МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

В. П. Шуленин

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Часть 1

ПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

Допущено УМО по классическому университетскому образованию в качестве учебника для студентов высших учебных заведений, обучающихся по направлениям ВПО 010400 «Прикладная математика и информатика» и 010300 «Фундаментальная информатика и информационные технологии»

ТОМСК
«Издательство НТЛ»
2012

УДК 519.2
Ш955

Шуленин В. П. Математическая статистика. Ч. 1. Параметрическая статистика: учебник. – Томск: Изд-во НТЛ, 2012. – 540 с.

Ш955

ISBN 978-5-89503-492-7

В учебнике в доступной форме изложены для первоначального изучения основные понятия и методы трех современных разделов математической статистики. В первой части – это классические результаты теории оценивания параметров и проверки статистических гипотез в рамках параметрических моделей. Во второй – описание статистических процедур, которые гарантируют качество принимаемых решений в рамках непараметрических моделей (при неизвестном функциональном характере распределения наблюдений). Эти процедуры основаны на использовании порядковых статистик, рангов и знаков наблюдений. Третья часть содержит описание статистических процедур, которые обладают устойчивостью (робастны) к отклонениям от исходных предпосылок в принятой статистической модели. Основные разделы сопровождаются задачами, упражнениями и большим числом примеров, которые иллюстрируют, а в ряде случаев и дополняют излагаемые результаты. Задачи, упражнения и дополнения, которые приводятся в конце каждой главы, могут служить материалом для практических занятий, а также для заданий по курсовым и дипломным работам.

Предназначен студентам и аспирантам вузов, научным работникам, а также может быть полезен преподавателям при разработке курсов лекций для магистрантов и аспирантов на факультетах прикладной математики и кибернетики.

УДК 519.2

Рецензенты:

Г. А. Медведев, доктор физико-мат. наук, профессор кафедры теории вероятностей и математической статистики Белорусского государственного университета;

В. А. Кочегуров, доктор техн. наук, профессор Томского политехнического университета

ISBN 978-5-89503-492-7

© В. П. Шуленин, 2012
© Оформление. Дизайн.
ООО «Издательство НТЛ», 2012

ПРЕДИСЛОВИЕ

Развитие научных исследований в настоящее время характеризуется всё более возрастающей ролью статистических процедур, их широким использованием в различных областях практики и требованиями качественной обработки экспериментальных данных. Однако их успешное применение сдерживается тем, что классические процедуры математической статистики обеспечивают заданное качество решений лишь при справедливости некоторых предположений о статистической модели, которая принята в рамках проводимого эксперимента. На практике часто возникают отклонения характеристик реальных данных эксперимента от предполагаемых в модели, и применение стандартных процедур в этих условиях может оказаться малоэффективным и даже может привести к существенным искажениям статистических выводов. Многие стандартные процедуры нормальной теории гарантируют качество решений лишь в рамках гауссовской модели наблюдений. В связи с этим возникает необходимость построения новых нетрадиционных методов обработки информации, «свободных от распределения» и устойчивых (или робастных) к возможным отклонениям характеристик реальных данных от предполагаемых в принятой модели.

Современное состояние и развитие математической статистики характеризуется тремя направлениями: параметрическая, непараметрическая и робастная статистика. Эти направления различают по типам рассматриваемых статистических моделей, которые отражают уровень и полноту априорной неопределенности в статистическом описании изучаемого случайного явления. Классические методы параметрической статистики используют в явном виде известный функциональный характер распределения вероятностей наблюдений. Многие из этих методов разработаны для гауссовской модели наблюдений, которая не всегда является адекватной моделью при решении практических задач. В таких случаях целесообразнее использовать методы непараметрической и робастной статистики, которые бурно развиваются и появились как альтернатива

классическим методам, в большинстве своем обладающим повышенной чувствительностью к исходным предпосылкам в статистических моделях.

Предлагаемое учебное издание состоит из трех частей. В первой части рассматриваются два основных раздела, связанных с задачами построения точечных и интервальных оценок параметров, а также проверки статистических гипотез в рамках параметрических моделей. Вторая часть посвящена изложению методов и отдельных результатов непараметрической статистики. Она содержит описание статистических процедур, «свободных от распределения», которые гарантируют качество принимаемых решений в рамках непараметрических моделей (при неизвестном функциональном характере распределения наблюдений). Эти процедуры основаны на использовании порядковых статистик, рангов и знаков наблюдений. Третья часть содержит описание статистических процедур, которые обладают устойчивостью (робастны) к отклонениям от исходных предпосылок в принятой статистической модели. Основные разделы сопровождаются задачами, упражнениями, библиографическими замечаниями и большим числом примеров, которые иллюстрируют, а в ряде случаев и дополняют излагаемые результаты. Приведены приложения, которые содержат некоторые неравенства, основные распределения вероятностей и их таблицы, общие теоремы и необходимые сведения из теории вероятностей и математической статистики, на которые имеются ссылки в основном тексте. Отметим, что в книге принята сплошная нумерация формул, определений, теорем, примеров и замечаний в пределах данного раздела, первая цифра которого отражает номер главы. Например, ссылка (3.4.1) соответствует первой формуле четвёртого раздела третьей главы. Ссылка (П.1.2) соответствует второй формуле в первом приложении. Для рисунков и таблиц используется самостоятельная нумерация в пределах раздела.

При изложении общих результатов и понятий использовались работы различных авторов. Все заимствованные результаты снабжены ссылками в тексте. Наряду с обобщением и систематизацией известных результатов, разбросанных по многочисленным публикациям, в книге приведены и новые результаты, ранее не публиковавшиеся в учебной литературе. Часть результатов автор использо-

вал при чтении курса лекций студентам факультета прикладной математики и кибернетики Томского университета.

Предназначается студентам и аспирантам вузов, научным работникам в области математической и прикладной статистики. Автор выражает надежду, что данное учебное издание принесет пользу широкому кругу специалистов, связанных с обработкой реальных статистических данных экспериментов и послужит критическому и осознанному применению непараметрических и робастных статистических процедур, которые описаны в специальных монографиях.

Выражаю благодарность моим студентам и дипломникам за помощь в оформлении рукописи, а также участникам совместного семинара лаборатории статистических методов СФТИ и кафедры теоретической кибернетики ФПМК Томского университета за доброжелательную и конструктивную критику. Благодарю рецензентов профессора Г.А. Медведева, профессора В.А. Кочегурова, анонимного рецензента УМО при МГУ, которые внимательно прочитали рукопись книги и сделали целый ряд полезных критических замечаний, которые были учтены при редакторской обработке рукописи. Особая благодарность моему научному руководителю профессору Ф.П. Тарасенко, с которым меня связывает многолетняя научная деятельность.

Автор заранее выражает свою признательность за любые замечания со стороны читателей книги. Их можно направить по адресу E-mail: shvp@fpmk.tsu.ru

В. Шуленин

Глава 1

ОСНОВНЫЕ ПОНЯТИЯ, ОПРЕДЕЛЕНИЯ И ОБОЗНАЧЕНИЯ

В этой общей главе излагаются основные понятия и элементы выборочной теории. В разд. 1.1 обсуждается суть задач математической статистики, конкретизируются понятия «выборка» и «статистическая модель». Типы статистических моделей приводятся в разд. 1.2 и обсуждается понятие «супермодель» для описания возможных отклонений от принятой модели. Основные формулы и понятия, используемые при изучении случайных величин, приводятся в разд. 1.3. Упражнения, задачи и дополнения приведены в разд. 1.4.

Теория вероятностей и математическая статистика – это самостоятельные и тесно связанные разделы математики, в которых разрабатываются методы изучения случайных явлений. При этом используют такие понятия, как *случайное событие* и *случайная величина*. Случайное событие изучают с помощью вероятности осуществления этого события в условиях опыта, а случайную величину характеризуют с помощью функции распределения вероятностей. В теории вероятностей разработаны правила, которые позволяют вычислять вероятности некоторых «сложных событий» через *известные* вероятности «простых событий». В отличие от теории вероятностей математическая статистика решает свои специфические задачи своими методами, которые основаны на использовании результатов экспериментов и методов теории вероятностей. Современное состояние и развитие математической статистики характеризуется тремя самостоятельными направлениями:

- параметрическая (классическая) статистика;
- непараметрическая статистика;
- робастная статистика.

Эти направления различают по типам рассматриваемых статистических моделей, которые отражают уровень и полноту априорной неопределенности в статистическом описании изучаемого случайного явления (см., например, [59]). Для того чтобы подчеркнуть специфику задач математической статистики и их отличие от задач теории вероятностей, конкретизируем понятие статистической модели, статистических данных и уточним, что понимается под вероятностным (или статистическим) описанием случайного явления. Для этого введем необходимые понятия и обозначения.

1.1. Понятие выборки и статистической модели, суть статистических задач

При изучении случайных явлений основным объектом изучения является случайная величина (с.в.) X, которая может быть одномерной или многомерной и может быть непрерывного, дискретного или смешанного типа. Следуя принятой традиции в научной литературе по статистике, всюду ниже будем обозначать случайные величины прописными латинскими буквами. Возможные состояния (или реализации) случайных величин будем обозначать соответствующими строчными буквами. Например, X – случайная величина, а ее возможное состояние (или реализацию) будем обозначать через x, или $\vec{X} = (X_1,...,X_n)$ – случайный вектор, а его реализация – $\vec{x} = (x_1,...,x_n)$. Для того чтобы охарактеризовать случайную величину, необходимо задать распределение вероятностей по ее возможным значениям. Для осуществления этого используется понятие функции распределения (ф.р.) вероятностей, которое является универсальным инструментом, пригодным для изучения любой случайной величины, одномерной или многомерной, и непрерывного, дискретного или смешанного типа. Для случайной величины X ее функцию распределения вероятностей везде ниже будем обозначать в виде F_X. Если X – одномерная случайная величина непрерывного типа с бесконечным числом возможных значений $\{x\}$ на действительной оси $R^1 = \{x : -\infty < x < +\infty\}$, то она характеризуется одномерной функцией распределения вероятностей, которая опре-

деляется в виде $F_X(x) = P(X \le x), x \in R^1$. Иногда удобнее описывать случайную величину X одномерной плотностью распределения вероятностей $f_X(x) = F'_X(x)$, $x \in R^1$. Для описания случайного вектора $\vec{X} = (X_1,...,X_n)$ используют функцию n переменных $x_1,...,x_n$, которая в точке $(x_1,...,x_n) \in R^n$, где R^n обозначает n-мерное евклидовое пространство, определяется с помощью вероятности совместного осуществления событий в фигурных скобках, то есть функция распределения случайного вектора $\vec{X} = (X_1,...,X_n)$ определяется в виде

$$F_{X_1,...,X_n}(x_1,...,x_n) = F_{\vec{X}}(x_1,...,x_n) = P\{X_1 \le x_1;...;X_n \le x_n\},$$
$$(x_1,...,x_n) \in R^n.$$

Итак, для того чтобы охарактеризовать случайную величину X, необходимо задать ее функцию распределения вероятностей F_X. Зная эту функцию, можно вычислить методами теории вероятностей, по крайней мере в принципе, вероятности любых событий, связанных с поведением случайной величины X. Однако при решении реальных задач на практике функция распределения вероятностей F_X обычно полностью или частично неизвестна. В связи с этим и возникают статистические задачи, которые по своей сути направлены на устранение этой «неизвестности» на основе информации, содержащейся в результатах опыта, полученных путем наблюдений над изучаемой случайной величиной X.

Отметим, что в реальных задачах степень «неизвестности» относительно ф.р. F_X может быть самой различной, она уточняется при формулировке статистической модели, которая играет важную роль при решении любой статистической задачи. Статистическую модель задают в виде пары множеств $(\mathrm{X}, \Im)$, где X – выборочное пространство, а $\Im$ – множество допустимых функций распределения вероятностей в условиях данного опыта, результатом которого является *наблюденная* реализация $\vec{x} = (x_1,...,x_n)$ *наблюдаемого* случайного вектора $\vec{X} = (X_1,...,X_n)$. Отметим, что множество $\Im$ задается на основе тех априорных сведений о распределении вероятно-

стей, которыми располагает экспериментатор до проведения опыта. Выборочное пространство X определяется условиями проведения опыта. В реальных задачах используют различные формы и способы получения статистических данных. Они подробно обсуждаются в общей теории статистики (см., например, работы [16, 29, 43, 50]). В математической статистике обычно предполагается, что статистические данные получены выборочным методом. Суть этого метода состоит в формировании выборочной совокупности, причем для всех объектов изучаемой совокупности обеспечивается равная вероятность быть включенным в выборочную совокупность. При математическом описании статистических данных их обычно представляют совокупностью случайных величин $\vec{X} = (X_1,...,X_n)$, которую коротко называют выборкой объема *n*, где *n* – число наблюдений, а X_i, $i = 1,...,n$, называют элементами выборки. Следуя принятой традиции, конкретную реализацию выборки $\vec{X} = (X_1,...,X_n)$ в опыте будем обозначать строчными буквами, то есть в виде $\vec{x} = (x_1,...,x_n)$.

Определение 1.1.1. Выборочным пространством $\mathrm{X} = \{\vec{x}\}$ называют множество всех возможных реализаций $\vec{x} = (x_1,...,x_n)$ векторной случайной величины $\vec{X} = (X_1,...,X_n)$, на котором задано ее распределение вероятностей, однозначно определяемое функцией распределения $F_{\vec{X}}(x_1,...,x_n) \in \mathfrak{I}$, $(x_1,...,x_n) \in \mathrm{X}$.

Отметим еще раз, что элементами (точками) выборочного пространства X являются конкретные реализации $\vec{x} = (x_1,...,x_n)$ случайного вектора $\vec{X} = (X_1,...,X_n)$. Отметим также, что обычно выборочное пространство $\mathrm{X} = \{\vec{x}\}$ является либо всем *n*-мерным евклидовым пространством R^n, либо его подмножеством, либо оно может состоять из отдельных точек R^n, если случайная величина X дискретная.

Итак, распределение вероятностей случайной величины $\vec{X} = (X_1,...,X_n)$ на множестве $\mathrm{X} = \{\vec{x}\}$ однозначно определяется ее функцией распределения $F_{\vec{X}}$. Множество допустимых функций

распределения случайной величины $\vec{X} = (X_1,...,X_n)$ в условиях данного статистического эксперимента обозначим через $\Im = \{F_{\vec{X}}\}$.

Таким образом, чтобы задать статистическую модель эксперимента, связанного с изучением с.в. X, необходимо конкретизировать выборочное пространство $\mathrm{X} = \{\vec{x}\}$ и множество допустимых функций распределения вероятностей $\Im = \{F_{\vec{X}}\}$. Для удобства дальнейших ссылок введем понятие статической модели в виде следующего определения.

Определение 1.1.2. Под статистической моделью эксперимента понимают пару $(\mathrm{X}, \Im)$, где $\mathrm{X} = \{\vec{x}\}$ – выборочное пространство и $\Im = \{F_{\vec{X}}\}$ – множество допустимых распределений в условиях данного эксперимента.

Отметим, что при решении многих статистических задач предполагается, что компоненты случайного вектора $\vec{X} = (X_1,...,X_n)$ являются независимыми и одинаково распределенными (н.о.р.) случайными величинами с такой же функцией распределения вероятностей, какой обладает изучаемая случайная величина X. При выполнении этих предположений говорят, что статистические данные получены в схеме повторных и независимых наблюдений над изучаемой с.в. X. В частности, предположение одинаковой распределенности формально означает, что выполняется равенство $P(X_i \le x) = P(X \le x)$, или $F_{X_i}(x) = F_X(x)$, $\forall i \in (1,...,n)$, а предположение независимости означает, что результат последующего опыта не зависит от результата предыдущего опыта. Математически эти предположения проявляются в том, что совместная плотность случайного вектора $\vec{X} = (X_1,...,X_n)$ факторизуется, то есть записывается в виде

$$f_{X_1,...,X_n}(x_1,...,x_n) = \prod_{i=1}^{n} f_X(x_i). \qquad \textbf{(1.1.3)}$$

Последовательность $X_1,...,X_n$ н.о.р. случайных величин коротко называют *выборкой объема n* из распределения F_X (или, точнее, *выборкой*, которая порождается распределением F_X). Информация,

содержащаяся в результатах опыта, служит основой для решения статистических задач.

Итак, при решении любой статистической задачи одинаково важную роль играют два понятия:

1) *статистические данные* $(X_1,...,X_n)$ – *выборка*;

2) *статистическая модель* $(\mathrm{X},\mathfrak{I})$.

Суть и цель любой статистической процедуры состоит в том, чтобы на основе информации, содержащейся в выборке $X_1,...,X_n$, *доопределить статистическую модель* $(\mathrm{X},\mathfrak{I})$. Эту цель кратко можно выразить следующим образом:

$$\begin{Bmatrix} X_1,...,X_n \\ стат.\,данные \end{Bmatrix} \Rightarrow \begin{Bmatrix} (\mathrm{X},\mathfrak{I}) \\ стат.\,модель \end{Bmatrix}. \qquad \textbf{(1.1.4)}$$

Замечание 1.1.5. Отметим, что математическое описание статистических данных подробно обсуждается в разд. 4.1. В указанном разделе вводится понятие упорядоченной статистики $X_{(\cdot)} = (X_{(1)},...,X_{(n)})$ и рангового вектора $\vec{R} = (R_1,...,R_n)$ выборки $\vec{X} = (X_1,...,X_n)$ и отмечается, что информация, содержащаяся в исходной выборке, условно может быть разделена на две части. Одна из них содержится в упорядоченной статистике $X_{(\cdot)} = (X_{(1)},...,X_{(n)})$, а другая – в ранговом векторе $\vec{R} = (R_1,...,R_n)$. Этот факт можно представить следующим выражением:

$$\{(X_1,...,X_n)\} \Leftrightarrow \left\{\left(X_{(1)},...,X_{(n)}\right);(R_1,...,R_n)\right\}, \qquad \textbf{(1.1.6)}$$

которое означает, что совместное распределение векторов $\vec{X}_{(\cdot)}$ и $\vec{R}$ определяется распределением вектора $\vec{X}$.

Таким образом, статистические процедуры, направленные на уточнение статистической модели $(\mathrm{X},\mathfrak{I})$, могут быть основаны либо на исходной выборке, либо на порядковых статистиках, либо на рангах наблюдений. Статистические процедуры, разрабатываемые для параметрических моделей (см. гл. 2 и 3), в большинстве своем основаны на исходной выборке $\vec{X} = (X_1,...,X_n)$, а непараметрические и робастные процедуры (см. гл. 4 – 10), как правило, основаны на порядковых статистиках либо на рангах и знаках наблюдений.

1.2. Типы статистических моделей

В математической статистике различают три общих типа моделей: параметрические, непараметрические и используемые при изучении и построении робастных статистических процедур, которые коротко называют супермоделями. Для этих типов моделей множество допустимых функций распределения $\Im$ будем соответственно обозначать в виде $\Im_\theta$, $\Im_N$, $\Im_R$. Наглядно это можно представить следующим образом:

$$\Im = \begin{cases} \to \Im_\theta = \{F_X : F_X(x;\theta),\ \theta \in \Theta\} \text{ – параметрическая модель;} \\ \to \Im_N = \{\text{мн-во аб. непр. р.}\} \text{ – непараметрическая модель;} \\ \to \Im_R = \{F_X :\ d(F_0, F_X) < \varepsilon\} \text{ – супермодель роб. статистики.} \end{cases} \quad \textbf{(1.2.1)}$$

Описание способов задания множеств $\Im_\theta, \Im_N, \Im_R$ приводится ниже.

1.2.А. Параметрические модели

Для статистических моделей параметрической (классической) статистики характерно, что множество допустимых функций распределений $\Im_\theta$ задается в виде параметрического семейства функций с конечным числом неизвестных параметров θ, которые принадлежат заданному параметрическому множеству Θ, то есть множество $\Im_\theta$ задается в виде

$$\Im_\theta = \{F_X : F_X(x\,;\theta),\ \theta \in \Theta\}. \quad \textbf{(1.2.2)}$$

Отметим, что в общем случае параметр θ может быть векторным. При задании параметрических моделей функциональный характер (тип) распределения считается известным, то есть вся неопределенность модели связана с неопределенностью параметров $\theta \in \Theta$. В связи с этим, статистические процедуры, реализуемые в рамках параметрических моделей, направлены на построение точечных и интервальных оценок этих параметров, а также на построение критериев проверки статистических гипотез относительно параметров θ на основе информации, содержащейся в выборке. Эти параметрические процедуры обсуждаются во второй и третьей главах.

1.2.Б. Непараметрические модели

Эти модели применяются при решении задач непараметрической статистики. Для них характерно, что множество допустимых функций распределений $\mathfrak{I}_N$ очень «широкое» и задается из некоторых общих соображений, типа непрерывности, симметрии, равенства и т.п. Отметим, что при конкретизации множества $\mathfrak{I}_N$ функциональный характер распределения F_X *не задается*, то есть считается, что ф.р. F_X *неизвестна*. Типичным примером является задача сравнения некоторого нового типа обработки по сравнению со стандартным. Например, пусть имеются статистические данные $X_1,...,X_n$, полученные в результате наблюдений над с.в. X с ф.р. F_X в стандартных условиях, и статистические данные $Y_1,...,Y_m$, полученные в результате наблюдений над с.в. Y с ф.р. F_Y в условиях, отличных от стандартных. По этим статистическим данным проверяется гипотеза о том, что изменение условий эксперимента не сказывается на его результатах. Эта гипотеза сводится к проверке факта равенства распределений F_X и F_Y. Если при этой проверке предполагается лишь равенство распределений, а их функциональный вид не конкретизируется и предполагается лишь, что они непрерывны, то эта задача однородности решается в рамках непараметрической статистической модели. Отметим, что многие задачи непараметрической статистики в математическом отношении сводятся к изучению поведения некоторых функционалов $T(F)$, заданных на множестве $\mathfrak{I}_N$, и связаны с построением оценок этих функционалов на основе информации, содержащейся в выборке. Эти задачи обсуждаются в гл. 4 и 5.

1.2.В. Модели робастной статистики. Понятие супермодели

Для того чтобы задать модель в виде параметрического множества $\mathfrak{I}_\theta$, требуется достаточно большое количество информации, позволяющее конкретизировать функциональный вид распределения вероятностей. При построении непараметрических моделей путем задания множества $\mathfrak{I}_N$ используется гораздо меньше информа-

ции. Однако при построении как параметрических, так и непараметрических моделей неизменно считается, что *все предположения этих моделей верны*, то есть они соответствуют реально наблюдаемому случайному явлению. В таких случаях говорят, что статистическая модель адекватна. Но всякая модель является лишь в лучшем случае удачной аппроксимацией, на практике неизбежно возникают отклонения от предположений модели. В связи с этим возникает необходимость изучения влияния таких отклонений на статистические выводы и желательно иметь сведения о том, какие отклонения и от каких предположений модели оказывают решающее влияние на конечный вывод при статистическом анализе и нарушение каких допущений является относительно безвредным. В частности, может возникнуть вопрос о применимости процедур, рассчитанных на нормальное распределение, в ситуациях, когда истинное распределение не является таковым. Ответам на подобные вопросы и построению статистических процедур, нечувствительных к отклонениям от модели, посвящено самостоятельное направление в математической статистике, которое получило название робастной статистики. Термин «робастность» был введен в статистическую литературу Боксом в 1953 году (см. работу [136]) и американским математиком Тьюки [188, 189], робастная статистика была выделена в «статистику третьего поколения» после параметрической и непараметрической статистики.

Особенности моделей робастной статистики состоят в том, что некоторые предположения принятой за основу идеальной модели в виде заданной ф.р. F_0 подвергаются сомнению. По этой причине принятая статистическая модель «расширяется» до «супермодели», которую обычно задают в виде некоторой окрестности идеальной модели F_0, конкретизируемой выбранной метрикой $d(F_0, F_X)$ на множестве допустимых ф.р. $\mathfrak{I}$, то есть супермодель задается в виде множества $\mathfrak{I}_R = \{F_X : d(F_0, F_X) < \varepsilon\}$, $\varepsilon > 0$. В литературе используют различные метрики $d(F_0, F_X)$ и описаны также другие подходы к заданию супермоделей, они будут подробно обсуждаться в разд. 6.3.

Итак, в теории робастной статистики вместе с понятием базовой (идеальной) статистической модели вводится дополнительное по-

нятие, названное супермоделью. Понятие супермодели служит для того, чтобы описать некоторую базовую модель, в которую мы верим и выбираем ее в качестве основы, а также описать возможные в условиях реального эксперимента отклонения от предположений базовой модели. При этом основная цель состоит в разработке таких статистических процедур, которые должны «работать достаточно хорошо» как при идеальной модели F_0, так и в рамках принятой супермодели $\Im_R$. Понятие «работать достаточно хорошо» каждый раз требует дополнительного уточнения (эти вопросы обсуждаются в разд. 6.4. В математическом отношении построение таких процедур сводится к изучению свойств некоторых функционалов $T(F)$, заданных на множестве $\Im_R$, а также к изучению свойств оценок этих функционалов как при идеальной модели F_0, так и в рамках принятой супермодели $\Im_R$.

1.3. Вероятностное описание случайных величин

В данном разделе приведем основные понятия и формулы, которые используются при статистическом описании случайных величин. Для того чтобы охарактеризовать случайную величину X, необходимо задать распределение вероятностей по ее возможным значениям. Для осуществления этого используется понятие функции распределения (ф.р.) вероятностей, которая везде ниже будет обозначаться в виде F_X. Одномерная случайная величина X с возможными значениями на действительной оси $R^1 = \{x : -\infty < x < +\infty\}$ характеризуется функцией распределения вероятностей F_X, которая определяется в виде

$$F_X(x) = P(X \le x),\ x \in R^1. \qquad \textbf{(1.3.1)}$$

Если случайная величина $\vec{X} = (X_1,...,X_n)$ n-мерная, то ее ф.р. является функцией n переменных $x_1,...,x_n$ и в точке $(x_1,...,x_n) \in R^n$, где R^n – n-мерное евклидовое пространство, определяется с помощью вероятности совместного осуществления событий в фигурных скобках, то есть в виде

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = P\{X_1 \le x_1;\ldots;X_n \le x_n\},$$
$$(x_1,\ldots,x_n) \in R^n. \qquad \textbf{(1.3.2)}$$

Отметим, что при изучении случайных величин непрерывного типа в некоторых случаях удобнее использовать вместо функции распределения вероятностей плотность распределения вероятностей (п.р.в.), которую везде ниже для одномерного случая будем обозначать как $f_X(x),\ x \in R^1$, а для n-мерного случая в виде

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{\vec{X}}(x_1,\ldots,x_n),\ (x_1,\ldots,x_n) \in R^n.$$

Отметим также, что в тех случаях, когда это не приводит к путанице, индекс X в обозначениях функции распределения и плотности будет опускаться.

1.3.А. Непрерывный тип

Пусть X – одномерная непрерывная случайная величина с возможными значениями на действительной оси R^1 и с абсолютно непрерывной ф.р. вероятностей $F_X(x),\ x \in R^1$. Тогда для нее существует функция $f_X(x),\ x \in R^1$, которую называют плотностью распределения вероятностей, такая, что выполняются следующие выражения:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx,\ x \in R^1, \qquad \textbf{(1.3.3)}$$

и

$$f_X(x) = \frac{dF_X(x)}{dx},\ x \in R^1. \qquad \textbf{(1.3.4)}$$

Отметим, что всякая плотность распределения вероятностей является неотрицательной функцией, то есть $f_X(x) \ge 0,\ \forall x \in R^1$, удовлетворяющей условию нормировки

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1. \qquad \textbf{(1.3.5)}$$

Обратно, любая неотрицательная функция $f_X(x)$, удовлетворяю-

щая условию нормировки (1.3.5), может быть интерпретирована как плотность распределения вероятностей некоторой случайной величины. Отметим также, что дифференциал ф.р. $F_X(x)$ равен произведению плотности $f_X(x)$ на дифференциал независимого аргумента, то есть

$$dF_X(x) = f_X(x)dx. \qquad \textbf{(1.3.6)}$$

Это непосредственно следует из (1.3.4). Отметим, что плотность распределения вероятностей $f_X(x)$ случайной величины X в любой точке $x \in R^1$ определяется в виде предела отношения вероятности события, состоящего в том, что случайная величина X попадет в бесконечно малый интервал $(x, x+\Delta x]$, к величине интервала Δx, при $\Delta x \to 0$, то есть в виде

$$f_X(x) = \lim_{\Delta x \to 0} \frac{P\{x < X \le x + \Delta x\}}{\Delta x} =$$

$$= \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x} = \frac{dF_X(x)}{dx},\ x \in R^1. \qquad \textbf{(1.3.7)}$$

Справедливы следующие свойства ф.р. $F_X(x)$ с.в. X:

1. $0 \le F_X(x) \le 1,\ x \in R^1$.
2. $F_X(x_1) \le F_X(x_2)$, для $x_1 < x_2$.
3. $\lim_{x \to -\infty} F_X(x) = 0,\ \lim_{x \to +\infty} F_X(x) = 1$.
4. $P\{a \le X \le b\} = F_X(b) - F_X(a)$.
5. $F_X(x)$ – непрерывная справа функция, то есть

$$\forall \varepsilon > 0 \quad \lim_{\varepsilon \to 0} F_X(x+\varepsilon) = F_X(x+0) = F_X(x),\ x \in R^1. \qquad \textbf{(1.3.8)}$$

Перечисленные свойства ф.р. F_X следуют из ее определения в виде (1.3.1).

Общие свойства плотности $f_X(x)$ с.в. X следующие:

1. $f_X(x) \ge 0,\ x \in R^1$.
2. $\int\limits_{-\infty}^{+\infty} f_X(x)dx = 1$.

3. $f_X(x)dx = dF_X(x)$.

4. $P\{a \le X \le b\} = \int_a^b f_X(x)dx = \int_a^b dF_X(x) = F_X(b) - F_X(a)$. **(1.3.9)**

Эти свойства также непосредственно следуют из формул (1.3.4) и (1.3.7).

Итак, исчерпывающая информация о случайной величине X содержится в ее функции распределения F_X. Зная эту функцию, можно вычислить вероятности любых событий, связанных со случайной величиной X. При изучении случайной величины X на практике также часто используют различные числовые характеристики, которые вычисляются с помощью функции распределения вероятностей. Наиболее часто на практике применяют начальные и центральные моменты.

Определение 1.3.10. Пусть X – случайная величина с ф.р. $F_X(x)$ и плотностью $f_X(x)$, $x \in R^1$. Начальный момент k-го порядка, $k = 1, 2, \ldots$, определяется в виде

$$\alpha_k(F) = M[X^k] = \int_{-\infty}^{+\infty} x^k f_X(x)dx = \int_{-\infty}^{+\infty} x^k dF_X(x),\ k = 1, 2, \ldots. \quad \textbf{(1.3.11)}$$

Отметим, что начальный момент первого порядка является математическим ожиданием с.в. X, то есть $\alpha_1 = M(X)$.

Определение 1.3.12. Центральный момент k-го порядка, $k = 1, 2, \ldots$, определяется в виде

$$\mu_k(F) = M[(X - \alpha_1)^k] = \int_{-\infty}^{+\infty} (x - \alpha_1)^k f_X(x)dx = \int_{-\infty}^{+\infty} (x - \alpha_1)^k dF_X(x),$$
$$k = 1, 2, \ldots. \quad \textbf{(1.3.13)}$$

Отметим, что центральный момент второго порядка является дисперсией с.в. X, то есть $\mu_2 = M[(X - \alpha_1)^2] = D(X)$. Начальные и центральные моменты с.в. X, в свою очередь, используются для определения других числовых характеристик, например таких, как коэффициент асимметрии

$$\gamma_1(F) = \mu_3(F) / \mu_2^{3/2}(F) \quad \textbf{(1.3.14)}$$

и коэффициент эксцесса

$$\tilde{\gamma}_2(F) = \mu_4(F)/\mu_2^2(F),\ \gamma_2(F) = \mu_4(F)/\mu_2^2(F)\ - 3. \qquad \textbf{(1.3.15)}$$

Для изучения случайных величин в теории вероятностей используется большое число различных числовых характеристик (см., например, [14, 23]). Предполагая, что читатель знаком с этими вопросами, и для удобства дальнейших ссылок, приведем здесь лишь некоторые формулы и те числовые характеристики, которые будут использоваться ниже. В дальнейшем часто будут использоваться формулы, позволяющие вычислять распределение вероятностей функции случайного аргумента и формулы вычисления математического ожидания функции от случайной величины.

Пусть с.в. X имеет ф.р. $F_X(x)$ и плотность $f_X(x)$, $x \in R^1$. Пусть далее $y = \varphi(x)$ – непрерывная, дифференцируемая, строго монотонная функция и $x = \varphi^{-1}(y)$ – однозначная обратная для нее функция, которая также непрерывная и строго монотонная. Тогда ф.р. $F_Y(y)$ случайной величины $Y = \varphi(X)$ вычисляется по формуле

$$F_Y(y) = P(Y \le y) = P\{\varphi(X) \le y\} = P\{X \le \varphi^{-1}(y)\} = F_X(\varphi^{-1}(y)),$$

$$y \in R^1, \qquad \textbf{(1.3.16)}$$

и, следовательно, плотность $f_Y(y)$ случайной величины $Y = \varphi(X)$ вычисляется по формуле

$$f_Y(y) = f_X(\varphi^{-1}(y))\left|\frac{d\varphi^{-1}(y)}{dy}\right| = f_X(\varphi^{-1}(y))\left|\frac{1}{\varphi'(\varphi^{-1}(y))}\right|,\ y \in R^1. \qquad \textbf{(1.3.17)}$$

Рассмотрим пример, когда случайные величины X и Y связаны линейно, то есть $Y = aX + b$. В этом случае имеем

$$F_Y(y) = P(Y \le y) = P\{aX + b \le y\} = \begin{cases} F_X\{(y-b)/a\}, & a > 0, \\ 1 - F_X\{(y-b)/a\}, & a < 0. \end{cases} \qquad \textbf{(1.3.18)}$$

Плотность $f_Y(y)$ с.в. Y, согласно (1.3.17), запишется в виде

$$f_Y(y) = f_X(\varphi^{-1}(y))\left|\frac{d\varphi^{-1}(y)}{dy}\right| = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right). \qquad \textbf{(1.3.19)}$$

Отметим, что если функция $y=\varphi(x)$ не является строго монотонной, тогда обратная функция не является однозначной, то есть двум (или большему числу) различным значениям x может соответствовать одно значение y. В этом случае формулу (1.3.17) следует применить на каждом участке монотонности функции $y=\varphi(x)$, а затем сложить полученные плотности, чтобы получить результирующую плотность. Например, пусть надо найти плотность распределения вероятности случайной величины $Y=\varphi(X)=X^2$ (см. также упр. 1.4.22). В этом случае функция $y=\varphi(x)=x^2$ не является строго монотонной и имеет две обратные дифференцируемые строго монотонные функции $x_1=\varphi_1^{-1}(y)=\sqrt{y}$ и $x_2=\varphi_2^{-1}(y)=-\sqrt{y}$, $y>0$. Далее, формула (1.3.16) в данном случае для $0<y<\infty$ запишется в виде

$$F_Y(y)=P(Y\le y)=P\{X^2\le y\}=P\{-\sqrt{y}\le X\le\sqrt{y}\}=F_X(\sqrt{y})-F_X(-\sqrt{y}).$$

Отсюда получаем плотность случайной величины $Y=\varphi(X)=X^2$ в виде

$$f_Y(y)=f_X(\sqrt{y})/2\sqrt{y}+f_X(-\sqrt{y})/2\sqrt{y}=$$
$$=f_X(\varphi_1^{-1}(y))\left|\frac{d\varphi_1^{-1}(y)}{dy}\right|+f_X(\varphi_2^{-1}(y))\left|\frac{d\varphi_2^{-1}(y)}{dy}\right|,\ 0<y<\infty.$$

В общем случае для функции $y=\varphi(x)$, область определения которой разбивается на k интервалов монотонности, получаем формулу

$$f_Y(y)=\sum_{i=1}^{k}f_X(\varphi_i^{-1}(y))\left|\frac{d\varphi_i^{-1}(y)}{dy}\right|,\ y\in R^1. \qquad \textbf{(1.3.20)}$$

Отметим, что для вычисления математического ожидания случайной величины вида $Y=\varphi(X)$ не требуется вычислять плотность $f_Y(y)$, поскольку справедливы следующие равенства:

$$M(Y)=M[\varphi(X)]=\int_{-\infty}^{+\infty}y\,f_Y(y)dy=\int_{-\infty}^{+\infty}\varphi(x)f_X(x)dx=\int_{-\infty}^{+\infty}\varphi(x)dF_X(x). \qquad \textbf{(1.3.21)}$$

1.3.Б. Дискретный тип

Пусть X – одномерная дискретная случайная величина, возможные значения которой $\tilde{x}_1,\tilde{x}_2,...$ отстоят друг от друга и являются элементами конечного либо счетного множества. На практике часто с.в. X имеет конечное число k возможных состояний $\tilde{x}_1,...,\tilde{x}_k$, которые называют также точками сосредоточения вероятностей.

Для того чтобы охарактеризовать дискретную случайную величину X необходимо задать распределение вероятностей по ее возможным значениям $\tilde{x}_1,...,\tilde{x}_k$, то есть конкретизировать вероятности $p_i = P(X = \tilde{x}_i)$, $i = 1,...,k$. Кратко это записывают в виде

$$\begin{pmatrix} X \\ \vec{p} \end{pmatrix} : \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ p_1,...,p_k \end{pmatrix}, \ \sum_{i=1}^{k} p_i = 1. \qquad \textbf{(1.3.22)}$$

Отметим, что возможные состояния $\tilde{x}_1,...,\tilde{x}_k$ в (1.3.22) обычно располагают для удобства в порядке возрастания, то есть $\tilde{x}_1 < ... < \tilde{x}_k$. Приведем выражение для ф.р. $F_X(x)$, $x \in R^1$, дискретной с.в. X, заданной распределением вероятностей в виде (1.3.22). Для удобства аналитической записи ф.р. $F_X(x)$ введем понятие единичной функции Хевисайда, которая определяется в виде

$$C(x-y) = \begin{cases} 1, & x \geq y, \\ 0, & x < y. \end{cases} \qquad \textbf{(1.3.23)}$$

Используя определение функции распределения вероятностей в виде (1.3.1), применяя теорему сложения вероятностей и учитывая, что события $(X = \tilde{x}_i)$, $i = 1,...,k$, попарно несовместны, получаем выражение для ф.р. $F_X(x)$ в виде

$$F_X(x) = \sum_{i:\,\tilde{x}_i \leq x} p_i = \sum_{i=1}^{k} p_i C(x - \tilde{x}_i), \ x \in R^1. \qquad \textbf{(1.3.24)}$$

Отметим, что ф.р. $F_X(x)$ дискретной с.в. X, заданной распределением (1.3.22), является разрывной ступенчатой функцией со значениями $(p_1 + ... + p_i)$ в точках ее возможных состояний $\tilde{x}_i$, $i = 1,...,k$.

Замечание 1.3.25. Плотность распределения вероятностей используют для изучения случайных величин непрерывного типа. Однако это понятие можно формально ввести и для случайных величин дискретного типа. Чтобы осуществить это, введем понятие дельта-функции Дирака. Дельта-функцию определяют в виде

$$\delta(x - x_0) = \begin{cases} \infty, & x = x_0, \\ 0, & x \neq x_0. \end{cases} \qquad \textbf{(1.3.26)}$$

Отметим (см., например, [33] и упр. 1.4.93), что для любой непрерывной в точке x_0 функции $\varphi(x)$ выполняется выражение

$$\int_{-\infty}^{+\infty} \varphi(x)\delta(x-x_0)dx = \varphi(x_0)\,. \qquad \textbf{(1.3.27)}$$

Отметим также, что единичная функция (1.3.23) не является аналитической функцией, однако формально можно использовать следующие выражения:

$$dC(x-y) = \delta(x-y)dx\,,\ \int_{-\infty}^{y} \delta(x)dx = \int_{-\infty}^{+\infty} C(y-x)\delta(x-0)dx = C(y)\,. \qquad \textbf{(1.3.28)}$$

Итак, формальное выражение для плотности дискретной с.в. X, заданной распределением (1.3.22), с учетом (1.3.24), записывается в виде

$$f_X(x) = \sum_{i=1}^{k} p_i\delta(x-\tilde{x}_i)\,,\ x \in R^1\,. \qquad \textbf{(1.3.29)}$$

Это выражение удобно использовать при аналитических выкладках в тех случаях, когда плотность $f_X(x)$ входит в подынтегральную функцию.

Пример 1.3.30. Пусть X – дискретная случайная величина, заданная распределением (1.3.22). Тогда математическое ожидание этой случайной величины вычисляется по формуле

$$M(X) = \sum_{i=1}^{k} p_i\tilde{x}_i\,.$$

Эту формулу можно получить из общего определения математического ожидания с.в. X, которое записывается с помощью интеграла Стилтьеса (см. упр. 1.4.93) в виде

$$M(X) = \int_{-\infty}^{+\infty} x dF_X(x)\,.$$

В самом деле, используя (1.3.24) либо (1.3.29) и учитывая (1.3.27), получаем

$$M(X) = \int_{-\infty}^{+\infty} x dF_X(x) = \int_{-\infty}^{+\infty} x\sum_{i=1}^{k} p_i\delta(x-\tilde{x}_i)dx =$$

$$= \sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x\,\delta(x-\tilde{x}_i)dx = \sum_{i=1}^{k} p_i\tilde{x}_i.$$

1.3.В. Двумерный вариант, дискретный тип

Пусть X и Y – одномерные дискретные случайные величины, заданные своими распределениями вероятностей:

$$\begin{pmatrix} X \\ p \end{pmatrix} : \begin{pmatrix} \tilde{x}_1, ..., \tilde{x}_n \\ p_X(\tilde{x}_1), ..., p_X(\tilde{x}_n) \end{pmatrix},$$

$$p_X(\tilde{x}_i) = P\{X = \tilde{x}_i\},\ i = 1,...,n,\ \sum_{i=1}^{n} p_X(\tilde{x}_i) = 1,$$

и

$$\begin{pmatrix} Y \\ p \end{pmatrix} : \begin{pmatrix} \tilde{y}_1, ..., \tilde{y}_m \\ p_Y(\tilde{y}_1), ..., p_Y(\tilde{y}_m) \end{pmatrix},$$

$$p_Y(\tilde{y}_j) = P\{Y = \tilde{y}_j\},\ j = 1,...,m,\ \sum_{j=1}^{m} p_Y(\tilde{y}_j) = 1.$$

Двумерная дискретная случайная величина (X, Y) характеризуется своими возможными состояниями, которые задаются в виде матрицы размера $(n \times m)$ с элементами $\{\tilde{x}_i, \tilde{y}_j\}$, $i = 1,...,n$, $j = 1,...,m$, и двумерным распределением вероятностей

$$p_{XY}(\tilde{x}_i, \tilde{y}_j) = P\{X = \tilde{x}_i, Y = \tilde{y}_j\},\ i = 1,...,n,\ j = 1,...,m.$$

Отметим, что событие $(X = \tilde{x}_i)$, $i = 1,...,n$, может быть представлено в виде

$$(X = \tilde{x}_i) = (X = \tilde{x}_i)\,(Y = \tilde{y}_1) + ... + (X = \tilde{x}_i)\,(Y = \tilde{y}_m).$$

События $(Y = \tilde{y}_1),..., (Y = \tilde{y}_m)$ попарно несовместные и, следовательно, согласно теореме сложения вероятностей получаем

$$p_X(\tilde{x}_i) = P(X = \tilde{x}_i) = P\{(X = \tilde{x}_i)(Y = \tilde{y}_1)\} + ... + P\{(X = \tilde{x}_i)(Y = \tilde{y}_m)\}.$$

Таким образом, маргинальные вероятности, которые определяют распределения вероятностей с.в. X и Y, вычисляются по формулам

$$p_X(\tilde{x}_i) = \sum_{j=1}^{m} p_{XY}(\tilde{x}_i, \tilde{y}_j),\ i = 1,...,n, \qquad \textbf{(1.3.31)}$$

и аналогично

$$p_Y(\tilde{y}_j) = \sum_{i=1}^{n} p_{XY}(\tilde{x}_i, \tilde{y}_j),\ j = 1,...,m. \qquad \textbf{(1.3.32)}$$

Определение 1.3.33. Условные вероятности того, что $X = \tilde{x}_i$, $i = 1,...,n$, при условии, что $Y = \tilde{y}_j$, $j = 1,...,m$, определяются в виде

$$P\{X = \tilde{x}_i \,|\, Y = \tilde{y}_j\} = \frac{P\{X = \tilde{x}_i, Y = \tilde{y}_j\}}{P\{Y = \tilde{y}_j\}} = \frac{p_{XY}(\tilde{x}_i, \tilde{y}_j)}{p_Y(\tilde{y}_j)} = p_{X|Y}(\tilde{x}_i \,|\, \tilde{y}_j),$$
$$p_Y(\tilde{y}_i) \neq 0.$$

Определение 1.3.34. Случайные величины X и Y называют независимыми, если для всех значений i и j условные вероятности совпадают с безусловными, то есть выполняются выражения

$$p_{X|Y}(\tilde{x}_i \,|\, \tilde{y}_j) = p_X(\tilde{x}_i),\ p_{Y|X}(\tilde{y}_j \,|\, \tilde{x}_i) = p_Y(\tilde{y}_j),\ i = 1,...,n,\ j = 1,...,m,$$

или, что эквивалентно, если совместные (двумерные) вероятности равны произведениям маргинальных (одномерных) вероятностей, то есть

$$p_{XY}(\tilde{x}_i, \tilde{y}_j) = p_X(\tilde{x}_i) p_Y(\tilde{y}_j),\ i = 1,...,n,\ j = 1,...,m.$$

1.3.Г. Двумерный вариант, непрерывный тип

Пусть (X, Y) – двумерная непрерывная случайная величина с возможными значениями на плоскости

$$R^2 = \{(x, y) : -\infty < x < +\infty, -\infty < y < +\infty\}.$$

По аналогии с одномерным вариантом обозначим двумерную (или совместную) плотность двумерной с.в. (X, Y) через $f_{XY}(x, y)$, а двумерную функцию распределения вероятностей обозначим через $F_{XY}(x, y)$. Функцию распределения вероятностей в двумерном случае определяют в виде

$$F_{XY}(x, y) = P\{X \le x, Y \le y\} = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) dx dy, \qquad \textbf{(1.3.35)}$$

где $f_{XY}(x, y)$ – двумерная плотность, которая по аналогии с одномерным вариантом определяется выражением

$$f_{XY}(x,y)=\lim_{\Delta x\to 0,\Delta y\to 0}\frac{P\{x<X\le x+\Delta x,\ y<Y\le y+\Delta y\}}{\Delta x\Delta y}. \quad \textbf{(1.3.36)}$$

Для двумерной ф.р. $F_{XY}(x,y)$ выполняются следующие свойства:

1. $F_{XY}(-\infty,-\infty)=F_{XY}(-\infty,y)=F_{XY}(x,-\infty)=0$.

2. $F_{XY}(+\infty,+\infty)=1$.

3. Маргинальные ф.р. случайных величин X и Y определяются формулами

$$F_X(x)=F_{XY}(x,+\infty) \text{ и } F_Y(y)=F_{XY}(+\infty,y).$$

4. $F_{XY}(x,y)$ – непрерывная справа двумерная функция по каждой переменной, то есть в каждой точке $(x,y)\in R^2$ выполняются равенства

$$F_{XY}(x+0,y)=F_{XY}(x,y+0)=F_{XY}(x,y).$$

Отметим также, что для двумерной плотности $f_{XY}(x,y)$ справедливо выражение

$$f_{XY}(x,y)=\frac{\partial^{(2)}F_{XY}(x,y)}{\partial x\partial y},\ (x,y)\in R^2, \quad \textbf{(1.3.37)}$$

а условие нормировки записывается в виде

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y)dxdy=1.$$

Маргинальные плотности случайных величин X и Y (см. упр. 1.4.1) определяются формулами

$$f_X(x)=\int_{-\infty}^{+\infty} f_{XY}(x,y)dy,\ f_Y(y)=\int_{-\infty}^{+\infty} f_{XY}(x,y)dx. \quad \textbf{(1.3.38)}$$

Для определения понятия независимости непрерывных случайных величин X и Y вводят понятие условной плотности и условной функции распределения вероятностей. Определим $\forall x\in R^1$ событие $A=(X\le x)$ и событие $B=(y<Y\le y+\Delta y)$, где $y\in R^1$ выбрано таким образом, чтобы $f_Y(y)\ne 0$. Согласно определению условной вероятности событий, имеем $P(A\,|\,B)=P(AB)/P(B)$, поэтому условная функция распределения с.в. X при условии, что с.в. Y попадет в бесконечно малый интервал $(y<Y\le y+\Delta y]$ запишется

в виде

$$P\{X \le x \mid y < Y \le y + \Delta y\} = \frac{\int_{-\infty}^{x} \int_{y}^{y+\Delta y} f_{XY}(u,v)\,du\,dv}{\int_{y}^{y+\Delta y} f_Y(u)\,du}.$$

Разделив числитель и знаменатель предыдущего выражения на Δy и переходя к пределу при $\Delta y \to 0$, получим условную функцию распределения с.в. X при условии, что $Y = y$, в виде

$$F_{X|Y}(x \mid y) = \lim_{\Delta y \to 0} P\{X \le x | y < Y \le y + \Delta y\} =$$

$$= P\{X \le x \mid y\} = \frac{\int_{-\infty}^{x} f_{XY}(u,y)\,du}{f_Y(y)}. \qquad \textbf{(1.3.39)}$$

Дифференцируя по x условную функцию распределения с.в. X, получим ее условную плотность $f_{X|Y}(x \mid y)$ в виде

$$f_{X|Y}(x \mid y) = \frac{dF_{X|Y}(x \mid y)}{dx} = \frac{f_{XY}(x,y)}{f_Y(y)},\ f_Y(y) \ne 0. \qquad \textbf{(1.3.40)}$$

Аналогично определяется условная плотность с.в. Y в виде

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)},\ f_X(x) \ne 0. \qquad \textbf{(1.3.41)}$$

Пример 1.3.42. Пусть X – одномерная непрерывная случайная величина с ф.р. $F_X(x)$, плотностью $f_X(x)$, $x \in R^1$. Рассмотрим новую случайную величину $Y = \varphi(X)$, где $y = \varphi(x)$ – непрерывная дифференцируемая строго монотонная функция, для которой существует единственная обратная функция φ^{-1}. Тогда ф.р. случайной величины $Y = \varphi(X)$ вычисляется по формуле (1.3.16), а плотность вычисляется по формуле (1.3.17). Убедимся, что с помощью дельта-функции, определенной в (1.3.26), формула (1.3.17) может быть записана в виде

$$f_Y(y) = \int_{-\infty}^{+\infty} f_X(x)\delta(y - \varphi(x))\,dx = f_X(\varphi^{-1}(y))\left|\frac{1}{\varphi'(\varphi^{-1}(y))}\right|.$$

В самом деле, если $Y=\varphi(X)$, то с.в. Y при любом значении x с.в. X имеет единственное возможное значение $\varphi(x)$ и условная вероятность этого значения при данном x равна единице. Таким образом, условная функция распределения случайной величины Y при условии, что $X=x$, является вырожденной функцией распределения, то есть $F_{Y|X}(y|x)=C(y-\varphi(x))$, и, следовательно, плотность с.в. Y при условии, что $X=x$, является δ-функцией, то есть $f_{Y|X}(y|x)=\delta(y-\varphi(x))$. Отсюда следует, что совместная плотность случайных величин X и $Y=\varphi(X)$ запишется в виде

$$f_{XY}(x,y)=f_{Y|X}(y|x)f_X(x)=f_X(x)\delta(y-\varphi(x)).$$

Далее, согласно (1.3.38), с учетом предыдущего выражения, маргинальная плотность случайной величины Y запишется в виде

$$f_Y(y)=\int\limits_{-\infty}^{+\infty}f_{XY}(x,y)dx=\int\limits_{-\infty}^{+\infty}f_X(x)\delta(y-\varphi(x))dx.$$

После замены переменной интегрирования в виде $y-\varphi(x)=t$ правая часть предыдущего выражения запишется в виде

$$\int\limits_{-\infty}^{+\infty}f_X(x)\delta(y-\varphi(x))dx=\int\limits_{-\infty}^{+\infty}f_X(\varphi^{-1}(y-t))\frac{1}{\varphi'(\varphi^{-1}(y-t))}\delta(t-0)dt.$$

Отсюда, с использованием (1.3.27), следует формула (1.3.17).

Обобщение формулы (1.3.17) для вычисления распределения вероятностей функции двух непрерывных случайных величин проводится следующим образом (см. также упр. 1.4.75). Пусть (X_1,X_2) – непрерывная двумерная случайная величина с плотностью $f_{X_1X_2}(x_1,x_2)$, $(x_1,x_2)\in R^2$. Рассмотрим случайные величины $Y_1=\varphi_1(X_1,X_2)$ и $Y_2=\varphi_2(X_1,X_2)$, которые получены с помощью преобразований $y_i=\varphi_i(x_1,x_2)$, $i=1,2$, над случайной величиной (X_1,X_2). Предполагается, что эти преобразования взаимно однозначно отображают x_1x_2-плоскость на y_1y_2-плоскость и допускают единственные решения $x_i=\varphi_i^{-1}(y_1,y_2)$, $i=1,2$, которые непрерывны и имеют первые частные производные. Якобиан преобразований $y_i=\varphi_i(x_1,x_2)$, $i=1,2$, обозначается через J и определяется в виде

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial u} & \dfrac{\partial x_2}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial \varphi_1^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial \varphi_1^{-1}(y_1, y_2)}{\partial y_2} \\ \dfrac{\partial \varphi_2^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial \varphi_2^{-1}(y_1, y_2)}{\partial y_2} \end{vmatrix}. \qquad \textbf{(1.3.43)}$$

Пусть далее $S = \{[x_1; x_1 + \Delta x_1) \times [x_2; x_2 + \Delta x_2)\}$ – некоторая достаточно малая окрестность точки (x_1, x_2), а R – образ этой окрестности, получаемый при отображении с помощью функций φ_1 и φ_2, и пусть $P\{S\}$ – вероятность того, что случайный вектор (X_1, X_2) примет значение из множества S, тогда

$$P\{S\} = \int_{x_1}^{x_1+\Delta x_1} \int_{x_2}^{x_2+\Delta x_2} f_{X_1X_2}(x_1, x_2)dx_1dx_2 = \iint_S f_{X_1X_2}(x_1, x_2)dx_1dx_2 .$$

Переходя в этом интеграле к переменным y_1 и y_2, получим

$$P\{S\} = \iint_S f_{X_1X_2}(x_1, x_2)dx_1dx_2 =$$

$$= \iint_R f_{X_1X_2}\{\varphi_1^{-1}(y_1, y_2), \varphi_2^{-1}(y_1, y_2)\} \mid J \mid dy_1dy_2 = P\{R\},$$

где $P\{R\}$ – вероятность того, что случайный вектор (Y_1, Y_2) примет значение из множества R.

Отсюда следует, что плотность распределения вероятностей двумерной случайной величины (Y_1, Y_2) записывается в виде

$$f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(\varphi_1^{-1}(y_1, y_2), \varphi_2^{-1}(y_1, y_2)) \cdot |J|, \ (y_1, y_2) \in R^2. \qquad \textbf{(1.3.44)}$$

Понятия двумерных, условных и маргинальных плотностей позволяют сформулировать следующее определение независимости с.в. X и Y.

Определение 1.3.45. Случайные величины X и Y называют независимыми, если их условные плотности совпадают с безусловными, то есть

$$f_{X|Y}(x \mid y) = f_X(x) \text{ или } f_{Y|X}(y \mid x) = f_Y(y).$$

Отметим, что эквивалентное определение независимости с.в. X и Y формулируется следующим образом. Случайные величины X и

Y называют независимыми, если их совместная плотность $f_{XY}(x,y)$ факторизуется, то есть для нее выполняется следующее равенство:

$$f_{XY}(x,y) = f_X(x)\, f_Y(y). \qquad \textbf{(1.3.46)}$$

Факт независимости случайных величин X, Y и «силу», или «тесноту», их связи описывают с помощью таких числовых характеристик, как ковариация и коэффициент корреляции. Приведем определения этих числовых характеристик.

Определение 1.3.47. Ковариация двух случайных величин X и Y определяется в виде

$$COV(X,Y) = M[(X - M(X))(Y - M(Y))] = M(XY) - M(X)M(Y),$$

где $M(X)$ и $M(Y)$ – математические ожидания случайных величин X и Y, а $M(XY)$ – математическое ожидание произведения случайных величин X и Y, вычисляемое по формуле

$$M(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_{XY}(x,y)dxdy. \qquad \textbf{(1.3.48)}$$

Отметим, что ковариация случайной величины X с самой собой равна дисперсии, то есть $COV(X,X) = D(X)$. Отметим также, что если случайные величины X и Y независимые, то $COV(X,Y) = 0$. Это следует из определения (1.3.47) и формулы (1.3.46) (см. также упр. 1.4.22).

Определение 1.3.49. Коэффициентом корреляции случайных величин X и Y называют число ρ_{XY}, которое вычисляют по формуле

$$\rho_{XY} = \frac{COV(X,Y)}{\sqrt{D(X)D(Y)}}, \qquad \textbf{(1.3.50)}$$

и если $\rho_{XY} = 0$, то случайные величины X и Y называют некоррелированными.

Отметим, что если случайные величины X и Y линейно связаны, то есть $Y = aX + b$, то $\rho_{XY} = +1$ при $a > 0$ и $\rho_{XY} = -1$ при $a < 0$ (см. упр. 1.4.10). В общем случае для коэффициента корреляции выполняются неравенства $-1 \le \rho_{XY} \le +1$. В этом легко убедиться,

вычислив математическое ожидание случайной величины

$$\left[\lambda\cdot\frac{X-M(X)}{\sqrt{D(X)}}+\frac{Y-M(Y)}{\sqrt{D(Y)}}\right]^2,$$

где λ – действительное число. Используя свойства математического ожидания, получаем

$$M\left[\lambda\cdot\frac{X-M(X)}{\sqrt{D(X)}}+\frac{Y-M(Y)}{\sqrt{D(Y)}}\right]^2=\lambda^2+2\lambda\rho_{XY}+1\geq 0.$$

Очевидно, что $\lambda^2+2\lambda\rho_{XY}+1\geq 0$ при всех вещественных λ тогда и только тогда, когда $\rho_{XY}^2-1\leq 0$, и, следовательно, $-1\leq\rho_{XY}\leq+1$.

Отметим также, что при изучении характера связи между случайными величинами X и Y используют понятие функций регрессий. Приведем определение.

Определение 1.3.51. Условное математическое ожидание $M(X|Y=y)$ случайной величины X при условии, что $Y=y$, являющееся функцией y и вычисляемое по формуле

$$M(X|Y=y)=\int_{-\infty}^{+\infty}xf_{X|Y}(x\,|\,y)dx,\ y\in R^1, \qquad \textbf{(1.3.52)}$$

называется функцией регрессии X на Y. Аналогично, регрессия Y на X определяется в виде условного математического ожидания $M(Y|X=x)$, которое является функцией $x\in R^1$.

Замечание 1.3.53. Отметим, что понятия независимости и некоррелированности не являются тождественными. Из независимости случайных величин X и Y, то есть из равенства $COV(X,Y)=0$, следует некоррелированность, то есть $\rho_{XY}=0$. Обратное утверждение в общем случае неверно, исключение составляет, например, гауссовский случай, для которого из некоррелированности следует независимость (см. упр. 1.4.22). Отметим также, что для двумерной нормальной с.в. (X,Y) функции регрессии $M(X|Y=y)$ и $M(Y|X=x)$ являются линейными функциями (см. упр. 2.15.16), причем для независимых случайных величин X и Y они ортогональны, и, если угол пересечения между ними убывает до нуля, то это говорит об усилении зависимости между случайными величинами X и Y.

1.3.Д. Двумерный вариант, смешанный тип

Пусть (X,Y) – двумерная случайная величина смешанного типа, то есть одна ее компонента X является дискретной случайной величиной с конечным числом состояний, заданная распределением вида (1.3.22), другая компонента Y является непрерывной случайной величиной с функцией распределения $F_Y(y)$ и плотностью $f_Y(y)$, $y \in R^1$. Обозначим двумерную функцию распределения с.в. (X,Y) через $F_{XY}(x,y)$ и определим условные плотности для непрерывной случайной величины Y при условии, что дискретная случайная величина X приняла *i*-е состояние, $i = 1,...,k$, в виде

$$f_{Y|X}(y|\tilde{x}_i) = \frac{1}{p_X(\tilde{x}_i)} \lim_{\Delta y \to 0} P\{y < Y \le y + \Delta y, X = \tilde{x}_i\} / \Delta y,$$

$$i = 1,...,k,\ y \in R^1. \qquad \textbf{(1.3.54)}$$

Определение 1.3.55. Двумерную случайную величину (X,Y) называют случайной величиной смешанного типа, если для нее существуют одномерные условные плотности $f_{Y|X}(y|\tilde{x}_i)$, $i = 1,...,k$, $y \in R^1$, определяемые формулой (1.3.54), такие, что двумерная функция распределения с.в. (X,Y) записывается в виде

$$F_{XY}(x,y) = \sum_{i=1}^{k} p_X(\tilde{x}_i) C(x - \tilde{x}_i) \int_{-\infty}^{y} f_{Y|X}(y|\tilde{x}_i) dy,\ (x,y) \in R^2, \quad \textbf{(1.3.56)}$$

где $C(x - \tilde{x}_i)$ – единичная функция Хевисайда, определенная в (1.3.23).

Отметим, что маргинальная ф.р. $F_X(x)$ дискретной случайной величины X равна $F_{XY}(x,+\infty)$ и записывается в виде

$$F_X(x) = \sum_{i=1}^{k} p_X(\tilde{x}_i) C(x - \tilde{x}_i),\ x \in R^1, \qquad \textbf{(1.3.57)}$$

а маргинальная ф.р. $F_Y(y)$ непрерывной случайной величины Y равна $F_{XY}(+\infty, y)$ и записывается в виде

$$F_Y(y)=\sum_{i=1}^{k} p_X(\tilde{x}_i)\int_{-\infty}^{y} f_{Y|X}(y|\tilde{x}_i)dy,\ y\in R^1. \qquad \textbf{(1.3.58)}$$

Итак, можно сделать следующий вывод. Двумерная случайная величина (X,Y) смешанного типа, где X – дискретная случайная величина, заданная распределением

$$\begin{pmatrix} X \\ p \end{pmatrix}:\begin{pmatrix} \tilde{x}_1,.....,\tilde{x}_k \\ p_X(\tilde{x}_1),...,p_X(\tilde{x}_k) \end{pmatrix},\ \sum_{i=1}^{k} p_X(\tilde{x}_i)=1,$$

а Y – непрерывная случайная величина с функцией распределения $F_Y(y)$ и плотностью $f_Y(y)$, $y\in R^1$, полностью определяется парами $[p_X(\tilde{x}_i), f_{Y|X}(y|\tilde{x}_i)]$, $i=1,...,k$, где условные плотности $f_{Y|X}(y|\tilde{x}_i)$ определены в (1.3.54). Для наглядности вероятностного описания случайной величины (X,Y) смешанного типа можно привести следующую интерпретацию (см. монографию [66, с.59]). Сначала вся единичная вероятностная масса распределяется на отдельные количества $p_X(\tilde{x}_i)$, $i=1,...,k$, по состояниям $\tilde{x}_1,...,\tilde{x}_k$ дискретной случайной величины X. Затем эти количества вероятности непрерывно распределяются вдоль вертикальных линий $x=\tilde{x}_i$, $i=1,...,k$, таким образом, чтобы в любой точке $(\tilde{x}_i, y)$ на i-й линии плотность была бы равна $p_X(\tilde{x}_i)\, f_{Y|X}(y|\tilde{x}_i)$.

Замечание 1.3.59. Мы обсудили случай, когда дискретная случайная величина X имеет конечное число состояний. Обобщение на случай счетного числа состояний не вызывает принципиальных трудностей.

На будущее отметим, что математическое ожидание произведения случайных величин X и Y смешанного типа вычисляется по формуле

$$M(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_{XY}(x,y)dxdy=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\sum_{i=1}^{k} p_X(\tilde{x}_i)\delta(x-\tilde{x}_i)f_{Y|X}(y|\tilde{x}_i)dxdy=$$

$$=\int_{-\infty}^{+\infty}\left[\sum_{i=1}^{k} p_X(\tilde{x}_i)\int_{-\infty}^{+\infty} x\delta(x-\tilde{x}_i)dx\right] y\, f_{Y|X}(y|\tilde{x}_i)dy=\int_{-\infty}^{+\infty}\sum_{i=1}^{k}\tilde{x}_i p_X(\tilde{x}_i)\, y\, f_{Y|X}(y|\tilde{x}_i)dy=$$

$$=\sum_{i=1}^{k}\tilde{x}_i p_X(\tilde{x}_i)\cdot\int_{-\infty}^{+\infty} y\, f_{Y|X}(y|\tilde{x}_i)dy. \qquad \textbf{(1.3.60)}$$

Обсудим теперь критерий независимости компонент двумерной случайной величины (X,Y) смешанного типа. Отметим, что если условные плотности $f_{Y|X}(y|\tilde{x}_i)$ непрерывной случайной величины Y не зависят от номера состояния i дискретной случайной величины X, то есть, если $f_{Y|X}(y|\tilde{x}_i) = f_Y(y),\ \forall y \in R^1$, то выражение (1.3.56) записывается в виде $F_{XY}(x,y) = F_X(x)\,F_Y(y)$ и это означает, что случайные величины X и Y независимы.

Таким образом, в терминах плотностей распределения вероятностей для двумерной случайной величины (X,Y) смешанного типа равенство

$$f_{Y|X}(y|\tilde{x}_i) = f_Y(y),\ \forall y \in R^1, \qquad \textbf{(1.3.61)}$$

выполняемое для всех $i = 1,...,k$, является критерием независимости случайных величин X и Y. Можно сказать даже более точно, что равенство (1.3.61), выполняемое для всех $i = 1,...,k$, является необходимым и достаточным условием независимости X и Y (см. [66, с. 59]).

Отметим, что случайные величины смешанного типа будут использованы нами при изучении статистической связи между наблюдениями и их рангами в разд. 4.6.

1.3.E. Многомерный вариант, непрерывный тип

Пусть $\vec{X} = (X_1,...,X_n)$ – n-мерная случайная величина (или n-мерный случайный вектор) непрерывного типа с возможными значениями в n-мерном евклидовом пространстве R^n. Обозначим ф.р. этой случайной величины через $F_{X_1,...,X_n}(x_1,...,x_n) = F_{\vec{X}}(x_1,...,x_n)$ и ее плотность обозначим через $f_{X_1,...,X_n}(x_1,...,x_n) = f_{\vec{X}}(x_1,...,x_n)$, $(x_1,...,x_n) \in R^n$. По аналогии с предыдущими случаями ф.р. случайной величины $\vec{X} = (X_1,...,X_n)$ определяется в виде

$$F_{\vec{X}}(x_1,...,x_n) = P\{X_1 \le x_1;...;X_n \le x_n\}.$$

Определение 1.3.62. Случайный вектор $\vec{X} = (X_1,...,X_n)$ называют n-мерной случайной величиной непрерывного типа, если для нее существует функция $f_{\vec{X}}(x_1,...,x_n)$, $(x_1,...,x_n) \in R^n$, называемая совместной плотностью распределения вероятностей случайной величины $\vec{X} = (X_1,...,X_n)$, такая, что выполняются следующие выражения:

$$F_{\vec{X}}(x_1,...,x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{\vec{X}}(x_1,...,x_n)dx_1...dx_n \qquad \textbf{(1.3.63)}$$

и

$$f_{\vec{X}}(x_1,...,x_n) = \frac{\partial^n}{\partial x_1...\partial x_n} F_{\vec{X}}(x_1,...,x_n). \qquad \textbf{(1.3.64)}$$

Таким образом, плотность $f_{\vec{X}}(x_1,...,x_n)$ для n-мерной случайной величины $\vec{X} = (X_1,...,X_n)$ представляет собой симметричную смешанную производную n-го порядка функции распределения $F_{\vec{X}}(x_1,...,x_n)$. Из определения (1.3.62) вытекают следующие свойства:

1. $0 \le F_{\vec{X}}(x_1,...,x_n) \le 1$, $\forall (x_1,...,x_n) \in R^n$.

2. $F_{\vec{X}}(x_1,...,x_n)$ является неубывающей функцией по каждой из своих переменных x_i, $i = 1,...,n$.

3. $F_{\vec{X}}(x_1,...,x_n)$ – непрерывная справа функция по каждой из своих переменных x_i, $i = 1,...,n$.

4. $F_{\vec{X}}(x_1,...,x_n) = 0$, если, по крайней мере, одна из переменных x_i, $i = 1,...,n$, равна $-\infty$, то есть

$$F_{\vec{X}}(-\infty, x_2,...,x_n) = ... = F_{\vec{X}}(x_1,...,x_{n-1},-\infty) = 0.$$

5. $F_{\vec{X}}(x_1,...,x_n) = 1$, если все переменные x_i, $i = 1,...,n$, равны $+\infty$, то есть $F_{\vec{X}}(+\infty,...,+\infty) = 1$.

6. Маргинальная ф.р. случайной величины X_1 определяется в виде $F_{X_1}(x_1) = F_{\vec{X}}(x_1,+\infty,...,+\infty)$, причем другие маргинальные ф.р. $F_{X_i}(x_i)$, $i = 1,...,n$, определяются аналогично.

Для плотности $f_{\vec{X}}(x_1,...,x_n)$ справедливы следующие свойства:

1. $f_{\vec{X}}(x_1,...,x_n) \geq 0$, $\forall (x_1,...,x_n) \in R^n$.

2. Условие нормировки

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f_{\vec{X}}(x_1,...,x_n)dx_1...dx_n = 1. \quad \textbf{(1.3.65)}$$

3. Пусть событие $A \in R^n$, тогда

$$P(A) = P\{(X_1,...,X_n) \in A\} = \int_A ... \int f_{\vec{X}}(x_1,...,x_n)dx_1...dx_n =$$

$$= \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f_{\vec{X}}(x_1,...,x_n) I_A(x_1,...,x_n)dx_1...dx_n, \quad \textbf{(1.3.66)}$$

где $I_A(x_1,...,x_n)$ – индикатор события A, то есть $I_A(x_1,...,x_n) = 1$, если $(x_1,...,x_n) \in A$, и $I_A(x_1,...,x_n) = 0$, если $(x_1,...,x_n) \notin A$.

4. Маргинальные плотности с.в. X_i, $i = 1,...,n$, вычисляются по формуле

$$f_{X_i}(x_i) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} f_{\vec{X}}(x_1,...,x_n)dx_1...dx_{i-1}dx_{i+1}...dx_n, \; i = 1,...,n. \quad \textbf{(1.3.67)}$$

Отметим, что математическое ожидание некоторой измеримой функции $g(\cdot)$ с.в. $(X_1,...,X_n)$ вычисляется по формуле

$$M[g(X_1,...X_n)] = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} g(x_1,...,x_n) f_{\vec{X}}(x_1,...,x_n)dx_1...dx_n, \quad \textbf{(1.3.68)}$$

и в частном случае, если $g(x_1,...,x_n) = x_i$, $i = 1,...,n$, то

$$M(X_i) = \int_{-\infty}^{+\infty} x_i f_{X_i}(x_i)dx_i = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} x_i f_{\vec{X}}(x_1,...,x_n)dx_1...dx_n,$$

$$i = 1,...,n. \quad \textbf{(1.3.69)}$$

Определение 1.3.70. Случайные величины $X_1,...,X_n$ называют независимыми, если их совместная ф.р. (либо совместная плотность) факторизуется, то есть

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n}F_{X_i}(x_i)$$

или
$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n}f_{X_i}(x_i). \qquad \textbf{(1.3.71)}$$

Замечание 1.3.72. В данном разделе мы привели лишь некоторые общие свойства распределений вероятностей, которые используются при изложении основного материала. Сведения о конкретных распределениях, описывающих дискретные случайные величины, приведены в Приложении 1, а специальные непрерывные распределения вероятностей приведены в Приложении 2 (см. также разд. 1.4).

1.4. Упражнения, задачи и дополнения

1.4.1. Пусть $Z=(X,Y)$ – непрерывная двумерная случайная величина с плотностью $f_{XY}(x,y)$, $(x,y)\in R^2$, и пусть $f_X(x)$, $x\in R^1$, $f_Y(y)$, $y\in R^1$, – маргинальные плотности случайных величин X и Y.

А) Покажите, что маргинальные плотности случайных величин X и Y вычисляются по формулам

$$\int_{-\infty}^{+\infty}f_{XY}(x,y)dx=f_Y(y),\ \int_{-\infty}^{+\infty}f_{XY}(x,y)dy=f_X(x). \qquad \textbf{(1.4.2)}$$

Б) Убедитесь, что для двумерной плотности $f_{XY}(x,y)$ и условных плотностей $f_{X|Y}(x\,|\,y)$, $f_{Y|X}(y\,|\,x)$ выполняются условия нормировки, то есть справедливы следующие равенства:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}f_{XY}(x,y)dxdy=1,\ \int_{-\infty}^{+\infty}f_{X|Y}(x\,|\,y)dx=1,\ \int_{-\infty}^{+\infty}f_{Y|X}(y\,|\,x)dy=1. \qquad \textbf{(1.4.3)}$$

В) Убедитесь, что для независимых случайных величин X и Y с плотностями $f_X(x)$, $x\in R^1$, и $f_Y(y)$, $y\in R^1$, соответственно функция распределения их суммы, то есть с.в. $U=X+Y$, вычисляется по формуле

$$F_U(u)=\int_{-\infty}^{+\infty}F_X(u-v)dF_Y(v), \qquad \textbf{(1.4.4)}$$

а плотность как

$$f_U(u) = \int_{-\infty}^{+\infty} f_X(u-v) f_Y(v) dv\,. \qquad \textbf{(1.4.5)}$$

Г) Убедитесь, что для независимых случайных величин X и Y с плотностями $f_X(x)$, $x \in R^1$, и $f_Y(y)$, $y \in R^1$, соответственно плотность распределения вероятностей их отношения, то есть с.в. $U = X/Y$, вычисляется по формуле

$$f_U(u) = \int_{-\infty}^{+\infty} f_X(uv) f_Y(v) |v| dv\,. \qquad \textbf{(1.4.6)}$$

Решение. А) Согласно определению ф.р. $F_X(x)$ случайной величины X с плотностью $f_X(x)$, $x \in R^1$, определяется выражением вида

$$F_X(x) = \int_{-\infty}^{x} f_X(x) dx\,.$$

С другой стороны, используя свойства двумерной функции распределения, получаем

$$F_X(x) = F_{XY}(x, +\infty) = \int_{-\infty}^{x} \left(\int_{-\infty}^{+\infty} f_{XY}(x,y) dy \right) dx = \int_{-\infty}^{x} f_X(x) dx\,.$$

Отсюда следует, что

$$\int_{-\infty}^{+\infty} f_{XY}(x,y) dy = f_X(x)\,.$$

Аналогично проверяется второе равенство в (1.4.2).

Б) Так как

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x,y) dx dy\,,$$

то, используя свойство двумерной функции распределения, согласно которому $F_{XY}(+\infty, +\infty) = 1$, имеем

$$1 = F_{XY}(+\infty, +\infty) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y) dx dy\,.$$

Учитывая далее, что условная плотность $f_{X|Y}(x|y) = f_{XY}(x,y)/f_Y(y)$, и используя (1.4.2) , получаем

$$\int_{-\infty}^{+\infty} f_{X|Y}(x|y) dx = \int_{-\infty}^{+\infty} \frac{f_{XY}(x,y)}{f_Y(y)} dx = \frac{1}{f_Y(y)} \int_{-\infty}^{+\infty} f_{XY}(x,y) dx = \frac{1}{f_Y(y)} \cdot f_Y(y) = 1\,.$$

Аналогично проверяется третье равенство в (1.4.3).

В) Так как X и Y – независимые случайные величины, то $f_{XY}(x,y)=f_X(x)f_Y(y)$. Запишем совместную плотность случайных величин $U=X+Y$ и $V=Y$. Для этого используем преобразования в виде замены переменных $u=x+y$, $v=y$. Обратные преобразования запишутся в виде $x=u-v$, $y=v$. Якобиан преобразований

$$J=\begin{vmatrix}\dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v}\\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v}\end{vmatrix}=\begin{vmatrix}1 & -1\\ 0 & 1\end{vmatrix}=1.$$

Далее, согласно (1.4.77), совместная плотность случайных величин $U=X+Y$ и $V=Y$ запишется в виде

$$f_{UV}(u,v)=f_X(u-v)f_Y(v)\cdot|J|,$$

отсюда с использованием (1.4.2) получаем плотность с.в. U в виде свертки, то есть

$$f_U(u)=\int\limits_{-\infty}^{+\infty}f_{UV}(u,v)dv=\int\limits_{-\infty}^{+\infty}f_X(u-v)f_Y(v)dv.$$

Используя определение функции распределения вероятностей, получаем

$$F_U(u)=\int\limits_{-\infty}^{u}f_U(z)dz=\int\limits_{-\infty}^{u}\Big(\int\limits_{-\infty}^{+\infty}f_X(z-v)f_Y(v)dv\Big)dz=$$

$$=\int\limits_{-\infty}^{+\infty}f_Y(v)\Big(\int\limits_{-\infty}^{u}f_X(z-v)dz\Big)\,dv=\int\limits_{-\infty}^{+\infty}F_X(u-v)dF_Y(v).$$

Г) Так как X и Y – независимые случайные величины, то $f_{XY}(x,y)=f_X(x)f_Y(y)$. Запишем совместную плотность случайных величин $U=X/Y$ и $V=Y$. Для этого используем преобразования в виде замены переменных $u=x/y$, $v=y$. Обратные преобразования запишутся в виде $x=uv$, $y=v$. Якобиан преобразований

$$J=\begin{vmatrix}\dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v}\\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v}\end{vmatrix}=\begin{vmatrix}v & u\\ 0 & 1\end{vmatrix}=v.$$

Далее, согласно (1.4.77), совместная плотность случайных величин $U=X/Y$ и $V=Y$ запишется в виде

$$f_{UV}(u,v)=f_X(uv)f_Y(v)\cdot|J|,$$

отсюда с использованием (1.4.2) получаем плотность с.в. U в виде

$$f_U(u) = \int_{-\infty}^{+\infty} f_{UV}(u,v)dv = \int_{-\infty}^{+\infty} f_X(uv) f_Y(v) \mid v \mid dv .$$

1.4.7. Пусть $Z = (X,Y)$ – непрерывная двумерная случайная величина с плотностью $f_{XY}(x,y)$, $(x,y) \in R^2$, и пусть $f_X(x)$, $x \in R^1$, $f_Y(y)$, $y \in R^1$, – маргинальные плотности случайных величин X и Y. Убедитесь, что математическое ожидание суммы случайных величин X и Y равно сумме их математических ожиданий, то есть $M(X+Y) = M(X) + M(Y)$.

Решение. Используя формулу (1.3.68) для вычисления математического ожидания функции случайных величин, имеем

$$M(X+Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x+y) f_{XY}(x,y)dxdy =$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x \cdot f_{XY}(x,y)dxdy + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y \cdot f_{XY}(x,y)dxdy =$$

$$= \int_{-\infty}^{+\infty} x\left(\int_{-\infty}^{+\infty} f_{XY}(x,y)dy\right)dx + \int_{-\infty}^{+\infty} y\left(\int_{-\infty}^{+\infty} f_{XY}(x,y)dx\right)dy =$$

$$= \int_{-\infty}^{+\infty} x f_X(x)dx + \int_{-\infty}^{+\infty} y f_Y(y)dy = M(X) + M(Y).$$

1.4.8. Пусть случайные величины X и Y связаны линейно, то есть $Y = aX + b$, где a и b константы. Убедитесь, что $D(Y) = a^2 D(X)$.

Решение. Согласно определению, дисперсия с.в. X равна $D(X) = M[X - M(X)]^2 = M(X^2) - [M(X)]^2$. Используя это определение и свойства математического ожидания, получаем

$$D(Y) = D(aX+b) = M[(aX+b)^2] - [M(aX+b)]^2 =$$

$$= M[a^2 + 2abX + b^2X^2] - [b + aM(X)]^2 =$$

$$= b^2 + 2abM(X) + a^2M(X^2) - b^2 - 2abM(X) - a^2[M(X)]^2 =$$

$$= a^2\{M(X^2) - [M(X)]^2\} = a^2 D(X).$$

1.4.9. Пусть V и W – случайные величины и с.в. X, Y линейно связаны с ними, то есть $X = aV + b$ и $Y = dW + c$. Убедитесь, что $COV(X,Y) = adCOV(V,W)$.

Решение.

$$COV(X,Y) = M\{[X - M(X)][Y - M(Y)]\} =$$

$$M\{[aV + b - aM(V) - b][dW + c - dM(W) - c]\} =$$

$$= adM\{[V - M(V)][W - M(W)]\} = adCOV(V,W).$$

1.4.10. Пусть случайные величины X и Y связаны линейно, то есть $Y = aX + b$, где a и b константы. Убедитесь, что при такой связи коэффициент корреляции ρ_{XY} равен $(+1)$ *при* $a > 0$ (положительная корреляция) и равен (-1) *при* $a < 0$ (отрицательная корреляция), то есть

$$\rho_{XY} = \frac{COV(X,Y)}{\sqrt{D(X)D(Y)}} = \frac{a}{|a|} = \begin{cases} +1, & a > 0, \\ -1, & a < 0. \end{cases} \qquad \textbf{(1.4.11)}$$

Решение. Согласно свойствам математического ожидания и дисперсии, имеем

$$M(Y) = aM(X) + b,\ D(Y) = a^2 D(X) \text{ и } \sqrt{D(Y)} = |a|\sqrt{D(X)}.$$

Далее, ковариация случайных величин X и Y равна

$$COV(X,Y) = M\{[X - M(X)][Y - M(Y)]\} =$$
$$M\{[X - M(X)][aX + b - aM(X) - b]\} = a \cdot M\{[X - M(X)]^2\} = a \cdot D(X).$$

С учетом полученных выражений имеем

$$\rho_{XY} = \frac{COV(X,Y)}{\sqrt{D(X)D(Y)}} = \frac{a \cdot D(X)}{|a|\sqrt{D(X)}\sqrt{D(X)}} = \frac{a}{|a|} = \begin{cases} +1, & a > 0, \\ -1, & a < 0. \end{cases}$$

Таким образом, коэффициент корреляции ρ_{XY}, для которого из неравенства Коши – Буняковского следуют неравенства $-1 \le \rho_{XY} \le +1$, характеризует линейную стохастическую связь между случайными величинами X и Y, причем чем ближе коэффициент корреляции ρ_{XY} к $(+1)$ или к (-1), тем ближе группируются выборочные точки $(x_1, y_1), ..., (x_n, y_n)$ вдоль прямой $y = ax + b$ на корреляционном поле двумерной случайной величины (X, Y).

1.4.12. Убедитесь, что выполняются следующие равенства:

А) $D(X + Y) = D(X) + D(Y) + 2\rho_{XY}\sqrt{D(X)D(Y)}$;

Б) $M\left(a_0 + \sum_{i=1}^{n} a_i X_i\right) = a_0 + \sum_{i=1}^{n} a_i M(X_i)$;

В) $D\left\{a_0 + \sum_{i=1}^{n} a_i X_i\right\} = \sum_{i=1}^{n} a_i^2 D(X_i) + \sum_{i=1}^{n} \sum_{j=1,\, i \ne j}^{n} a_i a_j COV(X_i, X_j)$;

Г) Для независимых случайных величин X_i и X_j $(i \ne j)$ дисперсия их суммы равна сумме их дисперсий, то есть

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

Решение. Используя определение дисперсии, ковариации, коэффициента корреляции и свойства математического ожидания, получаем

А) $D(X+Y)=M[(X+Y)^2]-[M(X+Y)]^2=$

$=M(X^2)+2M(XY)+M(Y^2)-[M(X^2)]-2M(X)M(Y)-[M(Y^2)]=$

$=D(X)+D(Y)+2COV(X,Y)=D(X)+D(Y)+2\rho_{XY}\sqrt{D(X)D(Y)}$.

Б) $M\left(a_0+\sum_{i=1}^{n}a_iX_i\right)=\int_{-\infty}^{+\infty}(a_0+\sum a_i\xi)f_{X_i}(\xi)d\xi=$

$=a_0\int_{-\infty}^{+\infty}f_{X_i}(\xi)d\xi+\sum_{i=1}^{n}a_i\int_{-\infty}^{+\infty}\xi\cdot f_{X_i}(\xi)d\xi=a_0+\sum_{i=1}^{n}a_iM(X_i).$

В) $D\left\{a_0+\sum_{i=1}^{n}a_iX_i\right\}=M\left\{\left[(a_0+\sum a_iX_i)-M(a_0-\sum a_iX_i)\right]^2\right\}=$

$=M\left\{\left[\sum a_iX_i-\sum a_iM(X_i)\right]^2\right\}=M\left\{\left[\sum a_i(X_i-M(X_i))\right]^2\right\}=$

$=M\left\{\sum_{i=1}^{n}a_i^2[X_i-M(X_i)]^2+\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}a_ia_j[X_i-M(X_i)][X_j-M(X_j)]\right\}=$

$=\sum_{i=1}^{n}a_i^2D(X_i)+\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}a_ia_jCOV(X_i,X_j)$. **(1.4.13)**

Учитывая, что $COV(X,X)=D(X)$, формула (1.4.13) запишется в виде

$$D\left\{a_0+\sum_{i=1}^{n}a_iX_i\right\}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jCOV(X_iX_j).$$

Отметим, что

$$\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}a_ia_jCOV(X_iX_j)=2\sum_{i=1}^{n-1}\sum_{j=2,i<j}^{n}a_ia_jCOV(X_iX_j). \quad \textbf{(1.4.14)}$$

Г) В частном случае при $a_0=0$, $a_1=a_2=...=a_n=1$ из (1.4.13) получаем

$$D\left\{\sum_{i=1}^{n}X_i\right\}=\sum_{i=1}^{n}D(X_i)+\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}COV(X_i,X_j). \quad \textbf{(1.4.15)}$$

Далее, из формулы (1.4.15) следует, что если случайные величины X_i и X_j $(i\neq j)$ *независимы* и, следовательно, *некоррелированы*, то $COV(X_i,X_j)=0$ $(i\neq j)$, и тогда дисперсия суммы *независимых* случайных величин равна сумме их дисперсий, то есть

$$D\left\{\sum_{i=1}^{n}X_i\right\}=\sum_{i=1}^{n}D(X_i). \quad \textbf{(1.4.16)}$$

1.4.17. Пусть $Z=(X,Y)$ – непрерывная двумерная случайная величина с плотностью $f_{XY}(x,y)$, $(x,y)\in R^2$. Пусть далее $f_{X|Y}(x\,|\,y)$ условная плотность и $f_X(x)$, $x\in R^1$, $f_Y(y)$, $y\in R^1$, – маргинальные плотности случайных величин X и Y. Убедитесь, что справедлива формула полного математического ожидания в виде

$$M\{M(X\,|\,Y)\}=M(X). \qquad \textbf{(1.4.18)}$$

Решение. Рассмотрим случайную величину $\varphi_X(Y)=M(X\,|\,Y)$ и вычислим ее математическое ожидание

$$M\{\varphi_X(Y)\}=M\{M(X\,|\,Y)\}=\int\limits_{-\infty}^{+\infty}M(X\,|\,y)f_Y(y)dy=\int\limits_{-\infty}^{+\infty}\left(\int\limits_{-\infty}^{+\infty}xf_X(x\,|\,y)dx\right)f_Y(y)dy=$$

$$=\int\limits_{-\infty}^{+\infty}\left(\int\limits_{-\infty}^{+\infty}x\frac{f_{XY}(x,y)}{f_Y(y)}dx\right)f_Y(y)dy=\int\limits_{-\infty}^{+\infty}x\left(\int\limits_{-\infty}^{+\infty}f_{XY}(x,y)dy\right)dx=\int\limits_{-\infty}^{+\infty}xf_X(x)dx=M(X).$$

1.4.19. Покажите, что для независимых случайных величин X и Y выполняется равенство

$$\frac{D(XY)}{\{M(X)M(Y)\}^2}=V_X^2\cdot V_Y^2+V_X^2+V_Y^2,$$

где $V_X=\sqrt{D(X)}/M(X)$ – коэффициент вариации случайной величины X.

1.4.20. А) Пусть случайная величина X имеет нормальное распределение с параметрами a и σ^2, коротко этот факт записывают в виде $L(X)=N(a;\sigma^2)$, и эта запись означает, что плотность распределения вероятностей $f_X(x)$ с.в. X определяется в виде

$$f_X(x;a,\sigma^2)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right)=\frac{1}{\sigma}\varphi\left(\frac{x-a}{\sigma}\right),\ -\infty<x<+\infty, \qquad \textbf{(1.4.21)}$$

где $\varphi(\cdot)$ – стандартная нормальная плотность, определенная в (1.4.26). Убедитесь, что параметры a и σ^2 этого распределения соответственно равны математическому ожиданию и дисперсии случайной величины X, то есть $M(X)=a$ и $D(X)=\sigma^2$. Убедитесь также, что для стандартной нормальной случайной величины $\tilde{X}=(X-a)/\sigma$ все моменты нечетного порядка $(2k+1)$, где $k>0$ – целое число, равны нулю, то есть

$$M(\tilde{X}^{2k+1})=0,\ k=1,2,\ldots,$$

а моменты четного порядка $2k$ вычисляются по формулам

$$M(\tilde{X}^{2k}) = \frac{2^k}{\sqrt{\pi}}\Gamma\left(k+\frac{1}{2}\right),\ k=1,2,...,$$

где $\Gamma(\cdot)$ – гамма-функция, определяемая формулой (1.4.36).

Б) Пусть случайная величина X имеет экспоненциальное распределение с параметром $\lambda > 0$, то есть плотность распределения вероятностей $f_X(x)$ с.в. X определяется в виде $f_X(x) = \lambda\exp(-\lambda x)$, $0 \le x < \infty$, $\lambda > 0$, а функция распределения имеет вид $F_X(x) = 1 - \exp(-\lambda x)$, $0 \le x < \infty$, $\lambda > 0$. Вычислить математическое ожидание, дисперсию и квантиль x_p уровня $0 < p < 1$.

Решение. Рассмотрим сначала гауссовский случай, то есть вариант А. Используя определения математического ожидания и дисперсии, получаем

$$M(X) = \int_{-\infty}^{+\infty} x f_X(x)dx = \int_{-\infty}^{+\infty} x\,\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right)dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty}(\sigma y + a)e^{-y^2/2}\sigma dy =$$

$$= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y e^{-y^2/2}\,dy + \frac{a}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-y^2/2}\,dy = \frac{\sigma}{\sqrt{2\pi}}\cdot 0 + a\cdot 1 = a.$$

$$D(X) = \int_{-\infty}^{+\infty}(x - M(X))^2 f_X(x)dx = \int_{-\infty}^{+\infty}(x-a)^2\,\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right)dx =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\left(\frac{x-a}{\sigma}\right)^2\exp\left(-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2\right)d\left(\frac{x-a}{\sigma}\right) = \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y^2 e^{-y^2/2}\,dy =$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} y\,d(e^{-y^2/2}) = -\frac{\sigma^2}{\sqrt{2\pi}}\,y e^{-y^2/2}\Big|_{-\infty}^{+\infty} + \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-y^2/2}\,dy = 0 + \sigma^2\cdot 1 = \sigma^2.$$

Отметим, что если $L(X) = N(a;\sigma^2)$, то функция распределения $F_X(x)$ этой случайной величины X записывается в виде

$$F_X(x) = \Phi\left(\frac{x-a}{\sigma}\right),\ -\infty < x < +\infty,$$

где $\Phi(\cdot)$ – стандартная нормальная функция распределения, определенная в (1.4.27). Поэтому вероятность события, что с.в. $X \in [x_1, x_2]$, вычисляется по формуле

$$P\{x_1 \le X \le x_2\} = F_X(x_2) - F_X(x_1) = \Phi\left(\frac{x_2-a}{\sigma}\right) - \Phi\left(\frac{x_1-a}{\sigma}\right),$$

и, в частности при $x_1 = a - 3\sigma$ и $x_2 = a + 3\sigma$, получаем

$$P\{a - 3\sigma \le X \le a + 3\sigma\} = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 2\cdot 0{,}9986 - 1 \approx 0{,}997.$$

Таким образом, нормальная случайная величина с параметрами a и σ^2 практически достоверно лежит в интервале $(a-3\sigma, a+3\sigma)$. Этот факт часто в литературе называют правилом «трех сигм».

Вычислим теперь начальные моменты стандартной нормальной случайной величины $\tilde{X}=(X-a)/\sigma$, для которой $M(\tilde{X})=0$, $D(\tilde{X})=1$ и ее плотность определяется формулой (1.4.26). Все моменты нечетного порядка $(2k+1)$ равны нулю, то есть

$$M(\tilde{X}^{2k+1})=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty}x^{2k+1}e^{-\frac{1}{2}x^2}dx=0,\ k=1,2,...,$$

так как при их вычислении интегрируется нечетная функция в симметричных пределах. Моменты четного порядка $2k$ вычисляются по формулам

$$M(\tilde{X}^{2k})=\frac{2}{\sqrt{2\pi}}\int\limits_{0}^{\infty}x^{2k}e^{-\frac{1}{2}x^2}dx=\|\ x=\sqrt{2t}\ ,dx=dt/\sqrt{2t}\ \|=$$

$$=\frac{2^k}{\sqrt{\pi}}\int\limits_{0}^{\infty}t^{(k+1/2)-1}e^{-t}dt=\frac{2^k}{\sqrt{\pi}}\Gamma\left(k+\frac{1}{2}\right),\ k=1,2,....$$

Например, учитывая, что $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$, $\Gamma(1/2)=\sqrt{\pi}$, при $k=1$ получаем $M(\tilde{X}^2)=1$, а при $k=2$ – $M(\tilde{X}^4)=3$.

Б) Для экспоненциального распределения имеем следующие формулы.

$$M(X)=\int\limits_{0}^{+\infty}x\lambda\exp(-\lambda x)dx=-\int\limits_{0}^{+\infty}xd(\exp(-\lambda x))=$$

$$-x\exp(-\lambda x)\,|_0^{\infty}+\int\limits_{0}^{\infty}\exp(-\lambda x)dx=0-\frac{1}{\lambda}\exp(-\lambda x)\,|_0^{\infty}=\frac{1}{\lambda}.$$

$$M(X^2)=\int\limits_{0}^{+\infty}x^2\lambda\exp(-\lambda x)dx=-\int\limits_{0}^{+\infty}x^2d(\exp(-\lambda x))=$$

$$=-x^2\exp(-\lambda x)\,|_0^{\infty}+2\int\limits_{0}^{\infty}x\exp(-\lambda x)dx=\frac{2}{\lambda}M(X)=\frac{2}{\lambda^2}.$$

$$D(X)=M(X^2)-[M(X)]^2=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}.$$

Квантиль x_p уровня $0<p<1$ определяется выражением

$$F_X(x_p)=1-\exp(-\lambda x_p)=p.$$

Отсюда $x_p = -\frac{1}{\lambda}\ln(1-p)$, $0 < p < 1$.

1.4.22. Пусть (X,Y) – непрерывная двумерная случайная величина с плотностью $f_{XY}(x,y)$, $(x,y) \in R^2$, и пусть $f_X(x)$, $x \in R^1$, $f_Y(y)$, $y \in R^1$, – маргинальные плотности случайных величин X и Y. Убедитесь, что выполняются следующие утверждения:

А) из независимости случайных величин X и Y следует их некоррелированность;

Б) из некоррелированности случайных величин X и Y в общем случае не следует их независимость, исключение составляет гауссовский случай, для которого из некоррелированности следует независимость.

Решение. Рассмотрим сначала вариант А. Согласно определению (1.3.47), ковариация случайных величин X и Y равна $COV(X,Y)=M(XY)-M(X)M(Y)$, которая с учетом (1.3.48) и (1.3.11) записывается в виде

$$COV(X,Y) = M(XY) - M(X)M(Y) =$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{XY}(x,y)dxdy - \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy .$$

Далее, если случайные величины X и Y независимы, то $f_{XY}(x,y) = f_X(x)\cdot f_Y(y)$ и, следовательно,

$$M(XY) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{XY}(x,y)dxdy = \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy = M(X)M(Y).$$

Таким образом, из независимости случайных величин X и Y получаем, что $COV(X,Y)=0$ и, следовательно, $\rho_{XY}=0$, то есть из независимости следует некоррелированность случайных величин X и Y.

Б) Для доказательства того, что из некоррелированности случайных величин X и Y в общем случае не следует их независимость, приведем пример, взятый из [9, с.70]. Пусть (X,Y) – дискретная двумерная случайная величина с возможными состояниями $(-1,0)$, $(0,1)$, $(0,-1)$, $(1,0)$ и вероятности этих состояний $p_{XY}(\tilde{x}_i,\tilde{y}_j)=1/4$, $i,j=1,...,4$. Убедимся, что в данном случае $COV(X,Y)=0$. В самом деле,

$$M(XY) = \sum_{i=1}^{4}\sum_{j=1}^{4} \tilde{x}_i\tilde{y}_j p_{XY}(\tilde{x}_i,\tilde{y}_j) =$$

$$= (-1)\cdot 0\cdot(1/4) + 0\cdot 1\cdot(1/4) + 0\cdot(-1)\cdot(1/4) + 1\cdot 0\cdot(1/4) = 0 .$$

$$M(X) = \sum_{i=1}^{4} \tilde{x}_i p_X(\tilde{x}_i) = (-1)\cdot(1/4) + 0\cdot(1/2) + 1\cdot(1/4) = 0 .$$

Аналогично получаем $M(Y)=0$. Таким образом,

$$COV(X,Y)=M(XY)-M(X)M(Y)=0-0\cdot 0=0.$$

Итак, ковариация случайных величин X и Y равна нулю и, следовательно, $\rho_{XY}=0$, то есть эти случайные величины некоррелированы, но они зависимы, так как для них не выполняется равенство $p_{XY}(\tilde{x}_i,\tilde{y}_j)=p_X(\tilde{x}_i)\cdot p_Y(y_j)$, $i,j=1,...,4$. Например, для $\tilde{x}_1=-1$ и $\tilde{y}_1=0$ имеем $p_{XY}(\tilde{x}_1,\tilde{y}_1)=1/4$, $p_X(\tilde{x}_1)=1/4$, $p_Y(\tilde{y}_1)=1/2$.

Рассмотрим теперь гауссовский случай, то есть предполагаем, что $L(X,Y)=NN(a_X,a_Y,\sigma_X^2,\sigma_Y^2,\rho)$. Эта запись означает, что двумерная случайная величина $Z=(X,Y)$ имеет двумерное нормальное распределение вероятностей с плотностью

$$f_{XY}(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\times$$

$$\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-a_X}{\sigma_X}\right)^2-2\rho\left(\frac{x-a_X}{\sigma_X}\right)\left(\frac{y-a_Y}{\sigma_Y}\right)+\left(\frac{y-a_Y}{\sigma_Y}\right)^2\right]\right\},$$

$$(x,y)\in R^2, \qquad \textbf{(1.4.23)}$$

где $a_X=M(X)$, $a_Y=M(Y)$, $\sigma_X^2=D(X)$, $\sigma_Y^2=D(Y)$ и ρ – коэффициент корреляции. Пусть случайные величины X и Y некоррелированы, то есть $\rho_{XY}=0$. Тогда выполняется равенство

$$f_{XY}(x,y)=\frac{1}{\sqrt{2\pi}\,\sigma_X}\exp\left\{-\frac{1}{2}\left(\frac{x-a_X}{\sigma_X}\right)^2\right\}\frac{1}{\sqrt{2\pi}\,\sigma_Y}\exp\left\{-\frac{1}{2}\left(\frac{y-a_Y}{\sigma_Y}\right)^2\right\}=$$

$$=f_X(x)\cdot f_Y(y),$$

из которого следует независимость случайных величин X и Y.

1.4.24. Пусть X и Y – случайные величины с математическими ожиданиями $M(X)$, $M(Y)$ и дисперсиями $D(X)$, $D(Y)$. Пусть, далее, $\tilde{X}=(X-M(X))/\sqrt{D(X)}$, $\tilde{Y}=(Y-M(Y))/\sqrt{D(Y)}$ – стандартизованные варианты этих случайных величин, при этом их математические ожидания равны нулю, а дисперсии равны единице. Убедитесь, что коэффициент корреляции случайных величин X,Y равен ковариации их стандартизованных вариантов, то есть $\rho_{XY}=COV(\tilde{X},\tilde{Y})$.

Решение. Используя определение ковариации, коэффициента корреляции и свойства математического ожидания и учитывая, что $M(\tilde{X})=M(\tilde{Y})=0$, получаем

$$COV(\tilde{X},\tilde{Y}) = M(\tilde{X}\tilde{Y}) - M(\tilde{X})M(\tilde{Y}) = M\left\{\frac{X-M(X)}{\sqrt{D(X)}}\cdot\frac{Y-M(Y)}{\sqrt{D(Y)}}\right\} =$$

$$= \frac{1}{\sqrt{D(X)D(Y)}}\, M\{[X-M(X)][Y-M(Y)]\} = \frac{COV(X,Y)}{\sqrt{D(X)D(Y)}} = \rho_{XY}.$$

1.4.25. Пусть $L(X) = N(0,1)$. Эта краткая запись означает, что с.в. X имеет стандартное нормальное распределение вероятностей с плотностью

$$f_X(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},\ x \in R^1, \qquad \textbf{(1.4.26)}$$

и функцией распределения вероятностей $F_X(x) = \Phi(x)$, где $\Phi(x)$ – стандартная нормальная ф.р. вида

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt,\ x \in R^1. \qquad \textbf{(1.4.27)}$$

Пусть, далее, $L(X^2(n)) = \chi^2(n)$. Эта запись означает, что с.в. $X^2(n)$ имеет распределение хи-квадрат с n степенями свободы и ее плотность распределения вероятностей (см. П.2.78) записывается в виде

$$f_{X^2(n)}(x;n) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)},\ 0 < x < \infty. \qquad \textbf{(1.4.28)}$$

Убедитесь, что случайная величина X^2, равная квадрату стандартной нормальной величины X, имеет распределение хи-квадрат с числом степеней свободы равным единице, то есть $L(X^2) = L(X^2(1)) = \chi^2(1)$ и, следовательно,

$$f_{X^2}(x) = f_{X^2(1)}(x;1) = \frac{e^{-x/2}}{\sqrt{x}\cdot\sqrt{2\pi}},\ 0 < x < \infty. \qquad \textbf{(1.4.29)}$$

Решение. Используя определения функции распределения и плотности, получаем

$$F_{X^2}(x) = P\{X^2 \le x\} = P\{-\sqrt{x} \le X \le \sqrt{x}\} = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1,$$

$$f_{X^2}(x) = dF_{X^2}(x)/dx = \frac{2\varphi(\sqrt{x})}{2\sqrt{x}} = \frac{e^{-x/2}}{\sqrt{x}\cdot\sqrt{2\pi}} = f_{X^2(1)}(x;1),\ 0 < x < \infty.$$

Отсюда в частности следует, что квантиль $\chi^2_{p,1}$ уровня p, $0 < p < 1$, распределения хи-квадрат с одной степенью свободы равен квадрату квантиля уровня $(1+p)/2$ стандартного нормального распределения, то есть $\chi^2_{p,1} = \lambda^2_{(1+p)/2}$ (см. также упр. 3.9.62). В самом деле,

$p = F_{X^2}(\chi^2_{p,1}) = 2\Phi(\sqrt{\chi^2_{p,1}}) - 1$, или $\sqrt{\chi^2_{p,1}} = \Phi^{-1}[(1+p)/2] = \lambda_{(1+p)/2}$. Отсюда получаем требуемое равенство $\chi^2_{p,1} = \lambda^2_{(1+p)/2}$. Например, если $p = 0{,}95$, то $(1+p)/2 = 0{,}975$ и, следовательно, $\chi^2_{0{,}95\ ,1} = 0{,}384 = (1{,}96)^2 = \lambda^2_{0{,}975}$.

1.4.30. Пусть $L(X) = Be(n,m)$. Эта краткая запись означает, что с.в. X имеет бета-распределение (см. П.2.9) с целыми и положительными параметрами n и m, а ее плотность распределения вероятностей записывается в виде

$$f_X(x;n,m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} x^{n-1}(1-x)^{m-1},\ x \in (0,1). \qquad \textbf{(1.4.31)}$$

Отметим, что функция распределения $F_X(x)$, $0 < x < 1$, этой случайной величины может быть выражена через неполную бета-функцию $I_p(n,m)$, которая определяется в виде

$$I_p(n,m) = \frac{1}{B(n,m)} \int_0^p x^{n-1}(1-x)^{m-1} dx = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} \int_0^p x^{n-1}(1-x)^{m-1} dx, \qquad \textbf{(1.4.32)}$$

и для которой подстановка $t = 1 - x$ приводит к следующему тождеству:

$$I_p(n,m) = 1 - I_{1-p}(m,n). \qquad \textbf{(1.4.33)}$$

В самом деле, функция распределения $F_X(x)$, $0 < x < 1$, случайной величины X, которая имеет бета-распределение, записывается в форме

$$F_X(x;n,m) = \frac{1}{B(n,m)} \int_0^x t^{n-1}(1-t)^{m-1} dt = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} \int_0^x t^{n-1}(1-t)^{m-1} dt = I_x(n\ ,m),$$
$$n > 0\ , m > 0\ ,\ 0 < x < 1, \qquad \textbf{(1.4.34)}$$

где $B(n,m)$ – бета-функция, определяемая в виде

$$B(n,m) = \int_0^1 x^{n-1}(1-x)^{m-1} dx = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)}, \qquad \textbf{(1.4.35)}$$

и $\Gamma(\cdot)$ – гамма-функция, определяемая в виде

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx. \qquad \textbf{(1.4.36)}$$

Отметим, что $\Gamma(n) = (n-1)\Gamma(n-1)$, $\Gamma(n) = (n-1)!$, $\Gamma(1) = 1$, $\Gamma(2) = 1$, $\Gamma(1/2) = \sqrt{\pi}$ (см. упр. 4.11.44).

Убедитесь, что для бета-распределения с параметрами n и m математическое ожидание и дисперсия вычисляются по формулам

$$M(X)=n/(n+m),\ D(X)=nm/(n+m)^2(n+m+1). \qquad \textbf{(1.4.37)}$$

Решение. Вычислим начальный момент k-го порядка. Используя (1.4.31), получаем

$$M(X^k)=\int_0^1 x^k f_X(x;n,m)dx=\frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)}\int_0^1 x^k\cdot x^{n-1}(1-x)^{m-1}dx=$$
$$=\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}B(n+k\,,m)=\frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)}\,\frac{\Gamma(n+k)\Gamma(m)}{\Gamma(n+m+k)}=$$
$$=\frac{(m+n-1)!}{(m-1)!(n-1)!}\cdot\frac{(n+k-1)!(m-1)!}{(m+n+k-1)!}=\frac{(m+n-1)!}{(n-1)!}\cdot\frac{(n+k-1)!}{(m+n+k-1)!}. \qquad \textbf{(1.4.38)}$$

Отсюда, при $k=1$ получаем

$$M(X)=\frac{(m+n-1)!}{(n-1)!}\cdot\frac{n!}{(m+n)!}=\frac{n}{n+m}\ .$$

Далее, при $k=2$ получаем

$$M(X^2)=\frac{(m+n-1)!}{(n-1)!}\cdot\frac{(n+1)!}{(m+n+1)!}=\frac{n(n+1)}{(m+n)(m+n+1)}.$$

Используя найденные формулы, выражение для дисперсии запишем в виде

$$D(X)=M(X^2)-[M(X)]^2=\frac{n(n+1)}{(m+n)(m+n+1)}-\frac{n^2}{(m+n)^2}=\frac{nm}{(n+m)^2(n+m+1)}.$$

1.4.39. Пусть $L(X)=\Gamma(\alpha\,,\lambda)$. Эта краткая запись означает (см. П.2.18), что с. в. X имеет гамма-распределение с параметром формы $\alpha>0$ и масштабным параметром $\lambda>0$ и ее плотность распределения вероятностей записывается в виде

$$f_X(x;\alpha,\lambda)=\begin{cases}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}/\Gamma(\alpha)\,, & x>0,\\ 0\,, & x\le 0\,.\end{cases} \qquad \textbf{(1.4.40)}$$

А) Убедитесь, что распределение хи-квадрат с числом степеней свободы n совпадает с гамма-распределением при $\alpha=n/2$ и $\lambda=1/2$, то есть $\Gamma(n/2\,,1/2)\sim\chi^2(n)$.

Б) Убедиться, что функция распределения с. в. X записывается в виде

$$F_X(x;\alpha,\lambda)=\begin{cases}\gamma(\alpha\,,\lambda x)/\Gamma(\alpha)\,, & x>0,\\ 0\,, & x\le 0\,,\end{cases} \qquad \textbf{(1.4.41)}$$

где $\Gamma(\alpha)$ – гамма-функция и $\gamma(\alpha\,,x)$ обозначает неполную гамма-функцию,

определяемую (см. [15, (8.350)]) в виде

$$\gamma(\alpha\,,\,x)=\int_0^x t^{\alpha-1}e^{-t}dt\,. \qquad \textbf{(1.4.42)}$$

В) Убедитесь в справедливости следующих формул:

$$M(X^k)=\frac{\Gamma(k+\alpha)}{\lambda^k\Gamma(\alpha)},\ k=1\,,\,2\,,\ldots \qquad \textbf{(1.4.43)}$$

$$M(X)=\alpha/\lambda\,,\ \ M(X^2)=\alpha(\alpha+1)/\lambda^2,\ D(X)=\alpha/\lambda^2\,. \qquad \textbf{(1.4.44)}$$

Решение. А) Факт эквивалентности гамма-распределения и хи-квадрат-распределения, точнее выражение $\Gamma(n/2\,,1/2)\sim\chi^2(n)$, проверяется непосредственно с помощью формул (1.4.40) и (1.4.28).

Б) Используя определение функции распределения вероятностей, для $x>0$ получаем

$$F_X(x;\alpha,\lambda)=\int_0^x f_X(z;\alpha,\lambda)dz=\int_0^x[\lambda^\alpha z^{\alpha-1}e^{-\lambda z}/\Gamma(\alpha)]dz=\|\,\lambda z=t\big|_0^{\lambda x}\,,\,\lambda dz=dt\,\|=$$

$$=\frac{1}{\Gamma(\alpha)}\int_0^{\lambda x}\lambda^\alpha(t/\lambda)^{\alpha-1}e^{-t}d(t/\lambda)=\frac{1}{\Gamma(\alpha)}\int_0^{\lambda x}t^{\alpha-1}e^{-t}dt=\frac{\gamma(\alpha\,,\,\lambda x)}{\Gamma(\alpha)},\ 0<x<\infty\,.$$

В) Начальные моменты *k*-го порядка вычисляются по формулам

$$\alpha_k(\lambda)=M(X^k)=\int_0^{+\infty}x^k f_X(x;\alpha,\lambda)dx=\frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^{+\infty}x^{k+\alpha-1}e^{-\lambda x}dx=$$

$$=\|\,t=\lambda x\,,\,dx=dt/\lambda\,\|\,=\frac{\lambda^\alpha}{\Gamma(\alpha)\lambda^{k+\alpha-1}\cdot\lambda}\int_0^{+\infty}t^{k+\alpha-1}e^{-t}dt=\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\lambda^k},\ k=1\,,\,2\,,\ldots.$$

Далее, учитывая, что $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$, при $k=1$ получаем

$$\alpha_1(\lambda)=M(X)=\frac{\Gamma(\alpha+1)}{\lambda\cdot\Gamma(\alpha)}=\frac{\alpha\cdot\Gamma(\alpha)}{\lambda\cdot\Gamma(\alpha)}=\frac{\alpha}{\lambda},$$

а при $k=2$ –

$$\alpha_2(\lambda)=M(X^2)=\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2}=\frac{(\alpha+1)\cdot\alpha\cdot\Gamma(\alpha)}{\lambda^2\cdot\Gamma(\alpha)}=\frac{(\alpha+1)\alpha}{\lambda^2}.$$

Отсюда центральный момент второго порядка μ_2, или дисперсия, вычисляется по формуле

$$\mu_2=\alpha_2-\alpha_1^2=D(X)=\frac{\alpha^2+\alpha}{\lambda^2}-\frac{\alpha^2}{\lambda^2}=\frac{\alpha}{\lambda^2}\,.$$

Напомним, что параметр $\lambda > 0$ гамма-распределения с плотностью (1.4.40) является масштабным параметром, то есть если $L(X) = \Gamma(\alpha, \lambda)$, то $L(\lambda X) = \Gamma(\alpha, 1)$, поэтому все общие свойства этого семейства распределений могут быть изучены при $\lambda = 1$.

Отметим также, что гамма-распределение является воспроизводящим по параметру α (см., например, [66, с. 185]), то есть если $L(X_1) = \Gamma(\alpha_1, 1)$ и $L(X_2) = \Gamma(\alpha_2, 1)$, то $L(X_1 + X_2) = \Gamma(\alpha_1 + \alpha_2, 1)$ и, в частности, если $X_1, ... X_n$ – независимые и одинаково распределенные случайные величины, причем $L(X_i) = \Gamma(1, 1)$, $i = 1, ..., n$, то случайная величина $Z = (X_1 + ... + X_n)$ также имеет гамма-распределение с параметрами n и единица, то есть $L(Z) = \Gamma(n, 1)$ и $f_Z(z) = z^{n-1} e^{-z} / \Gamma(n)$, $z > 0$.

1.4.45. Распределение Снедекора – Фишера. Пусть $\chi_1^2(n)$ и $\chi_2^2(m)$ – независимые случайные величины, имеющие хи-квадрат-распределение с числами степеней свободы n и m соответственно (n и m – натуральные числа). Рассмотрим случайные величины

$$\varsigma_{nm} = \chi_1^2(n) / \chi_2^2(m) \qquad \textbf{(1.4.46)}$$

и

$$X_{nm} = \frac{\chi_1^2(n)/n}{\chi_2^2(m)/m} = \frac{m}{n}\varsigma_{nm}. \qquad \textbf{(1.4.47)}$$

А) Убедитесь, что функция распределения с.в. ς_{nm} записывается в виде

$$F_{\varsigma_{nm}}(x) = P(\varsigma_{nm} \le x) = \frac{1}{\Gamma(n/2)\Gamma(m/2)} \int_0^{\infty} dv \int_0^{vx} u^{(n/2)-1} v^{(m/2)-1} e^{-u-v} du \ , \ x > 0 \qquad \textbf{(1.4.48)}$$

и, следовательно, ее плотность $f_{\varsigma_{n,m}}(x; n, m)$ для положительных значений аргумента

$$f_{\varsigma_{n,m}}(x; n, m) = \frac{x^{(n/2)-1}}{B(n/2; m/2)} (1+x)^{-\frac{n+m}{2}} = \frac{\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)} x^{\frac{n}{2}-1} (1+x)^{-\frac{n+m}{2}}, \ x > 0. \qquad \textbf{(1.4.49)}$$

Б) Убедитесь также, что плотность распределения вероятности $f_{X_{n,m}}(x; n, m)$ случайной величины $X_{n,m}$ определяется равной нулю для $x \le 0$, а для положительных значений аргумента записывается в виде

$$f_{X_{n,m}}(x; n, m) = \frac{(n/m)^{n/2} x^{(n/2)-1}}{B(n/2 \ ; \ m/2)} \left(1 + \frac{n}{m} x\right)^{-\frac{n+m}{2}}, \ 0 < x < \infty. \qquad \textbf{(1.4.50)}$$

Отметим, что это распределение в литературе называют F-распределением Снедекора – Фишера, коротко это записывают в виде $L(X_{n,m}) = F(n, m)$, и эта

запись означает, что случайная величина $X_{n,m}$ имеет F-распределение с n и m степенями свободы и ее плотность определяется формулой (1.4.50).

В) Убедитесь, что при выполнении условия $-n < 2r < m$ для F-распределения Снедекора – Фишера начальный момент r-го порядка вычисляется по формулам

$$M(X_{nm}^r) = \left(\frac{m}{n}\right)^r \frac{\Gamma[(n/2)+r]\Gamma[(m/2)-r]}{\Gamma(n/2)\Gamma(m/2)},\ r = 1, 2, \ldots . \qquad \textbf{(1.4.52)}$$

Для среднего и дисперсии с.в. $X_{n,m}$ справедливы следующие формулы:

$$M(X_{nm}) = m/(m-2),\ D(X_{nm}) = \frac{2m(n+m-2)}{n(m-2)^2(m-4)}. \qquad \textbf{(1.4.53)}$$

Решение. Напомним, что распределение хи-квадрат с n степенями свободы совпадает с гамма-распределением $\Gamma(\alpha, \lambda)$ при $\alpha = n/2$ и $\lambda = 1/2$, то есть $L(\chi^2(n)) = \Gamma(n/2, 1/2)$. Плотность с.в. $\chi^2(n)$ записывается в виде

$$f_{\chi^2(n)}(x; n) = \frac{x^{(n/2)-1}}{2^{n/2}\Gamma(n/2)} e^{-\frac{x}{2}},\ 0 < x < \infty .$$

Определив множество $S = \{(x, y) : x > 0,\ y > 0,\ (x/y) \le z\}$, и учитывая независимость случайных величин $\chi_1^2(n)$ и $\chi_2^2(m)$, получаем

$$F_{\varsigma_{nm}}(z) = P(\varsigma_{nm} \le z) = \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\int_S x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} e^{-\frac{x}{2}} e^{-\frac{y}{2}} dxdy =$$

$$= \frac{1}{\Gamma(n/2)\Gamma(m/2)} \int_0^{+\infty} dv \int_0^{vz} u^{\frac{n}{2}-1} v^{\frac{m}{2}-1} e^{-u} e^{-v} du .$$

Отсюда получаем плотность с. в. ς_{nm} в виде

$$f_{\varsigma_{nm}}(z) = \frac{dF_{\varsigma_{nm}}(z)}{dz} = \frac{1}{\Gamma(n/2)\Gamma(m/2)} \int_0^{+\infty} (vz)^{\frac{n}{2}-1} v^{\frac{m}{2}-1} e^{-v} e^{-vz} v dv =$$

$$= \frac{z^{(n/2)-1}}{\Gamma(n/2)\Gamma(m/2)} \int_0^{+\infty} v^{\frac{n+m}{2}-1} e^{-v(1+z)} dv = \| \xi = v(1+z),\ v = \xi/(1+z),\ dv = d\xi/(1+z) \| =$$

$$= \frac{z^{(n/2)-1}(1+z)^{-(n+m)/2}}{\Gamma(n/2)\Gamma(m/2)} \int_0^{+\infty} \xi^{\frac{n+m}{2}-1} e^{-\xi} d\xi = \frac{\Gamma[(n+m)/2]\, z^{(n/2)-1}}{\Gamma(n/2)\Gamma(m/2)} (1+z)^{-(n+m)/2},$$

$$0 < z < \infty .$$

Б) Учитывая, что $X_{nm} = m\xi_{nm}/n$ и якобиан преобразования этих случайных величин $J = n/m$, использование (1.3.17) дает формулу (1.4.50). В самом деле,

$$f_{X_{n,m}}(x;n,m) = f_{\varsigma_{nm}}(nx/m)\cdot|J| = \frac{(n/m)^{n/2}x^{(n/2)-1}}{B(n/2\ ;m/2)}\left(1+\frac{n}{m}x\right)^{-\frac{n+m}{2}},\ 0<x<\infty.$$

В) Используя (1.4.50) и (1.4.73), получаем

$$M(X_{nm}^r) = \frac{(n/m)^{n/2}\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)}\int_0^{+\infty} x^{(n/2)+r-1}(1+nx/m)^{-(n+m)/2}dx =$$

$$\|\,\xi = nx/m\,,\ dx = md\xi/n\,\| =$$

$$=\left(\frac{m}{n}\right)^r \frac{(m/n)^{n/2}\cdot(n/m)^{n/2}\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)}\int_0^{+\infty}\xi^{(n/2)+r-1}(1+\xi)^{-[(n/2)+r]-[(m/2)-r]}d\xi =$$

$$=\left(\frac{m}{n}\right)^r \frac{\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)}B(n/2+r\,,m/2-r) = \left(\frac{m}{n}\right)^r \frac{\Gamma[(n/2)+r]\Gamma[(m/2)-r]}{\Gamma(n/2)\Gamma(m/2)},$$

$$r = 1, 2, \ldots\ .$$

1.4.54. Пусть $L(X) = Bi(n\,,p)$. Эта краткая запись означает что дискретная случайная величина X имеет биномиальное распределение с параметрами n и p, $0<p<1$, то есть с.в. X, равная числу успехов в серии из n независимых опытов с вероятностью успеха p, $0<p<1$, принимает целочисленные значения $0,\ 1,\ldots,\ n$ с вероятностями

$$p_n(k) = P(X=k) = C_n^k p^k(1-p)^{n-k},\ k = 0,\ 1,\ldots,\ n. \qquad \textbf{(1.4.55)}$$

А) Убедитесь в справедливости следующих формул:

$$M(X) = np\,,\ D(X) = np(1-p); \qquad \textbf{(1.4.56)}$$

Б) Функция распределения $F_X(x\,|\,n,p)$ биномиальной с.в. X в целочисленных точках $m = 0,\ 1,\ldots,\ n$ вычисляется через неполную бета-функцию $I_p(n,m)$, определенную в (1.4.32), по формуле

$$F_X(m\,|\,n,p) = 1 - I_p(m+1\,,n-m) = I_{1-p}(n-m\,,m+1),\ m = 0,\ 1,\ldots,\ n. \qquad \textbf{(1.4.57)}$$

В) Вероятности

$$p_n(k) = I_p(k\,,n-k+1) - I_p(k+1\,,n-k),\ k = 0,\ 1,\ldots,\ n. \qquad \textbf{(1.4.58)}$$

Решение. А) Пусть $L(X) = Bi(n,p)$. Представим биномиальную с.в. X в виде суммы независимых случайных величин с распределением Бернулли, то есть

$$X = \sum_{i=1}^{n} \xi_i , \qquad \textbf{(1.4.59)}$$

где $\xi_1,...,\xi_n$ – независимые и одинаково распределенные бернуллиевские случайные величины с параметром p, то есть

$$\xi_i = \begin{cases} 1, & \text{с вер. } p, \\ 0, & \text{с вер. } (1-p), \end{cases} \quad i = 1,..., n . \qquad \textbf{(1.4.60)}$$

Учитывая, что $M(\xi_i) = 1 \cdot p + 0 \cdot (1-p) = p$, $D(\xi_i) = p - p^2 = p(1-p)$, $i = 1,..., n$, а также используя свойства математического ожидания и дисперсии, получаем

$$M(X) = \sum_{i=1}^{n} M(\xi_i) = np , \; D(X) = \sum_{i=1}^{n} D(\xi_i) = np(1-p) .$$

Б) Функция распределения вероятностей $F_X(x \mid n, p)$ биномиальной с.в. X записывается в виде

$$F_X(x \mid n, p) = \begin{cases} 0, & \textit{если} \quad x < 0 , \\ \sum_{k=0}^{x} C_n^k p^k (1-p)^{n-k}, & 0 \le x \le n, \\ 1, & \textit{если} \quad x > n . \end{cases}$$

С использованием единичной функции Хевисайда (1.3.23) это выражение может быть переписано в компактной форме следующим образом:

$$F_X(x \mid n, p) = \sum_{k=0}^{n} C(x-k) p_n(k) = \sum_{k=0}^{n} C(x-k) C_n^k p^k (1-p)^{n-k} , \; x \in R^1 .$$

Пусть, далее, одно из $(n+1)$ возможных значений биномиальной с.в. X обозначено через m, $m = 0, 1,..., n$. Рассмотрим при фиксированных n и m функцию $F_X(m \mid n, p)$ параметра p, $0 < p < 1$. Дифференцируя данную функцию по p, после несложных преобразований (см. также упр. 4.11.7), получаем

$$\partial F_X(m \mid n, p) / \partial p = \sum_{k=0}^{m} C_n^k [k p^{k-1} (1-p)^{n-k} - (n-k) p^k (1-p)^{n-k-1}] =$$

$$= -\frac{n!}{m!(n-m-1)!} p^m (1-p)^{n-m-1} .$$

Интегрируя данное выражение, с учетом формул (1.4.32) и (1.4.33), получаем формулу для вычисления функции распределения вероятностей $F_X(x \mid n, p)$

биномиальной с. в. X в целочисленных точках $m = 0,\ 1,...,\ n$ в виде

$$F_X(m \mid n, p) = \frac{n!}{m!(n-m-1)!}\int_p^1 x^m (1-x)^{n-m-1}\,dx = 1 - I_p(m+1,\, n-m) =$$
$$= I_{1-p}(n-m,\, m+1),\ \ m = 0,\ 1,...,\ n.$$

Отсюда следует, что

$$F_X(x \mid n, p) = C(x-m)\cdot I_{1-p}(n-m,\, m+1),\ m = 0,\ 1,...,\ n,\ x \in R^1.$$

В) Используя (1.4.57), убеждаемся, что вероятности $p_n(k)$ возможных состояний биномиальной с. в. X вычисляются через неполную бета-функцию по формулам

$$p_n(k) = F_X(k \mid n, p) - F_X(k-1 \mid n, p) = 1 - I_p(k+1,\, n-k) - 1 + I_p(k,\, n-k+1) =$$
$$= I_p(k,\, n-k+1) - I_p(k+1,\, n-k),\ k = 0,\ 1,...,\ n.$$

Отметим, что гипергеометрическое распределение вероятностей (см. П.1.36) в предельном случае, когда число объектов N изучаемой совокупности стремиться к бесконечности, сходится к биномиальному распределению и при этом различие между получением наблюдений без возвращения (гипергеометрическое распределение) и с возвращением (биномиальное распределение) исчезает. Убедимся в этом. Пусть исследуемая совокупность состоит из N объектов, и пусть $M = Np$, $0 < p < 1$, обладают некоторым признаком (A) (например, признак (A) означает, что среди N деталей, $M = Np$ деталей стандартные). Случайным образом выбирают n $(n < N)$ объектов. Пусть случайная величина X обозначает число объектов, обладающих признаком (A), среди n выбранных на удачу объектов. Тогда вероятность того, что среди n выбранных объектов точно k будут обладать признаком (A), вычисляется по формуле

$$P(X = k) = \frac{C_{Np}^k C_{N(1-p)}^{n-k}}{C_N^n},\ k = 0,...,\min(n, M). \qquad \textbf{(1.4.61)}$$

Вычисление вероятностей по (1.4.61) является утомительным делом, и для этой цели обычно используют таблицы (см., например, [6]). В предельном случае, когда $N \to \infty$, но отношение M/N остается постоянным, то есть $M/N = p$, $0 < p < 1$, при фиксированном n, гипергеометрическое распределение сходится к биномиальному. В самом деле, переходя к факториалам, вероятности из (1.4.61) можно переписать в виде

$$P(X = k) = \frac{n!}{k!(n-k)!}\,\frac{Np(Np-1)\cdots(Np-k+1)\cdot(N-Np)\cdots(N-Np-n+k+1)}{N(N-1)\cdots(N-n+1)}.$$

Разделим числитель и знаменатель этой дроби на N^n. В результате получим

$$P(X=k)=C_n^k\frac{p\left(p-\frac{1}{N}\right)\cdots\left(p-\frac{k-1}{N}\right)\cdot(1-p)\cdots\left(1-p-\frac{n-k+1}{N}\right)}{\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)}.$$

Отсюда следует, что при $N\to\infty$ и фиксированных p, n и k выполняется приближенное равенство

$$P(X=k)\approx C_n^k p^k(1-p)^{n-k},\ k=0,...,n. \qquad \textbf{(1.4.62)}$$

Таким образом, гипергеометрическое распределение сходится к биномиальному распределению, причем качество аппроксимации (1.4.62) вполне удовлетворительно, когда число взятых на контроль объектов n не более одной десятой от общего числа объектов N, то есть при $n/N<0{,}1$.

Например, при $N=1000$, $M=20$, то есть $M/N=p=0{,}02$, и при $n=100$, $k=3$, гипергеометрическое распределение дает вычисленную по формуле (1.4.61) вероятность $P(X=3)=0{,}1918$, а биномиальное распределение при $p=0{,}02$ и $n=100$, $k=3$ дает вычисленную по формуле (1.4.62), вероятность $P_{100}(X=3)=0{,}1823$. Отметим, что гипергеометрическое распределение характеризуется параметрами N, n *и* p и если с.в. X имеет гипергеометрическое распределение, то кратко этот факт записывают в виде $L(X)=Гип(N,n,p)$.

1.4.63. Распределение Пуассона (см. П.1.141). Пусть $L(X)=Bi(n,p)$, то есть случайная величина X имеет биномиальное распределение с параметрами n и p, $0<p<1$, и она принимает целочисленные значения 0, 1,..., n с вероятностями $p_n(k)$, $k=0,\ 1,...,\ n$, определенными в (1.4.55). Напомним, что $M(X)=np$. Рассмотрим асимптотическое поведение вероятностей $p_n(k)$ при неограниченном увеличении числа опытов n и при стремлении вероятности p к нулю таким образом, что произведение np остается постоянной величиной, то есть $np=\lambda$. Убедитесь, что при такой асимптотике ($n\to\infty$, $p\to 0$, $np\to\lambda$) выполняется выражение

$$p_n(k)\to\frac{\lambda^k}{k!}e^{-\lambda},\ k=0,\ 1,\ ... \qquad \textbf{(1.4.64)}$$

Решение. Учитывая, что $np=\lambda$, представим вероятности $p_n(k)$ в виде

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

$$=\frac{n(n-1)...(n-k+1)}{n^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}=\left[1\cdot\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]\cdot\left(1-\frac{\lambda}{n}\right)^{-k}\cdot\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n.$$

Используя далее теоремы о вычислении пределов числовых последовательностей и учитывая, что

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n = e\ ,\ \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda},$$

получаем

$$\lim_{n\to\infty} p_n(k) = \lim_{n\to\infty}\left[1\cdot\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\cdot\lim_{n\to\infty}\frac{\lambda^k}{k!}\cdot\left(1-\frac{\lambda}{n}\right)^n =$$
$$= \frac{\lambda^k}{k!}\cdot 1\cdot 1\cdot e^{-\lambda} = \frac{\lambda^k}{k!}e^{-\lambda},\ k = 0,\ 1,\ \ldots\ .$$

Отметим, что полученное предельное распределение вероятностей по целочисленным значениям $k = 0,\ 1,\ \ldots$ называется распределением Пуассона (см. П.1.141). Для численной иллюстрации качества аппроксимации биномиальных вероятностей распределением Пуассона приведем пример. Пусть $p = 0{,}02$, тогда при $n = 100$ и $k = 3$ биномиальная вероятность $p_{100}(3) = C_{100}^3(0{,}02)^3(0{,}98)^{97} = 0{,}1823$. Распределение Пуассона при $\lambda = 2$ и $k = 3$ дает вероятность равную $2^3 e^{-2}/3! = 0{,}1804$. При увеличении n и уменьшении p качество аппроксимации улучшается.

1.4.65. Распределение Стьюдента (см. П.2.89) . Рассмотрим случайную величину T_k в виде отношения двух случайных величин X и $\chi^2(k)$, то есть

$$T_k = \frac{X}{\sqrt{\chi^2(k)/k}}. \qquad \textbf{(1.4.66)}$$

Предполагаем, что X и $\chi^2(k)$ – *независимые* случайные величины, причем с.в. X имеет стандартное нормальное распределение, то есть $L(X) = N(0\,,1)$, и эта запись означает, что ее плотность имеет вид

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\ ,\ -\infty < x < +\infty\ , \qquad \textbf{(1.4.67)}$$

а с.в. $\chi^2(k)$ имеет распределение хи-квадрат с числом степеней свободы k (см. П.2.78) и ее плотность имеет вид

$$f_{\chi^2(k)}(x;k) = \frac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)},\ 0 < x < \infty. \qquad \textbf{(1.4.68)}$$

Убедиться, что с.в. T_k имеет распределение Стьюдента с числом степеней свободы k, кратко это записывают в виде $L(T_k) = St(k)$, и эта запись означает,

что плотность с.в. T_k имеет вид

$$f_{T_k}(t;k)=\frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\,\Gamma(k/2)}(1+t^2/k)^{-\frac{k+1}{2}},\ -\infty<t<+\infty. \qquad \textbf{(1.4.69)}$$

Решение. Учитывая, что X и $\chi^2(k)$ – независимые случайные величины, их совместная плотность запишется в виде

$$f_{X\chi^2(k)}(x,y)=f_X(x)\cdot f_{\chi^2(k)}(y;k)=\frac{1}{2\sqrt{2\pi}\,\Gamma(k/2)}\left(\frac{y}{2}\right)^{\frac{1}{2}(k-2)}e^{-\frac{1}{2}(x^2+y)},$$
$$-\infty<x<+\infty,\ 0<y<+\infty.$$

Рассмотрим пару (T,S) новых случайных величин, причем $T=X/\sqrt{(S/k)}$, $S=\chi^2(k)$. Эти случайные величины получены с помощью преобразований переменных вида $t=x/\sqrt{(s/k)}$, $s=y$. Обратные преобразования записываются в виде $x=t\sqrt{s/k}$, $y=s$. Якобиан преобразований равен

$$J=\begin{vmatrix}\frac{\partial x}{\partial t} & \frac{\partial x}{\partial s}\\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial s}\end{vmatrix}=\begin{vmatrix}\sqrt{s/k} & t/2k\sqrt{s/k}\\ 0 & 1\end{vmatrix}=\sqrt{s/k}\quad.$$

Совместная плотность случайных величин T и S, согласно (1.4.77), запишется в виде

$$f_{TS}(t,s)=f_{X\chi^2(k)}(t\sqrt{s/k},s)\cdot|J|=\frac{1}{2\sqrt{2\pi}\,\Gamma(k/2)}\left(\frac{s}{2}\right)^{\frac{1}{2}(k-2)}e^{-\frac{1}{2}(1+\,t^2/k)s}\sqrt{s/k}=$$
$$=\frac{1}{2\sqrt{k\pi}\,\Gamma(k/2)}\left(\frac{s}{2}\right)^{\frac{1}{2}(k-1)}e^{-\frac{1}{2}(1+\,t^2/k)s},\ -\infty<t<+\infty,\ 0<s<+\infty.$$

Отсюда, используя (1.4.2), получаем выражение для плотности с.в. T_k в виде

$$f_{T_k}(t;k)=\int_0^\infty f_{TS}(t,s)ds=\frac{1}{2\sqrt{k\pi}\,\Gamma(k/2)}\int_0^\infty\left(\frac{s}{2}\right)^{\frac{1}{2}(k-1)}e^{-\frac{1}{2}(1+\,t^2/k)s}ds=$$
$$\left\|\,z=\left(1+\frac{t^2}{k}\right)\frac{s}{2},\ ds=dz\cdot 2/(1+\,t^2/k),\ s/2=z/(1+\,t^2/k)\,\right\|=$$
$$\frac{(1+t^2/k)^{-\frac{k+1}{2}}}{\sqrt{k\pi}\,\Gamma(k/2)}\int_0^\infty z^{\frac{1}{2}(k-1)}e^{-z}dz=\frac{\Gamma[(k+1)/2]\cdot(1+t^2/k)^{-\frac{k+1}{2}}}{\sqrt{k\pi}\,\Gamma(k/2)},\ -\infty<t<+\infty.$$

1.4.70. Убедитесь, что если $L(X)=St(k)$, то все начальные моменты нечетного порядка с.в. X равны нулю, а моменты четного порядка $2r$ при выполнении условия $-1<2r<k$ вычисляются по формулам

$$M(X^{2r})=k^r\frac{\Gamma[(1/2)+r]\Gamma[(k/2)-r]}{\Gamma(1/2)\Gamma(k/2)},\ r=1,2,... \qquad \textbf{(1.4.71)}$$

и, в частности,

$$M(X)=0,\ D(X)=k/(k-2). \qquad \textbf{(1.4.72)}$$

Решение. Напомним (см. [15, (8.384)]), что бета-функция определяется в виде

$$B(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt=\int_0^\infty t^{x-1}(1+t)^{-(x+y)}dt=\frac{\Gamma(x)\cdot\Gamma(y)}{\Gamma(x+y)}. \qquad \textbf{(1.4.73)}$$

Далее,

$$M(X^{2r})=\int_{-\infty}^{+\infty}x^{2r}f_X(x;k)dx=2\int_0^{+\infty}x^{2r}f_X(x;k)dx=$$

$$=\frac{2\Gamma[(k+1)/2]}{\Gamma(1/2)\Gamma(k/2)}\int_0^\infty x^{2r}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}dx=\|\,t=x^2/k,\ dx=\frac{\sqrt{k}dt}{2\sqrt{t}}\,\|=$$

$$=k^r\frac{\Gamma[(k+1)/2]}{\Gamma(1/2)\Gamma(k/2)}\int_0^\infty t^{r-(1/2)}(1+t)^{-\frac{k+1}{2}}dt=$$

$$=k^r\frac{\Gamma[(k+1)/2]}{\Gamma(1/2)\Gamma(k/2)}\int_0^\infty t^{\left(r+\frac{1}{2}\right)-1}(1+t)^{-\left(r+\frac{1}{2}\right)-\left(\frac{k}{2}-r\right)}dt.$$

Учитывая (1.4.73), интеграл в последнем выражении запишем через бета-функцию, в результате получаем окончательную формулу для вычисления четных моментов порядка $2r$ в виде

$$M(X^{2r})=k^r\frac{\Gamma[(k+1)/2]}{\Gamma(1/2)\Gamma(k/2)}\cdot B\left(r+\frac{1}{2},\frac{k}{2}-r\right)=$$

$$=k^r\frac{\Gamma[(k+1)/2]}{\Gamma(1/2)\Gamma(k/2)}\cdot\frac{\Gamma[(1/2)+r]\Gamma[(k/2)-r]}{\Gamma[(k+1)/2]}=$$

$$=k^r\frac{\Gamma[(1/2)+r]\Gamma[(k/2)-r]}{\Gamma(1/2)\Gamma(k/2)},\ r=1,2,... .$$

Отметим связь (см. [66, с.199]), которая существует между распределением Стьюдента и бета-распределением. Пусть с.в. T_k имеет распределение

Стьюдента с числом степеней свободы k, тогда с.в. $X = 1/(1+T^2/k)$ имеет бета-распределение, то есть,

если $L(T_k) = St(k)$, то $L(X) = L\{1/(1+T^2/k)\} = Be(k/2, 1/2)$. **(1.4.74)**

1.4.75. Пусть (X_1, X_2) – непрерывная двумерная случайная величина с плотностью $f_{X_1X_2}(x_1, x_2)$, которая определяется в виде

$$f_{X_1X_2}(x_1, x_2) = \begin{cases} e^{-x_1-x_2}, & x_1 \ge 0, x_2 \ge 0, \\ 0, & \text{в других случаях.} \end{cases}$$

Убедитесь, что плотность распределения вероятностей двумерной случайной величины (Y_1, Y_2), где $Y_1 = X_1 + X_2$ и $Y_2 = X_1 / X_2$, записывается в виде

$$f_{Y_1Y_2}(y_1, y_2) = \begin{cases} e^{-y_1} y_1/(1+y_2)^2, & y_1 \ge 0, y_2 \ge 0, \\ 0, & \text{в других случаях.} \end{cases}$$

Решение. Пусть (X_1, X_2) – непрерывная двумерная случайная величина с плотностью $f_{X_1X_2}(x_1, x_2)$, $(x_1, x_2) \in R^2$. Рассмотрим случайные величины $Y_1 = \varphi_1(X_1, X_2)$ и $Y_2 = \varphi_2(X_1, X_2)$, которые получены с помощью преобразований $y_i = \varphi_i(x_1, x_2)$, $i = 1, 2$, над случайной величиной (X_1, X_2). Предполагается, что эти преобразования взаимно однозначно отображают x_1x_2-плоскость на y_1y_2-плоскость и допускают единственные решения $x_i = \varphi_i^{-1}(y_1, y_2)$, $i = 1, 2$, которые непрерывны и имеют первые частные производные. Якобиан преобразований $y_i = \varphi_i(x_1, x_2)$, $i = 1, 2$, обозначается через J и определяется в виде

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial u} & \dfrac{\partial x_2}{\partial v} \end{vmatrix} = \begin{vmatrix} \dfrac{\partial \varphi_1^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial \varphi_1^{-1}(y_1, y_2)}{\partial y_2} \\ \dfrac{\partial \varphi_2^{-1}(y_1, y_2)}{\partial y_1} & \dfrac{\partial \varphi_2^{-1}(y_1, y_2)}{\partial y_2} \end{vmatrix}. \quad \textbf{(1.4.76)}$$

Тогда (см., например, [66, с. 69]) плотность распределения вероятностей двумерной случайной величины (Y_1, Y_2) записывается в виде

$$f_{Y_1Y_2}(y_1, y_2) = f_{X_1X_2}(\varphi_1^{-1}(y_1, y_2), \varphi_2^{-1}(y_1, y_2)) \cdot |J|, \ (y_1, y_2) \in R^2. \quad \textbf{(1.4.77)}$$

В данном примере преобразования $y_i = \varphi_i(x_1, x_2)$, $i = 1, 2$, записываются в такой форме $y_1 = x_1 + x_2$, $y_2 = x_1 / x_2$, а обратные преобразования имеют вид $x_1 = y_1/(1+y_2)$, $x_2 = y_1y_2/(1+y_2)$. Далее, якобиан J преобразований, которые

устанавливают взаимно однозначное соответствие между точками положительного квадранта x_1x_2-плоскости и точками положительного квадранта y_1y_2-плоскости, равен

$$J=\begin{vmatrix}\dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2}\\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2}\end{vmatrix}=\begin{vmatrix}\dfrac{1}{1+y_2} & -\dfrac{y_1}{(1+y_2)^2}\\ \dfrac{y_2}{1+y_2} & \dfrac{y_1}{(1+y_2)^2}\end{vmatrix}=y_1/(1+y_2)^2\,.$$

Таким образом, согласно (1.4.77), интересующая нас плотность запишется в виде

$$f_{Y_1Y_2}(y_1,y_2)=f_{X_1X_2}(\,y_1/(1+y_2)\,,\,y_1y_2/(1+y_2)\,)\cdot|J|=$$
$$=\begin{cases}e^{-y_1}y_1/(1+y_2)^2, & y_1\ge 0,\ y_2\ge 0,\\ 0, & \text{в других случаях.}\end{cases}$$

1.4.78. Пусть с.в. X имеет гамма-распределение с параметром формы $\alpha>0$ и масштабным параметром $\lambda>0$ (см. П.2.18). Кратко этот факт записывают в виде $L(X)=\Gamma(\alpha\,,\lambda)$, и эта запись означает, что ее плотность определяется выражением

$$f_X(x;\alpha,\lambda)=\begin{cases}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x}/\Gamma(\alpha)\,, & x>0,\\ 0\,, & x\le 0\,.\end{cases}\qquad\textbf{(1.4.79)}$$

Пусть далее X_1 и X_2 – независимые случайные величины. Убедитесь, что если $L(X_1)=\Gamma(\alpha_1,1)$ и $L(X_2)=\Gamma(\alpha_2,1)$, тогда $L(U)=L\{X_1/(X_1+X_2)\}=Be(\alpha_1,\alpha_2)$.

Решение. Учитывая (1.4.79) и независимость с.в. X_1 и X_2, запишем выражение для их совместной плотности в виде

$$f_{X_1X_2}(x_1,x_2)=\frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}e^{-x_1-x_2}\,,\ 0<x_1<\infty\,,\ 0<x_2<\infty\,.$$

Используем преобразование переменных в виде $u=x_1/(x_1+x_2)\,,\ v=x_2$. Обратные преобразования запишутся в виде $x_1=uv/(1-u)$ и $x_2=v$. С учетом этого получаем выражение для якобиана преобразований J в виде

$$J=\begin{vmatrix}\dfrac{\partial x_1}{\partial u} & \dfrac{\partial x_1}{\partial v}\\ \dfrac{\partial x_2}{\partial u} & \dfrac{\partial x_2}{\partial v}\end{vmatrix}=\begin{vmatrix}\dfrac{v(1-u)+uv}{(1-u)^2} & \dfrac{u}{1-u}\\ 0 & 1\end{vmatrix}=\frac{v}{(1-u)^2}\quad.$$

Согласно (1.4.77), совместная плотность случайных величин U и V запишется в такой форме

$$f_{UV}(u,v) = f_{X_1X_2}(uv/(1-u)\,,\,v)\cdot|\,J\,| = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}u^{\alpha_1-1}v^{\alpha_1+\alpha_2-1}(1-u)^{-(\alpha_1+1)}e^{-v/(1-u)},$$
$$0<u<1,\ 0<v<\infty.$$

Отсюда, интегрируя по v в пределах от нуля до бесконечности, убеждаемся, что случайная величина $U = X_1/(X_1+X_2)$ имеет бета-распределение. В самом деле,

$$f_U(u) = \int_0^\infty f_{UV}(u,v)dv = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^\infty u^{\alpha_1-1}v^{\alpha_1+\alpha_2-1}(1-u)^{-(\alpha_1+1)}e^{-v/(1-u)}dv =$$
$$= \|\, z = v/(1-u)\,,\ v = z(1-u)\,,\ dv = (1-u)dz\,\| =$$
$$= \frac{u^{\alpha_1-1}(1-u)^{-(\alpha_1+1)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^\infty (1-u)^{\alpha_1+\alpha_2-1}z^{\alpha_1+\alpha_2-1}e^{-z}(1-u)dz =$$
$$= \frac{u^{\alpha_1-1}(1-u)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^\infty z^{\alpha_1+\alpha_2-1}e^{-z}dz = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}u^{\alpha_1-1}(1-u)^{\alpha_2-1},\ 0<u<1.$$

1.4.80. Связи между F-распределением и бета-распределением. Пусть $L(X) = Be(n/2\,,\,m/2)$. Эта краткая запись означает, что плотность с.в. X записывается в виде

$$f_X(x) = \frac{x^{(n/2)-1}(1-x)^{(m/2)-1}}{B(n/2\,,\,m/2)},\ 0<x<1, \qquad \textbf{(1.4.81)}$$

а функция распределения вероятностей $F_X(x)$, $0<x<1$, выражается через неполную бета-функцию $I_p(n,m)$, определенную в (1.4.32), в виде

$$F_X(x) = I_x(n/2\,,\,m/2),\ 0<x<1. \qquad \textbf{(1.4.82)}$$

Рассмотрим случайную величину $Y = \varphi(X)$, где $y = \varphi(x) = mx/n(1+x)$, $0<x<1$.

А) Убедитесь, что с.в. Y имеет F-распределение с n и m степенями свободы, то есть $L(Y) = F(n\,,\,m)$ и эта запись означает, что плотность с.в. Y определяется формулой (1.4.50).

Б) Убедитесь, что если $L(Y) = F(n\,,\,m)$, то ф.р. случайной величины Y выражается через неполную бета-функцию $I_p(n,m)$ в виде

$$F_Y(y) = I_{ny/(m+ny)}(n/2\,,\,m/2),\ 0<y<\infty. \qquad \textbf{(1.4.83)}$$

В) Пусть $L(X) = Be(n/2\,,\,m/2)$ и $Y = mX/n(1+X)$. Тогда, согласно пункту Б, выполняется выражение $L(Y) = F(n\,,\,m)$. Убедитесь, что между квантилем $F_{p\,,(n,m)}$ уровня p, $0<p<1$, F-распределения числами степеней свободы

(n,m) и квантилем $b_{p,(n,m)}$ уровня p, $0<p<1$, бета-распределения с параметрами n и m существует связь вида

$$b_{p,(n/2,m/2)}=\frac{nF_{p,(n,m)}}{m+nF_{p,(n,m)}}. \qquad \textbf{(1.4.84)}$$

Решение. А) Используя формулу (1.3.17), запишем плотность случайной величины Y в виде

$$f_Y(y)=f_X(\varphi^{-1}(y))\cdot|J|=f_X(\varphi^{-1}(y))\,|d\varphi^{-1}(y)/dy|,$$

где J – якобиан преобразования переменных вида $y=\varphi(x)=mx/n(1+x)$, $0<x<1$. Учитывая, что обратное преобразование равно $x=\varphi^{-1}(y)=ny/(m+ny)$, $0<y<\infty$, получаем

$$J=dx/dy=d\varphi^{-1}(y)/dy=1/\varphi'(\varphi^{-1}(y))=nm/(m+ny)^2.$$

Таким образом,

$$f_Y(y)=f_X(\varphi^{-1}(y))\cdot|J|=f_X\left(\frac{ny}{m+ny}\right)\cdot\frac{nm}{(m+ny)^2}=$$

$$=\frac{nm}{(m+ny)^2}\frac{[ny/(m+ny)]^{(n/2)-1}[1-ny/(m+ny)]^{(m/2)-1}}{B(n/2\,,m/2)}=$$

$$=\frac{n^{n/2}m^{m/2}y^{(n/2)-1}}{(m+ny)^{(n+m)/2}B(n/2\,,m/2)}=\frac{(n/m)^{n/2}\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)}y^{\frac{n}{2}-1}\left(1+\frac{n}{m}y\right)^{-(n+m)/2},$$

$$0<y<\infty.$$

Б) Пусть $L(X)=Be(n/2\,,m/2)$ и $X=nY/(m+nY)$, где с.в. Y имеет F-распределение с n и m степенями свободы, то есть $L(Y)=F(n\,,m)$, тогда (1.4.82) перепишется в виде

$$I_x(n/2\,,m/2)=F_X(x)=P(X\le x)=P\{nY/(m+nY)\le x\}=$$
$$=P\{Y\le mx/(n-nx)\}=F_Y(mx/(n-nx)).$$

Отсюда, используя замену переменной $y=mx/(n-nx)$ и, следовательно, $x=ny/(m+ny)$, получаем

$$F_Y(y)=I_{ny/(m+ny)}(n/2\,,m/2).$$

В) Так как справедливо равенство $F_X(x)=F_Y(mx/(n-nx))$, то подставив вместо аргумента x квантиль $b_{p,(n,m)}$ уровня p, $0<p<1$, бета-распределения с параметрами n и m, получим

$$F_Y(mb_{p,(n/2,m/2)}/(n-nb_{p,(n/2,m/2)}))=p,$$

или $mb_{p,(n/2,m/2)}/(n-nb_{p,(n/2,m/2)}) = F_Y^{-1}(p) = F_{p,(n,m)}$.

