от X согласно (1.5.40) среднее значение условной дисперсии прогнозируемого значения случайной величины Y равно безусловной (априорной) дисперсии $\bar{D}_{Y|x} = D_Y$. Иначе говоря, наблюдение $X = x$ не добавляет никакой дополнительной информации о значении случайной величины Y по сравнению с информацией, определяемой априорным средним значением m_Y. С другой стороны, если случайные величины X и Y между собой связаны детерминированной функциональной зависимостью $Y \equiv m_Y(x)$, то $\eta^2_{Y|x} = 1$, так как согласно (1.5.40) $\bar{D}_{Y|x} = 0$ для произвольного значения $X = x$. В общем же случае для корреляционного отношения выполняется неравенство

$$0 \leqslant \eta^2_{Y|x} \leqslant 1.$$

В отличие от коэффициента корреляции ρ_{XY}, описывающего линейную статистическую зависимость между случайными величинами X и Y, корреляционное отношение $\eta^2_{Y|x}$ характеризует как линейную, так и нелинейную статистическую зависимость случайной величины Y от случайной величины X.

Согласно неравенству Буняковского–Шварца (1.5.33) для квадрата корреляционного момента случайных величин X и Y можно записать

$$K^2_{XY} = D_X D_Y \rho^2_{XY} = \left[\iint (x - m_X)(y - m_Y) f_X(x) f_Y(y \mid x)\,dxdy\right]^2 =$$
$$= \left[\int (x - m_X)[m_Y(x) - m_Y] f_X(x)\,dx\right]^2 \leqslant$$
$$\leqslant \int (x - m_X)^2 f_X(x)\,dx \int [m_Y(x) - m_Y]^2 f_X(x)\,dx =$$
$$= D_X D_m = D_X D_Y \eta^2_{Y|x}.$$

Это неравенство переходит в равенство, если величины $[m_Y(x) - m_Y]$ и $(x - m_X)$ связаны между собой детерминированной линейной зависимостью. Таким образом, справедливо соотношение

$$\rho^2_{XY} \leqslant \eta^2_{Y|x},$$

причем равенство будет только при детерминированной линейной зависимости (регрессии) между случайными величинами X и Y. Поэтому выражение, определяющее связь между квадратом коэффициента корреляции и корреляционным отношением, принимает вид

$$0 \leqslant \rho^2_{XY} \leqslant \eta^2_{Y|x} \leqslant 1. \tag{1.5.44}$$

При этом, с учетом выше отмеченного, справедливы соотношения:

а) $\rho_{XY}^2 = \eta_{Y/x}^2 = 0$, если случайные величины X и Y независимы; обратное неверно;

б) $\rho_{XY}^2 = \eta_{Y/x}^2 = 1$, если между случайными величинами X и Y имеется детерминированная линейная зависимость;

в) $\rho_{XY}^2 \leqslant \eta_{Y/x}^2 = 1$, если между случайными величинами X и Y имеется детерминированная нелинейная зависимость;

г) $\rho_{XY}^2 = \eta_{Y/x}^2 < 1$, если между случайными величинами X и Y имеется статистическая линейная зависимость;

д) $\rho_{XY}^2 < \eta_{Y/x}^2 < 1$, если между случайными величинами X и Y имеется статистическая нелинейная зависимость.

Уравнения регрессии и их статистические характеристики принципиально могут быть распространены и на дискретные случайные величины X_k и Y_k, $k = 1, 2, \ldots, q$.

Корреляционное отношение представляет собой некоторую величину, усредненную по всем возможным значениям случайной величины X. Поэтому оно не является мерой нелинейности для какого-либо конкретного вида зависимости X от Y. Однако значение величины $(\eta_{Y|x}^2 - \rho_{XY}^2)$ может служить индикатором нелинейности регрессии Y на X [15, 30].

Вместо независимой переменной $X = x$ можно рассматривать независимую переменную $Y = y$. Тогда условная плотность вероятности зависимой переменной X при фиксированном значении $Y = y$ описывается формулой (1.5.13). При этом кривые регрессии для условного среднего значения величины $x^* = m_X(y)$ определяются аналогично (1.5.34):

$$x^* = m_X(y) = \int_{-\infty}^{\infty} x f_X(x \mid y)\, dx. \qquad (1.5.45)$$

Уравнение (1.5.45) есть регрессия случайной величины X на Y. Все полученные выше соотношения для регрессии Y на X останутся справедливыми при замене x на y. При этом между корреляционными отношениями $\eta_{Y|x}^2$ и $\eta_{X|y}^2$ нет какой-либо простой зависимости [15, 57].

Две кривые регрессии $m_Y(x)$ и $m_X(y)$, определяемые соответственно (1.5.34) и (1.5.45), в общем случае не будут совпадать, как это показано на рис. 1.5.3. Если случайные величины некоррелированы, то кривые регрессий Y на X $[m'_Y(x) \equiv m_Y]$ и X на Y $[m'_X(y) \equiv m_X]$ будут представлять собой прямые, параллельные осям координат (пунктирные линии). Кривые регрессий $m_Y(x)$ и $m_X(y)$ будут совпадать тогда и только тогда, когда между случайными величинами X и Y будет детерминированная взаимно однозначная функциональная зависимость $Y = \varphi(x)$ и соответственно $X = \psi(y)$.

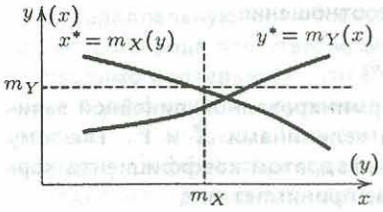


Рис. 1.5.3. Кривые регрессии $m_Y(x)$ и $m_X(y)$

1.5.3. Регрессия второго рода

Для наглядности изложения будем полагать, что случайные величины X и Y непрерывны, имеют совместную плотность вероятности $f(x, y)$ с конечными средними значениями $\mathrm{M}\{X\} = m_X$ и $\mathrm{M}\{Y\} = m_Y$, дисперсиями $D_X = \mathrm{M}\{(X - m_X)^2\}$ и $D_Y = \mathrm{M}\{(Y - m_Y)^2\}$, а также взаимным корреляционным моментом $K_{XY} = \mathrm{M}\{(X - m_X)(Y - m_Y)\}$.

Как отмечалось выше, для *регрессии второго рода заранее* (априори) *задается некоторый класс функций* $y = g(x; \mathrm{b})$ *с точностью до параметров* $\mathrm{b}(b_0, b_1, \ldots, b_p)$. Среди всех функций данного класса ищется такая, которая в соответствии с заранее выбранным критерием оптимальности дает лучшее представление (дает лучшую оценку) величины Y. В качестве критерия оптимальности наиболее часто используется критерий минимума среднего квадрата рассеяния случайной величины Y относительно прогнозируемого значения $g(X; \mathrm{b})$:

$$\mathrm{M}\{[Y - g(X; \mathrm{b})]^2\} = \widehat{D}_Y(\mathrm{b}) = \int\!\!\int [y - g(x; \mathrm{b})]^2 f(x, y)\, dx dy. \qquad (1.5.46)$$

Эта величина зависит от неизвестных составляющих векторного параметра $\mathrm{b}(b_0, b_1, \ldots, b_p)$. Минимизируя ее по значениям неизвестных параметров b_j, $j = 0, 1, \ldots, p$, находим систему $p+1$ уравнений для их определения

$$\partial \widehat{D}_Y(\mathrm{b}) / \partial b_j = 0, \quad j = 0, 1, 2, \ldots, p. \qquad (1.5.47)$$

Кривые регрессии второго рода, полученные с применением критерия минимума рассеяния оценки относительно прогнозируемой зависимости, называют *кривыми средней квадратической регрессии*.

1.5.3.1. Линейная средняя квадратическая регрессия

Простейшим видом кривой средней квадратической регрессии Y на X является *линейная средняя квадратическая регрессия* (в дальнейшем линейная регрессия)

$$y^* = g(x; b_0, b_1) = b_0 + b_1 x. \qquad (1.5.48)$$

При этом будем полагать, что случайные величины X и Y имеют конечные средние значения m_X и m_Y, дисперсии D_X и D_Y и корреляционный момент K_{XY}. Для нахождения параметров регрессии b_0 и b_1 выражение (1.5.46) перепишем в виде

$$\begin{aligned} \widehat{D}_Y(b_0, b_1) &= \mathrm{M}\{(Y - b_0 - b_1 X)^2\} = \\ &= \mathrm{M}\{[(Y - m_Y) - b_1(X - m_X) + (m_Y - b_0 - b_1 m_X)]^2\} = \\ &= D_Y + b_1^2 D_X + (m_Y - b_0 - b_1 m_X)^2 - 2 b_1 K_{XY}. \end{aligned} \qquad (1.5.49)$$

Решая уравнения $\partial \widehat{D}_Y(b_0, b_1)/\partial b_0 = 0$ и $\partial \widehat{D}_Y(b_0, b_1)/\partial b_1 = 0$ относительно коэффициентов b_0 и b_1, получим

$$b_1 = K_{XY}/D_X = \rho_{XY}(\sigma_Y/\sigma_X); \quad b_0 = m_Y - b_1 m_X, \qquad (1.5.50)$$

где, как и выше, $\rho_{XY} = K_{XY}/(\sigma_X\sigma_Y)$ — коэффициент корреляции случайных величин X и Y. Следовательно, уравнение линейной регрессии Y на X запишется в виде

$$y^* = m_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - m_X), \tag{1.5.51}$$

или

$$\frac{y^* - m_Y}{\sigma_Y} = \rho_{XY}\frac{x - m_X}{\sigma_X}. \tag{1.5.52}$$

Написанные выше уравнения прямых проходят через точку с координатами m_X, m_Y. Величина $b_1 \equiv b_{Y|x}$ — коэффициент линейной регрессии Y на X. Подставляя (1.5.50) в (1.5.49), находим минимальное значение рассеяния $\widehat{D}_Y(b_0, b_1)$

$$\min \widehat{D}_Y(b_0, b_1) = \mathrm{M}\{(Y - b_0 - b_1X)^2\} = \frac{D_X D_Y - K_{XY}^2}{D_X} = D_Y(1 - \rho_{XY}^2). \tag{1.5.53}$$

Выше выражением (1.5.43) было введено понятие квадрата корреляционного отношения $\eta_{Y|x}^2$. Для линейной регрессии определим связь $\eta_{Y|x}^2$ с квадратом коэффициента корреляции ρ_{XY}^2, для чего в формуле (1.5.43) следует положить $m_Y(x) = y^*$ из (1.5.51). В результате элементарных вычислений имеем

$$\eta_{Y|x}^2 = \rho_{XY}^2. \tag{1.5.54}$$

В подтверждение сказанному выше при рассмотрении регрессии первого рода из формулы (1.5.53) видно, что, если $\rho_{XY} = 0$, то минимальное значение рассеяния $\widehat{D}_Y(b_0, b_1)$ равно безусловной дисперсии D_Y. В этом случае знание $X = x$ не добавляет информации об оценке случайной величины Y, и наилучшей оценкой для Y является априорное среднее значение m_Y. Наоборот, если $\rho_{XY} = 1$, то $\min \widehat{D}_Y(b_0, b_1) = 0$, т.е. между случайными величинами X и Y существует детерминированная функциональная линейная зависимость.

Покажем, что рассматриваемая линейная средняя квадратическая регрессия наилучшим образом (в смысле минимума среднего квадратического отклонения) приближается к кривой условного среднего регрессии первого рода $m_Y(x)$, если условное среднее существует:

$$\begin{aligned}\mathrm{M}\{(Y - b_0 - b_1X)^2\} &= \mathrm{M}\{([Y - m_Y(x)] + [m_Y(x) - b_0 - b_1X])^2\} = \\ &= \mathrm{M}\{[Y - m_Y(x)]^2\} + \mathrm{M}\{[m_Y(x) - b_0 - b_1X]^2\} - \\ &\quad - 2\mathrm{M}\{[Y - m_Y(x)][m_Y(x) - b_0 - b_1X]\}.\end{aligned}$$

Так как при фиксированном $X = x$ математическое ожидание $\mathrm{M}\{Y - m_Y(x)\} = 0$, последнее слагаемое этого соотношения равно нулю. Поэтому справедлива запись

$$\mathrm{M}\{(Y - b_0 - b_1X)^2\} = \mathrm{M}\{[Y - m_Y(x)]^2\} + \mathrm{M}\{[m_Y(x) - b_0 - b_1X]^2\}.$$

Первое слагаемое в правой части не зависит от коэффициентов b_0 и b_1. Так как левая часть достигает минимального значения при b_0 и b_1, определяемыми (1.5.50), то второе слагаемое в правой части также будет минимальным при тех же значениях b_0 и b_1.

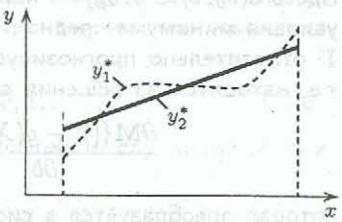


Рис. 1.5.4. Аппроксимация зависимости линейной регрессией

На рис. 1.5.4 в качестве иллюстрации приведены кривая регрессии первого рода $y_1^* = m_Y(x)$ и аппроксимирующая ее линейная регрессия второго рода $y_2^* = b_0 + b_1x$. Разность $y_1^* - y_2^* = m_Y(x) - b_0 - b_1x$ является расстоянием по вертикали между кривой регрессии $y_1^* = m_Y(x)$ и прямой $y_2^* = b_0 + b_1x$. Поэтому можно утверждать, что если кривая регрессии первого рода $y^* = m_Y(x)$ является прямой линией, то она совпадает с прямой средней квадратической регрессии (линейной регрессией второго рода).

В случае линейной регрессии X на Y все соотношения и выводы остаются справедливыми при замене y на x. В частности, уравнение линейной регрессии X на Y имеет вид

$$\frac{x^* - m_X}{\sigma_X} = \rho_{XY}\frac{y - m_Y}{\sigma_Y}. \tag{1.5.55}$$

При этом минимальное рассеяние $\widehat{D}_X(b_0, b_1)$

$$\min \widehat{D}_X(b_0, b_1) = D_X(1 - \rho_{XY}^2). \tag{1.5.56}$$

Обе прямые (1.5.52) и (1.5.55) проходят через центр $m_Y(x)$, $m_X(y)$, совпадают только при $\rho_{XY} = \pm 1$ и перпендикулярны при $\rho_{XY} = 0$.

Из рассмотренного выше следует, что для применения линейной средней квадратической регрессии и вычисления ее статистических характеристик достаточно знать первые и вторые моменты случайных величин X и Y и их взаимный коэффициент корреляции.

1.5.3.2. Нелинейная средняя квадратическая регрессия

Как правило, вид *нелинейной статистической зависимости* между случайными величинами X и Y выбирается, исходя из рассмотрения конкретной прикладной задачи. Таких зависимостей большое многообразие. При этом статистическая зависимость между случайными величинами X и Y может быть достаточно сложной. Однако в ограниченной области их значений прогноз величины Y в зависимости от $X = x$ может быть описан менее сложной функцией. В частности, любую непрерывную функцию $g(x)$, в том числе и нелинейную регрессию Y на X, на конечном интервале независимой переменной можно приблизить сколь угодно точно параболой p-го порядка:

$$y^* = g[x; \mathfrak{b}(b_0, b_1, \ldots, b_p)] = b_0 + b_1x + b_2x^2 + \cdots + b_px^p. \tag{1.5.57}$$

Здесь $\mathrm{b}(b_0, b_1, \ldots, b_p)$ — неизвестные коэффициенты, определяемые из условия минимума среднего квадрата отклонения случайной величины Y относительно прогнозируемого значения: $\mathrm{M}\{[Y - g(X;\mathrm{b})]^2\} = \min$, т.е. находятся из решения системы уравнений

$$\frac{\partial \mathrm{M}\{[Y - g(X;\mathrm{b})]^2\}}{\partial b_j} = 0, \quad j = 0, 1, \ldots, p,$$

которая преобразуется в систему уравнений

$$\mathrm{M}\{[Y - (b_0 + b_1 X + \ldots + b_p X^p)]X^j\} = 0, \quad j = 0, 1, \ldots, p.$$

Полагая, что начальные моменты $\mathrm{M}\{X^j Y^0\} \equiv m_{j,0}$ и смешанные моменты $\mathrm{M}\{X^j Y^1\} \equiv m_{j,1}$ конечны и априори известны, и вычисляя соответствующие моменты, входящие в последнее выражение, получим систему из $p+1$ линейных уравнений относительно неизвестных коэффициентов b_j ряда (1.5.57)

$$b_0 m_{j,0} + b_1 m_{j+1,0} + \ldots + b_p m_{j+p,0} = m_{j,1}, \quad j = 0, 1, \ldots, p. \qquad (1.5.58)$$

Система линейных уравнений (1.5.58) для оговоренных условий имеет единственное решение относительно неизвестных коэффициентов b_j. Естественно, нахождение неизвестных коэффициентов и, тем более, статистических характеристик оценки y^* существенно усложняется по мере увеличения числа слагаемых в (1.5.57). Кроме того, если возникает необходимость увеличить порядок параболы, систему уравнений приходится решать заново. Поэтому желательно иметь такой полином $g(x)$, с увеличением порядка которого предыдущие коэффициенты ряда оставались бы без изменения. Это возможно, если вместо системы одночленов $1, x, x^2, \ldots, x^p$ *взять систему ортогональных функций* (*полиномов*) $\varphi_n(x)$, удовлетворяющих соотношению

$$\int \varphi_n(x)\varphi_m(x) f(x)\, dx = \begin{cases} 1 & n = m, \\ 0 & n \neq m, \end{cases} \quad n, m = 0, 1, 2, \ldots, \qquad (1.5.59)$$

где $f(x)$ — плотность вероятности случайной величины X.

Ниже в п. 2.3 дано описание полиномов Эрмита и Лаггера, наиболее часто встречающихся в задачах теории вероятностей и математической статистики. С учетом (1.5.59) любая непрерывная функция $g(x;\mathrm{c})$ может быть записана в виде

$$g(x;\mathrm{c}) = c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_p\varphi_p(x), \qquad (1.5.60)$$

где $\mathrm{c}(c_0, c_1, \ldots, c_p)$ — постоянные коэффициенты, определяемые, как и выше, из условия

$$\frac{\partial}{\partial c_j}\mathrm{M}\{[Y - g(X;\mathrm{c})]^2\} = -2\mathrm{M}\{[Y - c_0\varphi_0(X) - c_1\varphi_1(X) - \ldots$$
$$\ldots - c_p\varphi_p(X)]X^j\} = 0, \quad j = 0, 1, \ldots, p.$$

После вычисления соответствующих моментов с учетом условия ортогональности (1.5.59) получаем формулы для вычисления коэффициентов c_j ряда (1.5.60):

$$c_j = \mathrm{M}\{\varphi_j(X)Y\}, \quad j = 0, 1, \ldots, p. \qquad (1.5.61)$$

При этом минимальное значение рассеяния оценки y^* случайной величины Y

$$\min \mathrm{M}\{[Y - g(X;\mathrm{c})]^2\} = \widehat{D}_Y = \mathrm{M}\{Y^2\} - c_0^2 - c_1^2 - \ldots - c_p^2. \qquad (1.5.62)$$

Из формулы (1.5.62) видно, что если известен полином порядка p и находится тот же полином порядка $p+1$, то необходимо только вычислить коэффициент $c_{p+1} = \mathrm{M}\{\varphi_{p+1}(X)Y\}$.

Пример 1.5.3. В качестве иллюстрации полученных соотношений рассмотрим полином второго порядка ($p = 2$). Вместо случайной величины X рассмотрим ее нормированное значение $Z = (X - m_X)/\sigma_X$, а в качестве ортогональных полиномов возьмем ненормированные *ортогональные полиномы Эрмита* $\varphi_n(z) = H_n(z)/\sqrt{n!}$, где $H_n(z)$ определяются формулой (2.3.7). Согласно (2.3.10) первые три ортогональных полинома Эрмита равны: $\varphi_0(z) = H_0(z) = 1$, $\varphi_1(z) = H_1(z) = z$, $\varphi_2(z) = H_2(z)/\sqrt{2} = (z^2 - 1)/\sqrt{2}$. Тогда прогнозируемое значение случайной величины Y (регрессия Y на Z) определяется выражением

$$y^* = g(z;\mathrm{c}) = c_0 + c_1 z + c_2(z^2 - 1)/\sqrt{2},$$

где согласно (1.5.61) коэффициенты c_j, $j = 0, 1, 2$, равны

$$c_0 = \mathrm{M}\{Y\} = m_Y;$$

$$c_1 = \mathrm{M}\left\{YZ = \frac{[(Y - m_Y) + m_Y](X - m_X)}{\sigma_X}\right\} = \sigma_Y \rho_{XY};$$

$$c_2 = \mathrm{M}\left\{Y\frac{(Z^2 - 1)}{\sqrt{2}} = \frac{Y}{\sqrt{2}}\left[\frac{(X - m_X)^2}{\sigma_X^2} - 1\right]\right\} = \frac{\mu_{21}(x, y)}{\sqrt{2}\sigma_X^2}.$$

Следовательно, уравнение квадратичной регрессии Y на X примет вид

$$y^* = m_Y + \frac{\sigma_X}{\sigma_Y}\rho_{XY}(x - m_X) + \frac{\mu_{21}(x, y)}{2\sigma_X^2}\left[\frac{(x - m_X)^2}{\sigma_X^2} - 1\right]. \qquad (1.5.63)$$

Нетрудно видеть, что если ограничиться линейным приближением (коэффициентами модели c_0 и c_1), то получаем выражение (1.5.51).

Вычислим минимальное рассеяние прогнозируемого значения случайной величины Y. Так как $\mathrm{M}\{Y^2\} = D_Y + m_Y^2$, в соответствии с (1.5.62) имеем

$$\widehat{D}_Y = D_Y\left(1 - \rho_{XY}^2 - \frac{\mu_{21}^2(x, y)}{2D_X^2 D_Y}\right). \qquad (1.5.64)$$

1.5.4. Многомерные случайные величины

Все изложенное выше легко обобщается на *многомерные* ($n > 2$) случайные величины или, иначе говоря, на многомерный случайный вектор $\mathbf{X}$ с координатами $X_1, X_2, \ldots, X_n$. Изложим это обобщение применительно к непрерывным случайным величинам X_i, $i = 1, 2, \ldots, n$.

Плотность вероятности многомерной n-мерной случайной величины

$\mathbf{X}(X_1, X_2, \ldots, X_n)$ аналогично (1.2.3) определяется соотношением

$$f_n(x_1, x_2, \ldots, x_n) = \lim_{\substack{\Delta x_1 \to 0 \\ \ldots \\ \Delta x_n \to 0}} \frac{\mathbf{P}\{x_1 \leqslant X_1 < x_1 + \Delta x_1, \ldots, x_n \leqslant X_n < x_n + \Delta x_n\}}{\Delta x_1 \Delta x_2 \ldots \Delta x_n}. \quad (1.5.65)$$

Многомерная плотность вероятности, как и одномерная плотность вероятности, — величина неотрицательная. *Многомерная функция распределения* $F_n(x_1, \ldots, x_n) = \mathbf{P}\{X_1 < x_1, \ldots, X_n < x_n\}$ и многомерная плотность вероятности связаны между собой формулами

$$F_n(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_n(y_1, \ldots, y_n)\, dy_1 \ldots dy_n; \quad (1.5.66)$$

$$f_n(x_1, \ldots, x_n) = \frac{\partial^n F_n(x_1, \ldots, x_n)}{\partial x_1 \ldots \partial x_n}. \quad (1.5.67)$$

Интегрируя n-мерную плотность вероятности по некоторым из переменных x_i, $i = 1, \ldots, n$, например по $x_{m+1}, x_{m+2}, \ldots, x_n$, получим m-мерную плотность вероятности по оставшимся переменным

$$f_m(x_1, \ldots, x_m) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_n(x_1, \ldots, x_m, x_{m+1} \ldots, x_n)\, dx_{m+1} \ldots dx_n. \quad (1.5.68)$$

Все сказанное в п. 1.5.1 относительно условных распределений двумерных случайных величин также обобщается и на условные распределения n-мерной случайной величины. Обобщаются и соотношения для начальных и центральных моментов. *Начальный многомерный момент n-мерной случайной величины $\mathbf{X}(X_1, \ldots, X_n)$ порядка $k_1 + \ldots + k_n$ определяется* как

$$m_{k_1, \ldots, k_n} = \mathbf{M}\{X_1^{k_1} \ldots X_n^{k_n}\}. \quad (1.5.69)$$

Соответственно *центральный многомерный момент порядка* $k_1 + \ldots + k_n$ имеет вид

$$\mu_{k_1, \ldots, k_n} = \mathbf{M}\{(X_1 - m_{X_1})^{k_1} \ldots (X_n - m_{X_n})^{k_n}\}, \quad (1.5.70)$$

где $m_{X_i} = \mathbf{M}\{X_i\}$ — среднее значение i-й компоненты n-мерной случайной величины, которое в обозначениях (1.5.69) соответствует тому, что один из индексов $k_i = 1$, а остальные индексы равны нулю. Если один из индексов в (1.5.70) положить равным двум (например, $k_i = 2$), а остальные приравнять нулю, то получим дисперсию $\mu_{2,0,\ldots,0} \equiv \mu_2(X_i) \equiv D_{X_i}$ случайной компоненты X_i. Если в (1.5.70) два индекса (например, $k_i = k_j = 1$) равны единице, а остальные равны нулю, получаем корреляционный момент $\mu_{x_i x_j} = K_{ij} \equiv K_{ji}$ случайных величин X_i и X_j, $i, j = 1, \ldots, n$, причем очевидно, что, если индексы совпадают, получаем формулу для дисперсии $K_{ii} = D_{X_i}$.

Совокупность корреляционных моментов образует *ковариационную матрицу* многомерной случайной величины

$$\mathbf{K} = \begin{pmatrix} K_{11} & \ldots & K_{1n} \\ \vdots & \ddots & \vdots \\ K_{n1} & \ldots & K_{nn} \end{pmatrix}. \quad (1.5.71)$$

Ковариационная матрица многомерной случайной величины симметрична относительно главной диагонали, являющейся дисперсиями составляющих многомерной случайной величины. Вместо корреляционной матрицы n-мерной случайной величины обычно рассматривают нормированную матрицу, составленную из коэффициентов корреляции (1.5.31) и называемую *корреляционной матрицей*

$$\mathbf{R} = \begin{pmatrix} 1 & \ldots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \ldots & 1 \end{pmatrix}, \quad (1.5.72)$$

где $\rho_{ij} = K_{ij} / \sqrt{D_{X_i} D_{X_j}}$.

Для многомерной случайной величины кроме рассмотренных выше моментов применяют показатели, которые характеризуют зависимость их от всей совокупности случайного вектора. Сюда можно отнести понятия частного коэффициента корреляции и множественного коэффициента корреляции [15, 31, 37 и др.].

Частный коэффициент корреляции опрделяет меру линейной зависимости между двумя случайными величинами X_i и X_j после устранения линейных изменений, вызванных влиянием других величин X_r, $r \neq i, j$. Применительно к случайным величинам X_1 и X_2 из некоторой совокупности $(X_1, X_2, \ldots, X_n)$ частный коэффициент корреляции, который часто обозначается как $\rho_{12.(34\ldots n)}$, может быть выражен через алгебраические дополнения R_{12}, R_{11}, R_{22} элемента ρ_{ij} корреляционной матрицы (1.5.72) в виде

$$\rho_{12.(34\ldots n)} = -\frac{R_{12}}{\sqrt{R_{11} R_{22}}}. \quad (1.5.73)$$

В общем случае частный коэффициент корреляции $\rho_{12.(34\ldots n)}$ отличается от парного коэффициента корреляции ρ_{12} величин X_1 и X_2. Поэтому по различию между $\rho_{12.(34\ldots n)}$ и ρ_{12} можно судить о том, зависимы величины X_1 и X_2 между собой или эта зависимость является следствием зависимости каждой из них от $X_3, X_4, \ldots, X_n$. Из (1.5.73) нетрудно видеть, что $-1 \leqslant \rho_{12.(34\ldots n)} \leqslant +1$, причем если величины $X_1, X_2, \ldots, X_n$ попарно некоррелированы ($\rho_{ij} = 0$, $R_{ij} = 0$, $i \neq j$), то частный коэффициент корреляции $\rho_{12.(34\ldots n)}$ также равен нулю. Перестановкой индексов получаются аналогичные выражения для частного коэффициента корреляции двух величин относительно остающихся $n - 2$ величин.

Применительно к трехмерной случайной величине X_1, X_2, X_3

$(n = 3)$ частный коэффициент корреляции между случайными величинами X_1 и X_2 определяется как

$$\rho_{12.(3)} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}. \qquad (1.5.74)$$

Пример 1.5.4. Пусть парные коэффициенты корреляции между случайными величинами X_1, X_2, X_3 равны $\rho_{12} = 0{,}12$; $\rho_{13} = 0{,}55$; $\rho_{23} = 0{,}99$. Вычисляя согласно (1.5.74) частные коэффициенты корреляции, имеем $\rho_{12.(3)} \approx 0{,}54$; $\rho_{13.(2)} \approx 0{,}53$; $\rho_{23.(1)} \approx 0{,}99$. Отсюда видно, что случайная величина X_3 влияет на корреляцию между случайными величинами X_1 и X_2.

Множественный коэффициент корреляции $\rho_{1.(234\ldots n)}$ является мерой линейной зависимости между случайной величиной X_1 и остальными случайными величинами $X_2, X_3, \ldots, X_n$. Множественный коэффициент корреляции определяется как обычный парный коэффициент корреляции между X_1 и его условным математическим ожиданием $\mathrm{M}\{X_1 | x_2, x_3, \ldots, x_n\}$ или, другими словами, $\rho_{1.(234\ldots n)}$ определяется как обычный парный коэффициент корреляции между X_1 и наилучшим линейным приближением (в смысле минимума рассеяния) X_1 по $X_2, X_3, \ldots, X_n$, т.е. регрессией величины X_1 по $X_2, X_3, \ldots, X_n$. При этом коэффициенты $a_2, a_3, \ldots, a_n$ такого линейного приближения определяются из соотношения

$$\min_a \mathrm{M}\left\{\left(X_1 - \sum_{i=2}^{n} a_i X_i\right)^2\right\}.$$

Аналитически квадрат множественного коэффициента корреляции определяется как

$$\rho^2_{1.(234\ldots n)} = 1 - \frac{R}{R_{11}} = 1 - (1-\rho^2_{13.(2)})(1-\rho^2_{14.(23)})\cdots$$
$$\cdots(1-\rho^2_{1n.(23\ldots(n-1))}). \qquad (1.5.75)$$

Здесь R — определитель матрицы (1.5.72); R_{11} — алгебраическое дополнение элемента ρ_{11} этой матрицы; $\rho_{12.(3)}, \rho_{12.(34)}, \cdots \rho^2_{12.(34\ldots(n-1))}$ — частные коэффициенты корреляции. При этом $0 \leqslant \rho^2_{1.(234\ldots n)} \leqslant 1$. Если $\rho^2_{1.(234\ldots n)} = 0$, то из (1.5.75) следует, что парный коэффициент корреляции ρ_{12} и все частные коэффициенты корреляции $\rho^2_{1p.(23\ldots(p-1))}$, $p = 3, 4, \ldots, n$, равны нулю, т.е. случайная величина X_1 некоррелирована ни с одной из остальных случайных величин. С другой стороны, если $\rho^2_{1.(234\ldots n)} = 1$, то хотя бы один из частных коэффициентов корреляции $\rho^2_{1p.(23\ldots(p-1))}$ должен быть равен единице, т.е. случайная величина X_1 равна некоторой линейной комбинации величин $X_2, X_3, \ldots, X_n$. При этом, если $n = 2$, то множественный коэффициент корреляции совпадает с парным коэффициентом корреляции $(\rho_{1.(2)} \equiv \rho_{12})$.

Так же, как и для частного коэффициента корреляции, перестановкой индексов получаются аналогичные выражения для множественного коэффициента корреляции случайной величины X_i, $i = 2, 3, \ldots, n$, относительно оставшихся $n-1$ величин.

Пример 1.5.5. Вычислим множественный коэффициент корреляции между случайной величиной X_1 и случайными величинами X_2 и X_3 применительно к данным предыдущего примера. Согласно (1.5.75) имеем

$$\rho_{1.(23)} = 1 - (1-\rho^2_{12})(1-\rho^2_{13.(2)}) = 1 - (1-0{,}12^2)(1-(-0{,}53)^2) \approx 0{,}364.$$

1.5.5. Двумерное и многомерное нормальное распределение

Двумерная плотность вероятности гауссовских случайных величин X и Y определяется выражением

$$f_2(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_X)^2}{\sigma_X^2} - \right.\right.$$
$$\left.\left. - \frac{2\rho(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2}\right]\right\}. \qquad (1.5.76)$$

Интегрируя $f_2(x,y)$ по переменной y или x, получим формулы одномерных распределений нормальной случайной величины в виде (1.4.1) со средними значениями m_X, m_Y и дисперсиями σ_X^2, σ_Y^2 соответственно для случайных величин X и Y. Параметр ρ в (1.5.76) является коэффициентом корреляции X и Y.

Если гауссовские случайные величины некоррелированы $(\rho = 0)$, из формулы (1.5.76) следует, что совместная плотность вероятности равна произведению плотностей вероятности каждой из величин. По определению такие величины называются независимыми. Следовательно, если две совместно гауссовские случайные величины некоррелированы, то они и независимы, т.е. некоррелированность двух совместно гауссовских случайных величин тождественна их независимости.

Рассмотрим условную плотность вероятности гауссовской случайной величины Y при условии $X = x$. Согласно (1.5.14) *условная плотность вероятности* случайной величины Y может быть представлена в виде

$$f_Y(y \mid x) = \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\frac{[y - m_Y - \rho(\sigma_Y/\sigma_X)(x-m_X)]^2}{2\sigma_Y^2(1-\rho^2)}\right\}. \qquad (1.5.77)$$

Из этого выражения видно, что условная случайная величина Y при фиксированном значении $X = x$ распределена по нормальному закону с условным средним значением и дисперсией:

$$m_Y(x) \equiv m_{Y|x} = m_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - m_X); \qquad (1.5.78)$$

$$D_Y(x) \equiv D_{Y|x} = \sigma_Y^2(1-\rho^2). \qquad (1.5.79)$$

Сравнивая уравнение (1.5.78) с уравнением линейной средней квадратической регрессии (1.5.51), убеждаемся в их идентичности: $y^* \equiv \equiv m_Y(x)$. Это свидетельствует о том, что условные средние значения зависимых между собой гауссовских случайных величин являются уравнениями линейной средней квадратической регрессии.

Таким образом, условное среднее значение $m_Y(x)$, являющееся наилучшим значением прогнозируемой случайной величины Y, представляет собой линейную функцию от конкретного значения случайной величины $X = x$. При этом условная дисперсия $D_{Y|x}$ не зависит от значения дисперсии, принимаемой случайной величиной X. Аналогичные свойства будут проявляться и у условного распределения случайной величины X при фиксированном значении второй случайной величины $Y = y$.

Две коррелированные (зависимые) совместно гауссовские случайные величины X, Y всегда можно привести к двум некоррелированным (независимым) гауссовским случайным величинам Z, V с помощью соответствующего функционального преобразования. В общем виде задача функциональных преобразований случайных величин изучается ниже в гл. 2. Здесь же рассматривается частный случай линейного преобразования для получения независимых нормально распределенных случайных величин. Для этого рассмотрим положительную квадратичную форму, стоящую в показателе экспоненты (1.5.76),

$$\frac{(x-m_X)^2}{\sigma_X^2} - \frac{2\rho(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \frac{(y-m_Y)^2}{\sigma_Y^2} = \lambda^2, \qquad (1.5.80)$$

где λ — некоторая постоянная. Выражение (1.5.80) представляет собой уравнение эллипса (уравнение постоянной плотности вероятности) с центром в точке $0'$ с координатами (m_X, m_Y). При этом одна из осей симметрии эллипса составляет угол α с осью абсцисс. Примерный вид эллипса приведен на рис. 1.5.5.

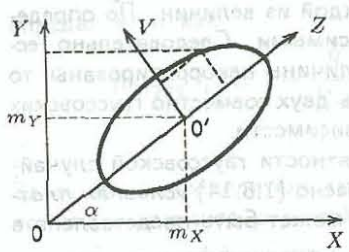


Рис. 1.5.5. Эллипс постоянной плотности вероятности

Очевидно, что корреляционный момент двух случайных величин не зависит от того, в каком порядке берутся эти случайные величины.

Если коэффициент корреляции равен нулю ($\rho = 0$), оси эллипса будут параллельны осям координат $0X$ и $0Y$. Определим угол α из условия, что в системе координат $Z0'V$ случайные величины Z и V некоррелированы. Координаты случайных точек в системе $Z0'V$ связаны с координатами этих же точек в системе $X0Y$ линейными соотношениями

$$Z = (x-m_X)\cos\alpha + (Y-m_Y)\sin\alpha;$$
$$V = -(x-m_X)\sin\alpha + (Y-m_Y)\cos\alpha.$$

Отсюда видно, что средние значения случайных величин Z и V равны нулю. Угол α находится из условия некоррелированности случайных величин Z и V: $\mathrm{M}\{ZV\} = K_{ZV} = 0$. Вводя центрированные случайные величины $X_0 = X - m_X$, $Y_0 = Y - m_Y$, имеем

$$K_{ZV} = -\mathrm{M}\{X_0^2\}\sin\alpha\cos\alpha - \mathrm{M}\{X_0Y_0\}\sin^2\alpha + \mathrm{M}\{X_0Y_0\}\cos^2\alpha +$$
$$+\mathrm{M}\{Y_0^2\}\sin\alpha\cos\alpha = -\frac{1}{2}(\sigma_X^2 - \sigma_Y^2)\sin 2\alpha + K_{XY}\cos 2\alpha = 0.$$

В результате получаем

$$\mathrm{tg}2\alpha = \frac{2K_{XY}}{\sigma_X^2 - \sigma_Y^2}. \qquad (1.5.81)$$

При этом плотность вероятности двух совместно гауссовских случайных величин после преобразования принимает вид

$$f(z,v) = \frac{1}{2\pi\sigma_Z\sigma_V}\exp\left[-\frac{z^2}{2\sigma_Z^2} - \frac{v^2}{2\sigma_V^2}\right]. \qquad (1.5.82)$$

Здесь σ_Z^2 и σ_V^2 — дисперсии случайных величин Z и V, связанные с дисперсиями и коэффициентом корреляции исходных случайных величин X и Y, определяются выражениями

$$\sigma_Z^2 = \mathrm{M}\{Z^2\} = \sigma_X^2\cos^2\alpha + \rho\sigma_X\sigma_Y\sin 2\alpha + \sigma_Y^2\sin^2\alpha;$$
$$\sigma_V^2 = \mathrm{M}\{V^2\} = \sigma_X^2\sin^2\alpha - \rho\sigma_X\sigma_Y\sin 2\alpha + \sigma_Y^2\cos^2\alpha. \qquad (1.5.83)$$

Запись двумерной плотности вероятности в виде (1.5.82) иногда называют *канонической формой* записи двумерной нормальной плотности вероятности.

Многомерная плотность вероятности нормальной n-мерной случайной величины $X\,(X_1, \ldots, X_n)$ определяется в общем виде формулой

$$f_n(x_1, \ldots, x_n) = \qquad (1.5.84)$$
$$= \frac{1}{\sigma_1 \ldots \sigma_n (2\pi)^{n/2}\sqrt{A}}\exp\left[-\frac{1}{2A}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{ij}\frac{(x_i - m_i)}{\sigma_i}\frac{(x_j - m_j)}{\sigma_j}\right].$$

Здесь, как и выше, m_i и σ_i^2 — среднее значение и дисперсия случайной величины X_i; A — определитель матрицы $\mathbf{R}$ n-го порядка, составленный из коэффициентов корреляции (1.5.72) ($A = \det\|\rho_{ij}\|$); A_{ij} — алгебраическое дополнение элемента ρ_{ij} определителя A.

Из формулы (1.5.84) и матриц корреляционных моментов и коэффициентов корреляции видно, что многомерная плотность вероятности гауссовской случайной величины полностью определяется средними значениями составляющих и их взаимными корреляционными моментами. Если составляющие X_i и Y_j, $i, j = 1, \ldots, n$, некоррелированы ($\rho_{ij} = 0$, $i \neq j$), то определитель A и алгебраическое дополнение A_{ij} матрицы

(1.5.72) соответственно равны $A = A_{ii} = 1$, а формула (1.5.84) преобразуется в произведение одномерных плотностей вероятности:

$$f_n(x, \ldots, x) = \frac{1}{(2\pi)^{n/2}\sigma_1 \ldots \sigma_n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m_i)^2}{\sigma_i^2}\right]. \qquad (1.5.85)$$

Таким образом, некоррелированность гауссовских случайных величин означает и их независимость, и наоборот.

Контрольные вопросы и задачи

1.1. Что позволяет оценить формула Байеса?

1.2. Каким функциональным соотношением связаны начальные моменты с характеристической функцией?

1.3. В каких задачах теории вероятностей используются биномиальное и пуассоновское распределения?

1.4. Получить формулы для среднего значения и дисперсии для биномиальной случайной величины через параметры бернуллиевской случайной величины.

1.5. Чем отличается корреляционное отношение от коэффициента корреляции?

1.6. По каналу связи, подверженному воздействию помех, передается один из возможных сигналов A_1 или A_2 с априорными вероятностями 0,8 и 0,2 соответственно. Вероятность ошибочного приема сигналов равна 0,4. Принят сигнал A_1. Определить апостериорные вероятности переданных сигналов.

1.7. Применительно к данным примера 1.1.1 построить функцию распределения и многоугольник распределения вероятностей непопадания и попадания выстрелов в цель соответственно одного, двух и трех раз.

1.8. Двухстороннее показательное распределение (*распределение Лапласа*) имеет плотность вероятности $f(x) = C\exp(-\alpha|x|)$, $-\infty < x < \infty$. Определить значение постоянной C, функцию распределения, а также вычислить дисперсию случайной величины X.

1.9. В задачах экономической статистики часто используют усеченное *распределение Парето* $F(x) = \mathbf{P}\{X < x\} = 1 - (x_0/x)^{\alpha}$, $(x > x_0,\ \alpha > 0)$. Определить плотность вероятности этого распределения и его среднее значение при $\alpha > 1$.

1.10. В каждой из $n = 4$ фиксированных точек во времени возможно превышение случайным по амплитуде импульсом некоторого порога с вероятностью $p = 0{,}8$. Определить вероятность превышения уровня один, два, три и четыре раза. Какова вероятность того, что уровень будет превышен не менее трех раз?

1.11. Полагая $p = 0{,}5$ и $n = 8$, проверить что биномиальное распределение симметрично относительно среднего значения. Построить многоугольник распределения.

1.12. Вычислить вероятность появления не менее $k = 12$ событий пуассоновской случайной величины с параметром $\lambda = 10$.

1.13. Прибор состоит из 480 компонент. В течение года из-за ненадежности компонент приходится заменять в среднем 48 компонент. Определить вероятность того, что за один месяц работы прибора придется менять не более трех компонент.

1.14. Показать, что среднее значение суммы n экспоненциально распределенных (1.4.20) случайных величин равно $n\lambda$, а плотность вероятности этой суммы является распределением Эрланга (1.4.19) с параметрами λ и $\nu = n$.

1.15. Время безотказной работы системы подчиняется экспоненциальному закону. Вычислить вероятность того, что система после девяти часов буде исправна, если известно, что среднее время до ремонта составляет сто часов.

1.16. Используя соотношение (1.5.43), показать, что для двумерной гауссовской случайной величины корреляционное отношение равно квадрату коэффициента корреляции.

ГЛАВА 2

Функциональные преобразования случайных величин

В прикладных исследованиях часто возникает задача функционального линейного и нелинейного преобразования случайных величин с известными законами распределения в другие случайные величины. В связи с этим для дальнейшего изучения преобразованных случайных величин необходимо определить их законы распределения и другие вероятностные характеристики.

2.1. Преобразования плотностей вероятности

2.1.1. Преобразование одномерных плотностей вероятности

Пусть известна плотность вероятности $f_X(x)$ случайной величины X и требуется найти плотность вероятности $f_Y(y)$ случайной величины $Y = \varphi(X)$, где $y = \varphi(x)$ — однозначная дифференцируемая детерминированная функция, для которой существует однозначная обратная функция $x = \psi(y)$. Так как случайные величины связаны однозначной детерминированной зависимостью, то из факта, что X заключено в достаточно малом интервале $[x, x+dx]$, достоверно следует, что случайная величина Y будет находиться в интервале $[y, y+dy]$ (рис. 2.1.1).

Отсюда следует, что вероятности этих двух событий равны, т.е. применительно к произвольным значениям переменных x и y выполняется условие

$$f_Y(y)|dy| = f_X(x)|dx|. \qquad (2.1.1)$$

Отсюда имеем

$$f_Y(y) = f_X[x = \psi(y)]\left|\frac{d[x = \psi(y)]}{dy}\right|. \qquad (2.1.2)$$

В последних выражениях знак модуля производных введен в связи с тем, что вероятность и плотность вероятности по определению действительные неотрицательные величины.

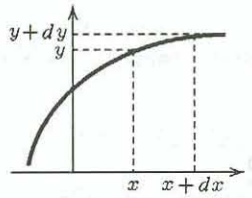


Рис. 2.1.1. Взаимнооднозначная зависимость

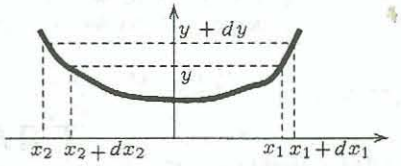


Рис. 2.1.2. Двузначное нелинейное преобразование

Обратная функция $x = \psi(y)$ может быть неоднозначной, т.е. одному значению y соответствует несколько значений x_i, $i = 1, 2, \ldots, k$; иначе, имеется несколько ветвей $x_i = \psi_i(y)$ обратной функции $\psi(y)$. Подобная ситуация возникает тогда, когда нелинейная функция $y = \varphi(x)$ не является монотонной.

Рассмотрим задачу применительно к двум ветвям обратной функции (рис. 2.1.2) $\psi_1(y)$ и $\psi_2(y)$. В данной ситуации выполнение неравенства

$$y \leqslant Y < y + dy \tag{2.1.3}$$

обеспечивается двумя несовместными событиями:

$$x_1 \leqslant X < x_1 + dx_1 \quad \text{или} \quad x_2 \leqslant X < x_2 + dx_2. \tag{2.1.4}$$

Поэтому вероятность выполнения неравенства (2.1.3) должна равняться сумме вероятностей каждого из неравенств (2.1.4):

$$f_Y(y)|dy| = f_X(x_1)|dx_1| + f_X(x_2)|dx_2|.$$

Выразив в правой части переменную x через обратные функции $\psi_1(y)$ и $\psi_2(y)$, окончательно получаем выражение для плотности вероятности

$$f_Y(y) = f_X(\psi_1(y))\left|\frac{d\psi_1(y)}{dy}\right| + f_X(\psi_2(y))\left|\frac{d\psi_2(y)}{dy}\right|. \tag{2.1.5}$$

При большем числе ветвей обратной функции в правой части формулы (2.1.5) следует брать сумму вероятностей по всем ветвям исходной переменной x.

Моменты преобразованной величины Y могут быть найдены без предварительного вычисления плотности вероятности $f_Y(y)$, воспользовавшись формулой математического ожидания произвольной функции (1.2.20). *Начальный и центральный моменты преобразованной случайной величины* Y определяются соответственно как

$$m_k(y) = \mathrm{M}\{Y^k\} = \int_{-\infty}^{\infty} \varphi^k(x) f_X(x)\, dx; \tag{2.1.6}$$

$$\mu_k(y) = \mathrm{M}\{(Y - m_Y)^k\} = \int_{-\infty}^{\infty} [\varphi(x) - m_Y]^k f_X(x)\, dx. \tag{2.1.7}$$

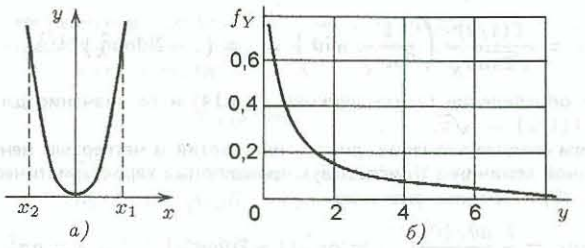


Рис. 2.1.3. Квадратичное преобразование

Вероятность преобразованной дискретной случайной величины Y совпадает с вероятностью преобразуемой дискретной случайной величины X, если фиксированному значению y_i, $i = 1, \ldots, n$, соответствует одно фиксированное значение x_i с вероятностью $p_i = \mathbf{P}\{X = x_i\}$, $i = 1, \ldots, n$, т.е.

$$\mathbf{P}\{Y = y_i\} = \mathbf{P}\{X = x_i\} = p_i, \quad i = 1, 2, \ldots, n. \tag{2.1.8}$$

Пример 2.1.1. В качестве иллюстрации полученных соотношений рассмотрим *квадратичное преобразование* непрерывной случайной величины X. Характеристику преобразования запишем в виде

$$y = ax^2, \quad a > 0. \tag{2.1.9}$$

Так как непрерывная случайная величина Y не может принимать отрицательные значения, плотность вероятности $f_Y(y) = 0$ для всех $y < 0$. При $y \geqslant 0$ обратная функция имеет две ветви (рис. 2.1.3,*а*):

$$\psi_1(y) = x_1 = \sqrt{y/a}; \quad \psi_2(y) = x_2 = -\sqrt{y/a}.$$

Поскольку $d\psi_1(y)/dy = (1/2)(ay)^{-1/2}$ и $d\psi_2(y)/dy = -(1/2)(ay)^{-1/2}$, то согласно формуле преобразования плотности вероятности (2.1.5) имеем

$$f_Y(y) = \frac{1}{2\sqrt{ay}}[f_X(\sqrt{y/a}) + f_X(-\sqrt{y/a})], \quad y \geqslant 0. \tag{2.1.10}$$

Если плотность вероятности $f_X(x)$ — четная функция, т.е. $f_X(x) = f_X(-x)$, то формула (2.1.10) упрощается:

$$f_Y(y) = \frac{1}{\sqrt{ay}} f_X(\sqrt{y/a}), \quad y \geqslant 0. \tag{2.1.11}$$

Пусть плотность вероятности случайной величины X является нормальной с нулевым средним значением и дисперсией σ_X^2:

$$f_X(x) = (\sqrt{2\pi}\sigma_X)^{-1} \exp(-x^2/2\sigma_X^2).$$

Тогда

$$f_Y(y) = \frac{1}{\sqrt{2\pi a y}\sigma_X} \exp\left[-\frac{y}{2a\sigma_X^2}\right], \quad y \geqslant 0. \tag{2.1.12}$$

График этой плотности вероятности для $a = 1$ и $\sigma_X^2 = 1$ приведен на рис. 2.1.3,*б*.

Характеристическая функция случайной величины с плотностью вероятности (2.1.12) определяется выражением

$$\theta_Y(\vartheta) = \frac{1}{\sqrt{2\pi a}\sigma_X} \int_0^{\infty} y^{-1/2} \exp\left[-\left(\frac{1}{2a\sigma_X^2} - i\vartheta\right) y\right] dy =$$

$$= \frac{\Gamma(1/2)}{\sqrt{2\pi}a\sigma_X}\left(\frac{1}{2a\sigma_X^2} - i\vartheta\right)^{-1/2} = (1 - 2i\vartheta a\sigma_X^2)^{-1/2}. \qquad (2.1.13)$$

Здесь учтено определение гамма-функции (1.4.14) и ее значение для $\alpha = -1/2$: $\Gamma(1+\alpha) = \Gamma(1/2) = \sqrt{\pi}$.

Вычислим среднее значение, дисперсию, третий и четвертый центральные моменты случайной величины Y, используя производные характеристической функции согласно (1.2.39). Среднее значение

$$m_Y = \frac{1}{i}\frac{d\theta_Y(\vartheta)}{d\vartheta}\bigg|_{\vartheta=0} = a\sigma_X^2(1 - 2i\vartheta a\sigma_X^2)^{-3/2}|_{\vartheta=0} = a\sigma_X^2.$$

Для вычисления дисперсии, третьего и четвертого центральных моментов аналогично найдем второй, третий и четвертый начальные моменты случайной величины Y:

$$m_2 = 3a^2\sigma_X^4; \quad m_3 = 15a^3\sigma_X^6; \quad m_4 = 105a^4\sigma_X^8.$$

Отсюда и в соответствии с выражениями (1.2.34), связывающими центральные моменты с начальными моментами, получим

$$D_Y = 2a^2\sigma_X^4; \quad \mu_3 = 8a^3\sigma_X^6; \quad \mu_4 = 54a^4\sigma_X^8.$$

Пример 2.1.2. Найдем функцию $y = \varphi(x)$ для преобразования случайной величины X с заданной плотностью вероятности $f_X(x)$ в случайную величину Y с равномерной плотностью вероятности $f_Y(y) = 1/(b-a)$ при $a \leqslant y \leqslant b$, где a и b — заданные крайние значения равномерной плотности вероятности.

Из условия равенства вероятностей

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\,dx = \int_{-\infty}^{y=\varphi(x)} f_Y(y)\,dy = \int_a^y \frac{dy}{b-a} = \frac{y-a}{b-a}$$

имеем

$$y = \varphi(x) = a + (b-a)F_X(x), \quad a \leqslant y \leqslant b.$$

Здесь $F_X(x)$ — нелинейная функция, совпадающая с функцией распределения исходной случайной величины X. Из последнего выражения видно, что для предельных значений случайной величины X $(-\infty, \infty)$, при которых выполняются предельные соотношения $F_X(-\infty) = 0$ и $F_X(\infty) = 1$, случайная величина Y принимает значения a и b соответственно. Следовательно, всякая непрерывная случайная величина с произвольной плотностью вероятности может быть преобразована в равномерно распределенную случайную величину с заданными крайними значениями.

Пример 2.1.3. В практических приложениях для получения случайных чисел с заданным законом распределения $f_Y(y) = dF_Y(y)/dy$ в качестве исходных величин используются случайные величины с равномерным распределением $f_X(x) = 1/(b-a)$, $a \leqslant x \leqslant b$, т.е. необходимо решить задачу, обратную предыдущей. Итак, найдем соответствующую этой задаче характеристику преобразования $y = \varphi(x)$.

Из равенства функций распределения случайных величин X и Y имеем уравнение

$$F_X(x) = \frac{x-a}{b-a} = F_Y(y) = \int_{-\infty}^{y=\varphi(x)} f_Y(y)\,dy,$$

которое надо решить относительно переменной y, что равносильно получению характеристики преобразования $y = \varphi(x)$.

Конкретизируем задачу применительно к отысканию характеристики преобразования для получения случайной величины Y с экспоненциальным распределением (1.4.20). Функция распределения этой случайной величины равна $F_Y(y) = 1 - \exp(-\lambda y)$, $0 \leqslant y < \infty$, $\lambda > 0$. Из решения уравнения

$$\frac{x-a}{b-a} = 1 - \exp(-\lambda y)$$

относительно переменной y находим характеристику преобразования случайной величины с равномерной плотностью вероятности в случайную величину с экспоненциальной плотностью вероятности:

$$y = \varphi(x) = \frac{1}{\lambda}\ln\frac{b-a}{b-x}.$$

Одним из наиболее распространенных функциональных преобразований является *линейное преобразование* вида

$$y = \varphi(x) = a + bx, \qquad (2.1.14)$$

где a и b — некоторые постоянные коэффициенты, характеризующие изменение среднего значения и среднего квадратического отклонения. Учитывая обратное преобразование $x = \psi(y) = (y-a)/b$ и соотношение (2.1.2), выражение для плотности вероятности случайной величины Y примет вид

$$f_Y(y) = \frac{1}{|b|}f_X\left(\frac{y-a}{b}\right). \qquad (2.1.15)$$

Характеристическая функция случайной величины (2.1.14) была получена выше (1.2.21).

Среднее значение m_Y и дисперсия σ_Y^2 линейно преобразованной случайной величины Y связаны с соответствующими моментами m_X и σ_X^2 случайной величины X выражениями

$$m_Y = a + bm_X; \quad \sigma_Y^2 = b^2\sigma_X^2. \qquad (2.1.16)$$

2.1.2. Преобразование многомерных плотностей вероятности

Исходные рассуждения, сделанные в предыдущем пункте, обобщаются и на *преобразование многомерных распределений*. Пусть известна совместная n-мерная плотность вероятности $f_{nX}(x_1, \ldots, x_n)$ случайных величин $\mathbf{X}(X_1, \ldots, X_n)$. Требуется найти n-мерную плотность вероятности $f_{nY}(y_1, \ldots, y_n)$ для случайных величин

$$\begin{aligned} Y_1 &= \varphi_1(X_1, \ldots, X_n), \\ Y_2 &= \varphi_2(X_1, \ldots, X_n), \\ &\ldots\ldots\ldots\ldots\ldots \\ Y_n &= \varphi_n(X_1, \ldots, X_n), \end{aligned} \qquad (2.1.17)$$

где $\varphi_i(\cdots)$ — кусочно непрерывные функции, $i = 1, \ldots, n$. Если существуют однозначные обратные функции

$$\begin{aligned} X_1 &= \psi_1(Y_1, \ldots, Y_n), \\ X_2 &= \psi_2(Y_1, \ldots, Y_n), \\ &\ldots\ldots\ldots\ldots\ldots \\ X_n &= \psi_n(Y_1, \ldots, Y_n), \end{aligned} \qquad (2.1.18)$$

то искомая плотность вероятности определяется формулой

$$f_{nY}(y_1,\ldots,y_n) = f_{nX}[\psi_1(y_1,\ldots,y_n),\ldots,\psi_n(y_1,\ldots,y_n)]|J_n|. \quad (2.1.19)$$

Здесь J_n — *якобиан преобразования* от случайных величин $X_1,\ldots,X_n$ к случайным величинам $Y_1,\ldots,Y_n$ (якобиан преобразования переменных при переходе от одной системы координат к другой системе координат):

$$J_n = \begin{vmatrix} \dfrac{\partial\psi_1}{\partial y_1} & \cdots & \dfrac{\partial\psi_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial\psi_n}{\partial y_1} & \cdots & \dfrac{\partial\psi_n}{\partial y_n} \end{vmatrix}. \quad (2.1.20)$$

Рассмотрим ряд частных случаев, часто встречающихся на практике. Пусть

$$y_1 = \varphi_1(x_1) = x_1; \quad y_2 \equiv y = \varphi_2(x_1,x_2), \quad (2.1.21)$$

причем обратные функции

$$x_1 = \psi_1(y_1) = y_1; \quad x_2 = \psi_2(y_1,y_2) \quad (2.1.22)$$

однозначны. В данном случае якобиан преобразования

$$J_2 = \begin{vmatrix} 1 & 0 \\ \dfrac{\partial\psi_2}{\partial y_1} & \dfrac{\partial\psi_2}{\partial y_2} \end{vmatrix} = \frac{\partial\psi_2}{\partial y_2}. \quad (2.1.23)$$

Поэтому совместная двумерная плотность вероятности случайных величин Y_1, Y_2 имеет вид

$$f_{2Y}(y_1,y_2) = f_{2X}[y_1,\psi_2(y_1,y_2)]\left|\frac{\partial\psi_2}{\partial y_2}\right|. \quad (2.1.24)$$

Из этой формулы получаем одномерную плотность вероятности для одной случайной величины $y \equiv y_2$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{2Y}(y_1,y)\,dy_1 = \int_{-\infty}^{\infty} f_{2X}[y_1,\psi_2(y_1,y)]\left|\frac{\partial\psi_2}{\partial y}\right| dy_1 =$$
$$= \int_{-\infty}^{\infty} f_{2X}[x_1,\psi_2(x_1,y)]\left|\frac{\partial\psi_2}{\partial y}\right| dx_1. \quad (2.1.25)$$

Полученная формула позволяет найти плотность вероятности суммы, разности, произведения и частного двух случайных величин.

1. Сумма $Y = X_1 + X_2$. Тогда $x_1 = \psi_1(y_1,y) = y_1$; $x_2 = \psi_2(y_1,y) = y - y_1$; $|\partial\psi_2/\partial y| = 1$. Следовательно, плотность вероятности суммы двух случайных величин

$$f_Y(y) = \int_{-\infty}^{\infty} f_{2X}(y_1,y-y_1)\,dy_1 = \int_{-\infty}^{\infty} f_{2X}(x_1,y-x_1)\,dx_1. \quad (2.1.26)$$

Если случайные величины X_1 и X_2 независимы, т.е. если $f_{2X}(x_1,x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1)f_{X_2}(y-x_1)\,dx_1. \quad (2.1.27)$$

2. Разность $Y = X_2 - X_1$. Формулы для плотности вероятности двух случайных величин получаются аналогично предыдущим формулам. Для зависимых случайных величин X_1 и X_2

$$f_Y(y) = \int_{-\infty}^{\infty} f_{2X}(y_1,y+y_1)\,dy_1 = \int_{-\infty}^{\infty} f_{2X}(x_1,y+x_1)\,dx_1. \quad (2.1.28)$$

Если X_1 и X_2 независимые случайные величины, то

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1)f_{X_2}(y+x_1)\,dx_1. \quad (2.1.29)$$

3. Произведение $Y = X_1 \times X_2$. При этом $x_2 = \psi_2(y_1,y) = y/x_1$; $|\partial\psi_2/\partial y| = 1/|x_1|$. Следовательно плотность вероятности произведения двух случайных величин

$$f_Y(y) = \int_{-\infty}^{\infty} f_{2X}(x_1,y/x_1)\frac{dx_1}{|x_1|}. \quad (2.1.30)$$

Применительно к независимым случайным величинам X_1 и X_2

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1)f_{X_2}(y/x_1)\frac{dx_1}{|x_1|}. \quad (2.1.31)$$

4. Частное $Y = X_2/X_1$. Тогда $\psi_2(y_1,y) = yx_1$, $|\partial\psi_2/\partial y| = |x_1|$ и поэтому плотность вероятности частного двух случайных величин

$$f_Y(y) = \int_{-\infty}^{\infty} f_{2X}(x_1,yx_1)|x_1|\,dx_1. \quad (2.1.32)$$

Для независимых случайных величин X_1 и X_2

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x_1)f_{X_2}(yx_1)|x_1|\,dx_1. \quad (2.1.33)$$

Пример 2.1.4. В качестве иллюстрации рассмотрим композицию двух независимых случайных величин X_1 и X_2 с равномерной плотностью вероятности $f_{X_1}(x_1) = 1/(b-a)$, $a \leqslant x_1 \leqslant b$, и нормальной плотностью вероятности $f_{X_2}(x_2) = (2\pi\sigma^2)^{-1/2}\exp[-(x_2-m)^2/2\sigma^2]$ соответственно. Применяя формулу сложения (2.1.27), имеем

$$f_Y(y) = \frac{1}{b-a}\frac{1}{\sqrt{2\pi}\sigma}\int_a^b \exp\left[-\frac{(y-x_1-m)^2}{2\sigma^2}\right] dx_1 =$$
$$= \frac{1}{b-a}\left[\Phi\left(\frac{b-(y-m)}{\sigma}\right) - \Phi\left(\frac{a-(y-m)}{\sigma}\right)\right],$$

где $\Phi(z)$, как и выше, интеграл нормального распределения (1.3.22).

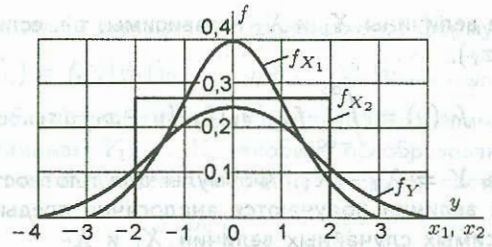


Рис. 2.1.4. Графики плотностей вероятности

Данный пример характерен для задачи суммирования систематической ошибки измерения какой-либо величины, распределение которой часто аппроксимируется равномерным распределением, и случайной нормально распределенной ошибки измерения той же величины. На графике (рис. 2.1.4) приведены распределения исходных случайных величин и суммарной случайной величины для $a = -2$, $b = 2$, $m = 0$, $\sigma = 1$.

Если случайные величины X и Y являются дискретными величинами, то в формулах (2.1.26)–(2.1.33) вместо интегралов надо понимать суммы с соответствующими пределами суммирования, зависящими от возможных значений величин X и Y.

Пример 2.1.5. Найдем закон распределения суммы $Y = X_1 + X_2$ двух независимых случайных величин X_1 и X_2, имеющих распределения Пуассона,

$$\mathbf{P}\{X_1 = n\} = p(n) = \frac{\lambda_1^n}{n!}\exp(-\lambda_1);$$

$$\mathbf{P}\{X_2 = k\} = p(k) = \frac{\lambda_2^k}{k!}\exp(-\lambda_2); \quad n, k = 0, 1, \ldots.$$

Заменяя интеграл в (2.1.27) суммой, имеем

$$p(n+k=m) = \sum_{n=0}^{\infty} \mathbf{P}\{X_1 = n\}\mathbf{P}\{X_2 = m-n\} =$$

$$= \sum_{n=0}^{\infty} \frac{\lambda_1^n}{n!}\exp(-\lambda_1)\mathbf{P}\{X_2 = m-n\}.$$

Так как $\mathbf{P}\{X_2 = m-n\} = 0$ для $n > m$, то верхний предел суммы равен m. Поэтому

$$p(m) = \sum_{n=0}^{m} \frac{\lambda_1^n}{n!}\mathrm{e}^{-\lambda_1}\frac{\lambda_2^{m-n}}{(m-n)!}\mathrm{e}^{-\lambda_2} = \frac{1}{m!}\mathrm{e}^{-(\lambda_1+\lambda_2)}\sum_{n=0}^{m}\frac{m!}{n!(m-n)!}\lambda_1^n\lambda_2^{m-n}.$$

Учитывая, что $C_m^n = \dfrac{m!}{n!(m-n)!}$ – биномиальные коэффициенты, а в соответствии с формулой бинома Ньютона $\sum\limits_{n=0}^{m} C_m^n\lambda_1^n\lambda_2^{m-n} = (\lambda_1+\lambda_2)^m$, получаем

$$\mathbf{P}\{Y = X_1 + X_2 = m\} = p(m = n+k) = \frac{(\lambda_1+\lambda_2)^m}{m!}\mathrm{e}^{-(\lambda_1+\lambda_2)}.$$

Из этой формулы видно, что сумма независимых случайных пуассоновских величин с параметрами λ_1 и λ_2 также является пуассоновской случайной величиной с параметром $\lambda = \lambda_1 + \lambda_2$.

Моменты преобразованных случайных величин $Y_1, \ldots, Y_n$ можно найти без предварительного определения их совместной плотности вероятности. Например, *начальный смешанный момент порядка* $k_1 + k_2 + \ldots + k_n$

$$\mathbf{M}\{Y_1^{k_1}\cdots Y_n^{k_n}\} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\varphi_1^{k_1}(x_1,\ldots,x_n)\cdots\varphi_n^{k_n}(x_1,\ldots,x_n)\times$$
$$\times f_{nX}(x_1,\ldots,x_n)\,dx_1\cdots dx_n. \qquad (2.1.34)$$

Одним из широко распространенных преобразований многомерных случайных величин является линейная сумма n случайных величин (1.2.22). Среднее значение суммы $\mathbf{M}\{X\}$ равно сумме средних значений слагаемых $\mathbf{M}\{X_i\} = m_{X_i}$:

$$m_X = \sum_{i=1}^{n} m_{X_i}. \qquad (2.1.35)$$

Вычислим дисперсию суммы (1.2.22), полагая слагаемые X_i независимыми случайными величинами с дисперсиями D_{X_i}. В общем виде дисперсия суммы случайных величин определяется выражением

$$D_X = \mathbf{M}\left\{\sum_{i=1}^{n}(X_i - m_{X_i})^2\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{M}\{(X_i - m_{X_i})(X_j - m_{X_j})\}.$$

Для независимых случайных величин справедливо соотношение

$$\mathbf{M}\{(X_i - m_{X_i})(X_j - m_{X_j})\} = \begin{cases}\mathbf{M}\{(X_i - m_{X_i})^2\} = D_{X_i}, & i = j;\\ 0, & i \neq j.\end{cases} \qquad (2.1.36)$$

Поэтому дисперсия суммы независимых случайных величин равна сумме дисперсий слагаемых:

$$D_X = \sum_{i=1}^{n} D_{X_i}. \qquad (2.1.37)$$

2.2. Распределения, порожденные гауссовскими случайными величинами

В математической статистике часто встречаются распределения хи-квадрат, Стьюдента, дисперсионного отношения и другие, порожденные гауссовскими случайными величинами в результате соответствующих преобразований последних.

2.2.1. Хи-квадрат (χ^2) распределение

Пусть каждая из независимых гауссовских случайных величин $X_1, \ldots, X_\nu$ имеет нулевое среднее значение и одинаковую единичную

дисперсию. Найдем плотность вероятности случайной величины Y, представляющей собой сумму квадратов упомянутых нормированных гауссовских случайных величин:

$$Y = \sum_{i=1}^{\nu} X_i^2. \tag{2.2.1}$$

Для этого рассмотрим характеристическую функцию $\theta(\vartheta)$ случайной величины Y, которая в силу независимости одинаково распределенных слагаемых в общем виде определяется формулой (1.2.24). В свою очередь характеристическая функция квадрата нормированной гауссовской случайной величины в соответствии с выражением (2.1.13), в котором необходимо положить $a = 1$ и $\sigma^2 = 1$, имеет вид

$$\theta_X(\vartheta) = (1 - 2i\vartheta)^{-1/2}.$$

В результате характеристическая функция случайной величины Y определяется формулой

$$\theta_Y(\vartheta) = (1 - 2i\vartheta)^{-\nu/2}. \tag{2.2.2}$$

При этом плотность вероятности случайной величины Y согласно (1.2.18) определяется из выражения

$$f_Y(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(1 - 2i\vartheta)^{-\nu/2}e^{-i\vartheta y}d\vartheta.$$

Вместо непосредственного вычисления интеграла воспользуемся полученным выше выражением (1.4.21) для характеристической функции гамма-распределения. Из него следует, что если характеристическая функция случайной величины описывается формулой (2.2.2), то соответствующая ей плотность вероятности определяется формулой (1.4.13), в которой необходимо положить $\beta = 2$, $\alpha+1 = \nu/2$. В результате получим

$$f_Y(y) = \frac{y^{(\nu/2)-1}}{2^{\nu/2}\Gamma(\nu/2)}e^{-y/2}, \quad y \geqslant 0. \tag{2.2.3}$$

Плотность вероятности (2.2.3) называется *плотностью вероятности хи-квадрат* (χ^2) *распределения* и часто обозначается как $f(\chi^2)$.

Функция хи-квадрат распределения определяется выражением

$$F_Y(y) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\int_0^y z^{(\nu/2)-1}e^{-z/2}\,dz. \tag{2.2.4}$$

Интеграл в (2.2.4) в элементарных функциях не вычисляется. В [8, 43, 51, 72 и др.] имеются таблицы функции распределения хи-квадрат для различных значений параметра ν. Для ряда вероятностей β там же даны таблицы критических значений $y_{\beta;\nu} \equiv \chi^2_{\beta;\nu}$, определяемых соотношением $\mathbf{P}\{Y \equiv \chi^2_\nu > \chi^2_{\beta;\nu}\} = \beta$; сокращенная часть этой таблицы приведена в табл. 2 Приложения 2.

Из формул плотности вероятности и функции распределения хи-квадрат распределения видно, что они определяются одним параметром ν, который называется *числом степеней свободы*. Так как распределение хи-квадрат является частным случаем более общего гамма-распределения (1.4.13) для $\alpha + 1 = \nu/2$ и $\beta = 2$, то из соотношений (1.4.22)–(1.4.24) получаем формулы для среднего значения, дисперсии, третьего и четвертого центральных моментов, коэффициентов асимметрии и эксцесса:

$$m = \nu; \quad D = 2\nu; \quad \mu_3 = 8\nu; \quad \mu_4 = 12\nu(\nu+4);$$
$$\gamma_1 = 4/\sqrt{2\nu}; \quad \gamma_2 = 12/\nu. \tag{2.2.5}$$

При $\nu \leqslant 2$ плотность вероятности хи-квадрат постоянно убывает, а при $\nu > 2$ имеется один максимум (мода), соответствующий точке $y = \nu - 2$.

На рис. 2.2.1 приведены графики плотности вероятности хи-квадрат распределения для ряда значений степеней свободы.

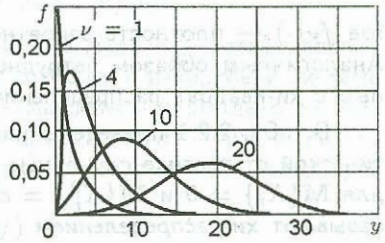


Рис. 2.2.1. Графики плотности вероятности хи-квадрат распределений

Аналогично биномиальному распределению, распределению Пуассона, гамма-распределению и нормальному распределению хи-квадрат распределение воспроизводит себя при композиции. Иначе говоря, если случайные величины $Y_1, \ldots, Y_q$ независимы и описываются хи-квадрат распределениями со степенями свободы $\nu_1, \ldots, \nu_q$, то случайная величина $Y = \sum_{j=1}^{q} Y_j$ распределена также по закону хи-квадрат с суммарной степенью свободы $\nu = \sum_{j=1}^{q} \nu_j$. Действительно,

$$\sum_{j=1}^{\nu_1} Y_j + \sum_{j=1}^{\nu_2} Y_j + \ldots + \sum_{j=1}^{\nu_q} Y_j = \sum_{j=1}^{\nu} Y_j = Y.$$

Полученное выражение для плотности вероятности хи-квадрат распределения (2.2.3) справедливо для суммирования квадратов независимых гауссовских случайных величин с единичной дисперсией. Путем несложного преобразования переменных можно получить плотность вероятности суммы квадратов независимых гауссовских случайных величин с произвольным значением дисперсии σ^2, для чего необходимо осуществить преобразование вида $z = \varphi(y) = \sigma^2 y$. При этом обратное преобразование $y = \psi(z) = z/\sigma^2$. Тогда согласно (2.1.2) плотность вероятности случайной величины $z \geqslant 0$ будет иметь вид

$$f(z) = f_Y(z/\sigma^2)\frac{1}{\sigma^2} = \frac{z^{(\nu/2)-1}\exp(-z/2\sigma^2)}{2^{\nu/2}\sigma^\nu\Gamma(\nu/2)}. \tag{2.2.6}$$

Таблица 2.2.1

Величина Z	Плотность вероятности $f(z)$	
$Z=\frac{1}{\nu}\sum_{i=1}^{\nu}X_i^2$	$f(z)=\frac{(\nu/2)^{\nu/2}z^{\nu/2-1}\exp(-\nu z/2\sigma^2)}{\sigma^{\nu}\Gamma(\nu/2)}$	(2.2.6а)
$Z=\sqrt{\sum_{i=1}^{\nu}X_i^2}$	$f(z)=\frac{2z^{\nu-1}\exp(-z^2/2\sigma^2)}{2^{\nu/2}\sigma^{\nu}\Gamma(\nu/2)}$	(2.2.6б)
$Z=\sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}X_i^2}$	$f(z)=\frac{2(\nu/2)^{\nu/2}z^{\nu-1}\exp(-\nu z^2/2\sigma^2)}{\sigma^{\nu}\Gamma(\nu/2)}$	(2.2.6в)

где $f_Y(\cdot)$ — плотность вероятности хи-квадрат распределения (2.2.3). Аналогичным образом нетрудно найти другие распределения, связанные с хи-квадрат распределением.

В табл. 2.2.1 приведены наиболее часто употребляемые в математической статистике случайные величины и их плотности вероятности для $\mathrm{M}\{X_i\}=0$ и $\mathrm{M}\{X_i^2\}=\sigma^2$. Распределение (2.2.6б) при $\sigma^2=1$ называют *хи-распределением* (χ-распределением).

Из распределений, приведенных в таблице, можно получить распределения плотности вероятности, широко используемые в теории вероятности и математической статистике. Так, если $Z=\sqrt{X_1^2+X_2^2}$, где X_1 и X_2 — независимые гауссовские случайные величины с нулевыми средними значениями и одинаковыми дисперсиями σ^2, то из распределения случайной величины (2.2.6б) получаем *распределение Релея*

$$f(z)=\frac{z}{\sigma^2}\exp\left(-\frac{z^2}{2\sigma^2}\right),\quad z\geqslant 0. \tag{2.2.7}$$

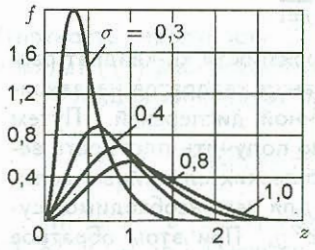


Рис. 2.2.2. Графики релеевской плотности вероятности

Формально распределение Релея описывает распределение длины случайного вектора, координаты которого в двумерной декартовой системе координат независимы и нормально распределены с нулевым средним значением и дисперсией σ^2. Физически плотность вероятности Релея описывает большой класс случайных величин, встречающихся в физике и других науках. В частности, закон Релея описывает распределение огибающей узкополосного гауссовского шума.

На рис. 2.2.2 приведены графики релеевской плотности вероятности.

Среднее значение, дисперсия и одномерные четные начальные моменты релеевской случайной величины равны

$$m_Z=\sigma\sqrt{\pi/2};\quad D(z)=(2-\pi/2)\sigma^2;\quad \mathrm{M}\{z^{2k}\}=\frac{2^{2k}(k!)^2(2k-1)!!}{(2k)!}\sigma^{2k},$$

где $(2k-1)!!=1\cdot 3\cdot 5\cdots(2k-1)$.

При $\nu=3$ из (2.2.6б) получаем *распределение Максвелла*

$$f(z)=\sqrt{\frac{2}{\pi}}\frac{z^2}{\sigma^3}\exp(-z^2/2\sigma^2),\quad 0\leqslant z<\infty.$$

Распределение Максвелла имеет конечные моменты любого порядка. В частности, среднее значение и дисперсия равны соответственно

$$m_Z=2\sqrt{2/\pi}\sigma;\quad D(z)=[(3\pi-8)/\pi]\sigma^2.$$

Формально распределение Максвелла определяет распределение длины случайного вектора, координаты которого в декартовой системе координат в трехмерном пространстве независимы и нормально распределены с нулевым средним значением и дисперсией σ^2. Физически это распределение достаточно хорошо описывает распределение скоростей молекул газа.

2.2.2. Распределение Стьюдента (t-распределение)

Пусть имеются $(\nu+1)$ независимые гауссовские случайные величины $X, X_1, X_2, \ldots, X_\nu$ с нулевыми средними значениями и одинаковыми дисперсиями. Найдем плотность вероятности частного двух случайных величин

$$T=X/Z, \tag{2.2.9}$$

где

$$Z=\sqrt{\frac{1}{\nu}\sum_{i=1}^{\nu}X_i^2}. \tag{2.2.10}$$

Согласно (2.1.33) плотность вероятности частного двух случайных величин в рассматриваемых обозначениях находится из выражения

$$f(t)=\int_{-\infty}^{\infty}|z|f_Z(z)f_X(tz)\,dz. \tag{2.2.11}$$

Поскольку полагается, что дисперсии случайных величин X и X_i, $i=1,\ldots,\nu$, одинаковы, частное (2.2.9) не изменится, если принять $\sigma=1$. В этом случае $f_X(x)$ — нормированная нормальная плотность вероятности с нулевым средним значением и единичной дисперсией, а

$f_Z(z)$ определяется выражением (2.2.6в). Подставляя эти плотности вероятности в формулу (2.2.11), получим

$$f(t) = \frac{2(\nu/2)^{\nu/2}}{\Gamma(\nu/2)\sqrt{2\pi}} \int_0^\infty z^\nu \exp\left[-\frac{t^2+\nu}{2} z^2\right] dz.$$

Введя новую переменную $y = (t^2+\nu)z^2/2$ и учитывая интеграл (1.4.14), искомая плотность вероятности преобразуется к виду

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)} \left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}. \qquad (2.2.12)$$

Эта плотность вероятности называется *плотностью вероятности распределения Стьюдента* или *t-распределения*. Плотность вероятности Стьюдента, как и плотность вероятности хи-квадрат, зависит от одного параметра ν, который также называется числом степеней свободы. Функция распределения Стьюдента

$$F(t) = \mathbf{P}\{T < t\} = \int_{-\infty}^{t} f(z)\, dz \qquad (2.2.13)$$

в явном виде не вычисляется; она табулирована [8, 51, 72 и др.]. В математической статистике чаще используются таблицы критических значений $t_{\beta;\nu}$, соответствующих уравнению

$$\mathbf{P}\{T > t_{\beta;\nu}\} = 1 - F(t_{\beta;\nu}) = \int_{t_{\beta;\nu}}^{\infty} f(t)\, dt = \beta. \qquad (2.2.14)$$

Часть этой таблицы для ряда значений параметров β и ν в сокращенном виде приведена в табл. 3 Приложения 2.

Характеристическая функция распределения Стьюдента так же, как и функция распределения для произвольного числа степеней свободы, в явном виде не записывается.

Из определения случайной величины T (2.2.9) и формулы (2.2.12) видно, что она изменяется от $-\infty$ до $+\infty$ и симметрична относительно $t = 0$. Следовательно, мода и медиана распределения равны нулю.

Начальный момент произвольного порядка определяется соотношением

$$m_k = \int_{-\infty}^{\infty} t^k f(t)\, dt = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \int_{-\infty}^{\infty} \frac{t^k dt}{(1+t^2/\nu)^{(\nu+1)/2}}.$$

Если $k < \nu$, интеграл в этом выражении является сходящимся в смысле существования предела

$$\lim_{\substack{A\to-\infty \\ B\to+\infty}} \int_A^B \varphi(y)\, dy.$$

При этом в последнем выражении A и B должны стремиться к своим пределам произвольным образом, а не в смысле главного значения интеграла, при котором $A = -B$ $(B \to \infty)$. Вследствие симметричности распределения все существующие моменты нечетного порядка $(k < \nu)$ равны нулю; в частности, среднее значение равно нулю при $\nu > 1$. Используя таблицы интегралов [21], для центральных моментов можно получить соотношение вида [37, 59 и др.]

$$\mu_{2k} = \frac{1\cdot 3\cdots(2k-1)\nu^k}{(\nu-2)(\nu-4)\cdots(\nu-2k)}, \quad 2k < \nu, \quad k = 1, 2, \ldots \qquad (2.2.15)$$

Отсюда находим формулы для дисперсии и четвертого центрального момента

$$\sigma^2 = \nu/(\nu-2) \text{ при } \nu > 2; \qquad \mu_4 = 3\nu^2/[(\nu-2)(\nu-4)] \text{ при } \nu > 4. \qquad (2.2.16)$$

При этом коэффициент эксцесса распределения Стьюдента

$$\gamma_2 = 6/(\nu-4), \quad \nu > 4. \qquad (2.2.17)$$

Из последнего выражения видно, что при $\nu \gg 1$ коэффициент эксцесса γ_2 стремится к нулю и, следовательно, распределение Стьюдента асимптотически переходит в нормальное распределение с нулевым средним значением и единичной дисперсией.

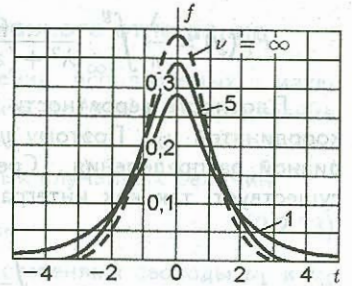


Рис. 2.2.3. Плотности вероятности распределения Стьюдента и нормальной плотности вероятности

На рис. 2.2.3 приведены графики плотности вероятности Стьюдента; там же для сравнения приведен график нормальной плотности вероятности с единичной дисперсией (пунктирная линия).

2.2.3. Распределение Коши

При $\nu = 1$ случайная величина Y (2.2.9) представляет собой отношение двух независимых случайных величин с нулевыми средними значениями и одинаковыми дисперсиями. Плотность вероятности этой случайной величины представляет собой частный случай плотности вероятности Стьюдента (2.2.12) при $\nu = 1$:

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty. \qquad (2.2.18)$$

Это распределение называется *распределением Коши*. В более общем виде распределение Коши записывается как

$$f_Y(y) = \frac{\lambda}{\pi[\lambda^2 + (y-y_0)^2]}, \quad -\infty < y < \infty. \qquad (2.2.19)$$

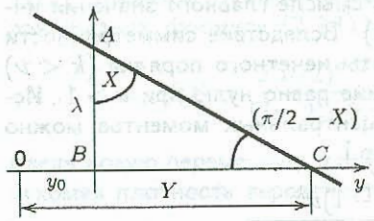


Рис. 2.2.4. Геометрическая модель случайной величины с распределением Коши

Распределение (2.2.19) получается из (2.2.18) с помощью линейного преобразования $y = \varphi(x) = y_0 + \lambda x$. Геометрическая модель случайной величины Y, распределенной по закону Коши, представлена на рис. 2.2.4. Пусть через точку A, находящуюся на расстоянии λ от фиксированной прямой $0y$, проведена произвольная прямая, пересекающая ось $0y$ в точке C под углом $(\pi/2 - X)$. Угол X является случайным, равномерно распределенным $[f_X(x) = 1/\pi]$ по интервалу $(-\pi/2, \pi/2)$. Величина $Y = 0C = y_0 + \lambda \operatorname{tg} X$ также является случайной. Используя соотношение для преобразования плотности вероятности (2.1.2), в котором $x = \psi(y) = \operatorname{arctg}[(y - y_0)/\lambda]$, получаем (2.2.19).

Функция распределения Коши определяется формулой

$$F_Y(y) = \frac{\lambda}{\pi}\int_{-\infty}^{y}\frac{dz}{\lambda^2 + (z - y_0)^2} = \frac{1}{2} + \frac{1}{\pi}\operatorname{arctg}\frac{y - y_0}{\lambda}. \qquad (2.2.20)$$

Плотность вероятности Коши симметрична относительно точки с координатой y_0. Поэтому y_0 является одновременно модой y_m и медианой распределения. Среднее значение случайной величины Y не существует, так как интеграл

$$\int_{-\infty}^{\infty}\frac{y\,dy}{\lambda^2 + (y - y_0)^2}$$

расходится. По этой же причине для распределения Коши не существуют и другие начальные и центральные моменты, в том числе не существует и дисперсии распределения Коши. Отсутствует также конечный предел среднего арифметического значения. Докажем это утверждение применительно к случайной величине X с плотностью вероятности (2.2.18), для чего рассмотрим предельное выборочное среднее для независимых элементов выборки

$$\bar{x} = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} x_k. \qquad (2.2.21)$$

Выразим характеристическую функцию выборочного среднего через характеристические функции слагаемых, учитывая соотношения (1.2.24) и (1.2.17). Характеристическая функция слагаемых суммы (2.2.21) определяется выражением

$$\theta_{X_k}(\vartheta) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\exp(i\vartheta x_k)}{1 + x_k^2}dx_k = \frac{2}{\pi}\int_{0}^{\infty}\frac{\cos\vartheta x_k}{1 + x_k^2}dx_k.$$

Используя соответствующие таблицы интегралов [21] и учитывая четность характеристической функции, имеем

$$\theta_{X_k}(\vartheta) = \exp(-|\vartheta|). \qquad (2.2.22)$$

Тогда характеристическая функция выборочного среднего случайной величины, распределенной по закону Коши,

$$\theta_X(\vartheta) = [\exp(-|\vartheta/n|)]^n = \exp(-|\vartheta|),$$

т.е. характеристическая функция среднего арифметического не зависит от числа слагаемых и совпадает с характеристической функцией слагаемого. Поэтому плотность вероятности арифметического среднего значения при беспредельном увеличении числа слагаемых совпадает с распределением Коши для одного слагаемого. Иначе говоря, арифметическое среднее значение при бесконечном числе слагаемых не стремится ни к какому пределу.

2.2.4. Распределение дисперсионного отношения

Одним из наиболее важных распределений, используемых в математической статистике и теории эксперимента, является распределение дисперсионного отношения.

Рассмотрим отношение двух независимых случайных величин

$$Y = Y_1/Y_2, \qquad (2.2.23)$$

распределенных по законам хи-квадрат со степенями свободы ν_1 и ν_2 соответственно. Плотность вероятности случайной величины Y определяется из соотношения (2.2.11), в которое вместо $f_Z(z)$ и $f_X(yz)$ необходимо подставить плотности вероятности случайных величин Y_1 и Y_2, определяемых формулой (2.2.3), со степенями свободы ν_1 и ν_2. После подстановок указанных плотностей вероятностей имеем

$$f(y) = \frac{2^{-(\nu_1+\nu_2)/2}y^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\int_0^{\infty} z^{(\nu_1+\nu_2)/2-1}\exp[-(1+y)z/2]\,dz.$$

Осуществляя замену переменной $t = (1 + y)z/2$ и учитывая интеграл (1.4.14), получим

$$f(y) = \frac{\Gamma[(\nu_1+\nu_2)/2]y^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(y+1)^{(\nu_1+\nu_2)/2}}, \quad y \geqslant 0. \qquad (2.2.24)$$

В дисперсионном анализе математической статистики рассматривается случайная величина

$$Z = \frac{Y_1}{\nu_1} : \frac{Y_2}{\nu_2} = \frac{\nu_2}{\nu_1}\cdot\frac{Y_1}{Y_2} \equiv F(\nu_1, \nu_2), \qquad (2.2.25)$$

плотность вероятности которой получается из формулы (2.2.24) путем

линейного преобразования случайных величин. В результате имеем

$$f(z)=\frac{\Gamma[(\nu_1+\nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\frac{z^{\nu_1/2-1}}{(\nu_2+\nu_1 z)^{(\nu_1+\nu_2)/2}},\quad z\geqslant 0. \quad (2.2.26)$$

Случайная величина Z (2.2.25) связана с отношением двух выборочных дисперсий (подробное рассмотрение этих величин дается ниже в п. 3.4)

$$s_1^2=\frac{1}{\nu_1}\sum_{i=1}^{\nu_1}y_{1i}^2;\quad s_2^2=\frac{1}{\nu_2}\sum_{i=1}^{\nu_2}y_{2i}^2, \quad (2.2.27)$$

определяемых по двум независимым выборкам y_1 $(y_{11},y_{12},\ldots)$ и y_2 $(y_{21},y_{22},\ldots)$ из гауссовской случайной величины с нулевым средним значением и дисперсией σ^2. В данной ситуации отношение s_1^2/s_2^2 не зависит от конкретного значения дисперсии и поэтому оно эквивалентно случайной величине Z (2.2.25). В связи с этим распределение (2.2.26) называется *распределением дисперсионного отношения*, а также *F-распределением*, *распределением Фишера* и *распределением Снедекора*.

Функция F-распределения с параметрами ν_1 и ν_2 в явном виде не выражается. Однако в [8, 43, 51, 72 и др.] приведены таблицы критических значений $z_\beta\equiv F_{\beta;\nu_1,\nu_2}$ F-распределения, определяемых из уравнения

$$\mathbf{P}\{F_{\nu_1,\nu_2}>F_{\beta;\nu_1,\nu_2}\}=\int_{F_{\beta;\nu_1,\nu_2}}^{\infty}f(z)\,dz=\beta, \quad (2.2.28)$$

для ряда значений вероятности β и числа степеней свободы ν_1 и ν_2. Частично эти данные повторены в табл. 4 Приложения 2.

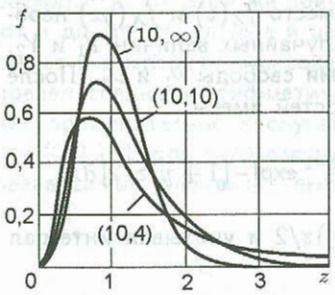


Рис. 2.2.5. Графики плотности вероятности F-распределения

F-распределение определяется двумя параметрами: числом степеней свободы числителя ν_1 и числом степеней свободы знаменателя ν_2 (2.2.25).

На рис. 2.2.5 приведены графики плотности вероятности F-распределения для трех значений параметров ν_1 и ν_2.

Мода распределения

$$z_m=\frac{\nu_1-2}{\nu_1}\frac{\nu_2}{\nu_2+2},\quad \nu_1>2. \quad (2.2.29)$$

Начальный момент k-го порядка находится из выражения

$$m_k=\frac{\Gamma[(\nu_1+\nu_2)/2]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\int_0^\infty z^{(\nu_1/2)+k-1}\frac{dz}{[1+(\nu_1/\nu_2)z]^{(\nu_1+\nu_2)/2}}.$$

Учитывая табличный интеграл [21]

$$\int_0^\infty\frac{x^{p-1}dx}{(1+qx)^{p+r}}=\frac{\Gamma(p)\Gamma(r)}{q^r\Gamma(p+r)},\quad p,q,r>0,$$

формула для начального момента k-го порядка примет вид

$$m_k=\frac{\Gamma[(\nu_1/2)+k]\Gamma[(\nu_2/2)-k]}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\left(\frac{\nu_2}{\nu_1}\right)^k,\quad \nu_2>2k. \quad (2.2.30)$$

Отсюда получаем выражения для среднего значения и дисперсии F-распределения:

$$m=\frac{\nu_2}{\nu_2-2},\quad \nu_2>2, \quad (2.2.31)$$

$$D=\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)},\quad \nu_2>4. \quad (2.2.32)$$

Аналогичным образом, учитывая связь начальных и центральных моментов (1.2.34), могут быть вычислены третий μ_3 и четвертый μ_4 центральные моменты, а следовательно, и коэффициенты асимметрии γ_1 и эксцесса γ_2.

2.3. Аппроксимация законов распределения

В задачах теории вероятностей и математической статистики часто не удается найти точные распределения, позволяющие вычислять соответствующие вероятности величин (событий). Кроме того, в некоторых случаях хотя и могут быть записаны точные формулы, но они оказываются настолько сложными (например, в виде сумм, произведений или интегралов), что вычисление соответствующих вероятностей также представляет достаточно сложную и громоздкую вычислительную процедуру. Поэтому на практике в подобных ситуациях осуществляется *аппроксимация законов распределения* (приближение известного или неизвестного точного распределения некоторой функцией). Наиболее широко используемым приемом является «подгонка» точного распределения простым распределением. В принципе любая неотрицательная функция, удовлетворяющая условиям нормировки и выбранному критерию приближения, может рассматриваться в качестве соответствующей аппроксимирующей кривой. В дальнейшем полагается, что анализируемая случайная величина непрерывна, а ее плотность вероятности — унимодальна.

При выборе аппроксимирующего распределения обычно руководствуются информацией об области возможных значений случайной величины (области определения) и его предельной формой. В этих условиях выбираются «гладкие» распределения, которые хорошо изучены и для которых имеются легко доступные таблицы вероятностей. При этом

часто из чисто качественных соображений приближенно можно указать класс или классы возможных распределений с точностью до значений их параметров (числовых характеристик). Кроме того, определенные сведения о законах распределения вероятностей несут его моменты, причем, как указывалось выше (п. 1.2.2), наибольшую информацию об одномерных законах распределения несут первые четыре момента: среднее значение, дисперсия, третий и четвертый центральные моменты. Так как найти моменты распределений значительно проще, чем определить сами законы распределения, возникает задача: используя информацию о моментах, аппроксимировать точное распределение некоторой функцией, обеспечивающей заданную точность приближения.

Принципиально существует несколько подходов к решению вопроса о рациональном способе аппроксимации законов распределения. Однако в прикладных задачах наибольшее применение получили два подхода. Первый из них основан на представлении точного распределения в виде ряда по ортогональным полиномам с использованием рационально выбранной весовой функции. При втором подходе точное распределение приближается кривыми (распределениями) Пирсона, использующими первые четыре момента точного распределения.

2.3.1. Аппроксимация одномерной плотности вероятности рядом по ортогональным полиномам

Система функций $Q_i(z)$ называется *ортогональной* относительно «весовой» функции $\varphi(z)$, определенной на интервале (α, β), если выполняется условие

$$\int_\alpha^\beta \varphi(z) Q_m(z) Q_n(z)\, dz = \begin{cases} a_n & \text{при } m = n, \\ 0 & \text{при } m \neq n. \end{cases} \qquad (2.3.1)$$

Величину a_n называют нормой функции $Q_n(z)$ относительно веса $\varphi(z)$. Если $a_n = 1$, то система функций $Q_n(z)$ называется *ортонормированной*.

Формально произвольную «гладкую» одномерную плотность вероятности $f(z)$ можно представить в виде ряда по ортогональным функциям:

$$f(z) = \varphi(z) \sum_{k=-\infty}^{\infty} c_k Q_k(z). \qquad (2.3.2)$$

Здесь коэффициенты c_k определяются из выражения, получающегося в результате умножения обеих частей (2.3.2) на $Q_n(z)$ и интегрирования в пределах (α, β) с использованием условия ортогональности (2.3.1). При этом все члены суммы (2.3.2) за исключением одного, соответствующего $k = n$, обращаются в нуль. В результате имеем

$$c_n = \frac{1}{a_n} \int_\alpha^\beta f(z) Q_n(z)\, dz. \qquad (2.3.3)$$

Если в качестве ортогональных функций выступают полиномы

$$Q_n(z) = \sum_{i=0}^{n} d_i z^i, \qquad (2.3.4)$$

где d_i — коэффициенты полинома, то такие полиномы называются *ортогональными полиномами*. В дальнейшем при аппроксимации законов распределения в качестве ортогональных функций рассматриваются ортогональные полиномы.

Применительно к ортогональным полиномам, подставляя (2.3.4) в (2.3.3), получим связь коэффициентов ряда (2.3.2) c_n с начальными моментами m_i аппроксимируемой плотности вероятности:

$$c_n = \frac{1}{a_n} \sum_{i=0}^{n} d_i \int_\alpha^\beta z^i f(z)\, dz = \frac{1}{a_n} \sum_{i=0}^{n} d_i m_i. \qquad (2.3.5)$$

В качестве весовой функции $\varphi(z)$, определяющей систему ортогональных полиномов, выбирают простейшую и хорошо изученную одномерную плотность вероятности, для которой имеются доступные таблицы распределения вероятностей. Для случайной величины, изменяющейся в пределах $(-\infty, \infty)$, в качестве весовой функции обычно выбирают нормальное распределение (1.4.1). Если область возможных значений случайной величины ограничена интервалом $(0, \infty)$, то предпочтительной весовой функцией является гамма-распределение (1.4.13). Параметры этих распределений, как правило, связывают с моментами аппроксимируемого распределения.

2.3.1.1. Аппроксимация распределений с использованием ортогональных полиномов Эрмита

Аппроксимируем одномерную плотность вероятности $f_X(x)$ случайной величины X с помощью суммы одномерных ортогональных полиномов Эрмита. Для этого рассмотрим нормированную случайную величину $Z = (X - m_X)/\sigma_X$, где m_X и σ_X — среднее значение и среднее квадратическое отклонение случайной величины X. Плотность вероятности нормированной случайной величины Z является линейным преобразованием от $f_X(x)$ и равна

$$f_Z(z) = \sigma_X f_X(\sigma_X z + m_X). \qquad (2.3.6)$$

Одномерные полиномы Эрмита n-го порядка определяются соотношением

$$H_n(z) = (-1)^n \mathrm{e}^{z^2/2} \frac{d^n}{dz^n} \mathrm{e}^{-z^2/2}, \quad n = 0, 1, 2, \ldots \qquad (2.3.7)$$

При n четном полином будет содержать члены переменной z только в четных степенях и, наоборот, при n нечетном полином будет иметь члены с нечетными степенями переменной.

Полиномы Эрмита удовлетворяют условию ортогональности (2.3.1):

$$\int_{-\infty}^{\infty} H_m(z)H_n(z)\varphi(z)\,dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} H_m(z)H_n(z)\mathrm{e}^{-z^2/2}dz = \begin{cases} n!, & m=n, \\ 0, & m\neq n. \end{cases} \qquad (2.3.8)$$

Здесь $\varphi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$ — нормированная нормальная плотность вероятности, фигурирующая в качестве весовой функции в (2.3.1) и (2.3.2). Результат (2.3.8) получается повторным интегрированием по частям.

Для полиномов Эрмита справедливо рекуррентное соотношение

$$H_{n+1}(z) = zH_n(z) - nH_{n-1}(z). \qquad (2.3.9)$$

Из (2.3.7) и последнего соотношения могут быть получены значения необходимых полиномов Эрмита. В частности, первые пять из них равны

$$H_0(z)=1; \quad H_1(z)=z; \quad H_2(z)=z^2-1;$$
$$H_3(z)=z^3-3z; \quad H_4(z)=z^4-6z^2+3. \qquad (2.3.10)$$

В соответствии с (2.3.2), (2.3.3) и (2.3.8) аппроксимируемую нормированную плотность вероятности можно представить в виде

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\sum_{n=0}^{\infty}\frac{b_n}{n!}H_n(z), \qquad (2.3.11)$$

где коэффициенты ряда с учетом (2.3.3) и (2.3.8) определяются как

$$b_n = \int_{-\infty}^{\infty} f_Z(z)H_n(z)\,dz. \qquad (2.3.12)$$

Перейдем от нормированной случайной величины Z к ненормированной случайной величине $X = \sigma_X Z + m_X$. В результате получим

$$f_X(x) = \varphi(x)\sum_{n=0}^{\infty}\frac{b_n}{n!}H_n\left(\frac{x-m_X}{\sigma_X}\right), \qquad (2.3.13)$$

где

$$b_n = \int_{-\infty}^{\infty} f_X(x)H_n\left(\frac{x-m_X}{\sigma_X}\right)dx; \qquad (2.3.14)$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left[-\frac{(x-m_X)^2}{2\sigma_X^2}\right]. \qquad (2.3.15)$$

Если исходная случайная величина является гауссовской, то для любого значения $n \geqslant 1$ справедливо тождество (для упрощения доказательства полагается, что $m_X = 0$, $\sigma_X^2 = 1$)

$$b_n = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2}(-1)^n\mathrm{e}^{x^2/2}\frac{d^n}{dx^n}(\mathrm{e}^{-x^2/2})\,dx =$$
$$= \frac{(-1)^n}{\sqrt{2\pi}}\frac{d^{n-1}}{dx^{n-1}}\mathrm{e}^{-x^2/2}\Big|_{-\infty}^{\infty} = 0.$$

В правой части последнего выражения учтено, что при $x \to \pm\infty$ функция $\exp(-x^2/2)$ и ее производные обращаются в нуль. Таким образом, для гауссовской случайной величины все коэффициенты ряда (2.3.13) при $n \geqslant 1$ равны нулю.

Из формулы (2.3.14) видно, что коэффициенты b_n ряда (2.3.13) связаны с моментами случайной величины X. При этом первые пять коэффициентов равны

$$b_0 = \int_{-\infty}^{\infty} f_X(x)\,dx = 1;$$
$$b_1 = \int_{-\infty}^{\infty}\frac{x-m_X}{\sigma_X}f_X(x)\,dx = 0;$$
$$b_2 = \int_{-\infty}^{\infty}\left[\frac{(x-m_X)^2}{\sigma_X^2}-1\right]f_X(x)\,dx = 0;$$
$$b_3 = \int_{-\infty}^{\infty}\left[\frac{(x-m_X)^3}{\sigma_X^3}-3\frac{x-m_X}{\sigma_X}\right]f_X(x)\,dx = \frac{\mu_3}{\sigma_X^3} = \gamma_1;$$
$$b_4 = \int_{-\infty}^{\infty}\left[\frac{(x-m_X)^4}{\sigma_X^4}-6\frac{(x-m_X)^2}{\sigma_X^2}+3\right]f_X(x)\,dx = \frac{\mu_4}{\sigma_X^4}-3 = \gamma_2.$$

где γ_1 и γ_2 — коэффициенты асимметрии и эксцесса соответственно.

Практически плотность вероятности необходимо знать с некоторой конечной точностью. Если в формуле (2.3.13) ограничиться r членами ряда, получим выражение, которое называют *рядом Эджворта*:

$$f_X(x) \approx \varphi(x)\left[1+\sum_{n=3}^{r}\frac{b_n}{n!}H_n\left(\frac{x-m_X}{\sigma_X}\right)\right]. \qquad (2.3.16)$$

Естественно, число слагаемых ряда зависит от требуемой точности аппроксимации.

Первые два коэффициента ряда b_3 и b_4 характеризуют наиболее существенные отклонения рассматриваемой плотности вероятности $f_X(x)$ от нормальной плотности вероятности $\varphi(x)$, для которой, как отмечалось выше, коэффициенты асимметрии и эксцесса равны нулю.

На практике при аппроксимации плотностей вероятности, не очень сильно отличающихся от нормальной плотности вероятности, часто ограничиваются учетом только коэффициентов асимметрии и эксцесса.

В этом случае формула (2.3.13) записывается в виде

$$f_X(x) \approx \varphi(x)\left[1+\frac{\gamma_1}{6}H_3\left(\frac{x-m_X}{\sigma_X}\right)+\frac{\gamma_2}{24}H_4\left(\frac{x-m_X}{\sigma_X}\right)\right]. \quad (2.3.17)$$

Поэтому для применения этой аппроксимации нужно тем или иным способом вычислить среднее значение m_X, дисперсию σ_X^2, третий μ_3 и четвертый μ_4 центральные моменты изучаемой случайной величины X. Однако необходимо иметь в виду, что при аппроксимации плотности вероятности в виде (2.3.16) или (2.3.17) нарушается свойство положительной определенности для плотности вероятности, т.е. при больших значениях переменной $|x|$ аппроксимирующая кривая может принимать отрицательные значения, недопустимые для плотности вероятности.

Полиномы Эрмита просто выражаются через производные от нормальной плотности вероятности $\varphi(z)$ и функции ее распределения $\Phi(z)$, определяемой (1.3.22):

$$\varphi^{(n)}(z) = \Phi^{(n+1)}(z) = (-1)^n H_n(z)\varphi(z). \quad (2.3.18)$$

Функция нормального распределения, плотность вероятности нормального распределения и первые пять производных плотности нормального распределения табулированы [8, 51]. Поэтому, если в формуле (2.3.17) перейти от полиномов Эрмита к производным от плотности вероятности нормального распределения, то можно записать

$$f_X(x) \approx \frac{1}{\sigma_X}\left[\Phi^{(1)}\left(\frac{x-m_X}{\sigma_X}\right)-\frac{\gamma_1}{6}\Phi^{(4)}\left(\frac{x-m_X}{\sigma_X}\right)+\frac{\gamma_2}{24}\Phi^{(5)}\left(\frac{x-m_X}{\sigma_X}\right)\right]. \quad (2.3.19)$$

В соответствии с (2.3.17) приближенные формулы для биномиального (1.3.27) и пуассоновского распределения (1.3.36) учитывают только коэффициенты асимметрии.

2.3.1.2. Аппроксимация распределений с использованием ортогональных полиномов Лагерра

Если случайная величина X принимает значения в интервале $(0,\infty)$, то приемлемая аппроксимация плотности вероятности $f_X(x)$ рядом по ортогональным полиномам Эрмита потребует большого числа его членов. В этой ситуации предпочтительно в качестве ортогональных полиномов использовать *полиномы Лагерра*, определяемые соотношением

$$L_n^{(\alpha)}(z) = \frac{z^{-\alpha}\mathrm{e}^z}{n!}\frac{d^n}{dz^n}(z^{n+\alpha}\mathrm{e}^{-z}), \quad n=0,1,\ldots;\ \alpha>-1;\ z>0. \quad (2.3.20)$$

Полиномы Лагерра в промежутке $(0,\infty)$ ортогональны с весовой функцией $\varphi(z) = z^\alpha \mathrm{e}^{-z}/\Gamma(\alpha+1)$, являющейся плотностью вероятности гамма-распределения (1.4.13) для нормированной случайной величины

$Z = X/\beta$:

$$\int_0^\infty \varphi(z)L_m^{(\alpha)}(z)L_n^{(\alpha)}(z)\,dz = \frac{1}{\Gamma(\alpha+1)}\int_0^\infty L_m^{(\alpha)}(z)L_n^{(\alpha)}(z)z^\alpha \mathrm{e}^{-z}dz = \begin{cases} \frac{1}{n!}\frac{\Gamma(n+\alpha+1)}{\Gamma(\alpha+1)}, & m=n, \\ 0, & m\neq n. \end{cases} \quad (2.3.21)$$

Первые четыре полинома Лагерра равны

$$L_0^{(\alpha)}(z)=1; \quad L_1^{(\alpha)}(z)=1+\alpha-z;$$
$$L_2^{(\alpha)}(z)=\frac{1}{2}[(\alpha+1)(\alpha+2)-2z(\alpha+2)+z^2]; \quad (2.3.22)$$
$$L_3^{(\alpha)}(z)=\frac{1}{6}[(\alpha+1)(\alpha+2)(\alpha+3)-3z(\alpha+2)(\alpha+3)+3z^2(\alpha+3)-z^3].$$

Применительно к ненормированной случайной величине $X=\beta Z$ аппроксимируемая плотность вероятности принимает вид

$$f_X(x) = \frac{(x/\beta)^\alpha \mathrm{e}^{-x/\beta}}{\beta\Gamma(\alpha+1)}\sum_{n=0}^\infty c_n L_n^{(\alpha)}(x/\beta), \quad (2.3.23)$$

где коэффициенты ряда определяются формулой

$$c_n = \frac{n!\Gamma(\alpha+1)}{\Gamma(n+\alpha+1)}\int_0^\infty f_X(x)L_n^{(\alpha)}(x/\beta)\,dx. \quad (2.3.24)$$

При этом первые четыре коэффициента ряда (2.3.23), связанные с параметрами α, β и начальными моментами m, m_2, m_3 плотности вероятности гамма-распределения (1.4.13), выступающей в качестве весовой функции, определяются соотношениями

$$c_0=1; \quad c_1=\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+2)}[1+\alpha-m/\beta];$$
$$c_2=\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+3)}\left[(\alpha+1)(\alpha+2)-2\frac{m}{\beta}(\alpha+2)+\frac{m_2}{\beta^2}\right];$$
$$c_3=\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+4)}\left[(\alpha+1)(\alpha+2)(\alpha+3)-3\frac{m}{\beta}(\alpha+2)(\alpha+3)+3\frac{m_2}{\beta^2}(\alpha+3)-\frac{m_3}{\beta^3}\right].$$

Аналогично вычисляются и другие коэффициенты ряда, но их выражения достаточно громоздки для решения прикладных задач.

При переходе от нормированной переменной к ненормированной переменной весовая функция совпадает с плотностью гамма-распределения (1.4.13).

Параметры α и β плотности вероятности (2.3.23) достаточно произвольны. Выберем их из условия, чтобы среднее значение и дисперсия

этой функции совпадали со средним значением m и дисперсией D исходной случайной величины X, т.е. чтобы параметры α и β были связаны с m и D выражениями (1.4.25). Подставляя эти значения параметров в формулы для коэффициентов ряда, имеем

$$c_0 = 1; \quad c_1 = c_2 = 0; \quad c_3 = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+4)}\left[\frac{m_2}{\beta^2}(\alpha+3) - \frac{m_3}{\beta^3}\right]. \qquad (2.3.25)$$

В результате аппроксимируемая плотность вероятности приближенно представляется выражением

$$f_X(x) \approx \frac{(x/\beta)^\alpha}{\beta\Gamma(\alpha+1)} e^{-x/\beta}\left\{1 + \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+4)}\left[\frac{m_2}{\beta^2}(\alpha+3) - \frac{m_3}{\beta^3}\right] L_3^{(\alpha)}(x/\beta)\right\};$$
$$\alpha = (m^2/D) - 1, \quad \beta = D/m. \qquad (2.3.26)$$

Если случайная величина X имеет гамма-распределение (1.4.13), то ее начальные моменты определяются (1.4.22). Подставляя значения m_2 и m_3 в (2.3.25), получаем $c_3 = 0$, т.е. искомая плотность вероятности $f_X(x)$, как и следовало ожидать, совпадает с аппроксимирующей функцией $\varphi(x)$.

Аналогично сказанному выше в п. 2.3.1.1 при аппроксимации плотности вероятности в виде (2.3.26) нарушается свойство положительной определенности, т.е. при больших значениях переменной x аппроксимирующая кривая может принимать отрицательные значения, недопустимые для плотности вероятности.

2.3.2. Распределения (кривые) Пирсона

Второй способ аппроксимации непрерывного неизвестного закона распределения $f(x)$ основан на использовании большого класса распределений (кривых), предложенных и исследованных К. Пирсоном в 1894 г. *Распределениями (кривыми, графиками) Пирсона* называются непрерывные распределения, плотности вероятности $f(x)$ которых являются решением дифференциального уравнения [8, 15, 37, 46, 47, 59] вида

$$\frac{df(x)}{dx} = \frac{x-c}{b_0 + b_1 x + b_2 x^2} f(x). \qquad (2.3.27)$$

Здесь b_0, b_1, b_2, c — постоянные параметры распределения, которые однозначно определяются первыми четырьмя начальными моментами аппроксимируемого распределения.

Из уравнения (2.3.27) видно, что для унимодальных распределений модой является абсцисса $x = c$, так как в этой точке производная равна нулю. Применительно к началу отсчета случайной величины X от среднего значения (т.е. полагая $m = 0$) постоянные b_0, b_1, b_2 и c выражаются при помощи моментов [59]:

$$b_0 = -\mu_2(4\beta_2 - 3\beta_1)/B; \quad b_1 = c = -\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2+3)/B;$$
$$b_2 = -(2\beta_2 - 3\beta_1 - 6)/B; \quad B = 2(5\beta_2 - 6\beta_1 - 9). \qquad (2.3.28)$$

Здесь

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \gamma_1^2; \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \gamma_2 + 3, \qquad (2.3.29)$$

где γ_1 и γ_2 — коэффициенты асимметрии и эксцесса.

Общее решение дифференциального уравнения (2.3.27) может быть представлено в виде

$$y = f(x) = y_0 \exp[v(x)],$$

где

$$v(x) = \int \frac{(x-c)\,dx}{b_0 + b_1 x + b_2 x^2}. \qquad (2.3.30)$$

Значение этого интеграла зависит от характера корней уравнения

$$b_0 + b_1 x + b_2 x^2 = 0. \qquad (2.3.31)$$

Корни этого уравнения

$$x_{1,2} = \frac{-b_1 \pm \sqrt{b_1^2 - 4b_0 b_2}}{2b_2} \qquad (2.3.32)$$

определяется его дискриминантом

$$D = b_1^2 - 4b_0 b_2 = b_1^2(1 - 1/\kappa), \qquad (2.3.33)$$

где

$$\kappa = \frac{b_1^2}{4b_0 b_2} = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)} \qquad (2.3.34)$$

— параметр, определяющий тип корней уравнения (2.3.31).

Условие $D > 0$ соответствует вещественным корням уравнения (2.3.31). При этом параметр κ должен принимать значения $\kappa > 1$ или $\kappa < 0$. Если $\kappa > 1$, то $D < b_1^2$, а вещественные корни уравнения будут одинаковыми по знаку. Если $\kappa < 0$, то $D > b_1^2$ и вещественные корни будут разными по знаку. Условие $D < 0$ $(0 < \kappa < 1)$ соответствует двум комплексным корням уравнения (2.3.31). Следовательно, вся область значений параметра κ разделяется на три подобласти: $-\infty < \kappa < 0$, $0 < \kappa < 1$ и $1 < \kappa < \infty$.

Таким образом, каждому значению величины κ соответствует определенное значение корней уравнения (2.3.31) и, следовательно, определенный результат интегрирования дифференциального уравнения. От характера корней x_1 и x_2 зависит также интервал переменной x, на котором задано соответствующее распределение $f(x)$; вне этого интервала плотность вероятности полагается равной нулю. В результате интегрирования получается некоторый частный вид функции $y = f(x)$, представляющий тот или иной вид кривой распределения Пирсона. Поэтому величина κ может быть названа критерием определения типа кривой Пирсона.

В соответствии с тремя подобластями критерия κ выделяют три главных типа распределения (I, IV и VI типы). Границы этих подобластей соответствуют переходным кривым распределения. В частности, при $\kappa = 0$ кривая Пирсона:

а) совпадает с нормальным распределением при $\beta_2 = 3$;

б) относится к II типу при $\beta_2 < 3$;

в) относится к VII типу при $\beta_2 > 3$.

Если $\kappa = 1$, то кривую относят к V типу; при $\kappa = \pm\infty$ кривую относят к III типу. Кроме того, из основных типов кривых выделяют еще пять частных типов. В результате весь класс кривых Пирсона сводится к 12 типам плюс нормальное распределение.

Так как параметр κ связан с величинами β_1 и β_2, а последние, в свою очередь, определяются коэффициентами асимметрии γ_1 и эксцесса γ_2, то в конечном итоге все возможные формы распределения Пирсона определяются разными соотношениями между коэффициентами асимметрии и эксцесса. Это обусловлено тем, что среднее значение и дисперсия определяют только положение и «ширину» кривой и потому не влияют на ее форму. Знак коэффициента асимметрии также не существенен для формы в том смысле, что два распределения, отличающиеся только знаками γ_1, будут зеркально отображающимися относительно вертикальной прямой, проходящей через точку $x = m$.

Для классификации кривых распределения К. Пирсон предложил пользоваться диаграммой в плоскости переменных β_1 и β_2, в которой каждой паре значений (β_1, β_2) соответствует определенная форма кривой Пирсона. На диаграмме, приведенной на рис. 2.3.1, показаны области, линии и точки кривых Пирсона [8, 59].

На диаграмме выделена верхняя граница (критическая область), в которой не существует распределений (в том числе и распределений Пирсона) с соответствующими парами конечных значений переменных (β_1, β_2). Уравнение верхней границы $\beta_2 - \beta_1 - 1 = 0$. Нижняя граница распределений Пирсона определяется уравнением $8\beta_2 - 15\beta_1 - 26 = 0$.

В табл. 2.3.1 [8] приведены уравнения семи кривых Пирсона, наиболее часто используемых в прикладных задачах, с указанием начала отсчета переменной и области ее определения. Кривая I типа имеет колоколообразную форму при положительных значениях g_1, g_2; J-образную форму — при разных знаках g_1, g_2; U-образную форму — при отрицательных значениях g_1, g_2. Параметр распределения y_0 находится из условия нормировки плотности вероятности, а другие параметры определяются из равенства моментов распределения Пирсона и четырех первых моментов исходной случайной величины. В [59] даны формулы всех возможных типов кривых Пирсона. При этом параметры кривых и пределы изменения переменной x связаны конкретными функциональными зависимостями с коэффициентами дифференциального уравнения (2.3.31), а также с четырьмя первыми моментами случайной величины X.

В [15, 46] также приведены все 12 типов кривых Пирсона. Многие непрерывные распределения, в том числе нормальное $(\beta_1 = 0, \beta_2 = 3)$,

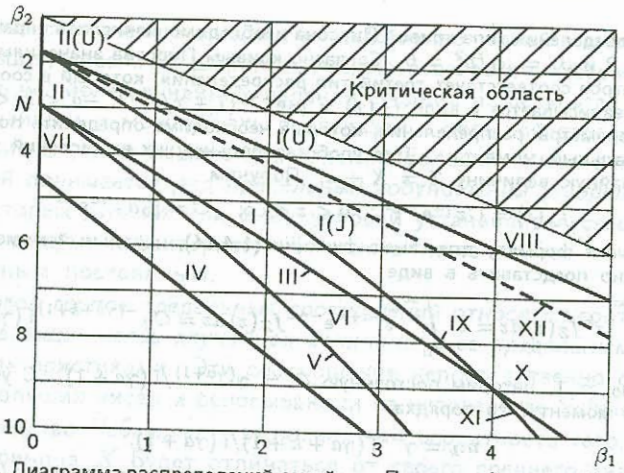


Рис. 2.3.1. Диаграмма распределений семейства Пирсона. Символы U и J означают U-образную и J-образную кривые. N — нормальное распределение

Таблица 2.3.1

Уравнения кривых Пирсона

Тип	Уравнение $y = f(x)$	Начало отсчета	Область определения
I	$y_0\left(1+\frac{x}{a_1}\right)^{g_1}\left(1-\frac{x}{a_2}\right)^{g_2}$	Мода	$-a_1 < x < a_2$
II	$y_0\left(1-\frac{x^2}{a^2}\right)^{g}$	Мода (среднее)	$-a < x < a$
III	$y_0 e^{-\gamma x}\left(1+\frac{x}{a}\right)^{\gamma a}$	Мода	$-a < x < \infty$
IV	$y_0 e^{-\nu \operatorname{arctg}(x/a)}\left(1+\frac{x^2}{a^2}\right)^{-g}$	Среднее $+\ \nu a/(2g-2)$	$-\infty < x < \infty$
V	$y_0 e^{-\nu/x} x^{-g}$	Начало кривой	$0 < x < \infty$
VI	$y_0 (x-a)^{g_2} x^{-g_1}$	Точка, отстоящая на $\lvert a\rvert$ от начала кривой	$0 < x < \infty$
VII	$y_0\left(1+\frac{x^2}{a^2}\right)^{-g}$	Среднее (мода)	$-\infty < x < \infty$

гамма, хи-квадрат, Стьюдента, дисперсионного отношения и некоторые другие распределения, рассмотренные выше, являются частными случаями семейства кривых Пирсона.

Замечательное свойство зависимости параметров распределений Пирсона только от четырех первых моментов позволяет использовать эти кривые для аппроксимации большого класса неизвестных законов распределения. Необходимые моменты находят либо теоретически, либо экспериментально.

Пример 2.3.1. Известны значения трех центральных моментов: $\mu_2 = 8$, $\mu_3 = 32$ и $\mu_4 = 384$. По кривым Пирсона, приведенным на рис. 2.3.1, определим тип распределения и вычислим его параметры.

Для определения типа кривой Пирсона необходимо вычислить параметры $\beta_1 = \mu_3^2/\mu_2^3 = 2$ и $\beta_2 = \mu_4/\mu_2^2 = 6$. Согласно кривым Пирсона значениям вычисленных параметров соответствует третий тип распределения, который в соответствии с табл. 2.3.1 записывается в виде $f_X(x) = y_0 e^{-\gamma x}(1 + x/a)^{\gamma a}$; $-a < x < \infty$. Здесь y, a, γ — параметры распределения, которые необходимо определить по известным трем центральным моментам. Для удобства дальнейших вычислений рассмотрим новую случайную величину $Z = X + a$. Получим

$$f_Z(z) = Cz^{\gamma a}e^{-\gamma z}, \quad 0 < z < \infty; \quad C = y_0 a^{-\gamma a}e^{\gamma a}.$$

Учитывая формулу для гамма-функции (1.4.14), начальный момент k-го порядка можно представить в виде

$$m_k = \int_0^\infty z^k f_Z(z)\,dz = C\int_0^\infty z^{\gamma a+k}e^{-\gamma z} f_Z(z)\,dz = C\gamma^{-(\gamma a+k+1)}\Gamma(\gamma a + k + 1).$$

Так как $m_0 = 1$, находим постоянную $C = \gamma^{(\gamma a+1)}/\Gamma(\gamma a + 1)$. С учетом этого начальный момент k-го порядка

$$m_k = \gamma^{-k}\Gamma(\gamma a + k + 1)/\Gamma(\gamma a + 1).$$

Принимая во внимание рекуррентное соотношение для гамма-функции (1.4.17), первые три начальных момента имеют вид

$$m \equiv m_1 = (\gamma a+1)/\gamma; \quad m_2 = (\gamma a+1)(\gamma a+2)/\gamma^2; \quad m_3 = (\gamma a+1)(\gamma a+2)(\gamma a+3)/\gamma^3.$$

При этом второй и третий центральные моменты определяются выражениями

$$\mu_2 = m_2 - m^2 = (\gamma a + 1)/\gamma^2; \quad \mu_3 = m_3 - 3m_2 m + 2m^3 = 2(\gamma a + 1)/\gamma^3.$$

Решая эти уравнения относительно неизвестных параметров γ и a, имеем

$$\gamma = 2\mu_2/\mu_3; \quad a = (4\mu_2^3 - \mu_3^2)/2\mu_3\mu_2.$$

Подставляя сюда численные значения второго и третьего центральных моментов, получим $\gamma = 1/2$, $a = 2$. Поэтому выражение для распределения случайной величины Z примет вид

$$f_Z(z) = \frac{z}{2^2\Gamma(2)}\exp\left(-\frac{z}{2}\right) = \frac{z^{4/2-1}}{2^{4/2}\Gamma(4/2)}\exp\left(-\frac{z}{2}\right).$$

Сравнивая полученное выражение с (2.2.3) нетрудно видеть, что они совпадают, если в (2.2.3) параметр $\nu = 4$. Таким образом, искомое распределение является хи-квадрат распределением с четырьмя степенями свободы. При этом значение четвертого центрального момента хи-квадрат распределения $\mu_4 = 12\nu(\nu + 4)$ для $\nu = 4$ совпадает с заданным значением $\mu_4 = 384$.

2.4. Предельные соотношения (теоремы) теории вероятностей

Одной из важнейших задач теории вероятностей является установление закономерностей случайных явлений, событий и т.п., происходящих с вероятностями, близкими к единице. При этом особую роль играют закономерности, возникающие в результате наложения большого числа независимых или слабо зависимых случайных факторов. Установление подобных закономерностей связано с *законом больших чисел*, являющимся основополагающим законом теории вероятностей. Действительно, в каждом отдельном явлении неизбежны случайные отклонения от некоторого среднего значения. В массе явлений случайные отклонения отдельных явлений взаимно погашаются, выравниваются.

Именно эта устойчивость средних и представляет собой содержание закона больших чисел. В широком смысле этот закон утверждает, что при большом числе случайных явлений средний их результат перестает быть случайным и может быть предсказан с большой степенью определенности. В узком смысле под *законом больших чисел в теории вероятностей понимается ряд предельных соотношений* (*теорем*), в каждом из которых для тех или иных условий устанавливается факт приближения средних характеристик большого числа опытов к некоторым определенным постоянным.

К первой группе предельных соотношений относятся соотношения, устанавливающие связь случайной величины с ее предельными числовыми характеристиками. Эти соотношения непосредственно связаны с законом больших чисел и основываются на неравенстве Чебышева.

Неравенство Чебышева утверждает, что вероятность того, что случайная величина X будет отличаться от своего среднего значения m по модулю не менее, чем на некоторое положительное число $\varepsilon > 0$, ограничена сверху величиной D/ε^2:

$$\mathbf{P}\{|X - m| \geqslant \varepsilon\} \leqslant D/\varepsilon^2. \tag{2.4.1}$$

Здесь $D = \sigma^2$ — дисперсия случайной величины X. Действительно, по выше сказанному определению для конкретного значения x справедливо неравенство $|X - m|/\varepsilon \geqslant 1$. Поэтому

$$\mathbf{P}\{|X - m| \geqslant \varepsilon\} = \int_{|X-m|\geqslant\varepsilon} f(x)\,dx \leqslant \frac{1}{\varepsilon^2}\int_{|X-m|\geqslant\varepsilon}(x-m)^2 f(x)\,dx \leqslant$$
$$\leqslant \frac{1}{\varepsilon^2}\int_{-\infty}^{\infty}(x-m)^2 f(x)\,dx = D/\varepsilon^2.$$

Из неравенства Чебышева (2.4.1) следует, что

$$\mathbf{P}\{m - \varepsilon < X < m + \varepsilon\} \leqslant 1 - D/\varepsilon^2, \tag{2.4.2}$$

т.е. чем меньше дисперсия случайной величины, тем больше вероятность того, что ее значения не выйдут за пределы $[m - \varepsilon, m + \varepsilon]$. Если обозначить $\varepsilon = k\sigma$, неравенство Чебышева принимает вид

$$\mathbf{P}\{|X - m| \geqslant k\sigma\} \leqslant 1/k^2. \tag{2.4.3}$$

Отсюда следует, что отклонение случайной величины, значительно превышающее среднее квадратическое отклонение, маловероятно.

Неравенство Чебышева дает верхнюю границу вероятности модуля отклонения случайной величины от своего среднего значения для произвольного распределения с конечными средним значением m и дисперсией D. Практически рассматриваемая вероятность отклонения значительно меньше величины, определяемой неравенством Чебышева. Например, для гауссовской случайной величины вероятность того, что случайная величина будет больше $m \pm 3\sigma$, равна приблизительно 0,003.

В то же время из неравенства (2.4.3) для $k = 3$ ($\varepsilon = 3\sigma$) следует, что вероятность того, что отклонения случайной величины от ее среднего значения выйдут за пределы трех средних квадратических отклонений, не может быть больше $1/9 \approx 0{,}111$.

Если случайная величина X принимает только положительные значения, то можно получить неравенство вида

$$\mathbf{P}\{X \geqslant \varepsilon\} \leqslant m/\varepsilon. \tag{2.4.4}$$

Это неравенство доказывается аналогично (2.4.1). Действительно, учитывая, что $x \geqslant \varepsilon$, получаем

$$\mathbf{P}\{X \geqslant \varepsilon\} = \int_{\varepsilon}^{\infty} f(x)\,dx \leqslant \frac{1}{\varepsilon}\int_{\varepsilon}^{\infty} xf(x)\,dx \leqslant \frac{1}{\varepsilon}\int_{0}^{\infty} xf(x)\,dx = \frac{m}{\varepsilon}.$$

Полагая, как и выше, $\varepsilon = km$, где $0 < k < \infty$, неравенство (2.4.4) можно записать в виде

$$\mathbf{P}\{X \geqslant km\} \leqslant \frac{1}{k}. \tag{2.4.5}$$

На основе неравенства Чебышева доказываются две основополагающие теоремы: Бернулли и теорема об арифметическом среднем.

Итак, пусть в результате опыта произошло k положительных исходов из n возможных. Величину k/n называют *частотой события*. Согласно *теореме Бернулли* частота события при увеличении числа независимых опытов стремится к его вероятности. Более строго эта теорема формулируется следующим образом: вероятность того, что частота события k/n отличается по модулю от своего среднего значения p не меньше, чем на ε, стремится к нулю при $n \to \infty$, как бы мало $\varepsilon > 0$ не было. Эту теорему часто также называют *законом больших чисел*.

Согласно биномиальному распределению среднее значение и дисперсия числа положительных исходов соответственно равны $m(k) = np$ и $D(k) = np(1-p)$, где p — вероятность одного положительного исхода. Тогда среднее значение и дисперсия случайной величины (k/n), представляющей собой частоту события, определяются как

$$\mathrm{M}\{k/n\} = p; \quad D(k/n) = \mathrm{M}\{(k/n - p)^2\} = p(1-p)/n.$$

При этом максимальное значение дисперсии соответствует $p = 1/2$ и равно $D_{\max} = 1/(4n)$. Согласно неравенству Чебышева

$$\mathbf{P}\{|(k/n) - p| \geqslant \varepsilon\} \leqslant 1/(4n\varepsilon^2).$$

В пределе при $n \to \infty$ для любого малого $\varepsilon > 0$, имеем

$$\lim_{n\to\infty} \mathbf{P}\{|(k/n) - p| \geqslant \varepsilon\} = 0, \tag{2.4.6}$$

т.е. частота $p = k/n$ появления событий в серии из n независимых опытов сходится по вероятности к вероятности появления этого события в одном опыте. Неравенство (2.4.6) эквивалентно неравенству

$$\lim_{n\to\infty} \mathbf{P}\{|(k/n) - p| < \varepsilon\} = 1.$$

Второй основополагающей теоремой теории вероятностей, использующей неравенство Чебышева и закон больших чисел, является *теорема об арифметическом среднем*. Согласно этой теореме, если независимые случайные величины X_i, $i = 1, \ldots, n$, одинаково распределены и имеют конечные средние значения $\mathrm{M}\{X_i\} = m_X$ и дисперсии $\mathrm{M}\{(X_i - m_X)^2\} = D_X$, то, какова бы ни была постоянная $\varepsilon > 0$, справедливо предельное выражение

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_X\right| < \varepsilon\right\} = 1. \tag{2.4.7}$$

Действительно, так как дисперсия суммы независимых одинаково распределенных случайных величин $Y = (1/n)\sum\limits_{i=1}^{n} X_i$ равна $D_Y = D_X/n$, то согласно неравенству Чебышева (2.4.1)

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_X\right| \geqslant \varepsilon\right\} \leqslant \frac{D_X}{n\varepsilon^2}. \tag{2.4.8}$$

При $n \to \infty$ и любом $\varepsilon > 0$ для арифметического среднего справедливо предельное соотношение

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - m_X\right| \geqslant \varepsilon\right\} = 0,$$

которое эквивалентно выражению (2.4.7).

Соотношение (2.4.7) известно также под названием *закона больших чисел*. Оно служит основанием для приближенного определения средних значений случайных величин по экспериментальным данным. Эти соотношения можно обобщить для начальных и центральных моментов произвольного порядка, а также числовых характеристик при функциональных преобразованиях случайных величин, если они имеют соответствующие конечные моменты. В частности, для дисперсии случайной величины, вычисляемой на основе опытных данных, справедливо предельное соотношение

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} (X_i - m)^2 - D_X\right| < \varepsilon\right\} = 1. \tag{2.4.9}$$

Применительно к корреляционному моменту попарно независимых случайных величин X_i и Y_i, $i = 1, \ldots, n$, со средними значениями m_X и m_Y следствием теоремы об арифметическом среднем является

предельное соотношение

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - m_X)(Y_i - m_Y) - K_{XY}\right| < \varepsilon\right\} = 1, \qquad (2.4.10)$$

где K_{XY} — корреляционный момент, определяемый согласно (1.5.25).

Ко второй группе соотношений относятся предельные законы распределения, базирующиеся на ряде теорем. Основной из них является центральная предельная теорема, которая посвящена установлению условий, при которых возникает нормальный закон распределения. Как отмечалось выше, нормальный закон распределения наиболее часто встречается в природе случайных явлений; он возникает при суммировании достаточно большого числа независимых случайных величин, каждая из которых в отдельности сравнительно мало влияет на сумму. Доказательство этого факта впервые было дано А.М. Ляпуновым в теореме, называемой *центральной предельной теоремой*, которая может быть сформулирована следующим образом.

Пусть $X_1, \ldots, X_n$ — независимые случайные величины с конечными средними значениями m_j и дисперсиями D_j, $j = 1, \ldots, n$, и выполняется условие

$$\lim_{n\to\infty} \sum_{j=1}^{n} \mathrm{M}\{|X_j - m_j|^3\} \Big/ \left(\sum_{j=1}^{n} D_j\right)^{3/2} = 0.$$

Тогда распределение суммы случайных величин

$$X = \sum_{j=1}^{n} X_j$$

асимптотически нормально со средним значением и дисперсией, равными

$$m = \sum_{j=1}^{n} m_j; \quad D = \sum_{j=1}^{n} D_j. \qquad (2.4.11)$$

Приведем вывод этой теоремы для одинаково распределенных слагаемых. При этом, естественно,

$$m = nm_0; \quad D = nD_0, \qquad (2.4.12)$$

где m_0 и D_0 — среднее значение и дисперсия отдельного слагаемого. Рассмотрим случайную величину

$$Y_j = (X_j - m_0)/\sqrt{D_0 n}. \qquad (2.4.13)$$

Центральные моменты случайной величины Y_j связаны с центральными

моментами μ_k случайной величины X_j соотношением

$$\mathrm{M}\{Y_j^k\} = \frac{1}{D_0^{k/2} n^{k/2}} \mathrm{M}\{(X_j - m_j)^k\} = \frac{\mu_k}{D_0^{k/2} n^{k/2}}.$$

Сумма случайных величин Y_j

$$Y = \sum_{j=1}^{n} Y_j = \frac{1}{\sqrt{D_0 n}} \sum_{j=1}^{n} (X_j - m_0) = \frac{X - m}{\sqrt{D_0 n}} \qquad (2.4.14)$$

имеет нулевое среднее значение и единичную дисперсию. Найдем предельное значение характеристической функции случайной величины Y при $n \to \infty$. Для этого характеристическую функцию $\theta_0(\vartheta)$ случайной компоненты Y_j представим в виде ряда (1.2.41), в котором начальные моменты m_k необходимо заменить на центральные μ_k:

$$\theta_0(\vartheta) = \sum_{k=0}^{\infty} \frac{1}{k!} (\mathrm{i}\vartheta)^k \mathrm{M}\{Y_j^k\} = \sum_{k=0}^{\infty} (\mathrm{i}\vartheta)^k \frac{\mu_k}{k! D_0^{k/2} n^{k/2}}. \qquad (2.4.15)$$

Характеристическая функция $\theta_n(\vartheta)$ суммы независимых слагаемых равна произведению характеристических функций слагаемых. Поэтому, учитывая, что $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = D_0$, имеем

$$\theta_n(\vartheta) = \left[\sum_{k=0}^{\infty} \frac{\mu_k (\mathrm{i}\vartheta)^k}{k! D_0^{k/2} n^{k/2}}\right]^n = \left[1 - \frac{\vartheta^2}{2n} - \mathrm{i}\frac{\mu_3 \vartheta^3}{6 D_0^{3/2} n^{3/2}} + \frac{\mu_4 \vartheta^4}{24 D_0^2 n^2} + \ldots\right]^n.$$

Осуществим предельный переход ($n \to \infty$), для чего сначала рассмотрим логарифм характеристической функции

$$\ln \theta_n(\vartheta) = n \ln\left[1 - \frac{\vartheta^2}{2n} - \mathrm{i}\frac{\mu_3 \vartheta^3}{6 D_0^{3/2} n^{3/2}} + \frac{\mu_4 \vartheta^4}{24 D_0^2 n^2} + \ldots\right].$$

Разлагая $\ln(1 + z)$ в степенной ряд (1.3.17), получим

$$\ln \theta_n(\vartheta) = -\frac{\vartheta^2}{2} - \mathrm{i}\frac{\mu_3 \vartheta^3}{6 D_0^{3/2} n^{1/2}} + \frac{\mu_4 \vartheta^4}{24 D_0^2 n} - \frac{\vartheta^4}{8n} - \frac{\mu_3^2 \vartheta^6}{72 D_0^3 n^2} + \ldots \qquad (2.4.16)$$

При $n \gg 1$ ($n \to \infty$) в правой части последнего выражения все члены за исключением первого члена стремятся к нулю. В результате имеем

$$\lim_{n\to\infty} \ln \theta_n(\vartheta) = -\vartheta^2/2.$$

Отсюда

$$\lim_{n\to\infty} \theta_n(\vartheta) = \exp(-\vartheta^2/2). \qquad (2.4.17)$$

Согласно (1.4.7) этой характеристической функции соответствует нормальная плотность вероятности с нулевым средним значением и еди-

ничной дисперсией, т.е.

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2).$$

Учитывая линейную связь между случайными величинами X и Y (2.4.14), получаем асимптотическое (предельное) выражение для плотности вероятности суммы n ($n \gg 1$) одинаково распределенных случайных величин с конечными средними значениями и конечными дисперсиями:

$$f_X(x) \approx \frac{1}{\sqrt{2\pi D}} \exp\left[-\frac{(x-m)^2}{2D}\right], \qquad (2.4.18)$$

где среднее значение m и дисперсия D определяются выражениями (2.4.12).

Выше в п. 1.3 с использованием формулы Стирлинга было найдено предельное соотношение (1.3.18), аналогичное (2.4.18), применительно к биномиальному распределению для достаточно большого числа n возможных положительных исходов (локальная формула Муавра–Лапласа). Совпадение вполне объяснимо, так как биномиально распределенная случайная величина является суммой независимых случайных положительных исходов (единиц) из n возможных исходов.