Отсюда имеем требуемый результат в виде (1.4.84).

1.4.85. Распределение Снедекора – Фишера (см. П.2.94). Пусть $\chi_1^2(n)$ и $\chi_2^2(m)$ независимые случайные величины, имеющие хи-квадрат-распределение с числами степеней свободы n и m соответственно (n и m – натуральные числа). Рассмотрим случайную величину

$$X_{nm} = \frac{\chi_1^2(n)/n}{\chi_2^2(m)/m}.$$

Получите выражение для функции распределения $F_{X_{n,m}}(x;n,m)$ случайной величины $X_{n,m}$.

Решение. Согласно определению ф.р. $F_{X_{n,m}}(z;n,m)$ случайной величины $X_{n,m}$ в точке z вычисляется через вероятность $P\{X_{n,m} \le z\}$. Для вычисления этой вероятности определим множество S следующим образом: $S = \{(x,y): x>0,\ y>0,\ (x/y)<z\}$. Используя (1.3.44) и (1.4.28) и учитывая, что $\chi_1^2(n)$ и $\chi_2^2(m)$ – независимые случайные величины, получаем

$$F_{X_{n,m}}(z;n,m) = P\{X_{n,m} \le z\} = \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\limits_S \int x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} e^{-(x+y)/2} dxdy =$$

$$= \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\limits_0^\infty dy \int\limits_0^{yz} x^{n/2-1} y^{m/2-1} e^{-x/2} e^{-y/2} dx =$$

$$= \| \ t\big|_0^z = x/y\ ,\ x = ty\ ,\ dx = ydt \ \| =$$

$$= \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\limits_0^\infty dy \int\limits_0^z t^{n/2-1} y^{(n+m)/2\ -1}\ e^{-ty/2}\ e^{-y/2} dt =$$

$$= \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\limits_0^z t^{n/2-1} dt \int\limits_0^\infty y^{\frac{n+m}{2}-1} e^{-\frac{y(t+1)}{2}} dy =$$

$$= \| \frac{y(t+1)}{2} = \xi\big|_0^\infty\ ,\ y = \frac{2\xi}{(t+1)},\ dy = \frac{2\xi}{(t+1)} \| =$$

$$= \frac{1}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int\limits_0^z t^{n/2-1}(1+t)^{-(n+m)/2} \cdot 2^{(n+m)/2} dt \int\limits_0^\infty \xi^{\frac{n+m}{2}-1} e^{-\xi} d\xi =$$

$$= \frac{\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)} \int\limits_0^z t^{n/2-1}(1+t)^{-(n+m)/2} dt\,.$$

Отметим, что отсюда, с учетом (1.4.32) и (1.4.33), следует (1.4.83).

1.4.86. Связи между гамма-распределением, бета-распределением и F-распределением Снедекора – Фишера.

А) Пусть X_1 и X_2 независимые случайные величины. Убедитесь, что если $L(X_1)=\Gamma(\alpha_1,\lambda)$, $L(X_2)=\Gamma(\alpha_2,\lambda)$ и $U=X_1/X_2$, тогда $L(U)=F(2\alpha_1,2\alpha_2)$.

Б) Убедитесь, что если $L(X_1)=\Gamma(\nu_1,\lambda)$, $L(X_2)=\Gamma(\nu_2,\lambda)$ и $\xi=X_1/(X_1+X_2)$, тогда $L(\xi)=Be(\nu_1,\nu_2)$.

Решение. А) Так как X_1 и X_2 – независимые случайные величины, имеющие гамма-распределение, то есть $L(X_1)=\Gamma(\nu_1,\lambda)$ и $L(X_2)=\Gamma(\nu_2,\lambda)$, то их совместная плотность запишется в виде

$$f_{X_1X_2}(x_1,x_2)=f_{X_1}(x_1)\cdot f_{X_2}(x_2)=\frac{\lambda^{\alpha_1}x_1^{\alpha_1-1}e^{-\lambda x_1}}{\Gamma(\alpha_1)}\cdot\frac{\lambda^{\alpha_2}x_2^{\alpha_2-1}e^{-\lambda x_2}}{\Gamma(\alpha_2)}.$$

Перейдем от совместной плотности случайных величин X_1 и X_2 к совместной плотности $f_{UV}(u,v)$ случайных величин U и V с помощью преобразования переменных в виде $u=x_1/x_2$ и $v=x_2$. Обратные преобразования $x_1=uv$, $x_2=v$ и якобиан преобразований $J=v$. Используя (1.4.77), получаем

$$f_{UV}(u,v)=f_{X_1}(uv)\cdot f_{X_2}(v)\cdot|J|=\frac{\lambda^{\alpha_1}(uv)^{\alpha_1-1}e^{-\lambda uv}}{\Gamma(\alpha_1)}\cdot\frac{\lambda^{\alpha_2}v^{\alpha_2-1}e^{-\lambda v}}{\Gamma(\alpha_2)}\cdot v.$$

Отсюда, используя (1.4.2), получаем плотность случайной величины $U=X_1/X_2$ в виде

$$f_U(u)=\int_0^{+\infty}f_{UV}(u,v)dv=\int_0^{+\infty}v\,\frac{\lambda^{\alpha_1}(uv)^{\alpha_1-1}e^{-\lambda uv}}{\Gamma(\alpha_1)}\cdot\frac{\lambda^{\alpha_2}v^{\alpha_2-1}e^{-\lambda v}}{\Gamma(\alpha_2)}dv=$$

$$=\frac{\lambda^{\alpha_1}\lambda^{\alpha_2}u^{\alpha_1-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^{+\infty}v^{\alpha_1+\alpha_2-1}e^{-\lambda v(1+u)}dv=$$

$$=\|\,z=\lambda v(1+u),\ v=z/\lambda(1+u),\ dv=dz/\lambda(1+u)\,\|=$$

$$=\frac{u^{\alpha_1-1}(1+u)^{-(\alpha_1+\alpha_2)}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^{+\infty}z^{\alpha_1+\alpha_2-1}e^{-z}dz=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}u^{\alpha_1-1}(1+u)^{-(\alpha_1+\alpha_2)},\ 0<u<\infty.$$

Б) Согласно условию, имеем $L(X_1)=\Gamma(\nu_1,\lambda)=\chi^2(2\nu_1)$ и $L(X_2)=\Gamma(\nu_2,\lambda)=\chi^2(2\nu_2)$. С учетом этого замечания и формулы (1.4.46), а также обозначив $2\nu_1=n$, $2\nu_2=m$, запишем с.в. ξ в виде

$$\xi=\frac{X_1}{X_1+X_2}=\frac{\chi_1^2(n)}{\chi_1^2(n)+\chi_2^2(m)}=\frac{\varsigma_{nm}}{1+\varsigma_{nm}}.$$

Отметим, что $0 \le \xi \le 1$, так что плотность вероятности с.в. ξ равна нулю вне интервала $(0\,,1)$. Когда ς_{nm}, определенная в (1.4.46), возрастает от 0 до ∞, ξ монотонно возрастает от 0 до 1. Неравенство $\xi \le x$ эквивалентно неравенству $\varsigma_{nm} \le x/(1-x)$ и, следовательно, для функции распределения с.в. ξ получаем выражение

$$F_{\xi}(x) = P(\xi \le x) = P(\varsigma_{nm} \le x/(1-x)) = F_{\varsigma_{nm}}(x/(1-x))\,.$$

Отсюда, с учетом (1.4.49), получаем плотность с. в. ξ в виде

$$f_{\xi}(x) = \frac{1}{(1-x)^2} f_{\varsigma_{nm}}\left(\frac{x}{1-x}\right) =$$

$$= \frac{1}{(1-x)^2} \frac{\Gamma[(n+m)/2](x/(1-x))^{(n/2)-1}}{\Gamma(n/2)\Gamma(m/2)} (1+x/(1-x))^{-(n+m)/2} =$$

$$= \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1}(1-x)^{\nu_2-1},\ 0 < x < \infty\,.$$

1.4.87. Пусть $L(X) = N(0\,,1)$, $L(T_m) = St(m)$, $L(X_{nm}) = F(n\,,m)$, а случайные величины T_m и X_{nm} определены в виде

$$T_m = \frac{X}{\sqrt{\chi^2(m)/m}} \text{ и } X_{nm} = \frac{\chi^2(n)/n}{\chi^2(m)/m}\,.$$

Убедитесь, что между квантилем $F_{p\,,(1,m)}$ уровня p, $0 < p < 1$, F-распределения с числами степеней свободы $(1,m)$, и квантилем $t_{p\,,m}$ уровня p, $0 < p < 1$, распределения Стьюдента с m степенями свободы существует связь вида

$$F_{p\,,(1,m)} = t^2_{(1+p)/2\,,\,m}\,. \qquad \textbf{(1.4.88)}$$

Решение. Напомним (см. (1.4.29)), что $L(X^2) = L(X^2(1)) = \chi^2(1)$, поэтому из определения случайных величин T_m и X_{nm} следует, что $T_m^2 \sim X_{1m}$, то есть $L(T_m^2) = F(1\,,m)$. С учетом этого замечания, и учетом факта симметрии распределения Стьюдента, получаем

$$F_{X_{1m}}(x) = P(X_{1m} \le x) = P(T_m^2 \le x) = P(-\sqrt{x} \le T_m \le \sqrt{x}) =$$

$$= F_{T_m}(\sqrt{x}) - F_{T_m}(-\sqrt{x}) = 2F_{T_m}(\sqrt{x}) - 1\,.$$

Отсюда следует, что

$$2F_{T_m}(\sqrt{F_{p\,,(1,m)}}) - 1 = p\,,\ 0 < p < 1\,,$$

или $$\sqrt{F_{p,(1,m)}} = F_{T_m}^{-1}[(1+p)/2] = t_{(1+p)/2,\,m}$$

и окончательно следует (1.4.88).

1.4.89. Усеченные распределения (см. [36, с.273]). Пусть X – с.в. с функцией распределения $F_X(x)$ и плотностью $f_X(x)$, $-\infty < x < +\infty$. Рассматривается интервал $a < X \le b$ и из последовательности повторных наблюдений над с.в. X отбрасываются все наблюдения, при которых были зафиксированы значения величины $X \le a$ или $> b$. Остающиеся наблюдения характеризуются *усеченным* распределением с функцией распределения

$$F_X(x \mid a < X \le b) = \begin{cases} 0, & \text{при } x \le a, \\ \dfrac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)}, & \text{при } a < x \le b, \\ 1, & \text{при } x > b. \end{cases}$$

Плотность *усеченного* распределения $f_X(x \mid a < x \le b)$ равна нулю вне интервала (a,b) и для $x \in (a,b)$ она определяется в виде

$$f_X(x \mid a < x \le b) = f_X(x) / \int_a^b f_X(z)dz\,,\ a < x < b\,.$$

Отметим, что как a, так и b могут быть бесконечными величинами. Приведем пример использования усеченных распределений. При решении задач экономической статистики налоговые органы обычно интересуются распределением годовых доходов лишь тех лиц, годовой доход которых превосходит некоторый уровень x_0, установленный законами о налогообложении. Это распределение обычно описывают *распределением Парето*, которое задается выражением $P(X > x) = (x/x_0)^{\alpha}$, $x > x_0$, $\alpha > 0$. Плотность этого распределения записывается в виде

$$f_X(x) = \begin{cases} \dfrac{\alpha}{x_0}\left(\dfrac{x_0}{x}\right)^{\alpha+1}, & \text{при } x > x_0\,, \\ 0, & \text{при } x \le x_0\,. \end{cases}$$

При $\alpha > 1$ математическое ожидание равно $M(X) = \alpha x_0/(\alpha - 1)$. Отметим, что распределение Парето есть усечение на интервале (x_0, ∞) степенного распределения с параметром α, имеющего плотность $f_X(x,\alpha) = \alpha x^{-(\alpha+1)}$, $0 < x < \infty$.

1.4.90. Связь между распределениями Коши и Лапласа. Отметим, что распределение Коши (см. П.2.56) входит в семейство распределений Стьюдента (см. П.2.89) как частный случай при числе степеней свободы равном единице.

Если с.в. X имеет распределение Коши с параметрами λ и μ, кратко это записывают в виде $L(X) = Ko(\lambda, \mu)$ и эта запись означает (см. П.2.56), что плотность с.в. X записывается в виде

$$f_X(x;\lambda,\mu) = \frac{1}{\pi} \cdot \frac{\lambda}{\lambda^2 + (x-\mu)^2} = \frac{1}{\pi\lambda} \cdot \frac{1}{1 + [(x-\mu)/\lambda]^2}, \quad -\infty < x < +\infty,$$

а характеристическая функция равна $\exp\{\mu it - \lambda |t|\}$. Отметим, что распределение Коши унимодальное (но не сильно одновершинное, см. упр. 9.8.2) и симметричное относительно точки $x = \mu$, которая совпадает с модой и медианой этого распределения. Распределение Коши не имеет моментов, квартили равны $\mu + \lambda$ и $\mu - \lambda$, интер-квартильный размах равен λ. Вид характеристической функции показывает, что распределение Коши воспроизводимо по своим параметрам, то есть свертка двух распределений Коши дает распределение Коши. Эта теорема сложения записывается в виде $f_{X_1}(x;\lambda_1,\mu_1) * f_{X_2}(x;\lambda_2,\mu_2) = f_X(x;\lambda_1+\lambda_2,\mu_1+\mu_2)$ и означает, что если $L(X_1) = Ko(\lambda_1,\mu_1)$ и $L(X_2) = Ko(\lambda_2,\mu_2)$, то $L(X_1 + X_2) = Ko(\lambda_1+\lambda_2, \mu_1+\mu_2)$. Отсюда следует интересное свойство распределения Коши: *Если* $X_1,...X_n$ *н.о.р. случайные величины, причем* $L(X_i) = Ko(\lambda,\mu)$, $i = 1,...,n$, *то среднее арифметическое* $\overline{X} = (X_1 + ... + X_n)/n$ *имеет то же распределение Коши, то есть* $L(\overline{X}) = Ko(\lambda,\mu)$.

Рассмотрим интеграл

$$\frac{1}{2}\int_{-\infty}^{+\infty} e^{itx-|x|}dx = \int_0^{\infty} \cos tx \cdot e^{-x}dx = \frac{t\sin tx - \cos tx}{1+t^2}\Bigg|_0^{\infty} = \frac{1}{1+t^2}.$$

Это выражение можно рассматривать как характеристическую функцию, соответствующую плотности Лапласа $f(x) = 2^{-1}e^{-|x|}$ (см. П.2.51). Эта характеристическая функция интегрируема на интервале $(-\infty, \infty)$, в результате получаем

$$\frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{e^{-itx}}{1+t^2}dt = e^{-|x|}.$$

С помощью линейного преобразования плотность Лапласа запишется в виде

$$f(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}}, \quad -\infty < x < +\infty,$$

а ее характеристическая функция определяется формулой

$$\exp\{\mu i t\}/(1+\lambda^2 t^2).$$

1.4.91. Пусть $X_1,...X_n$ – н.о.р. случайные величины, причем $L(X_i)=N(0\,,\sigma^2)$, $i=1,...,n$. Рассмотрим случайную величину $\xi=(X_1^2+...+X_n^2)$. Убедитесь, что плотность с.в. ξ записывается в виде

$$f_\xi(x)=\frac{1}{\sigma^n\,2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2\sigma^2}},\ 0<x<\infty.$$

Решение. Так как $L(X_i)=N(0\,,\sigma^2)$, то $L(X_i/\sigma)=N(0\,,1)$, и, следовательно, с.в. ξ можно записать в виде $\xi=X^2(n)/\sigma^2$, где с.в. $X^2(n)$ имеет хи-квадрат распределение с n степенями свободы, то есть $L(X^2(n))=\chi^2(n)$. Далее, используя (1.3.17) и (1.4.28), получаем

$$f_\xi(x)=\frac{1}{\sigma^2}f_{\chi^2(n)}\left(\frac{x}{\sigma^2}\right)=\frac{1}{\sigma^n\,2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}e^{-\frac{x}{2\sigma^2}},\ 0<x<\infty.$$

1.4.92. Пример, взятый из [36, с.267]. Пусть $X_1,...X_n$ – н.о.р. случайные величины, причем $L(X_i)=N(0\,,\sigma^2)$, $i=1,...,n$. Рассмотрим случайную величину $\xi=\sqrt{n}X_1/(X_1^2+...+X_n^2)^{1/2}$. Отметим, что $\xi^2\le n$, поэтому плотность с.в. ξ будет равна нулю вне интервала $(-\sqrt{n}\,,\sqrt{n})$.

Убедитесь, что плотность с.в. ξ записывается в виде

$$f_\xi(x)=\frac{1}{\sqrt{n\pi}}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\left(1-\frac{x^2}{n}\right)^{(n-3)/2},\ -\sqrt{n}\le x\le\sqrt{n}.$$

Решение. Числитель и знаменатель в представлении с.в. ξ являются зависимыми случайными величинами, поэтому ее распределение не может быть получено по схеме получения распределения Стьюдента. Рассмотрим новую случайную величину $\varsigma=\sqrt{n-1}\,X_1/(X_2^2+...+X_n^2)^{1/2}$. По построению (см. формулу (1.4.66)) эта случайная величина имеет распределение Стьюдента с числом степеней свободы $(n-1)$. Обозначим функцию распределения с.в. $T(n)$, имеющей распределение Стьюдента с числом степеней свободы n через $F_{T(n)}(x)$, а ее плотность через $f_{T(n)}(x)$. Отметим, что с.в. ς выражается через с.в. ξ в виде $\varsigma=\xi\sqrt{n-1}/\sqrt{n-\xi^2}=\sqrt{n-1}\,X_1/(X_2^2+...+X_n^2)^{1/2}$. Далее, можно убедиться (детали см. в [36, с.267]), что если ξ возрастает от $-\sqrt{n}$ до $\sqrt{n}$ то, ς постоянно возрастает от $-\infty$ до $+\infty$. Следовательно, неравенство $\xi\le x$ эквивалентно неравенству $\varsigma\le x\sqrt{n-1}/\sqrt{n-x^2}$. С учетом этого замечания,

получаем

$$F_\xi(x) = P\{\xi \le x\} = P\{\varsigma \le x\sqrt{n-1}/\sqrt{n-x^2}\} = F_\varsigma(x\sqrt{n-1}/\sqrt{n-x^2}) =$$
$$= F_{T(n-1)}\left(\sqrt{\frac{n-1}{n}}\frac{x}{\sqrt{1-(x^2/n)}}\right).$$

Отсюда, учитывая (1.4.69), получаем выражение для плотности с.в. ξ в виде

$$f_\xi(x) = \sqrt{\frac{n-1}{n}}\left(1-\frac{x^2}{n}\right)^{-3/2} \cdot f_{T(n-1)}\left(\sqrt{\frac{n-1}{n}}\frac{x}{\sqrt{1-(x^2/n)}}\right) =$$
$$= \frac{1}{\sqrt{n\pi}}\frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\left(1-\frac{x^2}{n}\right)^{(n-3)/2}, \ -\sqrt{n} \le x \le \sqrt{n}.$$

Отметим, что при $n=2$ плотность $f_\xi(x)$ с.в. ξ имеет «*U-образную*» форму, то есть имеет минимум в точке среднего значения $x=0$. При $n=3$ плотность постоянна, то есть имеет равномерное распределение на интервале $(-\sqrt{3}, \sqrt{3})$. При $n>3$ это распределение унимодальное и симметричное относительно среднего значения $x=0$. При всех n среднее значение этого распределения равно нулю, а дисперсия равна единице.

1.4.93. Рассмотрим функции $\varphi(x)$ и $F(x)$, определенные на интервале $[a,b]$. Пусть $a = x_0 < x_1 < ... < x_{n-1} < x_n = b$ – разбиение интервала $[a,b]$ на интервалы $\Delta_i = x_i - x_{i-1}$, $\xi_i \in \Delta_i$, $i = 1,...,n$, $\lambda = \max(\Delta_1,...,\Delta_n)$. Обозначим множество всевозможных разбиений интервала $[a,b]$ через $\{\Delta_1,...,\Delta_n\}$. Предполагаем, что $\varphi(x)$ – непрерывная на $[a,b]$ функция, а функция $F(x)$ – функция с ограниченной вариацией, то есть для неё выполняется выражение

$$\max_{\{\Delta_1,...,\Delta_n\}} \sum_{i=1}^{n} |F(x_i) - F(x_{i-1})| < \infty. \qquad \textbf{(1.4.94)}$$

Определенный интеграл Римана от функции $\varphi(x)$ на интервале $[a,b]$ определяется в виде предела интегральных сумм Римана функции $\varphi(x)$ на $[a,b]$, то есть

$$\int_a^b \varphi(x)dx = \lim_{\lambda \to 0} \sum_{i=1}^{n} \varphi(\xi_i)(x_i - x_{i-1}). \qquad \textbf{(1.4.95)}$$

Обобщением интеграла Римана является интеграл Стилтьеса, предложенный им в 1894 году в виде

$$\int_a^b \varphi(x)dF(x) = \lim_{\lambda\to 0}\sum_{i=1}^{n}\varphi(\xi_i)[F(x_i) - F(x_{i-1})], \qquad \textbf{(1.4.96)}$$

где «интегрирующая» функция $F(x)$ имеет ограниченную вариацию, то есть для неё выполняется выражение (1.4.94). Отметим, что если функция $F(x)$ дифференцируема на интервале $[a,b]$, то интеграл Стилтьеса совпадает с интегралом Римана (при условии, что он существует), то есть выполняется равенство

$$\int_a^b \varphi(x)dF(x) = \int_a^b \varphi(x)F'(x)dx.$$

Интеграл Стилтьеса на вещественной прямой R^1 определяется с помощью предела при $a \to -\infty$ (и или $b \to +\infty$). При этом говорят, что интеграл Стилтьеса от функции $\varphi(x)$ по функции $F(x)$ в бесконечных пределах существует, если выполняется выражение

$$\int_{-\infty}^{\infty} |\varphi(x)|\, dF(x) < \infty.$$

Рассмотрим пример, когда «интегрирующая» функция $F(x)$ является вырожденной в точке x_0 функцией распределения вероятностей случайной величины X, то есть она приписывает единичную вероятностную массу точке x_0 и записывается в виде $F_X(x) = C(x - x_0)$, $x \in R^1$, где $C(x - x_0)$ – единичная функция Хевисайда, определяемая в виде

$$C(x - x_0) = \begin{cases} 1, & x \ge x_0 \\ 0, & x < x_0 \end{cases}. \qquad \textbf{(1.4.97)}$$

Отметим, что $C(x - x_0)$ не является дифференцируемой функцией, однако формально можно записать следующее выражение:

$$dC(x - x_0) = \delta(x - x_0)dx, \qquad \textbf{(1.4.98)}$$

где $\delta(x - x_0)$ – дельта-функция Дирака, определяемая в виде

$$\delta(x - x_0) = \begin{cases} \infty, & x = x_0 \\ 0, & x \ne x_0 \end{cases}. \qquad \textbf{(1.4.99)}$$

Согласно (1.4.96), имеем

$$\int_a^b \varphi(x)dC(x - x_0) = \lim_{\lambda\to 0}\sum_{i=1}^{n}\varphi(\xi_i)[C(x_i - x_0) - C(x_{i-1} - x_0)] = \varphi(x_0).$$

В предыдущем выражении последнее равенство выполняется, так как, равенство $[c(x_i - x_0) - c(x_{i-1} - x_0)] = 1$ выполняется лишь для случая, когда $x_0 \in \Delta_i$ с учетом того, что $\varphi(x)$ – непрерывная функция на интервале $[a,b]$, а при $\lambda \to 0$ $\xi_i \to x_0$ и $\varphi(\xi_i) \to \varphi(x_0)$. Таким образом, для любой непрерывной функции $\varphi(x)$ в точке x_0 справедливо выражение

$$\int_a^b \varphi(x)\delta(x - x_0)dx = \varphi(x_0). \qquad \textbf{(1.4.100)}$$

и, следовательно, для случайной величины X с вырожденным в точке x_0 распределением имеем $M[\varphi(X)] = \varphi(x_0)$.

Глава 2

ХАРАКТЕРИСТИКА ПОДХОДОВ И МЕТОДОВ ТЕОРИИ ОЦЕНОК ПАРАМЕТРОВ

В данной главе обсуждаются общие свойства точечных и интервальных оценок параметров. В разд. 2.3 – 2.5 рассматривается функция правдоподобия и количество информации Фишера, а также неравенство Рао – Крамера и достаточные статистики. Понятие асимптотической эффективности оценок параметров обсуждается в разд. 2.6. Методы построения оценок параметров и их общие свойства приводятся в разд. 2.7 – 2.10. Методы построения интервальных оценок параметров приводятся в разд. 2.12 – 2.14. Упражнения, задачи и дополнения приведены в разд. 2.15. При написании данной главы использовались работы [5, 7, 22, 24, 36, 37].

Задачи математической статистики возникают в тех случаях, когда ф.р. F_X изучаемой с.в. X частично либо полностью неизвестна. Различная степень неизвестности относительно ф.р. F_X, которая возникает при решении конкретных практических задач, описывается при задании статистической модели $(\mathrm{X}, \Im)$ путем конкретизации выборочного пространства X и множества $\Im = \{F_X\}$ – допустимых в условиях данного эксперимента функций распределения вероятностей. Основные общие типы статистических моделей для случая, когда выборка формируется путем n повторных и независимых опытов, были обсуждены в разд. 1.2, и при этом было выделено три типа моделей: параметрические, непараметрические и модели робастной статистики. Для каждого из этих типов моделей возможно большое разнообразие при формулировке конкретных статистических задач, которые будут подробно обсуждаться ниже, здесь же обсудим общие подходы, понятия и методы на примерах решения статистических задач в рамках параметрической модели.

Можно выделить два основных и тесно связанных раздела решения статистических задач:

- теория оценивания параметров,
- теория проверки статистических гипотез.

2.1. Основные понятия теории оценок параметров

В математике используют достаточно общее понятие «*параметр*» (этот термин происходит от греческого слова παραμετρων – отмеривающий) и используют его для обозначения величины, с помощью которой различают между собой элементы некоторого множества. В математической статистике при задании статистической модели $(X, \Im)$ элементами множества $\Im$ являются функции распределения вероятностей, которые различают с помощью понятия «параметр». При этом понятие «параметр» используют в следующих двух вариантах.

В первом варианте это понятие используется при задании множества $\Im$ допустимых ф.р. в виде параметрического семейства функций, то есть в виде

$$\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}, \qquad \textbf{(2.1.1)}$$

причем функциональный характер ф.р. $F_X(x,\theta)$ задается, а параметр θ, который может быть как скалярным, так и векторным, считается неизвестным, но известно, что θ принадлежит заданному параметрическому множеству Θ.

Во втором варианте понятие «параметр» используют для определения числовых характеристик распределений, являющихся элементами множества $\Im$, то есть параметр θ представляется в виде функционала $T(F)$, который является измеримым отображением множества $\Im$ в действительную ось R^1. Этот факт записывают в виде

$$\theta = T(F),\ F \in \Im :\ \Im \to R^1. \qquad \textbf{(2.1.2)}$$

Отметим, что при задании параметрического семейства в виде (2.1.1) параметр θ может иметь самый различный смысл. Он может иметь явно выраженную вероятностную интерпретацию, например

гауссовское семейство распределений характеризуется векторным параметром $\vec{\theta} = (\theta_1, \theta_2)$, где θ_1 – математическое ожидание, а θ_2 – дисперсия нормальной случайной величины. Далее, параметр θ может быть связан с определением носителя распределения вероятностей, например при задании равномерного распределения в интервале $[0, \theta]$; или он может характеризовать форму распределения в семействе гамма-распределений (см. П.2.18); или может быть связан с параметром сдвига или с масштабным параметром заданного семейства распределений и т.д. Подчеркнем еще раз, что принципиальным моментом в задании параметрического семейства (2.1.1) является тот факт, что *функциональный вид распределения считается известным*. Именно эта информация существенно используется при разработке различных методов оценивания неизвестных параметров θ на основе информации, доставляемой выборкой. Функциональный характер распределения используется при реализации метода максимального правдоподобия (см. разд. 2.10), метода моментов (см. разд. 2.9) и метода минимума расстояний (разд. 2.11).

Можно выделить два подхода к построению оценок неизвестных параметров:

- точечные оценки параметров,
- интервальные оценки (доверительные интервалы).

Ниже мы рассмотрим основные методы построения оценок параметров и их общие свойства для случая, когда модель задана в виде (2.1.1).

Обсудим теперь кратко второй вариант использования понятия «параметр». Принципиальным отличием этого варианта от первого является тот факт, что распределение вероятностей F_X изучаемой случайной величины X считается полностью *неизвестным*. Исключением могут быть лишь такие общие предположения, как принадлежность ф.р. F_X к классу непрерывных или симметричных распределений. Далее, интересующий нас параметр θ во втором варианте часто имеет явно выраженную вероятностную интерпретацию. Обычно параметр θ характеризует некоторую числовую характеристику распределения F_X и записывается в виде функционала $T(F)$,

заданного на множестве допустимых функций распределений $\Im = \{F_X\}$, то есть записывается в виде (2.1.2). Основным методом оценивания параметров второго типа является метод подстановки, при реализации которого не используется информация о функциональном характере распределений, а используется лишь эмпирическая функция распределения, построенная по выборке (см. разд. 2.7 и 2.8).

2.2. Свойства точечных оценок параметров

Пусть имеется выборка $X_1,...,X_n$ из распределения F_X, которое является элементом параметрического семейства функций $\Im_\theta$ вида (2.1.1). То есть предполагается, что распределение $F_X(x,\theta)$ имеет известный функциональный вид, но зависит от неизвестного параметра θ (скалярного или векторного), который является элементом заданного параметрического множества Θ. Требуется по выборке $X_1,...,X_n$ построить оценку истинного значения θ_0 параметра θ. В данном разделе обсудим точечные оценки параметров, а интервальные оценки рассмотрим ниже, в разд. 2.12 – 2.14. Введем необходимые определения.

Определение 2.2.1. Статистикой $T(X_1,...,X_n)$ называют всякую случайную величину, которая является функцией только от выборки $X_1,...,X_n$.

Определение 2.2.2. Точечной оценкой параметра θ называют такую статистику $T_n = T(X_1,...,X_n)$, которая не зависит от оцениваемого параметра θ и значение которой при заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ принимается в качестве приближения к истинному значению θ_0 оцениваемого параметра θ.

Отметим, что везде ниже вместо слов «оценка T_n для истинного значения θ_0 параметра θ» будем кратко говорить: «оценка T_n для параметра θ». Отметим также, что для оценивания интересующего нас параметра θ можно предложить различные точечные оценки и для того, чтобы выбрать лучшую из них, нужно конкретизировать

критерии качества оценок, которые должны быть основаны на различных мерах близости между оценкой и истинным значением оцениваемого параметра. Так как всякая оценка $T(X_1,...,X_n)$ является случайной величиной, то судить о том «хорошо» или «плохо» оценка $T(X_1,...,X_n)$ отражает истинное значение оцениваемого параметра можно только в вероятностных терминах, положив в основу общее требование о том, чтобы распределение вероятности случайной величины $T(X_1,...,X_n)$ концентрировалось бы около истинного значения оцениваемого параметра. Чем больше степень концентрации распределения (или чем меньше некоторая характеристика разброса распределения), тем лучше соответствующая точечная оценка. Формализованное выражение этого общего требования приводит к таким понятиям, как несмещенность, состоятельность, среднеквадратическая ошибка (СКО) и эффективность. Обсудим эти понятия.

Определение 2.2.3. Статистику $T(X_1,...,X_n)$ называют несмещенной точечной оценкой для параметра θ, если математическое ожидание оценки совпадает с оцениваемым параметром, то есть если выполняется следующее выражение:

$$M[T(X_1,...,X_n)] = \theta,\ \forall\theta \in \Theta. \qquad \textbf{(2.2.4)}$$

Отметим, что если равенство (2.2.4) не выполняется, то оценку $T(X_1,...,X_n)$ называют смещенной и величину

$$b(\theta) = M[T(X_1,...,X_n)] - \theta$$

называют смещением оценки $T(X_1,...,X_n)$.

Определение 2.2.5. Среднеквадратической ошибкой оценки $T(X_1,...,X_n)$ параметра θ называют величину, равную математическому ожиданию квадрата уклонения оценки от оцениваемого параметра, то есть

$$CKO[T(X_1,...,X_n)] = M\{[T(X_1,...,X_n) - \theta]^2\}. \qquad \textbf{(2.2.6)}$$

Отметим, что среднеквадратическая ошибка несмещенной оценки совпадает с ее дисперсией, поскольку

$$CKO[T(X_1,...,X_n)] = D[T(X_1,...,X_n)] + b^2(\theta). \qquad \textbf{(2.2.7)}$$

В справедливости (2.2.7) легко убедиться (см. упр. 2.15.19), расписав выражение

$$M\{[T(X_1,...,X_n) - M(T(X_1,...X_n)) + M(T(X_1,...,X_n)) - \theta]^2\}.$$

Замечание 2.2.8. Обычно при построении точечных оценок требуют, чтобы они удовлетворяли некоторым желательным свойствам. Например, требуют, чтобы они обладали бы свойством несмещенности или были бы эквивариантны к линейным преобразованиям наблюдений и т.д. В таких случаях говорят, что оценки принадлежат заданному классу. Далее, если для таких оценок в качестве критерия выбрана некоторая мера близости оценки к истинному значению оцениваемого параметра (например, выбрана среднеквадратическая ошибка) и если конкретная оценка минимизирует выбранную меру близости, то такую оценку называют оптимальной в данном классе оценок в смысле выбранного критерия качества. Отметим также, что оценку $V(X_1,...X_n)$ параметра θ называют эквивариантной относительно линейных преобразований (см., например, [20]), если для нее выполняется равенство $V(ax_1 + b,...,ax_n + b) = aV(x_1,...,x_n) + b$. Это свойство эквивариантности оценок относительно сдвига позволяет положить, без потери общности, $\theta = 0$ при изучении свойств распределений оценок.

В приведенном определении (2.2.3) несмещенной оценки использовалась операция усреднения путем вычисления математического ожидания оценки. По этой причине в литературе оценку $T_n(X_1,...,X_n)$, для которой выполняется (2.2.4), иногда называют *несмещенной в среднем*. Отметим, что наряду с понятием несмещенности в среднем также используют понятия модальной несмещенности и медианной несмещенности (см., например, [11а, с.110]).

Замечание 2.2.9. Сравнение различных точечных оценок часто проводится в асимптотике (см. разд. 2.6), то есть в случае, когда объем выборки достаточно большой. В частности, если условие (2.2.4) выполняется лишь при $n \to \infty$, то такую последовательность оценок $T_n(X_1,...,X_n)$ называют асимптотически несмещенной. К асимптотическим понятиям также относится понятие состоятельности оценок, которое на качественном уровне отражает естественное требование к увеличению точности оценивания при увеличении объема выборки. Приведем определение этого понятия.

Определение 2.2.10. Последовательность оценок $T_n(X_1,...,X_n)$ параметра θ называют состоятельной, если она сходится по вероятности к оцениваемому параметру θ, то есть если

$T_n(X_1,...,X_n) \xrightarrow{P} \theta$, $n \to \infty$, и эта запись означает, что выполняется следующее выражение:

$$\forall \varepsilon > 0 \quad P\{\,|T_n(X_1,...,X_n) - \theta| < \varepsilon\,\} \to 1, \quad n \to \infty. \qquad \textbf{(2.2.11)}$$

Замечание 2.2.12. Отметим, что понятие эффективных оценок и асимптотической эффективности оценок, используемое для сравнения состоятельных оценок при больших объемах выборки, будет введено ниже (см. разд. 2.6). Различные типы сходимости случайных величин, включая и сходимость по вероятности, а также полезный критерий состоятельности оценок, обсуждаются в Приложении 3.

2.3. Функция правдоподобия и количество информации Фишера

Среди различных критериев качества оценок наиболее часто используют среднеквадратическую ошибку. Кроме того, среди общих свойств оценок требование несмещенности также является желательным, поскольку оно означает, что, по крайней мере «в среднем», несмещенная оценка совпадает с истинным значением оцениваемого параметра. Поэтому, если мы ограничиваемся классом несмещенных оценок, при этом условии среднеквадратическая ошибка оценки совпадает с ее дисперсией, то естественно в качестве критерия приемлемости оценки использовать ее дисперсию. Отметим, что при построении оценок параметров для случая, когда модель задана в параметрической форме, при выполнении достаточно общих условий регулярности, существует нижняя граница дисперсии оценок. Эта граница устанавливается неравенством Рао – Крамера. (см. разд. 2.4). Для того чтобы охарактеризовать это неравенство и ввести понятие эффективных оценок, нам понадобятся некоторые дополнительные понятия, такие, как функция правдоподобия, количество информации Фишера и некоторые предварительные результаты, которые также будут использованы в дальнейшем и для других целей.

Итак, пусть статистическая модель задана в параметрической форме, то есть в виде (2.1.1). Это означает, что имеется выборка

$X_1,...,X_n$ из распределения $F_X(x,\theta)$, функциональный вид которого известен, то есть предполагаем, что $F_X(x,\theta) \in \Im_\theta$ и пусть плотность $f_X(x,\theta)$, $x \in R^1$, $\theta \in \Theta$, удовлетворяет условиям регулярности, то есть она непрерывна, дважды дифференцируема и ее носитель не зависит от параметра θ (обсуждение условий регулярности приводится в [38, с.113]). Требуется, используя данную информацию, по выборке $X_1,...,X_n$ построить оценку некоторой заданной и дифференцируемой функции $\tau(\theta)$, $\theta \in \Theta$. Оценка самого параметра θ является частным случаем этой задачи при $\tau(\theta) = \theta$. Для построения точечных оценок параметров в случае, когда модель задана в параметрической форме, разработаны различные методы, такие, как метод моментов (см. разд. 2.9), метод максимального правдоподобия (см. разд. 2.10) и метод минимума расстояний (см. разд. 2.11). Суть этих методов и основные свойства оценок, построенных этими методами, обсуждаются в указанных разделах ниже. Далее нам потребуются следующие определения.

Определение 2.3.1. Совместная плотность распределения вероятностей $f_{X_1,...,X_n}(x_1,...,x_n;\theta)$ случайных величин $X_1,...,X_n$, рассматриваемая как функция параметра θ при фиксированной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$, называется функцией правдоподобия.

В литературе функция правдоподобия обычно обозначается через $L(x_1,...,x_n;\theta)$, или коротко в виде $L(\vec{x};\theta)$. Отметим, что при построении функции правдоподобия $L(\vec{x};\theta)$ аргументы $x_1,...,x_n$ (переменные величины) совместной плотности $f_{X_1,...,X_n}(x_1,...,x_n;\theta)$ случайного вектора $\vec{X} = (X_1,...,X_n)$ заменяются на наблюденную реализацию $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$. Отметим также, что $L(\vec{x};\theta)$ является наблюденной реализацией статистики правдоподобия $L(\vec{X};\theta)$. По этой причине оценка параметра θ, найденная методом максимального правдоподобия (см. разд. 2.10), будет зависеть от результатов наблюдений $\vec{x} = (x_1,...,x_n)$ и ее вычисленное значение $\hat{\theta}(x_1,...,x_n)$, при котором функция правдоподобия максимальна, яв-

ляется наблюденной реализацией случайной величины $\hat{\theta}(X_1,...,X_n)$. Для рассматриваемой параметрической модели случайные величины $X_1,...,X_n$ независимы и одинаково распределены, поэтому, согласно (1.1.72), функция правдоподобия записывается в виде

$$L(x_1,...,x_n;\theta) = f_{X_1,...,X_n}(x_1,...,x_n;\theta) = \prod_{i=1}^{n} f_X(x_i;\theta). \qquad \textbf{(2.3.2)}$$

Отметим, что логарифм функции правдоподобия, то есть выражение

$$\ln L(\vec{x},\theta) = \sum_{i=1}^{n} \ln f_X(x_i,\theta), \qquad \textbf{(2.3.2a)}$$

часто называют логарифмической функцией правдоподобия. Если функция правдоподобия $L(\vec{x};\theta)$ дифференцируема по скалярному параметру θ, то на основе статистики $L(\vec{X};\theta)$ вводят дополнительное понятие, которое называют функцией вклада выборки $\vec{X} = (X_1,...,X_n)$ и определяют ее следующим образом.

Определение 2.3.3. Случайную величину $U(\vec{X},\theta)$, равную логарифмической производной статистики правдоподобия $L(\vec{X};\theta)$, называют вкладом (или функцией вклада) выборки $\vec{X} = (X_1,...,X_n)$ и определяют ее в виде

$$U(\vec{X},\theta) = \frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i;\theta)}{\partial\theta}. \qquad \textbf{(2.3.4)}$$

Отметим, что *i*-е слагаемое в (2.3.4) называют вкладом отдельного наблюдения X_i, $i = 1,...,n$. Если параметр θ является векторным параметром размерности r, то есть $\vec{\theta} = (\theta_1,...,\theta_r)$, то под вкладом выборки $\vec{X} = (X_1,...,X_n)$ понимают случайный вектор

$(U_1(\vec{X};\vec{\theta}),...,U_r(\vec{X};\vec{\theta}))$, где $U_i(\vec{X};\vec{\theta}) = \partial \ln L(\vec{X};\vec{\theta})/\partial\theta_i$, $i = 1,...,r$.

Обсудим теперь некоторые общие свойства функции правдоподобия. Функция правдоподобия определяется совместной плотностью случайных величин $X_1,...,X_n$, которая рассматривается как функция параметра θ при фиксированной реализации $\vec{x} = (x_1,...,x_n)$

выборки $X_1,...,X_n$, поэтому она удовлетворяет условию нормировки и, следовательно, согласно (1.3.65), выполняется выражение вида

$$\int_{\mathrm{X}} ... \int L(x_1,...,x_n;\theta)dx_1...dx_n = 1. \qquad \textbf{(2.3.5)}$$

С учетом того, что плотность $f_X(x,\theta)$ удовлетворяет условиям регулярности, а это означает, что существуют две производные по θ функции $L(\vec{x};\theta)$ и носитель $f_X(x,\theta)$ не зависит от параметра θ, можем изменить порядок дифференцирования и интегрирования и из (2.3.5) получим

$$\int_{\mathrm{X}} ... \int \frac{\partial}{\partial\theta} L(\vec{x};\theta)\prod_{i=1}^{n} dx_i = 0. \qquad \textbf{(2.3.6)}$$

Выражение (2.3.6) можно переписать в виде

$$M\left[\frac{\partial}{\partial\theta}\ln L(\vec{X};\theta)\right] = \int_{\mathrm{X}} ... \int \frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} L(\vec{x};\theta)\prod_{i=1}^{n} dx_i = M[U(\vec{X};\theta)] = 0. \qquad \textbf{(2.3.7)}$$

Таким образом, математическое ожидание от логарифмической производной статистики правдоподобия, или от функции вклада выборки, равно нулю. Далее, дифференцируя обе части (2.3.7) по θ с последующей заменой порядка дифференцирования и интегрирования, получим следующее выражение:

$$\int_{\mathrm{X}} ... \int \left\{ \frac{\partial^2 \ln L(\vec{x};\theta)}{\partial\theta^2} + \left[\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} \right]^2 \right\} L(\vec{x};\theta)\prod_{i=1}^{n} dx_i = 0. \qquad \textbf{(2.3.8)}$$

Отсюда следует, что математическое ожидание от квадрата логарифмической производной статистики правдоподобия равняется с минусом математическому ожиданию от второй производной логарифма статистики правдоподобия, то есть выполняется равенство

$$M\left[\frac{\partial}{\partial\theta}\ln L(\vec{X};\theta)\right]^2 = -M\left[\frac{\partial^2}{\partial\theta^2}\ln L(\vec{X};\theta)\right]. \qquad \textbf{(2.3.9)}$$

В терминах функции вклада $U(\vec{X},\theta)$ равенство (2.3.9) перепишется в виде

$$M[U^2(\vec{X};\theta)] = -M[U'(\vec{X};\theta)]. \qquad \textbf{(2.3.9a)}$$

Определение 2.3.10. Количество информации Фишера о параметре θ, содержащееся в одном наблюдении, определяется в виде

$$i_1(\theta) = M\left[\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right]^2 = \int_{-\infty}^{+\infty}\left[\frac{\partial \ln f(x;\theta)}{\partial\theta}\right]^2 f(x;\theta)dx =$$

$$= -M\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1;\theta)\right] = -\int_{-\infty}^{+\infty}\left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\right] f(x;\theta)dx. \qquad \textbf{(2.3.11)}$$

Определение 2.3.12. Количество информации Фишера, содержащееся в выборке $X_1,...,X_n$ объема n относительно параметра θ, определяется в виде

$$i_n(\theta) = M\left[\frac{\partial}{\partial\theta}\ln L(\vec{X};\theta)\right]^2 = \int_{\mathrm{X}}...\int\left[\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta}\right]^2 L(\vec{x};\theta)\prod_{i=1}^{n}dx_i =$$

$$= -M\left[\frac{\partial^2}{\partial\theta^2}\ln L(\vec{X};\theta)\right] = -\int_{\mathrm{X}}...\int\frac{\partial^2 \ln L(\vec{x};\theta)}{\partial\theta^2} L(\vec{x};\theta)\prod_{i=1}^{n}dx_i. \qquad \textbf{(2.3.13)}$$

Отметим, что через функцию вклада $U(\vec{X},\theta)$, определенную в (2.3.4), количество информации Фишера $i_n(\theta)$ записывается в виде

$$i_n(\theta) = M[U^2(\vec{X};\theta)] = D[U(\vec{X};\theta)] = -M[U'(\vec{X};\theta)]. \qquad \textbf{(2.3.13a)}$$

Отметим также, что случайные величины $X_1,...,X_n$ независимы и одинаково распределены, поэтому для выборки $X_1,...,X_n$ объема n количество информации Фишера $i_n(\theta)$ проще вычислить, используя формулу (2.3.11), поскольку

$$i_n(\theta) = n \cdot i_1(\theta). \qquad \textbf{(2.3.14)}$$

Для проверки справедливости равенства (2.3.14) при указанных предположениях см. упр. 2.15.1.

Замечание 2.3.15. Пусть X – дискретная случайная величина с конечным числом k возможных состояний $\tilde{x}_1,...,\tilde{x}_k$, заданная распределением вероятностей вида

$$\begin{pmatrix} X \\ p \end{pmatrix} : \begin{pmatrix} \tilde{x}_1,.....,\tilde{x}_k \\ p_1(\theta),...,p_k(\theta) \end{pmatrix}, \ \sum_{i=1}^{k} p_i(\theta) = 1, \ \theta \in \Theta,$$

где вероятности $P_\theta\{X = \tilde{x}_i\}$ осуществления возможных состояний будут обозначаться через $p_i(\theta)$, или через $p(\tilde{x}_i;\theta)$, $i = 1,...,k$. Пусть имеется выборка $X_1,...,X_n$ объема n из совокупности объектов, признак X которых описывается функцией распределения вероятностей вида

$$F_X(x;\theta) = \sum_{i=1}^{k} p_i(\theta) \cdot C(x - \tilde{x}_i).$$

По наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ определяется дискретный вариационный ряд частот, то есть

$$(x_1,...,x_n) \Rightarrow \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ n_1,...,n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n.$$

По данным этого вариационного ряда частот функция правдоподобия запишется в виде

$$L(x_1,...,x_n;\theta) = P\{X_1 = \tilde{x}_1,...,X_n = \tilde{x}_n\} = p_1^{n_1}(\theta) \cdots p_k^{n_k}(\theta). \qquad \textbf{(2.3.16)}$$

Пример 2.3.17. Пусть $L(X) = Bi(1,\theta)$. Это предположение означает, что дискретная с.в. X имеет распределение Бернулли с параметром θ вида

$$\begin{pmatrix} X \\ P \end{pmatrix} : \begin{pmatrix} 0, & 1 \\ 1-\theta, & \theta \end{pmatrix},$$

где неизвестный параметр θ удовлетворяет неравенствам $0 \le \theta \le 1$. Имеется выборка $X_1,...,X_n$ из распределения Бернулли с ф.р. $F_X(x,\theta)$ (см. П.1.4) вида

$$F_X(x,\theta) = \begin{cases} 0, & x < 0, \\ 1-\theta, & 0 \le x < 1, \\ 1, & x \ge 1. \end{cases}$$

Для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия (2.3.16) (см. также пример 2.4.20) запишется в виде

$$L(x_1,...,x_n;\theta) = \theta^m (1-\theta)^{n-m},$$

где $m = \sum x_i$ равняется числу единиц в реализации $\vec{x} = (x_1,...,x_n)$. Логарифмическая производная функции правдоподобия записывается в виде

$$\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} = \frac{m}{\theta} - \frac{n-m}{1-\theta} = \frac{m-n\theta}{\theta(1-\theta)} = \frac{1}{\theta(1-\theta)}\sum_{i=1}^{n}(x_i-\theta). \qquad \textbf{(2.3.18)}$$

Отсюда $$M\left(\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta}\right)^2 = \frac{1}{\theta^2(1-\theta)^2} M\left(\sum_{i=1}^{n}(X_i-\theta)\right)^2 =$$

$$= \frac{1}{\theta^2(1-\theta)^2} M\left(\sum_{i=1}^{n}(X_i-\theta)^2 + \sum_{i\neq j}(X_i-\theta)(X_j-\theta)\right).$$

Учитывая далее, что $M(X_i)=\theta$, $D(X_i)=\theta(1-\theta)$, а также независимость X_i и X_j, $1\le i,j\le n$, получим

$$M\left(\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta}\right)^2 = \frac{1}{\theta^2(1-\theta)^2}\sum_{i=1}^{n} M(X_i-\theta)^2 = \frac{1}{\theta^2(1-\theta)^2} nD(X_1) = \frac{n}{\theta(1-\theta)}.$$

Таким образом, согласно (2.3.13), для распределения Бернулли $Bi(1,\theta)$ количество информации Фишера $i_n(\theta)$ в выборке $X_1,...,X_n$ объема n относительно параметра θ вычисляется по формуле

$$i_n(\theta) = \frac{n}{\theta(1-\theta)}. \qquad \textbf{(2.3.19)}$$

Заметим, что данное выражение также можно получить из (2.3.18) с учетом того, что случайная величина $m=\sum x_i$ имеет биномиальное распределение с параметрами n и θ, то есть $L(m)=Bi(n,\theta)$ и, следовательно, $M(m)=n\theta$, $D(m)=n\theta(1-\theta)$. В самом деле, из (2.3.18) получаем

$$i_n(\theta) = M\left(\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta}\right)^2 = M\left(\frac{m-n\theta}{\theta(1-\theta)}\right)^2 = \frac{D(m)}{\theta^2(1-\theta^2)} = \frac{n\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Отметим также, что формулу (2.3.19) можно получить и другим способом. Для этого воспользуемся критерием существования эффективной оценки вида (2.4.17) и формулой для вычисления ее дисперсии (2.4.18). В самом деле, согласно (2.3.18), логарифмическая производная функции правдоподобия записывается в виде

$$\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} = \frac{1}{\theta(1-\theta)}\sum_{i=1}^{n}(x_i-\theta) = \frac{n}{\theta(1-\theta)}\left(\frac{1}{n}\sum_{i=1}^{n}x_i - \theta\right).$$

Сравнивая данное выражение с (2.4.17), заключаем, что $A^*(\theta)=n/\theta(1-\theta)$ и, согласно (2.4.18), получаем $i_n(\theta)=A^*(\theta)=n/\theta(1-\theta)$. В заключение отметим, что вычисленное количество информации Фишера для различных моделей приведено в упр. 2.15.5, 2.15.6 и 5.8.57.

Замечание 2.3.20. Пусть с.в. X имеет ф.р. $F_X(x-\theta)$, $x \in R^1$, то есть предполагаем, что распределения параметрического семейства $\Im_\theta$ отличаются параметром сдвига $\theta \in R^1$ и оно задано в виде $\Im_\theta = \{F : F(x,\theta) = F_X(x-\theta)\}$, где функции распределения $F_X(x-\theta)$ с.в. X соответствует абсолютно непрерывная плотность $f_X(x-\theta)$, $x \in R^1$. Для параметрической модели с параметром сдвига количество информации Фишера $i_1(\theta)$ о параметре сдвига θ, содержащееся в одном наблюдении, определенное в общем случае в (2.3.11), не зависит от параметра $\theta \in R^1$. Оно определяется только типом распределения, поэтому переобозначим его через $I(f_X)$. Приведем необходимые для дальнейшего определения.

Определение 2.3.21. Если функция $-\ln f(x)$ выпукла в некотором открытом интервале (a,b), таком, что $-\infty \le a < b \le \infty$ и $\int_a^b f(x)dx = 1$, то плотность $f(x)$ называется сильно одновершинной. Такие плотности абсолютно непрерывны на (a,b), и функция $[-\ln f(x)]' = -f'(x)/f(x)$ не убывает (см. упр. 5.8.77).

Определение 2.3.22. Если плотность $f(x)$ абсолютно непрерывна и выполняется неравенство

$$I(f) = \int_{-\infty}^{\infty} \left(\frac{f'(x)}{f(x)} \right)^2 f(x)dx < \infty,$$

то говорят, что плотность $f(x)$ имеет конечное количество информации Фишера относительно параметра сдвига. Если плотность $f(x)$ не является абсолютно непрерывной, то полагают $I(f) = \infty$.

Количество информации Фишера $I(f_X)$ о параметре сдвига θ записывается в виде

$$I(f_X) = \int_{-\infty}^{\infty} \left(\frac{\partial \ln f_X(x-\theta)}{\partial \theta} \right)^2 dF_X(x-\theta) = \int_{-\infty}^{+\infty} \left(-\frac{f_X'(x)}{f_X(x)} \right)^2 f_X(x)dx =$$

$$= \int_0^1 \left(-\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \right)^2 du. \qquad \textbf{(2.3.23)}$$

Заметим, что логарифмическая производная $\partial \ln f(x,\theta)/\partial\theta$ характеризует относительную скорость изменения плотности в точке x. Поэтому математическое ожидание квадрата этой скорости, равное количеству информации Фишера, может быть качественно проинтерпретировано следующим образом.

Чем больше это математическое ожидание в данной точке θ_0, тем легче отличить θ_0 от соседних значений θ и, следовательно, тем точнее может быть оценен параметр θ при $\theta = \theta_0$. Справедливость этой интерпретации количества информации Фишера подтверждается при больших объемах выборки для регулярных параметрических моделей (см. [38, с. 112] и замечание 2.4.15).

Отметим также, что при вычислении меток линейных ранговых статистик (см. разд. 5.4 и монографию [12]) используют φ-функции вида

$$\varphi_f(u) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))},\ 0 < u < 1. \qquad \textbf{(2.3.24)}$$

Использование этих φ-функций также приводит к удобному выражению для вычисления количества информации Фишера $I(f_X)$ о параметре сдвига по формуле (2.3.26). В самом деле, если плотность $f(x)$ является сильно одновершинной, то функция $\varphi_f(u)$ – неубывающая, и если

$$\int_{-\infty}^{\infty} |f'(x)|\,dx < \infty, \text{ то } \int_0^1 \varphi_f(u)du = \int_{-\infty}^{\infty} (-f'(x))dx = 0, \qquad \textbf{(2.3.25)}$$

при условии $I(f) < \infty$, то функция $\varphi_f(u)$ интегрируема в квадрате и количество информации Фишера относительно параметра сдвига вычисляется (см. упр. 2.15.6) по формуле

$$I(f) = \int_0^1 \varphi_f^2(u)du. \qquad \textbf{(2.3.26)}$$

Значения $I(f_X)$ для некоторых стандартных распределений вычислены в упр. 2.15.6 и приведены в табл. 2.3.1.

Таблица 2.3.1

Количество информации Фишера $I(f_X)$ о параметре сдвига для стандартных распределений

Распределение	$F_{(1)}$ – нормальное	$F_{(2)}$ – логистическое	$F_{(3)}$ – Лапласа	$F_{(4)}$ – Коши
$I(f)$	1	1/3	1	1/2

Отметим, что для распределения Лапласа производная $f'(x)$ не существует в точке $x = 0$. Однако плотность распределения Лапласа $f(x)$ является абсолютно непрерывной, и при этом предположении формулы (2.3.21) и (2.3.26)

применимы. Заметим также, что для равномерного в интервале [0, 1] распределения производная $f'(x)=0$ для всех $x\neq 0,1$, так что формальное применение формулы (2.3.23) приводит к неверному значению $I(f)=0$. Поэтому в тех случаях, когда плотность не дифференцируема (равномерная плотность является таковой в точках $x=0$ и $x=1$), полагают количество информации Фишера равным бесконечности (см. [38, с.115]). В частности, для U-образных распределений плотности не дифференцируемы, и для них полагают $I(f)=\infty$ (см. рис. 8.9.19).

Замечание 2.3.27. Пусть случайные величины X и Y имеют непрерывные и дифференцируемые ф.р. $F_X(x)$ и $F_Y(x)$, $x\in R^1$. Рассмотрим случайную величину $Z=X+Y$ с плотностью f_Z, которая вычисляется с помощью интеграла свертки (см. упр. 1.4.1), то есть в виде

$$f_Z(x)=\int_{-\infty}^{+\infty} f_X(x-y)dF_Y(y). \qquad \textbf{(2.3.28)}$$

Можно показать (см. упр. 2.15.9), что выполняется неравенство $I(f_Z)\le I(f_X)$.

2.4. Неравенство Рао – Крамера

Пусть статистическая модель задана в параметрической форме. По выборке $X_1,...,X_n$ из распределения $F(x;\theta)\in\Im_\theta$ требуется оценить некоторую заданную и дифференцируемую функцию $\tau(\theta)$, $\theta\in\Theta$. Предполагаем, что плотность $f(x,\theta)$ удовлетворяет условиям регулярности, сформулированным выше, и пусть класс несмещенных оценок параметрической функции $\tau(\theta)$ обозначен через T_τ. Справедлива следующая теорема.

Теорема 2.4.1. В рамках параметрической модели при выполнении условий регулярности для любой несмещенной оценки $T(X_1,...,X_n)\in T_\tau$ дифференцируемой параметрической функции $\tau(\theta)$, $\theta\in\Theta$, выполняется неравенство Рао – Крамера, которое имеет вид

$$D[T(X_1,...,X_n)]\ge\frac{[\tau'(\theta)]^2}{i_n(\theta)},\ \forall T(\vec{X})\in T_\tau. \qquad \textbf{(2.4.2)}$$

Равенство в этом неравенстве достигается только тогда, когда логарифмическая производная статистики правдоподобия является ли-

нейной функцией от разности $T(X_1,...,X_n)-\tau(\theta)$, то есть

$$\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = A^*(\theta)[T(X_1,...,X_n)-\tau(\theta)], \qquad \textbf{(2.4.3)}$$

где $A^*(\theta)$ – некоторая функция, которая не зависит от наблюдений, но может зависеть от параметра θ.

Доказательство. Так как $T(X_1,...,X_n)$ – несмещенная оценка $\tau(\theta)$, то согласно (2.2.4) имеем

$$\tau(\theta) = \int_X ... \int T(x_1,...,x_n)L(\vec{x};\theta)\prod_{i=1}^{n} dx_i\ ,\ \theta\in\Theta\,. \qquad \textbf{(2.4.4)}$$

Учитывая условия регулярности и дифференцируя обе части (2.4.4) по параметру θ, получим

$$\tau'(\theta) = \int_X ... \int T(x_1,...,x_n)\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta}L(\vec{x};\theta)\prod_{i=1}^{n} dx_i\,. \qquad \textbf{(2.4.5)}$$

Последнее выражение (2.4.5), с учетом (2.3.7), можно переписать в виде

$$\tau'(\theta) = \int_X ... \int [T(x_1,...,x_n)-\tau(\theta)]\sqrt{L(\vec{x};\theta)}\,\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta}\sqrt{L(\vec{x};\theta)}\prod_{i=1}^{n} dx_i\,. \textbf{(2.4.6)}$$

Используя теперь неравенство Коши – Буняковского (см. (2.4.7)), получим

$$[\tau'(\theta)]^2 \le \int_X ... \int [T(x_1,...,x_n)-\tau(\theta)]^2 L(\vec{x};\theta)\prod_{i=1}^{n} dx_i \times$$

$$\times \int_X ... \int [\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta}]^2 L(\vec{x};\theta)\prod_{i=1}^{n} dx_i\,.$$

Далее, учитывая (2.3.13), окончательно имеем

$$[\tau'(\theta)]^2 \le D[T(X_1,...,X_n)]\cdot i_n(\theta),$$

отсюда следует неравенство (2.4.2). Для доказательства выражения (2.4.3) достаточно напомнить, что равенство в неравенстве Коши – Буняковского

$$\left(\int x(t)y(t)dt\right)^2 \le \int x^2(t)dt \cdot \int y^2(t)dt \qquad \textbf{(2.4.7)}$$

достигается тогда и только тогда, когда функции $x(t)$ и $y(t)$ пропорциональны. С учетом этого замечания, из (2.4.6) следует (2.4.3). Доказательство завершено.

Отметим, что неравенство Рао – Крамера (2.4.2) определяет нижнюю границу дисперсий всех несмещенных оценок функции $\tau(\theta)$ для регулярных параметрических моделей. Часто в литературе величину $[\tau'(\theta)]^2 / i_n(\theta)$ называют минимальной границей дисперсии (сокращенно МГД) оценки функции $\tau(\theta)$, а оценку, дисперсия которой при всех θ равна этой минимальной границе, называют МГД-оценкой, или эффективной оценкой $\tau(\theta)$. Для удобства дальнейших ссылок приведем следующее определение.

Определение 2.4.8. Для регулярной параметрической модели $\Im_\theta$ несмещенную оценку $T^*(X_1,...,X_n)$ заданной параметрической функции $\tau(\theta)$, $\theta \in \Theta$, называют эффективной, если ее дисперсия совпадает с нижней границей в неравенстве Рао – Крамера, то есть

$$D[T^*(X_1,...,X_n)] = [\tau'(\theta)]^2 / i_n(\theta). \qquad \textbf{(2.4.9)}$$

2.4.А. Критерий эффективности

Отметим, что для заданной регулярной параметрической модели и для заданной функции $\tau(\theta)$ эффективная оценка для $\tau(\theta)$ существует не всегда. Критерием существования эффективной оценки $T^*(X_1,...,X_n)$ для параметрической функции $\tau(\theta)$ является выражение (2.4.3). Таким образом, эффективная оценка существует, если логарифмическая производная статистики правдоподобия (или функция вклада выборки, определенная в (2.3.3)) является линейной функцией разности $[T^*(X_1,...,X_n) - \tau(\theta)]$, то есть может быть приведена к виду

$$\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = A^*(\theta)[T^*(X_1,...,X_n) - \tau(\theta)], \qquad \textbf{(2.4.10)}$$

причем дисперсия эффективной оценки, кроме формулы (2.4.9),

обычно гораздо проще может быть вычислена (см. упр. 2.15.2) по формуле

$$D[T^*(X_1,...,X_n)] = \left|\frac{\tau'(\theta)}{A^*(\theta)}\right|. \qquad \textbf{(2.4.11)}$$

Замечание 2.4.12. Отметим, что эффективная оценка $T^*(X_1,...,X_n)$ является оптимальной (см. замечание 2.2.8) и выражение (2.4.10) часто называют критерием оптимальности Рао – Крамера. Отметим также, что логарифмическая производная статистики правдоподобия, или функция вклада выборки, однозначно определяется заданной параметрической моделью, поэтому выражение (2.4.10), если оно выполняется, то является единственным, и это означает, что эффективная оценка может существовать только для определенной параметрической функции $\tau(\theta)$ и для других функций от параметра θ она не существует. Неравенство Рао – Крамера (2.4.2) было рассмотрено для случая, когда оценка $T(X_1,...,X_n)$ была несмещенной оценкой для $\tau(\theta)$. Обсудим теперь случай, когда оценка $\tilde{T}(X_1,...,X_n)$ – смещенная и пусть смещение $b(\theta) = M[\tilde{T}(X_1,...,X_n)] - \tau(\theta)$ является дифференцируемой функцией θ. Тогда, записав вместо формулы (2.4.4), выражение

$$\int_X ... \int \tilde{T}(x_1,...,x_n) L(\vec{x};\theta) \prod_{i=1}^{n} dx_i = \tau(\theta) + b(\theta),\ \theta \in \Theta, \qquad \textbf{(2.4.13)}$$

и повторив все рассуждения, которые были использованы при доказательстве неравенства (2.4.2), получим его обобщение для смещенной оценки $\tilde{T}(X_1,...,X_n)$ функции $\tau(\theta)$ в виде

$$D[\tilde{T}(X_1,...,X_n)] \geq \frac{[\tau'(\theta) + b'(\theta)]^2}{i_n(\theta)}. \qquad \textbf{(2.14.14)}$$

Замечание 2.4.15. Для удобства дальнейших ссылок приведем основные формулы для частного случая, когда в рамках параметрической и регулярной модели $\Im_\theta$ оценивается не функция $\tau(\theta)$, а сам параметр θ. В этом случае для несмещенной оценки $T(X_1,...,X_n)$ параметра θ неравенство Рао – Крамера записывается в виде

$$D[T(X_1,...,X_n)] \geq 1/i_n(\theta) \qquad \textbf{(2.4.16)}$$

и, следовательно, нижняя граница дисперсии несмещенной оценки $T(X_1,...,X_n)$ параметра θ равна обратной величине количества информации Фишера о параметре θ, содержащейся в выборке $X_1,...,X_n$ объема n. Критерий существования эффективной оценки (2.4.10) в данном случае записыва-

ется в виде

$$\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = A^*(\theta)[T^*(X_1,...,X_n) - \theta], \qquad \textbf{(2.4.17)}$$

а ее дисперсия вычисляется по формуле

$$D[T^*(X_1,...,X_n)] = 1/i_n(\theta) = \left|1/A^*(\theta)\right|. \qquad \textbf{(2.4.18)}$$

Отметим, что при оценивании параметра θ модуль коэффициента пропорциональности в критерии (2.4.17) совпадает с количеством информации Фишера о параметре θ, содержащемся в выборке, то есть $|A^*(\theta)| = i_n(\theta)$. Далее, если $\tilde{T}(X_1,...,X_n)$ – смещенная оценка параметра θ и её смещение равняется $b(\theta)$, то неравенство Рао – Крамера (2.4.14) запишется в виде

$$D[\tilde{T}(X_1,...,X_n)] \geq \frac{[1+b'(\theta)]^2}{i_n(\theta)}. \qquad \textbf{(2.4.19)}$$

Используя выражение (2.4.18), можно привести, следуя [38, с.112], такую качественную интерпретацию количества информации Фишера. Если дисперсия оценки служит мерой «трудности» оценивания параметра θ, то обратная величина, равная количеству информации Фишера, измеряет «легкость» с которой может быть оценен параметр θ.

Пример 2.4.20. Рассмотрим бернуллиевскую случайную величину X, которая имеет биномиальное распределение $Bi(n=1,p)$ с параметрами $n=1$ и $p=\theta$, то есть с.в. X задана распределением

$$\begin{pmatrix} X \\ p \end{pmatrix} : \begin{pmatrix} 0, & 1 \\ 1-\theta, & \theta \end{pmatrix}.$$

Пусть $X_1,...,X_n$ – выборка из распределения $Bi(1,\theta)$. Требуется оценить неизвестный параметр θ. Используя критерий существования эффективной оценки (2.4.17), убедимся, что в данном случае выборочное среднее $T^*(X_1,...,X_n) = n^{-1}\sum_{i=1}^{n} X_i$ является несмещенной и эффективной оценкой параметра θ. В самом деле,

$$M[T^*(X_1,...,X_n)] = n^{-1}\sum_{i=1}^{n} M(X_i) = n^{-1}\sum_{i=1}^{n}[0\cdot(1-\theta)+1\cdot\theta] = \theta.$$

Далее, наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ состоит из последовательности нулей и единиц. Вероятность любой конкретной реализации равна произведению величин θ и $1-\theta$, причем каждой единице в последовательности соответствует сомножитель θ, а каждому нулю соответст-

вует сомножитель $1-\theta$. Таким образом, для конкретной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия запишется в виде

$$L(\vec{x};\theta)=P\{X_1=x_1,...,X_n=x_n\}=\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{x_i}=\theta^{\sum x_i}(1-\theta)^{n-\sum x_i},\ x_i=0,1,$$

и, следовательно,

$$\ln L(\vec{x};\theta)=\sum x_i\ln\theta+(n-\sum x_i)\ln(1-\theta).$$

Отсюда имеем

$$\frac{\partial\ln L(\vec{x};\theta)}{\partial\theta}=\frac{\sum x_i}{\theta}-\frac{n-\sum x_i}{1-\theta}=\frac{\sum x_i-\theta\sum x_i-n\theta+\theta\sum x_i}{\theta(1-\theta)}=\frac{n}{\theta(1-\theta)}\left[\frac{1}{n}\sum_{i=1}^{n}x_i-\theta\right].$$

Сравнивая полученное выражение с (2.4.17), заключаем, что эффективная оценка $T^*(X_1,...,X_n)$ параметра θ существует, причем

$$T^*(X_1,...,X_n)=n^{-1}\sum_{i=1}^{n}X_i,\ A^*(\theta)=i_n(\theta)=n/\theta(1-\theta)$$

и дисперсия, согласно (2.4.18), равна

$$D[T^*(X_1,...,X_n)]=1/i_n(\theta)=\left|1/A^*(\theta)\right|=\theta(1-\theta)/n.$$

Пример 2.4.21. Пусть с.в. X имеет нормальное распределение $N(0,\theta^2)$ с параметрами $M(X)=0$ и $D(X)=\theta^2$, то есть ее плотность

$$f_X(x;\theta)=\frac{1}{\sqrt{2\pi}\theta}\exp\left(-\frac{x^2}{2\theta^2}\right),\ x\in R^1.$$

Пусть имеется выборка $X_1,...,X_n$ из распределения $N(0;\theta^2)$. Требуется построить оценку параметрической функции $\tau(\theta)=\theta^2$.

Используя критерий существования эффективной оценки (2.4.3), убедимся, что в данном случае для параметрической функции $\tau(\theta)=\theta^2$ несмещенная и эффективная оценка имеет вид

$$T^*(X_1,...,X_n)=n^{-1}\sum_{i=1}^{n}X_i^2.$$

В самом деле, несмещенность проверяется непосредственно путем вычисления математического ожидания:

$$M[T^*(X_1,...,X_n)]=n^{-1}\sum_{i=1}^{n}M(X^2{}_i)=M(X^2)=\theta^2=\tau(\theta).$$

Далее, для конкретной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ функция

правдоподобия запишется в виде

$$L(\vec{x};\theta)=\prod_{i=1}^{n} f_X(x_i;\theta)=(2\pi)^{-n/2}\theta^{-n}\exp\left(-\frac{1}{2\theta^2}\sum_{i=1}^{n}x_i^2\right)$$

и, следовательно,

$$\ln L(\vec{x};\theta)=-\frac{n}{2}\ln(2\pi)-n\ln\theta-\frac{1}{2\theta^2}\sum_{i=1}^{n}x_i^2\,.$$

Отсюда имеем

$$\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta}=-\frac{n}{\theta}+\frac{1}{\theta^3}\sum_{i=1}^{n}x_i^2=\frac{n}{\theta^3}\frac{1}{n}\sum_{i=1}^{n}x_i^2-\frac{n\theta^2}{\theta^3}=\frac{n}{\theta^3}\left[\frac{1}{n}\sum_{i=1}^{n}x_i^2-\theta^2\right].$$

Сравнивая полученное выражение с (2.4.3), заключаем, что эффективная оценка параметрической функции $\tau(\theta)=\theta^2$ равна $T^*(X_1,...,X_n)=n^{-1}\sum_{i=1}^{n}X_i^2$, $A^*(\theta)=n/\theta^3$ и ее дисперсия, согласно (2.4.11), равна $D[T^*(X_1,...,X_n)]=$ $=\left|\tau'(\theta)/A^*(\theta)\right|=2\theta^4/n$. Отметим, что для самого параметра θ (то есть для стандартного отклонения нормальной с.в. X) эффективная оценка не существует. Отметим также, что при известном математическом ожидании $M(X)=a$ выборочная дисперсия $S_a^2=n^{-1}\sum(X_i-a)^2$ является несмещенной и эффективной оценкой дисперсии.

Пример 2.4.22. Пусть X – дискретная случайная величина с распределением Пуассона (см. П.1.41), то есть $L(X)=\Pi y(\theta)$. Эта запись означает, что с.в. X пробегает значения $x=0,1,2,...$ с вероятностями

$$f_X(x,\theta)=P_\theta(X=x)=\frac{\theta^x e^{-\theta}}{x!},\ x=0,1,2,...,\ \theta>0.$$

Пусть $X_1,...,X_n$ – выборка из распределения $\Pi y(\theta)$. Используя критерий существования эффективной оценки (2.4.17), убедимся, что в данном случае для параметра θ эффективная оценка равна $T^*(X_1,...,X_n)=n^{-1}\sum_{i=1}^{n}X_i$. В самом деле, для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия запишется в виде

$$L(\vec{x};\theta)=\prod_{i=1}^{n} f_X(x_i;\theta)=\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{(x_i)!}=e^{-n\theta}\theta^{\sum x_i}\prod_{i=1}^{n}\frac{1}{(x_i)!}.$$

Отсюда имеем

$$\ln L(\vec{x};\theta)=-n\theta+\ln\theta\sum x_i-\sum\ln(x_i)!$$

и $$\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} = -n + \theta^{-1}\sum x_i = \frac{n}{\theta}\left[\frac{1}{n}\sum_{i=1}^{n} x_i - \theta\right].$$

Сравнивая полученное выражение с (2.4.17), заключаем, что эффективная оценка $T^*(X_1,...,X_n)$ параметра θ существует, причем $T^*(X_1,...,X_n) = n^{-1}\sum_{i=1}^{n} X_i$, $A^*(\theta) = n/\theta$ и её дисперсия, согласно (2.4.18), равна

$$D[T^*(X_1,...,X_n)] = 1/i_n(\theta) = \left|1/A^*(\theta)\right| = \theta/n. \qquad \textbf{(2.4.23)}$$

Отметим, что количество информации Фишера в выборке $X_1,...,X_n$ относительно параметра θ распределения Пуассона равно $i_n(\theta) = n/\theta$. Отметим также на будущее, что функция вклада $U(\vec{X},\theta)$, определенная в (2.3.4), в данном случае

$$U(\vec{X},\theta) = \frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = \frac{n}{\theta}(\overline{X} - \theta) \qquad \textbf{(2.4.24)}$$

и ее дисперсия равняется количеству информации Фишера, то есть

$$D[U(\vec{X},\theta)] = i_n(\theta) = n/\theta. \qquad \textbf{(2.4.25)}$$

Следовательно,

$$\frac{U(\vec{X},\theta)}{\sqrt{D[U(\vec{X},\theta)]}} = \frac{\dfrac{\partial}{\partial\theta}\ln L(\vec{X},\theta)}{\sqrt{D\left(\dfrac{\partial}{\partial\theta}\ln L(\vec{X},\theta)\right)}} =$$

$$= \frac{1}{\sqrt{i_n(\theta)}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i,\theta) = \sqrt{\frac{n}{\theta}}(\overline{X} - \theta). \qquad \textbf{(2.4.26)}$$

Формула (2.4.26) будет использована в разд. 2.13 для построения асимптотического доверительного интервала для параметра θ распределения Пуассона. Заметим, что найденная эффективная оценка

$$T^*(X_1,...,X_n) = n^{-1}\sum_{i=1}^{n} X_i$$

совпадает с оценкой максимального правдоподобия $\hat{\theta}$ (см. разд. 2.10) параметра θ распределения Пуассона, то есть $T^*(X_1,...,X_n) = \hat{\theta}$. Приведем числовой пример. Пусть с.в. X характеризует число поврежденных стеклянных изделий в одном контейнере. Предполагая, что $L(X) = \Pi y(\theta)$, постройте эффективную оценку параметра θ (или оценку максимального правдоподобия па-

раметра θ) по наблюденным данным $n = 500$ контейнеров, которые представлены следующим вариационным рядом частот:

x_i :	0	1	2	3	4	5	6	7
n_i :	199	169	87	31	9	3	1	1

В первой строке указано количество x_i поврежденных изделий в одном контейнере, во второй строке приведена частота n_i – число контейнеров, содержащих x_i поврежденных изделий. Эффективная оценка, совпадающая с оценкой максимального правдоподобия $\hat{\theta}$ параметра θ распределения Пуассона, равна $\hat{\theta} = n^{-1}\sum n_i x_i = (0\cdot 190 + 1\cdot 169 + ... + 7\cdot 1)/500 = 1$.

Замечание 2.4.27. Интегрируя (2.4.10) и обозначая интеграл от функции $A^*(\theta)$ через $A(\theta)$, получим

$$\ln L(\vec{x};\theta) = t(\vec{x})A(\theta) + P(\theta) + R(x_1,...,x_n). \qquad \textbf{(2.4.28)}$$

Отсюда следует, что плотность распределения $f_X(x;\theta)$ должна иметь вид

$$f_X(x;\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\}, \qquad \textbf{(2.4.29)}$$

где $t(\vec{x}) = \sum_{i=1}^{n} B(x_i)$, $R(x_1,...,x_n) = \sum_{i=1}^{n} D(x_i)$, $P(\theta) = nC(\theta)$.

Семейство распределений вида (2.4.29) в литературе часто называют экспоненциальным (см. также определение 2.5.33). Многие распределения, применяемые на практике (см., например, [24]), принадлежат экспоненциальному семейству. Это семейство играет важную роль при обсуждении эффективных (оптимальных) оценок, достаточных статистик и равномерно наиболее мощных критериев. Отметим, что вклад выборки $U(\vec{X};\theta)$ для экспоненциальной модели с плотностью (2.4.29) равен

$$U(\vec{X};\theta) = A'(\theta)\sum_{i=1}^{n} B(X_i) + nC'(\theta) = nA'(\theta)\left[\frac{1}{n}\sum_{i=1}^{n} B(X_i) + \frac{C'(\theta)}{A'(\theta)}\right].$$

Сравнивая данное выражение с критерием существования эффективной оценки вида (2.4.10), заключаем, что для регулярной экспоненциальной модели с плотностью (2.4.29) эффективная оценка $T^*(\vec{X})$ параметрической функции $\tau(\theta) = -C'(\theta)/A'(\theta)$ существует, она записывается в виде $T^*(\vec{X}) = n^{-1}\sum B(X_i)$, причем $A^*(\theta) = nA'(\theta)$ и её дисперсия равна

$$D[T^*(\vec{X})] = \{-C'(\theta)/A'(\theta)\}'/nA'(\theta).$$

Отметим также, что для регулярной экспоненциальной модели с плотностью (2.4.29) количество информации Фишера $i_1(\theta)$ в одном наблюдении относительно параметра θ можно вычислить по формуле

$$i_1(\theta) = \tau'(\theta)A'(\theta) = \{C'(\theta)A''(\theta)/A'(\theta)\} - C''(\theta).$$

В самом деле, согласно (2.4.9) и (2.4.11), выполняется равенство $\{[\tau'(\theta)]^2 / ni_1(\theta)\} = \{\tau'(\theta)/A^*(\theta)\}$, из которого с учетом того, что $A^*(\theta) = nA'(\theta)$, следует $i_1(\theta) = \tau'(\theta)A'(\theta)$.

Замечание 2.4.30. Для параметрически заданной *регулярной статистической модели* $(\mathrm{X}, \Im_\theta)$, в рамках которой по выборке $X_1,...,X_n$ из распределения $F(x;\theta) \in \Im_\theta$, $\theta \in \Theta$, требуется оценить некоторую заданную дифференцируемую функцию $\tau(\theta)$, неравенство Рао – Крамера (2.4.2) определяет потенциальную точность оценивания, то есть дает в классе несмещенных оценок функции $\tau(\theta)$, $\theta \in \Theta$, возможную нижнюю границу дисперсии оценок. Достижение этой границы при использовании некоторой оценки $T^*(\vec{X})$ приводит к понятию эффективной оценки (или МГД-оценки, или оптимальной оценки в соответствии с критерием Рао – Крамера). Эффективная оценка $T^*(\vec{X})$ параметрической функции $\tau(\theta)$ существует не всегда. Критерием существования эффективной оценки $T^*(\vec{X})$ для $\tau(\theta)$ служит выражение (2.4.10), согласно которому разность $\{T^*(\vec{x}) - \tau(\theta)\}$ является линейной функцией от логарифмической производной первого порядка функции правдоподобия, то есть от выражения $\partial \ln L(\vec{x};\theta)/\partial\theta = \{\partial L(\vec{x};\theta)/\partial\theta\}/L(\vec{x};\theta)$. В тех случаях, когда не существует такой эффективной оценки, то есть нельзя представить логарифмическую производную функции правдоподобия в виде (2.4.10) ни для какой статистики $T(\vec{X})$, можно указать другую нижнюю границу для дисперсии несмещенных оценок функции $\tau(\theta)$, бóльшую, чем дает неравенство Рао – Крамера (2.4.2). Эта большая граница определяется неравенством Бхаттачария (П.3.20). Достижение нижней границы в этом неравенстве приводит к понятию оптимальной оценки $T^{**}(\vec{X})$ для $\tau(\theta)$ в соответствие с критерием Бхаттачария (П.3.21). Условие достижимости этой большей границы (или критерий оптимальности Бхаттачария) также записывается в виде выражения, согласно которому разность $\{T^{**}(\vec{x}) - \tau(\theta)\}$ записывается в виде линейной функции, но уже от производных высшего порядка функции правдоподобия, то есть от выражений $\{\partial L(\vec{x};\theta)/\partial\theta\}/L(\vec{x};\theta)$, $\{\partial^2 L(\vec{x};\theta)/\partial\theta^2\}/L(\vec{x};\theta)$, …, $\{\partial^r L(\vec{x};\theta)/\partial\theta^r\}/L(\vec{x};\theta)$, при некотором $r \geq 2$. Выражение для неравенства Бхаттачария и точная формулировка критерия существования оптимальной оценки $T^{**}(\vec{X})$ для $\tau(\theta)$ приведены в (П.3.18), см. также работы [24, с. 28; 22, с. 50]. Отметим, что по-

ведение нижней допустимой границы для среднеквадратических уклонений оценок (или их дисперсий) *в нерегулярном случае*, то есть в случае, когда плотность $f(x,\theta)$ не дифференцируемая по θ или когда информация Фишера полагается равной бесконечности, изучается в [7, с. 203].

Замечание 2.4.31. Обсуждаемые выше критерии оптимальности Рао – Крамера (2.4.10) и Бхаттачария (П.3.22) относятся к случаю скалярного параметра θ. Приведем их краткое обобщение на случай векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r)$ размерности r. Как и прежде, для заданной параметрической и регулярной статистической модели $(X, \Im_\theta)$ требуется по выборке $X_1,...,X_n$ из распределения $F(x;\vec{\theta}) \in \Im_\theta$, $\vec{\theta} \in \Theta$, построить оценку заданной числовой функции $\tau(\vec{\theta})$, дифференцируемой по всем переменным $\theta_1,...,\theta_r$. Обобщение на векторный случай выглядит следующим образом (см., например, [24, с. 32; 22, с.51]).

Если $T(\vec{X})$ – несмещенная оценка числовой функции $\tau(\vec{\theta})$, то при всех $\vec{\theta} \in \Theta$ выполняется неравенство

$$D_\theta[T(\vec{X})] \geq \sum_{i,j=1}^{r} g_{ij}(\theta) c_i(\theta) c_j(\theta), \qquad \textbf{(2.4.32)}$$

где $$g_{ij}(\theta) = M_\theta\left(\frac{\partial \ln L(\vec{X},\vec{\theta})}{\partial \theta_i} \cdot \frac{\partial \ln L(\vec{X},\vec{\theta})}{\partial \theta_j} \right),\ i,j = 1,...,r, \qquad \textbf{(2.4.33)}$$

а коэффициенты $c_i(\theta)$ определяются системой уравнений

$$\sum_{j=1}^{r} g_{ij}(\theta) c_j(\theta) = \partial\tau(\vec{\theta}) / \partial\theta_i,\ i = 1,...,r. \qquad \textbf{(2.4.34)}$$

Если матрица $I_n(\vec{\theta}) = \| g_{ij}(\theta) \|_1^r$ не вырождена и $I_n^{-1}(\vec{\theta}) = \| g^{ij}(\theta) \|_1^r$ – обратная к ней матрица, то неравенство (2.4.32) эквивалентно неравенству

$$D_\theta[T(\vec{X})] \geq b(\vec{\theta}) I_n^{-1}(\theta) b^T(\vec{\theta}) = \sum_{i,j=1}^{r} g^{ij}(\theta) \{\partial\tau(\vec{\theta}) / \partial\theta_i\} \{\partial\tau(\vec{\theta}) / \partial\theta_j\}, \qquad \textbf{(2.4.35)}$$

где $b(\vec{\theta}) = (\partial\tau(\vec{\theta}) / \partial\theta_1, ..., \partial\tau(\vec{\theta}) / \partial\theta_r)$. Знак равенства в неравенствах (2.4.32) и (2.4.35) имеет место тогда и только тогда, когда

$$T(\vec{X}) - \tau(\vec{\theta}) = \sum_{i=1}^{r} c_i(\theta) \frac{\partial \ln L(\vec{X},\vec{\theta})}{\partial \theta_i}. \qquad \textbf{(2.4.36)}$$

Отметим, что если функция правдоподобия $L(\vec{x},\vec{\theta})$ дважды дифференцируема по $\vec{\theta}$, то дифференцирование тождества $\int L(\vec{x},\vec{\theta}) d\vec{x} \equiv 1$ сначала по θ_i, а

потом по θ_j, приводит к равенству

$$g_{ij}(\theta) = -M_\theta\left(\frac{\partial^2 \ln L(\vec{X};\vec{\theta})}{\partial\theta_i\partial\theta_j}\right) = -nM_\theta\left(\frac{\partial^2 \ln f(X_1;\vec{\theta})}{\partial\theta_i\partial\theta_j}\right),\ i,j = 1,...,r\,. \qquad \textbf{(2.4.37)}$$

Отметим также, что для случая векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r)$ матрицу $I_n(\vec{\theta}) = \| g_{ij}(\theta) \|_1^r$ называют информационной матрицей выборки, и она является аналогом количества информации $i_n(\theta)$, определенного в (2.3.13), а матрицу $I_1(\vec{\theta})$ называют информационной матрицей одного наблюдения, и она является аналогом количества информации $i_1(\theta)$, определенного в (2.3.11). Причем по аналогии со случаем скалярного параметра, для схемы повторных независимых наблюдений, согласно (2.4.37), в случае векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r)$ выполняется равенство $I_n(\vec{\theta}) = nI_1(\vec{\theta})$, которое является обобщением равенства (2.3.14).

2.5. Достаточные статистики

Рассмотренные выше понятия, такие, как несмещенность, состоятельность и эффективность, относятся к общим и часто используемым характеристикам, с помощью которых описывают свойства оценок параметров. При более глубоком изучении свойств оценок рассматривают понятие достаточности, введенное Фишером в 1922 году. Использование достаточных статистик является одним из эффективных способов построения оптимальных оценок. Достаточные статистики позволяют осуществить редукцию исходных статистических данных, то есть позволяют представить их в наиболее сжатом и компактном виде, который не приводит к потере информации. В рамках заданной модели факт существования достаточной статистики и ее конкретный вид устанавливается с помощью критерия факторизации, который приводится ниже в виде теоремы факторизации Неймана – Фишера. Мы сначала приведем определение достаточной статистики и его обсуждение с иллюстративными примерами для случая, когда модель задана в параметрической форме, а затем кратко обсудим другие варианты статистических моделей.

Итак, пусть $\vec{X} = (X_1,...,X_n)$ – выборка из распределения F_X, которое является элементом параметрического семейства

$\Im_\theta = \{F : F_X(x;\theta), \theta \in \Theta\}$, и пусть $\vec{x} = (x_1,...,x_n)$ – конкретная реализация выборки $X_1,...,X_n$.

Определение 2.5.1. Статистику $T(X_1,...,X_n)$ называют достаточной для модели $\Im_\theta$, или для параметра θ, если условная плотность (или вероятность) $L(\vec{x}|t;\theta) = f_{\vec{X}|T(\vec{X})=t}(\vec{x}|t;\theta)$ случайного вектора $\vec{X} = (X_1,...,X_n)$ при условии, что $T(\vec{X}) = t(\vec{x})$ не зависит от θ.

Отметим, что свойство достаточности на качественном уровне означает, что достаточная статистика $T(X_1,...,X_n)$ содержит в себе всю информацию о параметре θ, имеющуюся в выборке $X_1,...,X_n$, и поэтому все процедуры, связанные с оцениванием параметра θ или с проверкой гипотез относительно этого параметра, будут в итоге основаны на этой достаточной статистике. Достаточные статистики позволяют осуществить редукцию исходных статистических данных без потери информации, что особенно важно при обработке больших массивов статистической информации, хранение и передача которой связаны с большими материальными затратами.

Замечание 2.5.2. Понятие достаточной статистики, сформулированное для параметрической модели, естественным образом обобщается на случай произвольного семейства допустимых функций распределения вероятностей $\Im$. Итак, пусть $X_1,...,X_n$ – выборка из распределения $F_X \in \Im$ и $\vec{x}=(x_1,...,x_n)$ – ее конкретная реализация. Статистику $T(X_1,...,X_n)$ называют достаточной для $X_1,...,X_n$, или для семейства $\Im$, если условная совместная плотность (вероятность) случайных величин $X_1,...,X_n$ при условии, что $T(X_1,...,X_n)=t(x_1,...,x_n)$, обозначаемая через $f_{\vec{X}|T(\vec{X})=t}(x_1,...,x_n|t)=L(x_1,...,x_n|t)$, является одной и той же для всех распределений из $\Im$ или, другими словами, если условная плотность (вероятность) не меняется при смене распределения из множества $\Im$.

Пример 2.5.3. Пусть X – дискретная случайная величина, причем $L(X) = \Pi y(\theta)$. Эта символическая запись означает, что с.в. X имеет распределение Пуассона с неизвестным параметром $\theta = M(X)$, ее возможные состояния пробегают значения $x = 0,1,2,...$, а их вероятности вычисляются по формуле

$$P_\theta\{X = x\} = \frac{\theta^x e^{-\theta}}{x!},\ x = 0,1,2,... \qquad \textbf{(2.5.4)}$$

Параметрическая модель в данном случае (см. пример 1.1.3) определяется се-

мейством ф.р. $F_X(y;\theta)$ вида

$$F_X(y;\theta)=\sum_x C(y-x)\frac{\theta^x e^{-\theta}}{x!},\ y\in R^1,\ x=0,1,2,...,\ \theta\in(0,\infty).$$

Пусть, далее, $X_1,...,X_n$ – выборка, характеризуемая распределением Пуассона, то есть $L(X)=\Pi y(\theta)$ и $\vec{x}=(x_1,...,x_n)$ – ее конкретная реализация. Убедимся, что $T(X_1,...,X_n)=X_1+....+X_n$ является достаточной статистикой для данной модели или для параметра θ этой модели. Для этого надо показать, что условная вероятность

$$P_\theta\{X_1=x_1,...,X_n=x_n\,|\,T(\vec{X})=t(\vec{x})\}=\frac{P_\theta\{X_1=x_1,...,X_n=x_n;T(\vec{X})=t(\vec{x})\}}{P_\theta\{T(\vec{X})=t(\vec{x})\}} \quad \textbf{(2.5.5)}$$

не зависит от параметра θ. Для $T(\vec{X})=t(\vec{x})$ событие $\{\vec{X}=\vec{x}\}\subseteq\{T(\vec{X})=t(\vec{x})\}$, поэтому выполняется равенство

$$\{X_1=x_1,...,X_n=x_n;T(\vec{X})=t(\vec{x})\}=\{X_1=x_1,...,X_n=x_n\}, \quad \textbf{(2.5.6)}$$

и, следовательно, (2.5.5) запишется в виде

$$P_\theta\{X_1=x_1,...,X_n=x_n\,|\,T(\vec{X})=t(\vec{x})\}=\frac{P_\theta\{X_1=x_1,...,X_n=x_n\}}{P_\theta\{T(\vec{X})=t(\vec{x})\}}. \quad \textbf{(2.5.7)}$$

Далее, поскольку $X_1,...,X_n$ – н.о.р. случайные величины, то их совместное распределение имеет вид

$$P_\theta\{X_1=x_1,...,X_n=x_n\}=L(\vec{x};\theta)=\prod_{i=1}^n P_\theta(X_i=x_i)=\prod_{i=1}^n\frac{e^{-\theta}\theta^{x_i}}{(x_i)!}=\frac{e^{-n\theta}\theta^{\sum x_i}}{\prod(x_i)!}. \quad \textbf{(2.5.8)}$$

Учитывая теперь, что распределение Пуассона является воспроизводящим по параметру θ (см. П. 1.41), то есть для суммы $X_1+....+X_n$ выполняется выражение $L(\sum_{i=1}^n X_i)=\Pi y(n\theta)$, получаем

$$P_\theta\{T(\vec{X})=t(\vec{x})\}=\frac{e^{-n\theta}(n\theta)^t}{t!}\ \text{при}\ t=\sum_{i=1}^n x_i\ . \quad \textbf{(2.5.9)}$$

Используя (2.5.8) и (2.5.9), выражение (2.5.7) запишется в виде

$$P_\theta\{X_1=x_1,...,X_n=x_n\,|\,T(\vec{X})=t(\vec{x})\}=\frac{e^{-n\theta}\theta^{\sum x_i}}{\prod(x_i)!}\,\frac{(\sum x_i)!}{e^{-n\theta}(n\theta)^{\sum x_i}}=\frac{(\sum x_i)!}{\prod(x_i)!}\,\frac{1}{n^{\sum x_i}}$$

$$\text{при}\ t=\sum_{i=1}^n x_i\ , \quad \textbf{(2.5.10)}$$

и эта вероятность равна нулю при выполнении неравенства $t \neq \sum_{i=1}^{n} x_i$. Отметим, что полученная условная вероятность не зависит от параметра θ, она определяет полиномиальное распределение с числом испытаний $t = \sum_{i=1}^{n} x_i$ и равными $(1/n)$ вероятностями исходов. Таким образом, статистика $T(X_1,...,X_n) = X_1 + ... + X_n$ является достаточной для параметра θ пуассоновской модели.

Пример 2.5.11. Пусть X – непрерывная случайная величина, причем $L(X) = R(0,\theta)$. Эта запись означает, что с.в. X имеет равномерное распределение в интервале $[0,\theta]$ и, следовательно, ее плотность $f_X(x,\theta)$ и ф.р. $F_X(x,\theta)$ записываются в виде

$$f_X(x,\theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta, \\ 0, \text{ в других случ.}; \end{cases} \quad F_X(x,\theta) = \begin{cases} 0, & x < 0, \\ x/\theta, & 0 \le x \le \theta, \\ 1, & x > 0. \end{cases}$$

Пусть $X_1,...,X_n$ – выборка из равномерного распределения, то есть $L(X) = R(0,\theta)$ и $\vec{x} = (x_1,...,x_n)$ – ее наблюденная реализация. Далее, пусть $\vec{X}_{(.)} = (X_{(1)},...,X_{(n)})$ – упорядоченная статистика (см. разд. 4.1.А) и $\vec{x}_{(.)} = (x_{(1)},...,x_{(n)})$ – ее конкретная реализация, которую называют вариационным рядом. Величина $x_{(n)}$ является n-м элементом вариационного ряда $x_{(1)} < ... < x_{(n)}$, причем $x_{(n)} = \max(x_1,...,x_n)$ является реализацией случайной величины $X_{(n)}$, которую называют n-й порядковой статистикой. Убедимся, что для данной параметрической модели n-я порядковая статистика $X_{(n)}$ является достаточной статистикой для параметра θ. В самом деле, совместная плотность случайных величин $X_1,...,X_n$ равна $f_{X_1,...,X_n}(x_1,...,x_n;\theta) = 1/\theta^n$ при $x_{(n)} < \theta$ и равна нулю в других случаях. Используя определение (2.3.1) функции правдоподобия $L(\vec{x};\theta)$ и единичную функцию вида (1.4.97), получим

$$L(\vec{x};\theta) = f_{X_1,...,X_n}(x_1,...,x_n;\theta) = \theta^{-n} C(\theta - x_{(n)}) C(x_i),\ i = 1,...,n. \qquad \textbf{(2.5.12)}$$

Далее, (см. раздел 4.2.Б) плотность распределения n-й порядковой статистики имеет вид

$$f_{X_{(n)}}(x_{(n)}) = nF_X^{n-1}(x_{(n)},\theta) f_X(x_{(n)},\theta) = \theta^{-n} n x_{(n)}^{n-1} \text{ при } 0 < x_{(n)} < \theta.$$

Отсюда следует, что условная плотность случайных величин $X_1,...,X_n$ при

условии, что $X_{(n)} = x_{(n)}$, равна

$$f_{\vec{X}\left|X_{(n)}=x_{(n)}\right.}(x_1,...,x_n \left| x_{(n)};\theta\right) = 1/nx_{(n)}^{n-1} \text{ при } 0 < x_{(n)} < \theta \qquad \textbf{(2.5.13)}$$

и равна нулю в других случаях. Отметим, что правая часть в (2.5.13) формально зависит от θ, но неравенство $0 < x_{(n)} < \theta$ автоматически выполняется для всех значений θ, поэтому максимальная порядковая статистика $X_{(n)}$ есть достаточная статистика для параметра θ, который является правой границей интервала равномерного распределения.

Пример 2.5.14. Пусть $X_1,...,X_n$ – выборка из распределения F_X и $\vec{x} = (x_1,...,x_n)$ – ее наблюденная реализация. Далее, пусть $\vec{X}_{(.)} = (X_{(1)},...,X_{(n)})$ – упорядоченная статистика, реализацией которой является вариационный ряд $\vec{x}_{(.)} = (x_{(1)},...,x_{(n)})$ с элементами, удовлетворяющими неравенствам $x_{(1)} < ... < x_{(n)}$. Предполагаем, что ф.р. F_X принадлежит семейству $\Im$ абсолютно непрерывных распределений, имеющих плотность $f_X(x)$, $x \in R^1$. Убедимся, что упорядоченная статистика $\vec{X}_{(.)} = (X_{(1)},...,X_{(n)})$ является достаточной статистикой для данного семейства $\Im$. В самом деле, поскольку $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. F_X и плотностью $f_X(x)$, то их совместная плотность записывается в виде

$$f_{X_1,...,X_n}(x_1,...,x_n) = \prod_{i=1}^{n} f_X(x_i).$$

Далее, согласно (4.2.1), совместная плотность упорядоченной статистики $\vec{X}_{(.)} = (X_{(1)},...,X_{(n)})$ равна

$$f_{\vec{X}_{(.)}}(x_{(1)},...,x_{(n)}) = n! f_{\vec{X}}(x_{(1)},...,x_{(n)}) = n! \prod_{i=1}^{n} f_X(x_{(i)}),\ x_{(1)} < ... < x_{(n)}.$$

Отсюда заключаем, что условная вероятность случайных величин $X_1,...,X_n$, при условии, что $\vec{X}_{(.)} = \vec{x}_{(.)}$, равна $1/n!$, поскольку при заданных упорядоченных значениях наблюдений $x_{(1)} < ... < x_{(n)}$ возможные перестановки векторов $(X_{(i_1)},...,X_{(i_n)})$, общее число которых равно $n!$, являются равновероятными. Таким образом, упорядоченная статистика $\vec{X}_{(.)} = (X_{(1)},...,X_{(n)})$ является достаточной статистикой для $X_1,...,X_n$, или для семейства $\Im$.

Замечание 2.5.15. Приведенные выше примеры показали, что при заданном семействе распределений нахождение достаточных статистик может быть основано непосредственно на использовании определения. Однако на практике такой путь не является удобным, поскольку он требует сначала «догадаться»,

какой вид имеет достаточная статистика, а затем проанализировать поведение условной плотности (или вероятности) и убедиться, что она не зависит от θ. Более удобным для практических целей является критерий факторизации, который для заданной параметрической модели по виду функции правдоподобия позволяет судить о том, существует или нет достаточная статистика, и одновременно, если она существует, позволяет конкретизировать ее вид.

2.5.А. Критерий факторизации

Пусть модель задана в параметрической форме, то есть в виде $\Im_\theta = \{F : F_X(x;\theta), \theta \in \Theta\}$. Справедлива следующая факторизационная теорема Неймана – Фишера.

Теорема 2.5.16. Для того чтобы статистика $T(X_1,...,X_n)$ была достаточной для параметра θ в модели $\Im_\theta$, необходимо и достаточно, чтобы для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ функция правдоподобия $L(\vec{x};\theta)$ допускала представление в виде произведения двух неотрицательных функций φ и h, то есть

$$L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x}), \qquad \textbf{(2.5.17)}$$

причем функция φ может зависеть от θ, а от $\vec{x}$ зависит через реализацию $T(\vec{x})$ достаточной статистики $T(\vec{X})$, другая функция $h(\vec{x})$ от параметра θ не зависит (и даже может тождественно равняться единице).

Доказательство. Следуя работам [7, 22], приведем сначала доказательство для дискретного случая, а затем для непрерывного.

Пусть X – дискретная случайная величина с возможными значениями $\tilde{x}_1, \tilde{x}_2, ...$ и их вероятностями $P_\theta(X = \tilde{x}_i)$, $i = 1,2,...$, которые зависят от некоторого параметра θ. Параметрическая модель в данном случае определяется семейством ф.р. $F_X(x;\theta)$ вида

$$F_X(x;\theta) = \sum_i C(x - \tilde{x}_i)\, P_\theta(X = \tilde{x}_i).$$

Пусть, далее, $\vec{X} = (X_1,...,X_n)$ – выборка из распределения F_X и $\vec{x} = (x_1,...,x_n)$ – ее конкретная реализация. Предположим, что выражение (2.5.17) для функции правдоподобия $L(\vec{x};\theta)$ выполняется.

Убедимся, что в этом случае условная вероятность при условии, что $T(\vec{X}) = t(\vec{x})$, не зависит от параметра θ. В самом деле, для конкретной реализации $\vec{x} = (x_1, ..., x_n)$ имеем

$$P_\theta\{X_1 = x_1, ..., X_n = x_n \,|\, T(\vec{X}) = t(\vec{x})\} = \\ = \frac{P_\theta\{X_1 = x_1, ..., X_n = x_n; T(\vec{X}) = t(\vec{x})\}}{P_\theta\{T(\vec{X}) = t(\vec{x})\}}. \qquad \textbf{(2.5.18)}$$

Пусть $T(\vec{X}) = t(\vec{x})$. Отметим, что событие $\{\vec{X} = \vec{x}\} \subseteq \{T(\vec{X}) = t(\vec{x})\}$, поэтому выполняются равенства $\{\vec{X} = \vec{x}\,;\, T(\vec{X}) = t(\vec{x})\} = \{\vec{X} = \vec{x}\}$ и $P_\theta\{\vec{X} = \vec{x}\,;\, T(\vec{X}) = t(\vec{x})\} = P_\theta\{\vec{X} = \vec{x}\}$. С учетом последнего равенства формула (2.5.18) перепишется в виде

$$P_\theta\{X_1 = x_1, ..., X_n = x_n \,|\, T(\vec{X}) = t(\vec{x})\} = \frac{P_\theta\{X_1 = x_1, ..., X_n = x_n\}}{P_\theta\{T(\vec{X}) = t(\vec{x})\}}. \qquad \textbf{(2.5.19)}$$

Далее, числитель в (2.5.19) является функцией правдоподобия $L(\vec{x};\theta)$, а знаменатель может быть записан в виде

$$P_\theta\{T(\vec{x}) = t(\vec{x})\} = \sum_{y::t(\vec{y})=t(\vec{x})} L(\vec{y};\theta),$$

где суммирование осуществляется по всем таким $\vec{y}$, для которых $t(\vec{y}) = t(\vec{x})$. Для удобства дальнейшей записи правую часть предыдущего выражения обозначим в виде

$$\sum{}^{*} L(\vec{y},\theta) = \sum_{y::t(\vec{y})=t(\vec{x})} L(\vec{y};\theta).$$

С учетом этих замечаний формула (2.5.19) перепишется в виде

$$P_\theta\{X_1 = x_1, ..., X_n = x_n \,|\, T(\vec{X}) = t(\vec{x})\} = L(\vec{x};\theta) / \sum{}^{*} L(\vec{y};\theta).$$

Используя теперь (2.5.17), окончательно, после сокращения функции $\varphi(t,\theta)$, получаем

$$P_\theta\{X_1 = x_1, ..., X_n = x_n \,|\, T(\vec{X}) = t(\vec{x})\} = \\ = \frac{\varphi(t,\theta)h(\vec{x})}{\sum^{*} \varphi(t,\theta)h(\vec{y})} = h(\vec{x}) / \sum{}^{*} h(\vec{y}). \qquad \textbf{(2.5.20)}$$

Таким образом, условная вероятность

$$P_\theta\{X_1 = x_1,...,X_n = x_n \,|\, T(\vec{X}) = t(\vec{x})\},$$

определенная формулой (2.5.20), не зависит от параметра θ и, следовательно, $T(\vec{X})$ является достаточной статистикой для θ.

Обратное рассуждение. Пусть левая часть в (2.5.19) не зависит от θ и, следовательно, ее можно обозначить через $h(\vec{x})$. Убедимся, что при этом предположении выполняется (2.5.17). В самом деле, обозначив $P_\theta\{T(\vec{X}) = t(\vec{x})\}$ через $\varphi(T(\vec{x}),\theta)$, из (2.5.19) получаем

$$L(\vec{x};\theta) = P_\theta\{X_1 = x_1,...,X_n = x_n\} = \varphi(T(\vec{x}),\theta)h(\vec{x}).$$

Для дискретного случая доказательство завершено.

Рассмотрим теперь непрерывный случай. Пусть X – непрерывная случайная величина с ф.р. $F_X(x,\theta)$, плотностью $f_X(x,\theta)$, $x \in R^1$. Пусть, далее, $\vec{X} = (X_1,...,X_n)$ – выборка из распределения $F_X(x,\theta) \in \Im_\theta$, где $\Im_\theta$ – заданное параметрическое семейство. Как и в дискретном случае, убедимся сначала в том, что при выполнении (2.5.17) статистика $T(\vec{X})$ является достаточной статистикой для θ. Для этого, используя совместную плотность $f_{\vec{X}}(x_1,...,x_n;\theta)$ случайных величин $X_1,...,X_n$, получим сначала выражение для маргинальной плотности статистики $T(X_1,...,X_n)$. Предполагая, что статистика $T(X_1,...,X_n)$ является «гладкой» функцией, введем следующую замену переменных:

$$y_1 = T(\vec{x}),\ y_2 = y_2(\vec{x}),\ \ldots,\ y_n = y_n(\vec{x}),$$

здесь $\vec{x} = (x_1,...,x_n)$. Обратные преобразования обозначим через $x_i = x_i(y_1,...,y_n)$, $i = 1,...,n$, и пусть якобиан J этих преобразований отличен от нуля, то есть $J = \left|\dfrac{\partial x_i}{\partial y_i}\right| \neq 0$. Используя обобщение формулы (1.1.20) на многомерный случай, запишем выражение для совместной плотности случайных величин $\vec{Y} = (T(\vec{X}), y_2(\vec{X}),..., y_n(\vec{X}))$ в виде

$$f_{\vec{Y}}(y_1,...,y_n;\theta) = f_{\vec{X}}(x_1(y_1,...,y_n),...,x_n(y_1,...,y_n);\theta)|J|. \quad \textbf{(2.5.21)}$$

Отсюда, используя формулу (1.1.67), получим маргинальную плотность случайной величины $Y_1 = T(\vec{X})$ в виде

$$f_{T(\vec{X})}(y_1;\theta) = \int\limits_{R^{n-1}} ...\int . f_{\vec{Y}}(y_1,...,y_n;\theta)\prod_{i=2}^{n} dy_i =$$

$$= \int\limits_{R^{n-1}} ...\int . f_{\vec{X}}(x_1(y_1,...,y_n),...,x_n(y_1,...,y_n);\theta)|J|\prod_{i=2}^{n} dy_i .$$

Далее, из (2.5.21) и последнего выражения следует, что условная плотность $\vec{Y}$ при условии, что $T(\vec{X}) = t(\vec{x})$, имеет вид

$$f_{\vec{Y}|T(\vec{X})=t}(y_1,...,y_n|t;\theta) = \frac{f_{\vec{Y}}(y_1,...,y_n;\theta)}{f_{T(\vec{X})}(y_1;\theta)} =$$

$$= \frac{f_{\vec{X}}(x_1(y_1,...,y_n),...,x_n(y_1,...,y_n);\theta)|J|}{\int\limits_{R^{n-1}} ...\int f_{\vec{X}}(x_1(y_1,...,y_n),...,x_n(y_1,...,y_n);\theta)|J|dy_2...dy_n}$$

$$\text{при } y_1 = t . \qquad \textbf{(2.5.22)}$$

Предположим теперь, что (2.5.17) выполняется. Тогда (2.5.22) запишется в виде

$$f_{\vec{Y}|T(\vec{X})=t}(y_1,...,y_n|t;\theta) = \frac{\varphi(T(\vec{x});\theta)h(\vec{x})|J|}{\int\limits_{R^{n-1}} ...\int \varphi(T(\vec{x});\theta)h(\vec{x})|J|dy_2...dy_n} \text{ при } y_1 = t .$$

В правой части последнего выражения функция $\varphi(t;\theta)$, зависящая от θ, сокращается, и это означает, что условная плотность случайного вектора $\vec{Y}$, а следовательно, и вектора $\vec{X}$, при условии, что $T(\vec{X}) = t$, не зависит от θ, из чего следует, что $T(\vec{X})$ является достаточной статистикой для θ.

Обратное рассуждение. Пусть $f_{\vec{Y}|T(\vec{X})=t}(y_1,...,y_n|t;\theta)$ не зависит от θ и поскольку $y_1,...,y_n$ зависят от $\vec{x}$, то ее можно обозначить через $h(\vec{x})$. Далее, $f_{T(\vec{X})}(y_1;\theta)$ является функцией $y_1 = T(\vec{x})$ и θ, которую обозначим через $\varphi(T(\vec{x});\theta)$. Окончательно из (2.5.22) следует, что $L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x})$. Доказательство завершено.

Замечание 2.5.23. Использование критерия факторизации (2.5.17) и критерия существования эффективных оценок (2.4.17) позволяет установить связи между понятиями эффективная оценка и достаточная статистика. В самом деле, согласно (2.4.17), эффективная оценка $T^*(\vec{X})$ параметра θ существует, если выполняется выражение

$$\frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = A^*(\theta)[T^*(X_1,...,X_n) - \theta].$$

Отсюда следует, что функцию правдоподобия $L(\vec{x};\theta)$ можно записать в виде (2.4.28), или

$$L(\vec{x};\theta) = \exp\{\int_0^\theta A^*(\theta')[T^*(\vec{x}) - \theta']d\theta'\}.$$

Это выражение соответствует специальному варианту критерия факторизации (2.5.17), и, следовательно, $T^*(\vec{X})$ является достаточной статистикой для параметра θ.

Таким образом, можно сделать вывод, что всякая эффективная оценка одновременно является также достаточной статистикой, то есть требование достаточности является менее ограничительным, чем условие существования эффективной оценки. Эффективная оценка существует тогда и только тогда, когда имеется достаточная статистика, а достаточная статистика может существовать и при отсутствии эффективной оценки. Отметим, что для модели Коши (см. П.2.56) функция правдоподобия имеет вид

$$L(\vec{x};\theta) = \pi^{-n}\prod_{i=1}^{n}\{1/[1+(x_i-\theta)]\}.$$

Эту функцию правдоподобия нельзя привести к виду (2.5.17), то есть нельзя указать достаточную статистику для параметра θ, размерность которой меньше чем объем выборки n.

Замечание 2.5.24. Отметим, что любая непрерывная и монотонная функция $\psi(T(\vec{X}))$ от достаточной статистики $T(\vec{X})$ параметра θ также является достаточной статистикой для θ. В этом легко убедиться. Пусть $G(\vec{X}) = \psi(T(\vec{X}))$, тогда $T(\vec{X}) = \psi^{-1}(G(\vec{X}))$ и, следовательно, $L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x}) = \varphi(\psi^{-1}(G(\vec{x})),\theta)h(\vec{x})$, и из критерия факторизации (2.5.17) следует, что $G(\vec{X})$ является достаточной статистикой для параметра θ (см. также пример 2.5.26).

Пример 2.5.25. Возвращаясь к примеру 2.5.3, отметим, что функция правдоподобия для пуассоновской модели, согласно (2.5.8), имеет вид

$$L(\vec{x};\theta) = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod(x_i)!} = \varphi(\sum x_i;\theta)h(\vec{x}),$$

где $\varphi(\sum x_i;\theta)=e^{-n\theta}\theta^{\sum x_i}$ и $h(\vec{x})=1/\prod(x_i)!$. Следовательно, из критерия факторизации (2.5.17) заключаем, что $T(\vec{X})=\sum X_i$ является достаточной статистикой для параметра θ пуассоновской модели.

Аналогично, в примере 2.5.11 функция правдоподобия для распределения $R(0;\theta)$, согласно (2.5.12), имеет вид

$$L(\vec{x};\theta)=f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta)=\theta^{-n}C(\theta-x_{(n)})C(x_i)=\varphi(x_{(n)},\theta)h(\vec{x}),\ i=1,\ldots,n,$$

где $\varphi(x_{(n)},\theta)=\theta^{-n}C(\theta-x_{(n)})$ и $h(\vec{x})=C(x_i)$, $i=1,\ldots,n$. Таким образом, из критерия факторизации (2.5.17) заключаем, что максимальная порядковая статистика $X_{(n)}$ является достаточной статистикой для параметра θ, который является правой границей равномерного распределения в интервале $[0,\theta]$.

Пример 2.5.26. Пусть имеется выборка $X_1,\ldots,X_n$ из нормального распределения $L(X)=N(0,\sigma^2)$. Неизвестным параметром является дисперсия, то есть $\theta=\sigma^2$. Функция правдоподобия $L(\vec{x};\theta)$ имеет вид

$$L(\vec{x};\theta)=\prod_{i=1}^{n}f_X(x_i,\theta)=(2\pi\theta)^{-n/2}\exp\left\{-\frac{1}{2\theta}\sum_{i=1}^{n}x_i^2\right\}. \qquad \textbf{(2.5.27)}$$

Из критерия факторизации (2.5.17) следует, что $T(\vec{X})=\sum X_i^2$ является достаточной статистикой для параметра $\theta=\sigma^2$.

Отметим, что в данном примере, как и в других случаях, можно указать большое число достаточных статистик вида $G(\vec{X})=\psi(\sum X_i^2)$, где ψ – любая непрерывная и монотонная функция (см. замечание 2.5.24). Но все они являются эквивалентными в том смысле, что они содержат одно и то же количество информации об оцениваемом параметре θ. Кроме того, также могут существовать достаточные статистики, которые обеспечивают различную степень редукции исходных данных. В качестве примера для данного параметра $\theta=\sigma^2$ рассмотрим следующие статистики:

$$T_1(\vec{X})=(X_1,\ldots,X_n);\ T_2(\vec{X})=(X_1^2,\ldots,X_n^2);$$
$$T_3(\vec{X})=(X_1^2+\ldots+X_m^2,X_{m+1}^2+\ldots+X_n^2);\ T_4(\vec{X})=(X_1^2+\ldots+X_n^2).$$

Эти статистики являются достаточными для $\theta=\sigma^2$, но все они обеспечивают различную степень редукции исходных данных. Статистика $T_1(\vec{X})$ является тривиальной достаточной статистикой, поскольку она не приводит к редукции исходных данных $X_1,\ldots,X_n$, а статистика $T_4(\vec{X})$ обеспечивает максимальную степень редукции исходных данных. Такую статистику называют минимальной достаточной статистикой. Естественно, что мы заинтересованы

в использовании минимальных достаточных статистик. Детальное обсуждение этого важного понятия может быть найдено в классических монографиях по статистке (см., например, [7, 24, 37, 38, 66]).

Пример 2.5.28. Пусть имеется выборка $X_1,...,X_n$ из нормального распределения $L(X)=N(\mu,\sigma^2)$. Параметры μ и σ^2 неизвестны, то есть в данном случае имеем неизвестный векторный параметр $\vec{\theta}=(\theta_1,\theta_2)$, $\theta_1=\mu$ и $\theta_2=\sigma^2$. Функция правдоподобия $L(\vec{x};\theta)$ записывается в виде

$$L(\vec{x};\vec{\theta})=\prod_{i=1}^{n} f_X(x_i,\vec{\theta})=(2\pi\theta_2)^{-n/2}\exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right\}=$$

$$=(2\pi\theta_2)^{-n/2}\exp\left\{-\frac{1}{2\theta_2}\sum_{i=1}^{n}x_i^2+\frac{\theta_1}{\theta_2}\sum_{i=1}^{n}x_i-\frac{n\theta_1^2}{2\theta_2}\right\}. \qquad \textbf{(2.5.29)}$$

Из критерия факторизации следует, что для векторного параметра $\vec{\theta}=(\mu,\sigma^2)$ достаточная статистика также является вектором и имеет вид $T=(\sum X_i,\sum X_i^2)$. В данном случае эту статистику T обычно заменяют на более удобную и эквивалентную статистку $T^*=(\overline{X},S^2)$, где $\overline{X}=n^{-1}\sum X_i$ – выборочное среднее и $S^2=n^{-1}\sum(X_i-\overline{X})^2=n^{-1}\sum X_i^2-\overline{X}^2$ – выборочная дисперсия.

Пример 2.5.30. Пусть $X_1,...,X_n$ – выборка из распределения $L(X)=R(\theta_1,\theta_2)$. Векторный параметр $\vec{\theta}=(\theta_1,\theta_2)$ не известен. Функция правдоподобия $L(\vec{x};\theta)$ записывается в виде

$$L(\vec{x};\vec{\theta})=\begin{cases}\prod_{i=1}^{n} f_X(x_i,\vec{\theta})=(\theta_2-\theta_1)^{-n},\ \theta_1\le x_i\le\theta_2,\ i=1,...,n,\\ 0,\ \text{в других случаях,}\end{cases}$$

или с использованием единичной функции $L(\vec{x};\theta)$ запишется в виде

$$L(\vec{x};\theta)=C(\theta_2-x_{(n)})C(x_{(1)}-\theta_1)/(\theta_2-\theta_1)^n. \qquad \textbf{(2.5.31)}$$

Из критерия факторизации следует, что для векторного параметра $\vec{\theta}=(\theta_1,\theta_2)$, компоненты которого θ_1 и θ_2 являются соответственно левой и правой границами интервала равномерного распределения, достаточная статистика также является вектором и имеет вид $T=(T_1,T_2)$, где $T_1=X_{(1)}$ и $T_2=X_{(n)}$.

Замечание 2.5.32. В заключение данного раздела отметим, что достаточные статистики играют важную роль в процессе построения оптимальных оценок параметров. Мы привели лишь краткое обсуждение и иллюстратив-

ные примеры этого важного понятия. Дальнейшее обсуждение этих вопросов может быть найдено в литературе (см., например, [7, 24, 37, 38, 66]). При нахождении явного вида оптимальных оценок важную роль играет свойство полноты достаточной статистики, а также теорема Рао – Блекуэлла – Колмогорова, согласно которой всякая оптимальная оценка, если она существует, является функцией от достаточной статистики. Доказательство этой теоремы может быть найдено в указанной литературе (см. также [22, с. 48]).

2.5.Б. Эффективные оценки для экспоненциального семейства распределений

Можно выделить классы параметрических моделей (см., например, [24]), для которых всегда существует эффективная оценка некоторой параметрической функции. К таким классам относится семейство экспоненциальных моделей (см. формулу (2.4.29)), которые часто используются на практике.

Определение 2.5.33. Параметрическое семейство распределений вероятностей $\Im_\theta = \{F : F_X(x,\theta),\ \theta \in \Theta\}$ называют экспоненциальной моделью распределений, если плотность распределения вероятности $f_X(x;\theta)$ имеет вид

$$f_X(x;\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\}. \qquad \textbf{(2.5.34)}$$

Для экспоненциальной модели (2.5.34) функция вклада выборки, или логарифмическая производная статистики правдоподобия $L(\vec{X};\theta)$, записывается в виде

$$U(\vec{X},\theta) = \frac{\partial \ln L(\vec{X};\theta)}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i;\theta)}{\partial\theta} =$$

$$= A'(\theta)\sum_{i=1}^{n} B(X_i) + nC'(\theta) = nA'(\theta)\left[\frac{1}{n}\sum_{i=1}^{n} B(X_i) + \frac{C'(\theta)}{A'(\theta)}\right]. \qquad \textbf{(2.5.35)}$$

Сравнивая (2.5.35) и (2.4.10), заключаем, что последнее выражение будет совпадать с критерием эффективности, если параметрическая функция $\tau(\theta)$ и множитель $A^*(\theta)$ в (2.4.10) будут определяться следующими формулами:

$$\tau(\theta) = -C'(\theta)/A'(\theta),\ A^*(\theta) = nA'(\theta). \qquad \textbf{(2.5.36)}$$

Таким образом, если параметрическая модель $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$ экспоненциальная, то есть $f_X(x;\theta)$ определяется в виде (2.5.34) и удовлетворяет условиям регулярности (см. разд. 2.3), то для параметрической функции вида $\tau(\theta) = -C'(\theta)/A'(\theta)$ существует эффективная оценка

$$T^*(X_1,...,X_n) = \frac{1}{n}\sum_{i=1}^{n} B(X_i), \qquad \textbf{(2.5.37)}$$

дисперсия которой, согласно (2.4.11), вычисляется по формуле

$$D[T^*(X_1,...,X_n)] = \left|\frac{\tau'(\theta)}{nA'(\theta)}\right| = \left|\frac{C'(\theta)A''(\theta) - C''(\theta)A'(\theta)}{n[A'(\theta)]^3}\right|. \qquad \textbf{(2.5.38)}$$

Дополнительные сведения и примеры различных моделей экспоненциальных распределений могут быть найдены в [5, 7, 24, 38].

2.6. Асимптотическая эффективность оценок параметров

Для изучения свойств оценок параметров часто при их сравнении используют асимптотический подход, который основан на изучении асимптотических распределений оценок (см. [38, с. 305]). Многие оценки, используемые на практике, обычно являются несмещенными (или асимптотически несмещенными), состоятельными и имеют асимптотически нормальное распределение. Поэтому, естественно, что в качестве основы для их сравнения используется асимптотическая дисперсия.

Пусть интересующий нас параметр θ оценивается по выборке $\vec{X} = (X_1,...,X_n)$ из распределения F с помощью двух оценок $\hat{\theta}_1(X_1,...,X_n)$ и $\hat{\theta}_2(X_1,...,X_n)$. Предполагаем, что обе оценки состоятельные, асимптотически несмещенные и асимптотически нормальные, то есть при $n \to \infty$ выполняются выражения

$$L\left(\frac{\sqrt{n}(\hat{\theta}_i - \theta)}{\sigma(\hat{\theta}_i, F)}\right) = N(0;1),\ i = 1,2, \qquad \textbf{(2.6.1)}$$

где $\sigma^2(\hat{\theta}_i, F) = D_F[\sqrt{n}\hat{\theta}_i(\vec{X})]$ – асимптотическая дисперсия оценки

$\sqrt{n}\hat{\theta}_i(\vec{X})$, $i=1,2$. При выполнении этих предположений, для сравнения оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ используют асимптотическую относительную эффективность, которую определяют в виде отношения асимптотических дисперсий оценок, то есть

$$AO\Theta_F(\hat{\theta}_1:\hat{\theta}_2)=\frac{D_F(\hat{\theta}_2)}{D_F(\hat{\theta}_1)}=\frac{\sigma^2(\hat{\theta}_2,F)}{\sigma^2(\hat{\theta}_1,F)}. \qquad \textbf{(2.6.2)}$$

При изучении качества конкретной оценки $\hat{\theta}$ параметра θ также используют понятие асимптотической эффективности оценки (или абсолютной эффективности оценки), которую определяют в виде

$$A\Theta_F(\hat{\theta})=AO\Theta_F(\hat{\theta}:\hat{\theta}_{\text{эфф}})=\frac{\sigma^2(\hat{\theta}_{\text{эфф}},F)}{\sigma^2(\hat{\theta},F)}=\{\sigma^2(\hat{\theta},F)i_1(\theta)\}^{-1}, \qquad \textbf{(2.6.3)}$$

где $\hat{\theta}_{\text{эфф}}$ – эффективная оценка (или МГД-оценка, оценка с минимальной граничной дисперсией) параметра θ, дисперсия которой равна обратной величине количества информации Фишера $i_1(\theta)$, определенного в (2.3.10).

Замечание 2.6.4. Обоснованием введенного выражения асимптотической относительной эффективности оценок в виде (2.6.2) служат следующие рассуждения [38, с. 305]. При сравнении двух оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ параметра θ в асимптотике (то есть при $n\to\infty$) рассматривают свойства не самих оценок, а изучают свойства последовательностей оценок $\{\hat{\theta}_{1n}\}$ и $\{\hat{\theta}_{2n}\}$ при $n\to\infty$. Пусть $\{\hat{\theta}_{1n}\}$ – последовательность оценок для θ по наблюдениям $X_1,...,X_n$, такая, что выполняется выражение

$$\sqrt{n}\{\hat{\theta}_{1n}-\theta\}\xrightarrow{D}N\{0;\sigma^2(\hat{\theta}_1,F)\}. \qquad \textbf{(2.6.5)}$$

Далее, пусть $\{\hat{\theta}_{2n}\}$ – другая последовательность оценок для θ по $n'=n'(n)$ наблюдениям, и для нее также выполняется выражение

$$\sqrt{n}\{\hat{\theta}_{2n'}-\theta\}\xrightarrow{D}N\{0;\sigma^2(\hat{\theta}_2,F)\}. \qquad \textbf{(2.6.6)}$$

При сформулированных предположениях асимптотическая относительная эффективность последовательности $\{\hat{\theta}_{1n}\}$ относительно $\{\hat{\theta}_{2n}\}$ определяется в виде предела отношения объемов выборок, при которых предельные распре-

деления оценок $\hat{\theta}_{1n}$ и $\hat{\theta}_{2n'}$ совпадают, то есть

$$AO\Theta_F(\hat{\theta}_1 : \hat{\theta}_2) = \lim_{n\to\infty}\{n'(n)/n\} = \sigma^2(\hat{\theta}_2, F)/\sigma^2(\hat{\theta}_1, F). \qquad \textbf{(2.6.7)}$$

Справедливость второго равенства в (2.6.7) проверяется следующим образом. Так как выполняется равенство $\sqrt{n}\{\hat{\theta}_{2n'} - \theta\} = \sqrt{(n/n')}\sqrt{n}\{\hat{\theta}_{2n'} - \theta\}$, то из теоремы Слуцкого (П.3.9) следует, что $\sqrt{n}\{\hat{\theta}_{2n'} - \theta\}$ имеет такое же предельное распределение $N\{0\,; \sigma^2(\hat{\theta}_1, F)\}$, что и $\sqrt{n}\{\hat{\theta}_{1n} - \theta\}$ тогда и только тогда, когда существует предел $\lim_{n\to\infty}\{n/n'(n)\}$ и выполняется равенство $\sigma^2(\hat{\theta}_2, F)\cdot\lim_{n\to\infty}\{n/n'(n)\} = \sigma^2(\hat{\theta}_1, F)$, из которого следует второе равенство в (2.6.7).

Отметим, что асимптотическая относительная эффективность двух асимптотически нормальных оценок $\hat{\theta}_1$ и $\hat{\theta}_2$ параметра θ совпадает с асимптотической относительной эффективностью Питмена двух критериев, использование которых в процедуре Ходжеса – Лемана, приводит к оценкам $\hat{\theta}_1$ и $\hat{\theta}_2$ параметра θ (см. пример (5.2.24) и формулу (5.2.27)).

Отметим также, что для сравнения совокупности оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ при заданном симметричном распределении F используют понятие дефекта оценки (см., например, [122] и [100]). Пусть имеется конечный набор асимптотически нормальных и несмещенных оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ, построенных по выборке $X_1,...,X_n$ из распределения F, для которых выполняются выражения

$$L\left\{\frac{\sqrt{n}(\hat{\theta}_i - \theta)}{\sigma_F(\hat{\theta}_i)}\right\} = N(0,\ 1) \text{ при } n\to\infty,\ i = 1,...,k. \qquad \textbf{(2.6.8)}$$

Дефект оценки $\hat{\theta}_i$, $i = 1,...,k$, среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ при заданном распределении F определяется в виде

$$DE_F(\hat{\theta}_i) = 1 - \min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\}/\sigma_F^2(\hat{\theta}_i),\ i = 1,...,k. \qquad \textbf{(2.6.9)}$$

Отметим, что если среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ есть эффективная оценка $\hat{\theta}^*$ параметра θ при заданном распределении F, для которой $\sigma_F^2(\hat{\theta}^*) = 1/I(f)$, тогда $\min\{\sigma_F^2(\hat{\theta}_1),...,\sigma_F^2(\hat{\theta}_k)\} = 1/I(f)$ и, следовательно, в этом случае абсолютный дефект оценки $\hat{\theta}_i$ равен единица минус ее абсолютная эффективность, то есть

$$ADE_F(\hat{\theta}_i) = 1 - A\Theta_F(\hat{\theta}_i),\ i = 1,...,k. \qquad \textbf{(2.6.10)}$$

При изучении свойств робастности сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра сдвига θ в рамках супермодели $\Im$, состоящей из конечного набора симметричных распределений $\Im = \{F_1,...,F_r\}$, изучают поведение дефективностей оценок на плоскости двух распределений. По оси абсцисс обычно откладывают дефективность для базовой (идеальной модели, обычно гауссовской), а по оси ординат откладывают дефективность для альтернативной модели, входящей в супермодель $\Im = \{F_1,...,F_r\}$. При таком наглядном представлении дефективностей оценок на плоскости двух распределений предпочтение отдается той оценке, которая окажется ближе к началу координат (см., например, рис. 8.9.3, 8.9.4 и 8.9.21 – 8.9.24 в разд. 8.9). Если же мы хотим сделать вывод о предпочтительности оценки среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$ параметра θ в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$, то можно использовать евклидову метрику, которая, с использованием введенных обозначений, запишется в виде

$$d(\hat{\theta}_i;\Im) = \left\{\sum_{j=1}^{r}[DE_{F_j}(\hat{\theta}_i)]^2\right\}^{1/2}, \quad \text{или} \quad Ad(\hat{\theta}_i;\Im) = \left\{\sum_{j=1}^{r}[ADE_{F_j}(\hat{\theta}_i)]^2\right\}^{1/2},$$

$$i = 1,..., k. \qquad \textbf{(2.6.11)}$$

Предпочтение в рамках всей рассматриваемой супермодели $\Im = \{F_1,...,F_r\}$ отдается той оценке $\hat{\theta}_i$ среди сравниваемых оценок $\hat{\theta}_1,...,\hat{\theta}_k$, для которой вычисленное значение $d(\hat{\theta}_i;\Im)$ минимальное, то есть

$$d(\hat{\theta}_i;\Im) = \min\{d(\hat{\theta}_1;\Im),...,d(\hat{\theta}_k;\Im)\}. \qquad \textbf{(2.6.12)}$$

Примеры вычисления этих характеристик для оценок конкретных параметров будут приведены ниже в следующих разделах (см. также [100]).

2.7. Методы построения статистических оценок

В предыдущих разделах мы обсудили общие свойства оценок, которые используются при изучении их качества и при их сравнении. Прежде чем переходить к обсуждению основных методов построения точечных оценок параметров, мы рассмотрим сначала одну из основных, базовых задач, которая связана с построением по выборке $X_1,...,X_n$ оценки неизвестной ф.р. $F_X(x)$ в любой точке $x \in R^1$. Эту оценку называют эмпирической функцией распределе-

ния, и она служит основой при построении многих статистических процедур, в частности метода подстановок для нахождения выборочных оценок параметров, заданных в виде функционалов на множестве допустимых функций распределений $\Im = \{F_X\}$.

2.7.А. Эмпирическая функция распределения (непрерывный случай)

Пусть X – непрерывная случайная величина с ф.р. $F_X(x)$, которая неизвестна. Имеется выборка $X_1,...,X_n$ из распределения $F_X(x)$, $x \in R^1$. Требуется построить по выборке $X_1,...,X_n$ оценку неизвестной ф.р. $F_X(x) = P(X \le x)$ для любого $x \in R^1$. Для построения оценки вероятности события $(X \le x)$ при любом фиксированном $x \in R^1$ воспользуемся статистическим подходом к определению вероятности этого события. Учитывая, что $X_1,...,X_n$ – н.о.р. случайные величины, при этом $F_{X_i}(x) = F_X(x)$, $i = 1,...,n$, определим $\forall x \in R^1$ случайную величину

$$\mu_n(x) = [\text{числу } X_i \le x,\ i = 1,...,n]. \qquad \textbf{(2.7.1)}$$

Оценку неизвестной ф.р. $F_X(x)$ называют эмпирической функцией распределения, обозначают ее через $F_n(x)$ и определяют в виде относительной частоты события $(X \le x)$ в n независимых опытах $X_1,...,X_n$, то есть в виде

$$F_n(x) = \frac{\mu_n(x)}{n} = \frac{1}{n}\sum_{i=1}^{n} C(x - X_i) = \begin{cases} 0 & \text{при } x < X_{(1)}, \\ k/n & \text{при } X_{(k)} \le x < X_{(k+1)}, \\ 1 & \text{при } x \ge X_{(n)}, \end{cases}$$

$$k = 1,...,n-1, \qquad \textbf{(2.7.2)}$$

где $C(\cdot)$ – единичная функция, определенная в (1.3.23), и $X_{(1)},...,X_{(n)}$ – упорядоченная статистика (см. разд. 4.2). Отметим, что для непрерывной с.в. X с возможными значениями на действительной оси R^1 эмпирическая функция распределения $F_n(x)$ может

быть записана более компактно. Для этого следует определить дополнительные порядковые статистики $X_{(0)} = -\infty$ и $X_{(n+1)} = +\infty$, тогда (2.7.2) запишется в виде

$$F_n(x) = i/n \ \ при \ \ X_{(i)} \le x < X_{(i+1)},\ i = 0,1,...,n. \qquad \textbf{(2.7.2a)}$$

Отметим также, что случайная величина $\mu_n(x)$ для любого фиксированного $x \in R^1$ имеет биномиальное распределение с параметрами n и p, где n – объем выборки $X_1,...,X_n$ и p – вероятность «успеха». Под «успехом» понимается событие $(X_i \le x)$, которое осуществляется с вероятностью $P(X_i \le x) = F_{X_i}(x) = F_X(x),\ i = 1,...,n$. Коротко этот факт записывается в виде

$$L(\mu_n(x)) = Bi(n, F_X(x)),\ \forall x \in R^1. \qquad \textbf{(2.7.3)}$$

Эмпирическая функция распределения $F_n(x)$ обладает следующими общими свойствами.

1. Эмпирическая ф.р. $F_n(x)$, являясь для $\forall x \in R^1$ оценкой ф.р. $F_X(x)$, построенной по выборке $X_1,...,X_n$, обладает всеми общими свойствами ф.р. $F_X(x)$. В частности, значения эмпирической ф.р. $F_n(x)$ лежат в интервале от нуля до единицы, то есть $0 \le F_n(x) \le 1$ $\forall x \in R^1$. Далее, $F_n(x)$ – неубывающая и непрерывная справа функция. Отметим также, что $F_n(x)$ ступенчатая функция со скачками величиной $1/n$ в точках $x_{(i)},\ i = 1,...,n$, которые являются элементами вариационного ряда $x_{(1)} < ... < x_{(n)}$. График эмпирической функции $F_n(x)$, построенный по реализации $x_1,...,x_n$ выборки $X_1,...,X_n$, приведен на рис. 2.7.1. Отметим, что обратная функция для эмпирической функции $F_n(x)$ определяется в виде

$$x_{(np)} = F_n^{-1}(p) = \inf\{x : F(x) \ge p\},\ 0 < p < 1.$$

2. Эмпирическая ф.р. $F_n(x)$ является несмещенной оценкой ф.р. $F_X(x)$, то есть

$$M[F_n(x)] = F_X(x),\ \forall x \in R^1.$$

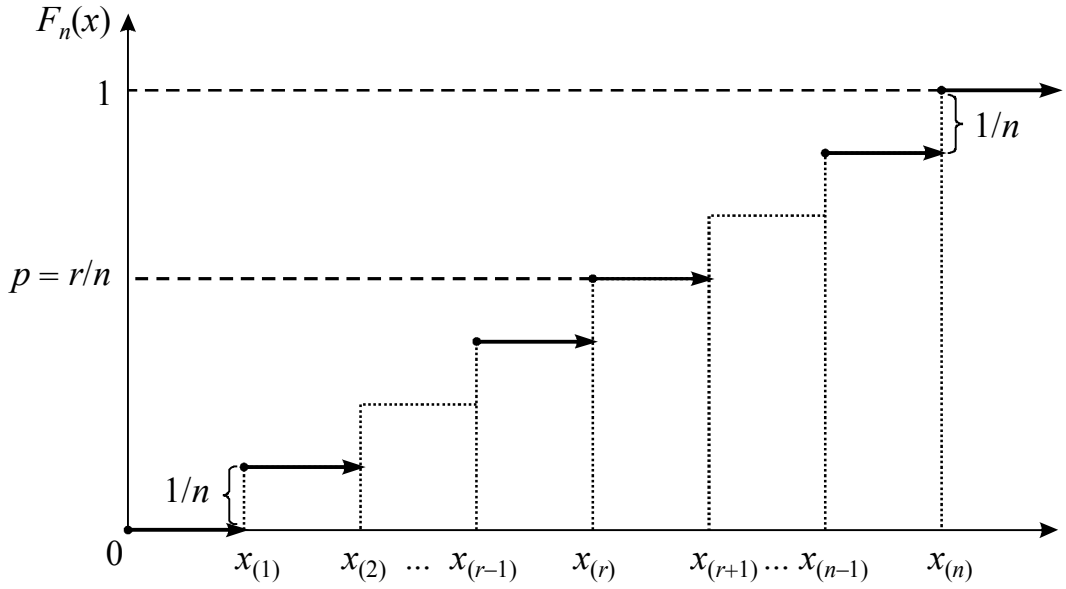

Рис. 2.7.1. Эмпирическая функция $F_n(x)$

Кроме того, ее дисперсия (детали см. в упр. 4.11.29) вычисляется по формуле

$$D(F_n(x)) = F_X(x)(1-F_X)/n\,,\ \forall x \in R^1\,.$$

3. Эмпирическая функция распределения $F_n(x)$ является случайной ступенчатой функцией, однако для любого *фиксированного* $x \in R^1$ она может рассматриваться как случайная величина дискретного типа с возможными значениями $0,\ 1/n,...,\ k/n,...,\ 1$, вероятности которых с учетом (2.7.3) вычисляются по формуле

$$P\{F_n(x) = k/n\} = P\{\mu_n(x) = k\} = C_n^k F_X^k(x)(1-F_X(x))^{n-k}\ ,$$
$$k = 0,1,...,n\,. \qquad \textbf{(2.7.4)}$$

4. Эмпирическая ф.р. $F_n(x)$ обладает «хорошими» асимптотическими свойствами, то есть при увеличении объема выборки n ее качество как оценки ф.р. $F_X(x)$ улучшается. В частности, эмпирическая ф.р. $F_n(x)$ является состоятельной оценкой ф.р. $F_X(x)$, то есть она сходится по вероятности к ф.р. $F_X(x)$, что записывается в виде $F_n(x) \xrightarrow{p} F_X(x)$ при $n \to \infty$ и эта запись означает выполнение следующего выражения:

$$\forall \varepsilon > 0 \;\; P\{|F_n(x) - F_X(x)| < \varepsilon\} \to 1 \text{ при } n \to \infty, \; \forall x \in R^1.$$

Справедливость этого выражения легко проверяется с помощью неравенства Чебышева с учетом формул $M[F_n(x)] = F_X(x)$, $D(F_n(x)) = F_X(x)(1 - F_X)/n, \forall x \in R^1$ (см. упр. 4.11.29). Далее, согласно теореме Муавра – Лапласа (см. П.1.17), эмпирическая ф.р. $F_n(x)$ имеет асимптотически нормальное распределение, то есть $\forall x \in R^1$ выполняется выражение

$$L\left\{\frac{\sqrt{n}(F_n(x) - F_X(x))}{\sqrt{F_X(x)(1 - F_X(x))}}\right\} = N(0,1) \text{ при } n \to \infty.$$

Отметим также, что асимптотическое поведение эмпирической ф.р. $F_n(x)$ характеризуется теоремой Гливенко (1933) и теоремой Колмогорова (1933). Согласно теореме Гливенко, отклонение эмпирической ф.р. $F_n(x)$ от ф.р. $F_X(x)$, описываемое величиной

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F_X(x)|, \qquad \textbf{(2.7.5)}$$

является сколь угодно малым с вероятностью единица при $n \to \infty$, то есть

$$P\{\lim_{n \to \infty} D_n = 0\} = 1. \qquad \textbf{(2.7.6)}$$

Далее, согласно теореме Колмогорова, для любой непрерывной ф.р. $F_X(x)$ асимптотическая функция распределения Колмогорова $F_{\sqrt{n}D_n}(y)$ случайной величины $\sqrt{n}D_n$ определяется в виде

$$\lim_{n \to \infty} P\{\sqrt{n}D_n \le y\} = F_{\sqrt{n}D_n}(y) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2y^2}, \; y > 0. \qquad \textbf{(2.7.7)}$$

Замечание 2.7.8. Отметим, что точное распределение статистики Колмогорова D_n при конечных объемах выборки $n \le 100$ подробно табулировано (см. например, [6]). Отметим также, что асимптотическое распределение $F_{\sqrt{n}D_n}(y)$ является приемлемой аппроксимацией для решения практических задач при объемах выборки $n \ge 40$ (см. упр. 4.11.49). Использование этого распределения и техники построения доверительных интервалов (см. разд. 2.12 – 2.14) позво-

ляет сделать полезные заключения относительно объема выборки n, при котором будет обеспечена заданная аппроксимация неизвестной ф.р. $F_X(x)$. В самом деле, пусть $t_{1-\alpha}$ квантиль уровня $1-\alpha$ функции распределения Колмогорова, то есть $F_{\sqrt{n}D_n}(t_{1-\alpha}) = 1-\alpha$, где α – маленькая величина, например $\alpha = 0,05$ или $\alpha = 0,01$. Тогда, при достаточно больших n, из (2.7.7) имеем

$$P\left\{|F_n(x) - F_x(x)| \le \frac{t_{1-\alpha}}{\sqrt{n}}\right\} = 1-\alpha \qquad \textbf{(2.7.9)}$$

или

$$P\left\{F_n(x) - \frac{t_{1-\alpha}}{\sqrt{n}} \le F_X(x) \le F_n(x) + \frac{t_{1-\alpha}}{\sqrt{n}}\right\} = 1-\alpha. \qquad \textbf{(2.7.10)}$$

Учитывая теперь, что $0 \le F_X(x) \le 1$, $\forall x \in R^1$, окончательно получаем доверительную полосу для функции распределения $F_X(x)$ в виде

$$P\left\{\max[0,\ F_n(x) - \frac{t_{1-\alpha}}{\sqrt{n}}] \le F_X(x) \le \min[1,\ F_n(x) + \frac{t_{1-\alpha}}{\sqrt{n}}]\right\} = 1-\alpha. \qquad \textbf{(2.7.11)}$$

Отметим, что использование полученной доверительной полосы позволяет вычислить необходимый объем выборки, для оценивания функции распределения с заданной точностью. В самом деле, пусть разность между F_n и F_X не превышает заданной величины Δ, то есть $|F_n(x) - F_X(x)| \le \Delta$, $\forall x \in R^1$. Тогда из формулы (2.7.9) следует, что при выполнении неравенства $n \ge (t_{1-\alpha}/\Delta)^2$ неизвестная ф.р. $F_X(x)$ будет накрываться полосой $(F_n(x) \pm \Delta)$ с доверительной вероятностью $1-\alpha$. Например, при заданной точности $\Delta = 0,05$ и доверительной вероятности 0,99 (то есть $\alpha = 0,01$ и $t_{0,99} = 1,628$), объем выборки должен быть более чем $(1,63/0,05)^2 \approx 1060$ наблюдений.

2.7.Б. Дискретный случай

Пусть X – дискретная случайная величина, заданная распределением вероятности (1.1.22), которое записывается в виде

$$\begin{pmatrix} X \\ p \end{pmatrix} : \begin{pmatrix} \tilde{x}_1, ..., \tilde{x}_k \\ p_1, ..., p_k \end{pmatrix}, \ \sum_{i=1}^{k} p_i = 1, \qquad \textbf{(2.7.12)}$$

где $\tilde{x}_1, ..., \tilde{x}_k$ – отстоящие друг от друга возможные состояния с.в. X, а $p_i = P(X = \tilde{x}_i)$, $i = 1, ..., k$, их вероятности. Согласно (1.3.24), функция распределения вероятностей с.в. X записывается в виде

$$F_X(x) = \sum_{i:\,\tilde{x}_i \le x} p_i = \sum_{i=1}^{k} p_i C(x - \tilde{x}_i),\ x \in R^1. \qquad \textbf{(2.7.13)}$$

На практике обычно вероятности $p_1,...,p_k$ неизвестны, поэтому ф.р. $F_X(x)$, определяемая формулой (2.7.13), не может быть конкретизирована. В таких случаях возникает статистическая задача, связанная с построением статистической оценки неизвестной функции распределения $F_X(x)$ по выборке $X_1,...,X_n$. Эту оценку, как и в непрерывном случае, называют эмпирической функцией распределения и обозначают через $F_n(x)$.

Для построения эмпирической функции распределения $F_n(x)$ введем понятия дискретных вариационных рядов частот и относительных частот.

Определение 2.7.14. Дискретный вариационный ряд частот определяется по наблюденной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$ из распределения $F_X(x)$ дискретной с.в. X, заданной распределением вероятностей (2.7.12), и записывается в виде

$$(x_1,...,x_n) \Rightarrow \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ n_1,...,n_k \end{pmatrix},\ \sum_{i=1}^{k} n_i = n, \qquad \textbf{(2.7.15)}$$

где n_i – частота появления состояния $\tilde{x}_i$, $i = 1,...,k$, в наблюденных данных $x_1,...,x_n$.

Отметим, что сумма частот $n_1,...,n_k$ равна объему выборки n. Отметим также, что частоты $n_1,...,n_k$ являются реализациями дискретных случайных величин $N_1,...,N_k$, которые зависимы и описываются полиномиальным распределением, то есть

$$P\{N_1 = n_1,...,N_k = n_k\} = \frac{n!}{n_1!...n_k!}\prod_{i=1}^{k} p_i^{n_i},\ \sum_{i=1}^{k} n_i = n. \qquad \textbf{(2.7.16)}$$

Определение 2.7.17. Дискретный вариационный ряд относительных частот определяется по конкретной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$ из распределения $F_X(x)$ дискретной с.в. X, заданной распределением вероятностей (2.7.12), и записывается в виде

$$\begin{pmatrix} X \\ \hat{p} \end{pmatrix} : \begin{pmatrix} \tilde{x}_1, \ldots\ldots\ldots, \tilde{x}_k \\ n_1/n, \ldots, n_k/n \end{pmatrix}, \ \sum_{i=1}^{k} (n_i/n) = 1, \qquad \textbf{(2.7.18)}$$

где n_i/n – относительная частота появления состояния $\tilde{x}_i$, $i = 1, \ldots, k$, в наблюденных данных $x_1, \ldots, x_n$.

Отметим, что относительные частоты $\hat{p}_i = n_i/n$, $i = 1, \ldots, k$, являются статистическими оценками вероятностей $p_i = P(X = \tilde{x}_i)$, $i = 1, \ldots, k$. Таким образом, дискретный вариационный ряд относительных частот (2.7.18) является статистическим аналогом распределения вероятностей (2.7.12) дискретной с.в. X, построенным по выборке $X_1, \ldots, X_n$.

Эмпирическая ф.р. $F_n(x)$ дискретной с.в. X вычисляется методом подстановки, то есть путем замены неизвестных вероятностей $p_1, \ldots, p_k$ в (2.7.13) на их оценки $\hat{p}_i = n_i/n$, $i = 1, \ldots, k$, и записывается в виде

$$F_n(x) = n^{-1} \sum_{i=1}^{k} n_i C(x - \tilde{x}_i) = \sum_{i=1}^{k} n_i C(x - \tilde{x}_i) \Big/ \sum_{i=1}^{k} n_i \, . \qquad \textbf{(2.7.19)}$$

Итак, эмпирическая ф.р. $F_n(x)$ в дискретном случае вычисляется по выборке $X_1, \ldots, X_n$, которая предварительно отображается в дискретный вариационный ряд частот (2.7.15) и записывается в виде среднего взвешенного единичных функций $C(x - \tilde{x}_i)$ с весами n_i, $i = 1, \ldots, k$, то есть в виде (2.7.19).