Найдем приближенное выражение для плотности вероятности суммы, когда условие $n \to \infty$ не выполняется, предполагая, как и выше, что слагаемые имеют одинаковые распределения.

Характеристическую функцию нормированной суммы (2.4.14) согласно (2.4.16) можно записать, ограничиваясь в первом приближении тремя членами разложения, в виде

$$\theta_n(\vartheta) \approx \exp\left(-\frac{\vartheta^2}{2}\right) \exp\left(-\mathrm{i}\gamma_1 \frac{\vartheta^3}{6\sqrt{n}} + \gamma_2 \frac{\vartheta^4}{24n}\right).$$

Здесь, как и выше, $\gamma_1 = \mu_3/D_0^{3/2}$ и $\gamma_2 = (\mu_4/D_0^2 - 3)$ — коэффициенты асимметрии и эксцесса отдельного слагаемого. Разлагая вторую экспоненту в степенной ряд и ограничиваясь двумя членами разложения, получим

$$\theta_n(\vartheta) \approx \exp\left(\frac{-\vartheta^2}{2}\right)\left[1 - \mathrm{i}\gamma_1 \frac{\vartheta^3}{6\sqrt{n}} + \gamma_2 \frac{\vartheta^4}{24n}\right]. \qquad (2.4.19)$$

Отсюда плотность вероятности случайной величины Y определяется выражением

$$f_Y(y) \approx \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\vartheta^2/2}\left[1 - \mathrm{i}\gamma_1 \frac{\vartheta^3}{6\sqrt{n}} + \gamma_2 \frac{\vartheta^4}{24n}\right] \mathrm{e}^{-\mathrm{i}\vartheta y}\, d\vartheta.$$

Этот интеграл разбивается на три интеграла, первый из которых равен

нормальной (гауссовской) функции

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\mathrm{i}\vartheta y - \vartheta^2/2)\, d\vartheta = \frac{1}{\sqrt{2\pi}} \exp(-y^2/2).$$

Остальные интегралы получаются из этого интеграла дифференцированием по параметру y:

$$\frac{1}{\sqrt{2\pi}} \frac{d^k}{dy^k} \mathrm{e}^{-y^2/2} = \frac{(-\mathrm{i})^k}{2\pi} \int_{-\infty}^{\infty} \vartheta^k \exp\left(-\mathrm{i}\vartheta y - \frac{\vartheta^2}{2}\right) d\vartheta.$$

Учитывая определение ортогональных полиномов Эрмита (2.3.7), последнее выражение можно записать в виде

$$\frac{\mathrm{i}^k}{2\pi} \int_{-\infty}^{\infty} \vartheta^k \exp\left(-\mathrm{i}\vartheta y - \frac{\vartheta^2}{2}\right) d\vartheta = \frac{1}{\sqrt{2\pi}} H_k(y) \mathrm{e}^{-y^2/2}.$$

В результате имеем

$$f_Y(y) \approx \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\left[1 + \frac{\gamma_1}{6\sqrt{n}} H_3(y) + \frac{\gamma_2}{24n} H_4(y)\right].$$

Так как случайные величины X и Y связаны между собой линейным соотношением (2.4.14), плотность вероятности суммы конечного числа одинаково распределенных случайных величин приближенно определяется формулой

$$f_X(x) \approx \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{(x-m)^2}{2D}\right)\left[1 + \frac{\gamma_1}{6\sqrt{n}} H_3\left(\frac{x-m}{\sqrt{D}}\right) + \frac{\gamma_2}{24n} H_4\left(\frac{x-m}{\sqrt{D}}\right)\right]. \qquad (2.4.20)$$

Сравнивая полученную формулу с (2.3.17), в которой под γ_1 и γ_2 понимаются коэффициенты асимметрии и эксцесса случайной величины X_i, можно сделать вывод, что при сложении случайных величин коэффициенты асимметрии и эксцесса суммы уменьшаются в $1/\sqrt{n}$ и $1/n$ соответственно.

Центральная предельная теорема вероятностей распространяется на многомерные случайные величины, а также справедлива (при определенных условиях) даже тогда, когда составляющие суммы не являются независимыми.

Пример 2.4.1. В качестве примера использования центральной предельной теоремы найдем распределение некоторой функции $g(X_1, \ldots, X_n)$, зависящей от n ($n \gg 1$) независимых случайных величин X_i со средними значениями $\mathrm{M}\{X_i\} = m_i$ и дисперсиями $\mathrm{M}\{(X_i - m_i)^2\} = D(X_i)$. Будем полагать, что функция $g(X_1, \ldots, X_n)$ имеет непрерывные производные первого и второго порядков в окрестности точки $\mathfrak{m}(m_1, \ldots, m_n)$. Разлагая функцию $g(X_1, \ldots, X_n)$ в окрестности точки $\mathfrak{m}(m_1, \ldots, m_n)$ в ряд Тейлора и ограничиваясь первыми двумя членами этого разло-

жения, имеем

$$g(X_1,\ldots,X_n)=g(m_1,\ldots,m_n)+\sum_{i=1}^{n}\left.\frac{\partial g}{\partial X_i}\right|_{m}(X_i-m_i)+R, \qquad (2.4.21)$$

где R — остаточный член, содержащий производные второго порядка. Согласно центральной предельной теореме сумма первых двух слагаемых при $n\gg 1$ приближенно имеет нормальное распределение со средним значением $m_g=g(m_1,\ldots,m_n)$ и дисперсией $D(g)=\sum\limits_{i=1}^{n}(\partial g/\partial X_i)^2|_m D(X_i)$.

Предельные соотношения, полученные на основании неравенства Чебышева, определяют границы вероятностей соответствующих случайных величин относительно их средних значений. Они, как правило, значительно превышают реальные вероятности. Вместе с этим для вычисления вероятностей сложных событий, зависящих от вероятностей элементарных событий, в математической статистике получены более точные границы, называемые *границами Чернова* [16]. Однако для их определения необходимо точно или приближенно знать законы распределения вероятностей элементарных событий. Границы Чернова широко используются в задачах определения верхних границ вероятностей ошибок при приеме дискретных кодированных сообщений. Формула (2.4.8), полученная на основании неравенства Чебышева, определяет верхнюю границу вероятности отклонения арифметического среднего относительно истинного среднего значения. Она, как правило, значительно превышает реальную вероятность этого отклонения.

В качестве иллюстрации рассмотрим границы Чернова при вычислении вероятностей верхних и нижних границ применительно к арифметическому среднему (выборочному среднему), чебышевские границы которых определяются (2.4.8).

Если независимые случайные величины X_i, $i=1,\ldots,n$, имеют конечные средние значения $\mathrm{M}\{X_i\}=m_X$ и дисперсии $\mathrm{M}\{(X_i-m_X)^2\}=D_X$, а также известно их распределение, то какова бы ни была постоянная $\varepsilon>0$, для верхних и нижних границ вероятностей арифметических средних справедливы выражения [16]

$$\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n}x_i\geqslant d\right\}\leqslant\left[\mathrm{M}\{\mathrm{e}^{\lambda(x-d)}\}\right]^n,\quad d>m_X+\varepsilon; \qquad (2.4.22)$$

$$\mathbf{P}\left\{\frac{1}{n}\sum_{i=1}^{n}x_i\leqslant d\right\}\leqslant\left[\mathrm{M}\{\mathrm{e}^{\lambda(x-d)}\}\right]^n,\quad d<m_X-\varepsilon. \qquad (2.4.23)$$

Здесь параметр λ есть корень уравнения

$$\frac{\mathrm{M}\{x\mathrm{e}^{\lambda x}\}}{\mathrm{M}\{\mathrm{e}^{\lambda x}\}}=d. \qquad (2.4.24)$$

При этом доказывается, что границы Чернова убывают экспоненциально с ростом объема выборки n, в то время как границы, полученные на основе неравенства Чебышева, убывает только как $1/n$. Вместе с этим еще раз необходимо отметить, что воспользоваться границей Чернова достаточно сложно, так как необходимо точно или приближенно знать одномерный закон распределения независимых одинаково распределенных составляющих X_i, $i=1,\ldots,n$.

Пример 2.4.2 [16]. Рассмотрим границу Чернова для совокупности бернуллиевских случайных величин

$$x_i=\begin{cases}1 & \text{с вероятностью } \mathbf{P}\{x_i=1\}=p,\\ 0 & \text{с вероятностью } \mathbf{P}\{x_i=0\}=q=1-p,\end{cases}\qquad i=1,2,\ldots,n. \qquad (2.4.25)$$

При этом среднее значение и дисперсия этих случайных величин равны $m_x=p$, $D_X=p(1-p)$. В этом случае

$$\mathrm{M}\{\mathrm{e}^{\lambda x}\}=\sum_{x=0}^{1}p(x)\mathrm{e}^{\lambda x}=(1-p)+p\mathrm{e}^{\lambda};\quad \mathrm{M}\{x\mathrm{e}^{\lambda x}\}=\sum_{x=0}^{1}xp(x)\mathrm{e}^{\lambda x}=p\mathrm{e}^{\lambda}.$$

Согласно (2.4.24) находим

$$d=\frac{p\mathrm{e}^{\lambda}}{(1-p)+p\mathrm{e}^{\lambda}}.$$

Решая это уравнение относительно параметра λ, имеем

$$\mathrm{e}^{\lambda}=\frac{d(1-p)}{p(d-1)},\quad \text{откуда}\quad \lambda=\ln\frac{d(1-p)}{p(1-d)}.$$

Так как $0\leqslant d\leqslant 1$, то $\lambda>0$ при $d>p$ и $\lambda<0$ при $d<p$.

Теперь можно вычислить

$$\mathrm{M}\{\mathrm{e}^{\lambda(x-d)}\}=\mathrm{M}\{\mathrm{e}^{\lambda x}\}\mathrm{e}^{\lambda d}=[(1-p)+p]\left[\mathrm{e}^{\lambda}\right]^d=$$

$$=\left[(1-p)+p\frac{d(1-p)}{p(1-p)}\right]\left[\frac{p(1-p)}{d(1-p)}\right]^d=\left(\frac{p}{d}\right)^d\left(\frac{1-p}{1-d}\right)^{1-d}.$$

Таким образом, используя (2.4.22) и (2.4.23), верхнюю и нижнюю границы для вероятности среднего арифметического биномиальных величин можно записать в виде

$$\mathbf{P}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}x_i\equiv\frac{k}{n}\right)-p\geqslant\varepsilon\right\}\leqslant\left[\left(\frac{p}{d}\right)^d\left(\frac{1-p}{1-d}\right)^{1-d}\right]^n,\quad 1\geqslant d=p+\varepsilon>p;$$

$$\mathbf{P}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}x_i\equiv\frac{k}{n}\right)-p\leqslant\varepsilon\right\}\leqslant\left[\left(\frac{p}{d}\right)^d\left(\frac{1-p}{1-d}\right)^{1-d}\right]^n,\quad 1\leqslant d=p-\varepsilon<p, \qquad (2.4.26)$$

где $k=\sum\limits_{i=1}^{n}x_i$ — число единиц из n возможных, k/n — частота появления единиц.

Вычислим $p_1=\mathbf{P}\{(k/n)-p\geqslant\varepsilon\}$, полагая $p=0{,}1$; $\varepsilon=0{,}4$; $n=13$. Подставляя эти величины в первую формулу (2.4.26), имеем $p_1\leqslant 1{,}3\cdot 10^{-3}$.

Учитывая, что $D_X=p(1-p)=0{,}1\cdot 0{,}9=0{,}09$, верхняя граница, основанная на законе больших чисел, согласно (2.4.8) равна $p_2=4{,}3\cdot 10^{-2}$, т.е. она более чем в 30 раз больше верхней границы Чернова. Истинная вероятность превышения числа биномиально распределенных случайных величин определяется формулой (1.3.11), в которой необходимо положить $k_2=n=13$, $k_1=[(d=p+\varepsilon)n]=[0{,}5\cdot 13]=[6{,}5]=6$, где квадратные скобки означают целую часть дробного числа, заключенного в них. Используя статистические таблицы [8, 41, 72 и др.], получаем более точное

значение искомой вероятности $p_3 = \sum_{k=5}^{13} C_{13}^k p^k (1-p)^{13-k} \approx 4 \cdot 10^{-4}$. Из этого примера видим, что верхняя граница Чернова значительно ближе к более точному значению, чем граница, получаемая на основе неравенства Чебышева, хотя требует знания закона распределения элементарных событий и более сложных вычислений.

Контрольные вопросы и задачи

2.1. В чем отличие нахождения законов распределения дискретных случайных величин от непрерывных случайных величин?

2.2. Показать, что случайные величины Y_1 и Y_2, которые взаимно однозначно связаны с независимыми случайными величинами X_1 и X_2 соотношениями $y_1 = \varphi_1(x_1)$ и $y_2 = \varphi_2(x_2)$, являются также независимыми.

2.3. Определить плотность вероятности случайной величины Y, связанной со случайной величиной X преобразованием $y = \varphi(x) = a\cos x$. Случайная величина X имеет плотность вероятность $f_X(x)$ на интервале $-\pi, \pi$. Конкретизировать полученное выражение для плотности вероятности Y применительно к прямоугольной плотности вероятности случайной величины X. Построить график полученной плотности вероятности Y.

2.4. В прямоугольной системе координат находится точка с независимыми случайными координатами X и Y, совместная плотность вероятности которой имеет вид $f_2(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$. Определить двумерную $f_2(r,a)$ и одномерные $f_R(r)$ и $f_A(a)$ плотности вероятности этой точки в полярных координатах r и a, где $r = \sqrt{x^2+y^2}$; $a = \operatorname{arctg}(y/x)$.

2.5. Изобразите кривую функции распределения случайной величины Y после нелинейного преобразования гауссовской случайной величины X с нулевым средним значением и единичной дисперсией посредством ограничения вида

$$\varphi(x) = \begin{cases} x & \text{при } x \geqslant 0, \\ 0 & \text{при } x < 0. \end{cases}$$

2.6. Какие основные распределения порождены гауссовскими случайными величинами?

2.7. К какому типу распределения Пирсона относится случайная величина, если коэффициенты асимметрии и эксцесса равны $\gamma_1 = \sqrt{2}$ и $\gamma_2 = 3$ соответственно?

2.8. На основе неравенства Чебышева (2.4.1) доказать справедливость предельного выражения (2.4.9) для выборочной дисперсии независимых случайных величин объема n.

2.9. Какова сущность неравенства Чебышева и центральной предельной теоремы для случайных величин?

2.10. Используя статистические таблицы (например, [8, 72]), вычислить вероятность отклонения среднего значения стьюдентской случайной величиной с $\nu = 10$ степенями свободы больше трех средних квадратических отклонений. Сравнить эту вероятность с вероятностью, получаемой из неравенства Чебышева.

2.11. Вычислить вероятность того, что случайная величина, распределенная по закону хи-квадрат с $\nu = 8$ степенями свободы, превышает значение km, где $k = 2$, $m = \nu$ — среднее значение случайной величины, и сравнить эту вероятность с вероятностью, получаемую из неравенства (2.4.5).

ЧАСТЬ II

Математическая статистика

> Эксперимент является безжалостным и суровым судьей работы теоретиков. Этот судья никогда не говорит теории *ДА*, в лучшем случае говорит *МОЖЕТ БЫТЬ*, а наиболее часто заявляет *НЕТ*. Если эксперимент согласуется с теорией, то для последней это означает *МОЖЕТ БЫТЬ*; если согласие отсутствует, то это означает *НЕТ*.
>
> *Альберт Эйнштейн. «Эволюция физики»*

ГЛАВА 3

Выборочный метод

3.1. Основные определения и задачи математической статистики

Математическая статистика — раздел прикладной математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и прикладных выводов. Предметом математической статистики является формальная математическая сторона статистических методов исследований, безразличная к специфической природе изучаемых объектов или событий. Математическая статистика базируется на понятиях и методах теории вероятностей, но решает свои специфические задачи своими методами.

Любая теория развивается в рамках некоторой модели, описывающей определенный круг реальных явлений, изучением которых и занимается данная теория. Задачей теории вероятностей является разработка методов нахождения вероятностей различных сложных событий для данной вероятностной модели. При этом полагаются известными исходные вероятности случайных событий (случайных величин), их средние значения, дисперсии, функции распределения, плотности вероятности и т.д. Однако в практических задачах положение совершенно иное: модель и соответствующие вероятностные характеристики неизвестны. Единственное, что можно сделать при изучении случайных явлений —

это ставить опыты и все характеристики, перечисленные выше, получить из этих опытов. При этом надо иметь в виду, что всякий эксперимент связан с ошибками наблюдений и измерений. Поэтому статистическая модель явления и характеристики, получаемые из опыта, являются приближенными. Следовательно, надо уметь оценить ошибку и достоверность (надежность) приближенного значения для полученной модели и статистических характеристик изучаемых событий (явлений, величин).

Кроме упомянутых выше задач определения характеристик (или параметров) модели и случайных величин, в математической статистике возникает задача проверки гипотез. Например, априори предполагается (выдвигается гипотеза), что изучаемые случайные величины подчиняются нормальному распределению. Необходимо по результатам опыта проверить, подтверждается ли данная гипотеза с заданной достоверностью. С подобными задачами приходится встречаться достаточно часто в различных прикладных задачах, когда наблюдаемые случайные величины аппроксимируются гауссовскими случайными величинами.

Итак, задача математической статистики состоит в том, чтобы по результатам поставленных экспериментов (наблюдений, измерений и т.п.) составить представление о той вероятностной ситуации, с которой сталкивается исследователь. Статистическая модель описывает также ситуации, когда в модели изучаемого явления (события) имеется та или иная неопределенность в задании вероятностных характеристик (параметров). *Математическая статистика* — это наука о статистических выводах. Она уточняет или выявляет структуру статистических моделей по результатам проводимых наблюдений. Методы математической статистики позволяют сделать не только выводы из поставленных экспериментов, но и организовать эти эксперименты наиболее рациональным образом. Необходимо также учесть, что в распоряжении исследователя обычно имеется ограниченный по объему (количеству) экспериментальный материал, по которому предстоит сделать заключение обо всей изучаемой вероятностной ситуации в целом.

Для определения ошибок измерения величин и вероятностей принятия тех или иных решений на основе наблюдаемых данных формируется определенная функция, называемая *статистикой*. Вид этой функции (статистики) определяется конкретной задачей математической статистики и удобством ее применения в прикладном плане.

Для упрощения рассматриваемых вопросов, как правило, будем полагать, если это не оговорено особо, что эксперименты между собой независимы, а условия их проведения постоянны. В этом допущении, конечно, имеется определенная идеализация, поскольку во времени условия эксперимента меняются.

Возникновение и развитие математической статистики определялось, в основном, потребностями практики. В настоящее время ее методы широко используются в различных областях науки и техники, в том числе в экономике, биологии, сельском хозяйстве и других науках, считавшихся далекими от математики.

Рассмотрим некоторые основные общепринятые определения. Исходным началом статистического исследования какой-либо случайной величины X является совокупность из n наблюдений, в результате которых случайная величина принимает конкретные значения

$$x_1, x_2, \ldots, x_n. \tag{3.1.1}$$

Эти значения называются *выборочными значениями* или *элементами выборки*. Множество всех возможных наблюдений (объектов) называется *генеральной совокупностью*. Множество отобранных наблюдений (объектов) $x_1, x_2, \ldots, x_n$ из генеральной совокупности называется *выборочной совокупностью* или *выборкой*. Число элементов выборочной совокупности (число выборочных значений) n называют *объемом выборки*. Если все элементы выборки представляют совокупность независимых случайных величин, выборку называют *независимой*. Если все элементы независимой выборки подчиняются одному и тому же распределению, выборку называют *однородной*.

Обычно генеральная совокупность содержит конечное множество возможных элементов (объектов), но оно, как правило, достаточно велико. Поэтому при теоретическом рассмотрении (для его упрощения) объем генеральной совокупности предполагается бесконечным. Подобное допущение оправдывается тем, что увеличение объема генеральной совокупности, начиная с некоторой достаточно большой величины, уже не сказывается на результатах обработки данных. Практически, если число элементов генеральной совокупности в сто и более раз превосходит объем выборки, генеральную совокупность можно полагать бесконечной.

Статистический метод исследования общих свойств генеральной совокупности каких-либо объектов на основе изучения свойств лишь части этих объектов, взятых в виде выборки, называется *выборочным методом*.

Сделанная выборка должна достаточно полно отражать особенности всех элементов (объектов) генеральной совокупности. Это очень важно, если генеральная совокупность имеет определенную неоднородность. Например, если измерения некоторой физической величины проведены при разных условиях, то в выборке должны быть представлены измерения, осуществленные при разных условиях. Иначе говоря, отобранная выборка будет представлять всю совокупность с приемлемой достоверностью при выполнении двух условий:

1) выборка должна быть *представительной* (многочисленной), чтобы в ней могли проявиться все закономерности, существующие в генеральной совокупности;

2) элементы выборки должны быть отобраны объективно, так, чтобы каждый элемент генеральной совокупности имел одинаковую возможность быть отобранным. Такую выборку называют *репрезентативной*.

Наиболее часто используются следующие способы отбора выборки.

Случайный отбор. Любой случайный эксперимент можно рассматривать как случайный выбор объекта из бесконечной генеральной совокупности. Иначе говоря, в результате проведения n независимых повторных экспериментов с некоторым объектом (явлением, системой, устройством и т.п.) получаем последовательность n наблюденных значений $x_1, x_2, \ldots, x_n$ некоторой случайной величины X. Эта последовательность наблюденных данных может рассматриваться как *случайная выборка* из бесконечной совокупности значений случайной величины X.

Серийный выбор. В качестве выборки берутся подряд объекты заданного объема n.

Механический или систематический отбор. Объекты отбираются через определенное число (например, каждый десятый, пятидесятый, сотый и т.п. объект).

Часто описанные выше способы отбора комбинируют в зависимости от конкретных условий эксперимента (исследования). В дальнейшем будем полагать, что рассматриваемая выборка получена в результате случайного отбора.

Изучение сходных объектов может производиться как по качественному, так и по количественному признакам.

При качественном изучении объектов определяется относительное число объектов, обладающих каким-либо качественным признаком (например, очень плохой, плохой, посредственный, хороший, очень хороший). При количественных признаках имеют дело с определением различных числовых характеристик или законов распределений генеральных совокупностей.

Расположим последовательность значений результатов наблюдений в возрастающем порядке, обозначая через $x_{(k)}$ k-е по величине наблюденное значение. Тогда последовательность (3.1.1) можно записать в виде

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(k)} \leqslant \ldots \leqslant x_{(n)}. \qquad (3.1.2)$$

Последовательность наблюденных значений исследуемой случайной величины, расположенных в возрастающем порядке, называется *вариационным рядом выборки* или просто *вариационным рядом*. Каждому элементу x_i выборки (3.1.1) соответствует определенное место в вариационном ряду. Место, занимаемое элементом выборки в вариационном ряде, называется *рангом элемента выборки*. Иногда k-й член вариационного ряда обозначают в виде $x^{(k)}$, $k = 1, 2, \ldots, n$. Члены вариационного ряда (3.1.2) в связи с их упорядочностью не являются взаимно независимыми величинами, даже если члены исходной выборки (3.1.1) являются взаимно независимыми величинами. Каждый член вариационного ряда называется *порядковой* (*ранговой*) статистикой: $x_{(k)}$ — k-я порядковая статистика; $x_{(1)}$ — наименьшее значение; $x_{(n)}$ — наибольшее значение; разность $x_{(n)} - x_{(1)}$ — *размах* или *широта* выборки. Если n — нечетное число: $n = 2h + 1$, то число $r \equiv x_{(h+1)} = x^*_{\mathrm{med}}$ называется *выборочной медианой*; она является приближением медианы x_{med},

которая определяется из решения уравнения $F(x_{\mathrm{med}}) = 1/2$. Если n — четное число ($n = 2h$), то выборочной медианой является число $r \equiv x^*_{\mathrm{med}} = (x_{(h)} + x_{(h+1)})/2$.

Вариационный ряд (3.1.2) является реализацией случайного вариационного ряда

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}. \qquad (3.1.3)$$

Элементы вариационного ряда дают наглядное представление о диапазоне возможных значений исходной случайной величины, степени ее разброса, центре группирования наблюдений и так далее. В связи с этим в прикладных задачах математической статистики представляет интерес знание распределений порядковых статистик и распределения размаха $U = X_{(n)} - X_{(1)}$. Наиболее просто выводятся выражения для распределений максимального и минимального значений выборки:

$$Y = \max\{X_1, X_2, \ldots, X_n\} \quad \text{и} \quad Z = \min\{X_1, X_2, \ldots, X_n\}.$$

Как уже отмечалось выше, в дальнейшем будем полагать, что выборка (3.1.1) является однородной и независимой.

Функция распределения максимального значения выборки определяется как

$$F_Y(y) = \mathbf{P}\{Y < y\} = \mathbf{P}\{X_1 < y, X_2 < y, \ldots, X_n < y\} = F^n_X(y), \qquad (3.1.4)$$

где $F_X(y) = \int_{-\infty}^{y} f_X(x)\,dx$ — функция распределения случайной величины X.

Аналогично находится *функция распределения минимального значения выборки*

$$F_Z(z) = \mathbf{P}\{Z < z\} = 1 - \mathbf{P}\{Z \geqslant z\} =$$
$$= 1 - \mathbf{P}\{X_1 \geqslant z, X_2 \geqslant z, \ldots, X_n \geqslant z\} = 1 - [1 - F_X(z)]^n. \qquad (3.1.5)$$

Полагая, что функция распределения случайной величины X дифференцируемая, *плотности вероятности максимального и минимального значений выборки* примут вид

$$f_Y(y) = n f_X(y) F^{n-1}_X(y); \quad f_Z(z) = n f_X(z)[1 - F_X(z)]^{n-1}. \qquad (3.1.6)$$

Для нахождения плотности вероятности произвольной порядковой статистики рассмотрим событие $t \leqslant T < t + dt$ и вычислим вероятность этого события $\mathbf{P}\{t \leqslant T < t+dt\} = f_T(t)\,dt$. Событие произойдет тогда и только тогда, когда одновременно (совместно) выполняются следующие условия: $(k-1)$ случайных величин X_i будут меньше t, одна случайная величина будет находиться в интервале $(t, t + dt)$, а $(n-k)$ случайных величин X_i будут больше $(t+dt)$. Вероятности этих совместных событий определяются как $F^{k-1}_X(t)$, $f_X(t)dt$ и $[1 - F_X(t)]^{n-k}$. Число способов,

которыми n наблюдений можно разбить на подобные три группы, равно

$$\frac{n!}{(k-1)!1!(n-k)!} = nC_{n-1}^{k-1} = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}. \quad (3.1.7)$$

Поэтому *плотность вероятности порядковой статистики* $T \equiv X_{(k)}$, $1 \leqslant k \leqslant n$, определяется как

$$f_T(t \equiv x_{(k)}) = nC_{n-1}^{k-1} F_X^{k-1}(t)[1 - F_X(t)]^{n-k} f_X(t). \quad (3.1.8)$$

Отсюда видно, что при $k = n$ и $k = 1$ следуют формулы (3.1.6).

Последняя формула позволяет получит плотность вероятности для выборочной медианы R. Если объем выборки — нечетное число ($n = 2h+1$), то выборочная медиана равна среднему члену вариационного ряда, т.е. $r \equiv k = h + 1$. Поэтому плотность вероятности выборочной медианы для подобного объема выборки

$$f_R(r) = \frac{n!}{[h!]^2} F_X^h(r)[1 - F_X(r)]^h f_X(r). \quad (3.1.9)$$

Если объем выборки — четное число: $n = 2h$, то выборочная медиана, как уже отмечалось выше, соответствует среднему арифметическому значению $r \equiv x^*_{\text{med}} = (x_{(h)} + x_{(h+1)})/2$. Для получения плотности вероятности выборочной медианы при четном числе объема выборки необходимо найти совместную (двумерную) плотность вероятности порядковых статистик $x_{(h)}$ и $x_{(h+1)}$, а затем согласно (2.1.26) вычислить *плотность вероятности выборочной медианы* $f_R(r)$, которая имеет вид [60]

$$f_R(r) = \frac{2 \cdot n!}{[(h-1)!]^2} \int_{-\infty}^{\infty} F_X^{h-1}(x)[1 - F_X(2r - x)]^{h-1} f_X(2r - x)\, dx. \quad (3.1.10)$$

Для определения функции распределения и плотности вероятности размаха $U = X_{(n)} - X_{(1)}$ необходимо знать совместную плотность вероятности для крайних членов вариационного ряда. В связи с этим рассмотрим совместную функцию распределения случайных величин $X_{(n)}$ и $X_{(1)}$

$$F_{YZ}(y, z) = \mathbf{P}\{X_{(1)} < z, X_{(n)} < y\} = 1 - \mathbf{P}\{X_{(1)} > z, X_{(n)} < y\} = \\ = 1 - [F_X(y) - F_X(z)]^n.$$

Дифференцируя это выражение дважды по y и z, получаем совместную плотность вероятности

$$f_{YZ}(y, z) = n(n-1)[F_X(y) - F_X(z)]^{n-2} f_X(y) f_X(z).$$

Учитывая, что $u = x_{(n)} - x_{(1)} \equiv y - z > 0$, функция распределения размаха определяется из выражения

$$F_U(u) = \mathbf{P}\{U < u\} = \iint_{0<y+z=u} f_{YZ}(y, z)\, dydz =$$

$$= \int_{-\infty}^{\infty} dz \int_{z}^{z+u} f_{YZ}(y, z)\, dy.$$

Подставляя в последний интеграл выше полученную формулу для совместной плотности вероятности $f_{YZ}(y, z)$, выражение для *функции распределения размаха* примет вид [8, 30, 60 и др.]

$$F_U(u) = n \int_{-\infty}^{\infty} [F_X(x + u) - F_X(x)]^{n-1} f_X(x)\, dx. \quad (3.1.11)$$

Дифференцируя это выражение по переменной u, получаем формулу для *плотности вероятности размаха* выборки

$$f_U(u) = n(n-1) \int_{-\infty}^{\infty} [F_X(x+u) - F_X(x)]^{n-2} f_X(x) f_X(x+u)\, dx. \quad (3.1.12)$$

Непосредственное использование написанных выше формул для порядковых статистик достаточно сложно. Однако применительно к гауссовской выборке имеются таблицы функций распределения экстремальных значений, медианы и размаха, а также таблицы крититических значений этих распределений [8, 41, 72]. Кроме этого, в математической статистике применительно к выборке достаточно большого объема ($n \gg 1$) были найдены асимптотические выражения для распределения порядковых статистик и распределения размаха выборки. Наиболее наглядные формулы получены для порядковых статистик при условии, что величины $k - 1$ и $n - k$ достаточно велики [8, 13, 36, 37 и др.]. В этом случае плотность вероятности порядковой статистики независимо от вида исходного распределения асимптотически нормальна, причем среднее значение и дисперсия статистики определяются достаточно просто, если известна плотность вероятности исходной случайной величины X. В частности, *асимптотическая плотность вероятности выборочной* медианы определяется как

$$f_R(r) \approx \frac{1}{\sqrt{2\pi}\sigma_R} \exp\left(-\frac{(r - x_{\text{med}})^2}{2\sigma_R^2}\right). \quad (3.1.13)$$

Здесь $\sigma_R^2 \approx 1/(4nf_{\text{med}}^2)$ — дисперсия выборочной медианы, где f_{med} — значение плотности вероятности случайной величины X при $x = x_{\text{med}}$: $f_{\text{med}} \equiv f_X(x = x_{\text{med}})$.

Если выборка является гауссовской со средним значением m_X и дисперсией σ_X^2, асимптотическое выражение плотности вероятности выборочной медианы имеет вид

$$f_R(r) = \frac{1}{\sqrt{2\pi}\sigma_R} \exp\left(-\frac{(r - x_{\text{med}})^2}{2\sigma_R^2}\right), \quad (3.1.14)$$

где $\sigma_R^2 = \sigma_X^2(\pi/2n)$ — дисперсия выборочной медианы.

Асимптотические формулы плотностей вероятности для крайних членов вариационного ряда и размаха даже при нормальном распределе-

нии случайной величины не являются нормальными. Они выражаются через переменные, функционально связанные с функцией распределения случайной величины X, и, как правило, достаточно сложны для непосредственного расчета. В [37, 57] приведены примеры вычисления асимптотических распределений экстремальных значений и размаха выборки для ряда распределений случайной величины X.

Для упрощения записи, особенно при вычислении статистических характеристик выборочных величин, случайные величины (функции) и их конкретные значения могут обозначаться одними и теми же символами, соответствующими конкретным значениям случайных величин (функций). Например, случайная величина X_i может обозначаться символом x_i. Подобная запись, хотя и не совсем корректна, часто используется в литературе по прикладной математической статистике. Такое упрощение обозначений не должно приводить к недоразумениям, поскольку смысл символа в записи соответствующего выражения будет достаточно ясно виден из контекста рассматриваемых вопросов.

3.2. Эмпирическая функция распределения

На основе выборки (3.1.1) или вариационного ряда (3.1.2) можно построить оценку (определить правило вычисления) функции распределения $F^*(x)$, часто называемую *выборочной* или *эмпирической функцией распределения*. Применительно к вариационному ряду оценка функции распределения определяется соотношением

$$F^*(x) = \begin{cases} 0 & \text{при } x < x_{(1)}; \\ k/n & \text{при } x_{(k)} \leqslant x < x_{(k+1)}, \quad k = 1, 2, \ldots, n-1; \\ 1 & \text{при } x \geqslant x_{(n)}. \end{cases} \tag{3.2.1}$$

Оценку функции распределения можно записать компактнее, если ввести *единичную функцию* (функцию единичного «скачка») (рис.3.2.1) вида

$$\kappa(x - x_0) = \begin{cases} 0 & \text{при } x < x_0, \\ 1 & \text{при } x \geqslant x_0. \end{cases} \tag{3.2.2}$$

Тогда

$$F^*(x) = \frac{1}{n}\sum_{k=1}^{n} \kappa(x - x_{(k)}) = \frac{1}{n}\sum_{i=1}^{n} \kappa(x - x_i). \tag{3.2.3}$$

Из-за случайного характера выборочных значений эмпирическая функция распределения $F^*(x)$ является случайной функцией, приближенно описывающей истинную функцию распределения $F(x)$ случайной величины X. Чтобы отличить $F^*(x)$ от $F(x)$, последнюю часто называют теоретической функцией распределения случайной величины.

Если все элементы выборки различны по своему значению, то величины всех «скачков» равны $1/n$, а типичный график функции $F^*(x)$ имеет вид, изображенный на рис. 3.2.2.

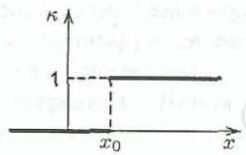


Рис. 3.2.1. Единичная функция

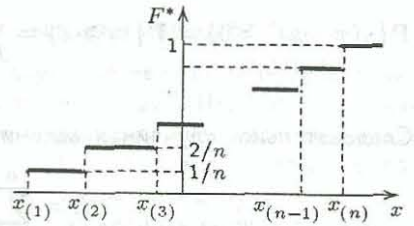


Рис. 3.2.2. Эмпирическая (выборочная) функция распределения

Итак, эмпирическая функция распределения (как и вариационный ряд) — это некоторая сводная характеристика выборки. Она однозначно определена и обладает всеми свойствами функции распределения, т.е. является неубывающей функцией и изменяется в интервале от нуля до единицы. Важным свойством эмпирической функции распределения является то, что при увеличении объема выборки n случайной величины X происходит сближение выборочной функции распределения с теоретической функцией распределения. Действительно, по определению функции распределения случайной величины X число $F(x)$ есть вероятность случайного события $\mathbf{P}\{X < x\}$, т.е. события, состоящего в том, что в результате опыта случайная величина X приняла значение, меньшее x. Число $F^*(x)$ есть относительная частота этого события при n опытах. Если не будет оговорено противное, то слово «частота» в дальнейшем будет означать относительную частоту. Замер каждого значения x_i в последовательности $x_1, x_2, \ldots, x_n$ можно рассматривать как опыт, в котором интересуемая нас случайная величина принимает некоторое значение. По теореме Бернулли (2.4.6) при большом объеме выборки ($n \gg 1$) частота события приближенно равна его вероятности, т.е. справедливо предельное соотношение

$$\lim_{n\to\infty} \mathbf{P}\{|F^*(x) - F(x)| < \varepsilon\} = 1, \tag{3.2.4}$$

где ε — любое положительное заранее заданное малое число.

Вычислим статистические характеристики оценки функции распределения: среднее значение и дисперсию, для чего обозначим

$$\Delta F(x) = F^*(x) - F(x) \tag{3.2.5}$$

— отклонение оценки функции распределения от ее теоретического значения.

Независимые случайные величины $\kappa(x - x_i)$ в (3.2.3) принимают два значения: единица или ноль с вероятностями соответственно:

$$\mathbf{P}\{\kappa(x - x_i) = 1\} = \mathbf{P}\{x_i < x\} = \int_{-\infty}^{x} f(x_i)\, dx_i = F(x) \equiv p; \tag{3.2.6}$$

$$\mathrm{P}\{\kappa(x-x_i)=0\}=\mathrm{P}\{x_i\geqslant x\}=\int_x^\infty f(x_i)\,dx_i=1-F(x)=1-p\equiv q. \tag{3.2.7}$$

Следовательно, случайная величина

$$K=\sum_{i=1}^{n}\kappa(x-x_i) \tag{3.2.8}$$

представляет собой сумму k единиц из n возможных и имеет биномиальное распределение (1.3.8) со средним значением

$$\mathrm{M}\{K\}=np=nF(x)$$

и дисперсией

$$D(K)=npq=nF(x)[1-F(x)].$$

В результате получаем, что среднее значение и дисперсия оценки функции распределения (3.2.3) равны

$$\mathrm{M}\{F^*(x)\}=F(x); \tag{3.2.9}$$

$$D[F^*(x)]=\frac{1}{n}F(x)[1-F(x)]. \tag{3.2.10}$$

Из (3.2.9) видно, что среднее значение рассматриваемой оценки функции распределения совпадает с ее теоретическим значением, т.е. систематическая составляющая ошибки оценки отсутствует. В этом случае говорят, что оценка *несмещенная*. Выражение (3.2.10) показывает, что дисперсия оценки функции распределения существенно зависит от значения x и достигает максимального значения

$$D[F^*(x)]_{\max}=0{,}25/n \tag{3.2.11}$$

при x, соответствующем $F(x)=0{,}5$. Иначе говоря, максимальная дисперсия оценки функции распределения соответствует медиане распределения $x=x_{\mathrm{med}}$.

3.3. Группированная выборка

При большом объеме выборки n (порядка нескольких сотен и более) использование как вариационного ряда, так и непосредственно выборки при проведении статистических исследований приводит к достаточно громоздким вычислениям. При этом могут потеряться некоторые существенные свойства статистического распределения («из-за деревьев не видно леса»). Поэтому при большом объеме выборки наблюдаемые значения группируются по заранее выбранным неперекрывающимся интервалам (разрядам).

3.3.1. Одномерная группированная выборка

Пусть в результате наблюдения случайной величины получена выборка достаточно большого объема $\mathbf{x}(x_1,x_2,\ldots,x_n)$. Наблюдаемые значения x_i группируются по ν заранее выбранным неперекрывающимся интервалам (разрядам) (x_{j-1},x_j), $j=1,2,\ldots,\nu$, где x_j — верхние границы интервалов. Длина (ширина) интервала равна $h_j=x_j-x_{j-1}$, $j=1,2,\ldots,\nu$. Часто все интервалы (или по крайней мере большинство из них) берутся одинаковыми по ширине (длине). После того как интервалы h_j выбраны, для заданной выборки $\mathbf{x}(x_1,x_2,\ldots,x_n)$ подсчитывают число n_j^* выборочных значений x_i, попавших в соответствующие интервалы (данные группируются), т.е. *образуется одномерная группированная выборка*. По значениям n_j^* вычисляются соответствующие частоты $p_j^*=n_j^*/n$, удовлетворяющие соотношению

$$\sum_{j=1}^{\nu}p_j^*=1,$$

и плотности выборочных значений в каждом интервале

$$\delta_j^*=p_j^*/h_j=n_j^*/(h_jn).$$

На основе полученных данных составляется таблица, в которой указывается номер разряда (интервала) j, границы разряда $[x_{j-1},x_j]$, численность разряда n_j^*, частоты p_j^* и плотности δ_j^* выборочных значений. Пример приведен в табл. 3.3.1.

Таблица 3.3.1

Параметр	Номер разряда j			
	1	2	...	ν
Границы разряда $[x_{j-1},x_j]$	$[x_0,x_1]$	$[x_1,x_2]$	...	$[x_{\nu-1},x_\nu]$
Численность разряда n_j^*	n_1^*	n_2^*	...	n_ν^*
Частота p_j^*	p_1^*	p_2^*	...	p_ν^*
Плотность δ_j^*	δ_1^*	δ_2^*	...	δ_ν^*

Группированные данные можно представить графически в виде ступенчатой кривой. По оси абсцисс откладываются соответствующие интервалы h_j и на каждом из них, как на основании, строится прямоугольник, высота которого равна плотности выборочных значений δ_j^*. Полученная ступенчатая кривая называется *гистограммой* (в переводе с греческого «гистос» означает ткань ленточного плетения).

На рис. 3.3.1 для иллюстрации приведена гистограмма для интервалов одинаковой длины $h_j=h=\mathrm{const}$. Поскольку площадь каждого прямоугольника гистограммы равна относительной частоте выборочных значений в соответствующем интервале, то по теореме Бернулли (2.4.6)

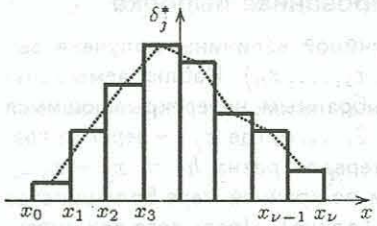


Рис. 3.3.1. Гистограмма выборочных значений и полигон частот

при $n \gg 1$ она будет сходиться к вероятности попадания значений случайной величины X в этот интервал h_j, т.е. p_j^* является оценкой вероятности попадания X в j-й интервал:

$$p_j = \int_{x_{j-1}}^{x_j} f(x)\,dx. \qquad (3.3.1)$$

Вычислим статистические характеристики частоты попадания выборочных значений в соответствующий интервал (x_{j-1}, x_j). Введем функцию $\eta_j(x)$, принимающую два значения:

$$\eta_j(x) = \begin{cases} 1 & \text{при } x_{j-1} \leqslant x \leqslant x_j, \\ 0 & \text{в остальных случаях.} \end{cases} \qquad (3.3.2)$$

Число элементов выборки n_j^*, попавших в интервал (x_{j-1}, x_j),

$$n_j^* = \sum_{i=1}^{n} \eta_j(x_i).$$

При этом вероятность того, что случайная величина η_j примет значение, равное единице, определяется выражением

$$\mathbf{P}\{\eta_j = 1\} = p_j = \int_{x_{j-1}}^{x_j} f(x)\,dx = F(x_j) - F(x_{j-1}).$$

Очевидно, что вероятность того, что случайная величина η_j примет нулевое значение, $\mathbf{P}\{\eta_j = 0\} = 1 - p_j$. Случайная величина η_j имеет биномиальное распределение. Поэтому при осуществлении n опытов среднее число единиц и дисперсия числа единиц соответственно равны

$$\mathbf{M}\{n_j^*\} = \sum_{i=1}^{n} \mathbf{M}\{\eta_j(x_i)\} = np_j;$$

$$D(n_j^*) = \mathbf{M}\left\{\left[\sum_{i=1}^{n} \eta_j(x_i) - np_j\right]^2\right\} = np_j(1-p_j).$$

Отсюда получаем, что оценка вероятности попадания случайной величины X в j-й интервал несмещенная:

$$\mathbf{M}\{p_j^*\} = n_j/n = p_j. \qquad (3.3.3)$$

При этом дисперсия оценки вероятности p_j^*

$$D(p_j^*) = \frac{1}{n^2} D(n_j) = \frac{1}{n} p_j(1-p_j). \qquad (3.3.4)$$

Так же, как и для функции распределения, максимум дисперсии оценки вероятности $D(p_j^*) = 0{,}25/n$ соответствует $p_j = 0{,}5$.

Из полученных выражений видно, что при достаточно больших объемах выборки ($n \gg 1$) частота попадания выборочных значений в соответствующий интервал приближенно равна вероятности попадания в этот же интервал самой случайной величины X.

Покажем, что если длина интервала h_j достаточно мала, а плотность вероятности $f(x)$ непрерывна, то вероятность попадания значений случайной величины X в интервал h_j приближенно равна $f(x_{0j})h_j$, где x_{0j} — середина соответствующего интервала. Действительно, разложим функцию распределения $F(x)$ в ряд Тейлора в окрестности точки $x_{0j} = (x_j + x_{j-1})/2$ и ограничимся двумя членами разложения:

$$F(x) \approx F(x_{0j}) + f(x_{0j})(x - x_{0j}).$$

При этом предполагается, что точки x_j и x_{j-1} достаточно близки, так что на этом интервале функцию распределения можно аппроксимировать линейной зависимостью. Отсюда значения функции распределения в этих точках определяются как

$$F(x_j) \approx F(x_{0j}) + f(x_{0j})(x_j - x_{0j});$$
$$F(x_{j-1}) \approx F(x_{0j}) + f(x_{0j})(x_{j-1} - x_{0j}).$$

Разность этих функций распределения

$$F(x_j) - F(x_{j-1}) \approx f(x_{0j})(x_j - x_{j-1}).$$

Отсюда плотность вероятности в точке x_{0j} приближенно можно представить в виде

$$f(x_{0j}) \approx [F(x_j) - F(x_{j-1})]/h_j = p_j/h_j. \qquad (3.3.5)$$

Аналогично подобное выражение можно записать и для оценки плотности вероятности в этой же точке:

$$f^*(x_{0j}) \approx [F^*(x_j) - F^*(x_{j-1})]/h_j = p_j^*/h_j = \delta_j^*. \qquad (3.3.6)$$

Найдем статистические характеристики оценки плотности вероятности. Среднее значение оценки

$$\mathbf{M}\{f^*(x_{0j})\} = \mathbf{M}\{p_j^*\}/h_j = p_j/h_j \approx f(x_{0j}), \qquad (3.3.7)$$

т.е. в первом приближении значение гистограммы в середине интервала приблизительно равно плотности вероятности в этой же точке. Дисперсия оценки плотности вероятности определяется выражением

$$D[f^*(x_{0j})] \approx \frac{D(p_j)}{h_j^2} = \frac{1}{h_j^2}\frac{p_j(1-p_j)}{n}. \qquad (3.3.8)$$

Если при группировке выборочных значений случайной величины по интервалам оказывается, что некоторые из них находятся на границах интервалов, то обычно принимают эти значения принадлежащим

двум соседним интервалам. При этом для каждого из этих интервалов число n_j^* увеличивается на 0,5.

Способ представления статистических данных с помощью гистограмм рекомендуется применять только для непрерывных случайных величин. Кроме того, представление статистических данных с помощью гистограмм обладает рядом очевидных недостатков:

- неопределенность в способе выбора интервалов;
- потеря информации при группировании данных, так как используются не сами выборочные значения x_i, а лишь частоты попадания в интервалы.

Особенно остро стоит вопрос выбора длины интервала h_j. Действительно, от выбора длины интервала зависит большая или меньшая выразительность полученных гистограмм. При слишком малых h_j из-за небольшого числа выборочных значений, попадающих в этот интервал, гистограмма будет содержать много случайного. Наоборот, при слишком больших интервалах h_j в гистограмме теряются индивидуальные черты сделанной выборки, и, следовательно, существенно сглаживаются особенности изучаемого явления. В отличие от гистограммы эмпирическая функция распределения менее чувствительна к группировке данных по промежуткам при ее построении. Как показывает практический опыт, для построения гистограмм обычно берут интервалы одинаковой длины ($h_j = h = \text{const}$), за исключением крайних интервалов h_1 и h_r. При этом желательно, чтобы в каждом из интервалов в среднем было не меньше десяти выборочных значений, число интервалов было порядка 10, а объем выборки должен быть порядка ста и более.

В методе гистограмм неизвестная плотность вероятности представляется в виде кусочно постоянного графика. Если плотность вероятности $f(x)$ достаточно гладкая функция, то лучшее приближение плотности вероятности можно получить, используя кусочно линейную аппроксимацию. Для этого ординаты гистограммы, соответствующие серединам интервалов, последовательно соединяют отрезками прямых. Подобная ломаная кривая называется *полигоном частот*. При этом, как было показано выше, значения полигона частот в точках перелома в первом приближении совпадают с теоретической плотностью вероятности. Примерный вид полигона частот приведен на рис. 3.3.1 в виде пунктирной кривой. Гистограмма также может быть аппроксимирована гладкой аналитической кривой на основе представления плотности вероятности в виде ряда по ортогональным полиномам (п. 2.3.1) или путем подбора соответствующих кривых Пирсона (п. 2.3.2) с предварительным вычислением первых четырех моментов на основе исходной выборки. Вопрос о достоверности аппроксимации обсуждается ниже при проверке гипотез согласия статистических оценок законов распределения предполагаемым теоретическим распределениям.

3.3.2. Двумерная группированная выборка

При наблюдении совокупности двух случайных величин X и Y регистрируются совместные (парные) выборочные значения x_i и y_i, $i = 1, 2, \ldots, n$. В случае большого объема ($n \gg 1$) выборок $\mathbf{x}$ и $\mathbf{y}$ аналогично одномерной выборке большого объема наблюдения x_i и y_i группируются по не перекрывающимися интервалам (разрядам) $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, k$, и $[y_{r-1}, y_r]$, $r = 1, 2, \ldots, q$. Эта выборка называется *двумерной группированной выборкой*.

Для каждой из случайных величин обозначим ширину интервалов как $h_j = x_j - x_{j-1}$ и $h_r = y_r - y_{r-1}$. Как правило, ширину интервалов h_j и h_r для каждой из случайных величин берут одинаковыми за исключением первых и последних интервалов.

По результатам наблюдений составляется табл. 3.3.2. В таблице вместо интервалов обычно приводятся средние (центральные) значения интервалов $x_j^* = (x_j + x_{j-1})/2$ и $y_r^* = (y_r + y_{r-1})/2$.

Таблица 3.3.2

Центр интервала x_j	Центр интервала y_r						
	y_1	y_2	...	y_r	...	y_q	Всего
x_1	n_{11}	n_{12}	...	n_{1r}	...	n_{1q}	$n_{.1}$
x_2	n_{21}	n_{22}	...	n_{2r}	...	n_{2q}	$n_{.2}$
...	...	...	...	...	...	...	...
x_j	n_{j1}	n_{j2}	...	n_{jr}	...	n_{jq}	$n_{.j}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kr}	...	n_{kq}	$n_{.k}$
Всего	$n_{.1}$	$n_{.2}$	...	$n_{.r}$	...	$n_{.q}$	n

В дальнейшем для сокращения записи центральные значения интервалов будем обозначать как x_j и y_r. В ячейке (клетке), находящейся на пересечении строки с x_j и столбца с y_r, указывается число выборочных значений n_{jr}, попавших в эту ячейку. Число n_{jr} называется *численностью ячейки*. В последних строке и столбце приводится суммарные числа $n_{j.}$ и $n_{.r}$ выборочных значений x_i и y_i, попавших в j-ю строку и в r-й столбец:

$$n_{j.} = \sum_{r=1}^{q} n_{jr}; \qquad n_{.r} = \sum_{j=1}^{q} n_{jr}. \tag{3.3.9}$$

Здесь и далее точка в нижнем индексе указывает на то, что рассматриваемая величина просуммирована по соответствующему индексу.

Для любого заданного значения случайной величины $X = x$ распределение случайной величины Y называется y-сечением. Оно представляет собой условное распределение случайной величины Y при заданном значении второй случайной величины $X = x$. Аналогично можно рассматривать x-сечение при заданном значении величины $Y = y$.

По выборочным значениям n_{jr} для каждой ячейки вычисляются соответствующие частоты и плотности выборочных распределений, определяемые соотношениями

$$p^*_{jr} = \frac{n_{jr}}{n}; \qquad \delta_{jr} = \frac{p^*_{jr}}{h_j h_r}. \tag{3.3.10}$$

По вычисленным выборочным значениям величин p^*_{jr} или δ^*_{jr} можно построить пространственную гистограмму. Для этого по осям координат x и y откладываются интервалы (x_{j-1}, x_j) и (y_{r-1}, y_r); в ячейке, соответствующей этим интервалам, как на основании, строится параллелограмм, высота которого равна p^*_{jr} или δ^*_{jr}.

Данные таблицы позволяют построить условные выборочные распределения в виде условных гистограмм. Например, условное выборочное распределение случайной величины Y в форме гистограмм условной частоты $p_{r|j}$ или условной плотности $\delta^*_{r|j}$ при фиксированном значении случайной величины $X = x_j$. При этом условные частота $p_{r|j}$ и плотность $\delta^*_{r|j}$ при фиксированном значении величины $X = x_j$ равны

$$p_{r|j} = \frac{n_{jr}}{n_{\cdot j}}; \qquad \delta_{r|j} = \frac{p^*_{r|j}}{h_j} = \frac{n_{jr}}{n_{\cdot j} h_r}. \tag{3.3.11}$$

Аналогично одномерной группированной выборке можно показать, что оценка вероятности попадания двумерной случайной величины (X, Y) в ячейку с координатами (x_j, y_r) несмещенная, т.е.

$$\mathrm{M}\{n_{jr}/n\} = p_{jr}. \tag{3.3.12}$$

Приведенная выше табл. 3.3.2 позволяет вычислить безусловные и условные группированные начальные и центральные моменты, а также группированный корреляционный момент (п. 3.4.3). Поэтому табл. 3.3.2 часто называют *корреляционной таблицей*.

В заключение этого параграфа отметим, что, так как истинные значения вероятностей p_j и p_{jr} неизвестны, то в формулах для дисперсий этих вероятностей в первом приближении вместо точных значений p_j и p_{jr} можно подставить их оценки p^*_j и p^*_{jr}.

3.4. Выборочные моменты

Аналогично тому, как теоретическому распределению $F(x)$ ставят в соответствие ее оценку (выборочное распределение) $F^*(x)$, теоретическим начальным m_k и центральным μ_k моментам ставят в соответствие их оценки (*выборочные начальные и центральные моменты*)

$$m^*_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k; \tag{3.4.1}$$

$$\mu^*_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - m^*_1)^k. \tag{3.4.2}$$

Правомерность подобного утверждения основана на теореме об арифметическом среднем, согласно которой выборочное среднее случайной величины с конечными моментами в пределе при достаточно большом объеме выборки стремится к своему среднему значению, т.е. выполняется предельное неравенство (2.4.7).

Между выборочными начальными и центральными моментами сохраняются те же соотношения, что и между теоретическими моментами; в частности, сохраняются и соотношения вида (1.2.34).

3.4.1. Выборочные начальные моменты

Среднее значение выборочного начального момента k-го порядка (3.4.1)

$$\mathrm{M}\{m^*_k\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{x_i^k\} = m_k, \tag{3.4.3}$$

т.е. выборочная оценка начального момента произвольного порядка является несмещенной оценкой (отсутствует систематическая ошибка).

Дисперсия выборочного начального момента k-го порядка определяется как

$$D(m^*_k) = \mathrm{M}\{(m^*_k - m_k)^2\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{(x_i^k - m_k)(x_j^k - m_k)\}.$$

В силу независимости выборочных значений

$$\mathrm{M}\{(x_i^k - m_k)(x_j^k - m_k)\} = \begin{cases} \mathrm{M}\{(x_i^k - m_k)^2\}, & i = j, \\ 0, & i \neq j. \end{cases}$$

Поэтому

$$D(m^*_k) = \frac{1}{n^2}\sum_{i=1}^{n}[\mathrm{M}\{x_i^{2k}\} - m_k^2] = \frac{m_{2k} - m_k^2}{n}. \tag{3.4.4}$$

Из начальных моментов в прикладном плане наибольший интерес представляет выборочный начальный момент первого порядка, называемый *выборочным средним значением* (*выборочным средним*):

$$m^*_1 \equiv m^* \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{3.4.5}$$

Согласно (3.4.3) и (3.4.4) выборочное среднее является несмещенной оценкой истинного среднего значения с дисперсией

$$D(\bar{x}) = \frac{m_2 - m^2}{n} = \frac{\mu_2}{n}, \tag{3.4.6}$$

где $\mu_2 \equiv D_X \equiv D$ — дисперсия случайной величины X.

Для рассматриваемой независимой выборки достаточно просто вычисляется третий центральный момент выборочного среднего

$$\mu_3(\bar{x}) = \mathrm{M}\{(\bar{x} - m)^3\} = \frac{1}{n^3}\sum_{i=1}^{n}(x_i - m)^3 = \frac{\mu_3}{n^2}, \qquad (3.4.7)$$

где μ_3 — третий центральный момент случайной величины X.

Аналогичным образом можно вычислить и другие моменты выборочного среднего. Однако эти вычисления становятся громоздкими. В частности, четвертый центральный момент выборочного среднего [37]

$$\mu_4(\bar{x}) = \frac{3\mu_2^2}{n^2} - \frac{\mu_4 - 3\mu_2^2}{n^3}. \qquad (3.4.8)$$

где μ_4 — четвертый центральный момент случайной величины X. Применительно к гауссовской выборке ($\mu_3 = 0$, $\mu_4 = 3\sigma^4$)

$$\mu_3(\bar{x}) = 0; \quad \mu_4(\bar{x}) = 3\sigma^4/n^2. \qquad (3.4.9)$$

3.4.2. Выборочная дисперсия

Выборочный центральный момент (3.4.2) второго порядка ($k = 2$) принято называть *выборочной дисперсией* и обозначать как

$$s^2 \equiv \mu_2^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (3.4.10)$$

При определении статистических характеристик выборочной дисперсии возможны случаи, когда априори известно среднее значение m. Тогда выборочная дисперсия s_1^2 определяется соотношением

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2. \qquad (3.4.11)$$

Из выражения

$$\mathrm{M}\{s_1^2\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{(x_i - m)^2\} = \mu_2$$

видно, что выборочная оценка дисперсии (3.4.11) является несмещенной оценкой дисперсии. Дисперсия этой оценки определяется из соотношения

$$D(s_1^2) = \mathrm{M}\{(s_1^2 - \mu_2)^2\} = \mathrm{M}\{(s_1^2)^2\} - \mu_2^2, \qquad (3.4.12)$$

где

$$\mathrm{M}\{(s_1^2)^2\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{(x_i - m)^2(x_j - m)^2\}.$$

Учитывая независимость выборочных значений x_i, имеем

$$\mathrm{M}\{(x_i - m)^2(x_j - m)^2\} = \begin{cases} \mathrm{M}\{(x_i - m)^4\} = \mu_4, & i = j, \\ \mathrm{M}\{(x_i - m)^2(x_j - m)^2\} = \mu_2^2, & i \neq j. \end{cases}$$

Поэтому

$$\mathrm{M}\{(s_1^2)^2\} = \frac{1}{n}[\mu_4 + (n-1)\mu_2^2].$$

Подставляя это выражение в (3.4.12), получаем формулу для дисперсии оценки дисперсии (3.4.11)

$$D(s_1^2) = \frac{\mu_4 - \mu_2^2}{n}. \qquad (3.4.13)$$

Применительно к гауссовской выборке $\mu_4 = 3\sigma^4$, поэтому для данного случая формула (3.4.13) упрощается и принимает вид

$$D(s_1^2) = 2\sigma^4/n. \qquad (3.4.14)$$

Рассмотрим более общую ситуацию, когда истинное среднее значение m неизвестно и оценка дисперсии осуществляется на основе вычисления выборочной дисперсии (3.4.10). Среднее значение оценки (3.4.10), учитывая независимость исходной выборки, определяется как

$$\mathrm{M}\{s^2\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{[(x_i - m) - (\bar{x} - m)]^2\} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{(x_i - m)^2\} - \mathrm{M}\{(\bar{x} - m)^2\} = \mu_2 - \frac{\mu_2}{n} = \frac{n-1}{n}\mu_2, \qquad (3.4.15)$$

т.е. выборочная дисперсия (3.4.10) является смещенной оценкой дисперсии. Смещение оценки дисперсии (3.4.10)

$$b(s^2) = -\mu_2/n. \qquad (3.4.16)$$

Дисперсия рассматриваемой оценки дисперсии определяется соотношением

$$D(s^2) = \mathrm{M}\{(s^2 - \mathrm{M}\{s^2\})^2\} = \mathrm{M}\{(s^2)^2\} - (\mathrm{M}\{s^2\})^2. \qquad (3.4.17)$$

При вычислении $\mathrm{M}\{(s^2)^2\}$ с целью упрощения записи в последующих преобразованиях без ограничения общности можно полагать, что среднее значение m равно нулю, так как это не меняет центрального момента произвольной случайной величины (центральный момент случайной величины не зависит от начала ее отсчета). Выражение для выборочной дисперсии (3.4.10) перепишем в виде

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \left[\frac{1}{n}\sum_{i=1}^{n}x_i\right]^2 = \frac{n-1}{n^2}\sum_{i=1}^{n}x_i^2 - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ i<j}}^{n}x_i x_j.$$

Здесь учтено, что

$$\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} x_i x_j;$$

$$\sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} x_i x_j = \sum_{i=1}^{n}\sum_{\substack{j=1\\ i>j}}^{n} x_i x_j + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i<j}}^{n} x_i x_j = 2\sum_{i=1}^{n}\sum_{\substack{j=1\\ i<j}}^{n} x_i x_j.$$

При этом в последней двойной сумме число слагаемых равно $n(n-1)/2$. Среднее значение квадрата выборочной дисперсии (второй начальный момент выборочной дисперсии s^2) определяется из соотношения

$$\mathrm{M}\{(s^2)^2\} = \mathrm{M}\left\{\frac{(n-1)^2}{n^4}\left(\sum_{i=1}^{n} x_i^2\right)^2 - \frac{4(n-1)}{n^4}\sum_{q=1}^{n} x_q^2 \sum_{i=1}^{n}\sum_{\substack{j=1\\ i<j}}^{n} x_i x_j + \right.$$
$$\left. + \frac{4}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1;i<j}^{n} x_i x_j\right)^2\right\}. \qquad (3.4.18)$$

Вычислим члены этого выражения по отдельности, учитывая независимость выборочных значений x_i. Имеем

$$\mathrm{M}\left\{\left(\sum_{i=1}^{n} x_i^2\right)^2\right\} = \mathrm{M}\left\{\sum_{i=1}^{n} x_i^4 + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i\neq j}}^{n} x_i^2 x_j^2\right\} = n\mu_4 + n(n-1)\mu_2^2;$$

$$\mathrm{M}\left\{\sum_{q=1}^{n}\sum_{i=1}^{n}\sum_{j=1;\ i<j}^{n} x_q^2 x_i x_j\right\} = 0;$$

$$\mathrm{M}\left\{\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\ i<j}}^{n} x_i x_j\right)^2\right\} = \sum_{i=1}^{n}\sum_{j=1;\ i<j}^{n} \mathrm{M}\{x_i^2\}\mathrm{M}\{x_j^2\} = \frac{n(n-1)}{2}\mu_2^2.$$

Подставляя эти соотношения в (3.4.18), получим

$$\mathrm{M}\{(s^2)^2\} = \frac{n-1}{n^3}[(n-1)\mu_4 + (n^2-2n+3)\mu_2^2]. \qquad (3.4.19)$$

При этом формула для дисперсии оценки дисперсии (3.4.17) принимает вид

$$D(s^2) = \frac{n-1}{n^3}[(n-1)\mu_4 + (3-n)\mu_2^2]. \qquad (3.4.20)$$

Применительно к гауссовской выборке ($\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$)

$$D(s^2) = \frac{2(n-1)}{n^2}\sigma^4. \qquad (3.4.21)$$

Суммарный квадрат ошибки оценки дисперсии (*рассеяние оценки дисперсии*) s^2 гауссовской величины

$$\widehat{D}(s^2) = D(s^2) + b^2(s^2) = \frac{2n-1}{n^2}\sigma^4. \qquad (3.4.22)$$

Из сравнения выражения (3.4.14) с выражениями (3.4.21) и (3.4.22) видно, что рассеяние и дисперсия смещенной оценки дисперсии меньше дисперсии ее несмещенной оценки. Это произошло за счет уменьшения, в среднем, оценки по сравнению с ее истинным значением.

В литературе по математической статистике [37] приведены приближенные выражения для третьего $\mu_3(s^2)$ и четвертого $\mu_4(s^2)$ центральных моментов выборочной дисперсии (3.4.10):

$$\mu_3(s^2) = \frac{\mu_6 - 3\mu_2\mu_4 - 6\mu_3^2 + 2\mu_2^3}{n^2} + o\left(\frac{1}{n^3}\right); \qquad (3.4.23)$$

$$\mu_4(s^2) = \frac{3(\mu_4 - \mu_2^2)^2}{n^2} + o\left(\frac{1}{n^3}\right). \qquad (3.4.24)$$

Здесь и далее символ $o(1/n^k)$ обозначает, что отброшенные члены имеют порядок $1/n^k$ и менее. Применительно к приведенным выше выражениям можно утверждать, что их погрешность имеет порядок $1/n^2$.

Для гауссовской выборки ($\mu_2 = \sigma^2$, $\mu_3 = 0$, $\mu_4 = 3\sigma^4$, $\mu_6 = 15\sigma^6$) приведенные выражения для третьего и четвертого центральных моментов выборочной дисперсии (3.4.10) упрощаются и принимают вид

$$\mu_3(s^2) = 8\sigma^6/n^2; \qquad (3.4.25)$$

$$\mu_4(s^2) = 12\sigma^8/n^2. \qquad (3.4.26)$$

Из формулы (3.4.15) видно, что для получения несмещенной оценки дисперсии оценку (3.4.10) следует умножить на коэффициент $n/(n-1)$, т.е. в качестве несмещенной оценки дисперсии необходимо взять величину

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (3.4.27)$$

Дисперсия этой оценки находится из соотношения

$$D(s_2^2) = \mathrm{M}\{(s_2^2 - \mu_2)^2\} = \mathrm{M}\{(s_2^2)^2 - \mu_2^2\}. \qquad (3.4.28)$$

При этом

$$\mathrm{M}\{(s_2^2)^2\} = \frac{n^2}{(n-1)^2}\mathrm{M}\{(s^2)^2\},$$

где $\mathrm{M}\{(s^2)^2\}$ определяется (3.4.19). В результате формула для дисперсии несмещенной оценки (3.4.27) примет вид

$$D(s_2^2) = \frac{1}{n(n-1)}[(n-1)\mu_4 + (3-n)\mu_2^2]. \qquad (3.4.29)$$

Применительно к гауссовской выборке формула для дисперсии оценки s_2^2 упрощается:

$$D(s_2^2) = \frac{2}{n-1}\sigma^4. \qquad (3.4.30)$$

Отношение рассеяния смещенной оценки дисперсии (3.4.10) к дисперсии несмещенной оценки (3.4.27) (применительно к гауссовской выборке)

$$\zeta = \frac{\widehat{D}(s^2)}{D(s_2^2)} = \left(1-\frac{1}{n}\right)\left(1-\frac{1}{2n}\right) < 1.$$

Однако при $n \gg 1$ все рассмотренные оценки дисперсии эквивалентны, а дисперсии оценок s^2 и s_2^2 стремятся к дисперсии несмещенной оценки с известным средним значением (3.4.14).

Вычислим корреляционный момент между выборочным средним $\bar{x}$ и выборочной дисперсией s_2^2, полагая для упрощения выводов $m = 0$. Учитывая независимость выборочных значений, имеем

$$K(\bar{x}, s^2) = \mathrm{M}\{\bar{x}(s_2^2 - \mu_2)\} = \frac{1}{n-1}\mathrm{M}\left\{\sum_{i=1}^{n}\bar{x}x_i^2 - n(\bar{x})^3\right\} =$$

$$= \frac{1}{n(n-1)}\left(1-\frac{1}{n}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\mathrm{M}\{x_i x_j x_l\} = \frac{\mu_3}{n}. \qquad (3.4.31)$$

Отсюда видно, что если распределение исходной выборки симметрично ($\mu_3 = 0$), то корреляционная связь между выборочными оценками $\bar{x}$ и s_2^2 отсутствует. В частности, это выполняется для гауссовской выборки. При этом для гауссовкой независимой выборки оценки $\bar{x}$ и s_2^2 между собой независимы [8, 30, 37, 52]. Это обусловлено тем, что гауссовские случайная величина $(\bar{x}-m)$ и случайная величина $(x_i-\bar{x})$, являющаяся составляющей выборочной несмещенной оценки дисперсии σ^2, некоррелированы. Действительно, полагая $m = 0$, имеем

$$\mathrm{M}\{(\bar{x}-m)(x_i-\bar{x})\} \equiv \mathrm{M}\{\bar{x}(x_i-\bar{x})\} = \mathrm{M}\left\{\frac{1}{n}\sum_{j=1}^{n}x_j\left(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i\right)\right\} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{x_i x_j\} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{x_i x_j\} = \frac{\sigma^2}{n} - \frac{1}{n^2}n\sigma^2 = 0. \qquad (3.4.32)$$

Поэтому можно утверждать, что для гауссовской независимой выборки оценки $\bar{x}$ и s_2^2 между собой независимы. Ниже в п. 4.9.2 утверждение о независимости оценок $\bar{x}$ и s_2^2 доказываются более строго.

3.4.3. Группированные оценки моментов

Как отмечалось выше, при большом объеме выборки ее элементы обычно объединяют в ν групп (разряды, интервалы), представляя выборочные значения в виде группированного вариационного ряда, примерный вид которого приведен в табл. 3.3.1. При этом оценки выборочного начального m_k^* и выборочного центрального μ_k^* моментов определяются по приближенным формулам

$$m_k^* \approx m_k' = \sum_{j=1}^{\nu} x_{0j}^k p_j^* = \sum_{j=1}^{\nu} x_{0j}^k (n_j^*/n); \qquad (3.4.33)$$

$$\mu_k^* \approx \mu_k' = \sum_{j=1}^{\nu}(x_{0j} - m_1^*)^k p_j^* = \sum_{j=1}^{\nu}(x_{0j} - m_1')^k (n_j^*/n), \qquad (3.4.34)$$

где m_k' и μ_k' — *группированные начальный и центральный моменты* k-го порядка; x_{0j} — среднее значение j-го интервала группирования; $p_j^* = (n_j^*/n)$ — относительная частота; n_j^* — численность разряда. При этом группированные выборочные центральные и начальные моменты связаны между собой соотношениями, аналогичными (1.2.34).

В математической статистике [8, 15, 37 и др.] приведены выражения, связывающие выборочные среднее $\bar{x}$, дисперсию s^2, третий и четвертый центральные моменты с соответствующими группированными моментами:

$$\begin{aligned} \bar{x} &\approx m_1'; \\ s^2 &\approx \mu_2' - \frac{h^2}{12}; \\ \mu_3^* &\approx \mu_3'; \\ \mu_4^* &\approx \mu_4' - \mu_2'\frac{h^2}{2} + \frac{7h^4}{240}. \end{aligned} \qquad (3.4.35)$$

Таким образом, в рассматриваемом приближении группированное выборочное среднее совпадает с выборочным средним, а выборочные группированный второй начальный момент, следовательно, и второй группированный центральный момент (группированная дисперсия) превышают соответствующие выборочные значения на величину $h^2/12$.

Поправки к группированным моментам называются *поправками Шеппарда*. Они становятся существенными, если поправка составляет заметную долю от среднего квадратического отклонения $\sigma = \sqrt{D}$. Если поправка составляет лишь малую долю от среднего квадратического отклонения, то ее можно не учитывать при вычислении соответствующих выборочных моментов.

Так как $\mathrm{M}\{n_j^*/n\} = \rho_j$, оценка выборочного среднего согласно (3.4.33) является несмещенной, а дисперсия ее оценки приближенно определяются формулой (3.4.6). Поскольку при группировании данных всегда $n \gg 1$, оценка выборочной дисперсии (3.4.34) будет асимптотически несмещенной; дисперсия оценки группированной дисперсии в первом приближении определяются формулой (3.4.20), а для гауссовской выборки — формулой (3.4.21).

Применительно к двумерной выборке, соответствующей случайным величинам X и Y, и группировании выборочных значений x_i и y_i, $i = 1, 2, \ldots, n$, в табл. 3.3.2 обычно вводятся дополнительная строка и дополнительный столбец. В них заносятся данные выборочных условных средних $\bar{x}_r$ и $\bar{y}_j$ соответственно по каждому столбцу и по каждой строке ($\bar{x}_r$ — среднее r-го x-сечения, $\bar{y}_j$ — среднее y-го сечения):

$$\bar{x}_r = \sum_{j=1}^{k} \frac{n_{jr}}{n_{\cdot r}} x_j; \quad \bar{y}_j = \sum_{r=1}^{q} \frac{n_{jr}}{n_{j\cdot}} y_r. \tag{3.4.36}$$

Кроме этого, по данным таблицы вычисляются следующие группированные выборочные моменты случайных величин X и Y:

- группированные безусловные выборочные средние:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} x_j = \frac{1}{n} \sum_{r=1}^{q} n_{\cdot r} \bar{x}_r; \tag{3.4.37}$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} y_j = \frac{1}{n} \sum_{j=1}^{k} n_{j\cdot} \bar{y}_j; \tag{3.4.38}$$

- группированные безусловные выборочные дисперсии:

$$s_X^2 = \frac{1}{n} \sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} (x_j - \bar{x})^2 = \frac{1}{n} \left(\sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} x_j^2 - n\bar{x}^2 \right); \tag{3.4.39}$$

$$s_Y^2 = \frac{1}{n} \sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} (y_r - \bar{y})^2 = \frac{1}{n} \left(\sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} y_r^2 - n\bar{y}^2 \right); \tag{3.4.40}$$

- группированные условные выборочные дисперсии

$$s_{X|y_r} = \sum_{j=1}^{k} \frac{n_{jr}}{n_{j\cdot}} (x_j - \bar{x}_r)^2; \tag{3.4.41}$$

$$s_{Y|x_j} = \sum_{r=1}^{q} \frac{n_{jr}}{n_{\cdot r}} (y_r - \bar{y}_j)^2. \tag{3.4.42}$$

Группированные безусловные и условные выборочные дисперсии (3.4.39)–(3.4.42) записаны в приближении без учета поправок Шеппарда.

Учитывая, что $\mathrm{M}\{n_{jr}/n\} = p_{jr}$, оценки безусловных выборочных средних величин X и Y в первом приближении являются несмещенными, а их дисперсии в том же приближении определяются формулой (3.4.6). Поскольку при группировании всегда $n \gg 1$, оценки безусловных дисперсий s_X^2 и s_Y^2 будут асимптотически несмещенными, а дисперсии их оценок в первом приближении определяются формулой (3.4.20) или формулой (3.4.21) применительно к гауссовским выборкам. Что касается условных выборочных средних и дисперсий, то при $n_j \gg 1$ и $n_r \gg 1$ можно воспользоваться изложенным для безусловных выборочных оценок средних и дисперсий.

Выборочный группированный корреляционный момент по аналогии с (1.5.29) определяется как

$$R_{XY} = \frac{1}{n} \sum_{j=1}^{k} \sum_{r=1}^{q} n_{jr} (x_j - \bar{x})(y_r - \bar{y}). \tag{3.4.43}$$

3.5. Косвенные оценки параметров

В технических приложениях достаточно часто приходится встречаться с двумя случаями так называемых *косвенных оценок параметров* (измерений неизвестных величин), основанных на измерениях других величин, связанных с первыми величинами точно известными функциональными зависимостями. В первом случае искомое значение величины функционально связано с измеренными величинами, для которых известны статистические характеристики их оценок. Во втором случае между измеряемыми величинами функциональная зависимость известна с точностью до некоторых неизвестных и подлежащих оценке параметров.

3.5.1. Функции от выборочных моментов

Пусть искомое значение некоторой величины θ находят на основе оценок других величин $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$, связанных с оцениваемой величиной θ известной функциональной зависимостью

$$\theta = \varphi(\vartheta_1, \vartheta_2, \ldots, \vartheta_m). \tag{3.5.1}$$

Будем полагать, что случайные величины (аргументы) $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$ оцениваются (измеряются) непосредственно на основе наблюдаемых выборочных значений. Такие оценки (измерения) называются *прямыми оценками* (измерениями). Аргументы по своей сути, как правило, представляют собой те или иные выборочные моменты соответствующей случайной величины. Обычно в качестве случайных аргументов $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$ фигурируют *выборочные средние и, реже, выборочные дисперсии* и другие выборочные моменты. Возникает задача вычисления среднего значения и дисперсии величины (параметра) θ по известным статистическим характеристикам аргументов $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$, т.е. необходимо найти

величины

$$m(\theta) = \mathrm{M}\{\theta\} = \mathrm{M}\{\varphi(\vartheta_1, \vartheta_2, \ldots, \vartheta_m)\}; \qquad (3.5.2)$$

$$D(\theta) = \mathrm{M}\{(\theta - \mathrm{M}\{\theta\})^2\}. \qquad (3.5.3)$$

По виду функциональной связи (3.5.1) различают косвенные оценки с линейной и нелинейной зависимостями между искомым параметром θ и аргументами $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$. Возможны косвенные оценки со смешанной функциональной зависимостью от аргументов (например, зависимость в виде некоторого полинома от величин $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$).

3.5.2. Линейные косвенные оценки

Для *линейной зависимости* между искомым параметром θ и аргументами $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$ имеем

$$\theta = \sum_{j=1}^{m} \beta_j \vartheta_j, \quad j = 1, 2, \ldots, m, \qquad (3.5.4)$$

где β_j — некоторый заранее известный постоянный коэффициент при j-м аргументе ϑ_j.

Среднее значение искомой величины θ

$$m(\theta) = \mathrm{M}\{\theta\} = \sum_{j=1}^{m} \beta_j \mathrm{M}\{\vartheta_j\} = \sum_{j=1}^{m} \beta_j m(\vartheta_j). \qquad (3.5.5)$$

Здесь $m(\vartheta_j) = \mathrm{M}\{\vartheta_j\}$ — средние значения случайных аргументов. Если оценки случайных аргументов несмещенные, то несмещенной будет и косвенная оценка параметра θ.

Для сокращения записи при вычислении дисперсии оценки параметра θ обозначим $\varepsilon_j = \vartheta_j - m(\vartheta_j)$. При этом

$$\mathrm{M}\{\varepsilon_j \varepsilon_l\} = \begin{cases} \mathrm{M}\{\varepsilon_j^2\} = D_j, & j = l, \\ \mathrm{M}\{\varepsilon_j \varepsilon_l\} = \sqrt{D_j D_l}\,\rho_{jl}, & j \neq l. \end{cases}$$

Здесь $D_j = \mathrm{M}\{[\vartheta_j - m(\vartheta_j)]^2\}$ — дисперсия случайного аргумента ϑ_j, а $\rho_{jl} = \mathrm{M}\{\varepsilon_j \varepsilon_l\}/\sqrt{D_j D_l}$ — коэффициент взаимной корреляции между случайными аргументами ϑ_j и ϑ_l. Тогда выражение для дисперсии косвенной оценки искомого параметра примет вид

$$D(\theta) = \mathrm{M}\{[\theta - m(\theta)]^2\} = \sum_{j=1}^{m} \sum_{l=1}^{m} \beta_j \beta_l \mathrm{M}\{\varepsilon_j \varepsilon_l\} =$$

$$= \sum_{j=1}^{m} \beta_j^2 D_j + \sum_{j=1}^{m} \sum_{l=1;\ j \neq l}^{m} \beta_j \beta_l \rho_{jl} \sqrt{D_j D_l}. \qquad (3.5.6)$$

Если оценки аргументов некоррелированы между собой, то выражение для дисперсии рассматриваемой оценки упрощается:

$$D(\theta) = \sum_{j=1}^{m} \beta_j^2 D_j. \qquad (3.5.7)$$

Поскольку истинные значения дисперсий и взаимных корреляционных моментов неизвестны, в качестве приближений можно взять их выборочные значения.

3.5.3. Нелинейные косвенные оценки

При анализе *нелинейных косвенных оценок* будем полагать, что функция (3.5.1) дифференцируема по аргументам. При этом задача вычисления статистических характеристик косвенной оценки параметра θ решается путем разложения нелинейной функции $\varphi(\vartheta_1, \vartheta_2, \ldots, \vartheta_m)$ в многомерный ряд Тейлора по аргументам $\vartheta_1, \vartheta_2, \ldots, \vartheta_m$ в окрестности точки, соответствующей их средним значениям $m(\vartheta_1), m(\vartheta_2), \ldots, m(\vartheta_m)$, и вычисления статистических характеристик членов разложения. Поскольку громоздкость вычисления существенно зависит от числа членов разложения и, особенно, от количества аргументов, ряд ограничивают соответствующим числом членов, позволяющим получить необходимую точность вычисления среднего значения и дисперсии косвенной оценки параметра θ. Как правило, если оставшиеся члены ряда обеспечивают точность статистических характеристик порядка $1/n^k$ (где n — объем выборки), то погрешность за счет отброшенных членов по порядку величины не превышает $1/n^{(k+1)}$.

Рассмотрим две наиболее часто встречающихся ситуации, когда косвенная нелинейная оценка зависит от одного и двух аргументов.

Итак, пусть оцениваемый параметр θ является нелинейной функцией одного параметра ϑ, т.е. $\theta = \varphi(\vartheta)$. Разложим эту функцию в ряд Тейлора по аргументу ϑ в окрестности среднего значения $m(\vartheta) = \mathrm{M}\{\vartheta\}$:

$$\theta = \varphi(\vartheta) = \varphi[m(\vartheta)] + \frac{\partial \varphi}{\partial \vartheta}[\vartheta - m(\vartheta)] + \frac{1}{2}\frac{\partial^2 \varphi}{\partial \vartheta^2}[\vartheta - m(\vartheta)]^2 + R_3.$$

Здесь и далее производные берутся в точке $\vartheta = m(\vartheta)$, а R_3 — остаточный член, который в форме Лагранжа определяется выражением

$$R_3 = \frac{1}{6}\frac{\partial^3 \varphi[m(\vartheta) + \nu(\vartheta - m(\vartheta))]}{\partial \vartheta^3}[\vartheta - m(\vartheta)]^3, \quad 0 < \nu < 1.$$

Полагая среднее значение остаточного члена по порядку величины меньше средних значений оставшейся суммы, находим среднее значение косвенной оценки параметра ϑ

$$m(\theta) = \mathrm{M}\{\theta\} = \varphi[m(\vartheta)] + \frac{1}{2}\frac{\partial^2 \varphi}{\partial \vartheta^2}\mathrm{M}\{[\vartheta - m(\vartheta)]^2\} = \varphi[m(\vartheta)] + \frac{1}{2}\frac{\partial^2 \varphi}{\partial \vartheta^2} D(\vartheta). \qquad (3.5.8)$$

Здесь $D(\vartheta)$ — дисперсия оценки случайного аргумента ϑ. Так как дисперсия выборочного начального момента первого порядка случайной

величины ϑ обратно пропорциональна объему выборки, можно утверждать, что при $n \gg 1$ полученная косвенная оценка асимптотически несмещенная.

В том же приближении дисперсия рассматриваемой косвенной оценки параметра θ определяется выражением

$$D(\theta) = \mathrm{M}\{[\theta - m(\theta)]^2\} = \left[\frac{\partial\varphi}{\partial\vartheta}\right]^2 D(\vartheta). \qquad (3.5.9)$$

Если случайным аргументом ϑ является выборочное среднее $\bar{x}$ (3.4.5) некоторой выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ со средним значением m_X и дисперсией D_X, то формулы для среднего значения и дисперсии косвенной оценки $\theta = \varphi(\bar{x})$ примут вид

$$m(\theta) = \varphi(m_X) + \frac{1}{2}\left.\frac{\partial^2\varphi}{\partial\bar{x}^2}\right|_{m_X} \frac{D_X}{n}; \qquad (3.5.10)$$

$$D(\theta) = \left(\frac{\partial\varphi}{\partial\bar{x}}\right)^2\Bigg|_{m_X} \frac{D_X}{n}. \qquad (3.5.11)$$

Из (3.5.10) наглядно видно, что оценка параметра, который нелинейно связан с выборочным средним аргумента, является асимптотически несмещенной.

Рассмотрим случай, когда косвенная оценка параметра θ нелинейно зависит от двух случайных аргументов ϑ_1 и ϑ_2, т.е. $\theta = \varphi(\vartheta_1, \vartheta_2)$. Разлагая эту функцию в двумерный ряд Тейлора в окрестности точки $m(\vartheta_1)$ и $m(\vartheta_2)$ и ограничиваясь тремя членами разложения, имеем

$$\theta = \varphi(\vartheta_1, \vartheta_2) = \varphi[m(\vartheta_1), m(\vartheta_2)] + \left[\frac{\partial\varphi}{\partial\vartheta_1}\delta_1 + \frac{\partial\varphi}{\partial\vartheta_2}\delta_2\right] + \\ + \frac{1}{2}\left[\frac{\partial^2\varphi}{\partial\vartheta_1^2}\delta_1^2 + 2\frac{\partial^2\varphi}{\partial\vartheta_1\partial\vartheta_2}\delta_1\delta_2 + \frac{\partial^2\varphi}{\partial\vartheta_2^2}\delta_2^2\right] + R_3.$$

Здесь $\delta_1 = \vartheta_1 - m(\vartheta_1)$; $\delta_2 = \vartheta_2 - m(\vartheta_2)$; R_3 — остаточный член; производные берутся в точке $[m(\vartheta_1), m(\vartheta_2)]$. При этом

$$\mathrm{M}\{\delta_1\} = \mathrm{M}\{\delta_2\} = 0; \quad \mathrm{M}\{\delta_1^2\} = D(\vartheta_1); \quad \mathrm{M}\{\delta_2^2\} = D(\vartheta_2),$$

а $\mathrm{M}\{\delta_1, \delta_2\} = \mu_{11}(\vartheta_1, \vartheta_2)$ — взаимный корреляционный момент между случайными аргументами ϑ_1 и ϑ_2. Полагая оценки аргументов несмещенными, среднее значение и дисперсия косвенной оценки параметра ϑ, нелинейно зависящего от двух случайных аргументов, в первом приближении определяются соотношениями

$$m(\theta) = \mathrm{M}\{\theta\} = \varphi[m(\vartheta_1), m(\vartheta_2)] + \frac{1}{2}\left[\frac{\partial^2\varphi}{\partial\vartheta_1^2}\mathrm{M}\{[\vartheta_1 - m(\vartheta_1)]^2\} + \right.$$
$$\left. + 2\frac{\partial^2\varphi}{\partial\vartheta_1\partial\vartheta_2}\mathrm{M}\{[\vartheta_1 - m(\vartheta_1)][\vartheta_2 - m(\vartheta_2)]\} + \frac{\partial^2\varphi}{\partial\vartheta_2^2}\mathrm{M}\{[\vartheta_2 - m(\vartheta_2)]^2\}\right] =$$

$$= \varphi[m(\vartheta_1), m(\vartheta_2)] + \frac{1}{2}\left[\frac{\partial^2\varphi}{\partial\vartheta_1^2}D(\vartheta_1) + 2\frac{\partial^2\varphi}{\partial\vartheta_1\partial_2}\mu_{11}(\vartheta_1, \vartheta_2) + \frac{\partial^2\varphi}{\partial\vartheta_2^2}D(\vartheta_2)\right]; \qquad (3.5.12)$$

$$D(\theta) = \mathrm{M}\left\{\left[\frac{\partial\varphi}{\partial\vartheta_1}[\vartheta_1 - m(\vartheta_1)] + \frac{\partial\varphi}{\partial\vartheta_2}[\vartheta_2 - m(\vartheta_2)]\right]^2\right\} =$$

$$= \left[\frac{\partial\varphi}{\partial\vartheta_1}\right]^2 D(\vartheta_1) + \left[\frac{\partial\varphi}{\partial\vartheta_2}\right]^2 D(\vartheta_2) + 2\frac{\partial\varphi}{\partial\vartheta_1}\frac{\partial\varphi}{\partial\vartheta_2}\mu_{11}(\vartheta_1, \vartheta_2). \qquad (3.5.13)$$

Если оценки случайных аргументов некоррелированы, то приведенные выше выражения несколько упрощаются и принимают вид

$$m(\theta) = \varphi[m(\vartheta_1), m(\vartheta_2)] + \frac{1}{2}\left[\frac{\partial^2\varphi}{\partial\vartheta_1^2}D(\vartheta_1) + \frac{\partial^2\varphi}{\partial\vartheta_2^2}D(\vartheta_2)\right]; \qquad (3.5.14)$$

$$D(\theta) = \left[\frac{\partial\varphi}{\partial\vartheta_1}\right]^2 D(\vartheta_1) + \left[\frac{\partial\varphi}{\partial\vartheta_2}\right]^2 D(\vartheta_2). \qquad (3.5.15)$$

Из выражений (3.5.12) и (3.5.14) видно, что в общем случае косвенная оценка параметра смещенная. Однако, учитывая обратно пропорциональную зависимость дисперсий оценок аргументов от объемов выборок, можно утверждать асимптотическую несмещенность косвенной оценки для достаточно больших объемов выборок при оценках аргументов.

В технических приложениях наиболее часто в качестве аргументов косвенной оценки фигурируют выборочные средние

$$\vartheta_1 \equiv \bar{x} = \frac{1}{n_X}\sum_{i=1}^{n_X} x_i; \quad \vartheta_2 \equiv \bar{y} = \frac{1}{n_Y}\sum_{i=1}^{n_Y} y_i$$

со средними значениями m_X, m_Y и дисперсиями $D(\bar{x}) = D_X/n_X$, $D(\bar{y}) = D_Y/n_Y$, где m_X, m_Y и D_X, D_Y — средние значения и дисперсии исходных выборок. Тогда формулы для вычисления среднего значения и дисперсии косвенной оценки для некоррелированных выборочных средних принимают вид

$$m(\theta) = \varphi(m_X, m_Y) + \frac{1}{2}\left.\left[\frac{\partial^2\varphi}{\partial\bar{x}^2}\frac{D_X}{n_X} + \frac{\partial^2\varphi}{\partial\bar{y}^2}\frac{D_Y}{n_Y}\right]\right|_{m_X, m_Y}; \qquad (3.5.16)$$

$$D(\theta) = \left.\left[\left(\frac{\partial\varphi}{\partial\bar{x}}\right)^2\frac{D_X}{n_X} + \left(\frac{\partial\varphi}{\partial\bar{y}}\right)^2\frac{D_Y}{n_Y}\right]\right|_{m_X, m_Y}. \qquad (3.5.17)$$

В [37] доказывается, что рассмотренные косвенные оценки, зависящие от выборочных моментов, при больших объемах выборок асимптотически нормальны со средним значением и дисперсией, определяемыми соответственно первым и вторым членами разложений в ряд Тейлора.

3.5.4. Примеры нелинейных косвенных оценок параметров

3.5.4.1. Выборочное среднее квадратическое отклонение

В качестве примера рассмотрим *выборочное среднее квадратическое отклонение* (с.к.о.), которое широко используется в прикладных задачах математической статистики в качестве характеристики точности измерения неизвестной величины. Выборочное с.к.о. применительно к неизвестному среднему значению выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ записывается в виде

$$s_2 = \sqrt{s_2^2} \equiv \vartheta = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.5.18)$$

где s_2^2 — несмещенная оценка дисперсии (3.4.27), характеризуемая дисперсией (3.4.29). После вычисления соответствующих производных согласно (3.5.8) и (3.5.9), имеем

$$m(s_2) = \sqrt{\mu_2}\left\{1 - \frac{1}{8n(n-1)}\left[(n-1)\frac{\mu_4}{\mu_2^2} + 3 - n\right]\right\}; \qquad (3.5.19)$$

$$D(s_2) = \frac{\mu_2}{4n(n-1)}\left[(n-1)\frac{\mu_4}{\mu_2^2} + 3 - n\right]. \qquad (3.5.20)$$

Применительно к гауссовской выборке ($\mu_2 = \sigma^2$, $\mu_4 = 3\sigma^4$)

$$m(s_2) = \sigma\frac{4n-5}{4(n-1)}; \qquad (3.5.21)$$

$$D(s_2) = \frac{\sigma^2}{2(n-1)}. \qquad (3.5.22)$$

Из (3.5.19) и (3.5.21) видно, что среднее значение с.к.о. даже в рассматриваемом первом приближении является смещенной оценкой. Применительно к гауссовской выборке смещение

$$b(s_2) = -\frac{\sigma}{4(n-1)}.$$

При $n \gg 1$ оценка с.к.о. становится асимптотически несмещенной. Для получения несмещенной оценки с.к.о. выражение (3.5.18) необходимо умножить на коэффициент

$$k_n = \left\{1 - \frac{1}{8n(n-1)}\left[(n-1)\frac{\mu_4}{\mu_2^2} + 3 - n\right]\right\}^{-1}, \qquad (3.5.23)$$

т.е. в качестве оценки с.к.о. рассматривать величину

$$s_{20} = k_n\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (3.5.24)$$

Применительно к гауссовской выборке коэффициент

$$k_n = \frac{4(n-1)}{4n-5}. \qquad (3.5.25)$$

Для гауссовской выборки может быть получено точное выражение для поправочного коэффициента k_n^*, учитывая формулу для плотности вероятности случайной величины s_2 в виде (2.2.6в), в которой число степеней свободы $\nu = n - 1$. Тогда среднее значение случайной величины s_2 определяется как

$$\mathrm{M}\{s_2\} = \frac{2(\nu/2)^{\nu/2}}{\sigma^\nu\Gamma(\nu/2)}\int_0^\infty z^\nu \exp\left(-\frac{\nu z^2}{2\sigma^2}\right)dz.$$

Осуществляя замену переменной $t = \nu z^2/2\sigma^2$, учитывая определение гамма-функции (1.4.14) и $\nu = n - 1$, получим

$$\mathrm{M}\{s_2\} = \sigma\sqrt{\frac{2}{n-1}}\frac{\Gamma(n/2)}{\Gamma\left((n-1)/2\right)}.$$

Из этого выражения видно, что оценка среднего квадратического отклонения смещенная. Для получения несмещенной оценки необходимо величину (3.5.18) умножить на коэффициент

$$k_n^* = \sqrt{\frac{n-1}{2}}\frac{\Gamma\left((n-1)/2\right)}{\Gamma(n/2)}. \qquad (3.5.26)$$

Значения коэффициентов k_n и k_n^* приведены в табл. 3.5.1, из которой видно, что $k_n \approx k_n^*$ с погрешностью менее 1,5 %. Этот частный результат свидетельствует о том, что приближенный метод нахождения статистических характеристик косвенных оценок вполне приемлем для прикладных задач математической статистики.

Таблица 3.5.1

n	3	4	5	6	7	10	15	20	30
k_n	1,143	1,090	1,067	1,053	1,044	1,029	1,018	1,013	1,009
k_n^*	1,128	1,085	1,064	1,051	1,042	1,028	1,018	1,013	1,009

Дисперсия оценки с.к.о.

$$D(s_{20}) = \mathrm{M}\{s_{20}^2\} - [\mathrm{M}\{s_{20}\}]^2 = \mu_2(k_n^2 - 1). \qquad (3.5.27)$$

Применительно к анализу гауссовской выборки

$$D(s_{20}) = \frac{8n-9}{(4n-5)^2}\sigma^2. \qquad (3.5.28)$$

Численными расчетами (хотя бы для гауссовской выборки) можно показать, что дисперсия несмещенной оценки с.к.о. больше не только дисперсии, но и рассеяния смещенной оценки с.к.о.

3.5.4.2. Выборочный коэффициент вариации (изменчивости)

В прикладных задачах в качестве меры относительной погрешности часто используют отношение среднего квадратического отклонения к истинному значению измеряемого параметра

$$v = \sqrt{\mu_2}/m. \tag{3.5.29}$$

В метрологии и математической статистике эту величину называют *коэффициентом вариации* или изменчивости случайной величины. Будем полагать, что среднее значение m является положительной величиной. *Выборочный коэффициент вариации* определим как

$$v^* = s_2/\bar{x} = \sqrt{\mu_2^*}/\bar{x}. \tag{3.5.30}$$

Пусть анализируемая выборка является гауссовской объема n. Выше (3.4.32) отмечалось, что выборочное среднее и выборочная дисперсия гауссовской случайной величины независимы. Тогда, вычисляя соответствующие производные согласно (3.5.14) и (3.5.15) и полагая при этом $\vartheta_1 = \mu_2^*$, $\vartheta_2 = \bar{x}$, $n \gg 1$, имеем

$$m(v^*) \approx \frac{\sigma}{m}\left[1 + \frac{1}{4n}\left(4\frac{\sigma^2}{m^2} - 1\right)\right]; \tag{3.5.31}$$

$$D(v^*) \approx \frac{\sigma^2(m^2 + 2\sigma^2)}{2m^4 n}. \tag{3.5.32}$$

Из (3.5.31) видно, что оценка коэффициента вариации асимптотически несмещенная.

Во всех выше приведенных выражениях фигурируют истинные значения соответствующих моментов, которые, как правило, не известны. Поэтому при вычислении статистических характеристик косвенной оценки в первом приближении при $n \gg 1$ вместо истинных значений параметров распределения можно взять их оценки (выборочные значения) m^*, s_2^2 и μ_4^*.

3.5.5. Косвенные оценки параметров, определяемые из системы уравнений. Метод наименьших квадратов

Как отмечалось выше, часто непосредственно измеряемые величины X и Y между собой связаны некоторой функциональной зависимостью $Y = \varphi(X; \mathrm{b})$ с неизвестными параметрами (коэффициентами) $\mathrm{b} = (b_0, b_1, \ldots, b_p)$. Необходимо на основе прямых измерений величин X и Y оценить эти параметры и вычислить их статистические характеристики.

Для однозначного определения параметров b_j, $j = 0, 1, 2, \ldots, p$, число n исходных уравнений должно быть, по крайней мере, не меньше, чем число неизвестных параметров $p + 1$. Для упрощения задачи будем полагать, что проведено n пар независимых измерений x_i и y_i. По имеющимся n парам значений (x_i, y_i) имеем n уравнений

$$y_i = \varphi(x_i; b_0, b_1, \ldots, b_p), \quad i = 1, 2, \ldots, n. \tag{3.5.33}$$

Если имеется столько исходных пар измерений, сколько неизвестных параметров $(n = p + 1)$, то определение параметров есть чисто алгебраическая задача, требующая решения системы, число уравнений которой равно числу неизвестных. Если произведено больше наблюдений, чем это необходимо $(n > p + 1)$, и если измерения величин x_i и y_i свободны от ошибок, то получающаяся система оказывается «переопределенной», так как число уравнений n больше числа неизвестных $p + 1$. В этом случае лишнее число уравнений можно отбросить, так как они не добавляют ничего нового к прежним уравнениям. Применительно к неточным измерениям и случайным величинам каждое новое измерение добавляет некоторую дополнительную информацию. Уравнения, в том виде, как они получились, являются несовместными, поскольку каждая система из $p + 1$ уравнений будет давать разные значения неизвестных параметров. Иначе говоря, получающаяся система из n уравнений (3.5.33) не может быть решена однозначно. Такую систему надо «подогнать» соответствующим способом к однозначной системе, т.е. получить систему из $p + 1$ совместных уравнений, в которой учитывалась бы вся совокупность из n измерений.

Рассмотрим разность между измеренными значениями y_i и «теоретическими» значениями $\varphi(x_i; b_0, b_1, \ldots, b_p)$, которые для полученной выборки следуют из известного вида функциональной зависимости

$$\alpha_i = y_i - \varphi(x_i; b_0, b_1, \ldots, b_p), \quad i = 1, 2, \ldots, n. \tag{3.5.34}$$

Эти разности называются *невязками*. Хотя невозможно найти значения неизвестных параметров $\mathrm{b}(b_0, b_1, \ldots, b_p)$, которые делали бы каждую из невязок равной нулю, всегда можно подобрать такие значения параметров b_j, которые делают сумму квадратов невязок минимальной:

$$G = \sum_{i=1}^{n} [y_i - \varphi(x_i; b_0, b_1, \ldots, b_p)]^2 = \min. \tag{3.5.35}$$

Минимум этого функционала эквивалентен его минимуму по каждому из параметров b_j, $j = 0, 1, \ldots, p$. Поэтому оценки параметров b_j^* находятся из системы уравнений

$$\frac{\partial G}{\partial b_j} = 0 \quad \text{при} \quad b_0 = b_0^*,\ b_1 = b_1^*, \ldots,\ b_p = b_p^* \quad (j = 0, 1, \ldots, p). \tag{3.5.36}$$

Метод определения неизвестных параметров функциональной связи между наблюдаемыми (измеряемыми) величинами на основе минимизации суммы квадратов невязок (разности между значениями, получен-

ными из измерений, и теоретическими данными) называется методом наименьших квадратов. Ниже (п. 4.5.1) будет показано, что метод наименьших квадратов является частным случаем метода максимального правдоподобия, когда наблюдения являются гауссовскими выборочными значениями. Исторически метод наименьших квадратов был разработан Гауссом применительно к выравниванию ошибок при астрономических и геодезических наблюдениях.

Нахождение оценок коэффициентов $\mathfrak{b}(b_0, b_1, \ldots, b_p)$ и их статистических характеристик для произвольной функциональной зависимости $Y = \varphi(X; \mathfrak{b})$, как правило, достаточно сложно. Задача существенно упрощается, если система уравнений (3.5.36) является линейной относительно оцениваемых коэффициентов $\mathfrak{b}(b_0, b_1, \ldots, b_p)$.

При определении статистических характеристик оценок коэффициентов обычно полагают, что отсутствуют систематические ошибки измерений величин (x_i, y_i), а ошибки определения или задания величин X пренебрежимо малы по сравнению с ошибками определения величин Y. При проведении опытов подобная ситуация встречается достаточно часто, когда величины X задаются с необходимой точностью. В этом случае $\mathbf{M}\{y_i\} = \varphi(x_i; b_0, b_1, \ldots, p)$. Кроме этого считается, что условная дисперсия случайной величины Y постоянна, т.е.

$$\mathbf{M}\{[\alpha_i = y_i - \varphi(x_i; b_0, b_1, \ldots, p)]^2\} = D_{Y|x} \equiv D.$$

Применим метод наименьших квадратов для оценки параметров в случае, когда уравнения (функциональные зависимости) являются линейными, а число неизвестных параметров равно двум. Итак, пусть система уравнений зависит от наблюдаемых двух случайных величин X_1 и Y. Для каждой совокупности выборочных значений (x_{1i}, y_i) имеем уравнения вида

$$b_0 + b_1 x_{1i} = y_i, \quad i = 1, 2, \ldots, n; \quad n \geqslant 2. \tag{3.5.37}$$

Здесь, как и выше, b_0, b_1 — искомые неизвестные параметры.

Для дальнейшего рассмотрения введем фиктивную величину $X_0 \equiv \equiv x_0 \equiv 1$. Тогда исходное уравнение (3.5.37) запишется в виде

$$b_0 x_{0i} + b_1 x_{1i} = y_i, \quad i = 1, 2, \ldots, n; \quad n \geqslant 2. \tag{3.5.38}$$

Согласно методу наименьших квадратов необходимо образовать функционал

$$G = \sum_{i=1}^{n} [y_i - (b_0 x_{0i} + b_1 x_{1i})]^2 \tag{3.5.39}$$

и решить систему уравнений $\partial G/\partial b_0 = dG/\partial b_1 = 0$ при $b_0 = b_0^*$, $b_1 = b_1^*$. Вычисляя частные производные (3.5.39) и приравнивая их к нулю, получаем систему линейных уравнений относительно неизвестных параметров в виде

$$\begin{aligned} b_0^* \sum_{i=1}^{n} x_{0i}^2 + b_1^* \sum_{i=1}^{n} x_{0i} x_{1i} &= \sum_{i=1}^{n} x_{0i} y_i; \\ b_0^* \sum_{i=1}^{n} x_{1i} x_{0i} + b_1^* \sum_{i=1}^{n} x_{1i}^2 &= \sum_{i=1}^{n} x_{1i} y_i. \end{aligned} \tag{3.5.40}$$

Эту систему уравнений в подобной записи называют системой *нормальных уравнений*, а также системой *условных уравнений*, поскольку число исходных уравнений n больше числа p неизвестных величин (параметров). При написании нормальных уравнений часто пользуются обозначениями Гаусса

$$\sum_{i=1}^{n} x_{0i}^2 = [x_0 x_0]; \quad \sum_{i=1}^{n} x_{0i} x_{1i} = [x_0 x_1] \equiv [x_1 x_0]; \quad \sum_{i=1}^{n} x_{0i} y_i = [x_0 y] \text{ и т.д.}$$

В обозначениях Гаусса нормальные уравнения принимают более компактный вид:

$$\begin{aligned} [x_0 x_0] b_0^* + [x_0 x_1] b_1^* &= [x_0 y]; \\ [x_1 x_0] b_0^* + [x_1 x_1] b_1^* &= [x_1 y]. \end{aligned} \tag{3.5.41}$$

Матрица, составленная из коэффициентов при неизвестных параметрах в системе уравнений, симметрична относительно главной диагонали, причем все элементы главной диагонали матрицы положительны.

Итак, число нормальных уравнений равно числу неизвестных b_j^*. Решение этой системы позволяет вычислить значения (оценки) неизвестных параметров. Наиболее кратко решение записывается с помощью определителей матрицы коэффициентов

$$b_0^* = \frac{A_0}{A}; \quad b_1^* = \frac{A_1}{A},$$

где A — главный определитель (детерминант) матрицы,

$$A = |\mathbf{A}| = \begin{vmatrix} [x_0 x_0] & [x_0 x_1] \\ [x_1 x_0] & [x_1 x_1] \end{vmatrix} = [x_0 x_0][x_1 x_1] - [x_0 x_1][x_1 x_0]. \tag{3.5.42}$$

Определители (дополнения) A_0, A_1 получаются из главного определителя A путем замены столбца с коэффициентами при b_0^* и b_1^* на столбец со свободными членами $[x_0 y]$ и $[x_1 y]$. В результате вычислений получаем формулы для оценок искомых параметров в обозначениях Гаусса:

$$b_0^* = \frac{[x_0 y][x_1 x_1] - [x_1 y][x_0 x_1]}{[x_0 x_0][x_1 x_1] - [x_0 x_1]^2} =$$

$$= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - b_1^* \bar{x};$$

$$b_1^* = \frac{[x_1 y][x_0 x_0] - [x_0 y][x_0 x_1]}{[x_0 x_0][x_1 x_1] - [x_0 x_1]^2} =$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad (3.5.43)$$

При вычислении статистических характеристик оценок (3.5.43), как и выше, будем полагать, что ошибки определения или задания величин X пренебрежимо малы по сравнению с ошибками определения величин Y. В таком приближении определитель A не зависит от величины Y, он является не случайной величиной. Определители A_j представляют собой линейные функции от случайной величины Y, их средние значения являются несмещенными оценками, так как считается, что $\mathrm{M}\{y_i\} = b_0 x_{0i} + b_1 x_{1i}$. Следовательно, несмещенными будут и оценки неизвестных параметров b_0 и b_1.

Для вычисления дисперсий оценок b_0, b_1

$$D(b_0) = D(A_0)/A^2, \quad D(b_1) = D(A_1)/A^2 \quad (3.5.44)$$

найдем дисперсии определителей A_0 и A_1. Согласно (3.5.34) случайная ошибка величины y_i определяется как $\alpha_i = y_i - \mathrm{M}\{y_i\}$. Тогда

$$D(A_0) = \mathrm{M}\{([x_0\alpha][x_1x_1] - [x_1\alpha][x_0x_1])^2\},$$

Так как выборочные значения y_i между собой независимы и полагается, что условная дисперсия величины α постоянна, т.е. $\mathrm{M}\{\alpha_i^2\} = D_{Y|x} \equiv D$, после элементарных вычислений соответствующих моментов находим выражение для условной дисперсии случайной величины A_0:

$$D(A_0) = D[x_1x_1]([x_0x_0][x_1x_1] - [x_0x_1]^2) =$$

$$= D\sum_{i=1}^{n} x_i^2 \left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Подставляя это соотношение в первую формулу (3.5.44), получим

$$D(b_0) = D\frac{[x_1x_1]}{[x_0x_0][x_1x_1] - [x_0x_1]^2} = D\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \quad (3.5.45)$$

Аналогично определяется дисперсия оценки параметра b_1:

$$D(b_1) = D\frac{[x_0x_0]}{[x_0x_0][x_1x_1] - [x_0x_1]^2} = \frac{D}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \quad (3.5.46)$$

Как правило, условная дисперсия D измеряемой величины Y не известна, поэтому ее приближенно можно заменить на выборочное значение несмещенной оценки

$$s_Y^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0^* x_{0i} - b_1^* x_{1i})^2, \quad n > 2. \quad (3.5.47)$$

В последнем соотношении применение множителя $1/(n-2)$ вместо привычного $1/n$ вызвано тем, что та же выборка была использована и для оценки двух неизвестных коэффициентов b_0 и b_1. Поэтому «эквивалентное» число независимых выборочных значений для получения несмещенной оценки дисперсии в данном случае равно $n-2$.

Пример 3.5.1. Вычислим коэффициенты b_0 и b_1 линейного уравнения (3.5.37) и их дисперсии, пренебрегая ошибками в задании переменной x_1 для $n = 5$. Опытные данные и значение фиктивной величины x_0 приведены в табл. 3.5.2.

Таблица 3.5.2

x_{0i}	1,00	1,00	1,00	1,00	1,00
x_{1i}	0,50	1,00	1,20	1,60	2,00
y_i	2,75	5,65	7,00	8,15	11,35

Оценки параметров b_0 и b_1 определяются по (3.5.43): $[x_0x_0] = \sum_{i=1}^{5} x_{0i}^2 = 5$; $[x_1x_1] = \sum_{i=1}^{5} x_{1i}^2 = 9{,}25$; $[x_0x_1] = \sum_{i=1}^{5} x_{0i}x_{1i} = 6{,}3$; $[yx_0] = \sum_{i=1}^{5} y_i x_{0i} = 34{,}9$; $[yx_1] = \sum_{i=1}^{5} y_i x_{1i} = 51{,}1659$; $[x_0x_0][x_1x_1] - [x_0x_1]^2 = 6{,}56$.

Подставляя эти значения в в (3.5.43), имеем $b_0^* \approx 0{,}075$; $b_1^* \approx 5{,}481$.

Для вычисления дисперсий оценок коэффициентов необходимо найти выборочное значение дисперсии

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{5}[y_i - (b_0^* x_{0i} - b_1^* x_{1i})]^2 \approx 0{,}669/3 \approx 0{,}223.$$

Поэтому $D(b_0^*) = 0{,}223 \cdot 9{,}25 : 6{,}56 \approx 0{,}314$; $D(b_1^*) = 0{,}223 \cdot 5 : 6{,}56 \approx 0{,}169$.

В общем случае линейной зависимости величины Y от $p+1$ независимых переменных $\mathbf{X}(X_0, X_1, \ldots, X_p)$

$$Y = \sum_{j=0}^{p} b_j X_j, \quad j = 0, 1, \ldots, p, \quad (3.5.48)$$

$$= \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - b_1^* \bar{x};$$

$$b_1^* = \frac{[x_1 y][x_0 x_0] - [x_0 y][x_0 x_1]}{[x_0 x_0][x_1 x_1] - [x_0 x_1]^2} =$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \quad (3.5.43)$$

При вычислении статистических характеристик оценок (3.5.43), как и выше, будем полагать, что ошибки определения или задания величин X пренебрежимо малы по сравнению с ошибками определения величин Y. В таком приближении определитель A не зависит от величины Y, он является не случайной величиной. Определители A_j представляют собой линейные функции от случайной величины Y, их средние значения являются несмещенными оценками, так как считается, что $\mathrm{M}\{y_i\} = b_0 x_{0i} + b_1 x_{1i}$. Следовательно, несмещенными будут и оценки неизвестных параметров b_0 и b_1.

Для вычисления дисперсий оценок b_0, b_1

$$D(b_0) = D(A_0)/A^2, \quad D(b_1) = D(A_1)/A^2 \quad (3.5.44)$$

найдем дисперсии определителей A_0 и A_1. Согласно (3.5.34) случайная ошибка величины y_i определяется как $\alpha_i = y_i - \mathrm{M}\{y_i\}$. Тогда

$$D(A_0) = \mathrm{M}\{([x_0\alpha][x_1x_1] - [x_1\alpha][x_0x_1])^2\},$$

Так как выборочные значения y_i между собой независимы и полагается, что условная дисперсия величины α постоянна, т.е. $\mathrm{M}\{\alpha_i^2\} = D_{Y|x} \equiv D$, после элементарных вычислений соответствующих моментов находим выражение для условной дисперсии случайной величины A_0:

$$D(A_0) = D[x_1x_1]([x_0x_0][x_1x_1] - [x_0x_1]^2) =$$

$$= D\sum_{i=1}^{n} x_i^2 \left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right].$$

Подставляя это соотношение в первую формулу (3.5.44), получим

$$D(b_0) = D\frac{[x_1x_1]}{[x_0x_0][x_1x_1] - [x_0x_1]^2} = D\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \quad (3.5.45)$$

Аналогично определяется дисперсия оценки параметра b_1:

$$D(b_1) = D\frac{[x_0x_0]}{[x_0x_0][x_1x_1] - [x_0x_1]^2} = \frac{D}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \quad (3.5.46)$$

Как правило, условная дисперсия D измеряемой величины Y не известна, поэтому ее приближенно можно заменить на выборочное значение несмещенной оценки

$$s_Y^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0^* x_{0i} - b_1^* x_{1i})^2, \quad n > 2. \quad (3.5.47)$$

В последнем соотношении применение множителя $1/(n-2)$ вместо привычного $1/n$ вызвано тем, что та же выборка была использована и для оценки двух неизвестных коэффициентов b_0 и b_1. Поэтому «эквивалентное» число независимых выборочных значений для получения несмещенной оценки дисперсии в данном случае равно $n-2$.

Пример 3.5.1. Вычислим коэффициенты b_0 и b_1 линейного уравнения (3.5.37) и их дисперсии, пренебрегая ошибками в задании переменной x_1 для $n = 5$. Опытные данные и значение фиктивной величины x_0 приведены в табл. 3.5.2.

Таблица 3.5.2

x_{0i}	1,00	1,00	1,00	1,00	1,00
x_{1i}	0,50	1,00	1,20	1,60	2,00
y_i	2,75	5,65	7,00	8,15	11,35

Оценки параметров b_0 и b_1 определяются по (3.5.43): $[x_0x_0] = \sum_{i=1}^{5} x_{0i}^2 = 5$; $[x_1x_1] = \sum_{i=1}^{5} x_{1i}^2 = 9{,}25$; $[x_0x_1] = \sum_{i=1}^{5} x_{0i}x_{1i} = 6{,}3$; $[yx_0] = \sum_{i=1}^{5} y_i x_{0i} = 34{,}9$; $[yx_1] = \sum_{i=1}^{5} y_i x_{1i} = 51{,}1659$; $[x_0x_0][x_1x_1] - [x_0x_1]^2 = 6{,}56$.

Подставляя эти значения в в (3.5.43), имеем $b_0^* \approx 0{,}075$; $b_1^* \approx 5{,}481$.

Для вычисления дисперсий оценок коэффициентов необходимо найти выборочное значение дисперсии

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{5}[y_i - (b_0^* x_{0i} - b_1^* x_{1i})]^2 \approx 0{,}669/3 \approx 0{,}223.$$

Поэтому $D(b_0^*) = 0{,}223 \cdot 9{,}25 : 6{,}56 \approx 0{,}314$; $D(b_1^*) = 0{,}223 \cdot 5 : 6{,}56 \approx 0{,}169$.

В общем случае линейной зависимости величины Y от $p+1$ независимых переменных $\mathbf{X}(X_0, X_1, \ldots, X_p)$

$$Y = \sum_{j=0}^{p} b_j X_j, \quad j = 0, 1, \ldots, p, \quad (3.5.48)$$

и числа опытов $n > p+1$ метод наименьших квадратов приводит к системе из $p+1$ линейных нормальных уравнений относительно оценок неизвестных параметров b_j^*, $j = 0, 1, \ldots, p$. При этом матрица коэффициентов при неизвестных параметрах имеет вид

$$A = \begin{vmatrix} [x_0x_0] & \ldots & [x_0x_p] \\ \vdots & \ddots & \vdots \\ [x_px_0] & \ldots & [x_px_p] \end{vmatrix}. \qquad (3.5.49)$$

Здесь, как и выше $X_0 \equiv x_0 \equiv 1$. Матрица (3.5.49) симметрична относительно главной диагонали, элементы которой являются положительными числами. Оценки неизвестных параметров находятся из выражений

$$b_j^* = A_j/A, \quad j = 0, 1, \ldots, p, \qquad (3.5.50)$$

где $A = |\mathbf{A}|$ — определитель (детерминант) матрицы (3.5.49), а A_j — определитель матрицы, получающейся из (3.5.49) путем замены j-го столбца матрицы на столбец, составленный из элементов $[yx_0]$, $[yx_1]$,..., $[yx_p]$. Как и выше, будем пренебрегать ошибками задания величин X_j и полагать, что оценка случайной величины Y несмещенная с постоянной условной дисперсией D. При этих условиях параметры b_j^* будут несмещенными оценками их истинных значений, а дисперсии оценок определяются из выражений

$$D(b_j) = D(A_{jj}/A) = Dc_{jj}, \quad j = 0, 1, \ldots, p. \qquad (3.5.51)$$

Здесь $c_{jj} = A_{jj}/A$, где A_{jj} — алгебраические дополнения элементов определителя A, получаемые путем удаления из матрицы определителя столбца и строки, на пересечении которых находится элемент $[x_jx_j]$. Для неизвестной условной дисперсии случайной величины Y берется ее выборочное значение

$$s^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} [y_i - (b_0^* + b_1^* x_{1i} + \ldots + b_p^* x_{pi})]^2. \qquad (3.5.52)$$

Полученные соотношения для оценок коэффициентов не зависят от распределения случайных величин X и Y. Поэтому принципиально они не позволяют находить различные вероятностные характеристики (кроме рассмотренных выше) оценок неизвестных параметров уравнений. Применительно к гауссовским случайным величинам ниже в п. 8.2, посвященном выборочной средней квадратичной регрессии, рассматриваются такие вероятностные характеристики.

Контрольные вопросы и задачи

3.1. В чем отличие задач математической статистики от задач теории вероятностей?

3.2. Записать асимптотическую плотность вероятности выборочной медианы распределения Стьюдента (2.2.12) для $n \gg 1$.

3.3. Построить гистограмму и многоугольник распределения для значений случайной величины, приведенных в табл. 5.4.1 на с. 252, применительно к пяти и десяти интервалам группирования одинаковой ширины. В качестве начала первого интервала и конца последнего интервала принять значения 6,4 и 8,4 соответственно.

3.4. На основе сгруппированных данных, приведенных в табл. 5.4.4 на с. 257, вычислите с учетом поправок Шеппарда (п. 3.4) выборочные среднее и дисперсию. Полученные данные сравните с данными, найденными в примере 5.4.1 ($\bar{x} = 7{,}33$; $s^2 = 0{,}21$).

3.5. Вычислить дисперсию выборочной несмещенной дисперсии суммы n независимых экспоненциально распределенных (1.4.20) случайных величин.

3.6. Вычислить дисперсии выборочного среднего и несмещенной выборочной дисперсии при анализе $n = 12$ независимых случайных величин, распределенных по закону Стьюдента (2.2.12).

3.7. На чем основано определение статистических характеристик косвенных нелинейных оценок?

3.8. Определить среднее значение и дисперсию косвенной нелинейной оценки $y = \bar{x}^2$, если среднее значение и дисперсия случайной величины X равны m_X и D_X, а объем выборки n.

3.9. В чем сущность метода наименьших квадратов?

3.10. В результате опыта получено $n = 5$ независимых выборочных значений (x_{1i}, x_{2i}, y_i): (1,0; 0,5; 5,2), (1,3; 1,0; 8,5), (1,5; 1,5; 9,5), (1,7; 2,0; 10,6) и (2,0; 2,5; 12,8). Определить оценки коэффициентов b_1^*, b_2^* линейной модели $y = b_1x_1 + b_2x_2$ и их дисперсии, пренебрегая ошибками в задании значений переменной x.

ГЛАВА 4

Основные понятия теории статистических оценок параметров распределений

4.1. Сущность задачи оценки параметров распределения

В общем виде задача оценки (измерения) параметров распределений может быть сформулирована следующим образом. Пусть наблюдается некоторая независимая, одинаково распределенная выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ объема n. Полагается, что распределение выборки известно с точностью до некоторого векторного параметра $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_q)$. Необходимо решить, какие значения (из заданного интервала возможных значений) принимают неизвестные параметры λ_j, $j = 1, \ldots q$. Другими словами, на основе обработки наблюдаемой выборки $\mathbf{x}$ необходимо произвести измерение, т.е. выработать оценку искомого многомерного параметра распределения $\boldsymbol{\lambda}$.

Оценка параметра распределения — это некоторая определенным образом выбранная система функций от наблюдаемой выборки $\mathbf{x}$. Значения этих функций при фиксированной выборке оценивают, т.е. определяют заданным способом неизвестные параметры распределения. Естественно полагать, что используемые для получения оценки методы при их многократном применении в большинстве случаев приведут к правильному заключению о значении оцениваемых параметров.

Термин «оценка», как это принято, имеет два значения. Во-первых, он используется для обозначения процесса или алгоритма измерения неизвестного параметра, и, во-вторых, оценкой называют измеренное значение неизвестного параметра. Однако такая ситуация в терминологии не приводит к недопониманию, поскольку обычно смысл, в котором употребляется термин «оценка», ясен из контекста.

В зависимости от требований, предъявляемых к процессу оценки и к самим оценкам параметров, возможны разнообразные методы оценивания. При этом каждая оценка характеризуется своим показателем качества, который в большинстве случаев указывает «меру близости» оценки к истинному значению оцениваемого параметра. Показатель качества оценки, в свою очередь, определяется выбором критерия качества оценки или, сокращенно, критерием оценки. Поэтому, прежде чем построить оценку, нужно выбрать критерий оценки.

Выбор критерия оценки зависит от конечной задачи, для которой используется оценка параметра. В связи с тем, что эти задачи могут значительно отличаться, не может быть единого критерия оценки и единственной оценки для данного параметра распределения. В частности, это обстоятельство затрудняет сравнение различных оценок. Ясно, что сравнение различных оценок возможно лишь при использовании одного и того же критерия оценки.

Во многих практических приложениях критерии оценок могут быть получены на основе интуитивных представлений о целевом назначении оценки. Вместе с тем значительный интерес представляет изучение формального подхода к выбору тех или иных критериев, поскольку именно такой подход позволяет яснее понять сущность и особенности проблемы оценок параметров распределений. Необходимо отметить, что к задаче оценки параметров распределений сводятся задачи оценок параметров случайных и детерминированных сигналов при приеме их на фоне помех.

Из-за конечного объема выборки любой оценке (т.е. алгоритму измерения неизвестного параметра или правилу решения) присущи ошибки, определяемые как критерием качества, так и условиями, при которых происходит процесс оценки. Поэтому задача оптимальной оценки параметра $\boldsymbol{\lambda}$ состоит в том, чтобы найти такой алгоритм определения (оценки) параметра $\boldsymbol{\lambda}^*$, при котором для заданного критерия оценки эти ошибки решения (оценки) были бы минимальными. Требование малости ошибок в общем случае не имеет однозначного смысла. Однако, если задан критерий оценки, то на его основе формируется показатель качества оценки, зависящий от ошибок, и задача получения оптимальной оценки сводится к нахождению процедуры решения (алгоритма получения оценки), которая минимизирует (или максимизирует) этот показатель качества. Другими словами, задача построения оптимальных (в соответствии с заранее выбранными критериями) оценивающих устройств состоит в том, чтобы, оперируя определенным образом над имеющимися данными $\mathbf{x}(x_1, x_2, \ldots, x_n)$, получить как можно большую (в смысле заданного критерия) информацию об интересующих наблюдателя параметрах распределения. Интуитивно ясно, что оценка параметра $\boldsymbol{\lambda}^*$ должна быть близка в смысле заданной меры к истинному значению оцениваемого параметра $\boldsymbol{\lambda}$, причем оптимальная оценка в соответствии с выбранным критерием должна минимизировать эту меру близости.

Для упрощения записи и рассуждений в дальнейшем будем полагать, что неизвестным параметром распределения является один параметр λ, хотя выводы, которые будут сделаны, останутся справедливыми (с очевидными изменениями) для совместной оценки нескольких параметров.

Для оценки одного параметра естественно получить одну функцию от наблюдаемой выборки, причем эта функция должна обладать некоторыми «хорошими» свойствами. В общем случае оценка неизвестного параметра может быть функцией от функции, т.е. будет являться функцио-

налом. К подобным ситуациям, например, могут приводить задачи оценки параметров, которые нелинейно связаны с моментами распределений.

Очевидно, что чем более полными сведениями располагает экспериментатор о характеристиках распределения, о возможных значениях оцениваемого параметра λ, тем легче и определеннее будет решаться задача синтеза устройства, обеспечивающего (в соответствии с заданным или выбранным критерием) минимальные ошибки оценки интересующего нас параметра.

Наиболее полные сведения о возможных значениях параметра λ даются апостериорной (послеопытной) вероятностью, если оцениваемый параметр принимает дискретные значения, или апостериорной плотностью вероятности применительно к непрерывному его изменению в некотором интервале (области) Λ, который, если не оговорено особо, простирается от $-\infty$ до $+\infty$. В дальнейшем будем полагать, что оцениваемый параметр на заданном интервале может принимать все возможные значения.

Апостериорная плотность вероятности $f_{\text{ps}}(\lambda)$ является условной плотностью вероятности параметра λ при условии, что вторая случайная величина принимает конкретное значение, совпадающее с выборочными значениями $\mathbf{x}(x_1, x_2, \ldots, x_n)$, т.е.

$$f_{\text{ps}}(\lambda) \equiv f_{\text{ps}}(\lambda; \mathbf{x}) \equiv f_\lambda(\lambda \mid \mathbf{x}).$$

Выражение для апостериорной плотности вероятности оцениваемого параметра λ аналогично (1.1.19) получается из теоремы об условных вероятностях двух случайных величин λ и $\mathbf{x}$:

$$f(\lambda, \mathbf{x}) = f_\lambda(\lambda) f(\mathbf{x} \mid \lambda) = f(\mathbf{x}) f_\lambda(\lambda \mid \mathbf{x}).$$

Отсюда имеем

$$f_\lambda(\lambda \mid \mathbf{x}) = f_\lambda(\lambda) f(\mathbf{x} \mid \lambda) / f(\mathbf{x}). \tag{4.1.1}$$

Здесь $f_\lambda(\lambda) \equiv f_{\text{pr}}(\lambda)$ — *априорная плотность вероятности* оцениваемого параметра λ; $f(\mathbf{x})$ — плотность вероятности многомерной выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$. Плотность вероятности $f(\mathbf{x})$ не зависит от оцениваемого параметра λ и может быть найдена из условия

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} f(\lambda, \mathbf{x})\, d\lambda = \int_{-\infty}^{\infty} f_{\text{pr}}(\lambda) f(\mathbf{x} \mid \lambda)\, d\lambda.$$

Условная плотность вероятности выборки наблюдаемых данных $\mathbf{x}$ (при условии, что оцениваемый параметр имеет значение λ) $f(\mathbf{x} \mid \lambda) = f(x_1, x_2, \ldots, x_n \mid \lambda)$, рассматриваемая как функция от оцениваемого параметра λ при фиксированной выборке $\mathbf{x}$, называется *функцией правдоподобия*. Эта функция при фиксированной выборке $\mathbf{x}$ показывает, насколько одно возможное значение параметра λ «более правдоподобно», чем другое. В таком понимании функция правдоподобия уже не

является плотностью вероятности и при фиксированном значении параметра λ представляет собой некоторое число, зависящее от конкретных выборочных значений. Из сказанного следует, что функция $f(\mathbf{x} \mid \lambda)$, рассматриваемая как функция от $\mathbf{x}$ при фиксированном λ, является условной плотностью вероятности, а рассматриваемая как функция от параметра λ при фиксированном значении $\mathbf{x}$ — функцией правдоподобия и обычно обозначается как $L(\lambda; \mathbf{x})$ или $L(\lambda)$. Функция правдоподобия играет фундаментальную роль в задачах математической статистики и статистической теории принятия решений. Функция правдоподобия для независимой выборки имеет вид

$$L(\lambda; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} L(\lambda; x_i). \tag{4.1.2}$$

Апостериорная плотность вероятности характеризует распределение оцениваемого параметра для конкретной фиксированной выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$. С учетом сказанного апостериорная плотность вероятности может быть записана в виде

$$f_{\text{ps}}(\lambda) = g f_{\text{pr}}(\lambda) L(\lambda; \mathbf{x}), \tag{4.1.3}$$

где коэффициент g, независящий от оцениваемого параметра, определяется из условия нормировки плотности вероятности:

$$\int_{-\infty}^{\infty} g f_{\text{pr}}(\lambda) f(\mathbf{x} \mid \lambda)\, d\lambda = 1.$$

Отсюда

$$g = \left[\int_{-\infty}^{\infty} f_{\text{pr}}(\lambda) f(\mathbf{x} \mid \lambda)\, d\lambda \right]^{-1}.$$

Таким образом, апостериорная плотность вероятности оцениваемого параметра $f_{\text{ps}}(\lambda)$ и функция правдоподобия $L(\lambda; \mathbf{x})$ являются функциями от принятой выборки. Поскольку выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ является конкретной реализацией случайной величины $\mathbf{X}(X_1, X_2, \ldots, X_n)$, то апостериорная плотность вероятности и функция правдоподобия по своей сути представляют собой конкретные значения функций из ансамбля возможных значений, т.е. являются условными функциями параметра.

В теории статистических оценок параметров распределений используют два вида оценок: интервальные (доверительные) оценки и точечные оценки (оценки в точке).

При *интервальной оценке* параметра распределения необходимо указать интервал, в котором с вероятностью, не меньше заданной, содержится истинное значение неизвестного параметра. Эта заданная вероятность называется *доверительной вероятностью*, а указанный интервал возможных значений оцениваемого параметра — *доверительным интервалом*. Верхняя и нижняя границы доверительного интервала, которые

называются *доверительными границами* (пределами), и сам доверительный интервал являются функциями наблюдаемой выборки.

При *точечной оценке* параметра распределения неизвестному параметру приписывают одно значение из интервала возможных его значений, т.е. на основе анализа наблюдаемой выборки $\mathbf{x}$ вырабатывается некоторая величина, которую используют в качестве истинного значения параметра. Применительно к оценке параметров распределения точечные оценки обладают целым рядом полезных свойств, которые обусловили их широкое применение в большом числе прикладных задач.

4.2. Точечные оценки параметров и их свойства

При точечной оценке параметра λ вынести решение (произвести оценку) — это значит каждой из возможных выборок $\mathbf{x}(x_1, x_2, \ldots, x_n)$ поставить в соответствие некоторую величину (функцию) $\lambda^* = \varphi(\mathbf{x})$ из интервала Λ возможных значений оцениваемого параметра. Из-за случайного характера оценки λ^* ее характеризуют *условной плотностью вероятности* $f_\lambda(\lambda^* \mid \lambda)$. Вид этой плотности вероятности определяет качество построения оценки. Полагая выборочные значения случайными, плотность вероятности $f_\lambda(\lambda^* \mid \lambda)$ при заданном правиле оценки $\lambda^* = \varphi(x_1, x_2, \ldots, x_n)$ может быть получена из условной плотности вероятности $f_{nX}(x_1, x_2, \ldots, x_n \mid \lambda)$ случайной величины $\mathbf{X}(X_1, X_2, \ldots, X_n)$ с помощью правила преобразования многомерных случайных величин, рассмотренного в п. 2.1.2.

Для нахождения условной плотности вероятности $f_\lambda(\lambda^* \mid \lambda)$ введем новые переменные

$$y_1 = x_1,\ y_2 = x_2, \ldots,\ y_{n-1} = x_{n-1},\ y_n = \lambda^* = \varphi(x_1, x_2, \ldots, x_n).$$

При этом будем полагать, что существует обратное однозначное преобразование $x_n = \psi(y_1, y_2, \ldots, y_n = \lambda^*)$. Тогда условная многомерная плотность вероятности $f_{nY}(y_1, y_2, \ldots, y_{n-1}, \lambda^* \mid \lambda)$ случайных величин $\mathbf{Y}(Y_1, Y_2, \ldots, Y_n)$ согласно (2.1.19) определяется выражением

$$f_{nY}(y_1, y_2, \ldots, y_{n-1}, \lambda^* \mid \lambda) = \\ = f_{nX}(y_1, y_2, \ldots, y_{n-1}, \varphi(y_1, y_2, \ldots, y_{n-1}, \lambda^*) \mid \lambda)|J_n|.$$

Здесь J_n — якобиан преобразования от переменных $\mathbf{x}(x_1, x_2, \ldots, x_n)$ к новым переменным $\mathbf{y}(y_1, y_2, \ldots, y_n = \lambda^*)$; в соответствии с (2.1.20) он имеет вид

$$J_n = \begin{vmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \partial\psi/\partial y_n \end{vmatrix} = \frac{\partial\psi(y_1, y_2, \ldots, y_{n-1}, y_n)}{\partial y_n}.$$

При этом якобиан преобразования не зависит от истинного значения оцениваемого параметра λ.

Искомая плотность вероятности оценки определяется из соотношения

$$f_\lambda(\lambda^* \mid \lambda) = \int \cdots \int f_{nY}(y_1, y_2, \ldots, y_{n-1}, \lambda^* \mid \lambda)\, dy_1 dy_2 \ldots dy_{n-1}.$$

Следует отметить, что непосредственное нахождение плотности вероятности $f_\lambda(\lambda^* \mid \lambda)$ во многих прикладных задачах весьма затруднительно. Поэтому, если есть основания предполагать, что плотность вероятности $f_\lambda(\lambda^* \mid \lambda)$ унимодальна и близка к симметричной, то в качестве характеристик сгруппированности оценки λ^* относительно истинного значения λ используют широко распространенные условные характеристики:

$$b(\lambda^* \mid \lambda) \equiv b(\lambda) = \mathrm{M}\{(\lambda^* - \lambda)\} = \\ = \int_\Lambda (\lambda^* - \lambda) f_\lambda(\lambda^* \mid \lambda)\, d\lambda^* = \int_{\mathbf{X}} [\varphi(x) - \lambda] f(\mathbf{x} \mid \lambda)\, d\mathbf{x} \qquad (4.2.1)$$

— *условное смещение оценки*;

$$\widehat{D}(\lambda^* \mid \lambda) \equiv \widehat{D}(\lambda) = \qquad (4.2.2)$$

$$= \mathrm{M}\{(\lambda^* - \lambda)^2\} = \int_\Lambda (\lambda^* - \lambda)^2 f_\lambda(\lambda^* \mid \lambda)\, d\lambda = \int_{\mathbf{X}} [\varphi(\mathbf{x}) - \lambda]^2 f(\mathbf{x} \mid \lambda)\, d\mathbf{x}$$

— *условное рассеяние оценки*;

$$D(\lambda^* \mid \lambda) \equiv D(\lambda) = \mathrm{M}\{(\lambda^* - \mathrm{M}\{\lambda^*\})^2\} = \qquad (4.2.3)$$

$$= \int_\Lambda (\lambda^* - \mathrm{M}\{\lambda^*\})^2 f_\lambda(\lambda^* \mid \lambda)\, d\lambda^* = \int_{\mathbf{X}} [\varphi(\mathbf{x}) - \mathrm{M}\{\varphi(\mathbf{x})\}]^2 f(\mathbf{x} \mid \lambda)\, d\mathbf{x}$$

— *условная дисперсия оценки*.

Здесь и далее интегрирование осуществляется по многомерной выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$.

Введенные выше условные характеристики оценки в общем случае могут зависеть от истинного значения оцениваемого параметра распределения λ. Для получения безусловных характеристик оценки их необходимо усреднить по априорной плотности вероятности $f_{\mathrm{pr}}(\lambda)$. В результате получаем *безусловные характеристики оценки*:

$$b_\lambda = \int_\Lambda b(\lambda) f_{\mathrm{pr}}(\lambda)\, d\lambda;$$

$$\widehat{D}_\lambda = \int_\Lambda \widehat{D}(\lambda) f_{\mathrm{pr}}(\lambda)\, d\lambda;$$

$$D_\lambda = \int_\Lambda D(\lambda) f_{\mathrm{pr}}(\lambda)\, d\lambda.$$

Принципиально может существовать множество функций от выборочных значений $\lambda_j^* = \varphi_j(x_1, x_2, \ldots, x_n)$, которые можно предложить в качестве оценок параметра λ. Поэтому возникает вопрос: как наилучшим образом использовать наблюденные данные и априорные сведения о характере закона распределения выборочных данных для получения оценок параметров? В связи с этим сразу возникает второй вопрос: что подразумевать под «наилучшими» оценками? Поскольку нельзя точно предсказать индивидуальное значение параметра в каждом частном случае, то о качестве оценки следует судить лишь по распределению значений оценки, которые обнаруживаются в длинном ряду наблюдений. Иначе говоря, о качестве оценки можно судить лишь на основе анализа выборочного распределения оценки. При этом естественно пытаться построить такую точечную оценку $\lambda^* = \varphi(x_1, x_2, \ldots, x_n)$, чтобы условная плотность вероятности $f_\lambda(\lambda^* \mid \lambda)$ была бы как можно более тесно сгруппирована вокруг значения λ.

Имеется несколько подходов к вопросу о желательных свойствах точечных оценок, из которых основными выделяют следующие свойства, сформулированные в терминах условных характеристик.

1. Весьма желательно, чтобы при выборке достаточно большого объема n ($n \to \infty$) оценка λ^* стремилась к истинному значению λ, т.е. оценка сходилась по вероятности к истинному значению:

$$\lim_{n\to\infty} \mathbf{P}\{|\lambda^* - \lambda| < \varepsilon\} = 1, \tag{4.2.4}$$

где ε — заранее заданное малое положительное число. Такую оценку называют *состоятельной*.

2. Для конечного объема выборки n должно выполняться условие

$$\mathbf{M}\{\lambda^*\} = \lambda, \tag{4.2.5}$$

т.е. чтобы разность между средним значением оценки и истинным значением параметра равнялась нулю:

$$b(\lambda) = \mathbf{M}\{\varphi(x_1, x_2, \ldots, x_n)\} - \lambda = 0.$$

Такую оценку называют *несмещенной*. Если не удается получить несмещенную оценку для конечного объема выборки, в крайнем случае, желательно, чтобы оценка была *асимптотически несмещенной* при $n \to \infty$:

$$\lim_{n\to\infty} b(\lambda) = \lim_{n\to\infty} [\mathbf{M}\{\varphi(x_1, x_2, \ldots, x_n)\} - \lambda] = 0.$$

3. Оценка $\lambda^* = \varphi(x_1, x_2, \ldots, x_n)$ по сравнению с другими возможными оценками должна обладать минимальным рассеянием (минимальным средним квадратом отклонения оценки λ^* относительно истинного значения λ):

$$\widehat{D}(\lambda) = \mathbf{M}\{(\lambda^* - \lambda)^2\} = \min. \tag{4.2.6}$$

Если λ^* оценка несмещенная, то записанное выше выражение определяет требование минимальной дисперсии. Оценки с минимальным рассеянием или минимальной дисперсией называются *эффективными*.