Замечание 2.7.20. Отметим, что при решении многих практических задач, возникающих при анализе сложных социально-экономических явлений, объемы выборок обычно очень большие. В связи с этим, исходные статистические данные на этапе первичной обработки часто представляют в группированной форме. Для этого наблюденная реализация $x_1, \ldots, x_n$ выборки $X_1, \ldots, X_n$ отображается в интервальный вариационный ряд частот (или относительных частот), то есть

$$(x_1, \ldots, x_n) \Rightarrow \begin{pmatrix} \Delta_1, \ldots, \Delta_k \\ n_1, \ldots, n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n, \qquad \textbf{(2.7.21)}$$

где n_i – частота попадания наблюденных данных $x_1, \ldots, x_n$ в интервал группировки Δ_i, $i = 1, \ldots, k$. Правила выбора интервалов группировки $\Delta_1, \ldots, \Delta_k$ обсуждаются в [16, 29, 56].

2.7.В. Многомерный вариант

Если (X,Y) – двумерная случайная величина непрерывного типа с ф.р. $F_{XY}(x,y)=P\{X\le x;Y\le y\}$, $(x,y)\in R^2$, то эмпирическая ф.р. $F_n(x,y)$, построенная по двумерной выборке $(X_1,Y_1),...,(X_n,Y_n)$ объема n, записывается в виде

$$F_n(x,y)=n^{-1}\sum_{i=1}^{n}I(X_i\le x)I(Y_i\le y)=n^{-1}\sum_{i=1}^{n}C(x-X_i)C(y-Y_i), \quad \textbf{(2.7.22)}$$

где $I(A)$ – индикатор события A.

В общем случае, если рассматривается l-мерный случайный вектор $\vec{X}_l=(X_1,...,X_l)$, то l-мерная эмпирическая ф.р. $F_n(x_1,...,x_l)$, построенная по l-мерной выборке $(X_{11},...,X_{1l}),...,(X_{n1},...,X_{nl})$ объема n, записывается в виде

$$F_n(x_1,...,x_l)=n^{-1}\sum_{i=1}^{n}\prod_{j=1}^{l}C(x_j-X_{ij}). \quad \textbf{(2.7.23)}$$

Свойство несмещенности эмпирической функции распределения $F_n(x_1,...,x_l)$ проверяется в упр. 4.11.36.

Подводя итог данного раздела, отметим, что эмпирическая ф.р. $F_n(x)$, является статистическим аналогом, или оценкой ф.р. $F_X(x)$, построенной по выборке $X_1,...,X_n$. Она обладает такими «хорошими» свойствами, как несмещенность и состоятельность. Точность этой оценки возрастает с увеличением объема выборки (см. упр. 4.11.29). Именно эти качества послужили основой для построения метода подстановки оценивания параметров, с помощью которого получены все выборочные оценки числовых характеристик распределений, представляемых в виде функционалов, определенных на множестве $\Im=\{F_X\}$.

2.8. Методы построения точечных оценок параметров

В теории построения оценок параметров выделяют два класса оценок. Их называют точечные и интервальные оценки. Мы рассмотрим сначала метод подстановок для нахождения точечных

оценок параметров, заданных в виде функционалов на множестве допустимых ф.р. $\Im = \{F_X\}$. После этого обсудим три основных метода построения точечных оценок параметров для случая, когда модель задана в параметрической форме. Это метод моментов, метод максимального правдоподобия и метод минимума расстояний. Каждый из этих методов имеет свои достоинства и недостатки, которые мы отметим после их описания и обсуждения.

2.8.А. Метод подстановки

Пусть интересующий нас параметр θ представляется в виде функционала $T(F)$ от ф. р. F_X, заданного на множестве допустимых функций распределения $\Im$, то есть

$$\theta = T(F),\ F \in \Im\colon\ \Im \to R^1. \qquad \textbf{(2.8.1)}$$

Предполагаем, что функциональный вид ф.р. F_X неизвестен и, следовательно, параметр θ не может быть вычислен. В таких ситуациях возникает статистическая задача, связанная с построением оценки параметра θ, на основе информации, содержащейся в выборке $X_1,...,X_n$. В данном случае статистические оценки интересующего нас параметра θ обычно строят методом подстановки. Суть этого метода состоит в том, что сначала на основе выборки $X_1,...,X_n$ строится эмпирическая функция распределения $F_n(x)$, а затем неизвестная ф.р. F_X в функционале $T(F)$ заменяется на эмпирическую функцию распределения $F_n(x)$. В результате выборочная оценка параметра θ будет зависеть от наблюдений $X_1,...,X_n$ и запишется в виде

$$\hat{\theta}(X_1,...,X_n) = T(F_n). \qquad \textbf{(2.8.2)}$$

Отметим, что если функционал $T(F)$ непрерывен относительно поточечной сходимости (см., например, [13, с. 370]), то есть $T(F_n) \to T(F)$ при $F_n(x) \to F(x)$ $\forall x \in R^1$ при $n \to \infty$, тогда $\lim\limits_{n\to\infty} T(F_n) = T(F) = \theta$ с вероятностью единица. Таким образом, выборочные оценки непрерывных функционалов, построенные мето-

дом подстановки, являются сильно состоятельными, и, следовательно, оценки $T(F_n)$ сходятся по вероятности к $T(F)$, то есть $T(F_n) \xrightarrow{P} T(F)$ при $n \to \infty$ (см. также теорему 4.4.1 в монографии [5, т.1, с.144]).

Пример 2.8.3. Многие числовые характеристики ф.р. F_X в обобщенном варианте могут быть представлены функционалами вида

$$T(F) = \int_{-\infty}^{+\infty} \varphi(x) dF(x), \qquad \textbf{(2.8.4)}$$

где $\varphi(x)$ – заданная непрерывная функция. Например, $\varphi(x) = x^k$, $k = 1, 2, \ldots$, соответствует начальным моментам α_k k-го порядка, см. определение 1.3.12. Другой класс числовых характеристик может быть представлен функционалами вида

$$S(F) = \int_{-\infty}^{+\infty} \psi(x, T(F)) dF(x), \qquad \textbf{(2.8.5)}$$

где $\psi(x, T(F))$ – заданная непрерывная функция и $T(F)$ – функционал, характеризующий параметр положения ф.р. F_X. Например, при $T(F) = \alpha_1(F)$ и $\psi(x, \alpha_1(F)) = (x - \alpha_1)^k$, $k = 1, 2, \ldots$, функционалы вида (2.8.5) соответствуют центральным моментам μ_k k-го порядка, см. формулу (1.3.15). Выборочные оценки этих функционалов, построенные методом подстановки, запишутся, с учетом формул (2.7.2) и (1.4.100), в виде

$$T(F_n) = \int_{-\infty}^{+\infty} \varphi(x) dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \varphi(x) \delta(x - X_i) dx = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i). \qquad \textbf{(2.8.6)}$$

Аналогично получаем выборочную оценку функционала $S(F)$ в виде

$$S(F_n) = \int_{-\infty}^{+\infty} \psi(x, T(F_n)) dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, T(F_n)). \qquad \textbf{(2.8.7)}$$

Примеры применения метода подстановки приводятся в упр. 2.15.14 и 2.15.15.

Пример 2.8.8. Рассмотрим различные варианты средних характеристик, которые часто используют при решении практических задач. В литературе по общей теории статистики (см., например, [16, 78]) различают два вида средних:

- аналитические характеристики среднего значения;
- структурные характеристики среднего значения.

Аналитические характеристики основаны на использовании понятия «среднее степенное». К структурным характеристикам относят такие понятия, как «медиана» и «мода». В данном примере приведем аналитические характеристики средних и их выборочные оценки как для непрерывного, так и для дискретного случаев. Структурные характеристики средних значений обсуждаются в упр. 2.15.24.

Пусть X – непрерывная случайная величина с ф.р. $F_X(x), x \in R^1$. Среднее степенное ν-го порядка (ν – целое) определяется в виде функционала $T_\nu(F)$, равного корню ν-й степени из начального момента α_ν ν-го порядка, то есть в виде

$$T_\nu(F) = \sqrt[\nu]{\alpha_\nu} = [M(X^\nu)]^{1/\nu} = \left\{ \int_{-\infty}^{+\infty} x^\nu dF_X(x) \right\}^{1/\nu}. \qquad \textbf{(2.8.9)}$$

Различные варианты средних соответствуют различным значениям параметра ν. В частности, при $\nu = 2$ получаем среднеквадратическое $T_2(F)$ в виде

$$T_2(F) = \sqrt{\alpha_2} = \left\{ \int_{-\infty}^{+\infty} x^2 dF_X(x) \right\}^{1/2}; \qquad \textbf{(2.8.10)}$$

при $\nu = 1$ получаем математическое ожидание, то есть

$$T_1(F) = M(X) = \int_{-\infty}^{+\infty} x dF_X(x); \qquad \textbf{(2.8.11)}$$

при $\nu = 0$, после раскрытия неопределенности типа (1^∞), см. упр. 2.15.23, получаем среднее геометрическое в виде

$$T_0(F) = \exp\left\{ \int_{-\infty}^{+\infty} \ln x \cdot dF_X(x) \right\}; \qquad \textbf{(2.8.12)}$$

при $\nu = -1$ получаем среднее гармоническое в виде

$$T_{-1}(F) = \left\{ \int_{-\infty}^{+\infty} x^{-1} dF_X(x) \right\}^{-1}. \qquad \textbf{(2.8.13)}$$

Выборочные оценки средних степенных, построенные по выборке $X_1,...,X_n$ из распределения $F_X(x)$ методом подстановки, запишутся в виде

$$T_\nu(F_n) = \left\{ \int_{-\infty}^{+\infty} x^\nu dF_n(x) \right\}^{1/\nu} = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^\nu \right\}^{1/\nu}. \qquad \textbf{(2.8.14)}$$

В частности, выборочная оценка среднеквадратического, обычно обозначаемая через S, запишется в виде

$$T_2(F_n) = S = \left\{ \int_{-\infty}^{+\infty} x^2 dF_n(x) \right\}^{1/2} = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right\}^{1/2}. \qquad \textbf{(2.8.15)}$$

Далее, выборочной оценкой математического ожидания является обычное среднее арифметическое, которое обозначается через $\overline{X}$ и записывается в виде

$$T_1(F_n) = \overline{X} = n^{-1} \sum_{i=1}^{n} X_i . \qquad \textbf{(2.8.16)}$$

Выборочная оценка среднего геометрического записывается в виде

$$T_0(F_n) = \exp\left\{ \int_{-\infty}^{+\infty} \ln x \cdot dF_n(x) \right\} = \exp\left\{ n^{-1} \sum_{i=1}^{n} \ln X_i \right\} = \exp\left\{ \ln\left(\prod_{i=1}^{n} X_i \right)^{1/n} \right\} =$$

$$= \left\{ \prod_{i=1}^{n} X_i \right\}^{1/n} = \sqrt[n]{X_1 \cdots X_n} . \qquad \textbf{(2.8.17)}$$

Выборочная оценка среднего гармонического имеет вид

$$T_{-1}(F_n) = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^{-1} \right\}^{-1} = n / \sum_{i=1}^{n} (1/X_i) . \qquad \textbf{(2.8.18)}$$

Рассмотрим теперь дискретный случай. Пусть X – дискретная случайная величина, заданная распределением вероятностей

$$\begin{pmatrix} X \\ p \end{pmatrix} : \begin{pmatrix} \tilde{x}_1, ..., \tilde{x}_k \\ p_1, ..., p_k \end{pmatrix}, \ \sum_{i=1}^{k} p_i = 1, \qquad \textbf{(2.8.19)}$$

Функция распределения вероятностей F_X этой случайной величины определяется выражением (2.7.13). В данном случае среднее степенное из (2.8.9) с учетом (2.7.13) запишется в виде

$$T_\nu(F) = \sqrt[\nu]{\alpha_\nu} = [M(X^\nu)]^{1/\nu} = \left\{ \int_{-\infty}^{+\infty} x^\nu dF_X(x) \right\}^{1/\nu} = \left\{ \int_{-\infty}^{+\infty} x^\nu d\left(\sum_{i=1}^{k} p_i C(x - \tilde{x}_i) \right) \right\}^{1/\nu} =$$

$$= \left\{ \sum_{i=1}^{k} p_i \int_{-\infty}^{+\infty} x^\nu \delta(x - \tilde{x}_i) dx \right\}^{1/\nu} = \left\{ \sum_{i=1}^{k} p_i \tilde{x}_i^\nu \right\}^{1/\nu} . \qquad \textbf{(2.8.20)}$$

Выборочные оценки этих средних степенных, построенные по выборке $X_1,...,X_n$, отображенной в дискретный вариационный ряд частот (2.7.15), запишутся как

$$T_\nu(F_n)=\left\{n^{-1}\sum_{i=1}^{k}n_i\tilde{x}_i^\nu\right\}^{1/\nu}. \qquad \textbf{(2.8.21)}$$

В частности, из (2.8.21) при $\nu=1$ следует, что выборочная оценка математического ожидания в дискретном случае записывается в виде среднего взвешенного:

$$T_1(F_n)=\bar{X}=\sum_{i=1}^{k}\tilde{x}_i n_i/\sum_{i=1}^{k}n_i. \qquad \textbf{(2.8.22)}$$

Далее, выборочная оценка среднего гармонического следует из (2.8.21) при $\nu=-1$ и записывается в виде

$$T_{-1}(F_n)=\sum_{i=1}^{k}n_i/\sum_{i=1}^{k}(n_i/\tilde{x}_i). \qquad \textbf{(2.8.23)}$$

Выборочная оценка среднего геометрического следует из (2.8.21) при $\nu=0$. После раскрытия неопределенности типа (1^∞) с помощью правила Лопиталя получаем

$$T_0(F_n)=\lim_{\nu\to0}T_\nu(F_n)=\lim_{\nu\to0}\left\{n^{-1}\sum_{i=1}^{k}n_i\tilde{x}_i^\nu\right\}^{1/\nu}=\exp\left\{\ln\lim_{\nu\to0}\left\{n^{-1}\sum_{i=1}^{k}n_i\tilde{x}_i^\nu\right\}^{1/\nu}\right\}=$$

$$=\exp\left\{\lim_{\nu\to0}\ln\left\{n^{-1}\sum_{i=1}^{k}n_i\tilde{x}_i^\nu\right\}^{1/\nu}\right\}=\exp\left\{\lim_{\nu\to0}\ln\left\{n^{-1}\sum_{i=1}^{k}n_i\tilde{x}_i^\nu\right\}/\nu\right\}=$$

$$=\exp\left\{n^{-1}\sum_{i=1}^{k}n_i\ln\tilde{x}_i\right\}=\exp\left(\ln\left(\prod_{i=1}^{k}\tilde{x}_i^{n_i}\right)^{1/n}\right)=\sqrt[n]{\tilde{x}_1^{n_1}\cdots\tilde{x}_k^{n_k}}=\left\{\prod_{i=1}^{k}\tilde{x}_i^{n_i}\right\}^{1/\sum n_i}. \qquad \textbf{(2.8.24)}$$

Замечание 2.8.25. Если исходные статистические данные $X_1,...,X_n$ представлены интервальным вариационным рядом частот (2.7.21), то есть

$$(X_1,...,X_n)\Rightarrow\begin{pmatrix}\Delta_1,...,\Delta_k\\ n_1,...,n_k\end{pmatrix},\ \sum_{i=1}^{k}n_i=n, \qquad \textbf{(2.8.26)}$$

то согласно принятому соглашению (см., например, [78]), выборочные оценки различных числовых характеристик вычисляют так же, как в дискретном случае, с той лишь разницей, что дискретные состояния $\tilde{x}_i$, $i=1,...,k$, заменяют на середины $\hat{x}_i$ интервалов группировки Δ_i, $i=1,...,k$. В частности, выбороч-

ные оценки различных вариантов средних вычисляются по данным интервального ряда частот (2.8.26) по формулам (2.8.21) – (2.8.24), в которых следует заменить дискретные состояния $\tilde{x}_i$ на середины интервалов группировки $\hat{x}_i$, $i = 1,...,k$.

Замечание 2.8.27. Отметим, что задача построения эмпирической функции распределения $F_n(x)$ в некоторой фиксированной точке $x_0 \in R^1$ также может быть отнесена к задаче оценивания параметров вида (2.8.1). В самом деле, для любой фиксированной точки $x_0 \in R^1$ ф.р. F_X в этой точке может быть представлена функционалом вида (2.8.4) при $\varphi(x) = C(x_0 - x)$, то есть

$$F_X(x_0) = \int_{-\infty}^{x_0} dF_X(x) = \int_{-\infty}^{+\infty} C(x_0 - x) dF_X(x). \qquad \textbf{(2.8.28)}$$

Формальное применение метода подстановки дает эмпирическую функцию распределения в точке x_0 в виде

$$F_n(x_0) = \int_{-\infty}^{+\infty} C(x_0 - x) dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} C(x_0 - x)\delta(x - X_i)dx = \frac{1}{n}\sum_{i=1}^{n} C(x_0 - X_i).$$

Отметим, что распределение вероятностей выборочных оценок $\hat{\theta}(X_1,...,X_n) = T(F_n)$ функционалов $\theta = T(F)$ при конечных объемах выборки n обычно имеет сложный вид, который зависит от характера функционала и от распределения F_X. В связи с этим, на практике обычно используют асимптотические распределения этих оценок. Метод изучения асимптотических распределений оценок вида $T(F_n)$ был предложен фон Мизесом в 1948 году (см. [175, 183]) и будет подробно обсуждаться ниже при рассмотрении конкретных вариантов определения параметров $\theta = T(F)$, см. разд. 7.1.

Замечание 2.8.29. Нормальное распределение вероятностей, которое характеризуется двумя параметрами, математическим ожиданием и дисперсией, наиболее часто используется на практике. В связи с этим, выборочные оценки математического ожидания и дисперсии, являющиеся частным случаем оценок (2.8.6) и (2.8.7), построенные по выборке $X_1,...,X_n$ и вычисляемые по формулам

$$\overline{X} = n^{-1}\sum_{i=1}^{n} X_i \text{ и } S^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad \textbf{(2.8.30)}$$

вызывают особый интерес в гауссовском случае. Статистические свойства $\overline{X}$ и S^2 подробно изучены в литературе (см., например, [22, с. 32]) и представлены в следующих, часто используемых в статистической практике, теоремах.

Теорема 2.8.31. Пусть $X_1,...,X_n$ – выборка из нормального распределения $N(a,\sigma^2)$. Тогда выборочное среднее $\overline{X}$ и выборочная дисперсия S^2 независимы и выполняются следующие выражения:

$$L(\sqrt{n}(\overline{X}-a)/\sigma) = N(0;1); \qquad \textbf{(2.8.32)}$$

$$L(\sqrt{n-1}(\overline{X}-a)/S) = St(n-1); \qquad \textbf{(2.8.33)}$$

$$L(nS^2/\sigma^2) = \chi^2(n-1). \qquad \textbf{(2.8.34)}$$

Теорема 2.8.35. Пусть $X_1,...,X_n$ и $Y_1,...,Y_m$ – две независимые выборки из нормального распределения $N(a,\sigma^2)$ и пусть $\overline{X}$, S_X^2 и $\overline{Y}$, S_Y^2 – соответствующие выборочные средние и дисперсии. Тогда при любых a и $\sigma^2 > 0$ случайная величина $\sqrt{\dfrac{mn(m+n-2)}{m+n}}\dfrac{\overline{X}-\overline{Y}}{\sqrt{nS_X^2+mS_Y^2}}$ имеет распределение Стьюдента с числом степеней свободы $(n+m-2)$.

Теорема 2.8.36. Пусть $X_1,...,X_n$ и $Y_1,...,Y_m$ – две независимые выборки из распределений $N(a_1,\sigma_1^2)$ и $N(a_2,\sigma_2^2)$. Далее, пусть S_X^2 и S_Y^2 – соответствующие выборочные дисперсии. Тогда случайная величина $F = \dfrac{n(m-1)\sigma_2^2 S_X^2}{m(n-1)\sigma_1^2 S_Y^2}$ имеет распределение Фишера с числами степеней свободы $(n-1)$ и $(m-1)$.

Доказательство приведенных теорем может быть найдено в [22, с. 32]. Используемые здесь символы для сокращения записи разных распределений пояснены в П.2.

2.9. Метод моментов

Этот метод применяется для построения точечных оценок параметров в случае, когда статистическая модель задана в параметрической форме, то есть в виде (2.1.1). Метод моментов относится к числу первых общих методов оценивания параметров, он был пред-

ложен К. Пирсоном еще в 1894 году. Суть этого метода состоит в приравнивании выборочных оценок моментов к соответствующим теоретическим моментам заданного распределения вероятностей, которое зависит от некоторого числа неизвестных параметров. Выбрав количество моментов, равное числу неизвестных параметров, и решая полученную систему уравнений относительно неизвестных параметров, получим оценки этих параметров, построенные методом моментов, которые коротко называют ММ-оценками.

Итак, пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F(x;\theta_1,...,\theta_r)$, которая задана с точностью до конечного числа r неизвестных параметров $\theta_1,...,\theta_r$. Предполагаем, что начальные моменты $\alpha_k(F)$ k-го порядка, $k=1,...,r$, ф.р. $F(x;\theta_1,...,\theta_r)$ существуют. Отметим, что эти моменты являются функциями параметров $\theta_1,...,\theta_r$ и вычисляются при известном функциональном виде распределения по формулам

$$\alpha_k(\theta_1,...,\theta_r)=\int_{-\infty}^{+\infty}x^k dF(x;\theta_1,...,\theta_r),\ k=1,...,r. \qquad \textbf{(2.9.1)}$$

Выборочные оценки моментов $\alpha_k(F)$, построенные по выборке $X_1,...,X_n$ методом подстановки, записываются в виде

$$\hat{\alpha}_k=\alpha_k(F_n)=\int_{-\infty}^{+\infty}x^k dF_n(x)=n^{-1}\sum_{i=1}^{n}X_i^k,\ k=1,...,r. \qquad \textbf{(2.9.2)}$$

Приравнивая теоретические и выборочные моменты, получим систему уравнений

$$\alpha_k(\theta_1,...,\theta_r)=\hat{\alpha}_k,\ k=1,...,r. \qquad \textbf{(2.9.3)}$$

Решение этой системы относительно $\theta_1,...,\theta_r$ дает ММ-оценки $\hat{\theta}_1,...,\hat{\theta}_r$ этих параметров.

Замечание 2.9.4. Отметим, что для реализации метода моментов требуется выбрать нужное число теоретических моментов. Это число должно равняться числу неизвестных параметров. Выбор нужного числа теоретических моментов не является однозначным, что приводит и к неоднозначности ММ-оценок. Обычно выбирают такие моменты, которые при заданном распределении несложно вычисляются и обеспечивают простоту решения системы

уравнений (2.9.3). Итак, существенным недостатком метода моментов является неоднозначность ММ-оценок.

Приведем пример. Пусть $L(X) = Пу(\lambda)$, это предположение означает, что дискретная с.в. X имеет распределение Пуассона с неизвестным параметром λ, то есть она пробегает значения $k = 0,1,2,...$, и их вероятности вычисляются по формуле $P(X = k) = \lambda^k e^{-\lambda} / k!$. В данном случае $M(X) = \lambda$ и $D(X) = \lambda$, то есть математическое ожидание и дисперсия совпадают. Поэтому оценка неизвестного параметра λ, построенная методом моментов по выборке $X_1,...,X_n$, может привести либо к выборочному среднему $\overline{X}$, либо к выборочной дисперсии S^2. Отметим также, что метод моментов может привести и к неразумной оценке (см. примеры в работе [5, с. 105]). Среди общих свойств ММ-оценок отметим, что в общем случае они не являются эффективными оценками, однако они являются:

а) состоятельными,

б) асимптотически несмещенными,

в) асимптотически нормальными оценками.

Эти свойства ММ-оценок следуют из аналогичных свойств выборочных оценок $\alpha_k(F_n)$ моментов $\alpha_k(F)$, $k = 1,2,...$, построенных методом подстановки. Например, состоятельность ММ-оценок устанавливается следующим образом. Перепишем систему уравнений (2.9.1) в виде $\alpha_k = \varphi_k(\theta_1,...,\theta_r)$, $k = 1,...,r$. Предположим, что соответствие между параметрами $\theta_1,...,\theta_r$ и моментами $\alpha_1,...,\alpha_r$, определяемое функциями $\varphi_1,...,\varphi_r$, взаимно однозначное и взаимно непрерывное, то есть существуют непрерывные функции $\psi_1,...,\psi_r$ такие, что $\theta_i = \psi_i(\alpha_1,...,\alpha_r)$, $i = 1,...,r$. Тогда решение системы (2.9.3) запишется в виде $\hat{\theta}_i = \psi_i(\hat{\alpha}_1,...,\hat{\alpha}_r)$, $i = 1,...,r$. Таким образом, ММ-оценки, являющиеся решением системы (2.9.3), по существу являются оценками, построенными методом подстановки путем замены в выражениях $\theta_i = \psi_i(\alpha_1,...,\alpha_r)$, $i = 1,...,r$, теоретических моментов $\alpha_1,...,\alpha_r$ их выборочными оценками $\hat{\alpha}_1,...,\hat{\alpha}_r$. Учитывая, что выборочные начальные моменты $\hat{\alpha}_k$ сходятся по вероятности к теоретическим моментам α_k, то есть $\hat{\alpha}_k \to^p \alpha_k$, $n \to \infty$, $k = 1,...,r$, согласно (П.3.8), заключаем, что $\hat{\theta}_i = \psi_i(\hat{\alpha}_1,...,\hat{\alpha}_r)$ сходятся по вероятности к $\theta_i = \psi_i(\alpha_1,...,\alpha_r)$, $i = 1,...,r$, то есть ММ-оценки являются состоятельными (см. также теорему (4.4.1) в работе [5, т.1, с.144]). Однако отметим еще раз, что в общем случае ММ-оценки не являются эффективными (см. пример 18.18 в [24, с. 98]). В вычислительном отношении ММ-оценки обычно получить проще, чем найти оценки другими методами. В частности, метод максимального правдоподобия, обсуждаемый ниже, часто приводит к необходимости решения нелинейных уравнений путем реализации итерацион-

ной процедуры, для которой в качестве начального приближения используют предварительно вычисленные состоятельные ММ-оценки. Асимптотическая нормальность ММ-оценок доказывается методом Мизеса в примере 7.2.6.

Пример 2.9.5. Пусть с.в. X имеет экспоненциальное распределение с параметром $\lambda > 0$, то есть $L(X) = \mathrm{Exp}(\lambda)$. Это предположение означает, что ф.р. $F_X(x,\lambda) = 1 - e^{-\lambda x}$, $0 \le x < \infty$. Имеется выборка $X_1,...,X_n$ из распределения $F_X(x,\lambda)$, требуется построить ММ-оценку неизвестного параметра λ. В данном случае имеется один неизвестный параметр λ, поэтому достаточно использовать лишь начальный момент первого порядка

$$\alpha_1(\lambda) = M(X) = \int_0^{+\infty} x d(1 - e^{-\lambda x}) = \lambda^{-1}.$$

Далее, выборочная оценка $\hat{\alpha}_1$ момента $\alpha_1(\lambda)$ равняется выборочному среднему $\overline{X}$, то есть $\hat{\alpha}_1 = \overline{X}$ и, следовательно, формула (2.9.3) в данном случае записывается в виде $\lambda^{-1} = \overline{X}$. Отсюда получаем, что ММ-оценка параметра λ экспоненциального распределения равняется обратной величине от выборочного среднего, то есть $\hat{\lambda} = 1/\overline{X}$.

Пример 2.9.6. Пусть $L(X) = R(a,b)$. Эта запись означает, что с.в. X имеет равномерное распределение в интервале $[a,b]$, то есть плотность $f_X(x;a,b) = 1/(b-a)$ для $x \in [a,b]$ и она равна нулю в других случаях. Требуется по выборке $X_1,...,X_n$ построить ММ-оценки параметров a и b. В данном случае, начальный момент первого порядка $\alpha_1 = M(X) = (a+b)/2$, начальный момент второго порядка $\alpha_2 = M(X^2) = (a^2 + ab + b^2)/3$ и центральный момент второго порядка $\mu_2 = \alpha_2 - \alpha_1^2 = D(X) = (b-a)^2/12$. Выборочные оценки этих моментов записываются в виде

$$\hat{\alpha}_1 = \overline{X} = n^{-1}\sum_{i=1}^n X_i,\ \hat{\alpha}_2 = n^{-1}\sum_{i=1}^n X_i^2,\ \hat{\mu}_2 = S^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2 = n^{-1}\sum_{i=1}^n X_i^2 - \overline{X}^2.$$

В данном примере имеется два неизвестных параметра a и b, поэтому нужно выбрать лишь два теоретических момента. Из соображений простоты решения системы уравнений выберем α_1 и μ_2. В результате система (2.9.3) запишется в виде

$$(a+b)/2 = \overline{X},$$

$$(b-a)^2/12 = S^2.$$

Решением этой системы являются ММ-оценки параметров a и b в виде $\hat{a} = \overline{X} - S\sqrt{3}$, $\hat{b} = \overline{X} + S\sqrt{3}$.

Рассмотрим теперь другой вариант задания равномерного распределения. Пусть $L(X) = R(\theta_1 - \theta_2; \theta_1 + \theta_2)$. В этом случае неизвестные параметры θ_1 и θ_2 этого распределения непосредственно выражаются через $\alpha_1 = M(X)$ и $\mu_2 = D(X)$. В самом деле, поскольку $\theta_1 - \theta_2 = a$, $\theta_1 + \theta_2 = b$, то $\theta_1 = (a+b)/2 = M(X)$ и $\theta_2 = (b-a)/2 = \sqrt{3(b-a)^2/12} = \sqrt{3D(X)}$.

Таким образом, ММ-оценки параметров θ_1 и θ_2 запишутся в виде $\hat{\theta}_1 = \overline{X}$, $\hat{\theta}_2 = S\sqrt{3}$. Отметим, что выбор в качестве теоретических моментов первых двух моментов α_1 и α_2 приводит к таким же оценкам параметров θ_1 и θ_2. В самом деле, в данном случае

$$\alpha_2 = M(X^2) = [(\theta_1 - \theta_2)^2 + (\theta_1 - \theta_2)(\theta_1 + \theta_2) + (\theta_1 + \theta_2)]/3 = (3\theta_1^2 + \theta_2^2)/3$$

и, следовательно, система (2.9.3) запишется в виде

$$\theta_1 = n^{-1}\sum_{i=1}^{n} X_i,\ (3\theta_1^2 + \theta_2^2)/3 = n^{-1}\sum_{i=1}^{n} X_i^2.$$

Решением данной системы являются ММ-оценки параметров θ_1 и θ_2 в виде $\hat{\theta}_1 = \overline{X}$, $\hat{\theta}_2 = [3(n^{-1}\sum X_i^2 - \overline{X}^2)]^{1/2} = S\sqrt{3}$.

Пример 2.9.7. Пусть $L(X) = \Gamma(\alpha, \lambda)$, то есть с.в. X имеет гамма-распределение с плотностью

$$f(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ 0 \le x < \infty, \qquad \textbf{(2.9.8)}$$

где $\Gamma(\alpha)$ – гамма-функция, определяемая в виде

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt. \qquad \textbf{(2.9.9)}$$

Отметим, что плотность $f(x;\alpha,\lambda)$ характеризуется двумя параметрами: параметром формы распределения $\alpha > 0$ и масштабным параметром $\lambda > 0$. Требуется по выборке $X_1, ..., X_n$ построить ММ-оценки параметров α и λ.

В данном случае начальные моменты *k*-го порядка, $k = 1, 2, ...$, вычисляются по формуле

$$\alpha_k(\alpha,\lambda) = M(X^k) = \int_0^{+\infty} x^k f(x;\alpha,\lambda) dx = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^{+\infty} x^{k+\alpha-1} e^{-\lambda x} dx =$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)\lambda^{k+\alpha-1} \cdot \lambda} \int_0^{+\infty} t^{k+\alpha-1} e^{-t} dt = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)\lambda^k}. \qquad \textbf{(2.9.10)}$$

Отсюда, при $k=1$ имеем

$$\alpha_1(\alpha,\lambda)=M(X)=\frac{\Gamma(\alpha+1)}{\lambda\cdot\Gamma(\alpha)}=\frac{\alpha\cdot\Gamma(\alpha)}{\lambda\cdot\Gamma(\alpha)}=\frac{\alpha}{\lambda},$$

а при $k=2$ получаем

$$\alpha_2(\alpha,\lambda)=M(X^2)=\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2}=\frac{(\alpha+1)\cdot\alpha\cdot\Gamma(\alpha)}{\lambda^2\cdot\Gamma(\alpha)}=\frac{(\alpha+1)\alpha}{\lambda^2}.$$

Далее, центральный момент второго порядка μ_2 равен

$$\mu_2=\alpha_2-\alpha_1^2=D(X)=\frac{\alpha^2+\alpha}{\lambda^2}-\frac{\alpha^2}{\lambda^2}=\frac{\alpha}{\lambda^2}.$$

Выбор в качестве теоретических моментов α_1 и μ_2 приводит запись системы (2.7.21) к виду

$$\alpha/\lambda=\overline{X},\ \alpha/\lambda^2=S^2.$$

Решением данной системы являются ММ-оценки параметров α и λ в виде

$$\hat{\alpha}=\overline{X}^2/S^2,\ \hat{\lambda}=\overline{X}/S^2. \qquad \textbf{(2.9.11)}$$

Отметим, что выбор в качестве теоретических моментов первых двух моментов α_1 и α_2 приводит, как и в предыдущем примере, к таким же ММ-оценкам параметров α и λ. В самом деле, при таком выборе теоретических моментов система уравнений (2.7.21) запишется в виде

$$\alpha/\lambda=\overline{X},\ \alpha(\alpha+1)/\lambda^2=n^{-1}\sum X_i^2.$$

Для решения этой системы разделим второе уравнение на квадрат первого, в результате получим

$$\frac{\alpha+1}{\alpha}=\frac{n^{-1}\sum X_i^2}{\overline{X}^2}.$$

Отсюда следует, что ММ-оценки параметров α и λ имеют вид

$$\hat{\alpha}=\frac{\overline{X}^2}{n^{-1}\sum X_i^2-\overline{X}^2}=\frac{\overline{X}^2}{S^2},\ \hat{\lambda}=\frac{\hat{\alpha}}{\overline{X}}=\frac{\overline{X}^2}{S^2\overline{X}}=\frac{\overline{X}}{S^2}.$$

Пример 2.9.12. Пусть $L(X)=N(\mu,\sigma^2)$. Требуется по выборке $X_1,...,X_n$ построить ММ-оценки параметров μ и σ^2. В данном случае $\alpha_1=M(X)=\mu$ и $\alpha_2=M(X^2)=\sigma^2+\mu^2$. Следовательно, система уравнений (2.7.21) запишется в виде

$$\mu=\overline{X}=n^{-1}\sum X_i\ ,\ \sigma^2+\mu^2=n^{-1}\sum X_i^2.$$

Решением этой системы являются ММ-оценки параметров μ и σ^2 в виде $\hat{\mu} = \bar{X}$ и $\hat{\sigma}^2 = n^{-1}\sum X_i^2 - \bar{X}^2 = n^{-1}\sum (X_i - \bar{X})^2 = S^2$.

2.10. Метод максимального правдоподобия

Этот метод, как и метод моментов, применяется для оценивания параметров статистических моделей, заданных в параметрической форме. Метод максимального правдоподобия основан на использовании понятия функции правдоподобия, которая определена в разд. 2.3. Оценки, построенные методом максимального правдоподобия, обладают многими «хорошими» свойствами и широко применяются на практике.

Итак, пусть статистическая модель задана в параметрической форме, то есть в виде (2.1.1). Это означает, что имеется выборка $X_1,...,X_n$ из распределения $F_X(x,\theta)$, функциональный вид которого известен, то есть предполагаем, что $F_X(x,\theta) \in \Im_\theta$ и пусть плотность $f_X(x,\theta)$, $x \in R^1$, $\theta \in \Theta$, непрерывна и трижды дифференцируема. Требуется, используя данную информацию, построить оценку параметра $\theta \in \Theta$ по выборке $X_1,...,X_n$. Оценку максимального правдоподобия параметра $\theta \in \Theta$ обозначим через $\hat{\theta}$ и коротко будем называть МП-оценкой. Напомним (см. определение 2.3.1), что для заданной параметрической модели $F_X(x,\theta) \in \Im_\theta$, $\theta \in \Theta$, функция правдоподобия $L(\vec{x};\theta)$ при конкретной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ является функцией только параметра $\theta \in \Theta$, которую часто для сокращения обозначений записывают в виде $L(\theta) = L(\vec{x};\theta)$. Функция правдоподобия $L(\vec{x};\theta)$ строится путем замены аргументов $x_1,...,x_n$ совместной плотности $f_{X_1,...,X_n}(x_1,...,x_n;\theta)$ случайных величин $X_1,...,X_n$ на реализацию $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$. Отметим, что функция правдоподобия $L(x_1,...,x_n;\theta) = L(\vec{x};\theta)$ является совместной плотностью распределения вероятностей случайных величин $X_1,...,X_n$, которая при фиксированной наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ рассматривается как функция параметра θ, поэтому ее

можно рассматривать в качестве меры того, насколько хорошо оценка параметра θ «объясняет» наблюденную реализацию $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$.

Определение 2.10.1. Оценкой максимального правдоподобия (МП-оценкой $\hat{\theta}$) параметра $\theta \in \Theta$ называют случайную величину $\hat{\theta}(X_1,...,X_n)$, для которой ее конкретное значение $\hat{\theta}(x_1,...,x_n)$ при заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ соответствует такой точке параметрического множества Θ, в которой совместная плотность $f_{X_1,...,X_n}(x_1,...,x_n;\theta)$ и функция правдоподобия $L(\vec{x};\theta)$ достигают максимума (см. рис. 2.10.1).

Итак, согласно определению, для МП-оценки $\hat{\theta}$ параметра $\theta \in \Theta$ выполняются следующее неравенство $L(\vec{x};\hat{\theta}) \geq L(\vec{x};\theta)$, $\forall \theta \in \Theta$, или выражение вида

$$L(\vec{x};\hat{\theta}) = \sup_{\theta \in \Theta} L(\vec{x};\theta). \qquad \textbf{(2.10.2)}$$

В тех случаях, когда функция правдоподобия $L(\vec{x};\theta)$ дифференцируема по θ и для каждой реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ ее максимум по θ достигается не на границе множества Θ, то МП-оценка (точнее, ее значение, вычисленное по наблюденной реализации $\vec{x} = (x_1,...,x_n)$) находится в результате решения уравнения

$$\frac{\partial L(\vec{x};\theta)}{\partial \theta} = 0. \qquad \textbf{(2.10.3)}$$

Отметим, что во многих случаях удобнее использовать не саму функцию правдоподобия $L(\vec{x};\theta)$, а ее логарифм, и поскольку точка максимума при этом не меняется, то обычно вместо уравнения (2.10.3) используют уравнение правдоподобия вида

$$\frac{\partial \ln L(\vec{x};\theta)}{\partial \theta} = 0, \text{ или } \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(x_i, T_n) = 0. \qquad \textbf{(2.10.4)}$$

Если параметр θ является векторным параметром размерности r, то есть $\vec{\theta} = (\theta_1,...,\theta_r)$, то для нахождения МП-оценок параметров

$\theta_1,...,\theta_r$ решают следующую систему уравнений:

$$\frac{\partial \ln L(\vec{x};\vec{\theta})}{\partial \theta_i} = 0,\ i = 1,...,r. \qquad \textbf{(2.10.5)}$$

Отметим еще раз, что конкретное значение МП-оценки $\hat{\theta}(x_1,...,x_n)$, вычисленное при заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ как решение уравнения (2.10.4), является реализацией случайной величины $\hat{\theta}(X_1,...,X_n)$, которая и является оценкой максимального правдоподобия параметра θ и для которой уравнение правдоподобия (2.10.4) с использованием функции вклада $U(\vec{X},\theta)$, определенной в (2.3.4), записывается в виде

$$U(\vec{x},\theta) = 0. \qquad \textbf{(2.10.6)}$$

2.10.А. Общие свойства МП-оценок

Оценки, построенные методом максимального правдоподобия, широко применяются на практике, поскольку они обладают привлекательными общими свойствами, которые формулируются следующим образом:

1. Если для параметра $\theta \in \Theta$ параметрической модели $\Im_\theta$ существует эффективная оценка $T^*(X_1,...,X_n)$, то МП-оценка $\hat{\theta}$ параметра θ совпадает с этой эффективной оценкой, то есть $\hat{\theta} = T^*(X_1,...,X_n)$. Справедливость этого свойства непосредственно следует из (2.10.4) с учетом критерия эффективности оценок (2.4.17), согласно которому $\partial \ln L(\vec{x};\theta)/\partial\theta = A(\theta)[T^*(x_1,...,x_n) - \theta]$.

2. В тех случаях, когда для параметра $\theta \in \Theta$ параметрической модели $\Im_\theta$ имеется достаточная статистика $T(X_1,...,X_n)$, то МП-оценка $\hat{\theta}$ параметра θ, если она существует и единственна, является функцией от достаточной статистики $T(X_1,...,X_n)$. Справедливость этого свойства непосредственно следует из критерия факторизации (2.5.17), согласно которому функция правдоподобия записывается в виде $L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x})$, и, следовательно, ее мак-

симизация по θ сводится к максимизации функции $\varphi(T(\vec{x}),\theta)$ по θ. А это и означает, что конкретное значение $\hat{\theta}(x_1,...,x_n)$ МП-оценки параметра θ будет функцией от $T(\vec{x})$, то есть $\hat{\theta}(x_1,...,x_n)=\psi\{T(\vec{x})\}$.

3. К общим полезным свойствам МП-оценок относится также их инвариантность к виду параметризации семейства распределений $\Im_\theta$. Это свойство проявляется в том, что оценка максимального правдоподобия некоторой непрерывной и монотонной функции φ от параметра θ равна этой функции от МП-оценки параметра θ, то есть $\hat{\varphi}(\theta)=\varphi(\hat{\theta})$ (см., например, [22, с. 68]).

4. Отметим, что МП-оценки могут быть представлены в виде функционала от эмпирической функции распределения $F_n(x)$, то есть в виде $\hat{\theta}=T(F_n)$, причем функционал $T(F)$ задается неявно с помощью выражения

$$\int_{-\infty}^{\infty}\psi(x,T(F))dF(x,\theta)=0,$$

в котором функция $\psi(x,\theta)$ согласована с плотностью $f(x,\theta)$ соотношением $\psi(x,\theta)=\partial\ln f(x,\theta)/\partial\theta$. В самом деле,

$$\int_{-\infty}^{\infty}\psi(x,T(F_n))dF_n(x)=\frac{1}{n}\sum_{i=1}^{n}\psi(x_i,T_n)=0,$$

а при $\psi(x,\theta)=\partial\ln f(x,\theta)/\partial\theta$ последнее равенство переписывается в виде

$$\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(x_i,T_n)=0.$$

Это равенство соответствует (2.10.4) для случая, когда $X_1,...,X_n$ – н.о.р. случайные величины, так как логарифм функции правдоподобия при этих предположениях записывается в виде

$$\ln L(\vec{x};\theta)=\sum_{i=1}^{n}\ln f(x_i,\theta).$$

Отметим, что наиболее полно исследованы асимптотические свойства МП-оценок, то есть в случае, когда объем выборки $n\to\infty$.

Асимптотические свойства МП-оценок обсуждаются ниже. Здесь же приведем примеры вычисления оценок максимального правдоподобия параметров для некоторых параметрических моделей.

Пример 2.10.7. А) Пусть $L(X)=N(\mu,\sigma^2)$. Имеется выборка $\vec{X}_1=(X_1,...,X_n)$ из распределения $F_X(x)=\Phi\{(x-\mu)/\sigma^2\}$, где $\Phi(\cdot)$ – стандартная нормальная функция распределения и $\mu=M(X)$, $\sigma^2=D(X)$ – неизвестные параметры. В данном случае параметрическая модель $\Im_\theta$ характеризует двухпараметрическое семейство гауссовских распределений, заданных с точностью до неизвестного векторного параметра $\vec{\theta}=(\theta_1,\theta_2)$, компоненты которого $\theta_1=\mu$ и $\theta_2=\sigma^2$ требуется оценить методом максимального правдоподобия по информации, содержащейся в выборке $\vec{X}_1=(X_1,...,X_n)$.

Для построения МП-оценок параметров μ и σ^2 запишем выражение функции правдоподобия. Согласно (2.3.2), с учетом того, что плотность с.в. X имеет вид $f_X(x;\mu,\sigma^2)=(2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/2\sigma^2\}$, функция правдоподобия запишется в виде

$$L(\vec{x};\mu,\sigma^2)=\prod_{i=1}^{n}f_X(x_i;\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\}. \quad \textbf{(2.10.8)}$$

Логарифмическая функция правдоподобия имеет вид

$$\ln L(\vec{x};\mu,\sigma^2)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Далее, система уравнений правдоподобия (2.10.5) в данном случае запишется в виде

$$\frac{\partial\ln L(\vec{x};\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0,$$

$$\frac{\partial\ln L(\vec{x};\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0.$$

Решением данной системы являются МП-оценки параметров μ и σ^2 в виде

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2. \quad \textbf{(2.10.9)}$$

Отметим, что найденные оценки максимального правдоподобия $\hat{\mu}$ и $\hat{\sigma}^2$ вида (2.10.9) для параметров μ и σ^2 в гауссовском случае совпадают с оценками этих параметров, построенных методом моментов (см. пример 2.9.12), а также

они совпадают с выборочными оценками параметров μ и σ^2, построенными методом подстановки (см. пример 2.8.3). Заметим также, что оценка максимального правдоподобия $\hat{\sigma}^2$ параметра σ^2 является смещенной оценкой, однако она асимптотически несмещенная оценка параметра σ^2 (см. упр. 2.15.13).

Б) Пусть $L(X) = \exp(\lambda)$. Эта запись означает, что с.в. X имеет экспоненциальное распределение с плотностью $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Пусть $X_1,...,X_n$ – выборка из экспоненциального распределения. Требуется оценить неизвестный параметр λ методом максимального правдоподобия по информации, содержащейся в выборке $\vec{X}_1 = (X_1,...,X_n)$. Для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия запишется в виде $L(\vec{x},\lambda) = \lambda^n \exp\{-\lambda \sum x_i\}$. Далее, $\ln L(\vec{x},\lambda) = n\ln\lambda - \lambda\sum x_i$ и уравнение правдоподобия запишется в виде

$$\frac{\partial \ln L(\vec{x},\lambda)}{\partial\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

Отсюда получаем $\hat{\lambda} = 1/n^{-1}\sum x_i$. Таким образом, оценка максимального правдоподобия параметра λ экспоненциального распределения равна обратной величине выборочного среднего, то есть $\hat{\lambda} = 1/\overline{X}$.

В) Пусть X – дискретная случайная величина с распределением Пуассона, то есть $L(X) = \Pi y(\theta)$. Эта запись означает, что с.в. X пробегает значения $x = 0,1,2,...$ с вероятностями

$$f_X(x,\theta) = P_\theta(X = x) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0,1,2,...,\ \theta > 0.$$

Пусть $X_1,...,X_n$ – выборка из распределения $\Pi y(\theta)$. Убедимся, что в данном случае эффективная оценка для параметра θ, равная $T^*(X_1,...,X_n) = n^{-1}\sum_{i=1}^{n} X_i$, (см. пример 2.4.22), совпадает с оценкой максимального правдоподобия этого параметра. В самом деле, для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия запишется в виде

$$L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{(x_i)!} = e^{-n\theta}\theta^{\sum x_i} \prod_{i=1}^{n} \frac{1}{(x_i)!}.$$

Отсюда имеем $\ln L(\vec{x};\theta) = -n\theta + \ln\theta\sum x_i - \sum\ln(x_i)!$ и уравнение правдоподобия запишется в виде

$$\frac{\partial \ln L(\vec{x};\theta)}{\partial\theta} = -n + \theta^{-1}\sum_{i=1}^{n} x_i = 0.$$

Решение данного уравнения дает МП-оценку параметра θ в виде $\hat{\theta} = \overline{X}$.

Г) Пусть $L(X) = Bi(n, p)$, где $p = P(A)$ – вероятность некоторого события A – «успех» – является неизвестным параметром биномиального распределения, то есть $\theta = p$. Проводится n независимых опытов, в которых «успех» наступил m раз. Требуется построить МП-оценку параметра $\theta = p$. В данном случае функция правдоподобия запишется в виде $L(m,\theta) = \theta^m (1-\theta)^{n-m}$ и $\ln L(m,\theta) = m\ln\theta + (n-m)\ln(1-\theta)$. Далее, $\partial \ln L(m,\theta)/\partial\theta = m/\theta - (n-m)/(1-\theta) = (m-n\theta)/\theta(1-\theta)$, и, следовательно, уравнение правдоподобия запишется в виде $m - n\theta = 0$. Отсюда получаем МП-оценку параметра $\theta = p$ в виде относительной частоты появления события A в n независимых опытах, то есть $\hat{\theta} = (m/n)$. Отметим, что $\hat{\theta} = (m/n)$ также является и эффективной оценкой параметра $p = P(A)$ биномиального распределения (см. примеры 2.3.17 и 2.4.20 и рис. 2.10.1, на котором приведены зависимости биномиального распределения $\tilde{L}(m,\theta) = C_n^m L(m,\theta)$ от параметра θ при заданных значениях n и m).

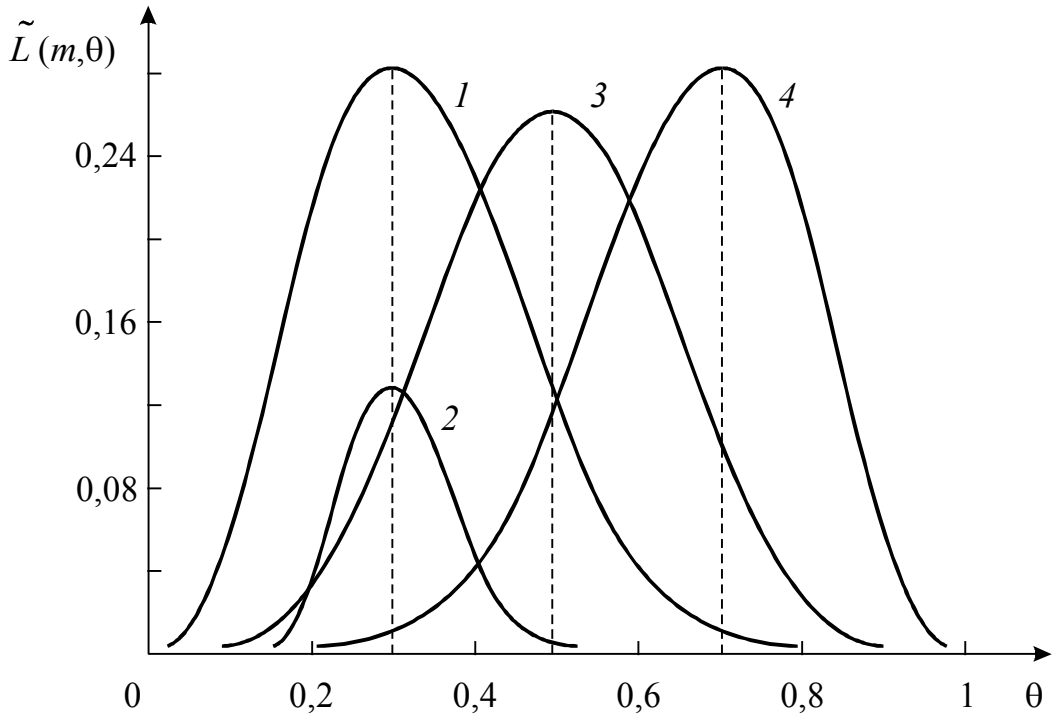

Рис. 2.10.1. Зависимости биномиального распределения от параметра θ при заданных n и m: кр. *1* соответствует $n = 10,\ m = 3$; кр. *2* – $n = 50,\ m = 15$; кр. *3* – $n = 10,\ m = 5$; кр. *4* – $n = 10,\ m = 7$

Д) Пусть $L(X) = R(0,\theta)$. Эта запись означает, что с.в. X имеет равномерное распределение в интервале $[0,\theta]$. Имеется выборка $X_1,...,X_n$ из равномерного распределения в интервале $[0,\theta]$. Требуется оценить параметр θ методом

максимального правдоподобия. Для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия (см. формулу (2.5.12)) запишется в виде

$$L(\vec{x},\theta) = \theta^{-n} C(\theta - x_{(n)}) C(x_i),\ i = 1,...,n,$$

где $C(\cdot)$ – единичная функция Хевисайда и $x_{(n)}$ – реализация n-й порядковой статистики $X_{(n)}$. Отметим, что данная функция правдоподобия $L(\vec{x},\theta)$ монотонно убывает по θ для $\theta \geq x_{(n)}$, является разрывной функцией и ее производная в точке $\theta = x_{(n)}$ не существует. Функция правдоподобия $L(\vec{x},\theta)$ достигает максимального значения при $\theta = x_{(n)}$. Таким образом, в данном случае МП-оценкой параметра θ является n-я порядковая статистика, то есть $\hat{\theta} = X_{(n)}$. Отметим, что она совпадает с достаточной статистикой (см. пример 2.5.11) и не является решением уравнения правдоподобия, так как функция правдоподобия $L(\vec{x},\theta)$ в данном случае не дифференцируема.

Пример 2.10.10. Пусть $L(X,Y) = NN(a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Эта запись означает, что двумерная случайная величина $Z = (X,Y)$ имеет двумерное нормальное распределение вероятностей с плотностью

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-a_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-a_X}{\sigma_X} \right) \left(\frac{y-a_Y}{\sigma_Y} \right) + \left(\frac{y-a_Y}{\sigma_Y} \right)^2 \right] \right\},$$

где $a_X = M(X)$, $a_Y = M(Y)$, $\sigma_X^2 = D(X)$, $\sigma_Y^2 = D(Y)$ и ρ – коэффициент корреляции. Имеется двумерная выборка $(X_1,Y_1),...,(X_n,Y_n)$ объема n. Требуется построить МП-оценки параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ по информации, содержащейся в выборке $(X_1,Y_1),...,(X_n,Y_n)$. Для заданной реализации $\vec{z} = (x_1,y_1),...,(x_n,y_n)$ двумерной выборки $(X_1,Y_1),...,(X_n,Y_n)$ функция правдоподобия запишется в виде

$$L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho) = \prod_{i=1}^{n} f_{XY}(x_i,y_i) = \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \right)^n \times$$

$$\times \exp\left\{ \frac{1}{2(1-\rho^2)} \left[\frac{1}{\sigma_X^2} \sum_{i=1}^{n} (x_i - a_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y} \sum_{i=1}^{n} (x_i - a_X)(y_i - a_Y) + \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (y_i - a_Y)^2 \right] \right\}.$$

Логарифмическая функция правдоподобия

$$\ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho) = -n\ln(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}) -$$
$$-\left\{\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i-a_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y}\sum_{i=1}^{n}(x_i-a_X)(y_i-a_Y) + \frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i-a_Y)^2\right]\right\}.$$

Система уравнений (2.10.5), записанная для данного примера, и ее решение в виде оценок максимального правдоподобия параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ приводятся в упр. 2.15.10. Отметим, что полученные оценки максимального правдоподобия параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ совпадают с выборочными оценками этих параметров, полученными методом подстановки (см. упр. 2.15.11– 2.15.15).

2.10.Б. Метод накопления для приближенного вычисления МП-оценок

Для нахождения оценок максимального правдоподобия $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ параметра θ параметрического семейства $\Im_\theta$ требуется решить уравнение правдоподобия (2.10.4) в скалярном случае либо систему уравнений (2.10.5) в векторном случае. Однако не всегда удается получить решение в явном виде. Тогда используют приближенные методы решений уравнений правдоподобия. Один из таких методов, предложенных Фишером, называют методом накопления. Этот метод приводит к реализации рекуррентной процедуры. Для простоты рассмотрим случай скалярного параметра θ. Отметим, что функция вклада $U(\vec{X},\theta)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ при заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ является функцией только параметра θ, то есть $U(\theta) = U(\vec{x},\theta)$. Предполагаем, что эту функцию можно разложить в ряд Тейлора. При сделанных предположениях дифференцируемости плотности $f_X(x,\theta)$, $x \in R^1$, $\theta \in \Theta$, это предположение выполняется. Выберем в качестве начального приближения МП-оценки $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ некоторую состоятельную оценку $\hat{\theta}_0$ параметра $\theta \in \Theta$. Например, в качестве такой оценки может быть выбрана ММ-оценка, которая обычно просто вычисляется по заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$.

Разложим теперь в ряд Тейлора правую часть уравнения правдоподобия $0 = U(\theta)$ в точке $\hat{\theta}_0$ и вычислим это разложение при $\theta = \hat{\theta}$. В результате получим

$$0 = U(\hat{\theta}_0) + (\hat{\theta} - \hat{\theta}_0)U'(\theta^*), \qquad \textbf{(2.10.11)}$$

где θ^* – некоторая точка между $\hat{\theta}$ и $\hat{\theta}_0$. Отметим, что при больших объемах выборки n все три точки $\hat{\theta}$, θ^* и $\hat{\theta}_0$ будут находиться вблизи истинного значения параметра θ, поэтому для приближенного вычисления МП-оценки в полученном выражении (2.10.11) можно заменить θ^* на $\hat{\theta}_0$ и значение $U'(\hat{\theta}_0) = U'(\vec{x}, \hat{\theta}_0)$ можно заменить средним значением. Напомним, что это среднее значение, согласно (2.3.13) и (2.3.14), равно с минусом количеству информации Фишера о параметре θ, содержащемуся в выборке $\vec{X}_1 = (X_1, ..., X_n)$, то есть

$$M[U'(\vec{X}, \hat{\theta}_0)] = M\left[\frac{\partial^2 \ln L(\vec{X}, \theta)}{\partial \theta^2}\Big|_{\theta=\hat{\theta}_0}\right] =$$

$$= -i_n(\hat{\theta}_0) = -n \cdot i_1(\hat{\theta}_0). \qquad \textbf{(2.10.12)}$$

В результате, из (2.10.11) получим первое приближение $\hat{\theta}_1$ для вычисления МП-оценки $\hat{\theta}$ в виде

$$\hat{\theta}_1 = \hat{\theta}_0 + U(\hat{\theta}_0)/n \cdot i_1(\hat{\theta}_0). \qquad \textbf{(2.10.13)}$$

Для вычисления второго приближения $\hat{\theta}_2$ описанную процедуру повторяют, выбрав в качестве начального приближения вычисленное по формуле (2.10.13) значение $\hat{\theta}_1$, и т.д. Таким образом, $(k+1)$-е приближение $\hat{\theta}_{k+1}$ по методу накопления вычисляют с помощью рекуррентной процедуры вида

$$\hat{\theta}_{k+1} = \hat{\theta}_k + U(\hat{\theta}_k)/n \cdot i_1(\hat{\theta}_k),\ k = 0,1,2,\ldots. \qquad \textbf{(2.10.14)}$$

Остановка вычислений по формуле (2.10.14) осуществляется по достижению заданной точности ε на $(k+1)$-м шаге при выполнении неравенства $|\hat{\theta}_{k+1} - \hat{\theta}_k| < \varepsilon$.

Пример 2.10.15. Пусть с.в. X имеет распределение Коши, то есть ф.р. имеет вид

$$F_X(x;\theta)=\frac{1}{2}+\frac{1}{\pi}\cdot\operatorname{arctg}(x-\theta),\ -\infty<x<+\infty,\ \theta\in R^1,$$

и плотность $f_X(x;\theta)$ равна

$$f_X(x;\theta)=\frac{1}{\pi}\cdot\frac{1}{1+(x-\theta)^2},\ -\infty<x<+\infty.$$

Требуется по выборке $\vec{X}_1=(X_1,...,X_n)$ из распределения $F_X(x;\theta)$ с плотностью $f_X(x;\theta)$ построить МП-оценку неизвестного параметра $\theta\in R^1$.

Для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}_1=(X_1,...,X_n)$ функция вклада $U(\vec{x},\theta)$ зависит только от параметра θ, то есть $U(\theta)=U(\vec{x},\theta)$. Эта функция, равная логарифмической производной функции правдоподобия, в данном случае определяется в виде

$$U(\theta)=\frac{\partial\ln L(\vec{x};\theta)}{\partial\theta}=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(x_i;\theta)=$$

$$=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\{-\ln\pi-\ln[1+(x-\theta)^2]\}=\sum_{i=1}^{n}\frac{2(x_i-\theta)}{1+(x_i-\theta)^2}.$$

Таким образом, уравнение правдоподобие (2.10.6) в данном случае запишется в виде

$$\sum_{i=1}^{n}\frac{2(x_i-\theta)}{1+(x_i-\theta)^2}=0.$$

Получить явное решение этого уравнения не представляется возможным, поэтому воспользуемся методом накопления для нахождения приближенной МП-оценки параметра θ. Используя формулу (2.10.13), получим первое приближение $\hat{\theta}_1$ для МП-оценки параметра θ. Для этого нам нужно выбрать в качестве начального приближения некоторую состоятельную оценку $\hat{\theta}_0$ параметра θ и вычислить количество информации Фишера $i_1(\theta)$. Для выбора начального приближения $\hat{\theta}_0$ приведем следующие соображения. Отметим, что распределение Коши симметрично относительно точки θ, и, следовательно, эта точка совпадает с медианой с.в. X, то есть $\theta=MED(X)$. Далее, выборочная медиана $\hat{x}_{1/2}=X_{([n/2]+1)}$, где $X_{(i)}$ обозначает i-ю порядковую статистику выборки $\vec{X}_1=(X_1,...,X_n)$, является состоятельной оценкой $\theta=MED(X)$ (см.

разд. 4.4.Б). Таким образом, в качестве начального приближения $\hat{\theta}_0$ выбираем выборочную медиану, то есть $\hat{\theta}_0 = \hat{x}_{1/2} = X_{([n/2]+1)}$. Далее, согласно (2.3.11), количество информации Фишера $i_1(\theta)$ равно

$$i_1(\theta) = M\left[\frac{\partial}{\partial\theta}\ln f(X_1;\theta)\right]^2 = -M\left[\frac{\partial^2}{\partial\theta^2}\ln f(X_1;\theta)\right] = -\int_{-\infty}^{+\infty}\left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}\right] f(x;\theta)dx.$$

Учитывая, что в данном случае

$$\frac{\partial \ln f(x;\theta)}{\partial\theta} = \frac{2(x-\theta)}{1+(x-\theta)^2} \text{ и } \frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2} = \frac{2(x-\theta)^2-2}{\{1+(x-\theta)^2\}^2},$$

получаем

$$i_1(\theta) = -\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{2(x-\theta)^2-2}{\{1+(x-\theta)^2\}^3}dx = -\frac{4}{\pi}\int_{-\infty}^{+\infty}\frac{(x^2-1)dx}{(1+x^2)^3}dx = \left(-\frac{4}{\pi}\right)\cdot\left(-\frac{\pi}{8}\right) = \frac{1}{2}.$$

Таким образом, согласно (2.10.13), первое приближение $\hat{\theta}_1$ для МП-оценки параметра θ записывается в виде

$$\hat{\theta}_1 = \hat{x}_{1/2} + \frac{4}{n}\sum_{i=1}^{n}\left(\frac{x_i - \hat{x}_{1/2}}{1+(x_i-\hat{x}_{1/2})^2}\right). \qquad \textbf{(2.10.16)}$$

Замечание 2.10.17. Отметим, что выборочная медиана $\hat{x}_{1/2} = X_{([n/2]+1)}$ является состоятельной и асимптотически нормальной оценкой параметра θ. Причем ее асимптотическая дисперсия при распределении Коши равна $D(\hat{x}_{1/2}) = 1/4nf^2(0) = \pi^2/4n$. Однако она не является асимптотически эффективной оценкой параметра θ для распределения Коши, так как ее асимптотическая дисперсия не достигает нижней границы в неравенстве Крамера – Рао. В самом деле, поскольку $\pi^2/4 > 2$, то

$$D(\hat{x}_{1/2}) = (\pi^2/4n) > (2/n) = 1/i_n(\theta).$$

В то же время, оценка максимального правдоподобия $\hat{\theta}_1$ является асимптотически эффективной оценкой параметра θ, то есть $\hat{\theta}_1 = \hat{\theta}_{\text{эфф}}$ при $n \to \infty$ и асимптотическая эффективность выборочной медианы $\hat{x}_{1/2} = X_{([n/2]+1)}$ при распределении Коши, согласно (2.6.3), равна

$$A\mathcal{Э}_F(\hat{x}_{1/2}) = AO\mathcal{Э}_F(\hat{x}_{1/2} : \hat{\theta}_{\text{эфф}}) = \frac{D_F(\hat{\theta}_{\text{эфф}})}{D_F(\hat{x}_{1/2})} = \frac{2}{n}\cdot\frac{4n}{\pi^2} = 8/\pi^2 = 0{,}8106. \qquad \textbf{(2.10.18)}$$

2.10.В. Асимптотические свойства МП-оценок

При изучении асимптотических свойств МП-оценок рассматриваются свойства последовательностей оценок при увеличении объема выборки n. Для того чтобы подчеркнуть этот факт, будем приписывать индекс n в обозначении оценки максимального правдоподобия, то есть $\hat{\theta}(X_1,...,X_n)=\hat{\theta}_n$ и также будем приписывать индекс n в обозначении функции вклада $U(\vec{X},\theta)$, то есть $U(\vec{x},\theta)=U(\theta)=$ $=U_n(\theta)$. Метод максимального правдоподобия широко применяется на практике. Одна из основных причин этого состоит в том, что при достаточно больших объемах выборки n, то есть в асимптотике, МП-оценки обладают многими «хорошими» свойствами. Отметим, что при изучении асимптотических свойств формально полагается, что объем выборки $n\to\infty$. На практике же объем выборки всегда конечен, однако во многих случаях асимптотические результаты являются вполне приемлемой аппроксимацией при конечных n. При выполнении достаточно общих условий, накладываемых на параметрическую статистическую модель, асимптотические свойства МП-оценок формулируются следующим образом:

1. МП-оценки асимптотически несмещенные. Отметим, что при конечных объемах выборки n МП-оценки могут иметь смещение, которое при $n\to\infty$ стремится к нулю (см. упр. 2.15.13).

2. МП-оценки сильно состоятельны, то есть при $n\to\infty$ МП-оценка $\hat{\theta}(X_1,...,X_n)$ параметра θ сходится с вероятностью единица к θ и, следовательно, сходится по вероятности к θ.

3. МП-оценки при выполнении условий теоремы 2.10.24, асимптотически нормальны и эффективны.

Доказательство этих асимптотических свойств МП-оценок может быть найдено в литературе (см., например, [7, 13, 22, 24, 36]). Обсудим подробнее эти свойства оценок максимального правдоподобия.

2.10.Г. Состоятельность оценок максимального правдоподобия

Следуя [24], приведем доказательство этого свойства. Пусть задана параметрическая статистическая модель в виде семейства распределений $\Im_\theta$, то есть предполагаем, что имеется выборка

$X_1,...,X_n$ из распределения $F_X(x,\theta)\in\Im_\theta$ с плотностью $f_X(x,\theta)$, $x\in R^1$, $\theta\in\Theta$. Как обычно, оценку максимального правдоподобия параметра $\theta\in\Theta$ обозначим через $\hat{\theta}$ и пусть θ_0 обозначает истинное значение неизвестного параметра $\theta\in\Theta$, а M_0 – соответствующее математическое ожидание, вычисляемое при ф.р. $F_X(x,\theta_0)$.

Согласно определению оценок максимального правдоподобия, для любого допустимого значения параметра $\theta\in\Theta$ выполняется неравенство

$$L(\vec{x};\hat{\theta})\geq L(\vec{x};\theta),\ \forall\theta\in\Theta. \qquad \textbf{(2.10.19)}$$

Здесь знак равенства указывает на возможную неоднозначность оценок максимального правдоподобия. Условия регулярности, при которых МП-оценка является единственной и состоятельной, подробно обсуждаются в [24, с. 65]. Отметим, что функция правдоподобия является неотрицательной функцией и $\ln x$ для $x>0$ является выпуклой вверх функцией, поэтому, согласно неравенству Иенсена (П.3.14), выполняется следующее неравенство:

$$M_0\left\{\ln\frac{L(\vec{X};\theta^*)}{L(\vec{X};\theta_0)}\right\}<\ln M_0\left\{\frac{L(\vec{X};\theta^*)}{L(\vec{X};\theta_0)}\right\}. \qquad \textbf{(2.10.20)}$$

Далее, математическое ожидание в правой части предыдущего неравенства, согласно свойству функции правдоподобия (2.3.5), равно единице, то есть

$$\int_{-\infty}^{+\infty}...\int_{-\infty}^{+\infty}\frac{L(\vec{x};\theta^*)}{L(\vec{x};\theta_0)}L(\vec{x};\theta_0)dx_1...dx_n=1,$$

и, следовательно, из неравенства (2.10.20) получаем неравенство

$$M_0\left\{\ln\frac{L(\vec{X};\theta^*)}{L(\vec{X};\theta_0)}\right\}<0.$$

Отсюда, после умножения обеих частей предыдущего неравенства на $1/n$, следует неравенство

$$M_0\left\{\frac{1}{n}\ln L(\vec{X};\theta^*)\right\}<M_0\left\{\frac{1}{n}\ln L(\vec{X};\theta_0)\right\} \qquad \textbf{(2.10.21)}$$

при условии, что математическое ожидание в правой части (2.10.21) существует. Обычно это условие выполняется. Отметим далее, что $X_1,...,X_n$ – н.о.р. случайные величины, поэтому, согласно (2.3.2а), для любого $\theta \in \Theta$ справедливо выражение

$$\frac{1}{n}\ln L(\vec{X};\theta) = \frac{1}{n}\sum_{i=1}^{n}\ln f_X(X_i,\theta), \qquad \textbf{(2.10.22)}$$

правая часть которого является суммой независимых и одинаково распределенных случайных величин с математическим ожиданием

$$M_0\{\ln f_X(X_1;\theta)\} = M_0\{n^{-1}\ln L(\vec{X};\theta)\}.$$

Таким образом, согласно усиленному закону больших чисел (П.3.35), левая часть в (2.10.22), то есть случайная величина $n^{-1}\ln L(\vec{X};\theta)$, сходится с вероятностью единица при $n\to\infty$ к своему математическому ожиданию. Отсюда из (2.10.20) следует, что при $n\to\infty$ с вероятностью единица

$$\frac{1}{n}\ln L(x;\theta^*) < \frac{1}{n}\ln L(\vec{x};\theta_0),$$

или $$\lim_{n\to\infty} P\{\ln L(x;\theta^*) < \ln L(\vec{x};\theta_0)\} = 1,\ \theta^* \neq \theta_0.$$

С другой стороны, из (2.10.19) при $\theta = \theta_0$ имеем

$$\ln L(\vec{x};\hat{\theta}) \geq \ln L(\vec{x};\theta_0).$$

Из двух последних выражений следует, что при $n\to\infty$ значение $L(\vec{x};\hat{\theta})$ не может отличаться от $L(\vec{x};\theta_0)$. Таким образом, если $L(\vec{x};\theta)$ является взаимно однозначной функцией от θ, то $P\{\lim_{n\to\infty}\hat{\theta}_n = \theta_0\} = 1$. Доказательство завершено.

Замечание 2.10.23. Отметим, что МП-оценки, обладая свойством состоятельности при достаточно общих условиях, не всегда обладают свойством несмещенности (см. упр. 2.15.13). Для подтверждения этого в общем случае напомним, что оценка максимального правдоподобия некоторой непрерывной и монотонной функции φ от параметра θ равна этой функции от МП-оценки

параметра θ, то есть $\hat{\varphi}(\theta) = \varphi(\hat{\theta})$. Однако в общем случае выполняется неравенство $M[\varphi(\hat{\theta})] \neq \varphi(M(\hat{\theta}))$, и, следовательно, если $\hat{\theta}$ несмещенная оценка параметра θ, то $\varphi(\hat{\theta})$ не может быть несмещенной оценкой для $\varphi(\theta)$. Если же объем выборки достаточно большой, точнее при $n \to \infty$, МП-оценки всегда асимптотически несмещенные (см. теорему 2.10.24).

2.10.Д. Эффективность и асимптотическая нормальность МП-оценок

Если свойство состоятельности МП-оценок выполняются при достаточно общих условиях, то для выполнения свойств оптимальности МП-оценок, связанных с асимптотической эффективностью и нормальностью, требуются дополнительные ограничения на параметрическую модель, которые конкретизируются в следующей теореме.

Теорема 2.10.24. Пусть для параметрической модели $\mathfrak{I}_\theta$ параметр θ является внутренней точкой параметрического множества Θ. Предположим, что выполнены следующие условия:

А) *существует производная* $\dfrac{\partial^k}{\partial\theta^k}\ln f_X(x,\theta)$, $k = 1,2,3$, *причем*

$$M\left|\frac{\partial}{\partial\theta}\ln f_X(x,\theta)\right| + M\left|\frac{\partial^2}{\partial\theta^2}\ln f_X(x,\theta)\right| < \infty;$$

Б) $\left|\dfrac{\partial^3}{\partial\theta^3}\ln f_X(x,\theta)\right| \le H(x)$, *где* $H(x)$ *не зависит* θ *и* $M[H(X)] < \infty$;

В) $0 < M\left(\dfrac{\partial}{\partial\theta}\ln f_X(x,\theta)\right)^2 < \infty$.

Тогда произвольная сильно состоятельная последовательность $\hat{\theta}_n$ оценок максимального правдоподобия асимптотически нормальна и эффективна.

Доказательство. При заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ функция вклада $U_n(\vec{X},\theta)$ является функцией лишь параметра θ, то есть $U_n(\vec{x},\theta) = U_n(\theta)$. Разложим эту функцию

в ряд Тейлора в точке истинного значения параметра θ и вычислим это разложение в точке $\hat{\theta}_n$. В результате этого уравнение правдоподобия $0 = U_n(\hat{\theta}_n)$ из (2.10.6) запишется в виде

$$0 = U_n(\theta) + (\hat{\theta}_n - \theta)U_n'(\theta) + \frac{1}{2!}(\hat{\theta}_n - \theta)^2 U_n''(\theta^*),$$

где θ^* – некоторая точка, лежащая между θ и $\hat{\theta}_n$. Отсюда получаем

$$U_n(\theta) = (\hat{\theta}_n - \theta)[-U_n'(\theta) - (\hat{\theta}_n - \theta)U_n''(\theta^*)/2]$$

или

$$\sqrt{n \cdot i_1(\theta)}\,(\hat{\theta}_n - \theta) = \frac{U_n(\theta)}{\sqrt{n \cdot i_1(\theta)}\left[-\dfrac{U_n'(\theta)}{n \cdot i_1(\theta)} - \dfrac{1}{2} \cdot \dfrac{(\hat{\theta}_n - \theta)U_n''(\theta^*)}{n \cdot i_1(\theta)}\right]}, \quad \textbf{(2.10.25)}$$

где функция вклада выборки, согласно (2.3.4), определяется в виде

$$U_n(\vec{X}, \theta) = \frac{\partial \ln L(\vec{X}; \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i; \theta)}{\partial \theta}$$

и $i_1(\theta)$ – количество информации Фишера о параметре θ, содержащееся в одном наблюдении X_1, вычисляемое, согласно (2.3.11), по формуле

$$i_1(\theta) = M\left[\frac{\partial}{\partial \theta} \ln f(X_1; \theta)\right]^2 = -M\left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1; \theta)\right].$$

Для простоты записи введем следующие обозначения:

$$\alpha_n = U_n(\vec{X}, \theta)/\sqrt{n \cdot i_1(\theta)} = \frac{1}{\sqrt{n \cdot i_1(\theta)}} \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i; \theta)}{\partial \theta}; \quad \textbf{(2.10.26)}$$

$$\beta_n = -U_n'(\vec{X}, \theta)/n i_1(\theta) = -\frac{1}{i_1(\theta)} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f_X(X_i; \theta)}{\partial \theta^2}; \quad \textbf{(2.10.27)}$$

$$\gamma_n = U_n''(\theta^*)/2n i_1(\theta) = \frac{1}{2 \cdot i_1(\theta)} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 \ln f_X(X_i; \theta)}{\partial \theta^3}\Big|_{\theta=\theta^*}. \quad \textbf{(2.10.28)}$$

С учетом этих обозначений, формула (2.10.25) перепишется в виде

$$\sqrt{n \cdot i_1(\theta)}\,(\hat{\theta}_n - \theta) = \frac{\alpha_n}{\beta_n - (\hat{\theta}_n - \theta)\gamma_n}. \qquad \textbf{(2.10.29)}$$

Далее, согласно условию (Б), имеем

$$\gamma_n \le \frac{1}{2 \cdot i_1(\theta)} \cdot \frac{1}{n} \sum_{i=1}^{n} H(X_i). \qquad \textbf{(2.10.30)}$$

Применяя закон больших чисел (см. теорему Чебышева (П.3.10)), заключаем, что $n^{-1} \sum H(X_i)$ в правой части (2.10.30) при $n \to \infty$ сходится по вероятности к конечному пределу и, следовательно, случайная величина γ_n стохастически ограничена. Учитывая, что $(\hat{\theta}_n - \theta)$ сходится по вероятности к нулю, заключаем, что

$$\gamma_n(\hat{\theta}_n - \theta) \xrightarrow{P} 0 \text{ при } n \to \infty. \qquad \textbf{(2.10.31)}$$

Далее, согласно закону больших чисел, имеем

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f_X(X_i;\theta)}{\partial \theta^2} \xrightarrow{P} M\left[\frac{\partial^2}{\partial \theta^2} \ln f(X_1;\theta)\right] = -i_1(\theta) \text{ при } n \to \infty,$$

и, следовательно, случайная величина β_n сходится по вероятности к единице, то есть

$$\beta_n \xrightarrow{P} 1 \text{ при } n \to \infty. \qquad \textbf{(2.10.32)}$$

Для изучения предельного поведения случайной величины α_n, которая является суммой независимых и одинаково распределенных случайных величин, при этом

$$M\left(\frac{\partial}{\partial \theta} \ln f_X(X_i,\theta)\right) = 0 \text{ и } M\left(\frac{\partial}{\partial \theta} \ln f_X(X_i,\theta)\right)^2 = i_1(\theta) > 0, \ \forall i = 1,...,n,$$

применим центральную предельную теорему (П.3.24). В результате получаем, что случайная величина α_n имеет асимптотически стандартное нормальное распределение, то есть $L(\alpha_n) \approx N(0;1)$ при $n \to \infty$. Таким образом, из выражения (2.10.29), с учетом (2.10.31) и

(2.10.32), применение теоремы Слуцкого (П.3.5) позволяет заключить, что случайная величина $\sqrt{n \cdot i_1(\theta)}(\hat{\theta}_n - \theta)$ также имеет асимптотически стандартное нормальное распределение, то есть справедливо выражение

$$L\{\sqrt{n \cdot i_1(\theta)}(\hat{\theta}_n - \theta)\} = N(0,1) \text{ при } n \to \infty. \qquad \textbf{(2.10.33)}$$

Отсюда следует, что оценка максимального правдоподобия $\hat{\theta}_n$ имеет асимптотически нормальное распределение с параметрами θ и $1/ni_1(\theta)$, то есть

$$L(\hat{\theta}_n) \approx N(\theta; 1/ni_1(\theta)) \text{ при } n \to \infty. \qquad \textbf{(2.10.34)}$$

Таким образом, оценка максимального правдоподобия $\hat{\theta}_n$ является асимптотически несмещенной оценкой параметра θ, и поскольку, ее асимптотическая дисперсия равна обратной величине количества информации Фишера $i_n(\theta) = n \cdot i_1(\theta)$, содержащегося в выборке $\vec{X}_1 = (X_1,...,X_n)$, то оценка максимального правдоподобия $\hat{\theta}_n$ является асимптотически эффективной. Доказательство завершено.

Замечание 2.10.35. Отметим, что асимптотическая дисперсия МП-оценок, равная обратной величине количества информации Фишера $i_n(\theta) = n \cdot i_1(\theta)$, зависит от неизвестного параметра θ. При решении ряда задач, например при построении асимптотических доверительных интервалов (см. разд. 2.13), возникает необходимость устранить эту неопределенность. Это достигается путем замены в $i_n(\theta) = n \cdot i_1(\theta)$ неизвестного значения параметра θ оценкой максимального правдоподобия $\hat{\theta}_n$. Отметим, что такая замена приводит к дополнительной ошибке, однако она имеет более высокий порядок малости, чем ошибка, которая возникает при замене точного распределения $\hat{\theta}_n$ асимптотически нормальным распределением. Поэтому, если $i_n(\theta) = n \cdot i_1(\theta)$ является непрерывной функцией по θ, то замена $i_n(\theta)$ на $i_n(\hat{\theta}_n)$, согласно теореме (П.3.67), не приводит к изменению асимптотического распределения оценки $\hat{\theta}_n$ и выражение (2.10.33) перепишется в виде

$$L\{\sqrt{n \cdot i_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta)\} \approx N(0\,;\,1) \text{ при } n \to \infty. \qquad \textbf{(2.10.36)}$$

Пример 2.10.37. Пусть $L(X)=\Gamma(\alpha,\lambda)$, эта запись означает, что с.в. X имеет гамма-распределение с параметрами $\lambda>0$, $\alpha>0$ и ее плотность определяется в виде

$$f_X(x;\alpha,\lambda)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\ 0\le x<\infty, \qquad \textbf{(2.10.38)}$$

где $\Gamma(\alpha)$ – гамма-функция (см. П.2.18).

Рассмотрим сначала случай, когда масштабный параметр λ известен (положим его равным единице), а неизвестным является параметр формы распределения $\alpha>0$, то есть предполагаем, что плотность имеет вид

$$f_X(x;\alpha)=\frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x},\ 0\le x<\infty.$$

Имеется выборка $\vec{X}_1=(X_1,...,X_n)$ из распределения с этой плотностью. Требуется построить оценку максимального правдоподобия параметра α.

Для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}_1=(X_1,...,X_n)$ логарифм функции правдоподобия в данном случае запишется в виде

$$\ln L(\vec{x},\alpha)=\sum_{i=1}^{n}\ln f_X(x_i,\alpha)=-n\ln\Gamma(\alpha)+(\alpha-1)\sum_{i=1}^{n}\ln x_i-\sum_{i=1}^{n}x_i.$$

Отсюда

$$U(\alpha)=\frac{\partial}{\partial\alpha}\ln L(\vec{x},\alpha)=-n\frac{d}{d\alpha}\ln\Gamma(\alpha)+\sum_{i=1}^{n}\ln x_i$$

и

$$\frac{\partial^2}{\partial\alpha^2}\ln L(\vec{x},\alpha)=-n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)=n\cdot i_1(\alpha).$$

Уравнение правдоподобия (2.10.4) в данном случае записывается в виде

$$0=-n\frac{d}{d\alpha}\ln\Gamma(\alpha)+\sum_{i=1}^{n}\ln x_i, \qquad \textbf{(2.10.39)}$$

и, следовательно, оценкой максимального правдоподобия $\hat{\alpha}$ параметра α является единственный корень уравнения

$$-n\frac{d}{d\alpha}\ln\Gamma(\alpha)=\sum_{i=1}^{n}\ln x_i,$$

решить которое в явном виде не представляется возможным. В данном случае, как и в примере (2.10.15), для приближенного вычисления оценки максимального правдоподобия $\hat{\alpha}$ параметра α воспользуемся методом накопления.

Для реализации этого метода выберем в качестве начального приближения $\hat{\alpha}_0$ состоятельную оценку параметра α, построенную методом моментов (см. пример (2.9.7)), которая при $\lambda = 1$ имеет вид

$$\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

С учетом этого первое приближение $\hat{\alpha}_1$ для МП-оценки параметра α, согласно (2.10.13), запишется в виде

$$\hat{\alpha}_1 = \hat{\alpha}_0 + U(\hat{\alpha}_0)/n \cdot i_1(\hat{\alpha}_0) = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\sum \ln X_i - n\partial \ln \Gamma(\alpha)/\partial\alpha\,|_{\alpha=\hat{\alpha}_0}}{n\partial^2 \ln \Gamma(\alpha)/\partial\alpha^2\,|_{\alpha=\hat{\alpha}_0}}.$$

Отметим, что МП-оценка $\hat{\alpha}$ параметра α, являющаяся решением уравнения (2.10.39), согласно теореме (2.10.24), будет асимптотически эффективной и асимптотически нормальной со средним значением α и дисперсией

$$D(\hat{\alpha}) = \frac{1}{n \cdot i_1(\alpha)} = \left[-nM\left(\frac{\partial^2}{\partial\alpha^2}\ln f_X(X_1,\alpha)\right)\right]^{-1} = \left[n\frac{d^2}{d\alpha^2}\ln\Gamma(\alpha)\right]^{-1}.$$

Таким образом, выражение (2.10.33) в данном случае имеет вид

$$L\{\sqrt{n\partial^2 \ln \Gamma(\alpha)/\partial\alpha^2}\cdot(\hat{\alpha}-\alpha)\} \approx N(0;1) \text{ при } n \to \infty.$$

Отметим, что оценка $\hat{\alpha}_0 = n^{-1}\sum X_i$, построенная методом моментов, является несмещенной и состоятельной оценкой параметра α, но она не является эффективной и асимптотически эффективной. В самом деле, при $\lambda = 1$ $M(\hat{\alpha}_0) = \alpha$, $D(\hat{\alpha}_0) = \alpha/n$ (см. пример 2.9.7). Видно, что дисперсия этой оценки не достигает нижней границы в неравенстве Рао – Крамера, равной $1/n \cdot i_1(\alpha) = [n \cdot d^2 \ln\Gamma(\alpha)/d\alpha^2] = D(\hat{\alpha})$, и поэтому оценка $\hat{\alpha}_0 = n^{-1}\sum X_i$ не является эффективной. Кроме того, асимптотическая эффективность этой оценки, вычисляемая по формуле (2.6.3), равна

$$A\mathcal{Э}_F(\hat{\alpha}_0) = AO\mathcal{Э}_F(\hat{\alpha}_0 : \hat{\theta}_{эф}) = D(\hat{\alpha})/D(\hat{\alpha}_0) = [\alpha \cdot d^2 \ln\Gamma(\alpha)/d\alpha^2]^{-1}.$$

Эта эффективность не зависит от объема выборки n, и поэтому она не может стремиться к единице при $n \to \infty$, и, следовательно, оценка $\hat{\alpha}_0$ не является и асимптотически эффективной.

Рассмотрим теперь случай, когда неизвестным параметром является масштабный параметр λ, а параметр формы α будем считать заданным. Таким образом, предполагаем, что имеется выборка $\vec{X}_1 = (X_1,...,X_n)$ из распределения с плотностью (2.10.38). Требуется построить оценку максимального правдоподобия масштабного параметра λ при известном параметре формы

α. Для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ функция правдоподобия запишется в виде

$$L(\vec{x};\alpha,\lambda) = \prod_{i=1}^{n} f_X(x_i;\alpha,\lambda) = \prod_{i=1}^{n}\left(\frac{\lambda^{\alpha} x_i^{\alpha-1} e^{-\lambda x_i}}{\Gamma(\alpha)}\right).$$

Отсюда

$$\ln L(\vec{x};\alpha,\lambda) = -n\ln\Gamma(\alpha) + n\alpha\ln\lambda + (\alpha-1)\sum_{i=1}^{n}\ln x_i - \lambda\sum_{i=1}^{n} x_i$$

и $$\frac{\partial \ln L(\vec{x};\alpha,\lambda)}{\partial\lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i\,,\ \frac{\partial^2 \ln L(\vec{x};\alpha,\lambda)}{\partial\lambda^2} = -\frac{n\alpha}{\lambda^2} < 0\,.$$

Таким образом, уравнение правдоподобия записывается в виде

$$\frac{\partial \ln L(\vec{x};\alpha,\lambda)}{\partial\lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0\,,$$

и, следовательно, МП-оценка параметра λ равна $\hat{\lambda} = \dfrac{n\alpha}{\sum X_i} = \alpha / \overline{X}$. Отметим, что полученная МП-оценка $\hat{\lambda} = \alpha / \overline{X}$ параметра λ для гамма-распределения при известном параметре α совпадает с ММ-оценкой этого параметра, построенной методом моментов (см. пример 2.9.7). Далее, количество информации Фишера равно

$$i_n(\lambda) = n \cdot i_1(\lambda) = -M\left(\frac{\partial^2 \ln L(\vec{x};\alpha,\lambda)}{\partial\lambda^2}\right) = \frac{n\alpha}{\lambda^2}\,.$$

Вычислим математическое ожидание и дисперсию оценки $\hat{\lambda}$. Для этого заметим, что если $L(X_i) = \Gamma(\alpha,\lambda)$, $\forall i = 1,...,n$, то сумма $X_1 + X_2 + ... + X_n = n\overline{X}$ также имеет гамма-распределение с параметрами $n\alpha$ и λ, то есть

$$f_{n\overline{X}}(x;n\alpha,\lambda) = \frac{\lambda^{n\alpha}}{\Gamma(n\alpha)} x^{n\alpha-1} e^{-\lambda x},\ 0 \le x < \infty\,,$$

Учитывая это замечание, получаем

$$M(\hat{\lambda}) = \int_0^{\infty} \frac{n\alpha}{x} f_{n\overline{X}}(x;n\alpha,\lambda)dx = \frac{n\alpha\cdot\lambda^{n\alpha}}{\Gamma(n\alpha)} \int_0^{\infty} x^{n\alpha-2} e^{-\lambda x} dx =$$

$$= \frac{n\alpha\cdot\lambda^{n\alpha-1}}{\Gamma(n\alpha)\lambda^{n\alpha-2}} \int_0^{\infty} t^{n\alpha-2} e^{-t} dt = \frac{n\alpha\cdot\lambda}{\Gamma(n\alpha)}\Gamma(n\alpha-1) = \frac{n\alpha\cdot\lambda\Gamma(n\alpha-1)}{(n\alpha-1)\Gamma(n\alpha-1)} = \frac{n\alpha}{(n\alpha-1)}\cdot\lambda\,.$$

Таким образом, $M(\hat{\lambda}) \neq \lambda$, и, следовательно, МП-оценка $\hat{\lambda}$ смещенная, но асимптотически несмещенная, так как $M(\hat{\lambda}) \to \lambda$ при $n \to \infty$. Смещение оценки легко устранить, рассмотрев несмещенную оценку вида $\tilde{\lambda} = \frac{n\alpha - 1}{n\alpha} \cdot \hat{\lambda} = \frac{n\alpha - 1}{\sum X_i}$. Очевидно, что $M(\tilde{\lambda}) = \lambda$. Далее, для вычисления дисперсии оценки $\hat{\lambda}$ сначала вычислим

$$M(\hat{\lambda}^2) = \int_0^\infty \left(\frac{n\alpha}{x}\right)^2 f_{n\bar{X}}(x; n\alpha, \lambda)dx = \frac{(n\alpha)^2 \cdot \lambda^{n\alpha}}{\Gamma(n\alpha)} \int_0^\infty x^{n\alpha-3} e^{-\lambda x} dx =$$

$$= \frac{(n\alpha)^2 \cdot \lambda^{n\alpha-1}}{\Gamma(n\alpha)\lambda^{n\alpha-3}} \int_0^\infty t^{n\alpha-3} e^{-t} dt = \frac{(n\alpha)^2 \cdot \lambda^2}{\Gamma(n\alpha)} \Gamma(n\alpha - 2) = \frac{(n\alpha\lambda)^2}{(n\alpha-1)(n\alpha-2)}.$$

Отсюда

$$D(\hat{\alpha}) = \frac{(n\alpha\lambda)^2}{(n\alpha-1)(n\alpha-2)} - \frac{(n\alpha\lambda)^2}{(n\alpha-1)^2} = \frac{(n\alpha\lambda)^2}{n\alpha-1}\left[\frac{1}{n\alpha-2} - \frac{1}{n\alpha-1}\right] = \frac{(n\alpha\lambda)^2}{(n\alpha-1)^2(n\alpha-2)}.$$

Следовательно, для несмещенной оценки $\tilde{\lambda} = \frac{n\alpha - 1}{n\alpha} \cdot \hat{\lambda}$ получаем дисперсию, определяемую формулой

$$D(\tilde{\lambda}) = \frac{(n\alpha-1)^2}{(n\alpha)^2} \cdot D(\hat{\lambda}) = \frac{(n\alpha-1)^2}{(n\alpha)^2} \cdot \frac{(n\alpha\lambda)^2}{(n\alpha-1)^2(n\alpha-2)} = \frac{\lambda^2}{n\alpha-2}.$$

Отметим, что из теоремы (2.10.24) следует асимптотическая эффективность и нормальность МП-оценки $\hat{\lambda} = \alpha / \bar{X}$ параметра λ. В частности, выражение (2.10.36) запишется в виде $L(\hat{\lambda}) \approx N(\lambda; n\alpha/\lambda^2)$ при $n \to \infty$. Кроме того, асимптотическое значение дисперсии оценки $\hat{\lambda}$ совпадает с дисперсией асимптотического распределения этой оценки, то есть $D(\hat{\alpha}) = (n\alpha\lambda)^2/(n\alpha-1)^2(n\alpha-2) \to n\alpha/\lambda^2$ при $n \to \infty$. Заметим, что отмеченный факт не всегда выполняется. Убедимся теперь, что несмещенная оценка $\tilde{\lambda}$ не является эффективной оценкой ни при каком конечном n, но она так же, как и оценка $\hat{\lambda}$, асимптотически эффективна. В самом деле, сравнивая дисперсию оценки $\tilde{\lambda}$ с нижней границей в неравенстве Рао – Крамера, получаем

$$[n \cdot i_1(\lambda) \cdot D(\tilde{\lambda})]^{-1} = \left[\frac{n\alpha}{\lambda^2} \cdot \frac{\lambda^2}{n\alpha - 2}\right]^{-1} = (n\alpha - 2)/n\alpha \ .$$

Ясно, что $(n\alpha - 2)/n\alpha < 1$ для любого конечного n и $(n\alpha - 2)/n\alpha \to 1$ при

$n \to \infty$. Отметим, что при $\alpha = 1$ гамма-распределение превращается в экспоненциальное распределение, то есть $L(X) = \Gamma(\alpha = 1, \lambda) = \mathrm{Exp}(\lambda)$. Плотность для экспоненциальной с.в. X записывается в виде $f_X(x; \alpha = 1, \lambda) = \lambda e^{-\lambda x}$, $0 \le x < \infty$, и, следовательно, МП-оценка параметра λ экспоненциального распределения равна $\hat{\lambda} = n / \sum X_i = 1/\overline{X}$.

Пример 2.10.40. В примере 2.10.10 были получены МП-оценки параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ двумерного нормального распределения. Общие асимптотические свойства этих оценок следуют из теоремы 2.10.24. Для иллюстрации применения этой теоремы рассмотрим для простоты нормированное двумерное нормальное распределение, для которого $a_X = a_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 1$, $\rho = M(XY)$, и конкретизируем асимптотическое распределение для МП-оценки $\hat{\rho}$ коэффициента корреляции ρ.

Для нормированного двумерного нормального распределения функция правдоподобия, при заданной реализации $\vec{z} = (x_1, y_1), ..., (x_n, y_n)$ двумерной выборки $(X_1, Y_1), ..., (X_n, Y_n)$, является лишь функцией коэффициента корреляции ρ, то есть $L(\vec{z}; \rho) = L(\rho)$. Далее, логарифмическая производная функции правдоподобия (см. формулу (5) в примере 2.10.10) имеет вид

$$\frac{\partial \ln L(\vec{z}; \rho)}{\partial \rho} = \frac{n\rho}{1-\rho^2} + \frac{1}{1-\rho^2}\sum_{i=1}^{n} x_i y_i - \frac{\rho}{(1-\rho^2)^2}\left[\sum_{i=1}^{n} x_i - 2\rho\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i{}^2\right].$$

Дифференцируя второй раз, получим

$$\frac{\partial^2 \ln L(\vec{z}; \rho)}{\partial \rho^2} = \frac{n(1+\rho^2)}{(1-\rho^2)^2} + \frac{4\rho}{(1-\rho^2)^2}\sum_{i=1}^{n} x_i y_i - \frac{1+3\rho^2}{(1-\rho^2)^3}\left[\sum_{i=1}^{n} x_i - 2\rho\sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i{}^2\right].$$

Отсюда, учитывая, что $M(X^2) = M(Y^2) = 1$ и $\rho = M(XY)$, получим

$$i_n(\rho) = n \cdot i_1(\rho) = -M\left(\frac{\partial^2 \ln L(\vec{z}; \rho)}{\partial \rho^2}\right) = -\frac{n(1+\rho^2)}{(1-\rho^2)^2} - \frac{4n\rho^2}{(1-\rho^2)^2} + \frac{2n(1+3\rho^2)}{(1-\rho^2)^2} = \frac{n(1+\rho^2)}{(1-\rho^2)^2}.$$

Таким образом, асимптотическая дисперсия МП-оценки $\hat{\rho}$ равна

$$D(\hat{\rho}) = 1/n \cdot i_1(\rho) = \frac{(1-\rho^2)^2}{n(1+\rho^2)},$$

и, согласно теореме (2.10.24), оценка $\hat{\rho}$ асимптотически эффективна и имеет асимптотически нормальное распределение, то есть выполняется следующее выражение:

$$L(\hat{\rho}) \approx N\{\rho\,; (1-\rho^2)^2 / n(1+\rho^2)\} \text{ при } n \to \infty.$$

Замечание 2.10.41. Примеры ситуаций, в которых МП-оценки не обладают свойством состоятельности, приводятся в [24, с. 91]. В частности, случай неодинаковых распределений, для которого функция правдоподобия записывается в виде $L(\vec{x};\theta_1,...,\theta_k) = f_1(x_1;\theta) f_2(x_2;\theta)\cdots f_n(x;\theta)$, где различные сомножители f_i, $i=1,...,n$, могут зависеть от различных функций от системы параметров $\theta_1,...,\theta_k$, обсуждается в [24, с. 92]. См. также примеры в книге [11а, с. 5–115].

2.11. Метод минимума расстояний

Предположим, что статистическая модель $(\mathrm{X},\Im_\theta)$ задана в параметрической форме. Здесь X – выборочное пространство и $\Im_\theta$ – множество допустимых распределений вероятностей, определяемое в виде $\Im_\theta = \{F : F_X(x,\theta), \theta\in\Theta\}$. Это предположение означает, что имеется выборка $X_1,...,X_n$ из распределения $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, $x\in R^1$, $\theta\in\Theta$, причем функциональный вид распределения задан с точностью до неизвестного параметра θ (скалярного либо векторного), который принадлежит заданному параметрическому множеству Θ. Требуется построить по выборке $X_1,...,X_n$ из распределения $F_X(x,\theta)$ оценку неизвестного параметра $\theta\in\Theta$.

Для построения оценок параметров статистической модели $(\mathrm{X},\Im_\theta)$ могут быть использованы различные методы, в частности метод моментов (см. разд. 2.9) и метод максимального правдоподобия (см. разд. 2.10). Обсудим еще один общий метод, который называют методом минимума расстояний. Оценки параметров, построенные методом минимума расстояний, в литературе кратко называют *MD-оценки* (*minimum distance estimators*). Суть метода минимума расстояний состоит в том, что на множестве непрерывных функций распределений $\Im$ для пары распределений $F,G\in\Im$, задается метрика (или расстояние) $\rho(F,G)$. Оценка параметра θ, построенная методом выбранного расстояния $\rho(F,G)$, определяется из условия минимума этого расстояния между эмпирической функцией распределения $F_n(x)$, построенной по выборке $X_1,...,X_n$, и функцией распределения $F_\theta(x) = F_X(x,\theta)$, принятой в модели $(\mathrm{X},\Im_\theta)$.

Таким образом, для выбранного расстояния $\rho(F,G)$, *MD-оценки* определяются в виде

$$\hat{\theta} = \arg\min_{\theta} [\rho(F_n, F_\theta]. \qquad \textbf{(2.11.1)}$$

При построении *MD-оценок* используют различные расстояния. В случае непрерывных распределений обычно используют расстояния супремального вида

$$\rho_1(F_n, F_X(x,\theta)) = \sup_{-\infty<x<+\infty} |F_n(x) - F_X(x,\theta)|\,\psi(F_X(x,\theta)) \qquad \textbf{(2.11.2)}$$

или расстояния интегрального вида

$$\rho_2(F,G) = \int_{-\infty}^{+\infty} [F(x) - G(x)]^2\,\omega(x)dx, \qquad \textbf{(2.11.3)}$$

где ψ и ω – заданные весовые функции.

В случае дискретных распределений обычно используется расстояние «хи-квадрат» вида (2.11.8), которое также может быть использовано и в непрерывном случае после применения метода группировки исходных наблюдений.

Замечание 2.11.4. Использование расстояний (2.11.3) при построении *MD-оценок* обсуждается в разд. 8.7. В данном разделе мы обсудим *MD-оценки*, построенные методом минимума расстояния «хи-квадрат», которые играют важную роль при решении задачи согласия со сложной гипотезой (см. разд. 3.5).

Для записи расстояния «хи-квадрат» вида (2.11.8) в непрерывном случае предварительно применяется метод группировки исходных наблюдений. При этом область возможных значений изучаемой с. в. X разбивают на заданное число k непересекающихся интервалов группировки $\Delta_1,...,\Delta_k$, где $\Delta_i = y_{i+1} - y_i$, $i = 1,...,k$. Далее, наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ отображается в дискретный вариационный ряд частот $n_i,...,n_k$, то есть

$$\vec{x} = (x_1,...,x_n) \Rightarrow \begin{pmatrix} \Delta_1,...,\Delta_k \\ n_1,...,n_k \end{pmatrix},\ \sum_{i=1}^{k} n_i = n. \qquad \textbf{(2.11.5)}$$

По данным вариационного ряда частот (2.11.5), вычисляется расстояние в виде

$$\sum_{i=1}^{k} c_i [(n_i / n) - p_i(\theta)]^2 , \qquad \textbf{(2.11.6)}$$

где $c_1,...,c_k$ – заданные весовые коэффициенты и $p_i(\theta)$, $i = 1,...,k$, – вероятности попадания с.в. X в интервалы группировки $\Delta_1,...,\Delta_k$, вычисляемые по формулам

$$p_i(\theta) = P\{X \in \Delta_i\} = \int_{y_i}^{y_{i+1}} dF_X(x;\theta) = F_X(y_{i+1};\theta) - F_X(y_i;\theta),$$
$$i = 1,...,k . \qquad \textbf{(2.11.7)}$$

Задание весов $c_1,...,c_k$ в формуле (2.11.6) в виде $c_i = n / p_i(\theta)$ соответствует расстоянию «хи-квадрат», вычисляемому по формулам

$$\chi^2 = \sum_{i=1}^{k} \frac{n}{p_i(\theta)} \left(\frac{n_i}{n} - p_i(\theta) \right)^2 = \sum_{i=1}^{k} \left[\frac{(n_i - np_i(\theta))^2}{np_i(\theta)} \right] = \sum_{i=1}^{k} \frac{n_i^2}{np_i(\theta)} - n . \qquad \textbf{(2.11.8)}$$

Минимизация этого расстояния по θ приводит к *MD-оценке* параметра θ, построенной методом минимума расстояния «хи-квадрат». Эта оценка является решением уравнения вида

$$\sum_{i=1}^{k} \frac{n_i^2}{p_i^2(\theta)} \frac{\partial p_i(\theta)}{\partial \theta} = 0 . \qquad \textbf{(2.11.9)}$$

В случае векторного параметра θ размерности r, то есть $\vec{\theta} = (\theta_1,...,\theta_r)$, *MD-оценки* находятся путем решения системы уравнений

$$\sum_{i=1}^{k} \frac{n_i^2}{p_i^2(\vec{\theta})} \frac{\partial p_i(\vec{\theta})}{\partial \theta_l} = 0 , \; l = 1,...,r . \qquad \textbf{(2.11.10)}$$

Замечание 2.11.11. Отметим, что решение системы (2.11.10) обычно вызывает трудности, поэтому часто используют другую систему вида

$$\sum_{i=1}^{k} \frac{n_i}{p_i(\vec{\theta})} \frac{\partial p_i(\vec{\theta})}{\partial \theta_l} = 0 , \; l = 1,..., r , \qquad \textbf{(2.11.12)}$$

решение которой обычно найти проще. Оценки, полученные путем решения системы (2.11.12), называют *MD-оценками*, построенными методом минимума *измененного расстояния* «хи-квадрат» [22]. Отметим, что *MD-оценки*, полученные методом минимума измененного расстояния «хи-квадрат», то есть являющиеся решением системы (2.11.12), обладают теми же асимптотическими свойствами, что и *MD-оценки*, полученные методом минимума расстояния «хи-квадрат» в результате решения системы (2.11.10), см. [36]. Отметим также, что *MD-оценки*, полученные методом минимума измененного расстояния «хи-квадрат», совпадают с так называемыми мультиномиальными оценками максимального правдоподобия (см. упр. 2.15.27). Эти оценки строятся не по исходным данным $X_1,...,X_n$, а по данным, которые предварительно обработаны с помощью метода группировки, то есть по данным вариационного ряда частот (2.11.5), и являются решением следующей системы уравнений:

$$\sum_{i=1}^{k} n_i \frac{\partial \ln p_i(\vec{\theta})}{\partial \theta_l} = 0\,,\ l = 1,...,r\,. \qquad \textbf{(2.11.13)}$$

Отметим также, что при достаточно общих условиях *MD-оценки*, так же как и оценки максимального правдоподобия, являются состоятельными, асимптотически нормальными и эффективными оценками, см. [36], а также разд. 8.7 и 8.8.

2.12. Интервальные оценки параметров. Доверительные интервалы

После вычисления по заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ точечной оценки $T_n = T_n(X_1,...,X_n)$ параметра θ в параметрической модели $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$, получаем число $t_n = t_n(x_1,...,x_n)$, которое принимается в качестве истинного значения θ_0 неизвестного параметра θ. Это число является реализацией случайной величины $T_n = T_n(X_1,...,X_n)$, и поэтому, в каждом конкретном случае, вычисленное значение оценки может отличаться от истинного значения параметра, и, следовательно, необходимо знать возможную погрешность, которая возникает при использовании данной оценки. Для описания качества оценивания неизвестного параметра θ используются такие общие понятия, как несмещенность, состоятельность, достаточность, эффективность. Эти поня-

тия были определены в разд. 2.1 – 2.5. Однако, характеризуя общие свойства оценок, эти понятия не дают представления о количественном значении точности оценивания. Для ответа на этот вопрос при построении точечных оценок требуется дополнительно рассчитывать такие количественные характеристики, как среднеквадратическая ошибка (СКО) оценки, дисперсия оценки или асимптотическая дисперсия. Альтернативный подход состоит в использовании интервальных оценок, для которых заданная характеристика точности непосредственно включена в процесс их построения. Итак, для построения интервальной оценки параметра θ по выборке $\vec{X}_1 = (X_1,...,X_n)$ требуется использовать две статистики $\hat{\theta}_1(X_1,...,X_n)$ и $\hat{\theta}_2(X_1,...,X_n)$, такие, чтобы выполнялось выражение

$$P\{\hat{\theta}_1(X_1,...,X_n) < \theta < \hat{\theta}_2(X_1,...,X_n)\} \geq 1-\alpha. \qquad \textbf{(2.12.1)}$$

Это выражение означает, что случайный интервал $(\hat{\theta}_1, \hat{\theta}_2)$ с границами $\hat{\theta}_1$ и $\hat{\theta}_2$, вычисляемыми по выборке $\vec{X}_1 = (X_1,...,X_n)$, накрывает неизвестный параметр θ с вероятностью, не меньшей заданной доверительной вероятности $\gamma = 1-\alpha$. Величину α называют уровнем значимости (см. разд. 3.4.Г о двойственности задач построения интервальных оценок и проверки статистических гипотез) и обычно используют небольшие и стандартные значения, например $\alpha = 0{,}10;\ 0{,}05;\ 0{,}01$. Отметим, что в экономической практике, при построении доверительных интервалов, например при $\alpha = 0{,}05$ и $\gamma = 0{,}95$, говорят о 95 %-м доверительном интервале. Техника построения доверительных интервалов тесно связана с построением критериев проверки статистических гипотез относительно значения параметра θ (см., например, [5, с. 191; 22, с. 163; 24, с. 277]). Этим, в частности, объясняется и запись доверительной вероятности γ в виде $\gamma = 1-\alpha$, где величина α является уровнем значимости критерия, или вероятностью ошибки первого рода. Подробности см. в разд. 3.4.Г и примерах 3.4.74 и 3.4.79.

Итак, при построении доверительного интервала для параметра θ необходимо определить такие две статистики $\hat{\theta}_1(X_1,...,X_n)$ и $\hat{\theta}_2(X_1,...,X_n)$, для которых при заданной доверительной вероятно-

сти $\gamma = 1 - \alpha$ выполняется выражение (2.12.1). Существуют различные подходы к построению интервальных оценок. Эти подходы существенно зависят от характера интересующего нас параметра θ и от уровня статистического описания наблюдений, то есть они определяются типом рассматриваемой статистической модели. Рассмотрим сначала параметрическую модель $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$ и обсудим один из общих подходов, который основан на использовании понятия «центральная статистика» см. [22, с. 83]. Другие подходы, в частности асимптотический подход, и непараметрические доверительные интервалы, которые также называют «свободными от распределения», рассматриваются в разд. 4.9 и 5.2.

Пусть задана параметрическая модель $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$. Это означает, что ф.р. $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, $x \in R^1$, задана с точностью до конечного числа неизвестных параметров. Предполагаем для простоты, что параметр θ скалярный. Пусть $\vec{X}_1 = (X_1,...,X_n)$ – выборка из распределения $F_X(x,\theta)$ и $\vec{x} = (x_1,...,x_n)$ – ее конкретная наблюденная реализация. Введем понятие центральной статистики для параметра θ параметрической модели $\Im_\theta$.

Определение 2.12.2. Статистику $S(\vec{X},\theta)$ называют центральной для параметра θ, если она удовлетворяет следующим двум условиям:

(А) Для каждой наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ $S(\vec{x},\theta)$ является непрерывной и строго монотонной функцией от параметра $\theta \in \Theta$.

(Б) Статистика $S(\vec{X},\theta)$ зависит от параметра θ, но ее распределение вероятностей от этого параметра не зависит.

Пусть для модели $\Im_\theta$ построена центральная статистика $S(\vec{X},\theta)$. Обозначим функцию распределения статистики $S(\vec{X},\theta)$ через $F_S(x)$. Эта функция распределения вероятностей, согласно условию (Б), не зависит от параметра θ. По этой причине всегда можно выбрать (большим числом способов) такие два числа s_1 и s_2 $(s_1 < s_2)$, чтобы выполнялось выражение

$$P\{s_1 < S(\vec{X},\theta) < s_2\} = \int_{S_1}^{S_2} dF_S(x) = 1-\alpha, \qquad \textbf{(2.12.3)}$$

где $\gamma = (1-\alpha)$ – заданная величина, причем $0 < \gamma < 1$. Далее, поскольку $S(\vec{x},\theta)$ для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ является, согласно условию (А), непрерывной и строго монотонной функцией от θ, то решение относительно θ уравнения $S(\vec{x},\theta) = s_1$ позволяет однозначно вычислить величину $\hat{\theta}_1(\vec{x})$, а решение уравнения $S(\vec{x},\theta) = s_2$ позволяет получить величину $\hat{\theta}_2(\vec{x})$, причем $\hat{\theta}_1(\vec{x}) < \hat{\theta}_2(\vec{x})$. Функция $S(\vec{x},\theta)$ строго монотонная по θ, поэтому неравенства $s_1 < S(\vec{x},\theta) < s_2$ и $\hat{\theta}_1(\vec{x}) < \theta < \hat{\theta}_2(\vec{x})$ эквивалентны, следовательно, выражение (2.12.3) перепишется в виде $P\{\hat{\theta}_1(\vec{X}) < \theta < \hat{\theta}_2(\vec{X})\} = 1-\alpha$, а это и означает, что $(\hat{\theta}_1(\vec{X}), \hat{\theta}_2(\vec{X}))$ является γ-доверительным интервалом для параметра θ.

Отметим, что при заданной величине $\gamma = 1-\alpha$, $0 < \gamma < 1$, выбор чисел s_1 и s_2 $(s_1 < s_2)$, удовлетворяющих выражению (2.12.3), не является однозначным, что приводит и к неоднозначности построения доверительного интервала описанным способом. Для обеспечения однозначности обычно числа s_1 и s_2 выбирают, исходя из выполнения следующих условий:

$$P\{S(\vec{X},\theta) \le s_1\} = P\{S(\vec{X},\theta) \ge s_2\} = \alpha/2.$$

Такой выбор чисел s_1 и s_2 приводит к доверительным интервалам, которые часто называют *центральными*, см., например, [22], и кажется естественным в тех случаях, когда ф.р. $F_S(x)$ центральной статистики $S(\vec{X},\theta)$ симметрична относительно нуля (см. пример 2.12.4). В этом случае числа s_1 и s_2 выбирают в следующем виде: $s_1 = -s_2$, причем s_2 является квантилем уровня $(1-\alpha/2)$ распределения $F_S(x)$, то есть $s_2 = F_S^{-1}(1-\alpha/2)$. Такой выбор чисел s_1 и s_2 удовлетворяет выражению (2.12.3) и при заданной доверительной вероятности $\gamma = (1-\alpha)$ приводит к минимальной длине доверительного интервала при симметричных ф.р. $F_S(x)$ (см. упр. 2.15.30).

Пример 2.12.4. Пусть $L(X)=N(a,\sigma^2)$. Имеется выборка $\vec{X}_1=(X_1,...,X_n)$ из распределения $F_X(x,\vec{\theta})=\Phi((x-a)/\sigma)$, где $\vec{\theta}=(a,\sigma^2)$ и $\Phi(\cdot)$ – стандартная нормальная функция распределения. Для данного примера мы рассматривали точечные оценки параметров a и σ^2, которые были построены различными методами. Оказалось, что все рассмотренные методы для данного примера приводят к одинаковым точечным оценкам параметров a и σ^2, которые соответственно называют выборочным средним и выборочной дисперсией и записываются в виде

$$\overline{X}=n^{-1}\sum_{i=1}^{n}X_i \text{ и } S^2=n^{-1}\sum_{i=1}^{n}(X_i-\overline{X})^2. \qquad \textbf{(2.12.5)}$$

Общие свойства этих точечных оценок см. в упр. 2.15.13. Обсудим теперь построение интервальных оценок параметров a и σ^2 для разных вариантов степени неопределенности относительно этих параметров.

Вариант 1. Для простоты рассмотрим сначала случай, когда параметр a является неизвестным, а параметр σ^2 предполагается известным. В этом случае центральная статистика записывается в виде $S(\vec{X},a)=\sqrt{n}(\overline{X}-a)/\sigma$. Отметим, что статистика $S(\vec{X},a)$ удовлетворяет условиям (А) и (Б) определения 2.12.2. В самом деле, для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}_1=(X_1,...,X_n)$ $S(\vec{x},a)$ является непрерывной и строго монотонной функцией по a. Далее, согласно теореме 2.8.31, выполняется выражение $L\{S(\vec{X},a)\}=N(0;1)$, то есть условие (Б) выполнено, следовательно, функция распределения статистики $S(\vec{X},a)$ совпадает со стандартной нормальной ф.р. $\Phi(x)$, то есть $F_S(x)=\Phi(x)$. Таким образом, выражение (2.12.3) в данном случае запишется в виде

$$P\{s_1<S(\vec{X},a)<s_2\}=\int_{S_1}^{S_2}d\Phi(x)=1-\alpha. \qquad \textbf{(2.12.6)}$$

Учитывая, что ф.р. $\Phi(x)$ симметрична относительно нуля, то есть $\Phi(x)=1-\Phi(-x)$, $\forall x\in R^1$, положим $s_1=-s_2$. С учетом этого замечания получаем

$$\int_{-S_2}^{S_2}d\Phi(x)=\Phi(s_2)-\Phi(-s_2)=2\Phi(s_2)-1=1-\alpha.$$

Отсюда имеем $\Phi(s_2)=1-\alpha/2$, и, следовательно, s_2 является квантилем уровня $(1-\alpha/2)$ стандартного нормального распределения, то есть $s_2=\Phi^{-1}(1-\alpha/2)$.

Для дальнейшего удобства введем принятое обозначение квантиля уровня q, $0<q<1$, стандартного нормального распределения $\Phi(x)$ в виде λ_q, то есть $\lambda_q=\Phi^{-1}(q)$, следовательно, $s_2=\lambda_{1-\alpha/2}$. Далее, разрешая для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}_1=(X_1,...,X_n)$ неравенства $-s_2<(\sqrt{n}(\overline{x}-a)/\sigma)<s_2$ относительно параметра a, получим $\overline{x}-s_2\sigma/\sqrt{n}<a<\overline{x}+s_2\sigma/\sqrt{n}$. Учитывая, что $s_2=\lambda_{1-\alpha/2}$, окончательно получаем из (2.12.6) доверительный интервал для параметра a в виде

$$P\{\overline{X}-\lambda_{1-\alpha/2}\sigma/\sqrt{n}<a<\overline{X}+\lambda_{1-\alpha/2}\sigma/\sqrt{n}\}=1-\alpha. \qquad \textbf{(2.12.7)}$$

Отметим, что полученный центральный γ-доверительный интервал для параметра a при известном параметре σ^2 имеет минимальную длину среди всех γ-доверительных интервалов, где $\gamma=(1-\alpha)$ (см. упр. 2.15.30).

Отметим также, что доверительные интервалы позволяют вычислять необходимое число экспериментов или объем выборки n, при котором с заданной доверительной вероятностью обеспечивается заданная точность оценивания параметра. Проиллюстрируем это на примере оценивания параметра a нормального распределения. Обозначим половину длины доверительного интервала $(\overline{X}\pm\Delta)$ из (2.12.7) через $\Delta=\lambda_{1-\alpha.2}\cdot\sigma/\sqrt{n}$. Отсюда легко вычислить необходимый объем выборки для обеспечения заданной точности Δ при оценивании параметра a по формуле $n=[\lambda^2_{1-\alpha/2}\cdot\sigma^2/\Delta^2]$, где $[\cdot]$ – обозначает наибольшую целую часть. Например, пусть $\sigma^2=5,0$ и $\alpha=0,05$, что соответствует доверительной вероятности $\gamma=0,95$ и при этом $\lambda_{0,975}=\Phi^{-1}(0,975)=1,96$. Для обеспечения заданной точности, например $\Delta=0,5$ с доверительной вероятностью $\gamma=0,95$, потребуется число наблюдений $n=[(1,96)^2\cdot 5/0,25]=77$.

Вариант 2. Рассмотрим теперь случай, когда оба параметра a и σ^2 неизвестны и требуется построить γ-доверительный интервал для параметра a. В данном случае центральная статистика, согласно теореме 2.8.31, записывается в виде $S(\vec{X},a)=\sqrt{n-1}(\overline{X}-a)/S$, где выборочное среднее $\overline{X}$ и выборочное стандартное отклонение S вычисляются с помощью формул (2.12.5). Отметим, что статистика $S(\vec{X},a)$ имеет распределение Стьюдента с числом степеней свободы $(n-1)$, то есть выполняется выражение $L\{S(\vec{X},a)\}=St(n-1)$ (см. П.2.89). Обозначим квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $(n-1)$ через $t_{1-\alpha/2,\,n-1}$. Рассуждения, аналогичные предыдущему варианту, с учетом того, что распределение Стьюдента симметрично относительно нуля, приводят к тому, что в доверительном интерва-

ле (2.12.7) следует заменить неизвестное в данном случае стандартное отклонение σ на ее выборочную оценку S и квантиль стандартного нормального распределения $\lambda_{1-\alpha/2}$ – на квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $(n-1)$, то есть на $t_{1-\alpha/2,\,n-1}$. В результате этих действий получаем γ-доверительный интервал, $\gamma=(1-\alpha)$, для параметра a в виде

$$P\{\bar{X}-t_{1-\alpha/2,\,n-1}\cdot S/\sqrt{n-1}<\ a\ <\bar{X}+t_{1-\alpha/2,\,n-1}\cdot S/\sqrt{n-1}\}=1-\alpha\,. \qquad \textbf{(2.12.8)}$$

Отметим, что при увеличении числа степеней свободы распределение Стьюдента приближается к нормальному распределению, поэтому, если объем выборки «большой», например $n>100$, то в доверительном интервале (2.12.8), без существенной погрешности можно использовать вместо квантиля распределения Стьюдента $t_{1-\alpha/2,\,n-1}$ квантиль стандартного нормального распределения $\lambda_{1-\alpha/2}$. Если объем выборки «маленький» $(n<30)$, то в доверительном интервале (2.12.8) обычно используют вместо выборочного стандартного отклонения S подправленную выборочную оценку стандартного отклонения $\tilde{S}=S\sqrt{n/(n-1)}$, в результате этого доверительный интервал запишется в виде

$$P\{\bar{X}-t_{1-\alpha/2,\,n-1}\cdot\tilde{S}/\sqrt{n}<a<\bar{X}+t_{1-\alpha/2,\,n-1}\cdot\tilde{S}/\sqrt{n}\}=1-\alpha\,. \qquad \textbf{(2.12.8а)}$$

Вариант 3. Построение доверительного интервала для параметра σ^2 при известном параметре a рассматривается в упр. 2.15.31. Рассмотрим теперь случай, когда оба параметра a и σ^2 неизвестны и требуется построить γ-доверительный интервал для параметра σ^2. В данном случае центральная статистика, согласно теореме (2.8.31), записывается в виде $S(\vec{X},\sigma^2)=nS^2/\sigma^2$, где S^2 – выборочная дисперсия, являющаяся точечной оценкой параметра σ^2, вычисляемая по формуле (2.12.5). Отметим, что для заданной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}_1=(X_1,...,X_n)$ $S(\vec{x},\sigma^2)$ является непрерывной и убывающей функцией по σ^2. Кроме того, статистика $S(\vec{X},\sigma^2)$ имеет распределение хи-квадрат с числом степеней свободы $(n-1)$, то есть $L\{S(\bar{X},\sigma^2)\}=\chi^2_{(n-1)}$, см. выражение (2.8.34). Таким образом, условия (А) и (Б) определения (2.12.2) выполнены. Построим центральный γ-доверительный интервал для параметра σ^2. Для этого воспользуемся выражением (2.12.3) и выберем числа s_1 и s_2 $(s_1<s_2)$ из условий

$$P\{\chi^2_{(n-1)}\le s_1\}=P\{\chi^2_{(n-1)}\ge s_2\}=\alpha/2\,,$$

Далее, обозначим квантиль уровня p, $0<p<1$, распределения хи-квадрат с

числом степеней свободы $(n-1)$ через $\chi^2_{p,\,n-1}$, тогда числа s_1 и s_2 будут соответственно равны: $s_1 = \chi^2_{\alpha/2,\,n-1}$, $s_2 = \chi^2_{1-\alpha/2,\,n-1}$. Подставив найденные значения s_1 и s_2 в (2.12.3), с учетом того, что $L\{S(\bar{X},\sigma^2)\} = \chi^2(n-1)$, получим для статистики $S(\vec{X},\sigma^2) = nS^2/\sigma^2$ следующее выражение:

$$P\{\chi^2_{\alpha/2,\,n-1} < nS^2/\sigma^2 < \chi^2_{1-\alpha/2,\,n-1}\} = 1-\alpha\,.$$

Разрешая неравенства в фигурных скобках относительно параметра σ^2, окончательно получаем γ-доверительный интервал, $\gamma = (1-\alpha)$, для параметра σ^2 в виде

$$((\chi^2_{1-\alpha/2,\,n-1})^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\,,\ (\chi^2_{\alpha/2,\,n-1})^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2)\,,$$

то есть для параметра σ^2 справедливо выражение

$$P\left\{\frac{nS^2}{\chi^2_{1-\alpha/2,\,n-1}} < \sigma^2 < \frac{nS^2}{\chi^2_{\alpha/2,\,n-1}}\right\} = 1-\alpha\,. \qquad \textbf{(2.12.9)}$$

Отметим, что для стандартного отклонения σ $\gamma = (1-\alpha)$-доверительный интервал запишется в виде

$$(\,\{(\chi^2_{1-\alpha/2,\,n-1})^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\}^{1/2}\,,\ \{(\chi^2_{\alpha/2,\,n-1})^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\}^{1/2}\,)\,. \qquad \textbf{(2.12.10)}$$

Замечание 2.12.11. Отметим, что доверительные интервалы (2.12.8) и (2.12.9) были построены для параметров a и σ^2 в отдельности. Утверждение о том, что пара значений (a,σ^2) накрывается прямоугольником

$$\{\bar{X} - t_{1-\alpha/2,\,n-1}\cdot S/\sqrt{n-1} < a < \bar{X} + t_{1-\alpha/2,\,n-1}\cdot S/\sqrt{n-1}\}\,,$$
$$\left\{\frac{nS^2}{\chi^2_{1-\alpha/2,\,n-1}} < \sigma^2 < \frac{nS^2}{\chi^2_{\alpha/2,\,n-1}}\right\}$$

с заданной доверительной вероятностью $\gamma = (1-\alpha)$, является ошибочным (см. [13, с. 410]), поскольку случайные величины $\sqrt{n}(\bar{X}-a)/S$ и S^2 являются зависимыми. Доверительную область для пары параметров (a,σ^2) можно построить, используя факт независимости выборочного среднего $\bar{X}$ и выборочной дисперсии S^2 (см. теорему 2.8.31). В самом деле, учитывая, что $\bar{X}$ и S^2 независимы, γ-доверительная область для пары параметров (a,σ^2) запишется в виде

$$\gamma = P\left\{ s_1 < \frac{\sqrt{n}(\overline{X} - a)}{\sigma} < s_2 \,,\, \beta_1 < \frac{nS^2}{\sigma^2} < \beta_2 \right\} =$$

$$= P\left\{ s_1 < \frac{\sqrt{n}(\overline{X} - a)}{\sigma} < s_2 \right\} \cdot P\left\{ \beta_1 < \frac{nS^2}{\sigma^2} < \beta_2 \right\}. \qquad \textbf{(2.12.11)}$$

Для определения чисел s_1, s_2 и β_1, β_2 каждый множитель правой части приравняем к величине равной $\sqrt{\gamma}$. В результате получим два выражения:

$$P\left\{ s_1 < \frac{\sqrt{n}(\overline{X} - a)}{\sigma} < s_2 \right\} = \sqrt{\gamma}\,; \qquad \textbf{(2.12.12)}$$

$$P\left\{ \beta_1 < \frac{nS^2}{\sigma^2} < \beta_2 \right\} = \sqrt{\gamma}\,. \qquad \textbf{(2.12.13)}$$

Учитывая, что статистика $\sqrt{n}(\overline{X} - a)/\sigma$ имеет симметричное распределение, положим $s_1 = -s_2$ и определим s_2 равным квантилю уровня $(1+\sqrt{\gamma})/2$ стандартного нормального распределения, то есть $s_2 = \lambda_{(1+\sqrt{\gamma})/2}$. В результате выражение (2.12.12) перепишется в виде

$$P\{-\lambda_{(1+\sqrt{\gamma})/2}\sigma/\sqrt{n} < (\overline{X} - a) < \lambda_{(1+\sqrt{\gamma})/2}\sigma/\sqrt{n}\,\} = \sqrt{\gamma}\,.$$

Далее, числа β_1, β_2 определим из условий

$$P\{\chi^2_{(n-1)} \le \beta_1\} = P\{\chi^2_{(n-1)} \ge \beta_2\} = (1 - \sqrt{\gamma}\,)/2\,.$$

В результате выражение (2.12.13) перепишется в виде

$$P\{nS^2/\chi^2_{(1+\sqrt{\gamma})/2,\,n-1} < \sigma^2 < nS^2/\chi^2_{(1-\sqrt{\gamma})/2,\,n-1}\} = \sqrt{\gamma}\,.$$

Таким образом, γ-доверительная область для пары параметров (a, σ^2) представляет собой область на плоскости (a, σ^2) вида

$$-\lambda_{(1+\sqrt{\gamma})/2}\sigma/\sqrt{n} < (\overline{X} - a) < \lambda_{(1+\sqrt{\gamma})/2}\sigma/\sqrt{n}\,,$$

$$nS^2/\chi^2_{(1+\sqrt{\gamma})/2,\,n-1} < \sigma^2 < nS^2/\chi^2_{(1-\sqrt{\gamma})/2,\,n-1}. \qquad \textbf{(2.12.15)}$$

Эта область ограничена параболой $\sigma^2 = n(\overline{X} - a)^2/\lambda^2_{(1+\sqrt{\gamma})/2}$ и двумя прямыми

$$\sigma^2 = nS^2/\chi^2_{(1+\sqrt{\gamma})/2,\,n-1} \text{ и } \sigma^2 = nS^2/\chi^2_{(1-\sqrt{\gamma})/2,\,n-1}.$$

Замечание 2.12.16. В рассмотренных вариантах построения доверительных интервалов параметров нормального распределения использовалось понятие «центральная статистика», причем при построении центральных статистик существенно использовалась информация о нормальности распределения, при котором выполняется теорема 2.8.31. Однако можно указать класс статистических моделей, для которых центральная статистика всегда существует и процесс ее построения не опирается на использовании информации о функциональном характере распределении. К таким моделям относятся модели, в которых множество допустимых распределений $\Im$ состоит из *непрерывных и монотонных* по параметру θ функций распределений $F_X(x,\theta)$. Основой для построения центральных статистик в таких моделях служит вероятностное интегральное преобразование (см. разд. 4.3.А), согласно которому для выборки $\vec{X}_1 = (X_1,...,X_n)$ из распределения $F_X(x,\theta)$ случайные величины $U_i = F_X(X_i,\theta)$, $i = 1,...,n$, имеют равномерное в интервале [0,1] распределение, то есть $L(U_i) = R(0;1)$, $\forall i = 1,...,n$. Предположение непрерывности и монотонности распределения $F_X(x,\theta)$ по параметру θ, а также использование вероятностного интегрального преобразования позволяет обеспечить выполнение условий (А) и (Б) определения (2.12.2). В качестве примера, взятого из [22], рассмотрим центральную статистику вида

$$S(\vec{X},\theta) = \sum_{i=1}^{n}(-\ln F_X(X_i,\theta)). \qquad \textbf{(2.12.17)}$$

Для этой статистики условие (А) выполнено, поскольку $F_X(x,\theta)$ монотонная функция по параметру θ. Далее, поскольку $U_i = F_X(X_i,\theta)$ и $L(U_i) = R(0;1)$ $\forall i = 1,...,n$ и $\forall\theta$, то распределение статистики $S(\vec{X},\theta)$ не зависит от параметра θ, то есть условие (Б) также выполнено. Кроме того, использование логарифмического преобразования позволяет легко конкретизировать распределение статистики $S(\vec{X},\theta)$. В самом деле, из того факта, что $L(U_i) = R(0;1)$ $\forall i = 1,...,n$, следует выполнение выражения $L(-\ln U_i) = \Gamma(1,1)$ $\forall i = 1,...,n$, где $\Gamma(\alpha,\lambda)$ – гамма распределение (см. табл П.2.18) с плотностью вида

$$f_X(x;\alpha,\lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x},\ 0 \le x < \infty, \qquad \textbf{(2.12.18)}$$

Таким образом, слагаемые в статистике $S(\vec{X},\theta)$ вида (2.12.17) независимы и каждое имеет распределение $\Gamma(1,1)$, поэтому сама статистика имеет распределение $\Gamma(n,1)$, то есть ее плотность равна $f_S(x;n,1) = x^{n-1}e^{-x}/\Gamma(n)$, $x > 0$. С учетом этого замечания, выражение (2.12.3) запишется в виде

$$P\{s_1 < \sum(-\ln F_X(X_i,\theta)) < s_2\} = \frac{1}{\Gamma(n)} \int_{S_1}^{S_2} x^{n-1}e^{-x}dx = 1-\alpha\,. \qquad \textbf{(2.12.19)}$$

Окончательно, для построения центрального γ-доверительного интервала, $\gamma=1-\alpha$, остается определить числа s_1 и s_2 из условий $P\{\xi\le s_1\}=P\{\xi\ge s_2\}=\alpha/2$, где с.в. ξ имеет гамма-распределение, то есть $L\{\xi\}=\Gamma(n,1)$ и разрешить относительно θ неравенства в фигурных скобках выражения (2.12.19). По всей вероятности, именно последняя операция и будет вызывать основные трудности при построении доверительного интервала для параметра θ при каждой конкретной ф.р. $F_X(x,\theta)$. Отметим, что приведенный метод основан на использовании интегрального вероятностного преобразования, которое будет также использовано и при построении непараметрических доверительных интервалов (см. разд. 4.9). Приведем теперь пример построения доверительного интервала с помощью описанного метода.

Пример 2.12.20. Пусть $L(X)=R(0,\theta)$. Эта запись означает, что с.в. X имеет равномерное распределение в интервале $[0,\theta]$, то есть ф.р. $F_X(x,\theta)$ этой случайной величины равна

$$F_X(x,\theta) = x/\theta\,,\ 0\le x\le\theta\,.$$

Имеется выборка $\vec{X}_1=(X_1,...,X_n)$ из распределения $F_X(x,\theta)$, требуется построить доверительный интервал для параметра $\theta>0$. Отметим, что в данном случае ф.р. $F_X(x,\theta)$ непрерывная и монотонно убывающая функция по θ при $\theta>0$ и, следовательно, в качестве центральной статистики может быть использована статистика $S(\vec{X},\theta)$ вида (2.12.17), которая для данного примера запишется в виде

$$S(\vec{X},\theta) = n\ln\theta - \sum_{i=1}^{n}\ln X_i\,. \qquad \textbf{(2.12.21)}$$

Далее, выражение (2.12.22) записывается в виде

$$P\{s_1 < n\ln\theta - \sum\ln X_i < s_2\} = \frac{1}{\Gamma(n)} \int_{S_1}^{S_2} x^{n-1}e^{-x}dx = 1-\alpha\,, \qquad \textbf{(2.12.22)}$$

где для построения центрального γ-доверительного интервала, $\gamma=1-\alpha$, числа s_1 и s_2 определяются из условий $P\{\xi\le s_1\}=P\{\xi\ge s_2\}=\alpha/2$, здесь с.в. ξ имеет гамма-распределение, то есть $L\{\xi\}=\Gamma(n,1)$. Окончательно, разрешая неравенства в фигурных скобках выражения (2.12.19), получим центральный γ-доверительный интервал для параметра θ в виде

$$P\{\exp(s_1/n)\sqrt[n]{X_1\cdots X_n} < \theta < \exp(s_2/n)\sqrt[n]{X_1\cdots X_n}\} = 1-\alpha, \qquad \textbf{(2.12.23)}$$

где s_1 – квантиль уровня $\alpha/2$, а s_2 – квантиль уровня $1-\alpha/2$ распределения $\Gamma(n,1)$ с плотностью $f(x) = x^{n-1}e^{-x}/\Gamma(n)$, $x > 0$.

Отметим, что в данном примере можно построить другую центральную статистику, исходя из тех соображений, что n-я порядковая статистика $X_{(n)}$ выборки $\vec{X}_1 = (X_1,...,X_n)$ из равномерного распределения в интервале $[0,\theta]$, являясь точечной оценкой параметра θ и достаточной статистикой для этого параметра (см. пример 2.5.11), имеет, согласно (4.2.4), ф.р. вида

$$F_{X_{(n)}}(x,\theta) = F_X^n(x,\theta) = (x/\theta)^n,\ 0 \le x \le \theta. \qquad (2.12.24)$$

Далее, согласно вероятностному интегральному преобразованию (4.3.16), случайная величина $U = F_{X_{(n)}}(X_{(n)},\theta)$ имеет для любого $\theta > 0$ равномерное в $[0,1]$ распределение, поэтому выполняется выражение $L\{(X_{(n)}/\theta)^n\} = R(0,1)$ $\forall\theta > 0$, а это означает, что условие (Б) определения (2.12.2) для статистики $S(\vec{X},\theta) = (X_{(n)}/\theta)^n$ выполнено. Кроме того, для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ функция $S(\vec{x},\theta) = (x_{(n)}/\theta)^n$ является непрерывной и убывающей функцией по θ, то есть условие (А) определения 2.12.2 также выполнено. Таким образом, для параметра θ равномерного распределения $R(0,\theta)$ центральная статистика имеет вид

$$S(\vec{X},\theta) = (X_{(n)}/\theta)^n, \qquad (2.12.25)$$

и, следовательно, выражение (2.12.3) с учетом того, что в данном случае s_1 равняется квантилю уровня $\alpha/2$ равномерного в $[0,1]$ распределения, то есть $s_1 = \alpha/2$, а s_2 равняется квантилю уровня $1-\alpha/2$ равномерного в $[0,1]$ распределения, то есть $s_2 = 1-\alpha/2$, запишется в виде

$$1-\alpha = P\{s_1 < S(\vec{X},\theta) < s_2\} = P\{\alpha/2 < (X_{(n)}/\theta)^n < 1-\alpha/2\}.$$

Учитывая теперь монотонность функции $S(\vec{x},\theta) = (x_{(n)}/\theta)^n$ по θ, окончательно получаем центральный γ-доверительный интервал, $\gamma = 1-\alpha$, для параметра θ равномерного распределения $R(0,\theta)$ в виде

$$P\{X_{(n)}/\sqrt[n]{1-\alpha/2} < \theta < X_{(n)}/\sqrt[n]{\alpha/2}\} = 1-\alpha. \qquad (2.12.26)$$

2.13. Асимптотический подход к построению доверительных интервалов

В разд. 2.12 были обсуждены интервальные оценки параметра θ параметрической модели $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$, при этом в качестве основного подхода использовался метод, основанный на использовании понятия «центральная статистика». Для построения центральных статистик была использована информация об известном функциональном характере ф.р. $F(x,\theta)$. Рассмотрим теперь интервальные оценки параметров второго типа, определяемых в виде (2.1.2), когда интересующий нас параметр θ является функционалом от ф.р. F, которая неизвестна. При этом будем предполагать, что объем выборки достаточно большой, такой, при котором применимы асимптотические результаты. Отметим, что выражение «объем выборки достаточно большой» является расплывчатым. Однако довольно часто асимптотические результаты являются вполне приемлемой аппроксимацией при конечных объемах выборки. В тех случаях, когда это будет возможно, мы будем указывать объемы выборок, при которых асимптотические результаты применимы на практике. Пусть интересующий нас параметр θ является функционалом от ф.р. F вида

$$\theta(F) = \int \varphi(x) dF(x), \qquad \textbf{(2.13.1)}$$

где $\varphi(x)$ – заданная непрерывная функция, причем такая, что

$$\sigma_F^2(\varphi) = \int \varphi^2(x) dF(x) - \left(\int \varphi(x) dF(x) \right)^2 < \infty. \qquad \textbf{(2.13.2)}$$

Имеется выборка $\vec{X}_1 = (X_1, ..., X_n)$ из неизвестного распределения $F \in \Im$, где $\Im$ – множество непрерывных допустимых распределений в условиях данного эксперимента. Точечная оценка параметра $\theta(F)$, построенная методом подстановки (см. пример 2.8.3), записывается в виде

$$\theta_n = \theta(F_n) = n^{-1} \sum_{i=1}^{n} \varphi(X_i). \qquad \textbf{(2.13.3)}$$

Согласно центральной предельной теореме П.3.37, при условии, что

$\sigma_F^2(\varphi) < \infty$, выборочная оценка θ_n имеет асимптотически нормальное распределение, то есть

$$L\left(\frac{\sqrt{n}(\theta_n - \theta(F))}{\sigma_F(\varphi)}\right) = N(0;1) \text{ при } n \to \infty. \qquad \textbf{(2.13.4)}$$

Отметим, что для статистики $\sqrt{n}(\theta_n - \theta(F))/\sigma_F(\varphi)$ условие (А) определения 2.12.2 может и не выполняться, поскольку $\sigma_F^2(\varphi)$ зависит от φ и θ, по этой причине она не может быть центральной статистикой для рассматриваемого параметра. Для построения центральной статистики обычно заменяют неизвестное значение $\sigma_F^2(\varphi)$ ее выборочной оценкой вида

$$\hat{\sigma}_F(\varphi) = \left\{ n^{-1} \sum_{i=1}^{n} [\varphi(X_i) - \theta_n]^2 \right\}^{1/2}. \qquad \textbf{(2.13.5)}$$

Такая замена оправдана, поскольку предполагается, что объем выборки достаточно большой. В результате получаем для параметра θ асимптотическую центральную статистику в виде

$$S(\vec{X}, \theta) = \sqrt{n}(\theta_n - \theta(F))/\hat{\sigma}_F(\varphi). \qquad \textbf{(2.13.6)}$$

Отметим, что замена $\sigma_F^2(\varphi)$ на $\hat{\sigma}_F^2(\varphi)$ в выражении (2.13.4) приводит к дополнительной ошибке, однако при больших n эта ошибка имеет более высокий порядок малости, чем ошибка, которая совершается при замене точного распределения оценки θ_n асимптотически нормальным, поэтому (см. замечание 2.10.35) для статистики (2.13.6) также выполняется выражение

$$L\left(\frac{\sqrt{n}(\theta_n - \theta(F))}{\hat{\sigma}_F(\varphi)}\right) = N(0;1) \text{ при } n \to \infty. \qquad \textbf{(2.13.7)}$$

Далее, для построения γ-доверительного интервала, где $\gamma = (1 - \alpha)$ – заданная доверительная вероятность, воспользуемся выражением (2.12.3), которое в данном случае запишется в виде

$$P\{s_1 < \sqrt{n}(\theta_n - \theta(F)/\hat{\sigma}_F(\varphi)) < s_2\} = \int_{S_1}^{S_2} d\Phi(x) = 1 - \alpha. \qquad \textbf{(2.13.8)}$$

Учитывая, что распределение центральной статистики (2.13.6) является симметричным, построим центральный доверительный интервал. Для этого положим $s_1 = -s_2$ и определим s_2 равным квантилю уровня $1-(\alpha/2)$ стандартного нормального распределения, то есть $s_2 = \lambda_{1-\alpha/2}$. В результате выражение (2.13.8) перепишется в виде

$$P\{-\lambda_{1-\alpha/2} < \sqrt{n}(\theta_n - \theta(F))/\hat{\sigma}_F(\varphi) < \lambda_{1-\alpha/2}\} = 1-\alpha .$$

Разрешая неравенства в фигурных скобках относительно θ, окончательно получаем центральный γ-доверительный интервал для рассматриваемого параметра θ в виде

$$P\{\theta_n - \lambda_{1-\alpha/2}\hat{\sigma}_F(\varphi)/\sqrt{n} < \theta(F) < \theta_n + \lambda_{1-\alpha/2}\hat{\sigma}_F(\varphi)/\sqrt{n}\} = 1-\alpha . \quad \textbf{(2.13.9)}$$

Пример 2.13.10. Пусть интересующий нас параметр θ характеризует начальный момент k-го порядка, $k = 1, 2, \ldots$, то есть

$$\theta(F) = \alpha_k = M(X^k) = \int x^k dF(x) .$$

Это означает, что функция $\varphi(x)$ равна $\varphi(x) = x^k$, $k = 1, 2, \ldots,$ и

$$\sigma_F^2(\varphi) = \int \varphi^2(x) dF(x) - \left(\int \varphi(x) dF(x)\right)^2 = \int x^{2k} dF(x) - \left(\int x^k dF(x)\right)^2 = \alpha_{2k} - \alpha_k^2 .$$

Точечная выборочная оценка начального момента k-го порядка, построенная по выборке $\vec{X}_1 = (X_1, \ldots, X_n)$, имеет вид $\hat{\alpha}_k = n^{-1}\sum_{i=1}^{n} X_i^k$, и выборочная оценка величины $\sigma_F(\varphi)$ равна $\hat{\sigma}_F(\varphi) = (\hat{\alpha}_{2k} - \hat{\alpha}_k^2)^{1/2}$. Таким образом, при больших объемах выборки центральный γ-доверительный интервал для начального момента k-го порядка записывается, согласно (2.13.9), в виде

$$P\{\hat{\alpha}_k - \lambda_{1-\alpha/2}\sqrt{(\hat{\alpha}_{2k} - \hat{\alpha}_k^2)/n} < \alpha_k < \hat{\alpha}_k + \lambda_{1-\alpha/2}\sqrt{(\hat{\alpha}_{2k} - \hat{\alpha}_k^2)/n}\} = 1-\alpha . \quad \textbf{(2.13.11)}$$

Пример 2.13.12. Пусть вероятность $p = P(A)$ некоторого события A нам неизвестна. Проводится серия из n независимых опытов, в результате которых событие A появилось k раз. Точечной оценкой неизвестной вероятности $p = P(A)$ является относительная частота $\hat{p} = k/n$. Построим интервальную оценку параметра p. Заметим, что k является дискретной случайной величиной с биномиальным распределением (см. П.1.8), то есть $L(k) = Bi(n; p)$ и, следовательно, $M(k) = np$, $D(k) = np(1-p)$. С учетом этого получаем, что

$M(\hat{p}) = p$ и $D(\hat{p}) = p(1-p)/n$. Далее, согласно теореме Муавра – Лапласа (см. П.1.15), случайная величина $\sqrt{n}(\hat{p}-p)/\sqrt{p(1-p)}$ имеет асимптотически стандартное нормальное распределение, то есть при $n \to \infty$ справедливо выражение

$$L\left(\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}\right) = N(0;1). \qquad \textbf{(2.13.13)}$$

Таким образом, при больших объемах выборки справедливо следующее выражение:

$$P\{\lambda_{\alpha/2} < \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} < \lambda_{1-\alpha/2}\} = 1-\alpha, \qquad \textbf{(2.13.14)}$$

где λ_q – квантиль уровня q, $0 < q < 1$, стандартного нормального распределения. Для построения γ-доверительного интервала, $\gamma = (1-\alpha)$, для параметра p перепишем в эквивалентной форме неравенства в фигурных скобках выражения (2.13.14). Учитывая, что $\lambda_{\alpha/2} = -\lambda_{1-\alpha/2}$ и полагая для сокращения записи $\lambda_{1-\alpha/2} = \lambda$, указанные неравенства запишем в виде $|\hat{p}-p| < \lambda\sqrt{p(1-p)/n}$ или $(\hat{p}-p)^2 - \lambda^2 p(1-p)/n < 0$, или, наконец, в виде $p^2(1+\lambda^2/n) - p(2\hat{p}+\lambda^2/n) + \hat{p}^2 < 0$. Многочлен второй степени относительно p в левой части последнего неравенства имеет два действительных корня

$$\hat{p}_1(k,n) = \frac{(2\hat{p}+\lambda^2/n) + \sqrt{(2\hat{p}+\lambda^2/n) - 4(1+\lambda^2/n)\hat{p}^2}}{2(1+\lambda^2/n)} =$$

$$= \frac{(\hat{p}+\lambda^2/2n) + \lambda\sqrt{(\hat{p}(1+\hat{p})/n + \lambda^2/4n^2}}{1+\lambda^2/n}$$

и

$$\hat{p}_2(k,n) = \frac{(2\hat{p}+\lambda^2/n) - \sqrt{(2\hat{p}+\lambda^2/n) - 4(1+\lambda^2/n)\hat{p}^2}}{2(1+\lambda^2/n)} =$$

$$= \frac{(\hat{p}+\lambda^2/2n) - \lambda\sqrt{(\hat{p}(1+\hat{p})/n + \lambda^2/4n^2}}{1+\lambda^2/n},$$

причем $0 < \hat{p}_1(k,n) < \hat{p}_2(k,n) < 1$. Далее, неравенства $\hat{p}_1(k,n) < p < \hat{p}_2(k,n)$ и неравенства в фигурных скобках выражения (2.13.14) эквивалентны, а это означает, что интервал $(\hat{p}_1(k,n), \hat{p}_2(k,n))$ является γ-доверительным интервалом, $\gamma = (1-\alpha)$, для параметра p, то есть справедливо следующее выражение:

$$P\{\hat{p}_1(k,n) < p < \hat{p}_2(k,n)\} = 1-\alpha. \qquad \textbf{(2.13.15)}$$

Отметим, что построение асимптотического доверительного интервала для параметра p с учетом поправки на непрерывность при аппроксимации биномиального распределения нормальным приводится в упр. 2.15.36. Отметим также, что при достаточно больших объемах выборки $(n>100)$ величинами λ^2/n, $\lambda^2/2n$, $\lambda^2/4n^2$ в выражении (2.13.15) можно пренебречь без потери в точности, в результате чего получаем более простое выражение для доверительного интервала. Учитывая это замечание и вспоминая, что $\lambda=\lambda_{1-\alpha/2}$, из (2.13.15) получаем центральный γ-доверительный интервал, $\gamma=(1-\alpha)$, для параметра p в виде

$$P\{\hat{p}-\lambda_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}<p<\hat{p}+\lambda_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\ \}=1-\alpha. \qquad \textbf{(2.13.16)}$$

Отметим, что доверительный интервал вида (2.13.16) можно получить и более коротким путем. Для этого, по аналогии с (2.13.6), используем асимптотическую центральную статистику для параметра p в виде $S(\vec{X},p)=\{\sqrt{n}(\hat{p}-p)/\sqrt{\hat{p}(1-\hat{p})}\}$, для которой (см. замечание 2.10.35) при $n\to\infty$ справедливо выражение

$$L\left(\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}}\right)=N(0\,,1), \qquad \textbf{(2.13.17)}$$

а это означает, что, в свою очередь, имеет место выражение вида

$$L\{\lambda_{\alpha/2}<\sqrt{n}(\hat{p}-p)/\sqrt{\hat{p}(1-\hat{p})}<\lambda_{1-\alpha/2}\}=1-\alpha,$$

из которого следует (2.13.16).