Необходимо заметить, что в литературе по математической статистике под эффективной оценкой часто понимают только несмещенную оценку с минимальной дисперсией. Если не удается получить эффективную оценку для конечного объема выборки, желательно, чтобы оценка была хотя бы *асимптотически эффективной* при $n \to \infty$.

4. Оценка $\lambda^* = \varphi(x_1, x_2, \ldots, x_n)$ должна обладать свойствами достаточности (являться достаточной статистикой). *Статистика* (*оценка*) называется *достаточной*, если в качестве ее понимается функция или функции наблюдаемых данных и все суждения об оцениваемом параметре могут быть вынесены на основании этой статистики без дополнительного обращения к принятым данным (выборочным значениям).

К сожалению, не всегда удается совместить все эти требования. Поэтому окончательный выбор желательных свойств оценки параметров распределения определяется спецификой решаемой задачи. Очевидно, что апостериорная плотность вероятности всегда является достаточной статистикой. Условие достаточности можно сформулировать в терминах функции правдоподобия.

Критерием достаточности оценки является возможность представления функции правдоподобия в виде произведения двух функций:

$$L(\lambda; \mathbf{x}) = h(\mathbf{x}) q(\lambda^* \mid \lambda), \tag{4.2.7}$$

где $h(\mathbf{x})$ — некоторая произвольная функция от выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$, не зависящая от оцениваемого параметра λ. Так как параметр λ не входит в функцию $h(\mathbf{x})$, то ее нельзя использовать для получения информации об оценке λ. Множитель $q(\lambda^* \mid \lambda)$ зависит от выборки $\mathbf{x}$ только через оценку $\lambda^* = \varphi(\mathbf{x})$, так что в $\varphi(\mathbf{x})$ должна содержаться вся информация об оцениваемом параметре λ.

При выполнении условия (4.2.7) апостериорная плотность вероятности (4.1.3) зависит от выборки $\mathbf{x}$ только через оценку $\varphi(\mathbf{x})$.

Рассмотрим несколько подробнее желаемые свойства несмещенности и минимума рассеяния оценки. При этом здесь и далее квадратный корень из рассеяния оценки будем называть *средней квадратической ошибкой* оценки.

Требование несмещенности или минимума смещения оценки тесно связано с требованием минимума дисперсии оценки. Какую оценку предпочесть: несмещенную, но с большей дисперсией, или смещенную, но с меньшей дисперсией, — зависит от того, для каких целей ищется оценка.

Так как смещение оценки является составной частью средней квадратической ошибки, то оно, как правило, может быть приемлемым до тех пор, пока мало по сравнению со средней квадратической ошибкой. Но когда, например, несколько равноценных оценок складываются независимо, смещение остается постоянным, в то время как составляющая

средней квадратической ошибки, обусловленная дисперсией оценки, уменьшается обратно пропорционально квадратному корню из числа независимых оценок. В этом случае средняя квадратическая ошибка полностью определяется смещением оценки. Вместе с тем не исключены случаи, когда за счет увеличения или введения смещения оценки целесообразно уменьшить ее дисперсию. В частности, как показано в п. 3.4.2, оценка дисперсии гауссовской случайной величины с неизвестным средним значением (3.4.10) хотя и является смещенной по сравнению с оценкой (3.4.27), но ее рассеяние (3.4.22) меньше дисперсии несмещенной оценки (3.4.30).

Наконец следует отметить, что в ряде случаев (например, как это рассмотрено в п. 3.4.2 при вычислении характеристик выборочных дисперсий) смещенную оценку можно сделать несмещенной. Так, если смещение оценки является линейной функцией истинного значения параметра $b(\lambda) = \alpha\lambda + \beta$, где α и β — любые действительные числа, и, следовательно, имеет место смещение, то, заменяя оценку λ^* на $\lambda_1^* = (\lambda^* - \beta)/(\alpha + 1)$, получаем несмещенную оценку. Действительно, так как $\mathbf{M}\{\lambda^*\} = \lambda + b(\lambda) = \lambda(\alpha+1) + \beta$, то $\mathbf{M}\{\lambda_1^*\} = [\mathbf{M}\{\lambda^*\} - \beta]/(\alpha+1) = \lambda$, т.е. оценка параметра λ_1^* является несмещенной.

Одним из основных требований при построении оценок является получение оценки с минимальной дисперсией или минимальным рассеянием (если они существуют). В связи с этим, как отмечалось выше, в математической статистике введено понятие эффективных оценок.

Применительно к смещенным оценкам параметров оценка λ_{ef} является эффективной, если среднее значение квадрата отклонения оценки от истинного значения оцениваемого параметра λ не превышает среднее значение квадрата отклонения любой другой оценки λ^*, т.е. выполняется неравенство

$$\widehat{D}_{\text{ef}}(\lambda) = \mathbf{M}\{(\lambda_{\text{ef}} - \lambda)^2\} \leqslant \mathbf{M}\{(\lambda^* - \lambda)^2\}.$$

Для несмещенной оценки $[b(\lambda) = 0]$ рассеяние оценки совпадает с ее дисперсией, и, следовательно, эффективная оценка определяется как оценка с минимальной дисперсией.

Г. Крамер [37] и С. Рао [53] независимо друг от друга получили выражения для нижних границ условных рассеяний и дисперсий оценок, которые являются рассеяниями и дисперсиями эффективных оценок при условии, что таковые существуют для данных параметров.

Приведем вывод этого выражения, полагая, что необходимые допущения справедливы.

Усредним выражение

$$[\varphi(\mathbf{x}) - \lambda]\frac{d}{d\lambda}\ln f(\mathbf{x} \mid \lambda) \equiv [\varphi(\mathbf{x}) - \lambda]\frac{d}{d\lambda}\ln L(\lambda; \mathbf{x}) \qquad (4.2.8)$$

по всевозможным значениям многомерной выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$, которая описывается условной плотностью вероятности $f(\mathbf{x} \mid \lambda)$. Учитывая известное соотношение для производной натурального логарифма $[\ln y]' = y'/y$, после усреднения получаем

$$\int_{\mathbf{X}} [\varphi(\mathbf{x}) - \lambda]\left[\frac{d}{d\lambda}\ln f(\mathbf{x} \mid \lambda)\right] f(\mathbf{x} \mid \lambda)\, d\mathbf{x} =$$
$$= \frac{d}{d\lambda}\int_{\mathbf{X}} \varphi(\mathbf{x}) f(\mathbf{x} \mid \lambda)\, d\mathbf{x} - \lambda\frac{d}{d\lambda}\int_{\mathbf{X}} f(\mathbf{x} \mid \lambda) d\mathbf{x}.$$

В силу свойства нормировки плотности вероятности последнее слагаемое равно нулю. Интеграл от первого слагаемого представляет среднее значение условной оценки

$$\int_{\mathbf{X}} \varphi(\mathbf{x}) f(\mathbf{x} \mid \lambda)\, d\mathbf{x} = \mathbf{M}\{\varphi(\mathbf{x})\} = \lambda + b(\lambda).$$

С учетом последнего выражения и введенной выше функции правдоподобия $L(\lambda; \mathbf{x}) \equiv f(\mathbf{x} \mid \lambda)$ усредненное значение (4.2.8) можно записать в виде

$$\mathbf{M}\{[\varphi(\mathbf{x}) - \lambda]\frac{d}{d\lambda}\ln L(\lambda; \mathbf{x})\} = 1 + \frac{db(\lambda)}{d\lambda}. \qquad (4.2.9)$$

Левая часть этого соотношения представляет собой среднее значение произведения двух случайных величин с конечными значениями первых двух моментов. Отмеченные условия обычно выполняются в прикладных задачах математической статистики. В связи с этим для квадрата математического ожидания произведения двух случайных величин справедливо неравенство Буняковского–Шварца (1.5.33), которое переходит в равенство, если случайные величины связаны между собой детерминированной линейной зависимостью. Поэтому для соотношения (4.2.9) можно записать

$$\left[1 + \frac{db(\lambda)}{d\lambda}\right]^2 \leqslant \widehat{D}(\lambda)\mathbf{M}\left\{\left[\frac{d}{d\lambda}\ln L(\lambda; \mathbf{x})\right]^2\right\},$$

или

$$\widehat{D}(\lambda) \geqslant \frac{[1 + db(\lambda)/d\lambda]^2}{\mathbf{M}\left\{\left[\frac{d}{d\lambda}\ln(\lambda; \mathbf{x})\right]^2\right\}}. \qquad (4.2.10)$$

Для несмещенных оценок и оценок с постоянным смещением рассеяние оценки совпадает с дисперсией оценки параметра и удовлетворяет *неравенству Рао–Крамера*

$$D(\lambda) \geqslant \frac{1}{\mathbf{M}\left\{\left[\frac{d}{d\lambda}\ln L(\lambda; \mathbf{x})\right]^2\right\}}. \qquad (4.2.11)$$

Необходимо еще раз отметить, что во всех соотношениях усреднение производится по многомерной выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$ наблюдаемых данных. Знак равенства в выражениях (4.2.10) и (4.2.11) достигается только для эффективных оценок.

Применительно к выражению (4.2.11) рассмотрим условия, при которых неравенство обращается в равенство, т.е. оценка параметра является эффективной несмещенной оценкой. Согласно неравенству Буняковского–Шварца для этого необходимо, чтобы случайные функции $d \ln L(\lambda; \mathbf{x})/d\lambda$ и $(\lambda^* - \lambda)$ были связаны детерминированной линейной зависимостью. Действительно, представим производную логарифма функции правдоподобия в виде

$$\frac{d}{d\lambda} \ln L(\lambda; \mathbf{x}) = \eta(\lambda)(\lambda^* - \lambda), \tag{4.2.12}$$

где $\eta(\lambda)$ — функция, которая не зависит от оценки λ^* и выборки наблюдаемых данных, но может зависеть от оцениваемого параметра λ (его «истинного значения»). При подстановке (4.2.9) и (4.2.12) в неравенство (4.2.11) оно (при $b(\lambda) = 0$) переходит в равенство. Однако представление производной логарифма функции правдоподобия в виде (4.2.12) возможно, если для оценки $\lambda^* = \varphi(\mathbf{x})$ выполняется *условие достаточности* (4.2.7), из которого следует, что

$$\frac{d}{d\lambda} \ln L(\lambda; \mathbf{x}) = \frac{d}{d\lambda} \ln q(\lambda^* \mid \lambda).$$

Следовательно, если производная логарифма функции правдоподобия линейно зависит от достаточной оценки, то коэффициент пропорциональности $\eta(\lambda)$ не зависит от выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$.

Таким образом, *критерием существования несмещенной эффективной оценки* является выполнение двух условий: оценка должна быть достаточной (4.2.7) и должно выполняться соотношение (4.2.12). Следует подчеркнуть еще раз, что условие (4.2.12) не может быть выполнено, если для оценки не удовлетворяется условие достаточности (4.2.7). Также следует отметить, что эффективная несмещенная оценка обязательно принадлежит к классу достаточных оценок, в то время как достаточная оценка не обязательно будет эффективной.

Аналогичные ограничения налагаются на существование *эффективных смещенных оценок*, при которых в выражении (4.2.10) знак неравенства переходит в равенство.

Полученное выше выражение для нижней границы рассеяния несмещенной оценки справедливо и для нижней границы дисперсии смещенной оценки, т.е.

$$D(\lambda) \geqslant \frac{[1 + db(\lambda)/d\lambda]^2}{\mathrm{M}\left\{\left[\dfrac{d}{d\lambda} \ln L(\lambda; \mathbf{x})\right]^2\right\}}. \tag{4.2.13}$$

Это неравенство переходит в равенство, если кроме условия достаточности оценки справедливо соотношение

$$\frac{d}{d\lambda} \ln L(\lambda; \mathbf{x}) = \eta(\lambda)(\lambda^* - \mathrm{M}\{\lambda^*\}), \tag{4.2.14}$$

где $\eta(\lambda)$ имеет тот же смысл, что и в выражении (4.2.12). Формула (4.2.13) выводится аналогично (4.2.10), если в исходном выражении (4.2.8) вместо $[\varphi(\mathbf{x}) - \lambda]$ рассматривать $[\varphi(\mathbf{x}) - \mathrm{M}\{\lambda^*\}]$. , Хотя правые части формул (4.2.10) и (4.2.13) формально идентичны, однако неравенства переходят в равенства при разных условиях, соответственно при (2.4.12) и (2.4.14).

Анализ выражений для рассеяния и дисперсии эффективных смещенных оценок (4.2.10) и (4.2.13) показывает, что могут существовать смещенные оценки, которые обеспечивают меньшие рассеяние и дисперсию оценки, чем несмещенные. Для этого необходимо, чтобы производная от смещения имела отрицательное значение.

Отметим, что характеристики эффективной оценки могут быть вычислены и для тех параметров, для которых эффективной оценки не существует. В этом случае полученные выше выражения для рассеяния и дисперсии оценки определяют нижнюю границу (недостижимую) для соответствующих характеристик оценки.

Для сравнения реальных оценок с эффективными оценками в математической статистике введено понятие *относительной эффективности оценок*, представляющее собой отношение среднего квадрата отклонения эффективной оценки относительно истинного значения параметра к среднему квадрату отклонения реальной оценки относительно истинного значения параметра:

$$E(\lambda^*) = \frac{\mathrm{M}\{(\lambda_{\mathrm{ef}} - \lambda)^2\}}{\mathrm{M}\{(\lambda^* - \lambda)^2\}} = \frac{D_{\mathrm{ef}}(\lambda)}{D(\lambda)}. \tag{4.2.15}$$

Здесь λ^* — реальная оценка, эффективность которой равна $E(\lambda^*)$; λ_{ef} — эффективная оценка. Из определения рассеяния эффективной оценки видно, что относительная эффективность изменяется в пределах

$$0 \leqslant E(\lambda^*) \leqslant 1.$$

Кроме понятия эффективных оценок, существует понятие *асимптотически эффективных оценок*. Под этим понятием подразумевается, что для достаточно большого объема выборки ($n \to \infty$) предельное значение относительной эффективности реальной оценки равно единице. Это означает, что при асимптотически эффективной оценке рассеяние оценки для заданного смещения определяется выражением (4.2.10) при замене неравенства на равенство.

Форму записи характеристик эффективной оценки можно привести к более удобному виду для прикладных вычислений. Для этого рас-

смотрим вторую производную логарифма функции правдоподобия

$$\frac{d^2}{d\lambda^2}\ln L(\lambda;\mathbf{x}) = \frac{d^2L(\lambda;\mathbf{x})/d\lambda^2}{L(\lambda;\mathbf{x})} - \left[\frac{d}{d\lambda}\ln L(\lambda;\mathbf{x})\right]^2.$$

Усредняя это равенство по выборке наблюдаемых данных $\mathbf{x}(x_1, x_2, \ldots, x_n)$ при фиксированном значении параметра λ, находим

$$\mathrm{M}\left\{\left[\frac{d}{d\lambda}\ln L(\lambda;\mathbf{x})\right]^2\right\} = -\mathrm{M}\left\{\frac{d^2}{d\lambda^2}\ln L(\lambda;\mathbf{x})\right\} + \mathrm{M}\left\{\frac{d^2L(\lambda;\mathbf{x})/d\lambda^2}{L(\lambda;\mathbf{x})}\right\}.$$

Во втором слагаемом правой части данного выражения перейдем от функции правдоподобия $L(\lambda;\mathbf{x})$ к условной плотности вероятности $f(\mathbf{x} \mid \lambda)$ и запишем операцию усреднения по условной плотности вероятности в явном виде. Получим (в силу нормировки)

$$\mathrm{M}\left\{\frac{d^2L(\lambda;\mathbf{x})/d\lambda^2}{L(\lambda;\mathbf{x})}\right\} = \int_{\mathbf{X}} \frac{d^2f(\mathbf{x} \mid \lambda)/d\lambda^2}{f(\mathbf{x} \mid \lambda)} f(\mathbf{x} \mid \lambda)\,d\mathbf{x} = \frac{d^2}{d\lambda^2}\int_{\mathbf{X}} f(\mathbf{x} \mid \lambda)\,d\mathbf{x} = 0.$$

Тогда формулы для дисперсий эффективных оценок можно записать в следующем виде:

- для смещенной оценки

$$D_{\mathrm{ef}}(\lambda) = -\frac{[1 + db(\lambda)/d\lambda]^2}{\mathrm{M}\{d^2\ln L(\lambda;\mathbf{x})/d\lambda^2\}}, \qquad (4.2.16)$$

- для несмещенной оценки

$$D_{\mathrm{ef}}(\lambda) = -\frac{1}{\mathrm{M}\{d^2\ln L(\lambda;\mathbf{x})/d\lambda^2\}}. \qquad (4.2.17)$$

Неотрицательную величину

$$I_n(\lambda) = \mathrm{M}\left\{\left[\frac{d}{d\lambda}\ln L(\lambda;\mathbf{x})\right]^2\right\} = -\mathrm{M}\left\{\frac{d^2}{d\lambda^2}\ln L(\lambda;\mathbf{x})\right\} \qquad (4.2.18)$$

называют *информацией по Фишеру*. Она определяет среднее количество информации об оцениваемом параметре, содержащейся в выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$. Поэтому формулу для дисперсии эффективной несмещенной оценки параметра распределения иногда записывают в виде

$$D_{\mathrm{ef}}(\lambda) = 1/I_n. \qquad (4.2.19)$$

Из этой формулы видно, что дисперсия эффективной оценки параметра обратно пропорциональна среднему количеству информации, содержащейся в выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$ об этом параметре.

Применительно к анализу независимой выборки объема n для функции правдоподобия справедливо соотношение (4.1.2). Полагая, что независимые выборочные значения одинаково распределены, выражения для дисперсий и рассеяний эффективных оценок упрощаются. Так, формула для дисперсии эффективной несмещенной оценки в зависимости от характера изменения случайной величины X принимает следующий вид:

- случайная величина X непрерывна на интервале $(-\infty, \infty)$:

$$D_{\mathrm{ef}}(\lambda) = \left\{n\int_X \left[\frac{\partial\ln L(\lambda;x)}{\partial\lambda}\right]^2 f(x \mid \lambda)dx\right\}^{-1}; \qquad (4.2.20)$$

- случайная величина X принимает ν дискретных значений с условными вероятностями $p(x_j \mid \lambda)$, где $j = 1, 2, \ldots, \nu$:

$$D_{\mathrm{ef}}(\lambda) = \left\{n\sum_{j=1}^{\nu}\left[\frac{d\ln L(\lambda;x_j)}{d\lambda}\right]^2 p(x_j \mid \lambda)\right\}^{-1}. \qquad (4.2.21)$$

В последних двух формулах $L(\lambda;x)$ и $L(\lambda;x_j)$ — функции правдоподобия для одного выборочного значения $x_i \equiv x$ и $x_{ij} \equiv x_j$ соответственно.

Определим связь коэффициента $\eta(\lambda)$ в соотношении (4.2.12), необходимом для существования эффективной оценки, с дисперсией несмещенной эффективной оценки (4.2.11). Для этого вычислим вторые моменты левой и правой части (4.2.12). Имеем

$$\mathrm{M}\{[d\ln L(\lambda;\mathbf{x})/d\lambda]^2\} = \eta^2(\lambda)D_{\mathrm{ef}}(\lambda).$$

Подставляя полученное соотношение в (4.2.11), находим

$$D_{\mathrm{ef}}(\lambda) = 1/\eta(\lambda). \qquad (4.2.22)$$

Таким образом, если оценка параметра является несмещенной, то коэффициент $\eta(\lambda)$ в формуле (4.2.12), определяющей условие существования эффективной оценки, обратно пропорционален дисперсии несмещенной оценки.

Понятия и формулы для эффективной оценки одного параметра обобщаются на совместную оценку нескольких неизвестных параметров [31, 37, 53 и др.]. Необходимые и достаточные условия существования *совместно эффективных оценок* аналогичны условиям существования оценки одного неизвестного параметра (4.2.7) и (4.2.12), в которых вместо λ надо понимать векторный параметр $\boldsymbol{\lambda}$. Аналогичны также понятия асимптотической эффективности оценок.

Применительно к несмещенным совместно эффективным оценкам составляющих векторного параметра $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_q)$ корреляционные моменты между оценками параметров λ_i и λ_j

$$K_{ij}(\lambda_{\mathrm{ef}}) = \mathrm{M}\{(\lambda_i^* - \lambda_i)(\lambda_j^* - \lambda_j)\} = A_{ij}/A. \qquad (4.2.23)$$

Здесь A_{ij} и A — алгебраическое дополнение и определитель матрицы, составленной из элементов

$$a_{ij} = -\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\ln L(\boldsymbol{\lambda};\mathbf{x})\right\}. \qquad (4.2.24)$$

Для совместно эффективной несмещенной оценке двух параметров λ_1 и λ_2 выражение (4.2.23) преобразуется в три формулы:

1) дисперсия оценки параметра λ_1

$$D(\lambda_1) = -\frac{1}{A}\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_2^2}\ln L(\lambda_1,\lambda_2;\mathbf{x})\right\}; \qquad (4.2.25)$$

2) дисперсия оценки параметра λ_2

$$D(\lambda_2) = -\frac{1}{A}\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_1^2}\ln L(\lambda_1,\lambda_2;\mathbf{x})\right\}; \qquad (4.2.26)$$

3) коэффициент взаимной корреляции оценок λ_1, λ_2

$$\rho(\lambda_1,\lambda_2) = \frac{K_{12}(\lambda_1,\lambda_2)}{\sqrt{D(\lambda_1)D(\lambda_2)}} = -\frac{\mathrm{M}\{\partial^2\ln L(\lambda_1,\lambda_2;\mathbf{x})/\partial\lambda_1\partial\lambda_2\}}{\sqrt{D(\lambda_1)D(\lambda_2)}}. \qquad (4.2.27)$$

Здесь

$$A = \mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_1^2}\ln L(\lambda_1,\lambda_2;\mathbf{x})\right\}\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_2^2}\ln L(\lambda_1,\lambda_2;\mathbf{x})\right\} - \left[\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_1\partial\lambda_2}\ln L(\lambda_1,\lambda_2;\mathbf{x})\right\}\right]^2, \qquad (4.2.28)$$

а усреднение осуществляется по всем возможным значениям случайной величины $\mathbf{X}$, т.е. по условной плотности вероятности $f(\mathbf{x}\mid\lambda_1,\lambda_2)$.

Из приведенных выше формул для совместно эффективных оценок видно, что если средние значения смешанных производных логарифма функции правдоподобия (4.2.24) равны нулю, то матрица, составленная из корреляционных моментов оценок, будет диагональной. При этом справедливы соотношения

$$A = \prod_{j=1}^{q} a_{jj}; \quad A_{ii} = \prod_{j=1;\ j\neq i}^{q} a_{jj}.$$

В результате получаем, что оценки параметров между собой независимы, а дисперсии эффективных совместных оценок совпадают с соответствующими дисперсиями эффективных раздельных оценок параметров

$$D_{\mathrm{ef}}(\lambda_i) = K_{ii}(\lambda_i,\lambda_i) = -\left[\mathrm{M}\left\{\frac{\partial^2}{\partial\lambda_i^2}\ln L(\boldsymbol{\lambda};\mathbf{x})\right\}\right]^{-1}, \quad i = 1,2,\ldots,q.$$

При этом необходимо отметить, что дисперсии совместно эффективных оценок при наличии корреляционной связи между оценками всегда будут больше дисперсий раздельных эффективных оценок.

4.3. Основные положения теории байесовских (оптимальных) оценок

4.3.1. Определения

Как отмечалось выше, из-за случайного характера наблюдаемой выборки при любом правиле решения (оценке) неизбежны ошибки в том смысле, что принятое решение (оценка) λ^* не совпадает с истинным значением параметра λ. Очевидно, что при различных правилах выбора решения ошибки различной величины будут появляться с различной вероятностью. При этом в зависимости от цели получения оценки последствия проявления ошибок могут быть различными. Поскольку всегда имеется отличная от нуля вероятность ошибки, необходимо тем или иным образом охарактеризовать качество различных оценок. С этой целью в теории решений введено понятие функции потерь. *Функция потерь* каждой комбинации из решения $\lambda^* = \varphi(\mathbf{x})$ и параметра λ приписывает определенную потерю $C(\lambda^*,\lambda)$. Обычно потери выбирают неотрицательными, а правильным решениям приписывают нулевые или отрицательные потери.

Физический смысл функции потерь состоит в том, что каждой возможной ошибке приписывается определенный неотрицательный вес («стоимость»). При этом в зависимости от цели, для которой находится оценка, наименее желательным ошибкам приписываются наибольшие веса. Выбор той или иной функции потерь производится в зависимости от конкретной специфики задачи, для которой находится оценка, и выходит за рамки теории статистических решений. К сожалению, не существует общего формального выбора функции потерь, и она в той или иной степени является субъективной. Определенный произвол в выборе функции потерь приводит к трудностям в использовании методов статистических решений.

Наиболее часто используются следующие функции потерь (рис. 4.3.1):

а) *простая*

$$C(\lambda^*,\lambda) = c_0 - \delta(\lambda^*-\lambda), \quad c_0 > 0, \qquad (4.3.1)$$

где $\delta(z)$ — дельта-функция (1.2.11);

б) *линейная по модулю*

$$C(\lambda^*,\lambda) = |\lambda^*-\lambda|; \qquad (4.3.2)$$

в) *квадратичная*

$$C(\lambda^*,\lambda) = (\lambda^*-\lambda)^2; \qquad (4.3.3)$$

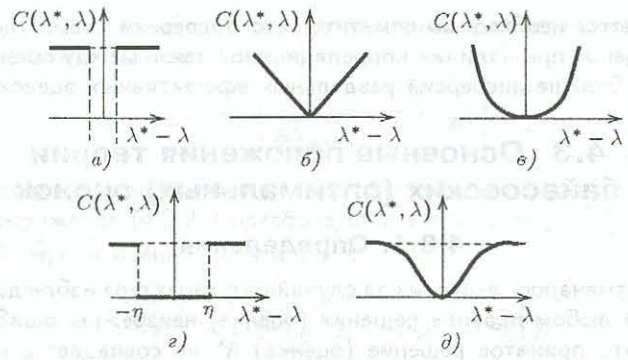


Рис. 4.3.1. Функции потерь: *а* — простая; *б* — линейная; *в* — квадратичная; *г* — прямоугольная; *д* — экспоненциальная

г) *прямоугольная*

$$C(\lambda^*,\lambda)=\begin{cases}0 & \text{при } |\lambda^*-\lambda|<\eta,\\ 1 & \text{при } |\lambda^*-\lambda|\geqslant\eta,\end{cases}\quad \eta>0; \tag{4.3.4}$$

д) *экспоненциальная* (функция потерь с насыщением)

$$C(\lambda^*,\lambda)=1-\exp\left[-\frac{(\lambda^*-\lambda)^2}{2\eta^2}\right]. \tag{4.3.5}$$

Приведенные функции потерь являются симметричными функциями разности $|\lambda^*-\lambda|$. При этом отклонение оценки параметра в одну и другую стороны относительно истинного значения оцениваемого параметра одинаково нежелательны. Вместе с этим не исключены задачи, где для специфических прикладных вопросов могут встречаться случаи, в которых отношение наблюдателя к знаку ошибки разное. Для таких ситуаций функции потерь будут несимметричными.

В силу случайного характера оценок $\lambda^*=\varphi(\mathbf{x})$ и параметра λ потери при любом правиле выбора решения являются случайными и не могут быть использованы для характеристики качества оценки (правила выбора решения). Для характеристики качества оценки можно принять среднее значение функции потерь, которое будет учитывать все возможные типы поведения оценки, все виды ошибок и относительную частоту их появления. Выбор для характеристики качества оценки среднего значения (а не другой статистической характеристики) функции потерь является хотя и произвольным, но рациональным.

Среднее значение (условное или безусловное) функции потерь называется *риском* (условным или безусловным). *Условный средний риск* $R(\lambda^*\mid\lambda)$ получается путем усреднения функции потерь по всевозможным значениям многомерной выборки наблюдаемых данных $\mathbf{x}(x_1,\ldots,x_n)$, характеризуемой условной плотностью вероятности $f(\mathbf{x}\mid\lambda)$:

$$R(\lambda^*\mid\lambda)=\int_{\mathbf{X}}C[\lambda^*=\varphi(\mathbf{x}),\lambda]f(\mathbf{x}\mid\lambda)\,d\mathbf{x}\equiv\int_{\mathbf{X}}C(\lambda^*,\lambda)f(\mathbf{x}\mid\lambda)\,d\mathbf{x}.$$

Из этого выражения и определения функции потерь следует, что более предпочтительными оценками будут те оценки, для которых условный риск минимален. Однако при разных значениях оцениваемого параметра λ условный риск будет иметь различные значения. Поэтому могут быть различными наиболее предпочтительные решающие правила (оценки). Значит, если известно априорное распределение значений оцениваемого параметра, то наилучшее правило выбора решения (оценку) целесообразно искать из условия минимума *безусловного среднего риска*

$$R(\lambda^*)=\int_{\Lambda}\int_{\mathbf{X}}C(\lambda^*,\lambda)f(\mathbf{x}\mid\lambda)f_{\mathrm{pr}}(\lambda)\,d\lambda d\mathbf{x}=$$
$$=\int_{\mathbf{X}}f(\mathbf{x})\left[\int_{\Lambda}C(\lambda^*,\lambda)f_{\mathrm{ps}}(\lambda;\mathbf{x})\,d\lambda\right]d\mathbf{x}, \tag{4.3.6}$$

где $f(\mathbf{x})$ — плотность вероятности выборки наблюдаемых данных.

Оценки, полученные по критерию минимального условного или безусловного (среднего) риска, называются условными и безусловными байесовскими оценками соответственно. Безусловная байесовская оценка часто называется просто байесовской оценкой. В дальнейшем под *байесовской оценкой* λ_m параметра λ будем понимать оценку, обеспечивающую минимальное значение безусловного среднего риска (4.3.6) при заданной функции потерь $C(\lambda^*,\lambda)$. Минимальное значение безусловного среднего риска, соответствующего байесовской оценке, называется *байесовским риском*

$$R_m=\mathbf{M}\left\{\int_{\Lambda}C(\lambda_m,\lambda)f_{\mathrm{ps}}(\lambda;\mathbf{x})\,d\lambda\right\}. \tag{4.3.7}$$

Здесь усреднение выполняется по выборке наблюдаемых данных.

Безусловный средний риск (4.3.6) может быть вычислен для любого заданного правила выбора решения (оценки), причем в силу определения байесовской оценки всегда выполняется условие

$$R_m\equiv R(\lambda_m)\leqslant R(\lambda^*).$$

Вычисляя средний риск (4.3.6) для различных оценок и сравнивая эти риски между собой и байесовским риском, можно судить, насколько одна оценка лучше другой и насколько какая-либо оценка близка к оптимальной (байесовской) оценке.

Так как сам риск (условный или безусловный) имеет различный физический смысл в зависимости от вида и физической интерпретации функции потерь $C(\lambda^*,\lambda)$, то смысл критерия оптимальности также зависит от вида функции потерь.

Поскольку плотность вероятности $f(\mathbf{x})$ есть неотрицательная функция, то минимизация выражения (4.3.6) по λ^* сводится к минимизации функции

$$R_{\text{ps}}(\lambda^*) = \int_{\Lambda} C(\lambda^*, \lambda) f_{\text{ps}}(\lambda; \mathbf{x})\, d\lambda \qquad (4.3.8)$$

при фиксированной выборке наблюдаемых данных. Функция $R_{\text{ps}}(\lambda^*)$ называется *апостериорным риском*. Если апостериорный риск дифференцируем по λ^*, то байесовская оценка λ_m может быть найдена как решение уравнения

$$dR_{\text{ps}}(\lambda^*)/d\lambda^*|_{\lambda^*=\lambda_m} = 0. \qquad (4.3.9)$$

При этом следует брать корень уравнения, обеспечивающий глобальный минимум (наименьший минимум) апостериорного риска.

Критерий минимума среднего риска основан на использовании полной априорной информации об оцениваемом параметре, т.е. он дает ответ на вопрос, каким образом надо использовать всю априорную информацию, чтобы получить наилучшую оценку. Однако отсутствие полной информации об оцениваемом параметре, которое имеет место в ряде прикладных задач, приводит к определенным трудностям (априорные трудности) в применении методов теории статистических решений. Известно несколько подходов к решению задачи нахождения оптимальных оценок при неизвестном априорном распределении оцениваемого параметра. Один из них заключается в отыскании байесовских оценок, инвариантных по отношению к достаточно широкому классу априорных распределений. В других случаях либо ограничиваются выбором оценки на основе минимизации условного риска, либо делают какие-либо предположения относительно априорного распределения оцениваемого параметра.

Если в качестве априорного распределения оцениваемого параметра взято наименее предпочтительное распределение, при котором байесовский риск будет максимален, получаемая оценка параметра называется *минимаксной*. Минимаксная оценка дает оптимальное решение только для самого наихудшего случая и определяет верхнюю границу байесовского риска, которую называют *минимаксным риском*. Хотя минимаксная оценка может приводить к большим потерям, чем другая оценка, она может быть полезной, если желательно застраховаться от потерь за счет оценивания в наиболее неблагоприятных априорных условиях.

В соответствии с определением минимаксная оценка может быть найдена следующим образом. Для произвольного априорного распределения в соответствии с заданной функцией потерь ищется байесовская оценка $\lambda_m = \varphi(\mathbf{x})$. Затем подбирается такое априорное распределение оцениваемого параметра, при котором минимальное значение среднего риска (байесовский риск) достигает максимума. Байесовская оценка при таком априорном распределении будет минимаксной.

Следует отметить, что строгое нахождение наименее предпочтительного априорного распределения сопряжено с большими математическими трудностями. Однако в подавляющем числе прикладных задач наименее предпочтительным распределением является равномерное (в заданном интервале) распределение.

Обсудим кратко свойства байесовских оценок для некоторых функций потерь, приведенных выше.

4.3.2. Простая функция потерь

Подставляя простую функцию потерь (4.3.1) в формулу (4.3.8) и используя фильтрующее свойство дельта-функции (1.2.13), получим выражение для апостериорного риска

$$R_{\text{ps}}(\lambda^*) = c_0 - f_{\text{ps}}(\lambda^*; \mathbf{x}). \qquad (4.3.10)$$

Апостериорный риск $R_{\text{ps}}(\lambda^*)$, а следовательно, и средний риск будут минимальны, если апостериорная плотность вероятности $f_{\text{ps}}(\lambda; \mathbf{x})$ для данной оценки принимает наибольшее значение из всех возможных значений λ. Иначе говоря, апостериорная плотность вероятности обращается в максимум при наличии одного максимума или в максимум максиморум (глобальный или абсолютный максимум) при наличии нескольких максимумов апостериорного распределения оцениваемого параметра. Это означает, что в качестве байесовской оценки параметра распределения должно быть взято наиболее вероятное значение (рис. 4.3.2) оцениваемого параметра, при котором выполняется условие

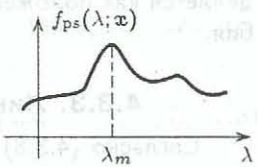


Рис. 4.3.2. Байесовская оценка параметра при простой функции потерь

$$f_{\text{ps}}(\lambda_m; \mathbf{x}) \geqslant f_{\text{ps}}(\lambda; \mathbf{x}), \quad \lambda_m, \lambda \in \Lambda. \qquad (4.3.11)$$

Если апостериорная плотность вероятности дифференцируема по параметру λ, то оценка $\lambda_m = \varphi(\mathbf{x})$ может быть найдена из решения уравнения

$$\left[\frac{df_{\text{ps}}(\lambda; \mathbf{x})}{d\lambda}\right]_{\lambda_m} = 0; \quad \left[\frac{d^2 f_{\text{ps}}(\lambda; \mathbf{x})}{d\lambda^2}\right]_{\lambda_m} < 0, \qquad (4.3.12)$$

причем из всех корней уравнения берется тот корень, для которого справедливо соотношение (4.3.11). Подставляя простую функцию потерь в (4.3.7), получаем выражение для байесовского риска:

$$R_m = c_0 - \mathrm{M}\{\max f_{\text{ps}}(\lambda; \mathbf{x})\} = c_0 - \mathrm{M}\{f_{\text{ps}}(\lambda_m; \mathbf{x})\}. \qquad (4.3.13)$$

Второй член в правой части этого выражения представляет собой (с точностью до некоторого постоянного коэффициента) среднюю вероятность правильного решения. Следовательно, байесовская оценка при простой функции потерь максимизирует вероятность правильного решения. Байесовский риск в данном случае пропорционален вероятности

неправильного решения. Поэтому байесовская оценка параметра при простой функции потерь минимизирует вероятность неправильного решения. При этом всем ошибкам приписывается одинаковый вес c_0, т.е. предполагается, что все ошибки нежелательны независимо от их значения.

В литературе байесовская оценка при простой функции потерь известна также под названием оценки по *максимуму* (*максимуму максиморуму*) *апостериорной плотности вероятности*.

Если априорная плотность вероятности $f_{\rm pr}(\lambda)$ постоянна в интервале возможных значений оцениваемого параметра, то согласно (4.1.3) апостериорная плотность вероятности с точностью до постоянного множителя совпадает с функцией правдоподобия $L(\lambda;\mathbf{x})$. Оценка по максимуму апостериорной плотности вероятности переходит в оценку максимального правдоподобия. *Оценка максимального правдоподобия* определяется как положение наибольшего максимума функции правдоподобия.

4.3.3. Линейная по модулю функция потерь

Согласно (4.3.8) апостериорный риск для функции потерь (4.3.2) определяется выражением

$$R_{\rm ps}(\lambda^*)=\int_{\Lambda}|\lambda^*-\lambda|f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda.$$

Освободимся от модуля в подынтегральном выражении, для чего интервал интегрирования $(-\infty,\infty)$ разобьем на два подынтервала: $-\infty<\lambda<\lambda^*$ и $\lambda^*<\lambda<\infty$, в каждом из которых можно отбросить знак модуля. Тогда предыдущее выражение перепишется в виде

$$R_{\rm ps}(\lambda^*)=\int_{-\infty}^{\lambda^*}(\lambda^*-\lambda)f_{\rm ps}(\lambda;\mathbf{x})\,dx-\int_{\lambda^*}^{\infty}(\lambda^*-\lambda)f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda.$$

Из условия экстремума функции $R_{\rm ps}(\lambda^*)$ получаем уравнение

$$\left[\frac{dR_{\rm ps}(\lambda^*)}{d\lambda^*}\right]_{\lambda_m}=\int_{-\infty}^{\lambda_m}f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda-\int_{\lambda_m}^{\infty}f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda=0,$$

или

$$\int_{-\infty}^{\lambda_m}f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda=\int_{\lambda_m}^{\infty}f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda=\frac{1}{2}. \qquad (4.3.14)$$

При этом для $\lambda^*=\lambda_m$ величина $\dfrac{d^2R_{\rm ps}(\lambda^*)}{d\lambda^{*2}}=2f_{\rm ps}(\lambda_m)>0$, т.е. λ_m соответствует минимуму апостериорного риска $R_{\rm ps}(\lambda^*)$. Из выражения (4.3.14) видно, что в качестве оценки параметра берется то его значение, при котором площади под кривой апостериорной плотности

вероятности слева и справа равны, т.е. байесовская оценка при данной функции потерь представляет собой медиану апостериорного распределения (рис. 4.3.3).

Анализируя выражение байесовского риска при линейной функции потерь

$$R_m=\int_{\mathbf{x}}\int_{-\infty}^{\infty}|\lambda_m-\lambda|f_{\rm ps}(\lambda;\mathbf{x})f(\mathbf{x})\,d\lambda d\mathbf{x}, \qquad (4.3.15)$$

видим, что байесовский риск равен минимальному среднему значению модуля отклонения медианы апостериорного распределения от истинного значения оцениваемого параметра.

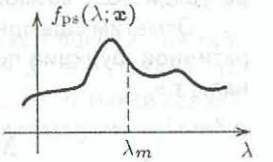


Рис. 4.3.3. Байесовская оценка параметра при линейной функции потерь

4.3.4. Квадратичная функция потерь

Для квадратичной функции потерь (4.3.3) апостериорный риск согласно (4.3.8)

$$R_{\rm ps}(\lambda^*)=\int_{-\infty}^{\infty}(\lambda^*-\lambda)^2f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda. \qquad (4.3.16)$$

Из условия экстремума функции $R_{\rm ps}(\lambda^*)$ получаем выражение для оценки

$$\lambda_m=\int_{-\infty}^{\infty}\lambda f_{\rm ps}(\lambda;\mathbf{x})\,d\lambda\equiv\lambda_{\rm ps}. \qquad (4.3.17)$$

Следовательно, в качестве оценки параметра λ_m при квадратичной функции потерь следует брать среднее значение апостериорного распределения («центр тяжести» апостериорного распределения) $\lambda_m=\lambda_{\rm ps}$ (рис. 4.3.4).

Апостериорный риск зависит от конкретной выборки и в связи с этим, как указывалось выше, является случайной величиной. Поэтому байесовский риск при квадратичной функции потерь совпадает с безусловным рассеянием оценки:

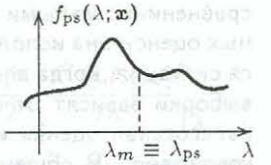


Рис. 4.3.4. Байесовская оценка параметра при квадратичной функции потерь

$$R_m=\widehat{D}_\lambda=\int_{\mathbf{x}}\int_{-\infty}^{\infty}(\lambda_m-\lambda)^2f_{\rm ps}(\lambda;\mathbf{x})f(\mathbf{x})\,d\lambda d\mathbf{x}=$$
$$=\int_{\mathbf{x}}\int_{-\infty}^{\infty}(\lambda_m-\lambda)^2f_{\rm pr}(\lambda)f(\mathbf{x}\mid\lambda)d\lambda d\mathbf{x}.$$

Оценка (4.3.17) находилась из условия минимума среднего риска, который пропорционален среднему квадрату ошибки. Это означает, что байесовская оценка при квадратичной функции потерь минимизирует безусловное рассеяние оценки. Если $\mathbf{M}\{\lambda_m\}=\lambda$ — условно несмещенная байесовская оценка, то при квадратичной функции потерь она

минимизирует дисперсию оценки. Иначе говоря, байесовская оценка при квадратичной функции потерь обеспечивает минимальное значение рассеяния оценки относительно истинного значения оцениваемого параметра среди всех возможных оценок параметра.

Отметим еще одно свойство байесовской оценки параметра при квадратичной функции потерь. Оценка (4.3.17) всегда безусловно несмещенная, т.е.

$$\mathrm{M}\{\lambda_{\mathrm{ps}}\} = \int_{-\infty}^{\infty} \lambda f_{\mathrm{pr}}(\lambda)\, d\lambda = \lambda_{\mathrm{pr}}. \qquad (4.3.18)$$

Действительно, подставляя в (4.3.17) значение апостериорной плотности вероятности и усредняя это выражение по выборке наблюдаемых данных $\mathbf{x}(x_1, x_2, \ldots, x_n)$, можно записать

$$\mathrm{M}\{\lambda_{\mathrm{ps}}\} = \int_{\mathbf{x}} \lambda_{\mathrm{ps}} f(\mathbf{x})\, d\mathbf{x} = \int_{\mathbf{x}} \int_{-\infty}^{\infty} \lambda f_{\mathrm{pr}}(\lambda) f(\mathbf{x} \mid \lambda)\, d\lambda d\mathbf{x}.$$

Меняя порядок интегрирования в последней части этого равенства и учитывая условие нормировки, приходим к выражению $\mathrm{M}\{\lambda_m\} = \lambda_{\mathrm{pr}}$.

Квадратичная функция потерь пропорциональна квадрату отклонения оценки от истинного значения параметра, т.е. всем ошибкам приписывается вес, возрастающий как квадрат их значений. Такого рода потери часто встречаются в различных приложениях математической статистики. Однако несмотря на то, что квадратичная функция потерь обладает рядом достоинств (удобна с математической точки зрения, в достаточной степени учитывает большее значение больших ошибок по сравнению с малыми ошибками и т.д.), в прикладных задачах оптимальных оценок она используется сравнительно редко. Исключением является ситуация, когда апостериорное распределение нормально, а элементы выборки зависят от оцениваемого параметра линейно. В этом случае байесовская оценка и оценка по методу минимума квадрата (п. 3.5.5) совпадают. В общем виде оцениваемый параметр связан нелинейно с анализируемыми выборочными значениями, а сама выборка имеет не гауссовское распределение. Поэтому для большинства реальных параметров распределений устройство получения оценки (4.3.17) оказывается достаточно сложным при его практическом осуществлении.

4.3.5. Прямоугольная функция потерь

Для прямоугольной функции потерь (4.3.4) характерно, что все ошибки, которые по модулю меньше заданного значения η, одинаково не опасны для наблюдателя и с его точки зрения не приводят к какому-либо ухудшению качества оценки параметра. Ошибки, мгновенные значения которых по модулю превышают значения η, одинаково нежелательны, и всем им приписывается одинаковый вес.

Подставляя (4.3.4) в (4.3.8) и разбивая интервал интегрирования на три подынтервала $(-\infty < \lambda < \lambda^* - \eta;\ \lambda^* - \eta \leqslant \lambda \leqslant \lambda^* + \eta$ и

$\lambda^* + \eta < \lambda < \infty)$, имеем

$$R_{\mathrm{ps}}(\lambda^*; \mathbf{x}) = 1 - \int_{\lambda^*-\eta}^{\lambda^*+\eta} f_{\mathrm{ps}}(\lambda; \mathbf{x}) d\lambda. \qquad (4.3.19)$$

Из этого выражения видно, что в качестве байесовской оценки при прямоугольной функции потерь надо выбирать значение $\lambda^* = \lambda_m$, для которого вероятность выполнения неравенства $|\lambda^* - \lambda_m| \leqslant \eta$ максимальна. Уравнение для оценки λ_m находим из условия экстремума апостериорного риска $R_{\mathrm{ps}}(\lambda^*)$:

$$\left[\frac{R_{\mathrm{ps}}(\lambda^*; \mathbf{x})}{d\lambda^*}\right]_{\lambda_m} = f_{\mathrm{ps}}(\lambda_m - \eta; \mathbf{x}) - f_{\mathrm{ps}}(\lambda_m + \eta; \mathbf{x}) = 0,$$

или

$$f_{\mathrm{ps}}(\lambda_m - \eta; \mathbf{x}) = f_{\mathrm{ps}}(\lambda_m + \eta; \mathbf{x}). \qquad (4.3.20)$$

Уравнение (4.3.20) показывает, что в качестве байесовской оценки параметра необходимо взять то его значение λ_m, при котором значения апостериорной плотности вероятности, соответствующие параметрам, отстоящим от оценки слева и справа на величину η, равны между собой (рис. 4.3.5).

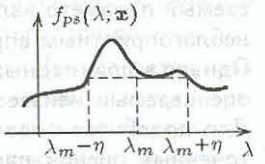


Рис. 4.3.5. Байесовская оценка параметра при прямоугольной функции потерь

Если апостериорная плотность вероятности $f_{\mathrm{ps}}(\lambda; \mathbf{x})$ дифференцируема, то для малых значений η можно ограничиться первыми двумя членами разложения в ряды Тейлора (относительно точки λ_m) апостериорных плотностей вероятностей, входящих в выражение (4.3.20):

$$f_{\mathrm{ps}}(\lambda_m \pm \eta; \mathbf{x}) \approx f_{\mathrm{ps}}(\lambda_m; \mathbf{x}) \pm \eta \left[\frac{df_{\mathrm{ps}}(\lambda; \mathbf{x})}{d\lambda}\right]_{\lambda_m} + \frac{\eta^2}{2}\left[\frac{d^2 f_{\mathrm{ps}}(\lambda; \mathbf{x})}{d\lambda^2}\right]_{\lambda_m}.$$

Подставляя последнее соотношение в (4.3.20), получаем уравнение

$$\left[\frac{df_{\mathrm{ps}}(\lambda; \mathbf{x})}{d\lambda}\right]_{\lambda_m} = 0,$$

которое совпадает с уравнением (4.3.12). Следовательно, при малых величинах «зоны нечувствительности» η байесовские оценки при простой и прямоугольной функциях потерь совпадают.

Байесовский риск (4.3.7) для прямоугольной функции потерь

$$R_m = 1 - \int_{\mathbf{x}} \left[\int_{\lambda_m-\eta}^{\lambda_m+\eta} f_{\mathrm{ps}}(\lambda; \mathbf{x})\, d\lambda\right] f(\mathbf{x})\, d\mathbf{x}. \qquad (4.3.21)$$

Величина (4.3.19) представляет собой апостериорную вероятность того, что в данной реализации (выборке) истинное значение оцениваемого параметра не заключено в интервале $\lambda_m \pm \eta$, т.е. случайная величина

$R_{\text{ps}}(\lambda^*)$ определяет апостериорную вероятность неправильного решения по данному критерию. При этом для байесовской оценки λ_m эта вероятность меньше, чем для любой другой оценки. Байесовский риск (4.3.21) определяет минимальную среднюю вероятность неправильного решения по данному критерию.

Таким образом, байесовская оценка при прямоугольной функции потерь максимизирует вероятность того, что истинное значение оцениваемого параметра будет лежать в пределах $\lambda_m \pm \eta$.

4.4. Прикладные методы получения точечных оценок параметров распределения

Как отмечалось в предыдущем параграфе, для нахождения оптимальных оценок параметров распределений предполагается, что оцениваемый параметр является случайным с точно известным или наиболее неблагоприятным априорным распределением оцениваемого параметра. Однако в прикладных задачах математической статистики, как правило, оцениваемый неизвестный параметр является неслучайной величиной. Для подобного рода задач выделяют три основных метода получения точечных оценок параметров распределений: метод моментов, метод максимального правдоподобия и метод минимума хи-квадрат (χ^2). Рассмотрим основные положения этих методов оценки и определим правила решения и статистические характеристики оцениваемых параметров.

4.4.1. Метод моментов

Пусть наблюдаются независимые однородные выборочные значения x_i, $i = 1, 2, \ldots, n$, непрерывной случайной величины X с условной плотностью вероятности $f(x \mid \lambda_1, \lambda_2, \ldots, \lambda_k)$, зависящей от k неизвестных параметров λ_j, $j = 1, 2, \ldots, k$. От этих же параметров будет зависеть и теоретические начальные моменты порядка q

$$m_q(\lambda_1, \lambda_2, \ldots, \lambda_k) = \int x^q f(x \mid \lambda_1, \lambda_2, \ldots, \lambda_k)\, dx. \qquad (4.4.1)$$

На основе наблюдаемой выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ формируются выборочные начальные моменты того же порядка q

$$m_q^* = \frac{1}{n}\sum_{i=1}^{n} x_i^q. \qquad (4.4.2)$$

Метод моментов основан на приравнивании теоретических начальных моментов выборочным начальным моментам одного и того же порядка. Естественно, что число сравниваемых моментов должно быть равно числу неизвестных параметров распределения, т.е. $q = 1, 2, \ldots, k$. Приравнивая полученные оценки начальных моментов m_q^* к теоретическим величинам $m_q(\lambda_1, \lambda_2, \ldots, \lambda_k)$, получаем соответствующую систему уравнений

$$m_q^* = m_q(\lambda_1, \lambda_2, \ldots, \lambda_k), \quad q = 1, 2, \ldots, k. \qquad (4.4.3)$$

Решая эту систему уравнений относительно неизвестных параметров, находим оценки λ_j^*, $j = 1, 2, \ldots, k$, как функции от выборочных начальных моментов m_q^*, а следовательно, от выборочных значений $\mathbf{x}(x_1, x_2, \ldots, x_n)$ случайной величины X:

$$\lambda_j^* = \psi_j(m_1^*, m_2^*, \ldots, m_k^*) = \varphi_j(x_1, x_2, \ldots, x_n), \quad j = 1, 2, \ldots, k. \qquad (4.4.4)$$

Метод моментов для оценки параметров распределений впервые был предложен К. Пирсоном в конце XIX века.

Преимуществом оценок параметров методом моментов по сравнению с другими методами является достаточно простые вычисления алгоритмов оценки параметров. Однако могут возникнуть математические сложности при определении статистических характеристик оценки.

В литературе по математической статистике [37] отмечается, что если существует единственное решение (4.4.4) и функции ψ_j и φ_j непрерывны, то получающаяся оценка λ_j^* является состоятельной оценкой параметра λ_j, хотя в общем случае оценки λ_j^* не являются эффективными. Состоятельность оценки параметров по методу моментов вытекает из состоятельности выборочных моментов

$$\lim_{n\to\infty} \mathbf{P}\{|m_q^* - m_q| < \varepsilon\} = 1, \qquad (4.4.5)$$

где ε — заранее заданное малое положительное число. Соотношение (4.4.5) доказывается аналогично предельному выражению (2.4.7).

Р. Фишер показал, что, как правило, асимптотическая эффективность оценки параметра по методу моментов значительно меньше единицы [37]. Кроме того рассматриваемой оценке присуще смещение оценки, которое принципиально можно исключить введением соответствующих поправок в виде множителя к выборочному начальному или центральному моментам.

4.4.2. Метод максимального правдоподобия

Метод максимального правдоподобия для оценки параметров распределений впервые был предложен и обоснован Р. Фишером в 1912 г. Рассмотрим сущность метода применительно к векторному параметру $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_k)$. Метод максимального правдоподобия оценки параметров заключается в том, что при фиксированной выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$ в качестве оценки неизвестного векторного параметра $\boldsymbol{\lambda}$ принимается такое его значение $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$, при котором функция правдоподобия $L(\boldsymbol{\lambda}; \mathbf{x})$ или ее натуральный логарифм $\ln L(\boldsymbol{\lambda}; \mathbf{x})$ достигают максиму-

ма. Поэтому для нахождения оценки решают *уравнения правдоподобия*

$$\left.\frac{\partial \ln L(\boldsymbol{\lambda};\mathbf{x})}{\partial \lambda_j}\right|_{\lambda_j^*} = 0, \quad j = 1,2,\ldots,k. \tag{4.4.6}$$

При решении уравнений правдоподобия следует отбросить решения в виде $\lambda_j^* = \text{const}$ и рассматривать лишь те уравнения, которые зависят от выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ и попадают в область допустимых значений параметра $\boldsymbol{\lambda}$. Каждое решение уравнений правдоподобия λ_j^*, зависящее от выборки, называется *оценкой максимального правдоподобия* параметра λ_j, $j = 1,2,\ldots,k$.

Выше были определены условия существования эффективных оценок в смысле оценок с минимальным рассеянием. Покажем, применительно к одному оцениваемому параметру λ, что, если условия (4.2.7) и (4.2.12) выполняются, эффективная оценка может быть получена методом максимального правдоподобия.

Обозначим оценку максимального правдоподобия как λ_m. Пусть выполняется условие достаточности оценки (4.2.7). Тогда уравнение правдоподобия примет вид

$$\frac{d \ln L(\lambda;\mathbf{x})}{d\lambda} = \frac{d}{d\lambda} q(\lambda^* \mid \lambda) = 0. \tag{4.4.7}$$

Поскольку условная функция $q(\lambda^* \mid \lambda)$ зависит только от двух величин λ^* и λ, то каждое решение уравнения правдоподобия будет детерминированной функцией от оценки, т.е.

$$\lambda_m = \zeta(\lambda^*). \tag{4.4.8}$$

Пусть теперь выполняется условие (4.2.12), которое применительно к оценке максимального правдоподобия с учетом соотношения (4.2.7) записывается как

$$\left.\frac{d \ln L(\lambda;\mathbf{x})}{d\lambda}\right|_{\lambda_m} = \left.\frac{d \ln q(\lambda^* \mid \lambda)}{d\lambda}\right|_{\lambda_m} = \eta(\lambda_m)(\lambda^* - \lambda_m) = 0.$$

Отсюда получаем, что оценка максимального правдоподобия равна эффективной оценке:

$$\lambda_m = \lambda^*. \tag{4.4.9}$$

Рассмотренные свойства оценки максимального правдоподобия одного параметра распространяются на совокупность совместно эффективных оценок, если таковые существуют.

Необходимо отметить, что метод максимального правдоподобия, как правило, приводит к менее сложным вычислениям по сравнению с методом моментов. Метод максимального правдоподобия, кроме отмеченных выше достоинств, связанных с получением эффективных оценок, оказывается особенно предпочтительным в случае малого объема выборки. Важным качеством оценок максимального правдоподобия является

также то, что при достаточно общих условиях уравнения правдоподобия имеют решения, дающие состоятельные, асимптотически нормальные и асимптотически эффективные оценки [37]. Кроме этого, оценки параметров по методу максимального правдоподобия не зависят от взаимно однозначного безынерционного (по оцениваемому параметру) нелинейного преобразования вида $\Psi[L(\lambda;\mathbf{x})]$, так как точка максимального правдоподобия остается инвариантной при этих преобразованиях. Действительно, уравнение

$$\left.\frac{d\Psi}{d\lambda}\right|_{\lambda_m} = \left[\frac{\partial \Psi}{\partial L}\frac{\partial L}{\partial \lambda}\right]_{\lambda_m} = 0$$

эквивалентно уравнению $dL(\lambda;\mathbf{x})/d\lambda = 0$ при $\lambda = \lambda_m$. Это свойство имеет большое значение при практической реализации алгоритма оценки параметра распределения.

4.4.3. Метод минимума хи-квадрат (χ^2)

Рассмотрим оценку параметра распределения по методу минимума хи-квадрат (минимума χ^2) применительно к анализу выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ непрерывной случайной величины X. Итак, пусть задана условная плотность вероятности $f(\mathbf{x} \mid \lambda)$ с точностью до одного неизвестного параметра λ. Сгруппируем выборку по k неперекрывающимся интервалам (x_{j-1}, x_j), где $j = 1,2,\ldots,k$. Обозначим через n_j^* число наблюдений в выборке, попавших в интервал (x_{j-1}, x_j). Теоретически в этот интервал должно попасть $np_j(\lambda)$ элементов выборки, где

$$p_j(\lambda) = \int_{x_{j-1}}^{x_j} f(x \mid \lambda)\, dx \tag{4.4.10}$$

— вероятность попадания случайной величины X в интервал (x_{j-1}, x_j), зависящая от значения оцениваемого параметра λ. На основании сгруппированности выборки и теоретического значения $np_j(\lambda)$ составим величину

$$\chi^2 = \sum_{j=1}^{k} \frac{1}{np_j(\lambda)}[n_j^* - np_j(\lambda)]^2, \tag{4.4.11}$$

которая при известном значении параметра λ, конечном объеме выборки n и числе интервалов k приближенно распределена по закону хи-квадрат с $\nu = (k-1)$ степенями свободы. Если параметр λ неизвестен (как это имеет место в рассматриваемом случае) и истинное значение параметра λ заменяется его оценкой λ^*, полученной на основе той же выборки, то число степеней свободы равно $\nu = (k-2)$. Положительная величина χ^2 будет минимальной тогда, когда значение оцениваемого параметра λ близко к истинному значению, от которого зависит сгруппированная выборка n_j^*, $j = 1,2,\ldots,k$.

Оценка параметра λ^* называется *оценкой по методу минимума хи-квадрат*, если она получается минимизацией по λ величины χ^2 (4.4.11). При этом необходимо решить уравнение $[d\chi^2/d\lambda] = 0$ при $\lambda = \lambda^*$, которое преобразуется к виду

$$\sum_{j=1}^{k}\left[\frac{n_j^* - np_j(\lambda)}{p_j(\lambda)} + \frac{(n_j^* - np_j(\lambda))^2}{2np_j^2(\lambda)}\right]\frac{dp_j(\lambda)}{d\lambda}\bigg|_{\lambda^*} = 0. \qquad (4.4.12)$$

Применительно к векторному параметру $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ для нахождения оценок параметров $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*)$ решается система уравнений

$$\sum_{j=1}^{k}\left[\frac{n_j^* - np_j(\boldsymbol{\lambda})}{p_j(\boldsymbol{\lambda})} + \frac{n_j^* - np_j(\boldsymbol{\lambda})}{2np_j^2(\boldsymbol{\lambda})}\right]\frac{\partial p_j(\boldsymbol{\lambda})}{\partial \lambda_q}\bigg|_{\boldsymbol{\lambda}^*} = 0, \quad q = 1, 2, \ldots, r. \qquad (4.4.13)$$

Уравнение (4.4.12), а тем более систему уравнений (4.4.13) часто очень трудно решить. Однако при достаточно большом объеме выборки n влиянием второго члена в квадратных скобках можно пренебречь. В этом случае пренебрежение вторым членом в квадратных скобках эквивалентно тому, что в (4.4.11) знаменатель полагается постоянным, т.е. уравнение (4.4.12) заменяется уравнением вида

$$\sum_{j=1}^{k}\frac{n_j^* - np_j(\lambda)}{p_j(\lambda)}\frac{dp_j(\lambda)}{d\lambda}\bigg|_{\lambda^*} = 0.$$

Учитывая условие нормировки $\sum p_j = 1$, последнее выражение приводится к виду

$$\sum_{j=1}^{k}\frac{n_j^*}{p_j(\lambda)}\frac{dp_j(\lambda)}{d\lambda}\bigg|_{\lambda^*} = 0. \qquad (4.4.14)$$

Аналогично изменяется и система уравнений (4.4.13)

$$\sum_{j=1}^{k}\frac{n_j^*}{p_j(\boldsymbol{\lambda})}\frac{\partial p_j(\boldsymbol{\lambda})}{\partial \lambda_q}\bigg|_{\boldsymbol{\lambda}^*} = 0, \quad q = 1, 2, \ldots, r. \qquad (4.4.15)$$

Метод оценки параметров, получаемый на основе решения уравнений (4.4.14) и (4.4.15), называется *видоизмененным методом хи-квадрат*. В [37] доказывается, что в данном приближении при достаточно общих условиях, как правило, выполняемых в прикладных задачах математической статистики, уравнения (4.4.14) и (4.4.15) имеют одну систему решений $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*)$, такую, что оценка $\boldsymbol{\lambda}^*$ сходится по вероятности к истинному значению $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ при $n \to \infty$. Величина χ^2, получающаяся при подстановке $\boldsymbol{\lambda}^*$ в исходное выражение (4.4.13), в пределе при $n \to \infty$ имеет распределение хи-квадрат (2.2.3) с $\nu = (k - r - 1)$ степенями свободы. Кроме того, доказывается, что оценки по минимуму хи-квадрат асимптотически нормальны и асимптотически эффективны.

Функция правдоподобия параметра λ сгруппированных выборочных значений имеет вид

$$L(\lambda; \mathbf{x}) = p_1^{n_1^*} p_2^{n_2^*} \cdots p_k^{n_k^*} = \prod_{j=1}^{k} p_j^{n_j^*}(\lambda).$$

Отсюда получаем выражение

$$\frac{d\ln L(\lambda; \mathbf{x})}{d\lambda}\bigg|_{\lambda^*} = \sum_{j=1}^{k}\frac{n_j^*}{p_j(\lambda)}\frac{dp_j(\lambda)}{d\lambda}\bigg|_{\lambda^*}, \qquad (4.4.16)$$

которое совпадает с уравнением (4.4.14). Из этого результата и из рассмотренных выше свойств оценок максимального правдоподобия следует: по своим асимптотическим свойствам оценки, получаемые по методу минимума хи-квадрат, весьма близки к оценкам максимального правдоподобия. Вместе с этим необходимо отметить, что для оценки параметров по методу минимума хи-квадрат объем выборки n и численность разрядов k должны быть достаточно большими. Практически считается достаточным, чтобы было $n \geqslant 50$, $k \geqslant 5$–8.

В заключение отметим, что метод хи-квадрат, как правило, не позволяет получить оценку в явном аналитическом виде в зависимости от исходной выборки. Кроме того, вычисление статистических характеристик оценок по методу хи-квадрат представляет определенную сложность по сравнению с вычислением тех же характеристик при оценке параметров методом максимального правдоподобия. Поэтому оценка параметров по методу хи-квадрат на практике применяется весьма редко.

4.5. Примеры нахождения оценок параметров распределений

4.5.1. Оценка среднего значения гауссовской случайной величины при известной дисперсии

Пусть имеется независимая выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ гауссовской случайной величины с известной дисперсией σ^2 и неизвестным средним значением m, которое необходимо оценить. При этом условная плотность вероятности для рассматриваемой случайной величины записывается в виде

$$f(x \mid m) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]. \qquad (4.5.1)$$

Оценка параметра m^* (выборочное среднее $\bar{x}$) по методу моментов определяется соотношением (3.4.5). При этом вне зависимости от вида

распределения случайной величины X оценка среднего значения несмещенная, а дисперсия оценки среднего значения определяется выражением (3.4.6). Поскольку элементы выборки распределены по нормальному закону, выборочное среднее также распределено по нормальному закону со средним значением m и дисперсией $D(m^*) = \sigma^2/n$. При этом функция правдоподобия параметра m для рассматриваемой независимой выборки имеет вид

$$L(m, \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]. \qquad (4.5.2)$$

Решая уравнение правдоподобия относительно неизвестного параметра m, получаем выражение, совпадающее с (3.4.5).

Обращение в максимум функции правдоподобия $L(m; \mathbf{x})$ или $\ln L(m; \mathbf{x})$ по параметру m эквивалентно обращению в минимум суммы квадратов $\sum(x_i - m)^2$. Таким образом, рассмотренный в п. 3.5.5 метод наименьших квадратов является частным случаем метода максимального правдоподобия, когда наблюдения (выборочные значения) распределены по нормальному закону.

Покажем, что для рассматриваемой оценки выполняются необходимые и достаточные условия (4.2.7) и (4.2.12) существования эффективных оценок. Для этого представим функцию правдоподобия (4.5.2) в виде двух сомножителей

$$L(m; \mathbf{x}) = \left[(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)\right] \exp\left(\frac{mn}{\sigma^2}\bar{x} - \frac{m^2 n}{2\sigma^2}\right).$$

Первый сомножитель не зависит от оцениваемого параметра m, а второй зависит от выборки только через оценку $m^* \equiv \bar{x}$. Следовательно, соотношение (4.2.7), необходимое для существования достаточной оценки, выполняется. Выполняется и условие (4.2.12), необходимое для существования эффективной оценки. Действительно, из (4.5.2) имеем

$$\frac{d\ln L(m; \mathbf{x})}{dm} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m) = \frac{n}{\sigma^2}(\bar{x} - m).$$

Таким образом, выборочное среднее гауссовской случайной величины является эффективной несмещенной оценкой, как по методу моментов, так и по методу максимального правдоподобия. При этом дисперсия рассматриваемой оценки $D_{\text{эф}}(m) = \sigma^2/n$, что, естественно, совпадает с ранее полученным выражением (3.4.6).

В качестве оценки среднего значения гауссовской случайной величины иногда применяют выборочную медиану, дисперсия которой при $n \gg 1$ согласно (3.1.14) равна $D(m^*) = (\sigma_X^2/n)(\pi/2)$, что в $\pi/2 \approx 1{,}57$ раза превышает дисперсию эффективной оценки этого же параметра.

4.5.2. Оценка дисперсии гауссовской случайной величины

Оценим дисперсию σ^2 гауссовской случайной величины с известным средним значением m, полагая, как и ранее, выборку независимой. Условная плотность вероятности записывается аналогично (4.5.1), где в левой части вместо m надо понимать σ^2. При этом оценка дисперсии $(\sigma^2)^*$ (выборочная дисперсия s_1^2) по методу моментов определяется соотношением (3.4.11). Оценка дисперсии для рассматриваемого случая является несмещенной, а ее дисперсия определяется выражением (3.4.14).

Функция правдоподобия параметра σ^2 для независимой гауссовской выборки имеет вид

$$L(\sigma^2; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\right]. \qquad (4.5.3)$$

Отсюда, решая уравнение правдоподобия $[d\ln L(\sigma^2; \mathbf{x})/d(\sigma^2) = 0]$, получим выражение для оценки дисперсии, совпадающее с (3.4.11), т.е. оценки дисперсии гауссовской случайной величины по методу моментов и максимуму функции правдоподобия совпадают. Следовательно, совпадают и их статистические характеристики.

Проверим выполнение условий (4.2.7) и (4.2.12), необходимых для существования достаточных и эффективных оценок. Учитывая выражение для оценки дисперсии (3.4.11), функция правдоподобия может быть представлена как

$$L(\sigma^2; \mathbf{x}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}(\sigma^2)^*\right].$$

Полагая в (4.2.7) $h(\mathbf{x}) = (2\pi)^{-n/2}$, убеждаемся, что условие достаточности оценки выполняется. Выполняется и условие (4.2.12), необходимое для существования эффективной оценки, так как

$$\frac{d\ln L(\sigma^2; \mathbf{x})}{d(\sigma^2)} = \frac{n}{2\sigma^4}[(\sigma^2)^* - \sigma^2],$$

в котором надо понимать $\eta(\sigma^2) = n/2\sigma^4$. При этом дисперсия эффективной оценки дисперсии $D_{\text{эф}}[(\sigma^2)^*] = 1/\eta = 2\sigma^4/n$ совпадает с выражением (3.4.14).

Таким образом, оценка дисперсии гауссовской случайной величины по независимой выборке с известным средним значением является несмещенной и эффективной оценкой.

Случайная величина

$$Y = \frac{ns_1^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i - m)^2}{\sigma^2}$$

представляет собой сумму квадратов независимых гауссовских случайных величин с нулевыми средними значениями и единичными дисперси-

ями. Следовательно, она распределена по закону хи-квадрат с $\nu = n$ степенями свободы. Поэтому выборочная дисперсия $s_1^2 \equiv (\sigma^2)^* = Y\sigma^2/n$ имеет плотность вероятности, определяемую соотношением (2.2.6а).

4.5.3. Совместная оценка среднего значения и дисперсии гауссовской случайной величины

Рассмотрим совместную оценку среднего значения m и дисперсии σ^2 на основе анализа независимой гауссовской выборки объема n.

Согласно методу моментов получаем два уравнения

$$m^* \equiv \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad m_2^* \equiv s^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Из второго выражения получаем формулу для оценки дисперсии (выборочной дисперсии s^2), совпадающую с соотношением (3.4.10). Статистические характеристики для оценок $\bar{x}$ и s^2 получены выше. В данном случае выборочное среднее — оценка несмещенная с дисперсией оценки $D(\bar{x}) = \sigma^2/n$; выборочная дисперсия имеет смещение (3.4.16), а дисперсия и рассеяние оценки дисперсии определяются выражениями (3.4.21) и (3.4.22).

Вычислим оценки среднего значения m и дисперсии σ^2 методом максимального правдоподобия, функция правдоподобия которых для рассматриваемой гауссовской выборки $L(m, \sigma^2; \mathbf{x})$ описывается правой частью выражений (4.5.2) и (4.5.3). Решая уравнения правдоподобия

$$\frac{\partial \ln L(m, \sigma^2; \mathbf{x})}{\partial m} = 0; \quad \frac{\partial \ln L(m, \sigma^2; \mathbf{x})}{\partial \sigma^2} = 0$$

относительно параметров m и σ^2, находим их оценки m^* и $(\sigma^2)^* \equiv s^2$, совпадающие с выражениями (3.4.5) и (3.4.10). Таким образом, для гауссовской независимой выборки совместные оценки по методу моментов и методу максимального правдоподобия эквивалентны. Следовательно, одинаковы и их статистические характеристики.

Выясним связь полученных оценок с эффективными оценками. Для этого вычислим согласно (4.2.25)–(4.2.28) средние значения производных от логарифма функции правдоподобия

$$\ln L(m, \sigma^2; \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2.$$

Имеем

$$\mathrm{M}\left\{\frac{\partial^2}{\partial(\sigma^2)^2}\ln L(m, \sigma^2; \mathbf{x})\right\} = -\frac{n}{2\sigma^4};$$

$$\mathrm{M}\left\{\frac{\partial^2}{\partial m^2}\ln L(m, \sigma^2; \mathbf{x})\right\} = -\frac{n}{\sigma^2};$$

$$\mathrm{M}\left\{\frac{\partial^2}{\partial m \partial(\sigma^2)}\ln L(m, \sigma^2; \mathbf{x})\right\} = 0.$$

Так как среднее значение смешанной производной логарифма функции правдоподобия равно нулю, дисперсии совместно эффективных несмещенных оценок равны дисперсиям эффективных несмещенных раздельных оценок. Подставляя полученные средние значения производных в формулы (4.2.25) и (4.2.26), находим выражения для дисперсий оценок, совпадающие с дисперсиями (3.4.6) и (3.4.14) раздельных оценок m^* и s_1^2 параметров m и σ^2.

Дисперсия эффективной смещенной оценки дисперсии (3.4.10) s^2 согласно (4.2.16)

$$D_{\mathrm{ef}}(s^2) = \frac{2(n-1)^2}{n^3}\sigma^4. \tag{4.5.4}$$

Из сравнения этого выражения с дисперсией эффективной оценки (3.4.11) $[D(s_1^2) = 2\sigma^4/n]$ видим, что дисперсия смещенной эффективной оценки дисперсии меньше дисперсии несмещенной эффективной оценки того же параметра.

Эффективность несмещенной оценки дисперсии s_2^2 при неизвестном среднем значении (3.4.27)

$$E(s_2^2) = \frac{D_{\mathrm{ef}}(s^2)}{D(s_2^2)} = \left(1 - \frac{1}{n}\right)^2 < 1, \tag{4.5.5}$$

где $D(s_2^2)$ определяется выражением (3.4.30). Однако при большом объеме выборки $(n \gg 1)$ оценка s_2^2 является асимптотически эффективной. При этих же условиях и оценка s^2 (3.4.10) также является асимптотически несмещенной и эффективной.

Как отмечалось выше (п. 3.4), выборочные среднее и дисперсия гауссовской выборки между собой независимы (см. (3.4.32)). Поэтому плотность вероятности совместной оценки параметров m и σ^2 равна произведению плотностей вероятности раздельных оценок этих параметров. Распределение выборочного среднего является нормальным со средним значением m и дисперсией σ^2/n. Выборочная дисперсия s^2 имеет распределение вида (2.2.6а) с $\nu = n-1$ степенями свободы. Последнее обусловлено тем, что число степеней свободы уменьшено за счет наличия одной линейной связи (3.4.5) между выборочными значениями. Ниже (п. 4.9.2) будет показано, что в данном случае сумма $Z = \sum(x_i - \bar{x})^2/\sigma^2$ эквивалентна сумме $(n-1)$ квадратов независимых случайных величин с нулевыми средними значениями и единичными дисперсиями и поэтому случайная величина Z распределена по закону хи-квадрат (2.2.3) с $\nu = n - 1$ степенями свободы.

4.5.4. Оценка параметров биномиального и пуассоновского распределений

4.5.4.1. Биномиальное распределение

Рассмотрим оценку вероятности p биномиального распределения

$$p_n(k) = C_n^k p^k (1-p)^{n-k}, \quad k \geqslant 0,$$

на основании анализа n независимых испытаний, причем в каждом испытании положительный исход наступает с постоянной вероятностью p и не наступает с вероятностью $q = 1-p$; k — число положительных исходов. В результате испытаний получено k^* положительных исходов. Приравнивая оценку числа положительных исходов ее теоретическому значению

$$m = \sum_{k=0}^{n} kp(k) = np,$$

находим правило оценки искомой вероятности

$$p^* = k^*/n. \tag{4.5.6}$$

Согласно закону больших чисел данная оценка по вероятности стремиться к своему истинному значению p, т.е. оценка (4.5.6) является состоятельной.

Дисперсия оценки вероятности p определяется соотношением

$$D(p) = D(k)/n^2.$$

Учитывая формулу для дисперсии числа положительных исходов (1.3.13), имеем

$$D(p) = p(1-p)/n. \tag{4.5.7}$$

К правилу оценки (4.5.6) параметра p приходим и при оценке его методом максимального правдоподобия. Действительно, функция правдоподобия параметра p имеет вид

$$L(p; k) = C_n^k p^k (1-p)^{n-k}.$$

Решая уравнение правдоподобия $[d \ln L(p; k)/dp = 0]$ относительно параметра p, получаем оценку в виде (4.5.6) с характеристиками оценки, приведенными выше.

Полагая в последнем выражении $h(\mathbf{x}) = C_n^k$, убеждаемся в выполнении условия достаточности (4.2.7) оценки (4.5.6). Выполняется и условие (4.2.12), необходимое для существования эффективной оценки:

$$\frac{d}{dp} \ln L(p; k^*) = \frac{d}{dp}[k^* \ln p + (n-k^*)\ln(1-p)] =$$
$$= \frac{k^* - np}{p(1-p)} = \frac{n}{p(1-p)}(p^* - p).$$

При этом согласно (4.2.19) выражение для дисперсии эффективной оценки $D_{\text{ef}}(p) = 1/\eta = p(1-p)/n$, естественно, совпадает с ранее полученным выражением (4.5.7).

Трудность применения формулы (4.5.7) для вычисления дисперсии оценки состоит в том, что истинное значение вероятности p неизвестно. Для преодоления этой трудности имеются различные пути. Одним из них является замена произведения $p(1-p)$ его максимальным значением, равным 0,25. Такой прием целесообразен для p, близких к 0,5. Вторым способом устранения указанного затруднения, является замена истинного значения параметра p его оценкой p^*, т.е. вместо истинного значения дисперсии $D(p) = p(1-p)/n$ взять приближенное значение $s^2(p) = p^*(1-p^*)/(n-1)$. Это вполне допустимо при большом объеме выборки, так как по закону больших чисел (предельная теорема Бернулли) оценка p^* близка к ее вероятности p. Однако, если величины p^* или $(1-p^*)$ близки к нулю, то при малом объеме выборки более точное приближение состоит в повторении N серий испытаний. В этом случае оценка параметра $p^* \equiv \bar{p}$ и выборочная дисперсия оценки $\bar{p}$ определяются как

$$\bar{p} = \frac{1}{nN} \sum_{i=1}^{n} k_i^*; \tag{4.5.8}$$

$$s_2^2(p) = \frac{1}{N-1} \sum_{i=1}^{N} (p_i^* - \bar{p})^2. \tag{4.5.9}$$

4.5.4.2. Распределение Пуассона

Рассмотрим оценку параметра λ при анализе g серий случайной величины, распределенной по пуассоновскому закону (1.3.29). Будем полагать, что в каждой серии испытаний получено k_j, $j = 1, 2, \ldots, g$, положительных исходов. Приравнивая выборочное среднее значение числа положительных исходов k^* теоретическому значению распределения Пуассона λ, получим формулу для оценки неизвестного параметра

$$\lambda^* = \frac{1}{g} \sum_{j=1}^{g} k_j = k^*. \tag{4.5.10}$$

Полученная методом моментов оценка параметра λ является несмещенной, так как $\mathrm{M}\{k_j\} = \lambda$. При этом дисперсия оценки параметра λ определяется выражением

$$D(\lambda) = D(k)/g = \lambda/g. \tag{4.5.11}$$

Функция правдоподобия параметра для принятых условий имеет вид

$$L(\lambda; k) = \prod_{j=1}^{g} \frac{\lambda^{k_j}}{k_j!} \mathrm{e}^{-g\lambda}. \tag{4.5.12}$$

Решая уравнение правдоподобия $[d\ln L(\lambda;k)/d\lambda = 0]$ относительно параметра λ, получаем оценку в виде (4.5.10) с характеристиками оценки, приведенными выше. Найденная оценка эффективная, так как выполняется условие (4.2.12) существования эффективной оценки:

$$\frac{d}{d\lambda}\ln L(\lambda;k^*) = \frac{g}{\lambda}(\lambda^* - \lambda).$$

4.5.5. Оценка параметров прямоугольного распределения

4.5.5.1. Оценка параметра a прямоугольного распределения

Рассмотрим оценку параметра a прямоугольного распределения

$$f(x) = \begin{cases} 1/a, & 0 \leqslant x \leqslant a, \\ 0, & x < 0,\ x > a. \end{cases} \tag{4.5.13}$$

при анализе n независимых одинаково распределенных выборочных значений $\mathbf{x}(x_1, x_2, \ldots, x_n)$.

Приравнивая среднее значение $m = a/2$ выборочному среднему $\bar{x}$, получаем оценку параметра a^* по методу моментов:

$$a^* = \frac{2}{n}\sum_{i=1}^{n} x_i. \tag{4.5.14}$$

Отсюда находим среднее значение оценки

$$\mathrm{M}\{a^*\} = \frac{2}{n}\sum_{i=1}^{n}\mathrm{M}\{x_i\} = a,$$

т.е. оценка параметра a рассматриваемого распределения несмещенная. Дисперсия оценки, учитывая независимость элементов выборки,

$$D_1(a^*) = \frac{4}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\left\{\left(x_i - \frac{a}{2}\right)\left(x_j - \frac{a}{2}\right)\right\} = \frac{a^2}{3n}. \tag{4.5.15}$$

Для определения оценки параметра a методом максимального правдоподобия рассматриваемую плотность вероятности запишем с использованием единичной функции (3.2.2)

$$f(x) = \frac{1}{a}\kappa(a - x), \quad 0 \leqslant x \leqslant a. \tag{4.5.16}$$

Тогда выражение для функции правдоподобия будет иметь вид

$$L(a, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, a) = \frac{1}{a^n}\prod_{i=1}^{n}\kappa(a - x_i). \tag{4.5.17}$$

Эта функция не дифференцируема, так как

$$\prod_{i=1}^{n}\kappa(a - x_i) = \begin{cases} 1 & a \geqslant x_{(n)}, \\ 0 & a < x_{(n)}. \end{cases}$$

Здесь, как и выше, $x_{(n)} = \max x_i$ — наибольший член вариационного ряда, $i = 1, 2, \ldots, n$.

Из-за недифференцируемости функции правдоподобия по параметру a решение уравнения правдоподобия для нахождения оценки не существует. Однако максимум функции правдоподобия имеется и соответствует

$$a^* = x_{(n)}. \tag{4.5.18}$$

Действительно, первый множитель (4.5.17) убывает с увеличением a, а второй множитель равен нулю при $a < x_{(n)}$ и остается постоянным (равным единице) при $a \geqslant x_{(n)}$. Поэтому максимальное значение функции правдоподобия будет соответствовать наименьшему возможному значению параметра a, т.е. $a_{\min} = x_{(n)}$. При этом естественно ожидать, что в среднем оценка (4.5.18) будет меньше истинного значения a, т.е. будет существовать смещение оценки.

Поскольку все наблюдения, включая наибольшее значение, не превосходят a, выражение (4.5.17) можно записать в виде

$$L(a;\mathbf{x}) = a^{-n}\kappa(a - a^*). \tag{4.5.19}$$

Покажем, что оценка (4.5.17) является достаточной статистикой. Для этого функцию правдоподобия (4.5.19) можно рассматривать как произведение двух функций: функции $h(x_1, x_2, \ldots, x_n) = 1$ и функции, определяемой правой частью (4.5.19) и зависящей от оцениваемого параметра a и оценки a^*. Таким образом, условие (4.2.7), необходимое для существования достаточной оценки, выполняется. Проверить выполнение условия существования эффективной оценки (4.2.12) не представляется возможным из-за недифференцируемости логарифма функции правдоподобия по оцениваемому параметру a.

Вычислим статистические характеристики оценки (4.5.18), для чего запишем функцию распределения оценки

$$\mathbf{P}\{a^* = x_{(n)} < u\} = F(u) = \mathbf{P}\{x_1 < u, x_2 < u, \ldots, x_n < u\} =$$

$$= \prod_{i=1}^{n}\mathbf{P}\{x_i < u\} = \begin{cases} (u/a)^n, & 0 \leqslant u \leqslant a, \\ 1, & u > a. \end{cases} \tag{4.5.20}$$

Функция распределения (4.5.20) определяет вероятность того, что ни одно из выборочных значений не превысит величины u, изменяющейся в пределах от 0 до a.

Используя (4.5.20), находим первые два начальных момента оцен-

ки a^*

$$\mathrm{M}\{a^*\} = \int_0^a u\,dF(u) = \int_0^a un\frac{u^{n-1}}{a^n}\,du = \frac{n}{n+1}a;$$

$$\mathrm{M}\{(a^*)^2\} = \int_0^a u^2\,dF(u) = \frac{n}{n+2}a^2.$$

Из первого соотношения можно сделать вывод, что оценка максимального правдоподобия параметра a^* является смещенной, так как

$$b(a) = -a/(n+1).$$

Однако при $n \gg 1$ $(n \to \infty)$ смещение оценки стремится к нулю, т.е. оценка параметра a по методу максимального правдоподобия является состоятельной и асимптотически несмещенной.

Дисперсия оценки рассматриваемого параметра

$$D_2(a^*) = \mathrm{M}\{(a^*)^2\} - [\mathrm{M}\{a^*\}]^2 = \frac{a^2 n}{(n+1)^2(n+2)}. \qquad (4.5.21)$$

Взяв отношение дисперсий (4.5.15) и (4.5.21)

$$\frac{D_1(a^*)}{D_2(a^*)} = \frac{(1+1/n)^2(n+2)}{3}, \qquad (4.5.22)$$

видим, что дисперсия оценки максимального правдоподобия параметра a значительно меньше дисперсии оценки этого же параметра по методу моментов.

Для получения несмещенной оценки параметра a в качестве оценки следует взять величину

$$a^*_{\text{н}} = \frac{n+1}{n}a^* = \frac{n+1}{n}x_{(n)}. \qquad (4.5.23)$$

При этом дисперсия несмещенной оценки параметра a

$$D(a^*_{\text{н}}) = \frac{a^2}{n(n+2)}. \qquad (4.5.24)$$

4.5.5.2. Оценка параметров a и c прямоугольного распределения

Рассмотрим оценку параметров a и c $(c > a)$ прямоугольного распределения

$$f(x; a, c) = \begin{cases} 1/(c-a), & a \leqslant x \leqslant c, \\ 0, & x > c,\ x < a \end{cases} \qquad (4.5.25)$$

при анализе n независимых одинаково распределенных выборочных значений $\mathbf{x}(x_1, x_2, \ldots, x_n)$.

Приравнивая первый начальный и второй центральный моменты рассматриваемого распределения их выборочным значениям, имеем систему из двух уравнений относительно параметров a и c

$$m = \frac{a+c}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}; \quad D = \frac{(c-a)^2}{12} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2.$$

Решая эту систему уравнений относительно неизвестных параметров a и c, получаем выражения для их оценок по методу моментов:

$$a^* = \bar{x}[1 - 3s^2/(\bar{x})^2]; \quad c^* = \bar{x}[1 + 3s^2/(\bar{x})^2]. \qquad (4.5.26)$$

Вычислить характеристики оценок в общем виде оказывается достаточно сложно. В то же время метод максимального правдоподобия позволяет найти совместные оценки параметров a и c, а также вычислить их статистические характеристики.

Аналогично (4.5.17) функцию правдоподобия параметров a и c можно записать в виде

$$L(a, c; \mathbf{x}) = \frac{1}{(c-a)^n}\prod_{i=1}^{n}\kappa(c - x_i)\kappa(x_i - a). \qquad (4.5.27)$$

Рассуждая подобно тому, как это сделано при оценке одного параметра в предыдущем пункте, функция правдоподобия (4.5.27) достигает своего максимального значения тогда, когда в качестве оценок параметров взяты величины

$$a^* = x_{(1)}; \quad c^* = x_{(n)}, \qquad (4.5.28)$$

соответствующие минимальному и максимальному выборочным значениям. Действительно, при этих статистиках функция правдоподобия принимает максимальное значение (рис. 4.5.1), так как одновременно выполняются условия

$$(c^* - a^*) = \min;$$

$$\kappa(c^* - x_i)\kappa(x_i - a^*) = 1 = \text{const}.$$

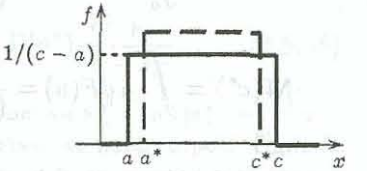


Рис. 4.5.1. Оценка параметров распределения

Аналогично (4.5.19) функцию правдоподобия (4.5.27) можно записать как

$$L(a, c; \mathbf{x}) = \frac{1}{(c-a)^n}\kappa(c - c^*)\kappa(a^* - a). \qquad (4.5.29)$$

Последнее выражение можно представить в виде (4.2.7), полагая $h(x_1, x_2, \ldots, x_n) = 1$. Поэтому полученная совместная оценка параметров a и c (4.5.28) является совместно достаточной статистикой. Так

же, как и в предыдущем пункте, проверить условие существование эффективных оценок (4.2.12) не представляется возможным из-за недифференцируемости функции правдоподобия. Необходимо заметить, что оценки параметров a и c, определяемые (4.5.28), статистически зависимы между собой. Однако, как показано в [30, 37], при $n \gg 1$ (в пределе при $n \to \infty$) они асимптотически независимы.

Вычислим статистические характеристики оценок (4.5.28), для чего необходимо найти их функции распределения.

Функция распределения случайной величины $a^* = x_{(1)}$ определяется как вероятность того, что ни одно из независимых выборочных значений не будет меньше наименьшего фиксированного значения $v \geqslant a$:

$$F(v) = 1 - \mathbf{P}\{x_{(1)} > v\} = 1 - \mathbf{P}\{x_1 > v, x_2 > v, \ldots, x_n > v\} = \\ = 1 - \prod_{i=1}^{n} \int_v^c f(x_i; a, c)\, dx_i = 1 - \frac{(c-v)^n}{(c-a)^n}, \quad a \leqslant v \leqslant c. \quad (4.5.30)$$

Функция распределения случайной величины $c^* = x_{(n)}$ определяется как вероятность того, что ни одно из независимых выборочных значений не превысит наибольшего фиксированного значения $u \leqslant c$, т.е.

$$F(u) = \mathbf{P}\{x_{(n)} < u\} = \mathbf{P}\{x_1 < u, x_2 < u, \ldots, x_n < u\} = \\ = \prod_{i=1}^{n} \int_a^u f(x_i; a, c)\, dx_i = \frac{(c-u)^n}{(c-a)^n}, \quad a \leqslant u \leqslant c. \quad (4.5.31)$$

Начальные моменты оценок:

$$\mathrm{M}\{a^*\} = \int_a^c v\, dF(v) = \frac{n}{(c-a)^n} \int_a^c v(c-v)^{n-1}\, dv = \frac{(na+c)}{(n+1)}; \quad (4.5.32)$$

$$\mathrm{M}\{c^*\} = \int_a^c u\, dF(u) = \frac{n}{(c-a)^n} \int_a^c u(u-a)^{n-1}\, du = \frac{(nc+a)}{(n+1)}. \quad (4.5.33)$$

Из последних соотношений видно, что оценки параметров a и c прямоугольного распределения смещенные. Однако при достаточно большом объеме выборки ($n \to \infty$) оценки становятся асимптотически несмещенными.

Для вычисления дисперсий оценок параметров a и c найдем их вторые начальные моменты

$$\mathrm{M}\{(a^*)^2\} = \int_a^c v^2\, dF(v) = \frac{n}{(c-a)^n} \int_a^c v^2(c-v)^{n-1}\, dv = \\ = \frac{n(n+1)a^2 + 2acn + 2c^2}{(n+1)(n+2)};$$

$$\mathrm{M}\{(c^*)^2\} = \int_a^c u^2\, dF(u) = \frac{n}{(c-a)^n} \int_a^c u^2(u-a)^{n-1}\, du = \\ = \frac{n(n+1)c^2 + 2acn + 2a^2}{(n+1)(n+2)}.$$

При этом (пренебрегая в первом приближении статистической зависимостью оценок a^* и c^*) дисперсии оценок параметров a и c равны между собой и определяются выражением

$$D(a^*) = D(c^*) = \frac{n(c-a)^2}{(n+1)^2(n+2)}. \quad (4.5.34)$$

Несмещенные оценки параметров $a^*_{\text{н}}$, $c^*_{\text{н}}$ связаны со смещенными оценками a^* и c^* соотношениями

$$a^* = \frac{na^*_{\text{н}} + c^*_{\text{н}}}{n+1}; \quad c^* = \frac{nc^*_{\text{н}} + a^*_{\text{н}}}{n+1},$$

вытекающими из выражений (4.5.32) и (4.5.33). Решая эти уравнения относительно параметров $a^*_{\text{н}}$ и $c^*_{\text{н}}$, получаем статистику (оценку) для несмещенных оценок параметров a и c:

$$a^*_{\text{н}} = \frac{na^* - c^*}{n-1}; \quad c^*_{\text{н}} = \frac{nc^* - a^*}{n-1}. \quad (4.5.35)$$

Непосредственным вычислением математических ожиданий левых и правых частей (4.5.35) нетрудно убедиться, что оценки $a^*_{\text{н}}$ и $c^*_{\text{н}}$ несмещенные.

Пренебрегая в первом приближении статистической зависимостью оценок a^* и c^*, дисперсии несмещенных оценок $a^*_{\text{н}}$ и $c^*_{\text{н}}$

$$D(a^*_{\text{н}}) = D(c^*_{\text{н}}) = \frac{n^2 D(a^*) + D(c^*)}{(n-1)^2} = D(a^*)\frac{n^2+1}{(n-1)^2}. \quad (4.5.36)$$

Таким образом, полученные несмещенные оценки в $(n^2+1)/(n-1)^2$ раз превышают дисперсии смещенных оценок этих же параметров. Однако при $n \gg 1$ дисперсии обеих совместных оценок практически совпадают.

4.6. Объединение нескольких оценок одного и того же параметра

В ряде практических приложений теории оценок возникает задача наилучшего использования k оценок $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$, полученных в системе из k измерителей неизвестного параметра. Под наилучшим (оптимальным) использованием частных оценок λ_j^*, $j = 1, 2, \ldots, k$, понимается построение оптимальной (в смысле заданного критерия) оценки λ^{**} неизвестного параметра на основе оценок $\boldsymbol{\lambda}^*$.

4.6.1. Объединение оценок при известных условных плотностях вероятности

Если известны $f_\lambda(\boldsymbol{\lambda}^* \mid \lambda)$ — условная плотность вероятности оценок $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$ при фиксированном значении неизвестного параметра λ и $f_{\mathrm{pr}}(\lambda)$ — априорная плотность вероятности λ, то условная плотность вероятности этого параметра при заданном наборе оценок $\boldsymbol{\lambda}^*$ (апостериорная плотность вероятности) может быть записана в виде

$$f(\lambda \mid \boldsymbol{\lambda}^*) = f_{\mathrm{ps}}(\lambda; \boldsymbol{\lambda}^*) = \frac{f_{\mathrm{pr}}(\lambda) f_\lambda(\boldsymbol{\lambda}^* \mid \lambda)}{\int_\Lambda f_{\mathrm{pr}}(\lambda) f_\lambda(\boldsymbol{\lambda}^* \mid \lambda)\, d\lambda}. \tag{4.6.1}$$

На основе апостериорной плотности вероятности может быть построена оптимальная оценка λ^{**}, т.е. найден оптимальный алгоритм использования частных оценок $\boldsymbol{\lambda}^*$. Для получения байесовской оценки λ_m при заданной функции потерь $C(\boldsymbol{\lambda}^*, \lambda)$ следует использовать изложенные в п. 4.3 общие положения теории статистических решений, подставляя в соответствующие выражения апостериорную плотность вероятности из (4.6.1).

При неизвестном априорном распределении параметра λ, *а также* в некоторых других случаях определение байесовского алгоритма объединения оценок невозможно или нецелесообразно. В такой ситуации может оказаться разумным на основе оценок $\boldsymbol{\lambda}^*$ формирование оценки максимального правдоподобия λ_m^* неизвестного параметра. Оптимальное использование оценок $\boldsymbol{\lambda}^*$ при этом сводится к вычислению функции правдоподобия $f_\lambda(\boldsymbol{\lambda}^* \mid \lambda)$ и к определению положения ее абсолютного максимума λ_m.

Относительно просто оптимальный (по методу максимального правдоподобия) алгоритм объединения оценок может быть найден, если эти оценки являются гауссовскими случайными величинами. Тогда, обозначая $\mathrm{M}\{\lambda_j^*\} = \lambda + b_j$, где b_j — смещение оценки, и полагая оценки независимыми между собой, функцию правдоподобия можно записать в виде

$$L(\lambda; \boldsymbol{\lambda}^*) = \prod_{j=1}^{k} (2\pi D_j)^{-1/2} \exp\left[-\frac{(\lambda_j^* - \lambda - b_j)^2}{2D_j}\right], \tag{4.6.2}$$

где $D_j \equiv D(\lambda_j)$ — дисперсия j-й оценки параметра.

Оценка максимального правдоподобия λ_m определяется как положение абсолютного максимума (4.6.2). Решая уравнение правдоподобия $[d \ln L(\lambda; \boldsymbol{\lambda}^*)/d\lambda = 0]$ относительно неизвестного параметра λ, имеем

$$\lambda_m = \sum_{j=1}^{k} \alpha_j (\lambda_j^* - b_j), \tag{4.6.3}$$

где

$$\alpha_j = \frac{1}{D_j}\left(\sum_{j=1}^{k} \frac{1}{D_j}\right)^{-1}. \tag{4.6.4}$$

Из последнего выражения видно, что сумма всех коэффициентов λ_j равна единице:

$$\sum_{j=1}^{k} \alpha_j = 1. \tag{4.6.5}$$

Согласно полученным выражениям оптимальный алгоритм использования оценок предполагает вычисление взвешенной суммы всех оценок *с весами, определяемыми* (4.6.4). *При этом вес частной оценки тем* меньше, чем больше ее дисперсия.

Учитывая предположение о независимости оценок, дисперсия объединенной оценки определяется выражением

$$D(\lambda_m) = \sum_{j=1}^{k} \alpha_j^2 D_j = \left(\sum_{j=1}^{k} \frac{1}{D_j}\right)^{-1}. \tag{4.6.6}$$

В частном случае, когда все оценки по точности одинаковы, т.е. $D_j = D$, из (4.6.4) получаем $\alpha_j = 1/k$, $D(\lambda_m) = D/k$.

4.6.2. Линейное объединение оценок

Если совместное распределение частных оценок $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$ неизвестно, а заданы лишь их первые два момента, то найти оптимальный (по методу максимального правдоподобия) алгоритм их объединения не представляется возможным. Однако в этом случае можно определить несмещенную оценку λ^{**}, которая минимизирует среднюю квадратическую ошибку в классе *линейных оценок*. Предположим, что оценка вычисляется по формуле

$$\lambda^{**} = \sum_{j=1}^{k} \beta_j (\lambda_j^* - b_j). \tag{4.6.7}$$

Весовые коэффициенты β_j должны удовлетворять двум условиям: объединенная оценка должна быть несмещенной и иметь минимальную дисперсию, которая применительно к независимым частным оценкам определяется соотношением

$$D[\lambda^{**}(\beta_1, \beta_2, \ldots, \beta_k)] = \sum_{j=1}^{k} \beta_j^2 D_j. \tag{4.6.8}$$

Для того чтобы формируемая оценка (4.6.7) была несмещенной, для весовых коэффициентов β_j должно выполняться равенство

$$\sum_{j=1}^{k} \beta_j = 1. \tag{4.6.9}$$

Так как весовые коэффициенты β_j между собой связаны линейной зависимостью (4.6.9), нахождение значений весовых коэффициентов, обеспечивающих минимум функции (4.6.8), может быть решено с помощью метода множителей Лагранжа. Применительно к рассматриваемой задаче (одна связь между переменными) вводится вспомогательная функция

$$F(\beta_1, \beta_2, \ldots, \beta_k) = D[\lambda^{**}(\beta_1, \beta_2, \ldots, \beta_k)] - g\left(\sum_{j=1}^{k} \beta_j - 1\right) \tag{4.6.10}$$

от независимых переменных $\beta_1, \beta_2, \ldots, \beta_k$, где g — множитель Лагранжа. Затем из системы уравнений

$$\frac{dF}{d\beta_j} = \frac{dD}{d\beta_j} - g\frac{d}{d\beta_j}\left(\sum_{j=1}^{k} \beta_j - 1\right) = 0, \quad j = 1, 2, \ldots, k, \tag{4.6.11}$$

с учетом уравнения связи (4.6.9) находят значения весовых коэффициентов β_j, обеспечивающие экстремум (в данном случае минимум) дисперсии объединенной оценки. Согласно описанной процедуре нахождения экстремума, имеем

$$2\beta_j D_j = g.$$

Откуда получаем

$$\beta_j = \frac{g}{2D_j}. \tag{4.6.12}$$

Согласно уравнению связи (4.6.9)

$$\sum_{j=1}^{k} \beta_j = \frac{g}{2}\sum_{j=1}^{k} \frac{1}{D_j} = 1.$$

Следовательно,

$$g = 2\left(\sum_{j=1}^{k} \frac{1}{D_j}\right)^{-1}. \tag{4.6.13}$$

Подставляя значение g в формулу (4.6.12), находим оптимальные значения весовых коэффициентов

$$\beta_j = \frac{1/D_j}{\sum\limits_{j=1}^{k} 1/D_j}, \tag{4.6.14}$$

совпадающие с (4.6.4), т.е. $\beta_j = \alpha_j$. Следовательно, дисперсия объединенной оценки также определяется формулой (4.6.6), точнее — ее правой частью.

Условие для оптимальных значений весовых коэффициентов часто записывают в виде отношений между ними:

$$\beta_1 : \beta_2 : \ldots : \beta_k = \frac{1}{D_1} : \frac{1}{D_2} : \ldots : \frac{1}{D_k}. \tag{4.6.15}$$

Дисперсии частных оценок параметров распределений зависят от соответствующих моментов этих распределений, которые, как правило, неизвестны. Поэтому в первом приближении вместо истинных значений моментов можно взять соответствующие выборочные моменты.

4.7. Интервальное оценивание

Рассмотренные выше точечные оценки параметров распределений, являясь случайными, как правило, не позволяют определить их достоверность. Это обусловлено, прежде всего, тем, что точности определения параметров зависят от соответствующих моментов распределений, истинные значения которых неизвестны. Поэтому вместо точных значений величин, входящих в выражения для дисперсий оценок параметров, используют выборочные значения этих величин, являющихся также случайными величинами. Если такие приближения и допустимы при достаточно большом объеме выборки (например, при $n \gg 1$), то при малом объеме выборки такой подход может привести к значительным ошибкам в определении точности оценок.

Отмеченные недостатки устраняются при интервальном методе оценивания параметров распределений, согласно которому определяется достоверность нахождения оцениваемого параметра в заданном интервале его значений.

Рассмотрим сущность *интервального оценивания* на примере одного неизвестного параметра распределения λ. В соответствии со смыслом интервального оценивания по сделанной выборке на основе некоторых правил должны быть найдены два значения λ_1 и λ_2 (пусть $\lambda_1 < \lambda_2$) такие, чтобы выполнялось условие

$$\mathbf{P}\{\lambda_1 \leqslant \lambda \leqslant \lambda_2\} = \gamma \equiv 1 - \alpha. \tag{4.7.1}$$

Числа λ_1 и λ_2 называются *доверительными границами* (доверительными пределами), при этом λ_1 — нижняя граница, а λ_2 — верхняя граница. Совокупность доверительных границ (λ_1, λ_2) представляет собой

доверительный интервал. При этом величину $\Delta = \lambda_2 - \lambda_1$ называют *шириной* (*длиной*) *доверительного интервала* для параметра λ. Число γ называют *доверительной вероятностью* (коэффициентом доверия). Числа λ_1 и λ_2 находятся (вычисляются) на основе имеющейся выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ и заданной доверительной вероятности γ. В результате можем записать

$$\lambda_1 = \psi_1(x_1, x_2, \ldots, x_n; \gamma); \quad \lambda_2 = \psi_2(x_1, x_2, \ldots, x_n; \gamma). \tag{4.7.2}$$

Следовательно, сами числа λ_1 и λ_2 также являются случайными величинами. Поэтому и доверительный интервал (λ_2, λ_1) тоже случаен. Доверительный интервал может покрывать истинное значение параметра λ с вероятностью γ или не покрывать его с вероятностью $\alpha = 1 - \gamma$. Именно в таком смысле и понимается вероятность события $\mathbf{P}\{\lambda_1 \leqslant \lambda \leqslant \lambda_2\}$, хотя в каждом конкретном измерении оцениваемый параметр λ, являясь неслучайным, попадает или не попадает в этот интервал с вероятностью, равной единице. Поэтому, если утверждается, что справедливо соотношение (4.7.1), то в конечном счете это утверждение будет верно в части случаев, равной $\gamma = 1 - \alpha$. Если полагается, что доверительная вероятность достаточно велика (например, $\gamma = 0{,}95$; 0,99 и т.д.), то практически достаточно достоверно попадание истинного значения параметра λ в доверительный интервал (λ_1, λ_2). В связи с этим число $(\lambda_1 + \lambda_2)/2$, являющееся серединой доверительного интервала, приближенно будет давать точечное значение неизвестного параметра с погрешностью до величины порядка $(\lambda_2 - \lambda_1)/2$. Если доверительная вероятность выбрана достаточно большой, то говорят, что параметр λ «почти всегда» или практически всегда лежит внутри доверительного интервала, и при единичной реализации выборочных данных пренебрегают тем, что результат оценивания доверительных границ λ_1 и λ_2 неточен.

Выбор конкретного значения доверительной вероятности в значительной степени зависит как от той цели, которая ставится при принятии решения, так и от условий проведения измерения искомого параметра. Например, пусть некоторый параметр механизма, влияющий на безопасность его эксплуатации, должен находится в пределах от λ_1 до λ_2. Для безаварийной работы этого механизма надежность (доверительная вероятность) оценки (попадание оцениваемого параметра в указанный интервал) должна быть достаточно большой (например, $\gamma \geqslant 0{,}9999$). С другой стороны, при большинстве статистических исследований, как правило, имеется много дополнительных источников ошибок, обусловленных недостоверностью теоретических предпосылок, упрощений, допущений и т.д. Поэтому составляющая ошибки из-за применения статистической обработки может быть соизмеримой или даже меньше указанных дополнительных составляющих ошибок. В связи с этим наиболее часто удовлетворяются величиной γ, лежащей в интервале 0,9...0,99.

Из соотношения (4.7.1) видно, что однозначность в выборе доверительных границ λ_1 и λ_2 отсутствует. Поэтому отсутствует и однозначность в вычислении доверительного интервала. Для устранения неоднозначности для каждой из границ необходимо задать значения величин α_1 и α_2, для которых справедливы соотношения

$$\mathbf{P}\{\lambda_1 < \lambda\} = 1 - \alpha_1; \quad \mathbf{P}\{\lambda_2 > \lambda\} = 1 - \alpha_2. \tag{4.7.3}$$

При этом должно выполняться условие

$$\alpha_1 + \alpha_2 = \alpha = 1 - \gamma. \tag{4.7.4}$$

Если величины α_1 и α_2 взяты равными, т.е. $\alpha_1 = \alpha_2 = \alpha/2 = (1 - \gamma)/2$, то доверительный интервал (λ_2, λ_1) называется *центральным*. В противном случае интервалы называют нецентральными. Следует заметить, что наличие центральных интервалов не означает равного отстояния границ λ_1 и λ_2 от истинного значения оцениваемого параметра λ.

Пусть на основе исходной выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ получена достаточная условная оценка

$$\lambda^* = \varphi(x_1, x_2, \ldots, x_n \mid \lambda), \tag{4.7.5}$$

включающая в себя всю информацию об оцениваемом параметре λ. Тогда соотношения (4.7.2) можно записать также в виде

$$\lambda_1 = \psi_1(\lambda^*; \gamma); \quad \lambda_2 = \psi_2(\lambda^*; \gamma). \tag{4.7.6}$$

Существует несколько подходов определения границ λ_1 и λ_2. При этом, естественно, для заданной доверительной вероятности γ желательно получить такие значения границ, при которых ширина доверительного интервала будет минимальной, т.е. чтобы выполнялось условие $(\lambda_2 - \lambda_1) = \min$. Как и выше, будем полагать, что выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ получена из непрерывного распределения $f(\mathbf{x}; \lambda)$, содержащего один неизвестный параметр λ. Пусть λ^* — оценка искомого параметра, а $f(\lambda^* \mid \lambda)$ — условная плотность вероятности оценки λ^* для данного значения оцениваемого параметра λ.

4.7.1. Байесовские интервальные оценки параметров

При байесовском подходе к определению границ интервальной оценки полагается, что оцениваемый параметр λ является случайной величиной с известным априорным распределением $f_{\text{pr}}(\lambda)$. Согласно формуле Байеса (1.1.19) условная плотность вероятности $f(\lambda \mid \lambda^*)$ может быть записана в виде (4.6.1). При этом величина

$$\mathbf{P}\{\lambda_1 \leqslant \lambda \leqslant \lambda_2 \mid \lambda^*\} = \int_{\lambda_1}^{\lambda_2} f(\lambda \mid \lambda^*)\, d\lambda = \gamma \equiv 1 - \alpha \tag{4.7.7}$$

представляет собой условную вероятность того, что истинное значение оцениваемого параметра λ лежит в интервале от λ_1 до λ_2 при найденном

на основе исходной выборки значении оценки λ^*. Эта вероятность представляет собой апостериорную вероятность события в отличие от априорной вероятности попадания оцениваемого параметра в этот же интервал:

$$\mathbf{P}_{\mathrm{pr}}\{\lambda_1 \leqslant \lambda \leqslant \lambda_2\} = \int_{\lambda_1}^{\lambda_2} f_{\mathrm{pr}}(\lambda)\, d\lambda = \gamma_{\mathrm{pr}}. \qquad (4.7.8)$$

Естественно, $\gamma \geqslant \gamma_{\mathrm{pr}}$.

Для заданной доверительной вероятности γ границы доверительного интервала λ_1 и λ_2 находятся согласно (4.7.3) и (4.7.4) из соотношений

$$\alpha_1 = \int_{-\infty}^{\lambda_1} f(\lambda \mid \lambda^*) d\lambda; \quad \alpha_2 = \int_{\lambda_2}^{\infty} f(\lambda \mid \lambda^*)\, d\lambda. \qquad (4.7.9)$$

Если при этом априорное распределение $f_{\mathrm{pr}}(\lambda)$ ограничено снизу и сверху величинами $\lambda_{1\mathrm{pr}}$ и $\lambda_{2\mathrm{pr}}$, то для верхней и нижней границ доверительного интервала должны выполняться условия

$$\lambda_{1\mathrm{pr}} \leqslant \lambda_1; \quad \lambda_2 \leqslant \lambda_{2\mathrm{pr}}. \qquad (4.7.10)$$

Пример 4.7.1. Рассмотрим байесовскую интервальную оценку среднего значения m гауссовской выборки объема n с известной дисперсией σ^2 на основе вычисления выборочного среднего $\bar{x}$ (3.4.5). Согласно п. 4.5.1 выборочное среднее имеет нормальное распределение вида

$$f(\bar{x} \mid m) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\frac{(\bar{x}-m)^2}{2\sigma^2/n}\right].$$

Пусть априорное распределение оцениваемого параметра m является прямоугольным (равномерным) в интервале от b до c:

$$f_{\mathrm{pr}}(m) = \begin{cases} 1/(c-b), & b \leqslant m \leqslant c, \\ 0, & m < b,\ m > c. \end{cases}$$

Тогда условная плотность вероятности оцениваемого параметра согласно (4.6.1) может быть записана в виде

$$f(m \mid \bar{x}) = \frac{\dfrac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\dfrac{n}{2\sigma^2}(m-\bar{x})^2\right]}{\Phi\left(\dfrac{c-\bar{x}}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\dfrac{b-\bar{x}}{\sqrt{\sigma^2/n}}\right)}, \quad b \leqslant \bar{x} \leqslant c.$$

Если задана доверительная вероятность γ, то верхняя и нижняя границы центрального доверительного интервала равны $m_{2,1} = \bar{x} \pm x_\gamma$, где величина x_γ определяется из уравнения

$$\Phi\left(\frac{x_\gamma}{\sqrt{\sigma^2/n}}\right) = \frac{\gamma}{2}\left[\Phi\left(\frac{c-\bar{x}}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\frac{b-\bar{x}}{\sqrt{\sigma^2/n}}\right)\right] + \frac{1}{2}.$$

При этом должны выполняться условия: $(\bar{x} - x_\gamma) \geqslant b$; $(\bar{x} + x_\gamma) \leqslant c$.

Вычислим границы центрального доверительного интервала для апостериорной доверительной вероятности $\gamma = 0{,}95$ применительно к выборке объема $n = 25$ с известной дисперсией $\sigma^2 = 1$ и значениями параметров симметричного априорного распределения оцениваемой величины m, равными $c = -b = 1$. Пусть на основе выборки вычислено выборочное среднее $\bar{x} = 0$. Подставляя указанные значения величин, входящих в последнее выражение, имеем $\Phi(t_\gamma = x_\gamma\sqrt{n}/\sigma) = 0{,}975$. По таблицам нормального распределения находим $t_\gamma \approx 1{,}96$. Следовательно, истинное значение m заключено в интервале $m_{2,1} = x_\gamma = \pm 1{,}96/5 = \pm 0{,}392$ с заданной апостериорной доверительной вероятностью $\gamma = 0{,}95$. При этом для заданных значений параметров априорного распределения априорная доверительная вероятность равна $\gamma_{\mathrm{pr}} = 0{,}4$.

4.7.2. Интервальное оценивание при неизвестном априорном распределении параметра

В большинстве прикладных задач математической статистики априорное распределение оцениваемого случайного параметра неизвестно или оцениваемый параметр является неизвестной постоянной величиной, которая для наблюдателя также может рассматриваться как случайная величина с неизвестным априорным распределением.

Рассмотрим условную плотность вероятности оценки $f_\lambda(\lambda^* \mid \lambda)$. При каждом фиксированном λ условная плотность вероятности $f_\lambda(\lambda^* \mid \lambda)$ задает распределение вероятностей величины λ^*. Для каждого значения λ^* аналогично (4.7.3) можно определить два числа $\lambda_{01} \leqslant \lambda^* \leqslant \lambda_{02}$, для которых справедливо равенство

$$\mathbf{P}\{\lambda_{01} \leqslant \lambda^* \leqslant \lambda_{02} \mid \lambda\} = \int_{\lambda_{01}}^{\lambda_{02}} f_\lambda(\lambda^* \mid \lambda)\, d\lambda^* = \gamma. \qquad (4.7.11)$$

Числа λ_{01} и λ_{02} зависят от истинного значения неизвестного параметра λ. Поэтому, если λ меняется, то точки (λ, λ_{01}) и (λ, λ_{02}) образуют в плоскости (λ, λ^*) две кривые, как показано на рис. 4.7.1. Пусть каждая из кривых пересекается с прямой, параллельной оси λ и проходящей через значение оценки λ^*, только в одной точке. Обозначим эти точки пересечения соответственно как $\lambda_1 = \lambda_1(\lambda, \gamma)$ и $\lambda_2 = \lambda_2(\lambda, \gamma)$ и пусть $C(\gamma)$ обозначает область (полосу), заключенную между двумя кривыми. Поскольку каждая точка (λ, λ^*) принадлежит области $C(\gamma)$, то для этой точки одновременно справедливы неравенства

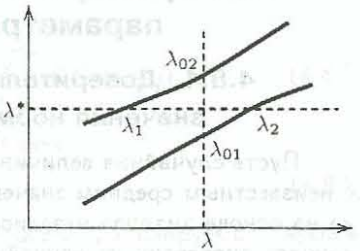


Рис. 4.7.1. Доверительные интервалы для неизвестного параметра

$$\lambda_{01}(\lambda, \gamma) \leqslant \lambda^* \leqslant \lambda_{02}(\lambda, \gamma); \qquad (4.7.12)$$

$$\lambda_1(\lambda^*, \gamma) \leqslant \lambda \leqslant \lambda_2(\lambda^*, \gamma). \qquad (4.7.13)$$

Поэтому равны и вероятности выполнения этих неравенств. Следовательно, при любом значении оцениваемого параметра λ из выражения (4.7.11) получаем, что при фиксированной оценке λ^*, полученной на основе выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$, истинное значение оцениваемого па-

раметра λ находится в границах, определяемых из соотношения

$$\mathbf{P}\{\lambda_1(\lambda^*,\gamma) \leqslant \lambda \leqslant \lambda_2(\lambda^*,\gamma)\} = \gamma. \qquad (4.7.14)$$

Отсюда видно, что переменный интервал (λ_1, λ_2) покрывает истинное значение оцениваемого параметра с заданной вероятностью γ.

Необходимо отметить, что если условная оценка параметра λ^* (4.7.6) нелинейно связана с истинным значением параметра λ, то переход от неравенства (4.7.12) к неравенству (4.7.13) может быть достаточно сложным. Кроме этого часто вместо плотности вероятности оценки λ^* удобнее использовать плотность вероятности некоторой нормированной величины, связанной с оценкой однозначной функциональной зависимостью. Как правило, выбирается такая функция $\Omega = \xi(\lambda, \lambda^*)$, выборочное значение которой не зависит от оцениваемого параметра λ. Подобную нормированную случайную величину Ω часто называют опорной случайной величиной. При этом желательно, чтобы для опорной случайной величины имелись таблицы функций распределения (например, нормальная, хи-квадрат, Стьюдента, дисперсионного отношения и др.).

В заключение данного параграфа необходимо отметить, что приведенные соотношения позволяют определить по заданным (полученным) каким-либо способам границам доверительного интервала доверительную вероятность (надежность) оцениваемого параметра.

4.8. Примеры интервального оценивания параметров распределений

4.8.1. Доверительный интервал для среднего значения нормального распределения

Пусть случайная величина X распределена по нормальному закону с неизвестным средним значением m, подлежащим интервальной оценке на основе анализа независимой выборки объема n. Рассмотрим два случая: дисперсия σ^2 случайной величины известна и неизвестна.

4.8.1.1. Известная дисперсия

Выборочное среднее $\bar{x}$ (3.4.5), основанное на анализе выборки объема n, также распределено по нормальному закону со средним значением $\mathrm{M}\{\bar{x}\} = m$ и дисперсией $D(\bar{x}) = \sigma^2/n$. При этом условные плотность вероятности и функция распределения имеют вид

$$f(\bar{x} \mid m) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\frac{(\bar{x}-m)^2}{2\sigma^2/n}\right]; \qquad (4.8.1)$$

$$F(\bar{x} \mid m) = \Phi\left(t = \frac{\bar{x}-m}{\sigma/\sqrt{n}}\right), \qquad (4.8.2)$$

где $\Phi(t)$ — интеграл от нормального распределения (1.3.22).

По заданной доверительной вероятности γ подберем такое положительное число

$$t_\gamma = x_\gamma/(\sigma\sqrt{n}), \qquad (4.8.3)$$

чтобы выполнялось условие

$$\mathbf{P}\left\{\left|t = \frac{\bar{x}-m}{\sigma/\sqrt{n}}\right| \leqslant t_\gamma = \frac{x_\gamma}{\sigma/\sqrt{n}}\right\} = 1 - 2\int_{t_\gamma}^{\infty} f(t)\,dt = 2\Phi(t_\gamma) - 1 = \gamma. \qquad (4.8.4)$$

Отсюда получаем уравнение для вычисления t_γ:

$$\Phi(t_\gamma) = \Phi\left(t_\gamma = \frac{x_\gamma}{\sigma/\sqrt{n}}\right) = \frac{1+\gamma}{2}. \qquad (4.8.5)$$

В силу симметричности распределения $f(t)$ относительно $t = 0$ выражение (4.8.4) можно переписать в равносильном виде, заменяя неравенство для модуля двойным неравенством

$$\mathbf{P}\left\{-t_\gamma \leqslant \frac{m-\bar{x}}{\sigma/\sqrt{n}} \leqslant t_\gamma\right\} = \mathbf{P}\left\{\bar{x} - \frac{t_\gamma\sigma}{\sqrt{n}} \leqslant m \leqslant \bar{x} + \frac{t_\gamma\sigma}{\sqrt{n}}\right\} = \gamma. \qquad (4.8.6)$$

Это соотношение показывает, что интервал $(\bar{x} \pm t_\gamma\sigma/\sqrt{n})$ с достоверностью γ является доверительным интервалом для среднего значения m гауссовской случайной величины с известной дисперсией σ^2.

Число t_γ представляет собой квантиль порядка $(1+\gamma)/2$ нормального распределения и подбирается по таблицам функции распределения $\Phi(t)$ или по таблицам вероятностей

$$\mathbf{P}\{t > t_\beta\} = 1 - \Phi(t_\beta) = \int_{t_\beta}^{\infty} f(t)\,dt = \beta, \qquad (4.8.7)$$

где применительно к рассматриваемой задаче

$$\beta = (1-\gamma)/2. \qquad (4.8.8)$$

Таким образом, если известна дисперсия анализируемой гауссовской случайной величины и оценивается ее выборочное среднее, то истинное среднее значение случайной величины с доверительной вероятностью γ лежит в интервале $\bar{x} \pm x_\gamma$. Смысл полученного результата состоит в следующем: с надежностью γ доверительный интервал

$$(\bar{x} - \sigma t_\gamma/\sqrt{n}; \bar{x} + \sigma t_\gamma/\sqrt{n})$$

покрывает неизвестный параметр m, т.е. выборочное среднее $\bar{x}$ дает оценку параметра m с точностью порядка $\pm x_\gamma = \pm\sigma t_\gamma/\sqrt{n}$ и надежностью γ. Доверительные границы и ширина доверительного интервала для рассматриваемого параметра соответственно равны

$$x_1 = \bar{x} - \sigma t_\gamma/\sqrt{n}; \quad x_2 = \bar{x} + \sigma t_\gamma/\sqrt{n}; \qquad (4.8.9)$$

$$\Delta = x_2 - x_1 = 2\sigma t_\gamma/\sqrt{n} = 2x_\gamma. \qquad (4.8.10)$$

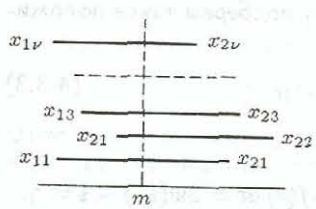


Рис. 4.8.1. Расположение доверительных интервалов для различных экспериментов

Из последнего выражения видно, что ширина доверительного интервала для заданных условий эксперимента и заданной надежности постоянна. Однако расположение доверительного интервала относительно среднего значения m зависит от случайного выборочного среднего $\bar{x}$ и меняется от эксперимента к эксперименту. На рис. 4.8.1 схематически друг над другом изображены доверительные интервалы, которые могут получаться для разных выборок j, $j = 1, 2, \ldots, \nu$, одного объема n, x_{1j} и x_{2j} — нижняя и верхняя границы доверительного интервала для j-го эксперимента.

Полученные соотношения позволяют определить минимальный объем выборки, соответствующий заданной доверительной вероятности γ и заданной ширине доверительного интервала $\Delta = 2x_\gamma$. Действительно, из выражения (4.8.3) находим

$$n = (t_\gamma \sigma / x_\gamma)^2, \tag{4.8.11}$$

где t_γ определяется по таблицам функции нормального распределения, исходя из формулы $\Phi(t_\gamma) = (1 + \gamma)/2$.

Пример 4.8.1. В качестве иллюстрации вычислим минимальный объем выборки n, обеспечивающий при доверительной вероятности $\gamma = 0{,}95$ и дисперсии $\sigma^2 = 0{,}25$ ширину доверительного интервала $\Delta = 2x_\gamma = 0{,}2$. По заданному значению доверительной вероятности по формуле (4.8.8) находим число $\beta = 0{,}025$, которому согласно табл. 1в Приложения 2 $\mathbf{P}\{t > t_\beta\} = 0{,}025$ соответствует значение аргумента $t_\gamma \equiv t_\beta = 1{,}96$. При этом необходимое минимальное число выборочных значений находим из (4.8.11): $n = 96$.

4.8.1.2. Неизвестная дисперсия

На основе выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ образуем статистику (случайную величину)

$$t_\nu = \frac{\bar{x} - m}{s_2/\sqrt{n}}, \tag{4.8.12}$$

являющуюся функцией от выборочного среднего $\bar{x}$ и выборочной дисперсии s_2^2. Перепишем последнее выражение в виде

$$t_\nu = \frac{\bar{x} - m}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma}\right)^2} \equiv \frac{z}{y}. \tag{4.8.13}$$

Здесь $z = (\bar{x} - m)/(\sigma/\sqrt{n})$ — нормально распределенная случайная величина с нулевым средним значением и единичной дисперсией; $\chi^2_{(n-1)} = (n-1)y^2$ — случайная величина, распределенная по закону хи-квадрат с $\nu = (n-1)$ степенями свободы. Выше (п. 3.4.2) было отмечено, что

выборочное среднее $\bar{x}$ и выборочная дисперсия s_2^2 гауссовской выборки независимы (3.4.32). Поэтому случайная величина t_ν имеет распределение Стьюдента $f(t)$ с $\nu = (n-1)$ степенями свободы (2.2.12).

Для заданной доверительной вероятности γ можем найти такое положительное число t_ν, чтобы выполнялось соотношение

$$\mathbf{P}\left\{\left|t_\nu = \frac{\bar{x} - m}{s_2/\sqrt{n}}\right| \leqslant t_\gamma\right\} = 1 - 2\int_{t_\gamma}^{\infty} f(t)\,dt = \gamma = 1 - 2\beta. \tag{4.8.14}$$

Здесь

$$\beta = \mathbf{P}\{t_\nu > t_{\beta;\nu} \equiv t_\gamma\} = \int_{t_{\beta;\nu}}^{\infty} f(t)\,dt = \frac{1-\gamma}{2}.$$

В табл. 3 Приложения 2 приведены величины $t_{\beta;\nu}$ для ряда значений параметра β и числа степеней свободы ν. В силу симметричности распределения $f(t)$ относительно $t = 0$ равенство (4.8.14) можно переписать в равносильном виде, заменяя неравенство для модуля двойным неравенством

$$\mathbf{P}\left\{-t_\gamma \leqslant \frac{m - \bar{x}}{s_2/\sqrt{n}} \leqslant t_\gamma\right\} = \mathbf{P}\left\{\bar{x} - \frac{t_\gamma s_2}{\sqrt{n}} \leqslant m \leqslant \bar{x} + \frac{t_\gamma s_2}{\sqrt{n}}\right\} = \gamma. \tag{4.8.15}$$

Последнее соотношение показывает, что интервал $(\bar{x} \pm t_\gamma s_2/\sqrt{n})$ с достоверностью γ есть доверительный интервал для среднего значения m случайной величины X.

Таблица 4.8.1

Отношение $\Delta_{ст}/\Delta_н$

n	3	5	7	10	15	20	30	50	70	100
$\gamma = 0{,}9$	1,76	1,29	1,18	1,11	1,07	1,05	1,03	1,02	1,01	1,00
$\gamma = 0{,}95$	2,19	1,42	1,25	1,15	1,09	1,07	1,04	1,03	1,02	1,01
$\gamma = 0{,}99$	3,85	1,79	1,44	1,26	1,16	1,11	1,07	1,04	1,03	1,02

Как отмечалось выше (п. 2.2.2), при $n \gg 1$ распределение Стьюдента асимптотически переходит в нормальное с теми же средним значением и дисперсией. Поэтому в практических приложениях часто доверительный интервал вычисляют исходя из приближения распределения Стьюдента нормальным распределением с дисперсией, равной выборочной несмещенной дисперсии s_2^2. При этом для заданной доверительной вероятности γ приближенный доверительный интервал $\Delta_н$ будет меньше точного доверительного интервала $\Delta_{ст}$. Разница тем больше, чем меньше объем выборки n и чем выше доверительная вероятность γ интервальной оценки. В табл. 4.8.1 приведено отношение $\Delta_{ст}/\Delta_н = t_{\gamma\,(ст)}/t_{\gamma\,(н)}$ для ряда значений n и γ, где $t_{\gamma\,(ст)}$ и $t_{\gamma\,(н)}$ — аргументы (критические значения) распределений Стьюдента и нормального распределения для $\beta = (1-\gamma)/2$. Из приведенных данных видно,

что использование нормального приближения при определении доверительного интервала (следовательно, и его границ) допустимо лишь при достаточно большом объеме выборки.

Пример 4.8.2. В качестве иллюстрации рассмотренной оценки найдем доверительный интервал с доверительной вероятностью $\gamma = 0{,}95$ при анализе выборки объема $n = 9$, для которой выборочные среднее и дисперсия соответственно равны $\bar{x} = 12{,}7$ и $s_2^2 = 0{,}25$. Из таблиц распределения Стьюдента для $\beta = (1-\gamma)/2 = 0{,}025$ и степеней свободы $\nu = (n-1) = 8$ находим $t_\gamma = t_{0{,}025;8} = 2{,}306$. Следовательно, доверительные границы и ширина доверительного интервала для приведенных данных соответственно равны: $x_1 \approx 12{,}32$; $x_2 \approx 13{,}08$; $\Delta \approx 0{,}76$. При известной дисперсии того же значения ($\sigma^2 = 0{,}25$) интервальная оценка выборочного среднего значения, вычисленная согласно (4.8.9) и (4.8.10) с использованием табл. 16 Приложения 2, дает следующие значения искомых величин: $t_\gamma = 1{,}96$; $x_1 \approx 12{,}37$; $x_2 \approx 13{,}03$; $\Delta \approx 0{,}66$. Таким образом, незнание дисперсии привело к увеличению ширины доверительного интервала (применительно к рассмотренным исходным данным) примерно в 1,15 раза.

4.8.2. Доверительный интервал для дисперсии нормального распределения

Пусть случайная величина X распределена по нормальному закону с неизвестной дисперсией σ^2, подлежащей интервальной оценке на основе анализа независимой выборки объема n.

Рассмотрим два случая: среднее значение m случайной величины известно и неизвестно.

4.8.2.1. Известное среднее значение m

Выборочная дисперсия s_1^2 определяется соотношением (3.4.11). Образуем случайную величину

$$\chi_n^2 = \frac{s_1^2 n}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{x_i - m}{\sigma} \right)^2, \qquad (4.8.16)$$

прямо пропорциональную сумме квадратов n независимых нормально распределенных случайных величин с нулевыми средними значениями и единичными дисперсиями. Подобная случайная величина имеет распределение хи-квадрат (2.2.3) с $\nu = n$ степенями свободы. Это распределение зависит только от объема выборки n. Распределение для выборочной дисперсии s_1^2 может быть получено из распределения хи-квадрат (2.2.3) при помощи линейного преобразования

$$s_1^2 = \sigma^2 \chi_n^2 / n. \qquad (4.8.17)$$

Выражение для плотности вероятности выборочной дисперсии приведено в (2.2.6а), где $z = s_1^2$.

Подберем по заданной доверительной вероятности γ положительные числа $\chi^2_{\alpha_1,n}$ и $\chi^2_{\alpha_2,n}$ так, чтобы выполнялось условие

$$\mathbf{P}\{\chi^2_{\alpha_1,n} \leqslant \chi_n^2 = (ns_1^2/\sigma^2) \leqslant \chi^2_{\alpha_2,n}\} =$$
$$= \mathbf{P}\{\chi_n^2 > \chi^2_{\alpha_1,n}\} - \mathbf{P}\{\chi_n^2 > \chi^2_{\alpha_2,n}\} = \gamma. \qquad (4.8.18)$$

Неравенство в фигурных скобках равносильно неравенству

$$\frac{1}{\chi^2_{\alpha_1,n}} \geqslant \frac{\sigma^2}{ns_1^2} \geqslant \frac{1}{\chi^2_{\alpha_2,n}}.$$

Следовательно, условие (4.8.18) можно записать в виде

$$\mathbf{P}\left\{ \frac{ns_1^2}{\chi^2_{\alpha_2,n}} \leqslant \sigma^2 \leqslant \frac{ns_1^2}{\chi^2_{\alpha_1,n}} \right\} = \gamma. \qquad (4.8.19)$$

Последнее выражение связывает между собой выборочную дисперсию, объем выборки, доверительную вероятность и доверительные границы для оценки истинного значения дисперсии.

Для принятого условия равенства вероятностей того, что оцениваемый параметр будет лежать левее нижней границы $\chi^2_{\alpha_1,n}$ и правее верхней границы $\chi^2_{\alpha_2,n}$, последние определяются соответственно из выражений

$$\int_{\chi^2_{\alpha_1,n}}^{\infty} f(\chi_n^2)\, d\chi_n^2 = \frac{1+\gamma}{2};$$
$$\int_{\chi^2_{\alpha_2,n}}^{\infty} f(\chi_n^2)\, d\chi_n^2 = \frac{1-\gamma}{2}. \qquad (4.8.20)$$

Пример 4.8.3. В качестве иллюстрации полученных соотношений вычислим доверительный интервал для оценки дисперсии по выборке объема $n = 9$ при доверительной вероятности $\gamma = 0{,}95$. Применительно к приведенным данным $\mathbf{P}\{\chi_n^2 > \chi^2_{\alpha_1,n}\} = (1+\gamma)/2 = 0{,}975$, $\mathbf{P}\{\chi_n^2 > \chi^2_{\alpha_2,n}\} = (1-\gamma)/2 = 0{,}025$. По табл. 2 Приложения 2 для заданных значений n и γ находим $\chi^2_{\alpha_1,n} = \chi^2_{0{,}975;9} = 2{,}70$; $\chi^2_{\alpha_2,n} = \chi^2_{0{,}025;9} = 19{,}02$. Следовательно, истинное значение дисперсии с доверительной вероятностью $\gamma = 0{,}95$ лежит в интервале $0{,}473\, s_1^2 \leqslant \sigma^2 \leqslant 3{,}333\, s_1^2$, т.е. ширина доверительного интервала оцениваемой дисперсии равна $\Delta_1 = 2{,}86\, s_1^2$.

4.8.2.2. Неизвестное среднее значение m

Изучим второй случай нахождения доверительного интервала для дисперсии нормального распределения при неизвестном среднем значении m. В этом случае выборочная дисперсия определяется соотношением (3.4.27), где выборочное среднее значение $\bar{x}$ находится из (3.4.5). Рассмотрим сумму квадратов нормированных гауссовских случайных величин, имеющих нулевое среднее значение и дисперсию $[(n-1)/n]$:

$$\chi^2_{(n-1)} = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 = \frac{n-1}{\sigma^2} s_2^2. \qquad (4.8.21)$$

Эта сумма в противоположность случаю известного среднего не является суммой независимых слагаемых, так как между n случайными величи-

нами $(x_i - \bar{x})$ имеется линейное соотношение вида

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0. \qquad (4.8.22)$$

Таким образом, случайные величины $(x_i - \bar{x})$ являются линейно зависимыми гауссовскими случайными величинами. Однако $\chi^2_{(n-1)}$ можно преобразовать в сумму квадратов $(n-1)$ статистически независимых нормированных случайных величин. Тогда функция распределения случайной величины $\chi^2_{(n-1)}$ будет иметь тот же вид, что и функция распределения случайной величины χ^2_n (4.8.16) для суммы n квадратов статистически независимых нормированных гауссовских случайных величин. Единственное различие между функциями распределений случайных величин $\chi^2_{(n-1)}$ и χ^2_n состоит в том, что число степеней свободы ν для хи-квадрат распределения случайной величины $\chi^2_{(n-1)}$ равно $\nu = n-1$ вместо $\nu = n$ для распределения случайной величины χ^2_n. Соответственно и плотность вероятности выборочной дисперсии при неизвестном среднем значении также определяется выражением (2.2.6а) для $\nu = n-1$. Вероятность выполнения неравенства (4.8.19), определяющего доверительные границы для истинного значения дисперсии при заданном уровне надежности γ, записывается в виде

$$\mathbf{P}\left\{\frac{(n-1)s_2^2}{\chi^2_{\alpha_2,(n-1)}} \leqslant \sigma^2 \leqslant \frac{(n-1)s_2^2}{\chi^2_{\alpha_1,(n-1)}}\right\} = \gamma, \qquad (4.8.23)$$

где $\chi^2_{\alpha_1,(n-1)}$ и $\chi^2_{\alpha_2,(n-1)}$ определяются аналогично (4.8.20) при подстановке вместо $f(\chi^2_n)$ плотности вероятности $f(\chi^2_{(n-1)})$.

Уменьшение числа степеней свободы распределения исследуемой статистики в зависимости от количества связей между независимыми выборочными значениями более подробно будет рассмотрено ниже в п. 4.9.2.

Применительно к данным иллюстративного примера 4.8.2 предыдущего пункта нормированные значения границ интервальной оценки дисперсии с неизвестным средним значением равны $\chi^2_{0,025;8} = 2{,}18$ и $\chi^2_{0,975;8} = 17{,}53$. Поэтому, учитывая (4.8.23), истинное значение дисперсии заключено в интервале $0{,}456\,s_2^2 \leqslant \sigma^2 \leqslant 3{,}670 s_2^2$, что соответствует относительной ширине доверительного интервала $\Delta_2/s_2^2 = 3{,}214$. Сравнивая это значение ширины доверительного интервала с аналогичной величиной $\Delta_1/s_1^2 = 2{,}860$, вычисленной в примере предыдущего пункта, видим, что из-за незнания среднего значения анализируемой выборки происходит увеличение (применительно к принятым количественным значениям исходных данных) ширины доверительного интервала примерно на 12 %.

Так как среднее квадратическое отклонение s случайной величины функционально и однозначно связано с дисперсией s^2, то доверительные границы для среднего квадратического отклонения s равны положительному значению квадратного корня из доверительных границ для дисперсии.

4.8.3. Доверительный интервал для вероятности p биномиального закона распределения

Пусть наблюдается независимая выборка объема n, распределенная по биномиальному закону (1.3.8) с неизвестным значением элементарной вероятности p. Как показано в п. 4.5.4, общее число успехов k^* и соответственно выборочная частота успехов (оценка неизвестной вероятности p) $p^* = k^*/n$ являются достаточными статистиками. Статистика числа положительных исходов имеет биномиальное распределение

$$p_n(k^*) = C_n^{k^*} p^{k^*}(1-p)^{n-k^*}, \quad k^* = 0, 1, .., n.$$

По наблюденной частоте $p^* = k^*/n$, которая является конкретным значением случайной величины P^*, и заданной доверительной вероятности γ можно определить нижнюю и верхнюю доверительные границы p_1 и p_2 для истинного значения вероятности p. Эти границы определяются из соотношения

$$\mathbf{P}\{p_1 \leqslant p \leqslant p_2\} = 1 - \mathbf{P}\{p > p_2\} - \mathbf{P}\{p < p_1\} \geqslant \gamma. \qquad (4.8.24)$$

Вероятности $\mathbf{P}\{p > p_2\}$ и $\mathbf{P}\{p < p_1\}$ должны являться функциями k^*, n и γ: $p_1 = \psi_1(\gamma, p^* = k^*/n)$, $p_2 = \psi_2(\gamma, p^* = k^*/n)$. Замена равенства после фигурных скобок на знак $\geqslant$ обусловлена тем, что биномиальное распределение является дискретным. При вычислении границ учтем, что доверительные интервалы являются центральными.