Замечание 2.13.18. Мы уже отмечали, что использование доверительных интервалов позволяет вычислять необходимое число экспериментов или объем выборки n, при котором с заданной доверительной вероятностью обеспечивается заданная точность оценивания параметра. Используя доверительный интервал (2.13.16), обсудим этот вопрос на примере оценивания параметра p, характеризующего неизвестную вероятность некоторого события A, то есть $p=P(A)$. Обозначим половину длины доверительного интервала (2.13.16) через $\Delta=\lambda_{1-\alpha/2}\cdot\sqrt{\hat{p}(1-\hat{p})/n}$. Отсюда легко вычислить, при заданной доверительной вероятности $\gamma=(1-\alpha)$ и заданной точности Δ, необходимый объем выборки по формуле

$$n=[\lambda^2_{1-\alpha/2}\hat{p}(1-\hat{p})/\Delta^2]=\left[\lambda^2_{1-\alpha/2}\frac{m_0}{\Delta^2 n_0}\left(1-\frac{m_0}{n_0}\right)\right], \qquad \textbf{(2.13.19)}$$

где n_0 – объем пробной (предварительной) выборки и m_0 – число появления события A в серии из n_0 опытов. Пусть, например, $\alpha=0{,}01$ (то есть довери-

тельная вероятность $\gamma = 0,99$ и $\lambda_{0,995} = \Phi^{-1}(0,995) = 2,58$), пусть далее задана величина $\Delta = 0,01$ и $\hat{p} = m_0 / n_0 = 0,4$, тогда требуемый объем выборки будет равен $n = [(2,58)^2 0,4 \cdot 0,6/(0,01)^2] = 15976$. При $\alpha = 0,05$ для обеспечения данной точности потребуется число экспериментов $n = 9220$. Отметим, что другие варианты построения γ-доверительного интервала для параметра p приводятся в упр. 2.15.36 и 2.15.39.

2.14. Доверительные интервалы, основанные на распределениях точечных оценок параметров

Мы уже отмечали, что возможны различные подходы к построению доверительных интервалов и они определяются характером оцениваемого параметра и видом статистической модели. В случае параметрической гауссовской модели в подходе, основанном на использовании понятия «центральная статистика», была использована информация о нормальности распределения. Для произвольной параметрической модели вида $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$ один из общих подходов, при условии, что $F_X(x,\theta)$ – непрерывная и монотонная функция по параметру θ, был основан на использовании центральных статистик и вероятностного интегрального преобразования (см. замечание 2.12.19). Другой общий подход основан на использовании точечных оценок $\hat{\theta}_n$ параметра θ и их распределений. Естественно, что мы заинтересованы в использовании таких точечных оценок, при которых для заданной доверительной вероятности интервалы будут обладать наименьшей длиной. Если используются несмещенные оценки, то интервалы будут тем короче, чем меньше дисперсии оценок, поэтому эффективные и асимптотически эффективные оценки приводят к интервалам с наименьшей длиной.

Итак, пусть $T_n = T(X_1,...,X_n)$ – точечная оценка параметра θ, построенная по выборке $X_1,...,X_n$ из распределения $F_X(x,\theta) \in \Im_\theta$, и пусть ф.р. случайной величины T_n обозначена через $F_{T_n}(x,\theta)$. Предположим, что распределение с.в. T_n непрерывно и ф.р. $F_{T_n}(x,\theta)$ является *непрерывной и монотонной* функцией по пара-

метру θ. Далее, для любого $\theta \in \Theta$ и для заданной доверительной вероятности $\gamma = 1 - \alpha$ определим два числа $t_i = t_i(\theta)$, $i = 1, 2$, причем $t_1 < t_2$, такие, чтобы для точечной оценки $T_n = T(X_1, ..., X_n)$ выполнялось выражение

$$P\{t_1 < T_n < t_2\} = \int_{t_1}^{t_2} dF_{T_n}(x, \theta) = F_{T_n}(t_2, \theta) - F_{T_n}(t_1, \theta) = 1 - \alpha. \quad \textbf{(2.14.1)}$$

Уже отмечалось, что выбор чисел t_1 и t_2, удовлетворяющих этому выражению, можно осуществить большим числом способов, поэтому обычно строят центральный доверительный интервал, а для этого числа t_1 и t_2 выбирают из условий

$$P\{T_n \le t_1\} = F_{T_n}(t_1, \theta) = \alpha/2, \; P\{T_n \ge t_2\} = 1 - F_{T_n}(t_2, \theta) = \alpha/2. \quad \textbf{(2.14.2)}$$

Учитывая теперь, что ф.р. $F_{T_n}(x, \theta)$ является *непрерывной и монотонной* функцией по параметру θ, разрешая (2.14.2) относительно θ для заданной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X}_1 = (X_1, ..., X_n)$, а следовательно, и для конкретного значения $t = T(\vec{x})$ статистики $T_n = T(X_1, ..., X_n)$, из выражения $F_{T_n}(t, \theta) = \alpha/2$ – число $\hat{\theta}_1(\vec{x})$, а из выражения $F_{T_n}(t, \theta) = 1 - \alpha/2$ определим число $\hat{\theta}_2(\vec{x})$, причем эти числа будут удовлетворять неравенству $\hat{\theta}_1(\vec{x}) < \hat{\theta}_2(\vec{x})$, если функция $F_{T_n}(x, \theta)$ возрастает по параметру θ, и неравенство изменится на противоположное, если ф. р. $F_{T_n}(x, \theta)$ убывает по параметру θ.

Таким образом, для случая, когда ф.р. $F_{T_n}(x, \theta)$ точечной оценки $T_n = T(X_1, ..., X_n)$ *убывает* по параметру θ, центральный γ-доверительный интервал для этого параметра запишется в виде

$$P\{\hat{\theta}_1(\vec{X}) < \theta < \hat{\theta}_2(\vec{X})\} = 1 - \alpha, \quad \textbf{(2.14.3)}$$

где числа $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ определяются из условий

$$F_{T_n}(t, \hat{\theta}_1(\vec{x})) = 1 - \alpha/2, \; F_{T_n}(t, \hat{\theta}_2(\vec{x})) = \alpha/2. \quad \textbf{(2.14.4)}$$

Если же ф.р. $F_{T_n}(x,\theta)$ точечной оценки $T_n = T(X_1,...,X_n)$ *возрастает* по параметру θ, то центральный γ-доверительный интервал для этого параметра также запишется в виде (2.14.3), но числа $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ определяются уже из других условий, а именно:

$$F_{T_n}(t,\hat{\theta}_1(\vec{x})) = \alpha/2\,,\ F_{T_n}(t,\hat{\theta}_2(\vec{x})) = 1-\alpha/2\,. \qquad \textbf{(2.14.5)}$$

Пример 2.14.6. Вернёмся к примеру 2.12.23 построения доверительного интервала для параметра θ равномерного распределения $R(0\,,\theta)$. Используем в качестве точечной оценки параметра θ n-ю порядковую статистику $X_{(n)}$ выборки $\vec{X}_1 = (X_1,...,X_n)$ из распределения $R(0\,,\theta)$. Напомним, что, согласно (2.12.27), функция распределения $F_{T_n}(x,\theta)$ точечной оценки $T_n = T(X_1,...,X_n)$ в виде порядковой статистики $X_{(n)}$ в данном случае записывается в виде $F_{X_{(n)}}(x,\theta) = (x/\theta)^n$, $0 \le x \le \theta$. Отметим, что эта функция непрерывная и убывающая по параметру θ и, следовательно, для определения чисел $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ в доверительном интервале (2.14.3) нужно воспользоваться условиями (2.14.4), которые в данном случае записываются в виде $(x_{(n)}/\hat{\theta}_1)^n = 1-\alpha/2$, отсюда $\hat{\theta}_1(\vec{x}) = x_{(n)}/\sqrt[n]{1-\alpha/2}$ и $(x_{(n)}/\hat{\theta}_2)^n = \alpha/2$, а $\hat{\theta}_2(\vec{x}) = x_{(n)}/\sqrt[n]{\alpha/2}$. Таким образом, центральный γ-доверительный интервал (2.14.3) в данном случае приобретает вид

$$P\{\, X_{(n)}/\sqrt[n]{1-\alpha/2} < \theta < X_{(n)}/\sqrt[n]{\alpha/2}\,\} = 1-\alpha\,, \qquad \textbf{(2.14.7)}$$

который совпадает с (2.12.26).

Обсудим теперь вариант построение асимптотических доверительных интервалов параметра θ для произвольной параметрической модели $\Im_\theta = \{F : F_X(x,\theta),\, \theta \in \Theta\}$, который также основан на использовании точечных оценок $\hat{\theta}_n$ параметра θ и их асимптотических распределений. Напомним, что для регулярных параметрических моделей оценки параметров, построенные методом максимального правдоподобия, являются асимптотически нормальными и эффективными (см. теорему 2.10.24), поэтому, если в качестве точечной оценки $\hat{\theta}_n$ параметра θ используется оценка максимального правдоподобия, то построенный асимптотический доверительный интервал будет иметь наименьшую длину при заданной доверительной вероятности.

Итак, пусть в качестве точечной оценки параметра θ используется оценка $\hat{\theta}_n$ максимального правдоподобия, тогда, согласно выражению (2.10.33), случайная величина $\sqrt{ni_1(\theta)}(\hat{\theta}_n - \theta)$ имеет стандартное нормальное распределение, где $i_1(\theta)$ – количество информации Фишера, определенное в (2.3.11), которое в общем случае зависит от оцениваемого параметра θ. По этой причине заменим $i_1(\theta)$ на $i_1(\hat{\theta}_n)$, в результате получим асимптотическую центральную статистику для параметра θ в виде $\sqrt{ni_1(\hat{\theta}_n)}(\hat{\theta}_n - \theta)$, которая, согласно замечанию (2.10.35), также имеет асимптотически стандартное нормальное распределение. Таким образом, при $n \to \infty$ выражение (2.12.3) запишется в виде

$$P\{\sqrt{ni_1(\hat{\theta}_n)}\,|\,\hat{\theta}_n - \theta\,| < \lambda_{1-\alpha/2}\} = \Phi(\lambda_{1-\alpha/2}) - \Phi(-\lambda_{1-\alpha/2}) = \\ = 2\Phi(\lambda_{1-\alpha/2}) - 1 = 2(1-\alpha/2) - 1 = 1-\alpha,$$

где $\lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ – квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения $\Phi(x)$. Отсюда получаем асимптотический γ-доверительный интервал для параметра θ в виде

$$P\{\hat{\theta}_n - \lambda_{1-\alpha/2}\ /\sqrt{ni_1(\hat{\theta}_n)} < \theta < \hat{\theta}_n + \lambda_{1-\alpha/2}\ /\sqrt{ni_1(\hat{\theta}_n)}\} = 1-\alpha. \qquad \textbf{(2.14.8)}$$

Отметим, что полученный доверительный интервал (2.14.8) является центральным и при достаточно больших объемах выборки имеет минимальную длину среди всех интервалов с доверительной вероятностью $\gamma = 1-\alpha$. Отметим также, что для заданной параметрической модели $\Im_\theta = \{F : F_X(x,\theta),\ \theta \in \Theta\}$ асимптотическая центральная статистика для параметра θ, обозначим ее через $\tilde{S}(\vec{X},\theta)$, может быть также записана непосредственно в терминах функции вклада $U(\vec{X},\theta)$, определенной в (2.3.4), или в терминах функции правдоподобия $L(\vec{X},\theta)$, определенной в (2.3.1), то есть она может быть записана в виде

$$\tilde{S}(\vec{X},\theta) = \frac{U(\vec{X},\theta)}{\sqrt{D[U(\vec{X},\theta)]}} = \frac{\dfrac{\partial}{\partial\theta}\ln L(\vec{X},\theta)}{\sqrt{D\left(\dfrac{\partial}{\partial\theta}\ln L(\vec{X},\theta)\right)}} = \sqrt{i_n(\theta)}\cdot\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(X_i,\theta). \qquad \textbf{(2.14.9)}$$

Отметим, что статистика $\tilde{S}(\vec{X},\theta)$ имеет асимптотически стандартное нормальное распределение, то есть $L\{\tilde{S}(\vec{X},\theta)\} = N(0,1)$ при $n \to \infty$, и этот факт может быть использован при построении асимптотических доверительных интервалов.

Пример 2.14.10. Пусть X – дискретная случайная величина с распределением Пуассона (см. П.1.41), то есть $L(X) = \Pi y(\theta)$. Эта запись означает, что

с.в. X пробегает значения $x = 0,1,2,...$ с вероятностями

$$f_X(x,\theta) = P(X = x) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0,1,2,...,\ \theta > 0.$$

Пусть $X_1,...,X_n$ – выборка из распределения $\Pi y(\theta)$. Использование критерия существования эффективной оценки (2.4.17) позволяет убедиться, что в данном случае для параметра θ эффективная оценка равна $T^*(X_1,...,X_n) = n^{-1}\sum_{i=1}^{n} X_i = \bar{X}$ (см. пример 2.4.22). Далее, учитывая, что в данном случае функция вклада $U(\vec{X},\theta)$ равна $U(\vec{X},\theta) = n(\bar{X} - \theta)/\theta$ (см. формулу 2.4.24), а также учитывая, что $D[U(\vec{X},\theta)] = i_n(\theta) = n/\theta$ (см. формулу 2.4.25), получаем выражение для статистики $\tilde{S}(\vec{X},\theta)$ из (2.14.9) в виде

$$\tilde{S}(\vec{X},\theta) = \sqrt{\frac{n}{\theta}}(\bar{X} - \theta). \qquad \textbf{(2.14.11)}$$

Далее, учитывая, что статистика $\tilde{S}(\vec{X},\theta)$ асимптотически нормальна, то есть $L\{\tilde{S}(\vec{X},\theta)\} = N(0,1)$ при $n \to \infty$, получим

$$P\{\sqrt{n}\,|\,\bar{X} - \theta\,|/\sqrt{\theta} < \lambda_{1-\alpha/2}\} = 1 - \alpha.$$

Отсюда, полагая $\theta = z^2$, $z > 0$, и разрешая неравенство в фигурных скобках относительно $z > 0$, получим центральный и асимптотический γ-доверительный интервал, $\gamma = 1 - \alpha$, для параметра θ распределения Пуассона в виде

$$P\{\sqrt{\bar{X} + \lambda^2_{1-\alpha/2}/4n} - \lambda_{1-\alpha/2}/2\sqrt{n} < \theta < \sqrt{\bar{X} + \lambda^2_{1-\alpha/2}/4n} + \lambda_{1-\alpha/2}/2\sqrt{n}\} = 1 - \alpha, \qquad \textbf{(2.14.12)}$$

где $\lambda_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ – квантиль уровня $(1 - \alpha/2)$ стандартного нормального распределения $\Phi(x)$.

Второй вариант использовании статистики $\tilde{S}(\vec{X},\theta)$, состоящий в замене дисперсии точечной оценки $\bar{X}$, зависящей от параметра θ, на ее выборочную оценку $\hat{\theta}_n = \bar{X}$, приводит к статистике $\sqrt{n}(\bar{X} - \theta)/\sqrt{\bar{X}}$, которая также асимптотически нормальна (см. замечание 2.10.35). Учитывая это замечание, получим

$$P\{\sqrt{n}\,|\,\bar{X} - \theta\,|/\sqrt{\bar{X}} < \lambda_{1-\alpha/2}\} = 1 - \alpha,$$

отсюда следует, что центральный и асимптотический γ-доверительный интервал, $\gamma = 1 - \alpha$, для параметра θ распределения Пуассона запишется в виде

$$P\{\,\bar{X} - \lambda_{1-\alpha/2}\sqrt{\bar{X}/n} < \theta < \bar{X} + \lambda_{1-\alpha/2}\sqrt{\bar{X}/n}\,\} = 1 - \alpha. \qquad \textbf{(2.14.13)}$$

Отметим, что полученный доверительный интервал является частным случаем интервала (2.14.8), с учетом того, что для данного примера точечная оценка параметра θ равна $\hat{\theta}_n = \overline{X}$ и количество информации Фишера $i_1(\theta)$, содержащееся в одном наблюдении относительно параметра θ, равно $i_1(\theta) = 1/\theta$.

Построенные доверительные интервалы (2.14.12) и (2.14.13) являются асимптотическими. Проиллюстрируем теперь применение описанной выше методики, основанной на использовании точечных оценок, при конечных объемах выборки на примере построения доверительного параметра θ распределения Пуассона. Пусть, как и выше, используется точечная оценка, равная $\hat{\theta}_n = \overline{X}$. Для построения доверительного интервала вида (2.14.3) необходимо записать выражение для функции распределения точечной оценки $\hat{\theta}_n = \overline{X}$ и убедиться, что она является непрерывной и монотонной функцией по θ. Для осуществления этого заметим, что если $L(X) = \Pi y(\theta)$, то, следовательно, и $L(n\overline{X}) = \Pi y(n\theta)$, и функция распределения точечной оценки $\hat{\theta}_n = \overline{X}$ запишется в виде

$$F_{\overline{X}}(x, \theta) = \sum_{i=0}^{\infty} C(x-i) e^{-n\theta} (n\theta)^i / i!.$$

Далее убедимся, что для заданной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X}_1 = (X_1, ..., X_n)$, а следовательно, и для заданной реализации $\overline{x} = k/n$, $k = 0, 1, 2, ...$, случайной величины $\overline{X}$, функция

$$F_{\overline{X}}(k/n; \theta) = \sum_{i=0}^{k} e^{-n\theta} (n\theta)^i / i!$$

непрерывна и убывает по θ. В самом деле,

$$\partial F_{\overline{X}}(k/n; \theta)/\partial\theta = \sum_{i=0}^{k} (1/i!) \cdot \{-ne^{-n\theta}(n\theta)^i + i(n\theta)^{i-1} ne^{-n\theta}\} =$$

$$= ne^{-n\theta} \sum_{i=0}^{k} \{(n\theta)^{i-1}/(i-1)! - (n\theta)^i/i!\} =$$

$$= ne^{-n\theta}\{[0-1] + [1-n\theta] + [n\theta - (n\theta)^2/2!] + [(n\theta)^2/2! - (n\theta)^3/3!] + ... +$$

$$+ [(n\theta)^{k-1}/(k-1)! - (n\theta)^k/k!]\} = -ne^{-n\theta}(n\theta)^k/k! < 0.$$

Таким образом, $\partial F_{\overline{X}}(k/n; \theta)/\partial\theta = -ne^{-n\theta}(n\theta)^k/k! < 0$, и это означает, что функция $F_{\overline{X}}(k/n; \theta)$ $\forall\theta > 0$ убывает по θ, и, следовательно, для построения доверительного интервала (2.14.3) в виде $P\{\hat{\theta}_1(\vec{X}) < \theta < \hat{\theta}_2(\vec{X})\} = 1 - \alpha$, нужно

использовать формулы (2.14.4), то есть числа $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ в данном случае определяются из следующих условий:

$$F_{\bar{X}}(k/n;\hat{\theta}_1)=\sum_{i=0}^{k}e^{-n\theta}(n\hat{\theta}_1)^i/i!=1-\alpha/2,$$

$$F_{\bar{X}}(k/n;\hat{\theta}_2)=\sum_{i=0}^{k}e^{-n\theta}(n\hat{\theta}_2)^i/i!=\alpha/2. \qquad \textbf{(2.14.14)}$$

Отметим, что построенный доверительный интервал (2.14.3) вида $P\{\hat{\theta}_1(\vec{X})<\theta<\hat{\theta}_2(\vec{X})\}=1-\alpha$, где числа $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ определяются из условий (2.14.14), описан в [66, с. 377]. Отметим также, что применение данной методики при построении доверительного интервала для параметра p распределения Бернулли $Bi(1,p)$ впервые было реализовано Клоппером и Пирсоном в 1934 году и приведено в упр. 2.15.36.

2.15. Упражнения, задачи и дополнения

2.15.1. Убедитесь, что для регулярных параметрических моделей $\Im_\theta$ информация Фишера относительно параметра θ для независимых наблюдений аддитивна.

А) Пусть случайные величины X и Y – независимы, с регулярными плотностями $f_X(x,\theta)$ и $f_Y(x,\theta)$, $x\in R^1$, $-\infty<\theta<\infty$, то есть плотности дважды дифференцируемы и их носитель не зависит от параметра θ. Пусть, далее, $I(f_X)$, $I(f_Y)$, $I(f_{XY})$ – информация Фишера, содержащаяся соответственно в случайных величинах X, Y и (XY) относительно параметра θ. При выполнении условий регулярности имеем

$$M[\partial\ln f_X(x;\theta)/\partial\theta]=0$$

и
$$I(f_X)=M[\partial\ln f_X(X;\theta)/\partial\theta]^2=-M[\partial^2\ln f_X(X;\theta)/\partial\theta^2]=$$
$$=D[\partial\ln f_X(X;\theta)/\partial\theta].$$

Необходимо убедиться, что $I(f_{XY})=I(f_X)+I(f_Y)$.

Указание. Используйте определение

$$I(XY)=M[\partial\ln f_{XY}(XY;\theta)/\partial\theta]^2=M\{[\partial\ln f_X(X;\theta)/\partial\theta]+[\partial\ln f_Y(Y;\theta)/\partial\theta]\}^2$$

и учтите, что для независимых случайных величин X и Y математическое ожидание смешанного произведения равно нулю, так как

$$M[\partial\ln f_X(x;\theta)/\partial\theta]=0.$$

Б) Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x;\theta)$ и регулярной плотностью $f_X(x;\theta)$, $\theta\in\Theta$, $x\in R^1$. Покажите, что количество информации Фишера относительно параметра θ, содержащееся в выборке $X_1,...,X_n$, равно произведению объема выборки на количество информации Фишера относительно параметра θ, содержащееся в одном наблюдении, то есть выполняется равенство $i_n(\theta)=n\cdot i_1(\theta)$.

Решение. Если $M(X)=0$, то $M(X^2)=D(X)$. Учитывая это замечание и формулу (2.3.7), получим

$$i_n(\theta)=M\left\{\frac{\partial\ln L(\vec{X};\theta)}{\partial\theta}\right\}^2=D\left\{\frac{\partial\ln L(\vec{X};\theta)}{\partial\theta}\right\}=D\left\{\sum_{i=1}^{n}\frac{\partial\ln f_X(X_i;\theta)}{\partial\theta}\right\}=$$

$$=\sum_{i=1}^{n}D\left\{\frac{\partial\ln f_X(X_i;\theta)}{\partial\theta}\right\}==\sum_{i=1}^{n}\left\{\int_{-\infty}^{+\infty}\{\partial\ln f_X(x;\theta)/\partial\theta\}^2 f_X(x;\theta)dx\right\}=n\cdot i_1(\theta).$$

2.15.2. Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x;\theta)$ и плотностью $f_X(x;\theta)$, $\theta\in\Theta$, $x\in R^1$. Предполагаем, что данная параметрическая модель регулярна, то есть плотность $f_X(x;\theta)$ непрерывна, дифференцируема и ее носитель не зависит от параметра θ. Обозначим класс несмещенных оценок заданной дифференцируемой функции $\tau(\theta)$ через $\aleph$, то есть $M[T(\vec{X})]=\tau(\theta)$, $\forall T(\vec{X})\in\aleph$.

В классе $\aleph$ несмещенных оценок функции $\tau(\theta)$ оценку $T^*(\vec{X})\in\aleph$ называют эффективной, если ее дисперсия совпадает с минимальной границей в неравенстве Рао – Крамера (2.4.4), то есть ее дисперсия равна

$$D[T^*(X_1,...,X_n)]=[\tau'(\theta)]^2/i_n(\theta). \qquad \textbf{(2.15.3)}$$

Используя критерий существования эффективной оценки (2.4.10), который записывается в виде

$$\partial\ln L(X;\theta)/\partial\theta=A^*(\theta)\cdot[T^*(\vec{X})-\tau(\theta)], \qquad \textbf{(2.15.4)}$$

покажите, что дисперсия эффективной оценки кроме формулы (2.15.3) также может быть вычислена по формуле

$$D[T^*(X_1,...,X_n)]=\left|\frac{\tau'(\theta)}{A^*(\theta)}\right|. \qquad \textbf{(2.15.4а)}$$

Решение. Возведем обе части (2.15.4) в квадрат и вычислим от обеих частей математическое ожидание. В результате получим

$$M\{\partial\ln L(X;\theta)/\partial\theta\}^2=M\{[A^*(\theta)]^2\cdot[T^*(\vec{X})-\tau(\theta)]^2\}.$$

Отсюда имеем

$$i_n(\theta) = [A^*(\theta)]^2 \, D[T^*(X_1,...,X_n)].$$

Далее, подставив это выражение в (2.15.3), получим

$$D^2[T^*(X_1,...,X_n)] = [\tau'(\theta)]^2 / [A^*(\theta)]^2.$$

Отсюда следует, что $D[T^*(X_1,...,X_n)] = |\tau'(\theta) / A^*(\theta)|$. Отметим, что из равенства формул для вычисления дисперсии эффективной оценки, то есть из равенства $\{[\tau'(\theta)]^2 / i_n(\theta)\} = \{\tau'(\theta) / A^*(\theta)\}$ следует, что количество информации Фишера $i_n(\theta)$ также можно вычислить, используя (2.15.4), по формуле

$$i_n(\theta) = \tau'(\theta) A^*(\theta). \qquad \textbf{(2.15.4б)}$$

2.15.5. Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x;\theta)$ и плотностью $f_X(x;\theta)$, $\theta \in \Theta$, $x \in R^1$. Примем для часто используемых на практике параметрических моделей следующие обозначения:

1. Пусть $L(X) = N(\mu,\sigma^2)$. Эта краткая запись означает, что с.в. X имеет нормальное распределение с параметрами $\mu = M(X)$ и $\sigma^2 = D(X)$, плотность которого записывается в виде

$$f_X(x) = (1/\sqrt{2\pi\sigma^2}) \cdot \exp\{-(x-\mu)^2 / 2\sigma^2\}, \; x \in R^1.$$

2. Пусть $L(X) = Bi(n,p)$. Эта краткая запись означает, что с.в. X имеет биномиальное распределение с параметрами n и p, то есть с.в. X – дискретного типа, равная числу «успехов» в серии из n независимых опытов с вероятностью «успеха» p $(0 < p < 1)$, и характеризуется возможными состояниями $k = 0,...,n$ и их вероятностями

$$P_n(k) = P\{X = k\} = C_n^k p^k (1-p)^{n-k}, \; k = 0,...,n.$$

3. Пусть $L(X) = \Pi y(\lambda)$. Эта краткая запись означает, что дискретная с.в. X характеризует число появлений некоторого события в заданный интервал времени или в заданной области пространства, то есть с.в. X принимает целочисленные значения $k = 0,1,2,...$ и имеет распределение Пуассона с параметром λ, то есть $P(X = k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$, $k = 0,1,2,....$

4. Пусть $L(X) = \Gamma(\alpha,\lambda)$. Эта краткая запись означает, что с.в. X имеет гамма-распределение с параметрами $\lambda > 0$ и $\alpha > 0$, плотность этого распределения записывается в виде

$$f_X(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \; 0 \le x < \infty,$$

где $\Gamma(\alpha)$ – гамма-функция, λ – масштабный параметр, α – параметр формы распределения.

Убедитесь в справедливости формул для количества информации Фишера $i_n(\theta)$ о параметре θ, для параметрических моделей, приведенных в следующей таблице (см. также примеры 2.3.17, 2.4.22 и 2.10.37).

Модель	$N(\theta,\sigma^2)$	$N(\mu,\theta^2)$	$\Gamma(\alpha,\theta)$	$Пу(\theta)$	$Bi(n,\theta)$
$i_n(\theta)$	n/σ^2	$2n/\theta^2$	$\alpha n/\theta^2$	n/θ	$n/[\theta(1-\theta)]$

Решение. Согласно (2.3.12), количество информации Фишера $i_n(\theta)$ в выборке $\vec{X}=(X_1,...,X_n)$ о параметре θ определяется в виде

$$i_n(\theta)=M\{[\partial \ln L(\vec{X};\theta)/\partial\theta]^2\}=-M\{\partial^2 \ln L(\vec{X};\theta)/\partial\theta^2\}.$$

Напомним (см. упр. 2.15.1), что для схемы повторных и независимых наблюдений выполняется равенство $i_n(\theta)=ni_1(\theta)$. Приведем вычисление $i_n(\theta)$ для биномиального распределения. Пусть в n независимых опытах, проведенных в рамках схемы испытаний Бернулли, «успех» с неизвестной вероятностью $p=\theta$ появился m раз, тогда функция правдоподобия запишется в виде $L(m;\theta)=\theta^m(1-\theta)^{n-m}$ и ее первая и вторая логарифмические производные соответственно определяются выражениями

$$\partial \ln L(m;\theta)/\partial\theta=(m-n\theta)/\theta(1-\theta),$$
$$\partial^2 \ln L(m;\theta)/\partial\theta^2=\{m(2\theta-1)-n\theta^2\}/\theta^2(1-\theta)^2.$$

Учитывая далее, что случайная величина m имеет биномиальное распределение, то есть $L(m)=Bi(n,\theta)$ и, следовательно, $M(m)=n\theta$, получаем

$$i_n(\theta)=-M\{\partial^2 \ln L(\vec{X};\theta)/\partial\theta^2\}=-M\{m(2\theta-1)-n\theta^2\}/\theta^2(1-\theta)^2=$$
$$=-\{n\theta(2\theta-1)-n\theta^2\}/\theta^2(1-\theta)^2=n/\theta(1-\theta).$$

Далее, согласно (2.4.9) и (2.4.11), выполняется равенство $\{[\tau'(\theta)]^2/i_n(\theta)\}=\{\tau'(\theta)/A^*(\theta)\}$, из которого следует, что $i_n(\theta)=\tau'(\theta)A^*(\theta)$. Для модели $N(\mu,\theta^2)$ при $\tau(\theta)=\theta^2$ коэффициент пропорциональности $A^*(\theta)=n/\theta^3$ (см. пример 2.4.21) и, следовательно,

$$i_n(\theta)=\tau'(\theta)A^*(\theta)=(\theta^2)'(n/\theta^3)=2n/\theta^2.$$

Вычисление $i_n(\theta)$ для модели $\Gamma(\alpha,\theta)$ приведено в примере 2.10.37, а для модели $Пу(\theta)$ – в примере 2.4.22.

2.15.6. Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x;\theta)$ и плотностью $f_X(x;\theta)$, $\theta\in\Theta$, $x\in R^1$. Рассмотрим случай, когда параметр θ является параметром сдвига, то есть $f_X(x;\theta)=f_X(x-\theta)$, $\theta\in R^1$, $x\in R^1$.

А. Убедитесь, что количество информации Фишера $i_1(\theta)$ о параметре сдвига θ, содержащееся в одном наблюдении, определяется только типом распределения $F_X(x)$ с плотностью $f_X(x)$, то есть не зависит от параметра θ и поэтому $i_1(\theta)$ можно переобозначить через $I(f)$ и вычислить по формулам

$$I(f)=\int_0^1 \varphi_f^2(F^{-1}(t))dt=\int_{-\infty}^{+\infty}\varphi_f'(x)dF_X(x), \qquad \textbf{(2.15.7)}$$

где $$\varphi_f(x)=-f_X'(x)/f_X(x). \qquad \textbf{(2.15.8)}$$

Б. Вычислите количество информации Фишера $I(f)$ (см. табл. 2.3.1) о параметре сдвига θ для следующих типов распределений:

1. Логистический тип с плотностью $f_X(x)=e^{-x}/(1+e^{-x})^2$, $-\infty<x+\infty$.

2. Коши с плотностью $f_X(x)=1/\pi(1+x^2)$, $-\infty<x+\infty$.

3. Лапласа с плотностью $f_X(x)=(1/2)\cdot e^{-|x|}$, $-\infty<x+\infty$.

В) Убедиться, что в случае оценивания параметра сдвига θ логистического распределения с плотностью $f_X(x)=e^{-(x-\theta)}/(1+e^{-(x-\theta)})^2$, $-\infty<x+\infty$, минимальная граница дисперсии, то есть нижняя граница в неравенстве Рао – Крамера (2.4.16), равна $3/n$, дисперсия выборочного среднего $D(\bar X)=\pi^2/3n$, а асимптотическая дисперсия выборочной медианы $\hat x_{1/2}$ – $D(\hat x_{1/2})=4/n$. Абсолютная эффективность выборочного среднего при логистическом распределении $A\Im_F(\bar X)=\{I(f)D(\bar X)\}^{-1}=\{(n/3)\cdot(\pi^2/3n)\}^{-1}=9/\pi^2=0{,}91$. Для выборочной медианы $\hat x_{1/2}$ при логистическом распределении асимптотическая абсолютная эффективность

$$A\Im_F(\hat x_{1/2})=\{I(f)D(\hat x_{1/2})\}^{-1}=\{(n/3)\cdot(4/n)\}^{-1}=3/4=0{,}75.$$

Для распределения Коши с плотностью $f_X(x)=1/\pi(1+(x-\theta)^2)$, $-\infty<x+\infty$, нижняя граница в неравенстве Рао – Крамера (2.4.16) равна $2/n$, а асимптотическая дисперсия выборочной медианы $D(\hat x_{1/2})=\pi^2/4n$, следовательно, асимптотическая абсолютная эффективность выборочной медианы при распределении Коши $A\Im_F(\hat x_{1/2})=\{I(f)D(\hat x_{1/2})\}^{-1}=\{(n/2)\cdot(\pi^2/4n)\}^{-1}=$ $=8/\pi^2=0{,}81$.

Г) Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_r(x-\theta)$, которая принадлежит множеству $\Im_r$ распределений Стьюдента (см. П.2.89) с r степенями свободы, то есть $F_r(x)\in\Im_r$. Убедиться, что информация Фишера $I(f_r)$ в одном наблюдении относительно параметра сдвига θ вычисляется по формуле

$$I(f_r)=\left(\frac{r+1}{r}\right)^2\left(\frac{r}{r+2}\right)\sqrt{\frac{r}{r+4}}\,\frac{A(r)}{A(r+4)}=$$

$$=\left(\frac{r+1}{r}\right)^2\left(\frac{r}{r+2}\right)\frac{\Gamma((r+1)/2)\Gamma((r+4)/2)}{\Gamma(r/2)\Gamma((r+5)/2)},$$

где $A(r)=\Gamma((r+1)/2)/\sqrt{r\pi}\,\Gamma(r/2)$ и $\Gamma(\cdot)$ – гамма-функция.

Решение. А) Используя определение (2.3.10), получим

$$i_1(\theta)=M\{[\partial\ln f_X(X_1;\theta)/\partial\theta]^2\}=\int_{-\infty}^{+\infty}[\partial\ln f_X(x-\theta)/\partial\theta]^2\,dF_X(x-\theta)=$$

$$=\int_{-\infty}^{+\infty}[-f_X'(x-\theta)/f_X(x-\theta)]^2\,dF_X(x-\theta)=\int_{-\infty}^{+\infty}[-f_X'(y)/f_X(y)]^2\,dF_X(y)=$$

$$=\int_0^1[-f'(F^{-1}(t))/f(F^{-1}(t))]^2\,dt=\int_0^1\varphi_f^2(F^{-1}(t))dt=I(f).$$

Далее, $$\int_{-\infty}^{+\infty}\varphi'(x)dF_X(x)=\int_{-\infty}^{+\infty}f_X(x)d\varphi(x)=f_X(x)\varphi(x)\,|_{-\infty}^{+\infty}+$$

$$+\int_{-\infty}^{+\infty}\varphi(x)[-f_X'(x)/f_X(x)]dF_X(x)=\int_0^1\varphi^2(F^{-1}(t))dt=I(f).$$

Б) Вычислим теперь количество информации Фишера $I(f)$ о параметре сдвига θ для различных типов распределений вероятностей.

1. Пусть $f_X(x)=e^{-x}/(1+e^{-x})^2$, $-\infty<x+\infty$, тогда $F_X(x)=1/(1+e^{-x})$, $\varphi_f(x)=-f_X'(x)/f_X(x)=(1-e^{-x})/(1+e^{-x})$ и $\varphi(F^{-1}(t))=2t-1$, $0\le t\le 1$.

Таким образом, для логистического распределения имеем

$$I(f)=\int_0^1\varphi^2(F^{-1}(t))dt=\int_0^1(2t-1)^2dt=1/3,$$

и, следовательно, нижняя граница дисперсии в неравенстве Рао – Крамера (2.4.16) равна $1/nI(f)=3/n$.

2. Пусть $f_X(x) = 1/\pi(1+x^2)$, $-\infty < x + \infty$, тогда функция $\varphi_f(x)$ запишется в виде

$$\varphi_f(x) = -f'_X(x)/f_X(x) = 2x/(1+x^2).$$

Таким образом, для распределения Коши имеем

$$I(f) = \int_{-\infty}^{+\infty} \varphi_f^2(x) dF_X(x) = \frac{4}{\pi} \int_{-\infty}^{+\infty} \frac{x^2 dx}{(1+x^2)^3} = \frac{4}{\pi} \cdot \frac{\pi}{8} = \frac{1}{2},$$

и, следовательно, нижняя граница дисперсии в неравенстве Рао – Крамера (2.4.16) равна $2/n$. Для распределения Коши асимптотическая дисперсия выборочной медианы $D(\hat{x}_{1/2}) = 1/4nf^2(0) = \pi^2/4n$.

3. Пусть $f_X(x) = 2^{-1}e^{-|x|} = \frac{1}{2}\begin{cases} e^x, x < 0, \\ e^{-x}, x \geq 0; \end{cases}$ $F_X(x) = \begin{cases} 2^{-1}e^x, & x < 0, \\ 1 - 2^{-1}e^{-x}, & x \geq 0. \end{cases}$

Тогда функция $\varphi_f(x)$ запишется в виде

$$\varphi(F^{-1}(t)) = \begin{cases} -1, & 0 \leq t < 1/2, \\ +1, & 1/2 \leq t \leq 1. \end{cases}$$

Таким образом, для распределения Лапласа имеем

$$I(f) = \int_0^1 \varphi_f^2(F^{-1}(t))dt = 2\int_{1/2}^1 dt = 1, D(\hat{x}_{1/2}) = 1/4nf^2(0) = 1/n.$$

Г) Для $F(x) = F_r(x)$

$$I(f_r) = \int_{-\infty}^{\infty} \left(-\frac{f_r'(x)}{f_r(x)} \right)^2 f_r(x)dx,$$

где $f_r(x)$ – плотность распределения Стьюдента с r степенями свободы, которая записывается в виде

$$f_r(x) = A(r)(1 + (x^2/r))^{-(r+1)/2}, \; A(r) = \Gamma((r+1)/2)/\sqrt{r\pi}\,\Gamma(r/2), \; x \in R^1.$$

Вычислим сначала производную плотности $f_r(x)$ и отношение $-f'_r(x)/f_r(x)$. Непосредственно получаем

$$f'_r(x) = (-(r+1)/2)A(r)(1 + (x^2/r))^{(-(r+1)/2)-1}(2x/r),$$
$$-f'_r(x)/f_r(x) = ((r+1)/r) \cdot x(1 + x^2/r)^{-1}.$$

Используя эти выражения, получим

$$I(f_r)=\int\limits_{-\infty}^{\infty}\left(-\frac{f_r'(x)}{f_r(x)}\right)^2 f_r(x)dx=\left(\frac{r+1}{r}\right)^2 A(r)\int\limits_{-\infty}^{\infty}x^2(1+\ x^2/r)^{-\frac{r+1}{2}-2}dx=$$

$$\|\,x=\sqrt{r/(r+4)}\,y\,\|==\left(\frac{r+1}{r}\right)^2 A(r)\left(\frac{r}{r+4}\right)\sqrt{\frac{r}{r+4}}\int\limits_{-\infty}^{\infty}y^2(1+\ y^2/(r+4))^{-\frac{(r+4)+1}{2}}dy.$$

Далее, учитывая, что (см. упр. 1.4.70)

$$A(r)\int\limits_{-\infty}^{\infty}x^2(1+\ x^2/r)^{-\frac{r+1}{2}}dx=r/(r-2),\ r>2,$$

из предыдущего выражения получаем

$$I(f_r)=\left(\frac{r+1}{r}\right)^2\frac{A(r)}{A(r+4)}\left(\frac{r}{r+4}\right)\sqrt{\frac{r}{r+4}}\cdot A(r+4)\int\limits_{-\infty}^{\infty}y^2(1+\ y^2/(r+4))^{-\frac{(r+4)+1}{2}}dy=$$

$$=\left(\frac{r+1}{r}\right)^2\frac{A(r)}{A(r+4)}\left(\frac{r}{r+2}\right)\sqrt{\frac{r}{r+4}}=\left(\frac{r+1}{r}\right)^2\left(\frac{r}{r+2}\right)\frac{\Gamma((r+1)/2)\Gamma((r+4)/2)}{\Gamma(r/2)\Gamma((r+5)/2)}.$$

Численные значения приведены в таблице

r	1	2	3	4	5	7	9	10	16	25	50	$r\to\infty$
$I(f_r)$	0,500	0,600	0,667	0,714	0,750	0,800	0,833	0,846	0,895	0,929	0,962	1,000

Отметим, что вычисленные значения информации Фишера для гауссовской модели с масштабным засорением $\Im_{\varepsilon,\tau}(\Phi)$ и для семейства $\Im_\lambda$-распределений Тьюки приведены в упр. 5.8.57.

2.15.9. Пусть с.в. X имеет ф.р. $F_X(x-\theta)$, $x\in R^1$, то есть предполагаем, что распределения параметрического семейства $\Im_\theta$ отличаются параметром сдвига $\theta\in R^1$ и оно задано в виде $\Im_\theta=\{F:F(x,\theta)=F_X(x-\theta)\}$, где функции распределения $F_X(x-\theta)$ с.в. X соответствует абсолютно непрерывная плотность $f_X(x-\theta)$, $x\in R^1$. Для параметрической модели с параметром сдвига количество информации Фишера $i_1(\theta)$ о параметре сдвига θ, содержащееся в одном наблюдении, определенное в общем случае в (2.3.11), не зависит от параметра $\theta\in R^1$ (см. упр. 2.15.6). Оно определяется только типом распределения, поэтому обозначим его через $I(f_X)$. Количество информации Фишера $I(f_X)$ о параметре сдвига θ записывается в виде

$$I(f_X)=\int\limits_{-\infty}^{+\infty}\left(\frac{f_X'(x)}{f_X(x)}\right)^2 f_X(x)dx<\infty\ .$$

Пусть далее с.в. Y имеет непрерывную и дифференцируемую ф.р. $F_Y(x)$, $x \in R^1$. Рассмотрим с.в. $Z = X + Y$ с плотностью f_Z, которая вычисляется с помощью интеграла свертки, то есть

$$f_Z(x) = \int_{-\infty}^{+\infty} f_X(x-y) dF_Y(y).$$

Показать, что $I(f_Z) \le I(f_X)$.

Решение. Так как плотность f_X – абсолютно непрерывная функция, то по теореме Фубини плотность f_Z также абсолютно непрерывная функция. Далее, производная плотности f_Z равна

$$f'_Z(x) = \int_{-\infty}^{+\infty} f'_X(x-y)\, dF_Y(y)$$

и, следовательно,

$$[f'_Z(x)]^2 = [\int_{-\infty}^{+\infty} f'_X(x-y) dF_Y(y)]^2 = [\int_{-\infty}^{+\infty} \frac{[f'_X(x-y)]}{\sqrt{f_X(x-y)}} \sqrt{f_Y(y)} \cdot \sqrt{f_X(x-y)} \sqrt{f_Y(y)} dy]^2 .$$

Применяя неравенство Коши – Буняковского, имеем

$$\frac{[f'_Z(x)]^2}{f_Z(x)} \le \int_{-\infty}^{+\infty} \frac{[f'_X(x-y)]^2}{f_X(x-y)} dF_Z(y).$$

Интегрируя это неравенство в бесконечных пределах, окончательно получаем

$$I(f_Z) = \int_{-\infty}^{+\infty} \left(\frac{f'_Z(x)}{f_Z(x)} \right)^2 f_Z(x) dx \le \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{[f'_X(x-y)]^2}{f_X(x-y)} dF_Z(y) dx =$$

$$= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \frac{[f'_X(x-y)]^2}{f_X(x-y)} dx \right) dF_Z(y) = I(f_X) \int_{-\infty}^{+\infty} dF_Z(y) = I(f_X).$$

Таким образом, неравенство $I(f_Z) \le I(f_X)$ доказано.

2.15.10. Пусть $L(X,Y) = NN(a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Эта запись означает, что двумерная случайная величина $Z = (X,Y)$ имеет двумерное нормальное распределение вероятностей с плотностью

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-a_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-a_X}{\sigma_X} \right) \left(\frac{y-a_Y}{\sigma_Y} \right) + \left(\frac{y-a_Y}{\sigma_Y} \right)^2 \right] \right\}, \ (x,y) \in R^2,$$

где $a_X = M(X)$, $a_Y = M(Y)$, $\sigma_X^2 = D(X)$, $\sigma_Y^2 = D(Y)$ и ρ – коэффициент корреляции. Имеется двумерная выборка $(X_1, Y_1),...,(X_n, Y_n)$ объема n. Требуется построить МП-оценки параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ по информации, содержащейся в выборке $(X_1, Y_1),...,(X_n, Y_n)$.

Решение. В данном примере для сокращения записи принята самостоятельная нумерация формул. Для заданной реализации $\vec{z} = (x_1, y_1),...,(x_n, y_n)$ двумерной выборки $(X_1, Y_1),...,(X_n, Y_n)$ функция правдоподобия запишется в виде

$$L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho) = \prod_{i=1}^{n} f_{XY}(x_i, y_i) = \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \right)^n \times$$

$$\times \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{1}{\sigma_X^2} \sum_{i=1}^{n} (x_i - a_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y} \sum_{i=1}^{n} (x_i - a_X)(y_i - a_Y) + \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (y_i - a_Y)^2 \right] \right\}.$$

Логарифмическая функция правдоподобия

$$\ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho) = -n \ln(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}) -$$

$$-\left\{ \frac{1}{2(1-\rho^2)} \left[\frac{1}{\sigma_X^2} \sum_{i=1}^{n} (x_i - a_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y} \sum_{i=1}^{n} (x_i - a_X)(y_i - a_Y) + \frac{1}{\sigma_Y^2} \sum_{i=1}^{n} (y_i - a_Y)^2 \right] \right\}.$$

Система уравнений (2.10.5) в данном случае записывается в виде

$$\frac{\partial \ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\partial a_X} =$$

$$= -\frac{1}{2(1-\rho^2)} \left[\frac{-2}{\sigma_X^2} \sum_{i=1}^{n} (x_i - a_X) + \frac{2\rho}{\sigma_X\sigma_Y} \sum_{i=1}^{n} (y_i - a_Y) \right] = 0; \qquad (1)$$

$$\frac{\partial \ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\partial a_Y} =$$

$$-\frac{1}{2(1-\rho^2)} \left[\frac{2\rho}{\sigma_X\sigma_Y} \sum_{i=1}^{n} (x_i - a_X) - \frac{2}{\sigma_Y^2} \sum_{i=1}^{n} (y_i - a_Y) \right] = 0; \qquad (2)$$

$$\frac{\partial \ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\partial \sigma_X} =$$

$$= -\frac{n}{\sigma_X} - \frac{1}{2(1-\rho^2)} \left[-\frac{2\sigma_X}{(\sigma_X^2)^2} \sum_{i=1}^{n} (x_i - a_X)^2 + \frac{2\rho}{\sigma_X^2\sigma_Y} \sum_{i=1}^{n} (x_i - a_X)(y_i - a_Y) \right] = 0; \qquad (3)$$

$$\frac{\partial \ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\partial \sigma_Y} =$$

$$= -\frac{n}{\sigma_Y} - \frac{1}{2(1-\rho^2)}\left[\frac{2\rho}{\sigma_X \sigma_Y^2}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) - \frac{2}{\sigma_Y^3}\sum_{i=1}^{n}(y_i - a_Y)^2\right] = 0; \qquad (4)$$

$$\frac{\partial \ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)}{\partial \rho} = \frac{n\rho}{1-\rho^2} + \frac{1}{(1-\rho^2)\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) -$$

$$-\frac{\rho}{(1-\rho^2)^2}\left[-\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - a_X)^2 - \frac{2\rho}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) + \frac{1}{\sigma_Y}\sum_{i=1}^{n}(y_i - a_Y)^2\right] = 0. \qquad (5)$$

Для нахождения МП-оценок параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ необходимо решить эту систему из пяти уравнений. Из уравнений (1) и (2) непосредственно следует, что $\sum_{i=1}^{n}(x_i - a_X) = 0$, $\sum_{i=1}^{n}(y_i - a_Y) = 0$ и, следовательно, МП-оценки параметров a_X, a_Y запишутся в виде

$$\hat{a}_X = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{a}_Y = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Далее, уравнение (3) перепишем в виде

$$-\frac{n(1-\rho^2)}{\sigma_X(1-\rho^2)} - \frac{1}{(1-\rho^2)\sigma_X}\left[-\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - a_X)^2 + \frac{\rho}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y)\right] = 0.$$

Отсюда получаем

$$\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - a_X)^2 = \frac{\rho}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) + n(1-\rho^2). \qquad (6)$$

Аналогично из (4) получаем

$$\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i - a_Y)^2 = \frac{\rho}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) + n(1-\rho^2). \qquad (7)$$

Складывая соответствующие члены (6) и (7), получим

$$\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - a_X)^2 - \frac{2\rho}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) + \frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i - a_Y) = 2n(1-\rho^2). \qquad (8)$$

Умножая теперь обе части (5) на $(1-\rho^2)$ и учитывая (8), получим

$$n\rho + \frac{1}{\sigma_X \sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) - \frac{\rho}{(1-\rho^2)} \cdot 2n(1-\rho^2) = 0,$$

или
$$\frac{1}{\sigma_X\sigma_Y}\sum_{i=1}^{n}(x_i-a_X)(y_i-a_Y)=n\rho, \tag{9}$$

или
$$\frac{\rho}{\sigma_X\sigma_Y}\sum_{i=1}^{n}(x_i-a_X)(y_i-a_Y)=n\rho^2. \tag{10}$$

Далее, из (6) с учетом (10) получим $\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i-a_X)^2=n\rho^2+n(1-\rho^2)$. Отсюда следует, что МП-оценка параметра σ_X^2 имеет вид

$$\hat{\sigma}_X^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{a}_X)^2.$$

Аналогично из (7) с учетом (10) получим

$$\frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i-a_Y)^2=n\rho^2+n(1-\rho^2),$$

следовательно, МП-оценка параметра σ_Y^2 имеет вид

$$\hat{\sigma}_Y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}_Y)^2.$$

Окончательно, из формулы (9) получаем МП-оценку параметра ρ в виде

$$\hat{\rho}=\frac{\sum(x_i-\hat{a}_X)(y_i-\hat{a}_Y)}{\sqrt{\sum(x_i-\hat{a}_X)^2\sum(y_i-\hat{a}_Y)^2}}.$$

Отметим, что полученные оценки максимального правдоподобия параметров a_X, a_Y, σ_X^2, σ_Y^2 и ρ совпадают с выборочными оценками этих параметров, полученными методом подстановки (см. упр. 2.15.11 – 2.15.15).

2.15.11. Пусть $\vec{X}=(X_1,...,X_n)$ – выборка из распределения $F_X(x)$ с плотностью $f_X(x)$, $x\in R^1$. Построить методом подстановки выборочную оценку $T(F_n)$ функционала $T(F)$ вида

$$T(F)=\int_{-\infty}^{+\infty}\varphi(x)dF(x),$$

где $\varphi(x)$ – заданная непрерывная функция и F_n – эмпирическая функция распределения. Убедитесь, что $T(F_n)$ является несмещенной и состоятельной оценкой функционала $T(F)$. Записать асимптотическое распределение $T(F_n)$.

Решение. Для построения выборочной оценки функционала $T(F)$ методом подстановки следует заменить в этом функционале неизвестную ф.р. $F_X(x)$ на эмпирическую ф.р. $F_n(x)$. В результате этих действий, с учетом (1.1.27) и (1.1.28), получаем

$$T(F_n) = \int_{-\infty}^{+\infty} \varphi(x) dF_n(x) = \int_{-\infty}^{+\infty} \varphi(x) d[n^{-1} \sum_{i=1}^{n} C(x - X_i)] =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} \varphi(x) \delta(x - X_i) dx = \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i) .$$

Далее, учитывая, что $X_1,...,X_n$ – н.о.р. случайные величины и, следовательно, $f_{X_i}(x) = f_X(x)$, $\forall i = 1,...,n$, получаем непосредственное доказательство несмещенности оценки $T(F_n)$ в виде

$$M[T(F_n)] = M[n^{-1} \sum_{i=1}^{n} \varphi(X_i)] = n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} \varphi(x) f_{X_i}(x) dx =$$

$$= n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} \varphi(x) dF_X(x) = T(F) .$$

Для доказательства состоятельности оценки $T(F_n)$ вычислим сначала ее дисперсию. Учитывая, что $X_1,...,X_n$ – н.о.р. случайные величины, получаем

$$D[T(F_n)] = D[n^{-1} \sum_{i=1}^{n} \varphi(X_i)] = n^{-2} \sum_{i=1}^{n} D[\varphi(X_i)] = \sigma_F^2(\varphi)/n ,$$

где $\sigma_F^2(\varphi)$ – константа, которая зависит от ф.р. F, функции φ и определяется в виде

$$\sigma_F^2(\varphi) = \int_{-\infty}^{+\infty} \varphi^2(x) dF(x) - (\int_{-\infty}^{+\infty} \varphi(x) dF(x))^2 = \int_{-\infty}^{+\infty} [\varphi(x) - T(F)]^2 dF(x) .$$

Предполагая далее, что $\sigma_F^2(\varphi) < \infty$, запишем неравенство Чебышева для оценки $T(F_n)$. Для $\forall \varepsilon > 0$ имеем

$$P\{|T(F_n) - T(F)| < \varepsilon\} \geq 1 - \frac{\sigma_F^2(\varphi)}{n \cdot \varepsilon^2} .$$

Отсюда получаем, что $P\{|T(F_n) - T(F)| < \varepsilon\} \to 1$ при $n \to \infty$ и, следовательно, $T(F_n) \xrightarrow{P} T(F)$, то есть $T(F_n)$ является состоятельной оценкой функционала $T(F)$.

Асимптотическое распределение оценки $T(F_n)$ следует из центральной предельной теоремы (см. П.3.27), согласно которой сумма независимых и одинаково распределенных случайных величин имеет нормальное распределение при увеличении числа слагаемых. Таким образом, при выполнении неравенства $\sigma_F^2(\varphi) < \infty$ и при $n \to \infty$ справедливо следующее выражение:

$$L\left(\frac{\sqrt{n}(T(F_n) - T(F))}{\sigma_F(\varphi)} \right) = N(0\,,1)\,.$$

Рассмотрим в качестве примера функцию $\varphi(x) = x^k$, $k = 1,2....$ При такой функции $\varphi(x)$ функционал $T(F)$ характеризует начальный момент *k*-го порядка в виде

$$\alpha_k = M(X^k) = \int_{-\infty}^{+\infty} x^k dF_X(x),\ k = 1,2...\ .$$

Выборочная оценка $\hat{\alpha}_k$ начального момента *k*-го порядка α_k записывается в виде

$$\hat{\alpha}_k = n^{-1} \sum_{i=1}^{n} X_i^k\ ,\ k = 1,2...\ .$$

Отметим, что $\hat{\alpha}_k$ является несмещенной и состоятельной оценкой α_k, то есть $M(\hat{\alpha}_k) = \alpha_k$, $k = 1,2...$, $\hat{\alpha}_k \to^P \alpha_k$ при $n \to \infty$, и ее асимптотическое распределение записывается в виде

$$L\left(\frac{\sqrt{n}(\hat{\alpha}_k - \alpha_k)}{\sqrt{\alpha_{2k} - \alpha_k^2}} \right) = N(0\,;1)\,,\ k = 1,2...\ ,\ n \to \infty\,.$$

В частности, при $k = 1$ получаем начальный момент 1-го порядка, который является математическим ожиданием с.в. X, то есть $\alpha_1 = M(X)$. Выборочная оценка $\hat{\alpha}_1$ математического ожидания $\alpha_1 = M(X)$ обычно обозначается через $\overline{X}$ и является средним арифметическим элементов выборки $X_1,...,X_n$, то есть $\hat{\alpha}_1 = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Итак, выборочное среднее $\overline{X}$ является несмещенной и состоятельной оценкой математического ожидания с.в. X, причем её дисперсия в n раз меньше дисперсии с.в. X, то есть

$$D(\overline{X}) = D(X)/n\,.$$

Кроме того, для любой ф.р. F с конечной дисперсией $D(X)$ выборочное среднее $\bar{X}$ имеет асимптотически нормальное распределение, то есть

$$L\left(\frac{\sqrt{n}(\bar{X}-M(X))}{\sqrt{D(X)}}\right)=N(0\,,1),\ n\to\infty.$$

Отметим также, что если выборка $X_1,...,X_n$ взята из *нормального распределения* $N(a,\sigma^2)$, тогда выборочное среднее $\bar{X}$ имеет нормальное распределение при любом *конечном* объеме выборки n (см. теорему 2.8.31), то есть справедливо выражение $L(\sqrt{n}(\bar{X}-a)/\sigma)=N(0;1),\ \forall n\in N$.

2.15.12. Пусть $\vec{X}=(X_1,...,X_n)$ – выборка из распределения $F_X(x)$ с плотностью $f_X(x)$, $x\in R^1$. Построим методом подстановки выборочную оценку $S(F_n)$ функционала $S(F)$ вида

$$S(F)=\int_{-\infty}^{+\infty}\psi(x,T(F))dF(x),$$

где $\psi(x,T(F))$ – заданная функция и $T(F)$ – функционал, характеризующий параметр положения ф.р. F_X. Рассмотрим частный случай функции $\psi(x,T(F))$ в виде $\psi(x,\alpha_1(F))=(x-\alpha_1)^k$, $\alpha_1=M(X)$, $k=1,2,...$, при котором функционалы $S(F)$ соответствуют центральным моментам μ_k k-го порядка, определяемым в виде

$$\mu_k=M[(X-\alpha_1)^k]=\int_{-\infty}^{+\infty}(x-\alpha_1)^k dF_X(x),\ k=1,2...\ .$$

Выборочная оценка функционала $S(F)$, построенная методом подстановки, с учетом (1.3.27) и (1.3.28), запишется как

$$S(F_n)=\int_{-\infty}^{+\infty}\psi(x,T(F_n))dF_n(x)=\int_{-\infty}^{+\infty}\psi(x,T(F_n))d[n^{-1}\sum_{i=1}^{n}C(x-X_i)]=$$

$$=\frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}\psi(x,T(F_n))\delta(x-X_i)dx=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,T(F_n)).$$

Выборочная оценка $\hat{\mu}_k$ центрального момента k-го порядка μ_k, $k=1,2...$, запишется в виде

$$\hat{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k\ ,\ k=1,2...,$$

и, в частности, выборочная дисперсия $\hat{\mu}_2$, обычно обозначаемая через S^2, примет вид

$$\hat{\mu}_2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 .$$

2.15.13. Пусть $\vec{X} = (X_1,...,X_n)$ – выборка из распределения $F_X(x)$ с плотностью $f_X(x)$, $x \in R^1$, причем $M(X) = a$ и $D(X) = \sigma^2$.

(А) Убедитесь, что выборочное среднее $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ является несмещенной оценкой математического ожидания $M(X) = a$.

(Б) Убедитесь, что выборочная дисперсия $S^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ является смещенной оценкой дисперсии $D(X) = \sigma^2$.

(В) Убедитесь, что подправленная выборочная дисперсия

$$\tilde{S}^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right\}$$

является несмещенной оценкой дисперсии $D(X) = \sigma^2$, то есть $M(\tilde{S}^2) = D(X)$.

Решение. Несмещенность выборочного среднего $\bar{X}$ была установлена в предыдущем упражнении и также может быть проверена непосредственно следующим образом:

$$M(\bar{X}) = n^{-1}\sum_{i=1}^{n} M(X_i) = n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} x dF_{X_i}(x) = n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} x dF_X(x) = n^{-1}\sum_{i=1}^{n} M(X) = a .$$

Для проверки того факта, что выборочная дисперсия S^2 является смещенной оценкой дисперсии $D(X) = \sigma^2$, прежде преобразуем эту оценку к более удобному для вычисления математического ожидания виду, в котором случайные величины X_i, $i = 1,...,n$ и $\bar{X}$ разнесены по разным слагаемым, то есть

$$S^2 = n^{-1}\sum_{i=1}^{n}[(X_i - a) - (\bar{X} - a)]^2 = n^{-1}\sum_{i=1}^{n}(X_i - a)^2 - 2n^{-1}\sum_{i=1}^{n}(X_i - a)(\bar{X} - a) +$$

$$+ n^{-1}\sum_{i=1}^{n}(\bar{X} - a)^2 = n^{-1}\sum_{i=1}^{n}(X_i - a)^2 - (\bar{X} - a)^2 .$$

Отсюда, учитывая, что $X_1,...,X_n$ – н.о.р. случайные величины и $D(\bar{X}) = D(X)/n$, получаем

$$M(S^2) = n^{-1}\sum_{i=1}^{n} M[(X_i - a)^2] - M[(\bar{X} - a)^2] = n^{-1}\sum_{i=1}^{n} D(X) - D(\bar{X}) =$$

$$= D(X) - D(X)/n = \frac{n-1}{n} D(X) \neq \sigma^2 .$$

Таким образом, $M(S^2) \neq D(X)$, то есть выборочная дисперсия S^2, является смещенной оценкой дисперсии $D(X)$, но она является асимптотически несмещенной оценкой дисперсии. Отметим, что при небольших объемах выборки ($n < 30$) обычно используют подправленную выборочную дисперсию

$$\tilde{S}^2 = n/(n-1)S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 ,$$

которая может быть получена методом «складного ножа» из выборочной дисперсии S^2 (см. замечание 6.5.22 и упр. 2.15.57) и которая является несмещенной оценкой дисперсии $D(X)$ при любом объеме выборки. В самом деле,

$$M[\tilde{S}^2] = n/(n-1)M[S^2] = \frac{n}{n-1}\cdot\frac{n-1}{n} D(X) = D(X) = \sigma^2 .$$

Отметим также, что при *известном* математическом ожидании $M(X) = a$ выборочная дисперсия

$$S_a^2 = n^{-1}\sum_{i=1}^{n}(X_i - a)^2$$

является несмещенной оценкой дисперсии $D(X)$, то есть

$$M(S_a^2) = n^{-1}\sum_{i=1}^{n} M[(X_i - a)^2] = D(X) = \sigma^2 .$$

2.15.14. Пусть $\vec{X} = (X_1, ..., X_n)$ – выборка из распределения $F_X(x)$ с плотностью $f_X(x)$, $x \in R^1$. Построить методом подстановки выборочную оценку коэффициента асимметрии

$$\gamma_1(F) = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\int (x-\alpha_1)^3 dF(x)}{\left\{\int (x-\alpha_1)^2 dF(x)\right\}^{3/2}}$$

и коэффициента эксцесса $\gamma_2(F) = \mu_4/\mu_2^2$, который с учетом того, что $\gamma_2(\Phi) = 3$, где Φ – стандартное нормальное распределение, обычно записывают в виде

$$\tilde{\gamma}_2(F) = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\int (x-\alpha_1)^4 dF(x)}{\left\{\int (x-\alpha_1)^2 dF(x)\right\}^2} - 3 .$$

Решение. Выборочный коэффициент асимметрии $\gamma_1(F_n)$, построенный по выборке $X_1,...,X_n$ методом подстановки, имеет вид

$$\gamma_1(F_n)=\frac{\int(x-\bar{X})^3 dF_n(x)}{\left\{\int(x-\bar{X})^2 dF_n(x)\right\}^{3/2}}=\frac{n^{-1}\sum(X_i-\bar{X})^3}{\left\{n^{-1}\sum(X_i-\bar{X})^2\right\}^{3/2}}=\frac{\sqrt{n}\sum(X_i-\bar{X})^3}{\left\{\sum(X_i-\bar{X})^2\right\}^{3/2}}.$$

Аналогично, выборочный коэффициент эксцесса $\gamma_2(F_n)$ запишется как

$$\gamma_2(F_n)=\frac{\int(x-\bar{X})^4 dF_n(x)}{\left\{\int(x-\bar{X})^2 dF_n(x)\right\}^{2}}=\frac{n^{-1}\sum(X_i-\bar{X})^4}{\left\{n^{-1}\sum(X_i-\bar{X})^2\right\}^{2}}=\frac{n\sum(X_i-\bar{X})^4}{\left\{\sum(X_i-\bar{X})^2\right\}^{2}}.$$

2.15.15. Имеется двумерная выборка $(X_1,Y_1),...,(X_n,Y_n)$ объема n из распределения $F_{XY}(x,y)$ с плотностью $f_{XY}(x,y)$, $(x,y)\in R^2$. Постройте методом подстановки выборочную оценку коэффициента корреляции $\rho_{XY}=\dfrac{COV(XY)}{\sqrt{D(X)D(Y)}}$.

Решение. Согласно (2.7.22), двумерная эмпирическая функция распределения $F_n(x,y)$, построенная по выборке $(X_1,Y_1),...,(X_n,Y_n)$, записывается в виде

$$F_n(x,y)=n^{-1}\sum_{i=1}^{n}C(x-X_i)C(y-Y_i).$$

Обозначим

$$dF_n(x,y)=(\partial^2F_n(x,y)/\partial x\partial y)dxdy=n^{-1}\sum_{i=1}^{n}\delta(x-X_i)\delta(y-Y_i)dxdy\ .$$

Подставив в выражение для ковариации

$$COV(XY)=\int\int(x-M(X))(y-M(Y)f(x,y)dxdy=$$
$$=\int\int(x-M(X))(y-M(Y))\{\partial^2F(x,y)/\partial x\partial y\}dxdy$$

вместо математических ожиданий $M(X)$ и $M(Y)$ их выборочные оценки $\bar{X}$ и $\bar{Y}$, а вместо неизвестной ф.р. $F(x,y)$ эмпирическую ф.р. $F_n(x,y)$, получим выборочную оценку ковариации в виде

$$COV_n(XY)=\int\int(x-\bar{X})(y-\bar{Y})dF_n(x,y)=$$
$$=n^{-1}\sum\int\int(x-\bar{X})(y-\bar{Y})\delta(x-X_i)\delta(y-Y_i)dxdy=n^{-1}\sum(X_i-\bar{X})(Y_i-\bar{Y}).$$

Далее, используя выборочные оценки S_X^2 и S_Y^2 дисперсий случайных величин

X и Y, построенные по выборке $(X_1,Y_1),...,(X_n,Y_n)$, окончательно получаем выборочную оценку $\hat{\rho}_{XY}$ коэффициента корреляции в виде

$$\hat{\rho}_{XY}=\frac{COV_n(XY)}{\sqrt{S_X^2S_Y^2}}=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}=\frac{\sum X_iY_i-n\overline{X}\overline{Y}}{\sqrt{(\sum X_i^2-n\overline{X}^2)(\sum Y_i^2-n\overline{Y}^2)}}.$$

2.15.16. Пусть $L(X)=N(a_X,\sigma_X^2)$ и $L(Y)=N(a_Y,\sigma_Y^2)$. Имеется двумерная выборка $(X_1,Y_1),...,(X_n,Y_n)$ объема n из двумерного нормального распределения, то есть предполагаем, что $L(X,Y)=NN(a_X,a_Y,\sigma_X^2,\sigma_Y^2,\rho)$, и эта запись означает, что совместная плотность $f_{XY}(x,y)$ двумерной случайной величины (X,Y) записывается в виде

$$f_{XY}(x,y)=(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}\times\exp\{-[\tilde{x}^2-2\rho\tilde{x}\tilde{y}+\tilde{y}^2]/2(1-\rho^2)\},\quad\textbf{(2.15.17)}$$

где $a_X=M(X)$, $a_Y=M(Y)$, $\sigma_X^2=D(X)$, $\sigma_Y^2=D(Y)$, ρ – коэффициент корреляции и $\tilde{x}=(x-a_X)/\sigma_X$, $\tilde{y}=(y-a_Y)/\sigma_Y$. Получите выражение для функции регрессии X на Y и постройте ее выборочную оценку по выборке $(X_1,Y_1),...,(X_n,Y_n)$ методом подстановки.

Решение. Получим сначала выражение для функции регрессии случайной величины X на Y, которая, согласно определению (1.1.51), записывается в виде условного математического ожидания вида

$$M(X\mid Y=y)=\int_{-\infty}^{+\infty}xf_{X|Y}(x\mid y)dx.$$

Для вычисления этого математического ожидания конкретизируем условную плотность $f_{X|Y}(x\mid y)$. Согласно определению (1.1.41), с учетом (2.15.17) и выражения $f_Y(y)=(\sqrt{2\pi}\sigma_Y)^{-1}\times\exp\{-\tilde{y}^2/2\}$, имеем

$$f_{X|Y}(x\mid y)=f_{XY}(x,y)/f_Y(y)=$$

$$=(\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2})^{-1}\times\exp\{-[\tilde{x}^2-2\rho\tilde{x}\tilde{y}+\tilde{y}^2]/2(1-\rho^2)+\tilde{y}^2/2\}=$$

$$=(\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2})^{-1}\times\exp\{-[\tilde{x}^2-2\rho\tilde{x}\tilde{y}+\tilde{y}^2-(1-\rho^2)\tilde{y}^2]/2(1-\rho^2)\}=$$

$$=(\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2})^{-1}\times\exp\{-(\tilde{x}-\rho\,\tilde{y})^2/2(1-\rho^2)\}=$$

$$=(\sqrt{2\pi}\,\sigma_X\sqrt{1-\rho^2})^{-1}\times\exp\{-(x-[a_X+\rho\sigma_X(y-a_Y)/\sigma_Y])^2/2\sigma_X^2(1-\rho^2)\}.$$

Отсюда заключаем, что условная плотность $f_{X|Y}(x\mid y)$ имеет вид нормальной плотности с параметрами $M(X\mid Y=y)$ и $D(X\mid Y=y)$, то есть записывается

в виде

$$f_{X|Y}(x \mid y) = \frac{1}{\sqrt{2\pi\, D(X \mid Y = y)}} \exp\left\{ -\frac{(x - M(X \mid Y = y))^2}{2D(X \mid Y = y)} \right\}, \qquad \textbf{(2.15.18)}$$

где $M(X \mid Y = y)$ – условное математическое ожидание, зависимость которого от y конкретизируется ниже, и $D(X \mid Y = y)$ – условная дисперсия с.в. X, которая не зависит от y и определяется в виде,

$$D(X \mid Y = y) = \sigma_X^2 (1 - \rho^2).$$

Отметим, что последнее выражение для условной дисперсии позволяет привести следующую важную интерпретацию коэффициента корреляции ρ. Напомним, что σ_X^2 обозначает безусловную дисперсию с.в. X (то есть это дисперсия X, когда значение Y неизвестно). С другой стороны, $D(X \mid Y = y)$ есть условная дисперсия с.в. X (то есть это дисперсия X, когда известно, что соответствующее значение $Y = y$). Далее, из выражения $D(X \mid Y = y) = \sigma_X^2 - \sigma_X^2 \rho^2$ следует, что уменьшение дисперсии X, вызванное знанием Y, есть $\sigma_X^2 - D(X \mid Y = y) = \sigma_X^2 \rho^2$. Отсюда получаем, что квадрат коэффициента корреляции ρ равен

$$\rho^2 = \{\sigma_X^2 - D(X \mid Y = y)\} / \sigma_X^2.$$

Таким образом, полученное равенство позволяет утверждать, что квадрат коэффициента корреляции характеризует долю дисперсии X, которая объясняется знанием Y. В тех случаях, когда $\rho = 0$, то есть случайные величины X и Y некоррелированы, тогда условная дисперсия совпадает с безусловной, то есть $D(X \mid Y = y) = \sigma_X^2$. Это равенство означает, что никакая доля дисперсии X не объясняется регрессией X на Y. Если же $\rho = \pm 1$, то условная дисперсия $D(X \mid Y = y) = 0$. В этом случае вся дисперсия X объясняется регрессией X на Y и зависимость между случайными величинами X и Y является точно линейной (см. также упр. 1.4.10).

Далее, $M(X \mid Y = y)$ – условное математическое ожидание с.в. X, при условии, что $Y = y$, которое является функцией от y, его называют функцией регрессии X на Y и с учетом (2.15.18), она определяется в виде

$$M(X \mid Y = y) = \int_{-\infty}^{+\infty} x f_{X|Y}(x \mid y) dx = [a_X + \rho \sigma_X (y - a_Y)/\sigma_Y] =$$

$$M(X) + \rho \sqrt{D(X)/D(Y)}\, (y - M(Y)).$$

Выборочная оценка полученного выражения для функции регрессии случай-

ной величины X на Y, построенная по выборке $(X_1,Y_1),...,(X_n,Y_n)$ методом подстановки, запишется в виде

$$\hat{M}(X \mid Y=y)=\overline{X}+\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}\cdot\frac{\sqrt{\sum(X_i-\overline{X})^2}}{\sqrt{\sum(Y_i-\overline{Y})^2}}\cdot(y-\overline{Y})=$$

$$=\overline{X}+\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(Y_i-\overline{Y})^2}\cdot(y-\overline{Y}),\ \text{где}\ \overline{X}=n^{-1}\sum_{i=1}^{n}X_i,\ \overline{Y}=n^{-1}\sum_{i=1}^{n}Y_i. \qquad \textbf{(2.15.18а)}$$

Аналогично получаем выборочную оценку регрессии Y на X (см. также формулу (2.15.50)) в виде

$$\hat{M}(Y \mid X=x)= =\overline{Y}+\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}(x-\overline{X})=$$

$$=\overline{Y}-\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}\cdot\overline{X}+\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}\cdot x=\hat{\theta}_1 x+\hat{\theta}_2, \qquad \textbf{(2.15.18б)}$$

где $$\hat{\theta}_2=\overline{Y}-\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}\cdot\overline{X},\ \hat{\theta}_1=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2}. \qquad \textbf{(2.15.18в)}$$

2.15.19. Покажите, что среднеквадратическая ошибка оценки $\hat{\theta}$ параметра θ равна сумме дисперсии оценки и квадрата ее смещения, то есть $CKO(\hat{\theta})=M[(\hat{\theta}-\theta)^2]= D(\hat{\theta})+b^2_{\hat{\theta}}(\theta)$.

Решение.

$$CKO(\hat{\theta})=M[(\hat{\theta}-\theta)^2]=M\{[\hat{\theta}-M(\hat{\theta})]+[M(\hat{\theta})-\theta]\}^2=$$

$$=M\{[\hat{\theta}-M(\hat{\theta})]^2\}+2[M(\hat{\theta})-\theta]M[\hat{\theta}-M(\hat{\theta})]+[M(\hat{\theta})-\theta]^2=D(\hat{\theta})+b^2_{\hat{\theta}}(\theta),$$

где $b_{\hat{\theta}}(\theta)=[M(\hat{\theta})-\theta]$ – смещение оценки $\hat{\theta}$ параметра θ, которое является функцией от оцениваемого параметра θ.

2.15.20. Пусть $L(X)=N(a,\sigma^2)$ и $\vec{X}=(X_1,...,X_n)$ – выборка из нормального распределения, параметры которого $M(X)=a$ и $D(X)=\sigma^2$ неизвестны, то есть мы имеем общую нормальную модель $N(\theta_1,\theta_2^2)$, которая задана в параметрической форме с точностью до неизвестного векторного параметра $\vec{\theta}=(\theta_1,\theta_2)$, причем $\theta_1=M(X)=a$, $\theta_2^2=D(X)=\sigma^2$ и плотность записывается в виде

$$f(x,\vec{\theta})=(1/\sqrt{2\pi}\,\theta_2)\exp\{-(x-\theta_1)^2/2\theta_2^2\},\ x\in R^1.$$

Убедитесь в справедливости следующих утверждений:

А) Выборочное среднее $\bar{X}$ является эффективной оценкой параметра $\theta_1 = a$, причем статистика $\bar{X}$ наилучшим образом оценивает неизвестное среднее нормальной модели независимо от того, известна дисперсия σ^2 или нет.

Б) Подправленная выборочная дисперсия $\tilde{S}^2 = (n-1)^{-1}\sum(X_i - \bar{X})^2$ не является эффективной оценкой дисперсии σ^2, однако она является оптимальной по критерию Бхаттачария (П.3.22) оценкой параметрической функции $\tau(\theta) = \theta_2^2$ (то есть дисперсии σ^2 общей нормальной модели $N(\theta_1, \theta_2^2)$ при $\theta_2^2 = \sigma^2$). Это означает, что дисперсия $D(\tilde{S}^2)$ подправленной выборочной дисперсии $\tilde{S}^2$ совпадает с нижней границей в неравенстве Бхаттачария (2.4.35) для случая векторного параметра.

В) Вычислить среднеквадратические ошибки выборочной дисперсии, подправленной выборочной дисперсии и убедиться в справедливости неравенства $CKO(S^2) < CKO(\tilde{S}^2)$.

Решение. А) Для общей нормальной модели $N(\theta_1, \theta_2^2)$ с плотностью вида

$$f(x, \vec{\theta}) = (1/\sqrt{2\pi}\,\theta_2)\exp\{-(x-\theta_1)^2/2\theta_2^2\},\ x \in R^1,$$

имеем

$$\ln f(x,\vec{\theta}) = \ln(1/\sqrt{2\pi}) - \ln\theta_2 - (x-\theta_1)^2/2\theta_2^2,\ \partial \ln f(x,\vec{\theta})/\partial\theta_1 = (x-\theta_1)/\theta_2^2,$$
$$\partial^2 \ln f(x,\vec{\theta})/\partial\theta_1^2 = -1/\theta_2^2,\ \partial^2 \ln f(x,\vec{\theta})/\partial\theta_1\partial\theta_2 = (-2/\theta_2^3)(x-\theta_1),$$
$$\partial \ln f(x,\vec{\theta})/\partial\theta_2 = (-1/\theta_2) + (x-\theta_1)^2/\theta_2^3,$$
$$\partial^2 \ln f(x,\vec{\theta})/\partial\theta_2^2 = (1/\theta_2^2) - (3/\theta_2^4)(x-\theta_1)^2.$$

Далее, при заданной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X} = (X_1, ..., X_n)$ функция правдоподобия запишется в виде

$$L(\vec{x};\theta_1,\theta_2) = (\sqrt{2\pi}\theta_2)^{-n}\exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right\}$$

и ее логарифм определяется выражением

$$\ln L(\vec{x};\theta_1,\theta_2) = -n\ln(\sqrt{2\pi}\theta_2) - \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2.$$

Используя данные выражения, непосредственно убеждаемся, что

$$\partial \ln L(\vec{x};\theta_1,\theta_2)/\partial\theta_1 = \frac{1}{\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1) = (n/\sigma^2)(\bar{X}-a).$$

Отсюда, согласно (2.4.36), заключаем, что выборочное среднее $\overline{X}$ является эффективной оценкой математического ожидания, то есть параметра $\theta_1 = a$.

Б) Напомним, что несмещенной и эффективной оценкой дисперсии в нормальной модели при известном математическом ожидании a является выборочная дисперсия $S_a^2 = n^{-1}\sum(X_i - a)^2$ (см. пример 2.4.21). Если же параметр $\theta_1 = a$ неизвестен, то выборочная дисперсия $S^2 = n^{-1}\sum(X_i - \overline{X})^2$ является смещенной оценкой дисперсии, а подправленная выборочная дисперсия $\tilde{S}^2 = (n-1)^{-1}\sum(X_i - \overline{X})^2$ является несмещенной оценкой дисперсии (см. упр. 2.15.13). Убедимся теперь, что подправленная выборочная дисперсия $\tilde{S}^2$ не является эффективной оценкой параметрической функции $\tau(\theta) = \theta_2^2$, сравнив ее дисперсию $D(\tilde{S}^2) = 2\sigma^4/(n-1)$ (см. пункт В), с нижней границей в неравенстве (2.4.35). Для определения этой границы сначала вычислим информационную матрицу $I_n(\vec{\theta}) = (g_{ij})$, $i, j = 1,..., r$, для выборки объема n, элементы $g_{ij}(\vec{\theta})$ которой, согласно (2.4.37), вычисляются по формулам

$$g_{ij}(\theta) = -M_\theta\left(\frac{\partial^2 \ln L(\vec{X};\vec{\theta})}{\partial\theta_i\partial\theta_j}\right) = -nM_\theta\left(\frac{\partial^2 \ln f(X_1;\vec{\theta})}{\partial\theta_i\partial\theta_j}\right),\ i, j = 1,..., r.$$

Для случая двумерного вектора $\vec{\theta} = (\theta_1, \theta_2)$ элементы $g_{ij}(\vec{\theta})$, $i, j = 1, 2$, информационной матрицы $I_n(\vec{\theta})$ равны:

$$g_{11} = -nM(\partial^2 \ln f(X_1, \vec{\theta})/\partial\theta_1^2) = -nM(-1/\theta_2^2) = n/\theta_2^2;$$

$$g_{12} = g_{21} = -nM(\partial^2 \ln f(X_1, \vec{\theta})/\partial\theta_1\partial\theta_2) = -nM\{(-2/\theta_2^3)(X_1 - \theta_1)\} = \\ = n(2/\theta_2^3)M(X_1 - \theta_1) = 0;$$

$$g_{22} = -nM(\partial^2 \ln f(X_1, \vec{\theta})/\partial\theta_2^2) = -nM\{(1/\theta_2^2) - (3/\theta_2^4)(X_1 - \theta_1)^2\} = \\ = (3n/\theta_2^4)M(X_1 - \theta_1)^2 - n/\theta_2^2 = 2n/\theta_2^2.$$

Таким образом, для общей нормальной модели $N(\theta_1, \theta_2^2)$ информационная матрица $I_n(\vec{\theta})$ имеет вид

$$I_n(\vec{\theta}) = \begin{pmatrix} n/\theta_2^2 & 0 \\ 0 & 2n/\theta_2^2 \end{pmatrix}.$$

Данная матрица невырожденная, так как ее определитель $|I_n(\vec{\theta})| = (2n^2/\theta_2^4) \neq 0$, и обратная для нее матрица $I_n^{-1}(\vec{\theta})$ равна

$$I_n^{-1}(\vec{\theta}) = \begin{pmatrix} 2n/\theta_2^2 & 0 \\ 0 & \theta_2^2/2n \end{pmatrix}.$$

Для данной обратной матрицы имеем $g^{22} = g_{22}^{-1} = \theta_2^2/2n = \sigma^2/2n$. С учетом этого выражения, правая часть (2.4.35) запишется в виде $g^{22}(\tau'(\sigma))^2 = 2\sigma^4/n$, что меньше чем $D(\tilde{S}^2) = 2\sigma^4/(n-1)$. Таким образом, подправленная выборочная дисперсия $\tilde{S}^2$ не является эффективной оценкой дисперсии σ^2, ее дисперсия не достигает нижней границы в (2.4.35).

Теперь воспользуемся аналогом критерия Бхаттачария (П.3.22) для векторного параметра $\vec{\theta} = (\theta_1, \theta_2)$. Используя выражение для логарифма функции правдоподобия, получаем

$$\frac{1}{L(\vec{x};\theta_1,\theta_2)}\frac{\partial L(\vec{x};\theta_1,\theta_2)}{\partial\theta_2} = \frac{\partial \ln L}{\partial\theta_2} = \frac{1}{\theta_2^3}\sum_{i=1}^{n}(x_i-\theta_1)^2 - (n/\theta_2^2) =$$

$$= \frac{1}{\theta_2^3}\sum_{i=1}^{n}(x_i-\overline{x})^2 + \frac{n}{\theta_2^3}(\overline{x}-\theta_1)^2 - n/\theta_2,$$

$$\frac{1}{L}\frac{\partial^2 L}{\partial\theta_1^2} = \frac{\partial^2 \ln L}{\partial\theta_1^2} + \left(\frac{\partial \ln L}{\partial\theta_1}\right)^2 = \frac{n^2}{\theta_2^4}(\overline{x}-\theta_1)^2 - n/\theta_2^2.$$

Отсюда

$$\frac{1}{L}\left(\frac{\theta_2^3}{n-1}\frac{\partial L}{\partial\theta_2} - \frac{\theta_2^4}{n(n-1)}\frac{\partial^2 L}{\partial\theta_1^2}\right) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 - \theta_2^2 = \tilde{S}^2 - \sigma^2.$$

Сравнивая последнее выражение с критерием оптимальности (П.3.22), заключаем, что подправленная выборочная дисперсия $\tilde{S}^2$ является оптимальной оценкой дисперсии σ^2, то есть в рамках нормальной модели $N(\theta_1, \theta_2^2)$ существует в классе несмещенных оценок дисперсии σ^2 (или параметрической функции $\tau(\theta) = \theta_2^2$) оценка, для которой выполняется неравенство $D_\theta[T(\vec{X})] \geq 2\sigma^4/(n-1)$, причем равенство в этом неравенстве достигается при использовании подправленной выборочной дисперсии $\tilde{S}^2$.

В) Согласно теореме (2.8.31), случайная величина nS^2/σ^2 имеет распределение хи-квадрат с числом степеней свободы $(n-1)$, то есть справедливо выражение $L(nS^2/\sigma^2) = \chi^2(n-1)$, и, следовательно, дисперсия случайной величины nS^2/σ^2 равна удвоенному числу степеней свободы, то есть $D[nS^2/\sigma^2] = 2(n-1)$. Отсюда следует, что $D[S^2] = 2\sigma^4(n-1)/n^2$. Далее, $M(S^2) = \sigma^2 - (\sigma^2/n)$, и, следовательно, смещение $b_{S^2}(\sigma^2)$ выборочной дис-

персии S^2 равно $b_{S^2}(\sigma^2) = -\sigma^2/n$. Таким образом, среднеквадратическая ошибка выборочной дисперсии равна

$$CKO(S^2) = D(S^2) + b^2_{S^2}(\sigma^2) = 2\sigma^4(n-1)/n^2 + \sigma^4/n^2 = \sigma^4 \cdot (2n-1)/n^2.$$

Далее, учитывая, что $\tilde{S}^2 = S^2 \cdot n/(n-1)$, получим

$$D(\tilde{S}^2) = [n/(n-1)]^2 \cdot D(S^2) = [n/(n-1)]^2 \cdot 2\sigma^4(n-1)/n^2 = \sigma^4 \cdot 2/(n-1).$$

Вспоминая, что $\tilde{S}^2$ – несмещенная оценка, то есть $M(\tilde{S}^2) = \sigma^2$ и, следовательно, ее смещение равно нулю, получим

$$CKO(\tilde{S}^2) = D(\tilde{S}^2) = \sigma^4 \cdot 2/(n-1).$$

Таким образом, так как выполняются неравенства $(2n-1)/n^2 = 2/n - 1/n^2 < 2/n < 2/(n-1)$, $\forall n \in N$, то среднеквадратическая ошибка выборочной дисперсии S^2, которая является *смещенной* оценкой параметра σ^2, меньше среднеквадратической ошибки подправленной выборочной дисперсии $\tilde{S}^2$, которая является *несмещенной* оценкой параметра σ^2, то есть $CKO(S^2) < CKO(\tilde{S}^2)$. Этот пример показывает, что хотя несмещенность является важным и желательным свойством оценок, но его не следует преувеличивать. В ряде случаев, как в данном примере, незначительное смещение может привести к уменьшению среднеквадратической ошибки.

Приведем еще один пример. Так как подправленная выборочная дисперсия $\tilde{S}^2$ является несмещенной оценкой дисперсии σ^2, то любая оценка вида $\lambda(n)\tilde{S}^2$, где $\lambda(n)$ – некоторая функция от объема выборки n, будет смещенной оценкой дисперсии. Однако при этом в классе оценок вида $\lambda(n)\tilde{S}^2$ можно указать оценку с минимальной среднеквадратической ошибкой, равной $\sigma^4 \cdot 2/(n+1)$, которая достигается при $\lambda(n) = (n-1)/(n+1)$, и это означает, что в классе оценок вида $\lambda(n)\tilde{S}^2$ минимальной среднеквадратической ошибкой обладает оценка вида

$$\hat{S}^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Убедимся в этом. В самом деле, для оценок вида $\lambda(n)\tilde{S}^2$ среднеквадратическая ошибка

$$CKO(\lambda(n)\tilde{S}^2) = M[(\lambda\tilde{S}^2 - \sigma^2)^2] =$$
$$= M[(\lambda\tilde{S}^2 - \lambda\sigma^2 + \lambda\sigma^2 - \sigma^2)^2] = M[(\lambda(\tilde{S}^2 - \sigma^2) + (\lambda-1)\sigma^2)^2] =$$
$$= M[\lambda^2(\tilde{S}^2 - \sigma^2)^2 + 2\lambda(\lambda-1)(\tilde{S}^2 - \sigma^2)\sigma^2 + (\lambda-1)^2\sigma^4] =$$

$$= \lambda^2 D(\tilde{S}^2) + (\lambda - 1)^2 \sigma^4 = \lambda^2 \sigma^4 \cdot 2/(n-1) + (\lambda - 1)^2 \sigma^4 =$$
$$= [2\lambda^2/(n-1) + (\lambda - 1)^2]\sigma^4 = \varphi(\lambda)\sigma^4,$$

где $\varphi(\lambda) = 2\lambda^2/(n-1) + (\lambda - 1)^2$. В классе оценок $\lambda(n)\tilde{S}^2$ найдем минимальную среднеквадратическую ошибку, используя следующее выражение

$$\min_{\lambda} CKO(\lambda(n)\tilde{S}^2) = \min_{\lambda} \varphi(\lambda)\sigma^4.$$

Для этого вычислим производную от функции $\varphi(\lambda)$ и приравняем ее нулю. В результате получим $\partial\varphi(\lambda)/\partial\lambda = 4\lambda/(n-1) + 2(\lambda - 1) = 0$. Отсюда имеем $2\lambda + \lambda n - \lambda - n + 1 = 0$, или $\lambda(n+1) = n - 1$. Таким образом, значение λ, при котором функция $\varphi(\lambda)$ достигает минимума, равно $\lambda^* = (n-1)/(n+1)$. Далее, минимум функции $\varphi(\lambda)$ при $\lambda^* = (n-1)/(n+1)$ равен

$$\varphi(\lambda^*) = \frac{2}{n-1}\frac{(n-1)^2}{(n+1)^2} + \left(\frac{n-1}{n+1} - 1\right)^2 = \frac{2}{n+1}.$$

Таким образом, в классе оценок $\lambda(n)\tilde{S}^2$ минимальной среднеквадратической ошибкой обладает оценка $\hat{S}^2$ и эта ошибка вычисляется по следующей формуле:

$$\min_{\lambda} CKO(\lambda(n)\tilde{S}^2) = \varphi(\lambda^*)\sigma^4 = 2\sigma^4/(n+1).$$

Вычислим теперь смещение $b_{\hat{S}^2}(\sigma^2)$ оценки $\hat{S}^2 = \tilde{S}^2(n-1)/(n+1)$ по формуле

$$b_{\hat{S}^2}(\sigma^2) = M(\hat{S}^2) - \sigma^2 = M(\tilde{S}^2)(n-1)/(n+1) - \sigma^2 =$$
$$= \sigma^2[(n-1)/(n+1) - 1] = -2\sigma^2/(n+1).$$

Отсюда следует, что $\hat{S}^2$ асимптотически несмещенная оценка дисперсии σ^2.

Таким образом, среднеквадратическая ошибка *смещенной* оценки $\hat{S}^2$ параметра σ^2 меньше среднеквадратической ошибки *несмещенной* оценки $\tilde{S}^2$ этого же параметра σ^2, то есть

$$\{CKO(\hat{S}^2) = 2\sigma^4/(n+1)\} < \{CKO(\tilde{S}^2) = 2\sigma^4/(n-1)\}.$$

2.15.21. Пусть $L(X) = \Pi y(\theta)$. Эта запись означает, что дискретная с.в. X имеет распределение Пуассона с параметром $\theta > 0$ и пробегает значения $x = 0,1,2,...$ с вероятностями

$$f_X(x,\theta) = P(X = x) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0,1,2,...,\ \theta > 0.$$

Предполагаем, что параметр $\theta = M(X)$ неизвестен. Пусть, далее, над с.в. X проведено одно наблюдение, то есть объем выборки $n = 1$. Убедитесь, что для функции $\varphi(\theta) = 1/\theta$ не существует несмещенной оценки.

Решение. Проведем рассуждения от противного. Предполагаем, что существует несмещенная оценка $T(X)$ функции $\varphi(\theta) = 1/\theta$. Тогда, согласно определению (2.2.3), должно выполняется выражение

$$\sum_{x=0}^{\infty} T(x)\, e^{-\theta}\theta^x / x! = 1/\theta \ , \ \forall\theta \in (0,\infty) .$$

Это выражение может быть переписано в виде

$$\sum_{x=0}^{\infty} T(x)\, \theta^{x+1} / x! = e^{\theta} = \sum_{x=0}^{\infty} \theta^x / x! .$$

Таким образом, чтобы левая и правая части этого выражения были бы равными, нужно, чтобы $T(x) = 1/\theta$, то есть нужно, чтобы оценка $T(X)$ была бы функцией от неизвестного параметра θ, но согласно определению (2.2.2), функция от наблюдений является оценкой параметра θ тогда, когда она не зависит от оцениваемого параметра θ. Итак, в данном случае не существует несмещенной оценки функции $\varphi(\theta) = 1/\theta$.

2.15.22. На практике применяют различные виды средних характеристик и их выборочные оценки. Среднее арифметическое, среднее геометрическое и среднее гармоническое соответствуют специальным случаям среднестепенных (см. пример 2.8.8). Классическим примером применения гармонического среднего является определение средней скорости движущегося объекта. Пусть расстояние S км между пунктами *А* и *Б* автомобиль преодолел со средней скоростью V_1 км/ч. Обратный путь автомобиль преодолел со средней скоростью V_2 км/ч. Определить среднюю скорость автомобиля на всем пути в $2S$ км.

Решение. Согласно формуле (2.8.18), средняя гармоническая скорость автомобиля на всем пути равна

$$V_{\text{ср}} = \frac{2}{V_1^{-1} + V_2^{-1}} .$$

В справедливости этой формулы можно убедиться и непосредственным образом. В самом деле, время t_1 преодоления расстояния S км между пунктами *А* и *Б* равно $t_1 = S/V_1$. На обратном пути будет затрачено время $t_2 = S/V_2$. Следовательно, общее время на преодоление расстояния $2S$ км будет равно $t_1 + t_2 = S/V_1 + S/V_2$. Отсюда средняя скорость автомобиля на всем пути в $2S$ км равна

$$V_{\text{ср}} = \frac{2S}{t_1 + t_2} = \frac{2S}{SV_1^{-1} + SV_2^{-1}} = \frac{2}{V_1^{-1} + V_2^{-1}}.$$

Напрмер, при $V_1 = 30$ км/ч, $V_2 = 60$ км/ч получаем $V_{\text{ср}} = 2/(1/30+1/60) = 40$ км/ч. Отметим, что средняя арифметическая скорость равна $(30+60)/2 = 45$ км/ч, что противоречит непосредственным рассуждениям, согласно которым $V_{\text{ср}} = 40$ км/ч.

2.15.23. Показать, что из формулы (2.8.9) при $\nu = 0$ следует (2.8.12).

Решение. Раскрывая неопределенность типа (1^{∞}) и используя правило Лопиталя, получим

$$T_0(F) = \lim_{\nu\to 0} T_\nu(F) = \lim_{\nu\to 0} \left\{ \int_{-\infty}^{+\infty} x^\nu dF_X(x) \right\}^{1/\nu} =$$

$$= \exp\left[\ln\left(\lim_{\nu\to 0} \left\{ \int_{-\infty}^{+\infty} x^\nu dF_X(x) \right\}^{1/\nu} \right)\right] = \exp\left(\lim_{\nu\to 0} \left\{ \frac{\ln \int_{-\infty}^{+\infty} x^\nu dF_X(x)}{\nu} \right\} \right) =$$

$$= \exp\left\{ \lim_{\nu\to 0} \frac{1/\left(\int x^\nu dF(x)\right) \cdot \int x^\nu \ln x dF(x)}{1} \right\} = \exp\left\{ \int_{-\infty}^{+\infty} \ln x \cdot dF_X(x) \right\}.$$

2.15.24. К структурным характеристикам основной тенденции признака X (или его «среднего» значения) относят медиану и моду. Приведем формулы и примеры вычисления выборочных оценок этих характеристик не по исходным данным $X_1,...,X_n$, а по данным, которые предварительно обработаны с помощью метода группировки и для конкретной реализация $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ представлены интервальным рядом частот $n_i,...,n_k$, то есть в виде

$$\vec{x} = (x_1,...,x_n) \Rightarrow \begin{pmatrix} \Delta_1,...,\Delta_k \\ n_1,...,n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n,$$

где $\Delta_1,...,\Delta_k$ – непересекающиеся интервалы группировки, на которые предварительно разбита область возможных значений с.в. X, и n_i – частота попадания наблюденных данных $\vec{x} = (x_1,...,x_n)$ в i-й интервал группировки $\Delta_i = \tilde{x}_i - \tilde{x}_{i-1}$, $i = 1,...,k$.

А) Выборочная оценка медианы по данным интервального ряда частот вычисляется (см., например, [78]) по формуле

$$\hat{x}_{\text{med}} = \tilde{x}_{i-1} + \left(\frac{n}{2} - \sum_{j=1}^{i-1} n_j \right) \frac{\Delta_i^*}{n_i^*},$$

где Δ_i^* – «медианный» интервал, для которого накопленные частоты в предыдущем $(i-1)$-м интервале меньше $n/2$, а накопленные частоты в i-м интервале больше $n/2$, то есть для медианного интервала Δ_i^* выполняются неравенства: $\sum_{j=1}^{i-1} n_j < n/2$ и $\sum_{j=1}^{i} n_j > n/2$; n_i^* – частота в медианном интервале и $\tilde{x}_{i-1}$ – его левая граница. Приведем пример.

Интервалы групп. $\{\Delta_i\}$	Частоты $\{n_i\}$	Накопленные частоты
10 – 15	6	6
15 – 20	9	15
20 – 25	20	***35***
25 – 30	**41**	**76**
30 – 35	26	102
35 – 40	21	123
40 – 45	14	137
45 – 50	5	142
50 – 55	1	143
	$n = 143$	

1. Вычислим $n/2 = 143/2 = 71{,}5$.

2. Определим медианный интервал $\Delta_i^* = [25\,,\,30]$.

3. Вычислим выборочную медиану

$$\hat{x}_{\text{med}} = \tilde{x}_{i-1} + \left(\frac{n}{2} - \sum_{j=1}^{i-1} n_j \right) \frac{\Delta_i^*}{n_i^*} = 25 + (71{,}5 - 35)5/41 = 29{,}5\,.$$

Б) Выборочная оценка моды по данным интервального ряда частот вычисляется (см., например, [78]) по формуле

$$\hat{x}_{\text{mod}} = \tilde{x}_{i-1} + \frac{(n_i - n_{i-1})\Delta_i^{**}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}\,,$$

где Δ_i^{**} – «модальный» интервал, соответствующий максимальной частоте n_i среди частот $n_1, ..., n_k$; $\tilde{x}_{i-1}$ – левая граница модального интервала Δ_i^{**} и n_i – частота интервала группировки $\Delta_i = \tilde{x}_i - \tilde{x}_{i-1}$, $i = 1, ..., k$. Для данных числового примера имеем

$$\hat{x}_{\text{mod}} = \tilde{x}_{i-1} + \frac{(n_i - n_{i-1})\Delta_i^{**}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} = 25 + \frac{(41-20)\cdot 5}{(41-20) + (41-26)} = 27{,}9\,.$$

2.15.25. Пусть имеется «выборка» $X_1,...,X_n$, получаемая в результате n независимых наблюдений над изучаемой случайной величиной X, причем каждое наблюдение получено с помощью измерительных приборов, обладающих различными характеристиками точности. Отметим, что обычно под выборкой понимают последовательность *независимых и одинаково* распределенных случайных величин, в данном же случае предполагается, что случайные величины $X_1,...,X_n$ независимые, имеют одинаковые математические ожидания, то есть $M(X_i)=\mu$, $i=1,...,n$, однако их дисперсии разные, то есть $D(X_i)=\sigma_i^2$, $i=1,...,n$. Такие наблюдения часто называют «*разноточными*». По этой причине термин выборка в начале данного примера взят в кавычки. Рассмотрим случайную величину Y в виде линейной комбинации случайных величин $X_1,...,X_n$, то есть

$$Y=\sum_{i=1}^{n}a_iX_i, \tag{1}$$

где a_i, $i=1,...,n$, – весовые коэффициенты. Требуется выбрать весовые коэффициенты $a_1,...,a_n$ таким образом, чтобы линейная комбинация (1) была бы несмещенной оценкой μ с минимальной дисперсией.

Решение. В данном упражнении используем самостоятельную нумерацию формул. Вычислим сначала математическое ожидание случайной величины Y и убедимся, что для обеспечения несмещенности оценки Y параметра μ требуется наложить на весовые коэффициенты $a_1,...,a_n$ условие вида

$$\sum_{i=1}^{n}a_i=1. \tag{2}$$

В самом деле,

$$M(Y)=M\left(\sum_{i=1}^{n}a_iX_i\right)=\sum_{i=1}^{n}a_iM(X_i)=\mu\sum_{i=1}^{n}a_i.$$

Таким образом, чтобы обеспечить несмещенность, то есть чтобы выполнялось равенство $M(Y)=\mu$, требуется выполнение условия (2), из которого следует, что

$$a_n=1-\sum_{i=1}^{n-1}a_i, \tag{3}$$

Вычислим теперь дисперсию случайной величины Y, а затем минимизируем ее по весовым коэффициентам $a_1,...,a_n$. Учитывая, что $X_1,...,X_n$ – независи-

мые случайные величины, и используя (3), получим

$$D(Y) = D\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 D(X_i) = \sum_{i=1}^{n-1} a_i^2 \sigma_i^2 + \left(1 - \sum_{i=1}^{n-1} a_i\right)^2 \sigma_n^2 =$$

$$= \sum_{i=1}^{n-1} a_i^2 \sigma_i^2 + \left(1 - 2\sum_{i=1}^{n-1} a_i + \left(\sum_{i=1}^{n-1} a_i\right)^2\right)\sigma_n^2 .$$

Далее, вычисляя частные производные по a_j, $j = 1,..., n-1$, и приравнивая их к нулю, получим

$$\frac{\partial D(Y)}{\partial a_j} = 2a_j\sigma_j^2 - 2\sigma_n^2 + 2\sigma_n^2\sum_{i=1}^{n-1} a_i = 2(a_j\sigma_j^2 - (1 - \sum_{i=1}^{n-1} a_i)\sigma_n^2) = 0,\ j = 1,...,n-1.$$

Отсюда, с учетом (3), имеем $a_j\sigma_j^2 - a_n\sigma_n^2 = 0$, $j = 1,...,n-1$, или

$$a_j = a_n \frac{\sigma_n^2}{\sigma_j^2},\ j = 1,...,\ n-1. \qquad (4)$$

Отметим, что формула (4) справедлива и для $j = n$, так как при $j = n$ получаем тождество $1 \equiv 1$. Суммируя далее обе части (4) по j, $j = 1,...,n$, и учитывая (2), получим

$$1 = a_n\sigma_n^2 \sum_{j=1}^{n} (1/\sigma_j^2).$$

Отсюда имеем

$$a_n = 1/\sigma_n^2 \sum_{j=1}^{n} (1/\sigma_j^2).$$

Подставив данное выражение в (4), получим

$$a_j = \left(\sigma_n^2 \sum_{j=1}^{n} (1/\sigma_j^2)\right)^{-1} \cdot \frac{\sigma_n^2}{\sigma_j^2} = \left(\sigma_j^2 \sum_{j=1}^{n} (1/\sigma_j^2)\right)^{-1},\ j = 1,...,n. \qquad (5)$$

Таким образом, если весовые коэффициенты $a_1,...,a_n$ определяются выражением (5), то оценка (1) в виде линейной комбинации случайных величин $X_1,...,X_n$ является несмещенной для параметра μ и имеет минимальную дисперсию.

Вычислим теперь минимальное значение дисперсии оценок вида (1). Для этого, подставив весовые коэффициенты, определяемые формулами (5), в вы-

ражение для дисперсии случайной величины Y, получим

$$D(Y)=\sum_{i=1}^{n}a_i^2\sigma_i^2=\sum_{i=1}^{n}\left(\sigma_i^2\sum_{j=1}^{n}(1/\sigma_j^2)\right)^{-2}\cdot\sigma_i^2=\sum_{i=1}^{n}(1/\sigma_i^2)\cdot\left(\sum_{j=1}^{n}(1/\sigma_j^2)\right)^{-2}=$$

$$=\left(\sum_{j=1}^{n}(1/\sigma_j^2)\right)^{-1}=\left(\sum_{j=1}^{n}\omega_j\right)^{-1},$$

где $\omega_j=1/\sigma_j^2$, $j=1,...,n$. С использованием этих обозначений, коэффициенты $a_1,...,a_n$, определяемые формулами (5), запишутся как

$$a_j=\omega_j/\sum_{j=1}^{n}\omega_j,\ j=1,...,n, \tag{6}$$

а формулы для оценки Y и ее дисперсии $D(Y)$ примут следующий вид:

$$Y=\sum_{i=1}^{n}a_iX_i=\sum_{i=1}^{n}\omega_iX_i/\sum_{j=1}^{n}\omega_j, \tag{7}$$

$$D(Y)=\sum_{i=1}^{n}a_i^2\sigma_i^2=\sum_{i=1}^{n}\left(\omega_i/\sum_{j=1}^{n}\omega_j\right)^2\frac{1}{\omega_i}=\sum_{i=1}^{n}\omega_i/\left(\sum_{j=1}^{n}\omega_j\right)^2=\left(\sum_{j=1}^{n}\omega_j\right)^{-1}. \tag{8}$$

Таким образом, несмещенная оценка параметра μ, обладающая минимальной дисперсией, которая построена по разноточным наблюдениям $X_1,...,X_n$, записывается в виде средневзвешенного (7) с весами ω_j, обратно пропорциональными дисперсиям σ_j^2, то есть $\omega_j=1/\sigma_j^2$, $j=1,...,n$.

2.15.26. Среднее и дисперсия в случае выборки из конечной совокупности. Пусть $x_1,...,x_N$ – конечная совокупность попарно различных элементов x_i, и пусть из этой совокупности извлекается *без возвращения* выборка $X_1,...,X_n$ объема n. Для любых попарно различных индексов $i_1,i_2,...,i_n$ имеем

$$P\{X_1=x_{i_1},...,X=x_{i_n}\}=\frac{(N-n)!}{N!}.$$

Далее, отсюда следует, что

$$P\{X_j=x_i\}=\frac{1}{N},\ i=1,...,N,\ i=1,...,n$$

и

$$P\{X_j=x_{i_j},X_k=x_{i_k}\}=\frac{1}{N(N-1)},\ j\neq k,\ i_j\neq i_k.$$

Обозначим математическое ожидание через

$$a = M[X_1] = \sum_{i=1}^{N} x_i P\{X_1 = x_i\} = \frac{1}{N}\sum_{i=1}^{N} x_i ,$$

дисперсию через

$$\sigma^2 = D[X_1] = M[X_1^2] - \{M[X_1]\}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - a^2$$

и выборочное среднее через $\bar{X} = (X_1 + \cdots + X_n)/n$. Убедитесь, что выполняются следующие равенства:

$$M[\bar{X}] = a ,\ M[\bar{X}^2] = a^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right),\ D[\bar{X}] = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Решение. Используя общие свойства математического ожидания, с учетом того, что $P\{X_j = x_i\} = 1/N$, $j = 1,...,N$, $i = 1,...,n$, и $P\{X_j = x_{i_j}, X_k = x_{i_k}\} = 1/N(N-1)$, $j \neq k$, $i_j \neq i_k$, имеем

$$M[X_i] = M[X_1],\ M[X_i^2] = M[X_1^2],\ M[X_i X_j] = M[X_1 X_2],\ i \neq j ,$$

$$M[\bar{X}] = \frac{1}{n} n M[X_1] = a .$$

Далее, распишем квадрат выборочного среднего в виде

$$\bar{X}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j = \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right).$$

Отсюда получаем $M[\bar{X}^2] = \{nM[X_1^2] + n(n-1)M[X_1 X_2]\}/n^2$. Вычислим теперь

$$M[X_1 X_2] = \sum_{i \neq j} x_i x_j P\{X_1 = x_i, X_2 = x_j\} = \frac{1}{N(N-1)}\sum_{i \neq j} x_i x_j =$$

$$= \frac{1}{N(N-1)}\left[\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right] =$$

$$= \{N^2 a^2 - N(a^2 + \sigma^2)\}/N(N-1) = a^2 - \sigma^2/(N-1) .$$

Используя последнее выражение и учитывая, что $M[X_1^2] = \sigma^2 + a^2$, получаем

$$M[\bar{X}^2] = \{nM[X_1^2] + n(n-1)M[X_1 X_2]\}/n^2 =$$

$$= \frac{\sigma^2 + a^2}{n} + \frac{n-1}{n}a^2 - \frac{(n-1)}{n(N-1)}\sigma^2 = a^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Отсюда окончательно получаем выражение для дисперсии в виде

$$D[\bar{X}] = M[\bar{X}^2] - \{M[\bar{X}]\}^2 = a^2 + \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) - a^2 = \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right).$$

Отметим, что если $N \to \infty$, то $D[\bar{X}] = \sigma^2 / n$, что согласуется с формулой для дисперсии выборочного среднего в непрерывном случае.

2.15.27. Рассмотрим параметрическую модель $(\mathrm{X}, \Im_\theta)$, где X – выборочное пространство и $\Im_\theta$ – параметрическое семейство функций, заданных с точностью до конечного числа неизвестных параметров $\theta \in \Theta$, то есть в виде $\Im_\theta = \{F : F_X(x, \theta), \theta \in \Theta\}$. Убедитесь, что *MD-оценки* параметров $\theta \in \Theta$, полученные методом минимума измененного расстояния «хи-квадрат», то есть являющиеся решением системы (2.11.11), совпадают с мультиномиальными оценками максимального правдоподобия, которые строятся не по исходным данным $X_1, ..., X_n$, а по данным вариационного ряда частот (2.11.4).

Решение. Пусть для определенности θ – векторный параметр размерности r, то есть $\vec{\theta} = (\theta_1, ..., \theta_r)$. Для случая непрерывных распределений мультиномиальные оценки максимального правдоподобия строятся не по исходным данным $X_1, ..., X_n$, а по данным, которые предварительно обработаны с помощью метода группировки и для конкретной реализация $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X} = (X_1, ..., X_n)$ представлены дискретным вариационным рядом частот $n_i, ..., n_k$, то есть в виде

$$\vec{x} = (x_1, ..., x_n) \Rightarrow \begin{pmatrix} \Delta_1, ..., \Delta_k \\ n_1, ..., n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n,$$

где $\Delta_1, ..., \Delta_k$ – непересекающиеся интервалы группировки, на которые предварительно разбита область возможных значений с.в. X, и n_i – частота попадания наблюденных данных $\vec{x} = (x_1, ..., x_n)$ в i-й интервал группировки Δ_i, $i = 1, ..., k$. Отметим, что выбор конкретных интервалов группировки $\Delta_1, ..., \Delta_k$ однозначно определяет вероятности $p_i(\theta) = P\{X \in \Delta_i\}$, $i = 1, ..., k$, и для каждой конкретной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X} = (X_1, ..., X_n)$ однозначно определяет реализацию $\vec{n} = (n_i, ..., n_k)$ вектора частот $\vec{N} = (N_1, ..., N_k)$. Распределение случайного вектора $\vec{N} = (N_1, ..., N_k)$ характеризуется полиномиальным (или мультиномиальным) распределением (см. П.1.49) вида

$$P_\theta\{\vec{N} = \vec{n}\} = P_\theta\{N_1 = n_1, ..., N_k = n_k\} = \frac{n!}{n_1! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i}(\theta).$$

Таким образом, частотная функция правдоподобия для группированных данных запишется (см. замечание 2.3.15) в виде

$$L(n_1,...,n_k;\vec{\theta}) = \frac{n!}{n_1!...n_k!}\prod_{i=1}^{k} p_i^{n_i}(\vec{\theta}).$$

Логарифмируя данную функцию и приравнивая к нулю ее частные производные, получим систему уравнений вида

$$\frac{\partial \ln L(\vec{n};\vec{\theta})}{\partial \theta_l} = \sum_{i=1}^{k} \frac{n_i}{p_i(\vec{\theta})} \frac{\partial p_i(\vec{\theta})}{\partial \theta_l} = 0 \ , \ l = 1,...,r. \qquad \textbf{(2.15.28)}$$

Решением данной системы являются так называемые (см., например, [22]) мультиномиальные оценки максимального правдоподобия. Окончательно, сравнивая системы (2.15.28) и (2.11.11), заключаем, что *MD-оценки* параметров $\theta \in \Theta$, полученные методом минимума измененного расстояния «хи-квадрат», совпадают с мультиномиальными оценками максимального правдоподобия.

2.15.29. Показать, что наилучшая аппроксимация биномиального распределения нормальным распределением записывается в виде

$$F_X(x) = P\{X \le x\} = \sum_{k=0}^{x} C_n^k p^k (1-p)^{n-k} = \Phi\left(\frac{x-np+c}{\sqrt{np(1-p)}}\right) + O\left(\frac{1}{\sqrt{n}}\right)$$

и обеспечивается при $c = 0,5$.

Решение. Отметим, что если с.в. X имеет биномиальное распределение, то есть $L(X) = Bi(n,p)$, то с.в. $n - X$ также имеет биномиальное распределение, но с параметрами n и $1-p$, то есть $L(n-X) = Bi(n,1-p)$, причем $P\{X \le x\} + P\{n - X \le n - x - 1\} = 1$.

Таким образом, используя приведенную выше аппроксимацию, получаем при $n \to \infty$ приближенное равенство

$$\Phi\left(\frac{x-np+c}{\sqrt{np(1-p)}}\right) + \Phi\left(-\frac{x-np+(1-c)}{\sqrt{np(1-p)}}\right) \approx 1.$$

В последнем выражении выберем постоянную величину c таким образом, чтобы оно из приближенного равенства превратилось в точное равенство. Учитывая далее, что стандартная нормальная функция распределения $\Phi(x)$ симметрична относительно нуля, то есть $\Phi(x) + \Phi(-x) = 1 \ \forall x \in R^1$, получаем

$$\Phi\left(\frac{x-np+c}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{x-np+(1-c)}{\sqrt{np(1-p)}}\right) = 0.$$

Отсюда следует, что $c = 1 - c$, или $c = 0{,}5$. Отметим, что величину $c = 0{,}5$ называют поправкой на непрерывность, которую применяют для повышения качества аппроксимации дискретных распределений статистик непрерывными распределениями (см. упр. 5.8.50).

2.15.30. Пусть $L(X) = N(0\,;1)$. Имеется выборка $\vec{X}_1 = (X_1,...,X_n)$ из распределения $F_X(x,\vec{\theta}) = \Phi((x-a)/\sigma)$, где $\vec{\theta} = (a,\sigma^2)$ и $\Phi(\cdot)$ – стандартная нормальная функция распределения. Предполагая, что σ^2 – известный параметр, убедитесь, что доверительный интервал (2.12.7) для параметра a имеет минимальную длину среди всех γ-доверительных интервалов вида ($\overline{X} - s_2\sigma/\sqrt{n}$, $\overline{X} - s_1\sigma/\sqrt{n}$), где $\gamma = (1-\alpha)$, $0 < \gamma < 1$, $s_1 < s_2$ – любые числа, удовлетворяющие, согласно (2.12.3), условию $\Phi(s_2) - \Phi(s_1) = 1 - \alpha$.

Решение. В данном случае центральная статистика (см. пример 2.12.4) имеет вид $S(\vec{X},a) = \sqrt{n}(\overline{X} - a)/\sigma$. Для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ функция $S(\vec{x},a)$ монотонно убывает по a, поэтому решения уравнений $S(\vec{x},a) = s_2$ и $S(\vec{x},a) = s_1$ относительно параметра a имеют вид $\hat{a}_1(\vec{x}) = \overline{x} - s_2\sigma/\sqrt{n}$ и $\hat{a}_2(\vec{x}) = \overline{x} - s_1\sigma/\sqrt{n}$. Поскольку функция $S(\vec{x},a)$ строго монотонная по a, то неравенства $s_1 < S(\vec{x},a) < s_2$ и $\hat{a}_1(\vec{x}) = \overline{x} - s_2\sigma/\sqrt{n} < a < \overline{x} - s_1\sigma/\sqrt{n} = \hat{a}_2(\vec{x})$ эквивалентны, и поэтому всякий доверительный интервал вида

$$(\overline{X} - s_2\sigma/\sqrt{n},\ \overline{X} - s_1\sigma/\sqrt{n}),$$

где $s_1 < s_2$ – любые числа, удовлетворяющие условию $\Phi(s_2) - \Phi(s_1) = 1 - \alpha$, является γ-доверительным интервалом, причем $\gamma = (1-\alpha)$. Отметим, что границы этого интервала являются случайными, однако его длина не случайна и равна $l(s_1,s_2) = (s_2 - s_1)\sigma/\sqrt{n}$. Естественно, что при заданной доверительной вероятности $\gamma = (1-\alpha)$ мы заинтересованы в том, чтобы доверительный интервал имел наименьшую длину. Для нахождения такого интервала нужно конкретизировать числа s_1 и s_2, которые обеспечивают минимум функции $l(s_1,s_2)$ при выполнении условия $\Phi(s_2) - \Phi(s_1) = 1 - \alpha$. Применим метод Лагранжа для нахождения условного экстремума функции двух переменных $l(s_1,s_2)$. Функция Лагранжа имеет вид

$$\psi(s_1,s_2,\lambda) = (s_2 - s_1)\sigma/\sqrt{n} + \lambda[\Phi(s_2) - \Phi(s_1) - (1-\alpha)],$$

где λ – неопределенный множитель. Вычисляя частные производные, получим следующую систему уравнений:

$$\frac{\partial\psi(s_1,s_2,\lambda)}{\partial s_1} = -\sigma/\sqrt{n} - \lambda\varphi(s_1) = 0,$$

$$\frac{\partial\psi(s_1,s_2,\lambda)}{\partial s_2} = \sigma/\sqrt{n} + \lambda\varphi(s_2) = 0,$$

$$\frac{\partial\psi(s_1,s_2,\lambda)}{\partial \lambda} = \Phi(s_2) - \Phi(s_1) - (1-\alpha) = 0,$$

где $\varphi(x) = \Phi'(x)$ – стандартная нормальная плотность. Из первых двух уравнений следует равенство $\varphi(s_1) = \varphi(s_2)$. Учитывая далее, что $s_1 < s_2$ и принимая во внимание, что плотность $\varphi(x)$ симметрична относительно нуля (то есть $\varphi(x)$ – четная функция), заключаем, что выполняется равенство $s_1 = -s_2$. Используя это равенство, а также соотношение $\Phi(x) = 1 - \Phi(-x)$, $\forall x \in R^1$, из последнего уравнения системы получаем, что $2\Phi(s_2) - 1 = 1 - \alpha$. Отсюда следует, что $s_2 = \Phi^{-1}(1-\alpha/2)$, то есть величина s_2 является квантилем уровня $(1-\alpha/2)$ стандартного нормального распределения, который был обозначен в примере 2.12.4 через $\lambda_{1-\alpha/2}$. Итак, $s_2 = \lambda_{1-\alpha/2}$, поэтому доверительный интервал ($\overline{X} - s_2\sigma/\sqrt{n}$, $\overline{X} - s_1\sigma/\sqrt{n}$), имеющий минимальную длину среди всех γ-доверительных интервалов, записывается в виде (2.12.7).

2.15.31. Пусть $L(X) = N(a,\sigma^2)$. Имеется выборка $\vec{X}_1 = (X_1,...,X_n)$ из распределения $F_X(x;a,\sigma^2) = \Phi((x-a)/\sigma)$, где $\Phi(\cdot)$ – стандартная нормальная функция распределения. Предполагаем, что параметр a известен, требуется по выборке $\vec{X}_1 = (X_1,...,X_n)$ построить γ-доверительный интервал для неизвестного параметра σ^2, γ – заданная доверительная вероятность, причем $0 < \gamma < 1$ и $\gamma = 1 - \alpha$.

Решение. В данном случае неизвестным параметром θ является дисперсия, то есть $\theta = \sigma^2$ и центральная статистика $S(\vec{X},\theta)$ записывается в виде

$$S(\vec{X},\theta) = \theta^{-1}\sum_{i=1}^{n}(X_i - a)^2 = SS/\theta,$$

где для сокращения записи через SS обозначена сумма квадратов отклонений случайных величин $X_1,...,X_n$ от их математического ожидания a, то есть

$$SS = \sum_{i=1}^{n}(X_i - a)^2. \qquad \textbf{(2.15..32)}$$

Отметим, что для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ функция $S(\vec{x},\sigma^2)$ является непрерывной и убывающей функ-

цией по σ^2 и центральная статистка $S(\vec{X},\theta) = SS/\theta$ имеет распределение хи-квадрат с числом степеней свободы n, то есть $L\{S(\vec{X},\sigma^2)\} = \chi^2(n)$.

Напомним, (см. П.2.73), что плотность распределения вероятностей случайной величины, имеющей распределение хи-квадрат с числом степеней свободы n, записывается в виде

$$f_{\chi^2(n)}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \; x > 0. \qquad \textbf{(2.15.33)}$$

Построим центральный γ-доверительный интервал для параметра σ^2 при известном параметре a. Для этого воспользуемся выражением (2.12.3) и выберем числа s_1 и s_2 $(s_1 < s_2)$ из условий

$$P\{\chi^2_{(n)} \le s_1\} = P\{\chi^2_{(n)} \ge s_2\} = \alpha/2.$$

Далее, обозначив квантиль уровня p, $0 < p < 1$, распределения хи-квадрат с числом степеней свободы n через $\chi^2_{p,n}$, определим числа s_1 и s_2 следующим образом: $s_1 = \chi^2_{\alpha/2,n}$, $s_2 = \chi^2_{1-\alpha/2,n}$. Подставив найденные значения s_1 и s_2 в (2.12.3), с учетом того, что $L\{S(\vec{X},\sigma^2)\} = \chi^2(n)$, получим для статистики $S(\vec{X},\sigma^2) = SS/\sigma^2$ следующее выражение:

$$P\{\chi^2_{\alpha/2,n} < SS/\sigma^2 < \chi^2_{1-\alpha/2,n}\} = 1-\alpha.$$

Разрешая неравенства в фигурных скобках относительно параметра σ^2 и вспоминая обозначение (2.15.32), окончательно получаем центральный γ-доверительный интервал, $\gamma = (1-\alpha)$, для параметра σ^2 в виде

$$((\chi^2_{1-\alpha/2,n})^{-1}\sum_{i=1}^{n}(X_i - a)^2, \; (\chi^2_{\alpha/2,n})^{-1}\sum_{i=1}^{n}(X_i - a)^2),$$

то есть для параметра σ^2 справедливо выражение

$$P\left\{\frac{SS}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 < \frac{SS}{\chi^2_{\alpha/2,n-1}}\right\} = 1-\alpha. \qquad \textbf{(2.15.34)}$$

Отметим, что в данном примере распределение центральной статистики $S(\vec{X},\sigma^2) = SS/\sigma^2$ не является симметричным. Поэтому построенный центральный доверительный интервал для параметра σ^2 вида (2.15.34), не обязательно имеет минимальную длину. Убедимся в этом. В самом деле, согласно (2.12.3), справедливо выражение

$$P\{s_1 < SS/\sigma^2 < s_2\} = 1-\alpha,$$

где числа s_1 и s_2 удовлетворяют условию

$$\int_{S_1}^{S_2} f_{\chi^2(n)}(x)dx = F_{\chi^2(n)}(s_2) - F_{\chi^2(n)}(s_1) = 1-\alpha. \qquad \textbf{(2.15.35)}$$

Отсюда следует, что γ-доверительным интервалом, $\gamma = 1-\alpha$, для параметра σ^2 в данном случае является любой интервал вида $(SS/s_2, SS/s_1)$, где числа s_1 и s_2 удовлетворяют (2.15.35). Отметим, что для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X}_1 = (X_1,...,X_n)$ длина интервала $(SS/s_2, SS/s_1)$ является функцией двух переменных s_1, s_2 и равна $l(s_1,s_2) = SS\cdot(s_2-s_1)/s_1s_2$. Таким образом, задача построения доверительного интервала с минимальной длиной сводится к нахождению чисел s_1 и s_2, при которых функция $l(s_1,s_2)$ достигает минимума при выполнении условия (2.15.34). Для нахождения условного экстремума функции $l(s_1,s_2)$ воспользуемся методом неопределенных множителей Лагранжа. Функция Лагранжа в данном случае записывается в виде

$$\psi(s_1,s_2,\lambda) = l(s_1,s_2) + \lambda[F_{\chi^2(n)}(s_2) - F_{\chi^2(n)}(s_1) - \gamma].$$

Вычисляя частные производные функции $\psi(s_1,s_2,\lambda)$ и приравнивая их нулю, получим систему уравнений вида

$$1+\lambda s_1^2 f_{\chi^2(n)}(s_1) = 0,$$
$$1+\lambda s_2^2 f_{\chi^2(n)}(s_2) = 0,$$
$$F_{\chi^2(n)}(s_2) - F_{\chi^2(n)}(s_1) = \gamma.$$

Из первых двух уравнений системы следует, что $s_1^2 f_{\chi^2(n)}(s_1) = s_2^2 f_{\chi^2(n)}(s_2)$. Далее положим $s_1 = \chi^2_{\alpha_1,n}$ и $s_2 = \chi^2_{1-\alpha_2,n}$ где $\alpha_1+\alpha_2 = \alpha$. С учетом этих обозначений и (2.15.33) получим уравнение вида

$$(\chi^2_{1-\alpha_2,n} / \chi^2_{\alpha_1,n})^{n/2+1} = \exp\{(\chi^2_{1-\alpha_2,n} - \chi^2_{\alpha_1,n})/2\}.$$

Из этого уравнения, учитывая, что $\alpha_1+\alpha_2 = \alpha$, можно определить величины α_1 и α_2, при которых длина интервала $(SS/\chi^2_{1-\alpha_2,n}, SS/\chi^2_{\alpha_1,n})$ будет минимальной. В работе [66, с. 294] отмечается, что доверительный интервал с наименьшей средней длиной получится, если выбрать разность

$(1/\chi^2_{\alpha_1,n})-(1/\chi^2_{1-\alpha_2,n})$ минимальной при условии, что числа $s_1=\chi^2_{\alpha_1,n}$ и $s_2=\chi^2_{1-\alpha_2,n}$, где $\alpha_1+\alpha_2=\alpha$, удовлетворяют (2.15.35). Таким образом, центральный интервал (2.15.34) в данном примере не является интервалом с наименьшей длиной.

2.15.36. Пусть $L(X)=Bi(n=1,p)$. Эта запись означает, что с.в. X имеет распределение Бернулли с параметром p, $0<p<1$. Имеется выборка $\vec{X}=(X_1,...,X_n)$. Требуется по этой выборке построить интервальную оценку неизвестного параметра $\theta=p$.

Решение. Хорошо известно (см. разд. 2.3), что состоятельная и эффективная точечная оценка параметра $\theta=p$ равна $\hat{\theta}=\overline{X}$. При конкретной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}=(X_1,...,X_n)$, с.в. $\overline{X}$ принимает свои возможные значения $\overline{x}=k/n$, $k=0,1,2,...,n$, при этом

$$F_{\hat{\theta}}(x,\theta)=\sum_{k=0}^{n}C(x-k)C_n^k\theta^k(1-\theta)^{n-k},\ x\in R^1,$$

и в целочисленных точках имеем

$$F_{\hat{\theta}}(k/n,\theta)=\sum_{i=0}^{k}C_n^i\theta^i(1-\theta)^{n-i},\quad k=0,1,2,...,n.$$

Далее можно убедиться, что функция $F_{\hat{\theta}}(k/n,\theta)$ монотонно убывает по θ при $k<n$. В самом деле,

$$\partial F_{\hat{\theta}}(k/n,\theta)/\partial\theta=\sum_{i=0}^{k}C_n^i\{\,i\theta^{i-1}(1-\theta)^{n-i}-(n-i)\theta^i(1-\theta)^{n-i-1}\,\}=$$

$$=\sum_{i=0}^{k}nC_n^i\{\,\theta^{i-1}(1-\theta)^{n-i}-\theta^i(1-\theta)^{n-i-1}\,\}=-nC_{n-1}^k\theta^k(1-\theta)^{n-k-1}<0,\ \forall k<n.$$

Таким образом, согласно методике, изложенной в разделе 2.14, центральный $(1-\alpha)$-доверительный интервал (2.14.3) записывается в виде $P\{\hat{\theta}_1(\vec{X})<\theta<\hat{\theta}_2(\vec{X})\}=1-\alpha$, где числа $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ определяются из условий (2.14.14), которые в данном случае для параметра $\theta=p$ определяются из выражений

$$F_{\hat{\theta}}(k/n,\hat{\theta}_2(\vec{x}))=\sum_{i=0}^{k}C_n^i\hat{\theta}_2{}^i(\vec{x})(1-\hat{\theta}_2(\vec{x}))^{n-i}=\alpha/2;\qquad\textbf{(2.15.37)}$$

$$F_{\hat{\theta}}(k/n,\hat{\theta}_1(\vec{x}))=\sum_{i=0}^{k}C_n^i\hat{\theta}_1{}^i(\vec{x})(1-\hat{\theta}_1(\vec{x}))^{n-i}=1-\alpha/2.\qquad\textbf{(2.15.38)}$$

Для вычисления $\hat{\theta}_1$ и $\hat{\theta}_2$ из этих выражений используют таблицы функции распределения биномиальной случайной величины (см., например, [6]). Отметим также, что эта функция распределения может быть выражена через неполную бета-функцию (1.4.57), которая также табулирована или может быть выражена (см. формулу (1.4.83)) в терминах табулированного F-распределения Снедекора – Фишера (см. упр. 2.15.39).

2.15.39. Интервальная оценка параметра p биномиального распределения, использующая квантили F-распределения. Пусть $L(X) = Bi(n, p)$. Это предположение означает, что с.в. X имеет биномиальное распределение с параметрами n и p, где n – число независимых опытов в схеме испытаний Бернулли и p – вероятность осуществления некоторого события, назовем его «успехом». Наблюденная реализация x этой биномиальной с.в. X может принимать одно из возможных значений $0,1,...,n$ с вероятностями $P_n(k) = C_n^k \theta^k (1-\theta)^{n-k}$, $k = 0,1,...,n,$ и предполагаем, что параметр $\theta = p$ неизвестен. Как было показано в примере (2.4.20), наилучшей точечной оценкой параметра θ является оценка максимального правдоподобия вида $\hat{\theta} = X/n$, численное значение которой в условиях конкретного опыта равно x/n, где x – наблюденное значение с.в. X (например, $x = k$, $k = 0,1,...,n$, – это число бракованных изделий среди n взятых на контроль из большой партии изделий). Требуется по этой выборке построить интервальную оценку неизвестного параметра θ.

Решение. В упражнении (1.4.54) было показано, что функция распределения биномиальной случайной величины X в целочисленных точках $k = 0,1,...,n$ вычисляется через неполную бета-функцию $I_p(n,m)$, определенную в (1.4.32), по формуле

$$P\{X \le k\} = \sum_{i=0}^{k} C_n^i \theta^i (1-\theta)^{n-i} = 1 - I_\theta(k+1, n-k),\ k = 0,1,...,n.$$

Далее, согласно формуле (1.4.83), функция распределения случайной величины $F(\nu_1, \nu_2)$, имеющей F-распределение с числами степеней свободы ν_1 и ν_2, выражается через неполную бета-функцию $I_p(n,m)$ в виде

$$P\{F(\nu_1, \nu_2) \le z\} = I_{\nu_1 z/(\nu_2 + \nu_1 z)}(\nu_1/2\ ,\ \nu_2/2).$$

Используя обозначения $\nu_1 = 2(k+1)$, $\nu_2 = 2(n-k)$, $\nu_1 z/(\nu_2 + \nu_1 z) = \theta$, выразим функцию $I_\theta(k+1, n-k)$ через F-распределение в виде

$$P\{F(2(k+1), 2(n-k)) \le \frac{n-k}{k+1} \cdot \frac{\theta}{1-\theta}\} = I_\theta(k+1, n-k).$$

С учетом этого получаем выражение

$$\sum_{i=0}^{k} C_n^i \theta^i (1-\theta)^{n-i} = 1 - P\{F(2(k+1), 2(n-k)) \le \frac{n-k}{k+1} \cdot \frac{\theta}{1-\theta}\},$$

которое позволяет использовать таблицы F-распределения при вычислении функции распределения биномиального закона. Из этого выражения также следует, что уравнение (2.15.37) в данном случае запишется в виде

$$1 - P\{F(2(k+1), 2(n-k)) \le \frac{n-k}{k+1} \cdot \frac{\hat{\theta}_2}{1-\hat{\theta}_2}\} = \alpha/2,$$

или

$$P\{F(2(k+1), 2(n-k)) \le \frac{n-k}{k+1} \cdot \frac{\hat{\theta}_2}{1-\hat{\theta}_2}\} = 1 - \frac{\alpha}{2}.$$

Далее, согласно определению квантиля заданного уровня $(1\text{-}\alpha/2)$, для F-распределения имеем

$$P\{F(2(k+1), 2(n-k)) \le F_{1-\alpha/2}(2(k+1), 2(n-k))\} = 1 - \alpha/2.$$

Отсюда получаем равенство

$$\frac{n-k}{k+1} \cdot \frac{\hat{\theta}_2}{1-\hat{\theta}_2} = F_{1-\alpha/2}(2(k+1), 2(n-k)),$$

из которого получаем формулу для вычисления $\hat{\theta}_2$ через квантиль $F_{1-\alpha/2}(n,m)$ уровня $(1-\alpha/2)$ для F-распределения с числами степеней свободы n и m в виде

$$\hat{\theta}_2 = \frac{(k+1) \cdot F_{1-\alpha/2}(2(k+1), 2(n-k))}{(n-k) + (k+1) \cdot F_{1-\alpha/2}(2(k+1), 2(n-k))}. \qquad \textbf{(2.15.40)}$$

Аналогично, применяя к (2.15.38) соотношение

$$\sum_{i=0}^{k} C_n^i \theta^i (1-\theta)^{n-i} = 1 - P\{F(2(k+1), 2(n-k)) \le \frac{n-k}{k+1} \cdot \frac{\theta}{1-\theta}\}$$

и учитывая, что $F_p(n,m) = 1/F_{1-p}(m,n)$, получаем формулу для вычисления $\hat{\theta}_1$ в виде

$$\hat{\theta}_1 = \frac{k \cdot F_{\alpha/2}(2k, 2(n-k+1))}{(n-k+1) + k \cdot F_{\alpha/2}(2k, 2(n-k+1))} =$$

$$= \frac{k}{(n-k+1) \cdot F_{1-\alpha/2}(2(n-k+1), 2k) + k}. \qquad \textbf{(2.15.41)}$$

Для иллюстрации полученных формул приведем числовой пример. Пусть из большой партии изделий взята выборка с возвращением объема $n=30$. В результате оказалось, что число бракованных изделий равно восьми, то есть $k=8$. Построить 95 %-й доверительный интервал для параметра θ, характеризующего долю бракованных изделий в большой партии изделий. Для $\alpha=0,05$ из таблицы квантилей F-распределения (см., например, [6, 9, 44]) находим

$$F_{1-\alpha/2}(2(k+1),2(n-k))=F_{0,975}(2(8+1),2(30-8))=F_{0,975}(18\,,44)=2,07\,,$$
$$F_{1-\alpha/2}(2(n-k+1)\,,2k)=F_{0,975}(2(30-8+1)\,,2\cdot 8)=F_{0,975}(46\,,16)=2,49\,.$$

Используя формулы (2.15.40) и (2.15.41), получаем

$$\hat{\theta}_1=\frac{8}{(30-8+1)\cdot 2,49+8}=0,123\ ,\hat{\theta}_2=\frac{(8+1)\cdot 2,07}{(30-8)+(8+1)\cdot 2,07}=0,459\,.$$

Таким образом, с вероятностью 0,95 истинное значение параметра θ, характеризующего долю бракованных изделий в большой партии изделий, накрывается интервалом (0,123 ; 0,459).

Отметим, что при больших объемах выборки n асимптотический доверительный интервал для параметра θ, равного вероятности «успеха» p в серии испытаний Бернулли, вычисляется по формуле (2.13.15), причем границы этого интервала $\hat{\theta}_1(\vec{x})$ и $\hat{\theta}_2(\vec{x})$ с учетом поправки на непрерывность (см. упр. 2.15.29) определяются согласно (2.14.14) из следующих выражений. Верхний предел $\hat{\theta}_2(\vec{x})$ доверительного интервала вычисляется с помощью уравнения

$$\Phi\left(\frac{x+0,5-n\hat{\theta}_2}{\sqrt{n\hat{\theta}_2(1-\hat{\theta}_2)}}\right)=\frac{\alpha}{2}\,,$$

которое сводится к квадратному уравнению относительно $\hat{\theta}_2$ вида

$$\frac{x+0,5-n\hat{\theta}_2}{\sqrt{n\hat{\theta}_2(1-\hat{\theta}_2)}}=\Phi^{-1}(\alpha/2)\,.$$

Решение этого уравнения при использовании обозначения $\lambda_{\alpha/2}=\Phi^{-1}(\alpha/2)$ записывается в виде

$$\hat{\theta}_2=\frac{1}{n+\lambda_{\alpha/2}^2}\left\{k+0,5+\frac{\lambda_{\alpha/2}^2}{2}-\lambda_{\alpha/2}\left[\frac{(k+0,5)(n-k-0,5)}{n}+\frac{\lambda_{\alpha/2}^2}{4}\right]^{1/2}\right\}. \quad \textbf{(2.15.42)}$$

Аналогично, нижний предел $\hat{\theta}_1(\vec{x})$ доверительного интервала находится с помощью уравнения

$$\Phi\left(\frac{x-0,5+n\hat{\theta}_1}{\sqrt{n\hat{\theta}_1(1-\hat{\theta}_1)}}\right)=1-\frac{\alpha}{2},$$

решение которого записывается в виде

$$\hat{\theta}_1=\frac{1}{n+\lambda^2_{1-\alpha/2}}\left\{k-0,5+\frac{\lambda^2_{1-\alpha/2}}{2}-\lambda_{1-\alpha/2}\left[\frac{(k-0,5)(n-k+0,5)}{n}+\frac{\lambda^2_{1-\alpha/2}}{4}\right]^{1/2}\right\}. \quad \textbf{(2.15.43)}$$

Для числового примера, рассмотренного выше, при $n=30$, $k=8$, для $\alpha=0,05$, $\lambda_{1-\alpha/2}=\lambda_{0,975}=1,96$, $\lambda_{\alpha/2}=\lambda_{0,025}=-1,96$ и из формулы (2.15.43) получаем

$$\hat{\theta}_1=\frac{1}{30+(1,96)^2}\left\{8-0,5+\frac{(1,96)^2}{2}-1,96\left[\frac{(8-0,5)(30-8+0,5)}{30}+\frac{(1,96)^2}{4}\right]^{1/2}\right\}=0,130.$$

Аналогично из (2.15.42) имеем

$$\hat{\theta}_2=\frac{1}{30+(1,96)^2}\left\{8+0,5+\frac{(1,96)^2}{2}+1,96\left[\frac{(8+0,5)(30-8-0,5)}{30}+\frac{(1,96)^2}{4}\right]^{1/2}\right\}=0,461.$$

Таким образом, для данного примера асимптотический 95 %-й доверительный интервал $(0,130;\ 0,461)$, вычисленный по асимптотическим формулам (2.15.42) и (2.15.43), лишь незначительно отличается от 95 %-го доверительного интервала (0,123; 0,459), вычисленного по точным формулам (2.15.40) и (2.15.41).

2.15.44. МНК-оценки. Пусть между переменными x и y существует функциональная зависимость вида $y=\varphi(x;\theta_1,...,\theta_k)$, причем параметры $\theta_1,...,\theta_k$ этой зависимости неизвестны. Требуется конкретизировать эту зависимость, то есть доопределить (или еще говорят оценить) неизвестные параметры $\theta_1,...,\theta_k$ на основе информации, содержащейся в экспериментальных данных $(x_1,y_1),...,(x_n,y_n)$, которые получены путем n измерений $y_1,...,y_n$ величины y при фиксированных значениях $x_1,...,x_n$ независимой переменной x. Учитывая, что всякий измерительный прибор имеет ошибки измерений, запишем математическую модель наблюдений в виде

$$Y_i=\varphi(x_i;\theta_1,...,\theta_k)+\varepsilon_i\ ,\ i=1,...,n, \quad \textbf{(2.15.45)}$$

ε_i характеризует случайные ошибки i-го измерения, а функция φ задана с точностью до конечного числа k неизвестных параметров $\theta_1,...,\theta_k$. Относительно случайных ошибок $\varepsilon_1,...,\varepsilon_n$ предполагается, что они удовлетворяют следующим ограничениям:

$$M(\varepsilon_i)=0,\ D(\varepsilon_i)=\sigma^2>0,\ Cov(\varepsilon_i,\varepsilon_j)=0,\ 1\le i<j\le n. \qquad \textbf{(2.15.46)}$$

Отметим, что неизвестная дисперсия σ^2 также является параметром модели наблюдений (2.15.45) и для уточнения совместного распределения ошибок $\varepsilon_1,...,\varepsilon_n$ требуется дополнительная информация (см. формулу (2.15.51)).

Итак, задача уточнения функциональной связи между переменными x и y сводится к задаче определения параметров $\theta_1,...,\theta_k$, то есть по экспериментальным данным $(x_1,y_1),...,(x_n,y_n)$ требуется оценить эти параметры.

Решение. Для решения этой задачи воспользуемся методом наименьших квадратов (МНК), согласно которому неизвестные параметры $\theta_1,...,\theta_k$ определяются из условия, чтобы сумма квадратов ошибок $\varepsilon_1,...,\varepsilon_n$ была бы минимальной, то есть из условия вида

$$\min_{\theta_1,...,\theta_k}\sum_{i=1}^{n}\varepsilon_i^2=\min_{\theta_1,...,\theta_k}\sum_{i=1}^{n}[y_i-\varphi(x_i;\theta_1,...,\theta_k)]^2. \qquad \textbf{(2.15.47)}$$

Определим для удобства записи многомерную функцию k переменных $\theta_1,...,\theta_k$, при заданном наборе экспериментальных данных $\{x_i,y_i\}$, в виде

$$S(\theta_1,...,\theta_k\,|\,\{x_i,y_i\})=\sum_{i=1}^{n}[y_i-\varphi(x_i;\theta_1,...,\theta_k)]^2. \qquad \textbf{(2.15.48)}$$

Обоснованием метода наименьших квадратов служит теорема Гаусса – Маркова (П.3.72), и его реализация сводится к задаче нахождения экстремума функции $S(\theta_1,...,\theta_k\,|\,\{x_i,y_i\})$. Согласно необходимым условиям существования экстремума многомерной функции, МНК-оценки параметров $\theta_1,...,\theta_k$ определяются из решения системы k нормальных уравнений вида

$$\frac{\partial S(\theta_1,...,\theta_k\,|\,\{x_i,y_i\})}{\partial\theta_j}=0,\ j=1,...,k. \qquad \textbf{(2.15.49)}$$

Отметим, что если функция $\varphi(x;\theta_1,...,\theta_k)$ линейная по параметрам $\theta_1,...,\theta_k$, то система нормальных уравнений (2.15.49) является системой линейных уравнений и может быть решена в явном виде.

Для иллюстрации метода наименьших квадратов рассмотрим пример линейной функции, то есть по экспериментальным данным $(x_1,y_1),...,(x_n,y_n)$ требуется подобрать прямую линию $y=\theta_1x+\theta_2$, которая наилучшим образом (в смысле минимума суммы квадратов ошибок измерений) согласуется с результатами эксперимента. Для решения этой задачи нужно определить параметры θ_1 и θ_2 методом наименьших квадратов. В данном случае функция $S(\theta_1,...,\theta_k\,|\,\{x_i,y_i\})$ из (2.15.48) приобретает вид

$$S(\theta_1,\theta_2 \mid \{x_i,y_i\}) = \sum_{i=1}^{n}[y_i - (\theta_1 x_i + \theta_2)]^2 ,$$

а система уравнений (2.15.49) записывается в такой форме

$$\frac{\partial S(\theta_1,\theta_2 \mid \{x_i,y_i\})}{\partial \theta_1} = -2\sum_{i=1}^{n}[y_i - (\theta_1 x_i + \theta_2)]x_i = 0 ,$$

$$\frac{\partial S(\theta_1,\theta_2 \mid \{x_i,y_i\})}{\partial \theta_2} = -2\sum_{i=1}^{n}[y_i - (\theta_1 x_i + \theta_2)] = 0$$

или ее можно переписать в виде следующей системы линейных уравнений:

$$\theta_1 \cdot \sum x_i^2 + \theta_2 \cdot \sum x_i = \sum x_i y_i ,$$
$$\theta_1 \cdot \sum x_i + \theta_2 \cdot n = \sum y_i .$$

Используя для решения систем линейных уравнений метод Крамера, получаем решение $(\hat{\theta}_1,\hat{\theta}_2)$ данной системы в виде

$$\hat{\theta}_1 = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} , \quad \hat{\theta}_2 = \frac{\begin{vmatrix} \sum x_i^2 & \sum x_i y_i \\ \sum x_i & \sum y_i \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{\sum x_i^2 \cdot \sum y_i - \sum x_i y_i \cdot \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} .$$

Часто эти оценки записывают в другой форме:

$$\hat{\theta}_1 = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \overline{x})y_i}{\sum (x_i - \overline{x})^2} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

и

$$\hat{\theta}_2 = \frac{\sum x_i^2 \cdot \sum y_i - \sum x_i y_i \cdot \sum x_i}{n\sum x_i^2 - (\sum x_i)^2} = \overline{y} - \hat{\theta}_1 \cdot \overline{x} ,$$

где

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Таким образом, оценкой прямой линии $y = \theta_1 x + \theta_2$, построенной методом наименьших квадратов по экспериментальным данным $(x_1,y_1),\ldots,(x_n,y_n)$, является *прямая наименьших квадратов* $\hat{y} = \hat{\theta}_1 x + \hat{\theta}_2$, которая записывается в виде

$$\hat{y} = \hat{\theta}_1 x + \hat{\theta}_2 = \overline{y} + \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}(x - \overline{x}) . \qquad \textbf{(2.15.50)}$$

Убедимся теперь, что точка $(\hat{\theta}_1,\hat{\theta}_2)$ является точкой минимума функции $S(\theta_1,\theta_2\,|\,\{x_i,y_i\})$, то есть в этой точке она достигает минимального значения. Для этого воспользуемся достаточными условиями существования экстремума многомерной функции. Учитывая, что для функции $S(\theta_1,\theta_2\,|\,\{x_i,y_i\})$ ее частные производные по переменным θ_1 и θ_2 записываются в виде

$$\partial S(\theta_1,\theta_2\,|\,\{x_i,y_i\})/\partial\theta_1 = -2\sum y_i x_i + 2\theta_1\cdot\sum x_i^2 + 2\theta_2\cdot\sum x_i,$$
$$\partial S(\theta_1,\theta_2\,|\,\{x_i,y_i\})/\partial\theta_2 = -2\sum y_i + 2\theta_1\cdot\sum x_i + 2\theta_2\cdot n,$$

получаем

$$\partial^2 S(\theta_1,\theta_2\,|\,\{x_i,y_i\})/\partial\theta_1^2 = 2\sum x_i^2 = A > 0,$$
$$\partial^2 S(\theta_1,\theta_2\,|\,\{x_i,y_i\})/\partial\theta_1\partial\theta_2 = 2\sum x_i = B,$$
$$\partial^2 S(\theta_1,\theta_2\,|\,\{x_i,y_i\})/\partial\theta_2^2 = 2n = C.$$

Отсюда следует, что

$$\Delta = AC - B^2 = 4n\sum x_i^2 - (2\sum x_i)^2 = 4\sum_{i<j}(x_i - x_j)^2 > 0.$$

Итак, $\Delta > 0$, $A > 0$, следовательно, согласно достаточным условиям существования экстремума функции многих переменных, функция $S(\theta_1,\theta_2\,|\,\{x_i,y_i\})$ достигает в точке $(\hat{\theta}_1,\hat{\theta}_2)$ минимального значения.

Отметим, что метод наименьших квадратов является самостоятельным методом оценивания параметров, однако если ошибки измерений $\varepsilon_1,...,\varepsilon_n$ в формуле (2.15.45) являются независимыми и нормальными случайными величинами, то МНК-оценки и МП-оценки, построенные методом максимального правдоподобия, совпадают. В самом деле, пусть $L(\varepsilon_i) = N(0,\sigma^2)$, $i = 1,...,n$, тогда наблюденные данные $y_1,...,y_n$ в (2.15.45) являются реализациями независимых случайных величин $Y_1,...,Y_n$ с нормальным распределением. Функция правдоподобия $L\{y_1,...y_n;\theta_1,...,\theta_k\}$ для заданной реализации $y_1,...,y_n$, рассматриваемая как многомерная функция параметров $\theta_1,...,\theta_k$, запишется в виде

$$L\{y_1,...y_n;\theta_1,...,\theta_k\} = \prod_{i=1}^{n} f_Y(y_i;\vec{\theta}) =$$
$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \prod_{i=1}^{n}\exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \varphi(x_i;\vec{\theta})\right)^2\right\}, \qquad \textbf{(2.15.51)}$$

а ее логарифм может быть представлен как

$$\ln L\{y_1,...y_n;\theta_1,...,\theta_k\} = \sum_{i=1}^{n} \ln f_Y(y_i;\vec{\theta}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - \varphi(x_i;\vec{\theta})]^2 .$$

Оценки максимального правдоподобия параметров $\theta_1,...,\theta_k$, находятся путем решения системы уравнений

$$\partial \ln L\{y_1,...y_n;\theta_1,...,\theta_k\}/\partial\theta_j = 0,\ j = 1,...,k, \qquad \textbf{(2.15.52)}$$

которая эквивалентна определению параметров $\theta_1,...,\theta_k$ из условия (2.15.47).