Согласно п.4.7.2 для любого заданного p оно попадает в доверительную область (полосу) C с вероятностью γ тогда и только тогда, когда наблюденная его статистика $p^* = k^*/n$ находится в этой же области (полосе). Справедливо и противоположное утверждение. Поэтому для нижней и верхней границ p_1 и p_2 справедливы соотношения

$$\mathbf{P}\{p < p_1\} = \mathbf{P}\{P^* > p^*\} \leqslant (1-\gamma)/2;$$
$$\mathbf{P}\{p > p_2\} = \mathbf{P}\{P^* < p^*\} \leqslant (1-\gamma)/2.$$

Из этих соотношений видно с учетом дискретности биномиального распределения оценки p^* для определения границ доверительного интервала p_1 и p_2 необходимо решить неравенства

$$\mathbf{P}\{P^* > p^*\} = \sum_{k=k^*}^{n} C_n^k p_1^k (1-p_1)^{n-k} \leqslant \frac{1-\gamma}{2}; \qquad (4.8.25)$$

$$\mathbf{P}\{P^* < p^*\} = \sum_{k=0}^{k^*} C_n^k p_2^k (1-p_2)^{n-k} \leqslant \frac{1-\gamma}{2} \qquad (4.8.26)$$

относительно величин p_1 и p_2. При этом необходимо, чтобы неравенства для p_1 и p_2 были как можно ближе к равенствам. Строгое решение задачи нахождения нижней и верхней доверительных границ для вероятности p достаточно сложно и решается, как правило, с использованием

соответствующих таблиц. В [8, 51, 72] приведены границы двухстороннего центрального доверительного интервала p_1 и p_2 для ряда значений доверительной вероятности γ в зависимости от объема выборки n числа положительных исходов k^*.

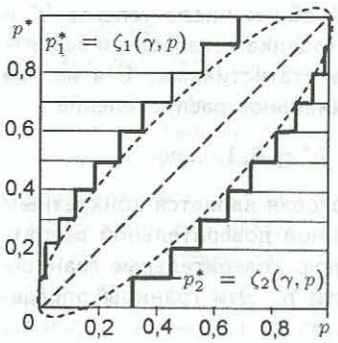


Рис. 4.8.2. Доверительные интервалы

Применительно к рассматриваемой задаче доверительные интервалы для неизвестной вероятности p можно представить в виде, приведенном на рис. 4.8.2 для $n = 10$. В силу дискретности частот $p^* = k^*/n$ кривые $p_1^* = \varsigma_1(\gamma, p)$ и $p_2^* = \varsigma_2(\gamma, p)$ являются ступенчатыми, причем высота всех ступеней равна $1/n$, а ширина каждой ступеньки равна длине соответствующего интервала значений p, при которых $p_1^* = \varsigma_1(\gamma, p)$ и $p_2^* = \varsigma_2(\gamma, p)$ сохраняют постоянное значение до очередного изменения скачком на $1/n$ [52]. Границами доверительного интервала p_1 и p_2 являются точки пересечения прямой $p^* = \text{const}$ ступенчатых кривых p_1^* и p_2^*.

В прикладных задачах обычно используют различные приближения, основанные на аппроксимации биномиального распределения нормальным распределением (1.3.18) со средним значением np и дисперсией $np(1-p)$ при условии достаточно больших значений np и $n(1-p)$.

Рассмотрим нормированную гауссовскую случайную величину

$$z = (k^* - np)/\sqrt{np(1-p)},$$

для которой при заданной вероятности $\gamma = 1 - \alpha$ и центральном интервале можем записать

$$\mathbf{P}\{-a \leqslant z \leqslant a\} = \gamma = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} \exp(-z^2/2)\, dz = 2\Phi(a) - 1, \quad (4.8.27)$$

где величина a определяется из выражения

$$\Phi(a) = (1+\gamma)/2.$$

Неравенство в фигурных скобках (4.8.27) эквивалентно двум неравенствам:

$$k^* \leqslant np + a\sqrt{np(1-p)}; \quad k^* \geqslant np - a\sqrt{np(1-p)}.$$

Учитывая $p^* = k^*/n$, граничные точки доверительного интервала p_1 и p_2 представляют собой корни квадратного уравнения

$$(p^* - p)^2 = a^2 p(1-p)/n, \quad (4.8.28)$$

которое преобразуется к виду

$$(1 + a^2/n)p^2 - 2(p^* + a^2/2n)p + p^{*2} = 0.$$

Корни этого уравнения

$$p_2, p_1 = \frac{p^* + a^2/2n \pm a\sqrt{p^*(1-p^*)/n + a^2/4n^2}}{1 + a^2/n}. \quad (4.8.29)$$

Полученные приближенные значения границ для неизвестной вероятности биномиального распределения дают приемлемую точность для практических приложений при np и $n(1-p)$ не менее 4 [13].

Приведенному приближенному решению можно дать геометрическую интерпретацию. Понимая под вероятностью p и ее оценкой p^* абсциссу и ординату точки в системе координат $p0p^*$, выражение (4.8.28) представляет собой уравнение эллипса, примерный вид которого показан на рис. 4.8.2. Эллипс проходит через точки (0,0) и (1,1) и касается вертикальных сторон квадрата. Форма эллипса зависит от параметра a, однозначно связанного с доверительной вероятностью γ и числа опытов n. Прямая с уравнением $p^* = \text{const}$ пересекает границы эллипса в двух точках p_1 и p_2, которые и являются границами доверительного интервала. Чем больше число опытов n, тем уже эллипс и тем меньше ширина доверительного интервала. В пределе для $n \to \infty$ эллипс переходит в отрезок прямой $p^* = p$, что соответствует нулевой ширине доверительного интервала при $\gamma \to 1$.

Кроме рассмотренного выше приближения, для самой грубой (но часто применяемой) оценки границ доверительного интервала полагают, что частота p^* (оценка неизвестной вероятности) распределена по нормальному закону со средним значением p^* и дисперсией $p^*(1-p^*)/n$. Тогда границы приближенно определяются выражением

$$p_2, p_1 \approx p^* \pm a\sqrt{p^*(1-p^*)/n}.$$

Пример 4.8.4. Для иллюстрации полученных выражений вычислим границы доверительного интервала оцениваемой элементарной вероятности p биномиального распределения для следующих исходных данных: $n = 20$, число положительных исходов $k^* = 6$ ($p^* = 0{,}3$), доверительная вероятность $\gamma = 0{,}95$ ($a = 1{,}96$). Точные значения границ доверительного интервала, найденные на основе соответствующих таблиц [8, 51], равны (0,199; 0,543). Приближенные формулы дают соответственно (0,173; 0,619) и (0,099; 0,501), т.е. при заданном объеме выборки и заданной доверительной вероятности доверительный интервал, вычисленный по приближенным формулам, превышает доверительный интервал, найденный на основе более строгого решения.

4.8.4. Доверительный интервал для параметра λ пуассоновского распределения

Задача интервальной оценки параметра λ пуассоновского распределения решается аналогично оценке параметра p биномиального распределения, рассмотренной выше в п. 4.8.3. Верхняя λ_2 и нижняя λ_1 границы центрального доверительного интервала с доверительной веро-

ятностью $\gamma = 1 - \alpha$ определяются решением уравнений

$$\sum_{k=0}^{k^*} \frac{\lambda_2^k \exp(-\lambda_2)}{k!} \leqslant \frac{1-\gamma}{2}; \quad \sum_{k=k^*}^{\infty} \frac{\lambda_1^k \exp(-\lambda_1)}{k!} \leqslant \frac{1-\gamma}{2}. \tag{4.8.30}$$

где, как и выше, k^* — экспериментально полученное число положительных исходов. Решение этих уравнений относительно неизвестных доверительных границ λ_1 и λ_2 возможно с применением функции распределения хи-квадрат [8, 51, 72], где также приведены соответствующие таблицы для нахождения нижней λ_1 и верхней λ_2 границ для ряда значений доверительной вероятности γ в зависимости от объема выборки n и числа положительных исходов k^*.

Приближенное решение задачи определения доверительных границ для параметра λ основано на аппроксимации числа положительных исходов $k^* \equiv \lambda^*$ нормальным распределением (при $\lambda \gg 1$) со средним значением $\mathrm{M}\{\lambda^*\} = \lambda$ и дисперсией $D(\lambda^*) = \lambda$ (при анализе одной серии наблюдений):

$$p(\lambda^*) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left[-\frac{(\lambda^* - \lambda)^2}{2\lambda}\right].$$

Аналогично предыдущей задаче, вводя нормированную переменную $z = (\lambda^* - \lambda)/\sqrt{\lambda}$ и учитывая соотношение вида (4.8.27), оценка параметра должна удовлетворять двум неравенствам:

$$\lambda^* \leqslant \lambda + a\sqrt{\lambda}; \quad \lambda^* \geqslant \lambda - a\sqrt{\lambda}.$$

Поэтому граничные точки доверительного интервала представляют собой корни квадратного уравнения

$$(\lambda^* - \lambda)^2 = a^2\lambda.$$

Корни этого уравнения

$$\lambda_2, \lambda_1 = \frac{a^2 + 2\lambda^* \pm a\sqrt{a^2 + 4\lambda^*}}{2}. \tag{4.8.31}$$

Для грубой оценки границ доверительного интервала полагают, как и выше, что оцениваемый параметр распределен по нормальному закону со средним значением и дисперсией, равными λ^*. Тогда приближенное выражение для границ доверительного интервала будет иметь вид

$$\lambda_2, \lambda_1 \approx \lambda^* \pm a\sqrt{\lambda^*}. \tag{4.8.32}$$

Полученные приближенные значения границ для неизвестного значения параметра λ пуассоновского распределения дают приемлемую точность для практических приложений при $\lambda \geqslant 5$.

Пример 4.8.5. Найдем границы доверительного интервала для параметра λ пуассоновского распределения для числа положительных исходов $k^* = 10$ и доверительной вероятности $\gamma = 0{,}95$ ($a = 1{,}96$). Точные значения границ, полученные на основе таблиц [8, 51, 72], равны (4,80; 18,39). Приближенные значения границ, вычисленные согласно (4.10.31) и (4.10.32), равны соответственно (5,43; 18,41) и (3,80; 16,20).

4.9. Степени свободы распределений случайных величин

Как отмечалось выше, при рассмотрении распределений, основанных на соответствующих преобразованиях гауссовских случайных величин, под числом степеней свободы понималось число, совпадающее с числом независимых выборочных значений n. Однако в большинстве задач математической статистики число независимых выборочных значений n и число свободных величин распределения $(n - g)$ не совпадают. Для определения распределений сумм квадратов случайных величин x_i, имеющих g линейных связей, используются ортогональные преобразования переменных.

4.9.1. Основные положения ортогонального преобразования

Линейное преобразование переменных $\mathbf{x}(x_1, x_2, \ldots, x_n)$

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1i}x_i + \ldots + a_{1n}x_n; \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2i}x_i + \ldots + a_{2n}x_n; \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{ni}x_i + \ldots + a_{nn}x_n \end{aligned} \tag{4.9.1}$$

называется *ортогональным*, если оно сохраняет инвариантной (неизменной) соотношение (форму) вида

$$\sum_{k=1}^{n} y_k^2 = \sum_{k=1}^{n} x_k^2. \tag{4.9.2}$$

При этом для коэффициентов a_{ij} должно выполняться условие

$$\sum_{k=1}^{n} a_{ki}a_{kj} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j. \end{cases} \tag{4.9.3}$$

Подставляя в выражение (4.9.2) вместо y_k их линейные преобразования и приравнивая коэффициенты при x_i^2 и x_ix_j с обеих сторон, получим другой вид условия ортогональности:

$$\begin{aligned} &a_{1i}^2 + a_{2i}^2 + \ldots + a_{ji}^2 + \ldots + a_{ni}^2 = 1; \\ &a_{1i}a_{1j} + a_{2i}a_{2j} + \ldots + a_{ni}a_{nj} = 0, \quad i \neq j;\ i, j = 1, 2, \ldots, n. \end{aligned} \tag{4.9.4}$$

Если исходные уравнения (4.9.1) умножить соответственно на a_{1i}, a_{2i}, ..., a_{ni} и затем сложить, то в силу условий ортогональности (4.9.3)

коэффициенты у всех x кроме x_i равны нулю, а коэффициент при x_i равен единице. В результате для обратного преобразования имеем

$$x_i = a_{1i}y_1 + a_{2i}y_2 + \ldots + a_{ji}y_j + \ldots + a_{ni}y_n, \quad i,j = 1,2,\ldots,n. \tag{4.9.5}$$

В силу условия (4.9.2) обратное преобразование снова будет ортогональным, и для него условия ортогональности примут вид

$$\sum_{k=1}^{n} a_{jk}a_{ik} = \begin{cases} 1 & \text{при } i = j, \\ 0 & \text{при } i \neq j, \end{cases} \quad i,j = 1,2,\ldots,n. \tag{4.9.6}$$

Ортогональное преобразование можно рассматривать как аналитическое выражение преобразования координат в n-мерном пространстве, которое осуществляется вращением осей прямоугольной системы координат около фиксированного начала координат. Поэтому согласно (4.9.2) расстояние $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ от начала координат до точки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ инвариантно (постоянно) относительно всякого подобного вращения.

Пусть $x_1, x_2, \ldots, x_n$ — независимые случайные гауссовские выборочные значения с нулевыми средними значениями и одинаковыми дисперсиями σ^2. Рассмотрим новые случайные величины $y_1, y_2, \ldots, y_n$, которые линейно связаны с исходной выборкой:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{ij}x_j + \ldots + a_{in}x_n, \quad i,j = 1,2,\ldots,n. \tag{4.9.7}$$

Так как преобразование (4.9.7) линейно, то новые случайные величины y_i также являются гауссовскими величинами с нулевыми средними значениями и дисперсиями

$$\mathrm{M}\{y_i^2\} = \sum_{j=1}^{n} a_{ij}^2 \mathrm{M}\{x_j^2\} = \sigma^2 \sum_{j=1}^{n} a_{ij}^2 = \sigma^2. \tag{4.9.8}$$

Преобразование (4.9.7) является ортогональным, поэтому в соответствии с (4.9.6) справедлива запись

$$\mathrm{M}\{y_i y_k\} = \sigma^2 \sum_{j=1}^{n} a_{ij}a_{kj} = \begin{cases} \sigma^2 & \text{при } i = k, \\ 0 & \text{при } i \neq k. \end{cases} \tag{4.9.9}$$

Таким образом, новые случайные величины y_i так же, как и исходные случайные величины x_i, являются независимыми нормально распределенными случайными величинами с нулевыми средними значениями и одинаковыми дисперсиями σ^2.

Так как геометрический смысл ортогонального преобразования соответствует вращению координатной системы около начала координат, то рассмотренное нормальное распределение инвариантно относительно такого вращения.

4.9.2. Степени свободы

Используем ортогональное преобразование применительно к рассмотренной выше в п. 4.8.2.2 оценке выборочной дисперсии гауссовской случайной величины при неизвестном среднем значении. С целью упрощения выводов без ограничения общности будем полагать, что истинное среднее значение $m_X = 0$, так как это не меняет значения выборочной дисперсии s_2^2 и эквивалентно прибавлению ко всем элементам выборки x_i некоторой постоянной величины.

Введем случайную величину

$$z = (n-1)s_2^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \tag{4.9.10}$$

Рассмотрим ортогональное преобразование (4.9.1), в котором в качестве первой строки возьмем выражение

$$y_1 = x_1/\sqrt{n} + x_2/\sqrt{n} + \ldots + x_n/\sqrt{n} = \sqrt{n}\bar{x}. \tag{4.9.11}$$

При этом для коэффициентов a_{1i} справедливо первое условие ортогональности (4.9.4): $a_{11}^2 + a_{12}^2 + \ldots + a_{1n}^2 = 1$. Принципиально можно подобрать коэффициенты a_{ik}, $i,k = 2,3,\ldots,n$, согласно (4.9.3) такие, что преобразование вида (4.9.1) с учетом (4.9.11) будет ортогональным [13, 37]. Согласно (4.9.2) можно записать

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=2}^{n} y_i^2 + y_1^2 = \sum_{i=2}^{n} y_i^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2.$$

Отсюда

$$\sum_{i=2}^{n} y_i^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{4.9.12}$$

Случайная величина

$$\chi^2_{(n-1)} = \frac{1}{\sigma^2}\sum_{i=2}^{n} y_i^2 = \frac{n-1}{\sigma^2}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{\sigma^2}s_2^2 \tag{4.9.13}$$

распределена по закону хи-квадрат с $\nu = (n-1)$ *степенями свободы*. При этом распределение выборочной дисперсии $s_2^2 = \sigma^2\chi^2_{(n-1)}/(n-1)$ определяется выражением (2.2.6а), в котором $\nu = (n-1)$.

Так как величины $y_1 = \bar{x}\sqrt{n}$, $y_2, y_3, \ldots, y_n$ независимы, то независимы и различные преобразования от этих величин. В частности, для этих величин справедливо тождество

$$\mathrm{M}\left\{y_1 \sum_{i=2}^{n} y_i^2\right\} = \mathrm{M}\{y_1y_2^2 + y_1y_3^2 + \ldots + y_1y_n^2\} = 0.$$

Величины y_1 и $\sum y_i^2$ при $i = 2, 3, .., n$ однозначно связаны с выборочным средним и выборочной дисперсией, поэтому можно утверждать, что выборочное среднее $\bar{x}$ и выборочная дисперсия s_2^2 также независимы между собой, т.е.

$$f_2(\bar{x}, s_2^2) = f_{\bar{x}}(\bar{x}) f_{s_2^2}(s_2^2). \tag{4.9.14}$$

Полученное выражение для распределения квадратичной формы от исходной независимой гауссовской выборки объема n обобщается на случай, когда имеется не одна, а несколько (g) линейных уравнений связи для выборочных значений. Соответствующим выбором коэффициентов a_{ij} подобно (4.9.11) уравнения связи можно свести к g выражениям вида (4.9.7).

Итак, пусть задано некоторое число $g < n$ линейных функций y_i, а весовые коэффициенты a_{ij} удовлетворяют условиям ортогональности (4.9.6)

$$\sum_{j=1}^{n} a_{ij} a_{kj} = \begin{cases} 1 & \text{при } i = k, \\ 0 & \text{при } i \neq k \end{cases} \tag{4.9.15}$$

для $i = 1, 2, \ldots, g;\ k = 1, 2, \ldots, g$.

В теории ортогональных преобразований доказывается [37 и др.], что можно найти $(n - g)$ строк чисел $a_{i1}, a_{i2}, \ldots, a_{in}$, где $i = g + 1, g+2, \ldots, n$, таких, чтобы линейные функции (4.9.7) были ортогональными. При этих предположениях квадратичная форма величин $x_1, x_2, \ldots, x_n$ может быть представлена в виде

$$Q(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i^2 - (y_1^2 + y_2^2 + \ldots + y_g^2).$$

Учитывая условие ортогональности (4.9.2), последнее выражение можно записать как

$$Q(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} y_i^2 - (y_1^2 + y_2^2 + \ldots + y_g^2) =$$

$$= y_{g+1}^2 + y_{g+2}^2 + \ldots + y_n^2 = \sum_{i=g+1}^{n} y_i^2. \tag{4.9.16}$$

Таким образом, введенная величина Q равна сумме квадратов $(n-g)$ независимых гауссовских случайных величин с нулевыми средними значениями и одинаковыми дисперсиями σ^2. Кроме того, эта сумма не зависит от $y_1, y_2, \ldots, y_g$. Поэтому случайная величина $U = Q/\sigma$ имеет распределение хи-квадрат (2.2.3) с параметром $\nu = n - g$.

Число $n - g$ является наименьшим числом независимых величин, к которым может быть приведена величина Q с помощью линейного преобразования. В статистических приложениях число $n - g$ называют *числом степеней свободы* распределения рассматриваемых случайных величин.

Понятия степеней свободы заимствованы из области физических наук, где число степеней свободы характеризует число независимых движений объекта в многомерном пространстве. Так, например, если объект имеет возможность двигаться только прямолинейно, он обладает одной степенью свободы; при возможном движении по плоскости в двух направлениях объект обладает двумя степенями свободы и т.д.

В математической статистике доказывается следующее положение [13, 15, 33, 37 и др.]. Если квадратичная форма нормальной случайной величины X имеет $r = (n - g)$ степеней свободы, то с помощью ортогонального преобразования она может быть представлена в виде сумм соответствующих независимых квадратичных форм, т.е. справедливо выражение

$$Q = \sum_{i=1}^{r} y_i^2 = Q_1 + Q_2 + \ldots + Q_k, \tag{4.9.17}$$

где

$$Q_1 = \sum_{i=1}^{r_1} y_i^2;\ Q_2 = \sum_{i=r_1+1}^{r_1+r_2} y_i^2; \ldots;\ Q_k = \sum_{i=r-r_k+1}^{r} y_i^2, \quad r = \sum_{i=1}^{k} r_i. \tag{4.9.18}$$

Здесь никакие две различные квадратичные формы Q_i и Q_j $(i, j = 1, 2 \ldots, k;\ i \neq j)$ не содержат общей переменной y.

4.10. Толерантные (допустимые) интервалы

Особым классом доверительных интервалов являются толерантные или допустимые интервалы. Суть толерантных интервалов состоит в следующем. На основе независимой выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$, взятой из некоторой генеральной совокупности, характеризуемой одномерной плотностью вероятности $f(x)$, необходимо найти такие границы переменной x, которые содержат долю распределения не меньше p с заданной доверительной вероятностью (надежностью) γ. Эти границы называются *толерантными границами*, а интервал значений случайной величины X, заключенный внутри них, называют *толерантным интервалом*. Толерантные границы, а соответственно и толерантные интервалы, могут быть как односторонними, так и двухсторонними. В качестве примера можно привести следующую ситуацию. Измеряется уровень радиосигнала. Необходимо по ограниченному числу экспериментов n определить такую нижнюю границу уровня, чтобы все возможные уровни радиосигналов (при «неизменных» условиях радиосвязи) были выше этого минимального уровня с вероятностью не менее заданной p при достоверности полученного результата γ (например, $p = 0{,}95$; $\gamma = 0{,}9$).

Рассмотрим несколько подробнее сформулированную задачу применительно к двусторонним толерантным интервалам. Итак пусть требуется определить двусторонние толерантные границы L_1 и L_2 (пусть $L_2 > L_1$), которые, естественно, зависят от заданных величин: доли распределения p, доверительной вероятности γ и объема выборки n.

Если бы все параметры распределения $f(x)$ были известны, то с вероятностью γ=1 границы L_1 и L_2 вычислялись бы на основе соотношения

$$\int_{L_1}^{L_2} f(x)\,dx = p.$$

Применительно к центральным интервалам границы L_1 и L_2 определяются из выражения

$$\int_{-\infty}^{L_1} f(x)\,dx = \int_{L_2}^{\infty} f(x)\,dx = \frac{1-p}{2}. \qquad (4.10.1)$$

Если параметры распределения не известны и определяются на основе конечной выборки, то границы L_1 и L_2 будут зависеть от выборочных значений неизвестных параметров, т.е. являются функциями от случайных величин (оценок параметров). Естественно, если объем выборки достаточно велик (в пределе $n \to \infty$), то, как правило, оценки неизвестных параметров являются асимптотически несмещенными и совпадают с истинными значениями, т.е. являются состоятельными. В данном асимптотическом случае распределение выборок будет известно точно, и границы L_1 и L_2 определяются согласно (4.10.1).

Конкретизируем постановку и решение задачи применительно к нормальному распределению $f(x)$ с неизвестными средним значением и дисперсией при анализе n независимых элементов выборки. Выше было показано, что состоятельными оценками указанных неизвестных параметров являются выборочное среднее $\bar{x}$ и выборочная дисперсия s^2. При этом они взаимно независимы и распределены, соответственно, по нормальному закону и закону (2.2.6а) с $\nu = n-1$ степенями свободы. По вычисленным выборочным характеристикам симметричные толерантные границы определяются выражениями

$$L_1 = \bar{x} - ks; \quad L_2 = \bar{x} + ks, \qquad (4.10.2)$$

где k — некоторая константа, которая выбирается таким образом, чтобы с доверительной вероятностью γ интервал (L_1, L_2) покрывал часть генеральной совокупности, не меньшую p. Константа k называется *доверительным пределом*. Так как нас интересует лишь доля от всей генеральной совокупности, то можно полагать, что истинное среднее значение равно нулю, а истинная дисперсия равна единице.

Приближенно метод нахождения толерантных интервалов состоит в следующем. Рассматривается случайная величина

$$A(\bar{x}, s, k) = \int_{\bar{x}-ks}^{\bar{x}+ks} f(x)\,dx. \qquad (4.10.3)$$

Требуется найти величину k, для которой выполняется соотношение

$$\mathbf{P}\{A(\bar{x}, s, k) > p\} = \gamma. \qquad (4.10.4)$$

Сложность решения задачи состоит в определении распределения случайной величины $A(\bar{x}, s, k)$.

Для определения константы k в [8, 57, 72] приведены таблицы для нескольких комбинаций доверительной вероятности γ, части генеральной совокупности p и различных значений объема выборки n. Кроме того, в [31, 57, 72] дана приближенная формула для вычисления константы k:

$$k = k_0\left(1 + \frac{k_\gamma}{\sqrt{2n}} + \frac{5k_\gamma^2 + 10}{12n}\right), \qquad (4.10.5)$$

где k_0 и k_γ — квантили, определяемые соотношениями

$$\Phi(k_0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{k_0} \exp(-t^2/2)\,dt = \frac{1+p}{2}; \qquad (4.10.6)$$

$$\Phi(k_\gamma) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{k_\gamma} \exp(-t^2/2)\,dt = \gamma. \qquad (4.10.7)$$

Следует отметить, что k_0 численно равно значению k при определении интервала для точно известных среднего значения и дисперсии нормальной генеральной совокупности. В [72] отмечается, что если $n \geqslant 160$, то приближенная формула (4.10.5) для $p \leqslant 0{,}999$ и $\gamma \leqslant 0{,}99$ даст значения k, которые совпадают с табличными в трех первых знаках. В прикладных задачах при определении константы k часто можно ограничиться погрешностью порядка трех процентов. При этом для $p \leqslant 0{,}95$ и $\gamma \leqslant 0{,}95$ объем выборки должен определяться величиной $n \geqslant 25$. Более точная, но и более сложная приближенная формула для вычисления константы k приведена в [8]; она достаточно точна уже при $n \geqslant 50$.

Применительно к односторонним толерантным интервалам нормальной генеральной совокупности $(-\infty, \bar{x} + ks)$ и $(\bar{x} - ks, \infty)$ в указанной выше литературе имеются методические указания и соответствующие таблицы для определения толерантного предела k.

Пример 4.10.1. Произведено $n = 200$ экспериментов измерения уровня принимаемых радиосигналов и вычислены выборочное среднее $\bar{x} = 10$ микровольт и выборочное среднее квадратическое отклонение $s = 2$ микровольт. Определим толерантные границы для доверительной вероятности $\gamma = 0{,}95$ и доли генеральной совокупности $p = 0{,}7$.

Согласно (4.10.6) и (4.10.7) имеем $\Phi(k_0) = 0{,}85$; $\Phi(k_\gamma) = 0{,}95$. По таблицам функции нормального распределения находим $k_0 = 1{,}04$ и $k_\gamma = 1{,}645$. Применяя

(4.10.5), получим

$$k = 1{,}04\left(1 + \frac{1{,}645}{\sqrt{2 \cdot 200}} + \frac{5 \cdot 1{,}645 + 10}{12 \cdot 200}\right) = 1{,}133.$$

Отсюда толерантные границы будут равны: $L_1 = 10 - 1{,}133 \cdot 2 \approx 7{,}7$ микровольт и $L_2 = 10 + 1{,}133 \cdot 2 \approx 12{,}3$ микровольт. Следовательно, с вероятностью $\gamma = 0{,}95$ можно считать, что в этих интервалах находится $100p = 70$ % всех возможных уровней принимаемых радиосигналов.

Контрольные вопросы и задачи

4.1. Чем отличаются точечные оценки параметров распределений случайных величин от интервальных оценок этих же параметров распределений случайных величин?

4.2. Назвать условия (соотношения), которые должны выполняться для получения эффективных точечных оценок параметров распределения случайных величин.

4.3. Что характеризуют условный и безусловный риски при точечной оценке параметра распределения случайной величины?

4.4. Пояснить смысл понятия степеней свободы для случайных величин.

4.5. В чем отличие толерантных (допустимых) интервалов от интервалов интервальной оценки параметров распределения?

4.6. Найти оценку максимального правдоподобия параметра λ экспоненциального распределения (1.4.20) при анализе независимой выборки объема n. Проверить выполнение условий для существования эффективной оценки параметра λ. Вычислить нижнюю границу дисперсии несмещенной оценки параметра λ.

4.7. Определить оценку параметра σ^2 релеевской случайной величины с плотностью вероятности (2.2.7) методами моментов и максимального правдоподобия при анализе независимой выборки объема n.

4.8. С доверительной вероятностью $\gamma = 0{,}95$ найти границы центрального доверительного интервала для параметра λ случайной величины с экспоненциальной плотностью вероятности (1.4.20) при анализе независимой выборки объема $n = 25$.

ГЛАВА 5

Проверка статистических гипотез

Проверка статистических гипотез является одной из основных задач математической статистики. При проверке статистических гипотез на основе выборочных значений случайных величин необходимо вынести суждение (с определенной степенью надежности) о верности или неверности тех или иных предположений (гипотез) относительно этих величин. Например, в двух сериях наблюдений получены две выборочные дисперсии, необходимо выяснить, относятся ли эти две серии наблюдений к одной генеральной совокупности или к разным совокупностям. Часто возникает следующая задача: на основе имеющейся выборки проверить гипотезу о принадлежности этой выборки нормальному распределению. Следует отметить, что статистические гипотезы принципиально отличаются от научных гипотез. Научная гипотеза — это разумное, обоснованное и развитое предположение, она является подлинно творческим актом исследователя и проверяется по результатам эксперимента. Например, научные гипотезы о существовании жизни на других планетах солнечной системы ничего общего не имеют со статистическими гипотезами.

Статистическая гипотеза — это утверждение относительно свойств тех или иных событий и случайных величин. Иначе говоря, статистические гипотезы связаны лишь с поведением наблюдаемых случайных величин. Как не все научные гипотезы требуется проверять статистическими методами, так и не все статистические гипотезы представляют научный интерес.

Гипотезы разделяют на параметрические и непараметрические. *Параметрические гипотезы* основаны на том, что рассматриваемые условные функции распределения известны с точностью до некоторых неизвестных параметров, в отношении которых проверяются те или иные предположения (гипотезы). Для *непараметрических гипотез* условные функции распределения могут быть известны лишь с точностью до достаточно широкого класса неизвестных параметров, в отношении которых гипотезы не проверяются, или вообще неизвестны.

В математической статистике при рассмотрении гипотез обычно выделяют *нулевую гипотезу* H_0, и проверка гипотез рассматривается с точ-

ки зрения принятия или отклонения этой гипотезы. Противоположные гипотезы называются *альтернативными*.

Рассмотрим основные понятия и критерии проверки статистических гипотез (в дальнейшем просто гипотез) применительно к двум альтернативным гипотезам H_0 и H_1. Пусть получена выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$, которая может быть порождена одним из взаимоисключающих событий A_0 или A_1. Событию A_0 соответствует гипотеза H_0, а событию A_1 — гипотеза H_1. Необходимо решить (принять решение): какая из гипотез соответствует этой выборке. Для предпочтения одной из гипотез требуется установить определенное правило (*критерий*) принятия того или иного решения относительно проверяемых гипотез. Следует оговорить, что любой эксперимент, строго говоря, не подтверждает и не отвергает гипотезу, а всего лишь не противоречит или противоречит ей. Однако в прикладных задачах математической статистики чаще употребляют термины «принять» и «отвергнуть» («отклонить») ту или иную гипотезу.

Возможны две ситуации при выборе *двуальтернативных гипотез*. В первой ситуации полагаются известными сравниваемые события, т.е. известны распределения для обеих гипотез. Такие гипотезы называются *простыми*. В противоположных ситуациях, когда для обеих гипотез или для одной из них вместо точно известных событий задаются их возможные области существования (задаются возможные распределения вероятностей гипотез), проверяемые гипотезы называют *сложными*.

Оптимальное (наилучшее) правило принятия решения формируется заранее до получения выборки в соответствии с выбранным критерием оптимальности.

5.1. Проверка простых гипотез

Применительно к простым нулевой и альтернативной гипотезам, соответствующим событиям A_0 и A_1, распределения вероятностей анализируемых выборок известны. Для синтеза оптимального правила выбора решения необходимо иметь определенные сведения относительно рассматриваемых событий. Во-первых, пусть известны *априорные вероятности событий (гипотез)* $p_{\text{pr}}(A_0) \equiv p_{\text{pr}}(H_0) = p_0$ и $p_{\text{pr}}(A_1) \equiv p_{\text{pr}}(H_1) = p_1 = 1 - p_0$ и, во-вторых, пусть заданы потери (стоимости, затраты), возникающие при принятии того или иного решения. Обозначим через Γ пространство всех возможных выборочных значений $\mathbf{X}(X_1, X_2, \ldots, X_n)$. При этом будем полагать, что это пространство непрерывно. Условные плотности вероятности выборки для обеих гипотез запишем в виде $f(\mathbf{x} \mid H_0)$ и $f(\mathbf{x} \mid H_1)$. Для принятия решения в пользу одной из гипотез пространство наблюдений Γ согласно заранее выбранному критерию оптимальности необходимо разделить на два подпространства Γ_0 и Γ_1, соответствующие гипотезам H_0 и H_1. Подпространство Γ_0 называют *областью принятия нулевой гипотезы*, а подпространство Γ_1 называют *критической областью* или *областью отклонения нулевой гипотезы*. При принятии решения в пользу одной или другой гипотез возможны

ошибки решения. Применительно к принятию или отклонению нулевой гипотезы возможны ошибки решения, обусловленные следующими ситуациями: выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ соответствует событию A_0 (гипотезе H_0), а принимается решение в пользу гипотезы H_1; выборка $\mathbf{x}(x_1, x_2, \ldots, x_n)$ соответствует событию A_1 (гипотезе H_1), а принимается решение в пользу гипотезы H_0. По своему смыслу критическая область должна включать все маловероятные выборочные значения при справедливости нулевой гипотезы.

Условная вероятность ошибки решения, связанная с отклонением нулевой гипотезы, когда она в действительности верна, называется *ошибкой первого рода*. Условная вероятность ошибки решения, связанная с принятием нулевой гипотезы, когда она неверна (верна альтернативная гипотеза H_1), называется *ошибкой второго рода*. В математической статистике ошибку первого рода называют *уровнем значимости критерия* для проверки гипотезы H_0 и обозначают α. Вместо ошибки второго рода, которую обычно обозначают β, рассматривают величину

$$w = 1 - \beta, \tag{5.1.1}$$

называемую *мощностью критерия* для проверки гипотезы H_0 против альтернативной гипотезы H_1. Мощность критерия численно равна вероятности принять гипотезу H_1, когда она в действительности верна. Уровень значимости α и ошибка второго рода β определяются выражениями:

$$\alpha = \int_{\Gamma_1} f(\mathbf{x} \mid H_0)\, d\mathbf{x}; \tag{5.1.2}$$

$$\beta = \int_{\Gamma_0} f(\mathbf{x} \mid H_1)\, d\mathbf{x}. \tag{5.1.3}$$

В задачах статистической радиолокации при приеме сигналов в шумах ошибку первого рода (уровень значимости) называют ложной тревогой, а мощность критерия — вероятностью обнаружения сигнала.

Суммарные условная p'_{e} и безусловная p_{e} вероятности ошибок равны соответственно:

$$p'_{\text{e}} = \alpha + \beta; \tag{5.1.4}$$

$$p_{\text{e}} = p_0\alpha + (1 - p_0)\beta. \tag{5.1.5}$$

Как отмечалось выше в п. 4.1, после проведения эксперимента и получения выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ наиболее полная информация о возможных событиях заключена в апостериорной вероятности этих событий. Для рассматриваемых дискретных событий согласно (4.1.3) *апостериорные вероятности гипотез* H_0 и H_1 определяются соотношениями

$$p_{\text{ps}}(H_0; \mathbf{x}) = \frac{p_0 L(H_0; \mathbf{x})}{f(\mathbf{x})}; \tag{5.1.6}$$

$$p_{\text{ps}}(H_1;\mathbf{x}) = \frac{p_1 L(H_1;\mathbf{x})}{f(\mathbf{x})}, \tag{5.1.7}$$

где

$$f(\mathbf{x}) = p_0 f(\mathbf{x} \mid H_0) + p_1 f(\mathbf{x} \mid H_1) \tag{5.1.8}$$

— безусловная плотность вероятности выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$; $L(H_0;\mathbf{x})$ и $L(H_1;\mathbf{x})$ — функции правдоподобия гипотез H_0 и H_1.

Подобно оптимальной байесовской оценке параметров распределений каждое решение при принятии гипотезы характеризуется некоторыми положительными потерями (затратами), определяемыми функциями потерь, аналогичными рассмотренным в п. 4.3. Для двухальтернативных гипотез *функцию потерь* (стоимость) можно записать как c_{ij}, $i, j = 0, 1$. Здесь первый индекс означает выбранную гипотезу, а второй — гипотезу, которая была правильной. Применительно к рассматриваемой нулевой гипотезе c_{10} — стоимость ошибки первого рода α, c_{01} — стоимость ошибки второго рода β. Естественно полагать, что потери при принятии правильных решений c_{00} и c_{11} должны быть меньше потерь за счет неправильных решений c_{10} и c_{01} соответственно, т.е. справедливы неравенства $c_{10} - c_{00} > 0$ и $c_{01} - c_{11} > 0$.

Средний риск (средние потери) при принятии решений аналогично (4.3.6) записываются в виде

$$R = p_0 \left[c_{00} \int_{\Gamma_0} f(\mathbf{x} \mid H_0)\, d\mathbf{x} + c_{10} \int_{\Gamma_1} f(\mathbf{x} \mid H_0)\, d\mathbf{x} \right] + \\ + p_1 \left[c_{11} \int_{\Gamma_1} f(\mathbf{x} \mid H_1)\, d\mathbf{x} + c_{01} \int_{\Gamma_0} f(\mathbf{x} \mid H_1)\, d\mathbf{x} \right]. \tag{5.1.9}$$

Здесь, как и выше, для компактности записи вместо обозначения n-кратного интегрирования по выборочному пространству используется один знак интеграла.

Согласно байесовскому критерию оптимальное правило выбора критической области Γ_1 определяется исходя из условия минимума среднего риска (5.1.9), который с учетом определений ошибок первого и второго рода можно записать в виде

$$R = p_0 c_{00} + p_1 c_{01} - \int_{\Gamma_1} [p_1(c_{01} - c_{11}) f(\mathbf{x} \mid H_1) - p_0(c_{10} - c_{00}) f(\mathbf{x} \mid H_0)]\, d\mathbf{x}.$$

Так как член $(p_0 c_{00} + p_1 c_{01})$ не зависит от выборки, то минимум среднего риска будет при условии, если

$$p_1(c_{01} - c_{11}) f(\mathbf{x} \mid H_1) - p_0(c_{10} - c_{00}) f(\mathbf{x} \mid H_0) \geqslant 0.$$

Отсюда получаем правило выделения критической области Γ_1 (точек пространства возможных выборочных значений), где отклоняется нулевая гипотеза,

$$\frac{f(\mathbf{x} \mid H_1)}{f(\mathbf{x} \mid H_0)} \geqslant \frac{p_0(c_{10} - c_{00})}{p_1(c_{01} - c_{11})}. \tag{5.1.10}$$

При фиксированной выборке $\mathbf{x}(x_1, x_2, \ldots, x_n)$ это отношение равно

$$l(\mathbf{x}) \equiv l(x_1, x_2, \ldots, x_n) = L(H_1;\mathbf{x})/L(H_0;\mathbf{x}). \tag{5.1.11}$$

Положительная функция $l(\mathrm{x})$ называется *отношением правдоподобия*. Так как функции правдоподобия являются достаточными статистиками, то и их отношение также является достаточной статистикой. Поэтому *байесовский критерий* (*оптимальный*) проверки нулевой гипотезы H_0 против альтернативы H_1 базируется на достаточной статистике отношения правдоподобия и заключается в сравнении этой статистики с критическим значением («порогом»)

$$h = \frac{p_0(c_{10} - c_{00})}{p_1(c_{01} - c_{11})}. \tag{5.1.12}$$

Из сказанного следует оптимальное (байесовское) решающее правило выбора гипотез

$$\begin{aligned} &l(\mathbf{x}) < h \text{ — принимается гипотеза } H_0; \\ &l(\mathbf{x}) \geqslant h \text{ — принимается гипотеза } H_1. \end{aligned} \tag{5.1.13}$$

При этом *байесовский риск* согласно (5.1.9)

$$R = p_0 c_{00} + p_1 c_{11} + p_0(c_{10} - c_{00})\alpha + p_1(c_{01} - c_{11})\beta. \tag{5.1.14}$$

К байесовскому правилу также можно прийти на основе анализа условного (апостериорного) риска R_{ps}, который аналогично (4.3.8) является суммой двух апостериорных рисков $R_{\text{ps}}(H_0)$ и $R_{\text{ps}}(H_1)$, обусловленных потерями при принятии гипотез H_0 и H_1 соответственно. Эти апостериорные риски равны

$$R_{\text{ps}}(H_0;\mathbf{x}) = c_{00} p_{\text{ps}}(H_0;\mathbf{x}) + c_{01} p_{\text{ps}}(H_1;\mathbf{x});$$
$$R_{\text{ps}}(H_1;\mathbf{x}) = c_{10} p_{\text{ps}}(H_0;\mathbf{x}) + c_{11} p_{\text{ps}}(H_1;\mathbf{x}).$$

Согласно байесовской теории правило выбора решения (в том числе и для принятия той или другой гипотезы) должно обеспечивать минимум апостериорного риска, т.е. из двух альтернативных гипотез следует выбрать ту, для которой апостериорный риск меньше. Решающее правило должно отнести выборку $\mathbf{x}(x_1, x_2, \ldots, x_n)$ к нулевой гипотезе H_0, если

$$R_{\text{ps}}(H_1;\mathbf{x}) > R_{\text{ps}}(H_0;\mathbf{x}). \tag{5.1.15}$$

При выполнении противоположного неравенства должно приниматься решение в пользу гипотезы H_1. Подставляя в (5.1.15) значения апостериорных рисков и учитывая связь апостериорных вероятностей с функ-

циями правдоподобия (5.1.6) и (5.1.7), получаем правило принятия решения (5.1.13).

Рассмотренное выше правило выбора решения не позволяет находить в явном виде границы областей Γ_0 и Γ_1. В то же время приведенные правила выбора решения дают возможность определить принадлежность выборки к той или иной области на основе сравнения отношения правдоподобия с порогом и вычислить соответствующие ошибки первого и второго рода. Для вычисления ошибок первого и второго рода рассмотрим отношение правдоподобия $l(x_1, x_2, \ldots, x_n)$, которое из-за случайности выборки является условной случайной величиной, зависящей от того, какой гипотезе в действительности принадлежит выборка. Отношение правдоподобия является измеряемой статистикой. При справедливости каждой из гипотез сформированная статистика $l(\mathbf{x})$ характеризуется известными законами распределения. Будем полагать, что этими законами являются условные плотности вероятности $f(l \mid H_0)$ и $f(l \mid H_1)$.

Применительно к нулевой гипотезе H_0 ошибки первого рода (уровень значимости) α и второго рода β определяются из соотношений

$$\alpha = p(H_1 \mid H_0) = \int_h^\infty f(l \mid H_0)\, dl; \tag{5.1.16}$$

$$\beta = p(H_0 \mid H_1) = \int_{-\infty}^h f(l \mid H_1)\, dl. \tag{5.1.17}$$

Таким образом, для вынесения решения в пользу той или иной гипотезы достаточно воспользоваться значением одной случайной величины — статистики отношения правдоподобия $l(x_1, x_2, \ldots, x_n)$.

Если в качестве функции потерь используется простая функция потерь, являющаяся аналогом (4.3.1), то потери при правильных решениях полагаются нулевыми ($c_{00} = c_{11} = 0$), а функции потерь за счет неправильных решений $c_{10} = c_{01} = c = 1$. В этом случае пороговая постоянная равна отношению априорных вероятностей гипотез H_0 и H_1:

$$h = p_0/p_1, \tag{5.1.18}$$

а оптимальное правило выбора решения (5.1.13) в пользу нулевой гипотезы H_0 примет вид

$$l(\mathbf{x}) < p_0/p_1. \tag{5.1.19}$$

Это правило выбора решения соответствует максимуму апостериорной вероятности нулевой гипотезы, а следовательно, и минимуму вероятности полной ошибки p_e (5.1.5). Оно называется *критерием максимальной апостериорной вероятности*, так как из (5.1.6) и (5.1.7) следует

$$\frac{p_{\text{ps}}(H_1; \mathbf{x})}{p_{\text{ps}}(H_0; \mathbf{x})} = \frac{p_1 L(H_1; \mathbf{x})}{p_0 L(H_0; \mathbf{x})}.$$

Средний риск, соответствующий критерию максимальной апостериорной вероятности, находится из (5.1.14):

$$R = p_0\alpha + p_1\beta. \tag{5.1.20}$$

В задачах теории передачи сообщений критерий максимальной апостериорной вероятности, минимизирующий суммарную ошибку решения (1.5.20), называют *критерием идеального наблюдателя*.

Если при формировании правила решения отсутствуют данные не только о функциях потерь, но и об априорных вероятностях гипотез, применяют *критерий максимального правдоподобия*. Согласно этому критерию принимается та гипотеза, для которой функция правдоподобия максимальна, т.е. в качестве порога выбирается величина $h = 1$. При этом нулевая гипотеза принимается, если

$$l(x_1, x_2, \ldots, x_n) < 1. \tag{5.1.21}$$

Формально критерий максимального правдоподобия вытекает из критерия максимальной апостериорной вероятности для $p_0 = p_1 = 1/2$.

При отсутствии априорной информации о вероятностях гипотез и о функциях потерь в математической статистике часто используется критерий Неймана–Пирсона. Согласно *критерию Неймана–Пирсона* подобласти Γ_0 и Γ_1 выбираются таким образом, чтобы ошибка первого рода (уровень значимости) не превышала заранее заданного значения α:

$$\int_{\Gamma_1} f(\mathbf{x} \mid H_0)\, d\mathbf{x} \leqslant \alpha = \text{const}. \tag{5.1.22}$$

При этом ошибка второго рода β должна быть минимальной, что эквивалентно выбору критической области с максимальной мощностью критерия

$$w = 1 - \beta = \int_{\Gamma_1} f(\mathbf{x} \mid H_1)\, d\mathbf{x} = \max. \tag{5.1.23}$$

Пороговое значение $h = h_\alpha$ для принятия решения по критерию Неймана–Пирсона определяется из уравнения

$$\alpha = p(H_1 \mid H_0) = \int_{h_\alpha}^\infty f(l \mid H_0)\, dl. \tag{5.1.24}$$

При этом ошибка второго рода при принятии решения в пользу нулевой гипотезы

$$\beta = p(H_0 \mid H_1) = \int_{-\infty}^{h_\alpha} f(l \mid H_1)\, dl. \tag{5.1.25}$$

В критерии Неймана–Пирсона используется достаточная статистика отношения правдоподобия. Поэтому можно утверждать, что из всех возможных критериев он обеспечивает максимальную мощность решения $w = 1 - \beta$ в пользу нулевой гипотезы при фиксированном значении ошибки первого рода α.

Все рассмотренные выше правила выбора решения (критерии) оптимальности основаны на вычислении отношения правдоподобия, а нахождение соответствующих вероятностей ошибок решения — на знании условных плотностей вероятности отношения правдоподобия. Поскольку непосредственное нахождение этих условных вероятностей представляет серьезные трудности, в качестве статистики можно взять любую функцию, однозначно связанную с отношением правдоподобия $l(\mathbf{x})$, для которой условные вероятности определяются, как правило, значительно проще. В качестве преобразования отношения правдоподобия наиболее конструктивным представляется *логарифм отношения правдоподобия* $l_0(\mathbf{x}) = \ln l(\mathbf{x}) = \ln L(\mathbf{x}; H_1) - \ln L(\mathbf{x}; H_0)$, который сравнивается с порогом ($h_0 = \ln h$ или $h_{0\alpha} = \ln h_\alpha$). Применительно к критерию Неймана–Пирсона нулевая гипотеза принимается, если

$$l_0(\mathbf{x}) = \ln l(\mathbf{x}) < h_{0\alpha} = \ln h_\alpha. \tag{5.1.26}$$

Пример 5.1.1. Проверим нулевую гипотезу H_0 о среднем значении гауссовской случайной величины $m = m_0$ против альтернативной гипотезы H_1 о значении $m = m_1$ для рассмотренных выше критериев проверки гипотез. При этом будем полагать, что анализируемая выборка объема n независима с известной дисперсией σ^2.

Функции правдоподобия параметра m для гипотез H_0 и H_1 равны

$$L(m_0; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_0)^2\right];$$

$$L(m_1; \mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m_i)^2\right].$$

Логарифм отношения правдоподобия

$$l_0(x_1, x_2, \ldots, x_n) = \frac{m_1 - m_0}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n(m_1^2 - m_0^2)}{2\sigma^2}$$

необходимо сравнить с порогом $h_0 = \ln h$. Нулевая гипотеза в пользу $m = m_0$ принимается, если

$$\frac{m_1 - m_0}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n(m_1^2 - m_0^2)}{2\sigma^2} < \ln h. \tag{5.1.27}$$

Это неравенство можно переписать в виде двух неравенств, зависящих от соотношения между m_0 и m_1. Если $m_0 < m_1$, то

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i < \frac{m_1 + m_0}{2} + \frac{\sigma^2 \ln h}{n(m_1 - m_0)} = h_1, \quad m_1 > m_0. \tag{5.1.28}$$

Если же $m_0 > m_1$, то $(m_1 - m_0) < 0$ и при делении левой и правой части (5.1.27) на $(m_1 - m_0)/\sigma^2$ знак неравенства меняется на обратный знак. В результате для $m_0 > m_1$ получаем правило выбора нулевой гипотезы

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i > \frac{m_1 + m_0}{2} - \frac{\sigma^2 \ln h}{n(m_0 - m_1)} = h_2, \quad m_0 > m_1. \tag{5.1.29}$$

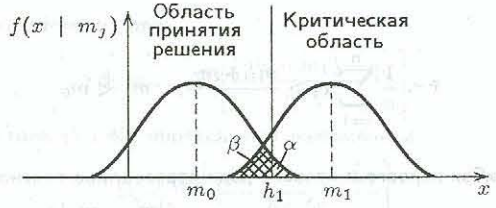


Рис. 5.1.1. Плотности вероятностей выборочного среднего

Из последних соотношений видно, что для принятия решения относительно среднего значения необходимо сравнить выборочное среднее с пороговым значением h_j, $j = 1, 2$, выбираемым исходя из рассмотренных выше критериев оптимальности. Как показано в п. 4.5.1, выборочное среднее является достаточной статистикой, распределено по нормальному закону со средним значением, совпадающим с истинным значение m_j, и дисперсией σ^2/n:

$$f(\bar{x} \mid m_j) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left[-\frac{(\bar{x} - m_j)^2}{2\sigma^2/n}\right], \quad j = 0, 1.$$

На рис. 5.1.1 приведены плотности вероятностей выборочного среднего для обеих гипотез при $m_1 > m_0$. Там же условно показано пороговое значение h_1 и заштрихованные площади, соответствующие ошибкам первого и второго рода при принятии решения в пользу нулевой гипотезы.

Вычислим ошибки первого и второго рода для случая $m_1 > m_0$. Ошибка первого рода (уровень значимости)

$$\alpha_1 = \int_{h_1}^{\infty} f(\bar{x} \mid H_0)\, d\bar{x} = 1 - \int_{-\infty}^{h_1} f(\bar{x} \mid H_0)\, d\bar{x} = 1 - \Phi\left(\frac{h_1 - m_0}{\sigma/\sqrt{n}}\right). \tag{5.1.30}$$

Ошибка второго рода определяется как

$$\beta_1 = \int_{-\infty}^{h_1} f(\bar{x} \mid H_1)\, d\bar{x} = \Phi\left(\frac{h_1 - m_1}{\sigma/\sqrt{n}}\right). \tag{5.1.31}$$

При этом суммарная ошибка

$$p_e' = \alpha_1 + \beta_1 = 1 - \Phi\left(\frac{h_1 - m_0}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{h_1 - m_1}{\sigma/\sqrt{n}}\right). \tag{5.1.32}$$

Если $m_1 < m_0$, то согласно правилу решения (5.1.29) ошибки первого и второго рода равны

$$\alpha_2 = \int_{-\infty}^{h_2} f(\bar{x} \mid H_0)\, d\bar{x} = \Phi\left(\frac{h_2 - m_0}{\sigma/\sqrt{n}}\right); \tag{5.1.33}$$

$$\beta_2 = \int_{h_2}^{\infty} f(\bar{x} \mid H_1)\, d\bar{x} = 1 - \Phi\left(\frac{h_2 - m_1}{\sigma/\sqrt{n}}\right). \tag{5.1.34}$$

Пусть в качестве правила принятия решения выбран критерий максимума отношения правдоподобия, для которого $\ln h = 0$. Тогда из (5.1.28) получаем, что для принятия решения о той или иной гипотезе выборочное среднее надо сравнить с порогом $h' = (m_1 + m_0)/2$, численно равным среднему значению из сравниваемых величин. Нулевая гипотеза о значении $m = m_0$ принимается, если

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i < \frac{m_1 + m_0}{2}, \quad m_1 > m_0, \tag{5.1.35}$$

или

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i > \frac{m_1 + m_0}{2}, \quad m_1 < m_0. \tag{5.1.36}$$

При этом для ошибок первого и второго рода справедливо соотношение

$$\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = 1 - \Phi\left(\frac{|m_1 - m_0|}{2\sigma/\sqrt{n}}\right). \tag{5.1.37}$$

Применительно к правилу Неймана-Пирсона ($m_1 > m_0$) для порогового значения введем обозначение $h'_{0\alpha} = m_0 + \tilde{h}_\alpha$. Тогда для заданной ошибки первого рода α согласно (5.1.24) величина $\tilde{h}_\alpha$ находится из выражения

$$\Phi\left(z_\alpha = \frac{\tilde{h}_\alpha}{\sigma/\sqrt{n}}\right) = 1 - \alpha. \tag{5.1.38}$$

Величина z_α является критическим значением нормального распределения и определяется как

$$\mathbf{P}\{Z > z_\alpha\} = \int_{z_\alpha}^{\infty} f(z)\,dz = 1 - \Phi(z_\alpha) = \alpha.$$

Решение в пользу нулевой гипотезы H_0 принимается, если выборочное среднее будет меньше порогового значения $h'_{0\alpha} = m_0 + \tilde{h}_\alpha$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i < m_0 + \tilde{h}_\alpha. \tag{5.1.39}$$

На основе полученного порогового значения $h'_{0\alpha} = m_0 + \tilde{h}_\alpha$ согласно (5.1.25) вычисляется ошибка второго рода

$$\beta_1 = 1 - \Phi\left[\frac{(m_1 - m_0) - \tilde{h}_\alpha}{\sigma/\sqrt{n}}\right]. \tag{5.1.40}$$

При этом мощность критерия (5.1.39)

$$w_1 = 1 - \beta_1 = 1 - \Phi\left[\frac{\tilde{h}_\alpha - (m_1 - m_0)}{\sigma/\sqrt{n}}\right]. \tag{5.1.41}$$

При $m_1 = m_0$ мощность критерия численно равна уровню значимости ($w_1 = \alpha$); если $(m_1 - m_0) \to \infty$, то $w_1 = 1$, при $m_1 = m_0 + \tilde{h} - \alpha$ мощность критерия $w_1 = 0{,}5$.

Конкретизируем полученные результаты для уровня значимости $\alpha = 0{,}025$. Из таблицы функции нормального распределения $\Phi(z_\alpha) = 0{,}975$ находим $z_\alpha = \tilde{h}_\alpha/(\sigma/\sqrt{n}) = 1{,}96$. Из последнего получаем значение порога $\tilde{h}_\alpha = 1{,}96\sigma/\sqrt{n}$.

Если $m_1 < m_0$, распределение выборочного среднего для нулевой гипотезы будет расположено правее распределения для альтернативной гипотезы. Применительно к критерию Неймана–Пирсона пороговое значение $h''_{\alpha 2} = m_0 - \tilde{h}_\alpha$, где величина $\tilde{h}_\alpha$ для заданного уровня значимости α находится также из решения уравнения (5.1.38). Решение в пользу нулевой гипотезы принимается, если выборочное среднее будет больше величины $h''_{\alpha 2} = m_0 - \tilde{h}_\alpha$, т.е. если

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i > m_0 - \tilde{h}_\alpha. \tag{5.1.42}$$

При этом ошибка второго рода

$$\beta_2 = 1 - \Phi\left[\frac{(m_0 - m_1) - \tilde{h}_\alpha}{\sigma/\sqrt{n}}\right], \tag{5.1.43}$$

а мощность критерия (5.1.42) определяется выражением

$$w_2 = \Phi\left[\frac{(m_0 - m_1) - \tilde{h}_\alpha}{\sigma/\sqrt{n}}\right]. \tag{5.1.44}$$

Аналогично случаю $m_1 > m_0$ при $m_1 = m_0$, $m_1 \to \infty$ и $m_1 = m_0 - \tilde{h}_\alpha$ мощность критерия (5.1.42) принимает значения, соответствующие $w_2 = \alpha$, $w_2 = 1$ и $w_2 = 0{,}5$.

Пример 5.1.2. Пусть анализируемая независимая выборка объема $n = 100$ представляет собой аддитивную смесь $x_i = \lambda a + y_i$, $i = 1, 2, \ldots, n$, детерминированной величины $a = 0{,}3$ и гауссовской случайной величины y_i с нулевым средним значением и дисперсией $\sigma^2 = 1$. Параметр λ может принимать два значения: $\lambda = 0$ (нулевая гипотеза H_0) и $\lambda = 1$ (альтернативная гипотеза H_1). По заданному уровню значимости $\alpha = 0{,}01$ определим правило принятия нулевой гипотезы и вычислим мощность критерия (вероятность обнаружения сигнала a). Функции правдоподобия для гипотез H_0 и H_1 определяются соотношениями

$$L(H_0;\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right];$$

$$L(H_1;\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (x_i - a)^2\right].$$

При этом логарифм отношения правдоподобия принимает вид $l_0(\mathbf{x}) = (na/\sigma^2)(\bar{x} - a/2)$. Так как логарифм отношения правдоподобия монотонно зависит от выборочного среднего $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, то для принятия той или иной гипотезы его необходимо сравнить с критическим значением (порогом) h_α, определяемым исходя из заданного уровня значимости α. Распределения статистик для рассматриваемых гипотез имеют вид

$$f(\bar{x} \mid H_0) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{\bar{x}^2}{2\sigma^2/n}\right];$$

$$f(\bar{x} \mid H_1 = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(\bar{x} - a)^2}{2\sigma^2/n}\right].$$

Критическое значение статистики h_α является критической точкой нормального распределения и определяется как

$$\alpha = \int_{h_\alpha}^{\infty} f(\bar{x} \mid H_0)\,d\bar{x} = 1 - \Phi\left(\frac{h_\alpha}{\sigma/\sqrt{n}}\right).$$

Мощность критерия (вероятность обнаружения сигнала a)

$$w = \int_{h_\alpha}^{\infty} f(\bar{x} \mid H_1)\,d\bar{x} = 1 - \int_{h_\alpha}^{\infty} f(\bar{x} \mid H_1)\,d\bar{x} = 1 - \Phi\left(\frac{h_\alpha - a}{\sigma/\sqrt{n}}\right).$$

Применительно к заданным числовым значениям из табл. 1в Приложения 2 имеем $h_\alpha\sqrt{n}/\sigma = 2{,}326$. Подставляя последнее значение в предыдущую формулу, получаем $w = 1 - \Phi(2{,}326 - 3) = \Phi(0{,}674 \approx 0{,}75$.

Пример 5.1.3. Проверим нулевую гипотезу H_0 о значении дисперсии гауссовской случайной величины $\sigma^2 = \sigma_0^2$ против альтернативной гипотезы H_1 о значении

$\sigma^2 = \sigma_1^2$. При этом будем полагать, что анализируемая выборка объема n независима с нулевым средним значением $m = 0$, а $\sigma_0^2 < \sigma_1^2$.

Функции правдоподобия для гипотез H_0 и H_1 определяются соотношениями

$$L(H_0; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n} x_i^2\right];$$

$$L(H_1; \mathbf{x}) = (2\pi\sigma_1^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n} x_i^2\right].$$

Логарифм отношения правдоподобия

$$l_0(\mathbf{x}) = \ln l(\mathbf{x}) = \ln\frac{L(H_1; \mathbf{x})}{L(H_0; \mathbf{x})} = n\ln\frac{\sigma_0}{\sigma_1} + \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_1^2\sigma_0^2}\sum_{i=1}^{n} x_i^2. \qquad (5.1.45)$$

Решение в пользу нулевой гипотезы $\sigma^2 = \sigma_0^2$ будет принято, если для измеряемой статистики выполняется неравенство

$$\sum_{i=1}^{n} x_i^2 < \frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2}\ln\left[\left(\frac{\sigma_1}{\sigma_0}\right)^n h\right] \equiv h', \qquad (5.1.46)$$

или

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 < \frac{1}{n}\frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2 - \sigma_0^2}\ln\left[\left(\frac{\sigma_1}{\sigma_0}\right)^n h\right] \equiv h'/n. \qquad (5.1.47)$$

Из последнего выражения видно, что в качестве измеряемой статистики берется выборочная дисперсия, которая и сравнивается с порогом h'/n.

Случайные величины $z = \sum(x_j/\sigma_j)^2$ для $j = 0, 1$ распределены по закону хи-квадрат (2.2.3) с $\nu = n$ степенями свободы. При этом применительно к проверяемой нулевой гипотезе для вычисления ошибки первого рода пороговое значение равно $\chi^2_{\alpha;n} = h'/\sigma_0^2$. Условные плотности вероятности случайной величины z для конкурирующих гипотез запишем в виде $f(z \mid H_0)$ и $f(z \mid H_1)$. Тогда ошибки первого и второго рода при принятии нулевой гипотезы определяются соотношениями

$$\alpha = \mathbf{P}\left\{\sum_{i=1}^{n} x_i^2 > h'\right\} \equiv \mathbf{P}\left\{z > \frac{h'}{\sigma_0^2}\right\} = \int_{\chi^2_{\alpha;n}}^{\infty} f(z \mid H_0)\, dz;$$

$$\beta = \mathbf{P}\left\{\sum_{i=1}^{n} x_i^2 < h'\right\} \equiv \mathbf{P}\left\{z < \frac{h'}{\sigma_1^2}\right\} = 1 - \int_{\chi^2_{\beta;n}}^{\infty} f(z \mid H_1)\, dz,$$

где $\chi^2_{\beta;n} = (\sigma_0^2/\sigma_1^2)\chi^2_{\alpha;n} = h'/\sigma_1^2$.

Конкретизируем значения порога h' и ошибок первого и второго рода для $\sigma_1^2/\sigma_0^2 = 3$, $n = 25$ при принятии решения по критерию максимума отношения правдоподобия ($h = 1$). Этим данным соответствуют $h' = (37{,}5\ln 3)\sigma_0^2$; $\chi^2_{\alpha;n} \approx 41{,}2$; $\chi^2_{\beta;n} \approx 13{,}7$. Таким образом, гипотеза H_0 в пользу $\sigma^2 = \sigma_0^2$ принимается, если $\sum x_i^2 < (37{,}5\ln 3)\sigma_0^2$. На основе таблиц хи-квадрат распределения для числа степеней свободы $n = 25$ определяются величины $\chi^2_{\alpha;n}$ и $\chi^2_{\beta;n}$, а затем и значения ошибок первого и второго рода: $\alpha \approx 0{,}023$ и $\beta \approx 0{,}035$. Если $\sigma_1^2 < \sigma_0^2$, аналогично предыдущему примеру для принятия нулевой гипотезы в пользу $\sigma^2 = \sigma_0^2$ в выражениях (5.1.46) и (5.1.47) знаки неравенства должны быть изменены на противоположные. Соответственно изменятся и формулы для ошибок первого и второго рода.

Приведенные выше соотношения для критериев проверки гипотез получены для непрерывных анализируемых случайных величин. Вместе с тем они обобщаются и на дискретные случайные величины. При этом для всех возможных дискретных выборочных значений пространство Γ и подпространства Γ_0 и Γ_1, соответствующие области принятия нулевой гипотезы H_0 и критической области (области отклонения нулевой гипотезы H_0), являются дискретными множествами.

Обобщение приведенных выше соотношений на случай анализа дискретных случайных величин сводится к замене условных плотностей вероятностей $f(\mathbf{x} \mid H_0)$ и $f(\mathbf{x} \mid H_1)$ на условные дискретные распределения вероятностей $p(\mathbf{x} \mid H_0)$ и $p(\mathbf{x} \mid H_1)$, а также замене соответствующих интегралов суммами (суммирование по всем возможным дискретным значениям случайной величины). Для дискретных случайных величин функции правдоподобия $L(H_0; \mathbf{x})$ и $L(H_1; \mathbf{x})$, отношение правдоподобия $l(x) = L(H_1; \mathbf{x})/L(H_0; \mathbf{x})$ и формируемые статистики будут дискретными функциями (величинами).

С целью использования известных и табулированных дискретных распределений (биномиального, Пуассона и др.) для вычисления ошибок первого рода α и второго рода β дискретную статистику приводят к неотрицательным целочисленным значениям. При этом критическое значение статистики, как правило, не является одним из возможных целочисленных значений статистики.

Принципиально, используя прием рандомизации, связанный с осуществлением специального случайного эксперимента, задачу с анализом дискретных данных можно свести к анализу непрерывных данных [13, 27, 46]. Однако процедура рандомизации достаточно искусственна, представляет в основном теоретический интерес и практически не используется в прикладной статистике.

В прикладной статистике при дискретных распределениях статистики обычно поступают следующим образом. Применительно к байесовскому критерию и критерию отношения правдоподобия получающееся критическое значение (нецелочисленное) статистики уменьшают до ближайшего допустимого целочисленного значения, тем самым увеличивая ошибку первого рода и уменьшая ошибку второго рода. При использовании критерия Неймана–Пирсона произвольно заданному уровню значимости (например, $\alpha = 0{,}05$) также, как правило, нельзя поставить в соответствие допустимое значение статистики. Поэтому и в этом случае в качестве критического значения статистики берут значение, соответствующее ближайшему наибольшему уровню значимости $\alpha' > \alpha$, увеличивая, как и выше, ошибку первого рода и уменьшая ошибку второго рода.

Пример 5.1.4. Дозиметрическим прибором (счетчиком частиц) в течение времени $T = 10$ секунд зарегистрировано $k^* = 20$ частиц. Необходимо проверить нулевую гипотезу H_0 о том, что результат наблюдения обусловлен радиационным фоном, а не суммой фона и радиационного излучения, если известно, что среднее число частиц за одну секунду равно $\nu_0 = 0{,}2$ и $\nu_1 = 0{,}8$ соответственно. Проверку гипотезы осуществим по критерию Неймана–Пирсона с уровнем значимости $\alpha = 0{,}05$.

Распределение вероятностей числа частиц k (положительных исходов) для обеих гипотез подчинено закону Пуассона (1.3.23) со средними значениями $\lambda_0 = \nu_0 T = 2{,}0$ и $\lambda_1 = \nu_1 T = 8{,}0$.

Функции правдоподобия для обеих гипотез записываются как

$$L(k; H_0) = \frac{\lambda_0^k}{k!}\exp(-\lambda_0); \quad L(k; H_1) = \frac{\lambda_1^k}{k!}\exp(-\lambda_1).$$

Так как логарифм отношения правдоподобия $l(k)$ определяется выражением

$$\ln l(k) = \ln[L(k; H_1)/L(k; H_0)] = k\ln(\lambda_1/\lambda_0) - (\lambda_1 - \lambda_0),$$

то в качестве статистики целесообразно использовать число положительных исходов (зарегистрированное число частиц) k, распределение которых подчиняется закону Пуассона с параметром λ_0 для нулевой гипотезы и с параметром λ_1 для альтернативной гипотезы. Критическое значение статистики k' должно удовлетворять уравнению

$$\mathbf{P}\{k > k'; H_0\} = \sum_{k=k'}^{\infty}\frac{\lambda_0^k}{k!}\exp(-\lambda_0) = \sum_{k=k'}^{\infty}\frac{2^k}{k!}\exp(-2) = \alpha = 0{,}05.$$

Согласно таблицам биномиального распределения для $\lambda = 2$ и заданного уровня значимости отсутствует целочисленное значение статистики. Наиболее близким к $\alpha = 0{,}05$ является $\alpha' = 0{,}053$, для которого целочисленное значение статистики равно $k' = 5$. Так как выборочное значение статистики $k^* = 20 > k' = 5$ отвергается нулевая гипотеза H_0 о том, что результат наблюдения обусловлен радиационным фоном. При этом вероятность отклонить альтернативную гипотезу H_1, когда она верна (ошибка второго рода),

$$\mathbf{P}\{k < k'; H_1\} = \sum_{k=0}^{k'}\frac{\lambda_1^k}{k!}\exp(-\lambda_1) = \sum_{k=0}^{5}\frac{8^k}{k!}\exp(-8) = \beta = 0{,}10.$$

В конкретных задачах для фиксированного значения ошибки первого рода α, кроме критерия Неймана–Пирсона, может быть предложен ряд критериев (правил выбора решения). Каждому из этих критериев будет соответствовать свое значение мощности критерия $w = 1 - \beta$. Критерий проверки гипотез, обладающий наибольшей мощностью по сравнению с другими критериями при одном и том же уровне значимости, называется *наиболее мощным критерием*. Поскольку критерий Неймана–Пирсона обеспечивает максимальную мощность решения в пользу нулевой гипотезы $w = 1 - \beta$ для заданного уровня значимости α, то он является наиболее мощным критерием проверки нулевой гипотезы.

Желательным свойством критерия проверки гипотез является его несмещенность. Критерий проверки гипотез называется *несмещенным*, если для заданных гипотез мощность критерия не меньше заданного уровня значимости, т.е. для мощности критерия выполняется условие

$$w = 1 - \beta \geqslant \alpha. \tag{5.1.48}$$

В противном случае он называется *смещенным* критерием. Например, критерий (5.1.39) для $m_1 < m_0$ будет смещенным, так как в этом случае мощность критерия меньше уровня значимости. В то же время критерий (5.1.39) для $m_1 > m_0$ будет несмещенным. Несмещенный критерий, *обеспечивающий наибольшую мощность среди других возможных несмещенных критериев*, называется *наиболее мощным несмещенным критерием*. При этом следует оговорить, что понятие несмещенности критерия в задачах проверки гипотез не связано с определением несмещенности при точечных оценках параметров распределений, рассмотренных выше.

Полученные выше соотношения применительно к критерию Неймана–Пирсона и ему подобным, в смысле фиксированного уровня значимости, *справедливы для любых значений* $0 \leqslant \alpha \leqslant 1$. При этом предполагалось, что величина α каким-то образом заранее выбрана и является малой, чтобы не отвергнуть нулевую гипотезу. Понятие «малости» ошибки первого рода, очевидно, является субъективным, если не определена «плата» за эту ошибку и ошибку второго рода. Если последнее не оговорено, то обычно величину ошибки первого рода выбирают из интервала $(0{,}001;\ 0{,}1)$ исходя из конкретно решаемой задачи. Необходимо иметь в виду, что при *фиксированном* объеме выборки n нельзя одновременно уменьшить ошибки первого и второго рода, так как с уменьшением α увеличивается β и, следовательно, уменьшается мощность критерия w. В пределе, когда $\alpha \to 0$, имеем $w \to 0$. Однако если имеется возможность увеличивать объем выборки n, то можно выбрать n таким, чтобы ошибки первого и второго рода были равны заданным значениям. При этом если α фиксировано, то при $n \gg 1$ (в пределе $n \to \infty$) все критерии будут обладать примерно *одинаковой мощностью*, в пределе стремящейся к единице $(w \to 1)$. Поэтому с увеличением n целесообразно одновременно уменьшать обе ошибки решения.

5.2. Проверка сложных гипотез

В *прикладной математической* статистике наиболее часто имеет место ситуация, для которой условное распределение выборки $f(x \mid H_0)$ нулевой гипотезы H_0 известно точно, а для альтернативной гипотезы $H \neq H_0$ условное распределение выборки точно неизвестно. Применительно к параметрическим гипотезам, например, нулевая гипотеза соответствует параметру распределения λ_0, а альтернативой гипотезе могут соответствовать $\lambda \neq \lambda_0$, $\lambda_0 > \lambda$ или $\lambda_0 < \lambda$.

В подобных ситуациях пространство *возможных* выборочных значений Γ разбивается на два подпространства Γ_0 и Γ_1 таким образом, чтобы критическая область Γ_1, где отклоняется нулевая гипотеза, *соответствовала* заданному уровню значимости (ошибке первого рода) α, т.е. критическая область Γ_1 находится из уравнения

$$\int_{\Gamma_1} f(\mathbf{x} \mid H_0)\,dx = \alpha.$$

Таким образом, нулевая гипотеза H_0 принимается, если конкретная выборка, полученная из опыта, не попадает в критическую область Γ_1. В противном случае *нулевая гипотеза* отклоняется. При этом ошибка

второго рода β и, естественно, мощность критерия $w = 1 - \beta$ будут зависеть от конкретной альтернативы. Для фиксированного значения ошибки первого рода $\alpha = \text{const}$ принципиально может существовать несколько правил выбора решения (критериев). Если из возможных критериев существует такой критерий, который для $\alpha = \text{const}$ обеспечивает максимальную мощность решения для всех возможных альтернатив, то такой критерий называется *равномерно наиболее мощным* (р.н.м.) критерием (правилом выбора решения). К сожалению, в большинстве прикладных задач равномерно наиболее мощный критерий выбора решения или не существует, или его трудно найти.

Так же, как и при проверке простой нулевой гипотезы H_0 против простой альтернативы H_1 и заданного уровня значимости α, в явном виде нельзя найти границы критической области Γ_1. В то же время на основе заданного (выбранного) уровня значимости α для разных критериев принятия решения можно определить пороговое значение h_α, превышение которого для данного критерия означает непринятие нулевой гипотезы с заданным уровнем значимости. Для этого необходимо выбрать (предложить, найти и т.д.) некоторую однозначную меру отклонения U опытных данных от данных, соответствующих проверяемой гипотезе H_0. Мера U представляет по существу статистику для проверки гипотезы. Из-за случайности самой выборки статистика U является случайной величиной, и применительно к нулевой гипотезе может быть определено ее распределение. Как правило, пытаются подобрать такую статистику, чтобы ее распределение сводилось к одному из известных и табулированных распределений (например, к распределениям типа нормального, хи-квадрат, Стьюдента, дисперсионного отношения). Будем полагать, что случайная величина U является непрерывной величиной и характеризуется условной плотностью вероятности $f(u \mid H_0)$. Тогда для заданного уровня значимости (ошибки первого рода) можно определить правило, разделяющее интервал возможных значений случайной величины на подынтервалы, соответствующие области принятия решения и критической области.

Если сформированная выборочная статистика u^* попадает в критическую область, то проверяемая гипотеза H_0 отвергается, и наоборот, если сформированная статистика u^* попадает в область принятия решения, то проверяемая гипотеза H_0 принимается с уровнем значимости α.

Конкретизируем сказанное выше применительно к параметрическим гипотезам, для которых условные плотности выборочных данных зависят от параметров. Итак, пусть нулевой гипотезе H_0 соответствует значение параметра λ_0, а альтернативной гипотезе может соответствовать одно из значений $\lambda \neq \lambda_0$ из интервала возможных значений Λ.

Альтернативы бывают *односторонние* (правосторонняя $\lambda_0 < \lambda$ и левосторонняя $\lambda < \lambda_0$) и *двусторонние* (одновременно $\lambda > \lambda_0$, $\lambda < \lambda_0$, $\lambda \neq \lambda_0$). Для каждой фиксированной альтернативы λ можно построить правило выбора решения (критерий) Неймана–Пирсона, т.е. сложную альтернативу можно рассматривать как совокупность простых альтернатив. Если выбранная на основе критерия Неймана–Пирсона критическая область зависит от произвольного значения параметра λ альтернативы, то не существует критической области, которая была бы наилучшей для всех значений параметра λ, следовательно, не существует р.н.м. критерия. Если для двусторонней альтернативы не удается найти р.н.м. критерий, то для односторонних альтернатив такой критерий существует. В последнем случае этот критерий должен совпадать с критерием Неймана–Пирсона, который для рассматриваемых монотонно изменяющихся отношений правдоподобия обеспечивает максимум мощности $w = 1 - \beta$ при принятии решения в пользу нулевой гипотезы H_0. Применительно к принятию простой гипотезы против сложной альтернативы вместо мощности критерия рассматривают *функцию мощности критерия* $w(\lambda) = 1 - \beta(\lambda)$, зависящую от текущего значения параметра λ. Если критическая область не зависит от конкретного значения параметра λ альтернативы и для всех значений λ выполняется условие

$$w(\lambda) = 1 - \beta(\lambda) \geqslant \alpha, \tag{5.2.1}$$

то такой критерий *называется несмещенным критерием* проверки гипотезы H_0 ($\lambda = \lambda_0$) против сложной альтернативы H_1 ($\lambda \neq \lambda_0$). Несмещенный критерий, функция мощности которого для всех альтернативных значений параметра больше функций мощности других возможных несмещенных критериев, называется *равномерным наиболее мощным* (р.н.м.) *несмещенным критерием*. В [27, 31, 37] приведены условия существования р.н.м. несмещенных критериев.

Так же, как при проверке простой гипотезы против простой альтернативы, необходимо задаться уровнем значимости α. В данной ситуации задача усложняется тем, что альтернатива неизвестна, и потому при выбранной малой величине ошибки первого рода ошибка второго рода может оказаться достаточно большой, в результате чего мощность критерия будет весьма малой. Поэтому на основе анализа конкретной задачи выясняют наиболее близкое значение альтернативы и для этой альтернативы решают задачу аналогично задаче проверки простой гипотезы против простой альтернативы.

Если распределения рассматриваемых гипотез, кроме параметров, подлежащих проверке, содержат ряд неизвестных параметров, то такие гипотезы, как отмечалось выше, являются непараметрическими. В этих ситуациях возможны два подхода. Если известны априорные распределения неизвестных параметров, исходные распределения для обеих гипотез следует усреднить по неизвестным параметрам. Если априорные распределения неизвестных параметров отсутствуют, то неизвестные параметры заменяются соответствующими им оценками.

Критерии проверки простых параметрических гипотез против сложных альтернатив тесно связаны с интервальными оценками, рассмотренными в п. 4.7. Обозначим согласно (4.7.6) через $\lambda_1 = \psi_1(\lambda^*, \gamma)$ и

$\lambda_2 = \psi_2(\lambda^*, \gamma)$ — центральные доверительные границы для истинного значения параметра λ_0, определяемые на основе оценки неизвестного параметра λ^* и доверительной вероятности $\gamma = 1 - \alpha$. Для этих доверительных границ справедлива запись

$$\mathbf{P}\{\lambda_1 \leqslant \lambda_0 \leqslant \lambda_2\} = \gamma = 1 - \alpha.$$

Двойное неравенство принципиально можно решить относительно оценки параметра λ^* и получить две функции $\varsigma_1(\lambda_0, \gamma)$ и $\varsigma_2(\lambda_0, \gamma)$, для которых остается справедливым неравенство

$$\mathbf{P}\{\varsigma_1 \leqslant \lambda^* \leqslant \varsigma_2\} = 1 - \alpha. \tag{5.2.2}$$

Так как величины ς_1 и ς_2 зависят от значения параметра λ_0, соответствующего нулевой гипотезе, их можно рассматривать как пороговые величины, а неравенство в (5.2.2) определяет правило принятия нулевой гипотезы с уровнем значимости α. Аналогично из правила принятия простой гипотезы против сложной альтернативы можно получить границы доверительного интервала. Следует отметить, что если существует р.н.м. критерий проверки гипотезы, то он соответствует наименьшему доверительному интервалу и наоборот [27].

Пример 5.2.1. Рассмотрим высказанные суждения на примере проверки гипотезы о среднем значении нормального распределения с учетом результатов, полученных в примере 5.1.1 применительно к использованию критерия Неймана–Пирсона. Так как в соответствии с критерием Неймана–Пирсона уровни значимости α задаются априори, то критические области («пороги») $m_0 + h_\alpha$ (для $m_1 > m_0$) и $m_0 - h_\alpha$ (для $m_1 < m_0$), естественно, не зависят от альтернативы m_1. Поэтому каждый из полученных критериев сравнения выборочного среднего с порогами согласно (5.1.39) и (5.1.42) является р.н.м. критерием только для своей критической области. Обозначенные разными индексами, функции мощности рассматриваемых односторонних критериев аналогично (5.1.41) и (5.1.44) соответственно равны

$$w_1(m) = 1 - \Phi\left(\frac{h_\alpha - (m - m_0)}{\sigma/\sqrt{n}}\right), \quad m > m_0; \tag{5.2.3}$$

$$w_2(m) = 1 - \Phi\left(\frac{h_\alpha - (m_0 - m)}{\sigma/\sqrt{n}}\right), \quad m < m_0. \tag{5.2.4}$$

Первая функция мощности критерия с увеличением m монотонно возрастает, а вторая — монотонно убывает; максимальные их значения равны единице. При $m = m_{1,2} = m_0 \pm h_\alpha$ они равны $1/2$, а для $m = m_0$ значения функций мощности критериев равны уровню значимости, т.е. $w_{1,2}(m_0) = \alpha$. На рис. 5.2.1 схематично изображены функции мощности $w(m)$. Для полного класса альтернатив $m \neq m_0$ не существует р.н.м. критерия.

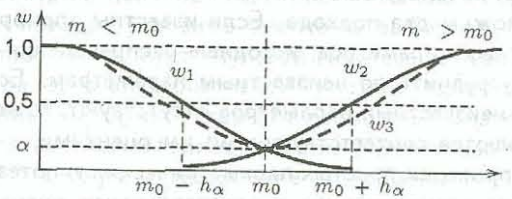


Рис. 5.2.1. Функции мощности при проверке гипотезы о среднем значении нормального распределения

Действительно, если в качестве критерия принятия нулевой гипотезы взять правило (5.1.39), равномерно наиболее мощное для $m > m_0$, то при $m < m_0$ функция мощности критерия $w_1(m) = 1 - \beta_1(m)$ меньше уровня значимости α, т.е. правило проверки нулевой гипотезы (5.1.39) смещенное. Аналогично, если в качестве критерия принятия нулевой гипотезы взять правило (5.1.42), равномерно наиболее мощное для $m < m_0$, то при $m_0 > m$ функция мощности критерия $w_2(m) = 1 - \beta_2(m)$ также меньше уровня значимости α, т.е. и в этом случае правило проверки нулевой гипотезы (5.1.39) смещенное. Таким образом, для полного класса альтернатив $m \neq m_0$ не существует р.н.м. критерия, так как всегда при одном и том же уровне значимости α каждый из двух односторонних критериев обеспечивает наибольшую мощность критерия только в своем диапазоне значений величины m.

Для альтернативы, включающей все значения параметра $m \neq m_0$, в качестве меры отличия выборочного среднего от гипотетического среднего m_0 можно рассматривать модуль отклонения выборочного среднего от его гипотетического значения, а в качестве критерия, определяющего критическую область, рассматривать соотношение

$$\left| \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i - m_0 \right| > h, \tag{5.2.5}$$

где, как и выше, h — пороговое значение, определяемое исходя из заданного уровня значимости критерия α:

$$\mathbf{P}\{|\bar{x} - m_0| > h\} = \alpha. \tag{5.2.6}$$

Следовательно, решение в пользу нулевой гипотезы H_0 определяется двойным неравенством

$$-h \leqslant \bar{x} - m_0 \leqslant h, \tag{5.2.7}$$

или

$$(m_0 - h) \leqslant \bar{x} \leqslant (m_0 + h). \tag{5.2.8}$$

Критерии (5.2.7) и (5.2.8) в отличие от рассмотренных ранее критериев являются двусторонними. Нулевая гипотеза H_0 не принимается, если абсолютная величина отклонения выборочного среднего от гипотетического m_0 больше выбранного порога h. Так как события $(\bar{x} - m_0) > h$ и $(\bar{x} - m_0) < h$ несовместны, правило определения критической области (5.2.6) перепишем в виде

$$\mathbf{P}\{(\bar{x} - m_0) > h\} + \mathbf{P}\{(\bar{x} - m_0) < -h\} = \alpha.$$

Учитывая, что выборочное среднее для нулевой гипотезы симметрично относительно m_0, обе вероятности в левой части последнего выражения равны между собой. Поэтому значение величины h, определяющей границы критической области, находится из уравнения

$$\mathbf{P}\{(\bar{x} - m_0) > h\} = \mathbf{P}\{\bar{x} > (h + m_0)\} =$$

$$= \int_{h+m_0}^{\infty} f(\bar{x} \mid H_0)\, d\bar{x} = 1 - \Phi\left(z_{\alpha/2} = \frac{h \equiv h_{\alpha/2}}{\sigma/\sqrt{n}}\right) = \frac{\alpha}{2}, \tag{5.2.9}$$

из которого окончательно получаем выражение для определения порогового значения $h = h_{\alpha/2}$

$$\Phi\left(z_{\alpha/2} = \frac{h_{\alpha/2}}{\sigma/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}. \tag{5.2.10}$$

Здесь $z_{\alpha/2}$ — нормированное значение порога, являющееся критическим значением нормального распределения.

Функция мощности, соответствующая правилу (5.2.5), определяется соотношением

$$w_3(m) = 1 - \beta(m) = 1 - \int_{m_0-h}^{m_0+h} f(\bar{x} \mid H_1)\, d\bar{x} =$$

$$= 1 - \Phi\left[\frac{h_{\alpha/2} - (m - m_0)}{\sigma/\sqrt{n}}\right] + \Phi\left[\frac{-h_{\alpha/2} - (m - m_0)}{\sigma/\sqrt{n}}\right]. \qquad (5.2.11)$$

Функция мощности критерия достигает минимума при $m = m_0$ и равна уровню значимости критерия:

$$w_3(m_0) = 2\left[1 - \Phi\left(\frac{h_{\alpha/2}}{\sigma/\sqrt{n}}\right)\right] = \alpha,$$

т.е. двусторонний критерий (5.2.8) является несмещенным критерием. Функция мощности двустороннего критерия $w_3(m)$ не превышает функций мощности односторонних критериев $w_1(m)$ и $w_2(m)$. Например, пусть $m > m_0$.

Для сравнения функций мощности учтем неравенство

$$\Delta(c) = \Phi(z_{\alpha/2} - c) - \Phi(-z_{\alpha/2} - c) - \Phi(z_\alpha - c) > 0, \qquad (5.2.12)$$

справедливое для произвольного $0 \leqslant c \leqslant \infty$ и $0 \leqslant \alpha \leqslant 1$. Это неравенство доказывается путем исследования поведения функции $\Delta(c)$ и ее первой производной $\Delta'(c)$ на интервале $0 \leqslant c \leqslant \infty$, учитывая при этом, что $\Phi(z_{\alpha/2}) = 1 - \alpha/2 > \Phi(z_\alpha) = 1 - \alpha$. Из последнего следует, что $z_{\alpha/2} > z_\alpha$. Наиболее просто в справедливости неравенства (5.2.12) можно убедиться расчетами применительно к конкретным числовым значениям величин c и α.

Подставляя в неравенство (5.2.12) соответствующие величины из (5.2.11) и (5.2.3), получаем $w_1(m) > w_3(m)$. На рис. 5.2.1 в виде пунктирной кривой изображена функция мощности $w_3(m)$ двустороннего критерия принятия нулевой гипотезы (5.2.8). В [27, 37] показано, что критерий (5.2.8) является не только несмещенным, но и двусторонним р.н.м. критерием.

Критерий (5.2.8) может быть получен из выражения центральных доверительных интервалов интервальной оценки среднего значения нормального распределения (п. 4.8.1). Двойное неравенство в (4.8.6) элементарно преобразуется в двойное неравенство вида

$$m - \frac{t_\gamma \sigma}{\sqrt{n}} \leqslant \bar{x} \leqslant m + \frac{t_\gamma \sigma}{\sqrt{n}}.$$

Полагая в выражении (4.8.6) $\gamma = 1 - \alpha$ и $t_\gamma = z_{\alpha/2}$, получаем правило решения в виде (5.2.8).

5.3. Прикладные задачи проверки параметрических гипотез

В рассматриваемых ниже задачах будем полагать, что альтернативные значения параметров отсутствуют, а решение о принятии или отказе от нулевой гипотезы осуществляется на основе использования независимой выборки.

5.3.1. Проверка гипотез о средних значениях выборок

5.3.1.1. Гипотеза о равенстве выборочного среднего заданному значению

Пусть дана независимая выборка объема n из некоторой генеральной совокупности нормально распределенной случайной величины с неизвестными средним значением и дисперсией. Проверяемая гипотеза

состоит в том, что среднее значение m этой генеральной совокупности равно некоторому значению m_0 против альтернативы $m \neq m_0$.

Образуем статистику

$$T = \frac{\bar{x} - m_0}{s/\sqrt{n}}, \qquad (5.3.1)$$

где, как и выше, $\bar{x}$ и s^2 — выборочные среднее и дисперсия, определяемые соответственно (3.4.5) и (3.4.27). При нулевой гипотезе статистика (5.3.1) имеет t-распределение (распределение Стьюдента) с $\nu = (n-1)$ степенями свободы. Конкретное выборочное значение статистики обозначим t^*. Учитывая симметричность t-распределения относительно $t = 0$, критическая область должна быть симметрична. Поэтому критическое значение статистики $\pm t_{\alpha/2;\nu}$ определяется из таблиц распределения Стьюдента согласно выражению

$$\mathbf{P}\{T > t_{\alpha/2;\nu}\} = \int_{t_{\alpha/2;\nu}}^{\infty} f(t)\,dt = \frac{\alpha}{2}. \qquad (5.3.2)$$

При этом нулевая гипотеза, утверждающая, что среднее значение генеральной совокупности равно m_0, принимается тогда, когда для выборочной статистики t^* выполняется соотношение

$$-t_{\alpha/2;\nu} \leqslant t^* = \frac{\bar{x} - m_0}{s/\sqrt{n}} \leqslant t_{\alpha/2;\nu}. \qquad (5.3.3)$$

Критерий проверки гипотезы о равенстве среднего значения генеральной совокупности m заданному значению m_0 на основе статистики (5.3.1) называется *t-критерием* или *критерием Стьюдента* проверки гипотезы.

5.3.1.2. Проверка гипотезы о равенстве средних значений

Пусть дана выборка из n_1 значений случайной величины X и выборка из n_2 значений случайной величины Y. Будем полагать, что дисперсии случайных величин одинаковы $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, но неизвестны. Проверяемая гипотеза H_0 состоит в том, что средние значения m_X и m_Y гауссовских случайных величин X и Y равны между собой $m_X = m_Y$ против альтернативы $m_X \neq m_Y$. Следовательно, критерий проверки гипотезы является двусторонним.

Наиболее полная информация о средних значениях случайной величины содержится в их выборочных значениях, распределенных по нормальным законам со средними значениями и дисперсиями соответственно равными m_X, m_Y, σ^2/n_1 и σ^2/n_2. При известной дисперсии σ^2 в качестве измеряемой статистики можно было бы взять величину

$$Z = \frac{\bar{x} - \bar{y}}{\sigma_{(X-Y)}} = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}},$$

которая для нулевой гипотезы $(m_X = m_Y)$ имеет нулевое среднее значение и единичную дисперсию. Аналогично примеру 5.2.1 для заданного уровня значимости можно было бы построить критерий проверки нулевой гипотезы. Поскольку истинное значение дисперсии неизвестно, воспользуемся объединенной ее оценкой (выборочной дисперсией), полученной из обеих выборок:

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right]. \qquad (5.3.4)$$

Так как две степени свободы использованы для определения выборочных средних значений $\bar{x}$ и $\bar{y}$, в последней формуле число степеней свободы по сравнению с суммарным объемом выборок $n = n_1 + n_2$ уменьшено на 2. Нормированная объединенная оценка дисперсии s^2/σ^2 распределена по закону хи-квадрат с $\nu = n_1 + n_2 - 2$ степенями свободы. Вместо величины Z образуем статистику

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s^2/n_1 + s^2/n_2}}, \qquad (5.3.5)$$

конкретное значение которой, получаемое в результате проведения эксперимента (выборочная статистика), обозначим как t^*. Применительно к нулевой гипотезе образованная статистика распределена по закону Стьюдента с тем же числом степеней свободы, что и объединенная выборочная дисперсия. При этом учтем, что плотность вероятности Стьюдента симметрична относительно $t = 0$. Для проверки нулевой гипотезы для заданных уровня значимости α и числа степеней свободы ν пороговые значения критерия (критические значения статистики) $\pm t_{\alpha/2;\nu}$ определяются по таблицам распределения Стьюдента согласно выражению (5.3.2).

Вместо критических значений статистики (5.3.5) иногда оперируют с критическими значениями разности выборочных средних:

$$(\bar{x} - \bar{y})_{\alpha/2;\nu} = -(\bar{x} - \bar{y})_{(1-\alpha/2);\nu} = t_{\alpha/2;\nu}\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}. \qquad (5.3.6)$$

Нулевая гипотеза о равенстве средних значений случайных величин с одинаковыми дисперсиями принимается с уровнем значимости α, если вычисленная для конкретных выборок статистика (5.3.5) удовлетворяет двойному неравенству

$$-t_{\alpha/2;\nu} < t^* = \frac{\bar{x} - \bar{y}}{\sqrt{s^2/n_1 + s^2/n_2}} < t_{\alpha/2;\nu}. \qquad (5.3.7)$$

В противоположном случае с уровнем значимости α гипотеза о равенстве средних значений отвергается.

Пример 5.3.1. Произведены две серии измерений при проведении некоторого эксперимента:

1 серия x_i: 15,7; 10,3; 12,6; 14,5; 12,6; 13,8; 11,9 ($n_1 = 7$),

2 серия y_i: 12,3; 13,7; 10,4; 11,4; 14,9; 12,6 ($n_2 = 6$).

На основе сравнения выборочных средних необходимо проверить гипотезу о том, что обе серии — это случайные выборки из одной и той же нормальной совокупности. Выборочные средние равны

$$\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n} x_i = 13{,}06; \quad \bar{y} = \frac{1}{n_2}\sum_{i=1}^{n} y_i = 12{,}55.$$

Разность (различие) выборочных средних $\bar{x} - \bar{y} = 0{,}51$.

Надо выяснить, случайно ли получено это различие или действительно оно обусловлено изменившимися условиями эксперимента для уровня значимости $\alpha = 0{,}05$. Уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $\nu = 13 - 2 = 11$ (согласно табл. 3 Приложения 2) соответствует критическое значение статистики $t_{0,025;11} = 2{,}201$. Из приведенных выборочных данных следует

$$\sum_{i=1}^{n_1=7}(x_i - 13{,}06)^2 = 18{,}98; \quad \sum_{i=1}^{n_2=6}(y_i - 12{,}55)^2 = 12{,}86; \quad s^2 = 2{,}89.$$

Подставляя найденные величины в (5.3.5), находим выборочное значение статистики $t^* = 0{,}54$. Поскольку полученное значение выборочной статистики t^* удовлетворяет двойному неравенству

$$t_{0,025;11} = -2{,}201 < t^* = 0{,}54 < t_{0,025;11} = 2{,}201,$$

принимается гипотеза о равенстве средних значений с уровнем значимости $\alpha = 0{,}05$. При этом границы критической области для разности выборочных средних в соответствии с (5.3.6) $(\bar{x} - \bar{y})_{0,025;11} = \pm 2{,}082$.

Проверим нулевую гипотезу о равенстве средних значений h гауссовских независимых выборок $\mathbf{x}_1(x_{11}, \ldots, x_{1n_1}), \ldots, \mathbf{x}_h(x_{h1}, \ldots, x_{hn_h})$ с одинаковыми, но неизвестными дисперсиями. Здесь n_i — объем i-й выборки.

В качестве статистики может быть взята величина, применяемая в однофакторном дисперсионном анализе [2, 15, 28 и др.]

$$F_{h-1,n-h} = \frac{\frac{1}{h-1}\sum_{i=1}^{h} n_i(x_{i\cdot} - x_{\cdot\cdot})^2}{\frac{1}{n-h}\sum_{i=1}^{h}\sum_{j=1}^{n_i}(x_{ij} - x_{i\cdot})^2}, \qquad (5.3.8)$$

где $n = \sum_{i=1}^{h} n_i$ — общий (суммарный) объем всей совокупности выборок, а $x_{i\cdot}$ и $x_{\cdot\cdot}$ — выборочные средние по каждой из выборок и общее среднее:

$$x_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}; \quad x_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i} x_{ij}.$$

Статистика (5.3.8) имеет F-распределение со степенями свободы числителя и знаменателя соответственно $\nu_1 = h - 1$ и $\nu_2 = n - h$, так как применительно к нулевой гипотезе выражения, стоящие в числителе и знаменателе, оценивают неизвестную дисперсию выборок. Нулевая гипотеза не отвергается с уровнем значимости α, если выборочное значение статистики $F^*_{\nu_1,\nu_2}$ не превышает его критического значения $F_{\alpha;\nu_1,\nu_2}$. Более доказательное обоснование распределения статистики

F_{ν_1,ν_2} приведено ниже при рассмотрении однофакторного дисперсионного анализа (п. 6.1).

5.3.2. Проверка гипотезы о равенстве дисперсий

На основе имеющихся двух выборок размера n_1 и n_2, соответствующих гауссовским случайным величинам X и Y, необходимо проверить гипотезу о равенстве их дисперсий $\sigma_X^2 = \sigma_Y^2$. Для неизвестных дисперсий сформируем две независимые оценки

$$s_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2; \quad s_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2. \tag{5.3.9}$$

При этом случайные величины

$$\mathbf{X}_X^2 = \frac{(n_1 - 1)s_X^2}{\sigma_X^2}; \quad \mathbf{X}_Y^2 = \frac{(n_2 - 1)s_Y^2}{\sigma_Y^2} \tag{5.3.10}$$

имеют хи-квадрат распределения с $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$ степенями свободы соответственно. Если верна гипотеза $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, то отношения

$$\mathbf{X}_X^2/\mathbf{X}_Y^2 = (\nu_1 s_X^2)/(\nu_2 s_Y^2); \quad \mathbf{X}_Y^2/\mathbf{X}_X^2 = (\nu_2 s_Y^2)/(\nu_1 s_X^2) \tag{5.3.11}$$

не зависят от истинного значения дисперсии σ^2. Поэтому для проверки рассматриваемой гипотезы в качестве статистики критерия можно взять отношение выборочных дисперсий (*статистика дисперсионного отношения* или *F-статистика*)

$$F_{\nu_1,\nu_2} = s_X^2/s_Y^2 \equiv F \tag{5.3.12}$$

или

$$F_{\nu_2,\nu_1} = s_Y^2/s_X^2 \equiv F'. \tag{5.3.13}$$

Применительно к проверяемой гипотезе статистики F и F' являются случайными величинами, распределенными по закону дисперсионного отношения (2.2.26) с числом степеней свободы (ν_1, ν_2) и (ν_2, ν_1) соответственно. Критерий проверки гипотезы, основанный на статистике отношения дисперсий, называется *критерием дисперсионного отношения* или *F-критерием*. Он должен быть двусторонним, так как возможны несовместные события $s_X^2 > s_Y^2$ или $s_X^2 < s_Y^2$. Применительно к F-статистике симметричная критическая область принимает два интервала $0 < F < F_{(1-\alpha/2);\nu_1,\nu_2}$ и $F > F_{\alpha/2;\nu_1,\nu_2}$, исходя из очевидных соотношений

$$\mathbf{P}\{F < F_{(1-\alpha/2);\nu_1,\nu_2} \equiv F_1\} = \alpha/2; \quad \mathbf{P}\{F > F_{\alpha/2;\nu_1,\nu_2} \equiv F_2\} = \alpha/2. \tag{5.3.14}$$

Аналогично для статистики $F' = 1/F$ критическая область определяется выражениями

$$\mathbf{P}\{F' < F_{(1-\alpha/2);\nu_2,\nu_1} \equiv F_1'\} = \alpha/2; \quad \mathbf{P}\{F' > F_{\alpha/2;\nu_2,\nu_1} \equiv F_2'\} = \alpha/2. \tag{5.3.15}$$

Найдем зависимость между верхними и нижними границами критических значений статистик F и F'. На основании соотношений (5.3.14) и (5.3.15) можно записать

$$\mathbf{P}\{F < F_1\} = \mathbf{P}\left\{\frac{1}{F} > \frac{1}{F_1}\right\} = \mathbf{P}\left\{F' > \frac{1}{F_1}\right\} = \frac{\alpha}{2} = \mathbf{P}\{F' > F_2'\}.$$

Этот результат показывает, что значение левой критической границы распределения F-статистики обратно пропорционально значению правой критической границе F'-статистики, т.е.

$$F_{(1-\alpha/2);\nu_1,\nu_2} = 1/F_{\alpha/2;\nu_2,\nu_1}.$$

Таким образом, необходимо найти только правые критические значения для статистик $F = s_X^2/s_Y^2$ и $F' = s_Y^2/s_X^2$. Можно также показать, что верхние и нижние критические значения статистик связаны соотношением

$$F_{(1-\alpha/2);\nu_2,\nu_1} = 1/F_{\alpha/2;\nu_1,\nu_2}.$$

Поэтому нулевая гипотеза о равенстве дисперсий принимается, если выборочная статистика удовлетворяет одному из неравенств

$$\begin{aligned} F_{\nu_1,\nu_2} &= (s_X^2/s_Y^2) \leqslant F_{\alpha/2;\nu_1,\nu_2} \text{ при } (s_X^2/s_Y^2) \geqslant 1; \\ F_{\nu_2,\nu_1} &= (s_Y^2/s_X^2) \leqslant F_{\alpha/2;\nu_2,\nu_1} \text{ при } (s_Y^2/s_X^2) \geqslant 1. \end{aligned} \tag{5.3.16}$$

Пример 5.3.2. Пусть на основе двух выборок объемов $n_1 = 11$ и $n_2 = 17$ вычислены значения выборочных дисперсий $s_X^2 = 34$ и $s_Y^2 = 17$.

Проверим гипотезу о равенстве дисперсий и определим границы области принятия решения для уровня значимости $\alpha = 0{,}10$. Значение выборочной статистики равно $F_{10,16}^* = (s_X^2/s_Y^2) = 2$, поэтому для принятия нулевой гипотезы должно удовлетворяться первое из неравенств (5.3.16). По заданным значениям степеней свободы $\nu_1 = n_1 - 1 = 10$, $\nu_2 = n_2 - 1 = 16$ и $\alpha/2 = 0{,}05$ по таблицам для дисперсионного отношения (табл. 4 Приложения 2) находим критическое значение статистики $F_{0,05;10,16} = 2{,}49$. Так как $F_{10,16}^* = 2 < 2{,}49$, с уровнем значимости $\alpha = 0{,}10$ принимается гипотеза о равенстве дисперсий.

Рассмотрим ситуацию, когда на основе имеющихся h выборок объемов $n_1, n_2, \ldots, n_h$, соответствующих гауссовским случайным величинам $X_1, X_2 \ldots, X_h$, необходимо проверить нулевую гипотезу о равенстве их дисперсий:

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_h^2 = \sigma^2.$$

По имеющимся выборкам согласно (3.4.27) можно вычислить несмещенные выборочные дисперсии $s_1^2, s_2^2, \ldots, s_h^2$ со степенями свободы $\nu_i = n_i - 1$, $i = 1, 2, \ldots, h$. Применительно к нулевой гипотезе $\mathbf{M}\{s_i^2\} = \sigma^2$.

Если нулевая гипотеза справедлива, то случайная величина

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{h}\nu_i s_i^2$$

представляет собой обобщенную (среднее значение) выборочную дисперсию со степенью свободы $\nu = \sum_{i=1}^{h} \nu_i$. При этом распределение частного s_i^2/s^2 зависит лишь от степеней свободы $\nu_1, \nu_2, \ldots, \nu_h$.

Для проверки нулевой гипотезы был предложен критерий, называемый *критерием Бартлета* [8, 15, 51, 72 и др.], основанный на статистике

$$B_{\nu_B} = \frac{1}{C}\left(\nu \ln s^2 - \sum_{i=1}^{h} \nu_i \ln s_i^2\right), \qquad (5.3.17)$$

где

$$C = 1 + \frac{1}{3(h-1)}\left(\sum_{i=1}^{h} \frac{1}{\nu_i} - \frac{1}{\nu}\right). \qquad (5.3.18)$$

При $\nu_i > 3$ распределение статистики B_{ν_B} приближенно является хи-квадрат распределением с $\nu_B = h - 1$ степенями свободы; поэтому определение критических значений $B_{\alpha;\nu_B}$ по заданному уровню значимости α осуществляется по таблицам критических значений хи-квадрат распределения. Так как $C \approx 1$, эту постоянную, как правило, можно не учитывать, если для $C = 1$ критическое значение статистики не является значимым при отклонении нулевой гипотезы. Критерий Бартлета весьма чувствителен к отклонениям истинных распределений величин $\nu_i s_i^2/\sigma_i^2$ от хи-квадрат распределения. Для более точного определения критических значений статистики (5.3.17) имеются соответствующие таблицы [8, 51, 72 и др.].

Пример 5.3.3. С уровнем значимости $\alpha = 0{,}05$ проверим гипотезу о равенстве дисперсий совокупности пяти выборок ($h = 5$), для которых выборочные значения дисперсий s_i^2 и их степени свободы ν_i представлены в табл. 5.3.1 во втором и третьем столбцах. В остальных столбцах приведены величины, необходимые для последующих вычислений. На основе приведенных данных имеем $s^2 = 69{,}455$; $\ln s^2 = 4{,}2406$; $C \approx 1{,}068$; $\nu_B = 4$; $B_4^* \approx 2{,}02$. Для заданного уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $\nu_B = 4$ согласно табл. 2 Приложения 2 критическое значение статистики $B_{0{,}05;4} = 9{,}49$. Так как $B_4^* \approx 2{,}02 < B_{0{,}05;4} = 9{,}49$, то нулевая гипотеза о равенстве дисперсий не отвергается.

Таблица 5.3.1

i	s_i^2	ν_i	$1/\nu_i$	$\nu_i s_i^2$	$\ln s_i^2$	$\nu_i \ln s_i^2$
1	35	8	0,125	280	3,5554	28,4432
2	92	6	0,167	552	4,5218	27,1308
3	65	4	0,250	260	4,1744	16,6976
4	100	5	0,200	500	4,6052	23,0260
5	70	10	0,100	700	4,2485	42,4850
$\sum$		33	0,842	2292		137,7826

Если объемы всех выборок одинаковы ($n_i \equiv n_0$, $i = 1, 2, \ldots, h$), то $\nu_i \equiv \nu_0$ и $\nu = h\nu_0$. При этом формула для статистики (5.3.17) существенно упрощается

$$B_{\nu_B} = \frac{1}{C} h\nu_0 \left(\ln s^2 - \frac{1}{h}\sum_{i=1}^{h} \ln s_i^2\right); \quad C = 1 + \frac{h+1}{3h\nu_0}. \qquad (5.3.19)$$

5.3.3. Проверка гипотезы об аномальных наблюдениях

При проведении экспериментов некоторые измеренные значения могут существенно отличаться от остальных. Если причина подобного отличия известна, показание, содержащее грубую ошибку, исключается из дальнейшей обработки результатов измерений. Решение отбросить какие-либо данные в конечном счете всегда субъективно. Наиболее строгий подход к проблеме анализа подобных аномальных данных — повторить измерения много раз. Если подобная аномалия повторится снова, вероятно, можно будет выяснить ее причину. Если же аномальное показание не повторится, то, даже оставляя его, оно мало скажется на дальнейших вычислениях из-за большого числа измерений. Однако при небольшом объеме опытных данных и отсутствии явных причин появления грубых ошибок целесообразно использовать статистические методы выявления грубой ошибки. Статистические методы, как правило, основаны на проверке нулевой гипотезы о том, что аномальное показание относится к тому же множеству, что остальные опытные данные. Естественно, для проверки этой гипотезы необходимо выбрать соответствующий критерий, согласно которому отвергается подозрительный (аномальный) результат с заданным уровнем значимости.

Итак, пусть в результате проведения эксперимента получена некоторая гауссовская выборка $\mathbf{x}(x_1, x_2, \ldots, x_m, \ldots, x_n)$ объема n. Одно из выборочных значений x_m существенно отличается от остальных. Необходимо решить, является оно грубой ошибкой или обусловлено случайной природой всех выборочных значений. Принципиально имеется несколько критериев проверки принадлежности аномального выборочного значения нулевой гипотезе. Однако наиболее простым и достаточно мощным является двусторонний несмещенный *критерий аномального наблюдения*, основанный на формировании статистики нормированного максимального абсолютного отклонения аномального показания от выборочного среднего, т.е. в качестве статистики критерия рассматривается величина

$$U_n = \frac{\max|x_i - \bar{x}|}{s}, \quad i = 1, 2, \ldots, n. \qquad (5.3.20)$$

Здесь, как и выше, выборочные среднее и дисперсия определяются выражениями

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Для статистики (5.3.20) получены законы распределения вероятностей и составлены таблицы критических значений распределения наибольшего по абсолютной величине нормированного выборочного отклонения в зависимости от объема выборки n и уровня значимости α [8], позволяющие найти по заданным значениям α и n критическое значение статистики $u_{\alpha,n}$, такое что

$$\mathbf{P}\{U > u_{\alpha;n}\} = \alpha. \tag{5.3.21}$$

Частично эти данные приведены в табл. 5 Приложения 2. Поэтому, если значение выборочной статистики u_n^*, полученное на основе опытных данных, не превышает критического значения $u_{\alpha,n}$, т.е. если

$$u_n^* = \frac{\max|x_i - \bar{x}|}{s} \leqslant u_{\alpha,n}, \tag{5.3.22}$$

то гипотеза о том, что аномальное значение x_m, $i = 1, 2, \ldots, n$, не является грубой ошибкой, принимается с уровнем значимости α. Следовательно, это аномальное значение (аномальное выборочное значение) должно быть оставлено для дальнейшего использования.

Пример 5.3.4. В результате опыта проведено шесть измерений и получены следующие данные: 42, 38, 20, 35, 30, 34. Здесь значение $x_3 = 20$ существенно отличается от остальных. Проверим в соответствии с критерием (5.3.22) гипотезу о том, что значение $x_3 = 20$ не является грубой ошибкой для уровня значимости критерия $\alpha = 0{,}05$. Из табл. 5 Приложения 2 для $n = 6$ и $\alpha = 0{,}05$ находим критическое значение статистики $u_{0,05;6} = 2{,}07$. Опытным данным соответствуют $\bar{x} = 33{,}17$, $s = 6{,}93$ и $u_6^* = 1{,}90$. Так как значение статистики, вычисленное на основе опытных данных, меньше критического значения, можно утверждать, что $x_3 = 20$ не является грубой ошибкой.

5.3.4. Проверка гипотезы о равенстве вероятностей событий

В результате двух серий независимых опытов объема n_1 и n_2 получено число положительных исходов (событий) k_1 и k_2, что соответствует частотам событий $p_1^* = k_1/n_1$ и $p_2^* = k_2/n_2$. Необходимо проверить гипотезу о равенстве двух вероятностей p_1 и p_2, т.е. $p_1 = p_2 = p$ против сложной альтернативы $p_1 \neq p_2$. Следовательно, критерий проверки гипотезы должен быть двусторонним.

Когда объемы выборок достаточно велики, так что выполняются условия типа $np \gg 1$ и $n(1-p) \gg 1$ (практически не менее 5), выборочные распределения частот приближенно можно полагать нормальными со средними значениями $m_1 = p_1$, $m_2 = p_2$ и дисперсиями $D_1 = p_1(1-p_1)/n_1$, $D_2 = p_2(1-p_2)/n_2$. Если проверяемая гипотеза верна, то разность $p_1^* - p_2^*$ будет нормально распределена с нулевым средним значением и суммарной дисперсией

$$D = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{5.3.23}$$

Поскольку истинное значение вероятности p неизвестно, вместо p используем ее оценку в виде суммарной частоты положительных исходов (событий)

$$p^* = \frac{k_1 + k_2}{n_1 + n_2} = \frac{k}{n}. \tag{5.3.24}$$

В качестве статистики критерия рассмотрим нормированную случайную величину

$$U = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(1/n_1 + 1/n_2)}}. \tag{5.3.25}$$

При нулевой гипотезе эта статистика распределена приближенно по нормальному закону с нулевым средним значением и единичной дисперсией ($m_u = 0$, $D_u = 1$). Учитывая четность распределения статистики относительно $u = 0$, критические значения статистик $\pm u_{\alpha/2}$ определяются из условия, аналогичного (5.3.2):

$$\mathbf{P}\{U > u_{\alpha/2}\} = \int_{u_{\alpha/2}}^{\infty} f(u \mid H_0)\, du = 1 - \Phi(u_{\alpha/2}) = \frac{\alpha}{2}. \tag{5.3.26}$$

Гипотеза о равенстве двух вероятностей принимается с уровнем значимости α, если для выборочной статистики u^*, вычисленной на основе опытных данных, выполняется двойное неравенство

$$-u_{\alpha/2} < u^* < u_{\alpha/2}. \tag{5.3.27}$$

Более строгое решение задачи проверки гипотезы о равенстве вероятностей событий может быть получено на основе критерия хи-квадрат при проверке гипотез согласия, рассмотренных ниже в п. 5.5.3.

Пример 5.3.5. Проведены две серии опытов объемов $n_1 = 27$ и $n_2 = 30$, в результате которых получено $k_1 = 12$ и $k_2 = 18$ положительных исходов, что соответствует частотам появления событий $p_1^* = 0{,}44$ и $p_2^* = 0{,}60$. Необходимо проверить гипотезу о равенстве вероятностей событий, оценками которых являются опытные значения частот, с уровнем значимости $\alpha = 0{,}05$. Для заданного уровня значимости согласно (5.3.26) по таблицам функций нормального распределения находим критические значения статистики $\pm u_{\alpha/2} = \pm u_{0,025} = \pm 1{,}96$. Значение статистики (5.3.25), вычисленное на основе опытных данных, равно $u^* = -1{,}21$. Поэтому согласно (5.3.27) гипотеза о равенстве вероятностей событий в двух сериях опытов принимается с уровнем значимости 0,05.

5.4. Проверка гипотез о виде распределений (гипотезы согласия)

Установления соответствия выборочных данных проверяемому гипотетическому распределению $F(x)$, называется *гипотезой согласия*. Все гипотезы согласия являются непараметрическими гипотезами. Способы проверки согласия результатов эксперимента с гипотезой о виде распределения базируются на стандартном подходе, описанным выше: выбирают некоторую неотрицательную статистику, служащую мерой отличия

выборочного (эмпирического) распределения $F^*(x)$ от гипотетического $F(x)$. Выбор этого показателя в значительной мере произволен, но необходимо, чтобы его распределение при справедливости проверяемой гипотезы было известно и не зависело от вида гипотетического распределения $F(x)$. В то же время требуется, чтобы математическое ожидание выбранной статистики существенно зависело от отличия выборочного распределения и гипотетического. Критерии проверки гипотез о виде распределений называют *критериями согласия*. Критерии согласия, как правило, являются односторонними (правосторонними) критериями. Ниже рассмотрены наиболее часто используемые критерии согласия выборочных данных гипотетическим распределениям.

Аналогично параметрическим гипотезам, гипотезы согласия разделяют на простые и сложные. *Гипотеза согласия является простой*, если гипотетическая функция распределения $F(x)$ полностью определена. *Гипотеза согласия относится к сложным гипотезам*, если функция распределения зависит от ряда неизвестных параметров. Например, если проверяется гипотеза о нормальной функции распределения с известными средним значением и дисперсией, то такая гипотеза называется простой. В противоположном случае, когда параметры гипотетического нормального распределения неизвестны, рассматриваемая гипотеза согласия является сложной. В сложных гипотезах согласия неизвестные параметры гипотетического распределения заменяют их оценками, полученными либо заранее, либо на основе той же самой выборки, которая используется и для поверки гипотезы согласия. Вид гипотетической функции распределения выбирается или на основе данных о механизме образования случайной величины, или путем соответствующей аппроксимации гистограммы распределения, в том числе и с помощью кривых Пирсона, учитывающих четыре первых выборочных момента. Однако следует особо отметить, что выборочная оценка функционального вида гипотетической функции распределения должна быть осуществлена по выборке, отличной от той, которая будет использоваться при проверке гипотезы согласия.

5.4.1. Критерий согласия хи-квадрат (критерий Пирсона)

В качестве меры отличия выборочной функции распределения $F^*(x)$ от гипотетической функции распределения $F(x)$ часто используется статистика, которая для проверяемой нулевой гипотезы приближенно подчиняется распределению хи-квадрат (*статистика хи-квадрат* (χ^2)). Правило проверки нулевой гипотезы на основе этой статистики называется *критерием хи-квадрат* (χ^2) или *критерием Пирсона*. Рассмотрим применение этого критерия для непрерывных случайных величин с гипотетической плотностью вероятности $f(x)$.

Итак, пусть в результате эксперимента получено n независимых наблюдений $\mathbf{x}(x_1, x_2, \ldots, x_n)$ случайной величины X. Аналогично п. 3.3

сгруппируем n наблюдений по k интервалам, называемым интервалами группировки, которые в совокупности образуют гистограмму частот, примерный вид которой показан на рис. 3.3.1. В j-й интервал $(x_{j-1}; x_j)$ попадает n_j^* наблюдений, $j = 1, 2, \ldots, k$. Для гипотетического распределения в этом же интервале должно быть n_j наблюдений:

$$n_j = np_j, \tag{5.4.1}$$

где

$$p_j = \int_{x_{j-1}}^{x_j} f(x)\,dx = F(x_j) - F(x_{j-1}) \tag{5.4.2}$$

— вероятность попадания случайной величины X в j-й интервал. Разность между экспериментальным и ожидаемым (теоретическим) числом наблюдений в j-м интервале равна $n_j^* - n_j$. В качестве статистики проверки гипотезы рассмотрим нормированный суммарный квадрат этой разности

$$\mathbf{X}_\nu^2 = \sum_{j=1}^{k} \frac{(n_j^* - np_j)^2}{np_j}. \tag{5.4.3}$$

Эта статистика — случайная величина с числом степеней свободы $\nu = k - 1$. Уменьшение числа степеней свободы на единицу по сравнению с числом интервалов связано с тем, что число наблюдений в последнем интервале группировки полностью определяется числом наблюдений всех остальных интервалов.

Конкретное значение $\mathbf{X}_\nu^{2*}$ для фиксированной выборки является выборочной статистикой. Если гипотеза о согласии выборочных данных гипотетической функции распределения справедлива, то согласно п. 3.3 среднее значение и дисперсия числа n_j^* определяются формулами np_j и $np_j(1 - p_j)$. С учетом этого среднее значение статистики $\mathbf{X}_\nu^2$

$$\mathrm{M}\{\mathbf{X}_\nu^2\} = \sum_{j=1}^{k}(1 - p_j) = k - 1. \tag{5.4.4}$$

Аналогично можно вычислить дисперсию случайной величины $\mathbf{X}_n^2$ [37]:

$$D(\mathbf{X}_\nu^2) = 2(k-1) + \frac{1}{n}\left[\sum_{j=1}^{n}\frac{1}{p_j} - k^2 - 2k + 2\right]. \tag{5.4.5}$$

Выражение (5.4.3) для статистики можно представить также в виде, удобном для вычисления

$$\mathbf{X}_\nu^2 = \frac{1}{n}\sum_{j=1}^{k}\frac{(n_j^*)^2}{p_j} - n. \tag{5.4.6}$$

Однако выражение (5.4.3) более наглядно, так как в нем присутствуют разности между наблюденными числами событий (случайных величин) n_j^* и их гипотетическими значениями $n_j = np_j$.

Применительно к нулевой гипотезе К. Пирсоном была доказана теорема, согласно которой функция распределения нормированной суммы квадратов отклонений числа наблюдений n_j^* от своих средних значений np_j при $n \to \infty$ стремится к функции распределения хи-квадрат (2.2.4) с $\nu = k-1$ степенями свободы. Распределение статистики (5.4.3) асимптотически свободно от гипотетического распределения и значений его параметров.

Критерию согласия хи-квадрат посвящено много исследований. В частности, определены погрешности, которые появляются при замене точного распределения случайной величины $\mathbf{X}_\nu^2$ на распределение хи-квадрат (χ_ν^2) для конечного объема выборки n. Оказывается, что вполне удовлетворительное приближение будет, если $np_j \geqslant 5$. Ошибка приближения особенно мала и не зависит от объема выборки, если значения np_j равны или близки между собой. Число событий np_j, ожидаемых в интервалах, может снижаться до трех и двух [13], если число интервалов достаточно велико (например, $k \geqslant 10$). Применение хи-квадрат распределения вместо точного распределения случайной величины $\mathbf{X}_\nu^2$ увеличивает надежность критерия, т.е. полученное критическое значение $\chi_{\alpha;\nu}^2$ на основе асимптотического хи-квадрат распределения оказывается несколько меньше точного критического значения. При этом особое внимание в отношении выполнения условия $np_j \geqslant 5$ необходимо уделять при одной степени свободы. При большом числе степеней свободы $\nu = k-1 \geqslant 30$ для статистики (5.4.3) можно использовать нормальное приближение со средним значением $\mathbf{M}\{\mathbf{X}_\nu^2\} = k$ и дисперсией $D(\mathbf{X}_\nu^2) = 2k$.

Необходимо заметить, что объединение наблюденных данных в группы связано с некоторой потерей информации о функции распределения, имевшейся в первоначальных данных. Кроме того, критерий хи-квадрат не чувствителен к знаку отклонения выборочной функции распределения $F_n^*(x)$ от гипотетической $F(x)$, что может привести к принятию неверного решения при проверке согласия по этому критерию. Поэтому желательно анализировать число знаков получившихся отклонений $(n_j^* - np)$ выборочного распределения от гипотетического. При возникновении сомнений о справедливости принятого решения целесообразно осуществить проверку согласия с помощью других критериев, описанных ниже.

Если r параметров $\boldsymbol{\lambda}(\lambda_1, \lambda_2, \ldots, \lambda_r)$, $r < k-1$, гипотетического распределения $F(x)$ неизвестны (H_0 — сложная гипотеза), то теоретические вероятности p_j не поддаются точному определению и вместо них в статистике (5.4.3) следует взять их точечные оценки $p_j^* \equiv p_j(\boldsymbol{\lambda}^*)$, зависящие от оценок неизвестных параметров $\boldsymbol{\lambda}^*(\lambda_1^*, \lambda_2^*, \ldots, \lambda_r^*)$, т.е. рассматривать статистику вида

$$\mathbf{X}_\nu^2 = \sum_{j=1}^{k} \frac{(n_j^* - np_j^*)^2}{np_j^*}. \tag{5.4.7}$$

Если на основе одних и тех же опытных данных одновременно с проверкой гипотезы подбираются параметры гипотетического распределения с целью достижения наилучшего согласования с гистограммой частот, число степеней свободы асимптотического распределения случайной величины (5.4.7) уменьшается на число неизвестных оцениваемых параметров гипотетического распределения. Иначе говоря, на теоретическую численность разрядов $n_j = np_j(\boldsymbol{\lambda}^*)$ будет наложено дополнительно r линейных условий типа (3.4.33) и (3.4.34), связанных с оценкой неизвестных параметров гипотетического распределения. В результате этого число степеней свободы должно быть снижено до $\nu = k-1-r$.

В прикладных задачах математической статистики число неизвестных параметров, как правило, не превышает двух, максимум трех. Например, если в качестве гипотетического распределения берется нормальное распределение с неизвестными средним значением и дисперсией, то за счет оценки этих неизвестных параметров число степеней свободы хи-квадрат распределения будет равно $\nu = k-3$. Принципиально оценки этих неизвестных параметров должны быть определены по сгруппированной выборке согласно (3.4.33) и (3.4.34). Однако в первом приближении (без поправок Шеппарда) они совпадают со значениями соответствующих выборочных моментов.

Для принятия нулевой гипотезы на основе распределения хи-квадрат по заданным значениям уровня значимости α и числа степеней свободы ν находится критическое значение статистики $\chi_{\alpha;\nu}^2$:

$$\mathbf{P}\{\chi^2 > \chi_{\alpha;\nu}^2\} = \int_{\chi_{\alpha;\nu}^2}^{\infty} f(\chi^2)\, d\chi^2 = \alpha, \tag{5.4.8}$$

с которым сравнивается выборочная статистика χ_ν^{2*}, вычисленная на основе опытных данных и гипотетического распределения согласно (5.4.3) или (5.4.7). Гипотеза о согласии эмпирической и гипотетической функций распределения принимается, если для выборочной статистики выполняется неравенство

$$\chi_\nu^{2*} = \sum_{j=1}^{k} \frac{(n_j^* - np_j^*)^2}{np_j^*} \leqslant \chi_{\alpha;\nu}^2. \tag{5.4.9}$$

При использовании критерия хи-квадрат возникает задача выбора интервалов группирования данных. Применительно к дискретной случайной величине задача определения интервалов возникает только в части объединения некоторых теоретических частот в дискретных точках. Поэтому задача определения границ интервалов актуальна лишь для

выборки, полученной из непрерывного распределения. В данном случае определение границ интервалов и число интервалов базируется на основе выборочных данных анализируемой случайной величины. При этом, как правило, из удобства расчетов интервалы берутся равными по ширине, за исключением крайних интервалов, если известно, что область существования случайной величины может быть бесконечной. При большом объеме выборки ($n \geqslant 100$) границы равных по ширине интервалов могут быть выбраны достаточно произвольно, исходя из условий, указанных выше по количеству выборочных значений в интервалах. При относительно малом объеме выборки (порядка нескольких десятков) положение границ центрального интервала и его ширина определяются соответственно средним значением и дисперсией распределения. Например, применительно к нормальному распределению границы интервалов могут быть выбраны из соотношения $x_j = \bar{x} \pm csj$, $j = 1, 2, \ldots, k$, где c — некоторое положительное число, определяющее полуширину интервала в долях от выборочного среднего квадратического отклонения. Обычно $c \leqslant 0{,}25$, т.е. ширина интервала группирования выборок применительно к нормальному распределению не должна быть больше половины выборочного среднего квадратического отклонения. Для упрощения расчетов вместо выборочных значений $\mathbf{x}(x_1, x_2, \ldots, x_n)$ удобнее оперировать с относительными величинами, для которых имеются соответствующие таблицы распределения вероятностей. Для нормально распределенной выборки целесообразно использовать относительные величины

$$z_i = (x_i - \bar{x})/s, \tag{5.4.10}$$

а в качестве одной из верхних границ взять $z_j = 0$, что соответствует верхней границе исходной случайной величины $x_j = z_j + \bar{x}$.

Пример 5.4.1. В табл. 5.4.1 представлена выборка из $n = 40$ независимых наблюдений.

Таблица 5.4.1

6,42	6,53	6,56	6,70	6,75	6,80	6,82	6,85	6,90	6,93
6,95	7,00	7,10	7,13	7,15	7,18	7,25	7,28	7,30	7,35
7,40	7,42	7,45	7,50	7,52	7,55	7,57	7,60	7,62	7,65
7,70	7,72	7,75	7,78	7,80	7,82	7,90	8,02	8,12	8,30

Для удобства выборочные значения упорядочены по возрастанию (образован вариационный ряд). С помощью рассмотренного критерия согласия хи-квадрат и уровня значимости $\alpha = 0{,}05$ проверим эту выборку на то, что она взята из генеральной совокупности с нормальным законом распределения. Для проверки гипотезы о нормальном распределении необходимо вычислить границы интервалов, для чего сначала определяются выборочное среднее и выборочное среднее квадратическое отклонение:

$$\bar{x} = \frac{1}{40}\sum_{i=1}^{40} x_i = 7{,}33; \quad s = \sqrt{\frac{1}{39}\sum_{i=1}^{40}(x_i - \bar{x})^2} = 0{,}46.$$

В табл. 5.4.2 сведены вычисления, необходимые для построения критерия проверки гипотезы. При этом длина (ширина) интервалов выбрана равной $\Delta x = 0{,}5s$ ($\Delta z = 0{,}5$), а отсчет их границ осуществлен от значения верхней границы центрального интервала $x_j = \bar{x}$ ($z_j = 0$). Первый и последний интервалы выбраны так, чтобы $n_j^* \geqslant 3$. В результате получилось шесть интервалов ($k = 6$). Вероятность попадания гипотетического нормального распределения в каждый из интервалов группировки

$$p_j = \mathbf{P}\{z_{j-1} < Z < z_j\} = F(z_j) - F(z_{j-1}), \quad j = 1, 2, \ldots, 6,$$

определяется по таблицам нормального распределения. Просуммировав значения последнего столбца, получим

$$\chi^{2*}_6 = \sum_{j=1}^{k=6} \frac{(n_j^* - n_j)^2}{n_j} = 1{,}746.$$

Таблица 5.4.2

Номер интервала j	Верхняя граница интервала z_j	Верхняя граница интервала x_j	p_i	$n_j = np_j$	n_j^*	$(n_j^* - n_j)^2/n_j$
1	−1,0	6,87	0,1587	6,3	8	0,459
2	−0,5	7,10	0,1498	6,0	5	0,167
3	0,0	7,33	0,1915	7,7	6	0,375
4	0,5	7,56	0,1915	7,7	7	0,064
5	1,0	7,79	0,1498	6,0	8	0,667
6	∞	∞	0,1587	6,3	6	0,014

В рассматриваемом случае число степеней свободы равно $\nu = k - 3 = 3$. Критическое значение критерия находим по таблицам критических значений хи-квадрат распределения (табл. 2 Приложения 2) $\chi^2_{0,05;3} = 7{,}81$. Так как $\chi^{2*}_3 = 1{,}746 < \chi^2_{0,05;3} = 7{,}81$, гипотеза о принадлежности приведенных выборочных данных нормальному распределению принимается с уровнем значимости $\alpha = 0{,}05$.

5.4.2. Критерий согласия Колмогорова

Критерий согласия Колмогорова применяют в задачах проверки гипотезы о виде распределения, когда функция распределения $F(x)$ непрерывна. *Статистикой критерия Колмогорова* является величина

$$D_n = \max|F^*(x) - F(x)|, \quad -\infty < x < \infty, \tag{5.4.11}$$

представляющая собой модуль максимального отклонения выборочной (эмпирической) функции распределения $F^*(x)$ от гипотетической (теоретической) функции распределения $F(x)$. Применительно к фиксированной выборке объема n выборочную статистику обозначим как d_n^*. В тех случаях, когда гипотеза H_0 о совпадении функций $F^*(x)$ и $F(x)$ верна, при большом объеме выборки ($n \gg 1$) происходит сближение этих функций распределения. При этом значение D_n не должно существенно отличаться от нуля. Отсюда следует, что критическую область критерия (область значений статистики, где отвергается гипотеза H_0), следует задавать в виде

$$D_n > d_{\alpha;n}. \tag{5.4.12}$$

Здесь $d_{\alpha;n}$ — некоторое пороговое значение случайной величины D_n, которое определяется исходя из заданного уровня значимости α

$(\mathbf{P}\{D_n > d_{\alpha;n}\} = \alpha)$. При выполнении неравенства вида (5.4.12) для выборочной статистики $(d_n^* > d_{\alpha;n})$ гипотеза о соответствии эмпирической и гипотетической функций распределения отвергается. Наоборот, при $d_n^* < d_{\alpha;n}$ гипотеза H_0 о равенстве $F^*(x)$ и $F(x)$ принимается.

Важным свойством распределения статистики $F(d_n) = \mathbf{P}\{D_n < d_n\}$ при нулевой гипотезе H_0 является то, что это распределение не зависит от функции распределения $F(x)$ случайной величины X [8, 27, 31 и др.]. Данный факт позволяет найти распределение статистики и составить соответствующие таблицы распределения этой статистики и критических значений $d_{\alpha;n}$ для наиболее простого прямоугольного распределения случайной величины X.

В табл. 6 Приложения 2 приведены таблицы критических значений $d_{\alpha;n}$ статистики D_n в зависимости от объема выборки n и уровня значимости α.

А.Н. Колмогоровым было доказано, что при большом объеме выборки $n \gg 1$ (в пределе при $n \to \infty$) независимо от вида закона распределения анализируемой случайной величины X для нулевой гипотезы случайная величина $Z = \sqrt{n}D_n$ имеет асимптотическую функцию распределения вида

$$K(z) = \mathbf{P}\{\sqrt{n}D_n < z\} =$$

$$= \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2z^2) = 1 + 2\sum_{j=1}^{\infty}(-1)^j \exp(-2j^2z^2). \quad (5.4.13)$$

Эту функцию распределения называют *функцией распределения Колмогорова*. Функция распределения Колмогорова $K(z)$ и критические значения этого распределения табулированы. Применительно к асимптотическому распределению (распределению Колмогорова) и заданному уровню значимости α асимптотические критические значения $z_\alpha = \sqrt{n}d_{\alpha;n}$ определяются из соотношения

$$\mathbf{P}\{D_n > d_{\alpha;n}\} = \mathbf{P}\{Z > z_\alpha\} = 1 - K(z_\alpha) = \alpha. \quad (5.4.14)$$

Как и выше, задаваясь уровнем значимости α, по соответствующим таблицам находится критическое значение z_α, на основе которого вычисляется граница критической области $d_{\alpha;n} = z_\alpha/\sqrt{n}$. Затем определяется модуль максимального отклонения d_n^* гипотетической и эмпирической функций распределения (выборочная статистика). Сравнивая выборочную статистику d_n^* с критическим значением $d_{\alpha;n}$, принимается или отвергается гипотеза о равенстве эмпирического и гипотетического распределений.

В табл. 5.4.3 приведены критические значения z_α распределения Колмогорова для ряда значений уровня значимости α.

Анализ показывает [8], что для нулевой гипотезы распределение случайной величины $Z = \sqrt{n}D_n$ при $n \geqslant 25$ слабо зависит от объема выборки n, т.е. практически совпадает с асимптотическим распределением. Этот факт дает основание использовать распределение Колмогорова

Таблица 5.4.3

Критические знания z_α распределения Колмогорова

α	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
z_α	0,828	0,895	0,974	1,073	1,224	1,358	1,520	1,627	1,950

для практических расчетов уже при $n \geqslant 25$ и $\alpha < 0{,}1$. Так, например, анализ точных и асимптотических критических значений для $\alpha = 0{,}05$ и $\alpha = 0{,}01$ показывает, что при $n \geqslant 25$ погрешность в замене точных критических значений на асимптотические не превышает 3 %. Кроме того, замена точных границ асимптотическими изменяет доверительный интервал $d_{\alpha;n}$ в сторону увеличения надежности критерия.

Для $n \gg 1$ определение критического значения z_α значительно упрощается, если воспользоваться двумя первыми членами разложения функции распределения Колмогорова (5.4.13). При этом

$$\alpha = 1 - K(z_\alpha) = -2\sum_{j=1}^{\infty}(-1)^j \exp(-2j^2z_\alpha^2) \approx 2\exp(-2nd_{\alpha;n}^2).$$

Отсюда выражение для границы критической области примет вид

$$d_{\alpha;n} \approx \sqrt{\ln(2/\alpha)/2n}. \quad (5.4.15)$$

Асимптотическая теория критерия согласия Колмогорова, строго говоря, справедлива для точно известного гипотетического распределения. Однако на практике гипотетический закон распределения, как правило, может быть выбран (известен) с точностью до некоторых параметров, которые определяются на основе исходной выборки. В этом случае критерий согласия Колмогорова будет давать приближенные решения. Естественно, приближение тем лучше, чем больше объем выборки. При этом, аналогично отмеченному выше, если вид гипотетического закона распределения неизвестен, то его оценка должна быть осуществлена лишь на основе другой выборки.

Критерий согласия Колмогорова в отличие от других возможных критериев согласия можно использовать для установления доверительных интервалов для непрерывной функции распределения. Действительно, из выражения

$$\mathbf{P}\{D_n = \max|F^*(x) - F(x)| > d_{\alpha;n}\} = \alpha$$

имеем

$$\mathbf{P}\{F^*(x) - d_{\alpha;n} \leqslant F(x) \leqslant F^*(x) + d_{\alpha;n}\} = 1 - \alpha = \gamma. \quad (5.4.16)$$

Таким образом, истинная функция $F(x)$ лежит целиком внутри полосы шириной $\pm d_{\alpha;n}$ около выборочной функции распределения $F^*(x)$. Соотношение (5.4.16) позволяет определить объем выборки, необходимый для аппроксимации функции распределения. Для того чтобы выборочная функция распределения $F^*(x)$ с доверительной вероятностью

$\gamma = 1 - \alpha$ отстояла от истинной не более, чем на величину $d_{\alpha;n}$, объем выборки согласно приближенному выражению (5.4.15) должен определяться как

$$n \geqslant \frac{1}{2d^2_{\alpha;n}} \ln \frac{2}{\alpha}. \qquad (5.4.17)$$

Например, для $\alpha = 0{,}05$ ($\gamma = 0{,}95$) и $d_{\alpha;n} = 0{,}1$ должно быть не менее 184 выборочных значений.

При большом объеме выборки (например, порядка нескольких сотен) вычисление эмпирической и гипотетической функций распределения для каждого выборочного значения x_i, $i = 1, 2, \ldots, n$, весьма громоздко. Поэтому при $n \gg 1$ критерий Колмогорова, несмотря на очевидную наглядность, используется редко. Для данной ситуации предложен ряд приближенных способов уменьшения трудоемкости проверки гипотезы по критерию Колмогорова. Наиболее приемлемым для практических задач является способ, основанный на разбиении исходной выборки по k интервалам, как правило, одинаковой ширины. В каждом из интервалов берется «представительное» значение аргумента x_j^*, $j = 1, 2, \ldots, k$, и для него вычисляются эмпирическое и гипотетическое значения функций распределения. В качестве x_j^* обычно выбирают средние значения аргументов в интервалах: $x_j^* = (x_{j-1} + x_j)/2$. При этом, учитывая непрерывность функции распределения $F(x)$ и ограничиваясь линейным характером изменения $F(x)$ в пределах каждого интервала, значения эмпирической функции распределения в точках x_j^* вычисляются по формуле

$$F^*(x_j^*) = \frac{1}{n}\left[\sum_{i=1}^{j-1} n_i^* + 0{,}5n_j^*\right] = F_j^* + \frac{n_j^*}{2n},\ j = 1, 2, \ldots, k. \qquad (5.4.18)$$

Здесь n_j^* — число выборочных значений, попавших в j-й интервал.