Итак, функция правдоподобия $L\{y_1,...y_n;\theta_1,...,\theta_k\}$ достигает максимума в том и только в том случае, если сумма $\sum_{i=1}^{n}[y_i - \varphi(x_i;\theta_1,...,\theta_k)]^2$ достигает минимума, то есть МНК-оценки и МП-оценки совпадают для случая, когда ошибки измерений $\varepsilon_1,...,\varepsilon_n$ в формуле (2.15.45) являются *независимыми и нормальными случайными величинами*.

Отметим, что в данном примере предполагалось, что случайная величина Y зависит от переменной x, которая *не является случайной величиной*. Случай, когда X и Y обе являются случайными величинами, подчиняющимися двумерному нормальному распределению, рассматривается в упражнении (2.15.16), в котором построена методом подстановки выборочная оценка регрессии X на Y по двумерной выборке $(X_1,Y_1),...,(X_n,Y_n)$ из нормального распределения (см. формулы (2.15.18б), (2.15.18в) и (2.15.50)).

2.15.53. Взвешенные наименьшие квадраты. Вернемся к предыдущему упражнению, в котором использовалась аддитивная модель наблюдений вида (2.15.45). Для этой модели предполагалось, что дисперсии ошибок измерений ε_i *одинаковы* для всех уровней x_i независимой переменной x. В литературе такие модели называют *гомоскедастическими*. Однако на практике не всегда такое предположение может быть обоснованным. Для того чтобы учесть зависимость дисперсии ошибок $D(\varepsilon_i)$ от x_i, рассматривают *гетероскедастические* модели, для которых эта зависимость учитывается с помощью мультипликативных постоянных, то есть дисперсия ошибок записывается в виде

$$D(\varepsilon_i) = \omega_i\sigma^2,\ 1 \le i \le n, \qquad \textbf{(2.15.54)}$$

где σ^2 – неизвестная дисперсия, как и в модели (2.15.45), а ω_i – известные веса. Отметим, что метод наименьших квадратов к гетероскедастической модели непосредственно неприменим. Для такой модели используют взвешенные наименьшие квадраты, к которым можно прийти с помощью следующих рассуждений. Рассмотрим случайные величины

$$Z_i = \frac{Y_i}{\sqrt{\omega_i}} = \frac{\varphi(x_i;\theta_1,...,\theta_k)}{\sqrt{\omega_i}} + \frac{\varepsilon_i}{\sqrt{\omega_i}},\ 1 \le i \le n. \qquad \textbf{(2.15.55)}$$

Отметим, что $D(\varepsilon_i/\sqrt{\omega_i}) = \omega_i\sigma^2/\omega_i = \sigma^2$. Следовательно, если положить

$$\varphi^*(x_i;\theta_1,...,\theta_k) = \varphi(x_i;\theta_1,...,\theta_k)/\sqrt{\omega_i}\ ,\ \varepsilon_i^* = \varepsilon_i/\sqrt{\omega_i}\ ,\ 1 \le i \le n\,,$$

то выражение (2.15.55) примет вид

$$Z_i = \varphi^*(x_i;\theta_1,...,\theta_k) + \varepsilon_i^*\ ,\ 1 \le i \le n\,. \qquad \textbf{(2.15.56)}$$

Таким образом, для величин Z_i мы получили гомоскедастическую модель (2.15.56), в которой ошибки $\varepsilon_1^*,...,\varepsilon_n^*$ удовлетворяют исходным допущениям (2.15.46) метода наименьших квадратов. При заданных значениях $z_i = y_i/\sqrt{\omega_i}$, $1 \le i \le n$, *оценки взвешенных наименьших квадратов* $\hat{\theta}_1^*,...,\hat{\theta}_k^*$ параметров $\theta_1,...,\theta_k$ находятся из условия $\min\limits_{\theta_1,...,\theta_k}\sum\limits_{i=1}^{n}[z_i - \varphi^*(x_i;\theta_1,...,\theta_k)]^2$, то есть при этих значениях параметров $\theta_1,...,\theta_k$ функция

$$S^*(\theta_1,...,\theta_k \mid \{x_i,z_i\}) = \sum_{i=1}^{n}[z_i - \varphi^*(x_i;\theta_1,...,\theta_k)]^2 = \sum_{i=1}^{n}\frac{1}{\omega_i}[y_i - \varphi(x_i;\theta_1,...,\theta_k)]^2$$

достигает минимального значения.

2.15.57. Имеется выборка $X_1,...,X_n$ из нормального распределения $N(a,\sigma^2)$, параметры a и σ^2 неизвестны. Оценками максимального правдоподобия параметров a и σ^2 являются соответственно выборочное среднее $\overline{X} = (1/n)\sum X_i$ и выборочная дисперсия $S_n^2 = n^{-1}\sum(X_i - \overline{X})^2$ (см. пример 2.10.7). Напомним, что оценка S_n^2 имеет смещение порядка $O(1/n)$, так как $M(S_n^2) = \sigma^2 - \sigma^2/n$. Убедиться, что применение метода «складного ножа» (см. замечание 6.5.22) к МП-оценке S_n^2 параметра σ^2 приводит к подправленной выборочной дисперсии вида

$$\tilde{S}_n^2 = [1/(n-1)]\sum(X_i - \overline{X})^2 = [1/(n-1)]\sum X_i^2 - [1/n(n-1)](\sum X_i)^2\,.$$

Решение. Учитывая, что выборочная дисперсия S_n^2 записывается в виде

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2\,,$$

рассмотрим статистики $S_{n-1,i}^2$, $i = 1,...,n$, которые построены по выборке объема $(n-1)$, полученной путем удаления элемента X_i, $i = 1,...,n$, из исходной выборки $X_1,...,X_n$, в виде

$$S^2_{n-1,i}=(n-1)^{-1}(X_1^2+...+X_{i-1}^2+X_{i+1}^2+...+X_n^2)-(n-1)^{-2}(X_1+...+X_{i-1}+X_{i+1}+...+X_n)^2=$$
$$=[1/(n-1)][(\sum X_i^2)-X_i^2]-[1/(n-1)^2][(\sum X_i)-X_i]^2,\ i=1,...,n.$$

Используя эти статистики, определим оценку $\hat{\theta}_n$ в виде

$$\hat{\theta}_n=\frac{1}{n}\sum_{i=1}^n S^2_{n-1,i}=\frac{1}{n}\sum_{i=1}^n X_i^2-\frac{1}{n(n-1)^2}\sum_{i=1}^n X_i^2-\frac{n-2}{n(n-1)^2}\left(\sum_{i=1}^n X_i\right)^2=\frac{(n-2)(\sum X_i^2-n\overline{X}^2)}{(n-1)^2}.$$

Далее, оценка $\hat{\theta}_n^*$ параметра σ^2, построенная методом «складного ножа» (см. 6.5.27), запишется как

$$\hat{\theta}_n^*=nS_n^2-(n-1)\hat{\theta}_n=\sum X_i^2-n^{-1}(\sum X_i)^2-(n-1)n^{-1}\sum X_i^2+$$
$$+[n(n-1)]^{-1}\sum X_i^2+[(n-2)/n(n-1)](\sum X_i)^2=$$
$$=(n-1)^{-1}\sum X_i^2-[1/n(n-1)](\sum X_i)^2=(n-1)^{-1}\sum(X_i-\overline{X})^2=\tilde{S}_n^2.$$

Таким образом, применяя метод «складного ножа» к МП-оценке S_n^2 параметра σ^2 нормального распределения, имеющей смещение порядка $O(1/n)$, мы получили подправленную выборочную дисперсию $\tilde{S}_n^2$, которая является несмещенной оценкой параметра σ^2 (см. упр. 2.15.13).

Глава 3

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

В данной главе обсуждаются общие понятия теории проверки статистических гипотез. В разд. 3.1 и 3.2 приводятся понятия статистического критерия, вероятностей ошибок и мощности критерия, а также примеры параметрических и непараметрических гипотез. В разд. 3.3 рассматриваются простые параметрические гипотезы и излагается подход Неймана – Пирсона к построению критериев. Параметрические сложные гипотезы обсуждаются в разд. 3.4. Критерии согласия описываются в разд. 3.5. Критерии однородности обсуждаются в разд. 3.6. Критерии для проверки гипотез независимости и случайности рассматриваются в разд. 3.7. Основные понятия дисперсионного анализа приведены в разд. 3.8. Упражнения, задачи и дополнения помещены в разд. 3.9. При написании данной главы использовались работы [5, 7, 22, 24, 30, 36, 37, 38].

3.1. Основные понятия теории проверки статистических гипотез

В математической статистике вместо общепринятого научного термина «гипотеза» используют понятие *статистическая гипотеза*. Под статистическими гипотезами понимают любые предположения относительно вида или общих свойств, или параметров распределений вероятностей наблюдаемых в эксперименте случайных величин. Статистические гипотезы разделяют на параметрические и непараметрические, а также на простые и сложные. Кроме того, вводят понятие основной (или нулевой) гипотезы, которая обычно обозначается через H_0, и понятие альтернативной гипотезы, или коротко альтернативы, которая обычно обозначается через H_1. Альтернативы, в свою очередь, различают на односторонние и дву-

сторонние. Отметим, что гипотезы H_0 и H_1 формально являются равноценными. Однако между ними можно отметить некоторую асимметрию, поскольку они вызывают различное отношение исследователей. Гипотеза H_0 обычно основывается на некоторых теоретических результатах и отражает устоявшуюся, принятую точку зрения, которая обычно формулируется на основе предположений равенства, симметрии или независимости. В противоположность этому, альтернативная гипотеза H_1 основывается на предположениях неравенства, отсутствия симметрии или появления зависимости, эти предположения и надеется подтвердить исследователь результатами опытов. В количественном отношении число различных вариантов при формулировке основной гипотезы относительно невелико (примеры приведены ниже). Далее, гипотеза H_0 обычно четко формулируется, и при этом распределение данных $X_1,...,X_n$ при условии, что гипотеза H_0 верна, является элементом однозначно заданного класса распределений. В противоположность этому, при формулировке альтернативных гипотез, мы часто ничего не можем сказать определенного относительно того, какой класс альтернативных распределений следует рассматривать. При этом число возможных альтернатив обычно огромно. В связи с этим математическое изучение критериев проверки гипотез H_0 и H_1 на основе статистических данных $X_1,...,X_n$ является более простой задачей при справедливости гипотезы H_0, изучение же свойств критериев при справедливости альтернативы H_1 обычно является более сложной задачей. Отсутствие однозначной определенности в формулировке альтернативы на практике проявляется в том, что отказу от гипотезы H_0 придается большее значение, чем ее принятию. В практической же деятельности мы обычно придерживаемся основной гипотезы до тех пор, пока результаты эксперимента не опровергают ее, и как только результаты опыта противоречат ей, основная гипотеза пересматривается.

Конкретизируем приведенные понятия. Пусть изучаемая случайная величина X имеет ф.р. $F_X \in \mathfrak{I}$, где $\mathfrak{I}$ – множество допустимых распределений вероятностей в условиях данного статисти-

ческого эксперимента, результатом которого является наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ из распределения F_X. Данный эксперимент характеризуется статистической моделью $\{\mathrm{X}, \Im\}$, где $\mathrm{X} = \{\vec{x}\}$ – выборочное пространство, на котором задается распределение вероятностей наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$ путем конкретизации совместной функции распределения вероятностей

$$F_{X_1,...,X_n}(x_1,...,x_n) = P\{X_1 \le x_1,...,X_n \le x_n\} = \prod_{i=1}^{n} F_X(x_i).$$

В тех случаях, когда множество допустимых распределений вероятностей $\Im$ представлено в виде параметрического семейства *заданных* функций с конечным числом *неизвестных* параметров θ, то есть в виде $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$, тогда в рамках такой параметрической модели $\{\mathrm{X}, \Im_\theta\}$ все статистические гипотезы являются *параметрическими* и формулируются в виде предположений относительно возможных значений неизвестного параметра θ, который в общем случае может быть векторным. Параметрическое множество Θ при этом разбивается на два множества Θ_0 и Θ_1, где Θ_0 – множество параметров, выделенное предположениями гипотезы H_0, а Θ_1 – множество параметров, выделенное предположениями альтернативы H_1, $\Theta_0 \cup \Theta_1 = \Theta$. Если же множество $\Im$ не может быть представлено в виде параметрического семейства функций, то для модели $\{\mathrm{X}, \Im\}$ статистические гипотезы формулируют в виде некоторых *общих предположений* относительно ф.р. $F_X \in \Im$ и называют *непараметрическими*. При этом в соответствии с предположениями гипотез H_0 и H_1 множество допустимых распределений $\Im$ разбивается на два класса $\Im_0$ и $\Im_1$, $\Im = \Im_0 \bigcup \Im_1$. Примеры формулировок различных вариантов параметрических и непараметрических гипотез приводятся ниже. Для удобства дальнейших ссылок приведем определения используемых понятий.

Определение 3.1.1. Статистическую гипотезу называют *параметрической*, если она формулируется на основе предположений о параметрах θ статистической модели $\{\mathrm{X}, \Im_\theta\}$.

Определение 3.1.2. Статистическую гипотезу называют *непараметрической*, если она содержит предположения относительно вида или некоторых общих свойств ф.р. $F_X \in \Im$ изучаемой в экспе-

рименте случайной величины X в рамках статистической модели $\{\mathrm{X}, \Im\}$.

Отметим важное обстоятельство, характеризующее отличие параметрической гипотезы от непараметрической. При формулировке параметрической гипотезы предполагается *известным функциональный характер ф.р.* $F_X(x,\theta) \in \Im_\theta$, а при формулировке непараметрической гипотезы *функциональный характер ф.р.* $F_X \in \Im$ *неизвестен.*

Определение 3.1.3. Статистическую гипотезу называют *простой*, если предположения этой гипотезы полностью определяют распределение вероятностей изучаемой с. в. X, что соответствует выделению отдельной точки на множестве $\Im$, в противном случае – гипотезу называют *сложной* и ее предположения выделяют некоторое подмножество на множестве $\Im$, например в виде параметрического семейства $\Im_\theta$.

3.1.А. Понятие статистического критерия и критической области

Отметим, что всякая проверка статистической гипотезы H_0 основана на использовании некоторого статистического критерия и заключается в том, чтобы, используя информацию, содержащуюся в статистических данных, обнаружить отклонения, если они имеются, от этой основной гипотезы H_0. Отклонения от гипотезы H_0 могут иметь самый различный характер. Эти отклонения уточняются при формулировке альтернативы H_1. В некоторых случаях характер отклонений является вполне ожидаемым, тогда используют так называемые односторонние альтернативы (см. 3.2.4). В тех случаях, когда отклонения от гипотезы H_0 являются непредсказуемыми и могут иметь произвольный характер, используют двусторонние альтернативы (см., например, 3.2.5).

Определение 3.1.4. Правило, которое позволяет обнаружить отклонения от основной гипотезы H_0 при заданной альтернативе H_1 на основе информации, содержащейся в выборке $\vec{X} = (X_1,...,X_n)$, называют *статистическим критерием* проверки гипотез H_0 и H_1.

Отметим, что в результате применения критерия выносится решение либо в пользу H_0, либо в пользу H_1, при этом можно либо прийти к правильному решению, либо совершить одну из двух ошибок: отвергнуть H_0, когда она верна (ошибка первого рода), или принять ее в ситуации, когда она неверна (ошибка второго рода). Вероятность отвергнуть гипотезу H_0, то есть принять альтернативу H_1 в ситуации, когда верна гипотеза H_0, называется *вероятностью ошибки первого рода*, она обычно обозначается через α и записывается в виде

$$\alpha = P\{H_1 \mid H_0\}. \qquad \textbf{(3.1.5)}$$

Вероятность отвергнуть альтернативу H_1, то есть принять гипотезу H_0 в ситуации, когда верна альтернатива H_1, называется *вероятностью ошибки второго рода*, она обычно обозначается через β и записывается в виде

$$\beta = P\{H_0 \mid H_1\}. \qquad \textbf{(3.1.6)}$$

Вероятность правильного принятия решения в пользу альтернативы H_1, когда она верна, называется *мощностью критерия*, обычно обозначается через W и записывается в виде

$$W = 1 - \beta = P\{H_1 \mid H_1\}. \qquad \textbf{(3.1.7)}$$

Обсудим теперь общую схему построения статистических критериев. Для построения конкретного критерия проверки гипотез H_0 и H_1 нужно до проведения эксперимента разбить выборочное пространство $\mathrm{X} = \{\vec{x}\}$ на два взаимно дополнительных множества X_0 и $\mathrm{X}_{1\alpha}$, $\mathrm{X}_0 \bigcap \mathrm{X}_{1\alpha} = \varnothing$ и $\mathrm{X}_0 \bigcup \mathrm{X}_{1\alpha} = \mathrm{X}$. Множество X_0 называют областью принятия гипотезы H_0, а множество $\mathrm{X}_{1\alpha}$ – областью отклонения гипотезы H_0, или критической областью размера α. Критическая область – это множество таких реализаций $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$, при которых отклоняется нулевая гипотеза. Если нулевая гипотеза справедлива и принят уровень значимости α, то с вероятностью α наблюденное значение $\vec{x} = (x_1,...,x_n)$ попа-

дает в критическую область $X_{1\alpha}$, и если в этом случае нулевая гипотеза будет ошибочно отклонена, то будет допущена ошибка первого рода.

Итак, благодаря особой роли основной гипотезы H_0 и отмеченной выше асимметрии между гипотезами H_0 и H_1, критическую область $X_{1\alpha}$ обычно выбирают из таких соображений, чтобы вероятность попасть выборочной точке $\vec{X} = (X_1,...,X_n)$ в эту область при справедливости гипотезы H_0, была бы не больше заданной, достаточно маленькой величины α, $0 < \alpha < 1$, то есть

$$P\{\vec{X} \in X_{1\alpha} \mid H_0\} \le \alpha \text{ для всех } F \in \Im_0. \qquad \textbf{(3.1.8)}$$

Замечание 3.1.9. Отметим, что при решении многих практических задач для величины α обычно используют стандартные значения: $\alpha = 0{,}1$, $\alpha = 0{,}05$, $\alpha = 0{,}01$, и называют эту величину α уровнем значимости критерия. Отметим также, что в литературе (см., например, [12, с. 30]) часто величину α называют *гарантированным уровнем значимости*, оставляя название *уровень значимости* для вероятности ошибки первого рода. С учетом этого замечания уровень значимости (или вероятность ошибки первого рода) представляет собой функционал $\alpha(F)$, заданный на множестве распределений $\Im_0$, которое определено предположениями гипотезы H_0, а гарантированный уровень значимости – это точная верхняя грань таких функционалов по всем $F \in \Im_0$, то есть $\sup P\{\vec{X} \in X_{1\alpha} \mid F \in \Im_0\}$. Эту верхнюю грань и называют *размером критерия*, или *размером критической области* $X_{1\alpha}$.

Определение 3.1.10. Размер критерия α, также называемый гарантированным уровнем значимости, является точной верхней гранью функционалов $\alpha(F) = P\{\vec{X} \in X_{1\alpha} \mid F \in \Im_0\}$ на множестве $\Im_0$, то есть $\alpha(F) : \Im_0 \to R^1$, $\alpha = \sup\{\alpha(F) : F \in \Im_0\}$, и эта верхняя грань совпадает с максимальной вероятностью ошибки первого рода.

После того когда выборочное пространство $X = \{\vec{x}\}$ разбито на два взаимно дополнительных множества X_0 и $X_{1\alpha}$, критерий проверки гипотез H_0 и H_1 формулируется в виде следующего правила:

«Если наблюденная реализация $\vec{x} = (x_1,...,x_n)$ *выборки* $\vec{X} = (X_1,...,X_n)$ *попала в критическую область* $\mathrm{X}_{1\alpha}$, *то гипотеза* H_0 *отвергается на уровне значимости* α (*принимается альтернатива* H_1), *если же* $x \in \mathrm{X}_0 = \overline{\mathrm{X}}_{1\alpha}$, *то гипотеза* H_0 *принимается как не противоречащая наблюденным данным* $\vec{x} = (x_1,...,x_n)$*».* **(3.1.11)**

Отметим, что выбор критической области $\mathrm{X}_{1\alpha}$ (или разбиение выборочного пространства $\mathrm{X} = \{\vec{x}\}$ на два взаимно дополнительных множества X_0 и $\mathrm{X}_{1\alpha}$) может быть проведен большим числом способов, поэтому необходимо сформулировать некоторые общие принципы, которые позволяли бы строить критерии проверки гипотез с нужными характеристиками качества. Выбор конкретного правила, или критерия проверки гипотезы H_0 против альтернативы H_1, сводится к заданию критического множества $\mathrm{X}_{1\alpha}$, и это множество обычно задается с помощью некоторой статистики критерия (или тестовой статистики), которую ниже будем обозначать через $T(X_1,...,X_n) = T(\vec{X})$. Соответствие между критической областью $\mathrm{X}_{1\alpha}$ и значением $T(\vec{x})$ статистики $T(\vec{X})$ при наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ записывается для двусторонних критериев в виде следующего выражения:

$$[\vec{x} \in \mathrm{X}_{1\alpha}] \Leftrightarrow [\, T(\vec{x}) \geq t_{1-\alpha/2} \;\; \textit{или} \;\; T(\vec{x}) \leq t_{\alpha/2} \,], \qquad \textbf{(3.1.12)}$$

а для односторонних критериев это соответствие записывается в виде

$$[\vec{x} \in \mathrm{X}^{+}_{1\alpha}] \Leftrightarrow [\, T(\vec{x}) \geq t_{1-\alpha} \,]; \qquad \textbf{(3.1.13)}$$

$$[\vec{x} \in \mathrm{X}^{-}_{1\alpha}] \Leftrightarrow [\, T(\vec{x}) \leq t_{\alpha} \,], \qquad \textbf{(3.1.14)}$$

где $t_{1-\alpha/2}$ и $t_{\alpha/2}$ соответственно называют верхним и нижним критическими значениями статистики $T(\vec{X})$ двусторонней критической области размера α. Отметим, что критерии, основанные на статистике $T(\vec{X})$, так же как и конкретный вид статистики $T(\vec{X})$, суще-

ственно зависят от характера высказанных гипотез H_0, H_1 и от принятого критерия оптимальности, положенного в основу построения критерия проверки гипотез. Согласно (3.1.12), критическая область размера α для критерия, основанного на статистике $T(\vec{X})$, при проверке гипотезы H_0 против двусторонней альтернативы H_1 (примеры смотри ниже) записывается в виде

$$\mathrm{X}_{1\alpha} = \{\vec{x} : t_{\alpha/2} \le T(\vec{x}) \ \textit{или} \ T(\vec{x}) \ge t_{1-\alpha/2}\}, \qquad \textbf{(3.1.15)}$$

где $t_{\alpha/2}$ – нижнее критическое значение статистики $T(\vec{X})$, $t_{1-\alpha/2}$ – верхнее критическое значение статистики $T(\vec{X})$ и α – уровень значимости критерия (или вероятность ошибки первого рода, которая возникает при отклонении верной гипотезы H_0). При проверке гипотезы H_0 против односторонних альтернатив H_1^+ и H_1^- (примеры смотри ниже) критические области размера α соответственно записываются в виде

$$\mathrm{X}_{1\alpha}^{+} = \{\vec{x} : T(\vec{x}) \ge t_{1-\alpha}\}, \ \mathrm{X}_{1\alpha}^{-} = \{\vec{x} : T(\vec{x}) \le t_{\alpha}\}. \qquad \textbf{(3.1.16)}$$

Таким образом, критерий (или правило вида (3.1.11)) для проверки гипотез H_0 и H_1 может быть сформулирован и в терминах тестовой статистики $T(\vec{X})$. В частности, для проверки гипотезы H_0 против *правосторонней* альтернативы H_1^+, критерий (3.1.11) с учетом (3.1.13) и (3.1.16) записывается в виде следующего правила:

«*Если наблюденное значение* $T(\vec{x})$ *статистики* $T(\vec{X})$, *вычисленное по наблюденной реализации* $\vec{x} = (x_1, ..., x_n)$ *выборки* $\vec{X} = (X_1, ..., X_n)$, *превышает критическое значение* $t_{1-\alpha}$, *то гипотеза* H_0 *отвергается на уровне значимости* α *и принимается альтернатива* H_1^+. *Если же вычисленное значение статистки критерия не превышает критическое значение* $t_{1-\alpha}$, *то нет оснований отвергать гипотезу* H_0, *то есть считается, что наблюденные данные* $\vec{x} = (x_1, ..., x_n)$ *не противоречат гипотезе* H_0».

Везде ниже критерий, основанный на статистике $T(\vec{X})$, будем кратко называть T-критерием. Для удобства дальнейших ссылок запишем сформулированное правило в компактной форме и присвоим ему номер ссылки.

«Если $T(\vec{x}) \geq t_{1-\alpha}$, *то гипотеза* H_0 *отвергается на уровне значимости* α, *если* $T(\vec{x}) < t_{1-\alpha}$, *то нет оснований отвергать* H_0*».* **(3.1.17)**

Замечание 3.1.18. Отметим, что для описания критической области критерия, основанного на статистике $T(\vec{X})$, удобно использовать критическую функцию $\varphi(\vec{x})$, которую для конкретной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ определяют в виде индикатора критической области, т.е. если $\varphi(\vec{x}) = 1$, то с помощью T-критерия гипотеза H_0 отвергается, а если $\varphi(\vec{x}) = 0$, то гипотеза H_0 принимается. В частности, для критерия (3.1.17) она определяется для конкретной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ в виде

$$\varphi(\vec{x}) = \begin{cases} 1, & \text{если } T(\vec{x}) \geq t_{1-\alpha}, \\ 0, & \text{если } T(\vec{x}) < t_{1-\alpha}. \end{cases} \qquad \textbf{(3.1.19)}$$

Отметим также, что использование критической функции удобно при построении общей теории критериев и, в частности, для построения рандомизированных критериев (см. разд. 3.3). Эти критерии строят в тех случаях, когда тестовая статистика имеет дискретное распределение, и поэтому не всегда возможно обеспечить равенство между размером критерия и его уровнем значимости. Такая ситуация характерна, например, для ранговых критериев. При сравнении же различных критериев по их мощности, в случае дискретных распределений тестовых статистик, прежде всего, необходимо обеспечить равенство для их уровней значимостей, это, в частности, достигается путем построения рандомизированных критериев (см. упр. 3.9.28) или путем сравнения критериев в асимптотике, когда точные распределения тестовых статистик аппроксимируется непрерывными распределениями.

Замечание 3.1.20. Отметим, что всякий статистический критерий, например в виде правила 3.1.17, не позволяет доказывать истинность гипотезы H_0. Когда наблюденные данные эксперимента согласуются с высказанной гипотезой, это еще не означает, что невозможно согласование этих же данных с другими гипотезами. Принятие гипотезы H_0 с помощью правила 3.1.17 вовсе не означает, что она единственно возможная, ее принятие означает лишь, что предположения гипотезы H_0 не противоречат наблюденным данным и таким же свойством, наряду с H_0, могут обладать и другие гипотезы. Таким обра-

зом, выводы, принимаемые на основе статистических критериев, формулируются в виде утверждений о том, согласуются ли экспериментальные данные с высказанной гипотезой H_0, или они противоречат ей, причем эти утверждения высказываются с определенной долей достоверности. В данном контексте вполне уместен афоризм: «истины нет, есть согласие».

3.1.Б. Вероятности ошибок и мощность критерия

Всякий статистический критерий характеризуется двумя типами ошибок. Ошибка первого рода возникает при отклонении верной гипотезы, а ошибка второго рода – при отклонении верной альтернативы. Вероятности этих ошибок были определены выражениями (3.1.5) и (3.1.6). Приведем формулы вычисления вероятностей ошибок на примере критерия, определенного правилом 3.1.17, для проверки гипотезы H_0 против *правосторонней* альтернативы H_1^+. Для вычисления вероятностей ошибок, а также для конкретизации способа задания критического значения $t_{1-\alpha}$ в правиле 3.1.17, обозначим плотность распределения вероятностей тестовой статистики $T(\vec{X})$ при справедливости гипотезы H_j через $f_T(x \mid H_j)$, а функцию распределения статистики $T(\vec{X})$ обозначим через $F_T(x \mid H_j)$, $j = 0, 1$.

Определение 3.1.21. Вероятность ошибки первого рода $\alpha = \alpha(F)$ для критерия вида 3.1.17 вычисляется по формуле

$$\alpha(F) = P\{T(\vec{X}) > t_{1-\alpha} \mid H_0\} = \int_{t_{1-\alpha}}^{+\infty} f_T(x \mid H_0)dx,\ F \in \Im_0, \quad \textbf{(3.1.22)}$$

или

$$\alpha(F) = 1 - P\{T(\vec{X}) \le t_{1-\alpha} \mid H_0\} = 1 - F_T(t_{1-\alpha} \mid H_0),\ F \in \Im_0. \quad \textbf{(3.1.23)}$$

Отметим, что критическое значение $t_{1-\alpha}$ статистики $T(\vec{X})$ вычисляется по *заданной* величине α из выражений (3.1.22) или (3.1.23). Из (3.1.23) также следует, что $t_{1-\alpha} = F_T^{-1}\{(1-\alpha) \mid H_0\}$, и это означает, что критическое значение $t_{1-\alpha}$ статистики $T(\vec{X})$ по смыслу является *квантилем уровня* $(1-\alpha)$ *функции распределения веро-*

ятностей $F_T(x \mid H_0)$ *тестовой статистики* $T(\vec{X})$ *при справедливости основной гипотезы* H_0.

Определение 3.1.24. Вероятность ошибки второго рода $\beta = \beta(F)$ для критерия вида 3.1.17 вычисляется по формуле

$$\beta(F) = P\{T(\vec{X}) < t_{1-\alpha} \mid H_1\} = \int_{-\infty}^{t_{1-\alpha}} f_T(x \mid H_1)dx = \\ = F_T(t_{1-\alpha} \mid H_1), F \in \Im_1. \qquad \textbf{(3.1.25)}$$

Определение 3.1.26. Вероятность правильного решения в пользу альтернативы H_1 при условии, что она верна, называют мощностью критерия $W = W(F)$. Для критерия вида 3.1.17 она вычисляется по формуле

$$W(F) = 1 - \beta(F) = P\{T(\vec{X}) \geq t_{1-\alpha} \mid H_1\} = \int_{t_{1-\alpha}}^{+\infty} f_T(x \mid H_1)dx = \\ = 1 - F_T(t_{1-\alpha} \mid H_1),\ F \in \Im_1. \qquad \textbf{(3.1.27)}$$

Для наглядности данных определений см. рис. 3.1.1

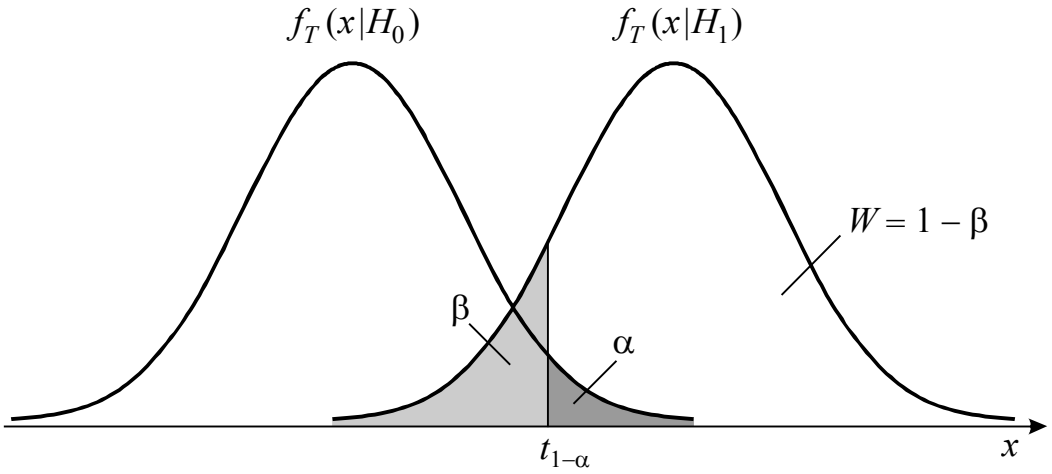

Рис. 3.1.1. Вероятности ошибок α и β для *T*-критерия проверки гипотез H_0, H_1^+

Замечание 3.1.28. Отметим, что при рассмотрении общих статистических моделей $\{\mathrm{X}, \Im\}$, где X – выборочное пространство, а $\Im$ – множество допустимых распределений вероятностей, мощность $W(F)$ является *функционалом*,

заданным на множестве альтернативных распределений $\Im_1 = \Im \setminus \Im_0$, то есть $W(F): \Im_1 \to R^1$. При рассмотрении же параметрических моделей $\{\mathrm{X}, \Im_\theta\}$, где $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ – заданное параметрическое семейство распределений, множество параметров Θ разбивается в соответствии с предположениями гипотезы H_0 и альтернативы H_1 на два множества Θ_0 и Θ_1, $\Theta_0 \cup \Theta_1 = \Theta$, и при изучении критериев проверки гипотез для таких параметрических моделей удобно обе вероятности ошибок первого и второго рода описывать с помощью одного понятия *функции мощности*. Для критерия проверки гипотез H_0 и H_1, основанного на статистике $T(\vec{X})$ с критической областью $\mathrm{X}_{1\alpha}$, функция мощности $W(\theta) = W(\mathrm{X}_{1\alpha};\theta)$, $\theta \in \Theta$ определяется в виде

$$W(\mathrm{X}_{1\alpha};\theta) = P_\theta\{\vec{X} \in \mathrm{X}_{1\alpha} \mid \theta \in \Theta\} = P_\theta\{T(\vec{X}) \geq t_{1-\alpha} \mid \theta \in \Theta\}. \qquad \textbf{(3.1.29)}$$

Если $\theta \in \Theta_0$, то $W(\mathrm{X}_{1\alpha};\theta)$ есть не что иное, как вероятность ошибки первого рода $\alpha(\theta)$, $\theta \in \Theta_0$, а гарантированный уровень значимости α, или размер критерия, равен максимальной вероятности ошибки первого рода, то есть $\alpha = \sup\{\alpha(\theta) : \theta \in \Theta_0\}$. Если же $\theta \in \Theta_1$, то $W(\mathrm{X}_{1\alpha};\theta)$ характеризует мощность T-критерия при заданном альтернативном значении параметра θ. Отметим, что для критерия, заданного критической функцией $\varphi(\vec{x})$ вида (3.1.19), функция мощности определяется как

$$W(\theta) = W(\mathrm{X}_{1\alpha};\theta) = P_\theta\{\textit{отвергнуть } H_0\} = P_\theta\{\varphi(X) = 1\} =$$

$$= M_\theta[\varphi(\vec{X})] = \int_{\mathrm{X}} \cdots \int \varphi(\vec{x}) \prod_{i=1}^{n} f_X(x_i;\theta) dx_i\,, \ \theta \in \Theta. \qquad \textbf{(3.1.30)}$$

Замечание 3.1.31. Отметим еще раз, что вероятность ошибки первого рода α часто также называют уровнем значимости критерия (см. замечание 3.1.9). Кроме того, стандартная процедура вида 3.1.17 проверки нулевой гипотезы H_0 с помощью T-критерия состоит в том, что по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ мы вычисляем наблюденное значение $T(\vec{x})$ статистики $T(\vec{X})$, и если значение $T(\vec{x})$ попадает в критическую область $\mathrm{X}_{1\alpha}$ данного критерия, то гипотеза H_0 отвергается на *заданном* уровне значимости α. Однако часто удобным является использование другой, более информативной процедуры. Вместо того чтобы работать с *фиксированным* уровнем значимости α и только соглашаться с отклонением или принятием гипотезы H_0, можно вычислять уровень значимости по наблюденному значению $T(\vec{x})$ статистики $T(\vec{X})$. Вычисленное при этом значение $\alpha_\partial(\vec{x})$ называют *достигнутым уровнем значимости*, или *минимальным уровнем значимости*, на котором можно отвергнуть нулевую гипотезу. Отметим,

что в литературе (см., например, [3, с. 439; 9, с. 113]) вместо термина *достигнутый уровень значимости* часто используют понятие «*P-значение критерия*». Формулы для вычисления *P-значения критерия* меняются в зависимости от рассматриваемых альтернативных гипотез. В частности, *P-значением* критерия, основанного на *левосторонней* статистике $T(\vec{X})$ с критической областью $X^{-}_{1\alpha} = \{\vec{x} : T(\vec{x}) \le t_\alpha\}$, называют значение функции распределения тестовой статистики $T(\vec{X})$ при справедливости гипотезы H_0 в точке, равной наблюденному значению $T(\vec{x})$ статистики $T(\vec{X})$, то есть

$$P\text{-значение} = P\{T(\vec{X}) \le T(\vec{x}) \mid H_0\} = F_T(T(\vec{x}) \mid H_0), \qquad \textbf{(3.1.31а)}$$

что совпадает с площадью под плотностью $f_T(x \mid H_0)$ слева от точки, равной наблюденному значению $T(\vec{x})$ тестовой статистики $T(\vec{X})$ (рис. 3.1.2).

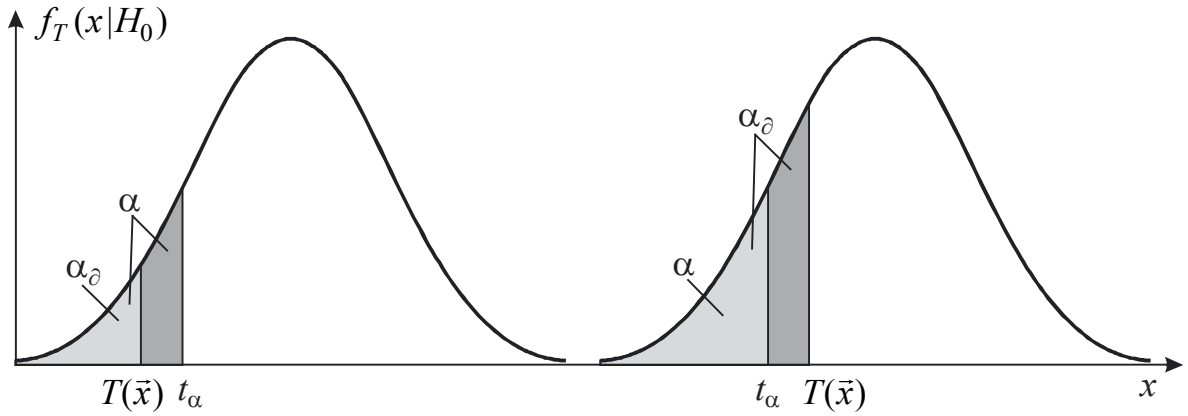

Рис. 3.1.2. Достигнутый уровень значимости α_∂ для *T*-критерия проверки гипотез H_0, H_1^-

Для критерия с правосторонней статистикой $T(\vec{X})$, определяемого критической областью $X^{+}_{1\alpha} = \{\vec{x} : T(\vec{x}) \ge t_{1-\alpha}\}$, *P-значение* критерия вычисляется по следующей формуле:

$$P\text{-значение} = P\{T(\vec{X}) \ge T(\vec{x}) \mid H_0\} = 1 - F_T(T(\vec{x}) \mid H_0), \qquad \textbf{(3.1.31б)}$$

то есть *P-значение* критерия численно равно площади под плотностью $f_T(x \mid H_0)$ справа от точки, равной наблюденному значению $T(\vec{x})$ тестовой статистики $T(\vec{X})$.

Отметим также, что для применения критерия, например вида 3.1.17, основанного на статистике $T(\vec{X})$, достаточно иметь лишь квантили *заданных* уровней распределения тестовой статистики $T(\vec{X})$ при гипотезе H_0, а для

вычисления достигнутого уровня значимости $\alpha_\partial(\vec{x})$ нужно знать функцию распределения вероятностей тестовой статистики $T(\vec{X})$ при гипотезе H_0 либо нужно иметь ее подробные таблицы. При использовании достигнутого уровня значимости правило 3.1.17 формулируется следующим образом: «Если $\alpha_\partial(\vec{x})$ «маленькая» величина, то гипотеза H_0 отвергается». Ответ на вопрос, что считать «маленькой величиной» для достигнутого уровня $\alpha_\partial(\vec{x})$, остается за исследователем. На практике часто «маленьким значением» считают величину $\alpha_\partial(\vec{x}) < 0{,}05$. Отметим, что чем большее значение $T(\vec{x})$ статистики $T(\vec{X})$ мы наблюдаем, тем меньше будет величина достигнутого уровня значимости $\alpha_\partial(\vec{x})$ и тем сильнее наблюденные данные $\vec{x} = (x_1,...,x_n)$ будут свидетельствовать против основной гипотезы H_0 при альтернативе H_1^+. Принятие решения с использованием фактически достигаемого уровня значимости не приводит к субъективному вмешательству статистика в решающую процедуру и не противоречит подходу Неймана – Пирсона, в котором уровень значимости задается до проведения эксперимента.

Определение 3.1.32. При проверке гипотезы H_0 против *правосторонней* альтернативы H_1^+ с помощью T-критерия, случайную величину $\alpha_\partial(\vec{X})$ называют *фактически достигнутым уровнем значимости*, если ее реализация $\alpha_\partial(\vec{x})$, равная вероятности того, что статистика $T(\vec{X})$ превышает наблюденное значение $T(\vec{x})$ при справедливости гипотезы H_0, вычисляется по формуле

$$\alpha_\partial(\vec{x}) = P\{T(\vec{X}) \geq T(\vec{x}) \mid H_0\} =$$

$$= \int_{T(\vec{x})}^{+\infty} f_T(x \mid H_0)dx = 1 - F_T(T(\vec{x}) \mid H_0). \qquad \textbf{(3.1.33)}$$

Отметим, что статистика $\alpha_\partial(\vec{X})$ при справедливости гипотезы H_0, то есть для $F \in \Im_0$, имеет равномерное распределение в интервале $[0,1]$, причем критическая область вида $\{\vec{x} : T(\vec{x}) \geq t_{1-\alpha}\}$ с уровнем значимости α эквивалентна критической области $\{\vec{x} : \alpha_\partial(\vec{x}) \leq \alpha\}$ с тем же самым уровнем значимости α.

Итак, достигнутый уровень значимости $\alpha_\partial(\vec{x})$ совпадает с *P-значением* критерия, и в частности, если он основан на правосто-

ронней статистике $T(\vec{X})$, то справедливо следующее выражение:

$$\alpha_{\partial}(\vec{x}) = P\text{-значение} = 1 - F_T(T(\vec{x}) \mid H_0). \qquad \textbf{(3.1.33а)}$$

Отметим еще раз, что понятие фактически достигаемого уровня значимости не противоречит «четкой теории Неймана – Пирсона» (см. *Прим. ред.* в книге [12, с. 33]).

Замечание 3.1.34. Все приведенные выше определения были сформулированы для случая проверки гипотезы H_0 против *правосторонней* альтернативы H_1^+. При проверке гипотезы H_0 против *левосторонней* альтернативы H_1^- либо против *двусторонней* альтернативы H_1, эти определения должны быть естественным образом изменены с учетом того, что для *левосторонней* альтернативы H_1^- критическая область определяется в виде (3.1.16), а для *двусторонней* альтернативы H_1 она определяется выражением (3.1.15). Например, *P-значение* $T(\vec{X})$-критерия проверки гипотез H_0 и H_1^-, определяемого критической областью $\mathrm{X}_{1\alpha}^- = \{\vec{x} : T(\vec{x}) \le t_\alpha\}$, вычисляется по формуле

$$P\text{-значение} = P\{T(\vec{X}) \le T(\vec{x}) \mid H_0\} = F_T(T(\vec{x}) \mid H_0), \qquad \textbf{(3.1.34а)}$$

а для двустороннего критерия проверки гипотез H_0 и H_1, определяемого критической областью $\mathrm{X}_{1\alpha} = \{\vec{x} : t_{\alpha/2} \le T(\vec{x})$ *или* $T(\vec{x}) \ge t_{1-\alpha/2}\}$, *P-значение* критерия вычисляется по формуле

$$P\text{-значение} = 2\min[P\{T(\vec{X}) \ge T(\vec{x}) \mid H_0\}, P\{T(\vec{X}) \le T(\vec{x}) \mid H_0\}]. \qquad \textbf{(3.1.34б)}$$

Окончательное решение на основе вычисленного *P-значения* критерия принимается следующим образом. Если *P-значение критерия меньше уровня значимости* α, *то гипотеза* H_0 *отвергается* (см. рис. 3.1.3).

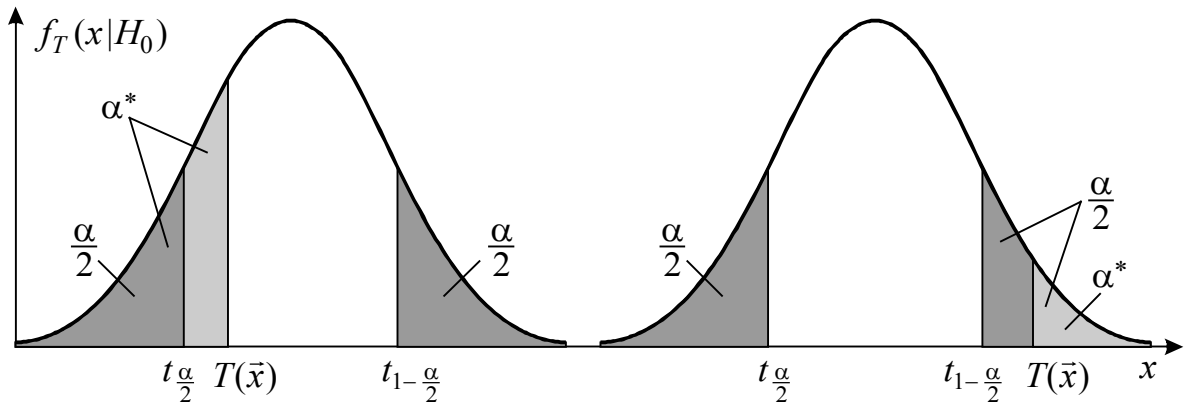

Рис. 3.1.3. Достигнутый уровень значимости $\alpha_{\partial} = 2\alpha^*$ для двустороннего T-критерия проверки гипотез H_0, H_1

Замечание 3.1.35. Отметим, что любое монотонное преобразование над тестовой статистикой критерия сохраняет его свойства. В частности, если T-критерий характеризуется критической областью вида $X_{1\alpha} = \{\vec{x} : T(\vec{x}) \geq t_{1-\alpha}\}$, то эквивалентный критерий, основанный на статистике $\psi[T(\vec{X})]$, где ψ – монотонная функция, характеризуется критической областью вида

$$X^*_{1\alpha} = \{\vec{x} : \psi[T(\vec{x})] \geq \psi(t_{1-\alpha})\}, \text{ или } X^*_{1\alpha} = \{\vec{x} : \psi[T(\vec{x})] \leq \psi(t_{\alpha})\}.$$

Обсудим теперь один из общих подходов к построению статистик критериев. Отмеченная выше асимметрия между основной гипотезой H_0 и альтернативной гипотезой H_1 приводит на практике к различным последствиям, вызванным ошибками первого и второго рода. Относительная значимость ошибок α и β, совершаемых при проверке гипотез H_0 и H_1, часто оказывается неодинаковой. Например, принятие партии с избыточным количеством дефектных изделий может оказаться более серьезной ошибкой и иметь серьезные последствия, чем возвращение «доброкачественной» партии изготовителю. В связи с этим, существует общепринятая практика (см., например, [5, с. 182]), согласно которой при решении многих практических задач вероятность ошибки первого рода α, возникающая при отклонении верной гипотезы H_0, является более серьезной ошибкой для экспериментатора. Конечно, при построении критериев было бы желательно, чтобы обе вероятности ошибок α и β были маленькими величинами. Однако эти ошибки связаны между собой, и сделать их одновременно маленькими величинами невозможно, уменьшение одной из них приводит к увеличению другой. По этой причине, один из общих принципов выбора статистики критерия формулируется следующим образом:

«Статистику $T(\vec{X})$ критерия следует выбирать таким образом, чтобы при заданной достаточно малой вероятности ошибки первого рода α, достигалась бы максимальная мощность W критерия». **(3.1.36)**

Реализация сформулированного принципа сводится к решению экстремальной задачи. В частности, при проверке простых параметрических гипотез (3.3.1) и (3.3.3) задача сводится к максимиза-

ции по критической области $\mathrm{X}_{1\alpha}$ мощности $W(\mathrm{X}_{1\alpha};\theta_1)$ при ограничении $W(\mathrm{X}_{1\alpha};\theta_0)=\alpha$. Решение этой экстремальной задачи было найдено Ю. Нейманом и Э. Пирсоном в 1933 году. Оно известно в литературе под названием фундаментальной леммы Неймана – Пирсона, обсуждение которой имеется во многих книгах (см., например, [22, 24, 66] и разд. 3.3).

3.1.В. Общие свойства критериев и их сравнение

Для описания общих свойств критериев проверки статистических гипотез, так же как и для оценок параметров, используют такие понятия, как несмещенность и состоятельность. Определим эти понятия для критериев на примере проверки параметрических гипотез статистических моделей $\{\mathrm{X},\Im_\theta\}$, где множество $\Im_\theta=\{F_X : F_X(x,\theta),\theta\in\Theta\}$ является заданным параметрическим семейством распределений. Рассмотрим сначала случай простых гипотез H_0 и H_1, т.е. предполагаем, что параметрическое множество Θ состоит только из двух элементов $(\theta_0,\theta_1)\in\Theta$ и $H_0:\theta=\theta_0$, а $H_1:\theta=\theta_1$.

Несмещенные критерии. Пусть для проверки сформулированных простых гипотез H_0 и H_1 используется *T*-критерий с критической областью $\mathrm{X}_{1\alpha}$, который характеризуется мощностью $W(\mathrm{X}_{1\alpha};\theta)=P_\theta\{T(\vec{X})\in\mathrm{X}_{1\alpha}\}$, $\theta\in\Theta=\{\theta_0,\theta_1\}$. Если при проверке этих гипотез с помощью *T*-критерия вероятность $P(H_1\,|\,H_1)$ правильного решения в пользу альтернативы H_1 в ситуации, когда она верна, меньше вероятности $P(H_1\,|\,H_0)$ принять альтернативу H_1 в ситуации, когда она неверна, то есть выполняется неравенство

$$W(\mathrm{X}_{1\alpha};\theta_1)=P_{\theta_1}\{T(\vec{X})\in\mathrm{X}_{1\alpha}\}<W(\mathrm{X}_{1\alpha};\theta_0)=P_{\theta_0}\{T(\vec{X})\in\mathrm{X}_{1\alpha}\}, \qquad \textbf{(3.1.37)}$$

то критерий называют *смещенным* и понятно, что такое свойство в общем случае нежелательно, так как смещенный критерий чаще выбирает нулевую гипотезу, когда она неверна, чем когда она верна.

Рассмотрим теперь случай сложных гипотез H_0^* и H_1^*, то есть в рамках параметрической модели $\{\mathrm{X}, \mathfrak{I}_\theta\}$ проверяется гипотеза $H_0^* : \theta \in \Theta_0$ против альтернативы $H_1^* : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$ с помощью *T*-критерия, который характеризуется критической областью $\mathrm{X}_{1\alpha}$ и функцией мощности $W(\mathrm{X}_{1\alpha}; \theta)$, $\theta \in \Theta$, определяемой в виде

$$W(\mathrm{X}_{1\alpha}; \theta) = P_\theta\{T(\vec{X}) \in \mathrm{X}_{1\alpha} \mid \theta \in \Theta\}. \qquad \textbf{(3.1.38)}$$

Определение 3.1.39. Критерий проверки сложных гипотез $H_0^* : \theta \in \Theta_0$ и $H_1^* : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$, характеризуемый функцией мощности $W(\mathrm{X}_{1\alpha}; \theta)$, $\theta \in \Theta$, называют *несмещенным*, если выполняются следующие условия:

$$W(\mathrm{X}_{1\alpha}; \theta) \le \alpha \ \textit{для всех}\ \theta \in \Theta_0; \qquad \textbf{(3.1.40)}$$

$$W(\mathrm{X}_{1\alpha}; \theta) \ge \alpha \ \textit{для всех}\ \theta \in \Theta_1. \qquad \textbf{(3.1.41)}$$

Замечание 3.1.42. Напомним, что для произвольной статистической модели $\{\mathrm{X}, \mathfrak{I}\}$, при проверке гипотезы $H_0 : F \in \mathfrak{I}_0$, против альтернативы $H_1 : F \in \mathfrak{I}_1$, $\mathfrak{I}_0 \bigcup \mathfrak{I}_1 = \mathfrak{I}$, вместо понятия *функция мощности* используется понятие *функционала мощности* $W(F) = W(\mathrm{X}_{1\alpha}; F)$, заданного на множестве допустимых распределений $\mathfrak{I}$. Причем $W(\mathrm{X}_{1\alpha}; F \mid F \in \mathfrak{I}_0)$ характеризует вероятность ошибки первого рода, а $W(\mathrm{X}_{1\alpha}; F \mid F \in \mathfrak{I}_1)$ – мощность критерия при альтернативном распределении $F \in \mathfrak{I}_1$. С использованием понятия функционала мощности $W(F) = W(\mathrm{X}_{1\alpha}; F)$, $F \in \mathfrak{I}$, критерий проверки сформулированных сложных гипотез H_0 и H_1 уровня значимости α называют *несмещенным*, если выполняется неравенство $W(\mathrm{X}_{1\alpha}; F \mid H_1) \ge \ge W(\mathrm{X}_{1\alpha}; F \mid H_0)$, или точнее, выполняются следующие условия:

$$W(\mathrm{X}_{1\alpha}; F \mid F \in \mathfrak{I}_0) \le \alpha; \qquad \textbf{(3.1.43)}$$

$$W(\mathrm{X}_{1\alpha}; F \mid F \in \mathfrak{I}_1) \ge \alpha. \qquad \textbf{(3.1.44)}$$

Таким образом, для несмещенного критерия вероятность отвергнуть нулевую гипотезу, когда она верна, не превышает уровень значимости α, а в тех случаях, когда гипотеза H_0 не верна (т. е. верна альтернатива H_1), гипотеза H_0 отвергается с вероятностью, большей α (т.е. вероятность принять верную альтернативу превышает уровень значимости α).

Состоятельные критерии. Свойство состоятельности любой статистической процедуры состоит в повышении качества процедуры при увеличении объема выборки $\vec{X} = (X_1,...,X_n)$, точнее при $n \to \infty$. Качество критериев характеризуется функцией (либо функционалом) мощности, которая зависит от уровня значимости α, от параметра, характеризующего степень различия гипотез H_0 и H_1, и от объема выборки n.

Определение 3.1.45. Критерий проверки сложных гипотез $H_0^*: \theta \in \Theta_0$ и $H_1^*: \theta \in \Theta_1,\ \Theta_0 \cup \Theta_1 = \Theta$, называют *состоятельным*, если его функция мощности $W(\mathrm{X}_{1\alpha};\theta),\ \theta \in \Theta$, стремиться к единице при увеличении объема выборки, то есть для состоятельного критерия выполняется выражение

$$\lim_{n\to\infty} W(\mathrm{X}_{1\alpha};\theta \mid \theta \in \Theta_1) = 1. \qquad \textbf{(3.1.46)}$$

Отметим, что для состоятельных критериев функция мощности $W(\theta) = W(\mathrm{X}_{1\alpha};\theta),\ \theta \in \Theta$, характеризующая вероятность принятия правильного решения в пользу альтернативы, когда она верна, для случая простых гипотез $H_0: \theta = \theta_0,\ H_1: \theta = \theta_1,\ \theta_0,\ \theta_1 \in \Theta$, стремится к ступенчатой функции при увеличении объема выборки, то есть $\lim_{n\to\infty} W(\mathrm{X}_{1\alpha};\theta_0) = \alpha$ и $\lim_{n\to\infty} W(\mathrm{X}_{1\alpha};\theta_1 \mid \theta_1 \neq \theta_0) = 1$, см. рис. 3.3.1. Таким образом, состоятельный критерий является асимптотически несмещенным критерием.

Равномерно наиболее мощные критерии. Для сравнения критериев между собой используют различные понятия, такие как: *равномерно наиболее мощный* (р.н.м.) критерий; *локально наиболее мощный* (л.н.м.) критерий для проверки сложных гипотез $H_0^*: \theta \in \Theta_0,\ H_1^*: \theta \in \Theta_1,\ \Theta_0 \cup \Theta_1 = \Theta$; и *наиболее мощный* (н.м.) критерий для проверки простых гипотез $H_0: \theta = \theta_0,\ H_1: \theta = \theta_1,\ \theta_0, \theta_1 \in \Theta$. Сравнение критериев при больших объемах выборки, то есть в асимптотике, осуществляется обычно с помощью понятия *асимптотическая относительная эффективность* Питмена, см. П.3.Е.

Пусть для проверки сложных гипотез H_0^* и H_1^* имеется два критерия с одинаковыми уровнями значимости α и критическими областями $X_{1\alpha}$ и $X_{1\alpha}^*$, то есть справедливо выражение $W(X_{1\alpha}^*;\theta \mid \theta \in \Theta_0) = W(X_{1\alpha};\theta \mid \theta \in \Theta_0) \le \alpha$.

Определение 3.1.47. Для проверки сложных гипотез $H_0^* : \theta \in \Theta_0$ и $H_1^* : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$ критерий с критической областью $X_{1\alpha}^*$ называется *равномерно наиболее мощным* (*р.н.м.*), если для *любого критерия* уровня значимости α с критической областью $X_{1\alpha}$ выполняются выражения

$$W(X_{1\alpha}^*;\theta \mid \theta \in \Theta_0) = W(X_{1\alpha};\theta \mid \theta \in \Theta_0) \le \alpha; \qquad \textbf{(3.1.48)}$$

$$W(X_{1\alpha}^*;\theta \mid \theta \in \Theta_1) \ge W(X_{1\alpha};\theta \mid \theta \in \Theta_1) \text{ для всех } \theta \in \Theta_1, \qquad \textbf{(3.1.49)}$$

причем строгое неравенство (3.1.49) выполняется хотя бы для одного $\theta \in \Theta_1$.

Таким образом, среди всех критериев уровня значимости α равномерно наиболее мощный критерий с критической областью $X_{1\alpha}^*$ обладает большей мощностью *для всех* $\theta \in \Theta_1$.

Отметим, что для простых гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, $\theta_0, \theta_1 \in \Theta$, параметрической модели $\{X, \Im_\theta\}$ вместо термина *равномерно наиболее мощный критерий* используют термин *наиболее мощный критерий*. Наиболее мощный критерий с критической областью $X_{1\alpha}^*$ для проверки простых гипотез H_0 и H_1 обладает большей мощностью среди всех критериев уровня значимости α, то есть

$$W(X_{1\alpha}^*;\theta \mid \theta = \theta_1) \ge W(X_{1\alpha};\theta \mid \theta = \theta_1). \qquad \textbf{(3.1.50)}$$

Отметим также, что в литературе (см., например, [24, с. 225]) критическую область, мощность которой не меньше мощности любой другой области того же размера, предназначенной для проверки гипотезы H_0 против альтернативной гипотезы H_1, называют наи-

лучшей критической областью (НКО), а критерий, основанный на НКО, называют наиболее мощным (н.м.) критерием.

Для построения наиболее мощных критериев при проверке простых параметрических гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, $\theta_0, \theta_1 \in \Theta$, развит подход Неймана – Пирсона, который обсуждается в разд. 3.4. Отметим, что (р.н.м.) критерии удается построить лишь в отдельных случаях параметрических моделей $\{\mathrm{X}, \mathfrak{I}_\theta\}$, в задачах проверки гипотез вида $H_0 : \theta \le \theta_0$, $H_1 : \theta > \theta_0$, где $\theta \in \Theta$ – вещественный параметр. Эти отдельные случаи и конкретные примеры построения р.н.м. критериев приводятся ниже (см. также упр. 3.9.40). В тех случаях, когда р.н.м. критерий не существует, часто удается построить *несмещенный* равномерно наиболее мощный критерий, либо *локально наиболее мощный* (л.н.м.) несмещенный критерий. Следуя работе [22], приведем определения этих понятий.

Определение 3.1.51. Для проверки сложных гипотез $H_0^* : \theta \in \Theta_0$ и $H_1^* : \theta \in \Theta_1$, $\Theta_0 \cup \Theta_1 = \Theta$ критерий с критической областью $\mathrm{X}_{1\alpha}$ и функцией мощности $W(\theta) = W(\mathrm{X}_{1\alpha}; \theta)$, $\theta \in \Theta$, называется *несмещенным равномерно наиболее мощным* (*н.р.н.м.*) критерием уровня значимости α, если выполняются следующие два выражения:

$$W(\mathrm{X}_{1\alpha}; \theta \mid \theta \in \Theta_0) \le \alpha \text{ для всех } \theta \in \Theta_0; \qquad \textbf{(3.1.52)}$$

$$W(\mathrm{X}_{1\alpha}; \theta \mid \theta \in \Theta_1) \ge \alpha \text{ для всех } \theta \in \Theta_1. \qquad \textbf{(3.1.53)}$$

Определим теперь понятие *локально наиболее мощный* (л.н.м.) критерий. Введем это понятие на примере проверки простой гипотезы $H_0 : \theta = \theta_0$ против сложной правосторонней альтернативы $H_1^+ : \theta > \theta_0$ с помощью критерия уровня значимости α с критической областью $\mathrm{X}_{1\alpha}$ и функцией мощности $W(\theta) = W(\mathrm{X}_{1\alpha}; \theta)$, $\theta \in \Theta$. Локально наиболее мощные критерии часто применяют на практике и их строят, для случаев «малых» отклонений от гипотезы $H_0 : \theta = \theta_0$, путем использования ограничений на поведение функции мощности $W(\theta) = W(\mathrm{X}_{1\alpha}; \theta), \theta \in \Theta$, в окрестности точки θ_0.

Локальные наиболее мощные критерии. Рассмотрим параметрическую модель $\{\mathrm{X}, \mathfrak{I}_\theta\}$, где $\mathfrak{I}_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ – заданное параметрическое семейство распределений. Пусть имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta)$ с плотностью $f_X(x;\theta)$, $x \in R^1$, и θ – неизвестный вещественный параметр. Требуется проверить простую гипотезу $H_0 : \theta = \theta_0$ против сложной правосторонней альтернативы $H_1^+ : \theta > \theta_0$.

Рассмотрим критерий уровня значимости α с критической областью $\mathrm{X}_{1\alpha}$ и функцией мощности $W(\theta) = W(\mathrm{X}_{1\alpha};\theta)$, $\theta \in \Theta$. Для этого критерия имеем

$$W(\theta) = W(\mathrm{X}_{1\alpha};\theta) = \int\limits_{\mathrm{X}_{1\alpha}} \cdots \int L(x_1,...x_n;\theta)dx_1 \cdots dx_n , \qquad \textbf{(3.1.54)}$$

где $L(x_1,...,x_n;\theta) = L(\vec{x};\theta) = \prod\limits_{i=1}^{n} f_X(x_i;\theta)$ – совместная плотность наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$. Предположим далее, что для любой критической области $\mathrm{X}_{1\alpha}$ размера α функция мощности $W(\theta)$, удовлетворяющая условию

$$W(\theta_0) = W(\mathrm{X}_{1\alpha};\theta_0) = \int\limits_{\mathrm{X}_{1\alpha}} \cdots \int L(x_1,...x_n;\theta_0)dx_1 \cdots dx_n = \alpha , \quad \textbf{(3.1.55)}$$

допускает разложение в ряд Тейлора в окрестности точки θ_0, то есть

$$W(\theta) = W(\theta_0) + W'(\theta_0)(\theta - \theta_0) + \frac{W''(\theta_0)}{2!}(\theta - \theta_0)^2 + ... +$$

$$+ \frac{W^{(n)}(\theta_0)}{n!}(\theta - \theta_0)^n + o((\theta - \theta_0)^n). \qquad \textbf{(3.1.56)}$$

Для простоты записи введем обозначение

$$L_k(\vec{x};\theta) = \partial^{(k)} L(\vec{x};\theta) / \partial\theta^k , \; k = 1, 2, ..., n, \qquad \textbf{(3.1.57)}$$

и знак многомерного интеграла в (3.1.55) заменим одномерным. Ограничиваясь двумя членами разложения в (3.1.56) с учетом (3.1.55), получаем приближенное значение для мощности $W(\theta) = W(\mathrm{X}_{1\alpha};\theta)$ в виде

$$W(\theta) = \alpha + W'(\theta_0)(\theta - \theta_0) + o(\theta - \theta_0)\,. \qquad \textbf{(3.1.58)}$$

Таким образом, согласно принципу 3.1.36, для построения л.н.м. одностороннего критерия проверки гипотез $H_0 : \theta = \theta_0$ и $H_1^+ : \theta > \theta_0$ необходимо максимизировать величину

$$W'(\theta_0) = \int\limits_{\mathrm{X}_{1\alpha}} L_1(\vec{x};\theta_0)d\vec{x}\ ,\ L_1(\vec{x};\theta_0) = \frac{\partial L(\vec{x};\theta)}{\partial\theta}\Big|_{\theta=\theta_0} \qquad \textbf{(3.1.59)}$$

при ограничении (3.1.55). Решение данной задачи, основанное на использовании леммы Неймана – Пирсона (см. разд. 3.3 и упр. 3.9.1), приводит к критической области л.н.м. одностороннего критерия проверки гипотез $H_0 : \theta = \theta_0$ и $H_1^+ : \theta > \theta_0$ в виде

$$\mathrm{X}_{1\alpha}^{+} = \{\vec{x} : L_1(\vec{x};\theta_0) / L(\vec{x};\theta_0) \geq t_{1-\alpha}\}\,, \qquad \textbf{(3.1.60)}$$

где критическое значение $t_{1-\alpha}$ статистики

$$\tilde{\Lambda}(\vec{X};\theta_0) = L_1(\vec{X};\theta_0) / L(\vec{X};\theta_0) \qquad \textbf{(3.1.61)}$$

является квантилем уровня $(1-\alpha)$ ф.р. $F_{\tilde{\Lambda}}(x \mid H_0)$ статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0 : \theta = \theta_0$, то есть $t_{1-\alpha} = F_{\tilde{\Lambda}}^{-1}\{(1-\alpha) \mid H_0\}$ и находится при заданном уровне значимости α из выражения (3.1.55), которое также может быть записано в виде

$$\int\limits_{t_{1-\alpha}}^{\infty} f_{\tilde{\Lambda}}(x;\theta_0)dx = \alpha\,. \qquad \textbf{(3.1.62)}$$

Здесь $f_{\tilde{\Lambda}}(x;\theta_0)$ обозначает плотность распределения вероятностей статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0 : \theta = \theta_0$. Рас-

пространение данного результата, то есть получение критической области вида (3.9.12) на случай проверки простой гипотезы $H_0 : \theta = \theta_0$ против сложной левосторонней альтернативы $H_1^- : \theta < \theta_0$ и сложной двусторонней альтернативы $H_1 : \theta \neq \theta_0$ приведено в упр. 3.9.1.

Замечание 3.1.63. Всякий T_n-критерий проверки статистических гипотез H_0, H_1, основанный на статистике $T_n(\vec{X})$, вычисляемой по выборке $X_1,...,X_n$ из распределения $F(x)$, характеризуется двумя вероятностями ошибок первого и второго рода. Вероятность ошибки первого рода используется для определения критической области тестовой статистики T_n путем задания уровня значимости α критерия. Вероятность ошибки второго рода приводит к понятию функции мощности критерия, которая отражает поведение вероятности правильного решения в пользу верной альтернативы и зависит от многих параметров, таких, как уровень значимости α, объем выборки n и параметр θ, который характеризует различие между проверяемыми гипотезами H_0, H_1. При сравнении двух состоятельных критериев с одинаковыми уровнями значимости α, с помощью поведения их функций мощности при конечных объемах выборок, возникают сложности в связи с большим числом параметров, характеризующим поведение функций мощности критериев. Кроме того, асимптотическое сравнение критериев также невозможно, так как для состоятельных критериев их функции мощности при фиксированных альтернативах вырождаются в единицу при неограниченном увеличении объема выборки. Для сравнения критериев предложены различные подходы (см., например, [183]). Наиболее часто для этой цели используют асимптотическую относительную эффективность Питмена (см. работы [24, 74, 183] и П.3.Е), которая вычисляется для последовательности контигуальных альтернатив, сходящихся к нулевой гипотезе при неограниченном увеличении объема выборки. На качественном уровне асимптотическая относительная эффективность двух критериев, обозначаемая через $ARE_F(T_n : S_n) = Э_F^2(T_n) / Э_F^2(S_n)$ (см. теорему П.3.101), характеризует отношение объемов выборок n_1 и n_2, при которых T_n-критерий и S_n-критерий проверки заданных гипотез H_0, H_1 с одинаковыми уровнями значимости обеспечивают одинаковую асимптотическую мощность против одной и той же последовательности альтернатив, сходящихся к нулевой гипотезе. Примеры сравнения критериев с помощью асимптотической относительной эффективности Питмена приводятся в разд. 5.2 – 5.4.

3.2. Примеры формулировок статистических гипотез

В данном разделе приведем примеры формулировок основной гипотезы H_0 и различных вариантов альтернативной гипотезы H_1.

3.2.А. Параметрические гипотезы

Рассмотрим сначала различные варианты задания гипотезы H_0 и альтернативы H_1 в рамках параметрической модели $\{\mathrm{X}, \Im_\theta\}$, где X – выборочное пространство и $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ – множество допустимых распределений вероятностей, которое представлено в виде *заданного* параметрического семейства распределений с точностью до *конечного* числа неизвестных параметров θ, принадлежащих *заданному* множеству параметров Θ. При рассмотрении случая двух *простых* параметрических гипотез модели $\{\mathrm{X}, \Im_\theta\}$ предполагается, что множество параметров Θ состоит только из двух точек θ_0 и θ_1, т.е. $\Theta = \{\theta_0, \theta_1\}$. Согласно этому предположению, допустимыми распределениями наблюдаемой с. в. X являются два распределения $F_X(x,\theta_0)$ и $F_X(x,\theta_1)$ с плотностями $f_X(x,\theta_0)$ и $f_X(x,\theta_1)$, $x \in R^1$. Требуется по выборке $X_1,...,X_n$, полученной в результате n повторных и независимых наблюдений над с. в. X, определить, какое из двух распределений истинно. Эта задача в терминах проверки *простых* параметрических гипотез записывается в следующем виде. Требуется проверить гипотезу $H_0 : \theta = \theta_0$ против одной из альтернатив:

$$H_1^+ : \theta = \theta_1, \text{ причем } \theta_1 > \theta_0 - \textit{правосторонняя альтернатива},$$

или

$$H_1^- : \theta = \theta_1 \text{ , причем } \theta_1 < \theta_0 - \textit{левосторонняя альтернатива}.$$

В тех случаях, когда обе гипотезы H_0 и H_1 являются *сложными*, задача формулируется следующим образом. Имеется выборка $X_1,...,X_n$ из распределения $F_X(x;\theta) \in \Im_\theta$, требуется построить кри-

терий проверки сложной гипотезы $H_0 : \theta \in \Theta_0$ против сложной альтернативы $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 \cup \Theta_1$, где Θ_0 и Θ_1 – множества параметров, определенные предположениями гипотез H_0 и H_1 .

Пример 3.2.1. Пусть $L(X) = N(\mu;\sigma^2)$. Эта краткая запись означает, что с.в. X имеет нормальное распределение вероятностей с неизвестными средним значением μ и дисперсией σ^2. Пусть далее имеется выборка $\vec{X} = (X_1,...,X_n)$ из этого распределения. В данном случае статистическая параметрическая модель $\{X, \Im_\theta\}$ характеризуется множеством распределений вида $\Im_\theta = \left\{F : F(x;\vec{\theta}), \vec{\theta} \in \Theta\right\}$, $\vec{\theta} = (\theta_1, \theta_2)$, причем $\theta_1 = \mu \in R^1$, $\theta_2 = \sigma^2 \in (0, +\infty)$ и ф.р.в. случайной величины X равна

$$F_X(x;\vec{\theta}) = \Phi\{(x-\mu)/\sigma\},$$

где $\Phi(x)$ обозначает функцию стандартного нормального распределения вида

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx, \; x \in R^1. \qquad \textbf{(3.2.2)}$$

Плотность распределения вероятности случайной величины X для этой модели имеет вид

$$f_X(x;\theta) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{1}{2\theta_2}(x-\theta_1)^2}, \; x \in R^1. \qquad \textbf{(3.2.3)}$$

Статистические задачи, которые возникают в рамках данной модели, сводятся к построению точечных и интервальных оценок неизвестных параметров $\theta_1 = \mu$ и $\theta_2 = \sigma^2$, а также к проверке гипотез относительно этих параметров на основе выборки $X_1,...,X_n$ из нормального распределения. Один из вариантов задач в рамках данной модели проверки простой гипотезы H_0 против простой и правосторонней альтернативы H_1^+ записывается в виде

$$H_0 : \mu = \mu_0, \; \sigma^2 \text{ – известный параметр},$$

$$H_1^+ : \mu = \mu_1, \; \mu_1 > \mu_0, \qquad \textbf{(3.2.4)}$$

или для двусторонней альтернативы эта задача запишется в такой форме

$$H_0 : \mu = \mu_0, \; \sigma^2 \text{ – известный параметр},$$

$$H_1 : \mu \neq \mu_0. \qquad \textbf{(3.2.5)}$$

Отметим, что гипотеза H_0 в (3.2.5) является параметрической и простой, а альтернатива H_1 – двусторонней и сложной. Если возможные отклонения от значения μ_0 параметра μ являются прогнозируемыми, то используются односторонние сложные альтернативы H_1^+, H_1^- и задача проверки гипотез кратко записывается в виде

$$H_0 : \mu = \mu_0,\ \sigma^2 \text{ – известный параметр,}$$

$$H_1^+ : \mu > \mu_0, \text{ (или } H_1^- : \mu < \mu_0). \qquad \textbf{(3.2.6)}$$

Отметим, что альтернативу H_1^+ часто называют правосторонней альтернативой, а H_1^- – левосторонней. Если параметр $\theta_2 = \sigma^2$ предполагается неизвестным, то гипотезы H_0 в (3.2.5) и (3.2.6) являются сложными.

Пример 3.2.7. Рассмотрим случай, когда X – дискретная случайная величина, которая имеет биномиальное распределение (см. П.1.8.) с параметрами n и p. Кратко это предположение будем записывать в виде $L(X) = Bi(n, p)$, где n – число опытов и p – вероятность некоторого события, назовем его «успехом». Для данного предположения статистическая модель запишется в виде $\Im_\theta = \{F : F(x;\theta),\ 0 < \theta < 1\}$, где ф.р.в. $F(x,\theta)$ определяются вероятностями

$$P_\theta(X = m) = C_n^m \theta^m (1-\theta)^{n-m},\ m = 0,1,...,n. \qquad \textbf{(3.2.8)}$$

Здесь $\theta = p$ – неизвестный параметр, характеризующий вероятность «успеха» и ф.р.в. $F(x,\theta)$ имеет вид

$$F(x,\theta) = \sum_{m=0}^{n} P_\theta(X = m) C(x - m), \qquad \textbf{(3.2.9)}$$

где $C(u)$ – единичная функция, определенная в (1.1.24). В рамках этой параметрической модели $\Im_\theta = \{F : F(x;\theta), 0 < \theta < 1\}$ задача проверки гипотезы относительно значения параметра $\theta = p$ против двусторонней альтернативы кратко записывается в виде

$$H_0 : p = p_0,$$

$$H_1 : p \neq p_0. \qquad \textbf{(3.2.10)}$$

Отметим, что в данном случае гипотеза H_0 простая, а двусторонняя альтернатива H_1 сложная.

Пример 3.2.11. Пусть с.в. X имеет пуассоновское распределение с неизвестным параметром $\theta \in (0,\infty)$. В данном случае параметрическая модель за-

дается семейством распределений $F(x,\theta)$, которые определяются вероятностями

$$P_\theta(X=m)=\frac{\theta^m}{m!}e^{-\theta},\ m=0,1,2,\dots\ , \qquad \textbf{(3.2.12)}$$

и записываются в виде

$$F(x,\theta)=\sum_{m=0}^{\infty}P_\theta(X=m)C(x-m)\,. \qquad \textbf{(3.2.13)}$$

Задача проверки простой гипотезы H_0 относительно значения параметра $\theta=M(X)$ против односторонней альтернативы H_1^+ имеет вид

$$H_0:\theta=\theta_0$$

$$H_1^+:\theta>\theta_0\,. \qquad \textbf{(3.2.14)}$$

Как и в задаче (3.2.10), в данном случае гипотеза H_0 простая, а правосторонняя альтернатива H_1^+ сложная.

3.2.Б. Непараметрические гипотезы

Рассмотрим непараметрические статистические модели $\{\mathrm{X},\Im_N\}$, где X – выборочное пространство и $\Im_N=\{F_X\}$ – множество допустимых распределений вероятностей, которое очень широкое и не может быть представлено параметрическим семейством известных функций с конечным числом неизвестных параметров. Этим и объясняется название этих моделей. Обычно множество $\Im_N=\{F_X\}$ задается из некоторых общих соображений типа равенства, непрерывности, симметрии, при этом функциональный вид ф.р.в. F_X *не конкретизируется*, то есть считается, что ф.р.в. F_X полностью *неизвестна*, исключением могут быть лишь такие общие предположения, как непрерывность. Отметим, что в статистической литературе выделяют типовые задачи проверки статистических гипотез (см., например, [24, с. 629; 22]). Их называют:

- задача согласия;
- задача однородности;
- задача независимости;
- задача случайности.

Приведем примеры формулировок этих задач.

Пример 3.2.15 (задача согласия). Пусть статистические данные представлены выборкой $X_1,...,X_n$ из распределения F_X, функциональный вид которого неизвестен. Требуется проверить гипотезу H_0, согласно которой функция распределения F_X равна некоторой заданной функция распределения F_0, то есть

$$H_0 : F_X = F_0 . \qquad \textbf{(3.2.16)}$$

Отметим, что эта гипотеза простая, так как F_0 полностью конкретизирована. Если гипотеза записывается в виде

$$H_0^* : F_X = F_0(x,\vec{\theta}), \qquad \textbf{(3.2.17)}$$

где ф.р. $F_0(x,\vec{\theta})$ задана с точностью до *неизвестного* (в общем случае векторного) параметра $\vec{\theta}$, то гипотезу H_0^* называют сложной. Критерии согласия для проверки гипотез (3.2.16) и (3.2.17) обсуждаются соответственно в разд. 3.5 и 3.6, см. также разд. 4.7.

Пример 3.2.18 (задача однородности). Пусть с.в. X имеет ф.р.в. F_X и с.в. Y имеет ф.р.в. F_Y. Статистические данные представлены выборкой $X_1,...,X_n$ из распределения F_X и выборкой $Y_1,...,Y_m$ из распределения F_Y. По этим статистическим данным требуется проверить гипотезу H_0, согласно которой ф.р.в. F_X и F_Y одинаковы, то есть

$$H_0 : F_X = F_Y . \qquad \textbf{(3.2.19)}$$

Отметим, что при формулировке этой непараметрической гипотезы функциональный вид распределений F_X и F_Y *не конкретизируется* и обычно предполагается лишь, что они непрерывны. Эта задача однородности сформулирована в двухвыборочном варианте. Отметим, что она естественным образом обобщается на заданное число k выборок и называется *k*-выборочной задачей однородности. В тех случаях, когда функциональный характер распределений F_X и F_Y известен с точностью до конечного числа неизвестных параметров, задача однородности формулируется в терминах равенства параметров этих распределений (см. разд. 3.7).

Пример 3.2.20 (задача независимости). Пусть $(X_1,Y_1),...,(X_n,Y_n)$ – выборка из двумерного распределения $F_{XY}(x,y)$. Требуется по этим данным проверить гипотезу H_0 о том, что случайные величины X и Y независимы, то есть

$$H_0 : F_{XY}(x,y) = F_X(x)F_Y(y) . \qquad \textbf{(3.2.21)}$$

Если функциональный вид ф.р.в. F_{XY} конкретизируется, например, предполагается, что ф.р.в. F_{XY} нормальная, то задача независимости решается в рамках параметрической модели и сводится к проверке гипотезы, согласно которой коэффициент корреляции равен нулю (см. разд. 3.7.Д). Если же функциональный вид ф.р.в. F_{XY} не конкретизируется, то задача независимости решается в рамках непараметрической модели (см. разд. 5.5 и 5.6).

Пример 3.2.22 (задача случайности). Рассмотрим случай, когда результат эксперимента характеризуется случайным вектором $\vec{X} = (X_1,...,X_n)$, который имеет функцию распределения $F_{\vec{X}}(x_1,...,x_n)$. При решении большинства статистических задач, связанных с изучением случайной величины X, обычно предполагается, что исходные статистические данные получены в серии повторных и независимых испытаний, упорядоченных каким-то образом (обычно по времени). В таких случаях статистические данные называют выборкой из некоторого распределения F_X, то есть предполагают, что компоненты вектора $\vec{X}$ получены независимо между собой из одной и той же генеральной совокупности, то есть являются *независимыми и одинаково распределенными случайными величинами* с функцией распределения F_X. Проверка сформулированных предположений на основе статистических данных называется задачей случайности, которая сводится в проверке гипотезы о том, что совместная функция распределения $F_{\vec{X}}(x_1,...,x_n)$ случайного вектора $\vec{X} = (X_1,...,X_n)$ факторизуется, то есть кратко сформулированная гипотеза записывается в виде

$$H_0 :\ F_{\vec{X}}(x_1,...,x_n) = \prod_{i=1}^{n} F_X(x_i), \qquad \textbf{(3.2.23)}$$

где F_X – некоторая одномерная функция распределения. Статистические критерии проверки сформулированных гипотез обсуждаются в разд. 3.7.Е и упр. 5.8.16.

3.3. Параметрические гипотезы. Подход Неймана – Пирсона

Обсудим сначала проверку *простых* параметрических гипотез, наиболее интересный для практики случай *сложных* гипотез обсуждается в разд. 3.4. Следуя [22], сначала рассмотрим проверку гипотез для непрерывных распределений, а потом рассмотрим дискретный вариант.

3.3.А. Непрерывный случай

Рассмотрим статистическую модель $\{\mathrm{X}, \Im_\theta\}$, для которой множество допустимых распределений определено в виде параметрического семейства *заданных* функций $F_X(x,\theta)$ с конечным числом *неизвестных* параметров θ, то есть

$$\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}, \qquad \textbf{(3.3.1)}$$

Θ – *заданное* множество параметров.

Рассмотрим для заданной модели $\{\mathrm{X}, \Im_\theta\}$ случай двух *простых* параметрических гипотез, предполагая, что множество параметров Θ состоит только из двух точек θ_0 и θ_1, т.е. $\Theta = \{\theta_0, \theta_1\}$.

Согласно принятому предположению, допустимыми распределениями наблюдаемой случайной величины X являются два распределения $F_X(x,\theta_0)$ и $F_X(x,\theta_1)$ с плотностями $f_X(x,\theta_0)$ и $f_X(x,\theta_1)$, $x \in R^1$. Требуется по выборке $X_1,...,X_n$, полученной в результате n повторных и независимых наблюдений над с. в. X, определить, какое из двух распределений истинно. Эта задача в терминах проверки *простых* параметрических гипотез записывается в следующем виде. Требуется проверить гипотезу

$$H_0 : \theta = \theta_0 \qquad \textbf{(3.3.2)}$$

против одной из односторонних альтернатив:

$$H_1^+ : \theta = \theta_1 \text{, причем } \theta_1 > \theta_0 - \text{правосторонняя альтернатива}, \qquad \textbf{(3.3.3)}$$

или

$$H_1^- : \theta = \theta_1 \text{, причем } \theta_1 < \theta_0 - \text{левосторонняя альтернатива}. \qquad \textbf{(3.3.4)}$$

Обсудим сначала решение данной задачи для случая проверки гипотезы H_0 против правосторонней альтернативы H_1^+ из (3.3.3).

Лемма Неймана – Пирсона. В рамках параметрической модели $\{\mathrm{X}, \Im_\theta\}$, $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$, для случая двух простых гипотез (3.3.2) и (3.3.3) существует наиболее мощный критерий, статистика которого $l(\vec{X})$ записывается при конкретной реализации

$\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ в виде отношения правдоподобия, то есть в виде

$$l(\vec{x}) = l(\vec{x};\theta_0,\theta_1) = \frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)} = \prod_{i=1}^{n} \frac{f_X(x_i;\theta_1)}{f_X(x_i;\theta_0)}. \qquad \textbf{(3.3.5)}$$

Критерий, основанный на статистике $l(\vec{X})$, определяется критической областью $\mathrm{X}^*_{1\alpha}$ вида

$$\mathrm{X}^*_{1\alpha} = \{\vec{x} : l(\vec{x}) \geq c\}, \qquad \textbf{(3.3.6)}$$

где граница c критической области определяется по заданной величине α из выражения

$$P\{l(\vec{X}) \geq c \mid H_0\} = \int\limits_{\vec{x}:\, l(\vec{x}) \,\geq\, c} \ldots \int L(x_1,...,x_n;\theta_0)dx_1 \cdots dx_n = \alpha. \qquad \textbf{(3.3.7)}$$

Доказательство. Везде ниже знаки многомерных интегралов вида

$$\int\limits_{\mathrm{X}} \ldots \int L(x_1,...,x_n;\theta)dx_1 \cdots dx_n$$

для простоты записи будем заменять знаками одномерного интеграла $\int\limits_{\mathrm{X}} L(\vec{x};\theta)d\vec{x}$. Следуя [22], рассмотрим другой критерий, статистика которого не определяется отношением правдоподобия (3.3.5), и пусть этот критерий основан на критической области $\mathrm{X}_{1\alpha}$ также размера α, но отличной от критической области $\mathrm{X}^*_{1\alpha}$, определенной в (3.3.6). Для доказательства леммы нужно убедиться в том, что при проверке H_0 против правосторонней альтернативы H_1^+ мощность критерия, определяемого критической областью $\mathrm{X}_{1\alpha}$, будет меньше мощности критерия, основанного на статистике в виде отношения правдоподобия (3.3.5), критическая область которого $\mathrm{X}^*_{1\alpha}$ определена в (3.3.6). Итак, согласно условию леммы, уровни значимости критериев, определяемых критическими областями $\mathrm{X}_{1\alpha}$ и $\mathrm{X}^*_{1\alpha}$, равны заданной величине α, то есть $P\{\vec{X} \in \mathrm{X}^*_{1\alpha} \mid H_0\} = P\{\vec{X} \in \mathrm{X}_{1\alpha} \mid H_0\} = \alpha$,

и, следовательно,

$$\int\limits_{X_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \alpha\,. \qquad \textbf{(3.3.8)}$$

Учитывая далее, что выполняется равенство

$$X_{1\alpha} = (X_{1\alpha} \cap X_{1\alpha}^*) \cup (X_{1\alpha} \cap \overline{X}_{1\alpha}^*) \qquad \textbf{(3.3.9)}$$

(рис. 3.3.1), из формулы (3.3.8) для критической области $X_{1\alpha}$ имеем

$$\alpha = \int\limits_{X_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X_{1\alpha}\cap X_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x} + \int\limits_{X_{1\alpha}\cap \overline{X}_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x}\,.$$

Отсюда получаем равенство

$$\int\limits_{X_{1\alpha}\cap \overline{X}_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x} = \alpha - \int\limits_{X_{1\alpha}\cap X_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x}\quad . \qquad \textbf{(3.3.10)}$$

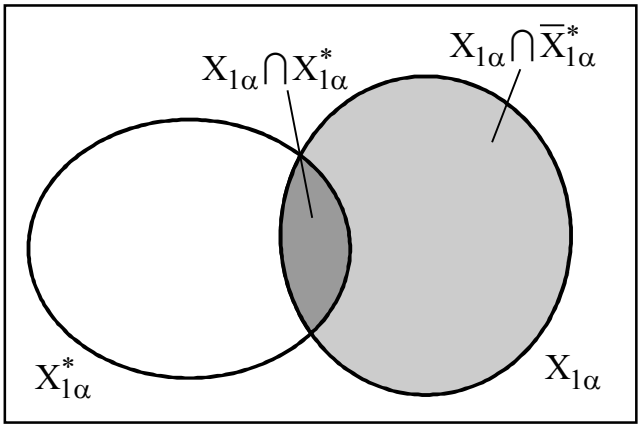

Рис. 3.3.1 Диаграмма Венна для $X_{1\alpha}$ и $X_{1\alpha}^*$

Аналогично для критической области $X_{1\alpha}^*$ имеем

$$\alpha = \int\limits_{X_{1\alpha}^*} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X_{1\alpha}^*\cap X_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} + \int\limits_{X_{1\alpha}^*\cap \overline{X}_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}$$

и отсюда получаем

$$\int\limits_{X_{1\alpha}^*\cap \overline{X}_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \alpha - \int\limits_{X_{1\alpha}^*\cap X_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}\,. \qquad \textbf{(3.3.11)}$$

Сравнивая (3.3.10) и (3.3.11), получаем следующее равенство:

$$\int\limits_{X_{1\alpha}\cap\overline{X}^{*}_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X^{*}_{1\alpha}\cap\overline{X}_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}. \qquad \textbf{(3.3.12)}$$

Рассмотрим теперь мощности критериев, определяемых критическими областями $X_{1\alpha}$ и $X^{*}_{1\alpha}$. Мощность критерия, определяемого критической областью $X_{1\alpha}$,

$$W(X_{1\alpha};\theta_1) = \int\limits_{X_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} =$$

$$= \int\limits_{X_{1\alpha}\cap X^{*}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} + \int\limits_{X_{1\alpha}\cap\overline{X}^{*}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}. \qquad \textbf{(3.3.13)}$$

Аналогично мощность критерия, определяемого критической областью $X^{*}_{1\alpha}$,

$$W(X^{*}_{1\alpha};\theta_1) = \int\limits_{X^{*}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} =$$

$$= \int\limits_{X^{*}_{1\alpha}\cap X_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} + \int\limits_{X^{*}_{1\alpha}\cap\overline{X}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}. \qquad \textbf{(3.3.14)}$$

Из правой части формулы (3.3.14) выразим первое слагаемое. В результате получим

$$\int\limits_{X^{*}_{1\alpha}\cap X_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} = W(X^{*}_{1\alpha};\theta_1) - \int\limits_{X^{*}_{1\alpha}\cap\overline{X}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}.$$

Подставив найденное выражение в (3.3.13), имеем

$$W(X_{1\alpha};\theta_1) = W(X^{*}_{1\alpha};\theta_1) +$$

$$+ \int\limits_{X_{1\alpha}\cap\overline{X}^{*}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x} - \int\limits_{X^{*}_{1\alpha}\cap\overline{X}_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}. \qquad \textbf{(3.3.15)}$$

Отметим, что вне множества $X^{*}_{1\alpha}$, то есть на множестве $\overline{X}^{*}_{1\alpha}$, согласно (3.3.6), выполняется неравенство $l(\vec{x}) < c$ и, следовательно, на множестве $X_{1\alpha}\cap\overline{X}^{*}_{1\alpha}$ интегрирования первого интеграла в (3.3.15)

выполняется неравенство $L(\vec{x};\theta_1) < cL(\vec{x};\theta_0)$. Аналогично на множестве $X^*_{1\alpha} \bigcap \overline{X}_{1\alpha}$ выполняется неравенство $L(\vec{x};\theta_1) \geq cL(\vec{x};\theta_0)$, или $-L(\vec{x};\theta_1) \leq -cL(\vec{x};\theta_0)$. С учетом этих замечаний, из формулы (3.3.15) получим следующее неравенство:

$$W(X_{1\alpha};\theta_1) < W(X^*_{1\alpha};\theta_1) + c\left\{\int\limits_{X_{1\alpha}\bigcap\overline{X}^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} - \int\limits_{X^*_{1\alpha}\bigcap\overline{X}_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}\right\}.$$

Отметим, что разность интегралов в фигурных скобках последнего неравенства равна нулю согласно (3.3.12), и это означает, что $W(X_{1\alpha};\theta_1) < W(X^*_{1\alpha};\theta_1)$. Доказательство завершено.

Пример 3.3.16. Вернемся снова к варианту нормального распределения, рассмотренному в примере (3.2.1), т.е. предполагаем, что $L(X) = N(\mu;\sigma^2)$, μ – неизвестный параметр, σ^2 – известный параметр. Требуется по выборке $\vec{X} = (X_1,...,X_n)$ из нормального распределения проверить гипотезы H_j, $j = 0,1,$ вида

$$H_0 : \mu = \mu_0,\ \sigma^2 \text{ – известный параметр},$$

$$H_1 : \mu = \mu_1,\ \mu_1 > \mu_0. \qquad \textbf{(3.3.17)}$$

Для построения наиболее мощного критерия проверки гипотез H_j, $j = 0,1$, воспользуемся леммой Неймана – Пирсона. В рамках данной параметрической модели функции правдоподобия при гипотезах H_j, $j = 0,1,$ определяются следующими выражениями:

$$L(\vec{x}|H_j) = \prod_{i=1}^{n} f_X(x_i|H_j) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_j)^2\right\},\ \ j=0,1. \qquad \textbf{(3.3.18)}$$

Таким образом, согласно (3.3.5), статистика критерия в виде отношения правдоподобия определяется выражением

$$l(\vec{x}) = \frac{L(\vec{x}\,|\,H_1)}{L(\vec{x}\,|\,H_0)} = \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right\}} =$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i-\mu_1)^2-(x_i-\mu_0)^2]\right\} = \exp\left\{-\frac{(\mu_1-\mu_0)n}{\sigma^2}\cdot n^{-1}\sum_{i=1}^{n}x_i - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2}\right\}.$$

Итак, отношение правдоподобия является *монотонной* функцией наблюденного значения статистики $\overline{X}$, поэтому (см. замечание 3.1.35), в качестве тестовой статистики критерия проверки гипотез (3.3.17) следует выбрать статистику вида

$$T(\vec{X}) = \overline{X} = n^{-1}\sum_{i=1}^{n} X_i \,. \qquad \textbf{(3.3.19)}$$

Учитывая, что $L(X_i \mid H_j) = N(\mu_j\,;\sigma^2)$, $j = 0,1$, $\forall i = 1,...,n$, заключаем, что случайная величина $T(\vec{X})$ при гипотезах H_j, $j = 0,1$, также имеет нормальное распределение, т.е. справедливы следующие выражения:

$$L\left\{T(\vec{X}) \mid H_j\right\} = N(\mu_j\,;\sigma_n^2),\ j = 0,1, \text{ где } \sigma_n^2 = \sigma^2/n,$$

и, следовательно, плотности распределения вероятностей статистики $T(\vec{X})$ при гипотезах H_j, $j = 0,1$, запишутся в виде

$$f_T(x \mid H_j) = (2\pi\sigma_n^2)^{-1/2} \exp\{-(x-\mu_j)^2/2\sigma_n^2\},\ j = 0,1\,.$$

Далее, согласно определению (3.1.21), вероятность ошибки первого рода α определяется выражением $\alpha = P\{T(\vec{X}) > t_{1-\alpha} \mid H_0\}$, или $\alpha = 1 - P\{T(\vec{X}) \le t_{1-\alpha} \mid H_0\} = 1 - F_T(t_{1-\alpha} \mid H_0)$. Учитывая, что для статистики $T(\vec{X})$ вида (3.3.19) ее функция распределения при справедливости гипотезы H_0 в точке $t_{1-\alpha}$ записывается в виде $F_T(t_{1-\alpha} \mid H_0) = \Phi\{(t_{1-\alpha} - \mu_0)/\sigma_n\}$, получаем $1 - \alpha = \Phi\{(t_{1-\alpha} - \mu_0)/\sigma_n\}$. Таким образом, критическое значение $t_{1-\alpha}$ статистики $T(\vec{X}) = \overline{X}$ вычисляется по формуле

$$t_{1-\alpha} = \lambda_{1-\alpha} \cdot \sigma_n + \mu_0\,, \qquad \textbf{(3.3.20)}$$

где $\sigma_n = \sigma/\sqrt{n}$ и $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$ – квантиль уровня $(1-\alpha)$ стандартного нормального распределения $\Phi(x)$. Отметим, что $\Phi(x)$ – симметричная относительно нуля функция распределения и, следовательно,

$$\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha) = -\lambda_\alpha\,. \qquad \textbf{(3.3.21)}$$

Таким образом, *наиболее мощный* критерий проверки гипотез (3.3.17), который среди всех критериев с уровнем значимости α обладает максимальной мощностью, основан на критической области

$$X_{1\alpha}^{+}\{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \ge \lambda_{1-\alpha}\} \qquad \textbf{(3.3.22)}$$

и записывается в виде следующего правила:

«*Если* $n^{-1}\sum_{i=1}^{n} x_i \geq \lambda_{1-\alpha}\cdot\sigma_n + \mu_0$, *то гипотеза* H_0 *из* (*3.3.17*) *отвергается на уровне значимости* α, *если выполняется обратное неравенство, то нет оснований отвергать* H_0». **(3.3.23)**

Далее, согласно определению (3.1.26), с учетом того, что ф.р. вероятностей статистики $T(\vec{X}) = \overline{X}$ при альтернативе H_1^+ из (3.3.17) имеет вид $F_T(t_{1-\alpha} \mid H_1) = \Phi\{(t_{1-\alpha} - \mu_1)/\sigma_n\}$, мощность W критерия (3.3.23) запишется в виде $W = 1-\beta = 1 - F_T(t_{1-\alpha} \mid H_1) = 1 - \Phi\{(t_{1-\alpha} - \mu_1)/\sigma_n\}$. Учитывая теперь, что выполняется равенство $\Phi(-x) = 1 - \Phi(x)$, $\forall x \in R^1$, и подставляя (3.3.20) в выражение для W, окончательно получаем

$$W = \Phi\{\sqrt{n}(\mu_1 - \mu_0)/\sigma - \lambda_{1-\alpha}\}. \qquad \textbf{(3.3.24)}$$

Отметим, что функция мощности $W(\delta)$, где $\delta = (\mu_1 - \mu_0)$ – параметр «различия» гипотез H_0 и H_1^+, монотонно возрастает по δ, приближаясь к единице. Кроме того, при $\delta = 0$ функция мощности равняется уровню значимости, то есть $W(0) = \Phi(-\lambda_{1-\alpha}) = \Phi(\Phi^{-1}(\alpha)) = \alpha$. Далее, при увеличении объема выборки n, мощность приближается к единице. Таким образом, $W(\delta) > \alpha$ *при* $\delta > 0$, и это означает, что критерий (3.3.23) *несмещенный*, кроме того, $W(\delta, n) \to 1$ при $n \to \infty$, и это означает, что критерий (3.3.23) *состоятельный*. Качественный график поведения функции мощности $W(\delta)$ в зависимости от параметра $\delta = (\mu_1 - \mu_0)$ при различных объемах выборки n приведен на рис. 3.3.1.

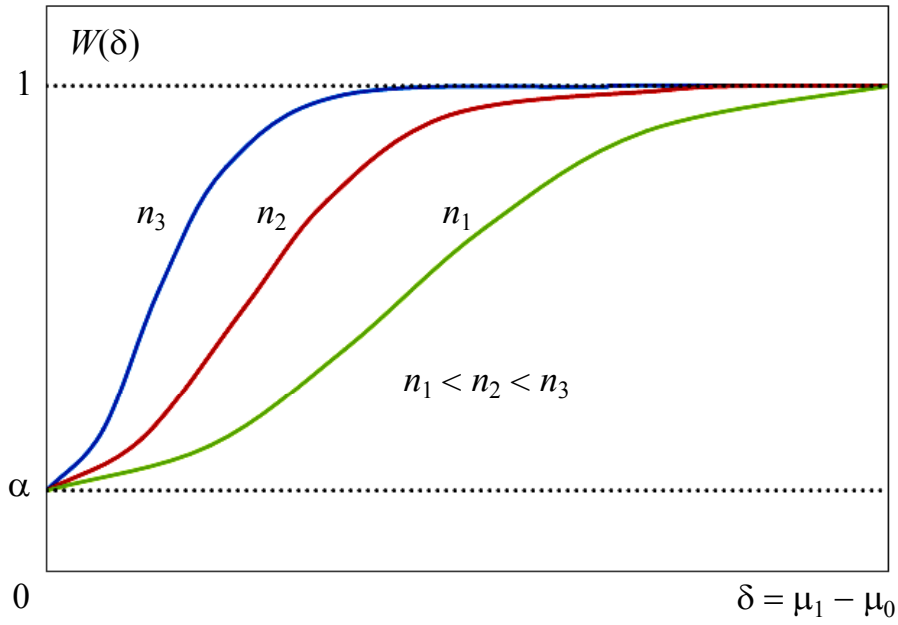

Рис. 3.3.1. Функции мощности $W(\delta)$ критерия (3.3.23)

Отметим также, что из выражения (3.3.24) можно получить формулу для вычисления необходимого объема выборки n, при котором критерий (3.3.23) будет обладать заданными вероятностями α_0 и β_0 ошибок первого и второго рода. Применяя к обеим частям (3.3.24) преобразование Φ^{-1}, получаем

$$n = \left[\frac{\{\Phi^{-1}(1-\beta_0) + \Phi^{-1}(1-\alpha_0)\}^2 \cdot \sigma^2}{(\mu_1 - \mu_0)^2} \right], \qquad \textbf{(3.3.25)}$$

где $[a]$ обозначает наибольшую целую часть числа a.

Например, при $\alpha_0 = \beta_0 = 0{,}001$ имеем $\Phi^{-1}(0{,}999) = 3{,}09$. Пусть далее $\sigma = 1$ и $\delta = (\mu_1 - \mu_0) = 1$. Для этих данных из (3.3.25) получаем $n = 39$. Для других исходных данных требуемый объем выборки может существенно измениться. Например, пусть $\sigma = 0{,}12$ и $\delta = (\mu_1 - \mu_0) = 0{,}02$, $\alpha_0 = 0{,}01$, $\beta_0 = 0{,}05$. Тогда $\Phi^{-1}(1-\alpha_0) = \Phi^{-1}(0{,}99) = 2{,}326$ и $\Phi^{-1}(1-\beta_0) = \Phi^{-1}(0{,}95) = 1{,}645$. Для этих данных из (3.3.25) получаем $n = 568$.

Замечание 3.3.26. Отметим, что вместо правила (3.3.23) решение относительно гипотезы H_0 из (3.3.17) можно принять и на основе вычисленного значения достигнутого уровня значимости α_∂, или *P-значения* критерия, основанного на статистике $T(\vec{X}) = \overline{X}$. Согласно (3.1.33), достигнутый уровень значимости при проверке гипотезы H_0 вычисляется по формуле

$$\alpha_\partial(\vec{x}) = P\{T(\vec{X}) > \overline{x} \mid H_0)\} = 1 - P\{T(\vec{X}) \le \overline{x} \mid H_0\} = 1 - F_T(\overline{x} \mid H_0) =$$
$$= 1 - \Phi\{\sqrt{n}(\overline{x} - \mu_0)/\sigma\}, \qquad \textbf{(3.3.27)}$$

где $\overline{x} = (x_1 + ... + x_n)/n$ – наблюденное значение статистики $T(\vec{X}) = \overline{X}$, вычисленное по наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $\vec{X} = (X_1, ..., X_n)$. Если вычисленное значение достигнутого уровня значимости α_∂ окажется «маленькой» величиной (например, $\alpha_\partial < 0{,}05$), то гипотезу H_0 следует отвергнуть.

Замечание 3.3.28. Отметим, что для проверки гипотезы $H_0 : \mu = \mu_0$ против левосторонней альтернативы $H_1^- : \mu = \mu_1$, $\mu_1 < \mu_0$, наиболее мощный критерий уровня значимости α основан на статистике $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ и определяется критической областью $X_{1\alpha}^- = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \le \lambda_\alpha\}$, где $\lambda_\alpha = \Phi^{-1}(\alpha)$ – квантиль уровня α стандартного нормального распределения. В данном случае *P-значение* критерия, основанного на *левосторонней* статистике $T(\vec{X}) = \overline{X}$, вычисляется по формуле

$$\textit{P-значение} = \alpha_\partial(\vec{x}) = P\{T(\vec{X}) \le \overline{x} \mid H_0)\} = F_T(\overline{x} \mid H_0) = \Phi\{\sqrt{n}(\overline{x} - \mu_0)/\sigma\}.$$

3.3.Б. Дискретный случай

Пусть X – дискретная случайная величина с возможными значениями $x_1, x_2, ...$ и их вероятностями $p(x_1;\theta), p(x_2;\theta), ...,\ \sum_i p(x_i;\theta) = 1$.

Предполагаем, что вероятности $p(x_1;\theta), p(x_2;\theta), ...$ возможных значений $x_1, x_2, ...$ зависят от некоторого неизвестного параметра θ, который принадлежит заданному параметрическому множеству Θ. Рассмотрим случай, когда это множество содержит только два элемента θ_0 и θ_1 т.е. предполагаем, что $\theta \in \Theta = \{\theta_0, \theta_1\}$. Имеется выборка $X_1, ..., X_n$ из распределения $F_X(x;\theta) = \sum_i p(x_i;\theta) C(x - x_i)$.

Требуется построить критерий с заданным уровнем значимости α для проверки простых гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$.

Для построения критерия проверки этих гипотез, согласно подходу Неймана – Пирсона, прежде всего, нужно вычислить для наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ выборки $X_1, ..., X_n$ отношение правдоподобия $l(\vec{x}) = L(\vec{x};\theta_1) / L(\vec{x};\theta_0)$ и конкретизировать критическую область $\mathrm{X}_{1\alpha} = \{\vec{x} : l(\vec{x}) \geq c\}$, т.е. нужно указать правило вычисления критического значения c статистики $l(\vec{X})$. Для построения критической области $\mathrm{X}_{1\alpha}$ критерия уровня значимости α все точки $\{\vec{x}\}$ выборочного пространства $\mathrm{X} = \{\vec{x}\}$ располагают в порядке возрастания величины отношения правдоподобия

$$l(\vec{x}) = L(\vec{x};\theta_1) / L(\vec{x};\theta_0) = \prod_{i=1}^{n} p(x_i;\theta_1) / \prod_{i=1}^{n} p(x_i;\theta_0) \qquad \textbf{(3.3.29)}$$

и в критическую область $\mathrm{X}_{1\alpha}$ включают максимальное число таких выборочных точек, для которых выполняется выражение

$$P_\theta(\vec{X} \in \mathrm{X}_{1\alpha} \mid \theta = \theta_0) = \sum_{\vec{x} \in \mathrm{X}_{1\alpha}} L(\vec{x};\theta_0) \leq \alpha\,. \qquad \textbf{(3.3.30)}$$

Однако при этом могут возникнуть трудности, в силу дискретности распределения выборки. Эти трудности связаны с тем, что не всегда можно получить точное значение уровня значимости α за счет

выбора границы c в неравенстве $l(\vec{x}) \geq c$, определяющем критическую область $X_{1\alpha}$. Эти трудности Э. Леман комментирует следующим образом [37, с.95]:

«Может случиться, что, включив очередную точку, мы еще не достигнем уровня α, а включив следующую, – превзойдем его. Точное значение α при этом или совсем не может быть получено, или же достигается присоединением к $X_{1\alpha}$ какой-либо из следующих по порядку точек. Эта трудность исчезает при переходе к рандомизированным критериям. С помощью рандомизации можно «расщепить» очередную точку, взяв в $X_{1\alpha}$ такую ее «часть», чтобы получить суммарную вероятность, в точности равную α, не нарушая принятого порядка точек».

Приведенные рассуждения формализуются при доказательстве фундаментальной леммы Неймана – Пирсона (см. [37]). Следуя [22], обсудим процедуру рандомизации, которая, несмотря на отмеченные трудности, позволяет построить рандомизированный критерий с заданным уровнем значимости.

Обозначим возможные значения статистики отношения правдоподобия $l(\vec{X})$ через $\{... < l_{k-1} < l_k < l_{k+1} < ...\}$. Далее, при заданном уровне значимости α рассмотрим два критерия с критическими областями

$$X^{**} = \{\vec{x} : l(\vec{x}) \geq l_k\} \text{ и } X^{*} = \{\vec{x} : l(\vec{x}) \geq l_{k+1}\} \qquad \textbf{(3.3.31)}$$

и соответственно с уровнями значимости

$$\alpha^{**} = \sum_{\vec{x} \in X^{**}} L(\vec{x};\theta_0) \text{ и } \alpha^{*} = \sum_{\vec{x} \in X^{*}} L(\vec{x};\theta_0), \qquad \textbf{(3.3.32)}$$

такими, для которых выполняются следующие неравенства:

$$\alpha^{*} < \alpha \leq \alpha^{**}. \qquad \textbf{(3.3.33)}$$

Если окажется, что в правой части (3.3.33) будет выполнено равенство $\alpha = \alpha^{**}$, то можно показать, как и в непрерывном случае (см., например, [22]), что наиболее мощный критерий уровня значимости α для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$ будет основан

на критической области $X^{**} = \{\vec{x} : l(\vec{x}) \geq l_k\}$, где l_k – критическое значение статистики $l(\vec{X})$, определяемое при заданном α из выражения $P_\theta(l(\vec{X}) \geq l_k \mid H_0) = \alpha$. Обсудим теперь случай, когда в (3.3.33) выполняются строгие неравенства. Прежде всего отметим, что критические множества X^{**} и X^{*} отличаются лишь одной точкой $\vec{x}$, в которой отношение правдоподобия $l(\vec{x})$ принимает значение l_k. Эту точку часто называют *выделенной точкой*. Вероятность того, что при справедливости гипотезы $H_0 : \theta = \theta_0$ статистика отношения правдоподобия $l(\vec{X})$ попадет в выделенную точку, равна

$$\alpha^{**} - \alpha^{*} = P_{\theta_0}\{l(\vec{X}) = l_k\}. \qquad \textbf{(3.3.34)}$$

Рандомизированная процедура, связанная с проведением дополнительного случайного эксперимента, осуществляется в случае, когда наблюденное отношение правдоподобия $l(\vec{x})$ попадает в выделенную точку, т.е. когда выполняется равенство $l(\vec{x}) = l_k$. Она предписывает проведение дополнительного случайного эксперимента с двумя исходами A и $\overline{A}$, вероятности которых соответственно равны γ и $1-\gamma$, где

$$\gamma = \frac{\alpha - \alpha^{*}}{\alpha^{**} - \alpha^{*}}. \qquad \textbf{(3.3.35)}$$

Согласно этой процедуре, в случае, когда выполняется равенство $l(\vec{x}) = l_k$, гипотезу $H_0 : \theta = \theta_0$ следует отвергнуть, если в результате случайного эксперимента реализуется событие A (с вероятностью γ), в противном случае, при реализации события $\overline{A}$ (с вероятностью $1-\gamma$), гипотеза H_0 принимается. Далее, если для наблюденного отношения правдоподобия $l(\vec{x})$ выполняется строгое неравенство $l(\vec{x}) > l_k$, то гипотезу H_0 следует отвергнуть, а при выполнении обратного неравенства $l(\vec{x}) < l_k$ ее следует принять. Обычно для удобства сформулированная процедура записывается в виде рандомизированного критерия, основанного на критической функции вида

$$\varphi(\vec{x}) = \begin{cases} 1 \text{ при } l(\vec{x}) > l_k \,, \\ \gamma \text{ при } l(\vec{x}) = l_k \,, \\ 0 \text{ при } l(\vec{x}) < l_k \,. \end{cases} \qquad \textbf{(3.3.36)}$$

Согласно (3.1.30), вероятность ошибки первого рода этого рандомизированного критерия с учетом (3.3.34) и (3.3.35) вычисляется по формуле

$$P(H_1 \mid H_0) = M_\theta[\varphi(\vec{X}) \mid \theta = \theta_0] =$$
$$= P_\theta\{l(\vec{X}) > l_k \mid \theta = \theta_0\} + \gamma \cdot P_\theta\{l(\vec{X}) = l_k \mid \theta = \theta_0\} =$$
$$= \alpha^* + \frac{\alpha - \alpha^*}{\alpha^{**} - \alpha^*}(\alpha^{**} - \alpha^*) = \alpha \,. \qquad \textbf{(3.3.37)}$$

Таким образом, рандомизированный критерий, основанный на критической функции (3.3.36), имеет уровень значимости, равный заданной величине α. Кроме того, мощность этого критерия вычисляется, с учетом (3.1.30), по формуле

$$W(\varphi\,;\theta_1) = P(H_1 \mid H_1) = M_\theta[\varphi(\vec{X}) \mid \theta = \theta_1] =$$
$$= P_\theta\{l(\vec{X}) > l_k \mid \theta = \theta_1\} + \gamma \cdot P_\theta\{l(\vec{X}) = l_k \mid \theta = \theta_1\} =$$
$$= \alpha_1^* + \frac{\alpha - \alpha^*}{\alpha^{**} - \alpha^*}(\alpha_1^{**} - \alpha_1^*) \,, \qquad \textbf{(3.3.38)}$$

где $$\alpha_1^{**} = \sum_{\vec{x} \in \mathrm{X}^{**}} L(\vec{x};\theta_1) \text{ и } \alpha_1^* = \sum_{\vec{x} \in \mathrm{X}^{*}} L(\vec{x};\theta_1) \qquad \textbf{(3.3.39)}$$

и $$\alpha_1^{**} - \alpha_1^* = P_{\theta_1}\{l(\vec{X}) = l_k\} \,. \qquad \textbf{(3.3.40)}$$

Отметим, что построенный рандомизированный критерий, основанный на критической функции (3.3.36), является н.м. критерием проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$ (см., например, [22] и упр. 3.9.27).

Пример 3.3.41. Приведем взятый из [9] пример одностороннего рандомизированного критерия, построенного по наблюдению, имеющему биномиальное распределение. Пусть случайная величина X имеет биномиальное распределение с параметрами (n,θ), т.е. $L(X) = Bi(n,\theta)$. Истинное значение

параметра θ неизвестно, но известно, что он принадлежит заданному параметрическому множеству Θ, т.е. $\theta \in \Theta = \{\theta_0, \theta_1\}$. Пусть в результате опыта с.в. X приняла значение x. Требуется на основе этих статистических данных построить критерий проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1$, для определенности полагаем, что $\theta_1 > \theta_0$.

Отметим, что в данном случае возможные значения $x = 0, 1, ..., (n-2), (n-1), n$ случайной величины X характеризуются вероятностями

$$p_\theta(x) = P_\theta\{X = x\} = C_n^x \theta^x (1-\theta)^{n-x}, \; x = 0, 1, ..., (n-2), (n-1), n \qquad \textbf{(3.3.42)}$$

и отношение правдоподобия записывается в виде

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = \frac{C_n^x \theta_1^x (1-\theta_1)^{n-x}}{C_n^x \theta_0^x (1-\theta_0)^{n-x}} = \left(\frac{\theta_1}{\theta_0}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-x}. \qquad \textbf{(3.3.43)}$$

Учитывая, что $\theta_1/\theta_0 > 1$ и $(1-\theta_1)/(1-\theta_0) < 1$, заключаем, что отношение правдоподобия $p_{\theta_1}(x)/p_{\theta_0}(x)$ становится больше при возрастании величины x. С учетом этого замечания, в критическую область $X_{1\alpha}$ критерия уровня значимости α для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$ следует включить такие возможные состояния $n, (n-1), (n-2), ...$, для которых вероятность того, что при справедливости гипотезы H_0 случайная величина X примет одно из этих значений, будет равняться заданной величине α. Однако, как мы уже отмечали выше, часто возникают трудности, связанные с дискретными распределениями, которые состоят в том, что не всегда возможно получить значение уровня значимости, равное заранее заданному числу α.

Приведем численный пример. Пусть $n = 10$, $\theta_0 = 1/4$ и $\alpha = 0{,}05$. Включим сначала в критическое множество следующие точки: $X^* = \{5, 6, 7, 8, 9, 10\}$. Тогда

$$\alpha = \sum_{x=5}^{10} C_{10}^x (1/4)^x (3/4)^{10-x} = 0{,}0781.$$

Если же выбрать критическое множество в виде следующих точек:

$$X^{**} = \{6, 7, 8, 9, 10\}, \text{ то } \alpha = \sum_{x=6}^{10} C_{10}^x (1/4)^x (3/4)^{10-x} = 0{,}0197.$$

Таким образом, критерий, построенный с помощью критического множества X^*, имеет уровень значимости $\alpha = 0{,}0781$ (что больше, чем $0{,}05$), а второй критерий с критическим множеством X^{**} имеет уровень значимости $\alpha = 0{,}0197$ (что значительно меньше, чем $0{,}05$). Отметим, что данная ситуа-

ция соответствует неравенствам (3.3.33), в которых выполняются строгие неравенства и, следовательно, для построения критерия с заданным уровнем значимости $\alpha = 0,05$ нужно использовать рандомизированную процедуру. Для построения рандомизированного критерия с заданным уровнем значимости (например, $\alpha = 0,05$) обозначим через α_k вероятность того, что при нулевой гипотезе наблюдение X попадет в критическое множество, состоящее из k упорядоченных точек. Для данного примера имеем $\alpha_6 = 0,0781$ (критическое множество X^*) и $\alpha_5 = 0,0197$ (критическое множество X^{**}). Эти два множества отличаются только одной точкой $x = 5$. Ниже такая точка называется *выделенной точкой*. Таким образом, в общем случае, $\alpha_k - \alpha_{k-1}$ есть вероятность того, что при нулевой гипотезе случайная величина X примет значение x, совпадающее с выделенной точкой. Для рассматриваемого примера $\alpha_6 - \alpha_5 = 0,0584$ есть вероятность того, что при нулевой гипотезе случайная величина X примет значение точки $x = 5$. Далее, определим число γ, вычисляемое по формуле

$$\gamma = \frac{\alpha - \alpha_{k-1}}{\alpha_k - \alpha_{k-1}} = \frac{0,05 - 0,0197}{0,0781 - 0,0197} = 0,519.$$

Рандомизированный критерий проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$ формулируется в виде следующего правила:

«Если наблюденное значение x случайной величины X попадает в критическую область, содержащую $(k-1)$ точек, то гипотеза H_0 отвергается, если наблюденное значение x попадает на выделенную точку (в данном примере $x = 5$), то гипотеза H_0 отвергается с вероятностью γ (в данном примере $\gamma = 0,519$). В других случаях нет оснований отвергать гипотезу H_0». **(3.3.44)**

Отметим, что для данного критерия вероятность отвергнуть гипотезу H_0 в ситуации, когда она верна, равна заданной величине α, так как имеем

$$P\{\text{отвергнуть } H_0 \mid H_0\} = \alpha_{k-1} + \gamma(\alpha_k - \alpha_{k-1}) = \alpha.$$

Для рассмотренного примера имеем

$$\alpha = 0,0197 + 0,519 \times 0,0584 = 0,050.$$

Замечание 3.3.45. Отметим, что процедура рандомизации критериев особенно полезна при сравнении их мощностей, так как при этом необходимо обеспечить равенство уровней значимости сравниваемых критериев. Для рассматриваемого примера мощность критерия с уровнем значимости $\alpha_6 = 0,0781$, основанного на критической области $\mathrm{X}^* = \{5, 6, ..., 10\}$, при альтернативе $\theta_1 = 1/2$, равна

$$W(\mathrm{X}^*; \theta_1 = 1/2) = P\{X \geq 5 \mid \theta_1 = 1/2\} = \sum_{x=5}^{10} C_{10}^x (1/2)^x (1/2)^{10-x} = 0,6231,$$

а для критерия с уровнем значимости $\alpha_5 = 0,0197$, основанного на критической области $\mathrm{X}^{**} = \{6, ..., 10\}$, при альтернативе $\theta_1 = 1/2$, мощность

$$W(\mathrm{X}^{**}; \theta_1 = 1/2) = P\{X \geq 6 \mid \theta_1 = 1/2\} = \sum_{x=6}^{10} C_{10}^x (1/2)^{10} = 0,3770.$$

Мощность рандомизированного критерия (3.3.44) с уровнем значимости $\alpha = 0,05$ при альтернативе $\theta_1 = 1/2$ для данного примера равна

$$P\{\textit{отвергнуть } H_0 \mid \theta_1 = 1/2\} =$$
$$= P\{X \geq 6 \mid \theta_1 = 1/2\} + P\{X = 5 \mid \theta_1 = 1/2\} \times P\{\textit{отвергнуть } H_0 \mid X = 5\} =$$
$$= 0,3770 + 0,2461 \times 0,519 = 0,5047.$$

В заключение отметим еще раз, что процедуру рандомизации обычно применяют при сравнении мощностей критериев для того, чтобы обеспечить равенство их уровней значимости. При использовании же критериев на практике обычно рандомизацию не применяют, а просто изменяют подходящим образом уровень значимости. Для рассмотренного примера используют либо $\alpha_6 = 0,0781$, либо $\alpha_5 = 0,0197$, или для принятия решения используют понятие достигнутого уровня значимости.

3.4. Параметрические сложные гипотезы

Рассмотрим параметрическую статистическую модель $\{\mathrm{X}, \Im_\theta\}$, для которой множество допустимых распределений вероятностей определено в виде параметрического семейства заданных функций $F_X(x, \theta)$ с конечным числом неизвестных параметров θ, то есть $\Im_\theta = \{ F_X : F_X(x, \theta), \theta \in \Theta \}$, где Θ – заданное множество параметров. В предыдущем разделе мы обсудили задачу построения наибо-

лее мощных критериев проверки *простых* гипотез H_0 и H_1 относительно параметра θ, решение которой основано на использовании подхода Неймана – Пирсона. Эта задача является важной в методологическом отношении и представляет теоретический интерес. На практике же наибольший интерес представляют задачи, в которых обе гипотезы H_0 и H_1, или одна из них, являются *сложными*. При проверке сложных гипотез мы заинтересованы в построении равномерно наиболее мощных (р.н.м.) критериев. Однако р.н.м. критерии существуют не всегда, их удается построить лишь в задачах проверки гипотез типа $H_0 : \theta \le \theta_0$ и $H_1 : \theta > \theta_0$ для некоторых статистических моделей, удовлетворяющих определенным ограничениям. Обсудим подробнее эти вопросы. Сначала рассмотрим случай проверки простой гипотезы против сложных альтернатив, а затем обсудим наиболее общий случай, когда обе гипотезы H_0 и H_1 являются сложными.

3.4.А. Статистические модели с монотонным отношением правдоподобия

Чтобы конкретизировать класс моделей, для которых существуют р.н.м. критерии при проверке гипотез $H_0 : \theta \le \theta_0$, $H_1 : \theta > \theta_0$ рассмотрим сначала, следуя [22], случай проверки простой гипотезы $H_0 : \theta = \theta_0$ против сложной альтернативы $H_1 : \theta \in \Theta_1$, $\Theta_1 = \Theta \setminus \{\theta_0\}$.

Выделим некоторое *фиксированное* значение $\theta_1 \in \Theta_1$ параметра θ при альтернативе и рассмотрим простые гипотезы $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1$. Для проверки этих простых гипотез, согласно фундаментальной лемме Неймана – Пирсона, существует основанный на статистике отношения правдоподобия наиболее мощный критерий заданного уровня значимости α с критической областью $X_{1\alpha}$, которая в общем случае зависит от θ_0 и θ_1 (см. разд. 3.3). Однако в некоторых случаях может оказаться, что критическая область $X_{1\alpha}$ *не будет зависеть* от θ_1 (см. пример 3.4.24), и это будет означать, что критерий с такой критической областью будет иметь максимальную мощность при *любой* допустимой альтернативе и, следо-

вательно, согласно определению (3.1.47), он будет р.н.м. критерием для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \in \Theta_1,\ \Theta_1 = \Theta \setminus \{\theta_0\}$. В тех случаях, когда критическая область $X_{1\alpha}$ *зависит* от θ_1, р.н.м. критерия для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \in \Theta_1,\ \Theta_1 = \Theta \setminus \{\theta_0\}$ не существует.

Приведем теперь общее достаточное условие существования р.н.м. критериев в случае односторонних альтернатив и вещественного параметра $\theta \in \Theta$. Пусть имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x,\theta),\ \theta \in \Theta$, требуется проверить гипотезу $H_0 : \theta = \theta_0$ против правосторонней альтернативы $H_1^+ : \theta > \theta_0$ либо против левосторонней альтернативы $H_1^- : \theta < \theta_1$. В данном случае параметрическое семейство $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ допустимых распределений, где множество параметров Θ представляется в виде $\Theta = \{\theta \ge \theta_0\}$ для альтернативы H_1^+ и $\Theta = \{\theta \le \theta_0\}$ для альтернативы H_1^-, имеет достаточную статистику $T(\vec{X})$ для параметра $\theta \in \Theta$ (см. разд. 2.5), если функция правдоподобия $L(x_1,...,x_n;\theta) = L(\vec{x}\,;\theta)$, согласно критерию факторизации (2.5.17), записывается для наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ в виде

$$L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x}). \qquad \textbf{(3.4.1)}$$

Далее, если снова выделить некоторое *фиксированное* значение $\theta_1 \in \Theta_1,\ \Theta_1 = \Theta \setminus \{\theta_0\}$ параметра θ при альтернативе и рассмотреть простые гипотезы $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$, то с учетом (3.4.1) статистика отношения правдоподобия $\lambda(\vec{X}) = L(\vec{X};\theta_1) / L(\vec{X};\theta_0)$ запишется в виде

$$\lambda(\vec{X}) = \varphi(T(\vec{X});\theta_1) / \varphi(T(\vec{X});\theta_0), \qquad \textbf{(3.4.2)}$$

то есть она является функцией от достаточной статистики $T(\vec{X})$. Если при этом окажется, что построенная статистика $\lambda(\vec{X})$ монотонно зависит от $T(\vec{X})$ (пусть для определенности $\lambda(\vec{X})$ монотонно возрастает от $T(\vec{X})$), то это означает, что критическая область

$\mathrm{X}_{1\alpha} = \{\vec{x} : \lambda(\vec{x}) \geq c\}$, определяющая критерий Неймана – Пирсона для проверки простых гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$, будет эквивалентна критической области, определяющейся неравенством $T(\vec{x}) \geq t_{1-\alpha}$, т.е.

$$\mathrm{X}_{1\alpha} = \{\vec{x} : \lambda(\vec{x}) \geq c\} \Leftrightarrow \mathrm{X}^*_{1\alpha} = \{\vec{x} : T(\vec{x}) \geq t_{1-\alpha}\}, \qquad \textbf{(3.4.3)}$$

где $t_{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения статистики $T(\vec{X})$ при гипотезе $H_0 : \theta = \theta_0$. Это распределение определяется видом статистики $T(\vec{X})$ и распределением $F_X(x;\theta_0)$ наблюдаемой с.в. X и не зависит от выбранного параметра θ_1. По этой причине критическая область $\mathrm{X}^*_{1\alpha} = \{\vec{x} : T(\vec{x}) \geq t_{1-\alpha}\}$ также не зависит от выбранного значения θ_1 параметра θ.

Таким образом, согласно (3.4.3), построенный критерий с критической областью $\mathrm{X}^*_{1\alpha} = \{\vec{x} : T(\vec{x}) \geq t_{1-\alpha}\}$ является н.м. критерием проверки простых гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1 > \theta_0$ и одновременно является р.н.м. критерием проверки *простой* гипотезы $H_0 : \theta = \theta_0$ против *сложной* правосторонней альтернативы H_1^+, поскольку его критическая область $\mathrm{X}^*_{1\alpha}$ *не зависит* от произвольно выбранного альтернативного значения θ_1 параметра θ. Аналогичные рассуждения могут быть проведены и для случая левосторонней альтернативы $H_1^- : \theta < \theta_0$, при этом следует лишь изменить неравенства в критических областях $\mathrm{X}_{1\alpha}$ и $\mathrm{X}^*_{1\alpha}$, определяющих соответственно критерий Неймана – Пирсона и критерий, основанный на достаточной статистике $T(\vec{X})$. Отметим, что приведенным рассуждениям можно придать и более общий характер. Для этого введем следующее определение (см., например, [37]).

Определение 3.4.4. Говорят, что семейство распределений $\mathfrak{I}_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ с плотностями $f_X(x,\theta)$ для параметрических моделей $\{\mathrm{X}, \mathfrak{I}_\theta\}$, где $\theta \in \Theta$ вещественный параметр, имеет *монотонное отношение правдоподобия*, если для любых θ и θ'

распределения $F_X(x,\theta)$ и $F_X(x,\theta')$ различны и существует функция $T(x)$ такая, что отношение $f_X(x,\theta')/f_X(x,\theta)$ является неубывающей функцией $T(x)$.

Отметим, что многие модели, используемые при решении статистических задач, обладают свойством монотонности отношения правдоподобия (см. упр. 3.9.24). С использованием определения 3.4.4, общий вывод относительно класса моделей, для которых существуют р.н.м. критерии при проверке простой гипотезы $H_0:\theta=\theta_0$ против односторонних и сложных альтернатив $H_1^+:\theta>\theta_0$, $H_1^-:\theta<\theta_1$ формулируется следующим образом:

Если семейство распределений $\Im_\theta=\{F_X:F_X(x,\theta),\theta\in\Theta\}$ *параметрической модели* $\{X,\Im_\theta\}$, *где* $\theta\in\Theta$ *– вещественный параметр, имеет монотонное отношение правдоподобия, то для него существует р.н.м. критерий для проверки простой гипотезы* $H_0:\theta=\theta_0$ *против односторонней и сложной альтернативы* H_1^+ (*либо против* H_1^-), *причем этот критерий совпадает с критерием Неймана – Пирсона для проверки простой гипотезы* $H_0:\theta=\theta_0$ *против произвольной фиксированной и простой альтернативы из* H_1^+ *либо из* H_1^-. **(3.4.5)**

Более общие результаты относительно свойств критериев для проверки сложных гипотез и альтернатив могут быть найдены в [7, 24, 37, 66].

Пример 3.4.6. Рассмотрим экспоненциальную модель, введенную в (2.5.33). Напомним, что параметрическая модель $\Im_\theta=\{F:F_X(x,\theta),\theta\in\Theta\}$ называется экспоненциальной, если плотность распределения вероятности $f_X(x;\theta)$ имеет вид $f_X(x;\theta)=\exp\{A(\theta)B(x)+C(\theta)+D(x)\}$. Отметим, что экспоненциальная модель имеет монотонное отношение правдоподобия, если $A(\theta)$ является неубывающей функцией, поскольку при этом условии отношение

$$f_X(x,\theta')/f_X(x,\theta)=\exp\{[A(\theta')-A(\theta)]\,B(x)+[C(\theta')-C(\theta)]\}$$

является неубывающей функцией $T(x)=B(x)$.

Далее, для экспоненциальной модели по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ функция правдоподобия $L(x_1,...,x_n;\theta) = L(\vec{x};\theta)$ запишется в виде

$$L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \exp\{A(\theta)\sum_{i=1}^{n} B(x_i) + nC(\theta)\} \cdot \exp\{\sum_{i=1}^{n} D(x_i)\}. \qquad \textbf{(3.4.7)}$$

Сравнивая это выражение с критерием факторизации (2.5.17) в виде $L(\vec{x};\theta) = \varphi(T(\vec{x}),\theta)h(\vec{x})$, заключаем, что $T(\vec{X}) = \sum_{i=1}^{n} B(X_i)$ является достаточной статистикой для параметра θ экспоненциальной модели. Отношение правдоподобия $\lambda(\vec{x}) = L(\vec{x};\theta_1)/L(\vec{x};\theta_0)$ в данном случае будет вычисляться по формуле

$$\lambda(\vec{x}) = \frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)} = \frac{\exp\{A(\theta_1)\sum B(x_i) + nC(\theta_1)\} \cdot \exp\{\sum D(x_i)\}}{\exp\{A(\theta_0)\sum B(x_i) + nC(\theta_0)\} \cdot \exp\{\sum D(x_i)\}} =$$

$$= \exp\{[A(\theta_1) - A(\theta_0)]\sum_{i=1}^{n} B(x_i) + n[C(\theta_1) - C(\theta_0)]\}.$$

Отсюда получаем, что статистика отношения правдоподобия $\lambda(\vec{X}) = L(X;\theta_1)/L(\vec{X};\theta_0)$ записывается в виде

$$\lambda(\vec{X}) = \exp\{[A(\theta_1) - A(\theta_0)]\,T(\vec{X}) + n[C(\theta_1) - C(\theta_0)]\} \qquad \textbf{(3.4.8)}$$

и, следовательно, если функция $A(\theta)$ строго монотонна, то отношение правдоподобия $\lambda(\vec{X})$ является монотонной функцией достаточной статистики $T(\vec{X})$. По этой причине для экспоненциальной модели всегда существует р.н.м. критерий для проверки простой гипотезы $H_0 : \theta = \theta_0$ против односторонних и сложных альтернатив $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$ соответственно с критическими областями $\mathrm{X}_{1\alpha}^+ = \{\vec{x} : \sum B(x_i) \geq t_{1-\alpha}\}$ и $\mathrm{X}_{1\alpha}^- = \{\vec{x} : \sum B(x_i) \leq t_\alpha\}$, при условии, что $A(\theta)$ – возрастающая функция. Если же функция $A(\theta)$ убывает, то неравенства в критических областях следует заменить на противоположные. В частности, для проверки гипотезы

$$H_0 : \mu = \mu_0,\ \sigma^2 \text{ – } \textit{известный параметр}, \qquad \textbf{(3.4.9)}$$

против односторонних альтернатив

$$H_1^+ : \mu > \mu_0,\ H_1^- : \mu < \mu_0 \qquad \textbf{(3.4.10)}$$

по выборке $X_1,...,X_n$ из нормального распределения $F_X(x;\mu) = \Phi\{(x-\mu)/\sigma\}$, где $\Phi(\cdot)$ – стандартное нормальное распределение, р.н.м. критерий уровня

значимости α для проверки гипотез H_0 и H_1^+ задается критической областью

$$\mathrm{X}_{1\alpha}^{+} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \geq \lambda_{1-\alpha}\},\ \lambda_{1-\alpha} = \Phi^{-1}(1-\alpha), \qquad \textbf{(3.4.11)}$$

а для проверки гипотез H_0 и H_1^- р.н.м. критерий уровня значимости α определяется критической областью

$$\mathrm{X}_{1\alpha}^{-} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \leq \lambda_{\alpha}\},\ \lambda_{\alpha} = \Phi^{-1}(\alpha). \qquad \textbf{(3.4.12)}$$

Здесь $\overline{x} = (x_1 + ... + x_n)/n$ – наблюденное значение статистики $\overline{X}$. Другие примеры построения р.н.м. критериев могут быть найдены в [22, с. 156–157].

3.4.Б. Критерий отношения правдоподобия

Рассмотрим теперь наиболее общий случай, когда обе гипотезы H_0 и H_1 являются сложными. Пусть имеется выборка $X_1,...,X_n$ из распределения $F_X(x;\theta) \in \Im_\theta$. Требуется построить критерий проверки сложной гипотезы

$$H_0 : \theta \in \Theta_0 \qquad \textbf{(3.4.13)}$$

против сложной альтернативы

$$H_1 : \theta \in \Theta_1,\ \Theta = \Theta_0 \cup \Theta_1. \qquad \textbf{(3.4.14)}$$

Для проверки этих гипотез используют *критерий отношения правдоподобия* (к.о.п.), статистика которого (см., например, [7, 24, 37, 66]) записывается в виде

$$\lambda(\vec{X}) = \frac{\sup\{L(\vec{X}\,;\theta) : \theta \in \Theta_1\}}{\sup\{L(\vec{X}\,;\theta) : \theta \in \Theta_0\}}. \qquad \textbf{(3.4.15)}$$

Эта статистика *правосторонняя*, то есть критерий отвергает гипотезу H_0 при «больших» значениях статистики $\lambda(\vec{X})$. Она является естественным обобщением статистики Неймана – Пирсона $l(\vec{X})$, определенной в (3.3.5), для случая простых гипотез.

Для обоснования целесообразности статистики $\lambda(\vec{X})$ напомним, что функция правдоподобия $L(x_1,...,x_n;\theta) = L(\vec{x}\,;\theta)$, определенная в (2.3.2), является совместной плотностью распределения вероятностей случайных величин $X_1,...,X_n$, которая при фиксированной на-

блюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ рассматривается как функция параметра θ, *поэтому ее можно рассматривать в качестве меры того, насколько хорошо параметр* θ *«объясняет» наблюденную реализацию* $\vec{x} = (x_1,...,x_n)$ *выборки* $X_1,...,X_n$. Таким образом, если верхняя грань $\sup\{L(\vec{x};\theta) : \theta \in \Theta_1\}$ велика по сравнению с верхней гранью $\sup\{L(\vec{x};\theta) : \theta \in \Theta_0\}$, то наблюденные данные $\vec{x} = (x_1,...,x_n)$ лучше всего объясняются каким-то значением параметра $\theta \in \Theta_1$, и наоборот.

При решении многих задач предполагается, что плотность $f_X(x;\theta)$ является непрерывной функцией параметра θ и размерность множества Θ_0 меньше размерности $\Theta = \Theta_0 \cup \Theta_1$, поэтому вместо статистики $\lambda(\vec{X})$ обычно используют эквивалентную статистику

$$\Lambda(\vec{X};\Theta_0) = \frac{\sup\{L(\vec{X};\theta) : \theta \in \Theta\}}{\sup\{L(\vec{X};\theta) : \theta \in \Theta_0\}}. \qquad \textbf{(3.4.16)}$$

Статистика $\Lambda(\vec{X};\Theta_0)$ *правосторонняя* и связана со статистикой $\lambda(\vec{X})$ выражением

$$\Lambda(\vec{X}) = \max\{1, \lambda(\vec{X})\},$$

так как

$$\sup\{L(\vec{X};\theta) : \theta \in \Theta\} = \max[\sup\{L(X;\theta) : \theta \in \Theta_0\}, \sup\{\vec{X};\theta\} : \theta \in \Theta_1] .$$

Замечание 3.4.17. Отметим, что в литературе часто (см. [9] и пример 3.4.24) вместо статистики $\Lambda(\vec{X};\Theta_0)$ используют обратную статистику

$$\Lambda^*(\vec{X};\Theta_0) = 1/\ \Lambda(\vec{X};\Theta_0). \qquad \textbf{(3.4.18)}$$

Критерии, основанные на статистиках $\Lambda^*(\vec{X};\Theta_0)$ и $\Lambda(\vec{X};\Theta_0)$, также называют критериями отношения правдоподобия.

Статистика $\Lambda^*(\vec{X};\Theta_0)$ является *левосторонней*, то есть нулевая гипотеза отвергается при «малых» значениях этой статистики и, следовательно, ее критическая область записывается в виде

$$X_{1\alpha} = \{\vec{x} : \Lambda^*(\vec{x};\Theta_0) \le t_\alpha\}. \qquad \textbf{(3.4.19)}$$

Критическое значение t_α статистики $\Lambda^*(\vec{X};\Theta_0)$ выбирается по заданному уровню значимости α из выражения

$$P(\vec{X} \in \mathrm{X}_{1\alpha} \mid H_0) = \int\limits_{\mathrm{X}_{1\alpha}} \cdots \int L(x_1,...,x_n;\theta)dx_1 \cdots dx_n \le \alpha \qquad \textbf{(3.4.20)}$$

или из выражения

$$P_\theta\{\Lambda^*(\vec{X};\Theta_0) \le t_\alpha\} = \int\limits_{-\infty}^{t_\alpha} f_{\Lambda^*}(x \mid H_0)dx = \alpha,\ \forall\theta \in \Theta_0, \qquad \textbf{(3.4.21)}$$

где $f_{\Lambda^*}(x \mid H_0)$ обозначает плотность распределения вероятности статистики $\Lambda^*(\vec{X};\Theta_0)$ при справедливости гипотезы $H_0 : \theta \in \Theta_0$.

Таким образом, при заданном параметрическом семействе распределений $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$ для проверки сложных гипотез $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \in \Theta_1$, $\Theta = \Theta_0 \cup \Theta_1$, нужно построить *критерий отношения правдоподобия* (к.о.п.), который основан на статистике $\Lambda^*(\vec{X};\Theta_0)$ или $\Lambda(\vec{X};\Theta_0)$. Для статистики $\Lambda^*(\vec{X};\Theta_0)$ критическая область $\mathrm{X}_{1\alpha} = \{\vec{x} : \Lambda^*(\vec{x};\Theta_0) \le t_\alpha\}$, определяется выражениями (3.4.19) – (3.4.21). Конечно, конкретный критерий будет зависеть от вида параметрического семейства $\Im_\theta = \{F : F_X(x,\theta), \theta \in \Theta\}$, однако для его построения всегда требуется выполнить следующие основные этапы:

1) по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ вычислить оценку максимального правдоподобия, МП-оценку параметра $\theta \in \Theta$;

2) вычислить оценку максимального правдоподобия, МП-оценку $\hat{\theta}_0$ параметра θ для случая, когда θ принимает значения только из множества Θ_0, то есть $\theta \in \Theta_0$;

3) составить отношение правдоподобия $\Lambda^*(\vec{x};\Theta_0) = L(\vec{x};\hat{\theta}_0)/L(\vec{x};\hat{\theta})$;

4) найти функцию ψ, строго возрастающую на множестве возможных значений статистики $\Lambda^*(\vec{X};\Theta_0) = L(\vec{X};\hat{\theta}_0)/L(\vec{X};\hat{\theta})$, такую, которая позволяла бы конкретизировать распределение статистики $\psi\{\Lambda^*(\vec{X};\Theta_0)\}$ при справедливости гипотезы $H_0 : \theta \in \Theta_0$ и определить квантиль уровня α (или уровня $(1-\alpha)$) этого распределения для определения критической области статистики $\psi\{\Lambda^*(\vec{X};\Theta_0)\}$.

Замечание 3.4.22. Отметим, что реализация четвертого этапа обычно вызывает трудности. Дело в том, что точное распределение статистики $\psi\{\Lambda^*(\vec{X};\Theta_0)\}$ удается получить лишь для некоторых моделей (например, для

нормальной, см. пример 3.4.41). Для большинства параметрических моделей точное распределение статистик $\psi\{\Lambda^*(\vec{X};\Theta_0)\}$ найти не удается. В таких случаях используют асимптотический вариант критерия отношения правдоподобия, который основан на использовании аппроксимации распределения статистики $-2\ln\{\Lambda^*(\vec{X};\Theta_0)\}$ распределением хи-квадрат. Эта аппроксимация имеет место при больших объемах выборки и при выполнении условий регулярности, обеспечивающих существование, единственность и асимптотическую нормальность оценок максимального правдоподобия (см. разд. 2.10 и разд. 3.4.В).

Пример 3.4.23. (нормальная модель с *известной* дисперсией). Рассмотрим пример построения *двустороннего* критерия проверки гипотез относительно среднего значения для нормально распределенных наблюдений при *известной* дисперсии. Предполагаем, что $L(X)=N(\mu;\sigma^2)$, μ – неизвестный параметр, $-\infty<\mu<+\infty$, σ^2 – известный параметр. Требуется по выборке $\vec{X}=(X_1,...,X_n)$ из нормального распределения проверить гипотезы H_j, $j=0,1,$ вида

$$H_0:\mu=\mu_0,\ \sigma^2 - \textit{известный параметр},$$

$$H_1:\mu\neq\mu_0. \qquad \textbf{(3.4.24)}$$

Таким образом, требуется проверить *простую* гипотезу H_0 против *сложной* альтернативы H_1 относительно неизвестного параметра $\theta=\mu\in\Theta$ параметрического семейства $F_X(x;\mu)=\Phi\{(x-\mu)/\sigma\}$, где $\Phi(\cdot)$ – стандартное нормальное распределение и множество параметров Θ в данном случае совпадает с действительной осью R^1. Для этого реализуем перечисленные выше четыре этапа.

Функция правдоподобия в данном случае по наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ запишется в виде

$$L(\vec{x};\mu)=\prod_{i=1}^{n}f_X(x_i;\mu)=(2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \qquad \textbf{(3.4.25)}$$

Так как параметр σ^2 предполагается известным, то наибольшее значение функции $L(\vec{x};\mu)$ достигается в точке $\overline{x}=(x_1+...+x_n)/n$, которая является наблюденным значением оценки максимального правдоподобия $\hat{\mu}=\overline{X}=(X_1+...+X_n)/n$ неизвестного параметра μ. Таким образом, при *фиксированной* наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ наибольшее значение $L(\vec{x};\overline{x})=\sup\{L(\vec{x};\mu):\mu\in R^1\}$ функции $L(\mu)=L(\vec{x};\mu)$ запишется в виде

$$L(\vec{x}\,;\overline{x}) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right\}.$$

Далее, так как гипотеза $H_0:\mu=\mu_0$, σ^2 – *известный параметр*, полностью определяет параметры нормального распределения, то

$$L(\vec{x}\,;\mu_0) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right\}.$$

С учетом приведенных выражений отношение правдоподобия $\Lambda^*(\vec{x}\,;\mu_0) = L(\vec{x}\,;\mu_0)/L(\vec{x}\,;\overline{x})$ запишется в виде

$$\Lambda^*(\vec{x}\,;\mu_0) = \frac{\exp\{-\sum(x_i-\mu_0)^2/2\sigma^2\}}{\exp\{-\sum(x_i-\overline{x})^2/2\sigma^2\}}.$$

Учитывая теперь, что $\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-\mu_0)=0$ и, следовательно,

$$\sum_{i=1}^{n}(x_i-\mu_0)^2 = \sum_{i=1}^{n}[(x_i-\overline{x})+(\overline{x}-\mu_0)]^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu_0)^2, \qquad \textbf{(3.4.26)}$$

окончательно получаем выражение для статистики отношения правдоподобия $\Lambda^*(\vec{X}\,;\mu_0) = L(\vec{X}\,;\mu_0)/L(\vec{X}\,;\hat{\mu})$ в виде следующей формулы:

$$\Lambda^*(\vec{X}\,;\mu_0) = \exp\{-(n/2\sigma^2)(\overline{X}-\mu_0)^2\}. \qquad \textbf{(3.4.27)}$$

Из полученного выражения (3.4.27) видно, что если $\overline{X}$ удаляется от μ_0 (влево или вправо), то статистика $(\overline{X}-\mu_0)^2$ увеличивается и, следовательно, статистка $\Lambda^*(\vec{X}\,;\mu_0)$ уменьшается. Вспоминая теперь, что эта статистика *левосторонняя*, то есть гипотеза H_0 отвергается, когда она принимает «маленькие» значения, заключаем, что гипотезу H_0 следует отвергнуть при «больших» значениях статистики $|\overline{X}-\mu_0|$, а это означает, что ее критическая область должна состоять из двух частей. Отметим далее, что $L(\overline{X}\,|\,H_0) = N(\mu_0\,;\sigma^2/n)$ и, следовательно, статистика $\sqrt{n}(\overline{X}-\mu_0)/\sigma$ имеет стандартное нормальное распределение, т.е. $L\{\sqrt{n}(\overline{X}-\mu_0)/\sigma\} = N(0\,,1)$. С учетом этого замечания и с учетом факта симметрии нормального распределения заключаем, что критерий отношения правдоподобия с уровнем значимости α для проверки гипотез $H_0:\mu=\mu_0$, σ^2 – *известный параметр* и $H_1:\mu\neq\mu_0$ основан на статистике

$$T(\vec{X}) = \sqrt{n}(\overline{X}-\mu_0)/\sigma, \qquad \textbf{(3.4.28)}$$

с критической областью $X_{1\alpha}$, состоящей из двух множеств:

$$\bar{x} \le \mu_0 + \lambda_{\alpha/2} \cdot \sigma/\sqrt{n} \quad \text{и} \quad \bar{x} \ge \mu_0 + \lambda_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \qquad \textbf{(3.4.29)}$$

где $\lambda_{\alpha/2}$ и $\lambda_{1-\alpha/2}$ – квантили стандартного нормального распределения соответственно уровней $\alpha/2$ и $1-\alpha/2$. Учитывая теперь, что $\lambda_{\alpha/2} = -\lambda_{1-\alpha/2}$, критическую область $X_{1\alpha}$, определенную в (3.4.29), можно записать в виде

$$X_{1\alpha} = \{\vec{x} : |T(\vec{x})| \ge \lambda_{1-\alpha/2}\} \text{ или } X_{1\alpha} = \{\vec{x} : \sqrt{n}\,|\bar{x} - \mu_0|/\sigma \ge \lambda_{1-\alpha/2}\}. \qquad \textbf{(3.4.30)}$$

Отметим, что *P-значение* двустороннего критерия проверки гипотез $H_0 : \mu = \mu_0$ и $H_1 : \mu \ne \mu_0$ с критической областью (3.4.30) вычисляется по формуле

$$\textit{P-значение} = \alpha_{\partial}(\vec{x}) = 2P\{T(\vec{X}) \ge \bar{x} \mid H_0)\} = 2\{1 - F_T(\bar{x} \mid H_0)\} =$$
$$= 2[\,1 - \Phi\{\sqrt{n}(\bar{x} - \mu_0)/\sigma\}\,].$$

Вычислим теперь мощность $W(X_{1\alpha};\mu_1)$ двустороннего критерия, основанного на статистике (3.4.28), с критической областью (3.4.30) при некотором альтернативном значении μ_1 параметра μ, которая равна вероятности попадания статистики критерия в критическую область при условии, что верна альтернатива $H_1 : \mu = \mu_1$, то есть

$$W(X_{1\alpha};\mu_1) = P\{\sqrt{n}(\bar{X} - \mu_0)/\sigma \le \lambda_{\alpha/2}\} + P\{\sqrt{n}(\bar{X} - \mu_0)/\sigma\} \ge \lambda_{1-\alpha/2}\}.$$

Отметим, что при альтернативе $H_1 : \mu = \mu_1$ случайная величина $\bar{X}$ имеет нормальное распределение, т.е. $L(\bar{X} \mid H_1) = N(\mu_1; \sigma^2/n)$, поэтому статистика $\sqrt{n}(\bar{X} - \mu_1)/\sigma$ имеет стандартное нормальное распределение, т.е. $L\{\sqrt{n}(\bar{X} - \mu_1)/\sigma \mid H_1) = N(0\,;1)$. Учитывая это замечание и обозначая параметр различия между гипотезами H_0 и H_1 через

$$\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma, \qquad \textbf{(3.4.31)}$$

получим выражение для мощности критерия в виде

$$W(X_{1\alpha};\mu_1) = P\{[\sqrt{n}(\bar{X} - \mu_1)/\sigma + \sqrt{n}(\mu_1 - \mu_0)/\sigma] \le \lambda_{1\alpha/2}\} +$$
$$+P\{[\sqrt{n}(\bar{X} - \mu_1)/\sigma + \sqrt{n}(\mu_1 - \mu_0)/\sigma] \ge \lambda_{1-\alpha/2}\} =$$
$$= P\{\sqrt{n}(\bar{X} - \mu_1)/\sigma \le \lambda_{\alpha/2} - \sqrt{n}(\mu_1 - \mu_0)/\sigma\} +$$
$$+1 - P\{[\sqrt{n}(\bar{X} - \mu_1)/\sigma \le \lambda_{1-\alpha/2} - \sqrt{n}(\mu_1 - \mu_0)/\sigma]\} = \Phi\{\lambda_{\alpha/2} - \delta\} + \Phi\{\lambda_{\alpha/2} + \delta\}.$$

Таким образом, мощность двустороннего критерия проверки гипотез $H_0:\mu=\mu_0$ и $H_1:\mu\neq\mu_0$ с критической областью (3.4.30) вычисляется по формуле

$$W(X_{1\alpha};\mu_1)=\Phi\{\lambda_{\alpha/2}-\delta\}+\Phi\{\lambda_{\alpha/2}+\delta\}, \qquad \textbf{(3.4.32)}$$

где $\lambda_{\alpha/2}=\Phi^{-1}(\alpha/2)$ и δ – параметр различия между гипотезами H_0 и H_1 – определен в (3.4.31). Например, если уровень значимости $\alpha=0,05$, то $\lambda_{\alpha/2}=\lambda_{0,025}=\Phi^{-1}(0,025)=-1,96$ и пусть $\delta=1$, тогда получим

$$W(X_{1\alpha};\mu_1)=\Phi\{-1,96-1\}+\Phi\{-1,96+1\}=\Phi\{-2,96\}+\Phi\{-0,96\}=$$
$$=0,0015+0,1685=0,1700.$$

Отметим, что функция мощности $W(X_{1\alpha};\mu)=W(\mu)$ двустороннего критерия, определенная выражением (3.4.32), является четной функцией от $(\mu-\mu_0)$ с минимумом в точке $\mu=\mu_0$. Кроме того, она всегда заключена между функциями $\Phi\{n^{1/2}(\mu-\mu_0)/\sigma-\lambda_\alpha\}$ и $\Phi\{n^{1/2}(\mu_0-\mu)/\sigma-\lambda_\alpha\}$, которые являются функциями мощности односторонних критериев, и в точке $\mu=\mu_0$ все три функции совпадают (см. рис. 3.4.1).