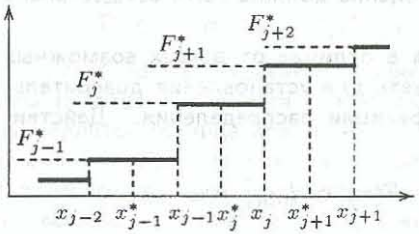


Рис. 5.4.1. Функция распределения

Примерный график эмпирической функции распределения приведен на рис. 5.4.1; он имеет скачки на границах интервалов в точках $x_{j\pm i}$, $i = 1, 2, \ldots$. Так как точка x_j^* является серединой j-го интервала, то значение эмпирической функции распределения в этой точке будет равно значению оценки функции распределения в точке x_{j-1} плюс половина оценки вероятности попадания значений случайной величины в j-й интервал. Принципиально в качестве «представительной» точки можно было бы взять любую точку в пределах интервала, в том числе и одну из границ интервала (верхнюю или нижнюю), с соответствующим учетом для нее поправки на эмпирическую функцию распределения. При группировании выборки по интервалам, естественно, теряется определенная доля информации о различии сравниваемых функций распределения.

Пример 5.4.2. Проверим гипотезу согласия наблюдений, представленных в табл. 5.4.1, нормальному распределению с выборочным средним $\bar{x} = 7{,}33$ и выборочным средним квадратическим отклонением $s = 0{,}46$ для уровня значимости $\alpha = 0{,}05$. Для объема рассматриваемой выборки $n = 40$ и заданного уровня значимости критическое значение статистики согласно (5.4.15) равно $d_{\alpha;n} = d_{0{,}05;40} = 0{,}215$. Для упрощения расчетов исходную выборку сгруппируем по $k = 10$ одинаковым по ширине интервалам, границы которых и средние значения интервалов x_j^* указаны во втором и третьем столбцах табл. 5.4.4. В четвертом столбце приведено число выборочных значений n_j^*, попадающих в соответствующий интервал, причем при попадании выборочного значения на границу интервалов, оно поровну относилось в оба интервала (в данном случае в каждый интервал прибавлялось по 1/2). В последних трех столбцах таблицы даны результаты расчетов значений эмпирической и теоретической функций распределения, а также модуль разности их значений. Значения эмпирической функции распределения вычислялись согласно (5.4.18), а значения теоретической (гипотетической) функции распределения находились по формуле

$$F(x_j^*) = \frac{1}{\sqrt{2\pi}s}\int_{-\infty}^{x_j^*} \exp\left(-\frac{x-\bar{x}}{2s^2}\right) dx = \Phi(2{,}17x_j^* - 15{,}9).$$

Таблица 5.4.4

j	$x_{j-1} - x_j$	x_j^*	n_j^*	$F^*(x_j^*)$	$F(x_j^*)$	$\lvert F^*(x_j^*) - F(x_j^*)\rvert$
1	6,4 – 6,6	6,5	3,0	0,0375	0,0359	0,0016
2	6,6 – 6,8	6,7	2,5	0,1063	0,0853	0,0210
3	6,8 – 7,0	6,9	6,0	0,2125	0,1762	0,0363
4	7,0 – 7,2	7,1	4,5	0,3438	0,3121	0,0317
5	7,2 – 7,4	7,3	4,5	0,4563	0,4761	0,0198
6	7,4 – 7,6	7,5	7,0	0,6000	0,6460	0,0460
7	7,6 – 7,8	7,7	7,0	0,7750	0,7910	0,0160
8	7,8 – 8,0	7,9	2,5	0,8938	0,8925	0,0013
9	8,0 – 8,2	8,1	2,0	0,9500	0,9535	0,0035
10	8,2 – 8,4	8,3	1,0	0,9875	0,9821	0,0054

Из таблицы видно, что максимальное значение модуля отклонения эмпирической функции распределения от гипотетической (выборочная статистика) равно $d_n^* = 0{,}0460$, которое меньше критического значения $d_{\alpha;n} = 0{,}215$. Следовательно, согласно критерию Колмогорова гипотеза о нормальном распределении исходной выборки может быть принята с уровнем значимости $\alpha = 0{,}05$.

5.4.3. Критерий Смирнова–Мизеса (критерий ω^2)

Согласно *критерию Смирнова–Мизеса* (ω^2) в качестве меры расхождения эмпирической функции распределения $F^*(x)$ и гипотетической функции распределения $F(x)$ (статистики) рассматривается средний квадрат отклонения по всем возможным значениям случайной величины, что, естественно, в противоположность критерию Колмогорова, как и в случае критерия хи-квадрат, приводит к некоторой потере информации о различии функций распределения. Будем полагать, что

гипотетическая функция распределения непрерывна. Рассмотрим выборочную статистику в виде

$$\omega^2 = \int_{-\infty}^{\infty} [F^*(x) - F(x)]^2 dF(x). \qquad (5.4.19)$$

Согласно определению эмпирической функции распределения $F^*(x)$ имеем

$$F^*(x) = \begin{cases} 0 & \text{при } x < x_{(1)}, \\ k/n & \text{при } x_{(k)} \leqslant x \leqslant x_{(k+1)}, \quad k = 1, 2, \ldots, n-1. \\ 1 & \text{при } x \geqslant x_{(n)}, \end{cases}$$

Разобьем всю область интегрирования на интервалы $(-\infty, x_{(1)})$, $(x_{(1)}, x_{(2)}), \ldots, (x_{(n)}, \infty)$, тогда выражение (5.4.19) преобразуется к виду

$$\omega^2 = \int_{-\infty}^{x_{(1)}} [0 - F(x)]^2 \, dF(x) + \\ + \sum_{k=1}^{n-1} \int_{x_{(k)}}^{x_{(k+1)}} \left[\frac{k}{n} - F(x)\right]^2 dF(x) + \int_{x_{(n)}}^{\infty} [1 - F(x)]^2 dF(x).$$

Вычисляя элементарные интегралы и объединяя члены, зависящие от $F(x)$, последнее соотношение можно представить в виде [8, 31, 57]

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{k=1}^{n} \left[F(x_{(k)}) - \frac{2k-1}{2n}\right]^2. \qquad (5.4.20)$$

Эмпирическая функция распределения является несмещенной оценкой действительной функции распределения с дисперсией, определяемой формулой (3.2.10). Поэтому для нулевой гипотезы среднее значение статистики (5.4.19)

$$\mathrm{M}\{\omega^2\} = \int_{-\infty}^{\infty} \mathrm{M}\{[F^*(x) - F(x)]^2\} \, dF(x) =$$

$$= \frac{1}{n} \int_0^1 F(x)[1 - F(x)] \, dF(x) = \frac{1}{n} \left[\frac{1}{2}F^2(x) - \frac{1}{3}F^3(x)\right]\Big|_0^1 = \frac{1}{6n}.$$

Аналогично вычисляется и дисперсия случайной величины ω^2 [31, 57]:

$$D(\omega^2) = \mathrm{M}\{\omega^4\} - (\mathrm{M}\{\omega^2\})^2 = \frac{4n-3}{180n^3}.$$

Из последних двух выражений видно, что для нулевой гипотезы среднее значение и дисперсия статистики ω^2 не зависят от гипотетической функции распределения и в пределе (при $n \gg 1$) стремятся к нулю. Вместо статистики ω^2 рассматривают статистику $Z = n\omega^2$, для которой $\mathrm{M}\{Z\} = 1/6$. Точное распределение статистики $Z = n\omega^2$ достаточно сложно, оно не зависит от гипотетической функции распределения. По сравнению с распределением Колмогорова распределение $Z = n\omega^2$ быстрее сходится к предельному значению. Уже при $n \geqslant 40$ точное распределение практически совпадает с предельным распределением, которое табулировано.

Таблица 5.4.5

Критические значения z_α для распределения $Z = n\omega^2$

α	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
z_α	0,147	0,184	0,241	0,347	0,461	0,620	0,736	1,168

Для этого предельного распределения вычислены критические значения статистики $z_\alpha = (n\omega^2)_\alpha$ для заданного уровня значимости α. В табл. 5.4.5 [57] приведен ряд критических значений статистики $z_\alpha = (n\omega^2)_\alpha$. Аналогично критерию Колмогорова при $n \gg 1$ процесс вычисления выборочной статистики для каждого значения $x_{(k)}$, $k = 1, 2, \ldots, n$, согласно (5.4.20) весьма громоздок.

5.5. Гипотезы однородности распределений

Имеются две независимые выборки: $\mathbf{x}_1(x_{11}, x_{12}, \ldots, x_{1n})$ объема n и $\mathbf{x}_2(x_{21}, x_{22}, \ldots, x_{2m})$ объема m. Требуется установить, являются ли они выборками из одного и того же распределения или нет. Пусть выборке $\mathbf{x}_1$ соответствует функция распределения $F_{X_1}(x)$, а выборке $\mathbf{x}_2$ — функция распределения $F_{X_2}(x)$. Необходимо проверить *гипотезу однородности* H_0: $F_{X_1}(x) = F_{X_2}(x)$.

Применительно к нормальному распределению в пп. 5.3.1 и 5.3.2 рассмотрены критерии Стьюдента и дисперсионного отношения, позволяющие по выборочным средним и выборочным дисперсиям проверить гипотезу о принадлежности двух исходных выборок одной и той же нормальной генеральной совокупности. Однако в прикладных задачах часто функции распределения сравниваемых выборок неизвестны, и потому интересны критерии проверки гипотезы однородности выборок, свободные от априорного знания функции распределения $F(x)$. Ниже рассмотрены три таких критерия.

5.5.1. Критерий однородности Смирнова

В соответствии с *критерием проверки однородности Смирнова* образуется статистика

$$D_{n,m} = \max |F^*_{X_1}(x) - F^*_{X_2}(x)|, \qquad (5.5.1)$$

где $F^*_{X_1}(x)$ и $F^*_{X_2}(x)$ — эмпирические (выборочные) функции распределения, построенные по выборкам $\mathbf{x}_1$ и $\mathbf{x}_2$ соответственно. Так же, как и критерий согласия Колмогорова и Смирнова–Мизеса, критерий проверки однородности Смирнова применяется в случае непрерывных

распределений. Как отмечалось выше, эмпирическая функция распределения с увеличением объема выборки приближается к теоретической. Если нулевая гипотеза H_0 верна, то с увеличением объема выборок n и m функции $F^*_{X_1}(x)$ и $F^*_{X_2}(x)$ оценивают одну и ту же неизвестную функцию распределения. В данном случае выборочное значение статистики $D^*_{n;m}$ не должно существенно отличаться от нуля. Поэтому, если выборочная статистика достаточно велика, то это свидетельствует против нулевой гипотезы. Аналогично критерию Колмогорова распределение статистики $F(d_{n,m}) = \mathbf{P}\{D_{n,m} < d_{n,m}\}$ при нулевой гипотезе H_0 не зависит от функций распределения $F(x)$ случайных величин X_1 и X_2. Для прикладных задач этот факт имеет важное значение, так как истинное распределение наблюдаемых случайных величин, как правило, бывает неизвестно, но приходится выяснять, изменяется или нет неизвестное распределение от выборки к выборке. Для применения критерия Смирнова необходимо выполнение только условия непрерывности, которое обычно вытекает из природы изучаемых явлений и не требует специальной проверки.

Н.В. Смирновым для нулевой гипотезы и произвольных значений n и m было получено распределение статистики $D_{n,m}$, которое достаточно сложно для непосредственных расчетов. В [8, 51, 72] приведены таблицы критических значений $d_{\alpha;n,m}$ статистики Смирнова $D_{n,m}$ такие, что $\mathbf{P}\{D_{n,m} \geqslant d_{\alpha;n,m}\} = \alpha$ для различных значений α, n и m. Н.В. Смирновым было доказано, что при достаточно больших объемах выборок n и m (в пределе $n \to \infty$, $m \to \infty$) величина

$$Z = \sqrt{nm/(n+m)}D_{n,m} \tag{5.5.2}$$

имеет распределение Колмогорова $K(z)$ в виде (5.4.13). Таким образом, в предельном (асимптотическом) случае распределение случайной величины (нормированной статистики Смирнова) Z не зависит от объема выборок.

Критические значения предельного распределения случайной величины (5.5.2) приведены в табл. 5.4.3. На основе полученного значения z_α находится критическое значение выборочной статистики

$$d_{\alpha;n,m} = z_\alpha\sqrt{(n+m)/nm}. \tag{5.5.3}$$

Следовательно, как и выше, задаваясь уровнем значимости α, по таблицам определяется точное или приближенное критическое значение выборочной статистики $d_{\alpha;n,m}$, которое сравнивается с выборочной статистикой $d^*_{n,m}$. Если выборочная статистика не превышает критического значения, то можно утверждать, что гипотеза об однородности исходных выборок может быть принята с заданным уровнем значимости.

5.5.2. Критерий однородности хи-квадрат

Наиболее часто применяемым критерием проверки гипотезы однородности является *критерий хи-квадрат*, который используется как для

дискретных, так и непрерывных распределений исходных выборок. Кроме того, критерий хи-квадрат позволяет проверять гипотезу однородности одновременно для любого конечного числа выборок, в то время как с помощью критерия Смирнова можно сравнивать между собой только две выборки. Сначала рассмотрим применение критерия хи-квадрат для проверки гипотезы однородности двух выборок.

Итак, пусть в результате наблюдений получены две выборки: $\boldsymbol{x}_1(x_{11}, x_{12}, \ldots, x_{1n_1})$ объема n_1 и $\boldsymbol{x}_2(x_{21}, x_{22}, \ldots, x_{2n_2})$ объема n_2. Аналогично рассмотренному выше применению критерия хи-квадрат для проверки согласия эмпирической и гипотетической функций распределения сгруппируем исходные выборки по k интервалам. Количество наблюдений, попавших в j-й интервал, $j = 1, 2, \ldots, k$, для выборок $\boldsymbol{x}_1$ и $\boldsymbol{x}_2$ обозначим соответственно как n^*_{1j} и n^*_{2j}. Если справедлива нулевая гипотеза, предполагающая, что обе выборки принадлежат одному и тому же известному распределению $F(x)$, то количество наблюдений, которые теоретически могли попасть в j-й интервал, для обеих выборок равны $n_{1j} = n_1 p_j$ и $n_{2j} = n_2 p_j$, где $p_j = F(x_j) - F(x_{j-1})$ — вероятность попадания случайной величины в j-й интервал. Тогда в качестве статистик можно взять случайные величины

$$\mathbf{X}^2_1 = \sum_{j=1}^{k} \frac{(n^*_{1j} - n_1 p_j)^2}{n_1 p_j}; \quad \mathbf{X}^2_2 = \sum_{j=1}^{k} \frac{(n^*_{2j} - n_2 p_j)^2}{n_2 p_j},$$

которые распределены по законам хи-квадрат с $\nu = k - 1$ степенями свободы. При этом сумма этих статистик

$$\mathbf{X}^2_\nu = \sum_{j=1}^{k} \left[\frac{(n^*_{1j} - n_1 p_j)^2}{n_1 p_j} + \frac{(n^*_{2j} - n_2 p_j)^2}{n_2 p_j} \right] \tag{5.5.4}$$

также распределена по закону хи-квадрат с $\nu = 2(k-1)$ степенями свободы.

Поскольку вероятности p_j обычно неизвестны, воспользуемся их оценками, которые при справедливости нулевой гипотезы равны

$$p^*_j = \frac{n^*_{1j} + n^*_{2j}}{n_1 + n_2}. \tag{5.5.5}$$

Тогда статистика (5.5.4) заменяется величиной

$$\mathbf{X}^2_\nu = \sum_{j=1}^{k} \left[\frac{(n^*_{1j} - n_1 p^*_j)^2}{n_1 p^*_j} + \frac{(n^*_{2j} - n_2 p^*_j)^2}{n_2 p^*_j} \right] =$$

$$= n_1 n_2 \sum_{j=1}^{k} \frac{(n^*_{1j}/n_1 - n^*_{2j}/n_2)^2}{n^*_{1j} + n^*_{2j}}. \tag{5.5.6}$$

Случайная величина $\mathbf{X}^2_\nu$ приближенно также имеет распределение

хи-квадрат, но уже с меньшим числом степеней свободы, так как из-за условия $\sum p_j = 1$ для оценок вероятностей p_j^* использовано $k-1$ линейное соотношение. В результате распределение (5.5.6) будет иметь $\nu = 2(k-1) - (k-1) = k-1$ степеней свободы.

Как и выше, задаваясь уровнем значимости α, по таблицам находится критическое значение выборочной статистики $\chi^2_{\alpha;\nu}$, соответствующее числу степеней свободы ν. Вычисляя выборочную статистику $\mathbf{X}^{2*}_{\nu}$ согласно (5.5.6) и сравнивая ее с критическим значением, проверяется нулевая гипотеза о принадлежности выборок $\mathbf{x}_1$ и $\mathbf{x}_2$ одному и тому же распределению. В частности, если $\mathbf{X}^{2*}_{\nu}$ меньше $\chi^2_{\alpha;\nu}$, то исходные выборки не противоречат нулевой гипотезе с заданным уровнем значимости α.

При проверке гипотезы о равенстве вероятностей событий, рассмотренной приближенно в п. 5.3.4, задача решается более строго с применением критерия хи-квадрат. В этом случае в формуле (5.5.6) необходимо положить $k = 2$, причем один из интервалов соответствует числу положительных исходов, а другой интервал — числу отрицательных исходов. При этом для данного частного случая число степеней свободы $\nu = 2 - 1 = 1$.

Пример 5.5.1. Рассмотрим исходные данные, приведенные в примере 5.3.4 по проверке гипотезы о равенстве вероятностей событий двух выборок с уровнем значимости $\alpha = 0{,}05$. Объемы выборок $n_1 = 27$ и $n_2 = 30$. Число положительных исходов, соответствующих первому интервалу, $n^*_{11} = 12$ и $n^*_{21} = 18$. Во втором интервале число выборочных значений $n^*_{12} = n_1 - n^*_{11} = 15$ и $n^*_{22} = n_2 - n_{21} = 12$. При этом $p_1^* = (n^*_{11} + n^*_{21})/(n_1 + n_2) \approx 0{,}53$; $p_2^* = (n^*_{12} + n^*_{22})/(n_1 + n_2) \approx 0{,}47$. Подставляя эти значения в (5.5.6), получим $\mathbf{X}^{2*}_{\nu=1} \approx 1{,}38$. Критическое значение статистики для $\alpha = 0{,}05$ и $\nu = 1$ равно $\chi^2_{0{,}05;1} = 3{,}84$. Нулевая гипотеза о равенстве вероятностей двух выборок не отклоняется, так как $\mathbf{X}^{2*}_{\nu=1} \approx 1{,}38 < \chi^2_{0{,}05;1} = 3{,}84$. Подобный вывод сделан и в примере 5.3.4.

Полученный результат для проверки однородности двух выборок легко обобщается на проверку гипотезы однородности нескольких выборок $\mathbf{x}_r$, $r = 1, 2, \ldots, m$, каждая объемом n_r. В этом случае статистика с числом степеней свободы $\nu = (k-1)(m-1)$ определяется выражением

$$\mathbf{X}^2_{\nu} = \sum_{r=1}^{m}\sum_{j=1}^{k} \frac{(n^*_{rj} - n_r p_j^*)^2}{n_r p_j^*} =$$

$$= n \sum_{r=1}^{m}\sum_{j=1}^{k} \frac{(n^*_{rj} - n_r n_j^*/n)}{n_r n_j^*} = n\left[\sum_{r=1}^{m}\sum_{j=1}^{k} \frac{(n^*_{rj})^2}{n_r n_j^*} - 1\right]. \qquad (5.5.7)$$

Здесь оценка вероятности

$$p_j^* = \frac{1}{n}\sum_{r=1}^{m} n^*_{rj} = \frac{n^*_{\cdot j}}{n}. \qquad (5.5.8)$$

Наличие точки у величины $n^*_{\cdot j}$ перед нижним индексом j указывает на то, что эта величина усреднена по переменной r.

Если в (5.5.7) положить $m = 2$, то получим выражение (5.5.6).

Пример 5.5.2. В табл. 5.5.1 приведены данные двух выборок одинакового объема $n_1 = n_2 = 50$, сгруппированные по $k = 6$ интервалам (разрядам). Проверим гипотезу об однородности этих выборок с уровнем значимости $\alpha = 0{,}05$.

Таблица 5.5.1

Номер интервала j	Численность разряда		$n^*_{1j} + n^*_{2j}$	$n^*_{1j} - n^*_{2j}$
	n^*_{1j}	n^*_{2j}		
1	2	1	3	1
2	5	7	12	−2
3	12	17	29	−5
4	15	13	28	2
5	12	9	21	3
6	4	3	7	1

Учитывая, что $n_1 = n_2 \equiv n$, формула (5.5.6) упрощается и принимает вид

$$\mathbf{X}^{2*}_{\nu} = \sum_{j=1}^{6} \frac{(n^*_{1j} - n^*_{2j})^2}{n^*_{1j} + n^*_{2j}}.$$

Число степеней свободы этой статистики $\nu = k - 1 = 5$. Подставляя в приведенную формулу исходные данные из таблицы, имеем $\mathbf{X}^2_{\nu=5} \approx 2{,}24$. Согласно таблицам хи-квадрат распределения критическое значение статистики для $\alpha = 0{,}05$ и $\nu = 5$ равно $\chi^2_{0{,}05;5} = 11{,}07$. Так как $\mathbf{X}^{2*}_{\nu=5} \approx 2{,}24 < \chi^2_{0{,}05;5} = 11{,}07$, можно утверждать, что с уровнем значимости $\alpha = 0{,}05$ полученные выборки однородны, т.е. принадлежат одной генеральной соокупности.

5.5.3. Критерий Вилкоксона

Критерий однородности Вилкоксона основан на порядковых статистиках, определение которых дано в п. 3.1. Применительно к задаче проверки гипотез в порядковых статистиках используются не сами значения наблюдаемых данных, а лишь их упорядоченность. Естественно, для нулевой гипотезы такие статистики и критерии проверки соответствующих гипотез не зависят от функций распределения изучаемых случайных величин и случайных событий. Порядковые статистики используют различные правила, основанные на сравнении выборочных данных между собой (больше, меньше). Рассмотренные выше критерии однородности Смирнова и хи-квадрат по своей сущности являются также порядковыми критериями, так как в них сравнивается число выборочных значений или число членов вариационного ряда выборок $\mathbf{x}_1(x_{11}, x_{12}, \ldots, x_{1n})$ и $\mathbf{x}_2(x_{21}, x_{22}, \ldots, x_{2m})$, попадающих в соответствующие интервалы.

Рассмотрим применение порядковых статистик для проверки гипотезы однородности двух выборок $\mathbf{x}(x_1, x_2, \ldots, x_n)$ и $\mathbf{y}(y_1, y_2, \ldots, y_m)$ с помощью критерия Вилкоксона. Для проверки нулевой гипотезы (выборки принадлежат одной совокупности) по этому критерию из двух исходных выборок составляется единый вариационный ряд, т.е. x_i, $i = 1, 2, \ldots, n$, и y_j, $j = 1, 2, \ldots, m$, располагаются в порядке возрастания их значений. Отбросив индексы, получим последовательность типа

$$\begin{array}{cccccccccccc} y & x & x & y & y & x & y & \ldots & y & x & x; \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & \ldots & N-2 & N-1 & N. \end{array}$$

Здесь $N = m + n$ — суммарный объем выборок. Снизу указаны порядковые номера (ранги) r_j, $j = 1, 2, \ldots, N$, общего вариационного ряда. В качестве элементов статистики Вилкоксона используется число *инверсий* одной из переменной, например x. Если какому-либо значению x предшествует некоторое y, то говорят, что эта пара дает одну инверсию. Так, например, в приведенной выше последовательности значения x, имеющие порядковые номера (ранги) 2 и 3, дают по одной инверсии с y, а значение x с порядковым номером 6 дает три инверсии с y. При нулевой гипотезе место расположения x_i и y_j в общей последовательности случайно. Таких последовательностей может быть $(n + m)!$. Если же выборки не однородны, то это условие нарушается.

Число инверсий для каждого порядкового номера общего вариационного ряда обозначим W_i, оно является случайной величиной. В *критерии Вилкоксона статистикой служит общее число инверсий* $W_{n,m}$ в последовательности. Величина $W_{n,m}$ также является случайной, конкретное ее значение обозначим как $w_{n,m}$. При этом нулевая гипотеза при двустороннем критерии отвергается тогда, когда выборочная статистика (выборочное число инверсий)

$$w^*_{n,m} = \sum_{i=1}^{n} w_i \tag{5.5.9}$$

выходит за критические границы распределения инверсий. Минимальное число инверсий переменной x соответствует ситуации, когда все x расположены левее y; суммарное число инверсий при этом равно нулю ($\min w^*_{n,m} = 0$). В противоположном случае, когда все x расположены правее y, число инверсий будет максимальным ($\max w^*_{n,m} = nm$). В литературе по математической статистике для нулевой гипотезы получены распределения вероятностей числа инверсий переменных в зависимости от объема выборок n и m. Эти распределения симметричны относительно среднего значения [8, 13, 27]

$$m_W = nm/2. \tag{5.5.10}$$

Дисперсия распределения определяется формулой

$$D_W = nm(n + m + 1)/12. \tag{5.5.11}$$

В [8] для разных значений n, m и уровня значимости α приведены таблицы нижних критических значений $w_{(1-\alpha/2);n,m}$, определяемых из выражения $\mathbf{P}\{W_{n,m} < w_{(1-\alpha/2);n,m}\} = \alpha/2$. Верхние критические значения определяются как

$$w_{\alpha/2;n,m} = nm - w_{(1-\alpha/2);n,m}. \tag{5.5.12}$$

Если $n, m \geqslant 4$, а $n + m \geqslant 20$, то число инверсий приближенно распределено по нормальному закону со средним значением и дисперсией, определяемыми формулами (5.5.10) и (5.5.11). При этом для заданного двустороннего уровня значимости α критическое значение $w'_{\alpha/2;n,m}$ находится из выражения (с учетом поправки на дискретность)

$$\mathbf{P}\left\{\left|W_{n,m} - \frac{nm}{2}\right| > w'_{\alpha/2;n,m}\right\} = 1 - \Phi\left(\frac{w'_{\alpha/2;n,m} - m_W + 0,5}{\sqrt{D_W}}\right) = \frac{\alpha}{2}. \tag{5.5.13}$$

Так как число инверсий должно быть целым числом, то в качестве критического значения $w'_{\alpha/2;n,m}$ берется наименьшее целое число $w'_{\alpha/2;n,m} = [w'_{\alpha/2;n,m}]$.

Итак, гипотеза об однородности двух выборок не отвергается, если выборочная статистика числа инверсий $w^*_{n,m}$ с уровнем значимости α лежит в интервале

$$w_{(1-\alpha/2);n,m} \leqslant w^*_{n,m} \leqslant w_{\alpha/2;n,m}. \tag{5.5.14}$$

Суммарное число инверсий переменной x относительно переменной y можно выразить через сумму порядковых номеров x в общем вариационном ряде. Действительно, если наименьший из всех x_i имеет ранг r_1, то число предшествующих y_j равно $r_1 - 1$ и, следовательно, значение x_i имеет $r_1 - 1$ инверсий. Рассуждая аналогично, получим, что сумма инверсий переменной x_i, $i = 1, 2, \ldots, n$, во всем вариационном ряде

$$w_{n,m} = (r_1 - 1) + (r_2 - 2) + \ldots + (r_n - n) = \sum_{i=1}^{n} r_i - \frac{n(n+1)}{2}.$$

Так как второй член этого выражения не зависит от конкретного опыта, в качестве статистики критерия Вилкоксона можно рассматривать величину

$$T_{n,m} = \sum_{i=1}^{n} r_i, \tag{5.5.15}$$

выборочное значение которой обозначим как $t^*_{n,m}$. Критерий, основанный на статистике (5.5.15), называют *ранговым критерием* проверки гипотезы однородности выборок. Поскольку статистики (5.5.9) и (5.5.15) связаны между собой линейной зависимостью, то линейной зависимостью связаны между собой статистические характеристики случайных величин $W_{n,m}$ и $T_{n,m}$, а также их критические значения для соответствующих выборочных статистик.

5.6. Порядковые критерии проверки независимости и случайности выборок

Гипотезы, основанные на порядковых критериях, широко используются для проверки гипотез независимости (взаимной независимости)

двух и более выборок с одинаковыми распределениями и случайности выборочных значений. В последнем случае под случайностью выборки понимается то, что выборочные значения между собой независимы. Применительно к двум выборкам одинакового объема n нулевая *гипотеза независимости* выборок соответствует соотношению $F_2(x_i, y_i) = F_X(x_i)F_Y(y_i)$, $i = 1, 2, \ldots, n$. Нулевая *гипотеза случайности* выборки отвечает выражению $F_n(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\cdots F(x_n)$. Необходимость проверки гипотезы независимости выборочных значений часто возникает в задачах, в которых элементами выборки являются измеренные значения какого-либо признака при изменяющихся условиях эксперимента, и требуется проверить гипотезу о влиянии на результаты измерений этих изменений.

Порядковые критерии (статистики) проверки гипотез независимости и случайности основаны на том, что при справедливости нулевых гипотез все выборочные значения x_i и y_i равноправны и не должны быть каким-либо способом упорядочены. При наличии статистической зависимости между случайными величинами X и Y между выборочными значениями также будет проявляться определенная «степень» упорядоченности. Поэтому критерии проверки гипотез о независимости и случайности выборок могут быть построены на основе статистик, имеющих некоторый показатель упорядоченности исходных данных. В математической статистике предложен ряд критериев и соответствующих им порядковых статистик для проверки гипотез независимости и случайности выборок. Ниже приведено описание трех из них, а именно критериев знаков, серий и инверсий.

5.6.1. Критерий знаков

Критерий знаков позволяет проверить нулевую гипотезу о том, что две исходные выборки взяты из двух взаимно независимых генеральных совокупностей с одинаковыми функциями распределения вероятностей. Итак, пусть в результате наблюдения получены две выборки $\mathbf{x}(x_1, x_2, \ldots, x_n)$ и $\mathbf{y}(y_1, y_2, \ldots, y_n)$ одинакового объема n. При этом будем полагать, что последовательность n пар наблюдений $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$, независима. На основе n пар измерений можно образовать n разностей $z_i = x_i - y_i$, из которых k_n разностей положительны и $(n - k_n)$ разностей отрицательны. Возможность равенства $x_i = y_i$ можно исключить (если $x_i = y_i$, то из исходного объема n можно вычесть число выборочных значений, соответствующее $z_i = 0$). В результате получается последовательность положительных «+» ($z_i > 0$) и отрицательных «−» ($z_i < 0$) данных.

В качестве иллюстрации в табл. 5.6.1 приведена последовательность таких измерений для $n = 17$. Здесь $k_n = 10$ — число положительных знаков (разностей), а $(n - k_n) = 7$ — число отрицательных знаков (разностей). Вопрос заключается в том, можно ли считать полученные в

Таблица 5.6.1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
x_i	5,5	5,1	5,7	5,2	6,8	4,8	6,5	5,4	5,8	4,9	6,6	5,0	5,3	5,9	6,1	6,2	6,9
y_i	5,1	4,6	6,0	4,7	5,7	6,2	5,5	5,0	4,8	5,2	5,9	6,7	6,1	4,9	6,9	6,6	4,5
z_i	+	+	−	+	+	−	+	+	+	−	+	−	−	+	−	−	+

опытах различия в количестве положительных и отрицательных знаков значимыми или эти различия следует отнести за счет естественного случайного рассеяния исходных данных.

Поэтому необходимо проверить нулевую гипотезу, согласно которой при каждом i оба результата наблюдения x_i и y_i являются реализациями независимых, одинаково распределенных случайных величин X_i и Y_i. При нулевой гипотезе вероятность того, что случайная разность $Z_i = X_i - Y_i$ окажется положительной «+» в точности равна вероятности того, что эта разность будет отрицательной «−». Иначе говоря, проверяемая гипотеза H_0 утверждает, что для каждого i справедливо выражение

$$\mathbf{P}\{Z_i > 0\} = \mathbf{P}\{Z_i < 0\} = p = 1/2. \qquad (5.6.1)$$

При сделанных предположениях относительно всех $Z_1, Z_2, \ldots, Z_n$ число положительных исходов подчиняется биномиальному закону распределения вероятностей (1.3.8). Применительно к рассматриваемой задаче $p = q$, в результате чего вероятность появления k положительных исходов из n возможных

$$p_n(k) = (1/2)^n C_n^k. \qquad (5.6.2)$$

Аналогичное распределение вероятностей имеет и число отрицательных исходов ($Z_i < 0$). Поэтому проверка нулевой гипотезы H_0 сводится к проверке значимости расхождения между частотами (оценками вероятностей) $p_1 = k_n/n$ и $p_2 = (n - k_n)/n$, с одной стороны, и гипотетической вероятностью $p = 1/2$, с другой. В общем виде для достаточно большого объема выборки и произвольного значения вероятности p биномиального закона распределения подобная задача проверки гипотезы рассмотрена в п. 5.3.4.

Применительно к рассматриваемой непараметрической задаче проверки гипотезы в качестве статистики *критерия знаков* рассматривается случайное число положительных K_n и отрицательных $(n - K_n)$ знаков, а критерий проверки гипотезы для заданного уровня значимости α отвергает гипотезу независимости выборок тогда, когда в опыте число положительных k_n^* или число отрицательных $(n - k_n^*)$ знаков превышает одну и ту же границу (критическое значение) $k_{\alpha/2;n}$. Так как события $Z_i > 0$ и $Z_i < 0$ несовместны, то

$$\mathbf{P}\{K_n > k_{\alpha/2;n}; (n - K_n) > k_{\alpha/2;n}\} = \\ = \mathbf{P}\{K > k_{\alpha/2;n}\} + \mathbf{P}\{(n - K) > k_{\alpha/2;n}\} = \alpha. \qquad (5.6.3)$$

Учитывая, что критическое значение $k_{\alpha/2;n}$ — число целое, оно определяется исходя из одновременного выполнения неравенств

$$\mathbf{P}\{K_n \geqslant k_{\alpha/2;n}\} = \left(\frac{1}{2}\right)^n \sum_{i=k_{\alpha/2;n}}^{n} C_n^i \geqslant \frac{\alpha}{2};$$
$$\mathbf{P}\{K_n \geqslant k_{\alpha/2;n+1}\} = \left(\frac{1}{2}\right)^n \sum_{i=k_{\alpha/2;n+1}}^{n} C_n^i < \frac{\alpha}{2}. \quad (5.6.4)$$

Иначе говоря, критическое значение числа положительных или отрицательных исходов — наибольшее целое число, для которого выполняется правая часть первого уравнения (5.6.4), или наименьшее целое число, для которого выполняется правая часть второго уравнения (5.6.4).

В табл. 7 Приложения 2 приведены критические значения числа положительных и отрицательных знаков $k_{\alpha/2;n}$ для $n \leqslant 50$ с уровнями значимости $\alpha = 0{,}1;\ 0{,}05;\ 0{,}01$.

Как отмечалось выше при рассмотрении биномиального распределения вероятностей, оно при достаточно большом объеме выборки ($n \gg 1$) и определенных ограничениях на вероятности p и $q = 1-p$ (np и $nq \gg 1$) может быть аппроксимировано нормальным распределением со средним значением np и дисперсией npq. С учетом поправки на дискретность распределения числа знаков асимптотическая функция распределения числа положительных знаков имеет вид

$$F(k_n) = \Phi\left(t_n = \frac{k_n + 0{,}5 - n/2}{0{,}5\sqrt{n}}\right), \quad (5.6.5)$$

где $\Phi(t)$ — табулированная функция нормального распределения. Тогда нормированное критическое значение $t_{\alpha/2;n}$ определяется из соотношения

$$\alpha/2 = 1 - \Phi(t_{\alpha/2;n}). \quad (5.6.6)$$

Нормированное критическое значение $t_{\alpha/2;n}$ связано с критическим значением $k'_{\alpha/2;n}$ выражением

$$t_{\alpha/2;n} = \frac{k'_{\alpha/2;n} - n/2 + 0{,}5}{0{,}5\sqrt{n}}. \quad (5.6.7)$$

При этом в качестве критического значения $k_{\alpha/2;n}$ необходимо взять наименьшее целое число $k_{\alpha/2;n} = [k'_{\alpha/2;n}]$. Практически для критерия знаков аппроксимация биномиального распределения вероятностей нормальной кривой вполне приемлема уже при $n \geqslant 50$.

Пример 5.6.1. Проверим гипотезу о независимости двух выборок, данные о которых приведены в табл. 5.6.1 ($n = 17$, $k_n = 10$, $n - k_n = 7$) с уровнем значимости $\alpha = 0{,}05$.

Определим критическое значение числа положительных знаков для заданных объемов выборок и уровня значимости. Из табл. 7 Приложения 2 для $\alpha = 0{,}05$ и $n = 17$ находим критическое значение выборочной статистики $k_{0{,}025;17} = 13$. Так как

число положительных исходов $k_n^* = 10$ и число отрицательных исходов $(n - k_n^*) = 7$ меньше критического значения, то нулевая гипотеза H_0 о независимости и одинаковом распределении вероятностей выборок не отвергается. К такому же выводу приходим, используя асимптотическое приближение распределения числа положительных исходов в виде нормального распределения. Уровню значимости $\alpha = 0{,}05$ соответствует $t_{0{,}025;17} = 1{,}96$ или $k'_{0{,}025;17} = 12{,}04$. При этом ближайшее минимальное целое число равно $k_{0{,}025;17} = 12$, что практически совпадает с точным критическим значением.

Критерий знаков кроме рассмотренной задачи проверки гипотезы независимости двух выборок может быть применен для проверки гипотез, когда случайные величины Z_i, принимающие два значения «+» или «−», не являются вообще разностями, что характерно для опытов, результаты которых трудно или нецелесообразно представлять в количественной форме (например, в психологических опытах, общественно-политических опросах и т.п.). Он может быть применен также и для проверки гипотезы случайности (независимости) значений одной выборки $\boldsymbol{x}$, если под z_i понимать разность $x_i - \bar{x}$, где $\bar{x}$ — выборочное среднее.

Несмотря на простоту применения критерия знаков, существенным его недостатком является то, что заключенная в элементах выборок (выборки) информация о проверяемых гипотезах, используется явно недостаточно. Поэтому критерий знаков обычно применяется при предварительном анализе.

5.6.2. Критерий серий

Более информативным порядковым критерием по сравнению с критерием знаков для проверки гипотез независимости и случайности выборок является *критерий серий*. Рассмотрим суть статистики, соответствующей этому критерию. Пусть в результате проведения опыта получена последовательность двух взаимно исключающих событий «+» и «−». Вместо символов «+» и «−» часто используются буквенные символы, например «a» и «b». Одним из примеров может служит последовательность, приведенная в табл. 5.6.1. Подобная ситуация получается в результате бросания монеты, в которой каждое наблюдение либо герб «+», либо решка «−».

Обозначим через m и n количество положительных и отрицательных исходов. Тогда при справедливости нулевой гипотезы о случайности выборочных значений последовательность из m положительных исходов может быть записана C_{m+n}^m способами. Если в результате опытов появление любого из C_{m+n}^m вариантов одинаково вероятно, то элементы «+» и «−» в последовательности расположены случайно. В противоположной ситуации (элементы последовательности в определенной мере неслучайны) в конкретной выборке будет проявляться (в среднем) концентрация элементов «+» или «−» в той или иной области последовательности (например, смещение элементов того или другого знака к какому-либо краю последовательности). Одной из возможных статистик для количественной оценки степени «перемешивания» элементов в

последовательности является число серий $K_{m,n}$, составленных из этих элементов. *Серией называется последовательность однотипных наблюдений, перед и после которой следуют наблюдения противоположного типа или вообще нет никаких наблюдений.* Выборочное число серий обозначим как $k_{m,n}$. Так, в последовательности, приведенной в табл. 5.6.1, символы «+» и «−» группируются, образуя ряд серий. В приведенной конкретной последовательности таких серий $k^*_{m,n} = 11$, причем число положительных исходов $m = 10$, а число отрицательных исходов $n = 7$. Естественно, число серий изменяется от опыта к опыту случайным образом и зависит от того, зависимы или независимы (случайны) элементы последовательности.

Критерий серий и соответствующая статистика $K_{m,n}$, как правило, применяются для проверки гипотезы случайности одной выборки, для чего выборочные значения x_i сравниваются с выборочным средним $\bar{x}$ ($z_i = x_i - \bar{x} > 0$ соответствует «+», $z_i = x_i - \bar{x} < 0$ соответствует «−») или выборочной медианой x^*_{med} ($z_i = x_i - x^*_{med} > 0$ и $z_i = x_i - x^*_{med} < 0$). Однако критерий серий может быть использован для проверки гипотез и в других задачах математической статистики, например для проверки однородности распределений двух выборок.

Итак, пусть в результате опыта получено $k^*_{m,n}$ серий. Требуется ответить на вопрос: значимо или не значимо это число серий для того, чтобы принять или отвергнуть нулевую гипотезу H_0 о случайном расположении элементов «+» и «−» в последовательности. Так как при гипотезе H_0 элементы «+» и «−» равноправны, то полученные выводы будут справедливы как для $m \leqslant n$, так и для $m \geqslant n$. Случайное число серий $K_{m,n}$ может принимать значения, равные $k_{m,n} = 2, 3, \ldots, (2m+1)$ при $m < n$ и $k_{m,n} = 2, 3, \ldots, 2m$ при $m = n$. Так, если $m = 3$, $n = 5$, то максимальное число серий $\max k_{m,n} = 7$ соответствует расположению знаков, например, в последовательности вида «− + − + − + −−». Однако, если нулевая гипотеза верна, то вероятность появления малого и большого числа серий маловероятно. Поэтому, если выборочная статистика $k^*_{m,n}$ близка к минимальному или максимальному числу серий, то нулевая гипотеза должна быть отвергнута. Для принятия такого решения с заданным уровнем значимости α необходимо знать закон распределения числа серий для нулевой гипотезы H_0. На основе этого закона и заданного уровня значимости определяются верхняя $k_{\alpha/2;m,n}$ и нижняя $k_{(1-\alpha/2);m,n}$ границы (критические значения) выборочной статистики. При этом полагается, что вероятности отвергнуть гипотезу, когда она верна, при $k_{m,n} < k_{(1-\alpha/2);m,n}$ или $k_{m,n} > k_{\alpha/2;m,n}$ полагаются равными $\alpha/2$. Поскольку приведенные неравенства (события) несовместны, то полная вероятность отвергнуть нулевую гипотезу, когда она верна, равна α. Таким образом, гипотеза случайности (выборки независимы и одинаково распределены) принимается, если для выборочной статистики $k_{m,n}$ выполняется двойное неравенство

$$k_{(1-\alpha/2);m,n} \leqslant k^*_{m,n} \leqslant k_{\alpha/2;m,n}. \qquad (5.6.8)$$

В литературе по математической статистике, например в [8, 27, 31 и др.] приведены законы распределения вероятностей числа серий в последовательности, которые достаточно сложны для непосредственных вычислений критических значений рассматриваемой статистики. При этом среднее значение $m(k_{m,n})$ и дисперсия $D(k_{m,n})$ числа серий определяются формулами

$$m(k_{m,n}) = 1 + 2mn/(m+n); \qquad (5.6.9)$$

$$D(k_{m,n}) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}. \qquad (5.6.10)$$

Для нахождения критических значений нижней границы $k_{(1-\alpha/2);m,n}$ и верхней границы $k_{\alpha/2;m,n}$ имеются соответствующие таблицы [8, 51, 72].

Если $m \gg 1$ и $n \gg 1$, то дискретное распределение вероятностей числа серий можно аппроксимировать нормальным распределением с параметрами (5.6.9) и (5.6.10). С учетом поправки на дискретность асимптотическая функция распределения числа серий определяется выражением

$$F(k_{m,n}) = \Phi\left(t_{m,n} = \frac{k_{m,n} + 0,5 - m(k_{m,n})}{\sqrt{D(k_{m,n})}}\right). \qquad (5.6.11)$$

Нормальное приближение вполне допустимо для практических расчетов, если m или n больше 20.

В ряде прикладных задач порядковые статистические данные можно представить так, чтобы число положительных и отрицательных знаков (исходов) были равными, т.е. чтобы $m = n$. При этом формулы и таблицы существенно упрощаются. В данной ситуации

$$m(k_{m=n}) = m + 1; \quad D(k_{m=n}) = \frac{m(m-1)}{2m-1}, \qquad (5.6.12)$$

т.е. при независимости выборочных значений среднее число серий примерно равно количеству выборочных значений $m = n$, что интуитивно вполне понятно. Применительно к данному условию $m = n$ гипотеза случайности последовательности принимается с уровнем значимости α, если выполняется условие

$$k_{(1-\alpha/2);m=n} \leqslant k^*_{m=n} \leqslant k_{\alpha/2;m=n}. \qquad (5.6.13)$$

В табл. 8 Приложения 2 приведены критические значения (нижняя и верхняя границы) числа серий $k_{(1-\alpha/2);m=n}$ и $k_{\alpha/2;m=n}$ для $m = n$.

Пример 5.6.2. Проверим гипотезу случайности (независимости) данных, приведенных в табл. 5.6.2 [6], с уровнем значимости $\alpha = 0{,}05$.

Образуем последовательность в виде элементов «+» и «−», сравнивая исходную выборку с выборочной медианой x^*_{med}. Согласно определению, данному в п. 3.1, выборочная медиана — это значение случайной величины, по сравнению с которым

равное число наблюдений одновременно будет меньше и больше x^*_{med}. Из рассмотрения таблицы видно, что выборочной медианой является $x^*_{\text{med}} = 5{,}6$, так как из 20 наблюдений 10 выборочных значений больше $x^*_{\text{med}} = 5{,}6$ и 10 выборочных значений меньше $x^*_{\text{med}} = 5{,}6$. При этом последовательность положительных и отрицательных знаков примет вид, приведенный в табл. 5.6.3, из которой видно, что в последовательности из 20 наблюдений имеется $k^*_{10,10} = 13$ серий.

Таблица 5.6.2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_i	5,5	5,1	5,7	5,2	4,8	5,7	5,0	6,5	5,4	5,8	6,8	6,6	4,9	5,4	5,9	5,4	6,8	5,8	6,9	5,5

Таблица 5.6.3

Знак	−−	+	−−	+	−	+	−	+++	−−	+	−	+++	−
№ серии	1	2	3	4	5	6	7	8	9	10	11	12	13

Из табл. 8 Приложения 2 для нулевой гипотезы при заданном уровне значимости $\alpha = 0{,}05$ и $m = n = 10$ находим критические значения нижней и верхней границ: $k_{0,975;10,10} = 6$, $k_{0,025;10,10} = 16$. Так как выборочное значение числа серий, равное 13, попадает в интервал, заключенный между 6 и 16 сериями, гипотеза о случайности исходной последовательности может быть принята с уровнем значимости 0,05.

Если использовать асимптотическое нормальное приближение для распределения числа серий с выборочным средним значением $m^*(k_{10,10}) = 11$ и выборочной дисперсией $D(k_{10,10}) = 4{,}74$, то согласно (5.6.11) и данным табл. 1в Приложения 2 для $\alpha/2 = 0{,}025$ и $1 - \alpha/2 = 0{,}975$ получаем соответственно $t_{0,025;10,10} = 1{,}96$ и $t_{0,975;10,10} = -1{,}96$. Эти данные эквивалентны ненормированным критическим значениям $k_{0,025;10,10} = [k'_{0,025;10,10} = 14{,}77] = 14$ и $k_{0,975;10,10} = [k'_{0,975;10,10} = 6{,}23] = 6$. Полученные приближенным методом границы критической области практически совпадают с точными значениями.

5.6.3. Критерий инверсий

Таблица 5.6.4

i	1	2	3	4	5	6
x_i	5	3	8	9	4	2
k_i	3	1	2	2	1	0

Для проверки гипотезы случайности выборок (выборки независимы и одинаково распределены) наиболее эффективным критерием является *критерий инверсий*. Понятие инверсий обсуждается выше в п. 5.5.3 при рассмотрении гипотезы однородности выборок по критерию Вилкоксона. Применительно к проверке случайности сущность критерия инверсий проиллюстрируем сначала на примере, для чего рассмотрим последовательность из шести выборочных значений, представленных в первых двух строках табл. 5.6.4.

Подсчитаем, сколько раз в этой последовательности имеет место неравенство $x_i > x_j$ при $i < j$, т.е. сколько раз каждое значение x_i будет больше всех последующих значений x_j. *Каждое неравенство*, как и выше, *называется инверсией*. Общее число инверсий для данного x_i обозначим через k_i. Из рассмотрения приведенной последовательности имеем: $x_1 > x_2, x_5, x_6$ $(k_1 = 3)$; $x_2 > x_6$ $(k_2 = 1)$; $x_3 > x_5, x_6$ $(k_3 = 2)$; $x_4 > x_5, x_6$ $(k_4 = 2)$; $x_5 > x_6$ $(k_5 = 1)$; x_6 не имеет инверсий $(k_6 = 0)$. Общее число инверсий $g^*_{n=6} = g^*_6 = \sum k_i = 9$. Величина $g^*_6 = 9$ является реализацией некоторой случайной величины

$G_n = G_6$. При нулевой гипотезе она зависит от объема выборки, поэтому в качестве статистики, как и выше (п. 5.5.3), рассматривается общее число инверсий G_n. Для принятия или отклонения нулевой гипотезы необходимо найти функцию распределения числа инверсий при условии справедливости нулевой гипотезы.

Итак, пусть имеется последовательность из n выборочных значений x_i $(i = 1, 2, \ldots, n)$. Подсчитаем, сколько раз $x_i > x_j$ для всех i при $i < j$, и обозначим это число k_i. Общее число этих неравенств (инверсий) g^*_n является выборочной статистикой критерия инверсий. Величина g_n представляет собой реализацию некоторой случайной величины G_n. Статистика G_n при нулевой гипотезе является естественной мерой «беспорядка» среди наблюдений. Крайние случаи, когда $x_1 < x_2 < \ldots < x_n$ и $x_1 > x_2 > \ldots > x_n$, можно рассматривать как свидетельство отсутствия подобного «беспорядка». В первом случае статистика принимает значение $g^*_n = 0$, а во втором случае она максимальна и равна $g^*_n = 1 + 2 + \ldots + (n-1) = n(n-1)/2$. Таким образом, выборочные статистики, близкие к этим крайним значениям, следует рассматривать как критические. В [27] изложен подход к определению функции распределения числа инверсий. При этом среднее значение и дисперсия числа инверсий равны

$$m(G_n) = \frac{n(n-1)}{4}; \quad D(G_n) = \frac{n(2n+5)(n-1)}{72}. \tag{5.6.14}$$

Из первой формулы видно, что среднее значение числа инверсий для нулевой гипотезы H_0 совпадает с серединой интервала $[0, n(n-1)/2]$.

В соответствии с уравнением $\mathbf{P}\{G_n > g_{\beta;n}\} = \beta$ в [6] приведена таблица критических значений числа инверсий $g_{\beta;n}$, для ряда значений уровня значимости β и объема выборки n. Она повторена в табл. 9 Приложения 2.

Итак, гипотеза о случайности выборочных значений не отвергается, если выборочная статистика числа инверсий g^*_n лежит в интервале

$$g_{(1-\alpha/2);n} \leqslant g^*_n \leqslant g_{\alpha/2;n}. \tag{5.6.15}$$

Критерий инверсий применяется примерно так же, как и критерии серий и знаков. Однако он более эффективен (более мощный) по сравнению с выше названными критериями при необходимости обнаружения монотонного изменения среднего значения выборок (обнаружения тренда). Физически это связано с тем, что статистика критерия учитывает сравнение всех выборочных значений между собой, что дает большую информацию по сравнению со статистикой знаков и серий. Однако критерий инверсий менее чувствителен при выявлении тренда типа флуктуаций.

При достаточно большом объеме выборок $n \gg 1$ (практически при $n \geqslant 10$ [27]) точное распределение числа инверсий можно аппроксимировать нормальным распределением со средним значением и дисперсией, определяемыми выражениями (5.6.14). Тогда критические границы

числа инверсий (их целые числа) определяются из соотношений

$$\frac{\alpha}{2} = 1 - \Phi\left(\frac{g_{\alpha/2;n} - m(G_n)}{\sqrt{D(G_n)}}\right); \quad \frac{\alpha}{2} = 1 - \Phi\left(\frac{g_{(1-\alpha/2);n} - m(G_n)}{\sqrt{D(G_n)}}\right). \tag{5.6.16}$$

Пример 5.6.3. Проверим независимость (случайность) выборок в последовательности из $n = 20$ наблюдений из примера 5.6.2, рассмотренного при обсуждении критерия серий. Зададим уровень значимости $\alpha = 0{,}05$. Подсчитаем число инверсий в анализируемой последовательности: $k_1 = k_3 = k_8 = 8$; $k_2 = k_4 = k_{15} = 3$; $k_5 = k_{13} = k_{14} = k_{16} = 0$; $k_6 = k_{12} = 6$; $k_7 = k_9 = k_{18} = k_{19} = 1$; $k_{10} = 4$; $k_{11} = 7$; $k_{17} = 2$. Общее число выборочной статистики инверсий равно $g^*_{20} = \sum k_i = 62$. По табл. 9 Приложения 2 для уровня значимости $\alpha = 0{,}05$ находим нижнюю и верхнюю границы критических значений статистики: $g_{(1-0{,}025);20} = g_{0{,}975;20} = 64$, $g_{0{,}025;20} = 125$. Так как значение выборочной статистики $g^*_{20} = 62$ не попадает в интервал, заключенный между 64 и 125, то гипотеза о независимости (случайности) выборочных значений должна быть отвергнута с уровнем значимости $\alpha = 0{,}05$. К подобному выводу приходим, если воспользоваться и приближенным решением, основанным на аппроксимации точного распределения числа инверсий нормальным приближением. Действительно, применительно к рассматриваемому примеру $\alpha/2 = 0{,}025$; $m(G_{20}) = 95$; $D(G_{20}) = 237{,}5$. В соответствии с (5.6.16) имеем $\Phi[(g_{0{,}025;20} - 95)/15{,}4] = 0{,}975$; $\Phi[(g_{0{,}975;20} - 95)/15{,}4] = 0{,}025$. Используя таблицы нормальной функции распределения, получаем $g_{0{,}025;20} = 125$ и $g_{0{,}975;20} = [64{,}82] = 64$. Таким образом, аппроксимация точного распределения приближенным нормальным распределением дает тот же вывод о гипотезе случайности представленной выборки. При использовании критерия серий гипотеза о независимости этих же выборочных значений была принята. Данный пример иллюстрирует различную чувствительность этих двух критериев проверки гипотезы случайности.

Контрольные вопросы и задачи

5.1. В чем отличие параметрических и непараметрических методов проверки гипотез?

5.2. В каких случаях используются критерии проверки гипотез максимальной апостериорной вероятности, максимального правдоподобия и Неймана–Пирсона?

5.3. Какой критерий называется равномерно наиболее мощным несмещенным критерием проверки сложных гипотез?

5.4. Как выбирается вид гипотетической функции распределения при проверке гипотезы согласия?

5.5. Применительно к гауссовской выборке с известным средним значением m привести правило принятия гипотезы $\sigma^2 = \sigma_0^2$ против альтернативы $\sigma^2 = \sigma_1^2 > \sigma_0^2$ по критерию Неймана–Пирсона с уровнем значимости $\alpha = 0{,}05$; объем выборки $n = 20$.

5.6. Применительно к независимой выборке объема n с экспоненциальным распределением (1.4.20) с помощью критерия максимального правдоподобия найти правило проверки гипотезы $\lambda = \lambda_0$ против альтернативы $\lambda = \lambda_1$.

5.7. Полагая, что случайные величины X и Y, выборочные значения которых приведены в табл. 5.6.1, распределены по нормальному закону, проверить (с уровнем значимости $\alpha = 0{,}05$) гипотезы о равенстве их средних значений и дисперсий.

5.8. Для уровня значимости $\alpha = 0{,}05$ на основе критерия хи-квадрат проверить гипотезу о согласии выборочных данных, приведенных в табл. 5.6.1 для случайной величины X, нормальному распределению.

5.9. Используя критерий хи-квадрат, проверить гипотезу однородности случайных величин X и Y, приведенных в табл. 5.6.1, для уровня значимости $\alpha = 0{,}1$ и пяти одинаковым по ширине интервалам группирования выборочных значений. Полагать, что полный интервал изменяется от 4,5 до 7,0.

5.10. В соответствии с критерием Вилкоксона, используя приближенный вид распределения статистики, проверить гипотезу однородности двух выборок, приведенных в таблице, для уровня значимости $\alpha = 0{,}05$.

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

5.11. С помощью порядкового критерия знаков проверить гипотезу о независимости выборочных значений x_i и y_i, приведенных в таблице задачи 5.10, с уровнем значимости $\alpha = 0{,}05$.

5.12. С помощью порядкового критерия знаков проверить гипотезу случайности выборочных значений x_i, приведенных в таблице задачи 5.10, для уровня значимости $\alpha = 0{,}05$ путем сравнения элементов выборки с выборочным средним $\bar{x}$. Точное решение сравнить с нормальным приближением статистики числа положительных и отрицательных знаков.

5.13. С помощью порядкового критерия серий для уровня значимости $\alpha = 0{,}1$ проверить гипотезу случайности (независимости) выборочных данных случайной величины X, приведенных в табл. 5.6.1, сравнивая их с выборочной медианой.

5.14. На основе порядкового критерия инверсий для уровня значимости $\alpha = 0{,}05$ проверить гипотезу случайности (независимости) выборочных данных случайной величины X, приведенных в таблице задачи 5.10.

5.15. Что позволяет установить критерий согласия Колмогорова кроме основного своего назначения?

ГЛАВА 6
Дисперсионный анализ

При проведении опытов и испытаний имеется ряд *факторов*, значения которых влияют на изменчивость средних значений случайных величин. Значения этих факторов называют *уровнями факторов*. Уровни факторов могут выражаться различным образом: в виде числовых значений, качественных характеристик или в виде определенной комбинации начальных условий и т.д. Например, при изучении влияния изменения электрического напряжения источника тока на нормальное функционирование некоторого устройства изучаемым фактором может являться напряжение источника тока, а уровнями фактора — его численные значения. В случае исследования влияния комплекса удобрений на урожайность растения фактором является комплекс удобрений, а уровнем фактора — различные сочетания химических элементов, входящих в этот комплекс (азота, калия, фосфора и т.п.). Иногда уровни фактора называют классами фактора [13] или способами обработки. Последний термин имеет прямое отношение к агротехническим приемам обработки почвы.

Прежде чем исследовать влияние тех или иных факторов на тот или иной признак изучаемого явления или устройства, желательно, хотя бы приближенно, установить: полученные экспериментальные данные обусловлены наличием зависимости признака (признаков) от уровней фактора (факторов) или они вызваны естественными случайными явлениями (ошибками). При этом такие предварительные исследования должны осуществляться на основе относительно небольших временны́х и других затрат.

Исследование влияния одного или нескольких факторов на изменчивость средних значений изучаемых случайных величин называется *факторным анализом*. Если изучается влияние одного фактора на изменчивость среднего значения, то такое исследование называется *однофакторным анализом*. Изучение влияния нескольких факторов на изменчивость среднего значения носит название *многофакторного анализа*.

Основу дисперсионного анализа составляет представление результата эксперимента в виде линейной зависимости от уровней изучаемых факторов и их взаимного влияния, а также ошибок, которые появляются по разным неконтролируемым причинам при проведении эксперимента.

На основе этой модели эксперимента конструируется (предлагается) некоторая *статистика*, в основе которой лежат суммы квадратов отклонений выборочного среднего по каждому из уровней от общего выборочного среднего и выборочных значений по каждому из уровней факторов от выборочных средних по этим же уровням. Путем анализа этой статистики выносится то или иное решение относительно влияния уровней факторов на изменчивость средних значений изучаемых случайных величин (событий, явлений). Для этого сравнивают полученные результаты с предполагаемыми результатами применительно к нулевой гипотезе H_0, согласно которой уровни фактора не влияют на изменчивость среднего значения.

Если нулевая гипотеза не подтверждается, необходимо перейти к подробному исследованию влияния отдельных уровней фактора, для чего производится оценка расхождения выборочных средних, полученных при исследованиях по отдельным уровням фактора. В этой ситуации проверка гипотезы о значимости расхождения средних значений проводится, как правило, для двух разных уровней фактора на основе статистик, соответствующих анализируемым данным.

Основу исследования влияния отдельных уровней фактора на выходной эффект составляет *дисперсионный анализ*.

При классическом дисперсионном анализе полагается, что исследуемые величины имеют нормальное распределение с постоянной дисперсией и средними значениями, которые могут отличаться для разных выборочных совокупностей. В качестве критерия проверки нулевых гипотез используется критерий дисперсионного отношения. Хотя, строго говоря, дисперсионный анализ справедлив для гауссовских случайных величин, с определенной, достаточно приемлемой для прикладных задач, точностью, он может быть использован и для негауссовских случайных величин [13], причем с увеличением объема выборок n (практически $n \geqslant 4$) для каждого из уровней фактора погрешность приближения будет небольшой. Если требуется большая точность выводов, а распределения имеющихся данных неизвестны, в качестве критериев проверки гипотез, как правило, применяются критерии, свободные от распределений. Как отмечалось выше в гл. 5, такие критерии называются непараметрическими критериями. К их числу относятся и ранговые критерии, для которых численные значения экспериментальных данных заменяются их рангами. Ранговые критерии применяются и тогда, когда в качестве исходных данных фигурируют качественные признаки типа «больше – меньше», «лучше – хуже» и т.д. Такой анализ называют *непараметрическим* (*ранговым*) *дисперсионным анализом* [68].

6.1. Однофакторный дисперсионный анализ

Рассмотрим сущность *дисперсионного анализа применительно к одному фактору* с h уровнями.

Пусть в результате эксперимента получено h серий выборок гауссовской случайной величины Y: $\mathbf{y}_1(y_{11}, y_{12}, \ldots, y_{1n_1})$, $\mathbf{y}_2(y_{21}, y_{22}, \ldots, y_{2n_2})$, $\ldots$, $\mathbf{y}_h(y_{h1}, y_{h2}, \ldots, y_{hn_h})$, соответствующих h уровням (классам) некоторого изучаемого фактора A. Объем каждой серии выборок равен n_i, $i = 1, 2, \ldots, h$. Дисперсии выборок полагаются равными σ^2, хотя и неизвестными. Обозначим: m_i — среднее значение изучаемой случайной величины Y на i-м уровне; m — общее среднее значение всей совокупности случайных величин $\mathbf{Y}(Y_1, \ldots, Y_n)$, где $n = \sum n_i$ — общее число выборочных значений. Как правило, наблюденные данные записываются в виде таблицы (табл. 6.1.1).

Таблица 6.1.1

Результаты опытов

Уровни фактора i	Результаты опыта j							Сумма по строкам	Выборочное среднее $y_{i.}$
1	y_{11}	y_{12}	y_{13}	$\ldots$	y_{1j}	$\ldots$	y_{1n_1}	$\sum y_{1j}$	$y_{1.}$
2	y_{21}	y_{22}	y_{23}	$\ldots$	y_{2j}	$\ldots$	y_{2n_2}	$\sum y_{2j}$	$y_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
i	y_{i1}	y_{i2}	y_{i3}	$\ldots$	y_{ij}	$\ldots$	y_{in_i}	$\sum y_{ij}$	$y_{i.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
h	y_{h1}	y_{h2}	y_{h3}	$\ldots$	y_{hj}	$\ldots$	y_{hn_h}	$\sum y_{hj}$	$y_{h.}$

Математическую модель наблюдаемого значения y_{ij} можно представить в виде линейной зависимости

$$y_{ij} = m_i + \xi_{ij} = m + a_i + \xi_{ij}, \tag{6.1.1}$$

где

$$m_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{M}\{y_{ij}\}; \quad m = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i} \mathrm{M}\{y_{ij}\} = \frac{1}{n}\sum_{i=1}^{h} n_i m_i \tag{6.1.2}$$

— среднее значение по строкам и общее среднее значение всей совокупности опытных данных соответственно; $a_i = (m_i - m)$ — величина, характеризующая влияние i-го уровня фактора на общее среднее значение; $\xi_{ij} = (y_{ij} - m_i)$ — независимые нормально распределенные случайные величины с нулевыми средними значениями $\mathrm{M}\{\xi_{ij}\} = 0$ и одинаковыми дисперсиями $\mathrm{M}\{\xi_{ij}^2\} = \sigma^2$. При этом, естественно, должно выполняться тождество

$$n_1 a_1 + n_2 a_2 + \ldots + n_h a_h = 0, \tag{6.1.3}$$

которое вытекает из условия

$$\mathrm{M}\left\{\frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i} y_{ij}\right\} = \mathrm{M}\left\{\frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i}(m + a_i + \xi_{ij})\right\} = m.$$

На основе полученных h выборок общим объемом n необходимо проверить нулевую гипотезу H_0 о том, что уровни фактора не влияют на общее среднее значение результата эксперимента, т.е. нулевая гипотеза соответствует равенствам

$$m_1 = m_2 = \ldots = m_h = m \tag{6.1.4}$$

или

$$a_1 = a_2 = \ldots = a_h = 0. \tag{6.1.5}$$

При отсутствии случайных отклонений («ошибок») ($\xi_{ij} = 0$) $y_{ij} \equiv m_i = m + a_i$; поэтому решение о справедливости нулевой гипотезы можно было бы принять на основе сравнения выборочных значений y_{ij}. Естественно, из-за наличия ошибок выборочные средние, являясь случайными величинами, отличаются друг от друга. Необходимо выяснить: вызвано это отличие («расхождение») чисто случайными причинами либо оно вызвано существенной зависимостью среднего значения от уровня исследуемого фактора A. Для этого проверяется нулевая гипотеза H_0, согласно которой для заданного уровня значимости α уровни фактора не должны влиять на среднее значение изучаемой случайной величины Y.

Для дальнейшего анализа введем следующие выборочные средние:

$$\bar{y}_i \equiv y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = m_i + \xi_{i.} \tag{6.1.6}$$

— выборочное среднее по уровням (строкам);

$$\bar{y} \equiv y_{..} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{h} n_i y_{i.} = m + \xi_{..} \tag{6.1.7}$$

— общее выборочное среднее (по всем уровням), где

$$\xi_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} \xi_{ij}; \quad \xi_{..} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{n_i} \xi_{ij} = \frac{1}{n}\sum_{i=1}^{h} n_i \xi_{i.} \tag{6.1.8}$$

— выборочное среднее по уровням (строкам) и общее выборочное среднее случайной величины ξ_{ij}. Здесь, как и выше, замена у случайных величин y и ξ индекса (индексов) точкой указывает на то, что случайная величина усреднена по переменной, соответствующей этому индексу (индексам).

Оценки составляющих линейной модели (6.1.1) m, a_i и ξ_{ij} могут быть получены методом наименьших квадратов, рассмотренным в п. 3.5.5. Применительно к оценке m^* среднего значения m она находится из решения уравнения

$$\frac{\partial}{\partial m}\sum_{i=1}^{h}\sum_{j=1}^{n_i}[y_{ij} - (m + a_i)]^2\Big|_{m=m^*} = -2\sum_{i=1}^{h}\sum_{j=1}^{n_i}[y_{ij} - (m + a_i)] =$$

$$= -2n(y_{..} - m^*) + 2\sum_{i=1}^{h} n_i a_i = 0.$$

Учитывая соотношение (6.1.1), получаем

$$m^* = y_{..} = (m + \xi_{..}). \tag{6.1.9}$$

Аналогично находим оценку параметра a_i:

$$a_i^* = y_{i.} - m^* = y_{i.} - y_{..} = (a + \xi_{i.}). \tag{6.1.10}$$

Из (6.1.1) вычисляем оценку случайной величины ξ_{ij}:

$$\xi_{ij}^* = y_{ij} - y_{i.}. \tag{6.1.11}$$

При сделанных предположениях о нормальном распределении изучаемых независимых случайных величин y_{ji} выборочные средние $y_{..}$ и $y_{i.}$ также распределены по нормальным законам с дисперсиями σ^2/n и σ^2/n_i соответственно.

Представим общую сумму квадратов отклонений всех наблюденных значений от общего среднего в виде

$$Q = \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{..})^2 = \sum_{i=1}^{h}\sum_{j=1}^{n_i}[(y_{ij} - y_{i.}) + (y_{i.} - y_{..})]^2 =$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2 + \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{i.} - y_{..})^2 + 2\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})(y_{i.} - y_{..}).$$

Последний член правой части равен нулю, так как

$$\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})(y_{i.} - y_{..}) = \sum_{i=1}^{h}(y_{i.} - y_{..})\sum_{j=1}^{n_i}(y_{ij} - y_{i.}) =$$

$$= \sum_{i=1}^{h}(y_{i.} - y_{..})(n_i y_{i.} - n_i y_{i.}) = 0.$$

Поэтому общая сумма квадратов отклонений Q является суммой двух составляющих:

$$Q = Q_A + Q_R = \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{..})^2 = \sum_{i=1}^{h}\sum_{j=1}^{n_i} y_{ij}^2 - ny_{..}^2 =$$

$$= \sum_{i=1}^{h} n_i(y_{i.} - y_{..})^2 + \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2. \tag{6.1.12}$$

Первая составляющая

$$Q_A = \sum_{i=1}^{h} n_i(y_{i.} - y_{..})^2 = \sum_{i=1}^{h} n_i y_{i.}^2 - ny_{..}^2 \tag{6.1.13}$$

представляет собой взвешенную сумму квадратов отклонений выборочных средних по каждому уровню $y_{i.}$ от общего выборочного среднего $y_{..}$. Вторая составляющая

$$Q_R = \sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{i.})^2 = \sum_{i=1}^{h}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{h} n_i y_{i.}^2 \tag{6.1.14}$$

отражает суммарное влияние случайных факторов по каждому из уравнений факторов.

Итак, если составляющая Q_A соответствует источнику изменчивости среднего значения, то составляющая Q_R соответствует источнику случайных отклонений, всегда присущих испытаниям и экспериментам. В силу тождества (6.1.12) обычно вычисляется общая сумма квадратов Q и сумма квадратов между уровнями Q_A. Сумма квадратов внутри уровней получается как

$$Q_R = Q - Q_A. \tag{6.1.15}$$

В связи с этим сумму Q_R часто называют остаточной суммой квадратов. Из формул (6.1.12) видно, что случайные величины Q, Q_A и Q_R имеют степени свободы соответственно

$$\nu = n - 1; \quad \nu_A = h - 1; \quad \nu_R = n - h. \tag{6.1.16}$$

При этом выполняется тождество для числа степеней свободы

$$\nu = \nu_A + \nu_R. \tag{6.1.17}$$

При сделанных предположениях о нормальном распределении независимых случайных величин y_{ji} с одинаковыми дисперсиями с помощью ортогональных преобразований согласно (4.9.17) и (4.9.18) случайные величины Q, Q_A, Q_R можно представить в виде сумм квадратов независимых случайных величин z_i с нулевыми средними значениями и одинаковыми дисперсиями σ^2:

$$Q = \sum_{i=1}^{n-1} z_i^2; \quad Q_A = \sum_{i=1}^{h-1} z_i^2; \quad Q_R = \sum_{i=1}^{n-h} z_i^2. \tag{6.1.18}$$

Случайные величины Q, Q_A и Q_R можно использовать для нахождения оценок соответствующих дисперсий:

$$s^2 = \frac{Q}{n-1} = \frac{1}{n-1}\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij} - y_{..})^2; \tag{6.1.19}$$

$$s_A^2 = \frac{Q_A}{h-1} = \frac{1}{h-1}\sum_{i=1}^{h} n_i (y_{i\cdot} - y_{\cdot\cdot})^2; \qquad (6.1.20)$$

$$s_R^2 = \frac{Q_R}{n-h} = \frac{1}{n-h}\sum_{i=1}^{h}\sum_{j=1}^{n_i} (y_{ij} - y_{i\cdot})^2. \qquad (6.1.21)$$

Первая из этих оценок называется оценкой общей дисперсии, вторая — оценкой дисперсии по уровням (группам, классам), третья — оценкой дисперсии внутри уровней (групп, классов).

Вычислим средние значения случайных величин Q, Q_A и Q_R применительно к нулевой гипотезе H_0, т.е. при предположении, что все выборочные значения имеют не только одинаковые дисперсии, но и одинаковые средние значения. Имеем

$$\mathrm{M}\{Q\} = \sum_{i=1}^{h}\sum_{j=1}^{n_i} \mathrm{M}\{(y_{ij} - y_{\cdot\cdot})^2\} = \sum_{i=1}^{h}\sum_{j=1}^{n_i} \sigma^2 \frac{n-1}{n} = \sigma^2(n-1); \qquad (6.1.22)$$

$$\mathrm{M}\{Q_A\} = \sum_{i=1}^{h} n_i \mathrm{M}\{(y_{i\cdot} - y_{\cdot\cdot})^2\} = \sum_{i=1}^{h} n_i \frac{\sigma^2}{n_i}\frac{h-1}{h} = \sigma^2(h-1); \qquad (6.1.23)$$

$$\mathrm{M}\{Q_R\} = \sum_{i=1}^{h}\sum_{j=1}^{n_i} \mathrm{M}\{(y_{ij} - y_{i\cdot})^2\} = \sum_{i=1}^{h}\sum_{j=1}^{n_i} \sigma^2 \frac{n_i - 1}{n_i} = \sigma^2(n-h). \qquad (6.1.24)$$

Из этих соотношений и формул (6.1.19)–(6.1.21) видно, что применительно к нулевой гипотезе каждая из величин s^2, s_A^2 и s_R^2 определяет одну и ту же несмещенную оценку дисперсии анализируемой случайной величины. Так как случайные величины Q_A и Q_R взаимно независимы, то взаимно независимы и оценки дисперсий s_A^2 и s_R^2.

В силу сделанного выше предположения о том, что выборочные значения y_{ij} независимы и распределены нормально с одинаковыми дисперсиями, случайная величина Q_R/σ^2 имеет хи-квадрат распределение с $\nu_R = (n-h)$ степенями свободы вне зависимости от характера проверяемой гипотезы. Распределение случайной величины Q_A/σ^2 зависит от вида гипотезы. Применительно к нулевой гипотезе эта случайная величина распределена по закону хи-квадрат с $\nu_A = (h-1)$ степенями свободы. В этом случае для проверки нулевой гипотезы можно воспользоваться статистикой дисперсионного отношения (F-статистикой)

$$F_{\nu_A,\nu_R} = \frac{s_A^2}{s_R^2} = \frac{n-h}{h-1}\frac{Q_A}{Q_R}, \qquad (6.1.25)$$

имеющей F-распределение вероятностей с числом степеней свободы

$\nu_A = (h-1)$ и $\nu_R = (n-h)$. Нулевая гипотеза отвергается, если величина выборочной статистики $F^*_{\nu_A,\nu_R}$ для заданного уровня значимости α превышает критическую границу этой статистики $F_{\alpha;h-1,n-h}$, т.е. нулевая гипотеза отвергается, если выполняется неравенство

$$F^*_{\nu_A,\nu_R} > F_{\alpha;h-1,n-h}, \qquad (6.1.26)$$

где $F_{\alpha;h-1,n-h}$ находятся по таблицам критических значений F-распределения

$$\mathbf{P}\{F_{\nu_A,\nu_R} > F_{\alpha;h-1,n-h}\} = \alpha.$$

Если нулевая гипотеза верна, то оценки дисперсий s_A^2 и s_R^2 должны различаться между собой лишь случайно и в пределе при $n \gg 1$ и $h \gg 1$ отношение (6.1.25) стремится к единице. При этом выборочная F-статистика будет показывать с большой вероятностью лишь несущественные расхождения между оценками дисперсий s_A^2 и s_R^2.

Вычислим среднее значение случайной величины s_A^2, если нулевая гипотеза неверна, т.е. хотя бы одно из средних значений m_i отличается от общего среднего значения m. Подставляя выражения (6.1.6)–(6.1.8) в (6.1.20) и вычисляя математическое ожидание получившегося выражения, имеем

$$\mathrm{M}\{s_A^2\} = \frac{1}{h-1}\sum_{i=1}^{h} n_i \mathrm{M}\{[(m_i + \xi_{i\cdot}) - (m + \xi_{\cdot\cdot})]^2\} =$$

$$= \frac{1}{h-1}\sum_{i=1}^{h} n_i \mathrm{M}\{(\xi_{i\cdot} - \xi_{\cdot\cdot})^2 + 2(\xi_{i\cdot} - \xi_{\cdot\cdot})(m_i - m) + (m_i - m)^2\} =$$

$$= \sigma^2 + \frac{1}{h-1}\sum_{i=1}^{h} n_i (m_i - m)^2. \qquad (6.1.27)$$

Из этого выражения видно, что если нулевая гипотеза неверна, то оценка (6.1.20) дисперсии σ^2 смещенная, причем смещение оценки

$$b(\sigma^2) = \frac{1}{h-1}\sum_{i=1}^{h} n_i (m_i - m)^2$$

увеличивается с возрастанием объема выборки для соответствующих уровней фактора. Среднее значение случайной величины Q_R не зависит от характера рассматриваемой гипотезы, так как

$$\mathrm{M}\{s_R^2\} = \frac{1}{n-h}\sum_{i=1}^{h}\sum_{j=1}^{n_i} \mathrm{M}\{(\xi_{ij} - \xi_{i\cdot})^2\} = \sigma^2.$$

Таблица 6.1.2

Схема вычисления однофакторного дисперсионного анализа

Изменчивость (источник изменчивости)	Сумма квадратов	Число степеней свободы ν	Оценка дисперсии
Общая	$Q=\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij}-y_{..})^2$	$n-1$	$s^2=\dfrac{Q}{n-1}$
По уровням (между уровнями)	$Q_A=\sum_{i=1}^{h}n_i(y_{i.}-y_{..})^2$	$h-1$	$s_A^2=\dfrac{Q_A}{h-1}$
Остаточная (внутриуровней)	$Q_R=\sum_{i=1}^{h}\sum_{j=1}^{n_i}(y_{ij}-y_{i.})^2$	$n-h$	$s_R^2=\dfrac{Q_R}{n-h}$

Поэтому величину

$$\frac{h-1}{n}\mathrm{M}\{(s_A^2-s_R^2)\}=\frac{1}{n}\sum_{i=1}^{h}n_i(m_i-m)^2 \qquad (6.1.28)$$

можно рассматривать как меру изменчивости неизвестных средних значений m_i по уровням i исследуемого фактора.

Выше изложенное применительно к ненулевой гипотезе качественно можно пояснить следующим образом. Если нулевая гипотеза неверна, то величина s_R^2, характеризующая случайные отклонения, с увеличением числа элементов выборки для каждого уровня в среднем стремится к σ^2. В то же время величина s_A^2 из-за расхождения средних значений увеличивается. В результате значение статистики будет возрастать и, естественно, выходить за критическую границу $F_{\alpha;h-1,n-h}$. Иначе говоря, если нулевая гипотеза неверна, то выборочная F-статистика зафиксирует существенное отличие между дисперсиями s_A^2 и s_R^2, что будет свидетельствовать о неправомочности гипотезы H_0. В итоге приходим к выводу о том, что уровни (классы, группы) изучаемого фактора оказывают влияние на среднее значение анализируемой случайной величины Y.

Действия при выполнении однофакторного дисперсионного анализа можно представить в виде табл. 6.1.2.

При отклонении нулевой гипотезы необходим дальнейший более детальный анализ: какие из уровней данного фактора оказывают существенное влияние на среднее значение изучаемых величин (событий). Эту проверку можно осуществить с помощью критерия Стьюдента (t-критерия), рассмотренного выше в п. 5.3.1 и позволяющего попарно проверить гипотезу о равенстве средних значений в конкретных двух сериях наблюдений, соответствующих двум уровням исследуемого фактора. Следует отметить, что применительно к двум уровням фактора (двум сериям выборок) F-критерий однозначно связан с критерием Стьюдента. Действительно, пусть имеются две выборки объемов n_1 и n_2. В этом

случае $h=2$, $n=n_1+n_2$, а частные и общее выборочные средние равны

$$y_{i.}=\frac{1}{n_1}\sum_{j=1}^{n_1}y_{1j};\quad y_{2.}=\sum_{j=1}^{n_2}y_{2j};\quad y_{..}=\frac{n_1y_{1.}+n_2y_{2.}}{n_1+n_2}. \qquad (6.1.29)$$

При этом выборочные дисперсии s_A^2 и s_R^2 определяются формулами

$$s_A^2=\frac{Q_A}{h-1}=n_1(y_{1.}-y_{..})^2+n_2(y_{2.}-y_{..})^2=\frac{n_1n_2}{n_1+n_2}(y_{1.}-y_{2.})^2; \qquad (6.1.30)$$

$$s_R^2=\frac{Q_R}{n-2}=\frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}(y_{1j}-y_{1.})^2+\sum_{j=1}^{n_2}(y_{2j}-y_{2.})^2\right]. \qquad (6.1.31)$$

Поэтому

$$F_{\nu_A,\nu_R}=F_{1,n-2}=\frac{s_A^2}{s_R^2}=\frac{(y_{1.}-y_{2.})^2}{[1/n_1+1/n_2)]s_R^2}. \qquad (6.1.32)$$

Правая часть F-статистики (6.1.32) равна квадрату t-статистики (статистики Стьюдента), определенной формулой (5.3.5). Следовательно, при двух уровнях фактора F-критерий дисперсионного анализа совпадает с двухсторонним критерием Стьюдента.

Пример 6.1.1. Рассмотрим пример применения однофакторного дисперсионного анализа применительно к определению влияния типа звукопоглощающего покрытия на затухание звука в исследуемом помещении. Изучаются пять уровней ($h=5$). При этом на каждом уровне в одних и тех же *условиях опыты повторяются* четыре раза ($n_i=4=\mathrm{const}$). В результате опыта для каждого из уровней получены значения затуханий сигналов, представленные в табл. 6.1.3 в условных единицах.

С уровнем значимости $\alpha=0{,}05$ проверим зависимость затухания от типа звукопоглощающего покрытия. Учитывая, что значение дисперсии не зависит от начала отсчета данных, для упрощения вычислений каждое показание опыта, как правило, уменьшают на целое число, близкое к общему выборочному среднему. Применительно к рассматриваемому примеру все показания уменьшим на 70. Тогда необходимые вычисления можно свести в табл. 6.1.4.

По формулам (6.1.7), (6.1.12), (6.1.13) и (6.1.15) *вычисляем*: $y_{..}=0{,}85$; $Q=5713-20(0{,}85)^2=5698{,}55$; $Q_A=3550{,}75-20(0{,}85)^2=3536{,}30$; $Q_R=Q-Q_A=5698{,}55-3536{,}30=2162{,}25$.

Значения оценок дисперсий s^2, s_A^2 и s_R^2, вычисленных согласно (6.1.19)–(6.1.21), *представлены* в табл. 6.1.5. При этом выборочная F-статистика для проверки нулевой гипотезы равна $F_{4,15}^*=s_A^2/s_R^2=6{,}13$. По таблицам критических значений F-распределения (табл. 4 Приложения 2) для уровня значимости $\alpha=0{,}05$ и числа степеней свободы $\nu_A=4$ и $\nu_R=15$ находим критическое значение статистики $F_{0{,}05;4,15}=3{,}06$. Так как $F_{4,15}^*=6{,}13>F_{0{,}05;4,15}=3{,}06$, нулевая гипотеза о том, что рассматриваемые типы покрытий не влияют *на значение затухания* звука, должна быть отвергнута для заданного уровня значимости $\alpha=0{,}05$. Однако она не подтверждается и для уровня значимости $\alpha=0{,}01$ ($F_{0{,}01;4,15}=4{,}89$).

Поскольку средние значения затухания сигналов звука для различных уровней (*типа покрытия*) отличаются друг от друга, выясним, насколько эти отличия

Таблица 6.1.3

Результаты опытов

Уровень фактора i	Тип покрытия	Результаты опыта y_{ij}	Сумма $\sum_{j=1}^{n_i} y_{ij}$	Выборочное среднее $y_{i.}$
1	A	60 69 50 35	214	53,50
2	B	98 96 91 66	351	87,75
3	C	67 67 55 42	231	57,75
4	D	79 64 81 70	294	73,50
5	E	90 70 79 88	327	81,75

$$\sum_{i=1}^{5} y_{ij} = 1417; \quad y_{..} = \frac{1}{5}\sum_{i=1}^{5} y_{i.} = 70{,}85$$

Таблица 6.1.4

Уровни i	y_{ij}				$\sum y_{ij}$	n_i	$y_{i.}$	$\sum y_{ij}^2$	$n_i(y_{i.})^2$
1	−10	−1	−20	−35	−66	4	−16,50	1726	1089,00
2	28	26	21	−4	71	4	17,75	1917	1260,25
3	−3	−3	−15	−28	−49	4	−12,25	1027	600,25
4	9	−6	11	0	14	4	3,50	238	49,00
5	20	0	9	18	47	4	11,75	805	552,25
				Сумма	−17	20	4,25	5713	3550,75

Таблица 6.1.5

Таблица дисперсионного анализа

Изменчивость	Сумма квадратов	Число степеней	Оценка дисперсии
Общая	$Q = 5698{,}55$	19	$s^2 = 299{,}924$
По уровням	$Q_A = 3536{,}30$	4	$s_A^2 = 884{,}075$
Остаточная	$Q_R = 2162{,}25$	15	$s_R^2 = 144{,}150$

существенны с уровнем значимости $\alpha = 0{,}05$. В качестве критерия проверки соответствующих гипотез равенства выборочных средних применим двухсторонний критерий Стьюдента. Так как полагается, что дисперсии всех выборок одинаковы, в качестве выборочной дисперсии возьмем остаточную дисперсию $s_R^2 = 144{,}150$. При этом распределение статистики Стьюдента

$$t_{(i-j)} = |y_{i.} - y_{j.}|/s_R,$$

где $y_{i.}$ и $y_{j.}$ — выборочные средние затуханий для i-го и j-го уровней, имеет $n = 15$ степеней свободы. Для принятого уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $\nu = 15$ двустороннее критическое значение статистики критерия Стьюдента в соответствии с табл. 4 Приложения 2 равно $t_{\alpha/2;\nu} = t_{0{,}025;15} = 2{,}13$. Выборочное значение статистики при сравнении A и B типов покрытий (1-го и 2-го уровней фактора) равно $t^*_{(1-2)} = |53{,}50 - 87{,}75|/\sqrt{144{,}150} = 2{,}85$. Аналогично находятся выборочные статистики для других пар уровней (других пар покрытий). Их значения сведены в табл. 6.1.6.

Сравним полученные выборочные статистики с критическим значением $t_{0{,}025;15} = 2{,}13$.

Таблица 6.1.6

Выборочные статистики

Пары уровней	1–2 A–B	1–3 A–C	1–4 A–D	1–5 A–E	2–3 B–C	2–4 B–D	2–5 B–E	3–4 C–D	3–5 C–E	4–5 D–E
Значение выборочной статитики t^*	2,85	0,35	1,67	2,35	2,50	1,19	0,50	1,31	2,00	0,69

В результате сравнения можно сделать следующие выводы:

- имеется существенное отличие влияния на затухание звука покрытия типа A по сравнению с покрытиями типа B и E;
- отсутствует значимое отличие влияния на затухание звука покрытия типа C по сравнению с покрытием типа A;
- несущественная разница в затуханиях звука наблюдается при применении покрытий типа D по сравнению с типом C и т.д.

6.2. Двухфакторный дисперсионный анализ

С увеличением числа факторов, т.е. при двух, трех и более факторах, процедура дисперсионного анализа принципиально не меняется, хотя вычисления значительно усложняются по мере увеличения числа факторов и числа уровней факторов.

Рассмотрим сущность *двухфакторного анализа*. Пусть изучается влияние на среднее значение анализируемых случайных величин двух факторов A и B. Будем полагать, что по фактору A все наблюдения делятся на h уровней $(A_1, \ldots, A_h)$, а по фактору B — на q уровней $(B_1, \ldots, B_q)$, т.е. все экспериментальные данные разбиваются на hq уровней (групп, классов).

6.2.1. Двухфакторный анализ с одним наблюдением в «ячейке» («клетке»)

Ячейкой (клеткой), соответствующей паре уровней (i, j), будем называть место в таблице опытных данных, находящееся на пересечении уровней $i = 1, \ldots, h$ и $j = 1, \ldots, q$. На первом этапе изучения сущности двухфакторного анализа для упрощения рассуждений рассмотрим несколько идеализированную ситуацию, когда исследуемые факторы A и B действуют на среднее значение опытных данных *независимо* (в этом случае говорят, что факторы аддитивны). Будем полагать, что в каждой ячейке осуществляется одно наблюдение, т.е. общее число наблюдений равно $n = hq$. Тогда наблюдение y_{ij} является выборочным значением, попадающим в ячейку, соответствующую уровню фактора A_i, $i = 1, \ldots, h$, и уровню фактора B_j, $j = 1, \ldots, q$. Наблюдения y_{ij} могут быть представлены в виде табл. 6.2.1.

Таблица 6.2.1

Наблюдения над факторами A и B

		Уровень j фактора B				Среднее по строкам $y_{i.}$
		1	2	...	q	
Уровень i фактора A	1	y_{11}	y_{12}	...	y_{1q}	$y_{1.}$
	2	y_{21}	y_{22}	...	y_{2q}	$y_{2.}$
	⋮	⋮	⋮	⋮	⋮	⋮
	h	y_{h1}	y_{h2}	...	y_{hq}	$y_{h.}$
Среднее по столбцам $y_{.j}$		$y_{.1}$	$y_{.2}$	...	$y_{.q}$	$y_{..}$

Как и при однофакторном анализе полагается, что выборочные значения y_{ij} распределены по нормальному закону с одинаковыми дисперсиями σ^2. Математическая модель наблюдаемых данных подобно однофакторному анализу можно представить в виде

$$y_{ij} = m + a_i + b_j + \xi_{ij}, \tag{6.2.1}$$

где m — общее среднее значение изучаемой случайной величины; a_i и b_j — величины, характеризующие влияния i-го уровня фактора A и j-го уровня фактора B на общее среднее значение m; ξ_{ij} — независимые нормально распределенные случайные величины с нулевым средним значением и одной и той же дисперсией σ^2. При этом должны выполняться условия

$$\sum_{i=1}^{h} a_i = 0; \quad \sum_{j=1}^{q} b_j = 0. \tag{6.2.2}$$

Оценки параметров m, a_i, b_j линейной модели (6.2.1), вычисляемые методом наименьших квадратов, определяются соотношениями

$$m^* = y_{..}; \quad a_i^* = y_{i.} - y_{..}; \quad b_j^* = y_{.j} - y_{..}, \tag{6.2.3}$$

где выборочные средние по факторам A, B ($y_{i.}$, $y_{.j}$) и общее выборочное среднее ($y_{..}$) соответственно равны

$$y_{i.} = \frac{1}{q}\sum_{j=1}^{q} y_{ij}; \tag{6.2.4}$$

$$y_{.j} = \frac{1}{h}\sum_{i=1}^{h} y_{ij}; \tag{6.2.5}$$

$$y_{..} = \frac{1}{hq}\sum_{i=1}^{h}\sum_{j=1}^{q} y_{ij} = \frac{1}{h}\sum_{i=1}^{h} y_{i.} = \frac{1}{q}\sum_{j=1}^{q} y_{.j}. \tag{6.2.6}$$

Сумму квадратов отклонений всех $n = hq$ наблюдений от их общего выборочного среднего подобно соотношению (6.1.12) однофакторного

анализа можно представить в виде

$$Q = \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{..})^2 = \sum_{i=1}^{h}\sum_{j=1}^{q} y_{ij}^2 - ny_{..}^2 =$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{i.} - y_{.j} + y_{..} + y_{i.} - y_{..} + y_{.j} - y_{..})^2 =$$

$$= q\sum_{i=1}^{h}(y_{i.} - y_{..})^2 + h\sum_{j=1}^{q}(y_{.j} - y_{..})^2 + \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{i.} - y_{.j} + y_{..})^2 =$$

$$= Q_A + Q_B + Q_R. \tag{6.2.7}$$

Здесь

$$Q_A = q\sum_{i=1}^{h}(y_{i.} - y_{..})^2 = q\sum_{i=1}^{h} y_{i.}^2 - ny_{..}^2 \tag{6.2.8}$$

— взвешенная сумма квадратов отклонений выборочных средних по уровням фактора A от общего выборочного среднего;

$$Q_B = h\sum_{j=1}^{q}(y_{.j} - y_{..})^2 = h\sum_{j=1}^{q} y_{.j}^2 - n(y_{..})^2 \tag{6.2.9}$$

— взвешенная сумма квадратов отклонений выборочных средних по уровням фактора B от общего выборочного среднего;

$$Q_R = \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{i.} - y_{.j} + y_{..})^2 = Q - Q_A - Q_B \tag{6.2.10}$$

— остаточная сумма квадратов отклонений. Выражения Q_A и Q_B часто называют суммами квадратов разностей по строкам и по столбцам табл. 6.2.1. Распределения случайных величин Q, Q_A, Q_B и Q_R имеют соответственно $(hq-1)$, $(h-1)$, $(q-1)$ и $(h-1)(q-1)$ степеней свободы. При этом аналогично однофакторному анализу доказывается, что случайные величины Q_A, Q_B и Q_R между собой взаимно независимы.

Можно показать, что при нулевых гипотезах H_A и H_B ($a_i = 0$, $b_j = 0$ для всех i и j) средние значения случайных величин Q, Q_A, Q_B и Q_R равны дисперсии σ^2. Поэтому при нулевых гипотезах случайные величины Q, Q_A, Q_B и Q_R могут быть использованы для несмещенной оценки дисперсии σ^2:

$$s^2 = \frac{Q}{hq-1} = \frac{1}{hq-1}\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{..})^2; \tag{6.2.11}$$

$$s_A^2 = \frac{Q_A}{h-1} = \frac{q}{h-1}\sum_{i=1}^{h}(y_{i.} - y_{..})^2; \qquad (6.2.12)$$

$$s_B^2 = \frac{Q_B}{q-1} = \frac{h}{q-1}\sum_{j=1}^{q}(y_{.j} - y_{..})^2; \qquad (6.2.13)$$

$$s_R^2 = \frac{Q_R}{(h-1)(q-1)} = \frac{1}{(h-1)(q-1)}\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{i.} - y_{.j} + y_{..})^2. \qquad (6.2.14)$$

Оценки дисперсии σ^2 по формулам (6.2.11)–(6.2.14) остаются в силе и в том случае, когда нулевые гипотезы о равенстве средних значений по одному или обоим факторам не подтверждаются, лишь бы истинное значение дисперсии во всех наблюдениях было одинаковым. Однако в этом случае только оценка (6.2.14) является несмещенной оценкой дисперсии σ^2, а остальные оценки дисперсии аналогично (6.1.27) являются смещенными.

В связи с изложенным выше при нулевой гипотезе величины Q/σ^2, Q_A/σ^2, Q_B/σ^2 и Q_R/σ^2 распределены по закону хи-квадрат с $\nu = (hq-1)$, $\nu_A = (h-1)$, $\nu_B = (q-1)$ и $\nu_R = (h-1)(q-1)$ степенями свободы соответственно. Поэтому для проверки нулевых гипотез можно использовать статистику дисперсионного отношения и соответствующий F-критерий проверки гипотез.

Для проверки нулевых гипотез H_A и H_B (проверки значимости расхождений в средних значениях по строкам и столбцам) вычисляются статистики дисперсионного анализа

$$F_A = \frac{s_A^2}{s_R^2} = \frac{Q_A/(h-1)}{Q_R/[(h-1)(q-1)]}; \qquad (6.2.15)$$

$$F_B = \frac{s_B^2}{s_R^2} = \frac{Q_B/(q-1)}{Q_R/[(h-1)(q-1)]}. \qquad (6.2.16)$$

Схема вычислений дисперсионного анализа для двух факторов с однократными измерениями в ячейке приведена в табл. 6.2.2.

Для заданного уровня значимости α выборочные статистики F_A^* и F_B^* сравниваются с критическими значениями $F_{\text{кр.}A} = F_{\alpha;\nu_A,\nu_R}$ и $F_{\text{кр.}B} = F_{\alpha;\nu_B,\nu_R}$, определяемыми по таблицам F-распределения (табл. 4 Приложения 2) для соответствующих степеней свободы числителя и знаменателя выражений (6.2.15) и (6.2.16). Если $F_A^* > F_{\text{кр.}A}$, то нулевая гипотеза H_A о независимости среднего значения от уровней фактора A отвергается. Аналогичный вывод делается относительно влияния уровней фактора B, если $F_B^* > F_{\text{кр.}B}$.

Пример 6.2.1. С уровнем значимости $\alpha = 0{,}05$ проверим гипотезы о влиянии типа антенны (фактор A) и ее расположения в здании (фактор B) на величину радиосигнала, принимаемого от некоторого источника радиоизлучения. Значения

Таблица 6.2.2

Схема вычисления двухфакторного дисперсионного анализа с однократными измерениями

Изменчивость (источник изменчивости)	Сумма квадратов	Число степеней свободы n	Оценка дисперсии
Общая	$Q = \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{..})^2$	$hq-1$	$s^2 = \frac{Q}{hq-1}$
Между уровнями фактора A	$Q_A = q\sum_{i=1}^{h}(y_{i.} - y_{..})^2$	$h-1$	$s_A^2 = \frac{Q_A}{h-1}$
Между уровнями фактора B	$Q_B = h\sum_{j=1}^{q}(y_{.j} - y_{..})^2$	$q-1$	$s_B^2 = \frac{Q_B}{q-1}$
Остаточная (внутри уровней)	$Q_R = \sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij} - y_{i.} - y_{.j} + y_{..})^2$	$(h-1)(q-1)$	$s_R^2 = \frac{Q_R}{(h-1)(q-1)}$

Таблица 6.2.3

Результаты опытов

		Фактор B (j)				Среднее по строкам $y_{i.}$
		1	2	3	4	
Фактор A (i)	1	7	9	21	15	13,00
	2	3	14	19	10	11,50
	3	8	10	20	17	13,75
Среднее по столбцам $y_{.j}$		6	11	20	14	$y_{..} = 12{,}75$

Таблица 6.2.4

Таблица вычислений

Источник изменчивости	Сумма квадратов	Число степеней свободы	Оценка дисперсии	Выборочная статистика
Общая	$Q = 364{,}3$	$\nu = 11$	$s^2 = 33{,}12$	
Уровни фактора A	$Q_A = 10{,}5$	$\nu_A = 2$	$s_A^2 = 5{,}25$	$F_A^* = 0{,}69$
Уровни фактора B	$Q_B = 308{,}3$	$\nu_B = 3$	$s_B^2 = 102{,}7$	$F_B^* = 13{,}55$
Внутри уровней	$Q_R = 45{,}5$	$\nu_R = 6$	$s_R^2 = 7{,}58$	

радиосигналов в условных единицах для ряда сочетаний уровней факторов A и B даны в табл. 6.2.3. Там же приведены вычисленные выборочные средние по строкам $y_{i.}$ и столбцам $y_{.j}$, а также общее выборочное среднее $y_{..}$.

В табл. 6.2.4 сведены вычисленные значения суммарных квадратов отклонений выборочных средних по каждому из факторов относительно общего выборочного среднего, число степеней свободы (в рассматриваемом примере $h = 3$, $q = 4$), оценки дисперсий и выборочные статистики F-критерия для каждого фактора.

Для заданного уровня значимости $\alpha = 0{,}05$ критические значения дисперсионных отношений, определяемые по табл. 4 Приложения 2, равны: для фактора A — $F_{0{,}05;2,6} = 5{,}14$; для фактора B — $F_{0{,}05;3,6} = 4{,}76$. Сравнивая критические значения статистик с полученными выборочными статистиками, имеем $F_A^* = 0{,}69 < F_{0{,}05;2,6} = 5{,}14$; $F_B^* = 13{,}55 > F_{0{,}05;3,6} = 4{,}76$. Следовательно, с уровнем значимости $\alpha = 0{,}05$ гипотеза H_A об отсутствии существенного влияния рассматриваемых

типов антенн на величину радиосигнала не отвергается. В то же время гипотеза H_B об отсутствии влияния расположения антенны в помещении на величину принимаемого сигнала с тем же уровнем значимости $\alpha = 0{,}05$ не подтверждается по данным проведенного эксперимента.

6.2.2. Двухфакторный анализ с многократными наблюдениями в «ячейке» («клетке»)

Задача двухфакторного анализа несколько усложняется, если количество выборочных значений g для каждой пары уровней (i, j) не менее двух. В подобном случае имеется возможность оценить эффект взаимодействия между уровнями исследуемых факторов A и B. *Под взаимодействием факторов понимается ситуация, когда эффект влияния одного из факторов зависит от значения другого фактора.*

Итак, пусть факторы A и B имеют уровни h и q соответственно, а количество выборочных значений в каждой ячейке (клетке) факторов одинаково и равно g. При сделанных предположениях экспериментальные данные можно представить в виде табл. 6.2.5.

Таблица 6.2.5

Таблица наблюдений для двух факторов

		Уровень фактора B (j)				Среднее по строкам $y_{i..}$
		1	2	...	q	
Уровень фактора A (i)	1	$y_{111}, \ldots, y_{11g}$	$y_{121}, \ldots, y_{12g}$	...	$y_{1q1}, \ldots, y_{1qg}$	$y_{1..}$
	2	$y_{211}, \ldots, y_{21g}$	$y_{221}, \ldots, y_{22g}$	...	$y_{2q1}, \ldots, y_{2qg}$	$y_{2..}$
	⋮	⋮	⋮	⋮	⋮	⋮
	h	$y_{h11}, \ldots, y_{h1g}$	$y_{h21}, \ldots, y_{h2g}$	...	$y_{hq1}, \ldots, y_{hqg}$	$y_{h..}$
Среднее по столбцам $y_{.j.}$		$y_{.1.}$	$y_{.2.}$	...	$y_{.q.}$	$y_{...}$

Линейная математическая модель эксперимента может быть представлена в виде

$$y_{ijk} = m + a_i + b_j + (ab)_{ij} + \xi_{ijk}, \quad i = 1, \ldots, h;\ j = 1, \ldots, q;\ k = 1, \ldots, g. \tag{6.2.17}$$

Здесь дополнительно к математической модели (6.2.1) введена величина $(ab)_{ij}$, характеризующая эффект взаимодействия уровней факторов на общее среднее значение. С учетом принятого условия о равенстве объемов g выборок в ячейках для членов линейной математической модели (6.2.17) a, b и (ab) должны выполняться соотношения

$$\sum_{i=1}^{h} a_i = 0; \quad \sum_{j=1}^{q} b_j = 0; \quad \sum_{i=1}^{h} (ab)_{ij} = \sum_{j=1}^{q} (ab)_{ij} = 0. \tag{6.2.18}$$

Тогда оценки параметров линейной модели (6.2.17), полученные с использованием метода наименьших квадратов, можно представить в виде

$$\begin{aligned} &m^* = y_{...}; \quad a_i^* = y_{i..} - y_{...}; \quad b_j^* = y_{.j.} - y_{...}; \\ &(ab)_{ij}^* = y_{ij.} - (m^* + a_i^* + b_j^*) = y_{ij.} - y_{i..} - y_{.j.} + y_{...}, \end{aligned} \tag{6.2.19}$$

где $y_{i..}$ и $y_{.j.}$ — выборочные средние по факторам A и B (по строкам и столбцам); $y_{...}$ — общее выборочное среднее; $y_{ij.}$ — выборочное среднее по ячейкам (клеткам), соответствующим парам уровней факторов (i, j). Они определяются формулами

$$\begin{aligned} &y_{i..} = \frac{1}{qg} \sum_{j=1}^{q} \sum_{k=1}^{g} y_{ijk}; \quad y_{.j.} = \frac{1}{hg} \sum_{i=1}^{h} \sum_{k=1}^{g} y_{ijk}; \\ &y_{...} = \frac{1}{hqg} \sum_{i=1}^{h} \sum_{j=1}^{q} \sum_{k=1}^{g} y_{ijk}; \quad y_{ij.} = \frac{1}{g} \sum_{k=1}^{g} y_{ijk}. \end{aligned} \tag{6.2.20}$$

Средние значения (математические ожидания) выборочных средних (6.2.20) равны

$$\begin{aligned} &\mathbf{M}\{y_{i..}\} \equiv m_i = m + a_i; \quad \mathbf{M}\{y_{.j.}\} \equiv m_j = m + b_j; \\ &\mathbf{M}\{y_{...}\} = m; \quad \mathbf{M}\{y_{ij.}\} \equiv m_{ij} = m + a_i + b_j + (ab)_{ij}. \end{aligned} \tag{6.2.21}$$

С учетом этих выражений средние значения оценок параметров (6.2.19) математической модели эксперимента совпадают с их истинными значениями

$$\mathbf{M}\{m^*\} = m; \quad \mathbf{M}\{a_i^*\} = a_i; \quad \mathbf{M}\{b_j^*\} = b_j; \quad \mathbf{M}\{(ab)_{ij}^*\} = (ab)_{ij}, \tag{6.2.22}$$

т.е. оценки параметров рассматриваемой модели эксперимента являются несмещенными оценками.

Рассмотрим некоторую гипотетическую ситуацию, при которой уровни факторов линейно влияют на общее среднее значение, т.е. нулевая гипотеза неверна. Рассмотрим зависимость величины m_{ij} (6.2.21) от уровней фактора A, представляя уровни фактора B в качестве параметра.

Если факторы не взаимодействуют $(ab)_{ij} = 0$, то из (6.2.21) видно, что для разных значений уровня фактора B прямые $m_{ij} = m + a_i + b_j$ будут параллельны. Графически это изображено на рис. 6.2.1,*а* для двух уровней фактора B. Если же факторы взаимодействуют $(ab)_{ij} \neq 0$, то согласно (6.2.21) эти прямые не будут параллельными. Примерный (возможный) вид прямых для двух уровней фактора B графически изображен на рис. 6.2.1,*б*.

Сумму квадратов отклонений всех $n = hqg$ наблюдений от их общего выборочного среднего можно представить в виде суммы взаимно

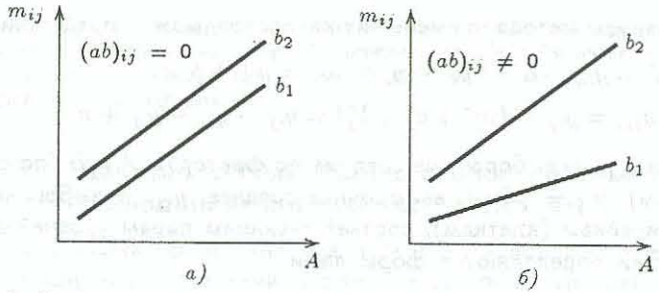


Рис. 6.2.1. Графическое представление влияния на среднее значение измеряемой случайной величины m_{ij} уровней факторов

независимых случайных величин

$$Q = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk} - y_{...})^2 = Q_A + Q_B + Q_{AB} + Q_R, \qquad (6.2.23)$$

где составляющие определяются выражениями

$$Q = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk} - y_{...})^2 = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk})^2 - hqg(y_{...})^2;$$

$$Q_A = qg\sum_{i=1}^{h}(y_{i..} - y_{...})^2 = qg\left[\sum_{i=1}^{h}(y_{i..})^2 - h(y_{...})^2\right];$$

$$Q_B = hg\sum_{j=1}^{q}(y_{.j.} - y_{...})^2 = hg\left[\sum_{j=1}^{q}(y_{.j.})^2 - q(y_{...})^2\right];$$

$$Q_{AB} = g\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2 =$$

$$= g\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij.})^2 - Q_A - Q_B - hqg(y_{...})^2;$$

$$Q_R = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk} - y_{ij.})^2 = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk})^2 - g\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij.})^2.$$

Случайные составляющие Q_A, Q_B и Q_R, как и выше, представляют собой взвешенные суммы квадратов отклонений выборочных средних по уровням факторов A и B от общего выборочного среднего и сумму квадратов отклонений каждой опытной величины от соответствующего выборочного среднего в каждой из ячеек. Случайная величина Q_{AB} является взвешенной суммой отклонений выборочных средних по ячейкам от общего выборочного среднего всей совокупности опытных данных. Она

Таблица 6.2.6

Схема вычисления двухфакторного дисперсионного анализа с многократными измерениями

Источник изменчивости	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Общая	$Q = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk} - y_{...})^2$	$hqg - 1$	$s^2 = \frac{Q}{hqg-1}$
Между уровнями фактора A	$Q_A = qg\sum_{i=1}^{h}(y_{i..} - y_{...})^2$	$h - 1$	$s_A^2 = \frac{Q_A}{h-1}$
Между уровнями фактора B	$Q_B = hg\sum_{j=1}^{q}(y_{.j.} - y_{...})^2$	$q - 1$	$s_B^2 = \frac{Q_B}{q-1}$
Взаимодействие факторов $A \times B$	$Q_{AB} = g\sum_{i=1}^{h}\sum_{j=1}^{q}(y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2$	$(h-1)(q-1)$	$s_{AB}^2 = \frac{Q_{AB}}{(h-1)(q-1)}$
Остаточная (внутри уровней)	$Q_R = \sum_{i=1}^{h}\sum_{j=1}^{q}\sum_{k=1}^{g}(y_{ijk} - y_{ij.})^2$	$hq(g-1)$	$s_R^2 = \frac{Q_R}{hq(\gamma-1)}$

характеризует влияние выборочных средних по парам уровней (i, j) факторов A и B на общее выборочное среднее экспериментальных данных. Распределения введенных величин имеют степени свободы $\nu = hqg - 1$, $\nu_A = h - 1$, $\nu_B = q - 1$, $\nu_{AB} = (h-1)(q-1)$ и $\nu_R = hq(g-1)$. Величины

$$s_A^2 = \frac{Q_A}{h-1}; \quad s_B^2 = \frac{Q_B}{q-1}; \quad s_{AB}^2 = \frac{Q_{AB}}{(h-1)(q-1)}; \quad s_R^2 = \frac{Q_R}{hq(g-1)}$$

могут рассматриваться в качестве оценок дисперсии σ^2 случайной составляющей ξ_{ijk}. При этом если величина s_R^2 вне зависимости от типа гипотезы представляет собой несмещенную оценку дисперсии σ^2, то остальные величины дают несмещенную оценку дисперсии σ^2 только в случае нулевых гипотез, для которой справедливы соотношения $a_i = b_j = (ab)_{ij} = 0$ $(i = 1, \ldots, h;\ j = 1, .., q)$.

Результаты двухфакторного дисперсионного анализа с многократными наблюдениями можно представить в виде табл. 6.2.6.

Так как для нулевых гипотез случайные величины Q_A/σ^2, Q_B/σ^2, Q_{AB}/σ^2 и Q_R/σ^2 распределены по закону хи-квадрат, для проверки нулевых гипотез применяется F-критерий. Статистика дисперсионного отношения для проверки гипотез H_A, H_B и H_{AB} определяется выражениями

$$F_A = \frac{s_A^2}{s_R^2}; \quad F_B = \frac{s_B^2}{s_R^2}; \quad F_{AB} = \frac{s_{AB}^2}{s_R^2}. \qquad (6.2.24)$$

Гипотезы H_A, H_B или H_{AB} отвергаются, если при заданном уровне

значимости α соответственно окажется, что выборочные статистики $F_A^* > F_{\text{кр}.A}$, $F_B^* > F_{\text{кр}.B}$ и $F_{AB}^* > F_{\text{кр}.AB}$, где критические значения статистик $F_{\text{кр}.A} = F_{\alpha;\nu_A,\nu_R}$, $F_{\text{кр}.B} = F_{\alpha;\nu_B,\nu_R}$ и $F_{AB} = F_{\alpha;\nu_{AB},\nu_R}$ находятся по таблицам F-распределения.

Пример 6.2.2. Для иллюстрации двухфакторного дисперсионного анализа рассмотрим некоторый гипотетический пример: с уровнем значимости $\alpha = 0{,}01$ выясним влияние на наблюдаемые величины фактора A с заданными четырьмя нормированными уровнями ($h = 4$; -3, -1, $+1$, $+3$) и фактора B с двумя уровнями ($q = 2$). Число измерений на каждом сочетании уровней $g = 6$. Результаты наблюдений сведены в табл. 6.2.7. В табл. 6.2.8 приведены вычисленные данные, необходимые для дисперсионного анализа.

Для заданного уровня значимости $\alpha = 0{,}01$ критические значения выборочной статистики дисперсионного отношения, определяемые по табл. 4 Приложения 2, соответственно равны: для фактора A — $F_{\text{кр}.A} = F_{0,01;3,40} = 4{,}31$; для фактора B — $F_{\text{кр}.B} = F_{0,01;1,40} = 7{,}31$; для взаимодействия факторов A и B — $F_{\text{кр}.AB} = F_{0,01;3,40} = 4{,}31$.

Поскольку выборочные F-статистики F_A^*, F_B^* и F_{AB}^*, приведенные в табл. 6.2.8, превышают соответствующие критические значения, то нулевые гипотезы о том, что факторы A и B и их взаимодействие не влияют на выходной эффект (на результаты наблюдения), отвергаются с уровнем значимости $\alpha = 0{,}01$.

На рис. 6.2.2 приведены графики линейной зависимости (уравнений линейных регрессий) $y_{ij.}^* = b_{0j}^* + b_{1j}^* A_i$, являющиеся оценками средних значений величин $y_{ij.}$ от нормированных уровней A_i фактора A для двух уровней фактора B. Оценки коэффициентов линейной регрессии b_{0j} и b_{1j}, вычисленные согласно выражениям (3.5.43), соответственно равны $b_{01}^* = 9{,}5$; $b_{02}^* = 20{,}9$; $b_{11}^* = 1{,}55$; $b_{12}^* = -3{,}03$.

Таблица 6.2.7

Результаты наблюдения

		Фактор B (j)		Среднее по строкам $y_{i..}$
		B_1	B_2	
Фактор A (i)	$A_1 = -3$	2 5 6 7 4 6 $y_{11.} = 5{,}0$	32 34 24 28 32 30 $y_{12.} = 30$	17,5
	$A_2 = -1$	9 8 5 6 7 10 $y_{21.} = 7{,}5$	19 20 31 28 22 24 $y_{22.} = 24$	15,8
	$A_3 = 1$	10 13 11 12 10 13 $y_{31.} = 11{,}5$	21 14 19 18 17 16 $y_{32.} = 17{,}5$	14,5
	$A_4 = 3$	13 17 11 12 15 16 $y_{41.} = 14$	9 12 14 15 10 12 $y_{42.} = 12$	13,0
Среднее по столбцам $y_{.j.}$		9,5	20,9	$y_{...} = 15{,}2$

Таблица 6.2.8

Таблица вычислений

Источник изменчивости	Сумма квадратов	Число степеней свободы	Оценка дисперсии	Выборочная статистка
Общая	$Q = 3225$	$\nu = 47$	$s^2 = 69$	
Между уровнями фактора A	$Q_A = 132$	$\nu_A = 3$	$s_A^2 = 44$	$F_A^* = 5{,}9$
Между уровнями фактора B	$Q_B = 1560$	$\nu_B = 1$	$s_B^2 = 1560$	$F_B^* = 208$
Взаимодействие между уровнями факторов A и B	$Q_{AB} = 1232$	$\nu_{AB} = 3$	$s_{AB}^2 = 411$	$F_{AB}^* = 55$
Остаточная (внутри уровней)	$Q_R = 301$	$\nu_R = 40$	$s_R^2 = 7{,}5$	

Из этих графиков видно, что, во-первых, имеется зависимость усредненных результатов наблюдений от значений уровней факторов A и B; во-вторых, кривые линейной регрессии явно не параллельны, что свидетельствует о взаимном влиянии факторов на результат наблюдения.

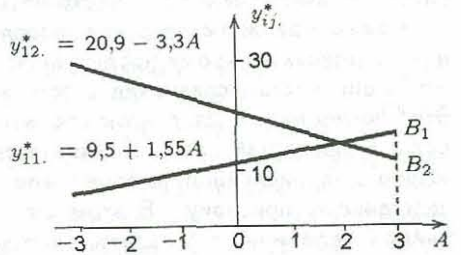


Рис. 6.2.2. Графики зависимостей результатов наблюдения от двух факторов A и B

С увеличением числа факторов принцип дисперсионного анализа остается неизменным, усложняется лишь процедура разделения суммарного квадрата отклонения выборочных средних от общего выборочного среднего на составляющие. Эти составляющие должны учитывать члены, обусловленные как самими факторами, так и парными, тройными и т.п. взаимодействиями между факторами. Если число исследуемых факторов равно k, то общее количество составляющих с учетом члена Q_R, обусловленного «ошибкой» (остаточной суммой),

$$N = 1 + k + C_k^2 + C_k^3 + \ldots + C_k^k = \sum_{m=0}^{k} C_k^m = 2^k. \qquad (6.2.25)$$

Например, количество членов разложения общей суммы квадратов для четырех и пяти факторов равно $N(4) = 16$ и $N(5) = 32$ соответственно.

Метод дисперсионного анализа разработан применительно к нормально распределенным случайным величинам. В то же время, если количество элементов выборки на каждом из уровней $n \geqslant 4$, то F-критерий можно применять даже в том случае, когда неизвестно, подчиняются результаты наблюдений нормальному распределению или нет [13]. При этом ошибка такого приближения будет вполне приемлема для прикладных задач, если у одинаково распределенных результатов наблюдений дисперсия имела конечное значение. Это обусловлено тем, что средние значения по факторам при суммировании случайных величин согласно центральной предельной теореме приближенно распределены по нормальному закону, даже если отдельные слагаемые значительно отличаются от гауссовских. При возведении выборочных средних значений по уровням в квадрат это отличие еще более сглаживается.

6.3. Непараметрический дисперсионный анализ

Как отмечалось выше, если распределения изучаемых случайных величин неизвестны или изучаются качественные признаки, то для проверки влияния уровней факторов на средние значения выходного эффекта наиболее предпочтительными являются ранговые критерии. В таких

ситуациях наблюдения (их часто называют индивидуумами) $x_1, x_2, \ldots, x_n$ можно «ранжировать», т.е. упорядочить по степени их предпочтения, приписав каждому наблюдению x_i порядковый номер r_i, соответствующий итогам сравнения с остальными наблюдениями x_j $(i \neq j)$. Этот номер называют *рангом* соответствующего *наблюдения*. Как правило, наибольший ранг присваивается наибольшему выборочному значению для численной выборки или наиболее предпочтительному качественному признаку. В этом случае рангом i-го наблюдения (случайной величины x_i) является номер места r_i, которое занимает x_i в вариационном ряду.

Если ранжированный ряд содержит n наблюдений, то сумма рангов должна удовлетворять условию

$$\sum_{i=1}^{n} r_i = n(n+1)/2,$$

которое выполняется, если все ранги различны (отсутствуют совпадающие ранги). Однако возможны случаи, когда в результате ранжирования в ряду будут совпадающие ранги, обусловленные тем, что некоторые значения или признаки могут быть неразличимы. Подобная ситуация, например, может быть обусловлена конечной точностью измерительных устроств при переводе численных значений выборки в вариационный ряд. В таких случаях исходная ранжированная последовательность должна быть перекодирована. Суть перекодирования состоит в том, что каждому из признаков с одинаковыми рангами приписывают ранг, равный среднему арифметическому значению тех мест, которые они делят. Например, пусть исходный ранжированный ряд из 11 признаков имеет вид: 4; 6; 4; 2; 3; 5; 5; 5; 1; 3; 5. Ранг 6, находящийся на втором месте, является наибольшим, поэтому второму признаку назначается ранг 11. Ранги со значениями 5 занимают места с 7 по 10; средний ранг для них (7+8+9+10)/4=8,5. Аналогичным образом определяются новые ранги и для других признаков. В результате преобразованный ряд записывается как 5,5; 11; 5,5; 2,0; 3,5; 8,5; 8,5; 8,5; 1,0; 3,5; 8,5. При этом для преобразованного ряда также выполняется условие $n(n+1)/2 = 66$.

Таким образом, исходная статистическая информация для подобных исследований может быть задана рангами наблюдений. Поэтому статистические методы исследования, использующие наблюдения в виде последовательности рангов, называются *ранговыми методами*; статистики, являющиеся функциями только рангов, называются *ранговыми статистиками*, а критерии, основанные на таких статистиках, — *ранговыми критериями*. Выше при проверке гипотезы об однородности двух выборок (двух признаков) по критерию Вилкоксона (п. 5.5.3) уже был рассмотрен подход, основанный на ранговой статистике одной из переменной общего вариационного ряда.

Рассмотрим использование этих критериев применительно к однофакторному и двухфакторному дисперсионному анализу.

6.3.1. Однофакторный непараметрический дисперсионный анализ

Пусть, как и в п. 6.1, в результате эксперимента получено h серий выборок гауссовской случайной величины $\mathfrak{y}_1(y_{11}, y_{12}, \ldots, y_{1n_1})$, $\mathfrak{y}_2(y_{21}, y_{22}, \ldots, y_{2n_2})$, $\ldots$, $\mathfrak{y}_h(y_{h1}, y_{h2}, \ldots, y_{hn_h})$, соответствующих h уровням (классам) некоторого изучаемого фактора A. Объем каждой серии выборок равен n_i, $i = 1, 2, \ldots, h$. При этом общее число выборочных значений $n = \sum n_i$. Эти наблюдения записываются в таблицу вида табл. 6.1.1 (без последних двух столбцов). Для проверки нулевой гипотезы H_0 об однородности выборок (выборки взяты из одной генеральной совокупности) полагается, что все h выборок можно рассматривать как одну (объединенную) выборку из общей генеральной совокупности. Проранжируем все наблюдения вместе (например, от меньшего значения к большему значению). Пусть r_{ij} — означает ранг величины y_{ij} в этой совместной ранжированной последовательности. В результате можно составить таблицу, содержащую вместо численных значений их ранги (табл. 6.3.1).

Таблица 6.3.1

Результаты ранжирования

Уровни фактора i	Результаты ранжироваания j							Сумма рангов R_i	Средний ранг $R_{i\cdot}$
1	r_{11}	r_{12}	r_{13}	$\ldots$	r_{1j}	$\ldots$	r_{1n_1}	R_1	$R_{1\cdot}$
2	r_{21}	r_{22}	r_{23}	$\ldots$	r_{2j}	$\ldots$	r_{2n_2}	R_2	$R_{2\cdot}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
i	r_{i1}	r_{i2}	r_{i3}	$\ldots$	y_{ij}	$\ldots$	r_{in_i}	R_i	$R_{i\cdot}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
h	r_{h1}	r_{h2}	r_{h3}	$\ldots$	r_{hj}	$\ldots$	r_{hn_h}	R_h	$R_{h\cdot}$

В последних двух столбцах таблицы приведены суммы рангов и средний ранг для каждого уровня $i = 1, 2, \ldots, h$:

$$R_i = \sum_{j=1}^{n_i} r_{ij}; \quad R_{i\cdot} = \frac{R_i}{n_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{i,j}. \tag{6.3.1}$$

Если между строками отсутствует систематическое различие, средние ранги по строкам не должны существенно отличаться друг от друга и от среднего ранга всей совокупности рангов

$$R_{\cdot\cdot} = \frac{n+1}{2}. \tag{6.3.2}$$

Рассмотрим обобщенную сумму квадратов отклонения средних рангов по строкам от общего среднего с учетом числа наблюдений n_i для разных уровней фактора

$$B = \sum_{i=1}^{h} n_i (R_{i\cdot} - R_{\cdot\cdot})^2.$$

Эта величина зависит от дисперсии рангов. Поэтому для того чтобы статистика для проверки гипотезы об однородности выборок не зависела от дисперсии рангов, ее нормируют на математическое ожидание выборочной дисперсии рангов:

$$D = \frac{1}{n(n-1)} \sum_{i=1}^{h} \sum_{j=1}^{n_i} \mathrm{M}\left[\left(R_{ij} - \frac{n+1}{2}\right)^2\right] =$$

$$= \frac{1}{n-1} \frac{(n^2-1)}{12} \sum_{i=1}^{h} n_i = \frac{n(n+1)}{12}.$$

В результате статистика для проверки гипотезы об однородности выборок принимает вид

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^{h} n_i (R_{i\cdot} - R_{\cdot\cdot})^2 = \frac{12}{n(n+1)} \sum_{i=1}^{h} \frac{R_i^2}{n_i} - 3(n+1). \quad (6.3.3)$$

Эту статистику называют *статистикой Крускала–Уоллиса* [62, 68]. В [51, 68] приведены таблицы критических значений этой статистики для трех выборок, при этом объем каждой из них не более пяти. Правило принятия решения на основе данной статистики называется *критеритем Крускала–Уоллиса*.

Для нулевой гипотезы при $n_i \gg 1$ $(n_i \to \infty)$, $i = 1, 2, \ldots, h$, статистика (6.3.3) асимптотически имеет хи-квадрат распределение с $\nu = h-1$ степенями свободы. Для конечных значений объемов выборок представление распределения статистики (6.3.3) в виде хи-квадрат распределения является приближенным. При этом надежность принятия нулевой гипотезы, увеличивается, так как действительный уровень значимости несколько меньше величины, получаемой на основе приближенного распределения статистики в виде хи-квадрат распределения.

Если в исходной табл. 6.3.1 имеются совпадающие ранги, то необходимо использовать средние ранги. В результате уточненная статистика примет вид [62, 68 и др.]

$$\tilde{Q} = \frac{Q}{1 - \sum_{k=1}^{g} T_k/(n^3 - n)}. \quad (6.3.4)$$

Здесь g — число групп совпадающих рангов, $T_k = (t_k^3 - t_k)$; t_k — объем группы совпадающих рангов с номером k. Если нет совпадающих рангов, то $t_k = 1$, $T_k = 0$ для всех уровней $i = 1, 2, \ldots, h$. При этом знаменатель (6.3.4) равен единице, в результате чего (6.3.4) переходит в (6.3.3).

Пример 6.3.1. В табл. 6.3.2 приведены численные данные для четырех уровней $(h = 4)$ фактора и одинакового их числа на каждом уровне $n_i = 5$ $(i = 1, 2, 3, 4)$. Необходимо с уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о том, что выборки, соответствующие уровням таблицы, однородны. Распределение выборок неизвестно.

Для решения задачи целесообразно применить ранговый критерий. В табл. 6.3.3 численные значения данных заменены их рангами. В связи с наличием совпадающих рангов табл. 6.3.3 следует перекодировать в табл. 6.3.4 с введением средних рангов.

Таблица 6.3.2

Уровни фактора	Результаты опыта				
1	106	108	98	106	106
2	88	94	116	93	112
3	110	104	105	114	88
4	105	109	88	109	94

Таблица 6.3.3

Уровни фактора	Результаты ранжирования				
1	7	8	4	7	7
2	1	3	13	2	11
3	10	5	6	12	1
4	6	9	1	9	3

Таблица 6.3.4

Уровни фактора	Результаты преобразованного ранжироваания					Сумма рангов	Средний ранг $R_{i\cdot}$
1	12	14	7	12	12	57	11,4
2	2	5,5	20	4	18	49,5	9,9
3	17	8	9,5	19	2	55,5	11,1
4	9,5	15,5	2	15,5	5,5	48	9,6

В двух крайних столбцах приведены суммы рангов и средние ранги по строкам (по каждому уровню фактора). Общее число наблюдений $n = 20$.

Подставляя в формулу (6.3.3) средние ранги из табл. 6.3.4, получим выборочное значение статистики $Q^* \approx 0{,}33$. Число групп совпадающих рангов $g = 5$, причем для каждой из групп имеем $T_1 = T_2 = 3^3 - 3 = 24$, $T_3 = T_4 = T_5 = 2^3 - 2 = 6$. Так как $\sum_{k=1}^{g=5} T_k/(n^3 - n) = 66/7980 \approx 0{,}008 \ll 1$, то значение выборочной статистики $\tilde{Q}^*$ не отличается от статистики Q^*. Число степеней свободы статистики (6.3.3) $\nu = h - 1 = 3$. По таблицам критических значений хи-квадрат распределения для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $\nu = 3$ находим критическое значение статистики $Q_{0{,}05;3} = 7{,}81$. Так как $Q_{0{,}05;3} = 7{,}81$ значительно больше $Q^* \approx 0{,}33$, нулевая гипотеза о том, что приведенные в табл. 6.3.2 данные относятся к одной генеральной совокупности, не отвергается.

Этот пример наглядно показывает, что при $n \gg 1$ для прикладных задач вполне допустимо применять статистику (6.3.3).

6.3.2. Двухфакторный непараметрический дисперсионный анализ

С увеличением числа факторов, т.е. при двух, трех и более факторах, процедура непараметрического дисперсионного анализа принципиально не меняется, хотя вычисления значительно усложняются по мере увеличения числа факторов и числа уровней факторов.

Рассмотрим сущность двухфакторного анализа. Пусть изучается влияние на среднее значение анализируемых случайных величин двух факторов A и B. Будем полагать, что по фактору A все наблюдения делятся на m уровней $(A_1, \ldots, A_m)$, а по фактору B — на n уровней $(B_1, \ldots, B_n)$, т.е. все экспериментальные данные разбиваются на mn клеток (ячеек).

Таблица 6.3.5

Таблица рангов факторов A и B

		Уровень i фактора B			
		1	2	...	n
Уровень j фактора A	1	r_{11}	r_{12}	...	r_{1n}
	2	r_{21}	r_{22}	...	r_{2n}
	⋮	⋮	⋮	⋮	⋮
	m	r_{m1}	r_{m2}	...	y_{mn}
Сумма рангов по столбцам R_i		R_1	R_2	...	R_n
Средний ранг по столбцам $R_{.i}$		$R_{.1}$	$R_{.2}$	...	$R_{.n}$

При этом будем полагать, что в каждой ячейке осуществляется одно наблюдение. Тогда наблюдение y_{ij} является выборочным значением, попадающим в ячейку, соответствующую уровню фактора A_j, $j = 1, \ldots, m$, и уровню фактора B_i, $i = 1, \ldots, n$. Наблюдения y_{ji} могут быть представлены в виде табл. 6.2.1, в которой i и j следует поменять местами и полагать $h = m$, $q = n$. При применении непараметрического двухфакторного анализа численные значения y_{ji} переводятся в ранги r_{ji}, причем ранжирование осуществляется по каждой строке. В результате табл. 6.2.1 преобразуется в таблицу рангов (табл. 6.3.5), причем в каждой строке значения рангов r_{ji} изменяются от 1 до n. Если в качестве исходных данных используются данные табл. 6.2.5, то ранжируются выборочные средние по каждой клетке.

В качестве нулевой гипотезы H_0 рассмотрим гипотезу о том, что уровни (признаки, обработки) фактора B не влияют на значения отклика по каждому из уровней фактора A, которые часто называют блоками.

В последних двух строках таблицы приведены сумма рангов по столбцам R_i и средний ранг по столбцам $R_{.i}$:

$$R_i = \sum_{j=1}^{m} r_{ji}; \quad R_{.i} = \frac{1}{m}\sum_{j=1}^{m} r_{ji}. \tag{6.3.5}$$

При этом средний ранг по каждой строке и сумма рангов всей совокупности рангов соответственно равны

$$R_{..} = \frac{n+1}{2}; \quad R = mR_{..} = \frac{m(n+1)}{2}. \tag{6.3.6}$$

Если справедлива нулевая гипотеза, то каждая строка рангов будет представлять случайную перестановку чисел от 1 до n, при этом таких перестановок $n!$ и все они равновероятны. Подобная задача рассматривается ниже в п. 7.2.2 для проверки гипотезы о коэффициенте согласованности (конкордации). В качестве статистики используется статистика

$$V = \frac{1}{D}\sum_{i=1}^{n}(R_i - R)^2 = \frac{12}{mn(n+1)}\sum_{i=1}^{n}\left[\sum_{j=1}^{m} r_{ji} - \frac{m(n+1)}{2}\right]^2 =$$

$$= \frac{12}{mn(n+1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} r_{ji}\right)^2 - 3m(n+1), \tag{6.3.7}$$

которую называют *статистикой Фридмана* [17, 51, 68]. Правило принятия решения на основе этой статистики называют *критерием Фридмана*. Для этой статистики существуют таблицы [2, 8, 51, 68 и др.], позволяющие по заданному уровню значимости α и малому числу уровней m и n определить критическое значение статистики для проверки нулевой гипотезы о том, что ранги, соответствующие уровням фактора B, независимы. При больших значениях m и n статистика (6.3.7) асимптотически подчиняется распределению хи-квадрат с $\nu = n - 1$ степенями свободы.

В прикладных задачах, как правило, замена точного распределения (6.3.7) его приближенным распределением в виде хи-квадрат вполне допустима и для небольших значений числа уровней m и n, увеличивая при этом, как отмечалось выше, надежность принятия нулевой гипотезы. Поэтому для проверки нулевой гипотезы H_0 о независимости рангов с односторонним уровнем значимости α, как и выше, по таблицам хи-квадрат распределения для степеней свободы $\nu = (n-1)$ находится верхнее критическое значение статистики $\chi^2_{\alpha;\nu}$. Это значение статистики сравнивается со значением выборочной статистики V^*, вычисленной согласно (6.3.7). Для этого в (6.3.7) подставляются выборочные значения R_i или r_{ji}, полученные на основе экспериментальных данных о рангах, приведенных в табл. 6.3.5. Гипотеза случайности рангов отклоняется, если для выборочной статистики и одностороннего уровня значимости α выполняется условие $V^* \geqslant \chi^2_{\alpha;\nu}$.

Если ранги каких-либо наблюдений (параметров) совпадают, то для этих наблюдений (параметров) назначаются ранги, равные среднему арифметическому из рангов смежных членов ряда. Формулы, учитывающие число совпадений, приведены ниже в п. 7.2.

Пример 6.3.2. В табл. 6.3.6 приведены ранги для пяти уровней ($n = 5$) фактора B и четырех уровней $m = 4$ фактора A. Необходимо с уровнем значимости $\alpha = 0{,}05$ проверить гипотезу о том, что ранги, соответствующие уровням фактора B таблицы, случайны.

Для решения задачи целесообразно применить ранговый критерий (6.3.7). В нижней строке табл. 6.3.6 приведены суммы рангов по столбцам. При этом сумма рангов всей совокупности и число степеней свободы для приведенных данных равны соответственно $R = 12$ и $\nu = n - 1 = 4$. Подставляя в формулу (6.3.7) необходимые величины, получаем значение выборочной статистики

$$V^* = \frac{12}{4 \cdot 5 \cdot 6}[(-3)^2 + (-2)^2 + 2^2 + 4^2 + (-1)^2] = 3{,}4.$$

Согласно табл. 2 Приложения 2 приближенное критическое значение статистики для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $\nu = 4$ равно $\chi^2_{0{,}05;4} = 9{,}49$. Так как величина выборочной статистики меньше критического значения, можно утверждать, что нулевая гипотеза о том, что уровни фактора B не влияют на измеряемые величины, не противоречит с уровнем значимости $\alpha = 0{,}05$.

К подобному выводу можно прийти, если учесть совпадающие ранги, применив формулы (7.3.37), (7.3.38) п. 7.2.

Таблица 6.3.6

Уровень j фактора A	Уровень i фактора B				
	1	2	3	4	5
1	2	1	3	5	4
2	2	3	5	4	1
3	3	2	1	4	5
4	2	4	5	3	1
Сумма рангов R_i	9	10	14	16	11

Контрольные вопросы и задачи

6.1. Какое предположение о распределении выборочных значений лежит в основе дисперсионного (классического) анализа?

6.2. В каких случаях следует применять непараметрический (ранговый) дисперсионный анализ?

6.3. Какой ситуации соответствует нулевая гипотеза применительно к уровням фактора при однофакторном дисперсионном анализе?

6.4. Что позволяет проверить двухфакторный дисперсионный анализ с многократными наблюдениями в ячейках по сравнению с однократными наблюдениями в ячейках?

6.5. В табл. 9.4.3 (стр. 404) приведены выборочные значения гауссовской случайной величины Y в пяти точках плана (для пяти уровней). Методом дисперсионного анализа убедиться в том, что выборочные средние по уровням $y_{i\cdot}$, $i = 1...5$, существенно отличаются от общего выборочного среднего $y_{\cdot\cdot}$ с уровнем значимости $\alpha = 0{,}01$.

6.6. Получены три выборки, в каждой из которых по семь наблюдений:

$$\mathfrak{y}_1(10{,}6;\,13{,}6;\,14{,}3;\,13{,}8;\,11{,}6;\,15{,}0;\,12{,}6),$$
$$\mathfrak{y}_2(11{,}3;\,13{,}7;\,11{,}2;\,12{,}9;\,14{,}5;\,10{,}8;\,14{,}1),$$
$$\mathfrak{y}_3(14{,}8;\,11{,}5;\,12{,}0;\,10{,}3;\,12{,}5;\,11{,}4;\,13{,}5).$$

С помощью рангового дисперсионного анализа проверить гипотезу об однородности выборок с уровнем значимости $\alpha = 0{,}05$.

ГЛАВА 7

Корреляционные зависимости

Большое место в прикладной математической статистике уделяется вопросам статистической зависимости (связи) между случайными величинами и случайными событиями. Хотя дальнейшее обсуждение ведется применительно к случайным величинам, если не оговорено особо, полученные выводы будут справедливы и для случайных событий. Исторически статистические зависимости изучались, как правило, между двумя величинами, а затем их результаты обобщались на многомерную ситуацию. Поэтому для упрощения рассматриваемых вопросов ниже в основном изучаются статистические зависимости между двумя величинами. Зависимости между величинами могут быть как чисто случайными, так и функциональными с неизвестными параметрами, определяемыми в ходе эксперимента и подверженными ошибкам наблюдений и измерений.

При рассмотрении статистической зависимости между величинами следует различать взаимозависимость и просто зависимость одной величины от другой (или других) величины. Например, для реального газа между давлением и температурой при постоянном объеме существует статистическая взаимозависимость. В то же время между физиологическим состоянием человека и погодными условиями существует только статистическая зависимость физиологического состояния человека от объективных данных погодных условий (температуры и влажности воздуха, давления атмосферы и т.д.). Исследование взаимозависимости между случайными величинами осуществляется корреляционными методами, а изучение зависимости одной случайной величины от другой является предметом регрессионного анализа.

Следует особо оговорить, что статистическая зависимость (взаимозависимость) между случайными величинами, как бы она ни была значима, никогда не может установить причинной связи между этими величинами. В приведенном примере зависимости самочувствия человека от погодных условий со статистической точки зрения формально можно изучать зависимость параметров погодных условий от параметров самочувствия человека. Однако вполне понятно, что самочувствие человека ни в коем случае не влияет на погоду. Поэтому необходимо еще раз подчеркнуть, что статистическая зависимость любого вида логически не связана с причиной зависимости изучаемых явлений и событий.

В данной главе рассматривается корреляционный анализ (параметрический и ранговый), называемый также корреляцией.

7.1. Корреляционный анализ (корреляция)

Корреляция — это зависимость между случайными величинами, которые не имеют детерминированную (строгую) функциональную зависимость. Как уже отмечалось выше (п. 1.5), зависимость между событиями (случайными величинами) проявляется в том, что условная вероятность одного из событий при фиксированных значениях других событий отличается от безусловной его вероятности. Для изучения подобной зависимости применяется *корреляционный (ковариационный) анализ (корреляция, или ковариация)*, в которой входят следующие практические приемы:

- составление корреляционной таблицы типа табл. 3.3.2;
- вычисление выборочных коэффициента корреляции и корреляционного отношения;
- проверку статистических гипотез относительно значимости статистической связи.