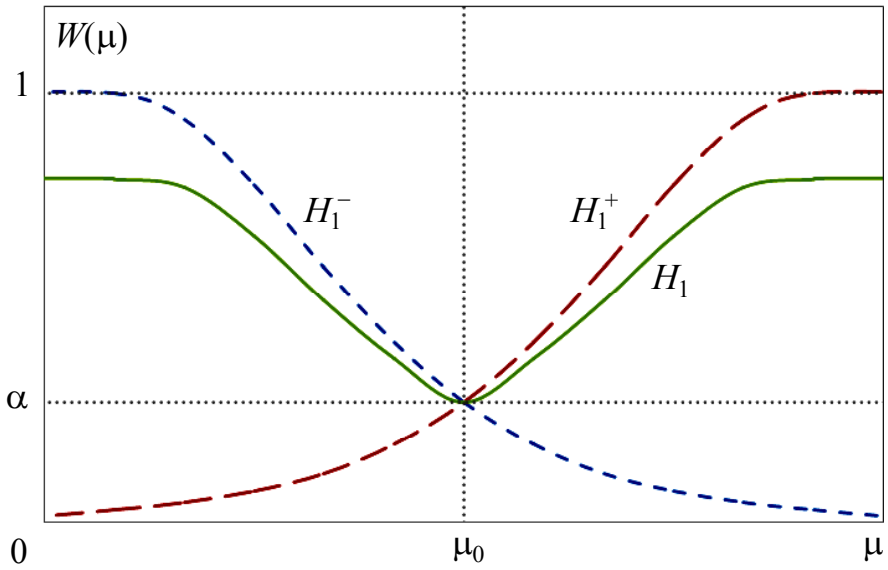

Рис. 3.4.1. Функции мощности односторонних и двустороннего критериев

Используя формулу (3.4.32), можно вычислить необходимый объем выборки n, при котором рассматриваемый критерий будет иметь заданную мощность $W_0(X_{1\alpha};\mu_1)=1-\beta_0$. Однако при решении этой задачи возникают аналитические трудности. Для их преодоления отметим, что при заданной ве-

личине δ одно из двух слагаемых в правой части (3.4.32) значительно меньше другого. Например, для $\delta = 1$ и $\alpha = 0{,}05$ вероятность $\Phi\{\lambda_{\alpha/2} + \delta\} = 0{,}1685$ более чем в 100 раз превосходит вероятность $\Phi\{\lambda_{\alpha/2} - \delta\} = 0{,}0015$. Поэтому, если $\delta > 0{,}5$, то членом $\Phi\{\lambda_{\alpha/2} - \delta\}$ в (3.4.32) можно пренебречь и в результате получим приближенную формулу

$$W(\mathrm{X}_{1\alpha};\mu_1) \approx \Phi\{\lambda_{\alpha/2} + \delta\} = \Phi\{\lambda_{\alpha/2} + \sqrt{n}(\mu_1 - \mu_0)/\sigma\}. \qquad \textbf{(3.4.33)}$$

Отсюда имеем

$$n \approx \sigma^2(\lambda_{1-\beta_0} + \lambda_{1-\alpha/2})^2/(\mu_1 - \mu_0)^2. \qquad \textbf{(3.4.34)}$$

Например, пусть $\sigma = 1{,}2$, $\mu_1 - \mu_0 = 0{,}2$, тогда при уровне значимости $\alpha = 0{,}01$, $\lambda_{1-\alpha/2} = \lambda_{0{,}995} = 2{,}58$, для обеспечения мощности $W_0(\mathrm{X}_{1\alpha};\mu_1) = 1 - \beta_0 = 0{,}95$, $\lambda_{1-\beta_0} = \lambda_{0{,}95} = 1{,}645$, согласно (3.4.34), потребуется объем выборки $n = 643$.

Замечание 3.4.35. Мы уже отмечали выше, что р.н.м. критерии существуют не всегда и они могут быть построены лишь для определенных статистических моделей при проверке гипотез типа: $H_0 : \theta \le \theta_0$ и $H_1 : \theta > \theta_0$. В частности, для достаточно общего класса экспоненциальных моделей, для которых плотность распределения вероятностей записывается в виде $f_X(x;\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\}$, при проверке простой гипотезы $H_0 : \theta = \theta_0$ против односторонних и сложных альтернатив $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$ р.н.м. критерии определяются соответственно критическими областями $\mathrm{X}_{1\alpha}^+ = \{\vec{x} : \sum B(x_i) \ge t_{1-\alpha}\}$ и $\mathrm{X}_{1\alpha}^- = \{\vec{x} : \sum B(x_i) \le t_\alpha\}$ при условии, что $A(\theta)$ строго монотонная функция. Если же при проверке простой гипотезы $H_0 : \theta = \theta_0$ рассматривать полный класс альтернатив $H_1 = H_1^+ \bigcup H_1^-$ (то есть класс двусторонних альтернатив $H_1 : \theta \ne \theta_0$), то р.н.м. критерии *не существуют*, так как каждый из односторонних критериев с критическими областями $\mathrm{X}_{1\alpha}^+$ и $\mathrm{X}_{1\alpha}^-$ будет иметь большую мощность по сравнению с любым двусторонним критерием с тем же уровнем значимости. При построении двусторонних критериев с нужным уровнем значимости α обычно поступают следующим образом:

1. Сначала находится статистика $T(\vec{X})$ для построения р.н.м. критериев с уровнями значимости α_2 и α_1 $(\alpha = \alpha_1 + \alpha_2)$ для проверки простой гипотезы $H_0 : \theta = \theta_0$ против односторонних альтернатив $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$ соответственно с критическими областями

$$\mathrm{X}_{1\alpha_2}^+ = \{\vec{x} : T(\vec{x}) \ge t_{1-\alpha_2}\} \text{ и } \mathrm{X}_{1\alpha_1}^- = \{\vec{x} : T(\vec{x}) \le t_{\alpha_1}\}. \qquad \textbf{(3.4.36)}$$

2. Затем, для построения двустороннего критерия с уровнем значимости α объединяют найденные критические области $X^{+}_{1\alpha_2}$ и $X^{-}_{1\alpha_1}$, в результате его критическая область размера $\alpha = \alpha_1 + \alpha_2$ записывается в виде

$$X_{1\alpha} = \{\vec{x} : (T(\vec{x}) \geq t_{1-\alpha_2}) \bigcup (T(\vec{x}) \leq t_{\alpha_1})\}. \qquad \textbf{(3.4.37)}$$

В тех случаях, когда распределение статистики $T(\vec{X})$ при гипотезе $H_0 : \theta = \theta_0$ симметрично относительно точки θ_0, обычно выбирают $\alpha_1 = \alpha_2 = \alpha/2$. При таком выборе уровней значимостей α_1 и α_2 двусторонние критерии часто называют «*симметричными*» (см., например, [22]) и для них выполняется условие $W'(X_{1\alpha};\theta_0) = 0$, обеспечивающее их *несмещенность*, т.е. выполнение неравенства $W(X_{1\alpha};\theta) \geq \alpha$ для всех $\theta \neq \theta_0$. Таким образом, двусторонние «симметричные» критерии с критической областью

$$X_{1\alpha} = \{\vec{x} : (T(\vec{x}) \leq t_{\alpha/2}) \bigcup (T(\vec{x}) \geq t_{1-\alpha/2})\} \qquad \textbf{(3.4.38)}$$

среди всех несмещенных критериев с уровнем значимости α обладают наибольшей мощностью, и поскольку критическая область (3.4.38) не зависит от альтернативных значений параметра θ, они являются несмещенными равномерно наиболее мощными (н.р.н.м.) критериями для проверки простой гипотезы $H_0 : \theta = \theta_0$ против двусторонней сложной альтернативы $H_1 : \theta \neq \theta_0$.

Для удобства дальнейших ссылок сформулированный вывод приведем в виде теоремы. Рассмотрим параметрическую модель $\{X, \Im_\theta\}$ с заданным параметрическим семейством распределений $\Im_\theta = \{F_X : F_X(x,\theta), \theta \in \Theta\}$. Предполагаем, что параметрическое множество Θ – это интервал действительной оси R^1 и θ_0 – внутренняя точка Θ. Предполагается также, что допустимые распределения $F_X(x;\theta)$ абсолютно непрерывны и функция правдоподобия $L(\vec{x};\theta)$ дифференцируема на Θ, причем для ее производных $\partial L(\vec{x};\theta)/\partial\theta = L_1(\vec{x};\theta)$ выполняются неравенства $|L_1(\vec{x};\theta)| < M(\vec{x})$, где $M(\vec{x})$ – интегрируемая функция на X. При выполнении этих условий операции дифференцирования и интегрирования можно менять местами, то есть выполняются равенства

$$\frac{\partial}{\partial\theta}\int_S L(\vec{x};\theta)d\vec{x} = \int_S [\partial L(\vec{x};\theta)/\partial\theta]d\vec{x} = \int_S L_1(\vec{x};\theta)d\vec{x}$$

для любого подмножества $S \subset X$ и $\forall\theta \in \Theta$.

Теорема 3.4.39. Для проверки простой гипотезы $H_0 : \theta = \theta_0$ против двусторонней сложной альтернативы $H_1 : \theta \neq \theta_0$ параметрической модели $\{X, \Im_\theta\}$, удовлетворяющей сформулированным выше

условиям, несмещенный равномерно наиболее мощный (н.р.н.м.) критерий уровня значимости α определяется критической областью

$$X_{1\alpha} = \{\vec{x} : L(\vec{x}\,;\theta_1) \geq cL(\vec{x}\,;\theta_0) + c_1 \partial L(x\,;\theta)/\partial\theta\,|_{\theta=\theta_0}\},$$

если она не зависит от $\theta_1 \neq \theta_0$ и постоянные $c \geq 0$ и c_1 однозначно определяются из условий

$$\int\limits_{X_{1\alpha}} L(\vec{x}\,;\theta_0)d\vec{x} = \alpha \quad \text{и} \quad \int\limits_{X_{1\alpha}} [\partial L(\vec{x}\,;\theta)/\partial\theta\,|_{\theta=\theta_0}]d\vec{x} = 0.$$

Доказательство этой теоремы проводится по аналогии с доказательством леммы Неймана – Пирсона (см. разд. 3.3) и может быть найдено в [37; 22, с. 159].

Пример 3.4.40. Проиллюстрируем применение приведенного выше правила построения двусторонних критериев на примере проверки гипотезы $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu \neq \mu_0$ для нормальной модели $N(\mu\,;\sigma^2)$ в предположении, что параметр σ^2 известен (см. пример 3.4.24). Согласно (3.4.11), для проверки гипотез $H_0 : \mu = \mu_0$ и $H_1^+ : \mu > \mu_0$ р.н.м. критерий уровня значимости α_2 задается критической областью

$$X_{1\alpha_2}^+ = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \geq \lambda_{1-\alpha_2}\},\ \lambda_{1-\alpha_2} = \Phi^{-1}(1-\alpha_2).$$

Далее, согласно (3.4.12), для проверки гипотез $H_0 : \mu = \mu_0$ и $H_1^- : \mu < \mu_0$ р.н.м. критерий уровня значимости α_1 задается критической областью

$$X_{1\alpha_1}^- = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \leq \lambda_{\alpha_1}\},\ \lambda_{\alpha_1} = \Phi^{-1}(\alpha_1).$$

Следовательно, для проверки гипотез $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ двусторонний критерий уровня значимости $\alpha = \alpha_1 + \alpha_2$ будет основан на критической области

$$X_{1\alpha} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \leq \lambda_{\alpha_1} \textit{ либо } \sqrt{n}(\overline{x} - \mu_0)/\sigma \geq \lambda_{1-\alpha_2}\}.$$

В данном случае выбор «симметричного» критерия, то есть выбор $\alpha_1 = \alpha_2 = \alpha/2$, приводит к критической области

$$X_{1\alpha} = \{\vec{x} : \sqrt{n}\,|\,\overline{x} - \mu_0\,|/\sigma \geq \lambda_{1-\alpha/2}\},\ \lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2),$$

которая совпадает с (3.4.30) и определяет н.р.н.м. критерий проверки гипотез $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$.

Пример 3.4.41 (нормальная модель с *неизвестной* дисперсией). Рассмотрим теперь пример построения *двустороннего* критерия проверки гипотез относительно среднего значения для нормально распределенных наблюдений при *неизвестной* дисперсии. Предполагаем, что $L(X) = N(\mu;\sigma^2)$, μ и σ – неизвестные параметры, $-\infty < \mu < +\infty$, $\sigma > 0$. Требуется по выборке $\vec{X} = (X_1,...,X_n)$ из нормального распределения проверить гипотезы H_j, $j = 0,1$, вида

$$H_0 : \mu = \mu_0,$$

$$H_1 : \mu \neq \mu_0. \qquad \textbf{(3.4.42)}$$

Таким образом, требуется проверить *простую* гипотезу H_0 против *сложной* и двусторонней альтернативы H_1 при неизвестном векторном параметре $\vec{\theta} = (\theta_1, \theta_2)$ нормального распределения, где $\theta_1 = \mu$ и $\theta_2 = \sigma^2$. В данном случае множество параметров $\Theta = \{(\theta_1, \theta_2) : -\infty < \theta_1 < +\infty, \theta_2 > 0\}$ есть множество точек верхней полуплоскости на плоскости (μ, σ) с координатными осями μ и σ, а множество $\Theta_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0, \theta_2 > 0\}$ – это полупрямая, совпадающая с прямой $\mu = \mu_0$ в первом квадранте координатной плоскости (μ, σ).

Для построения критерия отношения правдоподобия проверки сложных гипотез (3.4.42) реализуем описанные выше четыре этапа.

Этап первый. Как и в предыдущем примере, функция правдоподобия по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ запишется в виде

$$L(\vec{x};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \qquad \textbf{(3.4.43)}$$

Наибольшее значение $\sup\{L(\vec{x};\vec{\theta}) : \vec{\theta} \in \Theta\}$ на множестве Θ для функции $L(\vec{x};\mu,\sigma^2)$ достигается в точке $(\overline{x}, s^2)$, координаты которой являются наблюденными значениями оценок максимального правдоподобия $\hat{\mu} = \overline{X} = (X_1 + ... + X_n)/n$ и $\hat{\sigma}^2 = S^2 = n^{-1}\sum(X_i - \overline{X})^2$ параметров μ и σ^2, и записывается в виде

$$L(\vec{x};\overline{x},s^2) = \left[\frac{n}{2\pi\sum(x_i-\overline{x})^2}\right]^{n/2} \exp\{-n/2\}. \qquad \textbf{(3.4.44)}$$

Этап второй. Наибольшее значение $\sup\{L(\vec{x};\vec{\theta}) : \vec{\theta} \in \Theta_0\}$ на множестве Θ_0 для функции $L(\vec{x};\mu,\sigma^2)$ достигается в точке (μ_0, s_0^2), где s_0^2 – наблюденное значение оценки максимального правдоподобия $\hat{\sigma}_0^2 = S_0^2 = n^{-1}\sum(X_i - \mu_0)^2$

параметра σ^2 при гипотезе H_0, и записывается в виде

$$L(\vec{x};\mu_0,s_0^2)=\left[\frac{n}{2\pi\sum(x_i-\mu_0)^2}\right]^{n/2}\exp\{-n/2\}. \qquad \textbf{(3.4.45)}$$

Этап третий. С учетом приведенных выражений статистика отношения правдоподобия $\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)=L(\vec{X};\mu_0,\hat{\sigma}_0^2)/\ L(\vec{X};\hat{\mu},\hat{\sigma}^2)$ имеет вид

$$\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)=\left[\frac{\sum(X_i-\bar{X})^2}{\sum(X_i-\mu_0)^2}\right]^{-n/2}. \qquad \textbf{(3.4.46)}$$

Этап четвертый. Так как любая монотонная функция от $\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)$ может быть использована для проверки гипотез (3.4.42), то рассмотрим корень *n*-й степени от квадрата статистики $\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)$. В результате с учетом (3.4.26) получаем

$$\{\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)\}^{2/n}=\frac{\sum(X_i-\bar{X})^2}{\sum(X_i-\mu_0)^2}=\frac{\sum(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2+n(\bar{X}-\mu_0)^2}=$$
$$=\left[1+\frac{n(\bar{X}-\mu_0)^2}{\sum(X_i-\bar{X})^2}\right]^{-1}=\left[1+\frac{T^2(\vec{X})}{n-1}\right]^{-1}, \qquad \textbf{(3.4.47)}$$

где статистика $T(\vec{X})$ определяется в виде

$$T(\vec{X})=\sqrt{n-1}(\bar{X}-\mu_0)/S=\sqrt{n}(\bar{X}-\mu_0)/\tilde{S}. \qquad \textbf{(3.4.48)}$$

Здесь $\tilde{S}^2=\sum(X_i-\bar{X})^2/(n-1)$ – подправленная выборочная дисперсия, являющаяся несмещенной оценкой σ^2 при неизвестном среднем значении. Напомним, что статистика $\Lambda^*(\vec{X};\mu_0,\hat{\sigma}^2)$ *левосторонняя*, то есть гипотеза H_0 отвергается, когда она принимает «маленькие» значения, поэтому, согласно (3.4.38), гипотезу H_0 следует отвергнуть при «больших» значениях статистики $|T(\vec{X})|$. Напомним также, что статистика $T(\vec{X})$ при справедливости гипотезы H_0 из (3.4.42), согласно теореме 2.8.31, имеет распределение Стьюдента с числом степеней свободы $(n-1)$, то есть выполняется выражение

$$L\{\sqrt{n-1}(\bar{X}-\mu_0)/S\,|\,H_0\}=St(n-1). \qquad \textbf{(3.4.49)}$$

С учетом этого замечания, критерий отношения правдоподобия для проверки гипотез $H_0:\mu=\mu_0$, $H_1:\mu\neq\mu_0$ при условии, *что параметр* σ^2 *неизвестен*, приводит к эквивалентному критерию уровня значимости α, основанного на

статистике (3.4.48) с критической областью $X_{1\alpha}$, определяемой в виде

$$X_{1\alpha} = \{\vec{x} : \sqrt{n-1}\,|\,\overline{x} - \mu_0\,|\,/\,s \geq t_{1-\alpha/2,(n-1)}\}, \qquad \textbf{(3.4.50)}$$

где $t_{1-\alpha/2,(n-1)}$ – квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $(n-1)$.

3.4.B. Асимптотические свойства критерия отношения правдоподобия

Для реализации критерия отношения правдоподобия требуется знание точного распределения вероятностей статистики отношения правдоподобия при справедливости нулевой гипотезы. Мы уже отмечали (см. замечание 3.4.22), что свойства оптимальности к.о.п. при конечных объемах выборки, и в частности точное распределение вероятностей статистики отношения правдоподобия при справедливости нулевой гипотезы, удается найти лишь для некоторых моделей (например, для нормальной модели). Для большинства параметрических моделей эта задача вызывает трудности, которые обычно преодолеваются путем использования асимптотических результатов. Асимптотические методы играют важную роль при решении задач проверки гипотез, так же как и при решении задач оценивания параметров, и успешно применяются на практике в случаях, когда объем выборки достаточно большой. Асимптотическая теория к.о.п. достаточно хорошо развита и подробно изложена в монографиях [7, 37, 24, 66]. Мы приведем лишь отдельные результаты, которые понадобятся в дальнейшем.

Асимптотические свойства критериев проверки гипотез изучают в предположении, что объем выборки n неограниченно возрастает, при этом рассматривают свойства не отдельного критерия, а последовательности критериев, статистики которых зависят от неограниченно возрастающего объема выборки n. Для того чтобы подчеркнуть этот факт, при изучении асимптотических свойств статистикам критериев обычно приписывают индекс n.

Таким образом, при изучении асимптотических свойств критериев отношения правдоподобия статистики $l(\vec{X};\theta_0,\theta_1)$, $\Lambda(\vec{X};\Theta_0)$ и $\Lambda^*(\vec{X};\Theta_0)$, определенные формулами (3.3.5), (3.4.16) и (3.4.18), бу-

дем соответственно обозначать $l_n(\vec{X};\theta_0,\theta_1)$, $\Lambda_n(\vec{X};\Theta_0)$ и $\Lambda_n^*(\vec{X};\Theta_0)$. Для дальнейшего нам понадобится следующее определение (см. [66]).

Определение 3.4.51. Пусть $\Im$ – класс абсолютно непрерывных распределений. Если $\{F_1(x),F_2(x),G(x)\}\in\Im$, то среднее количество информации, или интеграл информации Кульбака для различения между распределениями $F_1(x)$ и $F_2(x)$ по наблюдению над распределением $G(x)$, определяется как

$$I(1\,;2)=\int_{-\infty}^{+\infty}\ln\frac{dF_1(x)}{dF_2(x)}dG(x). \qquad \textbf{(3.4.52)}$$

Для изучения асимптотических свойств критериев отношения правдоподобия рассмотрим параметрическую модель $\{\mathrm{X},\Im_\theta\}$, для которой множество допустимых распределений определено в виде параметрического семейства *заданных* функций $F_X(x,\theta)$ с конечным числом *неизвестных* параметров θ, то есть

$$\Im_\theta=\{\,F_X : F_X(x,\theta),\,\theta\in\Theta\,\},$$

где Θ – *заданное* множество параметров.

Простые параметрические гипотезы. Рассмотрим сначала случай простых гипотез, то есть для данной модели $\{\mathrm{X},\Im_\theta\}$ предполагаем, что множество параметров Θ состоит только из двух точек θ_0 и θ_1, т.е. $\Theta=\{\theta_0,\theta_1\}$. Требуется по выборке $X_1,...,X_n$ из распределения $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, $\theta\in\Theta=\{\theta_0,\theta_1\}$, $x\in R^1$, проверить простую гипотезу $H_0:\theta=\theta_0$ против простой альтернативы $H_1:\theta=\theta_1$.

Для решения сформулированной задачи с помощью критерия отношения правдоподобия приведем асимптотическое распределение статистики отношения правдоподобия $l_n(\vec{X};\theta_0,\theta_1)$, возможные значения которой $l_n(\vec{x})$ определены формулой (3.3.5) и для наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выборки $X_1,...,X_n$ записываются в виде

$$l_n(\vec{x}) = l_n(\vec{x};\theta_0,\theta_1) = \frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)} = \prod_{i=1}^{n} \frac{f_X(x_i;\theta_1)}{f_X(x_i;\theta_0)},\ n = 1, 2, \ldots . \quad \textbf{(3.4.53)}$$

Рассмотрим логарифм статистики отношения правдоподобия $l_n(\vec{X};\theta_0,\theta_1)$:

$$\ln l_n(\vec{X};\theta_0,\theta_1) = \ln \frac{L(\vec{X};\theta_1)}{L(\vec{X};\theta_0)} = \ln \prod_{i=1}^{n} \frac{f_X(X_i;\theta_1)}{f_X(X_i;\theta_0)} =$$

$$= \sum_{i=1}^{n} L^*(X_i),\ n = 1, 2, \ldots, \quad \textbf{(3.4.54)}$$

где $$L^*(x) = \ln\{f_X(x;\theta_1)/f_X(x;\theta_0)\}. \quad \textbf{(3.4.55)}$$

Убедимся, что при выполнении некоторых условий распределение последовательности $\ln l_n(\vec{X})$, $n = 1, 2, \ldots$, из (3.4.54) при справедливости гипотезы $H_0 : \theta = \theta_0$ сходится к нормальному распределению. Для этого используем понятие *количества информации по Кульбаку – Лейблеру*, которое для параметрической модели $\{\mathrm{X}, \Im_\theta\}$ характеризует различие между ф.р. $F_X(x;\theta)$ и $F_X(x;\theta')$ по наблюдению над $F_X(x;\theta)$ и определяется в виде

$$I(\theta,\theta') = M_\theta[\ln\{f_X(X_1;\theta)/f_X(X_1;\theta')\}] =$$

$$= \int_{-\infty}^{+\infty} \left[\ln \frac{f_X(x;\theta)}{f_X(x;\theta')}\right] f_X(x;\theta)dx. \quad \textbf{(3.4.56)}$$

Отметим, что при проверке простых гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ из формулы (3.4.54) с учетом (3.4.56) следуют выражения для математических ожиданий в виде

$$M_{\theta_1}[L^*(X_1)] = I(\theta_1,\theta_0) \quad \textbf{(3.4.57)}$$

и $$M_{\theta_0}[L^*(X_1)] = M_{\theta_0}[-L^*(X_1)] = -I(\theta_0,\theta_1). \quad \textbf{(3.4.58)}$$

Далее предположим, что дисперсия с.в. $L^*(X_1)$ конечна и положительна, то есть

$$0 < \sigma_0^2 = D_{\theta_0}[L^*(X_1)] < \infty. \quad \textbf{(3.4.59)}$$

Тогда, при выполнении условия (3.4.59), согласно центральной предельной теореме (см. П.3.37), при $n \to \infty$ имеем

$$P_{\theta_0}\{\ln l_n(\vec{X};\theta_0,\theta_1) \geq -nI(\theta_0,\theta_1) + z\cdot\sigma_0\sqrt{n}\} =$$

$$= P\left\{\frac{1}{\sigma_0\sqrt{n}}\sum_{i=1}^{n}[L^*(X_i) - M_{\theta_0}(L^*(X_i))] \geq z\right\} \to 1-\Phi(z). \quad \textbf{(3.4.60)}$$

Таким образом, логарифм статистики отношения правдоподобия имеет при гипотезе $H_0 : \theta = \theta_0$ асимптотически нормальное распределение, то есть при $n \to \infty$ справедливо выражение

$$L\{[\ln l_n(\vec{X};\theta_0,\theta_1) + nI(\theta_0,\theta_1)]/\sigma_0\sqrt{n}\} \mid H_0\} = N(0,1). \quad \textbf{(3.4.61)}$$

Отсюда в частности следует, что при больших объемах выборки критическое значение статистики $l_n(\vec{X};\theta_0,\theta_1)$ при проверке гипотез $H_0 : \theta = \theta_0$ и $H_1^+ : \theta > \theta_0$ записывается в виде

$$\exp\{-nI(\theta_0,\theta_1) + \sigma_0\sqrt{n}\cdot\lambda_{1-\alpha}\},\ \lambda_{1-\alpha} = \Phi^{-1}(1-\alpha). \quad \textbf{(3.4.62)}$$

Подводя итог, отметим, что асимптотическая критическая область размера α критерия отношения правдоподобия, основанного на статистике $l_n(\vec{X};\theta_0,\theta_1)$, для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1^+ : \theta > \theta_0$ записывается в виде

$$X_{1\alpha}^+ = \{\vec{x} : l_n(\vec{X};\theta_0,\theta_1) \geq \exp(-nI(\theta_0,\theta_1) + \sigma_0\sqrt{n}\cdot\lambda_{1-\alpha})\},$$
$$\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha), \quad \textbf{(3.4.63)}$$

а для проверки гипотез $H_0 : \theta = \theta_0$, $H_1^- : \theta < \theta_0$ она определяется выражением

$$X_{1\alpha}^- = \{\vec{x} : l_n(\vec{X};\theta_0,\theta_1) \leq \exp(-nI(\theta_0,\theta_1) + \sigma_0\sqrt{n}\cdot\lambda_{\alpha})\},$$
$$\lambda_{\alpha} = \Phi^{-1}(\alpha). \quad \textbf{(3.4.64)}$$

Для проверки гипотез $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ асимптотическая критическая область размера α выглядит следующим образом:

$$X_{1\alpha} = \{\vec{x} : |\, l_n(\vec{X};\theta_0,\theta_1)\,| \geq \exp(-nI(\theta_0,\theta_1) + \sigma_0\sqrt{n}\cdot\lambda_{1-\alpha/2})\}. \quad \textbf{(3.4.65)}$$

Замечание 3.4.66. Асимптотическая нормальность статистик отношения правдоподобия при нулевой гипотезе сохраняется и в дискретном случае, что позволяет преодолеть трудности, связанные с нахождением точного распределения статистик отношения правдоподобия (см. упр. 3.9.44).

Отметим, что если X – дискретная случайная величина с возможными значениями $x_1,..,.x_k$ и вероятностями $p(x_j;\theta)$, $j=1,...,k$, тогда количество информации Кульбака – Лейблера вычисляется по формуле

$$I(\theta_0,\theta_1)=\sum_{j=1}^{k} p(x_j;\theta_0)\ln\{p(x_j;\theta_0)/\,p(x_j;\theta_1)\} \qquad \textbf{(3.4.67)}$$

и

$$\sigma_0^2=\sum_{j=1}^{k} p(x_j;\theta_0)[\ln\{p(x_j;\theta_0)/\,p(x_j;\theta_1)\}]^2-[I(\theta_0,\theta_1)]^2. \qquad \textbf{(3.4.68)}$$

Подстановка выражений (3.4.67) и (3.4.68) в (3.4.62) позволяет вычислить критическое значение статистики отношения правдоподобия (3.3.29) в дискретном случае для проверки простых гипотез.

Простая гипотеза и сложная альтернатива. Обсудим теперь двусторонние критерии для случая простой гипотезы и сложной альтернативы. Прежде чем формулировать общие результаты об асимптотическом распределении статистик отношения правдоподобия, вернемся к примерам 3.4.24 и 3.4.41, в которых были описаны двусторонние критерии проверки гипотез относительно среднего значения нормальной модели для случаев известной и неизвестной дисперсии.

В примере 3.4.24 рассматривался случай проверки гипотез $H_0:\mu=\mu_0$, $H_1:\mu\neq\mu_0$ по выборке $X_1,...,X_n$ из нормального распределения $F_X(x;\mu)=\Phi\{(x-\mu)/\sigma\}$ при известной дисперсии σ^2. Используя (3.4.27), с учетом (3.4.18) запишем выражение для статистики отношения правдоподобия $\Lambda(\vec{X};\Theta_0)$, определенной в (3.4.16), в виде

$$\Lambda(\vec{X};\mu_0)=\exp\{(n/2\sigma^2)(\overline{X}-\mu_0)^2\}.$$

Отсюда следует, что

$$2\ln\Lambda(\vec{X};\mu_0)=[\sqrt{n}(\overline{X}-\mu_0)/\sigma]^2.$$

Далее учитывая, что $L\{\sqrt{n}(\overline{X}-\mu_0)/\sigma\,|\,H_0\}=N(0,1)$, заключаем,

что статистика $2\ln\Lambda(\vec{X};\mu_0)$ при гипотезе H_0 имеет распределение хи-квадрат с одной степенью свободы, то есть

$$L\{2\ln\Lambda(\vec{X};\mu_0)\,|\,H_0\}=\chi^2(1). \qquad \textbf{(3.4.69)}$$

Рассмотрим теперь случай неизвестной дисперсии σ^2. Из (3.4.47) получаем

$$2\ln\Lambda(\vec{X};\mu_0,\hat{\sigma}^2)=n\ln\left[1+\frac{T_n^2(\vec{X})}{n-1}\right], \qquad \textbf{(3.4.70)}$$

где статистика $T_n(\vec{X})$ определена в (3.4.48) и записывается в виде

$$T_n(\vec{X})=\sqrt{n-1}(\overline{X}-\mu_0)/S=\sqrt{n}(\overline{X}-\mu_0)/\tilde{S}.$$

Далее, согласно теореме (2.8.31), статистика $T_n(\vec{X})$ при справедливости гипотезы H_0 имеет распределение Стьюдента с числом степеней свободы $(n-1)$, то есть $L\{\sqrt{n-1}(\overline{X}-\mu_0)/S\,|\,H_0\}=St(n-1)$, и, следовательно, при $n\to\infty$, выполняется выражение $L\{T_n(\vec{X})\,|\,H_0\}=N(0,1)$. Учитывая теперь, что $(1/x)\ln(1+x)\to 1$ при $x\to 0$, и применяя теорему Слуцкого (см. П.3.9), из (3.4.70) окончательно получаем, что статистика $2\ln\Lambda(\vec{X};\mu_0,\hat{\sigma}^2)$ при справедливости гипотезы H_0 аппроксимируется статистикой $T_n^2(\vec{X})$, которая стремится по закону к распределению $\chi^2(1)$. Таким образом, при $n\to\infty$ справедливо выражение

$$L\{2\ln\Lambda(\vec{X};\mu_0,\hat{\sigma}^2)\,|\,H_0\}=\chi^2(1), \qquad \textbf{(3.4.71)}$$

то есть пришли к такому же результату, как и в случае, когда дисперсия σ^2 известна. Отметим, что результаты этих примеров являются точными прототипами того, что происходит в общем случае в задачах проверки гипотез с помощью двусторонних критериев при больших выборках.

Рассмотрим теперь случай проверки простой гипотезы и сложной альтернативы для векторного параметра $\vec{\theta}=(\theta_1,...,\theta_r)$. Предполагаем, что выполнены условия регулярности, обеспечивающие

существование, единственность и асимптотическую нормальность оценок максимального правдоподобия $\hat{\theta}_n = (\hat{\theta}_{1n},...,\hat{\theta}_{rn})$ параметра $\vec{\theta} = (\theta_1,...,\theta_r)$, см. Уилкс [66, с. 354, с. 418] и разд. 2.10. Приведем строгий результат в виде теоремы.

Теорема 3.4.72. Пусть требуется проверить гипотезу $H_0 : \vec{\theta} = \vec{\theta}_0$, где $\vec{\theta}_0 = (\theta_{10},...,\theta_{r0})$ – фиксированная внутренняя точка множества Θ. Тогда для больших выборок $(n \to \infty)$ при выполнении указанных условий регулярности асимптотическая критическая область критерия отношения правдоподобия записывается в виде

$$X_{1\alpha} = \{\vec{x} : -2\ln\Lambda^*(\vec{x}\,;\vec{\theta}_0) \geq \chi^2_{1-\alpha,\,r}\},$$

то есть при $n \to \infty$ выполняется выражение

$$P_\theta(\vec{X} \in X_{1\alpha} \mid H_0) = P_\theta(-2\ln\{\Lambda^*(\vec{X}\,;\vec{\theta}_0)\} \geq \chi_{1-\alpha,\,r} \mid H_0) \to \alpha.$$

Доказательство этой теоремы основано на использовании того факта, что при гипотезе $H_0 : \vec{\theta} = \vec{\theta}_0$ и при $n \to \infty$ выполняется следующее выражение: $L(-2\ln\{\Lambda^*(\vec{X}\,;\vec{\theta}_0)\} \mid H_0) = \chi^2(r)$. Этот факт является частным случаем более общего результата Уилкса [66, с. 418], полученного для случая сложной гипотезы и сложной альтернативы векторного параметра. Согласно этому результату, при нулевой гипотезе выполняется выражение

$$L(-2\ln\{\Lambda^*(\vec{X}\,;\Theta_0)\}) = \chi^2(s-q),$$

где s – число параметров при альтернативе H_1, а q – число параметров при гипотезе H_0. Доказательство теоремы может быть найдено в [22, с. 168] (см. также [24, с. 301] и [66, с. 426]).

3.4.Г. О связи задач проверки гипотез и интервальных оценок параметров

В рамках параметрических статистических моделей $\{X, \Im_\theta\}$, для которых множество допустимых распределений вероятностей $\Im_\theta$ определено в виде параметрического семейства заданных функций

$F_X(x,\theta)$ с конечным числом *неизвестных* параметров θ, то есть в виде $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$, где Θ – заданное множество параметров, все статистические задачи сводятся к построению точечных или интервальных оценок параметров θ, а также к построению критериев проверки гипотез относительно этих параметров. Все эти статистические задачи тесно связаны между собой. В частности, построение критериев отношения правдоподобия при проверке сложных гипотез требует нахождения оценок максимального правдоподобия неизвестных параметров. Связь между задачами проверки гипотез и построением доверительных интервалов обсудим на примере проверки простой гипотезы $H_0 : \theta = \theta_0$ против двусторонней альтернативы $H_1 : \theta \neq \theta_0$. Эта связь проявляется в том, что *гипотеза* $H_0 : \theta = \theta_0$ *принимается на уровне значимости* α *только в том случае, если* θ_0 *принадлежит доверительной области или* $(1-\alpha)$*-доверительному интервалу*. Убедимся, что задача построения доверительного интервала для параметра θ и задача проверки простой гипотезы относительно θ взаимно обратные, то есть, если для некоторой модели известно решение одной из них, то по описанному ниже алгоритму можно получить решение другой. При этом свойства оптимальности критериев и оценок параметров взаимосвязаны и наследуют друг друга, в частности, р.н.м. критерии (когда они существуют) соответствуют наикратчайшему доверительному интервалу. Напомним, что один из общих методов построения интервальных оценок параметра θ (см. разд. 2.12 – 2.14) основан на использовании понятия *центральной статистики* $T(\vec{X};\theta)$, которая монотонно зависит от θ, а ее распределение вероятностей от параметра θ не зависит. Использование такой статистики позволяет построить доверительный интервал для параметра θ. Отметим, что центральная статистика $T(\vec{X};\theta)$ по существу является тестовой статистикой, на которой основан критерий проверки гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$.

Итак, пусть по выборке $X_1,...,X_n$ из распределения $F_X(x;\theta)$ на заданном уровне значимости α проверяются гипотезы $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$ с помощью критерия, который основан на статистике

$T(\vec{X};\theta)$ с критической областью $\mathrm{X}_{1\alpha} = \mathrm{X}_{1\alpha}(\theta_0)$ размера α. Для простоты предположим, что центральная статистика $T(\vec{X};\theta)$ при нулевой гипотезе имеет симметричное распределение относительно точки θ_0. Пусть, далее, $\mathrm{X}_{0\alpha}(\theta_0) = \overline{\mathrm{X}}_{1\alpha}(\theta_0)$ – область принятия гипотезы $H_0 : \theta = \theta_0$ и, следовательно, выполняются равенства

$$P\{\vec{X} \in \mathrm{X}_{0\alpha}(\theta_0) \mid H_0\} = (1-\alpha) = P\{\mid T(\vec{X}, \theta_0 \mid \leq t_{1-\alpha/2} \mid H_0)\}.$$

Далее, на параметрическом множестве Θ определим семейство подмножеств $\{\hbar(\vec{x}), \vec{x} \in \mathrm{X}\}$, положив $\hbar(\vec{x}) = \{\theta : \vec{x} \in \mathrm{X}_{0\alpha}(\theta_0)\}$. Событие $\{\theta \in \hbar(\vec{X})\}$ означает, что случайное множество $\hbar(\vec{X})$ накрывает параметр θ. Отметим, что события $\{\theta \in \hbar(\vec{X})\}$ и $\{\vec{X} \in \mathrm{X}_{0\alpha}(\theta)\}$ эквивалентны по построению, то есть выполняется выражение $\{\theta \in \hbar(\vec{X})\} \Leftrightarrow \{\vec{X} \in \mathrm{X}_{0\alpha}(\theta)\}$ и, следовательно, их вероятности равны:

$$P\{\theta \in \hbar(\vec{X}) \mid H_0\} = P\{\vec{X} \in \mathrm{X}_{0\alpha}(\theta) \mid H_0\} = 1-\alpha. \qquad \textbf{(3.4.73)}$$

Отсюда заключаем, что $\hbar(\vec{X})$ является доверительной областью для параметра θ с доверительной вероятностью $(1-\alpha)$. Верно и обратное: если имеется семейство γ-доверительных множеств $\{\hbar(\vec{x}), \vec{x} \in \mathrm{X}\}$ для параметра θ, то множество $\mathrm{X}_{0,1-\gamma} = \{\vec{x} : \theta_0 \in \hbar_\gamma(\vec{x})\}$ определяет область принятия гипотезы $H_0 : \theta = \theta_0$ с уровнем значимости $\alpha = 1-\gamma$.

Пример 3.4.74. Для иллюстрации описанной двойственности задач построения интервальных оценок и критериев проверки гипотез, как и в примере 3.4.24, рассмотрим случай проверки гипотез $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ по выборке $X_1,...,X_n$ из нормального распределения $F_X(x;\mu) = \Phi\{(x-\mu)/\sigma\}$ при известной дисперсии σ^2. Для проверки гипотез $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ равномерно наиболее мощный несмещенный критерий определяется критической областью вида

$$\mathrm{X}_{1\alpha} = \{\vec{x} : \sqrt{n} \mid \overline{x} - \mu_0 \mid / \sigma \geq \lambda_{1-\alpha/2}\},$$

и, следовательно, область принятия гипотезы $H_0 : \mu = \mu_0$ записывается как

$$X_{0\alpha} = \{\vec{x} : \sqrt{n}\,|\,\overline{x} - \mu_0\,|\,/\,\sigma < \lambda_{1-\alpha/2}\},$$

где $\lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ – квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения. Множество $\hbar(\vec{x}) = \{\theta : \vec{x} \in X_{0\alpha}(\theta_0)\}$ в данном случае запишется в виде

$$\hbar(\vec{x}) = \{\mu : \sqrt{n}|\overline{x} - \mu|/\sigma < \lambda_{1-\alpha/2}\} = \{\mu : \overline{x} - \lambda_{1-\alpha/2}\cdot\sigma/\sqrt{n} < \mu < \overline{x} + \lambda_{1-\alpha/2}\cdot\sigma/\sqrt{n}\},$$

а выражение (3.4.73) в форме

$$P\{\sqrt{n}|\overline{X} - \mu|/\sigma < \lambda_{1-\alpha/2}\,|\,H_0\} = P\{\overline{X} - \lambda_{1-\alpha/2}\sigma/\sqrt{n} < \mu < \overline{X} + \lambda_{1-\alpha/2}\sigma/\sqrt{n}\} = 1-\alpha\,.$$

Итак, поскольку рассматриваемый критерий проверки гипотез $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ является н.р.н.м. критерием, то доверительный интервал $\{\vec{X} \pm \lambda_{1-\alpha/2}\cdot\sigma/\sqrt{n}\}$ является наикратчайшим, то есть он имеет минимальную длину среди всех интервалов с доверительной вероятностью $(1-\alpha)$ для неизвестного параметра μ нормального распределения (см. упр. 2.15.30).

Замечание 3.4.75. Общая формулировка задачи построения оценки параметра θ на основе статистики $T(X_1,...,X_n)$ критерия проверки гипотезы $H_0 : \theta = 0$, предложенная Ходжесом и Леманом (см. приложение П.3.Ж), выглядит следующим образом. Пусть $X_1,...,X_n$ – н.о.р. случайные величины с ф.р. $F_X(x-\theta)$, $F \in \Im$ и $T(X_1,...,X_n)$ – тестовая статистика критерия для проверки гипотез $H_0 : \theta = 0, H_1 : \theta \neq 0$. Далее, пусть $T(\vec{X};\theta)$ определяется путем замены в статистике $T(X_1,...,X_n)$ наблюдений X_i на $X_i - \theta$, $i = 1,...,n$, то есть $T(\vec{X};\theta) = T(X_1 - \theta,...,X_n - \theta)$. Предполагается, что при заданной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$ наблюденное значение $T(\vec{x};\theta)$ статистики $T(\vec{X};\theta)$ является невозрастающей функцией θ. Также предполагается, что статистика $T(\vec{X}) = T(X_1,...,X_n)$ свободна от распределения F_X при гипотезе $H_0 : \theta = 0$ и имеет симметричное относительно точки μ_0 распределение. Определим значения θ^* и θ^{**} с помощью выражений $\theta^* = \sup\{\theta : T(\theta) > \mu_0\}$, $\theta^{**} = \inf\{\theta : T(\theta) < \mu_0\}$. Для оценивания параметра θ может быть использовано любое значение из интервала $[\theta^*, \theta^{**}]$. Оценка Ходжеса – Лемана параметра θ определяется в виде

$$\hat{\theta} = (\theta^* + \theta^{**})/2\,. \qquad \textbf{(3.4.76)}$$

Отметим, что если $T(\vec{x};\theta)$ – непрерывная и строго монотонная функция, то оценка Ходжеса – Лемана находится из выражения $T(\vec{X};\hat{\theta}) = \mu_0$ и может рас-

сматриваться как оценка точки, соответствующей максимуму плотности распределения вероятности тестовой статистики $T(X_1,...,X_n)$ критерия проверки гипотезы $H_0:\theta=0$. Точка максимальной вероятности тестовой статистики при нулевой гипотезе называется модальной точкой. Например, при проверке гипотез $H_0:\theta=0, H_1:\theta\neq 0$ с помощью одновыборочного *t*-критерия Стьюдента, основанного на статистике $T(\vec{X};\theta)=\sqrt{n}(\overline{X}-\theta)/\tilde{S}$, выражение $T(\vec{X};\hat{\theta})=\mu_0$ переписывается в виде $\sqrt{n}\,\overline{X}/\tilde{S}-\sqrt{n}\,\hat{\theta}/\tilde{S}=0$. Отсюда получаем оценку $\hat{\theta}=\overline{X},$ и выражение (3.4.73) запишется в виде

$$P\{\sqrt{n}\,|\,\overline{X}-\theta\,|/\tilde{S}<t_{1-\alpha/2,(n-1)}\,|\,H_0\}=$$
$$=P\{\overline{X}-t_{1-\alpha/2,(n-1)}\tilde{S}/\sqrt{n}<\theta<\overline{X}+t_{1-\alpha/2,(n-1)}\tilde{S}/\sqrt{n}\}=1-\alpha.$$

Если же тестовая статистика T имеет дискретное распределение (например, является статистикой критерия знаков вида (5.1.6)), тогда $T(\vec{x};\theta)$ является ступенчатой и невозрастающей функцией и при четном объеме выборки $n=2k$ имеется две модальные точки: $X_{(k)}, X_{(k+1)}$, а оценка вычисляется по формуле $\hat{\theta}=(X_{(k)}+X_{(k+1)})/2$; если же $n=2k+1$, то есть объем выборки нечетный, то имеется одна модальная точка $X_{(k+1)}$ и $\hat{\theta}=X_{(k+1)}$. Таким образом, при использовании статистики критерия знаков, процедура Ходжеса – Лемана приводит к оценке параметра θ в виде выборочной медианы (детали см. в примере 5.1.20).

Отметим, что процедура Ходжеса – Лемана по своей сути имеет много общего с методом максимального правдоподобия. Интуитивная обоснованность этой процедуры состоит в том, что в качестве оценки параметра θ выбирается такая величина $\hat{\theta}$, при которой сдвинутая выборка $X_1-\hat{\theta}$,..., $X_n-\hat{\theta}$ «воспринимается тестовой статистикой» так же, как и выборка из совокупности с нулевой медианой.

Далее, если при заданном уровне значимости α для *T*-критерия проверки гипотез $H_0:\theta=0$, $H_1:\theta\neq 0$ определить значения $\hat{\theta}_1$ и $\hat{\theta}_2$ с помощью выражений

$$\hat{\theta}_1=\inf\{\theta: T(\theta)<t_{1-\alpha/2}\},\ \ \hat{\theta}_2=\sup\{\theta: T(\theta)>t_{\alpha/2}\}, \qquad \textbf{(3.4.77)}$$

в которых критические значения $t_{\alpha/2}$ и $t_{1-\alpha/2}$ тестовой статистики T при двусторонней альтернативе $H_1:\theta\neq 0$ определяются из равенств

$$P\{T(\vec{X})\le t_{\alpha/2}\,|\,H_0\}=P\{T(\vec{X})\ge t_{1-\alpha/2}\,|\,H_0\}=\alpha/2,$$

тогда интервал $[\hat{\theta}_1,\hat{\theta}_2]$ является $(1-\alpha)$-доверительным интервалом для пара-

метра θ, основанным на статистике $T(X_1,...,X_n)$, то есть выполняются равенства

$$P\{|\,T(\vec{X};\theta)|\le t_{1-\alpha/2}\,|\,H_0\}=P\{\hat{\theta}_1\le\theta\le\hat{\theta}_2\}=1-\alpha\,. \qquad \textbf{(3.4.78)}$$

Применение процедуры Ходжеса – Лемана для построения точечных и интервальных оценок параметра θ с использованием критерия знаковых рангов Уилкоксона при проверке гипотезы $H_0:\theta=0$, $F_X\in\Im_{S|0}$, приводится в примере 5.2.24.

Пример 3.4.79. Пусть $L(X)=N(\mu,\sigma^2)$. Имеется выборка $X_1,...,X_n$ из нормального распределения $F_X(x;\mu)=\Phi\{(x-\mu)/\sigma\}$, параметры μ и σ^2 неизвестны. Напомним, что оценками максимального правдоподобия параметров μ и σ^2 являются соответственно выборочное среднее $\overline{X}=\sum X_i/n$ и выборочная дисперсия $S^2=\sum(X_i-\overline{X})^2/n$, а несмещенной оценкой параметра σ^2 является подправленная выборочная дисперсия, которая имеет вид $\tilde{S}^2=\sum(X_i-\overline{X})^2/(n-1)$. Отметим, что $nS^2=\sum(X_i-\overline{X})^2=\tilde{S}^2(n-1)$ и, согласно теореме (2.8.31), выполняются выражения $L(nS^2/\sigma^2)=\chi^2(n-1)$ и $L\{(n-1)\tilde{S}^2/\sigma^2\}=\chi^2(n-1)$. Воспользуемся интервальной оценкой параметра σ^2, определяемой выражением (2.12.9), и построим двусторонний критерий проверки гипотез $H_0:\sigma^2=\sigma_0^2$, $H_1:\sigma^2\ne\sigma_0^2$. Для этого запишем сначала критические области размера α односторонних критериев для проверки гипотезы $H_0:\sigma^2=\sigma_0^2$ против односторонних альтернатив $H_1^+:\sigma^2>\sigma_0^2$ и $H_1^-:\sigma^2<\sigma_0^2$. Учитывая, что $L\{(n-1)\tilde{S}^2/\sigma^2\}=\chi^2(n-1)$, получаем критические области в виде

$$\mathrm{X}_{1\alpha}^{+}=\{\vec{x}:(n-1)\tilde{S}^2/\sigma_0^2\ge\chi^2_{1-\alpha,(n-1)}\},\ \mathrm{X}_{1\alpha}^{-}=\{\vec{x}:(n-1)\tilde{S}^2/\sigma_0^2\le\chi^2_{\alpha,(n-1)}\}.$$

Используя теперь общее правило (см. замечание 3.4.35 и выражение (3.4.37)) и объединяя эти критические области, получаем критическую область размера α двустороннего критерия, основанного на статистике $T(\vec{X},\sigma_0^2)=(n-1)\tilde{S}^2/\sigma_0^2$, для проверки гипотез $H_0:\sigma^2=\sigma_0^2$, $H_1:\sigma^2\ne\sigma_0^2$, в виде

$$\mathrm{X}_{1\alpha}=\{\vec{x}:(n-1)\tilde{S}^2/\sigma_0^2<\chi^2_{\alpha/2,(n-1)}\ \ \textit{или}\ \ (n-1)\tilde{S}^2/\sigma_0^2>\chi^2_{1-\alpha/2,(n-1)}\}.$$

Таким образом, с учетом двойственности задач построения интервальных оценок и проверки гипотез, согласно (3.4.73), выполняются равенства

$$P\{T(\vec{X},\sigma^2)\in\overline{\mathrm{X}}_{1\alpha}\,|\,H_0\}=P\left\{\frac{(n-1)\tilde{S}^2}{\chi^2_{1-\alpha/2,(n-1)}}\le\sigma^2\le\frac{(n-1)\tilde{S}^2}{\chi^2_{\alpha/2,(n-1)}}\right\}=1-\alpha\,.$$

Отметим, что для критерия, определяемого критической областью $\mathrm{X}^{+}_{1\alpha}$, его мощность и необходимый объем выборки n для обеспечения заданной мощности, вычисляются в упр. 3.9.55.

3.4.Д. Общие выводы относительно свойств одновыборочных критериев проверки параметрических гипотез

Подведем итог относительно общих свойств критериев проверки гипотез в рамках параметрических моделей $\{\mathrm{X}, \mathfrak{I}_\theta\}$, для которых множество допустимых распределений $\mathfrak{I}_\theta$ определено в виде параметрического семейства заданных функций $F_X(x,\theta)$ с конечным числом *неизвестных* параметров θ, то есть в виде $\mathfrak{I}_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$, где Θ – заданное множество параметров. При формулировке асимптотических результатов предполагается, что параметрическая модель $\{\mathrm{X}, \mathfrak{I}_\theta\}$ удовлетворяет условиям регулярности, точная формулировка которых приводится в [38] и [66, с. 353]. Предполагаем, что имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x,\theta)$ с плотностью $f_X(x,\theta)$, и ниже используются следующие обозначения. Совместная плотность наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$ обозначается в виде

$$L(x_1,...,x_n;\theta) = L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta),$$

а ее частные производные обозначаются как

$$L_k(\vec{x};\theta) = \partial^{(k)} L(\vec{x};\theta) / \partial\theta^k,\ k = 1,\ 2,...$$

Общие свойства критериев проверки гипотез в рамках параметрических моделей $\{\mathrm{X}, \mathfrak{I}_\theta\}$ формулируются следующим образом.

1. Согласно фундаментальной лемме Неймана – Пирсона, для проверки простых гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ критерий, основанный на статистике

$$l_n(\vec{x}) = l_n(\vec{x};\theta_0,\theta_1) = \frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)} = \prod_{i=1}^{n} \frac{f_X(x_i;\theta_1)}{f_X(x_i;\theta_0)},$$

является *состоятельным* (см. упр. 3.9.53) и *наиболее мощным* критерием среди всех критериев с заданным уровнем значимости α. Этот критерий задан критической областью, которая определяется выражением (3.3.6), см. также замечание 3.3.28.

2. Асимптотическая критическая область размера α критерия, основанного на статистике Неймана – Пирсона $l_n(\vec{X};\theta_0,\theta_1)$, для проверки гипотез $H_0:\theta=\theta_0$ и $H_1^+:\theta>\theta_0$ записывается в виде

$$\mathrm{X}_{1\alpha}^{+}=\{\vec{x}:l_n(\vec{X};\theta_0,\theta_1)\geq\exp(-nI(\theta_0,\theta_1)+\sigma_0\sqrt{n}\cdot\lambda_{1-\alpha})\},$$
$$\lambda_{1-\alpha}=\Phi^{-1}(1-\alpha);$$

для проверки гипотез $H_0:\theta=\theta_0$, $H_1^-:\theta<\theta_0$ она определяется в следующем виде:

$$\mathrm{X}_{1\alpha}^{-}=\{\vec{x}:l_n(\vec{X};\theta_0,\theta_1)\leq\exp(-nI(\theta_0,\theta_1)+\sigma_0\sqrt{n}\cdot\lambda_{\alpha})\},\ \lambda_{\alpha}=\Phi^{-1}(\alpha).$$

Для проверки гипотез $H_0:\theta=\theta_0, H_1:\theta\neq\theta_0$ асимптотическая критическая область размера α определяется выражением

$$\mathrm{X}_{1\alpha}=\{\vec{x}:|\,l_n(\vec{X};\theta_0,\theta_1)\,|\geq\exp(-nI(\theta_0,\theta_1)+\sigma_0\sqrt{n}\cdot\lambda_{1-\alpha/2})\},$$

где количество информации по Кульбаку – Лейблеру $I(\theta_0,\theta_1)$ определено в (3.4.56) и величина σ_0^2 вычисляется по формуле (3.4.59).

3. Для проверки простой гипотезы $H_0:\theta=\theta_0$ против односторонних и сложных альтернатив $H_1^+:\theta>\theta_0$, $H_1^-:\theta<\theta_1$ общий вывод (см. выражение (3.4.5)) относительно класса моделей, для которых существуют р.н.м. критерии формулируется следующим образом:

«Если семейство распределений $\Im_\theta$ параметрической модели $\{\mathrm{X},\Im_\theta\}$, где $\theta\in\Theta$ вещественный параметр, имеет *монотонное отношение правдоподобия*, то для него *существует равномерно наиболее мощный критерий* для проверки простой гипотезы $H_0:\theta=\theta_0$ против односторонней и сложной альтернативы H_1^+ (либо против альтернативы H_1^-). Этот критерий совпадает с критерием Неймана

– Пирсона для проверки простой гипотезы $H_0 : \theta = \theta_0$ против произвольной фиксированной и простой альтернативы H_1^+ либо H_1^-».

Отметим, что для экспоненциальных семейств распределений, имеющих монотонное отношение правдоподобия, для которых плотность распределения вероятностей определяется в виде

$$f_X(x;\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\},$$

всегда существуют равномерно наиболее мощные критерии для проверки простой гипотезы $H_0 : \theta = \theta_0$ против односторонних и сложных альтернатив $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$, которые соответственно определяются критическими областями

$$\mathrm{X}_{1\alpha}^+ = \{\vec{x} : \sum B(x_i) \geq t_{1-\alpha}\} \text{ и } \mathrm{X}_{1\alpha}^- = \{\vec{x} : \sum B(x_i) \leq t_\alpha\},$$

при условии, что $A(\theta)$ – возрастающая функция. Если же функция $A(\theta)$ убывает, то неравенства в критических областях следует заменить на противоположные.

4. Для проверки простой гипотезы $H_0 : \theta = \theta_0$ против двусторонней сложной альтернативы $H_1 : \theta \neq \theta_0$ параметрической модели $\{\mathrm{X}, \Im_\theta\}$ *р.н.м. критерии не существуют*, однако для проверки таких гипотез могут быть построены *несмещенные равномерно наиболее мощные критерии* и *локально наиболее мощные несмещенные критерии* (см. упр. 3.9.1).

Согласно теореме (3.4.39), для проверки гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$ *регулярной* параметрической модели $\{\mathrm{X}, \Im_\theta\}$ несмещенный равномерно наиболее мощный (н.р.н.м.) критерий уровня значимости α определяется критической областью

$$\mathrm{X}_{1\alpha} = \{\vec{x} : L(\vec{x}\,;\theta_1) \geq cL(\vec{x}\,;\theta_0) + c_1 L_1(\vec{x}\,;\theta_0)\}$$

при условии, что она не зависит от $\theta_1 \neq \theta_0$. Постоянные $c \geq 0$ и c_1 однозначно определяются из условий

$$\int\limits_{\mathrm{X}_{1\alpha}} L(\vec{x}\,;\theta_0)d\vec{x} = \alpha \text{ и } \int\limits_{\mathrm{X}_{1\alpha}} L_1(\vec{x}\,;\theta_0)d\vec{x} = 0.$$

5. Для проверки гипотез $H_0:\theta=\theta_0$, $H_1^+:\theta>\theta_0$ *локально наиболее мощный критерий* (см. упр. 3.9.1) основан на критической области $X_{1\alpha}^+$ размера α, которая определяется в виде

$$X_{1\alpha}^+=\{\vec{x}: L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0)\ge t_{1-\alpha}\},$$

где критическое значение $t_{1-\alpha}$ статистики $\tilde{\Lambda}(\vec{X};\theta_0)=L_1(\vec{X};\theta_0)/L(\vec{X};\theta_0)$ является квантилем уровня $(1-\alpha)$ ф.р. $F_{\tilde{\Lambda}}(x\,|\,H_0)$, статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0:\theta=\theta_0$, то есть $t_{1-\alpha}=F_{\tilde{\Lambda}}^{-1}\{(1-\alpha)\,|\,H_0\}$, и $t_{1-\alpha}$ находится при заданном уровне значимости α из выражения

$$\int_{t_{1-\alpha}}^{\infty} f_{\tilde{\Lambda}}(x;\theta_0)dx=\alpha.$$

Здесь $f_{\tilde{\Lambda}}(x;\theta_0)$ обозначает плотность распределения вероятностей статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0:\theta=\theta_0$.

Для проверки гипотез $H_0:\theta=\theta_0$, $H_1^-:\theta<\theta_0$ *локально наиболее мощный критерий* основан на критической области $X_{1\alpha}^-$ размера α, которая определяется в виде

$$X_{1\alpha}^-=\{\vec{x}: L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0)\le t_{\alpha}\},$$

где критическое значение t_{α} статистики $\tilde{\Lambda}(\vec{X};\theta_0)=L_1(\vec{X};\theta_0)/L(\vec{X};\theta_0)$ находится при заданном уровне значимости α из выражения

$$\int_{-\infty}^{t_{\alpha}} f_{\tilde{\Lambda}}(x;\theta_0)dx=\alpha.$$

Для проверки гипотезы $H_0:\theta=\theta_0$ против двусторонней альтернативы $H_1:\theta\ne\theta_0$ *несмещенный локально наиболее мощный* критерий основан на критической области $\tilde{X}_{1\alpha}$ размера α, определяемой в виде

$$\tilde{X}_{1\alpha}=\{\vec{x}: L_2(\vec{x};\theta_0)\ge c_1L_1(\vec{x};\theta_0)+cL(\vec{x};\theta_0)\},$$

где константы c и c_1 определяются соответственно из условий $W(\theta_0)=\alpha$, $W'(\theta_0)=0$.

6. Для больших объемов выборки, точнее при $n\to\infty$, имеем л.н.м. критерий проверки гипотезы $H_0:\theta=\theta_0$ против правосторонней альтернативы $H_1^+:\theta>\theta_0$, который задается асимптотической критической областью $X_{1\alpha}^+$ размера α, определяемой в виде

$$X_{1\alpha}^+=\{\vec{x}: L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0)\ge t_{1-\alpha}\}=\{\vec{x}: U(\vec{x};\theta_0)\ge \lambda_{1-\alpha}\sqrt{ni_1(\theta_0)}\},$$

где $\lambda_{1-\alpha}=\Phi^{-1}(1-\alpha)$, $i_1(\theta_0)$ – количество информации Фишера относительно θ_0, содержащееся в одном наблюдении и $U(\vec{x};\theta_0)$ – наблюденное значение функции вклада выборки $U(\vec{X};\theta)$, вычисленное по наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}=(X_1,...,X_n)$ при справедливости гипотезы $H_0:\theta=\theta_0$ по формуле

$$U(\vec{x};\theta_0)=L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0)=\partial\ln L(\vec{x};\theta)/\partial\theta\,|_{\theta=\theta_0}=$$

$$=\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(x_i;\theta)\,|_{\theta=\theta_0}.$$

Аналогично, для проверки гипотезы $H_0:\theta=\theta_0$ против левосторонней альтернативы $H_1^-:\theta<\theta_0$ л.н.м. критерий задается асимптотической критической областью $X_{1\alpha}^-$ размера α, определяемой в виде

$$X_{1\alpha}^-=\{\vec{x}: L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0)\le t_\alpha\}=\{\vec{x}: U(\vec{x};\theta_0)\le \lambda_\alpha\sqrt{ni(\theta_0)}\},$$

$$\lambda_\alpha=\Phi^{-1}(\alpha).$$

Объединяя результаты для односторонних альтернатив H_1^+ и H_1^-, в случае двусторонней альтернативы, т.е. для проверки гипотез $H_0:\theta=\theta_0$, $H_1:\theta\ne\theta_0$, с учетом того, что $\lambda_{\alpha/2}=-\lambda_{1-\alpha/2}$, асимптотически л.н.м. критерий уровня значимости α определяется критической областью

$$\tilde{X}_{1\alpha}=\{\vec{x}:|T(\vec{x})|\ge\lambda_{1-\alpha/2}\}=\{\vec{x}:|U(\vec{x};\theta_0)|\ge\lambda_{1-\alpha/2}\sqrt{ni_1(\theta_0)}\}.$$

7. Согласно теореме (3.4.72), при выполнении условий регулярности, обеспечивающих существование, единственность и асимптотическую нормальность оценок максимального правдоподобия $\hat{\theta}_n = (\hat{\theta}_{1n},...,\hat{\theta}_{rn})$ векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r)$ (см. Уилкс [66, с. 354, 418, 426]), критерий отношения правдоподобия размера α для проверки гипотезы $H_0 : \vec{\theta} = \vec{\theta}_0$ с учетом того, что при $n \to \infty$ выполняется выражение

$$P_\theta(\vec{X} \in \mathrm{X}_{1\alpha} \mid H_0) = P_\theta(-2\ln\{\Lambda^*(\vec{X}\,;\vec{\theta}_0)\} \geq \chi_{1-\alpha,\,r} \mid H_0) \to \alpha,$$

задается асимптотически критической областью вида

$$\mathrm{X}_{1\alpha} = \{\vec{x} : -2\ln\Lambda^*(\vec{x}\,;\vec{\theta}_0) \geq \chi^2_{1-\alpha,\,r}\}.$$

8. В случае векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r;\theta_{r+1},...,\theta_s)$ для проверки сложной гипотезы $H_0 : \theta_1 = \theta_{10}\,,...,\,\theta_r = \theta_{r0}$ против сложной альтернативы, что не все равенства имеют место, для определения критической области критерия отношения правдоподобия используется асимптотический результат (см. [66, с. 426] и [24, с. 301]). Согласно этому результату, при нулевой гипотезе выполняется выражение

$$L(-2\ln\{\Lambda^*(\vec{X}\,;\Theta_0)\}) = \chi^2(s-r),$$

где s – число параметров при альтернативе H_1, а r – число параметров при гипотезе H_0. Более полное обсуждение этих вопросов может быть найдено в монографиях [7, 37, 66].

3.5. Критерии согласия

Критерии согласия предназначены для решения задачи согласия, которая была сформулирована в примере 3.2.15. Эта задача относится к основным статистическим задачам на этапе первичной обработки собранной статистической информации и состоит в конкретизации распределения вероятностей $F_X(x)$ изучаемой с.в. X, которая осуществляется при построении статистической модели. Представление о неизвестной до проведения опыта ф.р. $F_X(x)$

можно получить и непосредственным образом, путем построения эмпирической ф.р. $F_n(x)$ по выборке $\vec{X} = (X_1,...,X_n)$ (см. разд. 2.7). Однако при построении статистических моделей часто возникает необходимость в использовании аналитического выражения для функции распределения $F_X(x)$. Например, для обоснованного применения методов нормальной теории необходимо иметь инструменты проверки гипотезы о том, что наблюдаемые данные $\vec{X} = (X_1,...,X_n)$ могут рассматриваться как выборка из нормального распределения. В качестве таких инструментов используют критерии согласия, которые применяют для проверки адекватности принятой статистической модели.

В настоящее время разработано достаточно большое количество критериев согласия, каждый из них имеет свои границы применимости и свои достоинства и недостатки. В этом разделе приведем обсуждение исторически первого критерия согласия, предложенного Пирсоном. Этот критерий получил название «хи-квадрат-критерий», он является универсальным в том смысле, что его можно применять как в непрерывном, так и в дискретном случаях. Этот критерий допускает простую модификацию для сложной гипотезы и наиболее часто используется на практике. Общий подход к построению непараметрических критериев согласия с обсуждением свойств таких критериев, как критерии Колмогорова, Смирнова, Реньи, Андерсона – Дарлинга, Крамера – Мизеса, которые ориентированы на решение задачи согласия в непрерывном случае и основаны на использовании свойств порядковых статистик, приводится в разд. 4.7.

3.5.А. χ^2-критерий согласия Пирсона в дискретном случае. Этот критерий был предложен К. Пирсоном в 1900 году, и в литературе за ним закрепилось название «хи-квадрат-критерий, χ^2-критерий». Обсудим сначала этот критерий для решения задачи согласия с простой гипотезой в дискретном случае. Отметим, что непрерывный случай сводится к дискретному с помощью метода группировки наблюдений.

Пусть проводится n независимых опытов, в каждом из которых может осуществиться одно из k попарно несовместных событий

$A_1,...,A_k$ с вероятностями $p_1,...,p_k$, причем сумма этих вероятностей равна единице. Обозначим вектор вероятностей событий $A_1,...,A_k$ через $\vec{p} = (p_1,...,p_k)$, а через $\vec{n} = (n_1,...,n_k)$ – вектор частот реализаций событий $A_1,...,A_k$ в n независимых опытах, $(n_1 + ... + n_k) = n$. Требуется по результатам этих n опытов проверить гипотезу о том, что вероятности $p_1,...,p_k$ равны заданным значениям $p_{1,0},...,p_{k,0}$.

Отметим, что данная схема испытаний эквивалентна n независимым наблюдениям $X_1,...,X_n$ над дискретной с.в. X с конечным числом k отстоящих друг от друга возможных состояний $\tilde{x}_1,...,\tilde{x}_k$, вероятности которых обозначены через $p_i = P(X = \tilde{x}_i)$, $i = 1,...,k$.

Итак, пусть имеется выборка $X_1,...,X_n$ объема n из совокупности объектов, признак X которых описывается функцией распределения вероятностей

$$F_X(x) = \sum_{i=1}^{k} p_i C(x - \tilde{x}_i),\ x \in R^1.$$

Требуется по этим данным проверить гипотезу H_0 о том, что вероятности $p_1,...,p_k$ возможных состояний $\tilde{x}_1,...,\tilde{x}_k$ дискретной с.в. X равны заданным значениям $p_{1,0},...,p_{k,0}$, то есть её функция распределения вероятностей имеет вид

$$F_0(x) = \sum_{i=1}^{k} p_{i,0} C(x - \tilde{x}_i),\ x \in R^1.$$

Эта гипотеза кратко записывается в виде $H_0 : F_X(x) = F_0(x)$, для всех $x \in R^1$, или через вероятности $p_1,...,p_k$ состояний $\tilde{x}_1,...,\tilde{x}_k$ в виде

$$H_0 : p_1 = p_{1,0},\quad p_2 = p_{2,0}\quad ,...,\quad p_k = p_{k,0}. \qquad \textbf{(3.5.1)}$$

Проверка гипотезы H_0 против альтернативы, что не все равенства в (3.5.1) выполняются, осуществляется с помощью χ^2-критерия. Для

вычисления тестовой статистики этого критерия сначала по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ строится дискретный вариационный ряд частот, то есть

$$(x_1,...,x_n) \Rightarrow \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ n_1,...,n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n, \qquad \textbf{(3.5.2)}$$

а затем, по данным этого вариационного ряда частот вычисляется статистика χ^2 в виде

$$\chi^2 = \sum_{i=1}^{k} [n_i - np_{i,0}]^2 / np_{i,0} = n\sum_{i=1}^{k} \left(\frac{n_i}{n} - p_{i,0} \right)^2 / p_{i,0} = \sum_{i=1}^{k} n_i^2 / (np_{i,0}) - n. \qquad \textbf{(3.5.3)}$$

Отметим, что статистика хи-квадрат характеризует различие между эмпирическими $n_1,...,n_k$ и гипотетическими частотами $np_{i,0}$, $i = 1,...,k,$ и определена в виде квадрата расстояния между вектором относительных частот $(n_1/n,...,n_k/n)$, полученных в результате опыта, и вектором гипотетических вероятностей $(p_{1,0},...,p_{k,0})$. Отличие этого расстояния от евклидовой метрики состоит лишь в том, что разные координаты векторов входят в него с разными весами. Отметим далее, что, согласно закону больших чисел, отношение $n_i / n \to^P p_i$, $i = 1,...,k$, при $n \to \infty$, поэтому выполняется выражение

$$\sum_{i=1}^{k} \left(\frac{n_i}{n} - p_{i,0} \right)^2 / p_{i,0} \to^P \sum_{i=1}^{k} \left(p_i - p_{i,0} \right)^2 / p_{i,0} \text{ при } n \to \infty$$

и, следовательно, при справедливости гипотезы H_0, правая часть этого выражения равняется нулю. Если основная гипотеза H_0 неверна, то статистика χ^2 имеет тенденцию увеличиваться и, следовательно, критическая область χ^2-критерия запишется в виде $X_{1\alpha} = \{\vec{x} : \chi^2 > c\}$. Хорошо известно (см., например, [22, с. 110]), что при больших объемах выборки (точнее, при $n \geq 50$) и при выполнении неравенства $\min(n_1,...,n_k) \geq 5$ статистика χ^2 при справедливости гипотезы H_0 имеет распределение хи-квадрат с числом сте-

пеней свободы $(k-1)$, то есть для неё выполняется следующее выражение:

$$L(\chi^2 \mid H_0) = \chi^2(k-1) \text{ при } n \to \infty. \qquad \textbf{(3.5.4)}$$

С учетом этого критическая область размера α статистики χ^2 запишется в виде

$$\mathrm{X}_{1\alpha} = \{\vec{x} : \chi^2 > \chi^2_{1-\alpha,\,(k-1)}\}, \qquad \textbf{(3.5.5)}$$

где $\chi^2_{1-\alpha,\,(k-1)}$ – квантиль уровня $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы $(k-1)$.

Таким образом, при больших объемах выборки χ^2-критерий проверки гипотезы H_0, определенной в (3.5.1), записывается в виде следующего правила:

«*Если вычисленное по данным вариационного ряда частот значение статистики* χ^2 *превышает ее критическое значение* $\chi^2_{1-\alpha,(k-1)}$, *то гипотеза* H_0 *отвергается на уровне значимости* α, *если* $\chi^2 < \chi^2_{1-\alpha,(k-1)}$, *то нет оснований отвергать гипотезу* H_0*».* **(3.5.6)**

Асимптотические свойства χ^2-критерия. К наиболее важным асимптотическим характеристикам критериев проверки гипотез относится понятие состоятельности критериев (см. определение 3.1.45). Для критерия хи-квадрат, определяемого критической областью (3.5.5), справедлива следующая теорема.

Теорема 3.5.7. Пусть вектор $\vec{p} = (p_1,...,p_k)$ характеризует вероятности возможных состояний $\tilde{x}_1,...,\tilde{x}_k$ дискретной с.в. X при альтернативе $H_1 : \vec{p} \neq \vec{p}_0$ и $\vec{p}_0 = (p_{1,0},...,p_{k,0})$ – вектор, заданный гипотезой (3.5.1). Тогда для любого вектора $\vec{p} \neq \vec{p}_0$, при увеличении объема выборки, то есть при $n \to \infty$, функция мощности $W_n(\mathrm{X}_{1\alpha} \mid H_1)$ хи-квадрат-критерия с критической областью (3.5.5) стремится к единице. Таким образом, χ^2-критерий является асимптотически несмещенным и состоятельным критерием.

Доказательство. Следуя работе [22], перепишем формулу (3.5.3) для статистики хи-квадрат-критерия в виде

$$\chi^2=\sum_{i=1}^{k}[n_i-np_i]^2/np_{i,0}+2\sum_{i=1}^{k}(n_i-np_i)(p_i-p_{i,0})/p_{i,0}+\sum_{i=1}^{k}(p_i-p_{i,0})^2/p_{i,0}.$$

Используя данное выражение, с учетом того, что

$$M(n_i|H_1)=np_i,\ M[(n_i-np_i)^2|H_1]=D(n_i|H_1)=np_i(1-p_i),\ i=1,...,k,$$

получаем математическое ожидание статистики хи-квадрат при альтернативе H_1 в виде

$$M(\chi^2\mid H_1)=n\sum_{i=1}^{k}(p_i-p_{i,0})^2/p_{i,0}+\sum_{i=1}^{k}p_i(1-p_i)/p_{i,0}.$$

Из последнего выражения следует, что при справедливости альтернативы H_1 и при больших объемах выборки n математическое ожидание статистики хи-квадрат будет иметь порядок роста n, и в частности при нулевой гипотезе, будет равно $M(\chi^2\mid H_0)=k-1$. Можно также убедиться, что дисперсия статистики хи-квадрат при альтернативе H_1 также имеет порядок роста n при $n\to\infty$ (детали см. в [22]). Отметим, что при нулевой гипотезе дисперсия статистики хи-квадрат вычисляется по формуле

$$D(\chi^2\mid H_0)=2(k-1)+\frac{1}{n}\sum_{i=1}^{k}(1/p_{i,0})-(k^2/n)\ -(2k/n)+(2/n)$$

и, в частности, выполняется равенство $\lim_{n\to\infty}D(\chi^2\mid H_0)=2(k-1)$. Далее, используя неравенство Чебышева (см. П.3.6), получаем

$$\begin{aligned}1-W_n(\mathrm{X}_{1\alpha}\mid H_1)&=P\{\chi^2<\chi^2_{1-\alpha,(k-1)}\mid H_1\}=\\&=P\{M(\chi^2\mid H_1)-\chi^2>M(\chi^2\mid H_1)-\chi^2_{1-\alpha,(k-1)}\mid H_1\}\le\\&\le P\{|M(\chi^2\mid H_1)-\chi^2|>|M(\chi^2\mid H_1)-\chi^2_{1-\alpha,(k-1)}|\}\le\\&\le D(\chi^2\mid H_1)/[M(\chi^2\mid H_1)-\chi^2_{1-\alpha,(k-1)}]^2=O(1/n).\end{aligned}$$

Отсюда следует, что $\lim_{n\to\infty}W_n(\mathrm{X}_{1\alpha}\mid H_1)=1$. Доказательство завершено.

Примеры, иллюстрирующие применение χ^2-критерия для проверки гипотезы о симметрии подбрасываемой монеты в опытах Бюффона и К. Пирсона, приводятся в упр. 3.9.67. Проверка гипотезы согласия с распределением Пуассона обсуждается в упр. 3.9.69. Применение χ^2-критерия для проверки гипотезы нормальности приводится в упр. 3.9.78 и 3.9.78а.

3.5.Б. χ^2-критерий согласия в непрерывном случае

Обозначим множество абсолютно непрерывных ф.р. через $\Im$. Пусть далее случайная величина X имеет функцию распределения $F_X(x) \in \Im$ с плотностью $f_X(x)$, $x \in R^1$. Требуется по выборке $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x) \in \Im$ проверить гипотезу о том, что неизвестная ф.р. $F_X(x)$ изучаемой с.в. X равняется заданной ф.р. $F_0(x)$. Эта задача согласия с *простой* гипотезой H_0, которая кратко записывается в виде

$$H_0 : F_X(x) = F_0(x), \text{ для всех } x \in R^1. \qquad \textbf{(3.5.8)}$$

Для проверки этой гипотезы с помощью χ^2-критерия нужно выполнить следующие действия. Сначала осуществляется группировка данных. Для этого область возможных значений с.в. X разбивается на заданное число k интервалов группировки $\Delta_1,...,\Delta_k$, где $\Delta_i = y_{i+1} - y_i$, $i = 1,...,k$. Затем наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ отображается в интервальный вариационный ряд частот $n_i,...,n_k$, то есть

$$\vec{x} = (x_1,...,x_n) \Rightarrow \begin{pmatrix} \Delta_1,...,\Delta_k \\ n_1,...,n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n. \qquad \textbf{(3.5.9)}$$

По данным этого интервального вариационного ряда частот вычисляется статистика χ^2-критерия в виде

$$\chi^2 = \sum_{i=1}^{k} [(n_i - np_i)^2 / np_i], \qquad \textbf{(3.5.10)}$$

где p_i, $i = 1,...,k$, – вероятности попадания с.в. X в интервалы группировки $\Delta_1,...,\Delta_k$ и вычисляются они при справедливости гипотезы H_0 по формулам

$$p_i = P\{X \in \Delta_i \mid H_0\} = \int_{y_i}^{y_{i+1}} dF_0(x) = F_0(y_{i+1}) - F_0(y_i),\ i = 1,...,k. \quad \textbf{(3.5.11)}$$

Для статистики хи-квадрат, вычисляемой по формуле (3.5.10) с учетом (3.5.11), также справедливо выражение (3.5.4), поэтому при больших объемах выборки критерий проверки простой гипотезы согласия (3.5.8) определяется критической областью вида

$$\mathrm{X}_{1\alpha} = \{\vec{x} : \chi^2 > \chi^2_{1-\alpha,\,(k-1)}\}, \quad \textbf{(3.5.12)}$$

где $\chi^2_{1-\alpha,\,(k-1)}$ – квантиль уровня $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы $(k-1)$, и наблюденное значение статистики χ^2 вычисляется по данным интервального ряда частот (3.5.9) по формуле (3.5.10) с учетом (3.5.11).

Замечание 3.5.13. Отметим, что критерий согласия хи-квадрат в непрерывном случае зависит от выбираемого нами числа k интервалов группировки. Эта процедура не является однозначной и требует аккуратного обращения. При построении интервалов группировки на практике обычно используют следующую рекомендацию. Если среди построенных интервалов группировки имеются «пустые», точнее, если не выполняется неравенство $\min(n_1,...,n_k) \geq 5$, то интервалы группировки укрупняют путем их объединения. Выполнение этого неравенства обеспечивает при больших объемах выборки $(n \geq 50)$ удовлетворительную аппроксимацию распределения статистики хи-квадрат при гипотезе H_0 в виде выражения (3.5.4). Более полно эта рекомендация, сформулированная К. Кокреном (см., например, [56, т. 1, с. 366]), выглядит следующим образом. Интервалы группировки на «хвостах» распределений объединяют таким образом, чтобы наименьшая ожидаемая частота была не меньше единице. Для максимизации чувствительности критерия размер интервалов группировки следует выбирать достаточно малым, чтобы ожидаемые частоты не превышали 12 в каждом интервале для выборки объема $n = 200$, 20 – для выборки объема $n = 400$ и 30 – для выборки объема $n = 1000$.

См. также [79] и работу [Б.Ю. Лемешко, С.Н. Постовалов. Рекомендации ГОСТ Р 50.1-033-2001 и ГОСТ Р 50.1-037-2002 как итог исследований вопросов применения критериев согласия // Труды 10-го юбилейного симпозиума по непараметрическим и робастным статистическим методам в кибернетике. – Томск: Изд-во НТЛ, 2004. С. 55–67].

3.5.В. χ^2-критерий согласия для сложной гипотезы

На практике наиболее часто встречается задача согласия со сложной гипотезой, определенной в (3.2.17), которая выделяет не единственное распределение, а некоторое подмножество распределений на множестве $\Im$, например в виде *параметрического* семейства $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$ либо в виде *непараметрического* множества $\Im_0$, содержащего непрерывные распределения с заданными квантилями (см. разд. 3.5.Г). Рассмотрим сначала задачу согласия со сложной параметрической гипотезой. Пусть статистические данные представлены выборкой $X_1,...,X_n$ из распределения $F_X(x;\theta)$, функциональный вид которого задается с точностью до неизвестного (в общем случае векторного) параметра $\theta \in \Theta$. Требуется по этой выборке $X_1,...,X_n$ проверить *сложную* гипотезу H_0^* вида

$$H_0^* : F_X(x;\theta) = F_0(x;\theta), \text{ для всех } x \in R^1,\ \theta \in \Theta. \qquad \textbf{(3.5.14)}$$

При построении критериев согласия проверки сложных гипотез обычно используется следующий общий подход, суть которого для удобства дальнейших ссылок приведем в виде замечания.

Замечание 3.5.15. При построении критериев согласия проверки сложных гипотез обычно сложная гипотеза предварительно сводится к простой гипотезе путем построения оценки $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ неизвестного параметра θ по информации, содержащейся в исходной выборке $X_1,...,X_n$. В результате этой операции сложная гипотеза H_0^* превращается в простую гипотезу $H_0 : F_X(x;\theta) = F_0(x;\hat{\theta})$, и ее проверка может быть осуществлена по схеме, описанной в разд. 3.5. Однако при этом тестовая статистика, и в частности статистика хи-квадрат, будет зависеть от оценки $\hat{\theta} = \hat{\theta}(X_1,...,X_n)$ неизвестного параметра θ (то есть её модифицированная форма запишется в виде $\chi^2(\hat{\theta})$) и, следовательно, распределение вероятностей модифицированной статистики будет определяться выбранным способом построения оценки $\hat{\theta}$. В тех случаях, когда распределение модифицированной статистики $\chi^2(\hat{\theta})$ при гипотезе H_0^* может быть найдено, этот подход позволяет решить поставленную задачу. Отметим, что при реализации этого подхода на практике при проверке

сложной гипотезы часто ошибочно для модифицированной статистики используют такие же критические значения статистик, как и в случае простой гипотезы. Так делать нельзя, так как при этом не обеспечивается заданный уровень значимости критерия проверки сложной гипотезы. Ниже обсуждается реализация этого подхода с использованием мультиномиальных оценок максимального правдоподобия для параметра $\theta \in \Theta$, которые строятся не по исходной выборке $X_1,...,X_n$, а по данным интервального ряда частот.

Пусть для определенности неизвестный параметр $\theta \in \Theta$ является векторным параметром размерности r, то есть $\vec{\theta} = (\theta_1,...,\theta_r)$. Для проверки гипотезы (3.5.14) с помощью хи-квадрат-критерия, как и в случае простой гипотезы (см. разд.3.5.Б), сначала осуществляется группировка данных и для этого область возможных значений с.в. X разбивается на заданное число k интервалов группировки $\Delta_1,...,\Delta_k$, где $\Delta_i = y_{i+1} - y_i$, $i = 1,...,k$, а затем наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ отображается в интервальный вариационный ряд частот $n_1,...,n_k$, определенный в (3.5.9). Сравнение наблюденных частот $n_1,...,n_k$, с гипотетическими частотами np_i, $i = 1,...,k$, при справедливости гипотезы H_0^* осуществляется с помощью статистики хи-квадрат. Для ее вычисления требуется предварительно определить вероятности p_i, $i = 1,...,k$, попадания с.в. X в интервалы группировки $\Delta_1,...,\Delta_k$. В данном случае эти вероятности вычисляются по формулам

$$p_i(\vec{\theta}) = P\{X \in \Delta_i \mid H_0^*\} = \int_{y_i}^{y_{i+1}} dF_0(x;\vec{\theta}) = F_0(y_{i+1};\vec{\theta}) - F_0(y_i;\vec{\theta}),\ i = 1,...,k,$$

и зависят от неизвестного векторного параметра $\vec{\theta}$. По этой причине они не могут быть использованы при вычислении статистики хи-квадрат по формуле (3.5.10). Поэтому вместо статистики (3.5.10), которая в данном случае является функцией неизвестных параметров, то есть записывается в виде

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{k} [(n_i - np_i(\vec{\theta}))^2 / np_i(\vec{\theta})],$$

вычисляют модифицированную статистику, в которой неизвестные параметры заменены их оценками, то есть

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{k} [(n_i - np_i(\hat{\theta}))^2 / np_i(\hat{\theta})], \qquad \textbf{(3.5.16)}$$

где вероятности $p_i(\hat{\theta})$, $i = 1,...,k$, вычисляются с использованием предварительно найденных оценок $\hat{\theta}$ по формулам

$$p_i(\hat{\theta}) = P\{X \in \Delta_i \mid H_0^*\} = \int_{y_i}^{y_{i+1}} dF_0(x;\hat{\theta}) = F_0(y_{i+1};\hat{\theta}) - F_0(y_i;\hat{\theta}),$$
$$i = 1,...,k, \qquad \textbf{(3.5.17)}$$

а $\hat{\theta} = (\hat{\theta}_1,...,\hat{\theta}_r)$ – мультиномиальные оценки максимального правдоподобия вектора $\vec{\theta} = (\theta_1,...,\theta_r)$, построенные по данным интервального ряда частот (3.5.8). Эти оценки являются решением системы уравнений

$$\sum_{i=1}^{k} n_i \frac{\partial}{\partial \theta_j} \ln p_i(\theta_1,...\theta_r) = 0, \; j = 1,...,r. \qquad \textbf{(3.5.18)}$$

Отметим, что задача нахождения асимптотического распределения модифицированной статистики $\chi^2(\hat{\theta})$ при справедливости гипотезы H_0^* и при использовании мультиномиальных оценок максимального правдоподобия $\hat{\theta} = (\hat{\theta}_1,...,\hat{\theta}_r)$ вектора $\vec{\theta} = (\theta_1,...,\theta_r)$ была решена Фишером в 1924 году. Он показал, что в результате модификации статистики ее распределение остается распределением хи-квадрат, но с измененным числом степеней свободы, равным $(k-1-r)$, где k – число интервалов группировки и r – размерность вектора $\vec{\theta} = (\theta_1,...,\theta_r)$, компоненты которого оцениваются по данным интервального ряда частот. Для удобства дальнейших ссылок приведем результат Фишера в виде теоремы.

Теорема 3.5.19. Пусть функции $p_j(\vec{\theta})$, $j = 1,...,k$, $\vec{\theta} = (\theta_1,...,\theta_r)$, $r < k-1$, удовлетворяют следующим условиям:

1) $\sum_{j=1}^{n} p_j(\vec{\theta}) = 1,\ \forall \vec{\theta} \in \Theta$;

2) $p_j(\vec{\theta}) \geq c > 0,\ \ j = 1,...,k$, и существуют непрерывные производные $\dfrac{\partial p_j(\vec{\theta})}{\partial \theta_k}$ и $\dfrac{\partial^2 p_j(\vec{\theta})}{\partial \theta_k \partial \theta_l},\ k,l = 1,...,k$;

3) Матрица $\left\| \dfrac{\partial p_j(\vec{\theta})}{\partial \theta_k} \right\|$ размера $k \times r$ имеет ранг $r\ \ \forall \vec{\theta} \in \Theta$.

Тогда, если $\hat{\theta} = (\hat{\theta}_1,...,\hat{\theta}_r)$ – мультиномиальная оценка максимального правдоподобия параметра $\vec{\theta} = (\theta_1,...,\theta_r)$, являющаяся решением системы уравнений (3.5.18), и статистика $\hat{\chi}^2 = \chi^2(\hat{\theta})$ вычислена по формуле (3.5.16), то асимптотическое распределение модифицированной статистики $\hat{\chi}^2$ при гипотезе H_0^* является распределением хи-квадрат с числом степеней свободы $(k-r-1)$, то есть $L(\hat{\chi}^2 \mid H_0) = \chi^2(k-r-1)$ при $n \to \infty$.

Доказательство этой теоремы может быть найдено в [36, с. 462–470].

Итак, согласно теореме 3.5.19, модифицированная статистика $\hat{\chi}^2$ при справедливости гипотезы H_0^* имеет асимптотическое распределение хи-квадрат с числом степеней свободы $(k-1-r)$, то есть справедливо следующее выражение:

$$L(\hat{\chi}^2 \mid H_0^*) = \chi^2(k-1-r). \qquad \textbf{(3.5.20)}$$

С учетом выражения (3.5.20), $\hat{\chi}^2$-критерий для проверки сложной гипотезы H_0^* записывается в виде следующего правила:

«*Если* $\hat{\chi}^2 \geq \chi^2_{1-\alpha,(k-1-r)}$, *то гипотеза* H_0^* *отвергается на уровне значимости* α, *если* $\hat{\chi}^2 < \chi^2_{1-\alpha,(k-1-r)}$, *то нет оснований отвергать гипотезу* H_0^* ». **(3.5.21)**

Подведем итог данного раздела, отметив еще раз, что χ^2-критерий является универсальным критерием, его можно применять как в непрерывном, так и в дискретном случаях и его достоинством является простая модификация при проверке сложной гипотезы H_0^*. Этот критерий является *асимптотически непараметрическим*. Существенным недостатком этого критерия является его зависимость от неоднозначной процедуры, связанной со способом разбиения области возможных значений непрерывной с.в. X на интервалы группировки $\Delta_1,...,\Delta_k$. Эта зависимость проявляется в том, что распределение вероятностей статистики χ^2 при гипотезе H_0 определяется числом k *выбираемых нами* интервалов группировки. При построении интервалов группировки на практике обычно используют следующую рекомендацию: *Если среди построенных интервалов имеются «пустые»*, *точнее*, *если не выполняется неравенство* $\min(n_1,...,n_k) \geq 5$, *то интервалы группировки укрупняют*. При соблюдении этой рекомендации и при объемах выборки $n \geq 50$, χ^2-критерий гарантирует заданный уровень значимости α, точнее, при его применении вероятность отвергнуть нулевую гипотезу, когда она верна, приближенно равняется величине α. Хи-квадрат-критерий является *состоятельным* для широкого класса альтернатив и *асимптотически несмещенным* критерием (см., например, [36, 66]).

Замечание 3.5.22. При построении критериев согласия проверки сложной гипотезы статистика хи-квадрат вида (3.5.10) зависит от неизвестного параметра θ. Модификация этой статистики $\chi^2(\theta)$ для проверки сложной гипотезы осуществляется путем замены неизвестного параметра θ на его оценку $\hat{\theta}$. При этом могут быть использованы разные способы оценивания параметров. Распределение вероятностей модифицированной статистики $\chi^2(\hat{\theta})$ зависит от используемой оценки $\hat{\theta}$. Теорема Фишера 3.5.19 утверждает, что при использовании мультиномиальных оценок максимального правдоподобия, являющихся решением системы уравнений (3.5.18), распределение вероятностей модифицированной статистики $\chi^2(\hat{\theta})$ при гипотезе H_0^* остается распределением хи-квадрат и в результате модификации уменьшается лишь число степеней свободы на величину, равную числу оцениваемых параметров. В непрерывном случае возникает естественное желание использовать для мо-

дификации хи-квадрат-статистики обычные оценки $\hat{\theta}^*(X_1,...,X_n)$ максимального правдоподобия, основанные не на частотах $n_1,...,n_k$ интервального ряда (3.5.9), а на исходной выборке $X_1,...,X_n$. Чернов и Леман в 1954 году исследовали асимптотическое распределение таким образом модифицированных статистик $\chi^2(\hat{\theta}^*)$ и показали, что при такой модификации выражение (3.5.20) не имеет места (см. также работу [79]). Тем не менее на практике, когда решение системы уравнений (3.5.18) вызывает трудности и объем выборки достаточно большой, часто используют для модификации хи-квадрат-статистики обычные оценки $\hat{\theta}^*(X_1,...,X_n)$ максимального правдоподобия (см. упр. 3.9.69 и 3.9.78). Эти оценки асимптотически эффективные, и их обычно проще вычислить, чем решить систему уравнений (3.5.18). Отметим также, что существует модификация статистики хи-квадрат в форме Большева – Джапаридзе – Никулина, в которую можно подставлять любую состоятельную оценку параметра.

Пример 3.5.23. Рассмотрим проверку гипотезы о нормальности распределения, которая записывается в виде $H_0^*: F_X(x;\theta_1,\theta_2) = \Phi\{(x-\theta_1)/\theta_2\}$, где Φ – стандартная нормальная функция распределения и параметр $\theta_1 = M(X)$, а параметр $\theta_2 = \sqrt{D(X)}$. Это сложная гипотеза согласия с нормальным распределением, заданная с точностью до неизвестного векторного параметра $\vec{\theta} = (\theta_1, \theta_2)$. Оценками максимального правдоподобия параметров θ_1 и θ_2, основанными на исходной выборке $X_1,...,X_n$, являются соответственно выборочное среднее $\overline{X}$ и выборочное стандартное отклонение S. Модифицированная статистика хи-квадрат, вычисленная по данным интервального ряда частот вида (3.5.9) с использованием оценки максимального правдоподобия $\hat{\theta}^* = (\overline{X}, S)$, в данном случае запишется в виде (3.5.16), то есть

$$\chi^2(\hat{\theta}^*) = \sum_{i=1}^{k} [(n_i - np_i(\hat{\theta}^*))^2 / np_i(\hat{\theta}^*)], \qquad \textbf{(3.5.24)}$$

где $p_i(\hat{\theta}^*) = P\{X \in \Delta_i \mid H_0^*\} = \Phi\{(y_{i+1} - \overline{X})/S\} - \Phi\{(y_i - \overline{X})/S\}$, $i = 1,...,k$,

и число k интервалов группировки $\Delta_1,...,\Delta_k$, $\Delta_i = y_{i+1} - y_i$, $i = 1,...,k$, выбирается с учетом правила «трех сигма», то есть выбрав за основу интервал $[\overline{X} - 3S, \overline{X} + 3S]$, носитель распределения делится на такое число k равных интервалов группировки, при котором выполняется неравенство $\min(n_1,...,n_k) \geq 5$. Отметим, что если с.в. $\chi^2(\nu)$ имеет хи-квадрат-распределение с ν степенями свободы и ν достаточно велико (например, $\nu > 30$), то

распределение с.в. $Z = \sqrt{2\chi^2(\nu)} - \sqrt{2\nu - 1}$ можно аппроксимировать стандартным нормальным распределением [47, с. 99]. Таким образом, при больших значениях k и при $n \to \infty$ справедливо выражение

$$L\{\sqrt{2\chi^2(\theta^*)} - \sqrt{2(k-3)-1} \mid H_0^*\} \approx N(0\,,1)\,.$$

С учетом этого выражения, при больших объемах выборки гипотеза H_0^* о нормальности распределения отвергается на уровне значимости α, если вычисленное значение статистики $\sqrt{2\chi^2(\theta^*)} - \sqrt{2(k-3)-1}$ превышает ее критическое значение $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$. В противном случае нет оснований отвергать гипотезу о нормальности распределения.

Приведем числовой пример. Используя критерий хи-квадрат, проверить гипотезу о том, что выборка $X_1,...,X_n$ объема $n = 112$, представленная наблюденными значениями $\overline{x} = 2,45$, $s^2 = 1,74$ и интервальным вариационным рядом частот (см. первые два столбца табл. 3.5.1), взята из нормальной совокупности, то есть кратко сложная гипотеза согласия записывается в виде $H_0 : L(X) = N(\mu, \sigma^2)$, параметры μ и σ^2 неизвестны. Для удобства в табл. 3.5.1 приведены также вычисленные ожидаемые частоты при гипотезе H_0, слагаемые для статистики хи-квадрат и ее наблюденное значение.

Таблица 3.5.1

Интервалы группировки $\{\Delta_i\}$	Наблюденные частоты $\{n_i\}$	Ожидаемые частоты $\{np_i\}$	$\dfrac{(n_i - np_i)^2}{np_i}$
$(-\infty, 0,5)$	1	7,85	5,98
[0,5, 1,0)	9	7,38	0,36
[1,0, 1,5)	23	11,20	12,43
[1,5, 2,0)	17	14,67	0,37
[2,0, 2,5)	13	16,58	0,77
[2,5, 3,0)	12	16,46	1,21
[3,0, 3,5)	10	14,00	1,14
[3,5, 4,0)	9	10,42	0,19
[4,0, 4,5)	9	6,61	0,86
[4,5, 5,0)	3	3,81	0,17
$[5,0, +\infty)$	6	3,02	2,94
Σ	112		$\hat{\chi}^2 = 26,4$

Решение. Для проверки сложной гипотезы нормальности модифицированная статистика хи-квадрат, определенная в (3.5.16), записывается в виде

$$\chi^2(\hat{\theta}^*) = \sum_{i=1}^{k} [(n_i - np_i(\hat{\theta}^*))^2 / np_i(\hat{\theta}^*)],$$

где $\{n_i\}$ – наблюденные частоты, а $\{np_i(\hat{\theta}^*)\}$ – ожидаемые гипотетические частоты и $\hat{\theta}^* = (\hat{\mu}, \hat{\sigma})$, $\hat{\mu} = \overline{X}$ и $\hat{\sigma}^2 = S^2$ – оценки максимального правдоподобия неизвестных параметров μ и σ^2. Для вычисления ожидаемых частот, вероятности $p_i(\hat{\theta}^*)$ в данном примере находятся по формулам

$$p_i(\hat{\theta}^*) = P\{X \in \Delta_i \mid H_0^*\} = \Phi\{(y_{i+1} - \overline{x})/s\} - \Phi\{(y_i - \overline{x})/s\},\ i = 1,...,k,$$

где k – число интервалов группировки (в данном примере $k = 11$). Например, вероятность попадания случайной величины X в первый интервал группировки $\Delta_1 = [-\infty, 0,5)$

$$p_1(\hat{\theta}^*) = P\{X \in \Delta_1 \mid H_0^*\} = \Phi\{(0,5 - 2,45)/\sqrt{1,74}\} = \Phi\{-1,478\} =$$
$$= 1 - \Phi\{1,478\} = 1 - 0,9299 = 0,0701,$$

во второй интервал группировки $\Delta_2 = [0,5, 1,0)$ – вычисляется аналогично:

$$p_2(\hat{\theta}^*) = P\{X \in \Delta_2 \mid H_0^*\} = \Phi\{(1,00 - 2,45)/\sqrt{1,74}\} - \Phi\{(0,5 - 2,45)/\sqrt{1,74}\} = 0,0659.$$

Вычисленные ожидаемые частоты приведены в третьем столбце таблицы. С использованием этих данных, наблюденное значение статистики хи-квадрат вычисляется по формуле

$$\chi^2(\hat{\theta}^*) = \sum_{i=1}^{k} [(n_i - np_i(\hat{\theta}^*))^2 / np_i(\hat{\theta}^*)] = \frac{(1-7,85)^2}{7,85} + ... + \frac{(6-3,02)^2}{3,02} = 5,98 + ... + 2,94 = 26,4$$

и приведено внизу четвертого столбца таблицы. Далее, так как число интервалов группировки $k = 11$ и по выборке оценивалось два неизвестных параметра μ и σ^2, то число степеней свободы в данном примере равно $(k - 1 - 2) = 11 - 3 = 8$. Используя таблицы интеграла вероятностей χ^2 в работе [6, табл. 2.1а], находим $\Pr\{\chi^2(8) > 26,4\} < 0,001$. Таким образом, *P-значение* $< 0,001$, из чего следует, что гипотеза о нормальности отвергается.

Отметим, что нормальное распределение наиболее часто используется на практике. По этой причине разработаны специальные критерии согласия, предназначенные для проверки гипотезы о нормальности распределения вероятностей наблюдаемых случайных величин $X_1,...,X_n$. Эти критерии кратко

называют критериями нормальности. Наиболее популярный критерий нормальности Шапиро – Уилка обсуждается в разд. 4.7. Сравнение функций мощности этого критерия с другими критериями согласия для широкого класса альтернатив приводится в работе [S.S. Shapiro, M.B. Wilk and Mrs. H.J. Chen. A comparative study of various tests for normality // J. Amer. Statist. Assoc. 1968. V. 63. No. 324. P. 1343–1372].

3.5.Г. Критерий квантилей

Пусть $\Im$ – класс непрерывных функций распределения вероятностей. Имеется выборка $\vec{X} = (X_1,...,X_n)$ из совокупности с непрерывной функцией распределения $F_X(x) \in \Im$. Пусть далее, для некоторого распределения $F_0(x) \in \Im$ определены заданное число $(m-1)$ квантилей $x_{i/m}$ уровней $i/m = p_i$, $i = 1,..., m-1$, то есть $F_0(x_{i/m}) = p_i$, $i = 1,..., m-1$, причем $-\infty < x_{1/m} < ... < x_{(m-1)/m} < +\infty$ и $0 < p_1 < ... < p_{m-1} < 1$. Обозначим через $\Im_0$ подкласс $\Im$, состоящий из распределений с указанными выше квантилями, то есть

$$\Im_0 = \{ F : F(x_{i/m}) = p_i,\ i = 1,...,m-1 \} \text{ и } F_0(x) \in \Im_0. \qquad \textbf{(3.5.25)}$$

Требуется по выборке $\vec{X} = (X_1,...,X_n)$ проверить *сложную* гипотезу H_0, согласно которой неизвестная нам функция распределения F_X имеет заданные выше квантили $x_{i/m}$ уровней $i/m = p_i$, $i = 1,..., m-1$, то есть кратко нулевая гипотеза записывается в виде $H_0 : F_X(x) \in \Im_0$. Эта гипотеза проверяется против альтернативы $H_1 : F_X(x) \in (\Im \setminus \Im_0)$.

Подчеркнем еще раз, что сформулированная гипотеза H_0 является *сложной и непараметрической*, она включает в себя все непрерывные распределения с указанными выше квантилями $x_{i/m}$ уровней $i/m = p_i$, $i = 1,..., m-1$. Для построения критерия проверки гипотез H_0, H_1 на основе хи-квадрат-критерия, прежде всего, нужно конкретизировать интервалы группировки. Эти интервалы определяют *естественным образом*, *порождаемым самой природой проверяемой гипотезы* H_0, в виде равновероятных интервалов

$I_i = (x_{(i-1)/m}, x_{i/m}]$, $i = 1,\dots, m$, $x_0 = -\infty$, $x_1 = +\infty$. Пусть далее, n_i – число элементов выборки $\vec{X} = (X_1,\dots,X_n)$, попавших в интервал I_i, $i = 1,\dots, m$. Если выборка извлечена из совокупности с распределением $F_X(x) \in \Im_0$ (то есть верна гипотеза H_0), то частоты $(n_1,\dots,n_m)$ образуют многомерную случайную величину, описываемую полиномиальным распределением

$$P\{(n_1,\dots,n_m) \mid H_0\} = \frac{n!}{n_1!\cdots n_m!}\left(\frac{1}{m}\right)^n,\ n_1 + \dots + n_m = n.$$

Таким образом, сформулированная гипотеза $H_0 : F_X(x) \in \Im_0$ эквивалентна утверждению, что вероятности исходов в полиномиальной схеме, построенной с помощью интервалов группировки $I_1,\dots,I_m$, равны заданным числам $p_{10},\dots,p_{m0}$, где $p_{i0} = p_i - p_{i-1} = 1/m$, $i = 1,\dots, m$, $p_0 = 0$, $p_m = 1$. С учетом этого замечания, для проверки гипотез H_0, H_1 может быть применен критерий хи-квадрат, основанный на статистике (3.5.10), которая в данном случае запишется в виде

$$\chi^2 = \frac{m}{n}\sum_{i=1}^{m}\left(n_i - \frac{n}{m}\right)^2. \qquad \textbf{(3.5.26)}$$

При больших объемах выборки n критическая область размера α для критерия проверки гипотез H_0, H_1 записывается в виде $\mathrm{X}_{1\alpha} = \{\vec{x} : \chi^2 > \chi^2_{1-\alpha,\,(m-1)}\}$. Этот критерий называют в литературе (см., например, [66, с.435]) *критерием квантилей*. Факт состоятельности критерия квантилей для любой альтернативы из класса $(\Im \setminus \Im_0)$ доказывается в [66, с. 439].

Отметим, что при построении критерия квантилей используется лишь факт равной вероятности попадания в интервалы группировки $I_1,\dots,I_m$ и не учитывается локальное поведение функции распределения в этих интервалах. По этой причине он не может различать все непрерывные распределения из класса $\Im_0$ с заданными квантилями и *оказывается несостоятельным для любой альтернативы из класса* $\Im_0$ (см., например, [66, с. 440]). Для устранения этого недостатка в литературе были предприняты попытки к «уменьшению»

класса $\Im_0$ неразличимых распределений с помощью критерия квантилей путем одновременного увеличения как числа наблюдений n, так и числа интервалов группировки m (число заданных квантилей $m-1$ большое, но остается конечным), однако при этом возникают серьезные и аналитические, и вычислительные трудности (см. [22, с. 119]). Некоторое продвижение в решении этих вопросов было достигнуто при изучении свойств симметрических статистик в схеме группировки, которые в частном случае приводят к *критерию пустых ящиков*, предложенному Дэвидом в 1950 году (см. [66, с. 440] и упр. 5.8.1).

Замечание 3.5.27. Отметим, что при числе интервалов $m=2$ и одном заданном квантиле $x_{1/2}$ уровня $p_1=1/2$, который называют медианой, *критерий квантилей* превращается в *критерий знаков* (см. также разд. 5.1). В этом случае гипотеза H_0 предполагает, что функция распределения $F_X(x)$, из которой взята выборка $\vec{X}=(X_1,...,X_n)$, имеет заданную медиану $x_{1/2}$. С учетом того, что $n_1+n_2=n$, статистика (3.5.26) в данном случае запишется в виде

$$\chi^2=\frac{2}{n}\left(n_1-\frac{n}{2}\right)^2+\frac{2}{n}\left(n-n_1-\frac{n}{2}\right)^2=\frac{4}{n}\left(n_1-\frac{n}{2}\right)^2, \qquad \textbf{(3.5.28)}$$

где n_1, n_2 – частоты попаданий элементов выборки соответственно в интервалы группировки $(-\infty, x_{1/2}]$ и $[x_{1/2}, +\infty)$. Итак, статистика критерия знаков зависит лишь от числа n_1, которое равняется числу отрицательных разностей $X_i-x_{1/2}$, $i=1,...,n$, этим и объясняется название критерия. Критерий знаков будет дополнительно обсуждаться в разделе (5.1), здесь же попутно отметим, что его часто используют для проверки гипотезы о том, что статистические данные $(X_1,Y_1),...,(X_n,Y_n)$ являются выборкой из некоторого двумерного распределения $F_{XY}(x,y)$, то есть предполагается, что они независимы и одинаково распределены с функцией распределения $F_{XY}(x,y)$. Эта гипотеза кратко записывается в виде $H_0: F_{XY}(x,y)=F(x)\cdot F(y)$, где $F(x)$ – некоторая одномерная функция распределения. Проверка сформулированной гипотезы с помощью критерия знаков (или критерия квантилей) осуществляется следующим образом. Обозначим $Z_i=X_i-Y_i$, $i=1,...,n$. При справедливости гипотезы H_0 выполняются выражения $P(Z_i<0)=P(Z_i>0)=1/2$, $\forall i=1,...,n$, и, следовательно, эта гипотеза эквивалентна утверждению о том, что выборка $Z_1,...,Z_n$ извлечена из совокупности с распределением $F_Z(z)$, которое имеет нулевую медиану, то есть $z_{1/2}=0$. Для проверки этой гипотезы может быть использован критерий, основанный на статистике (3.5.28), в которой число n_1 равняется числу отрицательных величин среди $Z_1,...,Z_n$. Критическая область

размера α этого критерия при больших значениях n записывается в виде $\mathrm{X}_{1\alpha} = \{\vec{z} : (4(n_1 - n/2)^2 / n) \geq \chi^2_{1-\alpha,(1)}\}$, так как справедливо выражение $L\{4(n_1 - n/2)^2 / n \mid H_0\} = \chi^2(1)$, $n \to \infty$ (см., например, [66]).

3.6. Критерии однородности

Задача однородности, сформулированная в примере (2.3.18), относится к основным статистическим задачам, которые часто возникают на практике, когда требуется проверить гипотезу об однородности статистических данных, полученных в разное время или в разных условиях при изучении одного и того же случайного явления. Типичными примерами являются задачи сравнения некоторого нового типа обработки по сравнению со стандартным. Такие задачи возникают, например, при внедрении нового технологического процесса, при оценке эффективности нового метода лечения, при использовании разных измерительных приборов, при сравнении качества продукции, выпускаемой на разных предприятиях и т.д. Задачу однородности решают в различных вариантах. В простейшем случае статистические данные представлены двумя выборками (независимыми или связанными), в общих случаях они представлены заданным числом k выборок, причем задача однородности может быть сформулирована в рамках как параметрических, так и непараметрических моделей.

В простейшем (двухвыборочном) варианте задача однородности решается на основе статистических данных, представленных двумя выборками, выборкой $X_1,...,X_{n_1}$ из распределения F_X и выборкой $Y_1,...,Y_{n_2}$ из распределения F_Y. По этим статистическим данным требуется проверить гипотезу H_0, согласно которой распределения F_X и F_Y одинаковы, то есть

$$H_0 : F_X = F_Y . \qquad \textbf{(3.6.1)}$$

Отметим, что эта гипотеза *непараметрическая*, так как при ее формулировке предполагается лишь равенство распределений, а функциональный вид распределений F_X и F_Y *не конкретизируется* и обычно лишь предполагается, что они непрерывны и выборки

$X_1,...,X_{n_1}$, $Y_1,...,Y_{n_2}$ независимы. Данная гипотеза может проверяться против различных альтернатив. Например, против односторонних альтернатив, утверждающих, что с.в. X «стохастически больше (меньше)» с.в. Y, то есть $H_1^+ : F_Y \geq F_X$, ($H_1^- : F_Y \leq F_X$), либо против двусторонней альтернативы $H_1 : F_X \neq F_Y$. Отметим также, что на практике наиболее часто используют альтернативы сдвига $H_1^* : F_Y(x) = F_X(x-\theta)$, $\theta \in R^1$, или альтернативы масштаба $H_1^{**} : F_Y(x) = F_X(x/\tau)$, $\tau \geq 1$.

Двухвыборочная задача однородности естественным образом обобщается на заданное число k выборок, она называется k-выборочной задачей однородности и формулируется следующим образом. Имеется k выборок $X_{1j},...,X_{n_j j}$ из распределений F_j, $j = 1,...,k$. Требуется проверить гипотезу

$$H_0 : \ F_1 = F_2 = \cdots = F_k \qquad \textbf{(3.6.2)}$$

против альтернативы, согласно которой не все равенства в (3.6.2) выполняются.

Задача однородности, сформулированная в рамках параметрической модели, наиболее часто на практике применяется для проверки гипотез относительно параметров нормальных распределений

$$F_X(x;\vec{\theta}_1) = \Phi\{(x-a_1)/\sigma_1\}, \vec{\theta}_1 = (a_1,\sigma_1)$$

и

$$F_Y(x;\vec{\theta}_2) = \Phi\{(x-a_2)/\sigma_2\}, \vec{\theta}_2 = (a_2,\sigma_2).$$

Мы обсудим сначала двухвыборочные и затем k-выборочные критерии для проверки гипотез относительно параметров нормальной модели. Критерии однородности для непараметрических моделей будут рассмотрены ниже в разд. 4.7.

3.6.А. Двухвыборочный критерий отношения дисперсий

Пусть случайные величины X и Y имеют нормальные распределения вероятностей, то есть $L(X) = N(a_1;\sigma_1^2)$ и $L(Y) = N(a_2;\sigma_2^2)$. Имеются две независимые выборки из этих распределений $X_1,...,X_{n_1}$ и $Y_1,...,Y_{n_2}$. Параметры a_j и σ_j^2, $j = 1,2$, неизвестны. Тре-

буется проверить гипотезу о равенстве дисперсий, то есть

$$H_0 : \sigma_1^2 = \sigma_2^2, \tag{3.6.3}$$

против односторонних альтернатив

$$H_1^+ : \sigma_1^2 > \sigma_2^2 \quad \text{или} \quad H_1^- : \sigma_1^2 < \sigma_2^2 \tag{3.6.4}$$

либо против двусторонней альтернативы

$$H_1 : \sigma_1^2 \neq \sigma_2^2. \tag{3.6.5}$$

Напомним, что несмещенные оценки параметров σ_j^2, $j = 1,2$, построенные соответственно по выборкам $X_1,...,X_{n_1}$ и $Y_1,...,Y_{n_2}$, записываются в виде

$$\tilde{S}_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2,\ \tilde{S}_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2,$$

причем, согласно теореме (2.8.31), выполняются выражения

$$L\{\nu_j \tilde{S}_j^2 / \sigma_j^2\} = \chi^2(\nu_j),\ \nu_j = n_j - 1,\ j = 1,2. \tag{3.6.6}$$

Рассмотрим случайную величину, которую часто называют *дисперсионным отношением*, в виде

$$F = \tilde{S}_1^2 / \tilde{S}_2^2. \tag{3.6.7}$$

Согласно теореме (2.8.36), случайная величина $F = \tilde{S}_1^2 / \tilde{S}_2^2$ при гипотезе $H_0 : \sigma_1^2 = \sigma_2^2$ имеет F-распределение Фишера с числами степеней свободы ν_1 и ν_2, то есть справедливо выражение

$$L\{\tilde{S}_1^2 / \tilde{S}_2^2 \mid H_0\} = F(\nu_1, \nu_2). \tag{3.6.8}$$

Отметим, что

$$\frac{\tilde{S}_1^2}{\tilde{S}_2^2} \sim \frac{\sigma_1^2 \chi_2(\nu_1)/\nu_1}{\sigma_2^2 \chi^2(\nu_2)/\nu_2} = \frac{\sigma_1^2}{\sigma_2^2} F(\nu_1, \nu_2),$$

или отношение $\tilde{S}_1^2 / \tilde{S}_2^2$ эквивалентно $(\sigma_{1.}^2 / \sigma_2^2) F(\nu_1, \nu_2)$, то есть

$$\tilde{S}_1^2 / \tilde{S}_2^2 \sim (\sigma_{1.}^2 / \sigma_2^2) F(\nu_1, \nu_2). \tag{3.6.9}$$

Выражение (3.6.8) позволяет построить критерий проверки гипотез $H_0:\sigma_1^2=\sigma_2^2,\ H_1:\sigma_1^2\neq\sigma_2^2$. В самом деле, для конкретизации статистики этого критерия следует получить выражение для статистики отношения правдоподобия. Можно убедиться (см., например, [9]), что статистика отношения правдоподобия $\Lambda^*(\vec{X};\Theta_0)$, определенная в (3.4.18), в данном случае будет функцией отношения $\tilde{S}_1^2/\tilde{S}_2^2$ и будет принимать «маленькие» значения в двух случаях, когда отношения $\tilde{S}_1^2/\tilde{S}_2^2$ будут малы или велики.

Итак, выбрав в качестве тестовой статистики отношение $\tilde{S}_1^2/\tilde{S}_2^2$, и используя выражение (3.6.8.), получаем критерий уровня значимости α проверки гипотез $H_0:\sigma_1^2=\sigma_2^2,\ H_1^+:\sigma_1^2>\sigma_2^2$ с критической областью вида

$$\mathrm{X}_{1\alpha}^{+}=\{(\vec{x},\vec{y}):(\tilde{s}_1^2/\tilde{s}_2^2)\geq F_{1-\alpha}(\nu_1,\nu_2)\}. \qquad \textbf{(3.6.10)}$$

Для проверки гипотез $H_0:\sigma_1^2=\sigma_2^2,\ H_1^-:\sigma_1^2<\sigma_2^2$ критическая область записывается в виде

$$\mathrm{X}_{1\alpha}^{-}=\{(\vec{x},\vec{y}):(\tilde{s}_1^2/\tilde{s}_2^2)\leq F_{\alpha}(\nu_1,\nu_2)\}, \qquad \textbf{(3.6.11)}$$

где $F_q(\nu_1,\nu_2)$ – квантиль уровня q, $0<q<1$, F-распределения Фишера с числами степеней свободы ν_1,ν_2 и $(\tilde{s}_1^2/\tilde{s}_2^2)$ – наблюденное значение статистики $\tilde{S}_1^2/\tilde{S}_2^2$, вычисленное по наблюденным реализациям $x_1,...,x_{n_1}$ и $y_1,...,y_{n_2}$ выборок $X_1,...,X_{n_1}$ и $Y_1,...,Y_{n_2}$.

Отметим (см. упр. 3.9.60), что нижние и верхние квантили F-распределения Фишера связаны выражением

$$F_q(\nu_1,\nu_2)=1/F_{1-q}(\nu_2,\nu_1),\ 0<q<1, \qquad \textbf{(3.6.12)}$$

поэтому, используя только верхние критические точки F-распределения Фишера, критическую область $\mathrm{X}_{1\alpha}^{-}$ можно представить в эквивалентной форме, то есть в виде

$$\mathrm{X}_{1\alpha}^{-}=\{(\vec{x},\vec{y}):(\tilde{s}_1^2/\tilde{s}_2^2)\leq F_{\alpha}(\nu_1,\nu_2)\}\Leftrightarrow\mathrm{X}_{1\alpha}^{+}=\{(\vec{x},\vec{y}):(\tilde{s}_2^2/\tilde{s}_1^2)\geq F_{1-\alpha}(\nu_2,\nu_1)\}.$$

Объединяя критические области односторонних критериев, согласно правилу (3.4.37), получаем для проверки гипотез $H_0 : \sigma_1^2 = \sigma_2^2,\ H_1 : \sigma_1^2 \neq \sigma_2^2$ двусторонний «симметричный» критерий уровня значимости α, критическая область которого определяется в виде

$$X_{1\alpha} = \{(\vec{x},\vec{y}) : (\tilde{s}_1^2 / \tilde{s}_2^2) \le F_{\alpha/2}(\nu_1,\nu_2) \text{ или } (\tilde{s}_1^2 / \tilde{s}_2^2) \ge F_{1-\alpha/2}(\nu_1,\nu_2)\}. \quad \textbf{(3.6.13)}$$

Эту критическую область можно записать только через «верхние» квантили *F*-распределения. В самом деле, используя (3.6.12), получаем

$$X_{1\alpha} = \{(\vec{x},\vec{y}) : (\tilde{s}_2^2 / \tilde{s}_1^2) \ge F_{1-\alpha/2}(\nu_2,\nu_1) \text{ или } (\tilde{s}_1^2 / \tilde{s}_2^2) \ge F_{1-\alpha/2}(\nu_1,\nu_2)\}. \textbf{(3.6.14)}$$

Используя (3.6.9) и выбрав «симметричные по вероятности» квантили $F_{\alpha/2}(\nu_1,\nu_2)$ и $F_{1-\alpha/2}(\nu_1,\nu_2)$ для *F*-распределения, получим выражение

$$P\left\{\frac{\sigma_1^2}{\sigma_2^2} F_{\alpha/2}(\nu_1,\nu_2) < \frac{\tilde{S}_1^2}{\tilde{S}_2^2} < \frac{\sigma_1^2}{\sigma_2^2} F_{1-\alpha/2}(\nu_1,\nu_2)\right\} = 1-\alpha.$$

Отсюда имеем $(1-\alpha)$-доверительный интервал для отношения дисперсий σ_1^2 / σ_2^2 в виде

$$P\left\{\frac{\tilde{S}_1^2}{\tilde{S}_2^2} \cdot \frac{1}{F_{1-\alpha/2}(\nu_1,\nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\tilde{S}_1^2}{\tilde{S}_2^2} \cdot \frac{1}{F_{\alpha/2}(\nu_1,\nu_2)}\right\} = 1-\alpha \quad \textbf{(3.6.15)}$$

или, учитывая (3.6.12), $(1-\alpha)$-доверительный интервал для отношения дисперсий σ_1^2 / σ_2^2, выраженный только через верхние квантили *F*-распределения, в виде

$$P\left\{(\tilde{S}_1^2 / \tilde{S}_2^2) / F_{1-\alpha/2}(\nu_1,\nu_2) < (\sigma_1^2 / \sigma_2^2) < (\tilde{S}_1^2 / \tilde{S}_2^2) F_{1-\alpha/2}(\nu_2,\nu_1)\right\} = 1-\alpha. \quad \textbf{(3.6.16)}$$

Мощность критерия с критической областью $X_{1\alpha}^+$ вида (3.6.10), при альтернативе $H_1^+ : \sigma_1^2 > \sigma_2^2$, для заданного значения отношения $(\sigma_1^2 / \sigma_2^2) = \tau$ равна

$$W(X_{1\alpha}^+;\tau) = P\left\{\frac{\tilde{S}_1^2}{\tilde{S}_2^2} \ge F_{1-\alpha}(\nu_1,\nu_2) \,|\, H_1^+\right\} = P\left\{\frac{\tilde{S}_1^2 / \sigma_1^2}{\tilde{S}_2^2 / \sigma_2^2} \ge \tau^{-1} F_{1-\alpha}(\nu_1,\nu_2) \,|\, H_1^+\right\}.$$

Учитывая (3.6.9), получаем выражение для мощности критерия в виде

$$W(\mathrm{X}^{+}_{1\alpha};\tau)=P\left\{F(\nu_1,\nu_2)\geq\tau^{-1}F_{1-\alpha}(\nu_1,\nu_2)\right\}. \qquad \textbf{(3.6.17)}$$

Вычисление мощности по формуле (3.6.17), а также вычисление необходимых объемов выборки n_1 и n_2, при которых обеспечивается заданная мощность, приводятся в упр. 3.9.61.

3.6.Б. Критерий равенства дисперсий

Пусть имеется заданное число k выборок $\vec{X}_j=(X_{1j},...,X_{n_j j})$, $j=1,...,k$, из нормальных распределений $F_X(x,\vec{\theta}_j)=\Phi\{(x-a_j)/\sigma_j\}$, где $\vec{\theta}_j=(a_j,\sigma_j)$ – вектор неизвестных параметров a_j и σ_j, $j=1,...,k$, и $n=n_1+\cdots+n_k$ – общее число наблюдений. Требуется проверить гипотезу H_0 о том, что дисперсии нормальных распределений равны, то есть гипотеза записывается в виде

$$H_0:\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2 \qquad \textbf{(3.6.18)}$$

против альтернативы H_1, согласно которой не все равенства гипотезы H_0 выполняются.

Для проверки этих сложных гипотез H_0, H_1 построим критерий отношения правдоподобия. По предположению, плотность случайных величин X_{ij}, $i=1,...,n_j$, $j=1,...,k$, имеет вид

$$f_X(x,\vec{\theta}_j)=\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left(-\frac{(x-a_j)^2}{2\sigma_j^2}\right),\ j=1,...,k.$$

Поэтому функция правдоподобия запишется как

$$L\{\{\vec{X}_1,...,\vec{X}_k\};\vec{\theta}_1,...,\vec{\theta}_k\}=$$

$$=\frac{1}{(2\pi)^{n/2}\sigma_1^{n_1}\cdots\sigma_k^{n_k}}\exp\left[-\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{X_{ij}-a_j}{\sigma_j}\right)^2\right], \qquad \textbf{(3.6.19)}$$

где $n = n_1 + \cdots + n_k$. Обозначим далее, как обычно, множество параметров через Θ, а множество параметров, соответствующих гипотезам H_0 и H_1, соответственно через Θ_0 и Θ_1, $\Theta = \Theta_0 \cup \Theta_1$. При построении статистики отношения правдоподобия найдем сначала наибольшее значение функции правдоподобия $L\{\cdot\}$ на параметрическом множестве Θ. Для этой цели удобнее использовать логарифм функции правдоподобия, который записывается в виде

$$\ln L\{\cdot\} = -\frac{n}{2}\ln 2\pi - n_1 \ln \sigma_1 - \cdots - n_k \ln \sigma_k - \frac{1}{2}\sum_{j=1}^{k}\frac{1}{\sigma_j^2}\sum_{i=1}^{n_j}(X_{ij} - a_j)^2 .$$

Отметим, что на множестве Θ нет никаких ограничений на параметры a_j и σ_j, кроме естественных $\sigma_j > 0$, $j = 1,...,k$, поэтому для нахождения наибольшего значения функции правдоподобия $L\{\cdot\}$ на параметрическом множестве Θ вычислим частные производные от $\ln L\{\cdot\}$ по параметрам a_j, σ_j и приравняем их нулю. В результате получим систему уравнений

$$\frac{\partial \ln L\{\cdot\}}{\partial a_j} = \frac{1}{\sigma_j^2}\sum_{i=1}^{n_j}(X_{ij} - a_j) = 0,\ j = 1,...,k,$$

$$\frac{\partial \ln L\{\cdot\}}{\partial \sigma_j} = -\frac{n_j}{\sigma_j} + \frac{1}{\sigma_j^3}\sum_{i=1}^{n_j}(X_{ij} - a_j)^2 = 0,\ j = 1,...,k.$$

Решая данную систему, найдем оценки максимального правдоподобия неизвестных параметров a_j и σ_j, $j = 1,...,k$, в виде

$$\hat{a}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = \overline{X}_{\bullet j},\quad j = 1,...,k,$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2,\quad j = 1,...,k.$$

Подставив найденные оценки в (3.6.19), получим наибольшее значение функции правдоподобия $L\{\cdot\}$ на параметрическом множестве Θ в виде

$$L\{\{\vec{X}_1,...,\vec{X}_k\};\Theta\} = \frac{1}{(2\pi)^{n/2}\hat{\sigma}_1^{n_1}\cdots\hat{\sigma}_k^{n_k}}\exp\left[-\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{X_{ij}-\overline{X}_{\bullet j}}{\hat{\sigma}_j}\right)^2\right] =$$

$$= -\frac{e^{-n/2}}{(2\pi)^{n/2}\hat{\sigma}_1^{n_1}\cdots\hat{\sigma}_k^{n_k}}.$$

Для вычисления наибольшего значения функции правдоподобия $L\{\cdot\}$ на параметрическом множестве Θ_0, определяемом нулевой гипотезой, продифференцируем $\ln L\{\cdot\}$ по параметрам a_j, σ и приравняем производные нулю. В результате получим систему уравнений

$$\frac{\partial \ln L\{\cdot\}}{\partial a_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n_j}(X_{ij}-a_j) = 0,\ j = 1,...,k,$$

$$\frac{\partial \ln L\{\cdot\}}{\partial \sigma} = -\frac{n}{\sigma} - \frac{1}{\sigma^2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-a_j)^2 = 0.$$

Решая данную систему, находим оценки максимального правдоподобия при нулевой гипотезе в виде

$$\hat{a}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}X_{ij} = \overline{X}_{\bullet j},\ j = 1,...,k,$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j})^2 = \frac{1}{n}\sum_{j=1}^{k}n_j\hat{\sigma}_j^2 = \sum_{j=1}^{k}n_j\hat{\sigma}_j^2 / \sum_{j=1}^{k}n_j.$$

Подставив найденные оценки в (3.6.19), получим наибольшее значение функции правдоподобия $L\{\cdot\}$ на параметрическом множестве Θ_0 в виде

$$L\{\{\vec{X}_1,...,\vec{X}_k\};\Theta_0\} =$$

$$= \left[(2\pi)^{n/2}\left(n^{-1}\sum_{j=1}^{k}n_j\hat{\sigma}_j^2\right)^{n/2}\right]^{-1}\exp\left[-\frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\frac{X_{ij}-\overline{X}_{\bullet j}}{\hat{\sigma}}\right)^2\right] =$$

$$= \frac{e^{-n/2}}{(2\pi)^{n/2}(n^{-1}\Sigma n_j \hat{\sigma}_j^2)^{n/2}}.$$

Таким образом, определенная в (3.4.18) статистика отношения правдоподобия

$$\Lambda^*\{\{\vec{X}_1,...,\vec{X}_k\};\Theta_0\} = L\{\{\vec{X}_1,...,\vec{X}_k\};\Theta_0\} / L\{\{\vec{X}_1,...,\vec{X}_k\};\Theta\},$$

в данном случае

$$\Lambda^*\{\{\vec{X}_1,...,\vec{X}_k\};\Theta_0\} = \frac{\hat{\sigma}_1^{n_1}\cdots\hat{\sigma}_k^{n_k}}{\left(n^{-1}\sum_{j=1}^{k} n_j\hat{\sigma}_j^2\right)^{n/2}} = \prod_{j=1}^{k}\hat{\sigma}_j^{n_j} / \hat{\sigma}^n .$$

Отсюда

$$-2\ln\Lambda^*\{\vec{X}_1,...,\vec{X}_k;\Theta_0\} = -2\sum_{j=1}^{k}\ln(\hat{\sigma}_j^{n_j} / \hat{\sigma}^n) =$$

$$= -\sum_{j=1}^{k} n_j \ln\hat{\sigma}_j^2 + n\ln\hat{\sigma}^2 = -\sum_{j=1}^{k} n_j \ln\{\hat{\sigma}_j^2 / \hat{\sigma}^2\}. \qquad \textbf{(3.6.20)}$$

Отметим, что при альтернативе H_1 имеется $2k$ неизвестных параметров $(a_1,...,a_k;\sigma_1,...,\sigma_k)$, а при гипотезе H_0 число неизвестных параметров $(a_1,...,a_k;\sigma)$ равняется $k+1$. Поэтому, согласно теореме Уилкса (см. [66] и разд. 3.4), при справедливости гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ статистика $-2\ln\Lambda^*$ при больших объемах выборки имеет распределение хи-квадрат с числом степеней свободы $k-1$, то есть справедливо выражение

$$L\{-2\ln\Lambda^* \mid H_0\} = \chi^2_{k-1} \text{ при } n\to\infty. \qquad \textbf{(3.6.21)}$$

Таким образом, при больших объемах выборки критерий размера α для проверки гипотез H_0 и H_1 основан на критической области вида

$$X_{1\alpha} = \{\{\vec{x}_1,...,\vec{x}_k\} : -2\ln\Lambda^*(\vec{x}_1,...,\vec{x}_k;\Theta_0) \ge \chi^2_{1-\alpha,(k-1)}\}. \qquad \textbf{(3.6.22)}$$

Замечание 3.6.23. Для проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ Бартлетт [Bartlett M.S. The effect of non-normality on the t-distribution // Proc. Cambridge Phil. Soc. 1935. V. 31. P. 223–231] предложил модифицированный вариант статистики отношения правдоподобия (3.6.20), в котором n_j заменяется на $n_j - 1$ и оценки максимального правдоподобия $\hat{\sigma}_j^2$, $j = 1,...,k$, и $\hat{\sigma}^2$ заменяются несмещенными оценками $\tilde{S}_j^2$ и S_W^2. В результате такой модификации статистика (3.6.20) запишется в виде

$$B\{\vec{X}_1,...,\vec{X}_k\} = -\sum_{j=1}^{k}(n_j - 1)\ln(\tilde{S}_j^2 / S_W^2) =$$

$$= (n-k)\ln S_W^2 - \{(n_1 - 1)\ln \tilde{S}_1^2 + \cdots + (n_k - 1)\ln \tilde{S}_k^2\}, \qquad \textbf{(3.6.23a)}$$

где $\tilde{S}_j^2$ – несмещенная оценка дисперсии σ_j^2, $j = 1,...,k$, определяемая выражением

$$\tilde{S}_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2, \; j = 1,...,k, \qquad \textbf{(3.6.24)}$$

и S_W^2 – несмещенная оценка дисперсии внутри k нормальных совокупностей (внутри k групп см. формулу (3.8.13)), определяемая в виде

$$S_W^2 = \frac{\sum_{j=1}^{k}(n_j - 1)\tilde{S}_j^2}{\sum_{j=1}^{k}(n_j - 1)} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2}{\sum_{j=1}^{k}(n_j - 1)} = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2. \qquad \textbf{(3.6.25)}$$

Отметим, что при $k = 2$ и $n_1 = n$, $n_2 = m$ несмещенная оценка дисперсии внутри групп S_W^2 совпадает с объединенной несмещенной оценкой $\tilde{S}_\Sigma^2$ дисперсии, построенной по двум выборкам $X_1,...X_n$, $Y_1,...Y_m$, определенной далее в (3.6.41), то есть

$$S_W^2 = \frac{(n-1)\tilde{S}_1^2 + (m-1)\tilde{S}_2^2}{n+m-2} = \tilde{S}_\Sigma^2.$$

Для повышения качества аппроксимации распределения статистики $B\{\vec{X}_1,...,\vec{X}_k\}$ при нулевой гипотезе распределением хи-квадрат Бартлетт предложил нормировать эту статистику на коэффициент C, вычисляемый по формуле

$$C = 1 + \{1/(n_1 - 1) + \cdots + 1/(n_k - 1) - 1/(n-k)\}/\{3(k-1)\}. \qquad \textbf{(3.6.26)}$$

С учетом этого замечания, критерий Бартлетта, основанный на статистике $B\{\vec{X}_1,...,\vec{X}_k\}/C$, определяется критической областью вида

$$\mathrm{X}_{1\alpha} = \{\,\{\vec{x}_1,...,\vec{x}_k\} : B\{\vec{x}_1,...,\vec{x}_k\}/C \geq \chi^2_{1-\alpha\,,\,(k-1)}\,\}\,. \qquad \textbf{(3.6.27)}$$

Отметим, что критерий Бартлетта *неустойчив* при отклонениях от нормальности, то есть не является робастным критерием к отклонениям от нормальной модели (см. разд. 6.2). Впервые на этот факт обратили внимание Бокс и Андерсон [137], которые показали, что если значение статистики $B\{\vec{X}_1,...,\vec{X}_k\}/C$ лежит в критическом множестве, то это могло произойти с одинаковым успехом как от нарушения предположения о нормальности, так и от того, что гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ не выполняется, то есть дисперсии различны.

Отметим также, что кроме критерия Бартлетта для проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ может быть использован критерий Хартли при условии, что объемы выборок равны, то есть $n_1 = n_2 = ... = n_k = n$. Статистика этого критерия записывается в виде отношения $\max\{S_1^2,...,S_k^2\}/\min\{S_1^2,...,S_k^2\}$. Таблицы критических значений этой статистики при уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$ приведены в [21, табл.151]. Входами в таблицу критических значений являются число групп k и число степеней свободы каждой групповой дисперсии, равное $(n-1)$. Если наблюденное значение статистики превышает критическое значение, то гипотеза H_0 о равенстве или однородности дисперсий отклоняется на заданном уровне значимости α и принимается альтернативная гипотеза, согласно которой $\sigma_i^2 \neq \sigma^2$ для некоторых значений индекса i. В тех случаях, когда одна из выборочных дисперсий $(S^2_{\max})$ значительно больше чем остальные (или когда число групп $k > 12$), для проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ против альтернативы $H_1 : \sigma^2_{\max} \neq \sigma^2$ используют критерий Кокрена. Статистика этого критерия записывается в виде отношения $S^2_{\max}/(S_1^2 + ... + S_k^2)$ и ее критические значения приводятся в [21, табл. 152; 44, с. 155]. Числовой пример проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ с использованием критерия Бартлетта приводится в упр. 3.9.94.

3.6.В. Двухвыборочные критерии проверки гипотезы о равенстве математических ожиданий нормальных распределений

Пусть $L(X) = N(a_1 ; \sigma_1^2)$ и $L(Y) = N(a_2 ; \sigma_2^2)$. Имеются две независимые выборки из этих распределений $X_1,...,X_n$ и $Y_1,...,Y_m$. Проверяется гипотеза о равенстве математических ожиданий нормальных распределений, то есть

$$H_0 : a_1 = a_2 \qquad \textbf{(3.6.28)}$$

против двусторонней альтернативы

$$H_1 : a_1 \neq a_2 \qquad \textbf{(3.6.29)}$$

либо против одной из односторонних альтернатив

$$H_1^+ : a_1 > a_2 \qquad \textbf{(3.6.30)}$$

или

$$H_1^- : a_1 < a_2 \qquad \textbf{(3.6.31)}$$

при следующих предположениях на дисперсиях:

а) обе дисперсии σ_1^2 и σ_2^2 известны и равны между собой, то есть $\sigma_1^2 = \sigma_2^2 = \sigma^2$;

б) обе дисперсии σ_1^2 и σ_2^2 известны, но не равны между собой, то есть $\sigma_1^2 \neq \sigma_2^2$;

в) обе дисперсии σ_1^2 и σ_2^2 неизвестны, но предполагается, что они равны между собой;

г) обе дисперсии σ_1^2 и σ_2^2 неизвестны и их равенство не предполагается.

Обсудим каждый из этих вариантов отдельно.

Вариант а. Для проверки гипотез $H_0 : a_1 = a_2$ и $H_1 : a_1 \neq a_2$ при условии, что дисперсии σ_1^2 и σ_2^2 *известны и равны* между собой, используется критерий, основанный на статистике вида

$$T(\vec{X}, \vec{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma^2 / n + \sigma^2 / m}} = \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{1/n + 1/m}}, \qquad \textbf{(3.6.32)}$$

где выборочные средние $\overline{X}$ и $\overline{Y}$, являясь достаточными статисти-

ками для параметров a_1 и a_2, вычисляются по формулам

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i. \qquad \textbf{(3.6.33)}$$

Отметим, что $L\{\overline{X}\} = N(a_1 ; \sigma^2 / n)$ и $L\{\overline{Y}\} = N(a_2 ; \sigma^2 / n)$. Таким образом, статистика $T(\vec{X}, \vec{Y})$ является линейной комбинацией независимых нормальных случайных величин $\overline{X}$ и $\overline{Y}$, и поэтому, при справедливости гипотезы $H_0 : a_1 = a_2$, она имеет стандартное нормальное распределение, то есть справедливо выражение

$$L(T(\vec{X}, \vec{Y}) \mid H_0) = N(0 ; 1). \qquad \textbf{(3.6.34)}$$

С учетом этого замечания, двусторонний критерий размера α для проверки гипотезы $H_0 : a_1 = a_2$ против альтернативы $H_1 : a_1 \neq a_2$, при условии, что дисперсии σ_1^2 и σ_2^2 известны и равны между собой, определяется критической областью вида

$$\mathrm{X}_{1\alpha} = \{ (\vec{x}, \vec{y}) : | T(\vec{x}, \vec{y}) | \geq \lambda_{1-\alpha/2} \}, \ \lambda_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2) \qquad \textbf{(3.6.35)}$$

и записывается в виде следующего правила:

«*Если модуль наблюденного значения $T(\vec{x}, \vec{y})$ статистики $T(\vec{X}, \vec{Y})$, вычисленного по наблюденным реализациям $x_1, ..., x_n$ и $y_1, ..., y_m$ выборок $X_1, ..., X_n$ и $Y_1, ..., Y_m$, превышает квантиль $\lambda_{1-\alpha/2}$ уровня $1 - \alpha/2$ стандартного нормального распределения, то гипотеза $H_0 : a_1 = a_2$ отвергается на уровне значимости α, в противном случае нет оснований отвергать эту гипотезу*». **(3.6.36)**

Отметим, что для проверки гипотезы $H_0 : a_1 = a_2$ против односторонних альтернатив $H_1^+ : a_1 > a_2$, $H_1^- : a_1 < a_2$ критерии с уровнем значимости α соответственно определяются критическими областями вида

$$\mathrm{X}_{1\alpha}^{+} = \{ (\vec{x}, \vec{y}) : T(\vec{x}, \vec{y}) \geq \lambda_{1-\alpha} \}, \ \lambda_{1-\alpha} = \Phi^{-1}(1 - \alpha); \qquad \textbf{(3.6.37)}$$

$$\mathrm{X}_{1\alpha}^{-} = \{ (\vec{x}, \vec{y}) : T(\vec{x}, \vec{y}) \leq \lambda_{\alpha} \}, \ \lambda_{\alpha} = \Phi^{-1}(\alpha). \qquad \textbf{(3.6.38)}$$

Вариант б. Для проверки гипотез $H_0 : a_1 = a_2$ и $H_1 : a_1 \neq a_2$ при условии, что дисперсии σ_1^2 и σ_2^2 *известны и не равны* между собой, используется критерий, основанный на статистике вида

$$T(\vec{X}, \vec{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}. \qquad \textbf{(3.6.39)}$$

Эта статистика также имеет стандартное нормальное распределение при гипотезе H_0 (см. (3.6.52)) и, следовательно, правило принятия решения аналогично правилу (3.6.36).

Вариант в. Двухвыборочный *t*-критерий Стьюдента. Для проверки гипотез $H_0 : a_1 = a_2$ и $H_1 : a_1 \neq a_2$ при условии, что дисперсии σ_1^2 и σ_2^2 *неизвестны*, *но предполагается*, *что они равны между собой*, используется критерий, основанный на статистике $T(\vec{X}, \vec{Y})$, определяемой в виде

$$T(\vec{X}, \vec{Y}) = \frac{\overline{X} - \overline{Y}}{\tilde{S}_\Sigma \sqrt{1/n + 1/m}} = \sqrt{\frac{nm}{n+m}} \left(\frac{\overline{X} - \overline{Y}}{\tilde{S}_\Sigma} \right), \qquad \textbf{(3.6.40)}$$

где $\tilde{S}_\Sigma^2$ – объединенная несмещенная оценка неизвестного параметра $\sigma^2 = \sigma_1^2 = \sigma_2^2$, построенная по выборкам $\vec{X} = (X_1, ..., X_n)$ и $\vec{Y} = (Y_1, ..., Y_m)$ и вычисляемая по формуле

$$\tilde{S}_\Sigma^2 = \frac{\tilde{S}_1^2 (n-1) + \tilde{S}_2^2 (m-1)}{(n-1) + (m-1)}. \qquad \textbf{(3.6.41)}$$

Здесь $\tilde{S}_1^2$ и $\tilde{S}_2^2$ – несмещенные оценки параметров σ_1^2 и σ_2^2, вычисляемые по выборкам $\vec{X} = (X_1, ..., X_n)$ и $\vec{Y} = (Y_1, ..., Y_m)$ с помощью следующих формул:

$$\tilde{S}_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 , \; \tilde{S}_2^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y})^2 . \qquad \textbf{(3.6.42)}$$

Заметим, что при проверке гипотезы $H_0 : a_1 - a_2 = \delta$, где δ – константа, статистику двухвыборочного *t*-критерия Стьюдента следует

записывать в виде

$$T(\vec{X},\vec{Y}) = \{\,(\overline{X}-\overline{Y})-\delta\,\}/\tilde{S}_{\Sigma}\sqrt{1/n+1/m}\,.$$

Отметим также, что статистика $T(\vec{X},\vec{Y})$ вида (3.6.40) эквивалентна статистике критерия отношения правдоподобия, определенной в общем случае формулой (3.4.18). Убедимся в этом. Для этого построим статистику критерия отношения правдоподобия в данном случае. Пусть $\vec{\theta} = \{a_1, a_2, \sigma^2\}$, тогда параметрические множества, соответствующие предположениям гипотез H_0 и H_1, запишутся в виде $\Theta_0 = \{\vec{\theta} : a_1 = a_2\}$ и $\Theta_1 = \{\vec{\theta} : a_1 \neq a_2\}$. Далее, логарифм функции правдоподобия в данном случае имеет вид

$$\ln L(\vec{x},\vec{y};\vec{\theta}) = -[(n+m)/2]\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-a_1)^2 + \sum_{j=1}^{m}(y_j-a_2)^2\right].$$

Нетрудно убедиться, что при наблюденных реализациях $x_1,...,x_n$ и $y_1,...,y_m$ выборок $X_1,...,X_n$ и $Y_1,...,Y_m$ максимум функции $\ln L(\vec{x},\vec{y};\vec{\theta})$ и, следовательно, максимум функции правдоподобия $L(\vec{x},\vec{y};\vec{\theta})$ на множестве $\Theta = \Theta_0 \cup \Theta_1$ достигаются при замене $\vec{\theta} = \{a_1, a_2, \sigma^2\}$ на наблюденное значение $\{\overline{x}, \overline{y}, s^2\}$ оценки максимального правдоподобия $\hat{\theta} = \{\overline{X}, \overline{Y}, S^2\}$,

где $$S^2 = \frac{1}{(n+m)}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2 + \sum_{j=1}^{m}(Y_j-\overline{Y})^2\right]. \qquad \textbf{(3.6.43)}$$

Далее, при справедливости гипотезы $H_0 : a_1 = a_2 = a$ рассматриваемая модель сводится к нормальной модели с одной выборкой (см. пример 3.4.41) и, следовательно, максимум функции правдоподобия на множестве $\Theta_0 = \{\vec{\theta} : a_1 = a_2\}$ достигается при $\hat{\theta} = \{\hat{a}, \hat{a}, \hat{\sigma}_0^2\}$, где

$$\hat{a} = \frac{1}{n+m}\left[\sum_{i=1}^{n}X_i + \sum_{j=1}^{m}Y_j\right] \text{ и } \hat{\sigma}_0^2 = \frac{1}{(n+m)}\left[\sum_{i=1}^{n}(X_i-\hat{a})^2 + \sum_{j=1}^{m}(Y_j-\hat{a})^2\right].$$

Используя далее тождества

$$\frac{1}{n+m}\sum_{i=1}^{n}(X_i-\hat{a})^2=\frac{1}{n+m}\sum_{i=1}^{n}(X_i-\overline{X})^2+\frac{n}{n+m}(\overline{X}-\hat{a})^2$$

и $$\frac{1}{n+m}\sum_{i=1}^{n}(Y_i-\hat{a})^2=\frac{1}{n+m}\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\frac{n}{n+m}(\overline{Y}-\hat{a})^2,$$

(см. формулу (3.4.26)) и действуя так же, как в примере 3.4.41, получаем выражение для статистики отношения правдоподобия $\Lambda^*\{(\vec{X},\vec{Y});\Theta_0\}$ в виде

$$(\Lambda^*\{(\vec{X},\vec{Y});\Theta_0\})^{2/(n+m)}=\left[1+\frac{\tilde{T}^2(\vec{X},\vec{Y})}{n+m-2}\right]^{-1},$$

где $$\tilde{T}(\vec{X},\vec{Y})=\frac{(\overline{X}-\overline{Y})-(a_1-a_2)}{S\sqrt{1/n+1/m}} \qquad \textbf{(3.6.44)}$$

и оценка S^2 максимального правдоподобия неизвестного параметра $\sigma^2=\sigma_1^2=\sigma_2^2$ определена в (3.6.43). Отметим, что маленькие значения статистики отношения правдоподобия $\Lambda^*\{(\vec{X},\vec{Y});\Theta_0\}$ достигаются в двух случаях, когда статистика $\tilde{T}(\vec{X},\vec{Y})$ принимает маленькие или большие значения, то есть критическая область должна состоять из двух частей. Таким образом, симметричный двусторонний критерий проверки гипотез $H_0: a_1=a_2$ и $H_1: a_1\neq a_2$ при условии, что дисперсии σ_1^2 и σ_2^2 неизвестны, но равны между собой, будет отвергать нулевую гипотезу, если *абсолютная величина* наблюденного значения $\tilde{T}(\vec{x},\vec{y})$ статистики $\tilde{T}(\vec{X},\vec{Y})$ превышает критическое значение. Отметим, что статистика $T(\vec{X},\vec{Y})$ эквивалентна статистике отношения правдоподобия $\tilde{T}(\vec{X},\vec{Y})$ и имеет при справедливости гипотезы $H_0: a_1=a_2$ распределение Стьюдента с числом степеней свободы $(n+m-2)$. В этом легко убедиться, представив статистику $T(\vec{X},\vec{Y})$ в виде

$$\left(\frac{(\overline{X}-\overline{Y})-(a_1-a_2)}{\sigma\sqrt{1/n+1/m}}\right)\Big/\sqrt{\tilde{S}_{\Sigma}^2/\sigma^2}\ . \qquad \textbf{(3.6.45)}$$

Числителем этой дроби является стандартная нормальная случайная величина, которая не зависит от случайной величины $\tilde{S}_{\Sigma}^2/\sigma^2$, стоящей под корнем в знаменателе этой дроби, и имеет распределение хи-квадрат с числом степеней свободы $(n+m-2)$. Таким образом, справедливо выражение

$$L\{T(\vec{X},\vec{Y})\,|\,H_0\} = St(n+m-2). \qquad \textbf{(3.6.46)}$$

С учетом этого замечания, двусторонний критерий размера α для проверки гипотезы $H_0: a_1 = a_2$ против альтернативы $H_1: a_1 \neq a_2$, при условии, что дисперсии σ_1^2 и σ_2^2 неизвестны, но равны между собой, определяется критической областью вида

$$\mathrm{X}_{1\alpha} = \{(\vec{x},\vec{y}) : |T(\vec{x},\vec{y})| \geq t_{1-\alpha/2,\,(n+m-2)}\}, \qquad \textbf{(3.6.47)}$$

или

$$\mathrm{X}_{1\alpha} = \left\{(\vec{x},\vec{y}) : |\overline{x}-\overline{y}| \geq t_{1-\alpha/2,(n+m-2)}\sqrt{\frac{(n+m)\{(n-1)\tilde{s}_1^2+(m-1)\tilde{s}_2^2\}}{nm(n+m-2)}}\right\}, \qquad \textbf{(3.6.47а)}$$

где $t_{1-\alpha/2,\,(n+m-2)}$ – квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы $(n+m-2)$ и $T(\vec{x},\vec{y})$ – вычисленное по наблюденным реализациям $x_1,...,x_n$ и $y_1,...,y_m$ выборок $X_1,...,X_n$ и $Y_1,...,Y_m$ значение статистики $T(\vec{X},\vec{Y})$, определенной в (3.6.40).

Отметим, что двустороннему критерию соответствуют, как обычно, два односторонних критерия размера α для проверки гипотез $H_0: a_1 = a_2$, $H_1^+: a_1 > a_2$ или $H_0: a_1 = a_2$, $H_1^-: a_1 < a_2$, соответственно определяемых критическими областями

$$\mathrm{X}_{1\alpha}^{+} = \{(\vec{x},\vec{y}) : T(\vec{x},\vec{y}) \geq t_{1-\alpha,\,(n+m-2)}\}; \qquad \textbf{(3.6.48)}$$

$$\mathrm{X}_{1\alpha}^{-} = \{(\vec{x},\vec{y}) : T(\vec{x},\vec{y}) \leq t_{\alpha,\,(n+m-2)}\}. \qquad \textbf{(3.6.49)}$$

Отметим также, что выражение (3.6.44) позволяет записать симметричный $(1-\alpha)$-доверительный интервал для разности (a_1-a_2), с учетом симметрии распределения Стьюдента, в следующем виде:

$$P\{(\overline{X}-\overline{Y})-t_{1-\alpha/2,(n+m-2)}\cdot\tilde{S}_{\Sigma}\sqrt{1/n+1/m}<(a_1-a_2)<$$
$$<(\overline{X}-\overline{Y})+t_{1-\alpha/2,(n+m-2)}\cdot\tilde{S}_{\Sigma}\sqrt{1/n+1/m}\ \}=1-\alpha. \qquad \textbf{(3.6.50)}$$

Можно убедиться (см., например, [22]), что двусторонний критерий, определяемый критической областью (3.6.47), является н.р.н.м. критерием уровня α для проверки гипотез $H_0: a_1 = a_2$, $H_1: a_1 \neq a_2$. Поэтому для разности $(a_1 - a_2)$ доверительный интервал

$$(\overline{X}-\overline{Y})\pm t_{1-\alpha/2,(n+m-2)}\cdot\sqrt{\frac{(n+m)\{(n-1)\tilde{S}_1^2+(m-1)\tilde{S}_2^2\}}{nm(n+m-2)}} \qquad \textbf{(3.6.50а)}$$

будет наикратчайшим, то есть будет иметь минимальную длину среди всех $(1-\alpha)$-доверительных интервалов для разности $(a_1 - a_2)$.

Рассмотрим числовой пример. Имеем две независимые выборки $\vec{X}=(X_1,...,X_n)$ и $\vec{Y}=(Y_1,...,Y_m)$ из нормальных совокупностей с равными дисперсиями. Пусть $n=41$ и $m=70$. По наблюденным реализациям $\vec{x}=(x_1,...,x_{41})$ и $\vec{y}=(y_1,...,y_{70})$ этих выборок вычислены наблюденные значения статистик: $\overline{x}=0{,}695$, $\tilde{s}_1=0{,}326$ и $\overline{y}=0{,}399$, $\tilde{s}_2=0{,}383$. Требуется проверить гипотезу $H_0: a_1 - a_2 = 0$ против односторонней альтернативы $H_1: a_1 - a_2 > 0$ с помощью *t*-критерия Стьюдента, основанного на статистике (3.6.40). Согласно (3.6.41), наблюденное значение объединенной выборочной дисперсии равно

$$\tilde{S}_{\Sigma}^2=\frac{\tilde{S}_1^2(n-1)+\tilde{S}_2^2(m-1)}{(n-1)+(m-1)}=\frac{(41-1)(0{,}326)^2+(70-1)(0{,}383)^2}{41+70-2}=0{,}132.$$

Согласно (3.6.40), наблюденное значение *t*-статистики вычисляется по формуле

$$T(\vec{x},\vec{y})=\frac{\overline{x}-\overline{y}}{\tilde{s}_{\Sigma}\sqrt{1/n+1/m}}=\frac{0{,}695-0{,}399}{\sqrt{0{,}132(1/41+1/70)}}=4{,}14.$$

Число степеней свободы в данном случае равно $n+m-2=41+70-2=109$. Для принятия решения вычислим *P-значение* критерия (см. упр. 3.9.74):

$$\textit{P-значение}\ =P\{T(\vec{X},\vec{Y})\geq T(\vec{x},\vec{y})\,|\,H_0\}=P\{St(109)\geq 4{,}14\}<0{,}001,$$

Это неравенство означает, что гипотезу $H_0: a_1 = a_2$ следует отвергнуть.

Вариант г. Задачи проверки гипотез $H_0 : a_1 = a_2$ и $H_1 : a_1 \neq a_2$ при условии, что дисперсии σ_1^2 и σ_2^2 *неизвестны и их равенство не предполагается*, часто возникают, например, при измерениях характеристик физических объектов двумя измерительными приборами, при этом в общем случае совсем не обязательно, что приборы имеют одинаковую точность измерений. Отметим, что в данном случае и нулевая и альтернативная гипотезы являются сложными. Попытка решения данной задачи, которая в литературе известна как *проблема Беренса – Фишера*, путем построения критерия отношения правдоподобия не дает результата как в предыдущем варианте, так как в этом случае получим статистику, распределение которой зависит от неизвестного отношения σ_1^2 / σ_2^2.

Обсудим решение данной задачи, основываясь на том факте, что для векторного параметра $\vec{\theta} = (a_1, a_2, \sigma_1^2, \sigma_2^2)$ полной достаточной статистикой служит $(\overline{X}, \overline{Y}, \tilde{S}_1^2, \tilde{S}_2^2)$, где $\overline{X}$ и $\overline{Y}$ – выборочные средние, вычисляемые по формулам (3.6.33), а $\tilde{S}_1^2$ и $\tilde{S}_2^2$ – подправленные выборочные дисперсии, вычисляемые по формулам (3.6.42). Отметим, что выборочные средние $\overline{X}$ и $\overline{Y}$ имеют нормальные распределения, то есть $L(\overline{X}) = N(a_1; \sigma_1^2 / n)$ и $L(\overline{Y}) = N(a_2; \sigma_2^2 / m)$. Учитывая далее, что выборки $X_1, ..., X_n$ и $Y_1, ..., Y_m$ независимые, получаем выражение для дисперсии разности случайных величин $\overline{X}$ и $\overline{Y}$ в виде

$$D[\overline{X} - \overline{Y}] = D(\overline{X}) + D(\overline{Y}) = \sigma_1^2 / n + \sigma_2^2 / m. \qquad \textbf{(3.6.51)}$$

С учетом этих замечаний получаем, что случайная величина $\{(\overline{X} - \overline{Y}) - (a_1 - a_2)\} / \sqrt{\sigma_1^2 / n + \sigma_2^2 / m}$ имеет стандартное нормальное распределение, то есть справедливо выражение вида

$$L\left(\frac{(\overline{X} - \overline{Y}) - (a_1 - a_2)}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}} \right) = N(0\,;1). \qquad \textbf{(3.6.52)}$$

Так как дисперсии σ_1^2 и σ_2^2 неизвестны, то для построения статистики критерия используют несмещенные оценки этих параметров в виде (3.6.42), то есть

$$\tilde{S}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\ ,\ \ \tilde{S}_2^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2\ .$$

В результате замены неизвестных дисперсий σ_1^2 и σ_2^2 их несмещенными оценками получим статистику вида

$$\hat{T}(\vec{X},\vec{Y}) = \frac{\overline{X} - \overline{Y}}{\sqrt{\tilde{S}_1^2/n + \tilde{S}_2^2/m}}. \qquad \textbf{(3.6.53)}$$

Отметим, что при конечных (и небольших) объемах выборки n и m использование статистики $\hat{T}(\vec{X},\vec{Y})$ связано с трудностями, которые состоят в том, что ее распределение зависит от неизвестного отношения σ_1^2/σ_2^2. При больших объемах выборки n и m эти трудности исчезают, так как статистика $\hat{T}(\vec{X},\vec{Y})$ имеет по теореме Слуцкого (см. П.3.9) и центральной предельной теореме (см. П.3.37) приближенно нормальное распределение. На практике, при конечных и небольших объемах выборки n и m, обычно используют аппроксимацию Уэлша (см. [9]), согласно которой при гипотезе $H_0 : a_1 = a_2$ статистика $\hat{T}(\vec{X},\vec{Y})$ имеет распределение Стьюдента с числом степеней свободы

$$k = \{c^2/(n-1) + (1-c)^2/(m-1)\}^{-1},\ c = \frac{\tilde{S}_1^2/(n+m)}{\tilde{S}_1^2/n + \tilde{S}_2^2/m}. \qquad \textbf{(3.6.54)}$$

Замечание 3.6.55. Отметим, что число степеней свободы k из (3.6.54) является величиной случайной и, следовательно, может быть числом не целым. В таких случаях при нахождении критических значений статистики $\hat{T}(\vec{X},\vec{Y})$ используют линейную интерполяцию по таблицам *t*-распределения Стьюдента. Отметим также, что при построении доверительных интервалов для разности $(a_1 - a_2)$ в качестве центральной статистики (см. разд. 2.12) используют статистику вида

$$\frac{(\overline{X} - \overline{Y}) - (a_1 - a_2)}{\sqrt{\tilde{S}_1^2/n + \tilde{S}_2^2/m}}, \qquad \textbf{(3.6.56)}$$

распределение которой при конечных и небольших объемах выборки n и m находится в хорошем согласии с аппроксимацией Уэлша. Отметим, что кри-

терии и доверительные интервалы, к которым приводит эта аппроксимация, называются решениями Уэлша проблемы Беренса – Фишера и в литературе их часто называют *t*-критериями Уэлша. Подробное обсуждение аппроксимации Уэлша приводится в [9, с. 280] и [6].

В заключение отметим, что точное решение проблемы Беренса – Фишера при конечных объемах выборки n и m найти не удается, однако с практической точки зрения аппроксимация Уэлша вполне удовлетворительна. При обсуждении качества аппроксимации Уэлша в работе [9, с. 280] отмечается, что критерий Стьюдента, определяемый критической областью (3.6.46), очень чувствителен к нарушению предположения о равенстве дисперсий σ_1^2 и σ_2^2, при неравных объемах выборок n и m.

Таким образом, если $n \neq m$ и нет полной уверенности в том, что $\sigma_1^2 = \sigma_2^2$, то для проверки гипотез $H_0 : a_1 = a_2$, $H_1 : a_1 \neq a_2$ правильнее использовать *t*-критерий Уэлша, основанный на статистике $\hat{T}(\vec{X}, \vec{Y})$, определяемый критической областью размера α в виде

$$X_{1\alpha} = \{ (\vec{x}, \vec{y}) : | \hat{T}(\vec{x}, \vec{y}) | \geq t_{1-\alpha/2,\,k} \}, \qquad \textbf{(3.6.57)}$$

где $t_{1-\alpha/2,\,k}$ – квантиль уровня $1-\alpha/2$ распределения Стьюдента с числом степеней свободы k, вычисляемым по формуле (3.6.54), при этом, по возможности, следует обеспечить равенство объемов выборок n и m.

3.6.Г. Парный *t*-критерий Стьюдента

В предыдущем разделе рассматривались критерии сравнения средних значений нормальных совокупностей по двум *независимым* выборкам из этих совокупностей при различных предположениях на дисперсии. Было показано, что в случае *неизвестных*, *но равных дисперсий*, следует использовать *двухвыборочный t-критерий Стьюдента*, основанный на статистике (3.6.40). Если же дисперсии *неизвестны и их равенство не предполагается*, то следует использовать *t-критерий Уэлша*, основанный на статистике (3.6.53). На практике часто возникает необходимость сравнивать два различных технологических процесса или два метода обработки. При решении этих задач желательно уменьшить или вообще исключить рассеяние, которое имеется между сравниваемыми совокупностями. Для осуществления этого при сравнении характеристик двух совокупностей используют связанные (зависимые) выборки и методы парных сравнений. Кратко обсудим эти методы.

Итак, пусть у каждого объекта изучаемой совокупности измеряется одна и та же характеристика до и после «обработки». Для определенности пусть $\{X_i\}$ – измерения до обработки, а $\{Y_i\}$ – после обработки. Например, признак X характеризует вес пациента до диеты, а признак Y тоже характеризует вес пациента, но уже после диеты. Наблюдается n пациентов, и результаты *парных наблюдений* представлены двумя *связанными* (*зависимыми*) выборками $X_1,...,X_n$ и $Y_1,...,Y_n$. Пара (X_i, Y_i) характеризует *i*-го пациента, $i = 1,...,n$. На практике используют различные варианты формирования пар. Например, можно измерять выборочным методом один и тот же признак объекта до и после обработки или можно заранее сформировать n пар объектов по принципу их максимального сходства по отношению к цели измерений, а затем случайным образом (например, подбрасывая симметричную монету) к каждой компоненте пары применить интересующую нас обработку. При этом обеспечивается почти полная однородность экспериментального материала, достигается возможность контроля над внешними факторами и увеличивается чувствительность эксперимента, направленного на оценку эффективности интересующей нас обработки (см. упр. 3.9.76).

При построении модели парных наблюдений предполагаем, что вес пациента после перехода на диету равен его весу до диеты плюс эффект от ее воздействия, то есть $Y_i = X_i + \theta + \varepsilon_i$, $i = 1,...,n$, где ε_i – ненаблюдаемые случайные величины, θ – эффект от воздействия диеты (эффект «обработки»). Пусть далее определены разности $V_i = Y_i - X_i$, $i = 1,...,n$. Тогда модель парных наблюдений, проводимых с целью выяснения эффекта обработки, запишется в виде $V_i = \theta + \varepsilon_i$, $i = 1,...,n$, где ε_i – независимые случайные величины, причем обычно предполагается, что $L(\varepsilon_i) = N(\theta; \sigma_V^2)$, $i = 1,...,n$, и параметр θ характеризует эффект «обработки». Для такой модели наблюдений статистические задачи связаны с построением оценок параметра θ и проверкой гипотез относительно этого параметра по наблюдаемым данным $\vec{X} = (X_1,...,X_n)$ и $\vec{Y} = (Y_1,...,Y_n)$, которые яв-

ляются уже связанными (зависимыми) выборками. Анализ повторных парных наблюдений с помощью знаковых рангов для более общей непараметрической модели $V_i = \theta + \varepsilon_i$, $i = 1,...,n$, для которой ε_i – независимые случайные величины с *произвольным* непрерывным и симметричным относительно нуля распределением, обсуждается в разд. 5.2. Итак, в рамках гауссовской модели наблюдений вида

$$V_i = \theta + \varepsilon_i,\ L(\varepsilon_i) = N(\theta;\sigma_V^2),\ i = 1,...,n, \qquad \textbf{(3.6.57a)}$$

требуется проверить нулевую гипотезу об отсутствии эффекта обработки, которая записывается как $H_0 : \theta = \theta_0$, где θ_0 – заданная постоянная величина. Эта гипотеза может проверяться против односторонних альтернатив $H_1^+ : \theta > \theta_0$, $H_1^- : \theta < \theta_0$ либо против двусторонней альтернативы $H_1 : \theta \neq \theta_0$. Для проверки этих гипотез используется *парный t-критерий Стьюдента*, который основан на статистке вида

$$T(\vec{X},\vec{Y}) = \sqrt{n}(\overline{V} - \theta_0)/\tilde{S}_V, \qquad \textbf{(3.6.58)}$$

где $\overline{V}$ – выборочное среднее разности $V = (Y - X)$ и $\tilde{S}_V^2$ – подправленная выборочная дисперсия этой разности, то есть

$$\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i,\ \tilde{S}_V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} V_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} V_i\right)^2\right\}.$$

Замечание 3.6.59. Отметим, что для принятой модели парных наблюдений ряд значений разностей $V_i,...,V_n$, при справедливости гипотезы $H_0 : \theta = \theta_0$, может рассматриваться как выборка из нормальной совокупности со средним значением θ_0 и неизвестной дисперсией σ_V^2. С учетом этого замечания, согласно теореме (2.8.31), распределение статистики $T(\vec{X},\vec{Y}) = \sqrt{n}(\overline{V} - \theta_0)/\tilde{S}_V$ при справедливости гипотезы $H_0 : \theta = \theta_0$ характеризуется распределением Стьюдента с числом степеней свободы $(n-1)$, то есть справедливо выражение $L\{T(\vec{X},\vec{Y}) \mid H_0\} = St(n-1)$. Отсюда заключаем, что *P-значение* критерия, основанного на статистике $T(\vec{X},\vec{Y}) = \sqrt{n}(\overline{V} - \theta_0)/\tilde{S}_V$, для проверки соответствующих гипотез вычисляется по формулам, приведенным в табл. 3.6.1.

Таблица 3.6.1

Нулевая гипотеза	Альтернатива	*P-значение*
$H_0:\theta=\theta_0$, σ^2 *неизвестно*	$H_1^+:\theta>\theta_0$	$P=\Pr\{St(n-1)\ge t(\vec{x},\vec{y})\mid H_0\}$
	$H_1^-:\theta<\theta_0$	$P=\Pr\{St(n-1)\le t(\vec{x},\vec{y})\mid H_0\}$
	$H_1:\theta\ne\theta_0$	$P=2\min[\Pr\{St(n-1)\ge t(\vec{x},\vec{y})\mid H_0\},$ $\Pr\{St(n-1)\le t(\vec{x},\vec{y})\mid H_0\}]$

В этой таблице через $St(n-1)$ обозначена случайная величина, имеющая распределение Стьюдента с числом степеней свободы $(n-1)$, а $t(\vec{x},\vec{y})$ – наблюденное значение статистики $T(\vec{X},\vec{Y})=\sqrt{n}(\overline{V}-\theta_0)/\tilde{S}_V$, вычисленное по наблюденным реализациям $\vec{x}=(x_1,...,x_n)$ и $\vec{y}=(y_1,...,y_n)$ выборок $\vec{X}=(X_1,...,X_n)$ и $\vec{Y}=(Y_1,...,Y_n)$. Окончательное решение с использованием *P-значения* принимается следующим образом. Если вычисленное *P-значение* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается.

Отметим, что в частном случае, при $\theta_0=0$, предположение гипотезы H_0 означает, что отсутствует эффект «обработки» и она кратко записывается в виде $H_0:\mu_V=0$, где $\mu_V=M\{(Y-X)\}$ – истинное среднее значение разности $(Y-X)$, оценка которого находится по результатам наблюдений $V_i,...,V_n$. Относительно дисперсии разности $\sigma_V^2=D\{(Y-X)\}$ предполагается, что она постоянна. Гипотеза $H_0:\mu_V=0$ может проверяться против односторонних альтернатив $H_1^+:\mu_V>0$, $H_1^-:\mu_V<0$ либо против двусторонней альтернативы $H_1:\mu_V\ne0$. Наблюденное значение $t(\vec{x},\vec{y})$ статистики $T(\vec{X},\vec{Y})$ записывается в данном случае в удобном для вычислений виде

$$t(\vec{x},\vec{y})=\frac{\sum v_i/n}{\sqrt{\dfrac{\sum v_i^2-(\sum v_i)^2/n}{n(n-1)}}}, \tag{3.6.60}$$

где $v_i=y_i-x_i$, $i=1,...,n$, и $\vec{x}=(x_1,...,x_n)$, $\vec{y}=(y_1,...,y_n)$ – наблюденные реализации выборок $\vec{X}=(X_1,...,X_n)$ и $\vec{Y}=(Y_1,...,Y_n)$.

Рассмотрим числовой пример. Пусть объем выборки $n=8$ и $\{x_i\}$, $\{y_i\}$, $i=1,...,8$, соответственно наблюденные значения некоторого признака объекта до и после обработки приведены в табл. 3.6.2. Требуется проверить гипотезы $H_0:\mu_V=0$, $H_1:\mu_V\ne0$. Наблюденное значение статистики равно

$$t(\vec{x},\vec{y}) = \frac{9,2/8}{\sqrt{\dfrac{20,04-(9,2)^2/8}{8(8-1)}}} = 2,80.$$

Пусть уровень значимости равен $\alpha = 0,05$. По таблицам распределения Стьюдента при числе степеней свободы $(n-1) = 7$ с учетом того, что рассматривается двусторонняя альтернатива $H_1 : \mu_V \neq 0$, определяем критическое значение $t_{1-\alpha/2,\,(n-1)} = t_{0,975,\,(7)} = 2,36$. Далее, так как $(T(\vec{x},\vec{y}) = 2,80) > (t_{0,975,\,(7)} = 2,36)$, то гипотезу $H_0 : \mu_V = 0$ отвергаем на уровне значимости $\alpha = 0,05$, то есть разница между связанными выборками является значимой на пятипроцентном уровне значимости.

Таблица 3.6.2

№	$\{y_i\}$	$\{x_i\}$	$\{v_i = y_i - x_i\}$	$\{v_i^2\}$
1	4,0	3,0	1,0	1,00
2	3,5	3,0	0,5	0,25
3	4,1	3,8	0,3	0,09
4	5,5	2,1	3,4	11,56
5	4,6	4,9	-0,3	0,09
6	6,0	5,3	0,7	0,49
7	5,1	3,1	2,0	4,00
8	4,3	2,7	1,6	2,56
$n = 8$			$\sum v_i = 9,2$	$\sum v_i^2 = 20,04$

Замечание 3.6.61. Подводя итог разд. 3.6.В и 3.6.Г, отметим, что при сравнении средних значений двух нормальных случайных величин используют три варианта *t*-критерия. Эти варианты отличаются: 1) предположениями о независимости двух выборок и 2) поведением дисперсий нормальных популяций. Если рассматриваемая случайная величина определена на двух различных популяциях W_1 и W_2 (их называют подпопуляции, *слои или страты*) из более широкой популяции W и в этом случае имеются *две независимые выборки* – по одной из каждой нормальной популяции, то *при равных дисперсиях* используют *двухвыборочный t-критерий Стьюдента*, основанный на статистике (3.6.47), а если *равенство дисперсий не предполагается*, то используют *t-критерий Уэлша*, основанный на статистике (3.6.57). Далее, если же предположение *независимости двух выборок не выполняется*, то есть рассматриваются *связанные* (*зависимые*) выборки, то используют *парный t-критерий*, основанный на статистике (3.6.58). Проверка гипотезы о равенстве двух дисперсий, основанная на связанных выборках, рассматривается в

упр. 3.9.77. Непараметрические критерии сравнения двух совокупностей, как по независимым, так и связанным выборкам без предположения нормальности, обсуждаются в разд. 5.1 и 5.2.

3.6.Д. Критерий однородности χ^2

Этот критерий предназначен для проверки гипотезы об однородности данных заданного числа m различных совокупностей, объекты которых характеризуются некоторым признаком X дискретного или непрерывного типа. Мы обсудим лишь дискретный вариант, поскольку непрерывный случай сводится к дискретному с помощью метода группировки.

Итак, пусть имеется m различных совокупностей, объекты которых характеризуются дискретным признаком X с заданным числом k возможных состояний $\tilde{x}_1,...,\tilde{x}_k$, то есть имеем дело с m дискретными случайными величинами, заданными распределениями вероятностей вида

$$\begin{pmatrix} X \\ \vec{p}_j \end{pmatrix} : \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ p_{1j},...,p_{kj} \end{pmatrix}, \ j = 1,...,m. \qquad \textbf{(3.6.62)}$$

Выборочным методом в каждой j-й совокупности, $j = 1,...,m,$ осуществляется $n_{\bullet j}$ наблюдений над признаком X. В результате имеем выборки $X_{1j},...,X_{n_{\bullet j} j}$, $j = 1,...,m$. Далее, наблюденные данные представляются дискретными вариационными рядами частот

$$\begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ n_{1j},...,n_{kj} \end{pmatrix}, \ j = 1,...,m, \qquad \textbf{(3.6.63)}$$

где n_{ij} – число реализаций i-го состояния $\tilde{x}_i$, $i = 1,...,k,$ в j-й совокупности, $j = 1,...,m$, причем

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}, \ j = 1,...,m, \ n_{\bullet\bullet} = \sum_{j=1}^{m} n_{\bullet j} = \sum_{j=1}^{m}\sum_{i=1}^{k} n_{ij} = n. \qquad \textbf{(3.6.64)}$$

Наблюденные частоты $\{n_{ij}\}$, $i = 1,...,k$, $j = 1,...,m,$ для удобства и наглядности обозначений могут быть представлены в следующей таблице.

Таблица 3.6.3

$i \setminus j$	1	…	j	…	m	$\sum$
1	n_{11}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
⋮	⋮				⋮	⋮
i	n_{i1}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
⋮	⋮				⋮	⋮
k	n_{k1}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
$\sum$	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	$n_{\bullet\bullet}$

По данным частот $\{n_{ij}\}$, $i=1,...,k$, $j=1,...,m$, требуется проверить гипотезу однородности $H_0^{\div}$ о том, что все наблюдения проводились над одной и той же дискретной случайной величиной X, заданной распределением вида

$$\begin{pmatrix} X \\ \vec{p} \end{pmatrix} : \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ p_1,...,p_k \end{pmatrix}, \ \sum_{i=1}^{k} p_i = 1.$$

Кратко эта гипотеза запишется в виде

$$H_0^{\div} : (p_{1j},...,p_{kj}) = (p_1,...,p_k), \ j=1,...,m, \qquad \textbf{(3.6.65)}$$

где $\vec{p} = (p_1,...,p_k)$ – некоторый неизвестный вектор вероятностей, причем $p_1 + ... + p_k = 1$. Отметим, что наблюденные частоты $\{n_{ij}\}$, $i=1,...,k$, $j=1,...,m$, являются реализациями случайных величин $\{N_{ij}\}$, описываемых полиномиальным распределением (П.1.49), поэтому $M(N_{ij} \mid H_0^{\div}) = n_{\bullet j} p_i$, $i=1,...,k$, $j=1,...,m$. С учетом этого замечания статистика χ^2, основанная на отклонениях опытных данных $\{n_{ij}\}$ от их гипотетических значений $\{n_{\bullet j} p_i\}$, $i=1,...,k$, $j=1,...,m$, запишется в виде

$$\chi_2(\vec{p}) = \sum_{i=1}^{k}\sum_{j=1}^{m} (n_{ij} - n_{\bullet j} p_i)^2 / (n_{\bullet j} p_i). \qquad \textbf{(3.6.66)}$$

На практике вероятности $p_1,...,p_k$ обычно неизвестны, поэтому для вычисления статистики χ^2 требуется предварительно построить

оценки по наблюденным частотам $\{n_{ij}\}$, $i=1,...,k$, $j=1,...,m$, этих вероятностей при справедливости гипотезы $H_0^{\div}$. Для этого используется метод максимального правдоподобия. Для данных частот $\{n_{ij}\}$, $i=1,...,k$, $j=1,...,m$, функция правдоподобия при гипотезе $H_0^{\div}$ запишется в виде

$$L(\{n_{ij}\};\vec{p}\mid H_0^{\div})=\prod_{i=1}^{k}\prod_{j=1}^{m}p_i^{n_{ij}}=\prod_{i=1}^{k}p_i^{n_{i\bullet}}, \qquad \textbf{(3.6.67)}$$

где $n_{i\bullet}$ вычисляется для данных частот $\{n_{ij}\}$ по формуле

$$n_{i\bullet}=\sum_{j=1}^{m}n_{ij}. \qquad \textbf{(3.6.68)}$$

Логарифмируя функцию правдоподобия и применяя метод неопределенных множителей Лагранжа, получаем (см. упр. 3.9.72) мультиномиальные оценки максимального правдоподобия $\hat{p}_i$ параметров p_i, $i=1,...,k$, в виде

$$\hat{p}_i=\frac{n_{i\bullet}}{n}, \qquad \textbf{(3.6.69)}$$

где $n_{i\bullet}$, $i=1,...,k$, определены в (3.6.68) и n – общее число наблюдений по всем m совокупностям, то есть

$$n=n_{\bullet 1}+...+n_{\bullet m}=\sum_{j=1}^{m}n_{\bullet j}=\sum_{j=1}^{m}\sum_{i=1}^{k}n_{ij}. \qquad \textbf{(3.6.70)}$$

Заменяя неизвестные значения параметров $p_1,...,p_k$ в статистике (3.6.66) на их оценки, получим модифицированную статистику хи-квадрат в виде

$$\chi^2(\hat{p})=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-n_{\bullet j}n_{i\bullet}/n)^2}{n_{\bullet j}n_{i\bullet}/n}=n\left(\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}}-1\right). \qquad \textbf{(3.6.71)}$$

Хорошо известно (см., например, [36]), что эта статистика при справедливости гипотезы $H_0^{\div}$ имеет при больших объемах выборки

распределение хи-квадрат с числом степеней свободы $(k-1)(m-1)$, то есть справедливо выражение

$$L(\chi^2(\hat{p}) \mid H_0^{\div}) = \chi^2((k-1)(m-1)) \text{ при } n \to \infty. \qquad \textbf{(3.6.72)}$$

Таким образом, гипотеза однородности $H_0^{\div}$ отвергается на уровне значимости α, если вычисленное по наблюденным частотам $\{n_{ij}\}$, $i = 1,...,k$, $j = 1,...,m$, значение $\chi^2_{\text{н}}$ модифицированной статистики $\chi^2(\hat{p})$ превышает ее критическое значение $\chi^2_{1-\alpha,(k-1)(m-1)}$, в противном случае нет оснований отвергать гипотезу однородности $H_0^{\div}$ вида (3.6.65).

Замечание 3.6.73. Отметим, что изложенную методику можно использовать и для проверки гипотезы о том, что k серий наблюдений произведены над одной и той же случайной величиной непрерывного типа, имеющей заданное параметрическое распределение, например нормальное, экспоненциальное и т.п. В этом случае предварительно используют метод группировки и находят мультиномиальные оценки максимального правдоподобия $\hat{\theta}$ параметров распределения при нулевой гипотезе, которая записывается в виде $H_0^{\div\div}$: все данные воспринимаются как одна выборка с групповыми частотами $n_{1\bullet},...,n_{k\bullet}$. В данном случае используется модифицированная статистика хи-квадрат вида

$$\chi^2(p(\hat{\theta})) = \sum_{i=1}^{k}\sum_{j=1}^{m}(n_{ij} - n_{\bullet j}\hat{p}_i)^2 / (n_{\bullet j}\hat{p}_i),$$

где $\hat{p}_i = p_i(\hat{\theta})$, $i = 1,...,k$, и $\hat{\theta}$ – мультиномиальные оценки максимального правдоподобия векторного параметра θ размерности r. Для статистики $\chi^2(p(\hat{\theta}))$ справедливо следующее выражение:

$$L(\chi^2(p(\hat{\theta})) \mid H_0^{\div\div}) = \chi^2(m(k-1)-r) \text{ при } n \to \infty \text{ (см., например, [22, 36]).}$$

Отметим также, что проверка *гипотезы однородности* (3.6.65) по данным табл. 3.6.3 может трактоваться и как проверка *гипотезы независимости* (3.7.32) по данным таблицы сопряженности атрибутивных признаков (см. разд. 3.7). Это утверждение является справедливым в тех случаях, когда в соответствии с уровнями одного классифицирующего признака изучаемая совокупность W разбивается на непересекающиеся подсовокупности $W_1,...,W_m$.

В этом случае гипотезу независимости признаков можно формулировать и как гипотезу об однородности классифицирующего признака по отношению к уровням другого признака (см. пример 3.7.40).

Пример 3.6.74. На предприятие поступают комплектующие изделия одного наименования от трех поставщиков A, B, C. Результаты проверки изделий приведены в табл. 3.6.4. Требуется проверить на уровне значимости $\alpha = 0{,}10$ гипотезу о том, что продукция поставщиков A, B, C однородна по качеству (гипотеза однородности). Или, другими словами, можно ли считать на основании наблюденных данных табл. 3.6.4, что качество изделий не зависит от поставщика? (гипотеза независимости). Обе сформулированные гипотезы проверяются с помощью критерия хи-квадрат, статистика которого записывается в виде (3.6.71).

При $k = 2$, $m = 3$ и $n = 130$ наблюденное (вычисленное) значение $\chi^2_{\text{в}}$ статистики $\chi^2(\hat{p})$ равно

$$\chi^2_{\text{в}} = n\{\sum_{i=1}^{k}\sum_{j=1}^{m}(n_{ij}^2 / n_{i\bullet}\cdot n_{\bullet j}) - 1\} =$$

$$= 130\{(29^2/30\cdot 120) + (38^2/40\cdot 120) + (53^2/60\cdot 120) +$$

$$+ (1^2/30\cdot 10) + (2^2/40\cdot 10) + (7^2/60\cdot 10) - 1\} = 2{,}546.$$

Таблица 3.6.4

Качество\ поставщик	A	B	C	$\sum$
Годные	29	38	53	120
Негодные	1	2	7	10
$\sum$	30	40	60	130

Далее, число степеней свободы равно $(k-1)(m-1) = (2-1)(3-1) = 2$ и при $\alpha = 0{,}10$ критическое значение $\chi^2_{1-\alpha,\,(k-1)(m-1)}$ статистики $\chi^2(\hat{p})$, определяемое по таблицам квантилей хи-квадрат распределения, равно $\chi^2_{0{,}90,\,(2)} = 4{,}61$. Так как выполняется неравенство $(\chi^2_{\text{в}} = 2{,}546) < (\chi^2_{0{,}90,\,(2)} = 4{,}61)$, то следует принять на уровне значимости $\alpha = 0{,}10$ нулевую гипотезу, согласно которой качество изделий не зависит от поставщика.

Пример 3.6.75. Обсудим проверку гипотезы о равенстве параметров двух биномиальных распределений. Пусть независимо проведено две серии испытаний с числом наблюдений $n_{1\bullet}$ и $n_{2\bullet}$ соответственно. В первой серии испытаний некоторое событие A произошло n_{11} раз, а во второй серии – n_{21} раз.

Требуется проверить нулевую гипотезу о том, что вероятность появления события A в обеих сериях одна и та же, то есть $H_0 : p_1 = p_2$. Результаты наблюдений для удобства записываются в табл. 3.6.5 размера 2×2.

Таблица 3.6.5

Серия \ Событие	A	$\overline{A}$	Сумма
1	n_{11}	n_{12}	$n_{1\bullet}$
2	n_{21}	n_{22}	$n_{2\bullet}$
Сумма	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet} = n$

Сформулированная гипотеза $H_0 : p_1 = p_2$ эквивалентна утверждению о том, что обе выборки получены из одной генеральной совокупности, то есть однородны, и проверяется с помощью критерия хи-квадрат, статистику которого (3.6.71) в данном случае удобнее вычислять по формуле

$$\chi^2(\hat{p}) = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{\bullet 1}n_{\bullet 2}n_{2\bullet}n_{1\bullet}}. \qquad \textbf{(3.6.76)}$$

Гипотеза $H_0 : p_1 = p_2$ принимается на уровне значимости α, если для вычисленного значения статистики (3.6.76) выполняется неравенство $\chi^2(\hat{p}) < \chi^2_{1-\alpha,(1)}$. Если же $\chi^2(\hat{p}) \geq \chi^2_{1-\alpha,(1)}$, то результаты наблюдений, представленные в табл. 3.6.5, противоречат гипотезе однородности и ее следует отклонить.

Замечание 3.6.77. Отметим, что выражение (3.6.72) обеспечивает нужную для практики точность при условии, что все ожидаемые частоты $\hat{n}_{ij} = (n_{i\bullet} \cdot n_{\bullet j} / n) > 3$, $i, j = 1,2,$ и общее число наблюдений $n > 20$. Для малых n при вычислении наблюденного значения статистики $\chi^2(\hat{p})$ по формуле (3.6.76) обычно n заменяют на $n-1$ и при этом должны выполняться неравенства $n_{1\bullet} > 5$ и $n_{2\bullet} > (n_{1\bullet}/3)$ (см. [19, с.308]). Например, при $n = 36$ и частотах $n_{11} = 3$, $n_{21} = 6$, $n_{12} = 10$, $n_{22} = 17$ минимальная из ожидаемых частот $\hat{n}_{11} = (n_{1\bullet} \cdot n_{\bullet 1} / n) = (13 \cdot 9 / 36) = 3,25 > 3$ и, следовательно, аппроксимация (3.6.72) применима. Отметим также, что эта аппроксимация может быть улучшена путем использования поправки на непрерывность. При этом числитель в (3.6.76) записывается в виде $n\{|\, n_{11}n_{12} - n_{12}n_{21}\,| - n/2\}^2$ (см., например, [9, с. 204] и упр. 5.8.58).

3.7. Критерии независимости и случайности

В данном разделе обсудим сначала критерии проверки гипотезы независимости, которая была сформулирована в примере 3.2.20, а затем рассмотрим критерии проверки гипотезы случайности (3.2.23).

Исходной информацией для решения задачи независимости являются статистические данные, полученные выборочным методом путем n независимых наблюдений над совокупностью объектов, каждый из которых характеризуется двумерным признаком (X,Y). Например, X характеризует рост, Y – вес человека; X – число выкуриваемых в день сигарет, признак Y – продолжительность жизни человека; X – кровяное давление, Y – вес человека и т.д. Отметим, что признаки X и Y могут быть как непрерывными, так и дискретными и могут быть как количественными, так и качественными (атрибутивными). Например, X – цвет глаз, Y – цвет волос человека; X – дневная выручка в магазине, Y – день недели; X – количество бракованных изделий за смену, Y – номер бригады работников и т.д.

Итак, пусть $(X_1,Y_1),...,(X_n,Y_n)$ – выборка из двумерного распределения $F_{XY}(x,y)$, $(x,y)\in R^2$. Требуется по этим данным проверить гипотезу H_0 о том, что случайные величины X и Y независимы, то есть

$$H_0 : F_{XY}(x,y) = F_X(x)F_Y(y). \qquad \textbf{(3.7.1)}$$

Критерии проверки данной гипотезы кратко называют критериями независимости, и они определяются характером рассматриваемой статистической модели. Сначала обсудим наиболее часто используемый на практике вариант, когда случайная величина (X,Y) характеризуется двумерным нормальным распределением (см. разд. 3.7.А). Статистический анализ данных таблиц сопряженности атрибутивных признаков X и Y обсуждается в разд. 3.7.Б (см. также разд. 3.8). Критерии независимости, основанные на рангах наблюдений, предназначенные для проверки гипотезы (3.7.1) в рамках непараметрических моделей обсуждаются в разд. 5.5.

3.7.А. Проверка гипотезы независимости в гауссовском случае

Пусть объекты изучаемой совокупности характеризуются двумерным признаком (X,Y). Имеется двумерная выборка $(X_1,Y_1),...,(X_n,Y_n)$ объема n из распределения $F_{XY}(x,y)$, полученная выборочным методом в результате n независимых наблюдений над объектами с двумерным признаком (X,Y). Предполагается, что статистическая модель наблюдений гауссовская, то есть двумерные случайные величины $(X_1,Y_1),...,(X_n,Y_n)$ независимы и одинаково распределены, причем $L(X_i,Y_i)=NN(a_X,a_Y,\sigma_X^2,\sigma_Y^2,\rho)$, $i=1,...,n$. Эта краткая запись означает, что совместная плотность $f_{XY}(x,y)$, $(x,y)\in R^2$, двумерной случайной величины (X,Y) записывается в виде

$$f_{XY}(x,y)=(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}\times\exp\{-[\tilde{x}^2-2\rho\tilde{x}\tilde{y}+\tilde{y}^2]/2(1-\rho^2)\}, \quad \textbf{(3.7.2)}$$

где параметры этого распределения имеют вполне определенный вероятностный смысл, а именно, $a_X=M(X)$, $a_Y=M(Y)$, $\sigma_X^2=D(X)$, $\sigma_Y^2=D(Y)$, $\rho-$коэффициент корреляции и для сокращения записи введены следующие обозначения: $\tilde{x}=(x-a_X)/\sigma_X$, $\tilde{y}=(y-a_Y)/\sigma_Y$. Напомним (см. упр. 1.4.22), что в гауссовском случае из факта некоррелированности, то есть из равенства $\rho=0$, следует независимость признаков X и Y, то есть следует равенство $f_{XY}(x,y)=f_X(x)\cdot f_Y(y)$. Поэтому, гипотеза независимости H_0, сформулированная в (3.7.1), в данном случае перепишется в виде

$$H_0:\rho=0. \quad \textbf{(3.7.3)}$$

В тех случаях, когда все отклонения от нулевой гипотезы представляют одинаковый интерес, рассматривают двустороннюю альтернативу

$$H_1:\rho\neq 0. \quad \textbf{(3.7.4)}$$

Данная задача проверки гипотез H_0 и H_1 сформулирована в рамках параметрической статистической модели со сложными гипотезой и альтернативой. Поэтому для ее решения следует построить крите-

рий отношения правдоподобия (к.о.п., см. разд. 3.4.Б). Введем обозначения. Пусть вектор неизвестных параметров обозначен через $\vec{\theta} = (a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho)$, тогда параметрические множества, определяемые предположениями гипотез H_0 и H_1, запишутся в виде $\Theta_0 = \{\vec{\theta} : \rho = 0\}$ и $\Theta_1 = \{\vec{\theta} : \rho \neq 0\}$. Для наблюденной реализации $\vec{z} = \{(x_1, y_1), ..., (x_n, y_n)\}$ двумерной выборки $\vec{Z} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ логарифмическая функция правдоподобия (см. упр. 2.15.10) запишется в виде

$$\ln L(\vec{z}; a_X, a_Y, \sigma_X^2, \sigma_Y^2, \rho) = -n \ln(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}) - \\ -\left\{\frac{1}{2(1-\rho^2)}\left[\frac{1}{\sigma_X^2}\sum_{i=1}^{n}(x_i - a_X)^2 - \frac{2\rho}{\sigma_X\sigma_Y}\sum_{i=1}^{n}(x_i - a_X)(y_i - a_Y) + \right.\right. \\ \left.\left. + \frac{1}{\sigma_Y^2}\sum_{i=1}^{n}(y_i - a_Y)^2\right]\right\}. \qquad \textbf{(3.7.5)}$$

Далее, максимизация функции правдоподобия без ограничений, то есть на параметрическом множестве $\Theta = \Theta_0 \bigcup \Theta_1$, приводит к оценкам максимального правдоподобия параметров (см. упр. 2.15.10) в виде $\hat{\theta} = (\hat{a}_X, \hat{a}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\rho})$, где $\hat{a}_X = \overline{X}$, $\hat{a}_Y = \overline{Y}$ – выборочные средние; $\hat{\sigma}_X^2 = n^{-1}\sum(X_i - \overline{X})^2$, $\hat{\sigma}_Y^2 = n^{-1}\sum(Y_i - \overline{Y})^2$ – выборочные дисперсии и выборочный коэффициент корреляции определяется выражением $\hat{\rho} = \sum(X_i - \overline{X})(Y_i - \overline{Y}) / n\hat{\sigma}_X^2 \cdot n\hat{\sigma}_Y^2$. Далее, при максимизации функции правдоподобия на параметрическом множестве Θ_0, то есть при $\rho = 0$, мы имеем две независимые выборки $X_1, ..., X_n$ и $Y_1, ..., Y_n$. Поэтому оценки максимального правдоподобия $\hat{\theta}_0$ можно получить, максимизируя функции правдоподобия для каждой из выборок $X_1, ..., X_n$ и $Y_1, ..., Y_n$ отдельно (см. пример 2.10.7). Итак, в случае независимых выборок, то есть при $\rho = 0$, получаем $\hat{\theta}_0 = (\hat{a}_X, \hat{a}_Y, \hat{\sigma}_X^2, \hat{\sigma}_Y^2, 0)$. Таким образом, логарифм статистики отношения правдоподобия (3.4.16) запишется в виде

$$\ln \Lambda(\vec{Z};\Theta_0) = \ln\left\{\frac{\sup\{L(\vec{Z}\,;\theta):\theta\in\Theta\}}{\sup\{L(\vec{Z}\,;\theta):\theta\in\Theta_0\}}\right\} = \ln L(\vec{Z}\,;\hat{\theta}) - \ln L(\vec{Z}\,;\hat{\theta}_0) =$$

$$= \{-n\ln(2\pi\hat{\sigma}_X^2\hat{\sigma}_Y^2) - 2^{-1}n\ln(1-\hat{\rho}^2) - n\} - \{-n\ln(2\pi\hat{\sigma}_X^2\hat{\sigma}_Y^2) - n\} =$$

$$= -2^{-1}n\ln(1-\hat{\rho}^2). \qquad \textbf{(3.7.6)}$$

Отсюда следует, что $\ln\Lambda(Z\,;\Theta_0)$ – возрастающая функция от $\hat{\rho}^2$, а это означает, что критерий отношения правдоподобия отвергает гипотезу независимости $H_0:\rho=0$ при больших значениях $|\hat{\rho}|$.

Таким образом, для построения критерия проверки гипотезы независимости $H_0:\rho=0$ необходимо конкретизировать критическую область статистики критерия в виде выборочного коэффициента корреляции

$$\hat{\rho} = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2\sum(Y_i-\overline{Y})^2}}, \qquad \textbf{(3.7.7)}$$

а для этого требуется знать распределение $\hat{\rho}$ при гипотезе H_0. Отметим, что точное распределение выборочного коэффициента корреляции $\hat{\rho}$ при гипотезе H_0 очень сложное (см., например, [36]), по этой причине на практике обычно используют статистику вида

$$T_n = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}, \qquad \textbf{(3.7.8)}$$

которая при больших объемах выборки имеет распределение Стьюдента с числом степеней свободы $(n-2)$, то есть для нее справедливо выражение

$$L(T_n\,|\,H_0) = St(n-2) \text{ при } n\to\infty. \qquad \textbf{(3.7.9)}$$

С учетом этого выражения и того, что $|T_n|$ – возрастающая функция от $|\hat{\rho}|$, при больших объемах выборки *двусторонний* критерий проверки гипотезы $H_0:\rho=0$ против альтернативы $H_1:\rho\neq 0$ определяется критической областью размера α в виде

$$\mathrm{X}_{1\alpha} = \{\,\vec{z}:|t_n|>t_{1-(\alpha/2),(n-2)}\,\}, \qquad \textbf{(3.7.10)}$$

где $t_{1-(\alpha/2),(n-2)}$ обозначает квантиль уровня $(1-\alpha/2)$ распределения Стьюдента с числом степеней свободы $(n-2)$ и наблюденное значение t_n статистики T_n вычисляется по формуле (3.7.8), в которой вместо выборочного коэффициента корреляции $\hat{\rho}$ стоит его наблюденное значение r, вычисленное по наблюденной реализации $\vec{z}=\{(x_1,y_1),...,(x_n,y_n)\}$ двумерной выборки $\vec{Z}=\{(X_1,Y_1),...,(X_n,Y_n)\}$.

Таким образом, критерий проверки рассматриваемых гипотез H_0 и H_1 записывается в виде следующего правила.

Если для наблюденного значения t_n статистики T_n выполняется неравенство $|t_n|>t_{1-(\alpha/2),(n-2)}$, то гипотеза H_0 отвергается на уровне значимости α, если выполняется обратное неравенство, то нет оснований отвергать гипотезу H_0. **(3.7.11)**

Замечание 3.7.12. Отметим, что проверка гипотезы H_0 может быть проведена и непосредственно на статистике $\hat{\rho}$. В этом случае критерий запишется в виде следующего правила.

Если для модуля наблюденного значения r статистики $\hat{\rho}$ выполняется неравенство $|r|>r_{1-(\alpha/2),(n-2)}$, то гипотеза H_0 отвергается на уровне значимости α, если выполняется обратное неравенство, то нет оснований отвергать гипотезу H_0. Здесь $r_{1-(\alpha/2),(n-2)}$ обозначает критические значения статистики $\hat{\rho}$, вычисляемые через квантили $t_{1-(\alpha/2),(n-2)}$ по формуле

$$r_{1-(\alpha/2),(n-2)}=\frac{t_{1-(\alpha/2),(n-2)}}{\sqrt{(n-2)+t^2_{1-(\alpha/2),(n-2)}}}. \qquad \textbf{(3.7.13)}$$

Примеры проверки гипотезы независимости приводятся в упр. 3.9.70)

3.7.Б. Проверка гипотезы о «силе» связи между случайными величинами *X* и *Y* в гауссовском случае

В тех случаях, когда гипотеза $H_0:\rho=0$ о независимости признаков X и Y отвергнута, то есть между этими признаками существует значимая стохастическая линейная связь, возникает необходимость оценить «силу» или «тесноту» этой связи. Для этой цели

проверяются гипотезы

$$\tilde{H}_0 : \rho = \rho_0,\ \tilde{H}_1 : \rho \neq \rho_0, \qquad \textbf{(3.7.14)}$$

где ρ_0 – заданная величина, причем $-1 < \rho_0 < +1$.

Для проверки гипотез $\tilde{H}_0$ и $\tilde{H}_1$ естественно воспользоваться выборочным коэффициентом корреляции $\hat{\rho}$. Однако при выполнении неравенства $\rho \neq 0$ простой формы распределения вероятностей для статистик $\hat{\rho}$ (или T_n) не существует. Таблицы распределения статистики $\hat{\rho}$, составленные Дэвидом, при конечных объемах выборки: $n \leq 25$, $n = 50$, $n = 100$, $n = 200$, $n = 400$ могут быть найдены в работе [David F.N. Tables of the Correlation Coefficient // Biometrical Tables for Statisticians. Cambridge University Press, 1966. V. 1]. При других объемах выборки обычно используют асимптотические аппроксимации. Для этой цели Фишер предложил нормализующее преобразование вида

$$Z = \frac{1}{2}\ln\frac{1+\hat{\rho}}{1-\hat{\rho}}. \qquad \textbf{(3.7.15)}$$

Хорошо известно (см., например, [44]), что распределение статистики Z при справедливости гипотезы H_0 является асимптотически нормальным, то есть при $n \to \infty$ выполняется выражение

$$L(Z \mid H_0) = N\left\{\frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)}; \frac{1}{n-3}\right\}, \qquad \textbf{(3.7.16)}$$

где для статистики Z математическое ожидание $\xi_0 = M(Z \mid \tilde{H}_0)$ вычисляется по формуле

$$\xi_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)}. \qquad \textbf{(3.7.17)}$$

С учетом выражения (3.7.16), статистика критерия проверки гипотез $\tilde{H}_0$ и $\tilde{H}_1$ вида

$$\Lambda = (Z - \xi_0)\sqrt{n-3}, \qquad \textbf{(3.7.18)}$$

имеет стандартное нормальное распределение при гипотезе $\tilde{H}_0$, то есть справедливо выражение

$$L(\Lambda \mid \tilde{H}_0) = N(0\,,1) \text{ при } n \to \infty\,. \qquad \textbf{(3.7.19)}$$

Таким образом, при больших объемах выборки критерий проверки гипотез $\tilde{H}_0$ и $\tilde{H}_1$, основанный на статистике $\Lambda = (Z - \xi_0)\sqrt{n-3}$, определяется критической областью

$$\mathrm{X}_{1\alpha} = \{\, \vec{z} : |\,\lambda\,| > \lambda_{1-\alpha/2}\,\}\,, \qquad \textbf{(3.7.20)}$$

где $\lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ – квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения и наблюденное значение λ статистики Λ вычисляется по формуле $\lambda = (z - \xi_0)\sqrt{n-3}$, в которой наблюденное значение z статистики Z вычисляется по формуле $z = \{\ln(1+r) - \ln(1-r)\}/2$, а наблюденное значение r выборочного коэффициента корреляции $\hat{\rho}$ вычисляется по наблюденной реализации $(x_1, y_1),...,(x_n, y_n)$ двумерной выборки $(X_1, Y_1),...,(X_n, Y_n)$ по формуле (3.7.7). *P-значение* критерия равно удвоенной площади под стандартной нормальной плотностью справа от наблюденного значения λ статистики Λ, то есть *P-значение* $= 2[1 - \Phi(\lambda)]$. Если *P-значение* меньше уровня значимости α, то гипотеза $\tilde{H}_0$ отвергается.

Рассмотрим числовой пример. Пусть имеется двумерная выборка $(X_1, Y_1),...,(X_n, Y_n)$ объема $n = 50$ из двумерного нормального распределения. Требуется проверить гипотезы $\tilde{H}_0 : \rho = 0{,}4$, $\tilde{H}_1 : \rho \neq 0{,}4$ по наблюденным статистическим данным $(x_1, y_1),...,(x_{50}, y_{50})$, согласно которым наблюденное значение выборочного коэффициента корреляции равно $r = 0{,}35$. Для проверки сформулированных гипотез нужно выполнить следующие действия.

1. В табл. 12 работы [44] находим значение z статистики Z

$$z = \{\ln(1+r) - \ln(1-r)\}/2 = \{\ln(1+0{,}35) - \ln(1-0{,}35)\}/2 = 0{,}3654\,.$$

2. Используя эту же таблицу, вычисляем ξ_0 по формуле (3.7.16)

$$\xi_0 = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)} = \frac{1}{2}\ln\frac{1+0{,}4}{1-0{,}4} + \frac{0{,}4}{2\cdot 49} = 0{,}4236 + \frac{0{,}4}{2\cdot 49} = 0{,}4277.$$

3. Вычисляем наблюденное значение λ статистики Λ по формуле

$$\lambda = (z - \xi_0)\sqrt{n-3} = (0{,}3654 - 0{,}4277)\sqrt{47} = -0{,}4271.$$

4. Критерий. Пусть уровень значимости $\alpha = 0{,}05$, тогда по таблицам квантилей стандартного нормального распределения (см., например, [6, 44]) определяем $\lambda_{1-\alpha/2} = \lambda_{0.975} = 1{,}960$. Далее, так как $(|\lambda| = 0{,}4271) < (\lambda_{0{,}975} = 1{,}960)$, то нет оснований возражать на уровне значимости $\alpha = 0{,}05$ против гипотезы $\tilde{H}_0 : \rho = 0{,}4$, то есть наблюденные данные $(x_1, y_1), ..., (x_{50}, y_{50})$, согласно которым $r = 0{,}35$, не противоречат этой гипотезе. *P-значение* критерия равно:

$$P\text{-значение} = 2[1 - \Phi(0{,}427)] = 2[1 - 0{,}6653] = 0{,}669.$$

3.7.В. Статистический анализ таблиц сопряженности атрибутивных признаков

Пусть изучаемая совокупность состоит из объектов, которые характеризуются двумерным атрибутивным признаком (X, Y), причем признак X принимает k различных уровней $\tilde{X}_1, ..., \tilde{X}_k$, а признак Y принимает m различных уровней $\tilde{Y}_1, ..., \tilde{Y}_m$ с вероятностями

$$p_{ij} = P\{X = \tilde{X}_i, Y = \tilde{Y}_j\},\ i = 1, ..., k,\ j = 1, ..., m.$$

Отметим, что сумма вероятностей p_{ij}, $i = 1, ..., k$, $j = 1, ..., m$, равна единице, то есть

$$\sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} = 1. \qquad \textbf{(3.7.21)}$$

Определим маргинальные вероятности

$$P\{X = \tilde{X}_i\} = p_{i\bullet} = \sum_{j=1}^{m} p_{ij},\ i = 1, ..., k,$$

и

$$P\{Y = \tilde{Y}_j\} = p_{\bullet j} = \sum_{i=1}^{k} p_{ij},\ j = 1, ..., m. \qquad \textbf{(3.7.22)}$$

Отметим, что

$$\sum_{i=1}^{k} p_{i\bullet} = \sum_{j=1}^{m} p_{\bullet j} = 1. \qquad \textbf{(3.7.23)}$$

Статистические данные представляется с помощью таблицы сопряженности 3.7.1 размера $k \times m$ для признаков X и Y, которая заполняется после группировки n наблюденных объектов по k уровням признака X и m уровням признака Y.

Таблица 3.7.1

$X \setminus Y$	$\tilde{Y}_1$	…	$\tilde{Y}_j$	…	$\tilde{Y}_m$	Σ
$\tilde{X}_1$	n_{11}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
⋮	⋮				⋮	⋮
$\tilde{X}_i$	n_{i1}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
⋮	⋮				⋮	⋮
$\tilde{X}_k$	n_{k1}	…	n_{kj}	…	n_{km}	$n_{k\bullet}$
Σ	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	$n_{\bullet\bullet}$

В этой таблице n_{ij}, $i = 1,...,k$, $j = 1,...,m$, обозначают частоты появления объектов с *i*-м уровнем $\tilde{X}_i$ признака X и с *j*-м уровнем $\tilde{Y}_j$ признака Y в серии из n наблюдений, проведенных выборочным методом. Общее число наблюдений n равно сумме частот, то есть

$$n_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = n, \qquad \textbf{(3.7.24)}$$

а суммы частот (или накопленные частоты) по строкам и по столбцам обозначены соответственно через

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij}, \quad i = 1,...,k, \quad n_{\bullet j} = \sum_{i=1}^{k} n_{ij}, \quad j = 1,...,m. \qquad \textbf{(3.7.25)}$$

Отметим, что на практике вероятности $\{p_{ij}\}$, $i = 1,...,k$, $j = 1,...,m$, которые являются параметрами полиномиального распределения вероятностей вида

$$P\left(\{n_{ij}\}, \{p_{ij}\}\right) = \frac{n!}{n_{11}! \ldots n_{km}!} \prod_{i=1}^{k} \prod_{j=1}^{m} p_{ij}^{n_{ij}}, \qquad \textbf{(3.7.26)}$$

обычно неизвестны. Итак, в рамках этой параметрической модели

могут быть решены различные статистические задачи, которые связаны с проверкой гипотез о вероятностях $\{p_{ij}\}$, $i=1,...,k$, $j=1,...,m$, и оценкой этих параметров. Эти задачи решаются на основе данных эксперимента, представленных в таблице сопряженности признаков X и Y. В частности, на основе этих данных могут быть решены задачи согласия и независимости, которые обсуждаются ниже (см. также задачу однородности в разд. 3.6).

3.7.Г. Задача согласия

Эта задача сводится к проверке гипотезы вида

$$H_0^*: p_{ij} = p_{ij}^{(0)},\ i=1,...,k,\ j=1,...,m, \qquad \textbf{(3.7.27)}$$

здесь $p_{ij}^{(0)}$ – заданные вероятности полиномиального распределения (3.7.26). Альтернативная гипотеза записывается в виде

$$H_1^*: p_{ij} \neq p_{ij}^{(0)},\ i=1,...,k,\ j=1,...,m. \qquad \textbf{(3.7.28)}$$

Проверка гипотезы H_0^* против альтернативы H_1^* осуществляется по данным таблицы сопряженности признаков с помощью критерия хи-квадрат, статистика которого в данном случае записывается в виде

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij} - np_{ij}^{(0)})^2}{np_{ij}^{(0)}}. \qquad \textbf{(3.7.29)}$$

Статистика χ^2 из (3.7.29) при справедливости гипотезы H_0^* имеет асимптотическое распределение хи-квадрат с числом степеней свободы $mk-1$, то есть при $n\to\infty$ справедливо следующее выражение:

$$L(\chi^2 \mid H_0^*) = \chi^2(mk-1). \qquad \textbf{(3.7.30)}$$

С учетом этого выражения, критическая область размера α критерия согласия проверки гипотезы H_0^*, основанного на статистике (3.7.29), запишется в виде

$$\mathrm{X}_{1\alpha} = (\{n_{ij}\}: \chi^2 > \chi^2_{1-\alpha,(mk-1)}), \qquad \textbf{(3.7.31)}$$

где $\chi^2_{1-\alpha,(mk-1)}$ – квантиль уровня $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы $(mk-1)$. Таким образом, правило проверки гипотез H_0^*, H_1^* из (3.7.27) и (3.7.28) формулируется следующим образом.

«Если наблюденное значение статистики χ^2 превышает квантиль уровня $1-\alpha$ распределения хи-квадрат с числом степеней свободы $(mk-1)$, то гипотеза H_0^* отвергается на уровне значимости α, в противном случае нет оснований отвергать эту гипотезу».

3.7.Д. Проверка гипотезы независимости атрибутивных признаков

Отметим, что на практике вероятности $\{p_{ij}\}$ появления объектов с признаками $(\tilde{X}_i, \tilde{Y}_j)$, $i=1,...,k$, $j=1,...,m$, являются неизвестными параметрами полиномиального распределения вероятностей (3.7.26). Таким образом, статистическая модель наблюдений, представленных в таблице сопряженности признаков, задается в параметрической форме с точностью до числа km неизвестных параметров $\{p_{ij}\}$, $i=1,...,k$, $j=1,...,m$. Требуется по данным таблицы сопряженности признаков проверить гипотезу независимости признаков X и Y, которая записывается в виде

$$H_0^{**}: P\{X=\tilde{X}_i; Y=\tilde{Y}_j\} = P\{X=\tilde{X}_i\}\cdot P\{Y=\tilde{Y}_j\},$$
$$i=1,...,k,\ j=1,...,m,$$

или более кратко, с использованием введенных ранее обозначений, она записывается в виде

$$H_0^{**}: p_{ij} = p_{i\bullet}\cdot p_{\bullet j},\ i=1,...,k,\ j=1,...,m. \qquad \textbf{(3.7.32)}$$

Если равенство в (3.7.32) нарушено хотя бы для одной пары (i,j), то признаки X и Y зависимы. Таким образом, альтернативная гипотеза утверждает, что признаки X и Y зависимы, и записывается в виде

$$H_1^{**}: p_{ij} \neq p_{i\bullet}\cdot p_{\bullet j},\ i=1,...,k,\ j=1,...,m, \qquad \textbf{(3.7.33)}$$

Отметим, что гипотеза H_0^{**} является сложной, так как параметры p_{ij}, $p_{i\bullet}$ и $p_{\bullet j}$, $i=1,...,k$, $j=1,...,m$, неизвестны. По этой причине, эту гипотезу, как обычно (см. разд. 3.5, замечание 3.5.15, а также [2; 9, с. 198]), предварительно сводят к простой гипотезе путем построения оценок максимального правдоподобия $\hat{p}_{i\bullet}$ и $\hat{p}_{\bullet j}$ для неизвестных параметров $p_{i\bullet}$ и $p_{\bullet j}$ по данным таблицы сопряженности признаков. С использованием этих оценок, гипотеза H_0^{**} проверяется с помощью статистики χ^2 вида (3.7.29), в которой вместо ожидаемых теоретических частот np_{ij}, $i=1,...,k$, $j=1,...,m$, при справедливости гипотезы H_0^{**}, используются оценки максимального правдоподобия этих частот, также вычисленные в предположении независимости признаков X и Y, то есть при справедливости гипотезы H_0^{**}.

Отметим, что оценки максимального правдоподобия $\hat{p}_{i\bullet}$ и $\hat{p}_{\bullet j}$ неизвестных вероятностей $p_{i\bullet}$ и $p_{\bullet j}$, вычисленные по данным таблицы сопряженности в предположении, что признаки X и Y независимы, записываются, как и следовало ожидать (детали см. в упр. 3.9.72), в виде относительных частот событий $X=\tilde{X}_i$, $i=1,...,k$ и $Y=\tilde{Y}_j$, $j=1,...,m$, то есть

$$\hat{p}_{i\bullet}=n_{i\bullet}/n,\ i=1,...,k,\ \ \hat{p}_{\bullet j}=n_{\bullet j}/n,\ j=1,...,m. \qquad \textbf{(3.7.34)}$$

Таким образом, оценки максимального правдоподобия теоретических частот np_{ij}, $i=1,...,k$, $j=1,...,m$, при справедливости гипотезы H_0^{**} запишутся в виде

$$n\hat{p}_{i\bullet}\cdot\hat{p}_{\bullet j}=n_{i\bullet}n_{\bullet j}/n,\ i=1,...,k,\ j=1,...,m. \qquad \textbf{(3.7.35)}$$

Отсюда следует, что статистика χ^2 вида (3.7.29) в модифицированной форме, полученной путем замены неизвестных гипотетических частот на их оценки максимального правдоподобия, запишется в виде

$$\hat{\chi}^2=\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij}-n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n}.$$

С учетом (3.7.24) и (3.7.25) получаем более удобный для вычислений вариант этой статистики в виде

$$\hat{\chi}^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} = n\sum_{i=1}^{k}\sum_{j=1}^{m}\left(\frac{n_{ij}^2}{n_{i\bullet}n_{j\bullet}} - \frac{2n_{ij}}{n} + \frac{n_{i\bullet}n_{\bullet j}}{n^2}\right) =$$

$$= n\left(\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{n_{ij}^2}{n_{i\bullet}n_{\bullet j}} - 1\right). \qquad \textbf{(3.7.36)}$$

Для построения критерия проверки гипотезы о независимости признаков X и Y с помощью модифицированной статистики $\hat{\chi}^2$ необходимо конкретизировать ее распределение вероятностей при гипотезе H_0^{**}. Согласно теореме (3.5.19), статистика $\hat{\chi}^2$ при гипотезе H_0^{**} имеет асимптотическое распределение хи-квадрат с числом степеней свободы $(k-1)(m-1)$, то есть при $n \to \infty$ справедливо выражение

$$L(\hat{\chi}^2 \mid H_0^{**}) = \chi^2((k-1)(m-1)). \qquad \textbf{(3.7.37)}$$

Замечание 3.7.38. Напомним, что число степеней свободы хи-квадрат распределения равно числу независимых слагаемых в сумме квадратов стандартных нормальных случайных величин, которое равно числу *независимых* наблюдений. В данном случае наблюдениями являются частоты $\{n_{ij}\}$, $i = 1,...,k$, $j = 1,...,m$, их общее число равно mk, а число независимых частот равно $mk-1$, так как имеется одна связь (3.7.24). Кроме того, при построении статистики $\hat{\chi}^2$ использовалось k оценок неизвестных параметров $p_{i\bullet}$, $i = 1,...,k$, построенных по $k-1$ независимым частотам $\{n_{i\bullet}\}$, так как имеется одна связь (первое равенство в (3.7.25)), и m оценок параметров $p_{\bullet j}$, $j = 1,..., m$, построенных по $m-1$ независимым частотам $\{n_{\bullet j}\}$, так как имеется одна связь (второе равенство в (3.7.25)). Таким образом, согласно теореме Фишера 3.5.19, число степеней свободы для распределения хи-квадрат модифицированной статистики (3.7.36) при справедливости гипотезы H_0^{**} равно

$$(mk-1) - (k-1) - (m-1) = (k-1)(m-1). \qquad \textbf{(3.7.39)}$$

Итак, правило проверки гипотезы независимости H_0^{**} формулируется следующим образом:

«Если $\hat{\chi}^2 > \chi^2_{1-\alpha,\,(k-1)(m-1)}$, *то гипотеза* H_0^{**} *отвергается на уровне значимости* α, *в противном случае, нет оснований отвергать гипотезу* H_0^{**}*».*

Здесь $\chi^2_{1-\alpha,\,(k-1)(m-1)}$ – квантиль уровня $1-\alpha$ распределения хи-квадрат с числом степеней свободы $(k-1)(m-1)$.

Пример 3.7.40. В этом примере убедимся, что гипотеза о независимости двух атрибутивных признаков таблицы сопряженности (2×2) может быть сформулирована в виде гипотезы о равенстве долей некоторого признака в двух совокупностях объектов. Пусть для определенности рассматриваются две партии товара, которые обозначены через A_1 – 1-я партия товара, A_2 – 2-я партия товара, и пусть признаки B_1 – брак, а B_2 – не брак. В данном примере признак A характеризует партию товара с двумя уровнями A_1 и A_2, признак B характеризует качество товара с двумя уровнями B_1 и B_2, Статистические данные приведены в таблице сопряженности 2×2 вида

$A \setminus B$	B_1 – брак	B_2 – не брак	Всего
A_1	n_{11}	n_{12}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$n_{2\bullet}$
Всего	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Гипотеза независимости признаков A и B в данном случае запишется в виде

$$H_0^{**}: p_{ij} = p_{i\bullet}\cdot p_{\bullet j},\ \ i,j=1,2. \qquad \textbf{(3.7.41)}$$

Рассмотрим два случая этой гипотезы:

1) $i=1$, $j=1$, $p_{11} = p_{1\bullet}\cdot p_{\bullet 1}$ или $p_{\bullet 1} = \dfrac{p_{11}}{p_{1\bullet}}$;

2) $i=2$, $j=1$, $p_{21} = p_{2\bullet}\cdot p_{\bullet 1}$ или $p_{\bullet 1} = \dfrac{p_{21}}{p_{2\bullet}}$.

Итак, при справедливости гипотезы H_0^{**} из (3.7.41) выполняется равенство $(p_{11}/p_{1\bullet}) = (p_{21}/p_{2\bullet})$. Выясним смысл этого равенства. Левая его часть характеризует долю бракованного товара в первой партии, так как

$$\frac{p_{11}}{p_{1\bullet}} = \frac{P\{\text{Брак; 1-я партия}\}}{P\{\text{1-я партия}\}} = P\{\text{Брак}|\text{1-я партия}\},$$

а вторая его часть характеризует долю бракованного товара во второй партии, так как

$$\frac{p_{21}}{p_{2\bullet}} = \frac{P\{\text{Брак; 2-я партия}\}}{P\{\text{2-я партия}\}} = P\{\text{Брак} \mid \text{2-я партия}\}.$$

Таким образом, гипотеза H_0^{**} из (3.7.41) о независимости признаков A и B в данном примере может быть переформулирована в виде утверждения о том, что доли признака (в данном случае доли брака) равны в обеих совокупностях, то есть в виде гипотезы однородности

$$\tilde{H}_0^{**} : P\{\text{Брак}|\text{1-я партия}\} = P\{\text{Брак} \mid \text{2-я партия}\}. \qquad \textbf{(3.7.42)}$$

Проверка гипотезы однородности $\tilde{H}_0^{**}$ осуществляется с помощью статистики $\hat{\chi}^2$, которая в данном случае запишется в виде

$$\hat{\chi}^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2 \cdot n}{n_{\bullet 1}n_{\bullet 2}n_{2\bullet}n_{1\bullet}}. \qquad \textbf{(3.7.43)}$$

Отметим, что при больших объемах выборки выполняется выражение $L(\hat{\chi}^2 \mid \tilde{H}_0^{**}) = \chi^2(1)$. На практике данную аппроксимацию применяют, если выполняется неравенство $n_{i\bullet}n_{\bullet j}/n \geq 5$ (см. [9, с. 204]). С учетом этого замечания критерий проверки гипотезы однородности $\tilde{H}_0^{**}$ запишется в виде следующего правила:

«Если вычисленное значение статистики $\hat{\chi}^2$ превышает квантиль $\chi^2_{1-\alpha,(1)}$ уровня $1-\alpha$ распределения хи-квадрат с одной степенью свободы, то гипотеза $\tilde{H}_0^{**}$ отвергается на уровне значимости α, в противном случае, нет оснований отвергать гипотезу $\tilde{H}_0^{**}$».

Пример 3.7.44. Рассмотрим числовой пример, статистические данные для которого взяты из [9, с. 201]. Данные о колониях пищевых бактерий, содержащихся в трех видах пирожных, в виде частот $\{n_{ij}\}$, $i, j = 1,2,3$, полученные в $n = 368$ опытах, приведены в таблице

Пирожное	Колонии			
	малые	средние	большие	Всего
Эклер	92 (104,620)	37 (33,764)	46 (36,617)	175
Наполеон	53 (52,011)	15 (16,785)	19 (18,204)	87
Ореховое	75 (63,370)	19 (20,451)	12 (22,179)	106
Всего	220	71	77	368

По наблюденным данным частот $\{n_{ij}\}$, $i, j = 1,2,3$, представленным в таблице, требуется проверить гипотезу, согласно которой утверждается, что не существует зависимости между видами пирожных и размерами бактериаль-

ных колоний, содержащихся в них. Отметим, что в этой таблице в круглых скобках приведены оценки ожидаемых частот, вычисленные по формулам $n_{i\bullet}n_{\bullet j}/n$, $i,j=1,2,3$. Например, при $i=j=1$ имеем $220\cdot175/368=104{,}620$. Итак, в данном примере $m=3$, $k=3$ и общее число наблюдений $n=368$. Далее, используя исходные данные частот и вычисленные оценки ожидаемых частот, вычисляем наблюденное значение статистики хи-квадрат по формуле

$$\hat{\chi}^2=\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(n_{ij}-n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n}=\frac{(92-104{,}62)^2}{104{,}62}+\ldots+\frac{(12-22{,}179)^2}{22{,}179}=$$
$$=1{,}5223+\ldots+4{,}6716=11{,}389.$$

При нулевой гипотезе эта статистика приближенно распределена по закону хи-квадрат с числом степеней свободы $(k-1)(m-1)=(3-1)(3-1)=4$. Используя таблицы интеграла вероятностей χ^2 в [6, с. 140], находим $P\{\chi^2(4)\geq 11{,}389\}\approx 0{,}022$. Таким образом, достигнутый уровень значимости или *P-значение* $=0{,}022<0{,}05$ и, следовательно, гипотезу о независимости отвергаем на уровне значимости $\alpha=0{,}05$, то есть следует признать, что распределение бактериальных колоний зависит от вида пирожных.

Пример 3.7.44а. Из 300 абитуриентов, поступивших в институт, 97 человек имели по математике оценку «5» в школе и 48 человек получили оценку «5» на вступительных экзаменах в вуз, причем только 18 человек имели оценку «5» и в школе и на вступительном экзамене. Проверить гипотезу независимости выставляемых оценок «5» в школе и в вузе.

Составим таблицу сопряженности двух признаков. Признак A – оценка в вузе, B – оценка в школе.

$A \setminus B$	«5»	«не 5»	Всего
«5»	18	30	48
«не 5»	79	173	252
Всего	97	203	300

Наблюденное (вычисленное) значение $\hat{\chi}^2_{\text{в}}$ статистики $\hat{\chi}^2$ из (3.7.43)

$$\hat{\chi}^2_{\text{в}}=\frac{(n_{11}n_{22}-n_{12}n_{21})^2\cdot n}{n_{\bullet 1}n_{\bullet 2}n_{2\bullet}n_{1\bullet}}=\frac{300(18\cdot173-30\cdot79)^2}{97\cdot203\cdot252\cdot48}=0{,}697.$$

При уровне значимости $\alpha=0{,}05$ критическое значение $\chi^2_{1-\alpha,(1)}$ статистики $\hat{\chi}^2$ равно $\chi^2_{0{,}95,(1)}=3{,}84$. Так как выполняется неравенство $(\hat{\chi}^2_{\text{в}}=0{,}697)<(\chi^2_{0{,}95,(1)}=3{,}84)$, то гипотезу о независимости выставляемых оценок «5» в школе и в вузе принимаем на уровне значимости $\alpha=0{,}05$.

Отметим, что в данном примере достигнутый уровень значимости (наименьший уровень, на котором может быть отвергнута нулевая гипотеза независимости признаков A и B, в пользу альтернативы об их зависимости) или $P\text{-значение} = P\{\chi^2(1) \geq 0,697\} \approx 0,79$.

Пример 3.7.44б. Для определения факта зависимости цвета волос жителей от их местожительства были обследованы выборочным методом три группы людей из районов A, B, C. Проверить на уровне значимости $\alpha = 0,05$ гипотезу о независимости цвета волос жителей от их местожительства по следующим данным.

Район	Цвет волос			
	Рыжий	Блондин	Брюнет	Всего
A	2 (4)	9 (6)	9 (10)	20
B	3 (6)	6 (9)	21 (15)	30
C	15 (10)	15 (15)	20 (25)	50
Всего	20	30	50	100

Отметим, что в этой таблице в круглых скобках приведены оценки ожидаемых частот, вычисленные по формулам $n_{i\bullet}n_{\bullet j}/n$, $i, j = 1,2,3$. Наблюденное (вычисленное) значение $\hat{\chi}^2_{\text{в}}$ статистики $\hat{\chi}^2$ равно

$$\hat{\chi}^2_{\text{в}} = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n} = \{(2-4)^2/4\} + \{(9-6)^2/6\} + \{(9-10)^2/10\} +$$
$$+\{(3-6)^2/6\} + \{(6-9)^2/9\} + \{(21-15)^2/15\} +$$
$$+\{(15-10)^2/10\} + \{(15-15)^2/15\} + \{(20-25)^2/25\} = 11.$$

При $k = 3$, $m = 3$, число степеней свободы $(k-1)(m-1) = 4$ и для $\alpha = 0,05$ критическое значение статистики $\hat{\chi}^2$ равно $\chi^2_{1-\alpha,(k-1)(m-1)} = \chi^2_{0,95,(4)} = 9,49$. Так как выполняется неравенство $(\hat{\chi}^2_{\text{в}} = 11) > (\chi^2_{0,95,(4)} = 9,49)$, то гипотезу о независимости цвета волос жителей от их местожительства отвергаем на уровне значимости $\alpha = 0,05$, то есть цвет волос жителей зависит от районов A, B, C их проживания.

3.7.E. Проверка гипотезы случайности

При решении большинства статистических задач неизменно предполагается, что исходные статистические данные $\vec{X} = (X_1,...,X_n)$ представляют собой *случайную выборку* из некоторого распределения F_X, то есть предполагается, что компоненты

X_i, $i=1,...,n$, вектора $\vec{X}$ являются *независимыми и одинаково распределенными* случайными величинами. Часто эти предположения могут быть оправданы условиями проведения статистического эксперимента, однако в некоторых случаях они требуют дополнительной проверки. Для осуществления этого решается задача случайности (см. разд. 3.2), которая сводится к проверке гипотезы о том, что совместная функция распределения $F_{\vec{X}}(x_1,...,x_n)$ случайного вектора $\vec{X}=(X_1,...,X_n)$ факторизуется, то есть основная гипотеза записывается в виде

$$H_0:\ F_{\vec{X}}(x_1,...,x_n)=\prod_{i=1}^{n}F_X(x_i), \qquad \textbf{(3.7.45)}$$

где F_X – некоторая одномерная функция распределения. Для проверки гипотезы случайности разработаны различные критерии, как для параметрических моделей (в частности, для нормальной модели мы обсудим критерий Аббе), так и для непараметрических, которые применимы для любых непрерывных распределений. Мы рассмотрим критерий инверсий (см. пример 3.7.49), см. также критерий серий (5.8.3) и упр. 5.8.16. Обсудим сначала критерий Аббе, который предназначен для проверки гипотезы H_0 о независимости одинаково нормально распределенных случайных величин $X_1,...,X_n$, параметры которых μ и σ^2 неизвестны. Отметим, что в литературе (см., например, [9]) критерий Аббе также называют критерием среднего квадратов последовательных разностей.

Критерий Аббе. По статистическим данным $X_1,...,X_n$ требуется проверить гипотезу H_0, которая записывается в виде

$$H_0: X_1,...,X_n \ \textit{независимы},\ \textit{причем}\ L(X_i)=N(\mu,\sigma^2),\ \forall i\in(1,...,n).$$

Отметим, что данная гипотеза является частным случаем гипотезы (3.7.45) и может быть записана также в виде

$$H_0:\ F_{\vec{X}}(x_1,...,x_n)=(\sqrt{2\pi}\sigma)^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

Обычно в качестве альтернативы H_1 для гипотезы H_0 рассматривают гипотезу, согласно которой каждое последующее наблюдение

положительно коррелировано с предыдущим. Отметим, что сформулированные гипотезы также можно записать и в терминах математических ожиданий, то есть в виде

$$H_0 : M(X_1) = M(X_2) = ... = M(X_n) = \mu,$$
$$H_1 : |M(X_{i+1}) - M(X_i)| > 0, \text{ для всех значений } i = 1,2,...,n-1. \quad \textbf{(3.7.46)}$$

Для построения статистики критерия проверки гипотез H_0 и H_1 используют две различные оценки параметра σ^2. В качестве одной несмещенной (см. упр. 2.15.13) оценки параметра σ^2 используется подправленная выборочная дисперсия вида

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2\right\}.$$

В качестве второй несмещенной оценки параметра σ^2 используется оценка вида

$$\hat{S}^2 = \frac{1}{2(n-1)}\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2,$$

которая при справедливости гипотезы H_0 о независимости и одинаковой нормальной распределенности случайных величин $X_1,...,X_n$ также является несмещенной оценкой параметра σ^2, так как

$$M(\hat{S}^2 \mid H_0) = \frac{1}{2(n-1)}\sum_{i=1}^{n-1} M[(X_{i+1} - \mu) - (X_i - \mu)]^2 =$$
$$= \frac{1}{2(n-1)}\sum_{i=1}^{n-1}\{D(X_{i+1}) - 2COV(X_{i+1}, X_i) + D(X_i)\} = \sigma^2.$$

Статистика критерия Аббе определяется с помощью отношения этих оценок параметра σ^2 и записывается в виде

$$S = \hat{S}^2 / \tilde{S}^2 = \frac{1}{2}\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Отметим, что математическое ожидание и дисперсия статистики $S = \hat{S}^2 / \tilde{S}^2$ при гипотезе H_0 вычисляются (см., например, [9]) по формулам

$$M(S \mid H_0) = 1,\ D(S \mid H_0) = \frac{n-2}{n^2-1},$$

причем асимптотическое распределение статистики

$$(S-1)\sqrt{(n^2-1)/(n-2)}$$

при гипотезе H_0 является стандартным нормальным, то есть справедливо выражение

$$L\{\,(S-1)\sqrt{(n^2-1)/(n-2)}\ \mid H_0\,\} = N(0\,;1) \text{ при } n \to \infty. \quad \textbf{(3.7.47)}$$

Отметим, что нормальная аппроксимация дает приемлемые для практики результаты уже для $n \geq 10$. Отметим также, что таблицы точного распределения статистики $S = \hat{S}^2 / \tilde{S}^2$ при гипотезе H_0 для различных объемов выборки $n = 4,...,60$ приводятся в работе [6]. При альтернативной гипотезе H_1, согласно которой каждое последующее наблюдение положительно коррелировано с предыдущим, все разности $X_{i+1} - X_i$ в среднем будут принимать меньшие значения, чем при гипотезе H_0, поэтому $M(\hat{S}^2 \mid H_1) < \sigma^2$, и это означает, что при альтернативе H_1 статистика $S = \hat{S}^2 / \tilde{S}^2$ имеет тенденцию принимать меньшие значения, чем при гипотезе H_0, то есть *статистика* $(S-1)\sqrt{(n^2-1)/(n-2)}$ *является левосторонней*. С учетом этого замечания, асимптотическая критическая область уровня значимости α критерия Аббе для проверки гипотез H_0 и H_1 при больших объемах выборки n записывается в виде $\mathrm{X}_{1\alpha}^{-} = \{\vec{x} : (s-1)\sqrt{(n^2-1)/(n-2)} \leq \lambda_\alpha\}$, $\lambda_\alpha = \Phi^{-1}(\alpha)$.

Пример 3.7.48. Рассмотрим числовой пример, статистические данные для которого взяты из [9, с. 210]. При определении плотности земли Кэвендиш приводит упорядоченные по времени получения следующие измерения: $x_1 = 5,36$, $x_2 = 5,29$, $x_3 = 5,58$, $x_4 = 5,65$, $x_5 = 5,57$, $x_6 = 5,53$, $x_7 = 5,62$, $x_8 = 5,29$, $x_9 = 5,44$, $x_{10} = 5,34$, $x_{11} = 5,79$, $x_{12} = 5,10$, $x_{13} = 5,27$, $x_{14} = 5,39$, $x_{15} = 5,42$, $x_{16} = 5,47$, $x_{17} = 5,63$, $x_{18} = 5,34$, $x_{19} = 5,46$, $x_{20} = 5,30$, $x_{21} = 5,75$, $x_{22} = 5,68$, $x_{23} = 5,85$. Для данного примера $n = 23$ и $\tilde{S}^2 = 0,03626$,

$\hat{S}^2 = 0,03097$, откуда наблюденное значение s статистики $S = \hat{S}^2 / \tilde{S}^2$ равно $s = 0,85412$. Далее, наблюденное значение z_n асимптотически нормальной статистики $Z_n = (S-1)\sqrt{(n^2-1)/(n-2)}$ равно

$$z_n = (s-1)\sqrt{(n^2-1)/(n-2)} = (0,85412-1)\cdot\sqrt{25,14286} = \\ = -0,14588\cdot 5,01427 = -0,732.$$

Пусть уровень значимости $\alpha = 0,05$, тогда $\lambda_\alpha = \lambda_{0,05} = \Phi^{-1}(0,05) = -1,645$. Так как наблюденное значение $z_n = -0,732$ превышает критическое значение $\lambda_{0,05} = -1,645$, то на уровне значимости $\alpha = 0,05$ следует признать, что наблюденные данные при определении плотности земли не противоречат гипотезе независимости и нормальной распределенности. Вычислим теперь *P-значение* критерия, основанного на *левосторонней* статистике $Z_n = (S-1)\sqrt{(n^2-1)/(n-2)}$.

$$\textit{P-значение} = P\{Z_n \le z_n \mid H_0\} = \Phi(z_n) = \Phi(-0,732) = 1-\Phi(0,732) = 1-0,768 = 0,232.$$

Полученное *P-значение* критерия, основанного на левосторонней статистике Z_n, значительно превосходит типичный уровень $\alpha = 0,05$, поэтому гипотезу H_0 о независимости и одинаковой нормальной распределенности следует принять.

Отметим, что для проверки гипотезы нормальности приведенных данных также может быть применен критерий Шапиро – Уилка (см. разд. 4.8). Для этого исходные данные запишем в виде вариационного ряда:

$$x_{(1)} = 5,10,\ x_{(2)} = 5,27,\ x_{(3)} = 5,29,\ x_{(4)} = 5,29,\ x_{(5)} = 5,30,\ x_{(6)} = 5,34,$$
$$x_{(7)} = 5,34,\ x_{(8)} = 5,36,\ x_{(9)} = 5,39,\ x_{(10)} = 5,42,\ x_{(11)} = 5,46,\ x_{(12)} = 5,47,$$
$$x_{(13)} = 5,53,\ x_{(14)} = 5,54,\ x_{(15)} = 5,57,\ x_{(16)} = 5,58,\ x_{(17)} = 5,62,\ x_{(18)} = 5,63,$$
$$x_{(19)} = 5,65,\ x_{(20)} = 5,68,\ x_{(21)} = 5,75,\ x_{(22)} = 5,79,\ x_{(23)} = 5,85.$$

Для вычисления статистики критерия Шапиро – Уилка по формуле (4.8.3) сначала по таблице X в работе [71] при $n = 23$ находим весовые коэффициенты:

$$a_{23} = 0,4542,\ a_{22} = 0,3126,\ a_{21} = 0,2563,\ a_{20} = 0,2139,\ a_{19} = 0,1787,\ a_{18} = 0,1480,$$
$$a_{17} = 0,1201,\ a_{16} = 0,0941,\ a_{15} = 0,0696,\ a_{13} = 0,0459,\ a_{12} = 0,0228.$$

Далее вычисляем сумму в числителе формулы (4.8.3), которая равна

$$\sum_{i=1}^{[n/2]} a_{n-i+1}(x_{(n-i+1)} - x_{(i)}) = 0,4542(5,85-5,10) + 0,3126(5,79-5,27) + \ldots \\ +0,0228(5,53-5,46) = 0,884.$$

Отметим, что $\sum x_i = 126,22$, выборочное среднее $\overline{x} = 126,22/23 = 5,49$, $\sum x_i^2 = 693,47$, а сумма квадратов $SS = \sum (x_i - \overline{x})^2$ равна

$$SS = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2 / n = 693,47 - (126,22)^2 / 23 = 0,798.$$

С учетом этих данных, наблюденное значение $w(n)$ статистики $W(n)$ из (4.8.3) равно

$$w(n) = \frac{\left\{\sum_{i=1}^{[n/2]} a_{n-i+1}(x_{(n-i+1)} - x_{(i)})\right\}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{(0,884)^2}{0,798} = 0,979.$$

Далее, из табл. XI в работе [71] при $n = 23$ находим $P\{W(n) \le 0,962 \mid H_0\} = 0,5$. Это означает, что квантиль уровня (0,5) распределения статистики $W(n)$ при справедливости нулевой гипотезы и при $n = 23$ равен 0,962. Учитывая, что статистика $W(n)$ левосторонняя и выполняется неравенство $0,962 < (w(n) = 0,979)$, гипотезу о нормальности распределения принимаем.

Отметим, что для вычисления *P-значения* $W(n)$-критерия используется аппроксимирующее распределение S_B Джонсона и вычисляется реализация нормальной случайной величины по формуле $z = \gamma + \eta \ln\left(\frac{w(n) - \varepsilon}{1 - w(n)}\right)$, где параметры γ, η и ε табулированы в [71]. Пусть $L(Z) = N(0, 1)$, тогда с использованием предыдущего выражения *P-значение* $W(n)$-критерия вычисляется по формуле $P\text{-значение} = P(Z \le z \mid H_0)$. Для рассматриваемого примера имеем

$$z = \gamma + \eta \ln\left(\frac{w(n) - \varepsilon}{1 - w(n)}\right) = -5,508 + 1,848 \ln\left(\frac{0,979 - 0.2157}{1 - 0,979}\right) = 1,04,$$

и $$P\text{-значение} = P(Z \le z \mid H_0) = P(Z \le 1,04) = \Phi(1,04) = 0,85.$$

Вычисленное *P-значение* $W(n)$-критерия также подтверждает вывод, что результаты эксперимента Кэвендиша не противоречат гипотезе о нормальности распределения .

Пример 3.7.49. Критерий инверсий для проверки гипотезы случайности. Этот критерий применяется для любой непрерывной функции распределения вероятностей и основан на использовании понятия инверсии, которую образуют наблюденные данные $x_1, ..., x_n$ в вариационном ряду $x_{(1)}, ..., x_{(n)}$, который в свою очередь является наблюденной реализацией упорядоченной статистики $\vec{X}_{(\cdot)} = (X_{(1)}, ..., X_{(n)})$ наблюдаемого случайного векто-

ра $\vec{X} = (X_1, ..., X_n)$. Если гипотеза случайности (3.7.45) имеет место, то компоненты вектора $\vec{X} = (X_1, ..., X_n)$ «равноправны» в том смысле, что они образуют случайную выборку, то есть являются независимыми и одинаково распределенными случайными величинами, поэтому наблюденные данные $x_1, ..., x_n$ не должны быть упорядочены ни в каком смысле. Это означает, что ситуацию, соответствующую гипотезе случайности H_0, можно охарактеризовать как «полный беспорядок». Следовательно, критерий проверки гипотезы H_0 можно построить на основе статистики, которая измеряет степень «беспорядка» исходных данных. Одна из таких статистик (см., например, [22]) подсчитывает число инверсий в случайном векторе $\vec{X} = (X_1, ..., X_n)$. Обозначим эту статистику через $T_n(\vec{X})$ и поясним ее смысл через процесс построения реализации $T_n(\vec{x})$ этой статистики, то есть ее наблюденного значения, которое вычисляется по наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ случайного вектора $\vec{X} = (X_1, ..., X_n)$. Введем необходимые определения.

Определение 3.7.50. Говорят, что в перестановке чисел $1, 2, ..., n$ любая пара чисел образует инверсию, если в этой перестановке большее число стоит впереди меньшего.

Например, при $n = 5$ перестановка $3, 5, 2, 1, 4$ содержит шесть инверсий: (3, 2), (3, 1), (5, 2), (5, 1), (5, 4), (2, 1). Отметим, что общее число перестановок чисел $1, 2, ..., n$ равно $n!$. Определим теперь понятие инверсии для пары чисел x_i и x_j наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ случайного вектора $\vec{X} = (X_1, ..., X_n)$.

Определение 3.7.51. Говорят, что пара чисел x_i и x_j наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ случайного вектора $\vec{X} = (X_1, ..., X_n)$ образует инверсию, если $i < j$, но x_i стоит правее x_j в вариационном ряду $x_{(1)} < , ..., < x_{(n)}$, то есть инверсия имеет место, когда наблюдению с *меньшим* номером соответствует *большее* значение.

Для вычисления общего числа инверсий в $\vec{x} = (x_1, ..., x_n)$ определим число инверсий y_i, образованных компонентой x_i, $i = 1, ..., n-1$, в наблюденной реализации $\vec{x} = (x_1, ..., x_n)$.

Определение 3.7.52. Число инверсий y_i, образованных компонентой x_i, $i = 1, ..., n-1$, в наблюденной реализации $\vec{x} = (x_1, ..., x_n)$ равно числу наблюдений с большими, чем у x_i, номерами и стоящими левее x_i в вариационном ряду, то есть

$$y_i = \{\textit{числу}\ (x_{i+1}, ..., x_n), \textit{которые} < x_i\},\ i = 1, ..., n-1. \qquad \textbf{(3.7.53)}$$

Например, пусть $n=5$ и пусть наблюденная реализация $x_1,...,x_5$ образует следующий вариационный ряд:

$$(x_{(1)}=x_3) < (x_{(2)}=x_5) < (x_{(3)}=x_2) < (x_{(4)}=x_1) < (x_{(5)}=x_4).$$

Тогда $y_1=3$, $y_2=2$, $y_3=0$, $y_4=1$. Общее число инверсий в данной реализации $x_1,...,x_5$ равно $y_1+y_2+y_3+y_4=3+2+0+1=6$.

Отметим, что числа инверсий y_i, образованные компонентой x_i, $i=1,...,n-1$, в наблюденной реализации $\vec{x}=(x_1,...,x_n)$ являются возможными состояниями дискретной случайной величины Y_i, $i=1,...,n-1$, которая характеризует число инверсий, образованных компонентой X_i в случайном векторе $\vec{X}=(X_1,...,X_n)$. Таким образом, возможными состояниями случайной величины Y_i являются целочисленные значения $0,1,...,n-i$, которые она может принимать с одной и той же вероятностью $1/(n-i+1)$, $i=1,...,n-1$, то есть ее распределение вероятностей при справедливости гипотезы случайности H_0 записывается в виде

$$\begin{pmatrix} Y_i \\ \vec{p} \end{pmatrix} : \begin{pmatrix} 0, & ..... & ,n-i \\ \dfrac{1}{n-i+1}, & ..., & \dfrac{1}{n-i+1} \end{pmatrix}, \; i=1,...,n-1. \qquad \textbf{(3.7.54)}$$

Отметим, что введенная случайная величина Y_i, $i=1,...,n-1$, определяется расположением компоненты X_i по отношению к $X_{i+1},...,X_n$ в упорядоченной статистике $\vec{X}_{(\cdot)}=(X_{(1)},...,X_{(n)})$ и не зависит от относительного расположения $X_{i+1},...,X_n$ между собой, то есть случайная величина Y_i, при любом $i=1,...,n-2$ не зависит от $Y_{i+1},...,Y_{n-1}$. Таким образом, случайные величины $Y_1,...,Y_{n-1}$ взаимно независимы.

Статистику критерия $T_n(\vec{X})$ для проверки гипотезы случайности H_0 определяют в виде общего числа инверсий для случайного вектора $\vec{X}=(X_1,...,X_n)$, то есть в виде

$$T_n(\vec{X})=\sum_{i=1}^{n-1} Y_i, \qquad \textbf{(3.7.55)}$$

где Y_i – случайная величина, характеризующая число инверсий, образованных компонентой X_i, $i=1,...,n-1$, в случайном векторе $\vec{X}=(X_1,...,X_n)$, определенная распределением (3.7.54). Отметим, что статистика $T_n(\vec{X})$ может принимать целочисленные значения $T_n(\vec{x})$, которые меняются от нуля до

$n(n-1)/2$. Минимальное значение равное нулю она принимает, если для наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выполняются неравенства $x_1<x_2<...<x_n$, и максимальное значение $n(n-1)/2=(n-1)+(n-2)+...+1$ принимает, если для наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выполняются неравенства $x_n<x_{n-1}<...<x_1$.

Учитывая, что статистика $T_n(\vec{X})$ является суммой независимых случайных величин $Y_1,...,Y_{n-1}$ с распределением (3.7.54), для вычисления ее моментов запишем выражение производящей функции случайной величины Y_i в виде

$$\varphi_i(z)=\sum_{r=0}^{n-i}P\{Y_i=r\}\,z^r=\frac{1}{(n-i+1)}(1+z+...+z^{n-i}),\ i=1,...,n-1.$$

Используя эту функцию, получаем выражения для математического ожидания и дисперсии случайной величины Y_i, $i=1,...,n-1$, в виде

$$M(Y_i)=\varphi_i^{/}(z)\,|_{z=1}=\frac{1}{n-i+1}(1+2z+...+(n-i)z^{n-i-1})\,|_{z=1}=\frac{(n-i)(n-i+1)}{(n-i+1)\cdot 2}=\frac{n-i}{2},$$

$$D(Y_i)=\varphi_i^{//}(z)\,|_{z=1}+M(Y_i)-\{M(Y_i)\}^2=\frac{(n-i)(n-i+2)}{12},\ i=1,...,n-1.$$

Учитывая эти результаты, независимость случайных величин $Y_1,...,Y_{n-1}$, используя формулы (см. упр. 5.8.56) вида

$$\sum_{i=1}^{n}i=n(n+1)/2,\ \ \sum_{i=1}^{n}i^2=n(n+1)(2n+1)/6, \qquad \textbf{(3.7.56)}$$

окончательно получаем с помощью (3.7.55), выражения для математического ожидания и дисперсии статистики $T_n(\vec{X})$ при справедливости гипотезы H_0:

$$M\{T_n(\vec{X})\}=\sum_{i=1}^{n-1}M(\eta_i)=n(n-1)/4,$$

$$D\{T_n(\vec{X})\}=\sum_{i=1}^{n-1}D(\eta_i)=(2n^3+3n^2-5n)/72. \qquad \textbf{(3.7.57)}$$

Таким образом, среднее значение статистики $T_n(\vec{X})$ при справедливости гипотезы случайности H_0 совпадает с серединой интервала $[0,\ n(n-1)/2]$. Наличие малого и большого числа инверсий в наблюденной реализации $\vec{x}=(x_1,...,x_n)$ свидетельствует против гипотезы H_0. По этой причине, крити-

ческая область $X_{1\alpha}$ размера α должна состоять из двух частей и в нее следует включить все целые точки этого интервала, достаточно удаленные от его середины $n(n-1)/4$, то есть критическая область запишется как

$$X_{1\alpha} = \{\vec{x} : |T_n(\vec{x}) - n(n-1)/4| \geq t_{1-\alpha/2}(n)\}.$$

Здесь $t_{1-\alpha/2}(n)$ – критическое значение статистики $T_n(\vec{X})$, которое выбирается из условия $P\{T_n(\vec{X}) \in X_{1\alpha} \mid H_0\} \leq \alpha$. Это условие также может быть переписано в следующем виде:

$$P\{n(n-1)/4 - t_{1-\alpha/2}(n) \leq T_n(\vec{X}) \leq n(n-1)/4 + t_{1-\alpha/2}(n) \mid H_0\} \geq 1-\alpha. \quad \textbf{(3.7.58)}$$

Критические значения $t_{1-\alpha/2}(n)$ статистики $T_n(\vec{X})$, являющиеся минимальными числами, удовлетворяющими выражению (3.7.58), при различных значениях $n = 2,3,...,12$, приводятся в [6]. Для больших объемов выборки n (на практике уже при $n > 10$) используют асимптотический вариант критерия. Этот критерий основан на нормированной статистике

$$\tilde{T}_n(\vec{X}) = \frac{T_n(\vec{X}) - M\{T_n(\vec{X})\}}{\sqrt{D\{T_n(\vec{X})\}}} = \{T_n(\vec{X}) - n(n-1)/4\} \cdot \{6/n^{3/2}\}, \quad \textbf{(3.7.59)}$$

которая при гипотезе H_0 имеет асимптотически стандартное нормальное распределение, то есть для нее справедливо следующее выражение:

$$L\{\tilde{T}_n(\vec{X}) \mid H_0\} = N(0;1) \text{ при } n \to \infty. \quad \textbf{(3.7.60)}$$

Отметим еще раз, что эта аппроксимация обеспечивает нужную для практики точность уже при $n \geq 10$.

Таким образом, асимптотическая критическая область размера α критерия, основанного на двусторонней статистике (3.7.59), запишется в виде

$$X_{1\alpha} = \{\vec{x} : |\tilde{T}_n(\vec{x})| \geq \lambda_{1-\alpha/2}\},\ \lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2), \quad \textbf{(3.7.61)}$$

и асимптотическое (практически уже при $n \geq 10$) *P-значение* критерия инверсий, основанного на этой статистике, вычисляется по формуле

$$\textit{P-значение} = 2\Pr\{\tilde{T}_n(\vec{X}) \geq \tilde{T}_n(\vec{x}) \mid H_0\} = 2[1 - \Phi\{\tilde{T}_n(\vec{x})\}]. \quad \textbf{(3.7.62)}$$

Если вычисленное по наблюденному значению $\tilde{T}_n(\vec{x})$ статистики $\tilde{T}_n(\vec{X})$ *P-значение* критерия окажется меньше уровня значимости α, то гипотезу случайности H_0 следует отвергнуть.

3.8. Основные понятия дисперсионного анализа

Задачи дисперсионного анализа (в англоязычной литературе сокращенно ANOVA) возникают при изучении влияния одного или нескольких факторов (признаков) на некоторый итоговый показатель. Если выделяется *только один фактор*, который влияет на итоговый показатель, и этот фактор принимает конечное число значений (уровней), то дисперсионный анализ называют *однофакторным*. В других случаях используют термины двухфакторный и многофакторный дисперсионный анализ. Типичным примером решения задач дисперсионного анализа является сравнение некоторых конечных результатов при различных способах действия (*при различных обработках*), направленных на достижение одной цели. Например, сопоставление результатов труда в различных бригадах, влияние различных лекарств на процесс выздоровления, сравнение выручки магазина в различные дни недели, сравнение урожайности при различных агротехнических приемах обработки почвы и т. д.

В разд. 3.7 рассматривалась задача проверки гипотезы однородности, или гипотезы о равенстве средних значений двух нормальных распределений по двум выборкам из этих распределений с помощью двухвыборочного t-критерия Стьюдента. Обобщение этой задачи на случай заданного числа $k > 2$ выборок приводит к статистическим процедурам, известным в литературе как дисперсионный анализ, который был разработан Фишером. Дисперсионный анализ является развитой системой статистических процедур с большим числом различных вариантов в постановках задач, которые в современном изложении основаны на использовании общей линейной модели наблюдений (см. [3, 5, 49, 69, 80]).

При решении задач дисперсионного анализа используют различные модели наблюдений. Для однофакторного дисперсионного анализа, когда выделяется только один классифицирующий признак, изучают два типа моделей, их в литературе называют модели I и II однофакторного дисперсионного анализа (см., например, [9, с. 283]). Для модели I (*фиксированные эффекты*) характерно, что математические ожидания выборочных значений в каждой выборке являются *неизвестными постоянными величинами*, равенство которых предполагается при формулировке основной гипотезы. Для

модели II (*случайные эффекты*) упомянутые математические ожидания истолковываются как *реализации независимых случайных величин*, дисперсия которых предполагается равной нулю при формулировке основной гипотезы. При решении задач двухфакторного дисперсионного анализа кроме моделей I и II изучают также *смешанные модели*, когда один фактор характеризуется фиксированными эффектами, а другой – характеризуется случайными эффектами.

Итак, пусть объекты изучаемой совокупности характеризуются некоторым итоговым показателем X и пусть $X_{1j},...,X_{n_j j}$ – выборочные значения, соответствующие j-й выборке объема n_j, $j=1,...,k$, полученной при j-м уровне A_j, $j=1,...,k$, факторного признака A. Отметим, что значения измеряемого признака X часто называют *откликом*, а факторный признак A называют классифицирующим признаком, в соответствии с которым изучаемая совокупность разбивается на k классов (или k групп, или k подпопуляций). Отметим также, что результирующий показатель X является количественным признаком, а фактор A может быть как количественным, так и качественным признаком. При осуществлении дисперсионного анализа исходных данных $\{X_{ij}\}$, $i=1,...,n_j$, $j=1,...,k$, обычно используют аддитивную модель наблюдений вида $X_{ij}=Z_j+\varepsilon_{ij}$, где Z_j и $\varepsilon_{1j},...,\varepsilon_{n_j j}$, $i=1,...,n_j$, $j=1,...,k$, – взаимно независимые случайные величины с нормальными распределениями. Для того чтобы проиллюстрировать различие моделей I и II приведем пример, взятый из [9, с. 283]. Предположим, что некий поставщик выпускает однотипные станки и интересуется вариациями качества деталей, изготовляемых на таких станках. Далее предположим, что потребитель приобретает k станков, выбрав их случайно из текущей продукции, а затем поставщик и потребитель анализируют качество деталей, изготовленных на этих станках. При этом потребителя интересуют только приобретенные им k станков, и он, естественно, будет пользоваться моделью I, т.е. будет считать значения Z_j неизвестными постоянными. Поставщика же в первую очередь интересует вся совокупность выпускаемых станков (а не

только k станков, поставленных потребителю) и, значит, ему полезно знать, какова дисперсия случайных величин Z_j, поэтому поставщику целесообразно пользоваться моделью II .

Итак, для модели I однофакторного дисперсионного анализа статистическая модель наблюдений основана на следующих предположениях. Исходные данные $\{X_{ij}\}$, $i=1,...,n_j$, $j=1,...,k$, представляют собой выборку, полученную в результате n независимых наблюдений над показателем X из k нормальных совокупностей со средними значениями $\mu_1,...,\mu_k$ и с равными, но неизвестными дисперсиями, то есть $\sigma_1^2=...=\sigma_k^2=\sigma^2$. Модель наблюдений записывается в виде

$$X_{ij}=\mu_j+\varepsilon_{ij},\ i=1,...,n_j,\ j=1,...,k,$$
$$n=\sum_{j=1}^{k}n_j,\ L(\varepsilon_{ij})=N(0;\sigma^2). \qquad \textbf{(3.8.1)}$$

В рамках модели II дисперсионного анализа предполагается, что

$$X_{ij}=Z_j+\varepsilon_{ij}, \qquad \textbf{(3.8.2)}$$

где Z_j и $\varepsilon_{1j},...,\varepsilon_{n_j j}$, $i=1,...,n_j$, $j=1,...,k$, – случайные величины, которые взаимно независимы и подчиняются нормальным распределениям, причем выполняются выражения

$$L(\varepsilon_{ij})=N(0;\sigma^2),\ L(Z_j)=N(0;\theta^2),\ i=1,...,n_j,\ j=1,...,k.$$

Таким образом, если рассматриваются лишь условные распределения величин $\{X_{ij}\}$ при фиксированном Z_j, то говорят, что имеет место модель I однофакторного дисперсионного анализа, если же речь идет о безусловном распределении $\{X_{ij}\}$, то говорят, что имеет место модель II. Модель I дисперсионного анализа с фиксированными эффектами используется тогда, когда экспериментатора интересуют подпопуляции, связанные именно с данными уровнями фактора. Причем при повторении эксперимента будут рассматриваться случайные выборки из тех же самых подпопуляций. Модель II дисперсионного анализа со случайными эффектами используется

тогда, когда подпопуляции, соответствующие различным уровням фактора, выбираются случайно из большого (бесконечного) числа подпопуляций. Поэтому при повторении эксперимента, скорее всего, мы будем иметь дело со случайными выборками из других подпопуляций.

В разд. 3.8.А мы ограничимся обсуждением лишь однофакторного дисперсионного анализа, когда все наблюдения классифицируются только по *одному признаку* и описываются *гауссовской моделью* I , определенной в (3.8.1). Если же предположения нормальности не могут быть обоснованы, то рассматривают непараметрические модели и используют свободные от распределений процедуры, основанные на рангах или знаках наблюдений. Обсуждение таких процедур приводится в разд. 5.6 и 5.7. Модель II однофакторного дисперсионного анализа со случайными эффектами подробно обсуждается в разд. 3.8.Г. Модели двухфакторного дисперсионного анализа рассматриваются в разд. 3.8.Д. Упражнения задачи и дополнения приведены в разд. 3.9.А.

3.8.А. Однофакторный дисперсионный анализ: модель I

Пусть объекты изучаемой совокупности (или популяции) W характеризуются некоторым результирующим показателем X. В соответствии с факторным признаком A (его еще называют классифицирующим признаком), который может принимать k значений $A_1,...,A_k$, вся совокупность W разбивается на k групп $W_1,...,W_k$ (или k слоев, или k подпопуляций $W_1,...,W_k$ популяции W). Статистическими данными являются наблюденные реализации $x_{11},...,x_{n_1 1},\ldots,x_{1k},...,x_{n_k k}$ k выборок $X_{11},...,X_{n_1 1}$, …, $X_{1k},...,X_{n_k k}$ из совокупностей $W_1,...,W_k$ с непрерывными распределениями изучаемого показателя X. Исходные данные кратко записывают в виде $\{X_{ij}\}$, $j=1,...,k$, $i=1,...,n_j$, они получены в результате n_j наблюдений за результирующим показателем X при каждом фиксированном j-м уровне A_j, $j=1,...,k$, фактора A и для удобства обычно заносятся в следующую таблицу «показатель – фактор».

Таблица 3.8.1

$X \setminus A$	A_1	…	A_j	…	A_k	Σ
1	X_{11}	…	X_{1j}	…	X_{1k}	
2	X_{21}	…	X_{2j}	…	X_{2k}	
⋮	⋮		⋮		⋮	
n_j	$X_{n_1 1}$	…	$X_{n_j j}$	…	$X_{n_k k}$	
Σ	$X_{\bullet 1}$	…	$X_{\bullet j}$	…	$X_{\bullet k}$	$X_{\bullet\bullet}$

Предполагаем, что модель наблюдений записывается в виде (3.8.1). Согласно этой модели, исходные данные $\{X_{ij}\}$, $i = 1,...,n_j$, $j = 1,...,k$, представляют собой выборку, полученную в результате n независимых наблюдений над показателем X из k нормальных совокупностей $W_1,...,W_k$ со средними значениями $\mu_1,...,\mu_k$ и с равными, но неизвестными дисперсиями $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. Эту модель наблюдений называют *нормальной* (*или гауссовской*) *моделью I однофакторного дисперсионного анализа с фиксированными эффектами*. Для удобства дальнейших ссылок выделим в явном виде и пронумеруем все предположения этой модели наблюдений:

$$X_{ij} = \mu_j + \varepsilon_{ij},\ i = 1,...,n_j,\ j = 1,...,k,\ n = (n_1 + \cdots + n_k), \qquad \textbf{(3.8.3)}$$

где **а)** $\mu_j = M(X \mid A = A_j)$, $j = 1,...,k$, – *постоянные* величины,

б) $\varepsilon_1,...,\varepsilon_n$ – *независимые* случайные величины,

в) $\varepsilon_1,...,\varepsilon_n$ – *нормальные* случайные величины, то есть $L(\varepsilon_{ij}) = N(0;\sigma^2)$,

г) дисперсии совокупностей $W_1,...,W_k$ *равны неизвестному параметру* σ^2, то есть $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$.

В рамках этой модели требуется убедиться в том, что изменение фактора A не влияет на итоговый показатель X. На статистическом языке эта задача сводится к проверке однородности наблюдаемых данных, которая в данном случае кратко записывается в виде проверки гипотез:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu,\ H_1 : \textit{не все}\ \mu_j\ \textit{равны}\ ,\ j = 1,...,k. \qquad \textbf{(3.8.4)}$$

Замечание 3.8.5. Отметим, что гауссовскую модель наблюдений (3.8.3) часто удобнее рассматривать в параметризованных вариантах. Используют два варианта параметризации, в зависимости от того, равны или нет объемы выборок $n_1,...,n_k$ из k совокупностей $W_1,...,W_k$. Если объемы выборок $n_1,...,n_k$ различны, то для параметризации модели (3.8.3) вводят средний уровень показателя X по всем совокупностям $W_1,...,W_k$ в виде

$$\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j\,,\ n = (n_1 + ... + n_k)\,. \qquad \textbf{(3.8.6)}$$

Далее вводятся дополнительные параметры $\xi_j = \mu_j - \mu$, которые характеризуют отклонение среднего уровня от j-й обработки, $j = 1,...,k$. Эти параметры называют *дифференциальными эффектами* обработки. Параметризованный вариант модели (3.8.3) записывается в виде

$$X_{ij} = \mu + \xi_j + \varepsilon_{ij}\,,\ i = 1,...,n_j\,,\ j = 1,...,k\,,\ n = (n_1 + \cdots + n_k)\,,$$
$$L(\varepsilon_{ij}) = N(0\,;\sigma^2)\,. \qquad \textbf{(3.8.7)}$$

Из (3.8.6) следует, что взвешенная сумма параметров $\xi_1,...\xi_k$ равна нулю, то есть

$$\sum_{j=1}^{k} n_j \xi_j = \sum_{j=1}^{k} n_j (\mu_j - \mu) = \sum_{j=1}^{k} n_j \mu_j - \mu \sum_{j=1}^{k} n_j = 0\,, \qquad \textbf{(3.8.8)}$$

и гипотеза $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ об однородности данных в терминах введенных параметров $\xi_1,...\xi_k$ переписывается в виде

$$H_0 : \xi_1 = \xi_2 = ... = \xi_k = 0\,. \qquad \textbf{(3.8.9)}$$

Отметим, что в модели (3.8.7) имеется $(k+1)$ параметров, однако общее количество независимых параметров остается равным k, так как взвешенная сумма параметров $\xi_1,...\xi_k$ равна нулю (см. 3.8.8). Другой вариант параметризации модели (3.8.3) используется, когда объемы выборок $n_1,...,n_k$ из k совокупностей $W_1,...,W_k$ одинаковы. Эти два варианта параметризации модели (3.8.3) связаны между собой, и оба способа идентичны, если все $n_1,...,n_k$ одинаковы (см. упр. 3.9.80).

Задачами дисперсионного анализа кроме проверки гипотез (3.8.4) и (3.8.9) также являются задачи построения точечных оценок параметров μ, σ^2, $\{\xi_j\}$ модели (3.8.7) и задачи построения доверительных интервалов для параметров $\{\mu_j\}$ и $\{\mu_i - \mu_j\}$, $i \neq j$. Для

решения этих задач удобно предварительно провести статистическую обработку наблюденных данных, записанных в таблицу «показатель – фактор». При статистической обработке данных $\{X_{ij}\}$, $j = 1,...,k$, $i = 1,...,n_j$, таблицы «показатель – фактор» используют следующие принятые обозначения и понятия. Общее число наблюдений (объем выборки) обозначают через $n = (n_1 + ... + n_k)$ и сумму наблюдений j-й группы обозначают через $X_{\bullet j} = X_{1j} + ... + X_{n_j j}$, $j = 1,...,k$, а общую сумму всех наблюдений обозначают в виде $X_{\bullet\bullet} = X_{\bullet 1} + ... + X_{\bullet k}$, то есть

$$X_{\bullet\bullet} = \sum_{j=1}^{k} X_{\bullet j} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}. \qquad \textbf{(3.8.10)}$$

Групповое выборочное среднее j-й группы определяется в виде

$$\overline{X}_{\bullet j} = \frac{X_{\bullet j}}{n_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}\ ,\ j = 1,...,k. \qquad \textbf{(3.8.11)}$$

Условное математическое ожидание с.в. X при условии, что фактор $A = A_j$, $j = 1,...,k$, обозначим через

$$\mu_j = M(X \mid A = A_j),\ j = 1,...,k. \qquad \textbf{(3.8.12)}$$

Отметим, что выборочное среднее j-й группы $\overline{X}_{\bullet j}$ является статистической оценкой условного математического ожидания μ_j, то есть

$$\hat{\mu}_j = \overline{X}_{\bullet j},\ j = 1,...,k. \qquad \textbf{(3.8.13)}$$

Для нормальной модели (3.8.3) эта оценка также является нормальной случайной величиной, то есть выполняется выражение

$$L(\overline{X}_{\bullet j}) = N(\mu_j, \sigma^2 / n_j),\ j = 1,...,k. \qquad \textbf{(3.8.14)}$$

Общее выборочное среднее определяется в виде

$$\overline{X}_{\bullet\bullet} = \frac{X_{\bullet\bullet}}{n} = \frac{1}{n} \sum_{j=1}^{k} X_{\bullet j} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}. \qquad \textbf{(3.8.15)}$$

Отметим, что общее выборочное среднее $\overline{X}_{\bullet\bullet}$ также может быть вычислено и на основе групповых средних $\overline{X}_{\bullet j}$ в виде их среднего взвешенного по объемам n_j, $j=1,...,k$, то есть в виде

$$\overline{X}_{\bullet\bullet}=\frac{1}{n}\sum_{j=1}^{k}n_j\frac{1}{n_j}\sum_{i=1}^{n_j}X_{ij}=\sum_{j=1}^{k}n_j\overline{X}_{\bullet j}/\sum_{j=1}^{k}n_j\,. \qquad \textbf{(3.8.16)}$$

Формулу (3.8.16) часто называют правилом сложения групповых выборочных средних $\overline{X}_{\bullet j}$.

При анализе общей и групповых дисперсий данных $\{X_{ij}\}$, $j=1,...,k$, $i=1,...,n_j$, таблицы «показатель – фактор» используют следующие понятия.

Общая сумма квадратов SS_T отклонений выборочных значений $\{X_{ij}\}$ относительно общего среднего $\overline{X}_{\bullet\bullet}$ определяется в виде

$$SS_T=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet\bullet})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2-X_{\bullet\bullet}^2/n\,. \qquad \textbf{(3.8.17)}$$

Внутригрупповая (*остаточная*) *сумма квадратов* SS_W (часто обозначается SS_R, *Rest – остаток*) отклонений выборочных значений $\{X_{ij}\}$ относительно групповых средних $\overline{X}_{\bullet j}$ определяется в виде

$$SS_W=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j})^2=\sum_{j=1}^{k}n_jS_j^2\,, \qquad \textbf{(3.8.18)}$$

где S_j^2 – групповая выборочная дисперсия j-й группы, определяемая в виде

$$S_j^2=n_j^{-1}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j})^2\,,\ j=1,...,k\,. \qquad \textbf{(3.8.19)}$$

Межгрупповая (*факторная*) *сумма квадратов* SS_B отклонений групповых средних $\overline{X}_{\bullet j}$ относительно общего среднего $\overline{X}_{\bullet\bullet}$ определяется в виде

$$SS_B = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 =$$

$$= \sum_{j=1}^{k} X_{\bullet j}^2 / n_j - X_{\bullet\bullet}^2 / n. \qquad \textbf{(3.8.20)}$$

Пояснение к обозначениям: B – Between (между), W – Within (внутри), R – Rest (остаток), T – Total (полный, общий).

Отметим, что однофакторные модели дисперсионного анализа с фиксированными эффектами вида (3.8.3) и (3.8.7) являются частными случаями общей линейной модели. Поэтому оценки параметров этих моделей могут быть получены методом наименьших квадратов. В частности, задача нахождения МНК-оценок параметров μ и $\{\xi_j\}$ модели (3.8.7) сводится к минимизации функции $S(\mu, \xi_j \mid \{x_{ij}\})$ по переменным μ и ξ_j, $j = 1,...,k$, при наблюденных данных $\{x_{ij}\}$ и при выполнении условия (3.8.8). Функции $S(\mu, \xi_j \mid \{x_{ij}\})$ определяется в виде

$$S(\mu, \xi_j \mid \{x_{ij}\}) = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \mu - \xi_j)^2 .$$

МНК-оценки параметров μ и $\{\xi_j\}$ являются решением системы уравнений

$$\partial S(\mu, \xi_j \mid \{x_{ij}\}) / \partial\mu = -2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \mu - \xi_j) = 0,$$

$$\partial S(\mu, \xi_j \mid \{x_{ij}\}) / \partial\xi_j = -2\sum_{i=1}^{n_j}(x_{ij} - \mu - \xi_j) = 0,\ j = 1,...,k .$$

Единственность этих оценок обеспечивается условием (3.8.8), и они записываются в виде

$$\hat{\mu} = \bar{X}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j\hat{\mu}_j , \qquad \textbf{(3.8.21)}$$

$$\hat{\xi}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} - \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet} = \hat{\mu}_j - \hat{\mu},\ j = 1,...,k. \quad \textbf{(3.8.22)}$$

Основой для проведения дисперсионного анализа при проверке гипотез (3.8.4) и (3.8.9) служит правило сложения введенных сумм квадратов. Это правило проявляется в том, что общая сумма квадратов распадается на две составные части, одна из которых связана с оценкой неизвестной дисперсии внутри групп, другая – с оценкой дисперсии между группами. Правило сложения введенных сумм квадратов записывается в виде

$$SS_T = SS_W + SS_B. \quad \textbf{(3.8.23)}$$

В справедливости этого правила легко убедиться. В самом деле, так как

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) = \sum_{j=1}^{k}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j}) =$$

$$= \sum_{j=1}^{k}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})(X_{\bullet j} - n_j \cdot X_{\bullet j}/n_j) = 0,$$

то
$$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(X_{ij} - \bar{X}_{\bullet j}) + (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})]^2 =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2 + \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = SS_W + SS_B.$$

Использование правила сложения (3.8.23) позволяет оценить различия между совокупностями $W_1,...,W_k$ путем сравнения вкладов в общую дисперсию, которые дают оба эти источника вариации внутри групп и между группами, это и послужило основой для введения F-статистики Фишера в виде отношения $F = S_B^2 / S_W^2$, где S_B^2 и S_W^2 – средние квадраты для соответствующих сумм квадратов отклонений, определяемые в виде

$$S_B^2 = SS_B/(k-1),\ S_W^2 = SS_W/(n-k). \qquad \textbf{(3.8.24)}$$

Отметим, что F-статистика Фишера основана на сопоставлении двух различных оценок неизвестной дисперсии σ^2. Одна из этих оценок (*средний квадрат S_W^2 суммы квадратов отклонений SS_W внутри групп*) является несмещенной оценкой неизвестной дисперсии σ^2, то есть $M\{S_W^2\} = \sigma^2$, и она работает вне зависимости от того, верна или нет гипотеза $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ (см. упр. 3.9.80). Другая оценка (*средний квадрат S_B^2 суммы квадратов отклонений SS_B между группами*) существенно использует это предположение, она «хорошо» оценивает параметр σ^2 (см. формулу (3.8.28) и упр. 3.9.80) только в том случае, когда справедлива гипотеза H_0. Сопоставление этих оценок в виде их отношения и приводит к F-статистике, которая позволяет определить, какая часть вариаций в экспериментальных данных вызывается различием в совокупностях $W_1,...,W_k$, а что может быть отнесено на счет случайных отклонений. Этот вывод основан на правиле сложения групповых выборочных дисперсий, которое после деления обеих частей выражения (3.8.23) на объем выборки n записывается в виде

$$SS_T/n = SS_W/n + SS_B/n \qquad \textbf{(3.8.25)}$$

и формулируется следующим образом.

Общая выборочная дисперсия данных $\{X_{ij}\}$, $j = 1,...,k$, $i = 1,...,n_j$, таблицы «показатель – фактор» равна сумме выборочной дисперсии внутри групп и выборочной дисперсии между группами.

Отметим, что выборочная дисперсия внутри групп SS_W/n записывается в виде среднего взвешенного по объемам групп n_j групповых выборочных дисперсий S_j^2, то есть в виде

$$SS_W/n = \sum_{j=1}^{k} n_j S_j^2 / \sum_{j=1}^{k} n_j\,. \qquad \textbf{(3.8.26)}$$

Аналогично, выборочная дисперсия между группами SS_B/n записывается в виде среднего взвешенного квадратов уклонений $(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})$, см. формулу (3.8.20). Отметим также, что выборочные дисперсии SS_W/n и SS_B/n являются смещенными оценками параметра σ^2. Несмещенной оценкой этого параметра является остаточный средний квадрат $S_W^2 = SS_W/(n-k)$, а оценки $S_B^2 = SS_B/(k-1)$ и $S_T^2 = SS_T/(n-1)$ обладают этим свойством только при справедливости гипотезы H_0 (см. замечание 3.8.27). По этой причине при построении F-статистики Фишера для проверки гипотез (3.8.4) используются независимые и несмещенные оценки (3.8.24) параметра σ^2 (см. также упр. 3.9.80).

Замечание 3.8.27. Использование модели (3.8.7) позволяет убедиться (см. упр. 3.9.80), что математическое ожидание оценки S_B^2 вычисляется по формуле

$$M\{S_B^2\} = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j \xi_j^2 . \qquad \textbf{(3.8.28)}$$

Из этой формулы следует, что при справедливости гипотезы $H_0 : \xi_1 = \xi_2 = \ldots = \xi_k = 0$ статистика S_B^2 является несмещенной оценкой параметра σ^2, то есть $M\{S_B^2 \mid H_0\} = \sigma^2$. Для наглядности и удобства дальнейших ссылок сведем приведенные формулы в табл. 3.8.2.

Таблица 3.8.2

Сумма кв. отклонений	Степени свободы	Средний квадрат	$M\{S_{(\cdot)}^2\}$
$SS_B = \sum_{j=1}^{k} n_j (\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2$	$(k-1)$	$S_B^2 = SS_B/(k-1)$	$M(S_B^2) = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j \xi_j^2$
$SS_W = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\bullet j})^2$	$(n-k)$	$S_W^2 = SS_W/(n-k)$	$M(S_W^2) = \sigma^2$
$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\bullet\bullet})^2$	$(n-1)$	$S_T^2 = SS_T/(n-1)$	$M(S_T^2) = \sigma^2 + \frac{1}{n-1}\sum_{j=1}^{k} n_j \xi_j^2$

Отметим, что при справедливости гипотезы $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ средние квадраты S_B^2 и S_W^2, то есть статистики

$$S_B^2 = SS_B/(k-1) = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2,$$

$$S_W^2 = SS_W/(n-k) = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2,$$

являются, согласно теореме Кокрена (П.3.77), независимыми случайными величинами, и, кроме того, их соответствующие суммы квадратов SS_B и SS_W имеют распределение хи-квадрат соответственно с числами степеней свободы $(k-1)$ и $(n-k)$, то есть (см., например, [69]) справедливы выражения

$$L(SS_B \mid H_0) = \chi^2(k-1) \text{ и } L(SS_W \mid H_0) = \chi^2(n-k).$$

Отсюда следует, что статистика $F = S_B^2 / S_W^2$ при гипотезе H_0 имеет F-распределение Фишера с числами степеней свободы $(k-1)$ и $(n-k)$ (см. П.2.94), то есть справедливо выражение

$$L\{F = S_B^2 / S_W^2 \mid H_0\} = F(k-1, n-k). \qquad (3.8.29)$$

Далее, если гипотеза H_0 не выполняется, то есть существуют два или несколько значений μ_j, отличных от параметра μ (и это означает, что будет верна альтернатива H_1), тогда будет выполняться неравенство $M\{S_B^2 \mid H_1\} > \sigma^2$ и отношение S_B^2 / S_W^2 уже не будет подчиняться F-распределению Фишера с числами степеней свободы $k-1$ и $n-k$, а будет смещено вправо, точнее, будет иметь нецентральное F-распределение. Отсюда следует, что статистика $F = S_B^2 / S_W^2$ является *правосторонней*, то есть ее критическая область размера α состоит из больших значений, точнее, находится справа от квантиля $F_{1-\alpha}(k-1, n-k)$ уровня $(1-\alpha)$ для F-распределения с числами степеней свободы $k-1$ и $n-k$.

Таким образом, критерий проверки гипотез (3.8.4), основанный на F-статистике Фишера, формулируется в виде следующего правила.

Если вычисленное по наблюденным данным $\{x_{i,j}\}$, $i=1,...,n_j$, $j=1,...,k$, *значение статистики* $F=S_B^2/S_W^2$ *превышает квантиль* $F_{1-\alpha}(k-1,n-k)$ *уровня* $(1-\alpha)$ *распределения Фишера с числами степеней свободы* $(k-1;n-k)$, *то гипотеза* $H_0:\mu_1=\mu_2=...=\mu_k=\mu$ *отвергается на уровне значимости* α, *в противном случае нет оснований отвергать гипотезу* H_0. **(3.8.30)**

Пример 3.8.31. Рассмотрим числовой пример. Пусть объекты изучаемой совокупности W характеризуются итоговым показателем X. В соответствии с четырьмя уровнями A_1, A_2, A_3, A_4 факторного признака A совокупность W разбита на четыре нормальные подсовокупности W_1, W_2, W_3, W_4 со средними значениями $\mu_1, \mu_2, \mu_3, \mu_4$, дисперсии которых предполагаются равными неизвестному параметру σ^2. Наблюденные данные $\{x_{ij}\}$, $i=1,...,7$, $j=1,...,4$, в виде реализаций выборок равного объема $n_j=7$, $j=1,...,4$, из нормальных совокупностей W_1, W_2, W_3, W_4 приведены в табл. 3.8.3.

Таблица 3.8.3

$X \setminus A$	A_1	A_2	A_3	A_4
1	72	74	78	69
2	69	72	74	67
3	63	70	70	66
4	59	69	58	64
5	59	66	58	62
6	53	58	56	58
7	51	52	56	54
Сумма Среднее	$X_{\bullet 1}=426$ 60,9	$X_{\bullet 2}=461$ 65,9	$X_{\bullet 3}=450$ 64,3	$X_{\bullet 4}=440$ 62,9

Требуется в рамках гауссовской модели (3.8.3) по наблюденным данным $\{x_{ij}\}$, $i=1,...,7$, $j=1,...,4$, приведенным в таблице (3.8.3), проверить гипотезу

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ об однородности данных совокупностей W_1, W_2, W_3, W_4. Для удобства вычисления наблюденного значения F-статистики Фишера по исходным данным табл. 3.8.3 заполним таблицу дисперсионного анализа. В данном примере число уровней фактора A равно $k = 4$, объемы выборок из совокупностей W_1, W_2, W_3, W_4 одинаковы, то есть $n_1 = n_2 = n_3 = n_4 = 7$, общий объем наблюдений $n = n_1 + n_2 + n_3 + n_4 = 28$. Далее, общая сумма наблюдений

$$X_{\bullet\bullet} = X_{\bullet 1} + X_{\bullet 2} + X_{\bullet 3} + X_{\bullet 4} = 426 + 461 + 450 + 440 = 1777.$$

Межгрупповая сумма квадратов отклонений

$$SS_B = \sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 =$$

$$= \sum_{j=1}^{k} n_j \bar{X}_{\bullet j}^2 + n\bar{X}_{\bullet\bullet}^2 - 2\bar{X}_{\bullet\bullet} \sum_{j=1}^{k} n_j \bar{X}_{\bullet j} = \sum_{j=1}^{k} X_{\bullet j}^2 / n_j - X_{\bullet\bullet}^2 / n =$$

$$= (426)^2/7 + (461)^2/7 + (450)^2/7 + (440)^2/7 - (1777)^2/28 = 94{,}9.$$

Общая сумма квадратов

$$SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - X_{\bullet\bullet}^2 / n =$$

$$= (72)^2 + (69)^2 + \ldots + (54)^2 - (1777)^2/28 = 1540{,}9.$$

Остаточная сумма квадратов внутри групп, согласно (3.8.12),

$$SS_W = SS_T - SS_B = 1540{,}9 - 94{,}9 = 1446{,}0.$$

Заполненная таблица дисперсионного анализа наблюдений данного примера выглядит следующим образом.

Таблица 3.8.4

Сумма кв. отклонений	Степени свободы	Средний квадрат	Наблюденное значение F-статистики
Межгрупповая $SS_B = 94{,}9$	$(k-1) = 3$	$SS_B/(k-1) = 31{,}6$	$F = \dfrac{SS_B/(k-1)}{SS_W/(n-k)} = \dfrac{31{,}6}{60{,}3} = 0{,}52$
Внутригрупповая $SS_W = SS_T - SS_B = 1446{,}0$	$(n-k) = 24$	$SS_W/(n-k) = 60{,}3$	
Общая $SS_T = 1540{,}9$	$(n-1) = 27$	$SS_T/(n-1) = 57{,}1$	

Пусть уровень значимости $\alpha = 0,05$. Используя таблицу квантилей F-распределения Фишера (см. [6, с. 208]), находим $F_{1-\alpha,(k-1;n-k)} = F_{0,95,(3,24)} = 3,01$. Так как выполняется неравенство $(F = 0,52) < (F_{0,95,(3,24)} = 3,01)$, то гипотезу $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ принимаем на уровне значимости $\alpha = 0,05$. Отметим, что $F_{0,5,(3,24)} = 0,81$, поэтому в данном случае P-*значение* $> 0,5$, что также убедительно говорит в пользу нулевой гипотезы. Так как основная гипотеза принята, то оценкой параметра μ является общее выборочное среднее, равное $\overline{X}_{\bullet\bullet} = X_{\bullet\bullet}/n = 1777/28 = 63,5$, а наилучшей выборочной оценкой дисперсии σ^2 является общий средний квадрат $S_T^2 = SS_T/(n-1) = 1540,9/27 = 57,1$.

Замечание 3.8.32. Если в результате проверки гипотеза $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ об однородности данных изучаемых нормальных совокупностей $W_1,...,W_k$ оказалась незначимой (то есть она не была отвергнута на заданном уровне значимости, как в рассмотренном примере 3.8.31), то в качестве несмещенной оценки параметра μ используется $\hat{\mu} = \overline{X}_{\bullet\bullet}$, а для несмещенной оценки параметра σ^2 – общий средний квадрат $\hat{\sigma}^2 = S_T^2 = SS_T/(n-1)$. Далее, в тех случаях, когда гипотеза $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ отклоняется, это означает, что существует зависимость между результирующим показателем X и меняющимся факторным признаком A, и чтобы получить количественное представление о степени влияния фактора A на показатель X, вычисляют оценки дифференциальных эффектов $\hat{\xi}_j = \overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet} = \hat{\mu}_j - \hat{\mu}$, $j = 1,...,k$, из (3.8.22) и коэффициент детерминации η^2, который отражает долю изменчивости показателя X, вызванную меняющимся фактором A. Выборочной и несмещенной оценкой дисперсии σ^2 в этом случае является остаточный средний квадрат $S_W^2 = SS_W/(n-k)$. Коэффициент детерминации η^2 определяют в виде отношения выборочной дисперсии SS_B/n между группами к общей выборочной дисперсии SS_T/n, то есть

$$\eta^2 = (SS_B/n)/(SS_T/n) = SS_B/SS_T. \qquad (3.8.33)$$

Если изменение фактора A не влияет на результирующий показатель X, тогда межгрупповая (факторная) выборочная дисперсия SS_B/n будет иметь малую долю в общей выборочной дисперсии SS_T/n и в этом случае отношение SS_B/SS_T будет стремиться к нулю. Если же изменение фактора A сильно влияет на результирующий показатель X, то отношение SS_B/SS_T будет стремиться к единице. Отметим, что в экономической практике, наряду с коэффициентом детерминации η^2, используют эмпирическое корреляционное

отношение η, которое определяется в виде

$$\eta = \sqrt{SS_B / SS_T} = \sqrt{1-(SS_W / SS_T)}\,. \qquad (3.8.34)$$

Эмпирическое корреляционное отношение, являясь положительным значением корня квадратного из коэффициента детерминации, удовлетворяет неравенствам $0 \le \eta \le 1$, причем, чем ближе η к 1, тем «теснее» связь результирующего показателя X и фактора A.

Отметим также, что величина влияния фактора оценивается с помощью коэффициента Снедекора (см., например, [65, с. 212]). При равных объемах выборок в группах, то есть при $n_1 = n_2 = ... = n_k = m$, этот коэффициент вычисляется по формуле $h_B^2 = (S_B^2 - S_W^2)/(S_B^2 + (m-1)S_W^2)$. Если же объемы выборок не являются равными, то в приведенной формуле вместо m следует поставить величину $\tilde{m} = [n - (n_1^2 + ... + n_k^2)/n]/(k-1)$, где n обозначает общее число наблюдений, то есть $n = n_1 + ... + n_k$ и k – число уровней фактора.

В тех случаях, когда *F*-критерий отклоняет гипотезу $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, это означает, что существует, по крайней мере, одна пара совокупностей W_i и W_j $(i \ne j)$, для которой $\mu_i \ne \mu_j$. Но, зная лишь значение *F*-статистики, мы не можем указать номера совокупностей i и j, для которых $\mu_i \ne \mu_j$. Чтобы ответить на этот вопрос, нужно проводить дополнительный анализ с целью узнать, средние значения каких совокупностей отличаются значимо. Это позволяют сделать процедуры множественного сравнения совокупностей, которые обсуждаются ниже (*S*-метод Шеффе, *T*-метод Тьюки и *t*-метод Стьюдента – Ньюмна – Кейлса (см., например, [3, 5, 9], а также критерий Дункана в [44, с. 157] и упр. 3.9.100).

3.8.Б. Доверительные интервалы

Для построения доверительных интервалов параметров μ_j, $j = 1,...,k$, гауссовской модели (3.8.3) отметим, что выборочное среднее *j*-й группы $\overline{X}_{\bullet j}$ является несмещенной и нормальной оценкой параметра μ_j, то есть $L(\overline{X}_{\bullet j}) = N(\mu_j, \sigma^2 / n_j)$, $j = 1,...,k$. Так как параметр σ^2 неизвестен, заменим его несмещенной оценкой S_W^2, определенной в (3.8.24), и рассмотрим центральную статистику (см. определение 2.12.2) для параметра μ_j в виде $(\overline{X}_{\bullet j} - \mu_j)/\sqrt{(S_W^2 / n_j)}$. Учиты-

вая, что для гауссовской модели (3.8.3) эта центральная статистика имеет распределение Стьюдента с числом степеней свободы $(n-k)$, то есть справедливо выражение

$$L\{(\overline{X}_{\bullet j}-\mu_j)/\sqrt{(S_W^2/n_j)}\}=St(n-k), \qquad \textbf{(3.8.35)}$$

получаем для любого μ_j, $j=1,...,k$, $(1-\alpha)$-доверительный интервал в виде

$$\overline{X}_{\bullet j}\pm t_{1-\alpha/2,(n-k)}\sqrt{S_W^2/n_j}\,,\ j=1,...,k, \qquad \textbf{(3.8.36)}$$

где $t_{1-\alpha/2,(n-k)}$ – квантиль уровня $(1-\alpha/2)$ распределения Стьюдента с числом степеней свободы $(n-k)$. Аналогично, учитывая, что $M\{\overline{X}_{\bullet i}-\overline{X}_{\bullet j}\}=\mu_i-\mu_j$, $i\neq j$, и $D\{\overline{X}_{\bullet i}-\overline{X}_{\bullet j}\}=\sigma^2(1/n_i-1/n_j)$, определим центральную статистику для параметра $(\mu_i-\mu_j)$ в виде $\{(\overline{X}_{\bullet i}-\overline{X}_{\bullet j})-(\mu_i-\mu_j)\}/\sqrt{S_W^2(1/n_i-1/n_j)}$, которая имеет распределение Стьюдента с числом степеней свободы $(n-k)$. С учетом этого получаем $(1-\alpha)$-доверительный интервал для сравнения двух средних $(\mu_i-\mu_j)$ в виде

$$(\overline{X}_{\bullet i}-\overline{X}_{\bullet j})\pm t_{1-\alpha/2,(n-k)}\sqrt{S_W^2(1/n_i-1/n_j)}\,,\ i,j=1,...,k,\ i\neq j. \qquad \textbf{(3.8.37)}$$

Замечание 3.8.38. Отметим, что доверительный интервал (3.8.28) отличается от доверительного интервала, построенного с помощью двухвыборочного критерия Стьюдента (см. формулу (3.6.50)) тем, что для оценивания параметра σ^2 используется оценка S_W^2, построенная *по всем наблюдениям из всех* k *групп* $W_1,...,W_k$, *в предположении, что дисперсии* $\sigma_1^2,...,\sigma_k^2$ *равны*. Это означает, что если есть сомнение в справедливости предположения о равенстве всех дисперсий $\sigma_1^2,...,\sigma_k^2$, но есть уверенность, что дисперсии σ_i^2 и σ_j^2 равны, то для построения доверительного интервала параметра $(\mu_i-\mu_j)$ следует использовать формулу (3.6.50). Если же нет уверенности в равенстве дисперсий σ_i^2 и σ_j^2, то для построения доверительного интервала параметра $(\mu_i-\mu_j)$ следует использовать центральную статистику вида (3.6.56) и аппроксимацию Уэлша.

Отметим также, что *F*-критерий (3.8.30) ориентирован на гауссовскую модель наблюдений (3.8.3), одно из предположений которой связано с равенством дисперсий $\sigma_1^2,...,\sigma_k^2$ (см. формулу (3.8.3г)). Это предположение может быть предварительно проверено (см. разд. 3.6), например, с помощью критерия Бартлетта. Однако следует иметь в виду, что критерий Бартлетта не является робастным к отклонениям от нормальности модели наблюдений. Вопросы робастности *F*-критерия к нарушению предположений модели (3.8.3) обсуждаются в разд. 5.6 и работах [106, 112]. См. также работу [*Keselman H.J.*, *Wilcox Rand R*. Tests for mean equality that do not require homogeneity of variances: do they really work? // Commun. Statist. – Simula. 2000. V. 29. No. 3. P. 875–895].

3.8.В. Метод множественных сравнений Шеффе

Дисперсионный анализ, проведенный с помощью *F*-критерия (3.8.30), может показать нам, что совокупности, из которых отбираются наблюдения, различаются по своим средним значениям. Если *F*-критерий отклоняет гипотезу $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, то это означает, что существует, по крайней мере, одна пара совокупностей W_i и W_j $(i \neq j)$, для которой $\mu_i \neq \mu_j$. Но, зная лишь значение *F*-статистики, мы не можем указать номера совокупностей i и j, для которых $\mu_i \neq \mu_j$. Для ответа на этот вопрос нужно проводить дополнительный анализ, чтобы узнать, средние значения каких совокупностей отличаются значимо. Конечно, можно было бы попытаться произвести серию парных сравнений совокупностей при помощи *t*-критерия Стьюдента (см. разд. 3.6.Г). Однако такой подход ошибочен (см. [47, с. 182]), так как при заданном уровне значимости α для *t*-критерия Стьюдента, *при проведении серий парных сравнений с помощью t-критерия вероятность отклонить верную гипотезу становится значительно больше, чем* α, то есть не обеспечивается заданный уровень значимости. Для решения возникающей проблемы разработаны различные процедуры. Наиболее общие решения поставленной задачи были предложены Шеффе и Тьюки (см., например, [9]).

Рассмотрим линейную комбинацию $\theta = (c_1\mu_1 + \cdots + c_k\mu_k)$ средних значений $\mu_1,...,\mu_k$, где $c_1,...,c_k$ – постоянные величины. Если сумма этих постоянных равняется нулю, то есть $c_1 + c_2 + \cdots + c_k = 0$, то ли-

нейную комбинацию средних значений называют *контрастом*. Обсудим вопрос построения критерия проверки гипотез $H_0^*: \theta = 0$, $H_1^*: \theta \neq 0$. Отметим, что несмещенной оценкой контраста θ является $\hat{\theta} = (c_1 \overline{X}_{\bullet 1} + \cdots + c_k \overline{X}_{\bullet k})$. Далее, так как $\overline{X}_{\bullet 1}, ..., \overline{X}_{\bullet k}$ независимы и $D(X_{\bullet j}) = \sigma^2 / n_j$, $j = 1, ..., k$, то $D(\hat{\theta}) = \sigma^2 (c_1^2 / n_1 + \cdots + c_k^2 / n_k)$. Напомним, что несмещенной оценкой параметра σ^2 является средний квадрат $S_W^2 = SS_W / (n-k)$, поэтому $\hat{D}(\hat{\theta}) = S_W^2 (c_1^2 / n_1 + \cdots + c_k^2 / n_k)$ является оценкой дисперсии $D(\hat{\theta})$.

Таким образом, для модели наблюдений (3.8.3) случайная величина $(\hat{\theta} - \theta) / \sqrt{\hat{D}(\hat{\theta})}$ при гипотезе $H_0^*: \theta = 0$ имеет распределение Стьюдента с числом степеней свободы $(n-k)$ и, следовательно, критерий проверки гипотез

$$H_0^*: \sum_{j=1}^{k} c_j \mu_j = 0,\ c_1 + c_2 + \cdots + c_k = 0,\ \ H_1^*: \sum_{j=1}^{k} c_j \mu_j \neq 0 \qquad \textbf{(3.8.41)}$$

может быть основан на доверительном интервале для параметра θ, который записывается в виде

$$\sum_{j=1}^{k} c_j \overline{X}_{\bullet j} \pm t_{1-\alpha/2,\,(n-k)} \{ S_W^2 \sum_{j=1}^{k} (c_j^2 / n_j) \}^{1/2}. \qquad \textbf{(3.8.42)}$$

Если этот интервал накрывает ноль, то гипотеза $H_0^*: \theta = 0$ принимается, в противном случае гипотеза H_0^* отвергается с уровнем значимости α. Отметим, что последовательное применение данного критерия для определения значимо различающихся пар средних значений совокупностей в тех случаях, когда *F*-критерий отклоняет гипотезу $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, не обеспечивает заданного уровня значимости α. В таких случаях следует использовать процедуры множественного сравнения.

Метод множественного сравнения Шеффе (*S*-метод, см. [80]) дает возможность провести серию парных сравнений совокупностей так, что вероятность хотя бы одного неверного заключения в точ-

ности будет равна α, то есть этот метод позволяет сохранить α в качестве общего уровня значимости при проверке серии гипотез

$$H_0^*: \mu_i - \mu_j = 0,\ \ H_1^*: \mu_i - \mu_j \neq 0,\ \ 1 \le i \neq j \le k.$$

Шеффе [80] показал, что для любой линейной комбинации $c_1\mu_1 + \cdots + c_k\mu_k$ (и, следовательно, для контраста $\theta = (c_1\mu_1 + \cdots + c_k\mu_k)$, $c_1 + c_2 + \cdots + c_k = 0$ тоже) выполняется выражение

$$P\{\ \hat{\theta} - S\sqrt{\hat{D}(\hat{\theta})} < \theta < \hat{\theta} + S\sqrt{\hat{D}(\hat{\theta})}\ \} = 1 - \alpha, \qquad \textbf{(3.8.43)}$$

где

$$S^2 = (k-1)\cdot F_{1-\alpha}(k-1\ ,\ n-k). \qquad \textbf{(3.8.44)}$$

Использование выражения (3.8.43) позволяет построить критерий проверки гипотез (3.8.41) с помощью доверительного интервала

$$\sum_{j=1}^{k} c_j \bar{X}_{\bullet j} \pm \{\ (k-1)\cdot S_W^2 \cdot F_{1-\alpha}(k-1\,,\,n-k) \sum_{j=1}^{k} (c_j^2 / n_j)\ \}^{1/2}. \qquad \textbf{(3.8.45)}$$

Если этот интервал не содержит ноль, то гипотеза $H_0^*: \theta = 0$ отвергается с уровнем значимости α. Отметим, что этот критерий может быть последовательно применен для каждого представляющего интерес контраста, причем общий для всех критериев уровень значимости остается равным α. Например, при проверке гипотезы $H_0^*: \mu_i - \mu_j = 0$ $(i \neq j)$ следует положить $c_i = 1$, $c_j = -1$, а все остальные из $c_1, \ldots, c_k$ должны быть равны нулю. В этом случае доверительный интервал (3.8.45) запишется в виде

$$(\bar{X}_{\bullet i} - \bar{X}_{\bullet j}) \pm \{F_{1-\alpha}(k-1\,,\,n-k)\cdot(k-1)\cdot S_W^2 \cdot (1/n_j + 1/n_j)\}^{1/2}. \qquad \textbf{(3.8.46)}$$

Метод множественного сравнения Тьюки (*T*-метод, см. [3]) применяется *только для контрастов* при условии, что объемы выборок *одинаковы*, то есть $n_1 = n_2 = \cdots = n_k = m$. Для проверки гипотез (3.8.41) строится доверительный интервал

$$\sum_{j=1}^{k} c_j \bar{X}_{\bullet j} \pm \frac{1}{2}\sqrt{S_W^2 / m}\cdot q_{1-\alpha} \cdot \sum_{j=1}^{k} |\, c_j\,|, \qquad \textbf{(3.8.47)}$$

где $q_{1-\alpha}$ – квантиль уровня $(1-\alpha)$ распределения стьюдентизован-

ного размаха с k и $(n-k)$ степенями свободы (см. приложение II, табл. 7 в [3]). Если интервал (3.8.47) не содержит ноль, то гипотеза $H_0^*: \theta = 0$ отвергается с уровнем значимости α. Этот критерий также может быть последовательно применен для любого контраста, причем общий для всех критериев уровень значимости остается равным α.