Рассмотрим точечные и интервальные оценки корреляционного момента K_{XY} и коэффициента корреляции ρ_{XY} двух случайных величин X и Y на основе n пар выборочных значений $(x_1, y_1), \ldots, (x_n, y_n)$ в предположении, что пары независимы между собой и подчиняются двумерному нормальному закону распределения с одинаковыми характеристиками. Корреляционный момент K_{XY} и коэффициент корреляции ρ_{XY} определяются выражениями (1.5.25) и (1.5.31) соответственно.

7.1.1. Выборочный корреляционный момент

Согласно предельному соотношению (2.4.10) для попарно независимых выборочных значений x_i и y_i объема n в качестве оценки корреляционного момента K_{XY} при известных средних значениях m_X и m_Y рассматриваемых случайных величин можно взять его выборочное значение

$$R_0 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_X)(y_i - m_Y). \tag{7.1.1}$$

Нетрудно увидеть, что математическое ожидание (среднее значение) оценки корреляционного момента равно его истинному значению $\mathrm{M}\{R_0\} = K_{XY}$, т.е. для принятых условий оценка корреляционного момента (7.1.1) несмещенная. Дисперсия рассматриваемой оценки определяется как

$$D(R_0) = \mathrm{M}\{(R_0 - K_{XY})^2\} = \mathrm{M}\{R_0^2\} - K_{XY}^2 =$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{(x_i - m_X)(y_i - m_Y)(x_j - m_X)(y_j - m_Y)\} - K_{XY}^2. \tag{7.1.2}$$

В последнем выражении для вычисления смешанного центрального момента четвертого порядка учтем, что для попарно независимых выборочных значений имеют место соотношения:

$$\mathrm{M}\{(x_i - m_X)(y_j - m_Y)\} = \begin{cases} \mathrm{M}\{(x_i - m_X)(y_i - m_Y)\} = K_{XY}, & i = j, \\ 0, & i \neq j, \end{cases}$$

$$\mathrm{M}\{(x_i - m_X)(x_j - m_X)\} = \begin{cases} D_X, & i = j, \\ 0, & i \neq j, \end{cases}$$

$$\mathrm{M}\{(y_i - m_Y)(y_j - m_Y)\} = \begin{cases} D_Y, & i = j, \\ 0, & i \neq j. \end{cases}$$

Кроме того, согласно [37, 52, 59 и др.] четвертый смешанный момент гауссовских случайных величин z_i, $i = 1, 2, 3, 4$, с нулевыми средними значениями определяется как

$$\mathrm{M}\{z_1 z_2 z_3 z_4\} = \mathrm{M}\{z_1 z_2\}\mathrm{M}\{z_3 z_4\} + \mathrm{M}\{z_1 z_3\}\mathrm{M}\{z_2 z_4\} + \mathrm{M}\{z_1 z_4\}\mathrm{M}\{z_2 z_3\}. \tag{7.1.3}$$

В результате выражение для $\mathrm{M}\{R_0^2\}$ в (7.1.2) можно представить в виде

$$\mathrm{M}\{R_0^2\} = \frac{n^2 K_{XY}^2 + n D_X D_Y + n K_{XY}^2}{n^2} = K_{XY}^2 + \frac{D_X D_Y (1 + \rho_{XY}^2)}{n}. \tag{7.1.4}$$

Подставляя это выражение в (7.1.2), получаем формулу для дисперсии оценки корреляционного момента при анализе n независимых пар гауссовских случайных величин

$$D(R_0) = D_X D_Y (1 + \rho_{XY}^2)/n. \tag{7.1.5}$$

Применительно к неизвестным средним значениям рассматриваемых случайных величин для получения несмещенной оценки корреляционного момента *выборочный корреляционный момент* определяется как

$$R_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{7.1.6}$$

Здесь, как и выше, $\bar{x}$ и $\bar{y}$ — выборочные средние случайных величин X и Y соответственно.

Найдем среднее значение выборочного корреляционного момента (7.1.6), для чего последнее выражение перепишем в виде

$$R_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - m_X) - (\bar{x} - m_X)][(y_i - m_Y) - (\bar{y} - m_Y)] =$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - m_X)(y_i - m_Y) - (\bar{x} - m_X)(y_i - m_Y) - (x_i - m_X)(\bar{y} - m_Y) + (\bar{x} - m_X)(\bar{y} - m_Y)]. \tag{7.1.7}$$

При определении среднего значения каждого из членов в квадратных скобках, как и выше, учтем условие по парной независимости выборочных значений. Тогда дополнительно к ранее найденным выражениям имеем

$$\mathrm{M}\{(\bar{x}-m_X)(y_i-m_Y)\}=\frac{1}{n}\sum_{j=1}^{n}\mathrm{M}\{(x_j-m_X)(y_i-m_Y)\}=\frac{1}{n}K_{XY};$$

$$\mathrm{M}\{(\bar{y}-m_Y)(x_i-m_X)\}=\frac{1}{n}K_{XY};$$

$$\mathrm{M}\{(\bar{x}-m_X)(\bar{y}-m_Y)\}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{(x_i-m_X)(y_j-m_Y)\}=$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{M}\{(x_i-m_X)(y_i-m_Y)\}=\frac{1}{n^2}\sum_{i=1}^{n}K_{XY}=\frac{1}{n}K_{XY}.$$

С учетом приведенных соотношений среднее значение выборочного корреляционного момента (7.1.7) преобразуется к виду

$$\mathrm{M}\{R_{XY}\}=\frac{1}{n-1}\sum_{i=1}^{n}\left(K_{XY}-\frac{1}{n}K_{XY}-\frac{1}{n}K_{XY}+\frac{1}{n}K_{XY}\right)=K_{XY}.$$

Таким образом, убеждаемся, что выборочный корреляционный момент R_{XY}, определяемый согласно (7.1.6), является несмещенной оценкой корреляционного момента K_{XY}.

Аналогичным образом может быть вычислена дисперсия выборочного корреляционного момента R_{XY}. Она определяется формулой, аналогичной (7.1.5):

$$D(R_{XY})=\frac{D_X D_Y(1+\rho_{XY}^2)}{n-1}. \qquad (7.1.8)$$

Для достаточно большого объема выборки ($n \gg 1$) рассмотренные оценки корреляционного момента и их дисперсии совпадают. Поэтому, если объем выборки достаточно большой, то согласно центральной предельной теореме распределение выборочного корреляционного момента приближенно можно полагать нормальным со средним значением, равным истинному значению корреляционного момента K_{XY}, и дисперсией, определяемой формулой (7.1.5). При этих допущениях аналогично п. 4.8.1.1 [выражение (4.8.6)] истинное значение корреляционного момента K_{XY} будет находиться в интервале

$$R_{XY}\pm t_\gamma\sqrt{D(R_{XY})/n}$$

с доверительной вероятностью γ. Величина t_γ определяется по таблицам функции распределения нормированной гауссовской случайной величины $\Phi(t_\gamma)=(\gamma+1)/2$.

Применительно к группированной двумерной выборке, представленной в корреляционной таблице 3.3.2, *группированный корреляционный момент* определяется как

$$\tilde{R}_{XY}=\frac{1}{n}\sum_{j=1}^{k}\sum_{r-1}^{q}n_{jr}(x_j-\bar{x})(y_r-\bar{y}). \qquad (7.1.9)$$

7.1.2. Выборочный коэффициент корреляции

Как отмечалось выше, мерой линейной зависимости между случайными величинами X и Y является коэффициент корреляции ρ_{XY}, определяемый согласно (1.5.31) как

$$\rho_{XY}=\frac{\mathrm{M}\{(X-m_X)(Y-m_Y)\}}{\sqrt{\mathrm{M}\{(X-m_X)^2\}\mathrm{M}\{(Y-m_Y)^2\}}}=\frac{K_{XY}}{\sigma_X\sigma_Y}.$$

В соответствии с этим выражением в качестве оценки коэффициента корреляции ρ_{XY} берется *выборочный коэффициент корреляции*

$$r_{XY}=\frac{R_{XY}}{s_X s_Y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}. \qquad (7.1.10)$$

Выборочный коэффициент корреляции r_{XY} обладает теми же свойствами, какими обладает и истинный коэффициент корреляции ρ_{XY}. Как и ρ_{XY}, значение выборочного коэффициента корреляции лежит в интервале $[+1,-1]$ и принимает одно из граничных значений только при наличии идеальной линейной связи между наблюдениями. Для точного вычисления среднего значения и дисперсии выборочного коэффициента корреляции необходимо определить закон распределения случайной величины r_{XY}, а затем по известным правилам вычислить необходимые статистические характеристики выборочного коэффициента корреляции.

Применительно к гауссовским случайным величинам в литературе [8, 30, 37 и др.] приведено достаточно сложное выражение плотности вероятности для выборочного коэффициента корреляции. Это распределение зависит только от объема выборки n и истинного значения коэффициента корреляции ρ_{XY}, причем плотность вероятности выборочного коэффициента корреляции симметрична при $\rho_{XY}=0$ и существенно несимметрична для больших по абсолютной величине значений ρ_{XY}, даже при достаточно большом объеме выборки n. Если $\rho_{XY}=0$, то плотность вероятности выборочного коэффициента кор-

реляции симметрична и имеет вид

$$f_n(r_{XY}) = \frac{\Gamma[(n-1)/2]}{\Gamma[(n-2)/2])} \frac{(1-r_{XY}^2)^{(n-4)/2}}{\sqrt{\pi}}, \quad (|r_{XY}| < 1, n \geqslant 3). \tag{7.1.11}$$

Для произвольного значения ρ_{XY} в аналитическом виде строго вычислить статистические характеристики оценки коэффициента корреляции не представляется возможным. В то же время, если объем выборки достаточно велик, в результате чего ошибки определения корреляционного момента и дисперсий будут малыми, то можно ввести в вычисления некоторые упрощения. Принципиально для определения статистических характеристик выборочного коэффициента корреляции можно было бы воспользоваться методом, изложенным в п. 3.5.3 для нелинейной косвенной оценки параметра и основанным на разложении функциональной зависимости в соответствующий ряд Тейлора. Применительно к выборочному коэффициенту корреляции, зависящему от трех выборочных моментов, функцию $r_{XY} = \varphi(R_{XY}, s_X^2, s_Y^2)$ можно разложить в трехмерный ряд Тейлора в окрестности истинных значений выборочных параметров и для достаточно большой выборки $n \gg 1$ в первом приближении ограничиться первыми двумя членами разложения. Однако и в этом случае получающееся выражение для вычисления дисперсии оценки оказывается достаточно громоздким. Поэтому можно использовать прием, основанный на малых отклонениях оценки относительно точного значения оцениваемого параметра.

Обозначим через Δ_X, Δ_Y и Δ отклонения выборочных дисперсий и корреляционного момента от их точных значений:

$$\Delta_X = s_X^2 - D_X; \quad \Delta_Y = s_Y^2 - D_Y; \quad \Delta = R_{XY} - K_{XY}.$$

Тогда

$$r_{XY} = \frac{K_{XY} + \Delta}{\sqrt{(D_X + \Delta_X)(D_Y + \Delta_Y)}} = \frac{K_{XY}}{\sigma_X \sigma_Y} \frac{1 + \Delta/K_{XY}}{\sqrt{\left(1 + \dfrac{\Delta_X}{D_X}\right)\left(1 + \dfrac{\Delta_Y}{D_Y}\right)}}.$$

В силу полагаемой малости величин $(\Delta_X/D_X) \ll 1$ и $(\Delta_Y/D_Y) \ll 1$ справедливо приближенное соотношение

$$\left[\left(1 + \frac{\Delta_X}{D_X}\right)\left(1 + \frac{\Delta_Y}{D_Y}\right)\right]^{-1/2} = \left(1 + \frac{\Delta_X}{D_X} + \frac{\Delta_Y}{D_Y} + \frac{\Delta_X \Delta_Y}{D_X D_Y}\right)^{-1/2} \approx$$
$$\approx \left(1 + \frac{\Delta_X}{D_X} + \frac{\Delta_Y}{D_Y}\right)^{-1/2} \approx 1 - \frac{1}{2}\left(\frac{\Delta_X}{D_X} + \frac{\Delta_Y}{D_Y}\right).$$

Полагая также, что (Δ/K_{XY}) тоже малая величина, выражение для

выборочного коэффициента корреляции примет вид

$$r_{XY} \approx \rho_{XY}\left(1 + \frac{\Delta}{K_{XY}}\right)\left[1 - \frac{1}{2}\left(\frac{\Delta_X}{D_X} + \frac{\Delta_Y}{D_Y}\right)\right] \approx$$
$$\approx \rho_{XY}\left[1 + \frac{\Delta}{K_{XY}} - \frac{1}{2}\left(\frac{\Delta_X}{D_X} + \frac{\Delta_Y}{D_Y}\right)\right] =$$
$$= \rho_{XY}\left(1 + \frac{R_{XY} - K_{XY}}{K_{XY}} - \frac{s_X^2 - D_X}{2D_X} - \frac{s_Y^2 - D_Y}{2D_Y}\right).$$

Окончательно имеем

$$r_{XY} \approx \rho_{XY}\left(1 + \frac{R_{XY}}{K_{XY}} - \frac{s_X^2}{2D_X} - \frac{s_Y^2}{2D_Y}\right). \tag{7.1.12}$$

Полученное приближенное выражение для выборочного коэффициента корреляции линейно зависит от выборочных значений корреляционного момента R_{XY} и дисперсий s_X^2 и s_Y^2.

Учитывая, что оценки корреляционного момента и дисперсий несмещенные, рассматриваемая оценка коэффициента корреляции оказывается также несмещенной, т.е. $\mathrm{M}\{r_{XY}\} = \rho_{XY}$, при условии выполнения указанных выше допущений относительно малости ошибок, что равносильно достаточно большому объему выборок. В общем же случае оценка коэффициента корреляции в виде (7.1.10) оказывается смещенной.

Дисперсия выборочного коэффициента корреляции в рассматриваемом приближении определяется из соотношения

$$D(r) = \mathrm{M}\{r^2\} - (\mathrm{M}\{r\})^2 \approx \rho_{XY}^2 \mathrm{M}\left\{\left(\frac{R_{XY}}{K_{XY}} - \frac{s_X^2}{2D_X} - \frac{s_Y^2}{2D_Y}\right)^2\right\} =$$
$$= \rho_{XY}^2 \mathrm{M}\left\{\frac{R_{XY}^2}{K_{XY}^2} + \frac{(s_X^2)^2}{4D_X^2} + \frac{(s_Y^2)^2}{4D_Y^2} - \frac{R_{XY} s_X^2}{D_X K_{XY}} - \frac{R_{XY} s_Y^2}{D_Y K_{XY}} + \frac{s_X^2 s_Y^2}{2D_X D_Y}\right\}. \tag{7.1.13}$$

Аналогично определению выборочного корреляционного момента R_0 для рассматриваемого приближения будем полагать, что средние значения случайных величин X и Y известны, а вместо корреляционного момента R_{XY} (7.1.6) примем R_0 (7.1.1). Как и выше, полагая $m_X = m_Y = 0$, можно вычислить необходимые моменты, входящие в выражение для дисперсии выборочного коэффициента корреляции r_{XY}. Момент $\mathrm{M}\{R_{XY}^2\}$ определяется (7.1.4), а при вычислении остальных моментов в соответствии с (7.1.3) учтем, что

$$\mathrm{M}\{X_i^4\} = \begin{cases} 3(\mathrm{M}\{X_i^2\})^2 = 3D_X^2 & \text{при } i = j, \\ \mathrm{M}\{X_i^2\}\mathrm{M}\{X_j^2\} = D_X^2 & \text{при } i \neq j; \end{cases}$$

$$\mathrm{M}\{X_i^2 Y_j^2\} = \begin{cases} \mathrm{M}\{X_i^2 Y_i^2\} = D_X D_Y + 2K_{XY}^2 & \text{при } i = j, \\ \mathrm{M}\{X_i^2\}\mathrm{M}\{Y_j^2\} = D_X D_Y & \text{при } i \neq j; \end{cases}$$

$$\mathrm{M}\{X_i^2X_jY_j\} = \begin{cases} 3\mathrm{M}\{X_i^2\}\mathrm{M}\{X_iY_i\} = 3D_XK_{XY} & \text{при } i=j, \\ \mathrm{M}\{X_i^2\}\mathrm{M}\{X_jY_j\} = D_XK_{XY} & \text{при } i\neq j. \end{cases}$$

Подобные выражения получаем и для $\mathrm{M}\{Y_i^4\}$ и $\mathrm{M}\{X_iY_iY_j^2\}$ при замене соответствующих индексов x на y. При вычислении средних значений случайных величин, стоящих в числителях (7.1.13), представим двойные суммы в виде двух сумм: одинарных при $i=j$ с числом членов n и двойных при $i\neq j$ с числом членов $n(n-1)$. Подставляя полученные моменты в исходные выражения, имеем

$$\mathrm{M}\{(s_X^2)^2\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{X_i^2X_j^2\} = \frac{n+2}{n}D_X^2;$$

$$\mathrm{M}\{(s_Y^2)^2\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{Y_i^2Y_j^2\} = \frac{n+2}{n}D_Y^2;$$

$$\mathrm{M}\{R_{XY}s_X^2\} = \frac{n+2}{n}K_{XY}D_X; \quad \mathrm{M}\{R_{XY}s_Y^2\} = \frac{n+2}{n}K_{XY}D_Y;$$

$$\mathrm{M}\{s_X^2s_Y^2\} = \frac{1}{n}(2K_{XY}^2 + nD_XD_Y).$$

Подставляя эти моменты в исходное выражение для дисперсии выборочного коэффициента корреляции (7.1.13), получим

$$D(r_{XY}) \approx \frac{(1-\rho_{XY}^2)^2}{n} \leqslant \frac{1}{n}. \qquad (7.1.14)$$

Применительно к гауссовским случайным величинам в [8, 72] приведены графики функций нижней $r_1(\rho_{XY}, n, \gamma)$ и верхней $r_2(\rho_{XY}, n, \gamma)$ границ доверительного интервала истинного значения коэффициента корреляции ρ_{XY} для доверительной вероятности $\gamma = 0{,}95$; 0,99 и объема выборок n от 3 до 400. Эти же графики можно использовать для отыскания верхнего и нижнего критических значений для выборочного коэффициента корреляции r_{XY} при проверке гипотезы о заданном значении истинного коэффициента корреляции ρ_{XY} для двусторонних уровней значимости $\alpha = 0{,}025$ и 0,05. Применительно к нулевой гипотезе $(\rho_{XY}=0)$ случайная величина $t = r_{XY}\sqrt{(n-2)/(1-r_{XY}^2)}$ имеет распределение Стьюдента с $\nu = n-2$ степенями свободы [2, 8, 31, 37]. Поэтому, если обозначить t_α – критическое значение t-распределения с ν степенями свободы, то величина

$$r_\alpha = \frac{t_\alpha}{\sqrt{t_\alpha^2 + n - 2}} \qquad (7.1.15)$$

будет являться односторонним критическим значением выборочного коэффициента корреляции. В [8, 72] для $\rho_{XY}=0$ (применительно к проверке гипотез о некоррелированности выборок) имеются таблицы критических значений выборочного коэффициента корреляции $\mathrm{P}\{|r_{XY}| > r_{\alpha;\nu}\} = \alpha$. В сокращенном виде они приведены в табл. 10 Приложения 2. Поэтому нулевая гипотеза $(\rho_{XY}=0)$ отвергается, если $|r_{XY}| > r_{\alpha;\nu}$.

Пример 7.1.1. По выборке объема $n=20$ получено выборочное значение коэффициента корреляции $r_{XY} = 0{,}2$. Убедиться, что $r_{XY} = 0{,}2$ не противоречит гипотезе $\rho_{XY}=0$ с уровнем значимости $\alpha = 0{,}05$.

Согласно табл. 10 Приложения 2 для числа степеней свободы $\nu = n-2 = 18$ и уровня значимости $\alpha = 0{,}05$ критическое значение статистики $r_{0,05;18} = 0{,}444$. Нулевая гипотеза $\rho_{XY}=0$ не отвергается с уровнем значимости $\alpha = 0{,}05$, так как $r_{XY} = 0{,}2 < r_{0,05;18} = 0{,}444$.

Иногда выборочный коэффициент корреляции применяют в качестве меры зависимости выборок. Это оправдано лишь тогда, когда случайные величины имеют нормальное распределение, так как для подобных случайных величин некоррелированность эквивалентна независимости. Однако, если объем выборки достаточно велик, т.е. если $n \gg 1$ (в пределе $n\to\infty$), функция распределения коэффициента корреляции будет асимптотически нормальна со средним значением ρ_{XY} и дисперсией, определяемой (7.1.14) [8, 31, 37 и др.]. Вместе с этим необходимо подчеркнуть еще раз, что для ρ_{XY}, близких к ± 1, подобная аппроксимация возможна лишь при очень больших значениях n.

Задача интервальной оценки существенно упрощается благодаря *нормализующему преобразованию* выборочного коэффициента корреляции, предложенному Р.А. Фишером, вида

$$z = \frac{1}{2}\ln\frac{1+r_{XY}}{1-r_{XY}}. \qquad (7.1.16)$$

Нелинейное преобразование (7.1.16) является взаимно однозначным преобразованием величин z и r_{XY}. Среднее значение m_Z, дисперсия $D(z)$, коэффициенты асимметрии γ_1 и эксцесса γ_2 определяются следующими формулами [8, 13, 37, 53 и др.]:

$$m_Z \approx \frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} + \frac{\rho_{XY}}{2(n-1)}; \qquad (7.1.17)$$

$$D(z) \equiv \sigma_Z^2 \approx 1/(n-3); \qquad (7.1.18)$$

$$\gamma_1 \approx \rho_{XY}^6/(n-3)^3; \qquad (7.1.19)$$

$$\gamma_2 \approx 2/(n-3). \qquad (7.1.20)$$

Из этих формул видно, что дисперсия и коэффициент эксцесса практически не зависят от значения ρ_{XY}, а коэффициент асимметрии уже при достаточно небольших значениях n близок к нулю. Необходимо также отметить, что второй (поправочный) член у m_Z при достаточно больших n мал по сравнению с первым слагаемым и поэтому вторым членом часто можно пренебречь. В связи со сказанным, плотность вероятности случайной величины Z при $n \geqslant 20$ близка к нормальной.

Решая уравнение (7.1.16) относительно выборочного коэффициента

корреляции, находим r_{XY} как функцию от z:

$$r_{XY} = \frac{\exp(2z) - 1}{\exp(2z) + 1}. \qquad (7.1.21)$$

Для упрощения расчетов в [2, 8, 35, 51 и др.] приведены таблицы преобразования выборочного коэффициента корреляции (7.1.16) и (7.1.21).

Выше (п. 4.8.1) применительно к выборочному среднему, имеющему нормальный закон распределения со средним значением m и дисперсией $D = \sigma^2/n$, была получена формула для определения доверительных границ по заданной доверительной вероятности γ. Применительно к введенной величине Z имеем

$$\mathbf{P}\{|(z - m_Z)/\sigma_Z| \equiv |t| < t_\gamma\} = \gamma, \qquad (7.1.22)$$

где t — нормированная нормально распределенная случайная величина. При этом параметр t_γ определяется из соотношения

$$\Phi(t_\gamma) = (1 + \gamma)/2 = 1 - \mathbf{P}\{t > t_\gamma\}.$$

Таким образом, доверительные границы для величины m_Z равны

$$z_1 = z - \sigma_Z t_\gamma = \frac{1}{2} \ln\left(\frac{1 + r_{XY}}{1 - r_{XY}}\right) - \frac{t_\gamma}{\sqrt{n-3}};$$
$$z_2 = z + \sigma_Z t_\gamma = \frac{1}{2} \ln\left(\frac{1 + r_{XY}}{1 - r_{XY}}\right) + \frac{t_\gamma}{\sqrt{n-3}}. \qquad (7.1.23)$$

Поскольку случайная величина Z функционально взаимно однозначно связана с выборочным коэффициентом корреляции r_{XY}, а среднее значение m_Z функционально взаимно однозначно связано с истинным значением коэффициента корреляции ρ_{XY}, доверительные интервалы для истинного коэффициента корреляции r_1 и r_2 можно выразить через доверительные интервалы z_1 и z_2 соответственно. Согласно (7.1.21) имеем

$$r_1 = \frac{\exp(2z_1) - 1}{\exp(2z_1) + 1}; \quad r_2 = \frac{\exp(2z_2) - 1}{\exp(2z_2) + 1}. \qquad (7.1.24)$$

В этих границах с заданной доверительной вероятностью γ заключено истинное значение коэффициента корреляции ρ_{XY}.

Нормализующее преобразование (7.1.16), строго говоря [8], справедливо для значений выборочного коэффициента корреляции, лежащего в пределах $0 \leqslant |r_{XY}| \leqslant 0{,}97$.

Если же $0{,}97 < |r_{XY}| < 1$, то для вычисления z следует воспользоваться формулой [8]

$$z = -\frac{1}{2}\left(\frac{1 - r_{XY}}{2} + \ln\frac{1 - r_{XY}}{2}\right), \quad 0{,}97 < |r_{XY}| < 1. \qquad (7.1.25)$$

На рис. 7.1.1 приведены графики плотности вероятности для выборочного коэффициента корреляции r_{XY} (*а*) и нормализующей величины z (*б*) в выборках из нормальной совокупности для $n = 10$ и трех значений истинного коэффициента корреляции ρ_{XY} [37].

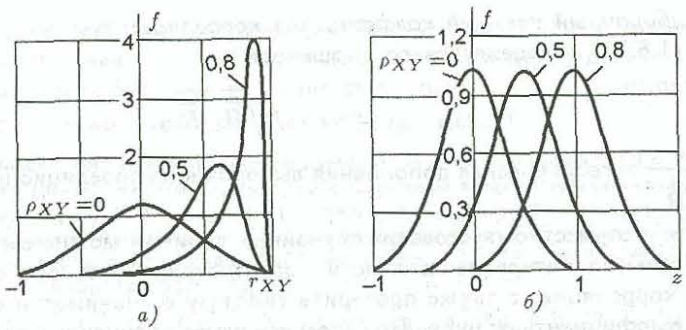


Рис. 7.1.1. Графики плотности вероятности выборочного коэффициента корреляции (а) и случайной величины z (*б*) для $n = 10$

Кривые подтверждают высказанные выше утверждения о несимметричности распределения выборочного коэффициента корреляции r_{XY} и нормализации распределения случайной величины Z.

Пример 7.1.2. По выборке, состоящей из $n = 25$ пар наблюдений, было найдено, что выборочный коэффициент корреляции равняется $r_{XY} = 0{,}6$. В каких доверительных границах находится истинное значение коэффициента корреляции ρ_{XY}, если предполагается, что наблюденные пары независимы и распределены нормально, а доверительная вероятность $\gamma = 0{,}95$.

В табл. 1б Приложения 2 для функции $\mathbf{P}\{t > t_\beta\} = \beta$, где $\beta = (1-\gamma)/2 = 0{,}025$, находим значение аргумента $t_\gamma = t_\beta = 1{,}96$. По приведенным выше формулам (7.1.16), (7.1.23) и (7.1.21) вычисляем: $z = 0{,}693$; $z_1 = 0{,}275$; $z_2 = 1{,}111$; $r_1 = 0{,}268$; $r_2 = 0{,}804$. Следовательно, истинное значение коэффициента корреляции лежит в интервале $0{,}268 \leqslant \rho_{XY} \leqslant 0{,}804$ с доверительной вероятностью $\gamma = 0{,}95$. Значения границ доверительного интервала, вычисленные для указанных исходных данных по графикам [8, 72], основанным на точном распределении выборочного коэффициента корреляции, равны 0,260 и 0,800 соответственно. Из приведенного примера видно, что доверительный интервал для приближенной интервальной оценки незначительно отличается от интервала, полученного на основе точных распределений, что свидетельствует о приемлемой точности статистических выводов при использовании приближенной оценки.

Выше в п. 1.5.4 введены понятия частного и множественного коэффициентов корреляции. Рассмотрим их выборочные аналоги [2, 15, 31 и др.]. Пусть наблюдается p-мерная выборка $\mathbf{x}(x_1, x_2, \cdots, x_p)$ случайной величины $\mathbf{X}(X_1, X_2, \cdots, X_p)$ с *выборочной корреляционной матрицей*

$$\widehat{\mathbf{R}} = \begin{pmatrix} r_{11} & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & r_{pp} \end{pmatrix}, \qquad (7.1.26)$$

где r_{ij}, $i, j = 1, 2, \cdots, p$, — выборочный (парный) коэффициент корреляции (7.1.10)

Выборочный частный коэффициент корреляции $r_{12(3,4,\cdots,p)}$ аналогично (1.5.73) определяется соотношением

$$r_{12.(3,4,\cdots,p)} = -\widehat{R}_{12}/\sqrt{\widehat{R}_{11}\widehat{R}_{22}}, \qquad (7.1.27)$$

где $\widehat{R}_{11}$ — алгебраические дополнения выборочной корреляционной матрицы $\widehat{\mathbf{R}}$.

Для p совместно гауссовских случайных величин можно вычислить доверительные интервалы для истинного значения частного коэффициента корреляции, а также проверить гипотезу о значимости отличия этого коэффициента от нуля. При этом все ранее полученные соотношения для распределения выборочного парного коэффициента корреляции справедливы и для распределения выборочных частных коэффициентов корреляции за исключением того, что вместо объема выборок n следует понимать величину $n-p$, причем, естественнно, $n > p$.

Пример 7.1.3. В результате опыта получены три гауссовские выборки, каждая объема $n = 25$. На основе этих данных согласно соотношениям, аналогичным (7.1.10), были вычислены выборочные парные коэффициенты корреляции $r_{12} = 0{,}15$; $r_{13} = 0{,}50$; $r_{23} = 0{,}25$. Учитывая соотношение вида (1.5.74), выборочные частные коэффициенты корреляции равны $r_{12.(3)} \approx 0{,}03$, $r_{13.(2)} \approx 0{,}48$, $r_{23.(1)} \approx 0{,}20$. Проверим гипотезу о том, что истинные частные коэффициенты корреляции $\rho_{12.(3)}$, $\rho_{32.(2)}$, $\rho_{23.(1)}$ несущественно отличаются от нуля с уровнем значимости $\alpha = 0{,}10$.

Для прроверки гипотез воспользуемся статистикой (7.1.15), в которой число степеней свободы равно $\nu = (n-3)-2 = 20$. При этом критическое значение этой статистики для $\alpha = 0{,}10$ и $\nu = 20$ согласно табл. 10 Приложения 2 $|r_{0,1;20}| = 0{,}36$. Сравнивая критическое значение с величинами выборочных частных коэффициентов корреляции нетрудно видеть, что несущественно отличаются от нуля с уровнем значимости $\alpha = 0{,}10$ только $\rho_{12.(3)}$ и $\rho_{23.(1)}$.

Выборочный множественный коэффициент корреляции $r_{1.(1,3,\cdots,p)}$ между случайной величиной X_1 и остальными случайными величинами $X_2, X_3, \cdots, X_p$ аналогично (1.5.75) определяется соотношением

$$r^2_{1.(23\cdots p)} = 1 - \frac{\widehat{R}}{\widehat{R}_{11}} = 1 - (1-r^2_{12})(1-r^2_{13.(2)})\ldots(1-r^2_{1p.(23\cdots(p-1))}), \qquad (7.1.28)$$

где $\hat{R}$ — определитель корреляционной матрицы $\widehat{\mathbf{R}}$.

Применительно к p совместно гауссовским случайным величинам выборочный множественный коэффициент корреляции, как правило, применяется для проверки гипотезы о том, что все выборочные парный и частный коэффициенты корреляции равны нулю ($\rho_{12} = 0$; $\rho^2_{1p.(23\cdots(p-1))} = 0$), т.е. нулевая гипотеза H_0 соответствует равенству нулю квадрата множественного коэффициента корреляции $\rho^2_{1.(23\cdots p)} = 0$. Статистикой для проверки этой гипотезы служит величина [2, 31]

$$F_{p-1,n-p} = \frac{r^2_{1.(2\cdots p)}/(p-1)}{\left(1-r^2_{1.(2\cdots p)}\right)/(n-p)}, \qquad (7.1.29)$$

которая имеет F-распределение с $\nu_1 = p-1$ и $\nu_2 = n-p$ степенями свободы. Нулевая гипотеза не отклоняется, если для заданного уровня значимости α выборочное значение статистики $F^*_{(p-1),(n-p)}$ не превышает ее критическое значение $F_{\alpha;(p-1),(n-p)}$.

Пример 7.1.4. Применительно к данным предыдущего примера ($n = 25, p = 3$) проверим нулевую гипотезу о том, что выборочный множественный коэффициент корреляции $r_{1.(23)}$ между случайной величиной X_1 и случайными величинами X_2 и X_3 с уровнем значимости $\alpha = 0{,}05$ несущественно отличается от нуля. Выборочный множественный коэффициент корреляции $r_{1.(23)}$ (7.1.28) и выборочное значение F-статистки (7.1.29) равны

$$r^2_{1.(23)} = 1-(1-r^2_{12})(1-r^2_{13.(2)}) = 1-(1-0{,}15^2)(1-0{,}48^2) \approx 0{,}248;$$

$$F^*_{2,22} = \frac{0{,}248/(3-1)}{(1-0{,}248)/(25-3)} \approx 3{,}63.$$

Критическое значение F-статистики для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы $\nu_1 = 2$ и $\nu_2 = 22$ согласно табл. 3 Приложения 2 $F_{0{,}05;2,22} = 3{,}47$. Так как выборочное значение статистики превышает критическое значение статистики, нулевая гипотеза о независимости случайной величины X_1 от случайных величин X_2 и X_3 с уровнем значимости $\alpha = 0{,}05$ не подтверждается.

При достаточно большой по объему двумерной выборке она группируется по соответствующим интервалам и представляется в виде корреляционной таблицы (см. табл. 3.3.2). На основе данных этой таблицы вычисляется *выборочный группированный коэффициент корреляции*

$$\tilde{r}_{XY} = \frac{\sum_{j=1}^{k}\sum_{l=1}^{q} n_{jl}(x_j-\bar{x})(y_l-\bar{y})}{\sqrt{\sum_{j=1}^{k} n_{j.}(x_j-\bar{x})^2 \sum_{l=1}^{q} n_{.l}(y_l-\bar{y})^2}}. \qquad (7.1.30)$$

Здесь k и q — количества интервалов переменных X и Y соответственно; n_{jl} — численность ячейки, $n_{j.}$ и $n_{.l}$ — число выборочных значений x_i и y_l, попавших в j-ю строку и l-й столбец таблицы.

7.1.3. Выборочное корреляционное отношение

Выборочный коэффициент корреляции (в том числе и группированный) характеризует статистическую линейную зависимость между двумя случайными величинами. Как показано выше в п. 1.5.2, для характеристики нелинейной статистической зависимости между двумя случайными величинами применяют корреляционное отношение.

Корреляционное отношение $\eta^2_{Y|x}$ случайной величины Y по случайной величине X определяется формулой (1.5.43), представляющей собой отношение усредненной по всем возможным значениям случайной величины X условного среднего случайной величины Y к ее безусловной дисперсии.

Необходимо еще раз подчеркнуть, что согласно неравенству (1.5.44) близость квадрата коэффициента корреляции к нулю не означает отсутствие нелинейной зависимости между случайными величинами.

Применительно к группированной выборке на основе корреляционной таблицы (типа табл. 3.3.2) аналогичным образом может быть записано выражение для *выборочного корреляционного отношения*

$$\tilde{\eta}^2_{Y|x} = \frac{\sum_{j=1}^{k} n_{j\cdot}(y_{j\cdot} - y_{\cdot\cdot})^2/n}{\sum_{l=1}^{q} n_{\cdot l}(y_l - y_{\cdot\cdot})^2/n} = \frac{\sum_{j=1}^{k} n_{j\cdot} y_{j\cdot}^2 - n y_{\cdot\cdot}^2}{\sum_{l=1}^{q} n_{\cdot l} y_l^2 - n y_{\cdot\cdot}^{\,2}} \qquad (7.1.31)$$

Здесь числитель является группированным усредненным рассеянием условных средних значений $y_{j\cdot}$ около безусловного среднего значения $y_{\cdot\cdot}$. Знаменатель есть группированная безусловная дисперсия случайной величины Y. Число интервалов по случайным величинам X и Y обозначено соответственно как k и q.

Аналогично вычисляется выборочное корреляционное отношение $\tilde{\eta}^2_{X|y}$ случайной величины X по случайной величине Y.

Пример 7.1.5. Рассмотрим гипотетический пример. Пусть на основе двумерной выборки случайных величин X и Y объема n образована корреляционная табл. 7.1.1. Вычислим выборочные группированный коэффициент корреляции r_{XY} между случайными величинами и корреляционное отношение $\tilde{\eta}^2_{Y|x}$ случайной величины Y по случайной величине X. В таблице также приведены данные, необходимые для дальнейших вычислений.

Согласно (3.4.37)–(3.4.40) имеем

$$\bar{x} = \frac{1}{40}\sum_{j=1}^{5}\sum_{l=1}^{7} n_{jl}x_j = 10; \quad s_X^2 = \frac{1}{40}\sum_{j=1}^{5}\sum_{l=1}^{7} n_{jl}x_j^2 - 10^2 = 0{,}187;$$

$$\bar{y} = \frac{1}{40}\sum_{j=1}^{5}\sum_{l=1}^{7} n_{jl}y_l = 7{,}15; \quad s_Y^2 = \frac{1}{40}\sum_{j=1}^{5}\sum_{l=1}^{7} n_{jl}y_l^2 - 7{,}15^2 = 12{,}5.$$

Таблица 7.1.1

Среднее значение интервала x_j	Среднее значение интервала y_l							$n_{j\cdot}$	$y_{j\cdot}$	$n_{j\cdot}y_{j\cdot}$	$n_{j\cdot}y_{j\cdot}^2$
	1	3	5	7	9	11	13				
6	–	–	–	1	2	3	2	8	10,5	84,00	882,0
8	–	2	4	2	1	–	–	9	5,44	48,96	266,8
10	2	3	1	–	–	–	–	6	2,67	16,02	42,7
12	–	2	3	3	1	–	–	9	5,60	51,03	269,0
14	–	–	–	1	2	2	3	8	10,75	86,00	924,5
$n_{\cdot l}$	2	7	8	7	6	5	5	40	–	–	2405,0

Подставляя вычисленные данные в (7.1.10) и (7.1.31), имеем $r_{XY} \approx 0{,}19$; $\tilde{\eta}^2_{Y|x} \approx 0{,}72$. Так как $r^2 \approx 0{,}04 < \tilde{\eta}^2_{Y|x} \approx 0{,}72$, можно утверждать, что зависимость случайной величины Y от случайной величины X носит явно нелинейный характер. При этом аппроксимирующая кривая этой зависимости имеет вид квадратичной параболы

$$y_{j\cdot}^* = 2{,}7 + 0{,}5(x_j - 10)^2.$$

В то же время аппроксимация, соответствующая линейной выборочной регрессии, рассматриваемой ниже в п. 8.2.1, определяется выражением

$$y_{j\cdot}^{**} = 7{,}15 + 0{,}03(x_j - 10).$$

На рис. 7.1.2 крестиками показаны «экспериментальные» точки $y_{j\cdot}$ и аппроксимирующие кривые $y_{j\cdot}^*$ (1) и $y_{j\cdot}^{**}$ (2). Из рисунка видно, что нелинейная зависимость случайной величины Y от случайной величины X близка к детерминированной. Линейная аппроксимация (регрессия) ни в коей мере не может характеризовать зависимость Y от X.

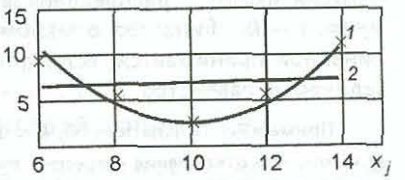


Рис. 7.1.2. Аппроксимация зависимостей

Выше в п. 1.5.2 было показано, что равенство нулю коэффициента корреляции ($\rho_{XY} = 0$) не означает отсутствие нелинейной зависимости между случайными величинами. Вместе с тем, если $\eta^2_{Y|x} = 0$, то корреляционные связи между случайными величинами отсутствуют. Отмечалось, также, что величина $(\eta^2_{Y|x} - \rho^2_{XY})$ может служить индикатором отличия рассматриваемой регрессии от линейной. Поэтому представляет интерес проверить нулевые гипотезы о том, что выборочные значения $(\tilde{\eta}^2_{Y|x})$ и $|\tilde{\eta}^2_{Y|x} - r^2_{XY}|$ с заданным уровнем значимости α не существенно отличаются от нуля. В литературе по математической статистике [2, 31, 72 и др.] рассматриваются статистики и критерии для проверки этих гипотез.

Для проверки гипотезы о некоррелированности выборочных значений наиболее часто применяется статитстика

$$F_{k-1;n-k} = \frac{\tilde{\eta}^2_{Y|x}/(k-1)}{(1-\tilde{\eta}^2_{Y|x})/(n-k)}, \qquad (7.1.32)$$

имеющая F-распределение со степенями свободы $\nu_1 = k-1$ и $\nu_2 = n-k$. Гипотеза о том, что выборочное значение корреляционного отношения несущественно отличается от нуля, не противоречит опытным данным, если для заданного уровня значимости α справедливо неравенство $F^*_{k-1,n-k} < F_{\alpha;k-1,n-k}$.

Пример 7.1.6. На основе группированной выборки объема $n = 60$ и числа интервалов $k = 10$ случайной величины X вычислено выборочное значение корреляционного отношения $\tilde{\eta}^2_{Y|x} = 0{,}3$. С уровнем значимости $\alpha = 0{,}05$ проверим нулевую гипотезу о том, что полученное значение корреляционного отношения несущественно отличается от нуля.

Согласно (7.1.32) $F^*_{9,50} \approx 2{,}86$, тогда как критическое значение статистики для проверки гипотезы ($\nu_1 = 9; \nu_2 = 50$) в соответствии с табл. 4 Приложения 2 $F_{0{,}05;9,50} \approx 2{,}07$. Так как $F^*_{9,50} \approx 2{,}86 > F_{0{,}05;9,50} \approx 2{,}07$, нулевая гипотеза не принимается с уровнем значимости $\alpha = 0{,}05$.

Для проверки гипотезы о том, что исследуемая зависимость мало отличается от линейной $|\tilde{\eta}^2_{Y|x} - r^2_{XY}|$ применяется статистика вида

$$F_{k-2,n-2} = \frac{|\tilde{\eta}^2_{Y|x} - r^2_{XY}|/(k-2)}{(1-\tilde{\eta}^2_{Y|x})/(n-k)} \qquad (7.1.33)$$

которая имеет F-распределение со степенями свободы $\nu_1 = k-2$ и $\nu_2 = n-k$. Гипотеза о малом отличии исследуемой зависимости от линейной принимается, если для заданного уровня значимости α справедливо неравенство $F^*_{k-2,n-k} < F_{\alpha;k-2,n-k}$.

Пример 7.1.7. Для $n = 60$, $k = 10$, $\hat{\eta}^2_{Y|x} = 0{,}3$ и $r^2_{XY} = 0{,}25$ проверим гипотезу о том, что отклонение регрессии $m_Y(x)$ от линейной зависимости несущественно для уровня значимости $\alpha = 0{,}05$. Согласно (7.1.33) выборочное значение статистики $F^*_{8,50} \approx 0{,}54$. При этом критическое значение статистики согласно табл. 4 Приложения 2 для $\nu_1 = 8$, $\nu_2 = 50$ и $\alpha = 0{,}05$ равно $F_{0{,}05;8,50} \approx 2{,}13$. Поскольку оно больше значения выборочной статистики, справедлива нулевая гипотеза о том, что отклонение регрессии $m_Y(x)$ от линейной зависимости несущественно для заданного уровня значимости $\alpha = 0{,}05$.

7.2. Ранговые методы исследования статистической связи

Как уже отмечалось выше, изучаемые объекты, события и т.д. могут отличаться друг от друга как количественными, так и качественными признаками (типа больше-меньше-равно, лучше-хуже и т.п.). Если количественные признаки можно измерить, то качественные признаки измерениям не поддаются. Особенно часто задача исследования качественных признаков наблюдается при психологических, социологических и общественно-политических исследованиях, а также при системном анализе других сложных процессов, явлений, систем и устройств. В подобных случаях наблюдения (их часто называют индивидуумами) $x_1, x_2, \ldots, x_n$ можно «ранжировать», т.е. упорядочить по степени их предпочтения, приписав каждому наблюдению x_i порядковый номер, соответствующий итогам сравнения с остальными наблюдениями x_j $(i \neq j)$. Этот номер, как отмечалось выше, называют *рангом* соответствующего *наблюдения*.

Если ранжированный ряд из n наблюдений содержит m признаков, то сумма рангов должна удовлетворять условию

$$\sum_{j=1}^{m}\sum_{i=1}^{n} = \frac{m(n+1)}{2}, \tag{7.2.1}$$

которое выполняется, если все ранги различны (отсутствуют совпадающие ранги). Однако возможны ситуации, когда в результате ранжирования в ряду будут совпадающие ранги, обусловленные тем, что признаки могут быть неразличимы по их важности. В таких случаях исходная ранжированная последовательность должна быть перекодирована. Сущность перекодирования рангов изложена выше на примере в гл. 6 (п. 6.3).

Если совокупность наблюдаемых данных обладает не одним, а двумя признаками, то для исследования влияния их друг на друга каждому наблюдению x_i, $i = 1, 2, \ldots, n$, приписывают два сравниваемых порядковых номера (два ранга) в соответствии с ранжировкой по обоим признакам. Аналогичным образом может рассматриваться задача исследования наблюдений, обладающих тремя и более качественными признаками.

7.2.1. Коэффициент ранговой корреляции

Достаточно эффективной мерой связи признаков служит коэффициент ранговой корреляции. Рассмотрим выборку из n независимых наблюдений двух признаков и каждому наблюдению припишем два порядковых номера (два ранга) в соответствии с ранжированием по обоим признакам. Результаты ранжирования можно записать в виде двух строк:

$r_{11}, r_{12}, \ldots, r_{1n}$ — ранжирование по первому признаку;

$r_{21}, r_{22}, \ldots, r_{2n}$ — ранжирование по второму признаку.

Тогда *выборочный коэффициент ранговой корреляции* r_s можно записать аналогично обычному выборочному коэффициенту взаимной корреляции r в виде

$$r_s = \frac{\sum_{i=1}^{n}(r_{1i} - R_{1.})(r_{2i} - R_{2.})}{\sqrt{\sum_{i=1}^{n}(r_{1i} - R_{1.})^2 \sum_{i=1}^{n}(R_{ri} - R_{2.})^2}}. \tag{7.2.2}$$

Здесь $R_{1.}$ и $R_{2.}$ — выборочные средние рангов по первому и второму признакам:

$$R_{1.} = \frac{1}{n}\sum_{i=1}^{n} r_{1i}; \quad R_{2.} = \frac{1}{n}\sum_{i=1}^{n} r_{2i}. \tag{7.2.3}$$

Как правило, ранжированная последовательность совпадает с рядом натуральных чисел 1, 2,..., n, для которого справедливы соотношения (1.3.39). Поэтому при ранжировании с использованием ряда натуральных чисел средние значения и сумма квадратов их отклонений от средних значений равны соответственно

$$R_{1.} = R_{2.} \equiv R_{..} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}; \tag{7.2.4}$$

$$q = \sum_{i=1}^{n}(r_{1i} - R_{1.})^2 = \sum_{i=1}^{n}(r_{2i} - R_{2.})^2 = \sum_{i=1}^{n} r^2_{1i} - nR^2_{..} = \frac{n^3 - n}{12}. \tag{7.2.5}$$

При этом

$$\sum_{i=1}^{n}(r_{1i} - r_{2i})^2 = \sum_{i=1}^{n}[(r_{1i} - R_{1.}) - (r_{2i} - R_{2.})]^2 =$$

$$= \sum_{i=1}^{n}[(r_{1i} - R_{1\cdot})^2 + (r_{2i} - R_{2\cdot})^2] - 2\sum_{i=1}^{n}(r_{1i} - R_{1\cdot})(r_{2i} - R_{2\cdot}) =$$

$$= \frac{n^3 - n}{6} - 2\sum_{i=1}^{n}(r_{1i} - R_{1\cdot})(r_{2i} - R_{2\cdot}).$$

Отсюда

$$\sum_{i=1}^{n}(r_{1i} - R_{1\cdot})(r_{2i} - R_{2\cdot}) = \frac{n^3 - n}{12} - \frac{1}{2}\sum_{i=1}^{n}(r_{1i} - r_{2i})^2.$$

Подставляя эти соотношения в выражение (7.2.2), получим формулу для вычисления коэффициента ранговой корреляции

$$r_s = 1 - \frac{6S_r}{n^3 - n}, \tag{7.2.6}$$

где

$$S_r = \sum_{i=1}^{n} d_i^2; \tag{7.2.7}$$

а

$$d_i = r_{1i} - r_{2i} \tag{7.2.8}$$

— разность между рангами i-й пары в последовательности, состоящей из n парных элементов, причем, как уже упоминалось выше, первые члены всех пар упорядочиваются отдельно от вторых членов.

Формула (7.2.6) предполагает, что в каждой из сравниваемых ранжированных последовательностей ранги не совпадают. При совпадении рангов в ранжированной последовательности, как отмечалось выше, этим рангам назначают ранги, равные среднему арифметическому из рангов тех мест, которые они делят в последовательности. Более точная формула для расчета коэффициента ранговой корреляции учитывает наличие групп g_1 и g_2 совпадающих рангов и число совпадений t_{1j} и t_{2j} в каждой из групп. Она имеет вид [2, 17, 62, 68 и др.]

$$r_s^* = \frac{(n^3 - n)/6 - \sum_{i=1}^{n}(r_{1i} - r_{2i})^2 - T_1 - T_2}{\sqrt{((n^3 - n)/6 - 2T_1)\,((n^3 - n)/6 - 2T_2)}}. \tag{7.2.9}$$

Здесь

$$T_1 = \sum_{j=1}^{g_1}\frac{t_{1j}^2 - t_{1j}}{12}; \quad T_2 = \sum_{j=1}^{g_2}\frac{t_{2j}^2 - t_{2j}}{12}, \tag{7.2.10}$$

Если величины T_1 и T_2 являются небольшими по сравнению с $(n^3 - n)/6$, для практических расчетов приемлема приближенная формула

$$r_s^{**} = 1 - \frac{1}{(n^3 - n)/6 - (T_1 + T_2)}\sum_{i=1}^{n}(r_{1i} - r_{2i})^2. \tag{7.2.11}$$

Пример 7.2.1. Имеются две ранжированные последовательности из 11 признаков, представленные в табл. 7.2.1. В первых двух строках последнего столбца написаны числа совпадающих рангов. В нижней строке последнего столбца приведена сумма квадратов разности рангов.

Таблица 7.2.1

r_{1i}	5,5	2,5	10	8	5,5	11	2,5	2,5	8	2,5	8	2+4
r_{2i}	2,5	2,5	8,5	6	8,5	10,5	6	2,5	6	2,5	10,5	4+2+3+2
$(r_{1i} - r_{2i})^2$	9	0	2,25	4	9	0,25	12,25	0	4	0	6,25	47

Вычислим коэффициенты ранговой корреляции по приближенным и точной формулами. Согласно (7.2.6), (7.2.9) и (7.2.11) имеем $r_s \approx 0{,}784$; $r_s^* \approx 0{,}773$; $r_s^{**} \approx 0{,}772$. Из этого примера видно, что приближение (7.2.11) вполне приемлемо для практических расчетов коэффициента ранговой корреляции при совпадающих рангах.

Если первая ранжированная последовательность r_{1i} совпадает с рядом натуральных чисел, то получается n пар $(1, r_{21}), (2, r_{22}), \ldots, (n, r_{2n})$. При этом выборочный коэффициент ранговой корреляции преобразуется к виду

$$r_s = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(i - r_{2i})^2. \tag{7.2.12}$$

Впервые формулу (7.2.12) применил в 1904 г. американский ученый Спирмен при проведении психологических исследований. Поскольку в математической статистике имеются и другие подходы (соответственно и другие формулы для вычисления коэффициентов ранговой корреляции), формулы (7.2.6) и (7.2.12) называют *коэффициентом ранговой корреляции по Спирмену*.

Если ранги сравниваемых последовательностей идентичны ($r_{1i} = r_{2i}$), то $r_s = 1$. В случае противоположных рангов

$$\sum_{i=1}^{n}(r_{1i} - r_{2i})^2 = (1 - n)^2 + (3 - n)^2 + \ldots + (n - 3)^2 + (n - 1)^2 =$$

$$= \sum_{i=1}^{n}[n - (2i - 1)]^2 = \sum_{i=1}^{n}[n^2 - 4(n + 1)i + 4i^2 + (2n + 1)].$$

Учитывая соотношения (1.3.39), имеем

$$\sum_{i=1}^{n}(r_{1i} - r_{2i})^2 = \frac{n^3 - n}{3}.$$

Подставляя это выражение в (7.2.6), получаем, что в случае противоположных рангов выборочный коэффициент ранговой корреляции $r_s = -1$. Таким образом, как и общепринятый коэффициент корреляции, коэффициент ранговой корреляции изменяется в пределах $-1 \leqslant r_s \leqslant 1$.

Для вычисления соответствующих вероятностей коэффициента ранговой корреляции имеются таблицы [8, 51 и др.].

Если совместное распределение признаков является нормальным, то выборочный коэффициент корреляции r_{XY} (7.1.10) при $n \gg 10$ связан с выборочным коэффициентом ранговой корреляции r_s соотношением [13]

$$r_{XY} \approx 2\sin(\pi r_s/6). \tag{7.2.13}$$

Коэффициент ранговой корреляции часто используется в качестве статистики для проверки гипотезы о независимости двух признаков (нулевая гипотеза H_0). В этом случае его называют *статистикой Спирмена*, а критерий проверки гипотезы H_0, основанный на этой статистике, — *критерием Спирмена*.

Вычислим среднее значение и дисперсию выборочного коэффициента ранговой корреляции r_s (7.2.12) применительно к нулевой гипотезе, т.е. к ситуации, при которой ранги независимы. С учетом соотношений (1.3.39) имеем

$$\mathrm{M}\{r_s\} = 1 - \frac{6}{n_3 - n}\left[\sum_{i=1}^{n} i^2 - 2\sum_{i=1}^{n}\mathrm{M}\{r_{2i}\} + \sum_{i=1}^{n}\mathrm{M}\{r_{2i}^2\}\right] =$$

$$= 1 - \frac{6}{n_3 - n}\left[2\sum_{i=1}^{n} i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} i\right)^2\right] = 0, \tag{7.2.14}$$

т.е. для нулевой гипотезы среднее значение коэффициента ранговой корреляции равно нулю. Для вычисления дисперсии коэффициента ранговой корреляции формулу (7.2.2) с учетом введенного обозначения (7.2.5) перепишем в виде

$$r_s = \frac{1}{q}\sum_{i=1}^{n}(r_{1i} - R_{..})(r_{2i} - R_{..}). \tag{7.2.15}$$

Так как для нулевой гипотезы $\mathrm{M}\{r_s\} = 0$, дисперсия r_s определяется выражением

$$D(r_s) = \mathrm{M}\{r_s^2\} = \frac{1}{q^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{M}\{\xi_i\xi_j\zeta_i\zeta_j\}. \tag{7.2.16}$$

Здесь обозначено

$$\xi_i = r_{1i} - R_{..}; \quad \xi_j = r_{1j} - R_{..}; \quad \zeta_i = r_{2i} - R_{..}; \quad \zeta_j = r_{2j} - R_{..}.$$

Представим сумму (7.2.16) в виде двух сумм: первая сумма содержит n слагаемых с индексами $i = j$, а вторая сумма содержит $n(n-1)$ слагаемых с индексами $i \neq j$. При этом для рассматриваемой нулевой гипотезы справедливы равенства

$$\mathrm{M}\{\xi_i\xi_j\zeta_i\zeta_j\} = \begin{cases} \mathrm{M}\{\xi_i^2\}\mathrm{M}\{\zeta_i^2\} = (\mathrm{M}\{\xi_i^2\})^2, & i = j, \\ \mathrm{M}\{\xi_i\xi_j\}\mathrm{M}\{\zeta_i\zeta_j\} = (\mathrm{M}\{\xi_i\xi_j\})^2, & i \neq j. \end{cases}$$

Учитывая (7.2.5), $\mathrm{M}\{\xi_i^2\} = q/n$. Так как $\sum \xi_i = 0$, для вычисления смешанного центрального момента $\mathrm{M}\{\xi_i\xi_j\}$ при $i \neq j$ воспользуемся очевидным тождеством

$$0 = \mathrm{M}\{\xi_j\sum_{i=1}^{n}\xi_i\} = \mathrm{M}\{\xi_j^2\} + (n-1)\mathrm{M}\{\xi_i\xi_j\}, \quad i \neq j.$$

Откуда имеем

$$\mathrm{M}\{\xi_i\xi_j\} = -\frac{1}{n-1}\mathrm{M}\{\xi_j^2\} = -\frac{q}{n(n-1)}.$$

Подставляя вычисленные значения моментов в (7.2.16), получим формулу для дисперсии выборочного коэффициента ранговой корреляции по Спирмену для нулевой гипотезы

$$D(r_s) = \frac{1}{n-1}. \tag{7.2.17}$$

Учитывая линейную функциональную зависимость между случайными величинами r_s и S_r в (7.2.6), нетрудно вычислить среднее значение и дисперсию случайной величины S_r:

$$\mathrm{M}\{S_r\} = \frac{n^3 - n}{6}; \quad D(S_r) = \frac{n^2(n+1)^2(n-1)}{36}. \tag{7.2.18}$$

Аналогично можно вычислить и другие центральные моменты выборочного коэффициента ранговой корреляции. В [31] доказывается, что точное распределение выборочного коэффициента ранговой корреляции для нулевой гипотезы симметрично и, следовательно, его нечетные центральные моменты равны нулю. Четвертый центральный момент

$$\mathrm{M}\{r_s^4\} = \frac{3}{n^2-1}\left[1 + \frac{12(n-2)(n-3)}{25n(n-1)^2}\right].$$

Вывод распределения случайной величины r_s основан на том факте, что множество рангов r_{2i} представляет собой перестановку ряда натуральных чисел 1, 2,..., n. Таких перестановок $n!$, и при нулевой гипотезе все они равновероятны. В [2, 8] даны таблицы значений вероятностей $\mathbf{P}\{S_r \geqslant s_{r;\beta,n}\} = \beta$ случайной величины S_r (7.2.7) для малых объемов выборок ($n \leqslant 10$), позволяющие найти верхние и нижние критические значения выборочной статистики величины S_r и рангового коэффициента корреляции (выборочной статистики Спирмена) (7.2.6), обеспечи-

вающие уровень значимости α' не более заданного α.

Если объем выборки достаточно велик (практически, если $n \geqslant 10$), то согласно [13, 31] распределение выборочного рангового коэффициента корреляции можно аппроксимировать нормальным распределением с нулевым средним значением и дисперсией (7.2.17), что значительно облегчает задачу нахождения критических верхнего $r_{s;\alpha/2}$ и нижнего $r_{s;(1-\alpha)/2}$ значений выборочного коэффициента ранговой корреляции r_s с заданным двусторонним уровнем значимости α. При этом надежность принятия нулевой гипотезы увеличивается, так как действительный уровень значимости несколько меньше величины, получаемой на основе нормального приближения. Согласно нормальному приближению $r_{s;(1-\alpha/2)} = -r_{s;\alpha/2}$, а верхнее критическое значение $r_{s;\alpha/2}$ определяется из уравнения

$$1 - \Phi(t_{\alpha/2} = r_{s;\alpha/2}/\sqrt{D(r_s)}) = \alpha/2$$

с помощью таблиц критических значений $t_{\alpha/2}$ нормального распределения (табл. 1в Приложения 2). Нулевая гипотеза о том, что ранжированные признаки независимы, не отвергается, если выполняется неравенство

$$|r_s| \leqslant t_{\alpha/2}/\sqrt{n-1}. \qquad (7.2.19)$$

Пример 7.2.2. Проверим гипотезу независимости выборочных значений двух случайных величин, представленных в табл. 5.6.1, для уровня значимости $\alpha = 0{,}05$. Результаты ранжирования численных значений x_i и y_i приведены в табл. 7.2.2.

Таблица 7.2.2

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$r_{X_i} = r_{1i}$	8	4	9	5	16	1	14	7	10	2	15	3	6	11	12	13	17
$r_{Y_i} = r_{2i}$	7	2	12	3	10	14	9	6	4	8	11	16	13	5	17	15	1
$(r_{1i} - r_{2i})^2$	1	4	9	4	36	169	25	1	36	36	16	169	49	36	25	4	256

Подставляя значения величин $n = 17$ и $(r_{1i} - r_{2i})^2$ в формулу коэффициента ранговой корреляции (7.2.6), получим $r_s = -0{,}07$. Для вычисления критических значений выборочной статистики воспользуемся нормальным приближением, для которого согласно (7.2.17) $D(r_s) = 1/16$. При этом верхнее критическое значение $r_{s;0{,}025}$ определяется из уравнения

$$1 - \Phi[t_{0{,}025} = (r_{s;0{,}025}/0{,}25)] = 0{,}025.$$

Из табл. 1в Приложения 2 находим, что вероятности 0,025 соответствует критическое значение $t_{0{,}025} = 1{,}96$. Отсюда получаем верхнее критическое значение статистики коэффициента ранговой корреляции $r_{s;0{,}025} = 0{,}25 \cdot 1{,}96 = 0{,}49$. Так как выборочное значение коэффициента ранговой корреляции $|r_s| = 0{,}07 < 0{,}49$, то нет оснований для того, чтобы отвергнуть гипотезу о взаимной независимости анализируемых выборок.

Кроме коэффициента ранговой корреляции Спирмена r_s (7.2.6) иногда применяется *ранговый коэффициент корреляции Кендалла* τ [8, 13, 27, 30]:

$$\tau = \frac{2S}{n(n-1)}; \quad S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \operatorname{sign}(r_j - r_i). \qquad (7.2.20)$$

Здесь знаковая функция $\operatorname{sign}(a)$ принимает два значения:

$$\operatorname{sign}(a) = \begin{cases} +1, & a > 0, \\ -1, & a < 0. \end{cases}$$

Ранговый коэффициент корреляции по Кендаллу изменяется в интервале $-1 \leqslant \tau \leqslant +1$. Величину τ называют *статистикой Кендалла*, а критерий проверки нулевой гипотезы H_0, основанный на этой статистике, называют *критерием Кендалла*. В [13, 51] приведены таблицы распределения статистики Кендалла, позволяющие по заданному уровню значимости α вычислить критические значения верхней и нижней границ выборочной статистики (7.3.20). В указанной выше литературе также показывается, что для нулевой гипотезы среднее значение статистики τ равно нулю ($\mathrm{M}\{\tau\} = 0$), а дисперсия этой статистики равна $D(\tau) = 2(2n+5)/[9n(n-1)]$, причем при $n \gg 1$ распределение статистики τ асимптотически стремится к нормальному распределению с нулевым средним значением и дисперсией $D(\tau) = 4/(9n)$. При этом отмечается, что критерии Спирмена и Кендалла асимптотически эквивалентны. Кроме того, в [13] показывается, что критерий Спирмена по сравнению с критерием Кендалла имеет приблизительно одинаковый уровень значимости и большую мощность, а вычисление выборочной статистики r_s *существенно проще вычисления* выборочной статистики τ. Поэтому можно утверждать, что коэффициент ранговой корреляции r_s по Спирмену предпочтительнее во многих отношениях по сравнению с коэффициентом ранговой корреляции τ по Кендаллу.

7.2.2. Коэффициент ранговой согласованности

Если число исследуемых признаков (наблюдений) m больше двух, то результаты ранжирования по этим признакам можно записать в виде матрицы, j-я строка которой содержит результат «ранжировки» по j-му признаку, $j = 1, 2, \ldots, m$, а столбцы соответствуют наблюдениям (индивидуумам, признакам и т.п.), принадлежащим исследуемой выборке:

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix}. \qquad (7.2.21)$$

Матрица содержит $m(m-1)/2$ пар признаков, и для каждой пары признаков может быть вычислен коэффициент ранговой корреляции. Следует также отметить, что в первой строке вместо $r_{11}, r_{12}, \ldots, r_{1n}$

можно было бы написать ранжированный ряд в виде натурального ряда чисел $1, 2, \ldots, n$.

В качестве выборочной меры связи m признаков Кендаллом предложено использовать *коэффициент согласованности* (согласия), называемый также *коэффициентом конкордации*,

$$W = \sum_{i=1}^{n}(R_i - R)^2 \Big/ \max \sum_{i=1}^{n}(R_i - R)^2. \qquad (7.2.22)$$

Здесь числитель представляет собой сумму квадратов отклонений сумм рангов по столбцам матрицы

$$R_i = \sum_{j=1}^{m} r_{ji} \qquad (7.2.23)$$

от их среднего значения

$$R = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ji} = \frac{m(n+1)}{2}. \qquad (7.2.24)$$

Из определения коэффициента согласованности следует, что он положителен и изменяется от 0 до 1.

Изучим сумму квадратов отклонения R_i от среднего значения R

$$\sum_{i=1}^{n}(R_i - R)^2 = \sum_{i=1}^{n} R_i^2 - nR^2. \qquad (7.2.25)$$

Максимальное значение этой величины будет тогда, когда ранжированные последовательности (наблюдения) одинаковы. Не уменьшая общности, пусть первая строка состоит из последовательности натуральных чисел. Тогда для одинаково ранжированных последовательностей, учитывая соотношение (1.3.39), имеем

$$R_i = \sum_{j=1}^{m} i = mi; \quad \sum_{i=1}^{n} R_i^2 = m^2 \sum_{i=1}^{n} i^2 = \frac{m^2 n(n+1)(2n+1)}{6}.$$

Подставляя эти выражения в (7.2.25), получаем максимальное значение, равное

$$\max \sum_{i=1}^{n}(R_i - R)^2 = \frac{m^2(n^3 - n)}{12}. \qquad (7.2.26)$$

Поэтому выражение для коэффициента согласованности W можно записать в виде

$$W = \frac{12 S_W}{m^2(n^3 - n)}, \qquad (7.2.27)$$

где

$$S_W = \sum_{i=1}^{n}(R_i - R)^2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{m} r_{ji} - \frac{m(n+1)}{2}\right]^2. \qquad (7.2.28)$$

Коэффициент согласованности W обычно используется для проверки гипотезы независимости признаков (нулевая гипотеза H_0). Признаки называются независимыми, если для наугад выбранного столбца матрицы (7.2.21) i-е ранги (порядковые номера) $r_{1i}, r_{2i}, \ldots, r_{ji}, \ldots, r_{mi}$ будут взаимно независимыми случайными величинами. При этом случайные величины R_i (7.2.23) также будут взаимно независимыми случайными величинами.

Применительно к нулевой гипотезе H_0 вычислим среднее значение случайных величин S_W и W. В рассматриваемой ситуации последовательности r_{ji} в каждой строке равновероятны со средним значением

$$m_j(r) = \mathrm{M}\{r_{ji}\} = (n+1)/2. \qquad (7.2.29)$$

и дисперсией

$$D_j(r) = \mathrm{M}\{[r_{ji} - m_j(r)]^2\} = (n^2 - 1)/12. \qquad (7.2.30)$$

Учитывая независимость рангов для нулевой гипотезы, среднее значение случайной величины S_W определяется выражением

$$\mathrm{M}\{S_W\} = \sum_{i=1}^{n}\mathrm{M}\{(R_i - R)^2\} =$$

$$= \sum_{i=1}^{n}\mathrm{M}\left\{\sum_{j=1}^{m}\sum_{k=1}^{m}\left(r_{ji} - \frac{n+1}{2}\right)\left(r_{ki} - \frac{n+1}{2}\right)\right\} =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{M}\left\{\left(r_{ji} - \frac{n+1}{2}\right)^2\right\} = \sum_{i=1}^{n}\sum_{j=1}^{m} D(r_{ji}) = \frac{nm(n^2-1)}{12}. \qquad (7.2.31)$$

При этом дисперсия случайной величины R_i для нулевой гипотезы

$$D = \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{M}\{(R_i - R)^2\} = \frac{nm(n+1)}{12}. \qquad (7.2.32)$$

Соответственно, для среднего значения коэффициента согласованности имеем

$$\mathrm{M}\{W\} = m_W = 1/m. \qquad (7.2.33)$$

Аналогично вычисляются дисперсии случайных величин W и S_W для

нулевой гипотезы; они равны [8]

$$D(W) = \frac{2(m-1)}{m^3(n-1)}; \tag{7.2.34}$$

$$D(S_W) = \frac{m(m-1)(n+1)(n^2-1)n^2}{72}. \tag{7.2.35}$$

Применительно к нулевой гипотезе в [2, 8 и др.] для малых величин n и m приведены таблицы точных значений распределения коэффициента согласованности S_W (7.2.28): $Q_W(s_W; m, n) = \mathbf{P}\{S_W > s_W\}$. В [2], кроме того, имеется таблица, позволяющая для уровня значимости $\alpha = 0,05$ и малых значений n и m вычислить критическое значение границы $s_{\text{кр.}W}(\alpha; m, n)$ выборочного коэффициента s_W. Поскольку статистики S_W и W связаны линейной зависимостью, вывод о принятии или отказа от нулевой гипотезы на основе статистики (7.2.28) справедлив и для статистики (7.2.27).

Вместо статистики (7.2.27) часто применяют статистику

$$V = \frac{1}{D}\sum_{i=1}^{n}(R_i - R)^2 = m(n-1)W =$$

$$= \frac{12}{mn(n+1)}\sum_{i=1}^{n}\left[\sum_{j=1}^{m} r_{ji} - \frac{m(n+1)}{2}\right]^2 =$$

$$= \frac{12S_W}{mn(n+1)} = \frac{12}{mn(n+1)}\sum_{i=1}^{n}\left(\sum_{j=1}^{m} r_{ij}\right)^2 - 3m(n+1), \tag{7.2.36}$$

которую, как отмечалось выше в п. 6.3, называют *статистикой Фридмана* [17, 51, 68], а правило принятия решения на основе этой статистики — *критерием Фридмана*. В [51] приведены критические значения статистики (7.2.36) для числа столбцов $n = 3$ и 4.

В прикладных задачах часто приходится встречаться с ситуациями, когда число наблюдений (индивидуумов) достаточно велико ($m \gg 10$), а таблицы соответствующих вероятностей отсутствуют. В подобных ситуациях для проверки нулевой гипотезы можно воспользоваться приближенным решением, основанным на том, что случайная величина (статистика) (7.2.36) асимптотически распределена по закону хи-квадрат с $\nu = (n-1)$ степенями свободы. Поэтому для проверки нулевой гипотезы H_0 о независимости признаков с односторонним уровнем значимости α, как и выше, по таблицам хи-квадрат распределения для степеней свободы $\nu = (n-1)$ находится верхнее критическое значение статистики $V_{\alpha;\nu} \equiv \chi^2_{\alpha;\nu}$. Это значение статистики сравнивается со значением выборочной статистики коэффициента V^*, вычисленного согласно (7.2.36) и линейно связанного с выборочным коэффициентом согласованности W^* (7.2.27). Для этого в (7.2.36) подставляются выборочные значения

величин R_i или r_{ji}, полученные на основе экспериментальных данных о рангах, указанных в матрице (7.2.21). Гипотеза случайности признаков (рангов) отклоняется, если для выборочной статистики V^* и одностороннего уровня значимости α выполняется условие $V^* \geqslant \chi^2_{\alpha;\nu}$.

Аналогично обсуждавшемуся выше (п. 5.4.1) применительно к гипотезе H_0, приближение точного распределения коэффициента согласованности распределением хи-квадрат вполне приемлемо для прикладных задач уже при $n \geqslant 5$. При этом полученное решение обеспечивает большую надежность в отношении справедливости нулевой гипотезы.

Коэффициент W (следовательно, и V) назван коэффициентом согласованности в связи с проведением экспертной оценки некоторого изделия (системы, события) по n параметрам экспертами, число которых m. Проверяется гипотеза о случайности (или не случайности) m мнений (рангов) экспертов. В данном случае в качестве наблюдений (событий, индивидуумов и т.п.) выступают n параметров изделия, а в качестве признаков фигурируют номера экспертов j, $j = 1, 2, \ldots, m$. Если при проведении экспертизы ранги, присвоенные экспертами по каждому из параметров, близки, а нулевая гипотеза в результате этого отвергается с достаточно малым уровнем значимости α, то говорят, что мнения экспертов согласованы с доверительной вероятностью $\gamma = 1 - \alpha$. Естественно, как и выше, остается открытым вопрос: что значит достаточно малый уровень значимости? Обычно при проведении экспертных оценок для отклонения гипотезы случайности мнений экспертов, уровень значимости выбирают менее 0,01.

Если ранги каких-либо наблюдений (параметров) совпадают, то для этих наблюдений (параметров) назначаются ранги, равные среднему арифметическому из рангов смежных членов ряда. При этом формула (7.2.27) принимает вид [2, 68 и др.]

$$\widetilde{W} = \frac{S_W}{m^2(n^3-n)/12 - m\sum_{j=1}^{m} T_j}. \tag{7.2.37}$$

Здесь

$$T_j = \frac{1}{12}\sum_{k=1}^{g_j}(t_{jk}^3 - t_{jk}), \tag{7.2.38}$$

а t_{jk} — число совпадений в k-й группе j-й строки.

Пример 7.2.3. В табл. 7.2.3 приведена матрица рангов по шести показателям ($n = 6$) системы десяти экспертов ($m = 10$). Проверим гипотезу о том, что мнения экспертов случайны (не согласованы) для уровня значимости $\alpha = 0,005$.

Для определения искомых характеристик по формулам (7.2.23) и (7.2.24) необходимо вычислить суммы рангов, их среднее значение, а также сумму квадратов отклонений суммы рангов по столбцам от их среднего значения (7.2.25). Среднее значение сумм рангов равно $R = m(n+1)/2 = 35$. В табл. 7.2.4 приведены суммы рангов и квадраты отклонения этих сумм от их среднего значения. По формулам (7.2.27) и (7.2.36) определим выборочные значения коэффициента согласованности W^* и статистики V^*:

$$W^* = \frac{12 \cdot 1125,5}{100 \cdot 210} = 0,643; \quad V^* = \frac{12 \cdot 1125,5}{10 \cdot 6 \cdot 7} = 32,16.$$

Таблица 7.2.3

Ранги экспертов R_{ji}

Номер эксперта (признака) j	Параметры системы (наблюдения)					
	1	2	3	4	5	6
1	1	3	2	5	6	4
2	1,5	4	1,5	6	5	3
3	1	4	2	5	6	3
4	1	4	2	6	5	3
5	1	4,5	3	2	6	4,5
6	1	6	5	2	3	4
7	1	5	2	4	3	6
8	1	3	2	5	6	4
9	1	3	2	5,5	5,5	4
10	1	2	3	5	6	4

Таблица 7.2.4

i	1	2	3	4	5	6	$\sum$
R_i	10,5	38,5	24,5	45,5	51,5	39,5	210
$(R_i - R)^2$	600,25	12,25	110.25	110,25	272,25	20,25	1125,5

Заданному уровню значимости $\alpha = 0{,}005$ и числу степеней свободы $\nu = n - 1 = 5$ согласно табл. 2 Приложения 2 соответствует критическое значение $\chi^2_{0{,}005;5} = 16{,}75$. Так как $V^* = 32{,}16 > \chi^2_{0{,}005;5} = 16{,}75$, то нулевая гипотеза о случайности мнений экспертов отвергается и, следовательно, можно утверждать, что мнения экспертов близки между собой с коэффициентом согласия $W^* = 0{,}643$ при доверительной вероятности $\gamma = 1 - \alpha = 0{,}995$. Значения коэффициентов согласованности несколько изменятся, если учесть совпадение рангов во второй, пятой и девятой строках таблицы рангов. При этом $T_2 = T_5 = T_9 = (2^3 - 2)/12 = 0{,}5$. Тогда согласно (7.2.37) имеем $\widetilde{W} \approx 0{,}648$; $\widetilde{V}^* \approx 32{,}42$. Это изменение незначительно, и поэтому решение о согласованности мнений экспертов с заданным уровнем значимости не изменяется.

Контрольные вопросы и задачи

7.1. В каких случаях выборочный коэффициент корреляции применяют в качестве меры зависимости двух выборочных совокупностей?

7.2. В чем сущность ранговых методов исследования статистической связи?

7.3. По двум выборкам из $n_1 = 16$ и $n_2 = 25$ пар независимых наблюдений гауссовских случайных величин вычислены выборочные коэффициеты корреляции $r_1 = 0{,}5$ и $r_2 = 0{,}7$ соответственно. Проверить гипотезу (с уровнем значимости $\alpha = 0{,}05$) о том, что обе выборки относятся к одной генеральной совокупности.

7.4. Получены три гауссовские выборки, каждая объема $n = 25$. На основе этих данных были определены выборочные парные коэффициенты корреляции $r_{12} = 0{,}15$; $r_{13} = 0{,}05$; $r_{23} = 0{,}98$:

а) вычислить выборочные частные коэффициенты корреляции $r_{12.(3)}$, $r_{13.(2)}$, $r_{23.(1)}$ и выборочный множественный коэффициент корреляции $r_{1.(23)}$;

б) указать доверительный интервал для $\rho_{12.(3)}$ с доверительной вероятностью $\gamma = 0{,}95$;

в) проверить гипотезу о том, что выборочный множественный коэффициент корреляции $r_{1.(23)}$ несущественно отличается от нуля с уровнем значимости $\alpha = 0{,}05$).

7.5. Выборочный коэффициент ранговой корреляции между двумя гауссовскими выборками равен $r_s = 0{,}8$. Чему равна оценка для истинного коэффициента корреляции?

7.6. Проверить с уровнем значимости $\alpha = 0{,}005$ гипотезу о том, что мнения 20 экспертов ($m = 20$) не согласованы при экспертной оценке по 15 признакам ($n = 15$), если выборочное значение коэффициента конкордации $W^* = 0{,}25$.

7.7. Имеются две ранжированные последовательности из 10 признаков, представленных в таблице.

r_{1i}	5,5	3,5	10,0	8,0	5,5	11	2,5	2,5	8	5,5
r_{2i}	4,5	2,5	6,0	10,0	5,5	5,0	5,5	6,0	2,5	11

Вычислить коэффициенты ранговой корреляции по формулам, не учитывающим и учитывающим совпадающие ранги.

7.8. Проверить гипотезу независимости выборочных значений двух случайных величин X и Y, представленных в таблице в виде ранжированных последовательностей, для уровня значимости значимости $\alpha = 0{,}05$.

r_{1i}	11	13	10	12	16	14	9	8	7	15
r_{2i}	15	12	16	9	8	10	14	17	11	9

ГЛАВА 8

Регрессионный анализ

8.1. Основные положения

В прикладных задачах математической статистики часто возникает необходимость прогнозировать (предсказывать) одну или несколько величин $y_1, y_2, \ldots, y_h$ на основе информации, доставляемой другими измеряемыми величинами $x_1, x_2, \ldots, x_p$, независимыми от $y_1, y_2, \ldots, y_h$. Последние статистически связаны с величинами $x_1, x_2, \ldots, x_p$, которые называют *факторами или предикторами* (т.е. предсказателями). Зависимые прогнозируемые величины $y_1, y_2, \ldots, y_h$ часто называют *откликами*. В [2] независимые и зависимые величины названы также объясняющими и оъясняемыми (величинами) соответственно.

Методы решения подобных задач носят название регрессионного анализа. Итак, *регрессионный анализ — это метод предсказания одних случайных величин по значениям других случайных величин*, связанных с первыми статистическими (стохастическими) зависимостями. Для понимания сущности регрессионного анализа будем полагать, что имеются две статистически связанные случайные величины X и Y. Необходимо установить статистическую функциональную зависимость прогнозируемого значения Y на основе измерения или задания величины $X = x$. Под статистической функциональной зависимостью понимается зависимость «в среднем». Рассмотренные выше в п. 1.5.3 регрессии первого и второго рода требуют определенных априорных знаний относительно статистических характеристик анализируемых случайных величин. Кроме этого в случае регрессии второго рода (с точностью до неизвестных и подлежащих оценке коэффициентов) должна быть известна статистическая функциональная зависимость. Однако в большинстве исследований эта информация отсутствует. Поэтому все необходимые сведения о статистической связи между случайными величинами могут быть получены только на основе опытных данных.

Пусть проведено n серий измерений, при которых определяются значения x_i и соответствующие им k значений y_{ik}, $k = 1, 2, \cdots, n_i$. Как и выше, полагается, что выборочные значения y_{ik} между собой независимы.

Предположим, что данные измерения располагаются, как указано в табл. 8.1.1 в первых двух столбцах.

Таблица 8.1.1

Независимая величина x_i	Зависимая величина y_i	Выборочное среднее $\bar{y}_i$	Математическое ожидание $\mathrm{M}\{y_i\} = g(x_i)$
x_1	$y_{11}, y_{12}, \ldots, y_{1k}$	$\bar{y}_1 \equiv y_{1\cdot}$	$\mathrm{M}\{y_1\} = g(x_1)$
x_2	$y_{21}, y_{22}, \ldots, y_{2k}$	$\bar{y}_2 \equiv y_{2\cdot}$	$\mathrm{M}\{y_2\} = g(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	$y_{n1}, y_{n2}, \ldots, y_{nk}$	$\bar{y}_n \equiv y_{n\cdot}$	$\mathrm{M}\{y_n\} = g(x_n)$

На основе профессиональных сведений выдвигается гипотеза относительно типа теоретической функции зависимости $\tilde{y} = g(x; b_0, b_1, \ldots, b_p)$, где $\mathrm{b}(b_0, b_1, \ldots, b_p)$ – независимые между собой неизвестные постоянные коэффициенты (параметры).

По экспериментальным данным таблицы необходимо построить оценку $y^* = g^*(x; \mathrm{b}^*)$ теоретической функции зависимости таким образом, чтобы $\mathrm{M}\{y\} = \tilde{y} \equiv g(x; \mathrm{b})$.

Итак будем полагать, что вид статистической функциональной зависимости между случайными величинами X и Y задан (или выбран), т.е. среднее значение случайной величины Y при фиксированном значении случайной величины $X = x$ (значение теоретической кривой регрессии $\tilde{y}$) определяется соотношением

$$\tilde{y} \equiv \mathrm{M}\{y\} = g(x; b_0, b_1, \ldots, b_p). \tag{8.1.1}$$

Случайную величину y_{ik}, описывающую результат k-го значения i-го опыта, как правило, представляют в виде

$$y_{ik} = g(x_i; b_0, b_1, \ldots, b_p) + \varepsilon_{ik}, \tag{8.1.2}$$

где ε_{ik} – случайная ошибка с нулевым средним значеним $\mathrm{M}\{\varepsilon_{ik}\} = 0$. Величину ε_{ik} обычно называется остатком или «остаточной» составляющей; она характеризует отклонение истинной зависимости от полученной в результате опыта выборочной регрессии y_i^*. Величина ε_{ik} обусловлена влиянием на формирование регрессии как случайных погрешностей при измерении случайной величины Y, так и влиянием неучтенных факторов, которые могут присутствовать при проведении эксперимента. Далее также полагается, что случайные величины ε_{ik} независимы между собой и независимы от неизвестных параметров регрессии b_j, $j = 0, 1, \ldots, p$. По имеющимся $n \geqslant p + 1$ сериям значений (x_i, y_{ik}) для $i = 1, 2, \ldots, n$ оценки неизвестных параметров регрессии согласно методу наименьших квадратов находят из условия обращения в минимум квадратичной формы

$$G = \sum_{i=1}^{n} \sum_{k=1}^{n_i} [y_{ik} - g(x_i; b_0, b_1, \ldots, b_p)]^2, \tag{8.1.3}$$

которая представляет собой сумму квадратов разностей между наблюдениями и их средними значениями (математическими ожиданиями).

Как отмечалось в п. 3.5.5, минимум функционала по значениям неизвестных параметров эквивалентен его минимуму по каждому из коэффициентов b_j, $j = 1, 2, \ldots, p$. Эти значения (оценки) b_j^* определяются из решения системы уравнений вида

$$\left.\frac{\partial G}{\partial b_j}\right|_{\mathbf{b}=\mathbf{b}^*} = 0 \quad \text{при} \quad \mathbf{b}^*(b_0^*, b_1^*, \ldots, b_p^*); \; j = 0, 1, \ldots, p. \tag{8.1.4}$$

На основе полученных оценок b_j^* и заданной функциональной зависимости (8.1.1) находится оценочное значение случайной величины y_i^* при заданном значении величины x_i:

$$y_i^* = g(x_i; b_0^*, b_1^*, \ldots, b_p^*). \tag{8.1.5}$$

Найденное оценочное значение y_i^* случайной величины Y является *выборочной средней квадратической регрессией* Y на X. Естественно, оценочное значение случайной величины y_i^*, как правило, не совпадает не только с наблюденным значением y_{ik}, но и с выборочным средним

$$y_{i\cdot} \equiv \bar{y}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} y_{ik}. \tag{8.1.6}$$

В теории выборочной регрессии в дальнейшем предполагается выполнение следующих условий.

1. При каждом значении x_i величины y_{ik} распределены по нормальному закону.

2. Условное среднее значение случайных величин y_{ik}

$$\mathrm{M}\{y_{ik}\} \equiv \mathrm{M}\{y_{ik} \mid x_i\} = m_Y(x_i) = g(x_i; b_0, b_1, \ldots, b_p), \tag{8.1.7}$$

где $\mathbf{b}(b_0, b_1, \ldots, b_p)$ — истинные (точные) значения неизвестных параметров.

3. Условная дисперсия случайных величин y_{ik}, соответствующая данному значению x_i, которую называют остаточной дисперсией, постоянна

$$\mathrm{M}\{[y_{ik} - g(x_i; b_0, b_1, \ldots, b_p)]^2\} = \mathrm{M}\{\varepsilon_{ik}^2\} = D_{Y|x} \equiv D = \mathrm{const} \tag{8.1.8}$$

или пропорциональна некоторой функции $q^2(x)$,т.е.

$$\mathrm{M}\{\varepsilon_{ik}^2\} = D_{Y|x} = Dq^2(x).$$

Однако в дальнейшем для упрощения рассматривается только первый случай. Особенности теории выборочной регрессии для условной дисперсии, зависящей от переменной x, изложены в [2, 23, 64 и др.].

Метод наименьших квадратов для нахождения оценок неизвестных параметров является частным случаем метода максимального правдоподобия, рассмотренного в п. 4.4.2. Действительно, функция правдоподобия параметров $\mathbf{b}(b_0, b_1, \ldots, b_p)$ при анализе n пар независимых выборочных значений (x_i, y_{ik}) и сделанных выше предположениях имеет вид

$$L(\mathbf{b} \mid \mathbf{y}) = \frac{1}{(2\pi D_Y)^{N/2}} \exp\left\{-\frac{1}{2D_Y}\sum_{i=1}^{n}\sum_{k=1}^{n_i}[y_{ik} - g(x_i; \mathbf{b})]^2\right\}. \tag{8.1.9}$$

Здесь

$$N = \sum_{i=1}^{n} n_i \tag{8.1.10}$$

— суммарное число выборочных значений случайной величины Y.

Функция правдоподобия (8.1.9) принимает максимальное значение при тех значениях оцениваемых параметров $\mathbf{b}^*$, для которых сумма в фигурных скобках будет минимальна, что полностью совпадает с (8.1.3). Поэтому, если существуют эффективные оценки рассматриваемых неизвестных параметров, то они могут быть найдены методом наименьших квадратов.

Для определения надежности полученных оценок параметров b_j^* и регрессии y^* следует определить их статистические характеристики (точности и доверительные границы при заданной доверительной вероятности). При сделанных предположениях оценки параметров b_j, полученные методом наименьших квадратов, являются «наилучшими» в том смысле, что они имеют нормальное распределение со средними значениями, равными искомым параметрам $\mathrm{M}\{b_j^*\} = b_j$ и с наименьшими возможными дисперсиями $D(b_j^*) = \min$. Кроме этого, случайная величина

$$\chi_\nu^2 = \frac{1}{D}\sum_{i=1}^{n}\sum_{k=1}^{n_i}[y_{ik} - g(x_i; b_0^*, b_1^*, \ldots, b_p^*)]^2 \tag{8.1.11}$$

имеет хи-квадрат распределение с $\nu = N - (p+1)$ степенями свободы. Уменьшение числа степеней свободы на $p+1$ по сравнению с N независимыми измерениями y_{ik} обусловлено тем, что $p+1$ из них использованы для определения $p+1$ параметров b_j.

Гипотеза относительно постоянства дисперсии D проверяется сравнением частных выборочных дисперсий

$$s_{Ri}^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(y_{ik} - y_{i\cdot})^2, \quad i = 1, 2, \ldots, n \tag{8.1.12}$$

по критерию Бартлета (см. п. 5.3.2). Если гипотеза постоянства дисперсий не отвергается с заданным уровнем значимости, то n дисперсий объединяются в оценку

$$s_R^2 = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik} - y_{i\cdot})^2}{\sum_{i=1}^{n} n_i - n}. \tag{8.1.13}$$

Принципиально относительно распределения величины x не делается никаких предположений. Регрессивный анализ производится одним и тем же способом независимо от того, является ли x случайной или неслучайной величиной и может ли принимать x все значения из некоторого интервала или только дискретные значения. Например, в экспериментах для установления связи между величинами y и x величина x часто является неслучайной, ее значения могут быть определены при планировании эксперимента и достаточно точно заданы априори.

В прикладных задачах достаточно часто приходится сталкиваться с ситуацией, когда вид регрессии точно не известен. Поэтому, принимая тот или иной вид регрессии и вычисляя неизвестные ее коэффициенты b_j^*, $j = 1, 2, \ldots, p$, необходимо (с заданным уровнем значимости) проверить гипотезу о том, что *принятая модель регрессии адекватна экспериментальным данным*. Как и в дисперсионном анализе, полная сумма квадратов отклонений выборочных значений относительно прогнозируемых значений разбивается на составляющие путем введения выборочных средних $y_{i\cdot}$. Для этого представим общую сумму квадратов отклонений y_{ik} от оценок y_i^* в виде

$$\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_i^*)^2 = \sum_{i=1}^{n}\sum_{k=1}^{n_i}[(y_{ik}-y_{i\cdot})+(y_{i\cdot}-y_i^*)]^2 =$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot})^2 + \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{i\cdot}-y_i^*)^2 + 2\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot})(y_{i\cdot}-y_i^*).$$

Так как

$$\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot})(y_{i\cdot}-y_i^*) = \sum_{i=1}^{n}(y_{i\cdot}-y_i^*)\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot}) = 0,$$

предыдущее соотношение представим как

$$Q = Q_A + Q_R. \tag{8.1.14}$$

Случайная составляющая

$$Q = \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_i^*)^2 \tag{8.1.15}$$

представляет собой суммарный квадрат отклонения измеренных значений y_{ik} относительно прогнозируемого значения y_i^* с числом степеней свободы $\nu = N - (p+1)$.

Случайная величина

$$Q_A = \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{i\cdot}-y_i^*)^2 = \sum_{i=1}^{n} n_i(y_{i\cdot}-y_i^*)^2 \tag{8.1.16}$$

представляет собой суммарный квадрат отклонения выборочных средних

случайной величины $y_{i\cdot}$ относительно прогнозируемого значения y_i^* с числом степеней свободы $\nu_A = n-(p+1)$ и характеризует *адекватность модели*. Для вычисления $y_{i\cdot}$ должно выполняться условие $n_i \geqslant 2$.

Случайная величина

$$Q_R = \sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot})^2 \tag{8.1.17}$$

представляет собой суммарный квадрат отклонения измеренных значений y_{ik} относительно выборочных средних $y_{i\cdot}$ с числом степеней свободы $\nu_R = N - n$ и характеризует суммарную ошибку измерения (наблюдения) случайной величины Y.

Из приведенных выше выражений для степеней свободы нетрудно видеть, что

$$\nu = \nu_A + \nu_R.$$

Так же, как в дисперсионном анализе (п. 6.1), можно показать, что применительно к нулевой гипотезе (гипотетический вид регрессии не противоречит опытным данным) выборочные дисперсии

$$s^2 = \frac{Q}{N-p-1}; \quad s_A^2 = \frac{Q_A}{n-p-1}; \quad s_R^2 = \frac{Q_R}{N-n} \tag{8.1.18}$$

оценивают одну и ту же дисперсию D. Величину s_A^2 называют *дисперсией адекватности модели*. Величину s_R^2 называют *дисперсией ошибки наблюдения*. Поэтому в качестве *критерия адекватности рассматриваемой регрессии (модели) экспериментальным данным применяется критерий дисперсионного отношения*, статистика для которого записывается как

$$F_{\nu_A,\nu_R} = \frac{s_A^2}{s_R^2} = \frac{Q_A/\nu_A}{Q_R/\nu_R} = \frac{\dfrac{1}{n-p-1}\displaystyle\sum_{i=1}^{n} n_i(y_{i\cdot}-y_i^*)^2}{\dfrac{1}{N-n}\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik}-y_{i\cdot})^2}. \tag{8.1.19}$$

Если для заданного уровня значимости α выборочное значение дисперсионного отношения $F^*_{\nu_A,\nu_R}$ не превышает критического значения $F_{\alpha;\nu_A,\nu_R}$, то считается, что для этих исходных данных гипотеза об адекватности модели не противоречит экспериментальным данным.

Полученная кривая регрессии $y_i^* = g(x_i; b_0^*, b_1^*, \ldots, b_p^*)$ описывает выборочное среднее искомой зависимости y^* как функцию значения аргумента x лишь в пределах его изменения. Если представляет интерес прогнозируемое значение зависимости $y_i^* = g(x_i; b_0^*, b_1^*, \ldots, b_p^*)$, выходящее за пределы изменения фактора x, то выбор гипотетической функции становится существенно важным. Это обусловлено тем, что если выбранная гипотетическая функция незначимо отличается от истинной функции в пределах наблюдений, она может значительно отличаться в

более широком интервале значений фактора x. Поэтому каждая экстраполяция заключает в себе элемент неопределенности, который невозможно оценить статистическими методами. В связи с этим при выборе гипотетической функциональной зависимости необходимо учитывать все профессиональные данные относительно связи между величинами.

Приведенные выше выражения несколько упрощаются, если для каждого значения независимой переменной x_i число измерений величины y_{ik} одинаково, т.е. если $n_i \equiv \tilde{n} = \text{const}$.

В прикладных задачах часто для каждого x_i можно получить только одно значение y_i. В подобной ситуации для проверки гипотезы об адекватности выбранной гипотетической модели экспериментальным данным статистика (8.1.19) не может быть использована. Для данного случая возможен приближенный метод проверки гипотезы адекватности модели, основанный на группировании экспериментальных данных [2]. В этом случае выбирается число интервалов группирования ν, которое должно быть не меньше числа неизвестных параметров регрессии $p+1$. Затем подсчитывается число n_j^* выборочных значений y_{ij}, попавших в j-й интервал группирования $[x_{j-1}, x_j]$. При этом формируется статистика дисперсионного отношения вида

$$F_{\nu-p-1,n-\nu} = \frac{\dfrac{1}{\nu-p-1}\displaystyle\sum_{j=1}^{n} n_j^*(y_{j\cdot}-y_j^*)^2}{\dfrac{1}{n-\nu}\displaystyle\sum_{j=1}^{\nu}\sum_{i=1}^{n_i}(y_{ij}-y_{j\cdot})^2}, \tag{8.1.20}$$

где

$$y_{j\cdot} = \frac{1}{n_j^*}\sum_{i=1}^{n_j^*} y_{ij} \tag{8.1.21}$$

— выборочное среднее величин y_{ij}, попавших в j-й интервал группирования.

Проверка гипотезы адекватности модели осуществляется как обычно путем задания уровня значимости критерия α, определения критического значения статистики $F_{\alpha;\nu-(p+1),n-\nu}$ и сравнения его с выборочной статистикой $F^*_{\nu-(p+1),n-\nu}$.

Для произвольной зависимости y от параметров $\mathfrak{b}(b_0, b_1, \ldots, b_p)$ оценка (вычисление) неизвестных параметров регрессии, а также проверка соответствующих гипотез относительно значимости параметров модели и ее адекватности, как правило, представляют значительные трудности. Поэтому из всей совокупности возможных зависимостей выделяют зависимости, являющиеся линейными функциями относительно параметров $b_0, b_1, \ldots, b_p$. Например, к числу подобных функций относятся

$$\tilde{y} = b_0 + b_1 x + b_2 x^2; \qquad \tilde{y} = b_0 + b_1 x + b_2 \log x \tag{8.1.22}$$

Особое внимание в регрессионном анализе уделяется линейной зависимости как от параметров регрессии $\mathfrak{b}(b_0, b_1, \ldots, b_p)$, так и от значений независимых переменных x_j (факторов):

$$\tilde{y} = b_0 + \sum_{j=1}^{p} b_j x_j. \tag{8.1.23}$$

Теория линейной регрессии разработана достаточно полно и наиболее широко используется при решении прикладных задач математической статистики.

По сложности задач нелинейные модели (модели, нелинейные по оцениваемым параметрам $\mathfrak{b}(b_0, b_1, \ldots, b_p)$) разделяются на два класса: внутренне линейные и внутренне нелинейные [23].

Для внутренне линейных моделей с помощью соответствующих преобразований нелинейную регрессию можно свести к линейной регрессии. Например, если рассматривается экспоненциальная модель вида

$$y = e^{b_0+b_1x_1+b_2x_2+\varepsilon}, \tag{8.1.24}$$

то, логарифмируя обе части уравнения, получим

$$z \equiv \ln y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon. \tag{8.1.25}$$

Преобразованная модель (8.1.25) имеет форму линейной регрессии. В подобной ситуации решение задачи нахождения оценок неизвестных параметров гипотетической нелинейной модели и проверка адекватности ее экспериментальным данным существенно упрощаются. При этом необходимо подчеркнуть, что для проведения соответствующего статистического анализа модели необходимо, чтобы величина ε была распределена по нормальному закону с нулевым средним значением и постоянной дисперсией.

Примерами внутренне нелинейных моделей регрессии являются

$$y = b_0 + b_1 e^{-b_2 x} + \varepsilon; \qquad y = b_0 + b_1 x + (b_2)^x + \varepsilon \tag{8.1.26}$$

Статистический анализ таких моделей представляет собой достаточно сложную задачу [23] и на практике используется весьма редко.

В ряде прикладных задач объем статистических данных оказывается недостаточным для проверки адекватности модели даже по приближенной статистике (8.1.20). Поэтому в этих случаях в качестве индикатора адекватности модели применяют величину [2, 35 и др.]

$$K_d = 1 - \frac{\tilde{Q}_A}{\tilde{Q}}, \tag{8.1.27}$$

называемую *выборочным коэффициентом детерминации* Y по X.

Здесь

$$\tilde{Q}_A = \sum_{i=1}^{n}(y_i - y_i^*)^2; \quad \tilde{Q} = \sum_{i=1}^{n}(y_i - \bar{y})^2; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{8.1.28}$$

Коэффициент детерминации изменяется от нуля до единицы. Чем ближе значение этого коэффициента к единицу, тем больше оснований полагать, что принятая модель близка к реальной регрессии. Коэффициент детерминации определяет долю общего рассеяния измеряемой величины Y, объясняемую выборочной регрессией.

8.2. Линейная средняя квадратическая выборочная регрессия

8.2.1. Парная линейная выборочная регрессия

В прикладных задачах регрессионного анализа наиболее широко применяется *парная линейная выборочная регрессия*, например Y по X.

В дальнейшем будем полагать, что проведено n пар независимых измерений, при которых определяются значение x_i и соответствующее ему значение y_i. По имеющимся парам значений (x_i, y_i) для $i = 1, 2, \ldots, n$ имеем гипотетическую (теоретическую) парную линейную функцию регрессии Y по X

$$\tilde{y} = b_0 + b_1 x, \tag{8.2.1}$$

где, как и выше, b_0 и b_1 — неизвестные коэффициенты.

В соответствии с предположениями, сделанными выше, считается, что:

- для заданного x_i и точных значений коэффициентов b_1 и b_2 случайные величины y_i имеют нормальное распределение;
- условное среднее значение случайных величин y_i $\mathrm{M}\{y_i \mid x_i\} = m_Y(x) = b_1 + b_2 x_i$;
- условная дисперсия случайных величин y_i не зависит от конкретного значения независимой переменной x_i, т.е. $\mathrm{M}\{(y_i - b_1 - b_2 x_i)^2\} = D_{Y|x} = D = \text{const}$.

Для дальнейшего анализа линейную зависимость прогнозируемой величины y_i^* от независимой переменной x_i удобнее представить в виде

$$\tilde{y} = b_0 + b_1(x - \bar{x}), \tag{8.2.2}$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

— выборочное среднее случайной величины X. В соответствии с методом наименьших квадратов в качестве оценок неизвестных параметров b_0 и b_1 необходимо взять такие их значения (оценки) b_0^* и b_1^*, которые

минимизируют функционал $G = \sum_{i=1}^{n}[y_i - b_0 - b_1(x_i - \bar{x})]^2$ и находятся из решения системы двух линейных уравнений

$$\frac{\partial G}{\partial b_0} = -2\sum_{i=1}^{n}[y_i - b_0^* - b_1^*(x_i - \bar{x})] = 0;$$

$$\frac{\partial G}{\partial b_1} = -2\sum_{i=1}^{n}[y_i - b_0^* - b_1^*(x_i - \bar{x})](x_i - \bar{x}) = 0.$$

Решая первое уравнение, имеем

$$nb_0^* = \sum_{i=1}^{n} y_i - b_1^*\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} y_i.$$

Отсюда получаем выражение для оценки параметра b_0:

$$b_0^* = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}. \tag{8.2.3}$$

Следовательно, оценка параметра b_0 равна выборочному среднему случайной величины Y. Из второго уравнения находим оценку параметра b_1:

$$b_1^* = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_Y}{s_X} r_{XY}. \tag{8.2.4}$$

Здесь s_X^2, s_Y^2 и r_{XY} — выборочные дисперсии и коэффициент взаимной корреляции случайных величин X и Y, определяемые согласно (3.4.27) и (7.1.10). При записи (8.2.4) учтено соотношение

$$\sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}).$$

Оценки коэффициентов b_0^* и b_1^* можно получить также из оценок, приведенных в п. 3.5, если в (3.5.43) положить $x_{1i} = x_i - \bar{x}$. Найденные оценки параметров b_0^* и b_1^* дают возможность сформировать уравнение для прогнозирования случайной величины Y при заданном значении независимой переменной $X = x$

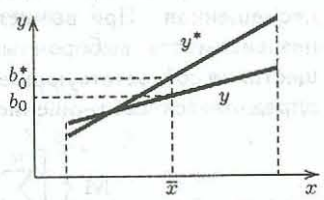


Рис. 8.2.1. Линейная регрессия

$$y^* = b_0^* + b_1^*(x - \bar{x}) = \bar{y} + r_{XY}\frac{s_Y}{s_X}(x - \bar{x}). \tag{8.2.5}$$

Прямая линия, задаваемая уравнением (8.2.5), называется *выборочной (парной) линейной регрессией* Y на X. По своему виду она идентична линейной регрессии (8.2.2), в которой вместо выборочных значений моментов случайных величин стоят их истинные значения.

На рис. 8.2.1 для иллюстрации приведены прямая регрессии $\tilde{y} = m_Y(x) = b_0 + b_1(x - \bar{x})$ при точных значениях параметров b_0, b_1 и ее оценка $y^* = b_0^* + b_1^*(x - \bar{x})$.

Вычислим точности условных оценок параметров b_0 и b_1, опуская в дальнейшем термин «условная». Учитывая (8.1.7), среднее значение оценки параметра b_0^*

$$\mathrm{M}\{b_0^*\} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{M}\{y_i\} = \frac{1}{n}\sum_{i=1}^{n}[b_0 + b_1(x_i - \bar{x})] = b_0, \qquad (8.2.6)$$

т.е. оценка параметра b_0 несмещенная. Учитывая независимость выборочных значений, а также (8.1.8), дисперсия оценки параметра b_0 определяется соотношением

$$D(b_0^*) = \mathrm{M}\{(b_0^* - b_0)^2\} = \frac{1}{n^2}\mathrm{M}\left\{\left(\sum_{i=1}^{n}[y_i - b_0 - b_1(x_i - \bar{x})]\right)^2\right\} =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{M}\{[y_i - b_0 - b_1(x - \bar{x})]^2\} = \frac{D}{n}. \qquad (8.2.7)$$

Среднее значение числителя (8.2.4) при фиксированных x_i преобразуется к виду

$$\mathrm{M}\left\{\sum_{i=1}^{n}(x_i - \bar{x})y_i\right\} = \sum_{i=1}^{n}(x_i - \bar{x})\mathrm{M}\{y_i\} =$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})[b_0 + b_1(x_i - \bar{x})] = b_1\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Тогда из формулы для оценки параметра b_1^* получаем $\mathrm{M}\{b_1^*\} = b_1$, т.е. оценка второго параметра рассматриваемой линейной регрессии также несмещенная. При вычислении дисперсии оценки параметра b_1^* учтем независимость выборочных значений, а также условие (8.1.8). Осуществляя соответствующие преобразования, дисперсия числителя (8.2.4) определяется соотношением

$$\mathrm{M}\left\{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \mathrm{M}\{y_i\})\right]^2\right\} =$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{M}\{(y_i - [b_0 + b_1(x_i - \bar{x})])^2\} = D\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

С учетом этого выражения формула для дисперсии оценки второго параметра линейной регрессии примет вид

$$D(b_1^*) = D \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (8.2.8)$$

Оценки параметров b_0^* и b_1^* линейно зависят от случайной величины y_i. Поэтому так же, как и y_i, они распределены по нормальному закону. Покажем, что условные оценки этих параметров взаимно некоррелированы. Для этого достаточно показать некоррелированность оценки параметра b_0 и числителя (8.2.4) оценки параметра b_1. Принимая во внимание независимость пар выборочных значений, имеем

$$\mathrm{M}\left\{\sum_{i=1}^{n}(y_i - b_0)\sum_{i=1}^{n}(x_i - \bar{x})[y_i - \mathrm{M}\{y_i\}]\right\} =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})\mathrm{M}\{(y_j - b_0)[(y_i - b_0) - b_1(x_i - \bar{x})]\} =$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})\mathrm{M}\{(y_i - b_0)^2\} - b_1\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{j=1}^{n}\mathrm{M}\{y_j - b_0\} = 0.$$

Таким образом, оценки параметров b_0^* и b_1^* некоррелированы, а следовательно, для нормального их распределения и независимы. Необходимо заметить, если в качестве модели линейной регрессии использовать уравнение (8.2.1), полученные оценки параметров модели b_0^{**} и b_1^{**} будут коррелированными, так как $b_0^{**} = \bar{y} - b_1^{**}\bar{x}$.

На основе полученных соотношений можно утверждать, что прогнозируемое значение y^* случайной величины Y, определяемое формулой (8.2.5), является несмещенной оценкой:

$$\mathrm{M}\{y^*\} = b_0 + b_1(x - \bar{x}) = m_Y(x). \qquad (8.2.9)$$

Иначе говоря, среднее значение оцениваемой линейной регрессии совпадает с условным средним значением случайной величины Y при фиксированном значении второй величины $X = x$.

Дисперсия оценки прогнозируемого значения y^*

$$D(y^*) = D(b_0^*) + (x - \bar{x})^2 D(b_1^*) = D\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \qquad (8.2.10)$$

Из последнего выражения видно, что дисперсия оценки прогнозируемого значения y^* минимальна при значении второй величины, совпадающей с выборочным средним. Потому неопределенность оценки прогноза возрастает с увеличением модуля разности между независимой переменной

x и выборочным средним $\bar{x}$.

В полученных выше формулах для дисперсий оценок b_0^*, b_1^* и y^* фигурирует истинное значение условной дисперсии D отклонения наблюдаемого значения y_i от условного среднего $m_Y(x_i) = b_0 + b_1(x_i - m_x)$. Поскольку она, как правило, неизвестна, вместо D берется выборочная условная дисперсия этого отклонения, определяемая соотношением

$$s^2_{Y|x} \equiv s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - y_i^*)^2 = \frac{1}{n-2}\sum_{i=1}^{n}[(y_i - \bar{y}) - (y_i^* - \bar{y})]^2 =$$

$$= \frac{1}{n-2}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - 2\sum_{i=1}^{n}(y_i - \bar{y})(y_i^* - \bar{y}) + \sum_{i=1}^{n}(y_i^* - \bar{y})^2\right]. \quad (8.2.11)$$

Здесь перед суммой стоит множитель $1/(n-2)$. Это обусловлено тем, что из общего числа исходных степеней свободы n (n независимых пар элементов выборки) две степени свободы использованы для определения оценок параметров b_0^* и b_1^*. Первая сумма в квадратных скобках равна $(n-1)s_Y^2$. Так как во второй сумме $y_i^* - \bar{y} = y_i^* - b_0^* = b_1^*(x_i - \bar{x})$, то

$$\sum_{i=1}^{n}(y_i - \bar{y})(y_i^* - \bar{y}) = b_1^*(n-1)\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) =$$

$$= b_1^*(n-1)s_X s_Y r_{XY} = (n-1)s_Y^2 r_{XY}^2.$$

В последнем выражении учтено соотношение (8.2.4). Аналогично вычисляется и третья сумма в (8.2.11):

$$\sum_{i=1}^{n}(y_i^* - \bar{y})^2 = (b_1^*)^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{s_Y^2}{s_X^2}r_{XY}^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)s_Y^2 r_{XY}^2.$$

В результате для условной выборочной дисперсии случайной величины y_i можно записать

$$s^2 = \frac{n-1}{n-2}s_Y^2(1 - r_{XY}^2). \quad (8.2.12)$$

Поэтому формулы для выборочных дисперсий оценок параметров b_0^*, b_1^* и прогнозируемого значения y^* примут вид

$$s^2(b_0^*) = \frac{1}{n}s^2; \quad (8.2.13)$$

$$s^2(b_1^*) = s^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2; \quad (8.2.14)$$

$$s^2(y^*) = s^2\left[\frac{1}{n} + (x - \bar{x})^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2\right]. \quad (8.2.15)$$

Для принятого предположения о нормальном распределении случайных величин y_i оценки b_0^*, b_1^* и y^*, линейно зависящие от y, также распределены по нормальному закону. В связи с этим статистики

$$t_{b_0} = \frac{b_0^* - b_0}{s(b_0^*)}; \quad t_{b_1} = \frac{b_1^* - b_1}{s(b_1^*)}; \quad t_Y = \frac{y^* - m_Y(x)}{s(y^*)} \quad (8.2.16)$$

имеют распределение Стьюдента с $\nu = n-2$ степенями свободы. На основании этих статистик по заданной доверительной вероятности γ можно найти доверительные границы и доверительные интервалы для точных значений параметров линейной регрессии b_0, b_1 и точного значения линейной регрессии $m_Y(x) = b_0 + b_1(x - \bar{x})$.

Применительно к выборочной регрессии y^* и заданной доверительной вероятности γ в соответствии с (8.2.16) центральные доверительные границы находятся из соотношения

$$\mathbf{P}\{y^* - s(y^*)t_\gamma \leqslant m_Y(x) \leqslant y^* + s(y^*)t_\gamma\} = \gamma, \quad (8.2.17)$$

где величина t_γ определяется по таблицам $\mathbf{P}\{t > t_\gamma\} = (1-\gamma)/2$ распределения Стьюдента (табл. 3 Приложения) для $\nu = n-2$ степеней свободы. Если рассматривать доверительный интервал

$$y^* \pm st_\gamma\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad (8.2.18)$$

как функцию от x, то последнее выражение определяет две ветви гиперболы, диаметром которой служит выборочная регрессия $y^* = b_0 + b_1(x - \bar{x})$. Доверительный интервал имеет наименьшую длину, когда значение независимой переменной x равно его выборочному среднему, и увеличивается по мере возрастания абсолютной разности $|x - \bar{x}|$.

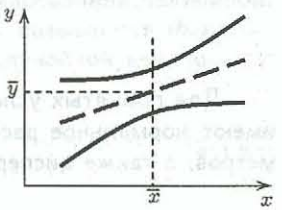


Рис. 8.2.2. Расположение доверительных границ регрессии

Рис. 8.2.2 иллюстрирует расположение доверительных границ.

Обобщим полученные выше результаты применительно к ситуации, при которой каждому значению x_i соответствует k значений y_{ik}, $k = 1, 2, \ldots, n_i$. При этом линейная функция регрессии, как и выше, описывается выражением (8.2.2), в котором $\bar{x}$ является взвешенным выборочным средним, т.е.

$$\bar{x} = \frac{\sum_{i=1}^{n} n_i x_i}{\sum_{i=1}^{n} n_i}. \quad (8.2.19)$$

Оценки параметров модели определяются из минимизации функци-

онала

$$G = \sum_{i=1}^{n}\sum_{k=1}^{n_i}[y_{ik} - b_0 - b_1(x_i - \bar{x})]^2. \qquad (8.2.20)$$

Применяя метод наименьших квадратов, выражения для оценок параметров модели примут вид, аналогичный (8.2.3) и (8.2.4):

$$b_0^* = \frac{\sum\limits_{i=1}^{n} n_i y_{i.}}{\sum\limits_{i=1}^{n} n_i} = \bar{y}; \qquad (8.2.21)$$

$$b_1^* = \frac{\sum\limits_{i=1}^{n} n_i y_{i.}(x_i - \bar{x})}{\sum\limits_{i=1}^{n} n_i (x_i - \bar{x})^2}. \qquad (8.2.22)$$

Здесь $y_{i.} \equiv \bar{y}_i$ — выборочное среднее случайной величины Y, определяемое согласно (8.1.6) для конкретного значения x_i; $\bar{y} \equiv y_{..}$ — общее выборочное среднее случайной величины Y:

$$y_{..} \equiv \bar{y} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n_i} y_{ik}}{\sum\limits_{i=1}^{n} n_i}. \qquad (8.2.23)$$

Для принятых условий оценки параметров модели несмещенные и имеют нормальное распределение. Выборочные дисперсии этих параметров, а также дисперсия выборочной линейной регрессии

$$y^* = b_0^* + b_1^*(x - \bar{x}) \qquad (8.2.24)$$

определяются выражениями, аналогичными (8.2.13)–(8.2.15):

$$s^2(b_0^*) = \frac{s^2}{\sum\limits_{i=1}^{n} n_i}; \qquad (8.2.25)$$

$$s^2(b_1^*) = s^2 \Big/ \sum_{i=1}^{n} n_i (x_i - \bar{x})^2; \qquad (8.2.26)$$

$$s^2(y^*) = s^2 \left[\frac{1}{\sum\limits_{i=1}^{n} n_i} + (x - \bar{x})^2 \Big/ \sum_{i=1}^{n} n_i (x_i - \bar{x})^2\right]. \qquad (8.2.27)$$

Здесь

$$s^2 = \frac{1}{\sum\limits_{i=1}^{n} n_i - 2}\sum_{i=1}^{n}\sum_{k=1}^{n_i}(y_{ik} - y_i^*)^2. \qquad (8.2.28)$$

— выборочная условная дисперсия, являющаяся оценкой неизвестной дисперсии D.

Рассматривая статистики вида (8.2.16), которые распределены по закону Стьюдента с $\nu = \sum\limits_{i=1}^{n} n_i - 2$ степенями свободы, по заданной доверительной вероятности γ находятся доверительные пределы (границы) для параметров выборочной регрессии. Применительно к выборочной регрессии доверительные интервалы определяются выражением

$$y^* \pm s t_\gamma \sqrt{\frac{1}{\sum\limits_{i=1}^{n} n_i} + \frac{(x - \bar{x})^2}{\sum\limits_{i=1}^{n} n_i (x_i - \bar{x})^2}}. \qquad (8.2.29)$$

8.2.2. Множественная линейная выборочная регрессия

При *множественной линейной выборочной регрессии* прогнозируемое значение случайной величины Y линейно зависит от векторной величины $\mathbf{X}(X_1, X_2, \ldots, X_p)$, которая, как и в случае парной регрессии, может быть либо случайной либо детерминированной величиной. Как отмечалось выше, составляющие вектора $\mathbf{X}$ называются *факторами*. В результате прогнозируемое значение y^* случайной величины Y можно представить в виде

$$y^* = b_0^* + \sum_{j=1}^{p} b_j^*(x_j - \bar{x}_j), \qquad (8.2.30)$$

где $\bar{x}_j = \sum\limits_{i=1}^{n} x_{ji}$ — среднее значение фактора x_j, n — объем выборки, а b_j^* — оценки неизвестных параметров, в которых $j = 1, 2, \ldots, p$.

Для получения оценок параметров регрессии воспользуемся методом наименьших квадратов, для чего необходимо минимизировать функционал

$$G = \sum_{i=1}^{n}\left[y_i - b_0 - \sum_{j=1}^{p} b_j(x_{ji} - \bar{x}_j)\right]^2 \qquad (8.2.31)$$

по значениям неизвестных параметров b_j. Для рассматриваемой линейной регрессии и числа опытов $n > p + 1$ метод наименьших квадратов приводит к системе из $p + 1$ линейных нормальных уравнений относительно оценок неизвестных параметров b_j^*, $j = 0, 1, \ldots, p$.

Применительно к множественной выборочной линейной регрессии в дальнейшем предполагается выполнение следующих условий.

1. При каждом наборе значений $\mathbf{x}_i(x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1, 2, \ldots, n$, величины y_i распределены по нормальному закону.

2. Условное среднее значение случайных величин y_i

$$\mathrm{M}\{y_i\} \equiv \mathrm{M}\{y_i \mid x_{1i}, x_{2i}, \ldots, x_{pi}\} = m_Y(x_{1i}, x_{2i}, \ldots, x_{pi}) = \\ = b_0 + \sum_{j=1}^{p} b_j(x_j - \bar{x}_j). \qquad (8.2.32)$$

где $\mathrm{b}(b_0, b_1, \ldots, b_p)$ — истинные (точные) значения неизвестных параметров.

3. Условная дисперсия случайных величин y_i постоянна

$$\mathrm{M}\left\{\left[y_i - b_0 - \sum_{j=1}^{p} b_j(x_{ji} - \bar{x}_j)\right]^2\right\} = D_{Y|x} \equiv D = \mathrm{const}\,. \qquad (8.2.33)$$

При этом, как и выше, полагается, что совокупности выборочных значений $(y_i; \mathbf{x}_i(x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1, 2, \ldots, n)$, независимы между собой. Для определения оценок коэффициентов b_0^*, b_j^*, $j = 1, 2, \ldots, p$, применительно к регрессии (8.2.30) можно воспользоваться результатами п. 3.5, а именно соотношениями (3.5.50). В (3.5.50) детерминант A и дополнения A_j матрицы коэффициентов определяются выражением (3.5.49), в котором элементы матрицы $[x_j x_l]$ заменяются элементами $[z_j z_l] \equiv [(x_j - \bar{x}_j)(x_l - \bar{x}_l)]$, $j, l, = 1, 2, \ldots, p$, а фиктивная переменная $x_0 \equiv 1$ заменяется фиктивной переменной $z_0 \equiv 1$. При этом матрица коэффициентов (3.5.49) при неизвестных параметрах преобразуется в матрицу

$$\mathbf{A} = \begin{pmatrix} [z_0 z_0] & \ldots & [z_0 z_p] \\ \vdots & \ddots & \vdots \\ [z_p z_0] & \ldots & [z_p z_p] \end{pmatrix}. \qquad (8.2.34)$$

Она симметрична относительно главной диагонали, элементы которой являются положительными числами. Оценки неизвестных параметров находятся из выражений

$$b_j^* = A_j/A, \quad j = 0, 1, \ldots, p. \qquad (8.2.35)$$

где $A = |\,\mathbf{A}\,|$ — определитель (детерминант) матрицы (8.2.34), а A_j — определитель матрицы, получающейся из (8.2.34) путем замены j-го столбца матрицы на столбец, составленный из элементов $[yz_0], [yz_1], \ldots, [yz_p]$. Как и выше, будем пренебрегать ошибками задания величин x_j и полагать, что оценка случайной величины Y несмещенная с постоянной условной дисперсией D. При этом параметры b_j^*

также будут несмещенными оценками истинных значений, а дисперсии оценок определяются из выражений

$$D(b_j^*) = D(A_{jj}/A) = Dc_{jj}, \quad j = 0, 1, \ldots, p. \qquad (8.2.36)$$

Здесь $c_{jj} = A_{jj}/A$, где A_{jj} — алгебраические дополнения элементов определителя A, получаемые путем удаления из матрицы определителя столбца и строки, на пересечении которых находится элемент $[z_j z_j]$. Для неизвестной условной дисперсии D случайной величины Y берется ее выборочное значение

$$s^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \left[y_i - (b_0^* + b_1^* x_{1i} + \cdot + b_p^*)\right]^2. \qquad (8.2.37)$$

Поэтому вместо дисперсии (8.2.36) следует взять ее выборочное значение

$$s^2(b_j^*) = s^2(A_{jj}/A) = s^2 c_{jj}, \quad j = 0, 1, \ldots, p. \qquad (8.2.38)$$

Аналогично парной выборочной регрессии оценки параметров распределены по нормальному закону, они несмещенные. Оценка параметра b_0^* независима от оценок других параметров b_j^*, $j = 1, 2, \ldots, p$, причем оценка параметра b_0^* и ее дисперсия определяются формулами (8.2.3) и (8.2.13).

Выборочная регрессия (прогнозируемое значение) y^* также распределена по нормальному закону, она независима от оценок параметров регрессии, а ее дисперсия

$$s^2(y^*) = s^2\left[\frac{1}{n} + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{jl}(x_j - \bar{x}_j)(x_l - \bar{x}_l)\right]. \qquad (8.2.39)$$

Статистики

$$t_{b_j} = \frac{b_j^* - b_j}{s(b_j^*)}; \quad j = 0, 1, \ldots, p; \quad t_Y = \frac{y^* - m_Y(x_1, x_2, \ldots, x_p)}{s(y^*)} \qquad (8.2.40)$$

имеют распределение Стьюдента с $\nu = n - (p+1)$ степенями свободы. На основе этих статистик по заданной доверительной вероятности γ можно определить доверительные границы и доверительный интервалы точных значений параметров b_j и точного значения самой линейной регрессии

$$m_Y(x_1, x_2, \ldots, x_p) = b_0 + \sum_{j=1}^{p} b_j(x_j - \bar{x}_j), \qquad (8.2.41)$$

Применительно к регрессии $m_Y(x_1, x_2, \ldots, x_p)$ и заданной доверительной вероятности γ центральные доверительные границы находятся из выражения

$$\mathbf{P}\{y^* - s(y^*)t_\gamma \leqslant m_Y(x_1, x_2, \ldots, x_p) \leqslant y^* + s(y^*)t_\gamma\} = \gamma, \qquad (8.2.42)$$

где t_γ определяется по таблицам $\mathrm{P}\{t > t_\gamma\} = (1-\gamma)/2$ критических значений распределения Стьюдента (табл. 3 Приложения 2) для $\nu = n - (p+1)$ степеней свободы.

Пусть проведено n серий измерений, при которых для каждого значения x_i определены соответствующие им k значений y_{ik}, $k = 1, 2, \ldots, n_i$. Случайную величину y_{ik}, описывающую результат k-го опыта, обычно представляют в виде

$$y_{ik} = b_o + \sum_{j=1}^{p} b_j(x_j - \bar{x}_j) + \varepsilon_{ik}, \qquad (8.2.43)$$

где ε_{ik} – случайная ошибка с нулевым средним значением $\mathrm{M}\{\varepsilon_{ik}\} = 0$. Как и выше, полагается, что случайные величины ε_{ik} независимы между собой и независимы от неизвестных параметров регрессии b_j, $j = 0, 1, \ldots, p$. По имеющимся $n \geqslant p+1$ сериям значений (x_i, y_{ik}) для $i = 1, 2, \ldots, n$ оценки неизвестных параметров регрессии согласно методу наименьших квадратов находят из условия обращения в минимум квадратичной формы

$$G = \sum_{i=1}^{n}\sum_{k=1}^{n_i}\left[y_{ik} - b_0 - \sum_{j=1}^{p} b_j(x_j - \bar{x}_j)\right]^2 \qquad (8.2.44)$$

Минимум функционала по значениям неизвестных параметров эквивалентен его минимуму по каждому из параметров b_j, $j = 0, 1, \ldots, p$. Оценки параметров b_j^* находятся из системы уравнений (8.1.4), а дисперсии их оценок вычисляется аналогично (8.2.38), в котором выборочная условная дисперсии определяется выражением (8.2.28).

Приведенные соотношения для прогнозируемых значений отклика получены в предположении независимых значений отстатков ε_{ik}. Вместе с этим в указанной выше литературе рассматривается и более общий случай, когда остатки коррелированы. В этом случае для нахождения оптимальных значений параметров регрессии применяется обобщенный метод наименьших квадратов, учитывающий соответствующие веса при невязках. При этом задача существенно усложняется, когда корреляционная матрица остатков априори неизвестна.

8.3. Преобразование нелинейной регрессии в линейную регрессию

Часто в прикладных задачах регрессия представляет собой нелинейную зависимость от параметров, и ее анализ становится обычно весьма сложным. Однако, как отмечалось выше, в ряде ситуаций оказывается возможным путем несложных преобразований соответствующих величин нелинейной регрессии относительно исходных факторов представить кривую регрессии как линейное соотношение между преобразованными величинами.

Таблица 8.3.1

$h(\tilde{y})$	$g(x)$		
	x	$1/x$	$\ln x$
$\tilde{y}$	$\tilde{y} = b_0 + b_1 x$	$\tilde{y} = b_0 + b_1/x$	$\tilde{y} = b_0 + b_1 \ln x$
$1/\tilde{y}$	$\tilde{y} = \dfrac{1}{b_1(x + b_0/b_1)}$	$\tilde{y} = \dfrac{1}{b_0} - \dfrac{b_1/b_0^2}{x + b_1/b_0}$	$\tilde{y} = \dfrac{1}{b_1(\ln x + b_0/b_1)}$
$\ln \tilde{y}$	$\tilde{y} = a\exp(b_1 x)$	$y = a\exp(b_1/x)$	$\tilde{y} = ax^{b_1}$
	$b_0 = \ln a$		

Далее рассмотрим вопросы линеаризации нелинейных регрессий применительно к парным зависимостям, т.е. когда исследуется зависимость y от одного фактора x.

Например, если кривая регрессии описывается уравнением $\tilde{y} = \gamma\exp(\beta x)$, то логарифмическое преобразование обеих величин приводит к уравнению $\ln\tilde{y} = \ln\gamma + \beta x$, т.е. точки $(\ln\tilde{y}, x)$ изображают уравнение прямой с параметрами $b_o \equiv \ln\gamma$ и $b_1 \equiv \beta$.

В общем случае этот принцип можно сформулировать следующим образом [64].

Если точки $(g(x_i), h(y_{ij}))$, соответствующие значениям $(x_i y_{ij})$, группируются вокруг прямой линии таким образом, что $h(y)$ распределена нормально со средним значением

$$\mathrm{M}\{h(y) \mid g(x)\} = b_0 + b_1[g(x) - \mathrm{M}\{g(x)\}] \qquad (8.3.1)$$

и условной дисперсией

$$D_{h(y)|g(x)} \equiv \sigma^2 = \mathrm{const}, \qquad (8.3.2)$$

то к преобразованным наблюдениям можно применить теорию линейной регрессии.

Функции преобразования выбираются на основе анализа графического представления экспериментальных данных и априорных теоретических данных относительно возможного поведения исследуемой регрессии. Вместе с этим следует отметить, что для одних и тех же экспериментальных данных можно предложить различные функциональные зависимости. Однако они будут приемлемы только для интервала изменения значений фактора, для которых проводился эксперимент, так как вне этих значений фактора истинная регрессия может значительно отличаться от выборочной регрессии. Поэтому при экстраполяции, т.е. при вычислении регрессии вне области экспериментальных значений фактора, выбираемый тип выборочной регрессии должен опираться на профессиональные знания относительно исследуемых явлений.

Задавая различные виды функций $g(x)$ и $h(\tilde{y})$, можно получить ряд интересных зависимостей $\tilde{y} = \psi(x, b_0, b_1)$.

В табл. 8.3.1 [64] приведены зависимости $\tilde{y} = \psi(x, b_0, b_1)$ для функций $g(x)$ и $h(\tilde{y})$, принимающих значения x, $1/x$, $\ln x$ и $\tilde{y}$, $1/\tilde{y}$, $\ln\tilde{y}$ соответственно. Из таблицы видно, что различного рода гиперболические,

степенные, показательные и логарифмические функции могут быть преобразованы в линейные зависимости.

Неопределенность в выборе вида (класса) нелинейной регрессии, вообще говоря, является достаточно сложной задачей. Иногда приходится рассматривать ряд различных функциональных зависимостей, каждую из них линеаризировать и из анализа полученных регрессий выбирать ту, которая в большей мере отражает реальную ситуацию. Вместе с этим в регрессивном анализе разработан ряд формализованных процедур подбора линеаризующего преобразования (метод Г. Бокса и Д. Кокса и другие методы) [23].

8.4. Проверка гипотез, связанных с выборочной линейной регрессией

Если есть сомнение в том, что истинный вид линейной регрессии может отличаться от выборочной линейной регрессии, то после вычисления оценок неизвестных параметров b_j^*, $j = 0, 1, 2, \ldots, p$, предполагаемой линейной регрессии, необходимо (с заданными уровнями значимости) проверить гипотезы:

- какие из оценок параметров модели незначительно отличаются от нуля и ими можно пренебречь по сравнению с оценками других параметров, которые существенно отличны от нуля;
- принятая модель регрессии адекватна экспериментальным данным.

Рассмотрим проверку этих гипотез применительно к множественной регрессии (8.2.30).

Оценка параметра b_j^*, $j = 0, 1, 2, \ldots, p$, должна незначительно отличаться от нуля, если его истинное значение равно нулю, т.е. если $b_j = 0$. Как отмечалось выше, оценки параметров модели (8.2.35) имеют нормальное распределение. Поэтому для *проверки гипотезы о значимости оценок параметров* b_j^*, $j = 0, 1, 2, \ldots, p$, формируются статистика

$$t_j = \frac{b_j^*}{s(b_j^*)}, \quad j = 0, 1, 2, \ldots, p, \tag{8.4.1}$$

имеющая распределение Стьюдента с числом степеней свободы $\nu = n - (p + 1)$.

В соответствии с заданным (или принятым) уровнем значимости α по таблицам распределения Стьюдента находятся двухсторонние критические значения $t_{\alpha/2;\nu}$. Оценкой параметров b_j^* можно пренебречь, если выборочная статистика t_j^* удовлетворяет неравенству

$$-t_{\alpha/2;\nu} \leqslant t_j^* \leqslant t_{\alpha/2;\nu} \quad \text{или} \quad |t_j^*| \leqslant t_{\alpha/2;\nu}. \tag{8.4.2}$$

Для проверки гипотезы адекватности регрессионной модели экспериментальным данным, как и в п. 8.1, будем полагать, что при каждом наборе $\mathbf{x}(x_{1i}, x_{2i}, \ldots, x_{pi})$, $i = 1, 2, \ldots, n$, измеряется n_i ($n_i \geqslant 2$) значений величины y_{ik}, $k = 1, 2, \ldots, n$, а в качестве статистики для проверки

гипотезы адекватности модели необходимо использовать соотношение (8.1.19).

В п. 8.3 отмечалось, что часто нелинейную регрессию $y = \varphi(x)$ можно преобразовать в линейную зависимость между двумя функциями $g(x)$ и $h(\tilde{y})$ в виде

$$h(\tilde{y}) = b_0 + b_1 g(x).$$

Поэтому для подобной ситуации, обозначая $z = h(\tilde{y})$ и $w = g(x)$, выборочное значение исходной регрессии y_{ik} преобразуется в линейную зависимость z_{ik}:

$$z_{ik} = b_0 + b_1 w_{ik} + \varepsilon_{ik}. \tag{8.4.3}$$

Дальнейший анализ по определению параметров этой регрессии, проверки соответствующих гипотез и вычисление доверительных пределов для заданной доверительной вероятности проводятся рассмотренными выше методами. Путем обратного преобразования $y = \psi(z)$ и $x = \vartheta(w)$ определяется выборочная регрессия $y^* = \varphi^*(x)$ и соответствующие ей доверительные пределы.

Пример 8.4.1. Рассмотрим пример, приведенный в [64], исследования зависимости прочности на разрыв g пяти ($h = 5$) пробных кусков цемента от времени выдержки t. Предполагается, что гипотетическая зависимость задается формулой

$$G = \beta_0 \exp(-\beta_1 / t),$$

где β_0 и β_1 — неизвестные параметры, подлежащие оценке по пяти сериям опытных данных, приведенных в табл. 8.4.1.

Логарифм прочности на разрыв

$$\lg G = \lg \beta_0 - (\beta_1 / t) \lg \mathrm{e}$$

является линейной функцией значения x, обратного времени выдержки. Обозначая логарифм прочности на разрыв через $\tilde{y}$, а величину, обратную времени выдержки, через x ($x = 1/t$), задача сводится к определению линейной регрессии $\tilde{y} = b_0 + b_1(x - \bar{x})$. Здесь $b_0 = \lg \beta_0$, $b_1 = -\beta_1 \lg \mathrm{e}$.

В табл. 8.4.1 приведены значения логарифмов прочности, их средние значения y_i, число степеней свободы $\nu_i = n_i - 1$ и дисперсия s_{Ri}^2 для каждой серии наблюдений. Общее число степеней свободы и объединенная дисперсия для всей совокупности опытных данных равны

$$\nu = \sum_{i=1}^{n=5} = 21 - 5 = 16; \quad s_R^2 = \frac{1}{\nu} \sum_{i=1}^{n=5} \sum_{k=1}^{n_i} (y_{ik} - y_{i\cdot})^2 = 0{,}001045.$$

Проверим с уровнем значимости $\alpha = 0{,}05$ гипотезу о постоянстве дисперсий для разных серий опытов (точек плана эксперимента) по критерию Бартлета согласно (5.3.17) и (5.3.18). Выборочная статистика Бартлета $B_{h-1=4}^* \approx 0{,}8$, в то время как ее критическое значение $B_{0{,}05;4} = 9{,}49$. Так как найденные пять оценок дисперсий незначимо отличаются одна от другой, можно принять гипотезу о том, что дисперсия независима от x. На основе приведенных выше формул получены оценки параметров и общий вид выборочной регрессии для x, соответствующих точкам плана эксперимента:

$\bar{x} = 0{,}336$; $b_0^* = \bar{y} = 1{,}434$; $b_1^* = 0{,}498$; $y^* = 1{,}434 - 0{,}498(x - 0{,}336) = 1{,}601 - 0{,}498x$.

Таблица 8.4.1

Зависимость прочности на разрыв от времени выдержки

Время выдержки, дни					
t_j	1	2	3	7	28
g_{ik}	13,0	21,9	29,8	32,4	41,8
	13,3	24,5	28,0	30,4	42,6
	11,8	24,7	24,1	34,5	40,3
	–	–	24,2	33,1	35,7
	–	–	26,2	35,7	37,3
Логарифм прочности на разрыв					
$x_i = 1/t$	1,000	0,500	0,333	1,143	1,036
y_{ik}	1,114	1,340	1,474	1,511	1,621
	1,124	1,389	1,447	1,483	1,629
	1,072	1,393	1,382	1,538	1,605
	–	–	1,384	1,520	1,553
	–	–	1,418	1,553	1,572
n_i	3	3	5	5	5
$y_{i\cdot}$	1,103	1,374	1,421	1,521	1,596
$\nu_i = n_i - 1$	2	2	4	4	4
s^2_{Ri}	0,000762	0,000871	0,001596	0,000715	0,001055
y^*_i	1,103	1,352	1,435	1,530	1,583
s_{yi}	0,0168	0,0081	0,0072	0,0084	0,0099
$y^*_i + 2{,}09s_{yi}$	1,138	1,369	1,450	1,548	1,604
$y^*_i - 2{,}09s_{yi}$	1,068	1,335	1,420	1,512	1,562
G^*_i	12,7	22,5	27,2	33,9	38,3
G^*_{1i}	13,7	23,4	28,2	35,3	40,2
G^*_{2i}	11,7	21,6	26,3	32,3	36,5

В табл. 8.4.1 даны значения выборочной регрессии y^*_i для x, соответствующих точкам плана эксперимента. На основе этих данных вычислены дисперсия адекватности модели $s^2_A = (1/(n-2))\sum_{i=1}^{n} n_i(y_{i\cdot} - y^*_i)^2 = 0{,}003682/3 = 0{,}001227$ и выборочная статистика дисперсионного отношения $F^*_{3;4} = 0{,}001227/0{,}001045 = 1{,}17$. Сравнивая это значение статистики с критическим значением (например, для уровня значимости $\alpha = 0{,}05$) $F_{0{,}05;3,4} = 6{,}59$, можно сделать очевидный вывод, что преобразованная линейная регрессия не противоречит опытным данным. Поэтому для дальнейших расчетов в качестве оценки неизвестной дисперсии можно воспользоваться оценкой (8.1.13). Ниже приведены значения условной выборочной дисперсии, среднеквадратических отклонений оценок параметров и дисперсии выборочной регрессии, вычисленные согласно (8.2.25)–(8.2.28):

$s^2 = 0{,}001073$; $s_{b_0} = 0{,}0071$; $s_{b_1} = 0{,}0229$; $s^2_y = 0{,}00005110 + 0{,}0005261(x - 0{,}336)^2$.

Для проверки гипотез значимости параметров регрессии согласно (8.4.1) определены статистики $t^*_{b_0} = 20{,}20$; $t^*_{b_1} = 21{,}75$. Сравнивая эти значения с их критическим значением (например, для уровня значимости $\alpha = 0{,}05$) $t_{\alpha/2;\nu} = t_{0{,}025;19} = 2{,}093$, можно сделать вывод о том, что значения параметров регрессии значимы и их следует оставить в уравнении выборочной регрессии.



Рис. 8.4.1. Зависимсть прочности бетона на разрыв от выдержки

В конце таблицы приведены выборочные значения среднего квадратического отклонения выборочной регрессии относительно ее истинного значения s_{yi} и центральные доверительные интервалы для доверительной вероятности $\gamma = 0{,}95$, вычисленные согласно (8.2.29): $y^*_i \pm 2{,}09s_{yi}$.

Учитывая, что $x = 1/t$ и $y = \lg G$ из уравнения выборочной линейной регрессии $y^* = 1{,}601 - 0{,}498x$ получается равенство для выборочной регрессии прочности G^*:

$$G^* = 39{,}9 \times 10^{-0{,}498/t} \approx 39{,}9\exp(-1{,}15/t)$$

В последних трех строках таблицы приведены значения прочности бетона, кг/см^2, вычисленные по выборочной регрессии G^*_i, а также соответственно верхняя G_{1i} и нижняя G_{2i} границы доверительного интервала для принятой доверительной вероятности $\gamma = 0{,}95$. Верхняя и нижняя границы доверительного интервала для кривой прочности бетона определялись путем вычисления антилогарифмов величин $y^*_i \pm 2{,}09s_{yi}$.

На рис. 8.4.1 приведена эмпирическая (выборочная) кривая регрессии (средняя кривая) для прочности бетона на разрыв (кг/см^2) в зависимости от времени выдержки (дни). Там же показаны доверительные границы для теоретической (истинной) кривой регрессии (верхняя и нижняя кривые) для доверительной вероятности $\gamma = 0{,}95$.

Полученная кривая регрессии описывает зависимость средней прочности (средняя кривая) на разрыв как функцию времени выдержки лишь в *в пределах наблюдений*, т.е. при времени выдержки от 1 до 28 дней. Если представляет интерес прогнозируемое значение прочности на разрыв, выходящее за пределы наблюдений, то выбор гипотетической функции становится существенно важным. Как отмечалось выше в п. 8.1 и п. 8.3, это обусловлено тем, что если выбранная гипотетическая функция регрессии незначимо отличается от истинной функции в пределах наблюдений, она может значительно отличаться в более широком интервале времени выдержки. В связи с этим при выборе гипотетической функциональной зависимости необходимо учитывать все профессиональные данные относительно связи между величинами. Применительно к данному примеру при $t \to \infty$, т.е. при $x = 1/t \to 0$ $\tilde{y} \to 1{,}601$, т.е. прочность на разрыв становится равной 39,9 кг/см^2, а доверительные границы этой асимптотической прочности равны 42,0 и 37,9 кг/см^2.

8.5. Временны́е ряды

Почти в каждой области человеческой деятельности встречаются явления, которые интересно и важно изучать в их развитии и изменении во времени. В качестве примеров подобных явлений можно рассматривать регистрацию значений различных физических, химических и других

величин через равные промежутки времени. В частности, при приеме коротких радиоволн уровень принимаемого сигнала при фиксированных местоположениях передатчика и приемника существенно зависит от времени суток, времени года, одиннадцатилетнего периода солнечной активности и других причин, представляя в общем виде сложную функциональную статистическую зависимость. При описании и анализе таких зависимостей естественно возникает целый ряд специфических статистических задач. Для описания и анализа подобных зависимостей, наблюдаемых во времени, разработано достаточно большое число методов. При этом некоторые из них для своего применения требуют большого объема выборок, т.е. адекватное описание ряда возможно лишь на основе совокупности наблюдений значений процесса в течение длительного периода времени. *Статистическое исследование зависимостей процессов и явлений, происходящих во времени, представляет собой специальный раздел прикладной статистики — анализ временных рядов.* Исследованиям временных рядов посвящено большое число фундаментальных работ. Основные теоретические и прикладные аспекты временных рядов изложены в [2, 32, 33, 62, 66 и др.].

В настоящем пункте дано краткое описание временных рядов и методов их анализа.

Во временных рядах наблюдения, как правило, зависимы, и характер этой зависимости может определяться положением наблюдений в последовательности.

Определения. *Временном рядом* называется последовательность наблюдений, упорядоченных во времени и представляющих собой совокупность значений

$$x(t_1), x(t_2), \ldots, x(t_n), \tag{8.5.1}$$

осуществленных в фиксированные дискретные моменты времени $t_1 < t_2 < \ldots < t_i < \ldots < t_n$. Принципиально переменная t может быть непрерывной величиной. Характерной чертой временного ряда является то, что порядок в последовательности $t_1, t_2, \ldots, t_n$ существенен, в то время как в случайной выборке $x_1, x_2, \ldots, x_n$ индексы служат лишь для удобства обозначения разных элементов выборки.

В дальнейшем будем полагать, что t_i — дискретная величина, а $x(t_i)$ — дискретный ряд, определенный в равноотстоящих друг от друга моментах времени. Наблюдения обычно производятся в дискретные моменты времени. Но даже при непрерывной регистрации наблюдаемых величин они, как правило, преобразуются, в дискретные последовательности, так как в современных условиях временные ряды изучаются с помощью ЭВМ. В качестве единицы времени выберем интервал между соседними моментами. В дальнейшем дискретные моменты времени для удобства и сокращения записи будут также нумероваться целыми числами и будут также обозначаться как $1, 2, \ldots, n$, а члены ряда — символами $x_1, x_2, \ldots, x_n$. Для значений ряда в моменты, предшествующие начальному моменту, введем обозначения x_{-1}, x_{-2} и т.д. Без потери общности временной интервал между наблюдениями (элементарный интервал) принимают за единицу, который часто называют шагом. Несколько (k) единичных интервалов (шагов) в прошлое называют лагом (запаздыванием) k. При представлении временных рядов естественно считать, что время течет в одном направлении; события, близкие по времени, должны быть сравнительно сильно связаны, а события, разделенные большим промежутком времени, не должны иметь сильной связи.

Во временных рядах наблюдения, как правило, зависимы между собой, и характер этой зависимости может определяться положением наблюдений в последовательности.

Принципиально вместо времени в качестве независимой переменной могут быть и другие переменные, например такие, как пространственные координаты.

Анализируемый отрезок временного ряда (8.5.1) можно рассматривать как одну из возможных реализаций случайного процесса. Цели изучения временных рядов могут быть различными. Основными практическими задачами, решаемыми с помощью описания и анализа временных рядов, являются следующие:

- выяснение механизма, порождающего ряд;
- прогнозирование (предсказание) будущих значений изучаемого процесса на основе ранее наблюденных значений;
- управление процессом, порождающим анализируемый временной ряд;
- сжатое описание характерных особенностей ряда.

На основе имеющихся последовательных наблюдений должна быть выбрана (предложена) математическая модель. Естественно математическая модель должна достаточно «хорошо» представлять механизм образования соответствующего ряда. Однако, будучи математической абстракцией, модель является лишь только приближением к реальному явлению (процессу). Степень достоверности (адекватности) модели реальной ситуации зависит от априорных знаний об исследуемом процессе и о той информации, которую можно получить с использованием статистического анализа.

Достаточно общей исходной статистической моделью временного ряда, как правило, служит модель, представляющая собой сумму (аддитивную смесь) детерминированной (закономерной) последовательности d_i и случайной («нерегулярной») последовательности ε_i:

$$x_i = d_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{8.5.2}$$

Кроме этого вида модели, рассматривают и другие; в частности, изучают мультипликативную модель вида

$$x_i = d_i \times \varepsilon_i,$$

которая путем логарифмирования левой и правой частей преобразуется в модель (8.5.2). Однако в дальнейшем анализируется аддитивная математическая модель временного ряда вида (8.5.2).

Применительно к модели (8.5.2) возможны различные варианты зависимости ее от времени: влияние времени может сказываться либо только на детерминированной последовательности d_i, либо только на случайной последовательности ε_i, либо, наконец, на обеих этих составляющих.

Детерминированная составляющая временного ряда d_i обусловлена определенными законами или причинами и вычисляется по некоторому правилу как функция времени. Детерминированную последовательность d_i часто называют систематической составляющей. Если эксперимент можно повторить многократно, то значения должны были бы оставаться одними и теми же, в то время как случайные составляющие и сами реализации временного ряда должны быть различными. Иначе говоря, полагается, что детерминированная составляющая есть математическое ожидание (среднее значение) временного ряда $\mathrm{M}\{y_i\} = d_i$.

Случайная составляющая временного ряда ε_i обязана наложением большого числа внешних факторов, не участвующих в образовании детерминированной составляющей и оказывающих каждый отдельно незначительное влияние на изменение результирующих статистических характеристик этой составляющей. Случайные составляющие ε_i могут включать в себя и ошибки наблюдения. При этом обычно считается, что случайные составляющие образуют стационарный нормальный случайный процесс с нулевым средним значением $\mathrm{M}\{\varepsilon_i\} = 0$, дисперсией $\mathrm{M}\{\varepsilon_i^2\} = D_\varepsilon$. В общем виде случайные составляющие могут полагаться как коррелированными, так и некоррелированными в зависимости от характера решаемой задачи. Случайную составляющую, получающуюся после вычитания из ряда его детерминированной составляющей, обычно называют остатком временного ряда. Если составляющие остатки независимы, их называют «ошибкам наблюдений (измерений)».

Основная задача анализа временных рядов состоит в выделении на основе знания отрезка временного ряда $x_1, x_2, \ldots, x_n$ детерминированной и случайной составляющих (компонент), а также в оценке их статистических характеристик.

В свою очередь, детерминированную составляющую в общем случае обычно представляют как сумму трех компонент

$$d_i = \varphi_i + s_i + c_i. \tag{8.5.3}$$

Здесь $\varphi(t)$ — долговременное (систематическое) изменение измеряемой величины, или тренд; $s(t)$ — сезонная компонента (кратковременные циклические колебательные относительно тренда с периодом p); $c(t)$ — долговременное циклическое колебание относительно тренда.

Вместе с этим следует отметить, что представление детерминированной составляющей в виде суммы трех компонент является достаточно условным. Кроме того, часто составляющая $d(t)$ содержит только некоторые из указанных выше компонент. Иногда при анализе временных рядов детерминированную составляющую представляют в виде

$$d_i = \varphi_i s_i + c_i, \tag{8.5.4}$$

т.е. тренд и сезонная компоненты образуют мультипликативную смесь.

Трендом временного ряда называют долговременную медленно и, как правило, монотонно изменяющуюся, компоненту детерминированной составляющей, описываемую той или иной неслучайной функцией $\varphi(t)$. Простейшей моделью тренда является линейная зависимость от времени вида $\varphi(t) = b_0 + b_1 t$, которая оказывается приемлемой во многих прикладных задачах анализа временных рядов. Более общей моделью тренда является полиномиальная зависимость от времени, а также нелинейные модели, подобные описанным выше. При этом следует отметить, что усложнение математического описания модели, как правило, не приводит к повышению точности описания тренда.

Сезонной компонентой называют функцию времени, обусловленную действием периодически повторяющимися факторов, причем период p неслучайной функции $s(t)$, описывающий эту компоненту, для упрощения дальнейшего анализа полагают кратным «сезонам» (день, месяц, год и т.д.). Из-за периодичности сезонной компоненты, естественно, в аналитическом выражении $s(t)$ должны присутствовать, как правило, тригонометрические функции, периодичность которых обусловлена содержательной сущностью задачи. Поскольку сезонная компонента часто служит главным источником краткосрочных колебаний временного ряда, поэтому ее выделение и исключение снижает вариацию (изменение во времени) остаточных компонент.

Циклической долговременной компонентой называют неслучайную функцию времени, проявляющую себя на протяжении длительного времени в результате действия факторов, обладающих большим последействием или медленными циклическими изменениями со временем. Примером такого явления служит влияние на многие процессы, происходящие в природе, изменения солнечной активности с периодичностью 11,2 года. Следует заметить, что в силу конечного по времени ряда циклическую компоненту иногда трудно отличить от тренда; поэтому часто их рассматривают как единое целое или циклическую компоненту рассматривают вместе со случайной составляющей. Примером циклической компоненты являются конечные отрезки ряда Фурье, представляющие собой конечные суммы пар синусоидальных и косинусоидальных составляющих. Такой парой, в частности, является функция $c_i = \alpha\cos\omega i + b\sin\omega i$, $0 < \omega < \pi$. Ее можно также представить как $c_i = \rho\cos(\omega i + \theta)$; период этой функции времени $T = 2\pi/\omega$, т.е. циклическая компонента повторяет свои значения всякий раз по прошествии времени T. Величина $\omega/(2\pi)$, обратная периоду T, является частотой, а коэффициенты $\rho = \sqrt{\alpha^2 + \beta^2}$ и $\theta = \operatorname{arctg}(\beta\sin\omega i/\alpha\cos\omega i)$ — ам-

плитудой и фазой соответственно. Как правило, период T значительно превышает длину анализируемого ряда, и поэтому при графическом представлении ряда визуально эта компонента не просматривается.

При анализе временного ряда предполагается, что поведение описанных выше компонент обусловлено различными группами причин. Ниже описываются основные направления анализа временных рядов. Подробное описание методов анализа временных рядов приведено в обширной литературе по математической статистике, в том числе и в [2, 32, 33, 62 и др.].

Модель временного ряда и его статистический анализ должны выбираться в соответствии с конкретными сведениями о причинах (механизмах), вызывающих изменение ряда во времени.

Определив оценки детерминированной и случайной составляющих, можно решать задачи прогноза будущих значений как самого временного ряда, так и его составляющих.

Табличное представление временного ряда и простейшие описательные статистики не позволяют понять характер изменения временного ряда. Поэтому первоначальный анализ временного ряда обычно начинается с построения и изучения его графика. На основе графика качественно можно определить наличие или отсутствие тренда, сезонных и циклических колебательных компонент, а также случайной составляющей. График дает возможность качественно определить и характер изменения компонент во времени.

Вторая стадия анализа временного ряда в значительной степени зависит от конкретной решаемой задачи, а также от характера изменения временного ряда и наличия тех или иных компонент детерминированной составляющей и может осуществляться различными способами. Например, возможны ситуации, когда априори известно, что анализируемый ряд содержит только часть из возможных компонент, в том числе ряд может состоять только из случайной составляющей. Поэтому, если имеются сомнения, что анализируемый ряд не содержит детерминированной составляющей, т.е. ряд представляет собой лишь случайную составляющую, проверяется гипотеза случайности. Существует множество критериев для проверки этой гипотезы, как правило, основанных на порядковых статистиках; некоторые порядковые критерии случайности обсуждаются в п. 5.6 и п. 7.2. Дальнейший анализ проводится в зависимости от справедливости той или иной гипотезы. Если гипотеза случайности (с выбранным или заданным уровнем значимости) не отвергается, дальнейший анализ ряда производится методами анализа дискретных случайных процессов. В противоположной ситуации изучается детерминированная составляющая и ее компоненты.

Анализ детерминированной составляющей. При этом анализе возможны ситуации, когда априори (с точностью до неизвестных параметров) известна функция тренда. Задачи статистических выводов в этом случае будут связаны именно с этими параметрами, которые оце-

ниваются методами регрессионного анализа. При этом необходимо отметить, что основные предположения, лежащие в основе регрессивного анализа, а именно нормальность распределений значений случайной составляющей и их независимость, редко выполняется в реальных ситуациях из-за наличия колебательной компоненты временного ряда. Поэтому полученные оценки коэффициентов модели тренда и доверительные интервалы для них необходимо рассматривать приближенными. Вместе с этим надо отметить, что в литературе, указанной выше, описываются прикладные методы регрессионного анализа и для коррелированных составляющих ряда. Если модель оказывается более неопределенной, то приходится применять более сложные непараметрические методы ее математического описания.

Сезонная компонента может представлять как самостоятельный интерес (связь, астрономия, экономика и др. науки), так и являться мешающим фактором. Часто эти периодические функции являются тригонометрическими функциями или могут быть выражены в виде линейной комбинации последних. Для выделения и устранения сезонной компоненты с известным периодом p имеется ряд способов. Принципиальным подходом при выделении сезонной компоненты является то, что сравниваются значения ряда через определенный период времени. Если априори известно, что детерминированная составляющая включает в себя только тренд $\varphi(t)$ и сезонную компоненту $s(t)$, то временной ряд принимает вид

$$x_i = \varphi_i + s_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{8.5.5}$$

Тогда для произвольного времени $t_i \equiv i$ справедлива запись $s_i = s_{i+p}$, где p — заранее известный период последовательности. По наблюденным значениям сначала методом регрессионного анализа определяется оценка тренда, а затем для каждого «сезона» i, $1 \leqslant i \leqslant p$, формируются разности вида

$$x_i - \varphi_i, \quad x_{i+p} - \varphi_{i+p}, \ldots, x_{i+mp} - \varphi_{i+mp}. \tag{8.5.6}$$

Усреднение этих разностей позволяет оценить сезонную компоненту

$$\bar{s}_i = \frac{1}{m+1} \sum_{l=0}^{m} [x_{i+lp} - \varphi_{i+lp}], \quad i = 1, 2, \ldots, p. \tag{8.5.7}$$

В литературе обсуждаются и другие оценки сезонной компоненты, в том числе на основе взвешенных средних, выборочной медианы, уменьшающие влияние резко выделяющихся наблюдений. В частности, если период сезонной компоненты не известен, для ее оценки используются методы, основанные на спектральном анализе временного ряда (метод периодограмм).

Когда вид детерминированной составляющей временного ряда не известен, альтернативой к нахождению детерминированной составляющей служит поиск полинома на основе метода «скользящих» средних.

Метод скользящих средних основан на усреднении случайных составляющих около выбранной точки временного ряда. Например, если в каждой точке временного ряда дисперсия постоянна и равна σ^2, то дисперсия среднего из k членов временного ряда уменьшается в k раз, т.е равна σ^2/k.

Сущность метода скользящего среднего состоит в следующем. Для первых k членов ряда строится полином степени q, не большей, чем $k-1$, и находится значение полинома в середине его области определения. Затем этот процесс повторяется для k членов, начиная со второго и до $(k+1)$-го. Как правило, выбирают нечетное число членов ряда, равное $k = 2m + 1$, в результате чего середина интервала является моментом времени, где производится сглаженная оценка временного ряда.

В общем виде, сглаженное значение временного ряда можно представить как [2, 32, 33, 62 и др.]

$$y_i = \sum_{j=-m}^{m} w_j x_{i+j}, \quad i = m+1, m+2, \ldots, n-m. \tag{8.5.8}$$

Здесь $x_{i-m}, x_{i-m+1}, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_{i+m}$ — значения исходного временного ряда; $w_j (j = -m, -m+1, ..., m)$ — «весовые» коэффициенты, определяемые аппроксимируемым полиномом, причем для них выполняется соотношение

$$\sum_{j=-m}^{m} w_j = 1.$$

Рассматриваемые методы скользящего среднего отличаются друг от друга лишь выбором параметра m и видом весовых коэффициентов w_j. Вид весовых коэффициентов («весов») определяется порядком «сглаживающего» полинома, а величина m обычно не превышает трех.

В прикладных задачах очень часто в качестве полинома берется простейший линейный полином вида (8.2.2), коэффициенты которого определены на основе заранее выбранного временного интервала. Учитывая, что значение средней точки простой линейной регрессии равно среднему арифметическому измеренных значений ряда, его сглаженное значение y_i можно представить в виде

$$y_i = \frac{1}{2m+1} \sum_{j=-m}^{m} x_{i+j}, \quad i = m+1, m+2, \ldots, n-m. \tag{8.5.9}$$

Иначе говоря, начальные значения ряда заменяются средними значениями на интервале времени, длина которого выбрана заранее. При этом сам интервал скользит вдоль ряда. Из (8.5.9) видно, что при линейном характере локальной аппроксимации весовые коэффициенты равны между собой, а их сумма равна единице.

Сглаженные значения y_i, как среднее $2m+1$ независимых переменных, имеют дисперсии $D_y = D_\varepsilon/(2m+1)$. Ковариация последовательных значений y_i определяется формулой

$$\mathrm{M}\{y_i y_{i\pm j}\} = \begin{cases} D_y(2m+1-|j|)/(2m+1)^2, & j = 0; \pm 1, \ldots, \pm 2m, \\ 0, & j = \pm(2m+1), \ldots. \end{cases} \tag{8.5.10}$$

из которой видно, что коррелированными составляющими сглаженного ряда являются составляющие отстоящие от момента i не более, чем на $\pm(m+1)$.

В [2, 32, 33, 62 и др.] метод скользящего среднего конкретизирован для полиномов различного порядка.

Рассмотренные выше методы скользящего среднего для выделения детерминированной составляющей ряда и ее компонент требуют (перед началом анализа) определения соответствующих функциональных зависимостей с точностью до их параметров. Однако данная процедура не всегда возможна и продуктивна. Поэтому часто для выделения детерминированной составляющей и получения прогноза применяют метод экспоненциального сглаживания. Этот метод часто называют методом Брауна. Основным требованием, предъявляемым к методу экспоненциального сглаживания, является достаточно медленное изменение детерминированной составляющей. При этом для упрощения анализа полагают, что значения случайной составляющей в разные моменты времени некоррелированы.

Метод экспоненциального сглаживания основан на представлении сглаженного значения ряда в момент i в виде суммы ранее наблюденных значений x_{i-j} с весовыми коэффициентами, которые уменьшаются по мере удаления в прошлое «время» $k = i - j$ экспоненциально (по закону убывающей геометрической прогрессии). Применительно к достаточно длинному ряду, когда крайние предшествующие значения его не влияют на значение ряда в момент наблюдения i (условно ряд полагается бесконечным), сглаженное значение ряда (экспоненциальное среднее) представляется в виде

$$y_i = \sum_{j=0}^{\infty} \lambda^j x_{i-j} = (1-\alpha) \sum_{j=0}^{\infty} \alpha^j x_{i-j}. \tag{8.5.11}$$

Здесь

$$\lambda^j = (1-\alpha)\alpha^j \tag{8.5.12}$$

— весовые коэффициенты («веса»), где $0 < \alpha < 1$ — константа сглаживания.

Вид весовых коэффициентов получается в результате минимизации функционала

$$Q = \sum_{j=0}^{\infty} \lambda^j (y_{i-j} - x_{i-j})^2 = \min \tag{8.5.13}$$

по параметру λ. Минимум функционала (8.5.12) достигается при равенстве нулю его производной по λ при условии

$$\sum_{j=0}^{\infty} \lambda^j = 1.$$

Этот метод называется *обобщенным (взвешенным) методом наименьших квадратов.*

Из рассмотрения полученного сглаженного ряда видно, что наблюдения входят в процедуру сглаживания с экспоненциально убывающими весами $\lambda^j = (1-\alpha)\alpha^j$. Качественно эти объясняется тем, что более близкие наблюдения к моменту оценки ряда воспринимаются с большим доверием, чем наблюдения, более отдаленные от момента оценки. При этом число членов ряда, «значимых» по своей величине, уменьшается по мере уменьшения значения константа сглаживания α. Если, например, $\alpha = 0{,}4$, то текущее наблюдение будет иметь вес 0,4, а веса предшествующих наблюдений составят соответственно 0,16; 0,064; 0,0256 и т.д. В отличие от рассмотренного выше метода скользящего среднего в методе экспоненциального сглаживания скользит только правый конец интервала усреднения. Учитывая, что веса экспоненциально убывают, то при достаточно большой длине ряда его прошлые значения быстро стремятся к нулю. При этом для рассматриваемого метода усреднения справедливо рекуррентное соотношение

$$y_i = (1-\alpha)x_i + \alpha y_{i-1}. \tag{8.5.14}$$

Для весовых коэффициентов выполняется условие

$$\sum_{j=1}^{\infty} \lambda^j = (1-\alpha)\sum_{j=1}^{\infty} \alpha^j = 1. \tag{8.5.15}$$

Так как оператор экспоненциального сглаживания линеен, он применим ко всем составляющим временного ряда, в том числе и к случайной составляющей ε_i:

$$\delta_i = (1-\alpha)\sum_{j=1}^{\infty} \alpha^j \varepsilon_{i-j}. \tag{8.5.16}$$

Для принятого допущения о независимости исходных значений случайной составляющей в различные моменты времени $\mathrm{M}\{\varepsilon_i\} = 0$ среднее значение сглаженной случайной составляющей $\mathrm{M}\{\delta_i\} = 0$, а ее дисперсия

$$D_\delta = \mathrm{M}\{\delta_i^2\} = \frac{1-\alpha}{1+\alpha} D_\varepsilon, \tag{8.5.17}$$

где D_ε — дисперсия случайной составляющей ε_i. Так как $0 < \alpha < 1$, дисперсия сглаженной случайной составляющей уменьшается по сравнению с дисперсией случайной составляющей исходного ряда. Это уменьшение тем больше, чем больше константа сглаживания α. Поэтому экспоненциальное сглаживание можно интерпретировать как цифровой низкочастотный фильтр, на вход которого поступает наблюдаемый ряд $x(t_1), x(t_2), \ldots, x(t_n)$, а на выходе фильтра формируется сглаженный ряд. Выбор значения константы λ зависит от статистических параметров ряда и характера его изменения во времени. Большое значение константы приводит к «сглаживанию» и изменяющейся детерминированной составляющей, что приводит к искажению ее действительного вида и уменьшению роли наиболее информативных последних членов ряда. Для уменьшения этого искажения λ целесообразно уменьшать. Поиск компромиссного значения константы λ является одной из принципиальных задач оптимизации процедуры экспоненциального сглаживания. Обычно наилучшее значение λ определяется экспериментально.

Применяя несколько раз процедуру сглаживания, а также подбирая соответствующим образом константу сглаживания, можно практически полностью исключить случайную составляющую, в результате чего останется только преобразованная детерминированная составляющая.

Для удаления детерминированной составляющей временного ряда используется метод, основанный на разностных операторах различных порядков. Поясним этот метод применительно к ряду

$$y_i = b_0 + b_1 i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{8.5.18}$$

детерминированная составляющая которого представляет собой линейную зависимость, а случайные составляющие, как и выше, полагаются некоррелированными с дисперсией D_ε. Разностными операторами первого и второго порядков называются соотношения вида $\Delta x_i = x_i - x_{i-1}$ и $\Delta^2 x_i = \Delta x_i - \Delta x_{i-1}$ соответственно. Используя оператор первого порядка, исходный ряд преобразуется к виду

$$y_i = \Delta x_i = b_0 + b_1 i + \varepsilon_i - b_0 - b_1(i-1) - \varepsilon_{i-1} = b_1 + \varepsilon_i - \varepsilon_{i-1}, \quad i = 1, 2, \ldots, n. \tag{8.5.19}$$

Из этого выражения видно, что, во-первых, преобразованный ряд y_i не содержит линейной составляющей; во-вторых, структура случайных составляющих $\delta = \varepsilon_i - \varepsilon_{i-1}$ также изменилась — они не являются независимыми. При этом нетрудно показать, что дисперсия и коэффициент корреляции между соседними членами новой последовательности равны $D_\delta = 2D_\varepsilon$ и $\rho_\delta = 0{,}5$ соответственно. Не соседние члены новой последовательности некоррелированы. Применяя оператор второго порядка, имеем

$$\tilde{y}_i = \Delta^2 x_i = \varepsilon_i - 2\varepsilon_{i-1} + \varepsilon_{i-2}, \quad i = 1, 2, \ldots, n.$$

Из последнего выражения видно, что новый преобразованный ряд $\tilde{y}$ уже не содержит детерминированной составляющей. При этом новая случайная составляющая также является коррелированной последовательностью

$$\tilde{\delta}_i = \varepsilon_i - 2\varepsilon_{i-1} + \varepsilon_{i-2}, \quad i = 1, 2, \ldots, n.$$

Дисперсия этого ряда $D_\delta = 6D_\varepsilon$, а коэффициент корреляции между членами новой последовательности определяется как $\rho_{i\pm1} = -1/6$, т.е. составляющие нового ряда некоррелированы, если суммарный временной интервал между членами ряда превышает два элементарных интервала временного ряда.

Анализ случайной составляющей. После определения (исключения) детерминированной составляющей ряда определяется случайная составляющая

$$\varepsilon_i = x_i - d_i, \quad i = 1, 2, \ldots, n. \tag{8.5.20}$$

Рассмотрим более подробно эту составляющую, возможные ее статистические модели, а также методы анализа. Знание этой информации позволяет с определенной надежностью прогнозировать значения ряда в будущее время.

Временной ряд, из которого исключена детерминированная составляющая или в нем ее не было, называется *стационарным временным рядом*. Различают стационарность в узком и широком смыслах. Ряд называется стационарным в узком смысле, если последовательные группы значений $\varepsilon_{i+1}, \varepsilon_{i+2}, \ldots, \varepsilon_{i+j}$ имеют одинаковое многомерное распределение при любых значениях i и j. Ряд называется стационарным в широком смысле, если среднее значение и дисперсия не зависят от времени

$$\mathrm{M}\{\varepsilon_i\} = m_\varepsilon = \mathrm{const}; \quad D_\varepsilon = \mathrm{M}\{(\varepsilon_i - m_\varepsilon)^2\} = \mathrm{const}, \tag{8.5.21}$$

а автоковариация K_j и автокорреляция ρ_j

$$\mathrm{M}\{(\varepsilon_i - m_\varepsilon)(\varepsilon_{i-j} - m_\varepsilon)\} = K_{-j} = K_j; \quad \rho_j = K_j/D_\varepsilon \tag{8.5.22}$$

являются четными функциями разности моментов наблюдения временного ряда. Стационарный случайный процесс обладает эргодическим свойством, если все его статистические параметры можно определить по одной реализации (последовательности) достаточно большой длительности n (в пределе $n \to \infty$). Простейшей математической моделью стационарного дискретного ряда является *дискретный белый шум*, который представляет собой стационарную независимую случайную последовательность с нулевым средним значением, автоковариацией

$$\mathrm{M}\{(\varepsilon_i - m_\varepsilon)(\varepsilon_{i-j} - m_\varepsilon)\} = K_j = \begin{cases} D_\varepsilon & \text{при } j = 0; \\ 0 & \text{при } j \neq 0 \end{cases} \tag{8.5.23}$$

и постоянной спектральной плотностью $G(\omega) = 1/2\pi$.

В дальнейшем полагается, что случайная составляющая представляет стационарный ряд в широком смысле и обладает эргодическим свойством. При этом выборочные среднее и дисперсия, а также j-й член *выборочной автокорреляции* (оценка автокорреляции между ε_i и ε_{i+j}) определяются соответственно как

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i; \quad s_\varepsilon^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2;$$
$$r_j = \frac{(n-1)\sum\limits_{i=1}^{n-j}(\varepsilon_i - \bar{\varepsilon})(\varepsilon_{i+j} - \bar{\varepsilon})}{(n-1-j)\sum\limits_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2}; \quad j = 1, 2, \ldots. \tag{8.5.24}$$

Приведенные соотношения аналогичны ранее полученным соотношениям при рассмотрении оценок параметров (статистических характеристик) случайной выборки. Выборочный коэффициент автокорреляции r_j называют также *сериальной корреляцией*, а совокупность коэффициентов r_j называют *коррелограммой*. Термин «коррелограмма» употребляется также для обозначения графика r_j как функции от j. Коррелограмма оказывается полезным инструментом при изучении структуры временных рядов.

Для членов ряда, разделенных s шагами, используют частные коэффициенты автокорреляции (см. пп. 1.5 и 7.1), которые устраняют влияние всех промежуточных членов ряда.

Для изучения структуры временных рядов широко используется спектральная плотность («спектр»), связанная с автокорреляцией преобразованием Фурье

$$g(\omega) = \sum_{k=-\infty}^{\infty}\rho_k \mathrm{e}^{-\mathrm{i}\omega k} = 1 + \sum_{k=1}^{\infty}\rho_k \cos\omega k. \tag{8.5.25}$$

В последнем соотношении учтено, что $\rho_j = \rho_{-j}$, а $\sin\omega k$ — нечетная функция от k. Спектр (8.5.25) — периодическая функция с периодом 2π. Он указывает на ритмы в ряду. Пики спектра, если они имеются, обычно отождествляются с гармоническими составляющими механизма, порождающего ряд.

Путем обратного преобразования Фурье автокорреляция определяется как

$$\rho_k = \frac{1}{\pi}\int_{-\pi}^{\pi} g(\omega)\mathrm{e}^{\mathrm{i}\omega k}\, d\omega. \tag{8.5.26}$$

Выбор для исследования структуры временного ряда коррелограммы или спектра зависит от ряда причин, в первую очередь от априорных знаний о механизмах, порождающих временной ряд. В экономических задачах предпочтение, как правило, отдается коррелограммам, которые более наглядны для анализа, в физических задачах наиболее широко используется спектральный анализ, хотя существуют области науки, где используются оба подхода (например, речевая акустика, биофизика и т.п.).

Для описания стационарных случайных временных рядов, их анализа и дальнейшего прогнозирования используется ряд математических моделей. Ниже приведено краткое описание математических моделей скользящего среднего и авторегрессии первых порядков. Модели более высокого порядка достаточно полно описаны в литературе, указанной выше. Эти модели оказываются весьма полезными во многих прикладных задачах.

Математической моделью процесса скользящего среднего первого порядка является

$$\delta_i = \varepsilon_i + \beta\varepsilon_{i-1}, \quad i = 1, 2, \ldots, n, \tag{8.5.27}$$

где ε_i — значения белого шума; $0 < \beta < 1$ — некоторый числовой коэффициент. Учитывая независимость случайных составляющих, ковариация рассматриваемого ряда равна

$$\mathrm{M}\{\delta_i\delta_{i-1}\} = \mathrm{M}\{\varepsilon_i\varepsilon_{i-1} + \beta\varepsilon_i^2 + \beta\varepsilon_i\varepsilon_{i-2} + \beta^2\varepsilon_{i-1}\varepsilon_{i-3}\} = \beta D_\varepsilon, \tag{8.5.28}$$

т.е. зависимыми оказываются только соседние величины δ_i и δ_{i+1}. Поэтому процессы скользящего среднего первого порядка являются простейшим обобщением последовательностей белого шума.

Математической моделью авторегрессии первого рода (называемой также марковским процессом) является

$$\delta_i = \eta\delta_{i-1} + \varepsilon_i, \tag{8.5.29}$$

где ε_i — случайная составляющая, имеющая нулевое среднее значение и дисперсию D_{ε_i}; η — некоторая постоянная величина, причем $|\eta| < 1$. При этом нетрудно вычислить дисперсию и автокорреляцию процесса порядка j

$$D_\delta = \frac{D_\varepsilon}{1-\eta^2}; \quad \rho_j = \frac{\mathrm{M}\{\delta_i\delta_{i\pm j}\}}{D_\delta} = \eta^j. \tag{8.5.30}$$

Из последней формулы видно, что, во-первых, автокорреляция экспоненциально убывает по мере взаимного удаления значений δ_i и δ_{i+j}, во-вторых, если величина $|\eta|$ близка к единице, то соседние значения ряда δ_i и δ_{i+j} сильно коррелированны, а дисперсия полученного ряда будет значительно больше дисперсии случайной составляющей ε_i.

В прикладной статистике описаны и другие более сложные математические модели случайной составляющей, а также их статистические характеристики.

Прогнозирование (экстраполяция). Во многих технических, экономических и других приложениях прикладной статистики часто возникают необходимость *прогнозирования* соответствующих явлений или параметров (показателей) функционирования наблюдаемых процессов и явлений на основе предыдущих наблюдений. Прогнозирование характеризует будущее развитие временного ряда. Оно основано на гипотезе о том, что сохраняются основные закономерности прошлого на период прогноза. В свою очередь эта гипотеза базируется на том факте, что факторы, порождающие временной ряд, достаточно инерционны, т.е. изменяются достаточно медленно со временем. Необходимо подчеркнуть, что не существует универсально предпочтительных методов прогнозирования. Поэтому выбор того или иного метода прогнозирования зависит от многих факторов, таких, как цель прогноза, число шагов вперед, для которых требуется прогноз, длина исходного временного ряда, наличие в наблюдаемом ряде тех или иных компонент детерминированной составляющей, наличие коррелированности значений случайной составляющей и т.п.

В зависимости от того, насколько шагов j вперед осуществляется прогноз, он условно разделяется на *краткосрочный* ($j \leqslant 3$), *среднесрочный* ($3 < j \leqslant 6$) и *долгосрочный* ($j > 6$). Обычно методами статистического анализа строят краткосрочные и среднесрочные прогнозы. Долгосрочное прогнозирование требует привлечения не только статистических экстраполяционных методов, но и других методов, в том числе и с использованием экспертных оценок.

Рассмотрим прогнозирование значения ряда y_τ в момент времени $t_\tau = t_{n+j} \equiv \tau = n + j$ по наблюденным значениям в предыдущие моменты времени $1, 2, \ldots, n$.

Если функция регрессии детерминированной составляющей известна, то, естественно, известно среднее значение прогнозируемой величины в момент времени τ. Это обстоятельство означает, что y_τ наилучшим образом прогнозирует значение ряда в момент времени τ в смысле минимума средней квадратичной ошибки прогноза.

Пример 8.5.1. В результате наблюдения некоторого процесса получена временная последовательность данных (временной ряд), представленная в табл. 8.5.1. На их основе на рис. 8.5.1 построен график временного ряда (сплошная линия с маркерами, соответствующими значениям временного ряда).

Из этого графика качественно видно, что временной ряд имеет как детерминированную составляющую в виде линейного тренда, так и случайную составляющую. Временной ряд запишем в виде

$$y_i = \varphi_i + \varepsilon_i = b_0 + b_1(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

где $\bar{x}$ — выборочное среднее. Как и выше, полагается, что элементы случайной составляющей ε_i являются независимыми гауссовскими величинами с нулевыми средними значениями и постоянной дисперсией.

Применяя формулы (8.2.3) и (8.2.4), находим выборочное среднее $\bar{x} = 10{,}5$ и оценки коэффициентов модели: $b_0^* = 15{,}34$; $b_1^* = 47$. Тогда оценка функции тренда запишется как

$$\varphi_i^* = 15{,}34 + 0{,}47(x_i - 10{,}5), \quad i = 1, 2, \ldots, n.$$

Таблица 8.5.1

x_i	1	2	3	4	5	6	7	8	9	10
y_i	11,84	11,52	11,11	12,40	13,01	11,52	12,69	15,35	13,52	15,98
x_i	11	12	13	14	15	16	17	18	19	20
y_i	16,30	16,11	16,77	18,13	16,38	18,29	19,13	17,54	20,10	19,10

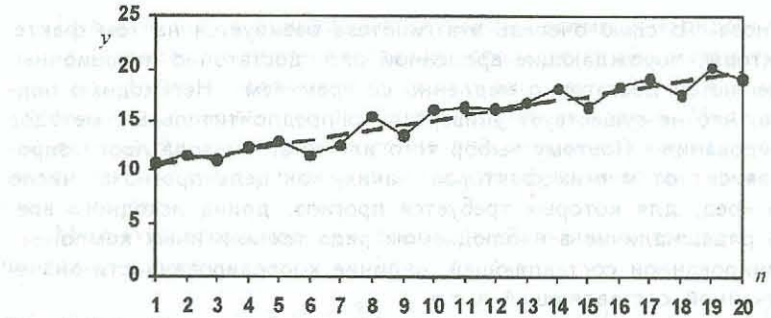


Рис. 8.5.1. График временного ряда

На рис. 8.5.1 пунктиром показана полученная оценка линейного тренда.

Верхняя и нижняя границы доверительного интервала определяются согласно (8.2.18). Для их вычисления необходимо найти по формуле (8.2.11) выборочную дисперсию $s^2_{Y|x} = 8{,}49$ ($s_{Y|x} = 2{,}91$) и задаться доверительной вероятностью γ (уровнем значимости $2\alpha = 1-\gamma$). Применительно к рассматриваемой задаче примем $\gamma = 0{,}95$ ($\alpha = 0{,}025$). Согласно табл. 3 Приложения критическое значение статистики Стьюдента $t_{\gamma=0{,}95} \equiv t_{\alpha=0{,}025} = 2{,}086$. В результате верхняя и нижняя границы доверительного интервала определяются соотношением

$$\varphi_i^* \pm 2{,}91 \cdot 2{,}086\sqrt{1/20 + \frac{(x_i - 10{,}5)^2}{665}} = \varphi_i^* \pm 6{,}067\sqrt{0{,}05 + 0{,}0015(x_i - 10{,}5)^2}.$$

Прогнозируемое значение временного ряда на j шагов и его доверительные границы определяется записанными выше выражениями, в которые вместо символа i надо поставить символ $n + j$.

Если функция регрессии детерминированной составляющей не известна, прогнозируемое значение временного ряда в момент времени $\tau = n + j$, представляется в виде взвешенной линейной суммы

$$y_\tau^* = \sum_{i=0}^{n} \lambda_i^* x_i.$$

Для оценки λ_i^* неизвестных коэффициентов λ_i требуют, чтобы прогноз был несмещенным, т.е. чтобы

$$\mathrm{M}\left\{\sum_{i=0}^{n} \lambda_i x_i - y_\tau^*\right\} = 0,$$

а дисперсия прогноза

$$\mathrm{M}\left\{\left(\sum_{i=0}^{n} \lambda_i x_i - y_\tau^*\right)^2\right\} = D_\varepsilon$$

была минимальной по сравнению с другими линейными оценками y_τ. Решение этой задачи осуществляется *обобщенным (взвешенным) методом наименьших квадратов*. В общем виде отыскание оптимальных оценок коэффициентов λ_i^* и их статистических характеристик представляет достаточно сложную задачу. Поэтому на практике применяют более простые методы получения прогнозируемых значений временного ряда.

Модель прогнозируемых значений достаточно длинного (теоретически бесконечного) ряда для $\tau \geqslant n+1$ часто можно представить в виде

$$y_\tau = \sum_{j=0}^{\infty} \lambda_j x_{\tau-j}, \quad \tau = n+1, n+2, \ldots. \tag{8.5.31}$$

Здесь $\lambda_0, \lambda_1, \ldots, \lambda_n$ – неизвестные коэффициенты ($0 < \lambda_j < 1$). Для неизвестных коэффициентов должны выполняться условия

$$0 < \lambda_j < 1; \quad \sum_{j=0}^{\infty} \lambda_j = 1. \tag{8.5.32}$$

Кроме этого, для упрощения задачи, как правило, задается определенная структурная связь между коэффициентами λ_j.

Одним из простейших методов прогнозирования, основанного на экспоненциальном сглаживании, является метод прогнозирования Брауна. В этом методе анализируемый процесс представляется полиномиальной моделью порядка p:

$$y_{\tau-n+j} = \sum_{i=0}^{p} \beta_i j^i, \quad j = 1, 2, \ldots. \tag{8.5.33}$$

Здесь полагается, что значение ряда в момент времени $t_n \equiv n$ является последним. Из (8.5.33) видно, что для прогноза начальная точка отсчета времени сдвигается в момент времени t_n, что существенно упрощает процедуру вычисления оценок коэффициентов модели.

Оценки коэффициентов β_i^* прогнозирующих полиномов, обеспечивающих минимум средней квадратической ошибки прогноза, определяются на основе обобщенного метода наименьших квадратов, т.е. из условия минимизации функционала

$$Q = \sum_{j=0}^{\infty} \lambda^j \left(y_{t-j} - \sum_{i=0}^{p} \beta_i j^i\right)^2 \tag{8.5.34}$$

Оценки коэффициентов прогноза β_i^* зависят от порядка полинома p, их оценок в момент времени t_n, значения константы сглаживания λ, а также от того насколько шагов j вперед осуществляется прогноз.

Доказывается (теорема Брауна), что оценки коэффициентов прогнозирующего полинома β_i^* линейно выражаются через сглаженные значения ряда y_{t-j}. Вычисление оценок коэффициентов прогноза (8.5.33) в общем виде представляет весьма сложную задачу, хотя для полиномов до третьего порядка оценки коэффициентов вычисляются достаточно несложно. При этом наиболее часто используется линейная ($p = 1$)

и квадратичная ($p = 2$) модели. В литературе приведены оценки этих коэффициентов для ряда значений порядка прогнозирующего полинома, а также получены рекуррентные формулы для вычисления прогнозирующих коэффициентов в момент времени $j + 1$ по их значениям в предыдущий момент времени j.

Если выполняются ранее принятые предположения, а именно наблюдения подчиняются нормальному закону с нулевым средним и постоянной дисперсией, можно сформировать статистику, подчиняющуюся распределению Стьюдента с $\nu = n - p$ степенями свободы. На основе этой статистики и заданного уровня значимости определяется доверительный интервал для прогноза y_τ.

В прикладных задачах прогнозирования достаточно широко применяется метод, основанный на простом соотношении вида

$$y_n^* = (1 - \lambda)x_n + \lambda y_{n-1}^*, \tag{8.5.35}$$

где $0 < \lambda < 1$ — весовые коэффициенты. Если константа λ известна, оценку y_n^* в момент времени t_n можно построить исходя из оценки, сделанной в момент времени t_{n-1}, путем прибавления к ней фактического наблюдения в момент t_n значения, умноженного на $(1-\lambda)$. Это значение может являться оценкой будущих членов ряда. Величина

$$e_n = x_n - y_{n-1}^* \tag{8.5.36}$$

является ошибкой, сделанной при построении в момент t_{n-1} прогноза на один шаг вперед, т.е. при прогнозировании величины x_n. Подставляя последнее соотношение в (8.5.35), получим

$$y_n^* = x_n - \lambda e_n = y_{n-1}^* + (1 - \lambda)e_n. \tag{8.5.37}$$

В результате прогноз в момент t_n является прогнозом y_n^*, сделанным в момент t_{n-1}, к которому добавляется последняя ошибка, умноженная на $(1-\lambda)$. Значение ошибки (8.5.36) используется для прогноза ряда не только на один, но и на несколько шагов вперед, т.е. в моменты времени $t_{n+1}, t_{n+2}, \ldots$. Прогнозируемые значения ряда определяются как

$$\begin{aligned} y_{n+1}^* &= y_n^* + (1 - \lambda)e_n; \\ y_{n+2}^* &= y_{n+1}^* + (1 - \lambda)e_n = y_n^* + 2(1 - \lambda)e_n; \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{aligned} \tag{8.5.38}$$

Пример 8.5.2. Пусть константа сглаживания $\lambda = 0{,}7$ и оценка значения временного ряда, сделанная на предыдущем шаге t_{n-1}, равна $y_n^* = 55$. В действительности оказалось, что значение временного ряда в момент времени t_n равно $x_n = 65$. Поэтому следующую оценку (на один шаг) получаем как

$$y_{n+1}^* = y_n^* + (1 - \lambda)e_n = 55 + (1 - 0{,}7)(65 - 55) = 58.$$

Различные методы прогнозирования разработаны и для случайной составляющей. Одной из них является экспоненциальная модель прогнозирования стационарного процесса с автокорреляционной функцией $\rho_k = \rho_1^k$, где ρ_1 — коэффициент автокорреляции соседних значений ряда, а k — число интервалов между значениями ряда. Подробное изложение методов прогнозирования случайной составляющей для разных моделей ее представления приведены в указанной выше литературе.

Контрольные вопросы и задачи

8.1. Какие задачи решаются регрессионным анализом?

8.2. В чем отличие выборочной средней квадратической регрессии от средней квадратической регрессии с известными моментами?

8.3. Как следует преобразовать нелинейные регрессии $\tilde{y} = \eta x^\beta$, $\tilde{y} = (\eta + \beta e^{-x})^{-1}$ в линейные регрессии?

8.4. Какие условия полагаются в теории выборочной средней квадратической регрессии?

8.5. Определить выборочную линейную регрессию $y^* = b_0^* + b_1^*(x - \bar{x})$ для значений случайных величин, приведенных ниже:

x_i	6,4	8,0	7,4	6,8	7,2	6,7	7,7	8,2
y_i	2,2	5,2	4,8	3,3	4,0	3,0	4,2	5,3

Для доверительной вероятности $\gamma = 0{,}9$ вычислить доверительные интервалы на концах и в середине значений случайной величины X.

8.6. В результате эксперимента в пяти точках плана произведено одинаковое число ($n_i = 3 = \text{const}$) измерений выходного эффекта y_{ik}, приведенных ниже в таблице.

x_i	1	2	3	4	5
y_{i1}	5,5	6,8	8,2	9,8	12,0
y_{i2}	6,4	6,9	8,4	10,4	11,8
y_{i3}	5,8	7,2	8,5	10,6	12,5

С уровнем значимости $\alpha = 0{,}05$ проверить гипотезу линейной зависимости выходного эффекта от значения фактора x.

8.7. Применяя метод прогнозирования, основанный на определении ошибки прогнозирования на последнем шаге n, вычислить прогнозируемое значение ряда в моменты времени $n+1$ и $n+2$, если известно, что в момент времени $n-1$ прогнозируемое значение ряда равно $y_n^* = 40$, хотя в действительности оно равнялось $y_n = 30$. Значение константы сглаживания $\lambda = 0{,}5$.

ГЛАВА 9
Основные понятия теории эксперимента

9.1. Сущность постановки задачи

Наблюдения, осуществляемые над явлениями и событиями окружающего мира, условно можно разделить на два больших класса: наблюдения над естественно происходящими явлениями (событиями) и наблюдения над искусственно воспроизводимыми явлениями, называемые *экспериментами* или *опытами*. При наблюдении над естественно происходящими явлениями с целью изучения их зависимостей от различного рода *факторов* (причин) исследователь, как правило, не имеет возможности не только многократно наблюдать изучаемые явления при одних и тех же условиях, но и целенаправленно количественно или качественно изменять *уровни* (*значения*) *факторов*, влияющих на исследуемые явления. В противоположность этому классу наблюдений эксперимент принципиально предусматривает возможность воспроизведения условий проведения опытов, многократного наблюдения изучаемых явлений, фиксации факторов и целенаправленного изменения уровней последних. Эксперимент является методом познания, при помощи которого в контролируемых и управляемых условиях исследуются интересуемые явления.

Экспериментальные методы исследования первоначально возникли в естествознании, а затем были распространены на изучение закономерностей других явлений, в том числе и на изучение общественных явлений. Для анализа результатов эксперимента стали применять методы математической статистики. По мере развития методов математической статистики и их прикладной стороны стало возможным расширить круг ее влияния и на процесс подготовки (планирования) эксперимента.

Планирование эксперимента — это постановка опытов по некоторой заранее составленной схеме, обладающей определенными оптимальными свойствами в соответствии с выбранными или заданными критериями оптимальности. В качестве критериев оптимальности при планировании эксперимента обычно выступают требования минимальных стоимости и продолжительности проведения опытов, а также достоверности получаемых экспериментальных выводов. Иначе говоря, задачей планирования эксперимента является получение как можно большей информации об изучаемом явлении при заданных «затратах».

Условно можно выделить следующие основные *этапы постановки эксперимента*:

1) формулирование исходных данных;
2) планирование;
3) проведение эксперимента;
4) анализ результатов и выводы.

Под исходными данными будем понимать:

- цель (задача) эксперимента;
- выбор зависимых переменных (откликов);
- выбор независимых переменных (факторов) и их значений (уровней).

Корректная, четкая и однозначная формулировка цели эксперимента во многом предопределяет дальнейшие этапы его планирования и проведения, предопределяет результативность выводов, способствует снижению затрат на осуществление эксперимента. Это особенно важно при проведении опытов со сложными, многофункциональными системами, зависящими от большого числа факторов, а также при осуществлении натурных экспериментов (натурных испытаний).

С формальной точки зрения прикладной математической статистики условно можно выделить два больших класса *целей эксперимента*.

1. Оценка влияния факторов на изучаемое явление; выявление из них наиболее *существенных* (причинных) *факторов* и их уровней. Применительно к научно-техническим задачам целью эксперимента является также сравнение опытных данных однородных объектов и технологических процессов.

2. Установление (уточнение) *математической* или *физической модели изучаемого явления* и оценка соответствия выбранной модели реальным опытным данным. В научно-технических задачах создание математической модели технологического или какого-либо другого процесса (явления) позволяет выбирать оптимальные режимы с целью получения максимального выходного продукта.

Основным математическим аппаратом для принятия или отклонения тех или иных решений (выводов) относительно полученных в результате эксперимента данных являются регрессионный анализ и дисперсионный факторный анализ проверки гипотез. Рассмотренные выше в гл. 3–5 вопросы оценок параметров распределений и проверки гипотез можно рассматривать как частные задачи сформулированных целей эксперимента.

После формулирования цели эксперимента необходимо выбрать зависимую переменную (или совокупность зависимых переменных), прямо или косвенно однозначно связанную с изучаемым явлением (событием). Зависимую переменную часто называют *выходным эффектом* объекта или *откликом* (*функцией отклика*). Если в большинстве научно-технических задач выбор наблюдаемых и измеряемых (контролируемых) зависимых переменных, как правило, не представляет затруднений, то

при исследованиях гуманитарных явлений выбор зависимых переменных, которые можно было бы измерить количественно, связан с определенными трудностями принципиального характера. При выборе независимых переменных (факторов) необходимо определить факторы, которые могут быть наиболее значимыми и которые в процессе опыта можно контролировать и изменять в интересуемых наблюдателя диапазонах допустимых значений. Необходимость выбора предполагаемых наиболее значимых факторов обусловлена тем, что с повышением их числа объем эксперимента существенно увеличивается, усложняются анализ опытных данных и процесс принятия решений. Поэтому, как бы ни казался на первый взгляд очевидным этап определения исходных данных, он является достаточно ответственной частью всего эксперимента, требует от исследователя тщательного изучения всех имеющихся у него априорных данных об изучаемом явлении. При постановке эксперимента часто приходится идти на определенный компромисс с целью удовлетворения противоречивых требований по объективности результатов исследований, с одной стороны, минимума числа опытов и наглядности представления их результатов для последующего анализа и принятия соответствующего решения, с другой стороны.

Вторым большим и не менее важным этапом эксперимента является его планирование. *При планировании эксперимента заранее составляется некоторая последовательность проведения опытов, в которой указывается порядок изменения факторов и их уровней.* До появления теории статистического планирования эксперимента (в дальнейшем — теории планирования эксперимента) проведение опытов, как правило, базировалось на интуиции исследователя и его предыдущих знаний в этой области. Однако создание теории планирования эксперимента позволило формализовать эту часть интеллектуального труда исследователя. В этой теории под планированием эксперимента понимается постановка опытов по определенной заранее составленной схеме (плану), обладающей оптимальными свойствами в соответствии с заранее выбранными критериями оптимальности. Цель теории планирования эксперимента состоит в том, чтобы предложить такой план, который по сравнению с традиционными способами проведения опытов позволил бы получить наибольшую информацию об изучаемом явлении при меньших затратах. Это обусловлено, в основном, тем, что получаемые в результате опыта данные всегда «затемняются» ошибками эксперимента. При помощи статистических методов обработки опытных данных, используя наиболее рациональный план проведения эксперимента, можно уменьшить эти ошибки.

Проиллюстрируем последнее утверждение простым примером точного взвешивания трех объектов A, B и C [49, 50]. План эксперимента при традиционном взвешивании приведен в табл. 9.1.1,*а*. В плане «+1» указывает, что объект взвешивания положен на весы, «−1» указывает на отсутствие объекта на весах. Согласно этому плану вначале

Таблица 9.1.1,*а*

Традиционная схема взвешивания

Номер опыта	A	B	C	y
1	−1	−1	−1	y_0
2	+1	−1	−1	y_1
3	−1	+1	−1	y_2
4	−1	−1	+1	y_3

Таблица 9.1.1,*б*

Матрица независимых переменных при взвешивании

Номер опыта	A	B	C	y
1	−1	−1	+1	y_1
2	−1	+1	−1	y_2
3	+1	−1	−1	y_3
4	+1	+1	+1	y_4

осуществляется «холостое» взвешивание y_0, в результате которого определяется нулевая точка весов, а затем по очереди взвешивается каждый из объектов y_i, $i = 1, 2, 3$.

На основании этих данных вычисляются оценки «веса» объектов. Например, оценка веса объекта A равна $A = y_1 - y_0$. Полагая ошибки измерения взвешиваний y независимыми с равными дисперсиями $\sigma^2(y)$, дисперсия оценки веса A

$$\sigma^2(A) = \sigma^2(y_1) + \sigma^2(y_0) = 2\sigma^2(y),$$

т.е. дисперсия оценки веса A равна удвоенной дисперсии ошибки взвешивания. Этому же значению равны дисперсии ошибок при определении весов объектов B и C.

Пусть тот же эксперимент осуществляется по плану, приведенному в табл. 9.1.1,*б*. Этот план представляет собой матрицу независимых переменных. Здесь в первых трех опытах последовательно взвешиваются объекты A, B и C, а в последнем опыте все три объекта взвешиваются вместе. По данному плану «холостое» взвешивание не проводится. Вес каждого объекта вычисляется по формулам

$$A = \frac{-y_1 - y_2 + y_3 + y_4}{2}; \quad B = \frac{-y_1 + y_2 - y_3 + y_4}{2};$$
$$C = \frac{+y_1 - y_2 - y_3 + y_4}{2}.$$

Здесь числители формально получены путем умножения элементов последнего столбца на элементы столбцов A, B и C. Поскольку результат взвешивания y_4 содержит оценки весов всех трех объектов, то в числителе каждой из формул вычитаются оценки весов других объектов. Число 2 в знаменателе обусловлено тем, что независимые оценки веса объекта входят в числитель дважды. Оценки результатов взвешивания независимы и характеризуются одинаковыми дисперсиями ошибок $\sigma^2(y)$. Из приведенных формул для оценок весов нетрудно получить дисперсии оценок. Применительно к оценке веса объекта A дисперсия оценки

$$\sigma^2(A) = \frac{\sigma^2(y_1) + \sigma^2(y_2) + \sigma^2(y_3) + \sigma^2(y_4)}{4} = \frac{4\sigma^2(y)}{4} = \sigma^2(y).$$

Аналогичным образом получаем $\sigma^2(B) = \sigma^2(y)$ и $\sigma^2(C) = \sigma^2(y)$.

Из этого иллюстративного примера видно, что нетрадиционный план эксперимента взвешивания обеспечивает дисперсию ошибки в два

раза меньше, чем традиционный способ. Увеличение точности опыта произошло из-за того, что во втором случае каждый вес вычислялся на основе четырех независимых опытов, в то время как при традиционном способе каждый вес вычислялся только на основе двух опытов. Вторая схема плана по существу является многофакторной, так как в ней оперируют всеми факторами.

Для правильного отображения объекта исследования необходимо оценить влияние на него различных факторов. Таких факторов может быть много, однако некоторые из *факторов* являются более *существенными* (причинными), чем другие, а ряд факторов вообще не оказывают какого-либо существенного влияния на выходной эффект изучаемого объекта. Так, например, в сложных технологических процессах число различных факторов может достигать несколько десятков и более. При этом степень их влияния на выходной эффект, как правило, весьма различна и лишь небольшое число факторов оказывает существенное влияние, а остальные воздействуют на выходной эффект практически незначительно. Не исключено, что при предварительном анализе могут быть предложены для рассмотрения такие факторы, которые вообще не влияют на выходной эффект. Задача эксперимента состоит в том, чтобы соответствующим образом выделить существенные факторы на «шумовом» фоне остальных факторов.

На первой стадии изучения всей предложенной совокупности факторов целесообразно использовать метод экспертных оценок на основе рассмотренного выше (п. 7.2) рангового критерия. Экспертные оценки, выполняемые специалистами в области данного эксперимента, позволяют получить, как правило, грубые оценки воздействия факторов на изучаемый объект (процесс) при меньших затратах времени и денежных средств. Каждый эксперт ранжирует предложенный перечень факторов, присваивая соответствующие ранги, причем ранг тем больше, чем существеннее, на взгляд эксперта, фактор. По результатам экспертного опроса составляется таблица суммарных рангов, в результате чего все предлагаемые для ранжирования факторы располагаются по степени их важности, причем, по мнению экспертов, наибольший суммарный ранг соответствует наиболее существенному фактору.



Рис. 9.1.1. Диаграмма рангов

Графически подобное ранжирование можно представить в виде диаграммы, примерный вид которой приведен на рис. 9.2.1. Факторы, суммарные ранги которых значительно меньше других факторов, можно интерпретировать как шумы и исключить их из дальнейшего анализа. Для того чтобы быть уверенным в том, что эксперты являются специалистами в рассматриваемой области эксперимента, проверяется гипотеза согласованности их мнений в соответствии с критерием (7.2.27), (7.2.36).

После грубого выделения предполагаемых существенных факторов производится более подробное изучение их влияния на выходной эффект с помощью эксперимента и использования для соответствующих выводов дисперсионного анализа.

Следующим этапом проведения эксперимента является его планировпние. Начало развитию теории планирования эксперимента было положено исследованиями английского статистика Р. Фишера, связанными с применением дисперсионного анализа для практического решения агробиологических проблем, направленных на повышение урожайности сельскохозяйственных культур. В 1935 г. Р. Фишер опубликовал монографию «Планирование эксперимента», в которой были изложены основные идеи нового направления прикладной статистики. Основополагающим моментом теории планирования эксперимента является введение в эксперимент принципа *рандомизации* (*случайности*) *при выборе уровней факторов*, сочетаний уровней факторов и получении выборочных данных. Считается, что из весьма большого количества положений, разработанных Р. Фишером в интересах математической статистики, принцип рандомизации, по-видимому, является одним из самых значительных его предложений.

При *классическом проведении эксперимента* (*классический факторный эксперимент – КФЭ*) для изучения влияния нескольких факторов на исследуемое явление предполагается, что последовательно изменяются уровни одного из причинных факторов при постоянстве уровней других факторов. При этом классическая теория ошибок основана на том, что случайные ошибки независимы и распределены нормально, т.е. рассеяние результатов эксперимента возникает благодаря постоянной системе случайных причин. Однако в реальных экспериментах, особенно при техническом эксперименте, обычно встречаются и другие источники рассеяния результатов наблюдения, включающие неоднородность экспериментальных данных и изменчивость уровней факторов (условий), не находящихся под контролем наблюдателя. В подобной ситуации неконтролируемые причины, как правило, не являются случайными и вызывают систематические ошибки, которые могут обуславливать относительно большое рассеяние опытных данных. Поэтому иногда трудно решить, что влияет на результат опыта: контролируемые и изменяемые факторы или неизвестное систематическое рассеяние данных. Итак, классический план проведения эксперимента (КФЭ) имеет два существенных недостатка.

Во-первых, классический принцип изменения уровней факторов, как правило, не позволяет количественно определить взаимодействие между уровнями факторов и может привести к неправильным выводам при определении уровней факторов, обеспечивающих достижение максимального эффекта изучаемого явления (отклика). Только одновременное изменение уровней всех факторов позволяет избежать подобной возможной ошибки.

Во-вторых, недостатком классического плана проведения эксперимента является опасность того, что за счет неконтролируемых причинных факторов возможно появление систематических ошибок, различных по своему значению для разных временных и пространственных интервалов проведения опытов. Кроме того, при осуществлении эксперимента никогда нельзя быть уверенным в том, что все существенно важные (причинные) факторы учтены. Возможны ситуации, когда некоторые факторы рассматривают как несущественные и ими умышленно пренебрегают. В результате таких неконтролируемых изменяющихся причинных факторов возможно принятие ошибочного решения относительно влияния каких-либо из них на изучаемое явление. Например, в процессе проведения опыта неучтенный неизвестный или ошибочно отброшенный причинный фактор может меняться таким образом, что конкретная комбинация уровней изучаемых факторов окажется ошибочно предпочтительной для достижения максимального выходного эффекта, хотя на самом деле к этому результату привела не комбинация исследуемых факторов, а неучтенный причинный фактор.

Сущность рандомизации эксперимента заключается в преобразовании неизвестного систематического рассеяния опытных данных в рассеяния, которые независимы и случайны по отношению к экспериментируемым объектам. Здесь и далее под объектами эксперимента будем понимать факторы и их уровни. Иначе говоря, при рандомизации эксперимента опыты по объектам необходимо разместить случайным образом. Правда, в данной ситуации предположение о нормальности экспериментальных данных, лежащее в основе дисперсионного анализа, не будет выполняться. При этом в пределах проводимых опытов на величину дисперсии ошибки влияют оба источника рассеяния: естественные нормально распределенные случайные ошибки и случайные ошибки, обусловленные неизвестным систематическим рассеянием. Поэтому при достаточно большом общем объеме выборки, рандомизация данных приводит к распределению отношения соответствующих дисперсий, близкому к F-распределению.

В теории планирования эксперимента рассматриваются различные способы рандомизации опытных данных. Сущность этих способов рандомизации и построение планов проведения опытов рассмотрим применительно к двум основным задачам планирования экспериментов: выявлению существенных (причинных) факторов и разработке математической модели, адекватно описывающей изучаемый процесс. Рассмотрение проведем применительно к специфике научно-технических задач.

9.2. Способы рандомизации

Как отмечалось выше, при проведении эксперимента необходимо осуществлять рандомизацию опытов. Рассмотрим некоторые, наиболее распространенные способы рандомизации.

Самым распространенным из способов рандомизации является полностью рандомизированный *полный факторный эксперимент* (ПФЭ). В соответствии с этим планом сначала определяется общее число опытов n, а затем порядок проведения этих опытов случайным образом распределяется по объектам эксперимента. Универсальный способ получения случайного порядка основывается на таблицах случайных чисел, приведенных в [8, 51 и др.]. ПФЭ обычно применяется тогда, когда экспериментальные данные не обнаруживают большой неоднородности, что наиболее характерно для лабораторных опытов при сохранении одних и тех же условий их проведения.

Если ожидается большая неоднородность условий при проведении эксперимента в пространстве и во времени, то с целью частичного исключения подобной неоднородности при дисперсионном анализе применяют более сложные планы экспериментов. Одним из них является так называемый *блочный факторный эксперимент* (БФЭ). Сущность этого плана состоит в том, что все опыты делятся на q групп (блоков), каждая из которых характеризуется достаточно однородными условиями проведения эксперимента. В пределах каждого блока объекты экспериментирования распределяются случайным образом. В этом случае на дисперсию ошибки влияет только рассеяние внутри блоков, поскольку рассеяния между блоками исключаются.

Пример 9.2.1. Изучим указанные выше способы рандомизации на следующем примере. Пусть поставлена задача экспериментально проверить влияние заданных четырех типов (разновидностей) A, B, C, D некоторого объекта X на выходной эффект. Здесь исследуемым фактором x является объект, а уровнями фактора — типы (разновидности) объекта, т.е. уровнями фактора являются $A \equiv x_1$, $B \equiv x_2$, $C \equiv x_3$, $D \equiv x_4$. Будем полагать, что каждый тип объекта испытывается пять раз ($n_i = 5 = \text{const}$), так что общее число опытов равно $n = 20$. Уровень значимости принятия нулевой гипотезы о том, что выходной эффект не зависит от типа объекта, выберем равным $\alpha = 0{,}05$. Одним из примеров подобной постановки задачи является проверка сравнительной эффективности четырех типов антенн при приеме радиосигналов.

Рассмотрим следующие две гипотетические модели проведения эксперимента при предположении, что тип объекта не влияет на выходной эффект. При этом предположении опытные данные x_j^*, $j = 1, 2, \ldots, 20$, для всех типов объекта являются выборочными значениями из одной и той же нормально распределенной генеральной совокупности со средним значением $m = 10$ и единичной дисперсией ($\sigma^2 = 1$). Получение опытных данных моделировалось с помощью таблицы независимых гауссовских случайных величин [8]. Они представлены во втором столбце табл. 9.2.1.

В первой модели будем полагать, что все типы объектов A, B, C и D испытываются в одинаковых условиях и отсутствуют систематические ошибки. Во второй модели предполагается, что условия проведения эксперимента изменяются таким образом, что «измеренная» величина x_j^* приобретает аддитивную составляющую, линейно возрастающую с увеличением номера опыта в виде $\lambda_j = 0{,}25j$. В результате выходной эффект равен $x_j = x_j^* + \lambda_j$. Аддитивная составляющая $\lambda_j = 0{,}25j$ и результирующее значение $x_j = x_j^* + \lambda_j$ приведены в третьем и четвертом столбцах табл. 9.2.1 соответственно. Согласно первой схеме в процессе проведения опытов отсутствуют систематические источники ошибок, поэтому в подобной ситуации приемлем классический факторный план проведения эксперимента (КФЭ), приведенный в пятом столбце таблицы. Проверим справедливость нулевой гипотезы H_0, согласно которой уровни исследуемого фактора (типы объектов) не являются существенными для выходного

Таблица 9.2.1

Таблица исходных данных

Номер опыта j	x_j^*	$\lambda_j = 0{,}25j$	$x_j = x_j^* + \lambda_j$	КФЭ x_i	ПФЭ x_i	БФЭ x_i
1	10,59	0,25	10,84	A	A	C
2	10,77	0,50	11,27	A	C	B
3	9,86	0,75	10,61	A	C	A
4	10,65	1,00	11,65	A	D	D
5	10,76	1,25	12,01	A	C	D
6	8,77	1,50	10,27	B	B	C
7	9,44	1,75	11,19	B	A	B
8	11,60	2,00	13,60	B	A	A
9	9,27	2,25	11,52	B	C	D
10	11,23	2,50	13,73	B	A	B
11	11,05	2,75	13,80	C	B	A
12	10,36	3,00	13,36	C	D	C
13	10,52	3,25	13,77	C	D	A
14	11,38	3,50	14,88	C	B	D
15	9,13	3,75	12,88	C	A	B
16	10,54	4,00	14,54	D	B	C
17	10,88	4,25	15,13	D	C	D
18	8,79	4,50	13,29	D	B	B
19	10,89	4,75	15,64	D	D	C
20	9,35	5,00	14,35	D	D	A

Таблица 9.2.2,*а*

Опытные данные КФЭ при отсутствии систематических ошибок

Уровень i	Номер элемента выборки j					$\sum_j x_{ij}$	$x_{i\cdot}$
	1	2	3	4	5		
$1 \to A$	10,59	10,77	9,86	10,65	10,76	52,63	10,526
$2 \to B$	8,77	9,44	11,60	9,27	11,23	50,31	10,062
$3 \to C$	11,05	10,36	10,52	11,38	9,13	52,44	10,488
$4 \to D$	10,54	10,88	8,79	10,89	9,35	50,45	10,090
$\sum_i \sum_j x_{ij} = 205{,}83; \quad x_{\cdot\cdot} = 10{,}2915$							

Таблица 9.2.2,*б*

Таблица дисперсионного анализа

Вид изменчивости	Q_k	ν_k	s_k^2	$F_{3,16}^*$
Между уровнями	$Q_X = 0{,}93$	$\nu_X = 3$	$s_X^2 = 0{,}31$	0,365
Внутри уровней	$Q_R = 13{,}67$	$\nu_R = 16$	$s_R^2 = 0{,}85$	
Общая	$Q = 14{,}60$	$\nu = 19$	$s^2 = 0{,}77$	

эффекта. Для проверки этого утверждения с помощью однофакторного дисперсионного анализа применительно к записанным «опытным» данным перепишем их в форме табл. 6.1.1. Они приведены в табл. 9.2.2,*а*. Осуществляя однофакторный дисперсионный анализ согласно схеме, приведенной в табл. 6.1.2, составим табл. 9.2.2,*б*.

Критическое значение дисперсионного отношения $F_{\alpha;\nu_1,\nu_2}$ для $\alpha = 0{,}05$ и $\nu_1 = 3$, $\nu_2 = 16$ определяемое по табл. 4 Приложения 2, $F_{0,05;3,16} = 3{,}24$. Так как $F_{3,16}^* = 0{,}365 < F_{0,05;3,16} = 3{,}24$, то решение о независимости выходного эффекта от типа испытуемого объекта (подтверждается нулевая гипотеза H_0) справедливо не только с уровнем значимости $\alpha = 0{,}05$, но и с уровнем значимости $\alpha = 0{,}25$, соответствующим согласно таблицам F-распределения [8] более «жесткой» критической границе $F_{0,25;3,16} = 1{,}51$.

Согласно второй модели в «эксперименте» присутствуют систематические ошибки, обусловленные неучтенным существенным фактором. Применительно к суммарным данным рассмотрим три плана проведения эксперимента: классический факторный эксперимент (КФЭ), полный факторный эксперимент (ПФЭ) и блочный факторный эксперимент (БФЭ). Порядки проведения опытов для ПФЭ и БФЭ согласно сказанным выше правилам рандомизации даны в шестом и седьмом столбцах табл. 9.2.1.

В соответствии с этими порядками проведения опытов и схемой однофакторного дисперсионного анализа составлены табл. 9.2.3, 9.2.4 и 9.2.5.

На основе этих таблиц были вычислены величины, необходимые для однофакторного дисперсионного анализа. Они даны в табл. 9.2.6.

Из анализа дисперсионных отношений применительно к рассматриваемому гипотетическому эксперименту можно сделать следующие выводы.

При КФЭ за счет систематических ошибок нулевая гипотеза H_0 не подтверждается, так как $F_{3,16}^* = 11{,}49 > F_{0,005;3,16} = 3{,}24$.

При ПФЭ и БФЭ систематическое изменение уровня существенного «неконтролируемого» фактора переводится в случайную ошибку, увеличивая ее дисперсию по сравнению с КФЭ соответственно в 2,61 и 3,06 раза; дисперсия случайной ошибки БФЭ по сравнению с дисперсий случайной ошибки ПФЭ увеличивается в 1,18 раза за счет дополнительного перевода систематической ошибки в случайную ошибку. Однако при этом выборочная статистика дисперсионного отношения БФЭ в 6,65 раз меньше соответствующей статистики ПФЭ. При этом для ПФЭ и БФЭ нулевая гипотеза не отвергается, так как выборочные статистики меньше критического значения.

Приведенный пример наглядно показывает эффективность рассмотренных двух видов рандомизации при планировании эксперимента. Вместе с этим эффективность рандомизации может быть повышена, так как даже БФЭ не всегда обеспечивает «равновероятные» возможности для опытов. Действительно, применительно к рассматриваемому примеру из табл. 9.2.1 перепишем первые четыре блока БФЭ в виде, показанном во втором (а) столбце табл. 9.2.7.

Из этой таблицы видно, что не все уровни (объекты) испытания находятся в «равноправном» положении. Например, уровень D стоит дважды первым (во втором и третьем блоках), уровень B — дважды вторым и т.д. Однако имеется возможность разместить уровни таким образом, чтобы в каждом столбце и в каждой строке уровень (объект опыта) появлялся только один раз. Например, уровни фактора можно разместить по строкам и столбцам так, как показано в третьем (б) столбце табл. 9.2.7.

Если в квадрате число строк и столбцов равно r, а уровни фактора по строкам и по столбцам не повторяются, то говорят, что данный квадрат называется *латинским квадратом* и имеет размер (объем) $r \times r$ (или порядок r). Приведенный в табл. 9.2.7 (столбец б) латинский квадрат имеет размер 4×4 (четвертый порядок).

План, в котором каждый вариант испытаний появляется один и только один раз в каждой строке и один и только один раз в каждом столбце, называется *планом латинского квадрата* [32, 50, 65, 67 и др.].

При планировании рандомизированного эксперимента конкретный латинский квадрат выбирается случайным образом из множества воз-

Таблица 9.2.3

Опытные данные КФЭ при наличии систематических ошибок

Уровень i	Номер элемента выборки j					$\sum_j x_{ij}$	$x_{i\cdot}$
	1	2	3	4	5		
$1 \to A$	10,84	11,27	10,61	11,65	12,01	56,38	11,276
$2 \to B$	10,27	11,19	13,60	11,52	13,73	60,31	12,062
$3 \to C$	13,80	13,36	13,77	14,88	12,88	68,69	13,738
$4 \to D$	14,54	15,13	13,29	15,64	14,35	72,95	14,590
$\sum\sum x_{ij} = 258{,}33; \quad x_{\cdot\cdot} = 12{,}9165$							

Таблица 9.2.4

Опытные данные ПФЭ при наличии систематических ошибок

Уровень i	Номер элемента выборки j					$\sum_j x_{ij}$	$x_{i\cdot}$
	1	2	3	4	5		
$1 \to A$	10,84	11,19	13,60	13,73	12,88	62,24	12,448
$2 \to B$	10,27	13,80	14,88	14,54	13,29	66,78	13,356
$3 \to C$	11,27	10,61	12,01	11,52	15,13	60,54	12,108
$4 \to D$	11,65	13,36	13,77	15,64	14,35	68,77	13,754
$\sum\sum x_{ij} = 258{,}33; \quad x_{\cdot\cdot} = 12{,}9165$							

Таблица 9.2.5

Опытные данные БФЭ при наличии систематических ошибок

Уровень i	Номер элемента выборки j					$\sum_j x_{ij}$	$x_{i\cdot}$
	1	2	3	4	5		
$1 \to A$	10,61	13,60	13,80	13,77	14,35	66,13	13,226
$2 \to B$	11,27	11,19	13,73	12,88	13,29	62,36	12,472
$3 \to C$	10,84	10,27	13,36	14,54	15,64	64,65	12,930
$4 \to D$	11,65	12,01	11,52	14,88	15,13	69,19	13,038
$\sum\sum x_{ij} = 258{,}33; \quad x_{\cdot\cdot} = 12{,}9165$							

Таблица 9.2.6

Сводная таблица дисперсионного анализа

Типплана	Q	Q_X	Q_R	s^2	s_X^2	s_R^2	$F^*_{3,16}$
КФЭ	50,54	34,48	16,06	2,66	11,49	1,00	11,49
ПФЭ	50,54	8,84	41,70	2,66	2,95	2,61	1,13
БФЭ	50,54	1,54	49,00	2,66	0,51	3,06	0,17

Таблица 9.2.7

Размещение уровней фактора по блокам

Блоки	Уровни фактора (а)	Уровни фактора (б)
1	$C\ B\ A\ D$	$C\ D\ A\ B$
2	$D\ C\ B\ A$	$B\ A\ D\ C$
3	$D\ B\ A\ C$	$D\ C\ B\ A$
4	$A\ D\ B\ C$	$A\ B\ C\ D$

Таблица 9.2.8

Стандартные латинские квадраты порядка 4 × 4

1	2	3	4
$A\ B\ C\ D$ $B\ A\ D\ C$ $C\ D\ B\ A$ $D\ C\ A\ B$	$A\ B\ C\ D$ $B\ C\ D\ A$ $C\ D\ A\ B$ $D\ A\ B\ C$	$A\ B\ C\ D$ $B\ D\ A\ C$ $C\ A\ D\ B$ $D\ C\ B\ A$	$A\ B\ C\ D$ $B\ A\ D\ C$ $C\ D\ A\ B$ $D\ C\ B\ A$

можных латинских квадратов данного порядка. Число возможных латинских квадратов r-го порядка достаточно велико. При этом принципиально путем перестановок столбцов и строк каждый квадрат может быть приведен к стандартному виду, в котором верхняя строка и левый столбец расположены в порядке $ABCD$ и т.д. В свою очередь из каждого стандартного латинского квадрата порядка r путем перестановок строк и столбцов может быть образовано $k = r!(r-1)!$ различных латинских квадратов этого же порядка.

Применительно к рассматриваемым латинским квадратам четвертого порядка (размерности 4×4) имеется четыре стандартных квадрата, приведенных в табл. 9.2.8. Путем соответствующих перестановок столбцов и строк каждого стандартного квадрата может быть образовано $k = 4!(4-1)! = 144$ латинских квадрата (не имеет смысла делать перестановку первой строки или первого столбца, так как в результате получится повторение уже имеющегося квадрата). При этом полное число квадратов равно $k' = 4 \cdot 144 = 576$. Латинский квадрат, представленный во втором (б) столбце табл. 9.2.7, путем соответствующих перестановок строк (первую строку — на третье место, а четвертую строку — на первое место) может быть сведен к четвертому стандартному квадрату табл. 9.2.8. Как отмечалось выше, число латинских квадратов существенно возрастает при увеличении их размера. Так, согласно [32], имеется 161 280 квадратов размера 5×5 и 812 851 200 квадратов размера 6×6.

Таким образом, для рандомизации плана эксперимента с помощью латинских квадратов необходимо случайным образом выбрать стандартный квадрат, а затем с помощью таблиц случайных перестановок или каким-либо другим случайным способом переставить строки и столбцы выбранного случайного стандартного квадрата. Однако в литературе по планированию эксперимента существуют полные таблицы латинских квадратов размера от 4×4 до 6×6 и примеры квадратов вплоть до размера 12×12. Для подобной ситуации способ рандомизации упрощается и состоит в выборе случайного латинского квадрата из соответствующей таблицы латинских квадратов.

Квадраты называют латинскими в связи с тем, что уровни фактора, как правило, обозначают латинскими буквами. Интересно отметить, что изучение свойств латинских квадратов, проведенное Эйлером в XVIII веке с чисто математическим интересом, нашло свое прикладное применение спустя два века.

Таблица 9.2.9
Неполноблочный сбалансированный план

Блоки	Уровни фактора			
	A	B	C	D
1	+	−	+	+
2	−	+	+	+
3	+	+	+	−
4	+	+	−	+

Таблица 9.2.10
Пример латинского квадрата размера 4 × 4 для трех факторов с четырьмя уровнями

Тип здания	Этажность здания			
	a	b	c	d
1	C	D	A	B
2	B	A	D	C
3	D	C	B	A
4	A	B	C	D

Возможны ситуации, при которых на рандомизацию приходится налагать определенные ограничения, обусловленные особенностями эксперимента. Например, в течение определенного времени из четырех объектов можно испытать только любые три из них. Применительно к латинским квадратам размера 4 × 4 в каждом из четырех блоков возможно лишь три уровня из четырех. В этом случае используются, так называемые, *неполноблочные сбалансированные планы*, один из которых приведен в табл. 9.2.9.

Знак «+» обозначает, что опыт проводится, знак «−» — что опыт не проводится. План называется *неполноблочным*, потому что этот план имеет меньше вариантов опытов, чем может поместиться в одном блоке, или, иначе говоря, в каждом блоке опущено по одному варианту опыта. План является *сбалансированным* в связи с тем, что одинаковое количество вариантов опытов в блоках не повторяется. Подобный план эксперимента за счет случайного выбора из множества подобных неполных квадратов полностью рандомизирован, хотя он и уступает полноблочному плану латинских квадратов по своей эффективности (компенсации систематических ошибок).

Приведенные полноблочные планы и неполноблочные сбалансированные планы можно рассматривать как двухфакторные. Второй фактор также имеет четыре уровня и уровни этого фактора также можно рандомизировать. В этом случае, осуществляя двухфакторный дисперсионный анализ, можно проверить не только нулевую гипотезу об отсутствии зависимости выходного эффекта от уровней изучаемого фактора, но также проверить и гипотезу о независимости выходного эффекта от номера блока или второго фактора. Этот второй фактор иногда называют *второстепенным* или *мешающим* [32, 50].

Планы на основе латинских квадратов могут усложняться в соответствии с теми или иными условиями проведения эксперимента.

Пусть исследуется зависимость показаний основного фактора от двух второстепенных факторов, число уровней которых совпадает с числом уровней основного фактора. На процесс рандомизации уровней могут быть наложены определенные ограничения, состоящие в том, что порядок испытания уровней основного фактора осуществляется независимо для каждого из второстепенных факторов.

Рассмотрим план трехфакторного латинского квадрата четвертого

Таблица 9.2.11
Латинские квадраты

C	D	A	B
B	A	D	C
D	C	B	A
A	B	C	D

δ	γ	β	α
γ	δ	α	β
β	α	δ	γ
α	β	γ	α

Таблица 9.2.12
Греко-латинский квадрат для четырех уровней

Тип здания	Этажность здания			
	a	b	c	d
1	$C\delta$	$D\gamma$	$A\beta$	$B\alpha$
2	$B\gamma$	$A\delta$	$D\alpha$	$C\beta$
3	$D\beta$	$C\alpha$	$B\delta$	$A\gamma$
4	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$

порядка на примере, являющемся развитием ранее рассмотренного случая оценки эффективности четырех типов приемных антенн (A, B, C, D). Антенны могут располагаться в одном из четырех типов зданий (1, 2, 3, 4) с четырьмя типами этажности (a, b, c, d). Здесь основным фактором являются типы антенн, а второстепенными факторами — типы зданий и варианты этажности. Латинский квадрат, соответствующий этому примеру, приведен в табл. 9.2.10. Он состоит из $n = 4^2 = 16$ точек плана и позволяет проверить гипотезы о влиянии уровней второстепенных факторов на эффективность антенн.

При увеличении количества факторов могут использоваться планы, называемые *греко-латинскими квадратами*. Сущность этого плана рассмотрим на дальнейшем развитии предыдущего примера. Изучим ситуацию, при которой кроме двух предыдущих второстепенных факторов дополнительно учитывается четвертый фактор — тип помещения, в котором может быть расположена приемная антенна. Пусть этот фактор также имеет четыре уровня $(\alpha, \beta, \gamma, \delta)$.

Уровни четвертого фактора $(\alpha, \beta, \gamma, \delta)$ можно так скомбинировать с уровнями основного фактора (типы приемных антенн), чтобы они появлялись один и только один раз в сочетании с каждым из уровней (A, B, C, D). Для этого рассмотрим два латинских квадрата четвертых порядков, составленных соответственно из латинских и греческих букв и приведенных в табл. 9.2.11.

Если наложить эти два латинских квадрата один на другой и составить третий квадрат, каждая клетка которого содержит как латинскую, так и греческую букву соответствующих клеток исходных квадратов, то получится новый квадрат, который применительно к рассматриваемому примеру с учетом латинского квадрата, приведенного в табл. 9.2.10, примет форму, представленную в табл. 9.2.12.

В этом квадрате каждая комбинация греческой и латинской буквы встречается только один раз. *Два исходных латинских квадрата, образующих греко-латинский квадрат, называются ортогональными.*

В плане, приведенном в табл. 9.2.12, как и выше, введено ограничение на рандомизацию четвертого фактора — уровни четвертого фактора $(\alpha, \beta, \gamma, \delta)$ рандомизированы только по строкам и столбцам плана.

Недостатком планов с использованием греко-латинских квадратов является слишком большая их насыщенность изучаемыми факторами

малого числа экспериментов. Так, применительно к греко-латинскому квадрату, приведенному в табл. 9.2.12, дисперсии между уровнями каждого из четырех изучаемых факторов будут оцениваться только с тремя степенями свободы. Кроме того, эксперимент значительно усложняется с увеличением опытов в каждой точке плана.

Принципиально на полученный греко-латинский квадрат можно наложить следующий ортогональный латинский квадрат. В получившемся квадрате каждая комбинация символов, находящихся в каждой ячейке (клетке) плана, не повторяется ни по строкам, ни по столбцам. Подобный план называют гипер-греко-латинским квадратом.

С увеличением числа факторов, числа уровней факторов и числа опытов в каждой точке плана сложность проведения эксперимента существенно возрастает. Поэтому в теории эксперимента разработан ряд планов, которые позволяют упростить проведение эксперимента без потери принципа рандомизации, например путем разделения изучаемых факторов по некоторым группам и др.

9.3. Статистическая модель изучаемого процесса (явления)

Для познания и управления различными процессами и явлениями объективного мира широко используются их модели, отражающие наиболее важные стороны явлений и их эволюцию. При создании модели исследователь обязан перечислить все те особенности явления, которые в дальнейшем будут приниматься во внимание. Изучаемое явление принуждает исследователя выбирать такие модели, которые позволят вполне адекватно передавать особенности изучаемого явления и его изменения во времени и в пространстве. Конечно, каждая модель изучаемого процесса огрубляет действительность, и задача исследователя состоит в первую очередь в том, чтобы предложить модель, передающую, с одной стороны, наиболее полно фактическую сторону явления, а с другой — дающую значительное приближение к действительности. К процессу моделирования явления предъявляются требования, связанные как с меньшими временными и экономическими затратами по сравнению с исследованиями на реальном явлении (объекте), так и с возможностью перенести результаты исследования модели на реальное явление.

В последние десятилетия в связи с развитием ЭВМ в качестве моделей используются их математические аналоги. Математическая модель явления или процесса является приближенным их описанием с помощью математической символики. Иначе говоря, *математическая модель — это математическая форма записи объективного закона изучаемого явления или процесса*. Разумеется, для одного и того же явления можно предложить несколько математических моделей. Все они имеют право

на существование до тех пор, пока не начнет сказываться существенное расхождение модели и действительности. Анализ математической модели позволяет вникнуть в сущность изучаемого процесса (явления).

При создании математической модели условно можно выделить четыре основных этапа.

Первый этап состоит в формулировании законов (функциональных соотношений), связывающих основные объекты (факторы). На этом этапе от исследователя требуется широкое знание фактов влияния различных факторов и их взаимосвязи на изучаемые явления. Этап, как правило, завершается записью в математических терминах качественных представлений о связях между объектами (факторами) модели.

Второй этап — это исследование конкретных задач на основе выдвинутого варианта математической модели. На данном этапе решается прямая задача в смысле получения в результате анализа модели выходных данных для дальнейшего их сопоставления с результатами наблюдений (опытов) изучаемых процессов (явлений).

Третий этап заключается в сравнении результатов наблюдений (опытов) с теоретическими данными, получаемыми из математической модели, с учетом точности проведенных наблюдений. Если отличие этих результатов выходит за предельные границы, то параметры модели уточняются. Последняя задача относится к классу обратных задач.

Четвертый этап создания математической модели состоит в построении новой, более совершенной математической модели на основе новых накопленных данных об изучаемом процессе (явлении).

В зависимости от априорных данных, имеющихся у исследователя, различают аналитические (теоретические) модели и статистические (имитационные) модели.

При создании *аналитических* (*теоретических*) моделей за основу берутся объективные закономерности моделируемых явлений в виде соответствующих уравнений, связывающих изучаемые явления с различными другими объектами-аргументами (факторами). Построение подобных моделей требует, как правило, проведения больших и трудоемких исследований, так как это обусловлено необходимостью выяснения природы большого числа частных процессов, протекающих внутри изучаемого явления. Так, например, математические модели, связанные с различными физическими явлениями, наиболее часто описываются при помощи дифференциальных уравнений с частными производными, интегро-дифференциальных уравнений, стохастических уравнений и т.п. Теоретические модели допускают экстраполяцию значений изучаемого явления для таких значений факторов-аргументов, которые по каким-либо причинам не могут быть воспроизведены экспериментально.

Статистические (*имитационные*) *модели* основаны на статистической обработке опытных данных, полученных на изучаемом объекте (процессе). При этом структура статистической модели, или, иначе, функциональная зависимость, положенная в основу статистической модели, хотя выбирается достаточно произвольно, но должна учитывать не

только цель исследования, но и априорные данные о поведении отклика в окрестности изменения исследуемых факторов-аргументов, а именно, имеются или отсутствуют экстремумы в заданной или выбранной окрестности изменения факторов-аргументов.

Статистическая модель должна соответствовать изучаемому процессу только количественно. Поэтому одна и та же статистическая модель может описывать различные процессы, так же как и сами процессы могут описываться различными статистическими моделями. В связи с этим статистические модели должны иметь относительно простую структуру. Наиболее часто они представляются с точностью до некоторых коэффициентов в виде полиномиальной зависимости значений изучаемого процесса y от причинных факторов x_i. Зависимость выходного эффекта от уровней факторов является *выборочной средней квадратической регрессией*, рассмотренной в гл. 8. Ниже приведена одна из возможных записей полиномиальной статистической зависимости:

$$y = \varphi(b_0, b_1, \ldots, b_p; x_1, \ldots, x_m) = \sum_{l=0}^{p} b_l \psi_l(x_1, \ldots, x_m), \qquad (9.3.1)$$

где $\varphi_l(x_1, \ldots, x_m)$ — известные функции вектора аргументов $\mathbf{x}(x_1, x_2, \ldots, x_m)$; $\mathbf{b}(b_0, b_1, \ldots, b_p)$ — вектор неизвестных коэффициентов, оценки которых определяются на основе опытных данных. Из (9.3.1) видно, что эта зависимость является линейной функцией неизвестных коэффициентов. Зависимую переменную статистической модели y, представляющую выходной эффект исследуемого процесса, как отмечалось выше, часто называют *функцией отклика* или просто *откликом*, а при геометрической интерпретации выходного эффекта — *поверхностью отклика*. Поэтому эту зависимость обычно называют *теоретической зависимостью* (*теоретическим откликом*).

Полученная модель позволяет вычислить «теоретическое» значение изучаемого процесса y и сравнить его с экспериментальным значением y^*. По результатам сравнения этих данных с учетом заранее выбранного критерия принимается или отвергается гипотеза о соответствии (*адекватности*) выбранной модели реальным опытным данным. Область применения статистических моделей ограничивается ближайшей окрестностью значений факторов, для которых проводятся эксперименты. Во многих случаях построение таких моделей можно выполнить ценой сравнительно небольших затрат времени и средств. Для получения функции отклика в теории планирования эксперимента разработано несколько специальных планов эксперимента, которые позволяют построить модель, по отклику близкую к действительности, при минимально возможном числе опытов.

В более общем виде изучаемый процесс может характеризоваться не одной, а несколькими функциональными зависимостями, подчеркивающими те или иные его свойства (показатели), т.е. вместо функции $y = \varphi(\mathbf{b}, \mathbf{x})$ необходимо найти μ зависимостей $y_h = \varphi_h(\mathbf{b}_h, \mathbf{x})$, $h = 1, \ldots \mu$.

Поскольку статистические модели могут содержать большое число факторов, то выделение из их числа наиболее существенных факторов методами, описанными в предыдущем пункте, является важным вспомогательным средством для эффективного построения статистических моделей, предназначенных для изучения различных процессов и явлений, в том числе и для их оптимизации и управления.

Для вычисления оценок коэффициентов $\mathbf{b}^*(b_0^*, b_1^*, \ldots, b_p^*)$ модели, как правило, применяются методы средней квадратической регрессии, рассмотренные в п. 8.2.2. Как и при регрессионном анализе будем полагать, что в различных точках экспериментального плана ошибки измерения отклика y распределены по нормальным законам с нулевыми средними значениями и одинаковыми дисперсиями, а ошибки при задании уровней факторов $\mathbf{x}(x_1, \ldots, x_m)$ пренебрежимо малы по сравнению с ошибками определения отклика y. Вполне очевидно, что число опытов n, необходимое для получения оценок коэффициентов модели, связано со сложностью выбранной статистической модели (например, со степенью полинома модели) и количеством переменных факторов. Число опытов, по крайней мере, должно быть не меньше числа неизвестных коэффициентов модели. Применительно к аппроксимации в виде (9.3.1) должно выполняться неравенство $n \geqslant (p+1)$. При этом должно существовать некоторое минимальное число точек плана, по опытам в которых можно будет вычислить оценки коэффициентов тех или иных моделей, причем минимальное число уровней каждого из факторов определяется максимальной его степенью. Так, например, если изучаемый отклик линейно зависит от одного фактора, то минимальное число уровней этого фактора равно двум; при квадратичной зависимости отклика от фактора — минимальное число уровней фактора равно трем и т.д., т.е. число уровней фактора должно быть на единицу больше его порядка в изучаемой зависимости. При использовании полных факторных экспериментов, т.е. если план основан на использовании всех возможных сочетаний уровней факторов, то применительно к двум уровням m факторов он обозначается как 2^m, при трех уровнях m факторов — 3^m и т.д. Современная теория планирования эксперимента позволяет определить точки плана, обеспечивающие минимальное число опытов для различного сочетания уровней факторов и их числа.

Пусть в каждой точке плана эксперимента проводится один опыт (осуществляется одно наблюдение), т.е. число точек плана N совпадает с числом опытов. Тогда число степеней свободы отклика (9.3.1) $\nu = N - (p+1)$. Если $\nu > 0$ $(N > p+1)$, т.е. число проведенных опытов больше числа неизвестных коэффициентов, то такие планы называют *ненасыщенными*. Если $\nu = 0$ $(N \leqslant p+1)$, т.е. если число проведенных опытов равно числу неизвестных коэффициентов модели, то подобные планы называют *насыщенными*. Если $\nu < 0$ $(N < p+1)$, т.е. если число проведенных опытов меньше числа неизвестных коэффициентов, такие планы называются *сверхнасыщенными*. Естественно, для сверхнасыщенных планов однозначно определить неизвестные коэффициенты

модели невозможно. В связи с этим в теории планирования разработаны специальные методы (планы), позволяющие из всех факторов выделить наиболее существенные факторы, для которых возможно применение ненасыщенных или, в крайнем случае, насыщенных планов.

Итак, пусть с помощью регрессионного анализа с использованием метода наименьших квадратов получены оценки коэффициентов модели b_l^*, $l = 0, 1, \ldots, p$, которые при сделанных предположениях имеют нормальные распределения. Подставляя в (9.3.1) вместо неизвестных коэффициентов их оценки, получаем выражение для оценки искомой статистической модели:

$$y^* = \sum_{l=0}^{p} b_l^* \psi_l(x_1, \ldots, x_m). \qquad (9.3.2)$$

Может оказаться, что некоторые из коэффициентов ряда (9.3.2) незначимы, т.е. близки к нулю, и ими можно пренебречь, тем самым упростив статистическую модель. Для этого применительно к каждой из оценок коэффициентов необходимо проверить нулевую гипотезу о том, что полученная оценка b_l^*, $l = 0, 1, \ldots, p$, существенно не отличается от нуля для заданного уровня значимости α. Будем полагать, что оценки коэффициентов модели имеют точное или приближенное нормальное распределение.

Пусть оценка коэффициента b_l^* характеризуется выборочной дисперсией $s^2(b_l^*)$, зависящей от измеренных и теоретических значений откликов. Для проверки гипотезы о значимости коэффициента b_l в зависимости (9.3.1) наиболее эффективной статистикой является t-статистика (статистика Стьюдента), определяемая соотношением (5.3.1), которая применительно к оценкам коэффициентов модели записывается в виде $t_{\nu_l} = b_l^*/s(b_l^*)$. Число степеней свободы ν_l данной статистики определяется не только числом неизвестных коэффициентов, но и конкретным планом эксперимента. Отличие коэффициента b_l^* от нуля считается несущественным для заданного уровня значимости α, если выборочная t-статистика по модулю не превышает критического значения $t_{\alpha/2;\nu_l}$, т.е. если

$$-t_{\alpha/2;\nu_l} \leqslant t_{\nu_l}^* = b_l^*/s(b_l^*) \leqslant t_{\alpha/2;\nu_l}, \qquad (9.3.3)$$

то данным коэффициентом ряда (9.3.2) можно пренебречь.

Полученная математическая модель (9.3.2) должна быть проверена на ее *адекватность* действительности, для чего экспериментально полученные данные y_j сравниваются с теоретическими значениями y_j^*, вычисляемыми согласно (9.3.2). Иначе говоря, необходимо в соответствии с заданным критерием проверить, насколько существенно отклонение теоретических значений от опытных данных. В качестве метода сравнения, как правило, используется линейный дисперсионный однофакторный анализ, достаточно подробное описание которого дано в п. 6.1. Применительно к рассматриваемой задаче проверки гипотезы фактором

является «адекватность модели», а уровнями фактора — N точек плана эксперимента i, $i = 1, 2, \ldots, N$. Для применения дисперсионного анализа целесообразно, чтобы в каждой точке плана (на каждом уровне) было не менее двух независимых измерений ($n \geqslant 2$) выходного эффекта (отклика) y, а число точек плана (число строк) хотя бы на единицу было больше числа неизвестных коэффициентов, т.е. чтобы выполнялись оба условия $n \geqslant 2$ и $N \geqslant (p+2)$.

Итак, пусть в каждой из N точек плана (с применением рандомизации) осуществлено одинаковое число опытов, равное n, так что общее число опытов равно Nn. Обозначим через y_{ij} значение измеряемого отклика в i-й точке плана (в i-й строке). Тогда

$$\bar{y}_i \equiv y_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} y_{ij} \qquad (9.3.4)$$

— выборочные средние опытных данных по уровням (строкам). В соответствии с однофакторным дисперсионным анализом аналогично (6.1.12) суммарный квадрат отклонения экспериментально полученных данных y_{ij}, $j = 1, \ldots, n$, от значений теоретической зависимости y_i^* можно представить в виде суммы двух независимых составляющих:

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - y_i^*)^2 = \sum_{i=1}^{N} n(y_{i\cdot} - y_i^*)^2 + \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - y_{i\cdot})^2 = Q_A + Q_R.$$

Здесь

$$Q_A = \sum_{i=1}^{N} n(y_{i\cdot} - y_i^*)^2; \quad Q_R = \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - y_{i\cdot})^2.$$

При адекватности модели случайные величины Q/σ^2, Q_A/σ^2 и Q_R/σ^2, где σ^2 — точное значение дисперсии случайного отклика y_{ij}, распределены по законам хи-квадрат со степенями свободы $\nu = Nn - (p+1)$, $\nu_A = N - (p+1)$ и $\nu_R = N(n-1)$ соответственно. При этом нетрудно убедиться в том, что $\nu = \nu_A + \nu_R$. Рассмотрим случайные величины $\sigma_A^2 = Q_A/\nu_A$ и $s_R^2 = Q_R/\nu_R$:

$$s_A^2 = \frac{1}{N-(p+1)} \sum_{i=1}^{N} n(y_{i\cdot} - y_i^*)^2; \qquad (9.3.5)$$

$$s_R^2 = \frac{1}{N(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{n} (y_{ij} - y_{i\cdot})^2. \qquad (9.3.6)$$

Случайная величина s_A^2 является оценкой суммарного квадрата отклонения выборочных средних опытных данных от теоретических значений в каждой точке плана эксперимента (по каждому уровню). Ее называют *дисперсией адекватности модели*. Случайная величина s_R^2 представляет

собой оценку суммарной дисперсии случайной ошибки измерения отклика y_{ij} и называется *дисперсией ошибки наблюдения*.

Так же, как и в п. 8.1, можно показать, что если предложенная *математическая модель адекватна действительности*, то случайные величины (9.3.5) и (9.3.6) оценивают одну и ту же дисперсию σ^2. Поэтому в качестве критерия проверки адекватности модели можно использовать F-статистику и соответствующий ей критерий принятия решения. Применительно к введенным обозначениям с уровнем значимости α нулевая *гипотеза об адекватности модели не отвергается, если выборочная статистика* $F^*_{\nu_A,\nu_R} = s^2_A/s^2_R$ *не превышает критического значения* $F_{\alpha;\nu_A,\nu_R}$, определяемого по соответствующим таблицам критических точек F-распределения, т.е. если выполняется неравенство

$$F^*_{\nu_A,\nu_R} = (s^2_A/s^2_R) \leqslant F_{\alpha;\nu_A,\nu_R}. \qquad (9.3.7)$$

Для более рационального планирования эксперимента по построению статистической модели натуральные значения факторов x^*_r целесообразно заменить их *нормированными значениями* (*«кодами»*) согласно выражению

$$x_r = \frac{x^*_r - x^*_{r\cdot}}{\Delta x^*_r/2} = \frac{x^*_r - (x^*_{r\max} + x^*_{r\min})/2}{(x^*_{r\max} - x^*_{r\min})/2}. \qquad (9.3.8)$$

Здесь $x^*_{r\max}$, $x^*_{r\min}$ и $x^*_{r\cdot}$ — максимальное, минимальное и среднее значения фактора в эксперименте; x_r — нормированное значение («код») фактора x^*_r; $\Delta x^*_r = x^*_{r\max} - x^*_{r\min}$ — диапазон изменения (варьирования) фактора x^*_r в эксперименте. При нормировании максимальные и минимальные значения всех переменных будут равны соответственно «+1» и «−1», а среднему значению соответствует значение (код) «0». В этом случае организация оптимального плана эксперимента и анализ результатов проведения опытов при введении нормированных переменных существенно упрощаются.

При планировании эксперимента часто возникает задача выбора оптимального значения интервала между уровнями δx^*_r для каждого из факторов в заданном диапазоне Δx^*_r. Для сокращения объема эксперимента естественно желание выбрать значение интервала δx^*_r как можно больше. Однако при этом есть опасность получить неадекватную статистическую модель изучаемого процесса. Слишком малое значение интервала δx^*_r за счет случайных ошибок может привести к тому, что функция отклика изменяется в пределах этого интервала незначительно, в результате чего может быть принято неправильное решение об отсутствии влияния уровней фактора на функцию отклика. Поэтому значение интервала δx^*_r должно выбираться таким, чтобы получающееся приращение функции отклика можно было с необходимой надежностью выделить на фоне случайных ошибок эксперимента.

Планирование эксперимента существенно зависит от задачи, которая ставится при создании математической модели: предназначена модель только для описания изучаемого процесса в окрестностях заданных уровней факторов или модель должна быть использована для отыскания оптимальных уровней факторов (оптимальных условий протекания процесса), обеспечивающих экстремум отклика. В последнем случае при поиске экстремальной точки отклика, в отличие от аналитического исследования, по результатам ряда опытов, специально поставленных около исходной точки, осуществляется локальное изучение поверхности отклика. Путем целенаправленного, многократного и последовательного изменения сочетания уровней факторов осуществляется движение к экстремуму. Обычно это движение осуществляется дискретно («шагами»). Существует ряд экспериментальных методов оптимизации, различающихся между собой способами определения направления движения (способами направления изменения сочетания уровней факторов) и организацией самого движения (организацией изменения уровней факторов). Наиболее распространенными способами поиска экстремальной точки отклика являются методы покоординатного движения (Гаусса–Зейделя), случайного поиска, градиента, крутого восхождения (Бокса–Уилсона), симплексный метод и метод эволюционного планирования. Ниже в основном изучаются планы применительно к созданию адекватной статистической модели для описания исследуемых процессов. Из методов поиска экстремальной точки отклика кратко рассмотрены лишь метод крутого восхождения, симплексный метод и метод эволюционного планирования, являющиеся простейшими из эффективных методов планирования экстремального эксперимента.

Как и в любой другой задаче, при разработке оптимального плана задаются критерием его оптимальности. В теории эксперимента предложено ряд подобных критериев оптимальности, основными из которых являются критерии ортогональности и ротатабельности.

В соответствии с *критерием ортогональности* требуется выбрать план для оценки неизвестных коэффициентов модели b^*_l, при котором они будут независимы, что существенно облегчает дальнейшие вычисления. Такой план эксперимента называется *ортогональным*.

Согласно *критерию ротатабельности* точки плана должны быть расположены таким образом, чтобы дисперсия оценки значений функции отклика в произвольной точке x, соответствующей области планирования, зависела бы только от расстояния этой точки до центра плана. Этот критерий с точки зрения точности оценки функции отклика обеспечивает равнозначность всех направлений от центра плана. Подобный план эксперимента называется *ротатабельным*.

В той или иной мере эти и большинство других критериев подразумевают, чтобы оптимальный план при заданной адекватности модели обеспечивался минимумом объема эксперимента. В частности, оптимальный план подразумевает, что если в результате анализа опытных данных гипотеза об адекватности предложенной модели действительному процессу не подтверждается, то для получения более совершенной

модели необходимо поставить опыты по такому плану, которые лишь бы уточняли коэффициенты модели более высокого порядка, оставляя без изменения коэффициенты предыдущей модели.

Одним из принципов последовательного приближения модели к действительности является так называемый принцип «от простого к сложному», т.е. на первоначальном этапе предлагается модель первого порядка, которая по мере необходимости заменяется на модель более высокого порядка.

9.4. Статистическая модель первого порядка

Простейшей *статистической моделью первого порядка (регрессией)*, зависящей от m независимых факторов (переменных) является *полиномиальная модель* вида

$$y = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=i+1}^{m} b_{ij} x_i x_j + \\ + \sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{k=j+1}^{m} b_{ijk} x_i x_j x_k + \ldots + b_{12\ldots m} x_1 x_2 \ldots x_m. \quad (9.4.1)$$

Здесь b_0 — значение функции отклика y в центре плана; b_i — коэффициенты, характеризующие влияние каждого из факторов; b_{ij} — коэффициенты, учитывающие влияние парного взаимодействия факторов; b_{ijk} — коэффициенты, отражающие влияние тройного взаимодействия факторов и т.д. Модель (9.4.1) называется *моделью первого порядка*, потому что факторы имеют степень не выше первой.

Ниже рассматриваются принципы построения планов эксперимента для получения оценок неизвестных коэффициентов полиномиальной модели первого порядка.

9.4.1. План полного факторного эксперимента

Рассмотрим принципы построения плана *полного факторного эксперимента* применительно к модели, зависящей от двух факторов x_1 и x_2:

$$y = b_0 x_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2. \quad (9.4.2)$$

Здесь $x_0 \equiv 1$ — фиктивная переменная, принимающая постоянное значение, равное единице; b_0 — значение функции отклика y в центре плана; b_1 и b_2 — коэффициенты, характеризующие степень влияния факторов x_1 и x_2 на значение функции отклика; b_{12} — коэффициент, учитывающий эффект влияния взаимодействия обоих факторов на значение функции отклика. Подобная модель первого порядка кроме свободного члена $b_0 x_0$ содержит два линейных члена $b_i x_i$ и один нелинейный член $b_{12} x_1 x_2$.

Таблица 9.4.1

Схема плана эксперимента 2^2

Точка плана i	План эксперимента				Результаты опыта y_i
	x_0	x_1	x_2	$x_3 = x_1 x_2$	
1	+1	−1	−1	+1	y_1
2	+1	−1	+1	−1	y_2
3	+1	+1	−1	−1	y_3
4	+1	+1	+1	+1	y_4

Для определения коэффициентов модели минимальное число уровней каждого фактора равно двум. С целью сокращения дальнейших записей обозначим $b_{12} \equiv b_3$, $x_1 x_2 \equiv x_3$. Для постановки опытов матрица планирования может быть представлена в виде, приведенном в табл. 9.4.1. Здесь схема планирования идентична схеме, представленной в табл. 9.1.1,*б*, в которой вместо третьего объекта взвешивания использована переменная «фактор», соответствующий влиянию взаимодействия уровней первых двух факторов и численно равный произведению уровней этих факторов. Кроме этого добавлен еще один столбец «фиктивной» переменной x_0 для оценки свободного члена b_0. Согласно приведенному плану опыты выполняются следующим образом: в первом опыте переменные-факторы x_1 и x_2 находятся на нижних уровнях; во втором опыте фактор x_1 находится на нижнем уровне, а фактор x_2 — на верхнем уровне и т.д. В этом плане в каждой его точке проводится по одному опыту, т.е. число опытов равно числу неизвестных коэффициентов. Подобные планы, как уже отмечалось выше, называются насыщенными, так как здесь все точки плана $N = 4$ используются для оценки $p + 1 = 4$ коэффициентов рассматриваемой статистической модели первого порядка.

Изучение данных, приведенных в табл. 9.4.1, показывает, что переменные x_0, x_1 и x_2 взаимно ортогональны, так как

$$\sum_{i=1;j\neq q}^{N=4} x_{ij} x_{iq} = 0, \quad j, q = 0, 1, 2, 3. \quad (9.4.3)$$

При этом также справедливы соотношения

$$\sum_{i=1}^{N=4} x_{ij}^2 = N; \qquad \sum_{i=1;j\neq 0}^{N=4} x_{ij} = 0. \quad (9.4.4)$$

Выражение в (9.4.3) представляет собой соответственно *условие ортогональности*. Первое соотношение в (9.4.4) называется *условием нормировки*, а второе — *условием симметричности* плана. Такой план называется *ортогональным планом*. Соотношения (9.4.3) и (9.4.4) справедливы для ортогональных планов любого порядка.

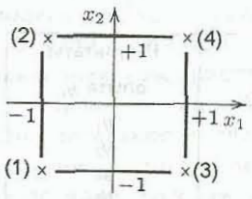


Рис. 9.4.1. Расположение точек плана модели первого порядка для двух факторов

Поскольку выражение (9.4.2) зависит лишь от четырех неизвестных коэффициентов, то для их определения вполне достаточно провести четыре рандомизированных опыта в четырех точках плана (на сочетаниях двух уровней каждого из факторов), как показано на рис. 9.4.1. Точки плана располагаются в вершинах квадрата с центром, совпадающим с центром плана. Каждому из сочетаний уровней факторов (каждой точке плана) соответствует определенное значение функции отклика. Подобный план является полным факторным экспериментом типа 2^2 (ПФЭ типа 2^2). Обычно план начинают с точки, соответствующей наименьшим значениям уровней факторов, т.е. с точки плана с координатами $(x_1 = -1,\ x_2 = -1)$.

Для нахождения коэффициентов модели в соответствии с методом наименьших квадратов, как и выше (п. 3.5.5), рассмотрим суммарный квадрат отклонения теоретического значения отклика от его действительного значения (суммарный квадрат невязок)

$$G = \sum_{i=1}^{N=4} \varepsilon_i^2 = \sum_{i=1}^{N=4} [y_i - (b_0 x_{0i} + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i})]^2.$$

Оценки коэффициентов модели b_l^* находятся в результате решения системы уравнений $[\partial G/\partial b_l] = 0,\ l = 0, 1, \ldots$, которые в форме нормальных уравнений записываются в виде

$$\sum_{i=1}^{N} y_i x_{0i} = b_0^* \sum_{i=1}^{N} x_{0i}^2 + b_1^* \sum_{i=1}^{N} x_{0i}x_{1i} + b_2^* \sum_{i=1}^{N} x_{0i}x_{2i} + b_3^* \sum_{i=1}^{N} x_{0i}x_{3i};$$

$$\sum_{i=1}^{N} y_i x_{1i} = b_0^* \sum_{i=1}^{N} x_{0i}x_{1i} + b_1^* \sum_{i=1}^{N} x_{1i}^2 + b_2^* \sum_{i=1}^{N} x_{1i}x_{2i} + b_3^* \sum_{i=1}^{N} x_{1i}x_{3i};$$

$$\sum_{i=1}^{N} y_i x_{2i} = b_0^* \sum_{i=1}^{N} x_{0i}x_{2i} + b_1^* \sum_{i=1}^{N} x_{1i}x_{2i} + b_2^* \sum_{i=1}^{N} x_{2i}^2 + b_3^* \sum_{i=1}^{N} x_{2i}x_{3i};$$

$$\sum_{i=1}^{N} y_i x_{3i} = b_0^* \sum_{i=1}^{N} x_{0i}x_{3i} + b_1^* \sum_{i=1}^{N} x_{1i}x_{3i} + b_2^* \sum_{i=1}^{N} x_{2i}x_{3i} + b_3^* \sum_{i=1}^{N} x_{3i}^2.$$

Учитывая условие ортогональности (9.4.3), имеем четыре независимых уравнения:

$$\sum_{i=1}^{N} y_i x_{0i} = b_0^* \sum_{i=1}^{N} x_{0i}^2; \quad \sum_{i=1}^{N} y_i x_{1i} = b_1^* \sum_{i=1}^{N} x_{1i}^2;$$

$$\sum_{i=1}^{N} y_i x_{2i} = b_2^* \sum_{i=1}^{N} x_{2i}^2; \quad \sum_{i=1}^{N} y_i x_{3i} = b_3^* \sum_{i=1}^{N} x_{3i}^2.$$

Отсюда получаем оценки коэффициентов модели

$$b_l^* = \sum_{i=1}^{N} y_i x_{li} \Big/ \sum_{i=1}^{N} x_{li}^2, \quad l = 0, 1, 2, 3. \tag{9.4.5}$$

В результате теоретическое значение отклика статистической модели запишется как

$$y^* = b_0^* x_0 + b_1^* x_1 + b_2^* x_2 + b_3^* x_3 \quad (b_3^* \equiv b_{12}^*; \quad x_0 \equiv 1; \quad x_3 \equiv x_1 x_2). \tag{9.4.6}$$

Вычислим статистические характеристики оценок коэффициентов, полагая, что в среднем модель адекватна реальным данным. В этом случае оценка отклика несмещенная, т.е.

$$\mathrm{M}\{y^*\} = y = b_0 x_0 + b_1 x_1 + b_2 x_2 + b_3 x_3.$$

Тогда

$$\mathrm{M}\{b_l^*\} = \frac{\sum_{i=1}^{N} \mathrm{M}\{y_i^*\} x_{li}}{\sum_{i=1}^{N} x_{li}^2} = \frac{b_l \sum_{i=1}^{N} x_{li}^2}{\sum_{i=1}^{N} x_{li}^2} = b_l, \quad l = 0, 1, 2, 3. \tag{9.4.7}$$

т.е. оценки коэффициентов модели также несмещенные. Корреляционный момент оценок коэффициентов модели

$$\mathrm{M}\{(b_l^* - b_l)(b_q^* - b_q)\} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{M}\{(y_i^* - y_i)(y_r^* - y_r)\} x_{li} x_{qr}}{\sum_{i=1}^{N} x_{li}^2 \sum_{r=1}^{N} x_{qr}^2}.$$

Учитывая независимость откликов,

$$\mathrm{M}\{(y_i^* - y_i)(y_r^* - y_r)\} = \begin{cases} 0 & \text{при } i \neq r, \\ \sigma^2(y) & \text{при } i = r, \end{cases}$$

где $\sigma^2(y)$ — дисперсия отклонения отклика относительно его истинного значения, а также условие ортогональности (9.4.3), имеем

$$\mathrm{M}\{(b_l^* - b_l)(b_q^* - b_q)\} = \begin{cases} 0 & \text{при } l \neq q, \\ \sigma^2(y) \Big/ \sum_{i=1}^{N} x_{li}^2 & \text{при } l = q. \end{cases}$$

Таким образом, оценки коэффициентов модели независимы, а их дисперсии определяются формулами

$$D(b_l) = \sigma^2(y) \Big/ \sum_{i=1}^{N} x_{li}^2, \quad l = 0, 1, 2, 3. \tag{9.4.8}$$

Применительно к насыщенному ортогональному плану справедливо условие нормировки (9.4.4), в результате чего нетрудно сделать вывод, что оценки коэффициентов не только независимы, но равны и их дисперсии. Это замечательное свойство ортогональных планов позволяет не проводить новых опытов для построения модели, если какие-либо коэффициенты модели оказываются незначимыми и приравненными нулю.

Если в статистической модели первого порядка отсутствуют нелинейные члены, обусловленные произведением факторов, то такой план становится *ортогональным линейным планом* вида

$$y = b_0x_0 + b_1x_1 + b_2x_2 + \ldots + b_mx_m. \qquad (9.4.9)$$

Этот план обладает свойством *ротатабельности*, так как в силу равенства дисперсий оценок коэффициентов модели и их независимости дисперсия оценки отклика (9.4.9)

$$D(y^*) = \frac{\sigma^2(y)}{N}\left(1 + \sum_{i=1}^{m} x_i^2\right) \equiv \frac{\sigma^2(y)}{N}(1+\rho^2), \qquad (9.4.10)$$

где ρ — радиус гиперсферы. Из (9.4.10) видно, что дисперсия оценки отклика на поверхности гиперсферы радиуса ρ не зависит от направления от центра плана и определяется только расстоянием прогнозируемой точки отклика до центра плана.

Полагая ошибки измерения опытных данных распределенными по нормальному закону, оценки коэффициентов модели также будут распределены по нормальному закону со средним значением b_l и дисперсией (9.4.8). Для проверки гипотезы о том, что коэффициенты модели отличны от нуля, можно было бы использовать статистику $z = b_l^*/\sqrt{D(b_l)}$, которая также распределена по нормальному закону. Поскольку истинное значение дисперсии $\sigma^2(y)$ неизвестно, вместо нее берется условное выборочное значение $s_A^2(y)$, определяемое (9.3.5), а в качестве выборочной статистики для проверки нулевой гипотезы о значимости коэффициентов модели принимается t-статистика (статистика Стьюдента) $t_l^* = b_l^*/s(b_l)$. Однако полученных четырех опытных данных y_i недостаточно для вычисления выборочной статистики даже при условии, что предложенная модель адекватна действительности, так как все четыре точки плана (все четыре y_i) использованы для оценки неизвестных коэффициентов модели. Этих данных тем более недостаточно для проверки адекватности принятой модели опытным данным.

В связи с изложенным выше, необходимо увеличить как число точек плана N (число уровней сочетаний факторов — строк плана), так и число опытов в каждой точке плана (число опытов в каждой строке плана). При этом дополнительные точки плана должны выбираться таким образом, чтобы сохранялась ортогональность плана эксперимента. Одним из способов улучшения плана, представленного в табл. 9.4.1, является план, в котором в качестве дополнительной точки выбран его центр.

Значение функции в центре плана можно использовать для качественной проверки адекватности модели. Например, если в центральной точке плана опытные данные (особенно их средние значения) значительно больше или меньше опытных данных в четырех окружающих ее точках, то это указывает на необходимость применения модели более высокого порядка. К подобному выводу можно придти и при анализе модуля разности между оценкой коэффициента b_0^* и опытными данными в центре плана. Если эта разность по модулю превышает некоторый порог на основе выбранного критерия и заданного уровня значимости, то линейную модель необходимо заменить моделью более высокого порядка.

Полагая число опытов в каждой точке плана равным n, выборочное среднее и выборочная дисперсия по каждой точке плана определяются известными формулами

$$y_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}; \quad s^2(y_i) = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - y_{i\cdot})^2. \qquad (9.4.11)$$

При этом среднее значение дисперсии ошибки измерения отклика по всем точкам плана s_R^2 определяется выражением (9.3.6).

При увеличении числа опытов в каждой точке плана, по сравнению с приведенным в табл. 9.4.1, в формулах для оценок коэффициентов модели (9.4.5) выборочные значения y_i должны быть заменены выборочными средними $y_{i\cdot}$, т.е. оценки коэффициентов модели должны определяться по формуле

$$b_l^* = \sum_{i=1}^{N} y_{i\cdot}x_{li} \Big/ \sum_{i=1}^{N} x_{li}^2, \quad l = 0,1,2,3. \qquad (9.4.12)$$

При этом дисперсии оценок коэффициентов модели (9.4.8) заменяются их выборочными значениями

$$s^2(b_l) = s_A^2(y_\cdot) \Big/ \sum_{i=1}^{N} x_{li}^2, \quad l = 0,1,2,3, \qquad (9.4.13)$$

где дисперсия $s_A^2(y)$ задается формулой (9.3.5), являющейся выборочной дисперсией отклонения выборочных средних от прогнозируемых значений, усредненной по всем точкам (по всем строкам) плана.

Задаваясь уровнем значимости α, определяются двусторонние критические значения выборочной статистики $t_{\text{кр}} = \pm t_{\alpha/2;\nu}$ для числа степеней свободы $\nu = N(n-1)$. Если модуль выборочной статистики удовлетворяет неравенству

$$|t_l^* = b_l^*/s(b_l)| \leqslant t_{\text{кр}} = t_{\alpha/2;\nu}, \qquad (9.4.14)$$

принимается решение, что коэффициент b_l несущественно отличается от нуля и им можно пренебречь в статистической модели (9.3.1).

Для проверки адекватности модели согласно (9.3.7) вычисляются

Таблица 9.4.2

Схема плана эксперимента 2^2

Точка плана i	План эксперимента x_0	x_1	x_2	$x_3 = x_1x_2$	Результаты опыта y_{i1}	y_{i2}	Выборочные среднее $y_{i.}$	дисперсия $s^2(y_i)$
1	+1	−1	−1	+1	y_{11}	y_{12}	$y_{1.}$	$s^2(y_1)$
2	+1	−1	+1	−1	y_{21}	y_{22}	$y_{2.}$	$s^2(y_2)$
3	+1	+1	−1	−1	y_{31}	y_{32}	$y_{3.}$	$s^2(y_3)$
4	+1	+1	+1	+1	y_{41}	y_{42}	$y_{4.}$	$s^2(y_4)$
5	+1	0	0	0	y_{51}	y_{52}	$y_{5.}$	$s^2(y_5)$

Таблица 9.4.3

Точка плана i	Результаты опыта y_{i1}	y_{i2}	$y_{i.}$	$s^2(y_i) = (y_{i1}-y_{i.})^2 + (y_{i2}-y_{i.})^2$	y_i^*	$(y_{i.}-y_i^*)^2$
1	30,5	31,5	31,0	0,5	30,9	0,01
2	39,0	37,0	38,0	2,0	37,9	0,01
3	32,5	31,5	32,0	0,5	31,9	0,01
4	43,0	47,0	45,0	8,0	44,9	0,01
5	36,4	36,0	36,2	0,1	36,4	0,04

по (9.3.5) и (9.3.6) дисперсии адекватности $s_A^2(y)$ и ошибок измерения $s_R^2(y)$, а затем находится выборочная статистика $F^*_{\nu_A,\nu_R} = s_A^2/s_R^2$, которая сравнивается с критическим значением $F_{\alpha;\nu_A,\nu_R}$. Если выполняется неравенство (9.3.7), гипотеза об адекватности модели не отвергается.

Применительно к двухфакторной модели (9.4.1) план эксперимента с минимальным числом точек плана представлен в табл. 9.4.2. Там же введены столбцы, в которых записываются опытные данные, а также вычисленные данные выборочных средних и выборочных дисперсий в точках плана (по строкам). Нетрудно убедиться, что план, приведенный в этой таблице, является ортогональным.

Пример 9.4.1. Пусть выход некоторого технологического процесса y зависит от двух факторов: продолжительности процесса t и его температуры T. Необходимо проверить гипотезу о том, что в интервалах t от 10 до 30 минут и T от 200 до 300 °C выходной эффект определяется зависимостью

$$y = a_0 + a_1 t + a_2 T + a_{12} tT. \qquad (9.4.15)$$

Введем нормированные переменные x_1 и x_2 согласно (9.3.8)

$$x_1 = \frac{t-(t_{\max}+t_{\min})/2}{(t_{\max}-t_{\min})/2} = \frac{t-20}{10}; \quad x_2 = \frac{T-(T_{\max}+T_{\min})/2}{(T_{\max}-T_{\min})/2} = \frac{T-250}{50}, \qquad (9.4.16)$$

а проверяемую зависимость (9.4.15) приведем к стандартному виду (9.4.2).

Для оценивания коэффициентов b, проверки гипотез значимости полученных значений коэффициентов в модели (9.4.2) и адекватности самой модели проведено минимально возможное число опытов согласно плану, приведенному в табл. 9.4.2. Опытные данные представлены во втором и третьем столбцах табл. 9.4.3. В четвертом и пятом столбцах приведены выборочные средние $y_{i.}$ и выборочные дисперсии $s^2(y_i)$ опытных данных в точках плана (по строкам). Из сравнения выборочного среднего в центре плана с выборочными средними в остальных точках плана нельзя утверждать с достаточной достоверностью, что модель первого порядка адекватна

или неадекватна опытным данным. При этом среднее значение выборочной дисперсии $s_R^2(y)$ согласно (9.3.6)

$$s_R^2(y) = \frac{1}{5}\sum_{i=1}^{5} s^2(y_i) = \frac{11{,}1}{5} = 2{,}22.$$

Распределение случайной величины $s_R^2(y)$ имеет $\nu_R = Nn - N = 5$ степеней свободы.

На основе выборочных средних значений в соответствии с формулами (9.4.12) вычислены оценки коэффициентов модели:

$$b_0^* = (31+38+32+45+36{,}2)/5 \approx 36{,}4; \quad b_1^* = (-31-38+32+45)/4 = 2{,}0;$$
$$b_2^* = (-31+38-32+45)/4 = 5{,}0; \quad b_{12}^* = (31-38-32+45)/4 = 1{,}5.$$

Здесь и далее учтено, что оценка коэффициента модели, соответствующая центру плана, определяется по пяти точкам плана, а остальные оценки коэффициентов модели — по четырем точкам плана.

В шестом и седьмом столбцах табл. 9.4.3 приведены вычисленные теоретические значения отклика y_i^* и квадрат его отклонения от выборочного среднего для каждой из указанных точек плана. При этом дисперсия адекватности согласно (9.3.5)

$$s_A^2(y) = \frac{1}{N-(p+1)}\sum_{i=1}^{N} n(y_{i.}-y_i^*)^2 = \frac{2\cdot 0{,}08}{5-4} = 0{,}16.$$

Распределение этой случайной величины имеет одну степень свободы ($\nu_A = 1$). При этом дисперсии выборочных коэффициентов модели согласно (9.4.13)

$$s^2(b_0) = s_A^2(y)/5 = 0{,}032; \quad s^2(b_1) = s^2(b_2) = s^2(b_{12}) = s^2(y_.)/4 = 0{,}04.$$

Проверим значимость полученных оценок коэффициентов для уровня значимости $\alpha = 0{,}1$, для чего образуем t-статистики (статистики Стьюдента), имеющие $\nu = 1$ степень свободы:

$$t_0^* = \frac{b_0^*}{s(b_0)} = \frac{36{,}4}{0{,}18} = 202; \quad t_1^* = \frac{b_1^*}{s(b_1)} = \frac{2{,}0}{0{,}2} = 10;$$
$$t_2^* = \frac{b_2^*}{s(b_2)} = \frac{5{,}0}{0{,}2} = 25; \quad t_{12}^* = \frac{b_{12}^*}{s(b_{12})} = \frac{1{,}5}{0{,}2} = 7{,}5. \qquad (9.4.17)$$

Для $\alpha = 0{,}1$ и $\nu = 1$ согласно табл. 3 Приложения 2 критическое значение выборочной t-статистики $t_{\text{кр}} = t_{0{,}05;1} = 6{,}13$. Сравнивая выборочные статистики с ее критическим значением, делаем вывод о значимости полученных оценок коэффициентов модели. Поэтому искомая статистическая модель для нормированных переменных записывается в виде

$$y^* = 36{,}4 + 2{,}0x_1 + 5{,}0x_2 + 1{,}5x_1x_2. \qquad (9.4.18)$$

Заменяя в последнем выражении нормированные переменные x_1 и x_2 их ненормированными аналогами t и T, получаем искомое уравнение регрессии для натуральных переменных

$$y^* = 36{,}4 + 2{,}0\left(\frac{t-20}{10}\right) + 5{,}0\left(\frac{T-250}{50}\right) + 1{,}5\left(\frac{t-20}{10}\right)\left(\frac{T-250}{50}\right) =$$
$$= 26{,}40 - 0{,}55t + 0{,}04T + 0{,}003tT. \qquad (9.4.19)$$

Проверим гипотезу об адекватности полученной модели опытным данным согласно F-критерию для уровня значимости $\alpha = 0{,}05$. Применительно к экспериментальным данным (числу степеней свободы $\nu_A = 1$, $\nu_R = 5$) критическое значение статистики $F_{\text{кр}} = F_{0{,}05;1;5} = 6{,}61$. Сравнивая это критическое значение с

Таблица 9.4.4

Схема плана эксперимента 2^3

Точки плана i	x_0	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	+1	−1	−1	−1	+1	+1	+1	−1
2	+1	−1	+1	−1	−1	+1	−1	+1
3	+1	+1	−1	−1	−1	−1	+1	+1
4	+1	+1	+1	−1	+1	−1	−1	−1
5	+1	−1	−1	+1	+1	−1	−1	+1
6	+1	−1	+1	+1	−1	−1	+1	−1
7	+1	+1	−1	+1	−1	+1	−1	−1
8	+1	+1	+1	+1	+1	+1	+1	+1

выборочной F-статисткой

$$F^*_{1,5} = \frac{s^2_A(y)}{s^2_R(y)} = \frac{0{,}16}{2{,}22} \approx 0{,}072, \qquad (9.4.20)$$

делается вывод, что с уровнем значимости $\alpha = 0{,}05$ статистическая модель (9.4.19) не противоречит опытным данным. Итак, уравнение регрессии (9.4.20) позволяет в интервале значений $t = (10...30)$ минут и $T = (200..300)\,^\circ$C, прогнозировать величину выходного продукта технологического процесса.

Пусть, например, при постоянстве откликов в крайних точках плана гипотетически выборочное среднее значение в центре плана $y_5 = 48$, что соответствует оценке коэффициента статистической модели в центре плана, равной $b^*_0 = 38{,}8$. Эти данные свидетельствуют о том, что в области центра плана может находиться максимум отклика. Полагая ошибки измерения отклика в этой точке плана такими же, как и ранее, и вычисляя аналогично описанной выше процедуре теоретические значения отклика, находим новое значение дисперсии адекватности модели $s^2_A(y) = 169{,}4$. При этом выборочное значение F-статистики также изменится: $F^*_{1,5} = 72{,}3$. Так как $F^*_{1,5} = 72{,}3 > F_{\text{кр}} = F_{0,05;1,5} = 6{,}61$, можно утверждать, что в этом случае полученная модель не адекватна действительности с уровнем значимости $\alpha = 0{,}05$. Поэтому требуется выбрать модель более высокого порядка, например статистическую модель второго порядка.

С увеличением числа факторов модель для функции отклика становится более сложной, хотя для полного факторного эксперимента задача построения полиномиальной модели первого порядка решается аналогично рассмотренному выше. Например, пусть отклик зависит от трех факторов x_1, x_2 и x_3:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3. \qquad (9.4.21)$$

В этой модели имеются восемь членов с восемью неизвестными коэффициентами: один свободный член b_0, три линейных члена типа b_ix_i, три члена с парными взаимодействиями типа $b_{ij}x_ix_j$ и один член с тройным взаимодействием $b_{123}x_1x_2x_3$. При варьировании каждым из трех факторов ($m = 3$) на двух уровнях минимальное число точек плана должно составлять $2^3 = 8$. Подобный план принято называть полным факторным экспериментом типа 2^3 (ПФЭ типа 2^3). Соответствующая ортогональная матрица точек плана может быть представлена в виде, приведенном в табл. 9.4.4.

Ядром этой матрицы является полный факторный эксперимент типа 2^2, взятый для двух переменных и повторенный дважды: один раз для значений $x_3 = -1$, другой раз для значений $x_3 = +1$. Элементы других столбцов получаются путем соответствующего перемножения первых трех столбцов матрицы планирования. Аналогичным образом могут быть получены матрицы планов полных факторных экспериментов типа 2^m, где m — число факторов. При геометрической интерпретации точки плана ПФЭ типа 2^3 расположены в вершинах куба. Число опытов полных факторных экспериментов $N = 2^m$ быстро растет с увеличением числа независимых факторов m (размерности факторного пространства m), в результате чего при больших m планы полных факторных экспериментов оказываются практически неприемлемыми.

В теории эксперимента разработан ряд методов, позволяющих сократить число точек плана, не нарушая свойства ортогональности. Эти методы основаны на соответствующем отборе части точек плана из общего числа 2^m точек плана ПФЭ. Одним из них является план под названием дробный факторный эксперимент (ДФЭ).

9.4.2. План дробного факторного эксперимента

План дробного факторного эксперимента (*ДФЭ*) основан на априорном предположении или знании того, что некоторые из взаимодействий факторов существенно не влияют на функцию отклика.

Рассмотрим принцип построения ДФЭ применительно к полиномиальной модели *первого порядка* для трех независимых факторов (9.4.21). Если взаимодействия уровней факторов отсутствуют, полиномиальная модель первого порядка становится *линейной моделью*:

$$y^* = b_0 + b_1x_1 + b_2x_2 + b_3x_3, \qquad (9.4.22)$$

зависящей только от четырех неизвестных коэффициентов, для определения которых достаточно провести четыре опыта вместо восьми, предусмотренных полным факторным экспериментом. Для того чтобы новый план был ортогональным, за его основу можно взять план ПФЭ типа 2^2, оставив без изменения уровни первых двух факторов x_1 и x_2, а в качестве уровней третьего фактора принять значение, равное произведению $x_3 = x_1x_2$. Этот план аналогичен плану ПФЭ типа 2^2, приведенному в табл. 9.4.1, в котором фактор x_3 принимает значения, соответствующие взаимодействию x_1x_2. Подобный сокращенный план содержит половину опытов от требуемого их числа 2^3; он называется полурепликой (1/2 репликой) от ПФЭ типа 2^3 и обозначается ДФЭ типа 2^{3-1}. Цифра 1 обозначает, что в дробном плане один из факторов заменяет одно взаимодействие. Применительно к рассматриваемой модели (9.4.22) дробный факторный план ПФЭ типа 2^{3-1} позволяет оценить свободный член b_0 и коэффициенты при линейных членах b_1, b_2 и b_3, полагая при этом, что остальные коэффициенты модели b_{12}, b_{23} и b_{123} равны или близки к нулю.

Таблица 9.4.5

Схема плана ДФЭ типа 2^{7-4} для семи факторов

Точки плана i	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	+1	−1	−1	−1	+1	+1	+1	−1
2	+1	−1	+1	−1	−1	+1	−1	+1
3	+1	+1	−1	−1	−1	−1	+1	+1
4	+1	+1	+1	−1	+1	−1	−1	−1
5	+1	−1	−1	+1	+1	−1	−1	+1
6	+1	−1	+1	+1	−1	−1	+1	−1
7	+1	+1	−1	+1	−1	+1	−1	−1
8	+1	+1	+1	+1	+1	+1	+1	+1

В общем случае для построения плана дробного факторного эксперимента типа 2^{m-p} (ДФЭ типа 2^{m-p}) из множества факторов m отбирают $m-p$ основных факторов или их взаимодействий и для них строят план полного факторного эксперимента. При этом p взаимодействий (p столбцов ПФЭ типа 2^m) заменяют p значимыми факторами.

При построении дробных факторных планов для числа факторов $m = 4, 5, 6, 7$ можно ограничиться восемью опытами, для чего в основу его построения берется полный факторный эксперимент типа 2^3 для факторов x_1, x_2 и x_3. Этот план дополняется столбцами, образованными поэлементным произведением столбцов плана 2^3: x_1x_2, x_1x_3, x_2x_3, $x_1x_2x_3$. Например, используя одно из четырех написанных выше выражений, можно построить четыре различных плана ДФЭ типа 2^{4-1} для образования уровней четвертого фактора x_4:

$$x_4 = x_1x_2; \quad x_4 = x_1x_3; \quad x_4 = x_2x_3; \quad x_4 = x_1x_2x_3. \tag{9.4.23}$$

По сравнению с $2^4 = 16$ точками плана ПФЭ дробный план типа 2^{4-1} состоит всего из $2^3 = 8$ точек.

Приведенные выше выражения и аналогичные выражения, применяемые при построении планов ДФЭ для большего числа факторов, называются *генераторами дробных планов*. Чем больше число факторов, тем больше будет вариантов ДФЭ и тем эффективнее становится применение ДФЭ по сравнению с ПФЭ. Так при семи независимых факторах ($m = 7$) для построения одного из вариантов плана ДФЭ типа 2^{7-4} в качестве генераторов можно использовать выражения

$$x_4 = x_1x_2; \quad x_5 = x_1x_3; \quad x_6 = x_2x_3; \quad x_7 = x_1x_2x_3. \tag{9.4.24}$$

В табл. 9.4.5 приведен вариант плана ДФЭ типа 2^{7-4} для семи независимых факторов, основой которого является ПФЭ типа 2^3, приведенный в табл. 9.4.4. По сравнению с $N = 2^7 = 128$ точками плана ПФЭ типа 2^7 полученный план ДФЭ типа 2^{7-4} состоит из $2^3 = 8$ точек. Данный дробный план является 1/16 репликой ПФЭ типа 2^7.

При использовании дробных планов в число значимых факторов могут быть включены и некоторые значимые парные взаимодействия и взаимодействия высших порядков, рассматриваемые как самостоятельные факторы.

При наличии в модели взаимодействий факторов оценки коэффициентов модели b_i^* при линейных членах остаются независимыми друг от друга. Однако эти оценки могут быть смешаны с взаимодействиями факторов. Оценки подобного рода называют смешанными. В литературе по прикладной статистике для построения планов ДФЭ разработаны правила смешивания оценок коэффициентов, основанные на понятии контраста плана. Под *контрастом плана* понимается соотношение, получающееся в результате умножения левой и правой частей выражений для генераторов плана на их левую часть, т.е. чтобы в результате этого умножения получилась матрица первого столбца, которая всегда равна единице и которую обозначают как $\mathbf{I}$. Применительно к дробному плану 2^{3-1} и генератору плана $x_3 = x_1x_2$, контраст

$$\mathbf{I} = x_3x_3 = x_3^2 = x_1x_2x_3. \tag{9.4.25}$$

Аналогично для дробного плана 2^{4-1}, если в качестве генератора плана используется выражение $x_4 = x_1x_2x_3$, контраст определяется как

$$\mathbf{I} = x_1x_2x_3x_4. \tag{9.4.26}$$

Для определения правила смешивания оценок коэффициентов модели для данного фактора необходимо умножить обе части контраста на этот фактор, считая, что $x_i^2 = 1$. В результате получается порядок *смешивания оценок коэффициентов модели* при применении данного плана. Так для плана ДФЭ типа 2^{3-1} с контрастом $\mathbf{I} = x_1x_2x_3$ получаем порядок смешивания для всех трех факторов x_1, x_2 и x_3:

$$x_1 = x_1^2x_2x_3 = x_2x_3; \quad x_2 = x_1x_2^2x_3 = x_1x_3;$$
$$x_3 = x_1x_2x_3^2 = x_1x_2. \tag{9.4.27}$$

Если применительно к модели (9.4.22) предположение об отсутствии взаимодействия неверно, то оценки коэффициентов линейной модели b_1, b_2, b_3 фактически будут являться совместными оценками нескольких коэффициентов модели, так как эти оценки будут обусловлены не только факторами x_1, x_2 и x_3, но их взаимодействиями x_1x_2, x_1x_3, x_2x_3 и $x_1x_2x_3$. Поэтому, если эффект взаимодействия факторов объективно имеется, полученные средние значения оценок коэффициентов модели b_1', b_2' и b_3' связаны с действительными коэффициентами выражениями

$$b_1' = b_1 + b_{23}; \quad b_2' = b_2 + b_{13}; \quad b_3' = b_3 + b_{12}. \tag{9.4.28}$$

Если в результате оценки адекватности модели или каких-либо других обстоятельств появилось сомнение в том, что взаимодействия факторов существенны, необходимо провести дополнительный эксперимент с четырьмя опытами в соответствии с матрицей плана ДФЭ типа 2^{3-1}, т.е. провести соответствующим образом спланированные четыре опыта для второй полуреплики от ПФЭ типа 2^3. Для этого необходимо приравнять $x_3 = -x_1x_2$. В результате взаимозависимость и других факторов

изменится аналогичным образом:

$$x_1 = -x_2x_3; \quad x_2 = -x_1x_3.$$

При этом средние значения совместных оценок коэффициентов модели b_1'', b_2'' и b_3'' будут связаны с действительными коэффициентами выражениями

$$b_1'' = b_1 - b_{23}; \quad b_2'' = b_2 - b_{13}; \quad b_3'' = b_3 - b_{12}. \tag{9.4.29}$$

Решая систему уравнений, составленную из (9.4.28) и (9.4.29), находятся оценки следующих шести коэффициентов модели:

$$b_1 = \frac{b_1' + b_1''}{2}; \quad b_2 = \frac{b_2' + b_2''}{2}; \quad b_3 = \frac{b_3' + b_3''}{2};$$
$$b_{23} = \frac{b_1' - b_1''}{2}; \quad b_{13} = \frac{b_2' - b_2''}{2}; \quad b_{12} = \frac{b_3' - b_3''}{2}. \tag{9.4.30}$$

Таким образом, общее число точек дробных планов, необходимых для раздельной оценки всех коэффициентов модели, должно быть не менее общего числа коэффициентов в предполагаемой статистической модели.

В зависимости от выбора генераторов получаются планы ДФЭ с различной разрешающей способностью, определяемой порядком контраста (числом факторов в контрасте). Планы ДФЭ с наибольшей разрешающей способностью называются *главными планами* ДФЭ, и этим планам следует отдавать предпочтение. Так, например, для плана ДФЭ типа 2^{4-1} предпочтение должно быть отдано генератору $x_4 = x_1x_2x_3$, имеющему контраст (9.4.26), для которого

$$x_1 = x_2x_3x_4; \quad b_1' = b_1 + b_{234};$$
$$x_2 = x_1x_3x_4; \quad b_2' = b_2 + b_{134};$$
$$x_3 = x_1x_2x_4; \quad b_3' = b_3 + b_{124};$$
$$x_4 = x_1x_2x_3; \quad b_4' = b_4 + b_{123}. \tag{9.4.31}$$

В приведенном плане отсутствуют смешения линейных факторов с парными взаимодействиями, которые являются наиболее существенными из всех взаимодействий. В практических задачах, как правило, взаимодействия более высокого порядка, чем парные взаимодействия, встречаются достаточно редко, и ими часто можно пренебречь. Поэтому главные планы для большого числа факторов ($m > 3$) обеспечивают большую разрешающую способность для раздельной оценки коэффициентов модели.

9.5. Планы для квадратичных моделей

Если рассмотренная выше полиномиальная модель первого порядка оказывается не адекватной действительности, переходят к модели более

высокого порядка. Простейшей из них является *квадратичная модель*, которая описывает поверхность отклика, имеющую в окрестности исследуемых значений причинных факторов не более одного экстремума. Применительно к двум независимым факторам эта модель принимает вид

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_2^2 + b_{12}x_1x_2. \tag{9.5.1}$$

В данной модели шесть неизвестных коэффициентов. Для оценки коэффициентов этой модели, содержащей независимые переменные типа x_i^2, каждая переменная в опыте должна принимать по крайней мере три различных значения. Принципиально для получения оценок коэффициентов модели (9.5.1) можно использовать план ПФЭ типа 3^2, т.е. необходимо провести опыты как минимум в девяти точках плана. По мере увеличения числа факторов количество точек плана для квадратичной модели существенно увеличивается до 3^m; так что при $m \geqslant 3$ применение плана ПФЭ становится практически нерациональным. Так как планы ПФЭ для квадратичной модели и модели более высокого порядка не являются ортогональными, они имеют существенный недостаток, связанный с тем, что получаемые оценки коэффициентов модели зависимы между собой. Это приводит к тому, что если в процессе оценок коэффициентов модели некоторые из них близки к нулю и их целесообразно исключить из модели или необходимо повысить порядок модели, то для получения новых оценок эксперимент необходимо провести снова в полном объеме.

Сократить число экспериментов можно, используя так называемые *центральные композиционные планы* (ЦКП). Среди большого разнообразия ЦКП наибольшее применение получили планы, основой которых («ядром») являются ортогональные планы ПФЭ типа 2^m или ДФЭ типа 2^{m-p} для линейных моделей, дополненные точками плана в его центре и *«звездными» точками* по осям координат факторов. Применительно к плану ПФЭ с m факторами звездные точки для независимых переменных задаются координатами:

$$x_1 \to (\pm\alpha, 0, 0, \ldots, 0);$$
$$x_2 \to (0, \pm\alpha, 0, \ldots, 0);$$
$$\ldots\ldots\ldots\ldots\ldots\ldots \tag{9.5.2}$$
$$x_m \to (0, 0, \ldots, 0, \pm\alpha).$$

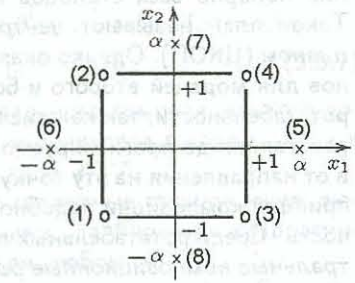


Рис. 9.5.1. Точки центрального композиционного плана для двух факторов

На рис. 9.5.1 показаны точки композиционного плана для двух факторов. Здесь в качестве ядра использованы точки ортогонального плана полного факторного эксперимента типа 2^2, обозначенные кружками. Крестиками обозначены дополнительные (звездные) точки плана, расположенные на осях координат на расстоянии α от центра плана.

Таблица 9.5.1

Матрица ЦКП второго порядка для $m = 2$

	Точка плана i	Факторы и их сочетания					
		x_0	x_1	x_2	x_1^2	x_2^2	x_1x_2
Ядро плана	1	+1	−1	−1	+1	+1	+1
	2	+1	−1	+1	+1	+1	−1
	3	+1	+1	−1	+1	+1	−1
	4	+1	+1	+1	+1	+1	+1
Звездные точки	5	+1	$+\alpha$	0	α^2	0	0
	6	+1	$-\alpha$	0	α^2	0	0
	7	+1	0	$+\alpha$	0	α^2	0
	8	+1	0	$-\alpha$	0	α^2	0
Центр плана	9	0	0	0	0	0	0

В табл. 9.5.1 приведена матрица планирования ЦКП для модели (9.5.1). Из приведенной матрицы плана следует, что условие ортогональности матрицы выполняется только для линейных членов полинома второго порядка. Поэтому независимыми оценками коэффициентов модели будут только коэффициенты при линейных членах.

Для произвольного числа факторов все звездные точки лежат на поверхности гиперсферы радиуса α. При этом общее число точек ЦКП

$$N = 2^{m-p} + 2m + m_0, \tag{9.5.3}$$

где m_0 — число опытов в центре плана. Величина «плеча» α зависит от выбранного критерия оптимальности построения плана ЦКП.

Первоначально Бокс и Уилсон предложили выбирать расстояние α от центра плана до звездных точек таким образом, чтобы получающийся план второго порядка был ортогональным, т.е. чтобы суммы произведений попарно всех столбцов независимых факторов были равны нулю. Такой план называют *центральным композиционным ортогональным планом* (ЦКОП). Однако оказалось, что ортогональность подобных планов для моделей второго и большего порядка не обеспечивает свойства ротатабельности, так как дисперсия оценки отклика зависит не только от расстояния до прогнозируемой точки от центра плана, но также зависит и от направления на эту точку. В связи с этим Бокс и Хантер разработали принцип композиции подобного плана, обеспечивающего его ротатабельность. Среди ротатабельных планов наибольшее применение нашли *центральные композиционные ротатабельные планы* (ЦКРП). Для критериев ротатабельности разработаны правила нахождения величины плеча α и правила составления матрицы соответствующего плана [7, 49, 50, 65].

9.5.1. Центральный композиционный ортогональный план

Применительно к ЦКОП Бокс и Уилсон предложили выбирать величину α таким образом, чтобы значения точек плана для нелинейных

членов модели обеспечивали ортогональность матрицы, т.е. чтобы сумма произведений уровней двух любых столбцов плана была равна нулю. Число точек в центре плана обычно берется равным единице: $m_0 = 1$. В этом случае общее число точек ЦКОП второго порядка, ядром которого являются ортогональные планы ДФЭ типа 2^{m-p},

$$N = 2^{m-p} + 2m + 1. \tag{9.5.4}$$

Из этого соотношения видно, что общее число точек ЦКОП существенно меньше числа точек для плана ПФЭ, причем эта разница тем больше, чем больше число изучаемых факторов. Так, например, при $p = 0$ и $m = 4$ план ПФЭ должен иметь $3^4 = 81$ точку, в то время как ЦКОП — 21 точку.

Для выполнения условий ортогональности в исходной модели второго порядка столбцы с квадратичными членами x_i^2 заменяются на столбцы с элементами $(x_i^2 - \beta)$, где в качестве β предложено брать величину, определяемую соотношением

$$\beta = \frac{1}{N}\sum_{i=1}^{N}(x_k^2)_i = \frac{2^{m-p} + 2\alpha^2}{N}, \tag{9.5.5}$$

где $(x_k^2)_i \equiv x_{ki}^2$ — квадратичный член k-го фактора, расположенный на i-й строке матрицы плана. При этом, естественно, нулевой член исходной модели «перекодируется» в величину

$$c_0 = b_0 + \beta\sum_{k=1}^{m} b_{kk}. \tag{9.5.6}$$

В результате «преобразованную» квадратичную модель в общем виде можно представить как

$$y = c_0 + b_1x_1 + \ldots + b_mx_m + b_{11}(x_1^2 - \beta) + \ldots + b_{mm}(x_m^2 - \beta) + \\ + b_{12}x_1x_2 + \ldots + b_{(m-1)m}x_{m-1}x_m. \tag{9.5.7}$$

В табл. 9.5.2 приведена матрица центрального композиционного ортогонального плана второго порядка для m факторов с ядром плана в виде дробного факторного эксперимента типа 2^{m-p}.

Из приведенной таблицы видно, что скалярные произведения любых двух столбцов матрицы, за исключением столбцов при квадратичных членах типа $(x_r^2 - \beta)$, равны нулю при любом α.

Вместе с этим согласно табл. 9.5.2 сумма произведений любых двух столбцов при квадратичных членах равна

$$2^{m-p}(1-\beta)^2 - 4\beta(\alpha^2 - \beta) + (2m-3)\beta^2 = 2^{m-p}\left(1 - \frac{2^{m-p} + 2\alpha^2}{N}\right)^2 - \\ -4\frac{2^{m-p} + 2\alpha^2}{N}\left(\alpha^2 - \frac{2^{m-p} + 2\alpha^2}{N}\right) + (2m-3)\left(\frac{2^{m-p} + 2\alpha^2}{N}\right)^2.$$

Таблица 9.5.2

Матрица ЦКОП второго порядка для m факторов

	Точка плана i	Факторы и их сочетания									
		x_0	x_1	...	x_m	$x_1^2-\beta$	...	$x_m^2-\beta$	x_1x_2	...	$x_{m-1}x_m$
Ядро плана	1	1	−1	...	+1	$1-\beta$	...	$1-\beta$	+1	...	1
	2	1	−1	...	+1	$1-\beta$	...	$1-\beta$	−1	...	1
	3	+1	+1	...	+1	$1-\beta$	...	$1-\beta$	−1	...	...
	4	+1	+1	...	+1	$1-\beta$	...	$1-\beta$	+1	...	...
	5	1	−1	...	−1	$1-\beta$	...	$1-\beta$	+1	...	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2^{m-p}	1	...	...	...	$1-\beta$	...	$1-\beta$	...	...	...
Звездные точки	$2^{m-p}+1$	1	$+\alpha$	...	0	$\alpha^2-\beta$	...	$-\beta$	0	...	0
	$2^{m-p}+2$	1	$-\alpha$	...	0	$\alpha^2-\beta$	...	$-\beta$	0	...	0
	$2^{m-p}+3$	1	0	...	0	$-\beta$	...	$-\beta$	0	...	0
	$2^{m-p}+4$	1	0	...	0	$-\beta$	...	$-\beta$	0	...	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	...	1	0	...	$+\alpha$	$-\beta$	...	$\alpha^2-\beta$	0	...	0
	$2^{m-p}+2m$	1	0	...	$-\alpha$	$-\beta$	...	$\alpha^2-\beta$	0	...	0
Центр плана	$N=2^{m-p}+$ $+2m+1$	1	0	...	0	$-\beta$	...	$-\beta$	0	0	0

Приравнивая эту сумму к нулю и решая получившееся уравнение относительно α, можно найти [51] функциональную зависимость α от количества факторов m, типа ядра плана эксперимента и общего числа точек плана N:

$$\alpha=\sqrt{2^{(m-p)/2-1}\sqrt{N-2^{(m-p)/2}}}. \tag{9.5.8}$$

Необходимо отметить, что в качестве ядра ЦКОП могут использоваться не все возможные виды дробных факторных планов. Это обусловлено тем, чтобы в матрице ядра плана у парных взаимодействий не было одинаковых столбцов. Иначе говоря, в качестве ядра в ЦКОП могут быть использованы только такие ДФЭ, которые позволяют получить несмешанные друг с другом коэффициенты модели при парных взаимодействиях. Анализ планов ДФЭ показывает, что для $m \leqslant 4$ ДФЭ, удовлетворяющих высказанным выше требованиям, отсутствуют. Подобные ограничения имеются на типы ДФЭ и для $m > 4$. В литературе по планированию эксперимента приведены возможные типы ДФЭ для разных значений числа факторов m и параметра p. В табл. 9.5.3 приведены значения параметров α и β для наиболее часто встречающихся ЦКОП [7, 49, 65].

Анализ результатов экспериментов, осуществленный согласно плану ЦКОП, аналогичен анализу результатов при проведении опытов по плану ПФЭ, рассмотренному в п. 9.4.1. Однако ряд элементов столбцов матрицы плана ЦКОП отличаются от значений ±1, характерных для ортогонального плана ПФЭ. Если в каждой точке плана ЦКОП осуществ-

Таблица 9.5.3

Параметры центральных композиционных ортогональных планов

	Число факторов m						
	2	3	4	5	6	7	8
Ядро плана	2^2	2^3	2^4	2^{5-1}	2^{6-1}	2^{7-1}	2^{8-2}
Число точек плана N	9	15	25	27	45	79	81
α	1,000	1,215	1,414	1,547	1,722	1,885	2,001
β	0,667	0,730	0,800	0,770	0,843	0,900	0,889

ляется по n опытов, то выборочное среднее и выборочная дисперсия по каждой строке плана определяются по формулам (9.4.11). Вычисление оценок коэффициентов модели c_0 и b_l, b_{ll}, b_{lq}, $l,q=1,2\ldots,m$, должно осуществляться на основе исходной формулы типа (9.4.12), а именно:

$$c_0^*=\frac{1}{N}\sum_{i=1}^{N}y_{i\cdot};\quad b_l^*=\frac{\sum\limits_{i=1}^{N}y_{i\cdot}(x_l)_i}{\sum\limits_{i=1}^{N}(x_l^2)_i};\quad b_{ll}^*=\frac{\sum\limits_{i=1}^{N}y_{i\cdot}(x_l^2-\beta)_i}{\sum\limits_{i=1}^{N}(x_l^2-\beta)_i^2};$$
$$b_{lq}^*=\frac{\sum\limits_{i=1}^{N}y_{i\cdot}(x_lx_q)_i}{\sum\limits_{i=1}^{N}(x_lx_q)_i^2}\quad (l,q=1,2,\ldots,m). \tag{9.5.9}$$

При этом оценка коэффициента b_0^* в исходной «непреобразованной» модели вычисляется согласно (9.5.6) как

$$b_0^*=c_0^*-\beta\sum_{l=1}^{m}b_{ll}^*. \tag{9.5.10}$$

Дисперсии полученных оценок коэффициентов модели находятся на основе формулы (9.4.13), в которую вместо x_{li}^2 необходимо подставить элементы столбца, соответствующего оцениваемому коэффициенту. Принципиально без изменения остаются процедуры проверки значимости коэффициентов модели и ее адекватности. Вместе с этим суммы в знаменателе (9.4.13) различны для разных коэффициентов. Поэтому в противоположность ортогональному плану ПФЭ оценка значимости коэффициентов модели будут осуществляться с различной точностью, которая также зависит от поворота системы координат, соответствующей m факторам. Последнее означает, что точность статистической модели плана ЦКОП на одинаковом расстоянии от центра плана ρ зависит от направления на прогнозируемую точку в факторном пространстве [7, 50, 65, 67].

9.5.2. Центральный композиционный ротатабельный план (ЦКРП)

Лучшим критерием оптимальности для построения квадратичных моделей является критерий ротатабельности, который, как отмечалось выше, гарантирует одинаковую дисперсию $s^2(y)$ выходного отклика y во всех направлениях на одинаковом расстоянии ρ от центра. Математическая теория построения ротатабельных планов второго порядка достаточно сложна, а полученные соотношения для вычисления коэффициентов полинома и их дисперсий значительно сложнее, чем при ЦКОП. Применительно к квадратичной модели (9.5.1) они приведены в [7, 65 и др.].

Для построения центральных ротатабельных планов в качестве ядра плана могут использоваться как планы ПФЭ типа 2^m, так и планы ДФЭ типа 2^{m-p}. При этом величина плеча α выбирается исходя из соотношения

$$\alpha = 2^{(m-p)/4}. \qquad (9.5.11)$$

Одной из особенностей ЦКРП является проведение нескольких (m_0) опытов в центре плана. Это вызвано тем, что для моделей второго и большего порядка наибольшую информацию об отличии отклика от модели первого порядка обеспечивает информация в центре плана. Поэтому именно в центре плана должно быть получено экспериментальное значение отклика с наименьшими ошибками. Математически это сводится к тому, чтобы дисперсия оценки отклика оставалась неизменной (или почти неизменной) для всех точек гиперсферы заданного радиуса ρ. В табл. 9.5.4 приведены параметры ряда центральных композиционных ротатабельных планов [49, 65].

В ЦКРП оценки коэффициентов модели при линейных членах и парных взаимодействиях не зависят от оценок остальных коэффициентов модели, в то время как оценки коэффициентов модели при квадратичных членах зависят не только между собой, но зависят и от оценки свободного члена. Это свойство ЦКРП позволяет независимо делать

Таблица 9.5.4

Параметры ЦКРП

Размерность полинома m	Ядро плана	Число звездных точек	m_0	N	α
2	2	4	5	13	1,414
3	2	6	6	20	1,682
4	2	8	7	31	2,000
5	2^5	10	10	52	2,378
	2^{5-1}	10	6	32	2,000
6	2^6	12	15	91	2,828
	2^{6-1}	12	9	53	2,378
7	2^7	14	21	163	3,333
	2^{7-1}	14	14	92	2,828

выводы относительно линейных членов и их взаимодействия, с одной стороны, и квадратичных членов и свободного члена — с другой.

Дальнейшие исследования в теории планирования эксперимента показали, что могут быть предложены критерии, позволяющие выбрать минимальное число точек плана в зависимости от требований, предъявляемых к вектору оценок коэффициентов модели. Так, например, критерий A-оптимальности позволяет построить план эксперимента, при котором минимизируется средняя дисперсия оценок коэффициентов модели. В свою очередь критерий D-оптимальности минимизирует объем эллипсоида рассеяния оценок коэффициентов модели. Исходя из конкретно решаемых экспериментальных исследований, в литературе [26, 32, 65, 67 и др.] изучается ряд других критериев построения оптимальных планов проведения экспериментов.

9.6. Планы экстремальных экспериментов

Одним из направлений экспериментального исследования во многих прикладных задачах естественных и других наук является нахождение значений уровней факторов и их сочетаний, обеспечивающих оптимальный выходной эффект в соответствии с выбранным или заданным критерием оптимальности. План экспериментального исследования, направленный на отыскание экстремума выходного отклика (функции регрессии), называется *планом экстремального эксперимента*. Постановка и проведение экстремальных экспериментов становится особенно актуальным при создании новых систем и устройств, а также при разработке перспективных химико-технологических процессов, обладающих определенными оптимальными свойствами. Например, применительно к системам передачи данных можно потребовать, чтобы создаваемая система обеспечивала максимальную пропускную способность передачи информации при фиксированных экономических затратах. Аналогично при разработке новых химико-технологических процессов необходимо экспериментально подобрать оптимальное сочетание процентного содержания составляющих элементов и внешних воздействий (температура, давление, время протекания процесса и т.п.), при которых выход продукции будет максимален. Решение этих и других подобных задач предусматривает два принципиально различных подхода планирования экстремального эксперимента.

Первый подход нахождения оптимальных условий для получения наилучшего (в заданном смысле) выходного эффекта основан на возможности создания статистической модели изучаемого процесса. Этот подход характерен для экспериментов, проводимых в лабораторных условиях.

Второй подход определения оптимальных условий для получения наилучшего (в заданном смысле) выходного эффекта основан на регулировании уровней факторов протекающего процесса непосредственно на объекте исследования. Данный подход используется, как правило,

в процессе функционирования систем и устройств, а также в процессе промышленного производства. Применительно к регулированию данного класса систем (объектов) и процессов характерны особенности, основными из которых являются следующие:

- весьма ограниченные возможности варьирования уровнями факторов из-за опасения нарушения функционирования систем и процессов;
- большое число неконтролируемых факторов («шумов»), создающих дополнительные ошибки при принятии решений.

В связи с отмеченной выше спецификой нахождения оптимальных условий для получения того или иного максимального эффекта при минимуме числа опытов применяются разные методы их планирования. Для планирования экстремальных экспериментов в лабораторных условиях наибольшее применение нашли метод крутого восхождения и симплексный метод, а для поиска оптимума в промышленных условиях — метод эволюционного планирования. В основу всех методов экстремального эксперимента положены принципы последовательных приближений.

9.6.1. Метод крутого восхождения (метод Бокса–Уилсона)

Методы экспериментального поиска экстремума выходного эффекта основаны на его изучении путем постановки соответствующих опытов около некоторой точки в m-мерном факторном пространстве. Анализ отклика в этой точке позволяет выработать правило последовательного движения (изменения уровней факторов) к экстремуму определенными «шагами». В литературе рассматривается ряд экспериментальных методов поиска экстремума, различающихся способом определения направления движения и организацией движения к экстремуму. Как отмечалось выше, при достаточно большом количестве независимых факторов практически все методы, за исключением метода крутого восхождения, имеют существенные недостатки, связанные с большими временными затратами при движении к экстремуму.

Рассмотрим сущность метода крутого восхождения применительно к отклику

$$y = \varphi(x_1, x_2), \tag{9.6.1}$$

зависящему от двух факторов x_1 и x_2, и представляющему некоторую поверхность в трехмерном пространстве. Соединяя точки (x_1, x_2), дающие одинаковые значения y, поверхность отклика можно изобразить с помощью линий или контуров равного отклика. На рис. 9.6.1 приведен возможный вид проекций $y = \varphi(x_1, x_2)$ на плоскость поверхности отклика типа «холма» с вершиной в точке O. Попасть в эту точку из точки A можно различными путями.

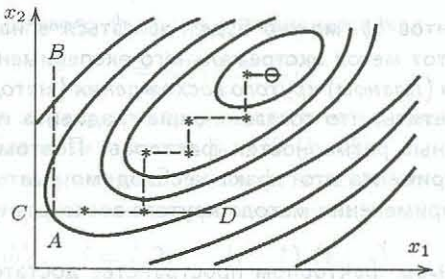


Рис. 9.6.1. Контуры равного отклика двумерной поверхности

В случае двухфакторного эксперимента последовательное движение к экстремуму-максимуму можно осуществить *методом* (*планом*) *покоординатного движения* (*метод Гаусса–Зейделя*). Сущность метода состоит в следующем. Оставляем без изменения один из факторов, например фактор x_1, а фактор x_2 изменяется в направлении AB до тех пор, пока увеличивается выходной отклик (достигается локальный максимум по переменной x_2 при $x_1 = \text{const}$). Значение (уровень) фактора x_2 «замораживается» и изменяется значение фактора x_1 в направлении CD до тех пор, пока увеличивается выход отклика (достигается локальный максимум по переменной x_1 при $x_2 = \text{const}$). Затем извилистый процесс движения к экстремуму-максимуму повторяется и далее до точки, движение из которой не увеличивает выходной эффект. Поскольку значения факторов изменяются не непрерывно, а дискретно, то точность попадания в точку максимума, естественно, зависит от шага дискретизации уровней факторов при постановке экспериментов. Несмотря на существенную простоту метода, не требующего формирования аналитической или статистической зависимости отклика от факторов, он практически не применяется при количестве факторов более 3–4.

Наиболее короткий путь к вершине «холма» это направление, перпендикулярное линиям одинаковых уровней отклика (направление градиента функции отклика).

Вектор-градиент непрерывной функции $y = \varphi(x_1, x_2, \ldots, x_m)$ записывается в виде

$$\operatorname{grad} y = \frac{\partial \varphi}{\partial x_1}\mathbf{i} + \frac{\partial \varphi}{\partial x_2}\mathbf{j} + \ldots, \tag{9.6.2}$$

где $\mathbf{i}$, $\mathbf{j}$, ...— единичные векторы (орты) в направлении координатных осей. При этом составляющие градиента являются частными производными функции отклика. Если статистическая модель отклика линейная, т.е. описывается уравнением (9.4.9), то коэффициенты $b_k = (\partial\varphi/\partial x_k)$, $k = 1, 2, \ldots, m$, представляют собой координаты вектор-градиента. В векторном анализе доказывается, что движение в направлении градиента является движением к экстремуму по кратчайшему наиболее крутому пути. Изменяя уровни факторов пропорционально найденным оцен-

кам коэффициентов b_k, можно будет двигаться в направлении градиента. Поэтому этот метод экстремального эксперимента Бокс и Уилсон назвали *методом (планом) крутого восхождения (метод Бокса-Уилсона)*. Необходимо отметить, что составляющие градиента и сам градиент зависят от выбранных размерностей факторов. Поэтому в процессе планирования эксперимента этот факт необходимо учитывать.

Процедура применения метода крутого восхождения состоит из следующих этапов.

1. В изучаемом факторном пространстве достаточно произвольно выбирается некоторая точка, удаленная от экстремума. В окрестности этой точки в предположении модели первого порядка проводится соответствующий эксперимент, на основе которого находятся коэффициенты модели и проверяется гипотеза об адекватности модели. После вычисления оценок коэффициентов и проверки адекватности модели определяется линейное приближение градиента в виде

$$\operatorname{grad} y = b_1^*\mathbf{i} + b_2^*\mathbf{j} + \ldots. \tag{9.6.3}$$

При изменении уровней факторов пропорционально значениям оценок линейных коэффициентов модели будет происходить движение в направлении градиента поверхности (функции) отклика.

При построении модели необходимо так выбирать интервалы варьирования факторов в натуральных единицах, чтобы величины значимых коэффициентов модели имели одинаковый порядок. Модель поверхности отклика, значения коэффициентов которой различаются несущественно, называется симметричной относительно коэффициентов. Если функция отклика существенно асимметрична, т.е. если какие-либо коэффициенты при линейных членах различаются на порядок от остальных коэффициентов, то необходимо поставить новую серию опытов, изменив для этих факторов интервалы варьирования.

2. Для осуществления движения по линии градиента выбирается «шаг» движения λ_r (изменения) фактора x_r^* в натуральных единицах. К сожалению, для выбора величины шага не существует формализованного решения. Он определяется исходя из физических соображений и разрешения определенного компромисса между точностью движения к экстремуму и необходимым объемом эксперимента. Небольшой шаг потребует значительного числа опытов при движении к оптимуму, большой шаг увеличивает вероятность пропуска области экстремума. При этом целесообразно выбирать исходное значение «шага» в натуральных единицах для фактора, который в наибольшей степени влияет на функцию отклика. Этот фактор часто называют базовым ($x_б^*$), он обычно соответствует наибольшему абсолютному значению произведения оценки коэффициента $|b_б^*|$ базового фактора на интервал его варьирования $\Delta x_б^* = (x_{б\max}^* - x_{б\min}^*)/2$. Например, если базовым фактором в натуральных единицах является i-й фактор, то $x_i^* \equiv x_б^*$, а коэффициент линеной модели (9.4.9) для этого фактора $b_i^* \equiv b_б^*$. Нормированное

значение шага λ базового фактора определяется как

$$\lambda = \lambda_б^*/(\Delta x_б^*|b_б^*|), \tag{9.6.4}$$

где $\lambda_б^*$ — значение шага базового фактора в натуральных единицах, которое обычно выбирается равным $\Delta x_б^*/2$ [1, 65].

Величина ненормированного шага по остальным факторам в натуральных единицах определяется соотношением

$$\lambda_j = \lambda\Delta x_j^*|b_j^*|, \quad j = 1, 2, \ldots, (i-1), (i+1), \ldots, m. \tag{9.6.5}$$

Координаты j-го фактора $x_{j;l}^*$ на l-м шаге по направлению градиента отсчитываются от первоначальной точки $x_{j;0}^*$, в качестве которой, как правило, выбирают центр плана либо одну из точек плана в направлении экстремума:

$$x_{j;l} = x_{j;0}^* + l\lambda\Delta x_j^* b_j^*. \tag{9.6.6}$$

Введенные выше соотношения показывают, что значения шагов λ_i, $i = 1, 2, \ldots, m$, соотносятся как модули величин $|\Delta x_j^* b_j^*|$:

$$\lambda_1 : \lambda_2 : \ldots : \lambda_m = |\Delta x_1^* b_1^*| : |\Delta x_2^* b_2^*| : \ldots : |\Delta x_m^* b_m^*|. \tag{9.6.7}$$

Если полученная модель первого порядка около исходной точки оказалась неадекватной или коэффициенты взаимодействия модели первого порядка существенны по сравнению с коэффициентами линейных членов, то изменяют параметры начального эксперимента. Обычно это сводится к уменьшению интервалов варьирования факторов и переносу начальной точки, что, естественно, требует постановки новой серии опытов.

3. Движение по градиенту осуществляется до тех пор, пока происходит изменение поверхности отклика в направлении к экстремуму. Как только обнаруживается, что отклик удаляется от экстремума, находится новое описание поверхности отклика в виде модели первого порядка. Описанный выше процесс оптимизации продолжается до точки факторного пространства (стационарной области), для которой принятая модель становится неадекватной и для описания которой требуется зависимость более высокого порядка (второго, а иногда и третьего порядка).

4. Для более точного описания поверхности отклика в стационарной области проводится большое число опытов. На основе этого описания находится точка факторного пространства, соответствующая экстремуму поверхности отклика.

5. При движении по градиенту часто осуществляются, так называемые, «мысленные опыты» (мысленное движение по градиенту), основанные на вычислении прогнозируемых значений поверхности отклика в точках факторного пространства, лежащих на пути к экстремуму от исходной точки. Число и порядок мысленных опытов зависит от конкретной задачи, и, как правило, оно ограничивается 3–5 точками. Для

проверки адекватности выбранной модели и пути движения к экстремуму в одних и тех же точках факторного пространства могут осуществляться как мысленные, так и реальные опыты.

После проведения первой серии опытов по крутому восхождению возможны следующие три варианта решений:

1) завершить опыты, если локальное значение экстремума функции отклика удовлетворяет исследователя;

2) построить линейный план следующего цикла крутого восхождения;

3) построить план более высокого порядка для детального изучения поверхности отклика и определения глобального экстремума функции отклика.

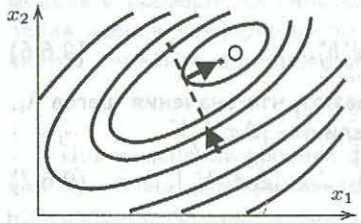


Рис. 9.6.2. Процедура метода крутого восхождения

На рис. 9.6.2 схематически показана процедура применения метода крутого восхождения применительно к двум независимым факторам x_1 и x_2. Движение из первоначальной точки осуществляется в направлении наибольшего изменения функции $y = \varphi(x_1, x_2)$ до достижения частного экстремума. Затем на основе опытов находится новое направление градиента и движение по нему происходит до достижения стационарной области.

Пример 9.6.1. Для иллюстрации метода крутого восхождения рассмотрим данные примера 9.4.1 применительно к построению статистической модели первого порядка, зависящей от двух натуральных факторов: $x_1^* \equiv t$, $x_2^* \equiv T$.

В табл. 9.4.3 приведены результаты опытов в каждой точке плана ПФЭ типа 2^2 и прогнозируемые значения отклика в этих точках. Функция отклика (уравнение регрессии) для нормированных и ненормированных (натуральных) значений переменных определяются выражениями (9.4.18) и (9.4.19) соответственно. При рассмотрении метода крутого восхождения составляют таблицу, в которую заносят исходные данные, расчеты значений шага и координат проведения мысленных и результаты реальных опытов. Эти данные отражены в табл. 9.6.1.

В таблице обозначены: x_i^* — значения факторов в натуральных единицах; $y.$ — среднее значение отклика, полученное опытным путем; y^* — прогнозируемое значение отклика; b_i^* — оценки коэффициентов модели первого порядка. В верхней части таблицы приведены исходные данные начального эксперимента в натуральных единицах. В средней части таблицы указан план проведения опытов, значения отклика в этих нормированных точках плана, величины оценок коэффициентов модели и модуль произведения $|b_i^*\Delta_i^*|$, $i = 1, 2$. По результатам анализа выбирается базовый фактор x_2 и величина шага $\lambda_2 = 25$ °C. При этом шаг оставшегося фактора округляется до значения $\lambda_1 = 2{,}0$ минут. В качестве начальной точки отсчета выбрана точка (+1,+1), соответствующая максимальному среднему значению отклика $y. = 44{,}9$ исходного плана ПФЭ типа 2^2.

Расчеты величины шага и координаты дальнейших опытов приведены в нижней части таблицы. В последних двух колонках нижней части таблицы приведены значения функции отклика реальных опытов и расчетные (прогнозируемые) значения. Из последних данных видно, что максимальное значение функции отклика (локальный максимум) достигается в одиннадцатом опыте, соответствующем $t = 42$ мин и

Таблица 9.6.1

Переменные		$x_1^* \equiv t$	$x_2^* \equiv T$ °C	$y.$	y^*
Основной уровень		20	250		
Интервал варьирования Δx^*		10	50		
Верхний уровень $x_{i\max}^*$		30	300		
Нижний уровень $x_{i\min}^*$		10	200		
Нормированные значения факторов		x_1	x_2		
Опыты ПФЭ типа 2^2	1	−1	−1	31,0	30,9
	2	−1	+1	38,0	37,9
	3	+1	−1	32,0	31,9
	4	+1	+1	45,0	44,9
	5	0	0	36,2	36,4
Оценки коэффициентов модели b_i^* ($b_0^* = 36{,}4$; $b_{12}^* = 1{,}5$)		2,0	5,0		
$\lvert b_i^*\Delta_i^*\rvert$		20	250		
Начальная точка отсчета		30 мин	300 °C		
Шаг при изменении x_2^* на 25 °C		2,0 мин	25 °C		
Округленная величина шага		2,0 мин	25 °C		
Опыты	6	32 мин	325 °C		49,0
	7	34 "	350 "		53,4
	8	36 "	375 "	61,7	55,8
	9	38 "	400 "		63,1
	10	40 "	425 "	70,2	68,4
	11	42 "	450 "	77,6	74,0
	12	44 "	475 "	74,4	77,9

$T = 450$ °C. По сравнению с исходной (первоначальной) точкой значение функции отклика возросло почти в 1,71 раза.

Для дальнейшего движения к экстремуму необходимо построить новый план ПФЭ типа 2^2 с центром в точке $t = 42$ мин и $T = 450$ °C. При этом, естественно, может измениться базовый фактор. Для того чтобы более точно выйти на следующий локальный максимум, шаг базового фактора необходимо уменьшить по сравнению с предыдущим циклом.

9.6.2. Симплексный метод

Симплексный метод (*план*) нахождения совокупности уровней факторов, обеспечивающих получение экстремального (оптимального) значения функции отклика, основан на проведении опытов в точках многомерного факторного пространства, расположенных в вершинах регулярного симплекса. В дальнейшем будем полагать, что искомым экстремумом является экстремум-максимум.

Регулярным симплексом (далее просто симплексом) называется множество $m + 1$ независимых точек, образующих выпуклую фигуру в m-мерном пространстве с равными расстояниями между всеми этими точками, называемыми вершинами симплекса. Число вершин симплекса всегда на единицу больше размерности m-мерного пространства. В одномерном пространстве симплексом является отрезок прямой, в двумерном пространстве (т.е. на плоскости) — равносторонний треуголь-

ник, в трехмерном пространстве — равнобедренная треугольная пирамида (тетраэдр) и т.д. Замечательным свойством симплекса является то, что из каждого симплекса можно построить новый симплекс того же порядка, отбросив одну из его вершин и заменив ее зеркальным отражением от противоположной грани.

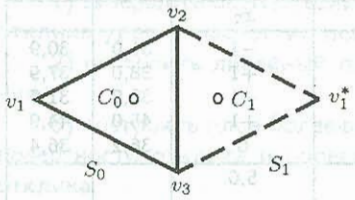


Рис. 9.6.3. Построение симплекса S_1 из симплекса S_0

На рис. 9.6.3 применительно к симплексу на плоскости (для двух факторов) S_0 с центром C_0 и тремя вершинами v_1, v_2 и v_3 показано построение нового симплекса того же третьего порядка S_1 с центром C_1. Симплекс S_1 получается путем замены вершины v_1 ее зеркальным отражением v_1^* от грани v_2v_3.

Многократное отражение симплекса приводит к шаговому движению центра симплекса по траектории некоторой ломаной линии. Координаты новой точки в векторной форме $\mathbf{v}_j^*$ представляют собой разность между удвоенным средним значением координат исходного симплекса C и координаты отбрасываемой j-й точки (вершины) [49, 50, 65]:

$$\mathbf{v}_j^* = \frac{2}{m}(\mathbf{v}_1 + \mathbf{v}_2 + \ldots + \mathbf{v}_{j-1} + \mathbf{v}_{j+1} + \ldots + \mathbf{v}_{m+1}) - \mathbf{v}_j = \\ = \frac{2}{m}\sum_{i=1}^{m+1}\mathbf{v}_i - \left(\frac{2}{m}+1\right)\mathbf{v}_j. \qquad (9.6.8)$$

Здесь j — номер отбрасываемой вершины исходного симплекса.

Для выявления направления движения к экстремуму в вершинах симплекса проводятся эксперименты. На основе сравнения значений функций отклика в вершинах симплекса выбирается точка (вершина), соответствующая сочетанию уровней факторов, при которых выходной эффект минимален. Отбрасывая эту точку, строится новый симплекс в направлении движения к оптимуму, и в новой точке (вершине) симплекса проводится дополнительный эксперимент. Если в этой точке значение отклика превышает значения откликов в других точках нового симплекса, то процесс движения к оптимуму повторяется. В противном случае необходимо возвратиться к предыдущему симплексу и отбросить вершину, для которой значение отклика следует по порядку за «наихудшей» вершиной симплекса. Затем процесс движения к оптимуму продолжается, как описано выше. В результате описанной процедуры происходит перемещение центра симплекса к оптимуму по ломаной линии, колеблющейся около направления градиента. Эмпирически было установлено, что оптимум считается достигнутым тогда, когда на протяжении нескольких шагов та или иная вершина сохраняет свое положение, т.е. одна и та же точка входит в последовательные симплексы N раз [65], где

$$N = 1{,}65m + 0{,}05m^2, \quad 2 \leqslant m \leqslant 30. \qquad (9.6.9)$$

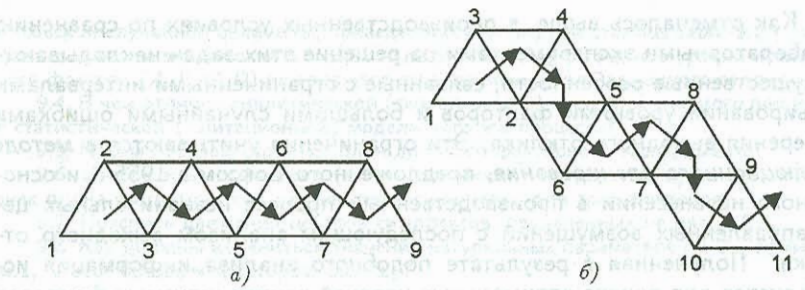


Рис. 9.6.4. Виды движения центра симплекса

В этом случае симплекс совершает вращение вокруг этой точки (вершины).

Виды движения центра симплекса к оптимуму при симплекс-планировании наглядно показаны на рис. 9.6.4 [65] на примере зависимости функции отклика от двух факторов, для которых симплексом является равносторонний треугольник с вершинами v_1, v_2 и v_3. В этих точках проводятся опыты, в результате которых получены значения функции отклика y_1, y_2 и y_3. Пусть наименьшее значение отклика соответствует вершине v_1. Схема рис. 9.6.4,*а* соответствует случаю, при котором последовательные значения функций отклика в «новых» вершинах симплексов больше всех предыдущих. Схема рис. 9.6.4,*б* соответствует ситуации, когда необходимо учитывать вторую вершину симплекса, в которой отклик по своему значению является следующим по сравнению с минимальным откликом.

Из описания симплексного метода планирования эксперимента видно, что для его проведения требуется минимально возможное число опытов. Объем вычислений также незначителен. При этом эффективность этого метода по сравнению с другими методами оптимизации тем выше, чем больше число факторов.

Основным недостатком симплекс-плана является то, что он не дает информации о влиянии каждого фактора на значение функции отклика; симплексный метод оптимизации практически весьма затруднительно использовать для исследования поверхности отклика.

9.6.3. Метод эволюционного планирования

Совокупность оптимальных значений уровней факторов, полученных в результате проведения лабораторных экспериментов, в производственных условиях может быть неоптимальной. Это обусловлено тем, что на процесс производства продукции помимо основных контролируемых факторов воздействует множество случайных неконтролируемых возмущений, в результате чего выходной отклик отличается от оптимума. Поэтому, не нарушая процесс производства, необходимо оценить это отклонение и предложить правила целенаправленного изменения значений уровней контролируемых факторов для получения оптимума откли-

ка. Как отмечалось выше, в производственных условиях по сравнению с лабораторными экспериментами на решение этих задач накладываются существенные особенности, связанные с ограниченными интервалами варьирования уровнями факторов и большими случайными ошибками измерения выходного отклика. Эти ограничения учитываются в *методе эволюционного планирования*, предложенного Боксом в 1955 г. и основанного на внесении в производственный процесс незначительных целенаправленных возмущений с последующим анализом выходного отклика. Полученная в результате подобного анализа информация используется для поиска оптимальных уровней контролируемых факторов (режимов производственного процесса).

Поскольку допустимы лишь малые изменения уровней факторов, возникает задача надежного выявления малых изменений полезного сигнала на фоне шума. Эта задача решается за счет n-кратного повторения независимых опытов в каждой точке плана, в результате чего дисперсия оценки уровня отклика уменьшается в n раз.

Согласно методу эволюционного планирования весь производственный процесс разбивается на отдельные фазы, состоящие из нескольких повторных циклов. Циклом называется однократная реализация плана эксперимента, а фазой — совокупность n реализаций (циклов) одного и того же плана исследований. Как отмечалось выше, в одной фазе осуществляется столько циклов, сколько необходимо для того, чтобы на фоне помех выявить эффект от возмущающих воздействий одного или нескольких факторов на функцию отклика. Если этот эффект (в соответствии с заданным критерием) значим, выбираются новые базовые значения для варьируемых факторов и реализуется новый план эксперимента.

После окончания каждой фазы эксперимента принимается одно из следующих решений:

1) изменить центр плана для смещения всего эксперимента в пространстве независимых факторов;

2) изменить интервалы (интервал) варьирования факторов;

3) изменить уровни ряда факторов, оставив без изменения оптимальные уровни остальных факторов, и начать их варьирование.

Весь процесс эволюционного планирования оформляется в виде специальных рабочих листков, в которые заносятся результаты наблюдений и обработки этих наблюдений. Необходимые расчетные формулы приведены в [65].

Контрольные вопросы и задачи

9.1. Расположить методы рандомизации по возрастанию степени их эффективности для исключения влияния изменения второстепенных (неконтролируемых) факторов на проверку нулевой гипотезы.

9.2. В чем смысл неполноблочного сбалансированного плана?

9.3. Используя стандартный латинский квадрат четвертого порядка, представленный в первом столбце табл. 9.2.8, и первые шестнадцать выборочных значений

гауссовской случайной величины, приведенных в четвертом столбце табл. 9.2.1, проверить с уровнем значимости $\alpha = 0{,}05$ нулевую гипотезу о том, что уровни исследуемого фактора (A, B, C, D) не являются существенными для выходного отклика.

9.4. В чем отличие аналитической (теоретической) модели изучаемого процесса от статистической (имитационной) модели того же процесса?

9.5. Каким образом строится план дробного факторного эксперимента?

9.6. Изобразить графически (в геометрической интерпретации) расположение точек плана статистической модели первого порядка для трех факторов.

9.7. Пояснить движения центров симплексов, приведенных на рис. 9.6.4.

9.8. Как должны измениться значения натуральных параметров t и T в примере 9.6.1, чтобы базовым фактором был x_1?

Список литературы

1. **Адлер Ю.П. и др.** Планирование эксперимента при поиске оптимальных условий. — М.: Наука, 1976.

2. **Айвазян С.А., Мхатарян В.С.** Прикладная статистика. Основы эконометрики. Т. 1, 2. — М.: ЮНИТИ-ДАНА, 2001.

3. **Айвазян С.А., Мхатарян В.С.** Прикладная статистика в задачах и упражнениях. — М.: ЮНИТИ-ДАНА, 2001.

4. **Андерсон Т.** Введение в многомерный статистический анализ. — М.: Физматгиз, 1963.

5. **Андерсон Т.** Статистический анализ временных рядов. — М.: Мир, 1989.

6. **Бендат Дж., Пирсол А.** Прикладной анализ случайных данных. — М.: Мир, 1989.

7. **Блохин В.Г. и др.** Современный эксперимент: подготовка, проведение, анализ результатов. — М.: Радио и связь, 1997.

8. **Большев Л.Н., Смирнов Н.В.** Таблицы математической статистики. — М.: Наука, 1983.

9. **Бриллинджер Д.** Временные ряды — М.: Мир, 1980.

10. **Бокс Дж., Дженкинс Г.** Анализ временных рядов. Прогноз и управление. — М.: Мир, 1974.

11. **Боровков А.А.** Математическая статистика. — Новосибирск: Наука, 1997.

12. **Бочаров П.П., Печенкин А.В.** Теория вероятностей и математическая статистика. — М.: Гардарика, 1998.

13. **Ван Дер Варден Б.Л.** Математическая статистика. — М.: Иностранная литература, 1960.

14. **Вентцель Е.С.** Теория вероятностей.— М.: Высшая школа, 2001.

15. **Вероятность** и математическая статистика: Энциклопедия. — М: Большая Российская энциклопедия, 1999.

16. **Возенкрафт Дж., Джекобс И.** Теоретические основы техники связи. —М.: Мир,1969.

17. **Гаек Я., Шидак З.** Теория ранговых критериев. — М.: Наука, 1971.

18. **Гайдышев И.** Анализ и обработка данных: Специальный справочник. — СПб: Питер, 2001.

19. **Гмурман В.Е.** Теория вероятностей и математическая статистика. — М.: Высшая школа, 2001.

20. **Гнеденко Б.В.** Курс теории вероятностей. — М.: Наука, 1988.

21. **Градштейн И.С.,Рыжик И.М.** Таблицы интегралов, сумм и произведений. — М.: Физматгиз, 1963.

22. **Дженкинс Г., Ваттс Д.** Спектральный анализ и его приложения. — М.: Мир, Вып. 1, 1971, — Вып. 2, 1972.

23. **Дрейпер Н., Смит Г.** Прикладной регрессионный анализ. Кн. 1.— М.: Финансы и статистика, 1983.

24. **Дубров А.М., Мхатарян В.С., Трошин Л.И.** Многомерные статистические методы. — М.: Финансы и статистика, 1998.

25. **Ермаков С.М. и др.** Математическая теория планирования эксперимента. — М.: Наука, 1983.

26. **Ермаков С.М., Жиглявский А.А.** Математическая теория оптимального эксперимента. — М.: Наука. 1988.

27. **Ивченко Г.И., Медведев Ю.И.** Математическая статистика. — М.: Высшая школа, 1984.

28. **Ивченко Г.И., Медведев Ю.И., Чистяков А.В.** Сборник задач по математической статистике. — М.: Высшая школа, 1989.

29. **Живописцев Ф.А., Иванов В.А.** Регрессионный анализ в экспериментальной физике. — М.: Изд-во МГУ, 1995.

30. **Кендалл М., Стьюарт А.** Теория распределений. — М.: Наука, 1966.

31. **Кендалл М., Стьюарт А.** Статистические выводы и связи. — М.: Наука, 1973.

32. **Кендалл М., Стьюарт А.** Многомерный статистический анализ и временные ряды. — М.: Наука, 1976.

33. **Кендалл М.** Временные ряды. — М.: Финансы и статистика, 1981.

34. **Коваленко И.Н., Филиппова А.А.** Теория вероятностей и математическая статистика. — М.: Высшая школа, 1982.

35. **Колемаев В.А., Калинина В.Н.** Теория вероятностей и математическая статистика. — М.: Инфра-М, 2001.

36. **Королюк В.С. и др.** Справочник по теории вероятностей и математической статистике. — Киев: Наукова думка, 1973.

37. **Крамер Г.** Математические методы статистики. — М.: Мир, 1975.

38. **Кремер Н.Ш.** Теория вероятностей и математическая статистика. — М.: ЮНИТИ-ДАНА, 2000.

39. **Куликов Е.И., Трифонов А.П.** Оценка параметров сигналов на фоне помех. — М.: Сов. радио, 1978.

40. **Куликов Е.И.** Методы измерения случайных процессов. — М.: Радио и связь, 1986.

41. **Куликов Е.И.** Прикладной статистический анализ. — М.: Радио и связь, 2003.

42. **Левин Б.Р.** Теоретические основы статистической радиотехники. — М.: Радио и связь, 1989.

43. **Лекиш И., Ляга И.** Основные таблицы математической статистики. — М.: Финансы и статистика, 1985.

44. **Львовский У.Н.** Статистические методы построения эмпирических формул. — М.: Высшая школа, 1988.

45. **Лукашин Ю.П.** Адаптивные методы краткосрочного прогнозирования. — М.: Статистика, 1979.

46. **Математическая** энциклопедия. Т. 3. — М.: Советская энциклопедия, 1985.

47. **Митропольский А.К.** Техника статистических вычислений. — М.: Наука, 1971.

48. **Монтгомери Д.К.** Планирование эксперимента и анализ данных. — Л.: Судостроение, 1980.

49. **Налимов В.В., Чернова Н.А.** Статистические методы планирования экстремальных экспериментов. — М.: Наука, 1965.

50. **Налимов В.В.** Теория эксперимента. — М.: Наука, 1971.

51. **Оуэн Д.В.** Сборник статистических таблиц. — М.: ВЦ АН СССР, 1966.

52. **Пугачев В.С.** Теория вероятностей и математическая статистика. — М.: Наука, 1979.

53. **Рао С.Р.** Линейные статистические методы и их применения. — М.: Наука, 1968.

54. **Румшинский Л.З.** Математическая обработка результатов эксперимента. — М.: Наука, 1971.

55. **Свешников А.А. и др.** Сборник задач по теории вероятностей, математической статистике и теории случайных функций. — М.: Наука, 1970.

56. **Севастьянов Б.А.** Курс теории вероятностей и математической статистики. — М.: Наука, 1982.

57. **Смирнов Н.В., Дунин-Барковский И.В.** Курс теории вероятностей и математической статистики для технических приложений. — М.: Наука, 1969.

58. **Справочник** по прикладной статистике. В 2-х тт. — М.: Финансы и статистика, т. 1 — 1989., т. 2 — 1990.

59. **Тихонов В.И.** Статистическая радиотехника. — М.: Радио и связь, 1982.

60. **Тихонов В.И., Харисов В.Н.** Статистический анализ и синтез радиотехнических устройств и систем. — М.: Радио и связь, 1991.

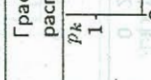
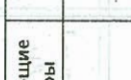
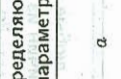
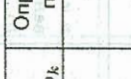
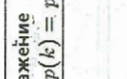
61. **Тихонов В.И., Шахтарин Б.И., Сизых В.В.** Случайные процессы, Примеры и задачи. Т. 1 — М.: Радио и связь, 2003.

62. **Тюрин Ю.Н., Макаров А.А.** Статистический анализ данных на компьютере. — М.: Инфра, 1998.
63. **Уилкс С.** Математическая статистика. — М.: Наука, 1967.
64. **Хальд А.** Математическая статистика с техническими приложениями. — М.: Иностранная литература, 1956.
65. **Хартман К. и др.** Планирование эксперимента в исследовании технологических процессов. — М.: Мир, 1977.
66. **Хеннен Э.** Анализ временных рядов.— М.: Наука, 1964.
67. **Хикс Ч.** Основные принципы планирования эксперимента. — М.: Мир, 1967.
68. **Холлендер М., Вульф Д.** Непараметрические методы статистики. — М.: Финансы и статистика, 1983.
69. **Чистяков В.П.** Курс теории вероятностей. — М.: Агар, 2000.
70. **Шварц Г.** Выборочный метод. — М.: Статистика, 1978.
71. **Шеффе Г.** Дисперсионный анализ. — М.: Наука, 1980.
72. **Янко Я.** Математико-статистические таблицы. — М.: Госстатиздат, 1961.
73. **Пагурова В.И.** Таблицы неполной гамма-функции. — М.: ВЦ АН СССР, 1963.

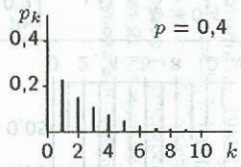
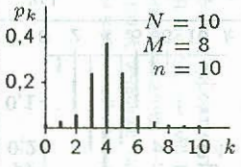
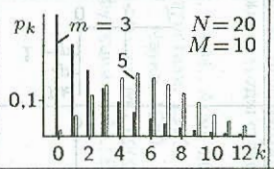
Приложение 1. Таблицы распределений

Законы распределения дискретной случайной величины

Таблица 1

Закон распределения	Область изменения значений	Аналитическое выражение закона распределения $p(k) = p_k$	Определяющие параметры	График закона распределения	Характеристическая функция $\theta(\vartheta)$
1. Вырожденный	$-\infty < k < \infty$	$\mathbf{P}\{K = a\} = 1$	a		$e^{i\vartheta a}$
2. Бернулли (бинарный)	$k = 0, 1$	$p(k) = p^k(1-p)^{1-k}$	p		$1 + p(e^{i\vartheta} - 1)$
3. Биномиальный	$k = 0, 1, 2, \ldots, n$	$p(k) = C_n^k p^k (1-p)^{n-k}$	n, p		$[1 + p(e^{i\vartheta} - 1)]^n$
4. Отрицательный биномиальный (Паскаля)	$k = 0, 1, 2, \ldots$	$p(k) = C_{k+n-1}^{n-1} p^n (1-p)^k$	n, p		$\dfrac{p^n}{[1-(1-p)e^{i\vartheta}]^n}$
5. Пуассона	$k = 0, 1, 2, \ldots$	$p(k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$	λ		$\exp[\lambda(e^{i\vartheta} - 1)]$

Окончание табл. 1.

Закон распределения	Область изменения значений	Аналитическое выражение закона распределения $p(k) = p_k$	Определяющие параметры	График закона распределения	Характеристическая функция $\theta(\vartheta)$
6. Геометрический (Фарри)	$k = 0, 1, 2, \ldots$	$p(k) = p(1-p)^k$	p		$\dfrac{p}{1-(1-p)e^{i\vartheta}}$
7. Гипергеометрический	$k = 0, 1, 2, \ldots,$ $\min(M, n)$	$p(k) = \dfrac{C_M^k C_{N-M}^{n-k}}{C_N^n}$	N, M, n		
8. Отрицательный гипергеометрический	$k = 0, 1, 2, \ldots,$ $N - M$	$p(k) = \dfrac{C_{k+m-1}^k C_{N-m-k}^{M-m}}{C_N^M}$ $m \leqslant M \leqslant N$	N, M, m		
9. Полиномиальный	$k_1 = 0, 1, \ldots, n;$ $\ldots$ $k_m = 0, 1, \ldots, n;$ $\sum_{i=1}^{m} k_i = n$	$p(k_1, k_2, \ldots, k_m) =$ $= \dfrac{n!}{k_1! k_2! \cdots k_m!} p_1^{k_1} \cdots p_m^{k_m};$ $p_1 + p_2 + \ldots + p_m = 1$	n и любые $n-1$ величин из $p_1, p_2, \ldots, p_n$		

28 — 8987

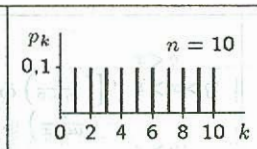
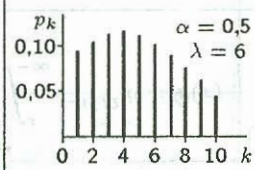
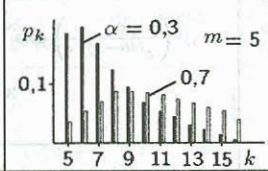
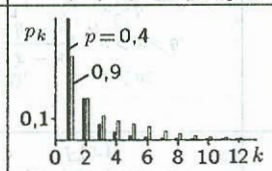
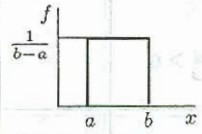
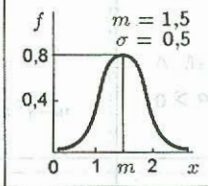
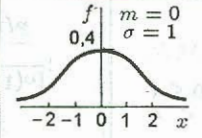
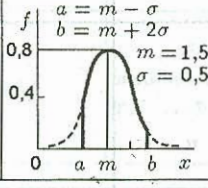
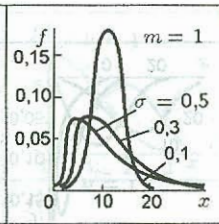
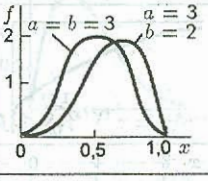
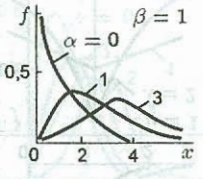
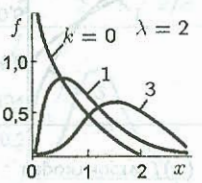
10. Равномерный	$k = 1, 2, \ldots, n$	$p(k) = \dfrac{1}{n}$	n		$\exp\left[\dfrac{i\vartheta(1-e^{i\vartheta n})}{n(1-e^{i\vartheta})}\right]$
11. Пойа	$k = 1, 2, \ldots;$ $\alpha > 0$	$p(k) = p_0 \left(\dfrac{\lambda}{1+\alpha\lambda}\right)^k \times$ $\times \dfrac{1(1+\alpha)\ldots[1+(k-1)\alpha]}{k!};$ $p_0 = p(0) = (1+\alpha\lambda)^{-1/\alpha}$	$\alpha > 0,$ $\lambda > 0$		$\dfrac{1}{[1+\alpha\lambda(1-e^{i\vartheta})]^{\lambda/\alpha}}$
12. Бореля–Таннера	$k = 1, 2, \ldots$	$p(k) = \dfrac{mk^{k-m-1}}{(k-m)!} e^{-\alpha k} \alpha^{k-m}$ $k = m, m+1, \ldots$	$0 < \alpha < 1$		
13. Логарифмическое	$k = 0, 1, 2, \ldots$	$-\dfrac{1}{\ln(1-p)} \dfrac{p^k}{k}$	$0 < p < 1$		$\dfrac{\ln(1-pe^{i\vartheta})}{\ln(1-p)}$

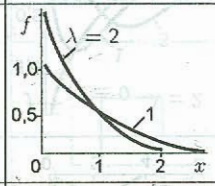
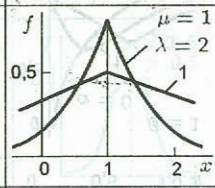
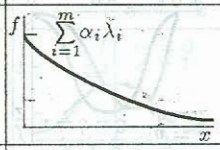
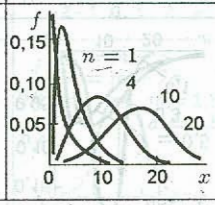
Таблица 2

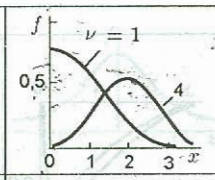
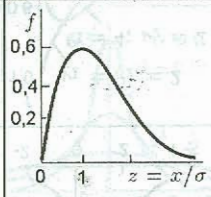
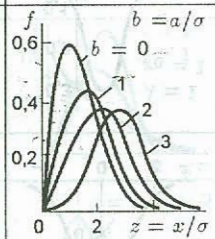
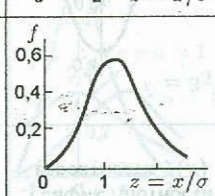
Законы распределения непрерывной случайной величины

Закон распределения	Область изменения значений	Аналитическое выражение плотности вероятности $f(x)$	Параметры $f(x)$	График плотности вероятности $f(x)$	Аналитическое выражение функции распределения $F(x)$	Характеристическая функция $\theta(\vartheta)$
1. Прямоугольный (равномерный)	$a<x<b$	$\frac{1}{b-a}$	a, b		$\begin{cases} 0, & x<a; \\ \frac{x-a}{b-a}, & a<x<b, \\ 1, & x>b \end{cases}$	$\frac{e^{ib\vartheta}-e^{ia\vartheta}}{i\vartheta(b-a)}$
2. Нормальный (гауссовский)	$-\infty<x<\infty$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$	m, σ		$\frac{1}{\sigma\sqrt{2\pi}}\times$ $\times\int_{-\infty}^{x}\exp\left(-\frac{(t-m)^2}{2\sigma^2}\right)dt=$ $=\Phi\left(\frac{x-m}{\sigma}\right)$	$\exp\left(im\vartheta-\frac{\sigma^2\vartheta^2}{2}\right)$
3. Нормальный стандартный	$-\infty<x<\infty$	$\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$	$m=0,$ $\sigma=1$		$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}dt=\Phi(x)$	$e^{-\vartheta^2/2}$
4. Нормальный усеченный	$a<x<b$	$\frac{c}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right],$ $c=\frac{1}{\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right)}$	$m, \sigma,$ a, b		$\begin{cases} 0, & x<a; \\ c\left[\Phi\left(\frac{x-m}{\sigma}\right)-\right. & \\ \left.-\Phi\left(\frac{a-m}{\sigma}\right)\right], & a<x<b; \\ 1, & x>b \end{cases}$	$c\exp\left(i\vartheta m-\frac{\vartheta^2\sigma^2}{2}\right)\times$ $\times\left[\Phi\left(\frac{b-m-i\vartheta\sigma^2}{\sigma}\right)-\right.$ $\left.-\Phi\left(\frac{a-m-i\vartheta\sigma^2}{\sigma}\right)\right]$

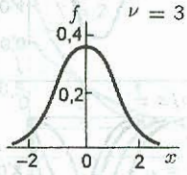
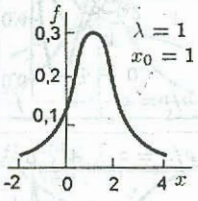
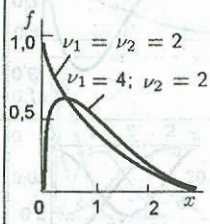
5. Логарифмически нормальный	$0<x<\infty$	$\frac{\log e}{x\sigma\sqrt{2\pi}}\times$ $\times\exp\left[-\frac{(\log x-m)^2}{2\sigma^2}\right]$	m, σ		$\Phi\left(\frac{\log x-m}{\sigma}\right)$	
6. Бета	$0<x<1$	$\frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1},$ $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$	a, b $a>0,$ $b>0$		$\begin{cases} 0, & x<0; \\ \frac{B_x(a,b)}{B(a,b)} & 0<x<1; \\ 1, & x>1, \end{cases}$ где $B_x(a,b)$ — неполная бета-функция	$\frac{\Gamma(a+b)}{\Gamma(a)}\times$ $\times\sum_{m=0}^{\infty}\frac{(i\vartheta)^m}{m!}\frac{\Gamma(a+m)}{\Gamma(a+b+m)}$
7. Гамма	$0<x<\infty$	$\frac{x^{\alpha}e^{-x/\beta}}{\beta^{\alpha+1}\Gamma(\alpha+1)}$	α, β $\alpha>-1$ $\beta>0$		$\frac{\Gamma\left(\alpha+1;\frac{x}{\beta}\right)}{\Gamma(\alpha+1)}$	$(1-i\beta\vartheta)^{-(\alpha+1)}$
8. Эрланга	$0<x<\infty$	$\frac{\lambda^{k+1}}{\Gamma(k+1)}x^{k}e^{-\lambda x}$	λ, k (k — целое)		$\frac{\Gamma(k+1;\lambda x)}{\Gamma(k+1)}$	$\left(1-\frac{i\vartheta}{\lambda}\right)^{-(k+1)}$

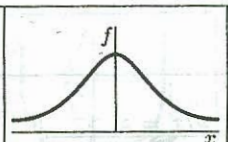
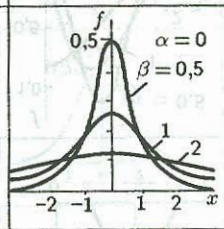
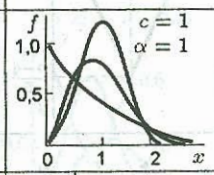
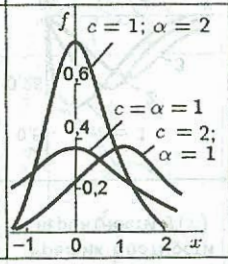
Продолжение табл. 2

Закон распределения	Область изменения значений	Аналитическое выражение плотности вероятности $f(x)$	Параметры $f(x)$	График плотности вероятности $f(x)$	Аналитическое выражение функции распределения $F(x)$	Характеристическая функция $\theta(\vartheta)$
9. Экспоненциальное	$0<x<\infty$	$\lambda e^{-\lambda x}$	λ		$1-e^{-\lambda x}$	$\frac{\lambda}{\lambda - i\vartheta}$
10. Лапласа	$-\infty<x<\infty$	$\frac{\lambda}{2}e^{-\lambda\lvert x-\mu\rvert}$	λ, μ		$\begin{cases} \frac{1}{2}e^{-\lambda(x-\mu)}, & -\infty<x<\mu; \\ 1-\frac{1}{2}e^{-\lambda(x-\mu)}, & \mu<x<\infty \end{cases}$	$\frac{\lambda^2 e^{i\vartheta\mu}}{\lambda^2+\vartheta^2}$
11. Гиперэкспоненциальное	$0<x<\infty$	$\sum_{k=1}^{m}\alpha_k\lambda_k e^{-\lambda_k x}$; $\alpha_k\lambda_k \geqslant 0$; $\sum_{k=1}^{m}\alpha_k = 1$	m, $\alpha_1, \ldots, \alpha_m$; $\lambda_1, \ldots, \lambda_m$		$1-\sum_{k=1}^{m}\alpha_k e^{-\lambda_k x}$	$\sum_{k=1}^{m}\frac{\alpha_k\lambda_k}{\lambda_k - i\vartheta}$
12. Хи-квадрат (χ^2-распределение)	$0<x<\infty$	$\frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}$	ν		$\frac{\Gamma\left(\frac{\nu}{2};\frac{x}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$	$(1-2i\vartheta)^{-\nu/2}$

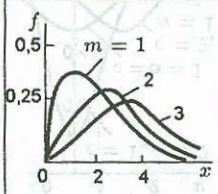
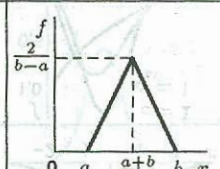
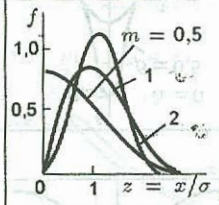
13. χ-распределение	$0<x<\infty$	$\frac{x^{\nu-1}e^{-x^2/2}}{2^{\nu/2-1}\Gamma\left(\frac{\nu}{2}\right)}$	ν		$\frac{\Gamma\left(\frac{\nu}{2};\frac{x^2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$	$\frac{\Gamma(\nu)e^{-\vartheta^2/4}D_{-\nu}(-i\vartheta)}{2^{\nu/2-1}\Gamma\left(\frac{\nu}{2}\right)}$, где $D_p(z)$ — функция гиперболического цилиндра
14. Релея	$0<x<\infty$	$\frac{x}{\sigma^2}e^{-x^2/2\sigma^2}$	σ		$1-e^{-x^2/2\sigma^2}$	$1+ia\vartheta\sqrt{\frac{\pi}{2}}W\left(\frac{a\vartheta}{\sqrt{2}}\right)$, $W(z)=e^{-z^2}\times$ $\times\left(1+\frac{2i}{\sqrt{\pi}}\int_0^z e^{t^2}dt\right)$
15. Райса (обобщенный закон Релея)	$0<x<\infty$	$\frac{x}{\sigma^2}\exp\left(-\frac{x^2+a^2}{2\sigma^2}\right)\times$ $\times I_0\left(\frac{ax}{\sigma^2}\right)$	a, σ		$\exp(-a^2/2\sigma^2)\times$ $\times\sum_{k=0}^{\infty}\frac{x}{(k!)^2}\left(\frac{a^2}{2\sigma^2}\right)^k\times$ $\times\Gamma\left(k+1;\frac{x^2}{2\sigma^2}\right)$	
16. Максвелла	$0<x<\infty$	$\sqrt{\frac{2}{\pi}}\frac{1}{\sigma^2}x^2e^{-x^2/2\sigma^2}$	σ		$\frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2};\frac{x^2}{2\sigma^2}\right)$	$(1-\sigma^2\vartheta^2)W\left(\frac{\sigma\vartheta}{\sqrt{2}}\right)+$ $+i\sigma\vartheta\sqrt{\frac{2}{\pi}}$

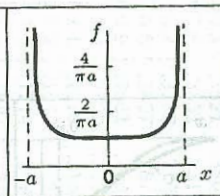
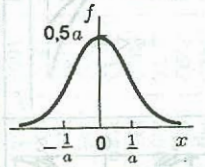
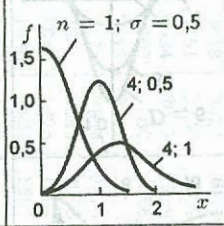
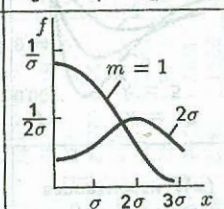
Продолжение табл. 2

Закон распределения	Область изменения значений	Аналитическое выражение плотности вероятности $f(x)$	Параметры $f(x)$	График плотности вероятности $f(x)$	Аналитическое выражение функции распределения $F(x)$	Характеристическая функция $\theta(\vartheta)$
17. Стьюдента (t-распределение)	$-\infty<x<\infty$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\times$ $\times\left(1+\frac{x^2}{\nu}\right)^{-\nu+1/2}$	ν		$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\times$ $\times\int_{-\infty}^{x}\left(1+\frac{t^2}{\nu}\right)^{-\nu+1/2}dt$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\times$ $\times\int_{-\infty}^{\infty}\left(1+\frac{t^2}{\nu}\right)^{-\nu+1/2}e^{i\vartheta t}dt$
18. Коши	$-\infty<x<\infty$	$\frac{1}{\pi}\frac{\lambda}{\lambda^2+(x-x_0)^2}$	λ, x_0		$\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}\frac{x-x_0}{\lambda}$	$\exp(ix_0\vartheta-\lambda\lvert\vartheta\rvert)$
19. F-распределение (Снедекора–Фишера)	$0<x<\infty$	$\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\times$ $\times\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}x^{\nu_1/2-1}\times$ $\times\left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}$	ν_1, ν_2			

20. Фишера z-распределение	$-\infty<x<\infty$	$\frac{2\nu_1^{\frac{\nu_1}{2}}\nu_2^{\frac{\nu_2}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\times$ $\times\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)e^{\nu_1 x}}{(\nu_2+\nu_1 e^{2x})^{\frac{\nu_1+\nu_2}{2}}}$	ν_1, ν_2			$\left(\frac{m_1}{m_2}\right)^{\frac{i\vartheta}{2}}\times$ $\times\frac{\Gamma\left(\frac{m_1+i\vartheta}{2}\right)\Gamma\left(\frac{m_2-i\vartheta}{2}\right)}{\Gamma\left(\frac{m_1}{2}\right)\Gamma\left(\frac{m_2}{2}\right)}$
21. Логистическое	$-\infty<x<\infty$	$\frac{1}{4\beta\operatorname{ch}^2[(x-\alpha)/2\beta]}$	α, β		$\frac{1}{1+\exp[(x-\alpha)/\beta]}$	$\frac{e^{i\alpha\vartheta}\pi\beta\vartheta}{\operatorname{sh}(\pi\beta\vartheta)}$
22. Вейбулла	$0<x<\infty$	$c\alpha x^{\alpha-1}e^{-cx^\alpha}$	$c>0,$ $\alpha>0$		$1-e^{-cx^\alpha}$	$1+i\vartheta\int_0^\infty e^{i\vartheta x-cx^\alpha}dx$
23. Двойное показательное	$-\infty<x<\infty$	$c\alpha\exp(-\alpha x-ce^{-\alpha x})$	$c>0,$ $\alpha>0$		$\exp(-ce^{-\alpha x})$	

Продолжение табл. 2

Закон распределения	Область изменения значений	Аналитическое выражение плотности вероятности $f(x)$	Параметры $f(x)$	График плотности вероятности $f(x)$	Аналитическое выражение функции распределения $F(x)$	Характеристическая функция $\theta(\vartheta)$
24. Показательно-степенное	$0<x<\infty$	$\frac{x^m}{m!}e^{-x}$	m		$\frac{\Gamma(m+1;x)}{\Gamma(m+1)}$	$(1-i\vartheta)^{-(m+1)}$
25. Симпсона (треугольное)	$a<x<b$	$\begin{cases} 0, & -\infty<x<a; \\ \frac{4(x-a)}{(b-a)^2}, & a<x<\frac{b+a}{2}; \\ \frac{4(b-x)}{(b-a)^2}, & \frac{a+b}{2}<x<b; \\ 0, & b<x<\infty \end{cases}$	a, b		$\begin{cases} 0, & -\infty<x<a; \\ \frac{2(x-a)^2}{(b-a)^2}, & a<x<\frac{a+b}{2}; \\ 1-\frac{2(x-a)^2}{(b-a)^2}, & \frac{a+b}{2}<x<b; \\ 1, & b<x<\infty \end{cases}$	$-\frac{4}{v^2(b-a)^2}\times$ $\times\left(e^{i\vartheta b/2}-e^{i\vartheta a/2}\right)^2$
26. Накагами (m-распределение)	$0<x<\infty$	$\frac{2}{\Gamma(m)}\left(\frac{m}{\sigma^2}\right)^m x^{2m-1}\times$ $\times\exp\left(-\frac{m}{\sigma^2}x^2\right),$ $m\geqslant 1/2$	m, σ		$\frac{\Gamma\left(m;\frac{mx^2}{\sigma^2}\right)}{\Gamma(m)}$	$\frac{\Gamma(2m)}{2^{m-1}\Gamma(m)}e^{-\frac{\vartheta^2\sigma^2}{8m}}\times$ $\times D_{-2m}\left(-\frac{i\vartheta\sigma}{\sqrt{2m}}\right)$

27. Арксинуса	$-a<x<a$	$\frac{1}{\pi\sqrt{\alpha^2-x^2}}$	a		$\begin{cases} 0, & x<-a; \\ \frac{1}{2}+\frac{1}{\pi}\arcsin\frac{x}{a}, & -a<x<a; \\ 1, & x>a \end{cases}$	$\frac{1}{\pi}\int_{-a}^{a}\frac{e^{i\vartheta x}}{\sqrt{a^2-x^2}}=$ $=J_0(av)$, где $J_0(av)$ — функция Бесселя нулевого порядка первого рода
28. $\mathrm{sch}^2 x$	$-\infty<x<\infty$	$\frac{\alpha}{2\,\mathrm{ch}^2\,\alpha x}$	a		$\frac{1}{2}+\frac{1}{2}\,\mathrm{th}\,\alpha x$	$\frac{\vartheta\pi}{2\alpha\,\mathrm{sh}\,\frac{\vartheta\pi}{2\alpha}}$
29. Модуля многомерного гауссовского вектора	$0<x<\infty$	$\frac{2x^{n-1}e^{-x^2/(2\sigma^2)}}{(2\sigma^2)^{n/2}\Gamma\left(\frac{n}{2}\right)}$	n, σ		$\frac{\Gamma\left(\frac{n}{2};\frac{x^2}{2\sigma^2}\right)}{\Gamma\left(\frac{n}{2}\right)}$	$\frac{\Gamma(n)e^{-\vartheta^2\sigma^2/4}D_{-n}(i\vartheta\sigma)}{2^{n/2-1}\Gamma\left(\frac{n}{2}\right)}$
30. Модуля гауссовской случайной величины	$0<x<\infty$	$\frac{1}{\sigma\sqrt{2\pi}}\times$ $\times\left\{\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]+\right.$ $\left.+\exp\left[-\frac{(x+m)^2}{2\sigma^2}\right]\right\}$	m, σ		$\Phi\left(\frac{x-m}{\sigma}\right)+\Phi\left(\frac{x+m}{\sigma}\right)-1$	$e^{-\sigma^2\vartheta^2/2}\times$ $\times\left[e^{i\vartheta m}\Phi\left(\frac{m}{\sigma}+i\vartheta\sigma\right)+\right.$ $\left.+e^{-i\vartheta m}\Phi\left(-\frac{m}{\sigma}+i\vartheta\sigma\right)\right]$

Окончание табл. 2

Закон распределения	Область изменения значений	Аналитическое выражение плотности вероятности $f(x)$	Параметры $f(x)$	График плотности вероятности $f(x)$	Аналитическое выражение функции распределения $F(x)$	Характеристическая функция $\theta(\vartheta)$
31. Хотеллинга (T^2-распределение)	$0 < x < \infty$	$\frac{\Gamma\left(\frac{n+1}{2}\right) x^{k/2-1}}{\Gamma\left(\frac{n-k+1}{2}\right)\Gamma\left(\frac{k}{2}\right)} \times \frac{(1+x/n)^{-(n+1)/2}}{n^{k/2}}$	$n \geqslant k \geqslant 1$			
32. Тихонова	$-\pi < x < \pi$	$\frac{1}{2\pi I_0(D)} e^{D\cos x}$	D		$\frac{1}{2\pi I_0(D)} \int_{-\pi}^{\pi} e^{D\cos t} dt$	
33. Парето	$x_0 < x < \infty$	$\frac{\alpha}{x_0}\left(\frac{x_0}{x}\right)^{\alpha+1}$	$x_0 > 0$, $\alpha > 0$		$1 - \left(\frac{x_0}{x}\right)^{\alpha}$	

Приложение 2. Статистические таблицы

Таблица 1а

Нормальная плотность вероятности

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2); \quad f(-x) = f(x).$$

x	0	1	2	3	4	5	6	7	8	9
0,0	,3989	,3989	,3989	,3987	,3986	,3984	,3982	,5280	,3977	,3973
0,1	,3969	,3965	,3961	,3956	,3951	,3945	,3939	,3932	,3925	,3918
0,2	,3910	,3902	,3894	,3885	,3876	,3867	,3857	,3847	,3836	,3825
0,3	,3814	,3802	,3790	,3778	,3765	,3752	,3739	,3726	,3715	,3697
0,4	,3683	,3668	,3653	,3637	,3621	,3605	,3589	,3572	,3555	,3538
0,5	,3521	,3503	,3485	,3467	,3448	,3429	,3410	,3391	,3372	,3352
0,6	,3332	,3312	,3292	,3271	,3251	,3230	,3209	,3187	,3166	,3144
0,7	,3123	,3101	,3078	,3056	,3034	,3011	,2989	,2966	,2943	,2920
0,8	,2897	,2874	,2850	,2827	,2803	,2780	,2756	,2732	,2709	,2685
0,9	,2661	,2637	,2613	,2589	,2565	,2541	,2516	,2492	,2468	,2444
1,0	,2420	,2395	,2371	,2347	,2323	,2299	,2275	,2251	,2227	,2203
1,1	,2178	,2155	,2131	,2107	,2083	,2059	,2033	,2012	,1987	,1965
1,2	,1942	,1919	,1895	,1872	,1849	,1826	,1804	,1781	,1758	,1736
1,3	,1714	,1692	,1669	,1647	,1626	,1604	,1582	,1561	,1539	,1518
1,4	,1497	,1476	,1456	,1435	,1415	,1394	,1374	,1354	,1334	,1315
1,5	,1295	,1276	,1257	,1238	,1219	,1200	,1182	,1163	,1145	,1127
1,6	,1109	,1092	,1074	,1057	,1040	,1023	,1006	,0989	,0973	,0957
1,7	,0941	,0925	,0909	,0893	,0878	,0863	,0848	,0833	,0818	,0804
1,8	,0789	,0775	,0761	,0748	,0734	,0721	,0707	,0694	,0681	,0669
1,9	,0656	,0639	,0632	,0619	,0608	,0596	,0584	,0573	,0562	,0551
2,0	,0540	,0529	,0519	,0508	,0498	,0488	,0478	,0468	,0459	,0449
2,1	,0440	,0431	,0422	,0413	,0404	,0396	,0387	,0379	,0371	,0363
2,2	,0355	,0347	,0339	,0332	,0325	,0317	,0310	,0303	,0296	,0290
2,3	,0283	,0277	,0271	,0264	,0258	,0252	,0246	,0241	,0235	,0229
2,4	,0224	,0219	,0213	,0208	,0203	,0198	,0194	,0189	,0184	,0180
2,5	,0175	,0171	,0167	,0163	,0158	,0155	,0151	,0147	,0143	,0139
2,6	,0136	,0132	,0129	,0126	,0122	,0119	,0116	,0113	,0110	,0107
2,7	,0104	,0101	,0099	,0096	,0094	,0091	,0088	,0086	,0084	,0081
2,8	,0079	,0077	,0075	,0073	,0071	,0069	,0067	,0065	,0063	,0061
2,9	,0059	,0058	,0056	,0055	,0053	,0051	,0050	,0048	,0047	,0046
3,0	,0044	,0043	,0042	,0041	,0039	,0038	,0037	,0036	,0035	,0034

Таблица 16

Функция нормального распределения

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-z^2/2)\, dz; \quad \Phi(-x) = 1 - \Phi(x).$$

x	0	1	2	3	4	5	6	7	8	9
0,0	,5000	,5040	,5180	,5120	,5160	,5190	,5239	,5279	,5319	,5359
0,1	,5398	,5398	,5438	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7703	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9901	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9950	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9983	,9983	,9984	,9984	,9985	,9985	,9986	,9986
3,0	,9987	,9987	,9987	,9988	,9988	,9989	,9989	,9989	,9990	,9999

Таблица 1в

Критические значения нормального распределения такие, что

$$P\{X > x_\beta\} = \frac{1}{\sqrt{2\pi}} \int_{x_\beta}^{\infty} \exp(-z^2/2)\, dz = \beta; \quad \Phi(-x_\beta) = 1 - \Phi(x_\beta).$$

β	0,0025	0,005	0,010	0,020	0,025	0,050	0,100	0,200	0,300
x_β	2,807	2,575	2,326	2,054	1,960	1,645	1,282	0,842	0,524

Таблица 2

Критические значения $\chi^2_{\beta,\nu}$ хи-квадрат распределения такие, что

$$P\{\chi^2_\nu > \chi^2_{\beta,\nu}\} = \beta.$$

ν	β							
	0,995	0,975	0,950	0,900	0,100	0,050	0,025	0,005
1	0,000	0,001	0,004	0,016	2,71	3,84	5,02	7,88
2	0,010	0,051	0,103	0,211	4,61	5,99	7,38	10,60
3	0,072	0,216	0,352	0,584	6,25	7,81	9,35	12,84
4	0,207	0,484	0,711	1,06	7,78	9,49	11,14	14,86
5	0,412	0,831	1,15	1,61	9,24	11,07	12,83	16,75
6	0,676	1,24	1,64	2,20	10,64	12,59	14,45	18,55
7	0,989	1,69	2,17	2,83	12,02	14,07	16,01	20,28
8	1,34	2,18	2,73	3,49	13,36	15,51	17,53	21,96
9	1,73	2,70	3,33	4,17	14,68	16,92	19,02	23,59
10	2,16	3,25	3,94	4,87	15,99	18,31	20,48	25,19
11	2,60	3,82	4,57	5,58	17,28	19,68	21,92	26,76
12	3,07	4,40	5,23	6,30	18,55	21,03	23,34	28,30
13	3,57	5,01	5,89	7,04	19,81	22,36	24,74	29,82
14	4,07	5,63	6,57	7,79	21,06	23,68	26,12	31,32
15	4,60	6,26	7,26	8,55	22,31	25,00	27,49	32,80
16	5,14	6,91	7,96	9,31	23,54	26,30	28,85	34,27
17	5,70	7,56	8,67	10,08	24,77	27,59	30,19	35,72
18	6,26	8,23	9,39	10,86	25,99	28,87	31,53	37,16
19	6,84	8,91	10,12	11,65	27,20	30,14	32,85	38,58
20	7,43	9,59	10,85	12,44	28,41	31,41	34,17	40,00
21	8,03	10,28	11,59	13,24	29,62	32,67	35,48	41,40
22	8,64	10,98	12,34	14,04	30,81	33,92	36,78	42,80
23	9,26	11,69	13,09	14,85	32,01	35,17	38,08	44,18
24	9,89	12,40	13,85	15,66	33,20	36,42	39,36	45,56
25	10,52	13,12	14,61	16,47	34,38	37,65	40,65	46,93
26	11,16	13,84	15,38	17,29	35,56	38,88	41,92	48,29
27	11,81	14,57	16,15	18,11	36,74	40,11	43,19	49,64
28	12,46	15,31	16,93	18,94	37,92	41,34	44,46	50,99
29	13,12	16,05	17,71	19,77	39,09	42,56	45,72	52,34
30	13,79	16,79	18,49	20,60	40,26	43,77	46,98	53,67
35	17,19	20,57	22,46	24,80	46,06	49,80	53,20	60,27
40	20,71	23,43	26,51	29,05	51,81	55,76	59,34	66,77
45	24,31	28,37	30,61	33,35	57,51	61,66	65,41	73,17
50	27,99	32,36	34,76	37,69	63,17	67,51	71,42	79,49
60	35,53	40,48	43,19	46,46	74,40	79,08	83,30	91,95
70	43,27	48,76	51,40	52,33	85,53	90,53	95,02	104,2
80	51,17	57,15	60,39	64,28	96,58	101,9	106,6	116,3
90	59,20	65,65	69,13	73,29	107,6	113,1	118,1	128,3
100	67,33	74,22	77,23	82,36	118,5	123,3	129,6	140,2
120	83,85	91,58	95,70	100,6	140,2	146,5	152,2	163,6

Таблица 3

Критические значения $t_{\beta,\nu}$ распределения Стьюдента такие, что

$$P\{t_\nu > t_{\beta,\nu}\} = \beta; \quad t_{(1-\beta),\nu} = -t_\beta$$

ν	β						
	0,250	0,200	0,100	0,050	0,025	0,010	0,005
1	1,000	1,376	3,078	6,314	12,70	31,821	63,657
2	0,815	1,061	1,886	2,920	4,303	6,925	9,925
3	,765	0,978	1,638	2,353	3,181	4,541	5,841
4	,741	,941	1,533	2,132	2,776	3,747	4,604
5	,727	,920	1,476	2,015	2,571	3,365	4,032
6	,718	,906	1,440	1,943	2,447	3,143	3,707
7	,711	,896	1,415	1,895	2,365	2,998	3,499
8	,706	,889	1,397	1,860	2,306	2,896	3,555
9	,703	,883	1,383	1,833	2,262	2,821	3,250
10	,700	,879	1,372	1,812	2,228	2,764	3,169
11	,697	,876	1,363	1,796	2,201	2,718	3,106
12	,695	,873	1,356	1,782	2,179	2,681	3,055
13	,694	,870	1,350	1,771	2,160	2,650	3,012
14	,692	,868	1,345	1,761	2,145	2,624	2,977
15	,691	,866	1,341	1,753	2,131	2,602	2,947
16	,690	,865	1,337	1,746	2,120	2,583	2,921
18	,688	,862	1,330	1,734	2,101	2,552	2,878
20	,687	,860	1,325	1,725	2,086	2,528	2,845
25	,684	,856	1,316	1,708	2,060	2,485	2,787
30	,683	,854	1,310	1,697	2,042	2,457	2,750
40	,681	,852	1,303	1,684	2,021	2,423	2,704
60	,679	,848	1,296	1,671	2,000	2,390	2,660
∞	,674	,842	1,282	1,645	1,960	2,326	2,575

Таблица 4

Критические значения $F_{\beta;\nu_1,\nu_2}$ F-распределения такие, что

$$P\{F_{\nu_1,\nu_2} > F_{\beta;\nu_1,\nu_2}\} = \beta.$$

Значения для $\beta = 0{,}05$ набраны в первых строках, а значения для $\beta = 0{,}01$ — во вторых строках.

ν_2	ν_1												
	1	2	3	4	5	6	7	8	9	10	12	20	50
1	161	200	216	225	230	234	237	239	241	242	244	248	252
	4050	5000	5400	5630	5760	5860	5930	5980	6920	6060	6100	6210	6300
2	18,5	19,0	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5
	98,5	99,0	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,5
3	10,3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,66	8,59
	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,	26,7	26,41
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,80	5,70
	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14,0	13,7
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,73	4,68	3,56	4,44
	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,89	9,55	9,24

Продолжение табл. 4

ν_2	ν_1												
	1	2	3	4	5	6	7	8	9	10	12	20	50
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,87	3,75
	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,40	7,09
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,44	3,32
	12,6	9,55	8,45	7,85	7,46	7,19	6,99	6.84	6,72	6,62	6,47	6,16	5,86
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,15	3,02
	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,36	5,07
9	5,12	4,25	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	2,94	2,80
	10,6	8,02	7,00	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,81	4,51
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,77	2,64
	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,41	4,12
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,95	2,79	2,65	2,40
	9,64	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,10	3,81
12	4,75	3,89	3,49	3,25	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,54	2,51
	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	3,86	3,57
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,39	2,24
	8,86	6,51	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,93	3,80	3,51	3,22
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,28	2,12
	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,26	2,97
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,19	2,04
	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,70	3,60	3,51	3,37	3,08	2,79
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,12	1,97
	8,10	5,85	4,94	4,43	4,10	4,87	3,70	3,56	3,46	3,37	3,23	2,94	2,64
25	4,24	3,39	3,00	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,01	1,87
	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	3,00	2,70	2,41
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	1,93	1,76
	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,55	2,25
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,84	1,66
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,37	2,06
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,95	1,78	1,60
	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,79	2,70	2,56	2,27	1,95
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,85	1,68	1,48
	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,37	2,07	1,73

Таблица 5

Критические значения $u_{\alpha,n}$ распределения наибольшего отклонения $U_n = \max|x_i - x|/s$ такие, что

$$P\{U_n > u_{\alpha,n}\} = \alpha.$$

α	$n=3$	4	5	6	7	8	9	10	12	14	16	18
0,01	1,41	1,73	1,97	2,16	2,31	2,43	2,53	2,62	2,75	2,86	2,95	3,02
0,05	1,41	1,71	1,92	2,07	2,18	2,27	2,35	2,41	2,52	2,60	2,67	2,73
0,10	1,41	1,69	1,87	2,00	2,09	2,17	2,39	2,29	2,39	2,46	2,52	2,58

Таблица 6

Критические значения $d_{\alpha,n}$ распределения статистики Колмогорова $D_n = \max|F^*(x) - F(x)|$ такие, что

$$P\{D_n > d_{\alpha,n}\} = \alpha.$$

α	$n=3$	4	5	6	7	8	9	10	12	14	16	18
0,01	0,829	,738	,668	,617	,576	,542	,513	,489	,449	,418	,392	,371
0,05	0,708	,624	,563	,519	,483	,454	,430	,409	,375	,349	,327	,309
0,10	0,636	,565	,509	,468	,436	,410	,387	,369	,338	,315	,291	,278

Таблица 7

Критические значения $k_{\alpha/2;n}$ числа положительных K_n и отрицательных $(n - K_n)$ знаков такие, что

$$P\{K_n > k_{\alpha/2,n}; (n - K_n) > k_{\alpha/2,n}\} = \alpha$$

n	α			n	α			n	α		
	0,10	0,05	0,01		0,10	0,05	0,01		0,10	0,05	0,01
5	5	5	–	21	15	16	17	37	24	25	27
6	6	6	–	22	16	17	18	38	25	26	28
7	7	7	–	23	16	17	19	39	26	27	28
8	7	8	8	24	17	18	19	40	26	27	29
9	8	8	9	25	18	18	20	41	27	28	30
10	9	9	10	26	18	19	20	42	27	28	30
11	9	10	10	27	19	20	21	43	28	29	31
12	10	11	11	28	19	20	22	44	28	29	31
13	10	11	12	29	20	21	22	45	29	30	32
14	11	12	13	30	20	21	23	46	30	31	33
15	12	12	13	31	21	22	24	47	30	31	33
16	12	13	14	32	22	23	24	48	31	32	34
17	13	13	15	33	22	23	25	49	31	32	34
18	13	14	15	34	23	24	25	50	32	33	35
19	14	15	16	35	23	24	26				
20	15	15	17	36	24	25	27				

Таблица 8

Критические значения (нижняя и верхняя границы) числа серий $k_{(1-\alpha/2);m=n}$ и $k_{\alpha/2;m=n}$ для $m = n = N/2$ такие, что

$$P\{k_{(1-\alpha/2);m=n=N/2} \leqslant K_n \leqslant k_{\alpha/2;m=n=N/2} = 1 - \alpha.$$

$m=n=N/2$	α			$m=n=N/2$	α		
	0,10	0,05	0,01		0,10	0,05	0,01
1				21	16 27	15 28	13 30
2	1 5	1 5	1 5	22	17 28	16 29	14 31
3	1 7	1 7	1 7	23	17 30	16 31	14 33
4	2 8	1 9	1 9	24	18 31	17 32	15 34
5	3 9	2 10	1 11	25	19 32	18 33	16 35
6	3 11	3 11	2 12	26	20 33	19 34	17 36
7	4 12	3 13	3 13	27	21 34	20 35	18 37
8	5 13	4 14	3 15	28	22 35	21 36	18 39
9	6 14	5 15	4 16	29	23 36	22 37	19 40
10	6 16	6 16	5 17	30	24 37	22 39	20 41
11	7 17	7 17	5 19	31	25 38	23 40	21 42
12	8 18	7 19	6 20	32	25 40	24 41	22 43
13	9 19	8 20	7 21	33	26 41	25 42	23 44
14	10 20	9 21	7 23	34	27 42	26 43	23 46
15	11 21	10 22	8 24	35	28 43	27 44	24 47
16	11 23	11 23	9 25	36	29 44	28 45	25 48
17	12 24	11 25	10 26	37	30 45	29 46	26 49
18	13 25	12 26	11 27	38	31 46	30 47	27 50
19	14 26	13 27	11 29	39	32 47	30 49	28 51
20	15 26	14 27	12 29	40	33 48	31 50	29 52

Таблица 9

Критические значения числа инверсий $g_{\alpha,n}$ такие, что $P\{G_n > g_{\alpha;n}\} = \alpha$

ν	α					
	0,99	0,975	0,95	0,05	0,025	0,01
	9	11	13	31	33	35
12	16	18	21	44	47	49
14	24	27	30	60	63	66
16	34	38	41	78	81	85
18	45	50	54	98	102	107
20	59	64	69	120	125	130
30	152	162	171	263	272	282
40	290	305	319	460	474	489
50	473	495	514	710	729	751
60	702	731	756	1013	1038	1067
70	977	1014	1045	1369	1400	1437
80	1299	1344	1382	1777	1815	1860
90	1668	1721	1766	2238	2283	2336
100	2083	2145	2198	2751	2804	2866

Таблица 10

Критические значения $r_{\alpha,\nu}$ выборочного коэффициента корреляции r_{XY} для $\rho_{XY} = 0$ такие, что

$$P\{|r_{XY}| > r_{\alpha;\nu}\} = \alpha.$$

ν	β				ν	β			
	0,05	0,025	0,01	0,005		0,05	0,025	0,01	0,005
1	0,988	0,997	0,999	0,9999	16	0,400	0,468	0,543	0,590
2	0,900	0,950	0,980	0,990	17	0,389	0,456	0,529	0,575
3	0,805	0,878	0,934	0,959	18	0,378	0,444	0,516	0,561
4	0,729	0,811	0,882	0,917	19	0,369	0,433	0,503	0,549
5	0,669	0,754	0,833	0,875	20	0,360	0,423	0,492	0,537
6	0,621	0,707	0,789	0,834	25	0,323	0,381	0,445	0,487
7	0,582	0,666	0,750	0,798	30	0,296	0,349	0,409	0,449
8	0,549	0,632	0,715	0,765	35	0,275	0,325	0,381	0,418
9	0,521	0,602	0,685	0,735	40	0,257	0,304	0,358	0,393
10	0,497	0,576	0,658	0,708	45	0,243	0,288	0,338	0,372
11	0,476	0,553	0,634	0,684	50	0,231	0,273	0,322	0,354
12	0,457	0,532	0,612	0,661	60	0,211	0,250	0,295	0,325
13	0,441	0,514	0,592	0,641	70	0,195	0,232	0,274	0,302
14	0,426	0,497	0,574	0,623	80	0,183	0,217	0,257	9,283
15	0,412	0,492	0,558	0,600	100	0,164	0,195	0,230	0,254

Ответы к контрольным вопросам и задачам

Глава 1

1.6. $p(H_1|A_1) = 0{,}86$; $p(H_2|A_1) = 0{,}14$.

1.8. $C = \alpha/2$; $\mu = 2/\alpha^2$.

$$F(x) = \begin{cases} 1/2\exp(\alpha x), & -\infty < x \leqslant 0; \\ 1 - 1/2\exp(-\alpha x), & 0 \leqslant x < \infty. \end{cases}$$

1.9. $f(x) = \begin{cases} (\alpha/x_0)(x_0/x)^{\alpha+1} & \text{при } x > x_0; \\ 0 & \text{при } x \leqslant x_0; \end{cases}$ $m = [\alpha/(\alpha-1)]x_0$, $\alpha > 1$.

1.10. $p_4(k=1) = 0{,}0256$; $p_4(k=2) = 0{,}1536$; $p_4(k=3) = 0{,}4096$; $p_4(k=4) = 0{,}4096$; $\mathbf{P}\{k \geqslant 3\} = p_4(k=3) + p_4(k=4) = 0{,}8192$. Указание: рекомендуется применить выражение (1.3.10), а в качестве исходной вероятности взять $p_4(k=0)$.

1.11. Распределение симметрично относительно $m = np = 4$ $(k=4)$, так как $p_8(0) = p_8(8) \approx 0{,}00392$; $p_8(1) = p_8(7) \approx 0{,}03125$; $p_8(2) = p_8(6) \approx 0{,}10937$; $p_8(3) = p_8(5) \approx 0{,}21875$; $p_8(4) \approx 0{,}27343$.

1.12. Так как параметр $\lambda = 10 \gg 1$, можно воспользоваться нормальным приближением со средним значением $m = D = \lambda$. Тогда $\mathbf{P}\{K \leqslant k = 12\} = 0{,}785$.

1.13. В течение одного месяца работы прибора потребуется заменить в среднем $\lambda = 48 : 12 = 4$ компоненты. Так как вероятность ненадежной работы одной компоненты $p = 4 : 480 \approx 0{,}0083$ — величина малая, а число компонент достаточно велико, для определения вероятности можно воспользоваться пуассоновским законом распределения: $\mathbf{P}\{K < 3\} = \sum_{k=0}^{k=2} p(k) = (1+4+8)\exp(-4) \approx 0{,}238$.

1.15. Среднее время безотказной работы системы $1/\lambda$ =100 часов. Поэтому вероятность того, что в течение времени t придется ремонтировать систему (вероятность того, что система откажет до момента t), $\mathbf{P}\{T = 10\} \approx 0{,}1$.

1.16. Указание: для доказательства равенства $\eta^2_{Y|x} = \rho^2_{XY}$ используйте условное среднее значение случайной величины Y.

Глава 2

2.1. Для дискретных случайных величин вероятности преобразуемой X и преобразованной Y случайных величин совпадают, если каждому Y_i соответствует одно значение X_i. Если одному какому-либо значению Y_j соответствует несколько значений X_r, $r = 1, 2, \ldots, k$, то $p(y_i) = \sum_{r=1}^{k} p(x_r)$.

2.2. Указание: для доказательства независимости преобразованных случайных величин $y_1 = \varphi_1(x_1)$ и $y_2 = \varphi_2(x_2)$ рассмотреть их двумерную плотность вероятности.

2.3.

$$f_Y(y) = \begin{cases} \dfrac{f_1[\psi_1(y)] + f_1[\psi_2(y)]}{\sqrt{a^2 - y^2}} & \text{при } -a \leqslant y \leqslant a, \\ 0 & \text{при } y < -a,\ y > a; \end{cases}$$

$$f_Y(y) = \begin{cases} 1/(\pi\sqrt{a^2 - y^2}) & \text{при } -a \leqslant y \leqslant a, \\ 0 & \text{при } y < -a,\ y > a. \end{cases}$$

2.4. $f_2(r, \alpha) = (r/2\pi\sigma^2)\exp(-r^2/2\sigma^2)$, $0 \leqslant r < \infty$, $0 \leqslant \alpha \leqslant 2\pi$; $f(r) = (r/\sigma^2)\exp(-r^2/2\sigma^2)$, $0 \leqslant r < \infty$; $f(\alpha) = 1/2\pi$, $0 \leqslant \alpha \leqslant 2\pi$.

2.5.

$$F_Y(y) = \begin{cases} 0 & \text{при } y < 0, \\ 0{,}5 & \text{при } y = 0, \\ \Phi(y) & \text{при } 0 < y < \infty. \end{cases}$$

2.7. Исходные данные соответствуют III типу распределений Пирсона: $f(x) = y_0 \exp(-\gamma x)(1 + x/a)^{\gamma a}$, $-a < x < \infty$ с началом отсчета от моды.

2.8. Указание: для доказательства учесть, что согласно (3.4.13) дисперсия выборочной дисперсии (3.4.11) определяется выражением $D(s_1^2) = (\mu_4 - \mu_2^2)/n$.

2.10. Вероятность того, что стьюдентская случайная величина по модулю будет больше $kD^{1/2} = 3{,}36$ ($D = \nu/(\nu - 2) = 1{,}25$), $\mathbf{P}\{X > k\sqrt{D}\} \approx 0{,}006$. Чебышевская верхняя граница равна 0,111.

2.11. Вероятность того, что случайная величина χ^2 будет больше $km = 16$, $\mathbf{P}\{X > km\} = 0{,}0424$. Чебышевская верхняя граница равна 0,5.

Глава 3

3.2. $f_R(r) = \dfrac{1}{\sqrt{2\pi\sigma_R^2}} \exp(-r^2/2\sigma_R^2)$, где $\sigma_R^2 = \dfrac{\pi\nu}{4n}\left(\dfrac{\Gamma(\nu/2)}{\Gamma[(\nu+1)/2]}\right)^2$.

3.4. Выборочные среднее и дисперсия группированной выборки равны $m^* \approx 7{,}3$ и $\mu_2^* \approx 0{,}211$. Поправка Шеппарда для дисперсии группированной оценки составляет величину $h^2/12 = 0{,}003$.

3.5. $D(s^2) = 2(4n - 3)/n(n - 1)\lambda^4$.

3.6. $D(\bar{x}) = 0{,}1$; $D(s^2) \approx 0{,}35$.

3.8. $m_Y = m_X^2 + D_X/n$; $D_Y = 4m_X^2 D_X/n$.

3.10. $y^* = 5{,}38x_1 + 0{,}84x_2$; $b_1^* = 5{,}38$; $b_2^* = 0{,}84$; $s^2(y) = 0{,}30$; $s^2(b_1) = 0{,}54$; $s^2(b_2) = 0{,}46$. Указание: для вычисления воспользоваться соотношениями (3.5.43), положив $x_0 \equiv x_2$, $b_0 \equiv b_2$.

Глава 4

4.2. Во-первых, оценка должна быть достаточной, т.е. функции правдоподобия представляется в виде произведения двух функций $L(\lambda^*, \mathbf{x}) = h(\mathbf{x})q(\lambda^*|\lambda)$, одна из которых $h(\mathbf{x})$ не зависит от оцениваемого параметра l, а вторая является условной функцией оценки λ^* и оцениваемого параметра λ: $q(\lambda^*|\lambda)$. Во-вторых, производная логарифма функции правдоподобия должна быть линейной функцией разности оценки и текущего значения оцениваемого параметра.

4.4. Если имеется независимая выборка объема n, образована квадратичная форма вида $\sum_{i=1}^n x_i^2$ и существует $g < n$ линейных функций от x_i, то исходную квадратичную форму путем ортогонального преобразования можно преобразовать в квадратичную форму $\sum_{i=1}^{n-g} y_i^2$, в которой отсутствуют линейные связи между случайными величинами y_i. В этом случае величину $\nu = n - g$ называют числом степеней свободы исходной квадратичной формы.

4.6. Функция правдоподобия параметра λ имеет вид $L(\lambda, \mathbf{x}) = \lambda^n \exp[-\lambda \times \times \sum_{i=1}^n x_i] = h(\mathbf{x})\lambda^n \exp[-\lambda \sum_{i=1}^n x_i]$, где $h(\mathbf{x}) = 1$, т.е. условие достаточности оценки (4.2.7) выполняется. Условие (4.2.12), необходимое для существования эффективной оценки, не выполняется, так как $\partial \ln L(\lambda, \mathbf{x})/\partial\lambda = (n/\lambda\lambda^*)(\lambda^* - \lambda)$. Нижняя граница дисперсии оценки равна $D(\lambda) > \lambda^2/n$.

4.7. $(\sigma^2)^* = (1/2n)\sum_{i=1}^n x_i^2$. Дисперсия оценки $D((\sigma^2)^*) = n/\sigma^4$. Аналогичная оценка получается на основе метода максимального правдоподобия. Из анализа функции правдоподобия следует выполнение условия существования эффективной оценки.

4.8. Истинное значение параметра λ заключено в интервале $0{,}066 \sum_{i=1}^n x_i \geqslant \geqslant \lambda \geqslant 0{,}029 \sum_{i=1}^n x_i$. Указание: воспользоваться нормальным приближением оценки параметра $\beta = 1/\lambda$.

Глава 5

5.5. $\sum_{i=1}^n (x_i - m)^2 < \chi_{0,05;25^2}\sigma_0^2 = 31{,}41\sigma_0^2$.

5.6. Нулевая гипотеза $\lambda = \lambda_0$ принимается, если для выборочного среднего выполняется неравенство $\bar{x} > \ln(\lambda_1/\lambda_0)/(\lambda_1 - \lambda_0)$.

5.7. Нулевые гипотеза о равенстве средних значений и о равенстве дисперсий не отвергаются. Указание: использовать критерии Стьюдента и дисперсионного отношения.

5.8. Гипотеза не отклоняется.

5.9. Гипотеза не отклоняется.

5.10. Нулевая гипотеза подтверждается.

5.11. Гипотеза не отклоняется.

5.12. Гипотеза не отклоняется.

5.13. Гипотеза не подтверждается.

5.14. Гипотеза не отклоняется.

Глава 6

6.2. Ранговый дисперсионный анализ следует применять в задачах, когда неизвестно распределение выборочных значений либо выборочные значения по уровням факторов представлены в виде рангов.

6.3. Нулевая гипотеза при однофакторном дисперсионном анализе соответствует тому, что уровни фактора не влияют на выходные (измеряемые) величины.

6.4. Двухфакторный дисперсионный анализ с многократными наблюдениями в ячейках позволяет проверить (оценить) эффект взаимодействия между уровнями исследуемых факторов.

6.5. Нулевая гипотеза отвергается.

6.6. Нулевая гипотеза подтверждается.

Глава 7

7.3. Нулевая гипотеза о том, что обе выборки относятся к одной генеральной совокупности, не отвергается. Указание: учесть нормализующее преобразование Фишера для выборочных коэффициентов корреляции и критерий проверки гипотезы о равенстве средних значений.

7.4. а) $r_{12.(3)} \approx 0{,}51$; $r_{13.(2)} \approx -0{,}49$; $r_{23.(1)} \approx 0{,}98$; $r_{1.(23)} = 0{,}26$;

б) $-0{,}50 \leqslant \rho_{12.(3)} \leqslant 0{,}35$.

в) Нулевая гипотеза не подтверждается.

7.6. Нулевая гипотеза не подтверждается.

7.7. $r_s = 0{,}222$; $r_s^* = 0{,}217$; $r_s^{**} = 0{,}218$.

7.8. Нулевая гипотеза отвергается.

Глава 8

8.3. Нелинейная регрессия $\tilde{y} = \eta x^\beta$ преобразуется в линейную путем введения новых переменных $z = \lg\tilde{y}$ и $u = \lg x$. Нелинейная регрессия $\tilde{y} = (\eta + \beta e^{-x})^{-1}$ преобразуется в линейную путем введения новых переменных $z = 1/\tilde{y}$ и $u = e^{-x}$.

8.5. $y^* = 4{,}0 + 1{,}66(x - 7{,}3)$. Значения доверительных интервалов в середине и на концах интервала возможных значений аргумента x: $y^* = 4{,}0 \pm 0{,}15$ для $x = 7{,}3$; $y^* = 2{,}51 \pm 0{,}23$ для $x = 6{,}4$; $y^* = 5{,}49 \pm 0{,}23$ для $x = 8{,}2$.

8.6. Гипотеза выборочной линейной зависимости не отвергается.

8.7. Ошибка прогноза на последнем шаге $e_n = x_n - y_n^* = 30 - 40 = -10$. Прогнозируемые значения ряда в моменты времени $n + 1$ и $n + 2$: $y_{n+1}^* = y_n^* + (1 - \lambda)e_n = 40 - (1 - 0{,}5) \cdot 10 = 35$; $y_{n+2}^* = y_n^* + 2(1 - \lambda)e_n = 40 - 10 = 30$.

Глава 9

9.1. Наиболее эффективным методом рандомизации является использование латинских квадратов, затем блочных факторных экспериментов (БФЭ) и, наконец, полных факторных экспериментов (ПФЭ).

9.3. Нулевая гипотеза не отвергается.

9.5. При построении плана дробного факторного эксперимента из множества факторов m отбирают $m - p$ факторов или их взаимодействий и для них строят планы полного факторного эксперимента.

9.6. При геометрической интерпретации точки плана статистической модели первого порядка для трех факторов (ПФЭ типа 23) расположены в вершинах куба.

9.7. На рис. 9.6.4,*а* прямолинейное движение центра симплекса обусловлено тем, что наименьшие значения откликов соответствуют последовательно точкам 1, 2, 3, 4, 5, 6, 7. На рис. 9.6.4,*б* движение центра симплекса вызвано тем, что наименьшие значения откликов последовательно соответствуют точкам 1, 3, 4, 2, 6, 5, 8, 7, 10.

9.8. Для того чтобы базовым фактором был x_1, необходимо увеличить интервал варьирования фактора $x_1^* \equiv t$ мин или уменьшить интервал варьирования фактора $(x_2^* \equiv T)$ C^o так, чтобы выполнялось условие $|\Delta x_1^* b_1^*| > |\Delta x_2^* b_2^*|$.

Предметный указатель

Адекватность модели 338, 394
Аксиомы теории вероятностей 7
Аппроксимация законов распределения 89

Вариационный ряд 114
Вектор-градиент 419
Вероятность гипотезы апостериорная 12, 221
– – априорная 12, 220
– произведения событий (формула умножения вероятностей) 8, 9
– события 6, 7
– – условная 8
– суммы событий (формула сложения вероятностей) 8, 9
Временной ряд 358
– – анализ 358
– – – методом скользящего среднего 364
– – – – экспоненциального сглаживания 365
– – компонента сезонная 361
– – – циклическая долговременная 361
– – прогнозирование 371
– – составляющая детерминированная 360
– – – случайная 360
– – стационарный 368
– – тренд 361
Выборка (выборочная совокупность) 113
– группированная одномерная 121
– – двумерная 125
– независимая 113
– однородная 113
– представительная 113
– репрезентативная 114
– случайная 114
Выборки объем 113
– размах (широта) 114
Выборочное среднее 127
– – квадратическое отклонение 140
Выборочный метод 113

Гамма-функция 39
– неполная 39
Генеральная совокупность 113
Генератор дробных планов 408
Гипотеза событий 11
– альтернативная 220
– – двухсторонняя 234
– – односторонняя 234
– двуальтернативная 220
– независимости 266
– непараметрическая 219
– нулевая 219
– однородности 259
– параметрическая 219
– простая 220
– сложная 220
– случайности 266
– согласия 247
– – простая 248
– – сложная 248
– статистическая 219
Гистограмма распределения 121
Граница Чернова 108
Группа событий полная 7

Дельта-функция 15
Дискретный белый шум 369
Дисперсионный анализ 277
– – однофакторный 277
– – двухфакторный 287
– – непараметрический 277, 297
– – – однофакторный 299
– – – двухфакторный 301
Дисперсия 21
– адекватности модели 339, 396
– выборочная 128
– косвенной оценки параметра 136–140
– оценки параметра распределения 155
– ошибки наблюдения 339, 396
– оценки параметра условная 155
– условная 53

– эффективной оценки параметра 163, 164
– – – векторного параметра 164
– – – двух параметров 164
Доверительная вероятность 153, 196
Доверительные границы 154, 195
Доверительный интервал (ширина) 153, 196
– – центральный 197

Закон больших чисел 100, 102, 103
– распределения вероятностей 13
– – – многомерный 46
Звездные точки плана 411

Единичная функция 118

Измерения прямые 135
Инверсия 264, 272
Информация по Фишеру 162
Интервальное оценивание параметра распределения 153, 195
– – – – байесовское 197
– – – – биномиального 207
– – – – нормального 200–206
– – – – пуассоновского 209

Квантиль 19
Ковариационная матрица 65
Композиция законов распределения 18
Контраст плана 409
Коррелограмма 369
Корреляционная матрица 65
– – выборочная 315
– таблица 126
Корреляционное отношение 56
– – выборочное 318
Корреляционный (ковариационный) анализ 306
Косвенная оценка параметра 135
– – – линейная 136
– – – нелинейная 137
– – параметров (коэффициентов) уравнения 143
Коэффициент асимметрии 21
– вариации 142
– – выборочный 142
– детерминации выборочный 341
– корреляции (парный) 51
– – выборочный 309
– – – группированный 317
– – множественный 66
– – – выборочный 316
– – ранговый Кендалла 326
– – – Спирмена 321, 323
– – частный 65
– – – выборочный 316
– согласованности (конкордации) 328
– эксцесса 21
Критерий проверки гипотез 220
– – – аномальных наблюдений 245
– – – байесовский (оптимальный) 223
– – – Бартлета 244
– – – дисперсионного отношения (F-критерий) 242
– – – идеального наблюдателя 225
– – – максимального правдоподобия 225
– – – максимальной апостериорной вероятности 224
– – – наиболее мощный 233
– – – – – несмещенный 233
– – – – – равномерный (РНМ) 234
– – – – – – несмещенный 235
– – – Крускала–Уоллиса 300
– – – Неймана–Пирсона 225
– – – несмещенный 235
Критерий проверки гипотез однородности 259
– – – – Вилкоксона 263, 264
– – – – ранговый 265
– – – – Смирнова 259
– – – – хи-квадрат 260
Критерий проверки гипотез согласия 248
– – – – Колмогорова 253
– – – – Смирнова–Мизеса 257
– – – – хи-квадрат (Пирсона) 248
– – – Кендалла 327
– – – ранговый 298
– – – Спирмена 324
– – – Стьюдента (t-критерий) 239
– – – Фридмана 303, 330
– ортогональности плана 397
– ротатабельности плана 397
– достаточности оценки 157
– эффективности оценки 160
Критические значения распределений 36
Кумулянты (семиварианты) 23

Латинский квадрат 385

Математическая модель (регрессия) 377, 390
– – аналитическая (теоретическая) 391
– – статистическая (имитационная) 391

– – – квадратическая 411
– – – первого порядка 398
– – – линейная 402
– – – – первого порядка 398
– – – полиномиальная 398
Математическая статистика 111, 112
Математическое ожидание 17
Медиана 19
– выборочная 114
Метод наименьших квадратов 143
– – – обобщенный 366, 373
Многоугольник распределения 13
Мода 19
Момент распределения 20
– – начальный 20
– – – смешанный 79
– – – двумерный 50
– – – выборочный 126
– – – – группированный 133,
– – многомерный начальный 64
– – – центральный 64
– – центральный 20
– – – двумерный 50
– – – выборочный 126
– – – – группированный 133
– – условный 53
– корреляционный (ковариация) 50
– – – выборочный 307
– – – – группированный 135, 308
Мощность критерия 221

Невязка 143
Неравенство Буняковского–Шварца 51
– Рао–Крамера 159
– Чебышева 101

Область критическая (отклонения нулевой гипотезы) 220
– принятия нулевой гипотезы 220
Ортогональная система функций 62, 90
Отношение правдоподобия 223
– – логарифм 226
Оценка параметра распределения 150
– – – байесовская 167
– – – достаточная 157
– – – максимальной апостериорной вероятности 170
– – – максимального правдоподобия 170, 176
– – – минимаксная 168
– – – несмещенная 120, 156
– – – – асимптотически 156
– – – состоятельная 156
– – – точечная 154
– – – эффективная 157
– – – – асимптотически 157, 161
– – – объединенная линейная 193
– – – методом моментов 174
– – – – минимума хи-квадрат 178
– – – – – видоизмененного 178
Оценки параметров распределения совместно-эффективные 163
Ошибка второго рода 221
– первого рода 221
– оценки средняя квадратическая 157

План факторного эксперимента блочный (БФЭ) 383
– – – дробный(ДФЭ) 407
– – – ненасыщенный 393
– – – полный (ПФЭ) 383, 398
– – – классический (КФЭ) 381
– квадратичной модели 411
– латинского квадрата 385
– неполноблочный сбалансированный 388
– ортогональный 397, 399
– – линейный 402, 407
– ротатабельный 397
– центральный композиционный (ЦКП) 411
– – – ортогональный (ЦКОП) 412, 413
– – – ротатабельный (ЦКРП) 412, 416
– экстремального эксперимента 417
– – – метод крутого восхождения (Бокса–Уилсона) 420
– – – – покоординатного движения (метод Гаусса–Зейделя) 419
– – – симплексный 423
– – – эволюционный 426
Плотность вероятности 14
– – апостериорная, априорная 152
– – выборочной медианы 116
– – – – асимптотическая 117
– – двумерная (совместная) 47
– – максимального значения выборки 115
– – минимального значения выборки 115
– – многомерная 63
– – оценки параметра условная 154

– – полимодальная 19
– – порядковой статистики 116
– – размаха выборки 117
– – унимодальная 19
– – условная 49
Полигон частот 124
Полиномы ортогональные 62, 91
– Лаггера 94
– Эрмита 63, 91
Поправки Шеппарда 133
Порядковая (ранговая) статистика 114
Порядковый критерий проверки гипотез 265
– – – – знаков 267
– – – – инверсий 272
– – – – серий 269
Предельные соотношения (теоремы) теории вероятностей 100, 103, 104
Преобразование нелинейной регрессии в линейную 353
– ортогональное 211
– случайных величин многомерных 75
– – – одномерных 71
– квадратичное 73
– линейное 75
– моментов 72, 79
– нормализующее Фишера 313
Проверка гипотезы адекватности модели (регрессии) 339, 394
– – значимости коэффициентов модели 354, 394
– – об аномальных наблюдениях 245
– – о равенстве вероятностей событий 246
– – – – дисперсий 240–244
– – – – средних значений 238–241

Ранг элемента выборки (наблюдения) 114, 298, 320
Рандомизация эксперимента 382
Распределение Бернулли (бинарное) 23
– биномиальное 25
– вероятности двумерное 46
– гамма 39
– дисперсионного отношения (F-распределение, распределение Фишера) 88
– Коши 85
– Лапласа 70
– логнормальное 44
– Максвелла 83
– нормальное 28, 35
– – двумерное 67
– – многомерное 69
– – условное 67
– Парето 70
– прямоугольное (равномерное) 41
– Пирсона (кривые Пирсона) 96
– Пуассона 31
– равномерное (дискретное) 34
– Релея 82
– Стьюдента (t-распределение) 84
– хи-квадрат 80
– экспоненциальное 40
– Эрланга 39
Рассеяние оценки дисперсии 131
– – параметра распределения условное 155
Регрессионный анализ 334
Регрессия первого рода 53
– второго рода 53, 59
– средняя квадратическая 59
– – – линейная 59
– – – нелинейная 61
– выборочная средняя квадратическая 54, 336, 392
– – – – линейная парная 342
– – – – – множественная 349
Риск оценки параметров 166
– – – апостериорный 168
– – – байесовский (безусловный) 167
– – – минимаксный 168
– – – условный 166
Риск проверки гипотез средний 222
– – – байесовский 223
Ряд Эджворта 93

Случайные величины 13
– – гауссовские 35, 36
– – двумерные 46
– – дискретные 13
– – многомерные (векторные) 46, 63
– – независимые 49
– – некоррелированные 51, 52
– – непрерывные 14
– – непрерывно-дискретные (смешанные) 16
– – ортогональные 51
Смещение оценки параметра условное 155
Событие 5
– достоверное 7

– невозможное 7
– независимое 8
– несовместное 7
– противоположное 7
– случайное 7
Среднее значение 19
– – условное 53
– квадратическое отклонение 21
Статистика 112, 277
– достаточная 157, 158
– ранговая (порядковое) 114
Степени свободы 81, 213, 215

Теорема Бернулли 29, 102
– об арифметическом среднем 103
– центральная предельная 104
Толерантные границы; интервалы 215
– доверительные пределы 216

Уравнение правдоподобия 176
Уравнения нормальные (условные) 145
Уровни факторов 276, 376
– – нормированные (коды) 396
Уровень значимости критерия 221

Факторный анализ 276
– – однофакторный 276
– – многофакторный 276
Факторы 276, 334
– второстепенные (мешающие) 388
– нормированные (коды) 396
– существенные 374, 377
Формула Муавра–Лапласа локальная 28
– Лапласа интегральная 29
– обратной вероятности событий (формула Байеса) 12
– полной вероятности событий 12
– Стирлинга 20
Функция мощности критерия 235
– отклика (выходной эффект) 334, 377, 392
– потерь 165, 222
– – квадратичная 165, 171
– – «линейная» по модулю 165, 170
– – простая 165, 169
– – прямоугольная 166, 172
– – экспоненциальная 166
– правдоподобия 152
Функция распределения 14
– – выборочная (эмпирическая) 118
– – – несмещенная 120
– – двумерная (совместная) 47
– – Колмогорова 254
– – максимального значения выборки 115
– – минимального значения выборки 115
– – размаха выборки 117
– – многомерная 64
– – условная 49
– характеристическая 17
Характеристики оценки параметра безусловные 155
– числовые 19

Частота событий 24, 102

Эксперимент (опыт) 376
– планирование 376, 378
– цели 377
– этапы постановки 377
Элементы (значения) выборки 113
Эффективность оценки относительная 161
– – – асимптотическая 161

Якобиан преобразования 76

Оглавление

Предисловие ... 3

Часть I. Основные положения теории вероятностей ... 5

Глава 1. События, случайные величины и их описание ... 5

1.1. События и вероятности событий ... 5
1.1.1. Определения ... 5
1.1.2. Вероятность произведения событий (теорема умножения вероятностей) ... 9
1.1.3. Вероятность суммы событий (теорема сложения вероятностей) 9
1.1.4. Формулы полной и обратной вероятности (формула Байеса) . 11
1.2. Случайные величины и их описание ... 13
1.2.1. Функция распределения, плотность вероятности и характеристическая функция ... 13
1.2.2. Числовые характеристики случайных величин ... 18
1.3. Законы распределения дискретных случайных величин ... 23
1.3.1. Закон распределения Бернулли ... 23
1.3.2. Биномиальный закон распределения ... 24
1.3.3. Распределение Пуассона ... 30
1.3.4. Равномерный закон распределения ... 34
1.4. Распределение непрерывных случайных величин ... 35
1.4.1. Нормальный закон распределения ... 35
1.4.2. Гамма-распределение ... 39
1.4.3. Прямоугольное (равномерное) распределение ... 41
1.4.4. Логарифмически нормальное (логнормальное) распределение 44
1.5. Совместные распределения вероятностей нескольких случайных величин ... 46
1.5.1. Двумерные случайные величины ... 46
1.5.2. Условные моменты. Регрессия первого рода ... 53
1.5.3. Регрессия второго рода ... 59
1.5.3.1. Линейная средняя квадратическая регрессия ... 59
1.5.3.2. Нелинейная средняя квадратическая регрессия ... 61
1.5.4. Многомерные случайные величины ... 63
1.5.5. Двумерное и многомерное нормальное распределение ... 67

Глава 2. Функциональные преобразования случайных величин ... 71

2.1. Преобразования плотностей вероятности ... 71
2.1.1. Преобразование одномерных плотностей вероятности ... 71
2.1.2. Преобразование многомерных плотностей вероятности ... 75
2.2. Распределения, порожденные гауссовскими случайными величинами 79
2.2.1. Хи-квадрат (χ^2) распределение ... 79
2.2.2. Распределение Стьюдента ... 83
2.2.3. Распределение Коши ... 85
2.2.4. Распределение дисперсионного отношения ... 87
2.3. Аппроксимация законов распределения ... 89
2.3.1. Аппроксимация одномерной плотности вероятности рядом по ортогональным полиномам ... 90

2.3.1.1. Аппроксимация распределений с использованием ортогональных полиномов Эрмита ... 91
2.3.1.2. Аппроксимация распределений с использованием ортогональных полиномов Лагерра ... 94
2.3.2. Распределения (кривые) Пирсона ... 96
2.4. Предельные соотношения (теоремы) теории вероятностей ... 100

Часть II. Математическая статистика ... 111

Глава 3. Выборочный метод ... 111

3.1. Основные определения и задачи математической статистики ... 111
3.2. Эмпирическая функция распределения ... 118
3.3. Группированная выборка ... 120
3.3.1. Одномерная группированная выборка ... 121
3.3.2. Двумерная группированная выборка ... 125
3.4. Выборочные моменты ... 126
3.4.1. Выборочные начальные моменты ... 127
3.4.2. Выборочная дисперсия ... 128
3.4.3. Группированные оценки моментов ... 133
3.5. Косвенные оценки параметров ... 135
3.5.1. Функции от выборочных моментов ... 135
3.5.2. Линейные косвенные оценки ... 136
3.5.3. Нелинейные косвенные оценки ... 137
3.5.4. Примеры нелинейных косвенных оценок параметров ... 140
3.5.4.1. Выборочное среднее квадратическое отклонение ... 140
3.5.4.2. Выборочный коэффициент вариации (изменчивости) .. 142
3.5.5. Косвенные оценки параметров, определяемые из системы уравнений. Метод наименьших квадратов ... 142

Глава 4. Статистические оценки параметров распределений ... 150

4.1. Сущность задачи оценки параметров распределения ... 150
4.2. Точечные оценки параметров и их свойства ... 154
4.3. Основные положения теории байесовских (оптимальных) оценок ... 165
4.3.1. Определения ... 165
4.3.2. Простая функция потерь ... 169
4.3.3. Линейная по модулю функция потерь ... 170
4.3.4. Квадратичная функция потерь ... 171
4.3.5. Прямоугольная функция потерь ... 172
4.4. Прикладные методы получения точечных оценок параметров распределения ... 174
4.4.1. Метод моментов ... 174
4.4.2. Метод максимального правдоподобия ... 175
4.4.3. Метод минимума хи-квадрат (χ^2) ... 177
4.5. Примеры нахождения оценок параметров распределений ... 179
4.5.1. Оценка среднего значения гауссовской случайной величины при известной дисперсии ... 179
4.5.2. Оценка дисперсии гауссовской случайной величины ... 181
4.5.3. Совместная оценка среднего значения и дисперсии гауссовской случайной величины ... 182
4.5.4. Оценка параметров биномиального и пуассоновского распределений ... 184
4.5.4.1. Биномиальное распределение ... 184
4.5.4.2. Распределение Пуассона ... 185

4.5.5. Оценка параметров прямоугольного распределения ... 186
4.5.5.1. Оценка параметра a прямоугольного распределения ... 186
4.5.5.2. Оценка параметров a и c прямоугольного распределения 188
4.6. Объединение нескольких оценок одного и того же параметра ... 191
4.6.1. Объединение оценок при известных условных плотностях вероятности ... 192
4.6.2. Линейное объединение оценок ... 193
4.7. Интервальное оценивание ... 195
4.7.1. Байесовские интервальные оценки параметров ... 197
4.7.2. Интервальное оценивание при неизвестном априорном распределении параметра ... 199
4.8. Примеры интервального оценивания параметров распределений ... 200
4.8.1. Доверительный интервал для среднего значения нормального распределения ... 200
4.8.1.1. Известная дисперсия ... 200
4.8.1.2. Неизвестная дисперсия ... 200
4.8.2. Доверительный интервал для дисперсии нормального распределения ... 204
4.8.2.1. Известное среднее значение m ... 204
4.8.2.2. Неизвестное среднее значение m ... 205
4.8.3. Доверительный интервал для вероятности p биномиального закона распределения ... 207
4.8.4. Доверительный интервал для параметра λ пуассоновского распределения ... 209
4.9. Степени свободы распределений случайных величин ... 211
4.9.1. Основные положения ортогонального преобразования ... 211
4.9.2. Степени свободы ... 213
4.10. Толерантные (допустимые) интервалы ... 215
Глава 5. Проверка статистических гипотез ... 219
5.1. Проверка простых гипотез ... 220
5.2. Проверка сложных гипотез ... 233
5.3. Прикладные задачи проверки параметрических гипотез ... 238
5.3.1. Проверка гипотез о средних значениях выборок ... 238
5.3.1.1. Гипотеза о равенстве выборочного среднего заданному значению ... 238
5.3.1.2. Проверка гипотезы о равенстве средних значений ... 239
5.3.2. Проверка гипотез о равенстве дисперсий ... 242
5.3.3. Проверка гипотезы об аномальных наблюдениях ... 245
5.3.4. Проверка гипотезы о равенстве вероятностей событий ... 246
5.4. Гипотезы о виде распределений (гипотезы согласия) ... 247
5.4.1. Критерий согласия хи-квадрат (критерий Пирсона) ... 248
5.4.2. Критерий согласия Колмогорова ... 253
5.4.3. Критерий Смирнова–Мизеса (критерий ω^2) ... 257
5.5. Гипотезы однородности распределений ... 259
5.5.1. Критерий однородности Смирнова ... 259
5.5.2. Критерий однородности хи-квадрат ... 260
5.5.3. Критерий Вилкоксона ... 263
5.6. Порядковые критерии проверки независимости и случайности выборок ... 265
5.6.1. Критерий знаков ... 266
5.6.2. Критерий серий ... 269

5.6.3. Критерий инверсий ... 272
Глава 6. Дисперсионный анализ ... 276
6.1. Однофакторный дисперсионный анализ ... 277
6.2. Двухфакторный дисперсионный анализ ... 287
6.2.1. Двухфакторный анализ с одним наблюдением в «ячейке» («клетке») ... 287
6.2.2. Двухфакторный анализ с многократными наблюдениями в «ячейке» («клетке») ... 292
6.3. Непараметрический дисперсионный анализ ... 297
6.3.1. Однофакторный непараметрический дисперсионный анализ ... 299
6.3.2. Двухфакторный непараметрический дисперсионный анализ ... 301
Глава 7. Корреляционные зависимости ... 305
7.1. Корреляционный анализ (корреляция) ... 306
7.1.1. Выборочный корреляционный момент ... 306
7.1.2. Выборочный коэффициент корреляции ... 309
7.1.3. Выборочное корреляционное отношение ... 317
7.2. Ранговые методы исследования статистической связи ... 320
7.2.1. Коэффициент ранговой корреляции ... 321
7.2.2. Коэффициент ранговой согласованности ... 327
Глава 8. Регрессионный анализ ... 334
8.1. Основные положения ... 334
8.2. Линейная средняя квадратическая выборочная регрессия ... 342
8.2.1. Парная линейная выборочная регрессия ... 342
8.2.2. Множественная линейная выборочная регрессия ... 349
8.3. Преобразование нелинейной регрессии в линейную регрессию ... 352
8.4. Проверка гипотез, связанных с выборочной линейной регрессией ... 354
8.5. Временные ряды ... 357
Глава 9. Основные понятия теории эксперимента ... 376
9.1. Сущность постановки задачи ... 376
9.2. Способы рандомизации ... 382
9.3. Статистическая модель изучаемого процесса (явления) ... 390
9.4. Статистическая модель первого порядка ... 398
9.4.1. План полного факторного эксперимента ... 398
9.4.2. План дробного факторного эксперимента ... 407
9.5. Планы для квадратичных моделей ... 410
9.5.1. Центральный композиционный ортогональный план ... 412
9.5.2. Центральный композиционный ротатабельный план ... 416
9.6. Планы экстремальных экспериментов ... 417
9.6.1. Метод крутого восхождения (метод Бокса–Уилсона) ... 418
9.6.2. Симплексный метод ... 423
9.6.3. Метод эволюционного планирования ... 425
Литература ... 428
Приложение 1. Таблицы распределений ... 431
Приложение 2. Статистические таблицы ... 443
Ответы к контрольным вопросам и задачам ... 451
Предметный указатель ... 455

Вышли в свет и имеются в продаже:

Методы обработки многомерных данных и временных рядов: Учебное пособие для вузов / А. А. Большаков, Р. Н. Каримов . – М.: Горячая линия–Телеком, 2007. – 522 с.: ил., ISBN 5-93517-287-9.

Рассмотрены основные методы обработки многомерных экспериментальных данных объектов числовой и нечисловой природы, разведочный анализ и представление данных. Приведено систематическое описание следующих методов многомерной статистической обработки: анализ главных компонент; каноническая корреляция; дискретно-косинусное преобразование и вейвлет-анализ; дискриминантный и факторный анализы; а также анализ соответствий и многомерное метрическое и неметрическое шкалирование. Изложены современные методы сингулярного разложения и вейвлет-анализа, используемые для обработки многокомпонентных временных рядов. Пособие иллюстрировано тщательно подобранными примерами, в том числе взятыми авторами из многолетней собственной практики решения реальных задач.

Для студентов высших учебных заведений, обучающихся по направлению подготовки 230100 (654600) – «Информатика и вычислительная техника», магистерской программе 550209 – «Автоматизация научных исследований, испытаний и эксперимента» направления 550200 – «Автоматизация и управление», будет полезна преподавателям и аспирантам.

Нейронные сети. STATISTICA Neural Networks: Методология и технологии современного анализа данных / Под ред. В. П. Боровикова. – 2-е изд., перераб. и доп. – М.: Горячая линия – Телеком, 2008. – 392 с., ил., ISBN 978-5-9912-0015-8.

Изложены нейросетевые методы анализа данных, основанные на использовании пакета STATISTICA Neural Networks (фирма производитель StatSoft), полностью адаптированного для русского пользователя. Даны основы теории нейронных сетей; большое внимание уделено решению практических задач, всесторонне рассмотрена методология и технология проведения исследований с помощью пакета STATISTICA Neural Networks – мощного инструмента анализа и прогнозирования данных, имеющего широкие применения в бизнесе, промышленности, управлении, финансах. Книга содержит множество примеров анализа данных, практические рекомендации по проведению анализа, прогнозирования, классификации, распознавания образов, управления производственными процессами с помощью нейронных сетей.

Для широкого круга читателей, занимающихся исследованиями в банковской сфере, промышленности, экономике, бизнесе, геологоразведке, управлении, транспорте и других областях.

Статистический анализ и синтез радиотехнических устройств и систем: Учебное пособие для вузов / В. И. Тихонов, В. Н. Харисов. – М.: Радио и связь, Горячая линия–Телеком, 2004. – 608 с.: ил., ISBN 5-256-01701-2.

Приводятся необходимые сведения из теории вероятностей и на их основе рассматриваются статистические методы анализа линейных систем и нелинейных радиотехнических устройств. На единой базе теории фильтрации экономно и единообразно изложены современные методы синтеза аналоговых и цифровых радиотехнических систем различного назначения, включая и адаптивные. Методика применения теоретических результатов к решению практических задач проиллюстрирована содержательными примерами.

Для студентов радиотехнических специальностей вузов, полезна инженерам и аспирантам.

Эконометрика. Решение задач с применением пакета программ GRETL / Куфель Тадеуш. Пер. с польск. И. Д. Рудинского . – М.: Горячая линия–Телеком, 2007. – 200 с.: ил., ISBN 5-93517-307-7.

Рассмотрены методы решения основных эконометрических задач с использованием пакета программ GRETL (GNU Regression Econometrics Time-series Library), предназначенного для практической реализации сложных вычислительных процедур эконометрического моделирования. Пакет программ GRETL и представленные в работе статистические данные доступны на интернет-сайте автора http://www.kufel.torun.pl.

Для студентов, аспирантов и преподавателей экономических вузов, а также для научных работников, ведущих экономические исследования.

Справки по телефону: (495) 737-39-27, WWW.TECHBOOK.RU