Замечание 3.8.48. В тех случаях, когда *F*-критерий отклоняет гипотезу $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, это означает, что существует, по крайней мере, одна пара совокупностей W_i и W_j $(i \neq j)$, для которой значимо выполняется неравенство $\mu_i \neq \mu_j$. Для выявления таких значимо различающихся пар совокупностей, общее число которых может равняться $C_k^2 = k!/(2! \cdot (k-2)!)$, обычно применяется на практике *S*-метод Шеффе в виде серии парных сравнений, то есть в виде последовательной проверке гипотез $H_0^*: \mu_i - \mu_j = 0$, $H_1^*: \mu_i - \mu_j \neq 0$, $1 \leq i \neq j \leq k$. Для того чтобы исключить противоречия и уменьшить число перебираемых пар совокупностей, на практике обычно поступают следующим образом. Сначала располагаются все рассматриваемые совокупности $W_1, ..., W_k$ в порядке возрастания их выборочных групповых средних. Затем проводится серия парных сравнений совокупностей в определенном порядке. Сначала – парное сравнение совокупности, обладающей наименьшим выборочным средним, с совокупностью, обладающей наибольшим выборочным средним. Затем, совокупность с наименьшим выборочным средним сравнивается со всеми последующими совокупностями, в порядке убывания их выборочных средних, и этот процесс продолжается до принятия основной гипотезы. Далее повторяется описанная процедура для совокупности со вторым по величине выборочным средним, то есть эта совокупность сравнивается со всеми последующими совокупностями в порядке убывания их выборочных средних. Эта процедура опять продолжается до принятия гипотезы $H_0^*: \mu_i - \mu_j = 0$, $i \neq j$. Затем повторяется этот процесс для совокупности, имеющей третье по величине выборочное среднее, и т.д. Применение описанной методики, которую будем называть *упорядоченной процедурой парных сравнений Шеффе*, приводиться в упр. 3.9.94 и 3.9.96.

Отметим, что процедура Шеффе для проверки гипотез $H_0^*: \theta = 0$, $H_1^*: \theta \neq 0$ может быть реализована и в стандартной форме в виде критерия, основанного на статистике

$$Z = \frac{\left(\sum_{j=1}^{k} c_j \bar{X}_{\bullet j}\right)^2}{(k-1)\cdot\left(SS_w/(n-k)\right)\sum_{j=1}^{k}(c_j^2/n_j)}. \qquad \textbf{(3.8.49)}$$

Для этой статистики выполняется выражение $L(Z \mid H_0^*) = F(k-1,\ n-k)$, с учетом которого критерий проверки гипотез H_0^*, H_1^* формулируется в виде следующего правила.

«Если наблюденное значение z статистики Z, вычисленное по формуле (*3.8.42*), *превышает ее критическое значение $F_{1-\alpha}(k-1,\ n-k)$, то гипотеза H_0^* отвергается на уровне значимости α, если $z < F_{1-\alpha}(k-1,\ n-k)$, то нет оснований отвергать гипотезу H_0^*».*

Отметим, что *S*-метод Шеффе и *T*-метод Тьюки обладают различными свойствами. Во-первых, *S*-метод Шеффе можно использовать для *любых* линейных комбинаций средних значений совокупностей (а не только для контрастов) и для случая, когда объемы выборок $n_1,...,n_k$ *различны*, а *T*-метод Тьюки можно применять *только для контрастов* и при условии, что объемы выборок $n_1,...,n_k$ *одинаковы*, то есть $n_1 = n_2 = \cdots = n_k = m$. Во-вторых, $(1-\alpha)$-доверительные интервалы, полученные методом Тьюки при парных сравнениях совокупностей, то есть при последовательной проверке гипотез $H_0^*: \mu_i - \mu_j = 0$, $i \neq j$, как правило, являются более короткими. Однако *S*-метод Шеффе дает более короткие доверительные интервалы для более сложных контрастов. Примеры, иллюстрирующие эти свойства процедур множественных сравнений, приводятся в [9, с. 293], см. также [80; 5, т. 2, с. 56; 44, с. 157].

3.8.Г. Однофакторный дисперсионный анализ: модель II

Пусть признак X характеризует качество объектов, которые хранятся на складе предприятия в большом количестве контейнеров. С целью определения качества всех объектов, хранящихся на складе, случайно выбирается заданное число k контейнеров и в каждом *j*-м контейнере, $j = 1,...,k$, берется выборка $X_{1j},...,X_{n_j j}$ объема n_j, $j = 1,...,k$. Математической моделью наблюдений над признаком X в условиях данного эксперимента является так называемая модель II дисперсионного анализа со случайными эффектами (*или модель компонентного анализа*). Эта модель формулируется следующим образом. Предполагается, что k групп (подпопуляций) случайно выбираются из

бесконечной совокупности всех возможных подпопуляций. Каждой группе присваивается номер от 1 до k, и j-я группа считается соответствующей j-му уровню фактора Y, который характеризует свойства рассматриваемых подпопуляций. Далее, из каждой группы случайно выбирается n_j объектов и рассматриваются значения $x_{1j},...,x_{n_j j}$, $j=1,...,k$, итогового показателя X. Предполагается, что наблюдаемые данные $X_{1j},...,X_{n_j j}$ распределены нормально со средним значением a_j и дисперсией σ^2, не зависящей от уровня j фактора Y, $j=1,...,k$, то есть для i-го элемента выборки в группе с номером j выполняется выражение $L(X_{ij} \mid j) = N(a_j, \sigma^2)$. Также предполагается, что $a_1,...,a_k$ представляют случайную выборку из совокупности объектов, признак X которых имеет нормальное распределение со средним значением μ и дисперсией σ_a^2, то есть справедливо выражение $L(a_j) = N(\mu, \sigma_a^2)$, $j=1,...,k$. Для удобства параметризации этой модели, так же как и для модели I вида (3.8.3), определяют j-й дифференциальный (главный) эффект фактора Y равенством $\alpha_j = a_j - \mu$. Отметим, что в отличие от *постоянных* дифференциальных эффектов $\xi_j = \mu_j - \mu$ фактора A в модели I с фиксированными эффектами (3.8.7), эффекты $\alpha_j = a_j - \mu$, $j=1,...,k$, рассматриваемой модели II являются *случайными величинами*, распределенными нормально с нулевым средним и дисперсией σ_a^2, то есть справедливы выражения $L(\alpha_j) = N(0, \sigma_a^2)$, $j=1,...,k$.

Итак, модель однофакторного дисперсионного анализа со случайными эффектами (коротко модель II) описывается наблюдениями вида

$$X_{ij} = a_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij},\ L(\alpha_j) = N(0, \sigma_a^2),\ L(\varepsilon_{ij}) = N(0, \sigma^2),$$
$$i = 1,...,n_j,\ j = 1,...,k, \qquad \textbf{(3.8.50)}$$

причем все *случайные величины* $\{\alpha_j\}$ и $\{\varepsilon_{ij}\}$ *в совокупности независимы*, поэтому выполняется выражение $L(X_{ij}) = N(\mu, \sigma^2 + \sigma_a^2)$, $i=1,...,n_j$, $j=1,...,k$. Отметим, что модель (3.8.50) характеризуется тремя параметрами. Параметр μ этой модели называют генеральным средним изучаемой совокупности, а параметры σ^2 и σ_a^2 – компонентами дисперсии, причем дисперсия σ^2 вызвана ошибками $\{\varepsilon_{ij}\}$ измерений, а параметр σ_a^2 связан со случайным выбором k подпопуляций. В случае модели I вида (3.8.7) нас интересовала оценка дифференциального эффекта ξ_j для j-го уровня фактора A и проверка гипотезы (3.8.9) о том, что все ξ_j равны нулю. Для модели II нас интересует не оценка отдельных эффектов α_j, а оценка дисперсии σ_a^2 распределения дифференци-

альных эффектов. Другими словами, целями дисперсионного анализа в условиях модели II, являются построение оценок параметров этой модели, в частности генерального среднего μ всей совокупности и двух компонент дисперсии σ^2 и σ_a^2, а также проверка гипотезы $H_0 : \sigma_a^2 = 0$, которая означает, что изменение фактора Y не вносит вклада в дисперсию. При статистической обработке наблюдаемых данных $\{X_{ij}\}$, $i = 1,...,n_j$, $j = 1,...,k$, вида (3.8.50) используются те же общие понятия, которые были введены при использовании модели I. В частности, остается справедливым правило сложения (3.8.23), согласно которому SS_T – общая сумма квадратов отклонений от $\bar{X}_{\bullet\bullet}$ распадается на две суммы, то есть

$$SS_T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2 + \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \\ = SS_W + SS_B . \qquad \textbf{(3.8.51)}$$

Как и для модели I, суммы квадратов SS_W и SS_B используются при построении F-статистики критерия проверки гипотезы $H_0 : \sigma_a^2 = 0$. Однако следует отметить, что средний квадрат $S_W^2 = SS_W/(n-k)$ внутри групп, как и прежде, остается несмещенной оценкой параметра σ^2, то есть $M(S_W^2) = \sigma^2$, но изменение модели сказывается на распределении среднего квадрата $S_B^2 = SS_B/(k-1)$ между группами и, в частности, на его математическом ожидании. Можно убедиться (см. упр. 3.9.101), что в рамках модели II, для которой наблюдения описываются выражением (3.8.50), математическое ожидание среднего квадрата между группами вычисляется по формуле

$$M(S_B^2) = \frac{1}{k-1}\sum_{j=1}^{k} n_j M(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \frac{1}{k-1}\left(n - \frac{1}{n}\sum_{j=1}^{k} n_j^2\right)\sigma_a^2 + \sigma^2 . \qquad \textbf{(3.8.52)}$$

Отметим, что в частном случае, когда объемы выборок в выбранных группах равны, то есть при $n_1 = n_2 = ... = n_k = m$ и $n = km$, эта формула принимает более простой вид:

$$M(S_B^2) = \frac{1}{k-1}\left(km - \frac{km^2}{km}\right)\sigma_a^2 + \sigma^2 = m\sigma_a^2 + \sigma^2 . \qquad \textbf{(3.8.53)}$$

Напомним (см. формулу (3.8.28) и упр. 3.9.80), что для модели I с *фиксированными дифференциальными эффектами* $\xi_j = \mu_j - \mu$, $j = 1,...,k$, математическое ожидание среднего квадрата $S_B^2 = SS_B/(k-1)$ между группами вычисляется по формуле

$$M(S_B^2) = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j \xi_j^2,$$

которая при $n_1 = n_2 = ... = n_k = m$ записывается в виде

$$M(S_B^2) = \sigma^2 + m\sum_{j=1}^{k} (\mu_j - \mu)^2 /(k-1). \qquad \textbf{(3.8.54)}$$

Сравнивая формулы (3.8.53) и (3.8.54), замечаем, что при $n_1 = n_2 = ... = n_k = m$ для модели I вместо σ_a^2 стоит сумма квадратов дифференциальных эффектов, деленная на $(k-1)$, то есть вместо σ_a^2 стоит $\{(\mu_1 - \mu)^2 + ... + (\mu_k - \mu)^2\}/(k-1)$. Ниже рассматривается сначала случай равных объемов выборок, то есть считается, что $n_1 = n_2 = ... = n_k = m$. Обсуждение общего случая неравных объемов выборок приводится в замечании 3.8.77 и детали могут быть найдены в [9, 69, 80].

Итак, средний квадрат внутри групп $S_W^2 = SS_W/(n-k)$ для обеих моделей I и II является несмещенной оценкой параметра σ^2, который характеризует рассеяние только внутри групп (внутри подпопуляций), вызванное ошибками измерений, причем статистика S_W^2 распределена как $\sigma^2\chi^2(km-k)/(km-k)$, то есть справедливо выражение

$$L\left(\frac{(km-k)\cdot S_W^2}{\sigma^2}\right) = \chi^2(km-k). \qquad \textbf{(3.8.55)}$$

Далее, наличие рассеяния между средними значениями рассматриваемых групп, то есть введение случайных параметров $\alpha_j = a_j - \mu$, $j = 1,...,k$, в модель II вида (3.8.50) приводит к изменению распределения среднего квадрата $S_B^2 = SS_B/(k-1)$ между группами. Обсудим этот вопрос подробнее. Используя формулу (3.8.50), получаем

$$\overline{X}_{\bullet j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n_j}\sum_{i=1}^{n_j}(\mu + \alpha_j + \varepsilon_{ij}) = \mu + \alpha_j + \overline{\varepsilon}_{\bullet j},\ j = 1,...,k, \qquad \textbf{(3.8.56)}$$

где $\overline{\varepsilon}_{\bullet j} = (\varepsilon_{1j} + ... + \varepsilon_{n_j j})/n_j$, $j = 1,...,k$. Учитывая теперь, что $\overline{X}_{\bullet j} = \mu + \alpha_j + \overline{\varepsilon}_{\bullet j}$, $j = 1,...,k$, и вспоминая, что $M(\alpha_j) = 0$, $M(\overline{\varepsilon}_{\bullet j}) = 0$ при $n_1 = n_2 = ... = n_k = m$ получаем

$$M(\overline{X}_{\bullet j}) = \mu,\ D(\overline{X}_{\bullet j}) = D(\alpha_j) + D(\overline{\varepsilon}_{\bullet j}) = \sigma_a^2 + \sigma^2/m. \qquad \textbf{(3.8.57)}$$

Учитывая далее, что случайные величины α_j и ε_{ij} нормально распределены и независимы, заключаем, что $\bar{X}_{\bullet 1},...,\bar{X}_{\bullet k}$ тоже нормально распределены, откуда следует, что $\sum(\bar{X}_{\bullet j}-\bar{X}_{\bullet\bullet})^2/(k-1)$, являющаяся оценкой дисперсии $\sigma_a^2+\sigma^2/m$, будет распределена как $(\sigma_a^2+\sigma^2/m)\chi^2(k-1)/(k-1)$. Следовательно, статистика S_B^2 распределена как $(\sigma^2+m\sigma_a^2)\chi^2(k-1)/(k-1)$, то есть справедливо выражение

$$L\left(\frac{(k-1)\cdot S_B^2}{\sigma^2+m\sigma_a^2}\right)=\chi^2(k-1). \qquad \textbf{(3.8.58)}$$

Из приведенных выражений (3.8.55) и (3.8.58) следует, что при выполнении равенства $\sigma_a^2=0$ статистика $F=S_B^2/S_W^2$, используемая для проверки гипотез (3.8.4) и (3.8.9) в рамках модели I, также имеет *F*-распределение с $(k-1)$ и $(mk-k)$ степенями свободы, то есть в рамках модели II при справедливости гипотезы $H_0:\sigma_a^2=0$ выполняется выражение

$$L(F=S_B^2/S_W^2\mid H_0)=F(k-1,\,n-k). \qquad \textbf{(3.8.59)}$$

Отсюда следует, *что F-критерий* (*3.8.30*), *основанный на статистике* $F=S_B^2/S_W^2$, *также может быть использован и для проверки гипотезы* $H_0:\sigma_a^2=0$ *в рамках модели II со случайными эффектами*. **(3.8.59а)**

Выпишем теперь точечные и интервальные оценки параметров μ, σ^2 и σ_a^2 модели II дисперсионного анализа со случайными эффектами. Из выражения (3.8.55) следует, что $M(S_W^2)\cdot(mk-k)/\sigma^2=(mk-k)$. Отсюда получаем, что средний квадрат внутри групп S_W^2 является несмещенной оценкой параметра σ^2, то есть $M(S_W^2)=\sigma^2$. Таким образом, параметр σ^2, характеризующий дисперсию случайных величин ε_{ij} в модели (3.8.50), оценивается с помощью остаточного среднего квадрата S_W^2, то есть

$$\hat{\sigma}^2=S_W^2=SS_W/(n-k)=\frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})^2. \qquad \textbf{(3.8.60)}$$

Для построения интервальных оценок параметра σ^2 используется центральная статистика для этого параметра в виде $S_W^2(km-k)/\sigma^2$, которая, согласно

(3.8.55), имеет хи-квадрат-распределение с числом степеней свободы $(km-k)$ и монотонно зависит от параметра σ^2. Следовательно, справедливо выражение

$$P\{\chi^2_{\alpha/2,(km-k)} < S_W^2(km-k)/\sigma^2 < \chi^2_{1-\alpha/2,(km-k)}\} = 1-\alpha,$$

из которого получаем $(1-\alpha)$-доверительный интервал для параметра σ^2 в виде

$$P\left\{\frac{(km-k)\cdot S_W^2}{\chi^2_{1-\alpha/2,(km-k)}} < \sigma^2 < \frac{(km-k)\cdot S_W^2}{\chi^2_{\alpha/2,(km-k)}}\right\} = 1-\alpha. \qquad \textbf{(3.8.61)}$$

Используя (3.8.53) и учитывая, что $M(S_W^2)=\sigma^2$, получаем $M(S_B^2)-M(S_W^2)=m\sigma_a^2$. Отсюда следует, что несмещенная оценка параметра σ_a^2 записывается в виде

$$\hat{\sigma}_a^2 = (S_B^2 - S_W^2)/m. \qquad \textbf{(3.8.62)}$$

Отметим, что к настоящему времени получены лишь приближенные доверительные интервалы для параметра σ_a^2. Для этой цели были использованы различные аппроксимации. Наилучшая аппроксимация, с использованием *F*-распределения, была предложена Броссом (см. [9, с. 297] , а также монографию [80]). Мы приведем сначала приближенные доверительные интервалы для параметра σ_a^2, основанные на гауссовской аппроксимации, которая приемлема при $k>30$, а затем приведем точные доверительные пределы для отношения σ_a^2/σ^2.

Вычислим дисперсию оценки $\hat{\sigma}_a^2 = (S_B^2 - S_W^2)/m$. Используя выражение (3.8.58), согласно (П.2.17) заключаем, что дисперсия случайной величины $(k-1)\cdot S_B^2/(\sigma^2+m\sigma_a^2)$ равна удвоенному числу степеней свободы распределения хи-квадрат, то есть равна $2(k-1)$, и, следовательно, дисперсия статистики S_B^2 вычисляется по формуле

$$D(S_B^2) = \frac{2(\sigma^2+m\sigma_a^2)^2}{k-1}. \qquad \textbf{(3.8.63)}$$

Далее, из (3.8.55) следует, что дисперсия статистики S_W^2 вычисляется следующим образом:

$$D(S_W^2) = \frac{2(\sigma^2)^2}{k(m-1)}. \qquad \textbf{(3.8.64)}$$

Используя формулы (3.8.63) и (3.8.64) с учетом независимости статистик S_B^2 и S_W^2, получаем

$$D(\hat{\sigma}_a^2) = D[(S_B^2 - S_W^2)/m] = \frac{2(\sigma^2 + m\sigma_a^2)^2}{m^2(k-1)} + \frac{2(\sigma^2)^2}{m^2 k(m-1)}.$$

Заменяя $\sigma^2 + m\sigma_a^2$ и σ^2 их оценками S_B^2 и S_W^2 соответственно, получим оценку дисперсии статистики $\hat{\sigma}_a^2 = (S_B^2 - S_W^2)/m$ в виде

$$\hat{D}(\hat{\sigma}_a^2) = \frac{2(S_B^2)^2}{m^2(k-1)} + \frac{2(S_W^2)^2}{m^2 k(m-1)} = 2[(S_B^2)^2/(k-1) + (S_W^2)^2/k(m-1)]/m^2. \qquad \textbf{(3.8.65)}$$

При выполнении неравенства $k > 30$ можно применить нормальную аппроксимацию (см. [69, с. 375]), согласно которой справедливо выражение

$$L\{(\hat{\sigma}_a^2 - \sigma_a^2)/\sqrt{\hat{D}(\hat{\sigma}_a^2)}\} = N(0,1). \qquad \textbf{(3.8.66)}$$

С учетом этого выражения, приближенный доверительный интервал для параметра σ_a^2 запишется в виде $\{\hat{\sigma}_a^2 \pm \lambda_{1-\alpha/2} \cdot \sqrt{\hat{D}(\hat{\sigma}_a^2)}\}$ или, окончательно, в такой форме:

$$\{(S_B^2 - S_W^2)/m \pm \lambda_{1-\alpha/2} \cdot \sqrt{2[(S_B^2)^2(k-1) + (S_W^2)^2/(km-k)]/m^2}\}, \qquad \textbf{(3.8.67)}$$

где $\lambda_{1-\alpha/2}$ – квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения, то есть $\lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$.

Приведем теперь точные доверительные пределы для отношения σ_a^2/σ^2. Действуя таким же образом, как при получении выражения (3.6.16), и учитывая, что несмещенная оценка S_W^2 параметра σ^2 распределена как $\sigma^2\chi^2(km-k)/(km-k)$, то есть справедливо выражение (3.8.55), а оценка S_B^2 для $(\sigma^2 + m\sigma_a^2)$ распределена как $(\sigma^2 + m\sigma_a^2)\chi^2(k-1)/(k-1)$, то есть справедливо выражение (3.8.58), получаем

$$P\left\{\frac{S_B^2}{S_W^2} \cdot \frac{1}{F_{1-\alpha/2}(k-1,\, km-k)} < \frac{\sigma^2 + m\sigma_a^2}{\sigma^2} < \frac{S_B^2}{S_W^2} \cdot F_{1-\alpha/2}(km-k, k-1)\right\} = 1-\alpha.$$

Отсюда следует, что $(1-\alpha)$-доверительный интервал для отношения σ_a^2/σ^2 записывается в виде

$$P\left\{\frac{1}{m}\left(\frac{S_B^2}{S_W^2} \cdot \frac{1}{F_{1-\alpha/2}(k-1,\, km-k)} - 1\right) < \frac{\sigma_a^2}{\sigma^2} < \frac{1}{m}\left(\frac{S_B^2}{S_W^2} \cdot F_{1-\alpha/2}(km-k, k-1) - 1\right)\right\} =$$

$$= 1-\alpha. \qquad \textbf{(3.8.68)}$$

Отметим, что отношение $(\sigma_a^2/\sigma^2)>0$, поэтому, если отношение S_B^2/S_W^2 окажется меньше чем $F_{1-\alpha/2}(k-1,km-k)$, то нижняя доверительная граница в (3.8.68) будет отрицательной, и поэтому ее следует положить равной нулю.

Приведем теперь точечную и интервальную оценку для параметра μ модели (3.8.50). Используя модель наблюдений вида (3.8.50), получим

$$\overline{X}_{\bullet\bullet}=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mu+\alpha_j+\varepsilon_{ij})=\mu+\overline{\alpha}_{\bullet}+\overline{\varepsilon}_{\bullet\bullet}, \qquad \textbf{(3.8.69)}$$

где $$\overline{\varepsilon}_{\bullet\bullet}=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\varepsilon_{ij}=\frac{1}{n}\sum_{j=1}^{k}n_j\overline{\varepsilon}_{\bullet j},\ \overline{\varepsilon}_{\bullet j}=\frac{1}{n_j}\sum_{i=1}^{n_j}\varepsilon_{ij},\ \overline{\alpha}_{\bullet}=\frac{1}{n}\sum_{j=1}^{k}n_j\alpha_j. \qquad \textbf{(3.8.70)}$$

Далее, согласно предположениям модели (3.8.50), выполняются равенства $M(\alpha_j)=M(\varepsilon_{ij})=0$ и, следовательно, также равенства $M(\overline{\alpha}_{\bullet})=M(\overline{\varepsilon}_{\bullet\bullet})=0$. Учитывая их, получаем

$$M(\overline{X}_{\bullet\bullet})=M(\mu+\overline{\alpha}_{\bullet}+\overline{\varepsilon}_{\bullet\bullet})=\mu.$$

Таким образом, несмещенной оценкой параметра μ модели II вида (3.8.50) является общее групповое выборочное среднее $\overline{X}_{\bullet\bullet}$, то есть

$$\hat{\mu}=\overline{X}_{\bullet\bullet}=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}\ . \qquad \textbf{(3.8.71)}$$

Вычислим теперь дисперсию $D(\overline{X}_{\bullet\bullet})=M(\overline{X}_{\bullet\bullet}^2)-\mu^2$. Используя формулу (3.8.69), согласно которой $\overline{X}_{\bullet\bullet}=\mu+\overline{\alpha}_{\bullet}+\overline{\varepsilon}_{\bullet\bullet}$, получим

$$\overline{X}_{\bullet\bullet}^2=(\mu+\overline{\alpha}_{\bullet}+\overline{\varepsilon}_{\bullet\bullet})^2=\mu^2+\overline{\alpha}_{\bullet}^2+\overline{\varepsilon}_{\bullet\bullet}^2+2\mu\overline{\alpha}_{\bullet}+2\mu\overline{\varepsilon}_{\bullet\bullet}+2\overline{\alpha}_{\bullet}\overline{\varepsilon}_{\bullet\bullet}.$$

Вспоминая далее, что для модели (3.8.50) все случайные величины α_j и ε_{ij} в совокупности независимы и $M(\alpha_j)=M(\varepsilon_{ij})=0$, получаем

$$M(\overline{X}_{\bullet\bullet}^2)=M(\mu^2+\overline{\alpha}_{\bullet}^2+\overline{\varepsilon}_{\bullet\bullet}^2+2\mu\overline{\alpha}_{\bullet}+2\mu\overline{\varepsilon}_{\bullet\bullet}+2\overline{\alpha}_{\bullet}\overline{\varepsilon}_{\bullet\bullet})=\mu^2+M(\overline{\alpha}_{\bullet}^2)+M(\overline{\varepsilon}_{\bullet\bullet}^2).$$

Далее, используя (3.8.70), имеем

$$M(\overline{\alpha}_{\bullet}^2)=D(\overline{\alpha}_{\bullet})=\frac{1}{n^2}\sum_{j=1}^{k}n_j^2\sigma_a^2=\frac{\sigma_a^2}{n^2}\sum_{j=1}^{k}n_j^2,$$

$$M(\overline{\varepsilon}_{\bullet\bullet}^2)=D(\overline{\varepsilon}_{\bullet\bullet})=\frac{1}{n^2}\sum_{j=1}^{k}n_j^2(\sigma^2/n_j)=\sigma^2/n.$$

Отсюда следует, что дисперсия статистики $\hat{\mu} = \bar{X}_{\bullet\bullet}$ запишется в виде

$$D(\bar{X}_{\bullet\bullet}) = M(\bar{X}_{\bullet\bullet}^2) - \mu^2 = M(\bar{\alpha}_{\bullet}^2) + M(\bar{\varepsilon}_{\bullet\bullet}^2) = \frac{\sigma_a^2}{n^2}\sum_{j=1}^{k} n_j^2 + \frac{\sigma^2}{n}. \qquad \textbf{(3.8.72)}$$

Отметим, что при равных объемах выборок, то есть при $n_1 = n_2 = ... = n_k = m$ и $n = mk$, формула (3.8.72) принимает следующий вид:

$$D(\bar{X}_{\bullet\bullet}) = (m\sigma_a^2 + \sigma^2)/mk. \qquad \textbf{(3.8.73)}$$

Итак, для модели (3.8.50) при равных объемах выборок, то есть при $n_1 = n_2 = ... = n_k = m$ и $n = mk$, случайная величина $(\bar{X}_{\bullet\bullet} - \mu)/\sqrt{D(\bar{X}_{\bullet\bullet})} = (\bar{X}_{\bullet\bullet} - \mu)/\sqrt{(m\sigma_a^2 + \sigma^2)/km}$ имеет стандартное нормальное распределение, то есть справедливо выражение

$$L\left(\frac{\bar{X}_{\bullet\bullet} - \mu}{\sqrt{(m\sigma_a^2 + \sigma^2)/km}}\right) = N(0,1). \qquad \textbf{(3.8.74)}$$

Заменяя в данном выражении дисперсию $m\sigma_a^2 + \sigma^2$ ее несмещенной оценкой S_B^2, имеющей $(k-1)$ степеней свободы, получаем, что случайная величина $(\bar{X}_{\bullet\bullet} - \mu)/\sqrt{S_B^2/km}$ имеет распределение Стьюдента с числом степеней свободы $k-1$, то есть справедливо выражение

$$L\left(\frac{\bar{X}_{\bullet\bullet} - \mu}{\sqrt{S_B^2/km}}\right) = St(k-1). \qquad \textbf{(3.8.75)}$$

Отсюда следует, что $(1-\alpha)$-доверительный интервал для параметра μ записывается в виде

$$(\ \bar{X}_{\bullet\bullet} \pm t_{1-\alpha/2,(k-1)} \cdot \sqrt{S_B^2/km}\ ). \qquad \textbf{(3.8.76)}$$

Замечание 3.8.77. Если объемы выборок $n_1,...,n_k$ различны, то дисперсионный анализ проводится по той же схеме, что и для случая равных объемов выборок, однако при этом необходимо учесть следующие изменения. При вычислении математического ожидания среднего квадрата между группами, вместо формулы (3.8.53) следует использовать формулу (3.8.52), согласно которой $M(S_B^2) = \tilde{m}\sigma_a^2 + \sigma^2$, где $\tilde{m} = [n - (n_1^2 + ... + n_k^2)/n]/(k-1)$ и $n = n_1 + ... + n_k$. Далее, если объемы выборок $n_1,...,n_k$ различны, то для модели II вида (3.8.50)

общее групповое выборочное среднее $\overline{X}_{\bullet\bullet}$ естественно остается несмещенной оценкой параметра μ и имеет нормальное распределение со средним μ и дисперсией $D(\overline{X}_{\bullet\bullet})$, вычисляемой по формуле (3.8.72) вида

$$D(\overline{X}_{\bullet\bullet}) = (n_1^2 + ... + n_k^2) \cdot \sigma_a^2 / n^2 \; + \; \sigma^2 / n \,, \; n = n_1 + ... + n_k \,,$$

при этом выполняется выражение

$$L\left(\frac{(\overline{X}_{\bullet\bullet} - \mu)}{\sqrt{\sigma_a^2 (n_1^2 + ... + n_k^2)/n^2 + \sigma^2 / n}} \right) = N(0,1) \,. \qquad \textbf{(3.8.78)}$$

Отметим также, что для неравных объемов выборок выражение (3.8.58) выполняется лишь приближенно. Это означает, что при $\sigma_a^2 = 0$ оно дает лишь приближенно распределение хи-квадрат с $n - k$ степенями свободы, поэтому F-критерий проверки гипотезы $H_0 : \sigma_a^2 = 0$ в рамках модели II при условии, что объемы выборок $n_1, ..., n_k$ различны, является лишь приближенным, однако на практике применяют обычно его, так как точный критерий слишком сложный. Если же есть возможность обеспечить объемы выборок равными, то при этих условиях F-критерий проверки гипотезы $H_0 : \sigma_a^2 = 0$ в рамках модели II является точным, то есть при справедливости предположений модели (3.8.50) он гарантирует заданный уровень значимости. Примеры вычисления объемов выборок, обеспечивающих заданные характеристики F-критерия в рамках модели II, а также вычисление его мощности, приводятся в упр. 3.9.105 и 3.9.106.

Вычислительная схема реализации двухфакторного дисперсионного анализа обсуждается в упр. 3.9.107. Отметим, что если число факторов больше единице и одни факторы соответствуют модели I с фиксированными эффектами, а другие – модели II со случайными эффектами, то вся модель называется *моделью со смешанными эффектами*. Детальное обсуждение дисперсионного анализа в рамках таких моделей может быть найдено в монографии Шеффе [80], см. также [3, 5, 20, 69].

Пример 3.8.79. Статистические данные для этого примера взяты из [9, с. 300]. Эксперимент описывается моделью II со случайными эффектами и характеризуется следующими параметрами: число групп $k = 22$, объем выборки в каждой группе $m = 4$, то есть общее число наблюдений $n = km = 88$ и $n - k = 66$. Вычисленное значение общего группового выборочного среднего $\overline{X}_{\bullet\bullet} = 45{,}659$ и заполненная таблица дисперсионного анализа выглядит следующим образом.

Таблица 3.8.5

Сумма кв. отклонений	Степени свободы	Средний квадрат	Наблюденное значение F-статистики
Межгрупповая $SS_B = 1919{,}773$	$(k-1)=21$	$S_B^2 = 91{,}418$	$F = \dfrac{S_B^2}{S_W^2} = \dfrac{91{,}418}{23{,}394} = 3{,}91$
Внутригрупповая $SS_W = SS_T - SS_B = 1544$	$(n-k)=66$	$S_W^2 = 23{,}394$	
Общая $SS_T = 3463{,}773$	$(n-1)=87$	$SS_T/(n-1) = 39{,}813$	

Применяя F-критерий (3.8.30) для проверки гипотезы $H_0 : \sigma_a^2 = 0$, с учетом того, что вычисленное значение статистики $F = S_B^2 / S_W^2 = 3{,}91$ и $F_{0{,}9995}(21, 66) = 2{,}94$, заключаем, что *P-значение* критерия меньше $0{,}0005$, то есть гипотезу $H_0 : \sigma_a^2 = 0$ следует отвергнуть. Согласно (3.8.60) и (3.8.62), точечные оценки параметров σ^2 и σ_a^2 соответственно равны: $\hat{\sigma}^2 = S_W^2 = 23{,}394$ и $\hat{\sigma}_a^2 = (S_B^2 - S_W^2)/m = (91{,}418 - 23{,}394)/4 = 17{,}006$. Далее, $(1-\alpha)$-доверительный интервал для параметра σ^2 вычисляется по формуле (3.8.61). При $\alpha = 0{,}10$ имеем $\chi^2_{0{,}95,(66)} = 85{,}96$, $\chi^2_{0{,}05,(66)} = 48{,}31$, поэтому $(1-\alpha)$-доверительный интервал для параметра σ^2 запишется в виде

$$P\left\{ \frac{66 \cdot 23{,}394}{85{,}96} < \sigma^2 < \frac{66 \cdot 23{,}394}{48{,}31} \right\} = 0{,}9 .$$

Таким образом, получаем 90 %-й доверительный интервал для параметра σ^2 в виде $(17{,}96\,;\,31{,}96)$.

Грубый доверительный интервал для параметра σ_a^2 определяется выражением (3.8.67) и записывается в виде $\{ (S_B^2 - S_W^2)/m \pm \lambda_{1-\alpha/2} \cdot \sqrt{\hat{D}(\hat{\sigma}_a^2)} \}$. При $\alpha = 0{,}10$, $\lambda_{0{,}95} = 1{,}645$, с учетом того, что $\hat{\sigma}_a^2 = (S_B^2 - S_W^2)/m = 17{,}006$ и

$$\hat{D}(\hat{\sigma}_a^2) = 2[(S_B^2)^2/(k-1) + (S_W^2)^2/(km-k)]/m^2 =$$
$$= 2[(91{,}418)^2/21 + (23{,}394)^2/66]/4^2 = 50{,}78 ,$$

получаем грубый 90 %-й доверительный интервал для параметра σ_a^2 в виде

$$\{ \hat{\sigma}_a^2 \pm \lambda_{1-\alpha/2} \cdot \sqrt{\hat{D}(\hat{\sigma}_a^2)} \} = \{ 17{,}006 \pm 1{,}645 \cdot \sqrt{50{,}78} \} \Rightarrow \{ 5{,}283\,;\, 28{,}729 \} .$$

Получим теперь точный доверительный интервал для отношения σ_a^2/σ^2. Согласно (3.8.68), точный $(1-\alpha)$-доверительный интервал для отношения σ_a^2/σ^2 при $\alpha=0{,}10$, $F_{0,95}(21,66)=1{,}72$, $F_{0,95}(66,21)=1{,}91$ запишется в виде

$$\left\{\frac{1}{4}\left(\frac{91{,}418}{23{,}394}\cdot\frac{1}{1{,}72}-1\right);\ \frac{1}{4}\left(\frac{91{,}418\cdot 1{,}91}{23{,}394}-1\right)\right\}\Rightarrow\{0{,}32\ ;\ 1{,}62\}.$$

Согласно (3.8.76), доверительный интервал для параметра μ при $\alpha=0{,}10$, $t_{0,95,(21)}=1{,}721$ запишется в виде

$$\bar{X}_{\bullet\bullet}\pm t_{1-\alpha/2,(k-1)}\cdot\sqrt{S_B^2/km}=$$

$$=45{,}659\pm 1{,}721\cdot\sqrt{91{,}418/88}=45{,}659\pm 1{,}754\Rightarrow\{43{,}9\ ;\ 47{,}4\}.$$

Замечание 3.8.80. Отметим, что основная цель дисперсионного анализа состоит в том, чтобы выяснить, есть ли влияние меняющегося фактора A на итоговый показатель X, или оно отсутствует. В тех случаях, когда это влияние установлено (то есть нулевая гипотеза об отсутствии эффектов обработки является значимой и отвергнута), то возникают задачи, связанные с оценкой величины влияния различных уровней фактора A на отклик X. Эти задачи сводятся к построению оценок параметров μ, $\{\xi_j\}$, σ^2 модели I дисперсионного анализа с фиксированными эффектами вида (3.8.7) и параметров μ, σ^2 и σ_a^2 модели II дисперсионного анализа со случайными эффектами вида (3.8.50). При решении этих задач основное внимание уделяется вопросам точности и достоверности полученных оценок, поэтому строятся доверительные интервалы для параметров моделей наблюдений и проводится оценка контрастов различных факторов. Отметим, что при этом все выводы о точности остаются справедливыми лишь *при адекватности гауссовской модели наблюдений*. Если же предположения нормальности не могут быть обоснованы, то следует воспользоваться процедурами, свободными от распределения, которые будут обсуждаться при рассмотрении непараметрических моделей дисперсионного анализа в разд. 5.6.

Отметим также, что при решении конкретных практических задач с использованием однофакторных моделей дисперсионного анализа могут возникать ситуации, когда нулевая гипотеза об отсутствии эффектов обработки не является значимой, то есть она не отвергается, и делается вывод о том, что изменение фактора Y не влияет на итоговый показатель X, хотя содержательные соображения указывают, что такое влияние должно быть и этот вывод противоречит сложившимся представлениям или результатам предыдущих исследований. Причиной такого явления может быть большой внутригрупповой разброс, на фоне которого действие фактора Y остается незамет-

ным. Этот разброс мог быть вызван действием другого неучтенного нами фактора. В таких случаях полезно продолжить исследование с использованием моделей двухфакторного дисперсионного анализа. Если мы можем выделить такой фактор, то его включают в двухфакторную модель, чтобы уменьшить статистическую неоднородность наблюдений и в результате этого обнаружить влияние на отклик X изучаемого нами фактора Y. В литературе (см., например, [65]) при переходе к двухфакторной модели изучаемый фактор Y называют *главным фактором*, а неучтенный нами фактор – «*мешающим фактором*». Конечно, не всегда дополнительное изучение однофакторной модели сводится к введению только одного «мешающего» фактора. Часто в качестве обобщения рассматривают и многофакторные модели (см., например, [3, 20, 69]). Для устранения нежелательного влияния мешающего фактора также используются специальные приемы получения статистических данных. Эти приемы направлены на то, чтобы уменьшить влияние мешающего фактора на отклик X за счет усреднения. Это достигается путем выбора специальной группы объектов наблюдений, для которых при фиксированном уровне фактора Y влияние мешающего фактора на изучаемый признак X уменьшается в результате усреднения. При этом в таблицу однофакторного анализа заносятся средние значения измеряемого признака X по выбранной группе объектов. Отметим, что при решении практических задач в результате проведения дополнительных исследований осуществляют переход не только от однофакторных моделей к двухфакторным, но и наоборот. Если при решении двухфакторной задачи оказывается, что влияние одного из факторов несущественно, то обычно задачу сводят к однофакторной как более простой в интерпретации.

3.8.Д. Двухфакторный дисперсионный анализ

Рассмотрим случай, когда на итоговый показатель X (или на отклик X) могут оказывать влияние два фактора A и B. При решении практических задач используют различные модели двухфакторного дисперсионного анализа, а именно, когда оба рассматриваемых фактора A и B с *фиксированными эффектами* (*модель* I) либо оба фактора *со случайными эффектами* (*модель* II), либо один фактор, например A с фиксированными эффектами, другой фактор B – со случайными эффектами (*смешанная модель*). Пусть фактор A имеет k дискретных уровней $A_1,...,A_k$, а фактор B – r дискретных уровней $B_1,...,B_r$. Комбинацию уровней (B_i, A_j) называют (i, j)-ячейкой, $i = 1,..., r$, $j = 1,...,k$. В литературе (см., например, [3]) рассматривают два типа отношений между факторами A и B, их называют *пересечением и группировкой*. Факторы A и B называют *пересекающимися*, если в плане эксперимента представлены *все возможные сочетания уровней факторов*, *общее число которых равно* rk. Далее, говорят, что фактор B группируется фактором A

(или фактор A *группирует* фактор B), если каждый уровень фактора A встречается в паре не более чем с одним уровнем фактора B. Мы обсудим вариант пересекающихся факторов. Обсуждение двухфакторной модели с группировкой (*или двухфакторной иерархической модели*), когда один фактор группирует другой, может быть найдено в [3, 80].

Предположим, что в каждой (i,j)-ячейке, которая соответствует i-му уровню B_i, $i=1,...,r$, фактора B и j-му уровню A_j, $j=1,...,k$, фактора A, имеется выборка $X_{ij1},...,X_{ijn_{ij}}$ объема n_{ij} из некоторого непрерывного распределения. Она может быть получена путем наблюдения признака X у n_{ij} случайно выбранных объектов изучаемой совокупности либо путем измерения n_{ij} раз признака X единственного экспериментального объекта в (i,j)-ячейке, который обладает i-м уровнем фактора B и j-м уровнем фактора A. При решении практических задач различают случаи, когда все объемы выборок n_{ij} одинаковы и равны числу n. В таких случаях эксперименты называют *повторяемыми*, а при $n=1$ их называют *неповторяемыми* (с одним наблюдением на (i,j)-ячейку). На практике наиболее часто используется аддитивная модель наблюдений, согласно которой значение отклика X_{ij} в (i,j)-ячейке является суммой самостоятельных вкладов соответствующих уровней каждого из факторов A и B, а также независимых от этих факторов случайных величин, связанных с ошибками измерений. Эти случайные величины характеризуют внутреннюю изменчивость отклика при фиксированных уровнях факторов, которая может порождаться различными причинами. Наиболее часто используются предположения нормальности, а в тех случаях, когда они не могут быть обоснованы, используют непараметрические модели. В данном разделе обсуждаются нормальные модели двухфакторного дисперсионного анализа. Обсуждение непараметрических моделей двухфакторного дисперсионного анализа приводится в разд. 5.6.

Для нормальной модели с *повторениями* наблюдаемые данные X_{ijl} в каждой (i,j)-ячейке представляют собой случайную выборку из нормального распределения, то есть $L(X_{ijl})=N(\mu_{ij},\sigma^2)$. Эти данные записываются в виде

$$X_{ijl}=\mu_{ij}+\varepsilon_{ijl},\ i=1,...,r,\ j=1,...,k,\ l=1,...,n, \qquad \textbf{(3.8.81)}$$

где μ_{ij} – средний отклик, соответствующий i-му уровню B_i, $i=1,...,r$, фактора B и j-му уровню A_j, $j=1,...,k$, фактора A, а ε_{ijl} – независимые в совокупности случайные величины, которые имеют нормальное распределение с нулевым средним и одинаковой дисперсией σ^2, то есть выполняется выражение $L(\varepsilon_{ijl})=N(0,\sigma^2)$ для всех $i=1,...,r$, $j=1,...,k$, $l=1,...,n$. Цель двухфак-

торного дисперсионного анализа состоит в том, чтобы построить оценки параметров $\{\mu_{ij}\}$, σ^2, а также проверить гипотезу об отсутствии взаимодействия уровней факторов A и B (такую модель называют *аддитивной*) и проверить гипотезы о равенстве средних значений по уровням факторов A и B. Простейшими и часто используемыми на практике моделями двухфакторного анализа являются *неповторяемые* (с одним наблюдением на (i, j)-ячейку, то есть $n = 1$) двухфакторные планы без взаимодействия. Для этих моделей изучают фиксированные эффекты факторов, случайные эффекты факторов и планы с рандомизированными блоками (смешанные модели).

Неповторяемые двухфакторные планы без взаимодействия. Фиксированные и случайные эффекты, смешанная модель. Следуя [3], в данном разделе и ниже, чтобы удобнее было различать варианты фиксированных и случайных эффектов, будем обозначать греческими буквами α_j, β_i *фиксированные* главные эффекты факторов A и B, а *случайные* эффекты этих факторов будем обозначать латинскими буквами a_j, b_i, $j = 1,...,k$, $i = 1,...,r$.

Неповторяемые двухфакторные планы (то есть с одним наблюдением на клетку) *с фиксированными эффектами* используют модель наблюдений вида

$$X_{ij} = \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \varepsilon_{ij},\ i = 1,...,r,\ j = 1,...,k. \qquad \textbf{(3.8.82)}$$

Для этой модели предполагается, что ошибки измерений ε_{ij} являются независимыми и одинаково распределенными случайными величинами, которые имеют нормальное распределение, то есть выполняется выражение $L(\varepsilon_{ij}) = N(0, \sigma^2)$ для всех $i = 1,...,r$, $j = 1,...,k$. Далее, общее среднее μ записывается как

$$\mu = \frac{1}{rk}\sum_{i=1}^{r}\sum_{j=1}^{k}\mu_{ij}\ , \qquad \textbf{(3.8.83)}$$

а параметры α_j, β_i и $(\alpha\beta)_{ij}$ определяются следующим образом. Параметр α_j характеризует j-й дифференциальный эффект фактора A (эффект столбца), который определяется в виде

$$\alpha_j = \sum_{i=1}^{r}(\mu_{ij}/r) - \mu = \mu_{\bullet j} - \mu\ ,\ j = 1,...,k, \qquad \textbf{(3.8.84)}$$

параметр β_i обозначает i-й главный эффект фактора B (эффект строки), который имеет вид

$$\beta_i = \sum_{j=1}^{k}(\mu_{ij}/k)\ - \mu = \mu_{i\bullet} - \mu\ ,\ i = 1,...,r. \qquad \textbf{(3.8.85)}$$

Величина $(\alpha\beta)_{ij}$ называется двухфакторным взаимодействием j-го уровня фактора A и i-го уровня фактора B. Эта величина учитывает дифференциальный эффект (i, j)-й комбинации факторов A, B

$$(\alpha\beta)_{ij} = \mu_{ij} + \mu - \left(\sum_{i=1}^{r} (\mu_{ij}/r) + \sum_{j=1}^{k} (\mu_{ij}/k) \right) = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu,$$
$$i = 1,...,r,\ j = 1,...,k.$$

Модель, в которой взаимодействия $(\alpha\beta)_{ij} = 0$ при всех $i = 1,...,r$, $j = 1,...,k$, называется *аддитивной*. Отметим также, что обычно двухфакторный дисперсионный анализ начинают с проверки нулевой гипотезы об аддитивности модели, которая записывается в виде H_{AB}: *все* $(\alpha\beta)_{ij} = 0$ (см. замечание 3.8.99). Гипотезы об отсутствии главных эффектов факторов A и B соответственно записываются как $H_A: \alpha_1 = ... = \alpha_k = 0$ и $H_B: \beta_1 = ... = \beta_k = 0$.

Итак, если $(\alpha\beta)_{ij} = 0$ для всех $i = 1,...,r$, $j = 1,...,k$, то такую модель называют аддитивным планом с двумя фиксированными эффектами (*аддитивная модель без повторений*). Согласно этой модели, наблюдаемые данные X_{ij} в (i, j)-ячейке представляются в виде суммы общего среднего μ, главного эффекта α_j фактора A (эффект столбца), главного эффекта β_i фактора B (эффект строки) и нормальных ошибок измерений ε_{ij}, для которых выполняется выражение $L(\varepsilon_{ij}) = N(0, \sigma^2)$ для всех $i = 1,...,r$, $j = 1,...,k$, то есть

$$X_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij},\ i = 1,...,r,\ j = 1,...,k. \qquad \textbf{(3.8.86)}$$

Неповторяемые двухфакторные планы со случайными эффектами используют модель наблюдений вида

$$X_{ij} = \mu + a_j + b_i + \varepsilon_{ij},\ i = 1,...,r,\ j = 1,...,k, \qquad \textbf{(3.8.87)}$$

где μ – генеральное среднее, a_j, b_i, ε_{ij} – взаимно независимые случайные величины, причем предполагается, что $L(\alpha_j) = N(0, \sigma_a^2)$, $L(b_i) = N(0, \sigma_b^2)$, а $L(\varepsilon_{ij}) = N(0, \sigma^2 + \sigma_{ab}^2)$. Отметим, что дисперсия ошибок измерений в данном случае равна сумме двух дисперсий σ_{ab}^2 и σ^2.

Смешанные двухфакторные планы без взаимодействия (*планы с рандомизированными блоками*) используют модель наблюдений вида

$$X_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij},\ i = 1,...,r,\ j = 1,...,k, \qquad \textbf{(3.8.88)}$$

где μ – генеральное среднее, параметр α_j характеризует дифференциальный эффект фактора A (с фиксированными эффектами), а параметры b_i и ε_{ij} являются взаимно независимыми случайными величинами с нормальными распределениями. Причем независимые величины b_i характеризуют случайные эффекты фактора B, и для них выполняется выражение $L(b_i) = N(0, \sigma_b^2)$, $i = 1,...,r$. Для независимых случайных величин ε_{ij}, которые характеризуют ошибки измерений, выполняется выражение $L(\varepsilon_{ij}) = N(0, \sigma^2)$ $i = 1,...,r$, $j = 1,...,k$. Для этой модели также предполагается, что между факторами A и B нет взаимодействия (модель без взаимодействия).

Фиксированные эффекты (модель I). Обсуждение моделей двухфакторного дисперсионного анализа начнем с рассмотрения самой простой модели неповторяемых двухфакторных планов с фиксированными эффектами, для которых аддитивная модель наблюдений записывается в виде (3.8.86). Эта модель часто используется в качестве обобщения однофакторной модели, для которой после дополнительного исследования оказалось, что нулевая гипотеза является незначимой, хотя это и противоречит сложившимся ранее представлениям при решении этой конкретной задачи. Причиной такого явления может быть большой внутригрупповой разброс, на фоне которого действие изучаемого фактора остается незаметным. Очень часто этот разброс вызывается не только случайными причинами, но также действием еще одного, неучтенного нами фактора. В таких случаях, при переходе от однофакторной модели к двухфакторной, используют следующую терминологию (см., например, [65]). Основной изучаемый фактор A называют *главным фактором*. Другой, неучтенный ранее фактор B, называют *мешающим фактором*. Пусть далее фактор A имеет k уровней, а фактор B – r уровней. Фактор B разбивает все объекты наблюдения на r *блоков*, каждый блок образует наблюдения, проведенные при фиксированном уровне фактора B. В блоке отклики могут значимо различаться только за счет применения к ним различных обработок, то есть за счет разных уровней фактора A. Уровни фактора A – *обработки*, в таблице отражаются по столбцам, уровни фактора B – *блоки*, отражаются по строкам табл. 3.8.6. Наблюдаемые данные $\{X_{ij}\}$, $i = 1,...,r$, $j = 1,...,k$, вида (3.8.87) заносятся в табл. 3.8.6 «блок – обработка». Традиционная терминология названия таблицы «блок – обработка» сложилась вследствие названий факторов A и B. Эта таблица содержит rk наблюдений (по одному наблюдению в каждой (i, j)-й ячейке этой таблицы). Отличие табл. 3.8.6 от табл. 3.8.1 однофакторного дисперсионного анализа заключается в том, что наблюдения в любом столбце таблицы не являются *однородными*, то есть они могут не образовывать выборки (если влияние мешающего фактора B значимо).

Таблица 3.8.6

$B \setminus A$	A_1	…	A_j	…	A_k	$\sum$
B_1	X_{11}	…	X_{1j}	…	X_{1k}	$X_{1\bullet}$
B_2	X_{21}	…	X_{2j}	…	X_{2k}	$X_{2\bullet}$
⋮	⋮		⋮		⋮	⋮
B_i	X_{i1}	…	X_{ij}	…	X_{ik}	$X_{i\bullet}$
⋮	⋮	…	⋮	…	⋮	⋮
B_r	X_{r1}		X_{rj}		X_{rk}	$X_{r\bullet}$
$\sum$	$X_{\bullet 1}$	…	$X_{\bullet j}$	…	$X_{\bullet k}$	$X_{\bullet\bullet}$

Основой для построения статистик критериев проверки гипотез $H_A : \alpha_1 = ... = \alpha_k = 0$ и $H_B : \beta_1 = ... = \beta_k = 0$ служит, как и в однофакторном дисперсионном анализе, правило сложения, согласно которому полная сумма квадратов отклонений величин X_{ij} от их общего среднего $\overline{X}_{\bullet\bullet}$ равна сумме слагаемых, обусловленных влиянием изменяющихся уровней факторов A, B и слагаемого, обусловленного случайной изменчивостью самих данных, то есть

$$\sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \overline{X}_{\bullet j} - \overline{X}_{i\bullet} + \overline{X}_{\bullet\bullet})^2 +$$

$$+\sum_{i=1}^{r}\sum_{j=1}^{k}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 + \sum_{i=1}^{r}\sum_{j=1}^{k}(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \overline{X}_{\bullet j} - \overline{X}_{i\bullet} + \overline{X}_{\bullet\bullet})^2 + k\sum_{i=1}^{r}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 + r\sum_{j=1}^{k}(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 .$$

Используя принятые обозначения, определим соответствующие суммы квадратов в виде

$$SS_T = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \overline{X}_{\bullet\bullet})^2 , \; SS_R = \sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \overline{X}_{\bullet j} - \overline{X}_{i\bullet} + \overline{X}_{\bullet\bullet})^2 ,$$

$$SS_B = k\sum_{i=1}^{r}(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 , \; SS_A = r\sum_{j=1}^{k}(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 . \qquad \textbf{(3.8.89)}$$

Правило сложения для этих сумм квадратов отклонений кратко записывается в виде

$$SS_T = SS_R + SS_A + SS_B . \qquad \textbf{(3.8.90)}$$

Отметим, что для нормальной модели (3.8.86), с учетом того, что $rk - 1 = (k-1)+(r-1)+(k-1)(r-1)$, согласно теореме Кокрена (П.3.77), SS_R, SS_A, SS_B

являются независимыми случайными величинами, причем для них выполняются следующие выражения:

$$L(SS_A/\sigma^2)=\chi^2(k-1),\ L(SS_B/\sigma^2)=\chi^2(r-1),$$
$$L(SS_R/\sigma^2)=\chi^2((k-1)(r-1))\ \text{и}\ L(SS_T/\sigma^2)=\chi^2(rk-1). \qquad \textbf{(3.8.91)}$$

При выполнении обычных условий на параметры α_j и β_i в виде $(\alpha_1+...+\alpha_k)=0$, $(\beta_1+...+\beta_r)=0$ однозначные МНК-оценки параметров μ,α_j,β_i находятся из минимума выражения $\sum(X_{ij}-\mu-\alpha_j-\beta_i)^2$. Оценки наименьших квадратов параметров μ,α_j,β_i записываются в виде

$$\hat{\mu}=\overline{X}_{\bullet\bullet}=\frac{1}{rk}\sum_{i=1}^{r}\sum_{j=1}^{k}X_{ij},\ \ \hat{\alpha}_j=\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet}=\frac{1}{r}\sum_{i=1}^{r}X_{ij}-\overline{X}_{\bullet\bullet},$$
$$\hat{\beta}_i=\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet}=\frac{1}{k}\sum_{j=1}^{k}X_{ij}-\overline{X}_{\bullet\bullet},\ i=1,...,r,\ j=1,...,k. \qquad \textbf{(3.8.92)}$$

Отметим, что для нормальной модели (3.8.86) эти оценки также имеют нормальные распределения, то есть справедливы следующие выражения:

$$L(\hat{\mu})=N(\mu,\sigma^2/rk),\ L(\hat{\alpha}_j)=N(\alpha_j,\sigma^2(k-1)/rk),\ L(\hat{\beta}_i)=N(\beta_i,\sigma^2(r-1)/rk).$$

Для построения F-статистик критериев проверки гипотез $H_A:\alpha_1=...=\alpha_k=0$ и $H_B:\beta_1=...=\beta_k=0$, как и в однофакторном дисперсионном анализе, используются две независимые оценки параметра σ^2. Одна из них основана на использовании остаточной суммы квадратов SS_R и записывается в виде

$$\hat{\sigma}_R^2=\frac{1}{(r-1)(k-1)}\sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij}-\overline{X}_{\bullet j}-\overline{X}_{i\bullet}+\overline{X}_{\bullet\bullet})^2. \qquad \textbf{(3.8.93)}$$

Оценка $\hat{\sigma}_R^2$ является несмещенной оценкой параметра σ^2, то есть $M(\hat{\sigma}_R^2)=\sigma^2$. В качестве другой оценки параметра σ^2 при проверке гипотез $H_A:\alpha_1=...=\alpha_k=0$, $H_B:\beta_1=...=\beta_k=0$ используются оценки, основанные соответственно на SS_A и SS_B, которые записываются в виде

$$\hat{\sigma}_A^2=SS_A/(k-1)=\frac{r}{k-1}\sum_{j=1}^{k}(\overline{X}_{\bullet j}-\overline{X}_{\bullet\bullet})^2,$$
$$\hat{\sigma}_B^2=SS_B/(r-1)=\frac{k}{r-1}\sum_{i=1}^{r}(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2. \qquad \textbf{(3.8.94)}$$

Можно убедиться (см. упр. 3.9.80), что математические ожидания этих оценок вычисляются по формулам

$$M(\hat{\sigma}_A^2) = \sigma^2 + \frac{r}{k-1}\sum_{j=1}^{k}\alpha_j^2,\ M(\hat{\sigma}_B^2) = \sigma^2 + \frac{k}{r-1}\sum_{i=1}^{r}\beta_i^2. \qquad \textbf{(3.8.95)}$$

Отметим, что оценка $\hat{\sigma}_A^2 = SS_A/(k-1)$ обладает свойством несмещенности только при справедливости гипотезы $H_A: \alpha_1 = ... = \alpha_k = 0$, то есть $M(\hat{\sigma}_A^2 \mid H_A) = \sigma^2$. Аналогично, оценка $\hat{\sigma}_B^2 = SS_B/(r-1)$ также является несмещенной оценкой параметра σ^2 только при справедливости гипотезы $H_B: \beta_1 = ... = \beta_k = 0$, то есть $M(\hat{\sigma}_B^2 \mid H_B) = \sigma^2$.

Итак, критерии проверки гипотезы $H_A: \alpha_1 = ... = \alpha_k = 0$ об отсутствии эффектов фактора A и гипотезы $H_B: \beta_1 = ... = \beta_k = 0$ об отсутствии эффектов фактора B соответственно основаны на статистиках

$$F_A = \frac{SS_A/(k-1)}{SS_R/(k-1)(r-1)},\ F_B = \frac{SS_B/(r-1)}{SS_R/(k-1)(r-1)}, \qquad \textbf{(3.8.96)}$$

каждая из которых, согласно (3.8.90), при справедливости основной гипотезы имеет F-распределение с числами степеней свободы ν_1 и ν_2, то есть для них справедливы выражения $L(F_A|H_A)=F(\nu_1,\nu_2)$, где $\nu_1 = k-1$, $\nu_2 = (k-1)(r-1)$, и $L(F_B \mid H_B) = F(\nu_1,\nu_2)$, где $\nu_1 = r-1$, $\nu_2 = (k-1)(r-1)$.

Таким образом, если наблюденное значение статистики $F_A(F_B)$ превышает критическое значение $F_{1-\alpha}(\nu_1,\nu_2)$, то гипотеза $H_A(H_B)$ – отвергается на уровне значимости α. Для сформулированных критериев *P-значение* определяется как площадь справа от наблюденного значения F-статистики под кривой плотности распределения Фишера $F(\nu_1,\nu_2)$, то есть, если *P-значение* $< \alpha$, то нулевая гипотеза отвергается (см. табл. 3.8.7).

Случайные эффекты (модель II). Неповторяемые двухфакторные планы со случайными эффектами используют модель наблюдений (3.8.87), которая записывается в виде

$$X_{ij} = \mu + a_j + b_i + \varepsilon_{ij},\ i = 1,...,r,\ j = 1,...,k,$$

где μ – генеральное среднее, a_j, b_i, ε_{ij} – взаимно независимые случайные величины, причем предполагается, что каждая из них имеет нормальное распределение с нулевым средним, но разными дисперсиями, то есть справедливы выражения $L(\alpha_j) = N(0, \sigma_a^2)$, $L(\beta_i) = N(0, \sigma_b^2)$, а $L(\varepsilon_{ij}) = N(0, \sigma^2 + \sigma_{ab}^2)$. Отметим еще раз, что дисперсия ошибок измерений в данном случае равна сумме двух дисперсий σ_{ab}^2 и σ^2. В этой модели выделяют четыре компоненты дисперсии: σ_a^2, σ_b^2, σ_{ab}^2 и остаточную дисперсию σ^2, которые связаны соответст-

венно с факторами A, B, (AB) и ошибками измерений $\{\varepsilon_{ij}\}$. Основой для проведения дисперсионного анализа в рамках данной модели II также служат суммы квадратов отклонений, определенные в (3.8.89). Целью двухфакторного дисперсионного анализа со случайными эффектами является проверка гипотез $H_0 : \sigma_a^2 = 0$ и $H_0 : \sigma_b^2 = 0$, а также построение оценок параметров σ_a^2, σ_b^2 и σ^2.

Можно убедиться (см., например, [3]), что для модели II математическое ожидание остаточного среднего квадрата (3.8.93) вычисляется по формуле $M(\hat{\sigma}_R^2) = M\{SS_R/(k-1)(r-1)\} = \sigma^2 + \sigma_{ab}^2$. Отсюда следует, что для оценки параметра σ^2 нужно использовать предположение вида $\sigma_{ab}^2 = 0$, что означает отсутствие совместного влияния факторов A и B. Далее, для модели II математические ожидания оценок (3.8.94) вычисляются по формулам

$$M\{SS_A/(k-1)\} = \sigma^2 + \sigma_{ab}^2 + r\sigma_a^2,\ M\{SS_B/(r-1)\} = \sigma^2 + \sigma_{ab}^2 + k\sigma_b^2. \quad \textbf{(3.8.97)}$$

Отсюда следует, что несмещенные оценки параметров σ_a^2 и σ_b^2 (при выполнении условия $\sigma_{ab}^2 = 0$) вычисляются по формулам

$$\hat{\sigma}_a^2 = \{SS_A/(k-1) - SS_R/(k-1)(r-1)\}/r,$$
$$\hat{\sigma}_b^2 = \{SS_B/(r-1) - SS_R/(k-1)(r-1)\}/k. \quad \textbf{(3.8.98)}$$

Для удобства дальнейших ссылок, следуя [3], приведем в табл. 3.8.7 критерии проверки гипотез $H_A : \alpha_1 = ... = \alpha_k = 0$ и $H_B : \beta_1 = ... = \beta_k = 0$ в рамках модели I для аддитивного неповторяемого двухфакторного плана (то есть без взаимодействия факторов), а также критерии проверки гипотез $H_0 : \sigma_a^2 = 0$ и $H_0 : \sigma_b^2 = 0$ в рамках модели II, основанные на статистиках F_A и F_B из (3.8.96).

Таблица 3.8.7

Модель I, фиксированные эффекты	$H_0: \alpha_1 = ... = \alpha_k = 0$	$H_0: \beta_1 = ... = \beta_k = 0$
Модель II, случайные эффекты	$H_0: \sigma_a^2 = 0$	$H_0: \sigma_b^2 = 0$
Если $F > F_{1-\alpha}(\nu_1, \nu_2)$, то гипотеза H_0 отвергается на уровне значимости α, в противном случае нет оснований возражать против гипотезы H_0	$F_A = \dfrac{SS_A/(k-1)}{SS_R/(k-1)(r-1)}$. $L(F_A \mid H_A) = F(\nu_1, \nu_2)$, $\nu_1 = k-1,\ \nu_2 = (k-1)(r-1)$	$F_B = \dfrac{SS_B/(r-1)}{SS_R/(k-1)(r-1)}$. $L(F_B \mid H_B) = F(\nu_1, \nu_2)$, $\nu_1 = r-1,\ \nu_2 = (k-1)(r-1)$

Смешанная модель. *Смешанные двухфакторные планы без взаимодействия* (*планы с рандомизированными блоками*) используют модель наблюдений (3.8.88), которая записывается в виде $X_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij}$, $i = 1,...,r$, $j = 1,...,k$, где μ – генеральное среднее, параметр α_j характеризует дифференциальный эффект фактора A (с фиксированными эффектами), а параметры b_i и ε_{ij} являются взаимно независимыми случайными величинами с нормальными распределениями. Причем независимые величины b_i характеризуют случайные эффекты фактора B, и для них выполняется выражение $L(b_i) = N(0, \sigma_b^2)$, $i = 1,...,r$. Для независимых случайных величин ε_{ij}, которые характеризуют ошибки измерений, выполняется выражение $L(\varepsilon_{ij}) = N(0, \sigma^2)$, $i = 1,...,r$, $j = 1,...,k$. Для этой модели также предполагается, что между факторами A и B нет взаимодействия, то есть $(\alpha\beta)_{ij} = 0$ при всех $i = 1,...,r$, $j = 1,...,k$ (модель без взаимодействия).

При выполнении обычных условий на параметры α_j в виде $(\alpha_1 + ... + \alpha_k) = 0$ однозначные МНК-оценки параметров μ, α_j вычисляются, как и для модели (3.8.87), по формулам (3.8.92) и записываются в виде

$$\hat{\mu} = \bar{X}_{\bullet\bullet} = \frac{1}{rk}\sum_{i=1}^{r}\sum_{j=1}^{k} X_{ij}\,,\ \hat{\alpha}_j = \bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet} = \frac{1}{r}\sum_{i=1}^{r} X_{ij} - \bar{X}_{\bullet\bullet}\,.$$

Несмещенной оценкой параметра σ^2 является оценка $\hat{\sigma}_R^2$ вида (3.8.93), то есть

$$\hat{\sigma}_R^2 = \frac{1}{(r-1)(k-1)}\sum_{i=1}^{r}\sum_{j=1}^{k}(X_{ij} - \bar{X}_{\bullet j} - \bar{X}_{i\bullet} + \bar{X}_{\bullet\bullet})^2\,.$$

Несмещенная оценка дисперсии σ_b^2 вычисляется по формуле (3.8.98) и записывается в виде

$$\hat{\sigma}_b^2 = \{SS_B/(r-1) - SS_R/(k-1)(r-1)\}/k\,.$$

F-критерии для проверки гипотез $H_0 : \alpha_1 = ... = \alpha_k = 0$ и $H_0 : \sigma_b^2 = 0$ приведены в табл. 3.8.7.

Отметим, что рассматриваемая смешанная модель описывает план эксперимента, который называют *планом с радомизированными блоками* (см. [3]). Эксперимент направлен на сравнение дифференциальных эффектов α_j, $j = 1,...,k$, при k способах обработки (или при k уровнях фактора A с фиксированными эффектами, описываемыми моделью I). Причем на отклик X_{ij} ока-

зывает влияние мешающий фактор B со случайными эффектами. Множество наблюдаемых объектов с фиксированным уровнем B_i, $i=1,...,r$, фактора B называют *блоком*, а его конкретный объект – *участком*. Обработки случайно распределяются по k уровням фактора A. Весь эксперимент повторяется r раз, то есть все k способов случайным образом распределяются в каждом из r блоков (фактор B со случайными эффектами, то есть соответствует модели II). Параметр σ_b^2 характеризует дисперсию между блоками. Отметим, что для данного плана эксперимента каждый способ обработки применяется только к одному участку внутри блока, по этой причине оценить взаимодействие «блок – обработка» невозможно. Поэтому в рамках этой модели и предполагается, что между факторами A и B взаимодействие отсутствует. В тех случаях, когда требуется оценить взаимодействие между факторами A и B, используют модель повторяемого плана с рандомизированными блоками, который называют *двухфакторным смешанным повторяемым планом* и для которого модель наблюдений (3.8.81) записывается в виде $X_{ijl}=\mu+\alpha_j+b_i+(\alpha b)_{ij}+\varepsilon_{ijl}$, $i=1,...,r$, $j=1,...,k$, $l=1,...,n$, $l=1,...,n$, где μ – генеральное среднее, параметр α_j характеризует дифференциальный эффект фактора A (с фиксированными эффектами), а параметры b_i, $(\alpha b)_{ij}$ и ε_{ij} являются взаимно независимыми случайными величинами с нормальными распределениями. Причем независимые величины b_i характеризуют случайные эффекты фактора B, и для них выполняется выражение $L(b_i)=N(0,\sigma_b^2)$, $i=1,...,r$. Для взаимодействий $(\alpha b)_{ij}$ выполняется выражение $L((\alpha b)_{ij})=N(0,\sigma_{ab}^2)$, $i=1,...,r$, $j=1,...,k$. Далее, для независимых случайных величин ε_{ijl}, которые характеризуют ошибки измерений, выполняется выражение $L(\varepsilon_{ijl})=N(0,\sigma^2)$, $i=1,...,r$, $j=1,...,k$, $l=1,...,n$. Обсуждение этой повторяемой модели для случаев фиксированных эффектов, случайных эффектов и смешанной модели может быть найдено в [3], см. также упр. 3.9.107.

Замечание 3.8.99. Если при анализе двухфакторной аддитивной модели без повторений возникают сомнения в аддитивности этой модели, то обычно в таких случаях поступают следующим образом. Рассматривают неаддитивную модель вида (3.8.82) и для нее проверяют гипотезу о том, что взаимодействие факторов A и B отсутствует. Эта гипотеза записывается в виде $H_{AB}: (\alpha\beta)_{ij}=0$, для всех $i=1,...,r$, $j=1,...,k$, и проверяется с помощью критерия Тьюки (см., например, [3, с. 256]), статистика которого записывается в виде $F_{AB}=(rk-k-r)SS_G/(SS_{AB}-SS_G)$, где

$$SS_G = \frac{\left(\sum_{i=1}^{r} \sum_{j=1}^{k} (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) \cdot X_{ij} \right)^2}{\sum_{i=1}^{r} (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 \cdot \sum_{j=1}^{k} (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2},$$

$$SS_{AB} = \sum_{i=1}^{r} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{\bullet j} - \bar{X}_{i\bullet} + \bar{X}_{\bullet\bullet})^2 .$$

Отметим, что при гипотезе H_{AB} справедливо выражение $L(F_{AB} \mid H_{AB}) = F(1,\ rk - k - r)$. С учетом этого замечания, *P-значение критерия Тьюки* равно площади справа от наблюденного значения статистики F_{AB} под кривой плотности *F*-распределения Фишера с числами степеней свободы 1 и $(rk - k - r)$. Если *P-значение* меньше уровня значимости α, то гипотеза H_{AB} об аддитивности модели (3.8.82) отвергается.

Пример 3.8.100. Приведем численный пример с использованием двухфакторной неповторяемой и аддитивной модели (3.8.86), оба фактора A и B с фиксированными эффектами. Наблюденные значения отклика X приведены в табл. 3.8.8.

Таблица 3.8.8

$B \setminus A$	A_1	A_2	A_3	Σ
B_1	11	10	13	34
B_2	9	10	10	29
B_3	12	14	8	34
B_4	15	11	7	33
Σ	47	45	38	130

В данном примере число уровней фактора B равно $r = 4$, а фактора A – $k = 3$. Заполним таблицу двухфакторного дисперсионного анализа. Для этого вычислим общую сумму квадратов отклонений

$$SS_T = \sum_{i=1}^{r} \sum_{j=1}^{k} (X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} X_{ij}^2 - \frac{X_{\bullet\bullet}^2}{rk} = 11^2 + 10^2 + 13^2 + \ldots + 7^2 - 130^2/12 = 61{,}667 .$$

Далее, сумма квадратов отклонений по столбцам

$$SS_A = r \sum_{j=1}^{k} (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \frac{1}{r} \sum_{j=1}^{k} X_{\bullet j}^2 - \frac{X_{\bullet\bullet}^2}{rk} = (47^2 + 45^2 + 38^2)/4 - 130^2/12 = 11{,}167 .$$

Сумма квадратов отклонений по строкам

$$SS_B = k\sum_{i=1}^{r}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 = \frac{1}{k}\sum_{i=1}^{r} X_{i\bullet}^2 - \frac{X_{\bullet\bullet}^2}{rk} = (34^2 + 29^2 + 34^2 + 33^2)/3 - 130^2/12 = 5{,}667.$$

Из (3.8.90) получаем

$$SS_R = SS_T - SS_A - SS_B = 61{,}667 - 11{,}167 - 5{,}667 = 44{,}833.$$

Итак, заполненная таблица двухфакторного дисперсионного анализа выглядит следующим образом.

Таблица 3.8.9

Источник вариации	Сумма кв. отклонений	Степени свободы	Средний квадрат	F-статистика
Между столбцами, фактор A	$SS_A = 11{,}167$	$\nu_A = k - 1 = 2$	$SS_A/\nu_A = 5{,}583$	$F_A = \dfrac{SS_A/\nu_A}{SS_R/\nu_R} = 0{,}747$
Между строками, фактор B	$SS_B = 5{,}667$	$\nu_B = r - 1 = 3$	$SS_B/\nu_B = 1{,}889$	$F_B = \dfrac{SS_B/\nu_B}{SS_R/\nu_R} = 0{,}253$
Ошибки, остаток R	$SS_R = 44{,}833$	$\nu_R = (k-1)\times$ $\times(r-1) = 6$	$SS_R/\nu_R = 7{,}472$	
Общая вариация	$SS_T = 61{,}667$			

Пусть уровень значимости $\alpha = 0{,}05$. Учитывая, что параметры $r = 4$, $k = 3$, определяем

$$F_{1-\alpha}(r-1,\ (k-1)(r-1)) = F_{0{,}95}(3\ ,\ 6) = 4{,}76,$$
$$F_{1-\alpha}(k-1,\ (k-1)(r-1)) = F_{0{,}95}(2\ ,\ 6) = 5{,}14.$$

Напомним, что если *P-значение* $< \alpha$, то нулевая гипотеза отвергается. В данном случае *P-значение* $> \alpha$ для обеих гипотез, то есть они не значимы и сохраняются. Обе нуль-гипотезы принимаются на уровне значимости $\alpha = 0{,}05$.

3.9. Упражнения, задачи и дополнения

3.9.1. Пусть имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta)$ с плотностью $f_X(x;\theta)$, $x \in R^1$ и $\theta \in \Theta$ – неизвестный вещественный параметр, Θ – заданное параметрическое множество. Проверяется простая гипотеза $H_0 : \theta = \theta_0$ против сложной правосторонней альтернативы $H_1^+ : \theta > \theta_0$ критерием уровня значимости α с критической областью $\mathrm{X}_{1\alpha}$ и функцией мощности $W(\theta) = W(\mathrm{X}_{1\alpha};\theta)$, $\theta \in \Theta$. Для этого критерия имеем

$$W(\theta) = W(\mathrm{X}_{1\alpha};\theta) = P_\theta\{\vec{X} \in \mathrm{X}_{1\alpha}\} = \int\limits_{\mathrm{X}_{1\alpha}} \cdots \int L(x_1,...x_n;\theta)dx_1 \cdots dx_n\,, \qquad \textbf{(3.9.2)}$$

где $L(x_1,...,x_n;\theta)$ – совместная плотность распределения вероятностей наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$, то есть

$$L(x_1,...,x_n;\theta) = L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)\,. \qquad \textbf{(3.9.3)}$$

Предположим далее, что для любой критической области $\mathrm{X}_{1\alpha}$ размера α функция мощности $W(\theta)$, удовлетворяющая условию

$$W(\theta_0) = W(\mathrm{X}_{1\alpha};\theta_0) = \int\limits_{\mathrm{X}_{1\alpha}} \cdots \int L(x_1,...x_n;\theta_0)dx_1 \cdots dx_n = \alpha\,, \qquad \textbf{(3.9.4)}$$

допускает разложение в ряд Тейлора в окрестности точки θ_0, то есть

$$W(\theta) = W(\theta_0) + W'(\theta_0)(\theta-\theta_0) + \frac{W''(\theta_0)}{2!}(\theta-\theta_0)^2 + ... +$$
$$+ \frac{W^{(n)}(\theta_0)}{n!}(\theta-\theta_0)^n + o((\theta-\theta_0)^n)\,. \qquad \textbf{(3.9.5)}$$

Для простоты записи ниже будем использовать следующие обозначения:

$$L_k(\vec{x};\theta) = \partial^{(k)}L(\vec{x};\theta)/\partial\theta^k\,,\ k = 1,2,...,n\,. \qquad \textbf{(3.9.6)}$$

А) Убедитесь, что локально наиболее мощный (л.н.м.) критерий проверки гипотезы $H_0 : \theta = \theta_0$ против правосторонней альтернативы $H_1^+ : \theta > \theta_0$ основан на критической области $\mathrm{X}_{1\alpha}^+$ размера α, которая определяется в виде

$$\mathrm{X}_{1\alpha}^+ = \{\vec{x} : L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0) \geq t_{1-\alpha}\}\,, \qquad \textbf{(3.9.7)}$$

где критическое значение $t_{1-\alpha}$ статистики

$$\tilde{\Lambda}(\vec{X};\theta_0) = L_1(\vec{X};\theta_0)/L(\vec{X};\theta_0) \qquad \textbf{(3.9.8)}$$

находится при заданном уровне значимости α из выражения

$$\int_{t_{1-\alpha}}^{\infty} f_{\tilde{\Lambda}}(x;\theta_0)dx = \alpha\,. \qquad \textbf{(3.9.9)}$$

Здесь $f_{\tilde{\Lambda}}(x;\theta_0)$ – плотность распределения вероятностей статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0:\theta=\theta_0$.

Б) Убедитесь, что л.н.м. критерий проверки гипотезы $H_0:\theta=\theta_0$ против левосторонней альтернативы $H_1^-:\theta<\theta_0$ основан на критической области $\mathrm{X}_{1\alpha}^-$ размера α, которая определяется в виде

$$\mathrm{X}_{1\alpha}^- = \{\vec{x}: L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0) \le t_\alpha\}, \qquad \textbf{(3.9.10)}$$

где критическое значение t_α статистики $\tilde{\Lambda}(\vec{X};\theta_0) = L_1(\vec{X};\theta_0)/L(\vec{X};\theta_0)$ находится при заданном уровне значимости α из выражения

$$\int_{-\infty}^{t_\alpha} f_{\tilde{\Lambda}}(x;\theta_0)dx = \alpha\,. \qquad \textbf{(3.9.11)}$$

В) Убедитесь, что *несмещенный* л.н.м. критерий проверки гипотезы $H_0:\theta=\theta_0$ против двусторонней альтернативы $H_1:\theta\ne\theta_0$ основан на критической области $\tilde{\mathrm{X}}_{1\alpha}$ размера α, определяемой в виде

$$\tilde{\mathrm{X}}_{1\alpha} = \{\vec{x}: L_2(\vec{x};\theta_0) \ge c_1 L_1(\vec{x};\theta_0) + cL(\vec{x};\theta_0)\}, \qquad \textbf{(3.9.12)}$$

где константы c и c_1 находятся соответственно из условий $W(\theta_0)=\alpha$, $W'(\theta_0)=0$.

Указание. Для доказательства справедливости (3.9.7) и (3.9.10) следует воспользоваться схемой доказательства леммы Неймана – Пирсона и приближенным выражением для функции мощности $W(\theta)$, содержащим два члена в разложении (3.9.5). Для доказательства (3.9.12) необходимо использовать приближенное выражение для функции мощности $W(\theta)$ в виде

$$W(\theta) = W(\theta_0) + W'(\theta_0)(\theta-\theta_0) + \frac{W''(\theta_0)}{2!}(\theta-\theta_0)^2 + \ldots \qquad \textbf{(3.9.13)}$$

с учетом того, что $W(\theta_0)=\alpha$ и $W'(\theta_0)=0$ – условие несмещенности л.н.м. критерия, критическая область которого определяется из условия максимума второй производной $W''(\theta_0)$ (см. также доказательство теоремы 4.5 в работе [22]).

3.9.14. Пусть $L(X) = N(\theta, \sigma^2)$. Эта запись означает, что плотность распределения вероятности случайной величины X нормальная с неизвестным параметром θ и известным параметром σ^2. Имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta) = \Phi((x-\theta)/\sigma)$ с плотностью $f_X(x;\theta) = (2\pi\sigma^2)^{-1/2}\exp\{-(x-\theta)^2/2\sigma^2\}$, $x \in R^1$, где $\Phi(x)$ – стандартная нормальная ф.р. Требуется построить л.н.м. критерий уровня значимости α для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \neq \theta_0$.

Решение. Совместная плотность наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$ в данном случае записывается в виде

$$L(x_1,...,x_n;\theta) = L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2\right\}.$$

Далее,

$$L_1(\vec{x};\theta) = \partial L(x;\theta)/\partial\theta = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \theta\right) L(\vec{x};\theta),$$

$$L_2(\vec{x};\theta) = \partial^2 L(x;\theta)/\partial\theta^2 = \{\ \frac{n^2}{\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \theta\right)^2 - n/\sigma^2\ \}\ L(\vec{x};\theta).$$

Таким образом,

$$\frac{L_1(\vec{x};\theta_0)}{L(\vec{x};\theta_0)} = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \theta_0\right) = n(\overline{x} - \theta_0)/\sigma^2,\ \ \frac{L_2(\vec{x};\theta_0)}{L(\vec{x};\theta_0)} = n^2(\overline{x} - \theta_0)^2/\sigma^4 - n/\sigma^2.$$

Отсюда следует с учетом (3.9.12), что критическая область $\tilde{X}_{1\alpha}$ определяется неравенством вида $T^2(\vec{x}) \geq k_1 T(\vec{x}) + k_2$, где $T(\vec{x}) = \sqrt{n}(\overline{x} - \theta_0)/\sigma$. Отметим, что при гипотезе $H_0 : \theta = \theta_0$ статистика $T(\vec{X}) = \sqrt{n}(\overline{X} - \theta_0)/\sigma$ имеет стандартное нормальное распределение, которое симметрично относительно нуля. Поэтому, для того чтобы было выполнено условие несмещенности критерия в виде $W'(\theta_0) = 0$, критическая область также должна быть симметричной, что будет иметь место только при $k_1 = 0$. Таким образом, критическая область несмещенного л.н.м. критерия размера α будет определятся неравенством вида $|T(\vec{x})| \geq t_{1-\alpha}$, в котором критическое значение $t_{1-\alpha}$ статистики $T(\vec{X})$, согласно условию $W(\theta_0) = \alpha$, с учетом того, что $L\{T(\vec{X}) \mid H_0\} = N\{\theta_0; \sigma^2/n\}$, определяется в виде $t_{1-\alpha} = \lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Таким образом, несмещенный л.н.м. критерий размера α для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \neq \theta_0$ определяется критической областью $\tilde{X}_{1\alpha} = \{\vec{x} : |T(\vec{x})| \geq \lambda_{1-\alpha/2}\}$ и совпадает с несмещенным р.н.м. критерием, построенным в примере 3.4.24.

3.9.15. Рассмотрим *регулярную* (см. разд. 2.3) параметрическую модель $\{\mathrm{X}, \Im_\theta\}$, для которой множество допустимых распределений определено в виде параметрического семейства *заданных* функций $F_X(x,\theta)$ с конечным числом *неизвестных* параметров θ, то есть $\Im_\theta = \{ F_X : F_X(x,\theta), \theta \in \Theta \}$, где Θ – *заданное* множество параметров. Пусть имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta)$ с плотностью $f_X(x;\theta)$, $x \in R^1$ и $\theta \in \Theta$ – неизвестный вещественный параметр.

А) Требуется построить асимптотические л.н.м. критерии для проверки гипотезы $H_0 : \theta = \theta_0$ против правосторонней альтернативы $H_1^+ : \theta > \theta_0$ (против левосторонней альтернативы $H_1^- : \theta < \theta_0$).

Б) Построить асимптотически л.н.м. несмещенный двусторонний критерий для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta \neq \theta_0$.

Решение. А) Для проверки гипотез $H_0 : \theta = \theta_0$, $H_1^+ : \theta > \theta_0$ критическая область, соответствующая л.н.м. критерию, определена выражением (3.9.7), в котором критическое значение $t_{1-\alpha}$ статистики $\tilde{\Lambda}(\vec{X};\theta_0) = L_1(\vec{X};\theta_0)/L(\vec{X};\theta_0)$ определяется при заданном уровне значимости α из выражения (3.9.9). Для вычисления критического значения $t_{1-\alpha}$ требуется знание плотности распределения вероятностей $f_{\tilde{\Lambda}}(x;\theta_0)$ статистики $\tilde{\Lambda}(\vec{X};\theta_0)$ при справедливости гипотезы $H_0 : \theta = \theta_0$. В случае, когда объем выборки неограниченно возрастает, плотность распределения вероятностей $f_{\tilde{\Lambda}}(x;\theta_0)$ для регулярных параметрических моделей $\{\mathrm{X}, \Im_\theta\}$ может быть аппроксимирована нормальной плотностью. В самом деле, для наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ имеем

$$\frac{L_1(\vec{x};\theta_0)}{L(\vec{x};\theta_0)} = \frac{\partial}{\partial\theta}\ln L(\vec{x};\theta)\,|_{\theta=\theta_0} = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f_X(x_i;\theta)\,|_{\theta=\theta_0} = U(\vec{x};\theta_0),$$

где $U(\vec{x};\theta_0)$ – наблюденное значение функции вклада выборки $U(\vec{X};\theta_0)$, определенной в (2.3.4), при справедливости гипотезы $H_0 : \theta = \theta_0$. Отметим, что, согласно (2.3.7) и (2.3.13а), справедливы выражения

$$M[U(\vec{X};\theta_0)] = 0,\ D[U(\vec{X};\theta_0)] = n i_1(\theta_0),$$

где $i_1(\theta_0)$ – количество информации Фишера относительно параметра θ_0, содержащееся в одном наблюдении и определенное в (2.3.11). С учетом этих замечаний, согласно центральной предельной тереме (П.3.24), при условии, что $0 < i_1(\theta_0) < \infty$, справедливо выражение

$$L\{ U(\vec{X};\theta_0)/\sqrt{n i_1(\theta_0)} \mid H_0 \} = N\{0;1\} \text{ при } n \to \infty. \qquad \textbf{(3.9.16)}$$

Таким образом, критическая область (3.9.7) размера α, определяющая л.н.м. критерий проверки гипотезы $H_0 : \theta = \theta_0$ против правосторонней альтернативы $H_1^+ : \theta > \theta_0$ запишется при $n \to \infty$ в виде

$$X_{1\alpha}^+ = \{\vec{x} : L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0) \geq t_{1-\alpha}\} = \{\vec{x} : U(\vec{x};\theta_0) \geq \lambda_{1-\alpha}\sqrt{ni_1(\theta_0)}\}, \quad \textbf{(3.9.16a)}$$

где $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$ – квантиль уровня $(1-\alpha)$ стандартного нормального распределения.

Аналогично, критическая область (3.9.10) размера α, определяющая л.н.м. критерий проверки гипотезы $H_0 : \theta = \theta_0$ против левосторонней альтернативы $H_1^- : \theta < \theta_0$, при $n \to \infty$ запишется в виде

$$X_{1\alpha}^- = \{\vec{x} : L_1(\vec{x};\theta_0)/L(\vec{x};\theta_0) \leq t_{\alpha}\} = \{\vec{x} : U(\vec{x};\theta_0) \leq \lambda_{\alpha}\sqrt{ni_1(\theta_0)}\}, \quad \textbf{(3.9.17)}$$

где $\lambda_{\alpha} = \Phi^{-1}(\alpha)$.

Б) Объединяя полученные результаты для односторонних альтернатив H_1^+ и H_1^- в соответствии с описанным выше правилом (3.4.37) , в случае двусторонней альтернативы, т.е. для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \neq \theta_0$ получаем, что асимптотически л.н.м. «симметричный» критерий уровня значимости α будет основан на критической области

$$\tilde{X}_{1\alpha} = \{\vec{x} : |U(\vec{x};\theta_0)| \geq \lambda_{1-\alpha/2}\sqrt{ni_1(\theta_0)}\}. \quad \textbf{(3.9.18)}$$

3.9.19. Постройте асимптотически л.н.м. критерий уровня значимости α проверки гипотез $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$ для модели Коши. Имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta)$ с плотностью

$$f_X(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \; x \in R^1. \quad \textbf{(3.9.20)}$$

Решение. Для наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $\vec{X} = (X_1,...,X_n)$ совместная плотность распределения вероятностей случайного вектора $\vec{X} = (X_1,...,X_n)$ запишется в виде

$$L(x_1,...,x_n;\theta) = L(\vec{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta) = \pi^{-n}\prod_{i=1}^{n}(1+(x_i-\theta)^2)^{-1}. \quad \textbf{(3.9.21)}$$

Далее, наблюденное значение $U(\vec{x};\theta_0)$ функции вклада выборки $U(\vec{X};\theta_0)$ при справедливости гипотезы $H_0 : \theta = \theta_0$ запишется в виде

$$U(\vec{x};\theta_0) = \partial \ln L(\vec{x};\theta)/\partial\theta\,|_{\theta=\theta_0} =$$

$$= \frac{\partial}{\partial\theta}\{-n\ln\pi - \sum_{i=1}^{n}\ln(1+(x-\theta)^2)\}\,|_{\theta=\theta_0} = 2\sum_{i=1}^{n}\frac{x_i-\theta_0}{1+(x_1-\theta_0)^2}. \qquad \textbf{(3.9.22)}$$

Учитывая теперь, что в данном случае количество информации Фишера $i_1(\theta_0)=1/2$, из (3.9.18) с использованием (3.9.22) получаем, что асимптотически л.н.м. критерий уровня значимости α проверки гипотез $H_0:\theta=\theta_0$, $H_1:\theta\neq\theta_0$ для модели Коши основан на критической области, которая определяется неравенством вида

$$\frac{2\sqrt{2}}{\sqrt{n}}\left|\sum_{i=1}^{n}\frac{x_i-\theta_0}{1+(x_i-\theta_0)^2}\right| \geq \lambda_{1-\alpha/2}\ ,\ \lambda_{1-\ \alpha/2}=\Phi^{-1}(1-\alpha/2). \qquad \textbf{(3.9.23)}$$

3.9.24. Отметим, что в литературе (см., например, [24, с. 225]) критическую область, мощность которой не меньше мощности любой другой области того же размера, предназначенной для проверки гипотезы H_0 против альтернативной гипотезы H_1, называют наилучшей критической областью (НКО), а критерий, основанный на НКО, называют наиболее мощным (н.м.) критерием. Если обе сравниваемые гипотезы относятся к значению параметра θ, для которого имеется достаточная статистика $T(\vec{X})$, то согласно критерию факторизации (2.5.17), функция правдоподобия $L(\vec{x};\theta)$ для наблюденной реализации $\vec{x}=(x_1,...,x_n)$ выборки $\vec{X}=(X_1,...,X_n)$ записывается в виде $L(\vec{x};\theta)=\varphi(T(\vec{x});\theta)h(\vec{x})$ и поэтому статистика Неймана – Пирсона в виде отношения правдоподобия будет функцией достаточной статистики $T(\vec{X})$, т.е.

$$\frac{L(\vec{X};\theta_1)}{L(\vec{X};\theta_0)} = \frac{\varphi(T(\vec{X});\theta_1)}{\varphi(T(\vec{X});\theta_0)}. \qquad \textbf{(3.9.25)}$$

Далее, если $\varphi(T(\vec{x});\theta_1)/\varphi(T(\vec{x});\theta_0)$ является монотонно неубывающей функцией от $T(\vec{x})$, то для случая $\theta_1>\theta_0$ н.м. критерий проверки простых гипотез $H_0:\theta=\theta_0$, $H_1:\theta=\theta_1$ будет определяться НКО размера α, которая записывается в виде $\mathrm{X}_{1\alpha}=\{\vec{x}:T(\vec{x})\geq t_{1-\alpha}\}$, где критическое значение $t_{1-\alpha}$ статистики $T(\vec{X})$ при заданном уровне значимости α вычисляется с помощью выражения $P_{\theta_0}\{T(\vec{X})\in\mathrm{X}_{1\alpha}\}=\alpha$. Отметим, что отношение $\varphi(T(\vec{x});\theta_1)/\varphi(T(\vec{x});\theta_0)$ является монотонно неубывающей функцией от $T(\vec{x})$ для $\theta_1>\theta_0$, если выполняется неравенство

$$\frac{\partial^2}{\partial\theta\,\partial T}\ln\varphi(T(\vec{x})\,|\,\theta)\geq 0, \qquad \textbf{(3.9.26)}$$

которое справедливо для многих (для почти всех) распределений, встречающихся в статистике [24, с. 225].

3.9.27. Покажите, что рандомизированный критерий, основанный на критической функции (3.3.36), является наиболее мощным критерием проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$ (см. [37, с. 97; 22, с. 144]).

3.9.28. Пусть случайная величина X имеет распределение Бернулли с вероятностью «успеха» θ, т.е. $L(X) = Bi(1;\theta)$. Предполагается, что параметр θ неизвестен. Имеется выборка $\vec{X} = (X_1,...,X_n)$ из распределения $F_X(x;\theta) = (1-\theta)C(x-0) + \theta C(x-1)$, $x \in R^1$. Требуется построить критерий уровня значимости α для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1 > \theta_0$.

Решение. Согласно условию, с.в. X может принимать только два возможных состояния 0 и 1 соответственно с вероятностью $1-\theta$ и вероятностью «успеха» θ, $0 \le \theta \le 1$. Отметим, что любая наблюденная реализация $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ состоит из последовательности нулей и единиц, т.е. для любой выборочной точки $\vec{x} = (x_1,...,x_n)$ имеем $x_i \in (0,1)$, $i = 1,...,n$. Далее, в данном случае функция правдоподобия запишется в виде

$$L(\vec{x};\theta_j) = \theta_j^k (1-\theta_j)^{n-k}, \; j = 0,1, \qquad \textbf{(3.9.29)}$$

где $k = k(\vec{x}) = \sum_{i=1}^{n} x_i$ – число «успехов», вычисленное по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$. Отсюда следует отношение правдоподобия $l(\vec{x}) = L(\vec{x};\theta_1)/L(\vec{x};\theta_0)$ в виде

$$l(\vec{x}) = \left(\frac{\theta_1}{\theta_0}\right)^k \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-k} = \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^k \left(\frac{1-\theta_1}{1-\theta_0}\right)^n. \qquad \textbf{(3.9.30)}$$

Далее, функция $\psi(\theta) = \theta/(1-\theta)$ возрастает на интервале $(0,1)$, поэтому при $\theta_1 > \theta_0$ $\psi(\theta_1)/\psi(\theta_0) > 1$ и, следовательно, отношение правдоподобия $l(\vec{x})$ монотонно возрастает при увеличении числа «успехов» k. Отсюда следует, что критическая область размера α, определяемая неравенством $l(\vec{x}) \ge c$, будет эквивалентна критической области вида

$$X_{1\alpha} = \{\vec{x} : k(\vec{x}) \ge k_\alpha\}, \qquad \textbf{(3.9.31)}$$

где k_α – критическое значение статистики $k(\vec{X})$, которое определяется с учетом того, что при справедливости гипотезы H_0 статистика $k(\vec{X})$ имеет биномиальное распределение, т.е. $L\{k(\vec{X}) \mid H_0\} = Bi(n,\theta_0)$. Итак, в данном примере неравенства (3.3.33) запишутся в виде

$$\sum_{i=k+1}^{n} C_n^i \theta_0^i (1-\theta_0)^{n-i} < \alpha \le \sum_{i=k}^{n} C_n^i \theta_0^i (1-\theta_0)^{n-i}. \qquad \textbf{(3.9.32)}$$

Далее, если в правом неравенстве будет выполнено равенство, т.е.

$$\alpha = \sum_{i=k(\alpha)}^{n} C_n^i \theta_0^i (1-\theta_0)^{n-i}, \qquad \textbf{(3.9.33)}$$

то критерий уровня значимости α характеризуется критической областью (3.9.31), для которой критическое значение $k_\alpha = k(\alpha)$ определяется из (3.9.33). Если же в (3.9.32) выполняются строгие неравенства, то для построения критерия с уровнем значимости α следует воспользоваться рандомизированной процедурой, описанной в разделе 3.3. Для этого введем следующие обозначения:

$$\alpha^* = \sum_{i=k+1}^{n} C_n^i \theta_0^i (1-\theta_0)^{n-i} \text{ и } \alpha^{**} = \sum_{i=k}^{n} C_n^i \theta_0^i (1-\theta_0)^{n-i}. \qquad \textbf{(3.9.34)}$$

С учетом этих обозначений неравенства (3.9.32) перепишутся в виде $\alpha^* < \alpha < \alpha^{**}$, и разность $(\alpha^{**} - \alpha^*)$ будет характеризовать вероятность того, что число «успехов» совпадет с выделенной точкой $k(\alpha)$, т.е.

$$(\alpha^{**} - \alpha^*) = P\{k(\vec{X}) = k(\alpha) \mid \theta = \theta_0\} = C_n^{k(\alpha)} \theta_0^{k(\alpha)} (1-\theta_0)^{n-k(\alpha)}. \qquad \textbf{(3.9.35)}$$

Далее, для заданного уровня значимости α определим число γ следующим образом:

$$\gamma = (\alpha - \alpha^*)/(\alpha^{**} - \alpha^*). \qquad \textbf{(3.9.36)}$$

Таким образом, общее выражение (3.3.36) для критической функции рандомизированного критерия в данном примере приводит к рандомизированной процедуре, которая предписывает принимать решение следующим образом: гипотеза $H_0 : \theta = \theta_0$ отвергается, если $k(\vec{x}) > k(\alpha)$, и принимается, если $k(\vec{x}) < k(\alpha)$; если $k(\vec{x}) = k(\alpha)$, то гипотеза H_0 отвергается с вероятностью $\gamma = (\alpha - \alpha^*)/(\alpha^{**} - \alpha^*)$ и принимается с вероятностью $(1-\gamma)$, в зависимости от исходов A и $\overline{A}$ дополнительного случайного эксперимента с вероятностями исходов γ и $(1-\gamma)$.

Итак, сформулированная процедура окончательно записывается в виде рандомизированного критерия с критической функцией вида

$$\varphi(\vec{x}) = \begin{cases} 1, \text{ при } k(\vec{x}) > k(\alpha), \\ \gamma, \text{ при } k(\vec{x}) = k(\alpha), \\ 0, \text{ при } k(\vec{x}) < k(\alpha). \end{cases} \qquad \textbf{(3.9.37)}$$

Вероятность ошибки первого рода для рандомизированного критерия, основанного на критической функции (3.9.37), равна

$$P(H_1 \mid H_0) = M_\theta[\varphi(\vec{X}) \mid \theta = \theta_0] =$$
$$= P_\theta\{k(\vec{X}) > k(\alpha) \mid \theta = \theta_0\} + \gamma \cdot P_\theta\{k(\vec{X}) = k(\alpha) \mid \theta = \theta_0\} =$$
$$= \alpha^* + \frac{\alpha - \alpha^*}{\alpha^{**} - \alpha^*}(\alpha^{**} - \alpha^*) = \alpha\,. \qquad \textbf{(3.9.38)}$$

Мощность этого критерия, согласно (3.3.38), вычисляется по формуле

$$W(\varphi;\theta_1) = P(H_1 \mid H_1) = M_\theta[\varphi(\vec{X}) \mid \theta = \theta_1] =$$
$$= P_\theta\{k(\vec{X}) > k(\alpha) \mid \theta = \theta_1\} + \gamma \cdot P_\theta\{k(\vec{X}) = k(\alpha) \mid \theta = \theta_1\} =$$
$$= \sum_{i=k(\alpha)+1}^{n} C_n^i \theta_1^i (1-\theta_1)^{n-i} + (\alpha - \alpha^*)\left(\frac{\theta_1}{\theta_0}\right)^{k(\alpha)}\left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-k(\alpha)}. \qquad \textbf{(3.9.39)}$$

Отметим, что на практике обычно используют нерандомизированные критерии и вместо заранее заданного уровня значимости α, согласно неравенствам $\alpha^* < \alpha < \alpha^{**}$, используют либо уровень значимости α^*, либо α^{**}. *P-значение* данного критерия вычисляется в упр. 3.9.73.

3.9.40. Покажите, что для проверки гипотезы $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu \neq \mu_0$ для нормальной модели $N(\mu;\sigma^2)$ в предположении, что параметр σ^2 известен, в классе двусторонних критериев уровня значимости $\alpha = \alpha_1 + \alpha_2$, определяемых критическими областями вида

$$\mathrm{X}_{1\alpha} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \le \lambda_{\alpha_1} \text{ , } \textit{либо } \sqrt{n}(\overline{x} - \mu_0)/\sigma \ge \lambda_{1-\alpha_2}\}, \qquad \textbf{(3.9.41)}$$

выбор «симметричного» критерия, то есть выбор $\alpha_1 = \alpha_2 = \alpha/2$, приводит к критической области

$$\tilde{\mathrm{X}}_{1\alpha} = \{\vec{x} : \sqrt{n} \mid \overline{x} - \mu_0 \mid /\sigma \ge \lambda_{1-\alpha/2}\}, \;\; \lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2), \qquad \textbf{(3.9.42)}$$

которая определяет *несмещенный равномерно наиболее мощный* (н.р.н.м.) критерий уровня значимости α.

Решение. Для некоторого альтернативного значения параметра $\mu \neq \mu_0$ функция мощности $W(\mathrm{X}_{1\alpha};\mu) = W(\mu)$ критерия с критической областью $\mathrm{X}_{1\alpha}$ запишется в виде

$$W(\mathrm{X}_{1\alpha};\mu) = P_\mu(\vec{X} \in \mathrm{X}^-_{1\alpha_1}) + P_\mu(\vec{X} \in \mathrm{X}^+_{1\alpha_2}),$$

где $$\mathrm{X}^-_{1\alpha_1} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \le \lambda_{\alpha_1}\},\ \lambda_{\alpha_1} = \Phi^{-1}(\alpha_1)$$

и $$\mathrm{X}^+_{1\alpha_2} = \{\vec{x} : \sqrt{n}(\overline{x} - \mu_0)/\sigma \ge \lambda_{1-\alpha_2}\},\ \lambda_{1-\alpha_2} = \Phi^{-1}(1-\alpha_2).$$

Тогда

$$P_\mu(\vec{X} \in \mathrm{X}^-_{1\alpha_1}) = P_\mu\{\,\bar{X} \le \mu_0 + \sigma\lambda_{\alpha_1}/\sqrt{n}\,\} =$$
$$= P_\mu\{\sqrt{n}(\bar{X}-\mu)/\sigma \le -\sqrt{n}(\mu-\mu_0)/\sigma + \lambda_{\alpha_1}\} = \Phi\{-\sqrt{n}(\mu-\mu_0)/\sigma + \lambda_{\alpha_1}\}.$$

Аналогично

$$P_\mu(\vec{X} \in \mathrm{X}^-_{1\alpha_2}) = P_\mu\{\,\bar{X} \ge \mu_0 + \sigma\lambda_{1-\alpha_2}/\sqrt{n}\,\} =$$
$$= P_\mu\{\sqrt{n}(\bar{X}-\mu)/\sigma \ge -\sqrt{n}(\mu-\mu_0)/\sigma + \lambda_{1-\alpha_2}\} =$$
$$= 1 - \Phi\{-\sqrt{n}(\mu-\mu_0)/\sigma + \lambda_{1-\alpha_2}\} = \Phi\{\sqrt{n}(\mu-\mu_0)/\sigma - \lambda_{1-\alpha_2}\}.$$

Таким образом, мощность критерия с критической областью $\mathrm{X}_{1\alpha}$ запишется в виде

$$W(\mathrm{X}_{1\alpha};\mu) = \Phi\{-\sqrt{n}(\mu-\mu_0)/\sigma + \lambda_{\alpha_1}\} + \Phi\{\sqrt{n}(\mu-\mu_0)/\sigma - \lambda_{1-\alpha_2}\},$$

или
$$W(\mathrm{X}_{1\alpha};\Delta) = \Phi\{-\sqrt{n}\,\Delta/\sigma + \lambda_{\alpha_1}\} + \Phi\{\sqrt{n}\,\Delta/\sigma - \lambda_{1-\alpha_2}\}, \qquad \textbf{(3.9.43)}$$

где $\Delta = \mu - \mu_0$. Исследуем поведение функции $W(\mathrm{X}_{1\alpha};\Delta)$ от параметра Δ. Для этого вычислим производные функции $W(\mathrm{X}_{1\alpha};\Delta)$ по Δ. Первая производная

$$W'(\mathrm{X}_{1\alpha};\Delta) = \frac{1}{\sigma}\sqrt{\frac{n}{2\pi}}\left[\exp\left\{-\frac{1}{2}\left(\frac{\sqrt{n}}{\sigma}\Delta - \lambda_{1-\alpha_2}\right)^2\right\} - \exp\left\{-\frac{1}{2}\left(-\frac{\sqrt{n}}{\sigma}\Delta + \lambda_{\alpha_1}\right)^2\right\}\right].$$

Отсюда находим, что $W'(\mathrm{X}_{1\alpha};\Delta) = 0$ в точке $\Delta^* = \sigma(\lambda_{\alpha_1} + \lambda_{1-\alpha_2})/2\sqrt{n}$. Далее, вторая производная записывается в виде

$$W''(\mathrm{X}_{1\alpha};\Delta) = \frac{n}{\sigma^2\sqrt{2\pi}}\left[\left(\lambda_{1-\alpha_2} - \sqrt{n}\Delta/\sigma\right)\exp\left\{-\frac{1}{2}\left(\frac{\sqrt{n}}{\sigma}\Delta - \lambda_{1-\alpha_2}\right)^2\right\} + \right.$$
$$\left. + (\sqrt{n}\Delta/\sigma + \lambda_{\alpha_1})\exp\left\{-\frac{1}{2}\left(\frac{\sqrt{n}}{\sigma}\Delta + \lambda_{\alpha_1}\right)^2\right\}\right].$$

Вычисляя вторую производную $W''(\mathrm{X}_{1\alpha};\Delta)$ в точке Δ^* и учитывая, что при малых α $(\lambda_{1-\alpha_2} - \lambda_{\alpha_2}) > 0$, получим

$$W''(\mathrm{X}_{1\alpha};\Delta^*) = (n/\sigma^2\sqrt{2\pi})(\lambda_{1-\alpha_2} - \lambda_{\alpha_1})\exp\{-(\lambda_{1-\alpha_2} - \lambda_{\alpha_2})^2/8\} > 0.$$

Таким образом, точка Δ^* является точкой локального минимума функции $W(\mathrm{X}_{1\alpha};\Delta)$. Далее, поскольку $\Delta^* = 0$ только при выполнении равенства

$\lambda_{1-\alpha_2} = -\lambda_{\alpha_1}$, из которого следует, что $\alpha_1 = \alpha_2 = \alpha/2$ и $W(\mathrm{X}_{1\alpha};0) = \Phi(\lambda_{\alpha_1}) + \Phi(-\lambda_{1-\alpha_2}) = \alpha_1 + \alpha_2 = \alpha$, то заключаем, что лишь для «симметричного» критерия с критической областью $\tilde{\mathrm{X}}_{1\alpha}$, определенной в (3.9.42), выполняется условие несмещенности критерия в виде неравенства $W(\mathrm{X}_{1\alpha};\mu) \geq \alpha$ для всех $\mu \neq \mu_0$. В остальных случаях (т.е. при $\alpha_1 \neq \alpha_2$) критерий с критической областью (3.9.41) при некоторых альтернативах имеет мощность, меньшую α, то есть является смещенным критерием. Таким образом, среди критериев с критическими областями (3.9.41) только «симметричный» критерий с критической областью (3.9.42) является *несмещенным*. Отметим, что его критическая область не зависит от альтернативного значения $\mu \neq \mu_0$ поэтому он является *несмещенным равномерно наиболее мощным* критерием для проверки гипотезы $H_0 : \mu = \mu_0$ против альтернативы $H_1 : \mu \neq \mu_0$ для нормальной модели $N(\mu\,;\sigma^2)$ в предположении, что параметр σ^2 известен.

3.9.44. Пусть проводится n независимых опытов, в каждом из которых может осуществиться одно из k событий $A_1,...,A_k$ с вероятностями $p_1,...,p_k$, причем сумма вероятностей равна единице. Обозначим через $\vec{p} = (p_1,...,p_k)$ вектор вероятностей событий $A_1,...,A_k$ и через $\vec{n} = (n_1,...,n_k)$ обозначим вектор частот реализаций событий $A_1,...,A_k$ в n независимых опытах, $(n_1 + ... + n_k) = n$. Отметим, что данная схема испытаний эквивалентна n независимым наблюдениям $X_1,...,X_n$ над дискретной случайной величиной X с возможными состояниями $\tilde{x}_1 = 1\,,...,\tilde{x}_k = k$ и их вероятностями $p_1,...,p_k$, то есть дискретная с.в. X задана распределением вероятностей в виде

$$\begin{pmatrix} X \\ \vec{p} \end{pmatrix} : \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ p_1,...,p_k \end{pmatrix}, \ \sum_{i=1}^{k} p_i = 1,$$

причем событие $(X = \tilde{x}_j = j)$ осуществляется, если осуществляется событие A_j, $j = 1,...,k$. При этом наблюденная реализация $\vec{n} = (n_1,...,n_k)$ вектора частот $\vec{N} = (N_1,...,N_k)$ определяется по наблюденной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$, то есть

$$(x_1,...,x_n) \Rightarrow \begin{pmatrix} 1,....,k \\ n_1,...,n_k \end{pmatrix}, \ \sum_{i=1}^{k} n_i = n,$$

Хорошо известно, что вектор частот $\vec{N} = (N_1,...,N_k)$ имеет полиномиальное распределение (П.1.49), то есть

$$P\{N_1 = n_1,...,N_k = n_k\,;\,\vec{p}\} = \frac{n!}{n_1!...n_k!}\, p_1^{n_1}...p_k^{n_k}.$$

Итак, требуется по выборке $X_1,...,X_n$ из распределения

$$F_X(x,\vec{p}) = \sum_{i=1}^{k} p_i C(x-i)$$

построить:

А) асимптотический критерий отношения правдоподобия для проверки простых гипотез H_0 и H_1, записываемых в виде

$$H_0 : p_1 = p_{10},...,p_k = p_{k0}, \;\; H_1 : p_1 = p_{11},...,p_k = p_{k1},$$

или кратко как

$$H_0 : \vec{p} = \vec{p}_0, \; H_1 : \vec{p} = \vec{p}_1, \text{ где } \vec{p}_0 = (p_{10},...,p_{k0}) \text{ и } \vec{p}_1 = (p_{11},...,p_{k1});$$

Б) асимптотический критерий отношения правдоподобия для проверки простой гипотезы H_0 против сложной альтернативы H_1^*, которые имеют вид

$$H_0 : \vec{p} = \vec{p}_0, \;\; H_1^* : \vec{p} \neq \vec{p}_0.$$

Решение. А) Для проверки простых гипотез $H_0 : \vec{p} = \vec{p}_0$, $H_1 : \vec{p} = \vec{p}_1$ к.о.п. основан на статистике Неймана – Пирсона (3.3.29), наблюденное значение которой, вычисленное по частотам $\vec{n} = (n_1,...,n_k)$, запишется как

$$l(\vec{n}\,;\vec{p}_0,\vec{p}_1) = \prod_{j=1}^{k} (p_{j1}/p_{j0})^{n_j}. \qquad \textbf{(3.9.45)}$$

Эта статистика «правосторонняя», и ее асимптотическое критическое значение, согласно (3.4.62), имеет вид

$$\exp\{-nI(\vec{p}_0,\vec{p}_1) + \sigma_0\sqrt{n}\cdot\lambda_{1-\alpha}\}, \; \lambda_{1-\alpha} = \Phi^{-1}(1-\alpha),$$

и, следовательно, асимптотическая критическая область $\mathrm{X}_{1\alpha}$ размера α для этой статистики, согласно (3.4.63), запишется в виде

$$\mathrm{X}_{1\alpha} = \{\vec{x} : l_n(\vec{n}\,;\vec{p}_0,\vec{p}_1) \geq \exp(-nI(\vec{p}_0,\vec{p}_1) + \sigma_0\sqrt{n}\cdot\lambda_{1-\alpha})\}, \qquad (3.9.46)$$

где $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha)$ – квантиль уровня $1-\alpha$ стандартного нормального распределения, $I(\vec{p}_0,\vec{p}_1)$ – количество информации по Кульбаку – Лейблеру, вычисляемое по формуле

$$I(\vec{p}_0,\vec{p}_1) = \sum_{j=1}^{k} p_{j0}\ln\{p_{j0}/p_{j1}\} \qquad \textbf{(3.9.47)}$$

и

$$\sigma_0^2 = \sum_{j=1}^{k} p_{j0}[\ln\{p_{j0}/p_{j1}\}]^2 - [I(\vec{p}_0,\vec{p}_1)]^2. \qquad \textbf{(3.9.48)}$$

Б) Для проверки простой гипотезы $H_0 : \vec{p} = \vec{p}_0$ против сложной альтернативы $H_1^* : \vec{p} \neq \vec{p}_0$ к.о.п. основан на статистике $\Lambda^*(\vec{X};\Theta_0) = 1/\ \Lambda(\vec{X};\Theta_0)$, определенной в (3.4.18). Для конкретизации этой статистики в данном примере напомним (см. разд. 2.10), что оценками максимального правдоподобия $\hat{p}_j$ параметров p_j, $j = 1,...,k$, являются относительные частоты появления событий $A_1,...,A_k$ в n независимых опытах, то есть $\hat{p}_j = n_j / n$, $j = 1,...,k$. С учетом этого замечания, наблюденное значение статистики отношения правдоподобия $\Lambda_n^*(\vec{N}\,;\vec{p}_0)$, вычисленное по частотам $\vec{n} = (n_1,...,n_k)$, запишется в виде

$$\Lambda_n^*(\vec{n}\,;\vec{p}_0) = \prod_{j=1}^{k} p_{j0}^{n_j} / \prod_{j=1}^{k} (n_j / n)^{n_j} = \prod_{j=1}^{k} (np_{j0} / n_j)^{n_j}\ .$$

Отсюда
$$-2\ln \Lambda_n^*(\vec{n}\,;\vec{p}_0) = 2\sum_{j=1}^{k} n_j \ln(n_j / np_{j0})\,. \qquad \textbf{(3.9.49)}$$

Отметим, что в данном примере роль неизвестного векторного параметра $\vec{\theta} = (\theta_1,...,\theta_r)$ играет вектор $\vec{p} = (p_1,...,p_k)$ размерности k, причем компоненты этого вектора связаны условием $p_1 + ... + p_k = 1$. Учитывая это, согласно теореме 3.4.72, заключаем, что асимптотическое распределение статистики $-2\ln \Lambda_n^*(\vec{N}\,;\vec{p}_0)$ при гипотезе $H_0 : \vec{p} = \vec{p}_0$ является распределением хи-квадрат с числом степеней свободы $k-1$, то есть выполняется выражение

$$L\{-2\ln \Lambda_n^*(\vec{N}\,;\vec{p}_0)\,|\,H_0\} = \chi^2(k-1)\,. \qquad \textbf{(3.9.50)}$$

Таким образом, асимптотическая критическая область к.о.п. размера α, основанного на статистике $-2\ln \Lambda_n^*(\vec{N}\,;\vec{p}_0)$, наблюденные значения которой вычисляются по формуле (3.9.49), имеет вид

$$\mathrm{X}_{1\alpha} = \{\vec{n}\ :\ 2\sum_{j=1}^{k} n_j \ln(n_j / np_{j0}) \geq \chi^2_{1-\alpha,\,(k-1)}\}\,. \qquad \textbf{(3.9.51)}$$

Заметим, что проверить гипотезу H_0 о том, что вероятности $p_1,...,p_k$ возможных состояний $\tilde{x}_1,...,\tilde{x}_k$ дискретной с.в. X равны заданным значениям $p_{10},...,p_{k0}$, то есть $H_0 : p_1 = p_{10}$, $p_2 = p_{20}$,..., $p_k = p_{k0}$ против альтернативы, что не все равенства выполняются, также можно с помощью χ^2-*критерия* (см. разд. 3.6). Для вычисления наблюденного значения статистики этого критерия сначала по конкретной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$ строится дискретный вариационный ряд частот, то есть

$$(x_1,...,x_n) \Rightarrow \begin{pmatrix} \tilde{x}_1,...,\tilde{x}_k \\ n_1,...,n_k \end{pmatrix},\ \sum_{i=1}^{k} n_i = n\,,$$

а затем по данным этого вариационного ряда частот вычисляется наблюденное значение статистики $\chi^2(\vec{N})$ по формуле

$$\chi^2(\vec{n}) = \sum_{i=1}^{k} [n_i - np_{i0}]^2 / np_{i0} = n \sum_{i=1}^{k} \left(\frac{n_i}{n} - p_{i0} \right)^2 / p_{i0} \,. \qquad \textbf{(3.9.52)}$$

Для этой статистики справедливо выражение $L(\chi^2(\vec{N}) \,|\, H_0) = \chi^2(k-1)$ при $n \to \infty$, которое обеспечивает нужную для практики точность уже при $n \geq 50$ и $\min\{n_1,...,n_k\} \geq 5$. Отметим, что χ^2-критерий и критерий отношения правдоподобия, основанный на статистике (3.9.49), асимптотически эквивалентны.

3.9.53. Состоятельность критерия отношения правдоподобия. Пусть $\vec{X} = (X_1,...,X_n)$ выборка из распределения $F_X(x;\theta)$ с плотностью $f_X(x;\theta)$, $x \in R^1$, $\theta \in R^1$. Показать, что для проверки гипотез $H_0 : \theta = \theta_0$ и $H_1 : \theta \neq \theta_0$ критерий отношения правдоподобия, основанный на статистике $\Lambda_n^*(\vec{X};\Theta_0)$ из (3.4.18), является состоятельным критерием.

Решение. Выберем произвольное фиксированное значение θ_1 параметра θ при альтернативе, т.е. $\theta_1 \neq \theta_0$, и для заданной реализации $\vec{x} = (x_1,...,x_n)$ выборки $X_1,...,X_n$, согласно (3.4.18), запишем наблюденное значение статистики $\Lambda_n^*(\vec{X};\Theta_0)$ в виде

$$\Lambda_n^*(\vec{x};\Theta_0) = \frac{L(\vec{x};\theta_0)}{L(\vec{x};\hat{\theta}_n)} = \frac{L(\vec{x};\theta_1)}{L(\vec{x};\hat{\theta}_n)} \cdot \frac{L(\vec{x};\theta_0)}{L(\vec{x};\theta_1)} .$$

Отсюда

$$-2\ln\Lambda_n^*(\vec{x};\Theta_0) = -2\ln\frac{L(\vec{x};\theta_1)}{L(\vec{x};\hat{\theta}_n)} + n(2/n)\left[\ln L(\vec{x};\theta_1) - \ln L(\vec{x};\theta_0)\right].$$

Таким образом, для статистики $\Lambda_n^*(\vec{X};\Theta_0)$ имеем

$$-2\ln\Lambda_n^*(\vec{X};\Theta_0) = -2\ln\frac{L(\vec{X};\theta_1)}{L(\vec{X};\hat{\theta}_n)} + 2n\left[\frac{1}{n}\sum_{i=1}^{n}\ln f_X(X_i;\theta_1) - \frac{1}{n}\sum_{i=1}^{n}\ln f_X(X_i;\theta_0)\right].$$

Далее, первое слагаемое в правой части предыдущего выражения по теореме (3.4.72) имеет распределение хи-квадрат с одной степенью свободы, а случайная величина в квадратных скобках второго слагаемого на основании закона больших чисел сходится по вероятности к положительной величине $[H(\theta_1) - H(\theta_0)]$, где $H(\theta) = M_{\theta_1}[\ln f_X(X_1;\theta)]$ – функция, достигающая максимальное значение при $\theta = \theta_1$, так как $H'(\theta_1) = 0$ и $H''(\theta_1) = -\,i(\theta_1) < 0$.

Итак, статистика $-2\ln\Lambda_n^*(\vec{X};\Theta_0)$ по вероятности неограниченно возрастает при $n\to\infty$. Отсюда заключаем, что при $n\to\infty$ мощность $W_n(\theta_1)$ критерия отношения правдоподобия, основанного на статистике $\Lambda_n^*(\vec{X};\Theta_0)$, стремиться к единице, то есть

$$W_n(\theta_1) = P_{\theta_1}\{-2\ln\Lambda_n^*(\vec{X};\Theta_0)\ge\chi^2_{1-\alpha,(1)}\}\to 1 \text{ при } n\to\infty. \qquad \textbf{(3.9.54)}$$

Таким образом, если ф.р. $F_X(x;\theta)$ регулярна в смысле первой и второй производной по θ, то критерий отношения правдоподобия, заданный критической областью $\mathrm{X}_{1\alpha}=\{\vec{x}:-2\ln\Lambda_n^*(\vec{x};\Theta_0)\ge\chi^2_{1-\alpha,(1)}\}$, является состоятельным критерием проверки гипотез $H_0:\theta=\theta_0$, $H_1:\theta\ne\theta_0$ (см. [66, с. 419]). Отметим, что при увеличении объема выборки функции мощности для всех состоятельных критериев стремятся к единице. По этой причине при асимптотическом сравнении состоятельных критериев рассматривают сходящиеся к нулевой гипотезе альтернативы (их называют контигуальные альтернативы), чтобы обеспечить значение предела функции мощности отличное от единицы, что и позволяет сравнивать состоятельные критерии между собой в асимптотике. В частности, контигуальные альтернативы используются для сравниваемых критериев при вычислении асимптотической относительной эффективности Питмена (см. П.3.Е).

3.9.55. Пусть $L(X)=N(\mu,\sigma^2)$. Имеется выборка $X_1,...,X_n$ из нормального распределения $F_X(x;\mu)=\Phi\{(x-\mu)/\sigma\}$, параметры μ и σ^2 неизвестны. Получите формулу для вычисления мощности одностороннего критерия проверки гипотез $H_0:\sigma^2=\sigma_0^2$, $H_1^+:\sigma^2>\sigma_0^2$, основанного на критической области

$$\mathrm{X}^+_{1\alpha}=\{\vec{x}:(n-1)\tilde{S}^2/\sigma_0^2\ge\chi^2_{1-\alpha,(n-1)}\}.$$

Получите также формулу для вычисления необходимого объема выборки n для обеспечения заданной мощности.

Решение. Для заданного альтернативного значения σ_1^2 параметра σ^2 мощность критерия с критической областью $\mathrm{X}^+_{1\alpha}$ запишется в виде

$$W(\mathrm{X}^+_{1\alpha};\sigma_1^2)=P\{(n-1)\tilde{S}^2/\sigma_0^2\ge\chi^2_{1-\alpha,(n-1)}\mid H_1^+\}=$$

$$=P\left\{\frac{\sigma_1^2}{\sigma_0^2}\frac{(n-1)\tilde{S}^2}{\sigma_1^2}\ge\chi^2_{1-\alpha,(n-1)}\mid H_1^+\right\}=P\left\{\frac{(n-1)\tilde{S}^2}{\sigma_1^2}\ge\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha,(n-1)}\mid H_1^+\right\}=$$

$$=P\left\{\chi^2(n-1)\ge\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha,(n-1)}\right\}. \qquad \textbf{(3.9.56)}$$

Например, пусть $n=20$, $\sigma_1^2/\sigma_0^2=3/2$, $\alpha=0{,}05$, $\chi^2_{0{,}95\,,\,(19)}=30{,}1$. При этих значениях мощность критерия

$$W(\mathrm{X}^+_{1\alpha};\sigma_1^2)=P\left\{\chi^2(19)\ge\frac{2}{3}\cdot 30{,}1\right\}=P\left\{\chi^2(19)\ge 20{,}0\right\}\approx 0{,}39.$$

Полученное выражение для мощности критерия позволяет вычислить объем выборки n для обеспечения заданной мощности $W=1-\beta$ из уравнения

$$P\left\{\chi^2(n-1)\ge\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha\,,\,(n-1)}\right\}=1-\beta,$$

которое может быть переписано в виде

$$\chi^2_{\beta\,,\,(n-1)}=\frac{\sigma_0^2}{\sigma_1^2}\chi^2_{1-\alpha\,,\,(n-1)},\ \text{или}\ \chi^2_{1-\alpha\,,\,(n-1)}/\chi^2_{\beta\,,\,(n-1)}=\sigma_1^2/\sigma_0^2. \qquad \textbf{(3.9.57)}$$

Используя таблицы хи-квадрат распределения, при заданных вероятностях ошибок α,β и при заданном отношении σ_1^2/σ_0^2, последнее уравнение может быть разрешено относительно n. Например, если $\sigma_1^2/\sigma_0^2=1{,}5$, $\alpha=0{,}05$ и $W=1-\beta=0{,}90$, то выбрав $n=101$, из таблиц хи-квадрат распределения находим $\chi^2_{0{,}95\,,\,(100)}=124{,}3$ и $\chi^2_{0{,}10\,,\,(100)}=82{,}4$. Отсюда $\chi^2_{0{,}95\,,\,(100)}/\chi^2_{0{,}10\,,\,(100)}=1{,}504$, что практически совпадает с заданным отношением $\sigma_1^2/\sigma_0^2=1{,}5$. Таким образом, при $n>100$ критерий будет иметь мощность больше 0,90. Отметим, что при больших объемах выборки можно воспользоваться нормальной аппроксимацией, предложенной Фишером, согласно которой (см., например, [9, с. 75]), квантиль уровня α распределения хи-квадрат с числом степеней свободы $\nu=n-1$ выражается через квантиль $\lambda_\alpha=\Phi^{-1}(\alpha)$ уровня α стандартного нормального распределения с помощью приближенной формулы

$$\chi^2_{\alpha\,,\,(\nu)}\approx(\sqrt{2\nu-1}+\lambda_\alpha)^2/2.$$

Используя эту аппроксимацию, из (3.9.57) получим приближенное равенство

$$(\sqrt{2\nu-1})+\lambda_{1-\alpha})^2/(\sqrt{2\nu-1}+\lambda_\beta)^2\approx\sigma_1^2/\sigma_0^2,$$

разрешая которое относительно числа степеней свободы $\nu=n-1$, получаем приближенную формулу для объема выборки n в виде

$$n\approx\frac{3}{2}+\frac{1}{2}\left(\frac{\lambda_{1-\alpha}+\lambda_{1-\beta}\sqrt{\sigma_1^2/\sigma_0^2}}{\sqrt{\sigma_1^2/\sigma_0^2}-1}\right)^2. \qquad \textbf{(3.9.58)}$$

Например, из этой формулы при $\sigma_1^2/\sigma_0^2 = 1{,}5$, $\alpha = 0{,}05$ и $W = 1-\beta = 0{,}90$, с учетом того, что $\lambda_{1-\alpha} = \lambda_{0{,}95} = 1{,}645$ и $\lambda_{1-\beta} = \lambda_{0{,}90} = 1{,}282$, получаем $n \approx 104$.

3.9.59. Пусть $L(X) = N(a_1;\sigma_1^2)$ и $L(Y) = N(a_2;\sigma_2^2)$. Имеются две выборки из этих распределений $X_1,...,X_{n_1}$ и $Y_1,...,Y_{n_2}$. Параметры a_j и σ_j^2, $j = 1,2$, неизвестны. Пусть $\tilde{S}_1^2$ и $\tilde{S}_2^2$ – независимые и несмещенные оценки для параметров σ_j^2, $j = 1,2$. Покажите, что статистика отношения правдоподобия $\Lambda^*(\vec{X};\Theta_0)$, определенная в (3.4.18), в данном случае проверки гипотез $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$ записывается в виде

$$\Lambda^*(\vec{X};\Theta_0) = \frac{(n_1+n_2)^{(n_1+n_2)/2}}{n_1^{n_1/2} n_2^{n_2/2}} \left[1+\frac{\nu_2\tilde{S}_2^2}{\nu_1\tilde{S}_1^2}\right]^{-n_1/2} \left[1+\frac{\nu_1\tilde{S}_1^2}{\nu_2\tilde{S}_2^2}\right]^{-n_2/2},$$

где число степеней свободы $\nu_j = n_j - 1$, $j = 1,2$. Отметим, что из полученного выражения следует, что малые значения статистики $\Lambda^*(\vec{X};\Theta_0)$ достигаются в двух случаях, если отношения $\tilde{S}_1^2/\tilde{S}_2^2$ будут малы или велики.

3.9.60. Убедитесь, что нижние и верхние квантили *F*-распределения Фишера связаны равенством

$$F_q(\nu_1,\nu_2) = 1/F_{1-q}(\nu_2,\nu_1),\ 0 < q < 1.$$

Решение. Используя определение квантиля заданного уровня $1-q$, получаем

$$P\{\ F(\nu_1,\nu_2) \le F_{1-q}(\nu_1,\nu_2)\ \} = P\{\ (\tilde{S}_1^2/\tilde{S}_2^2) \le F_{1-q}(\nu_1,\nu_2)\} = 1-q.$$

Отсюда следует выражение $P\{\ (\tilde{S}_2^2/\tilde{S}_1^2) > (1/F_{1-q}(\nu_1,\nu_2))\} = 1-q$, которое может быть записано в виде $P\{\ (\tilde{S}_2^2/\tilde{S}_1^2) \le (1/F_{1-q}(\nu_1,\nu_2))\} = q$. Далее, учитывая, что $P\{\ (\tilde{S}_2^2/\tilde{S}_1^2) \le F_q(\nu_2,\nu_1)\} = q$, и сравнивая два последних выражения, получаем $F_q(\nu_1,\nu_2) = 1/F_{1-q}(\nu_2,\nu_1)$. Таким образом, зная верхние квантили *F*-распределения Фишера, легко вычислить нижние. Например, нужно определить $F_{0{,}01}(18,15)$. Для этого по таблице (6) в [44, с. 138] находим $F_{0{,}99}(15,18) = 3{,}23$ и, следовательно,

$$F_{0{,}01}(18,15) = 1/F_{0{,}99}(15,18) = 1/3{,}23 = 0{,}31.$$

3.9.61. Пусть $L(X) = N(a_1;\sigma_1^2)$ и $L(Y) = N(a_2;\sigma_2^2)$. Имеются две независимые выборки из этих распределений $X_1,...,X_{n_1}$ и $Y_1,...,Y_{n_2}$. Параметры a_j и

$\sigma_j^2,\ j=1,2,$ неизвестны. Пусть $\tilde{S}_1^2$ и $\tilde{S}_2^2$ – независимые и несмещенные оценки для параметров $\sigma_j^2,\ j=1,2$. Критерий уровня значимости α для проверки гипотез $H_0:\sigma_1^2=\sigma_2^2,\ H_1^+:\sigma_1^2>\sigma_2^2$ определяется, согласно (3.6.10), критической областью

$$X_{1\alpha}^+=\{(\vec{x},\vec{y}):(\tilde{s}_1^2/\tilde{s}_2^2)\geq F_{1-\alpha}(\nu_1,\nu_2)\},$$

и его мощность при альтернативе $H_1^+:\sigma_1^2>\sigma_2^2$ для заданного значения отношения $\sigma_1^2/\sigma_2^2=\tau$, согласно (3.6.17), вычисляется по формуле

$$W(X_{1\alpha}^+;\tau)=P\left\{F(\nu_1,\nu_2)\geq\tau^{-1}F_{1-\alpha}(\nu_1,\nu_2)\right\}.$$

Требуется вычислить мощность критерия для следующих значений параметров:

а) $\alpha=0{,}05$, $\tau=5$, а $\nu_1=30$ и $\nu_2=10$;

б) $\alpha=0{,}05$, $\tau=1{,}5$, а $\nu_1=30$ и $\nu_2=10$.

Определите число степеней свободы ν_1 и ν_2 (или объемы выборок n_1 и n_2), при которых критерий будет иметь заданную мощность $1-\beta$ при заданном значении отношения $\sigma_1^2/\sigma_2^2=\tau$.

Решение. Для заданных параметров $\alpha=0{,}05$, $\nu_1=30$ и $\nu_2=10$ имеем $F_{0{,}95}(30{,}10)=2{,}70$ и мощность критерия при $\tau=5$

$$W(X_{1\alpha}^+;5)=P\{F(30\,,10)\geq(2{,}70/5)\}=P\{F(10\,,30)<(5/2{,}70)\}=$$
$$=P\{F(10\,,30)<1{,}85\}\approx 0{,}90.$$

Для значения $\tau=1{,}5$ мощность этого же критерия

$$W(X_{1\alpha}^+;1{,}5)=P\{F(30\,,10)\geq(2{,}70/1{,}5)\}=$$
$$=1-P\{F(30\,,10)\leq 2{,}16\}\approx 1-0{,}90=0{,}10.$$

Чтобы определить необходимые числа ν_1 и ν_2, при которых критерий будет иметь заданную мощность $1-\beta$, необходимо приравнять правую часть (3.6.17) заданной мощности $1-\beta$, то есть

$$P\left\{F(f_1,f_2)\geq\tau^{-1}F_{1-\alpha}(\nu_1,\nu_2)\right\}=1-\beta.$$

Учитывая далее, что $P\left\{F(f_1,f_2)\geq F_{\beta}(\nu_1,\nu_2)\right\}=1-\beta$, получаем

$$F_{\beta}(\nu_1,\nu_2)=\tau^{-1}F_{1-\alpha}(\nu_1,\nu_2).$$

Отсюда следует, что

$$\tau=F_{1-\alpha}(\nu_1,\nu_2)/F_{\beta}(\nu_1,\nu_2)=F_{1-\alpha}(\nu_1,\nu_2)\cdot F_{1-\beta}(\nu_2,\nu_1).$$

Например, если $\alpha = 0{,}05$, $1-\beta = 0{,}90$, то $F_{0{,}95}(80,80) = 1{,}45$, $F_{0{,}90}(80,80) = 1{,}33$ и произведение $F_{1-\alpha}(\nu_1,\nu_2)\cdot F_{1-\beta}(\nu_2,\nu_1) = 1{,}45\cdot 1{,}33 = 1{,}93$. Таким образом, при $\tau = 2$, для обеспечения критерием мощности $W(X^+_{1\alpha};2) = 1-\beta = 0{,}90$ необходимо взять $\nu_1 = \nu_2 \approx 80$.

3.9.62. Убедитесь, что квантиль уровня p, $0 < p < 1$, хи-квадрат-распределения с одной степенью свободы равен квадрату квантиля уровня $(1+p)/2$ стандартного нормального распределения.

Решение. Пусть $L(X) = N(0;1)$. Рассмотрим последовательность случайных величин $X_1,...,X_n$, которые независимы и имеют стандартное нормальное распределение, то есть $L(X_i) = N(0;1)$, $i = 1,...,n$. Рассмотрим с.в. $\chi^2(n) = (X_1^2 + X_2^2 + \cdots + X_n^2)$, которая имеет хи-квадрат-распределение с n степенями свободы, то есть $L\{\chi^2(n)\} = \chi^2_n$. Если $n = 1$, то $\chi^2(1) = X^2$. Пусть $\lambda_\alpha = \Phi^{-1}(\alpha)$ – квантиль уровня α стандартного нормального распределения. Отметим, что $\lambda_\alpha = -\lambda_{1-\alpha}$, $0 < \alpha < 1$. Согласно определению квантиля уровня α, имеем $P\{\lambda_\alpha \le X \le \lambda_{1-\alpha}\} = 1-2\alpha$ или, учитывая, что $\lambda_\alpha = -\lambda_{1-\alpha}$, получаем $P\{-\lambda_{1-\alpha} \le X \le \lambda_{1-\alpha}\} = 1-2\alpha$, откуда $P\{X^2 \le \lambda^{\,2}_{1-\alpha}\} = 1-2\alpha$. Положим $1-2\alpha = p$, тогда $1-\alpha = (1+p)/2$. С учетом этого обозначения последнее равенство запишется в виде $P\{X^2 \le \lambda^2_{(1+p)/2}\} = p$. Далее, пусть $\chi^2_{p,n}$ обозначает квантиль уровня p, $0 < p < 1$, хи-квадрат-распределения с n степенями свободы, тогда

$$P\{\chi^2(1) \le \chi^2_{p,1}\} = p\,.$$

Сравнивая два последних выражения, с учетом равенства $\chi^2(1) = X^2$, получаем $\chi^2_{p,1} = \lambda^2_{(1+p)/2}$, $0 < p < 1$.

3.9.63. Пусть с.в. X имеет стандартное нормальное распределение, то есть $L(X) = N(0;1)$, ее квантиль уровня p, $0 < p < 1$, обозначим через $\lambda_p = \Phi^{-1}(p)$. Пусть, далее, $X_1,...,X_n$ – независимые стандартные нормальные случайные величины, то есть $L(X_i) = N(0;1)$, $i = 1,...,n$. Рассмотрим следующие случайные величины:

$$\chi^2(n) = (X_1^2 + \cdots + X_n^2)/n; \qquad \textbf{(3.9.64)}$$

$$t(f) = X/\sqrt{\chi^2(f)/f}\,; \qquad \textbf{(3.9.65)}$$

$$F(f_1,f_2) = \frac{\chi^2(f_1)/f_1}{\chi^2(f_2)/f_2}\,. \qquad \textbf{(3.9.66)}$$

Случайная величина $\chi^2(n)$ имеет распределение хи-квадрат с числом степеней свободы n, то есть $L\{\chi^2(n)\} = \chi^2_n$, квантиль уровня p, $0 < p < 1$, с.в. $\chi^2(n)$ обозначим через $\chi^2_{p,n}$; случайная величина $t(f)$ имеет *t*-распределение Стьюдента с числом степеней свободы f, то есть $L\{t(f)\} = St(f)$, квантиль уровня p, $0 < p < 1$, с.в. $t(f)$ обозначим через $t_{p,f}$; случайная величина $F(f_1, f_2)$ имеет *F*-распределение Фишера с числами степеней свободы f_1 и f_2, ее квантиль уровня p, $0 < p < 1$, будем обозначать через $F_p(f_1, f_2)$. Сведения об этих распределениях вероятностей приведены в П.2.94. Убедитесь, что между этими случайными величинами выполняются следующие связи:

а) $F(f_1, \infty) = \chi^2(f_1)/f_1$;

б) при $f \to \infty$ отношение $\chi^2(f)/f \to \infty$ и, следовательно, *t*-распределение Стьюдента с числом степеней свободы f сходится к стандартному нормальному распределению;

в) $F(1, f) = t^2(f)$;

г) $F_p(1, f) = t^2_{(1+p)/2}(f)$, $0 < p < 1$;

Решение. Так как $M[\chi^2(f)] = f$ и $D[\chi^2(f)] = 2f$, то $M[\chi^2(f)/f] = 1$ и $D[\chi^2(f)/f] = 2/f$. Поэтому, когда $f \to \infty$, то отношение $\chi^2(f)/f$ близко к единице с большой вероятностью. Далее, если в формуле (3.9.66) положить $f_2 = \infty$, то отношение $\chi^2(f_2)/f_2$ становится равным единице, и в результате получаем

$$F(f_1, \infty) = \chi^2(f_1)/f_1.$$

Далее, при $f \to \infty$ отношение $\chi^2(f)/f \to 1$ и, следовательно, *t*-распределение Стьюдента с числом степеней свободы f сходится к стандартному нормальному распределению при $f \to \infty$.

Связь между *t*-распределением Стьюдента и *F*-распределением Фишера устанавливается следующим образом. Так как $L(X) = N(0;1)$, то $X^2 = \chi^2(1)$, и при $f_1 = 1$ из (3.9.66) имеем

$$F(1, f_2) = \frac{\chi^2(1)/1}{\chi^2(f_2)/f_2} = \frac{X^2}{\chi^2(f_2)/f_2}.$$

Сравнивая полученное выражение для $F(1, f_2)$ с $t(f) = X/\sqrt{\chi^2(f)/f}$, получаем равенство $F(1, f) = t^2(f)$. Учитывая, что *t*-распределение Стьюдента симметрично относительно нуля, можем убедиться (как и в упр. 3.9.62), что

между квантилями t-распределения Стьюдента и F-распределения Фишера существует связь, аналогичная связи между $\chi^2_p(1)$ и $\lambda_{(1+p)/2}$, а именно: $F_p(1,f) = t^2_{(1+p)/2}(f)$, $0 < p < 1$.

3.9.67. Используя χ^2-критерий согласия, проверьте гипотезу о симметричности подбрасываемой монеты:

А) по данным опытов Бюффона, в которых монета подбрасывалась $n = 4040$ раз и «герб» выпал 2048 раз, гипотезу проверьте на уровне значимости $\alpha = 0{,}05$;

Б) по данным опытов К. Пирсона, в которых монета подбрасывалась $n = 24000$ раз и «герб» выпал 12012 раз, гипотезу проверьте путем вычисления достигнутого уровня значимости α_∂ (см. определение 3.1.32).

В) Число выпадений герба при $n = 20$ подбрасываниях двух монет распределилось следующим образом: «0» ~ 4 раза; «1» ~ 8 раз; «2» ~ 8 раз. Согласуются ли эти результаты опыта с гипотезой о симметричности подбрасываемых монет. Гипотезу проверьте на уровне значимости $\alpha = 0{,}05$.

Г) Используя хи-квадрат-критерий, проверьте гипотезу о симметричности подбрасываемого шестигранного кубика, то есть $H_0 : p_1 = p_2 = ... = p_6 = 1/6$, против альтернативы H_1: не все равенства гипотезы выполняются, по следующим результатам опыта: «шестерка» выпала 9 раз; «пятерка» выпала 12 раз; «четверка» выпала 9 раз; «тройка» выпала 5 раз; «двойка» выпала 6 раз; «единица» выпала 9 раз.

Д) При $n = 120$ подбрасываниях шестигранного кубика «шестерка» выпала 15 раз; «пятерка» выпала 25 раз. Проверьте гипотезу о симметричности подбрасываемого шестигранного кубика на уровне значимости $\alpha = 0{,}01$.

Решение. Гипотеза о симметрии монеты сводится к проверке утверждения о том, вероятность выпадения «герба» $p_1 = p_{1,0}$ равняется вероятности выпадения «решки» $p_2 = p_{2,0}$ и обе они равны 0,5 , то есть гипотеза H_0 из (3.5.40) в данном случае имеет вид

$$H_0 : p_1 = 0{,}5 \ , \ p_2 = 0{,}5 .$$

Отметим, что в данном примере $k = 2$ и статистика χ^2 из (3.5.42) запишется в виде

$$\chi^2 = (n_1 - np_1)^2 / np_1 \ + (n_2 - np_2)^2 / np_2 \ , \qquad \textbf{(3.9.68)}$$

где n_1 – число выпадения «герба», а $n_2 = n - n_1$ – число выпадения «решки» в n опытах.

А) Рассмотрим данные Бюффона: $n = 4040$, $n_1 = 2048$ и $n_2 = 1992$. Вычисленное значение $\chi^2_в$ статистики χ^2 из (3.9.68) равняется

$\chi^2_{\text{в}} = (2048-2020)^2/2020 + (1992-2020)^2/2020 = 0,776$. Задав уровень значимости $\alpha = 0,05$, по таблицам квантилей распределения хи-квадрат с числом степеней свободы $(k-1) = (2-1) = 1$, находим квантиль уровня $(1-\alpha)$ этого распределения. При $\alpha = 0,05$ он равен $\chi_{1-\alpha\,;\,(k-1)} = \chi^2_{0,95\,;\,1} = 3,841$. Далее, поскольку выполняется неравенство $(\chi^2_{\text{в}} = 0,776) < (\chi^2_{0,95\,;\,1} = 3,841)$, то согласно правилу (3.5.43), заключаем, что опытные данные Бюффона не противоречат гипотезе о симметрии подбрасываемой монеты на уровне значимости $\alpha = 0,05$.

Б) Рассмотрим теперь данные К. Пирсона: $n = 24000$, $n_1 = 12012$ и $n_2 = 11988$. Наблюденное значение статистики χ^2, вычисленное по этим данным, равно

$$\chi^2_{\text{в}} = (12012-12000)^2/12000 + (11988-12000)^2/12000 = 0,024.$$

Пусть, далее, $\chi^2(1)$ обозначает случайную величину, имеющую распределение хи-квадрат с числом степеней свободы равным единице. Тогда, согласно (3.1.32), достигнутый уровень значимости α_{∂} равен $\alpha_{\partial} = P\{\chi^2(1) \geq 0,024\}$. Используя таблицу (2.1а) интеграла вероятности χ^2 в [6, с. 140], получаем $\alpha_{\partial} = 0,87$, и это говорит о том, что результаты опытов К. Пирсона находятся в очень хорошем согласии с гипотезой о симметричности подбрасываемой монеты.

В) При подбрасывании двух монет в одном опыте имеем четыре возможности: (РР), (ГР), (РГ), (ГГ). Следовательно, при симметричности монет имеем $p_1 = P\{\#(\Gamma) = 0\} = 1/4$, $p_2 = P\{\#(\Gamma) = 1\} = 1/2$, $p_3 = P\{\#(\Gamma) = 2\} = 1/4$. Далее, число опытов $n = 20$, $k = 3$, $n_1 = 4$, $n_2 = 8$, $n_3 = 8$. Наблюденное значение статистики хи-квадрат равно

$$\chi^2_{\text{в}} = \{(4-20/4)^2/5\} + \{(8-20/2)^2/10\} + \{(8-20/4)^2/5\} = 2,4.$$

При уровне значимости $\alpha = 0,05$ критическое значение $\chi^2_{1-\alpha,(k-1)} = \chi^2_{0,95,(2)} = 5,99$. Так как выполняется неравенство $(\chi^2_{\text{в}} = 2,4) < (\chi^2_{0,95\,,\,(2)} = 5,99)$, то гипотезу о симметрии монет принимаем на уровне значимости $\alpha = 0,05$.

Г) В данном случае $k = 6$ и наблюденное (вычисленное) значение статистики хи-квадрат

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \frac{(9-50/6)^2}{50/6} + \frac{(12-50/6)^2}{50/6} + \ldots + \frac{(9-50/6)^2}{50/6} = 3,76.$$

Используя таблицы интеграла вероятности χ^2, находим $\Pr\{\chi^2(5)\geq 3{,}76\}\approx 0{,}58$. Таким образом, *P-значение* критерия хи-квадрат равно

$$P\text{-значение} = \Pr\{\chi^2(5) \geq 3{,}76\} \approx 0{,}58,$$

то есть гипотезу H_0 принимаем, результат незначим и нет никакого подтверждения в несимметричности кубика.

Д) Пусть A_1 – событие, состоящее в появлении шести очков на верхней грани кубика, событие A_2 – в появлении пяти очков, событие A_3 – в появлении либо 1, либо 2, либо 3, либо 4. Далее, $p_1 = P(A_1) = 1/6$ и $n_1 = 15$, $p_2 = P(A_2) = 1/6$ и $n_2 = 25$, $p_3 = P(A_3) = 4/6$ и $n_3 = 120 - (15 + 20) = 80$. При $k = 3$ и этих данных наблюденное значение статистики равно $\chi^2_{\text{в}} = \{(15-120/6)^2/20\} + \{(25-120/6)^2/20\} + \{(80-80)^2/80\} = 2{,}5$. Для $\alpha = 0{,}01$ критическое значение $\chi_{1-\alpha,(k-1)} = \chi^2_{0{,}99,(2)} = 9{,}21$. Так как выполняется неравенство $(\chi^2_{\text{в}} = 2{,}5) < (\chi^2_{0{,}99,(2)} = 9{,}21)$, то гипотезу о симметрии кубика принимаем на уровне значимости $\alpha = 0{,}01$. Используя таблицы интеграла вероятности χ^2, находим $\Pr\{\chi^2(2) \geq 2{,}5\} \approx 0{,}28$. Таким образом, достигнутый уровень значимости или *P-значение* критерия хи-квадрат

$$P\text{-значение} = \Pr\{\chi^2(2) \geq 2{,}5\} \approx 0{,}28.$$

3.9.69. Пусть $L(X) = \Pi y(\theta)$, $H_1 : \sigma^2 \neq \sigma_0^2$. Эта краткая запись означает, что дискретная случайная величина X имеет распределение Пуассона с параметром θ, то есть ее возможными значениями являются неотрицательные целые числа $0, 1, 2, \ldots$ с вероятностями

$$P(X = j) = \frac{\theta^j}{j!} e^{-\theta}, \; j = 0, 1, 2, \ldots, \; 0 < \theta < \infty.$$

Имеется выборка $X_1, \ldots, X_n$, получаемая в n независимых наблюдениях над неотрицательной целочисленной случайной величиной X и $x_1, \ldots, x_n$ – ее наблюденная реализация. Требуется построить хи-квадрат-критерий согласия для проверки сложной гипотезы

$$H_0^* : F_X(x;\theta) = \sum_{j=1}^{\infty} \frac{\theta^j e^{-\theta}}{j!} C(x - j), \text{ для всех } x \in R^{1+}, \; 0 < \theta < \infty,$$

где θ – неизвестный параметр.

Решение. Рассмотрим события $A_1, \ldots, A_k$, где их число k определяется из условия, что выполнилось неравенство $\min(n_1, \ldots, n_k) \geq 5$. Здесь $n_1, \ldots, n_k$ – час-

тоты реализаций событий $A_1,...,A_k$ в n опытах, определенные по наблюденной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$, а сами события $A_1,...,A_k$ определены следующим образом:

$$A_j = \{j-1\},\ j=1,...,(k-1),\ A_k = \{k-1, k, ...\}.$$

Вероятности событий $A_1,...,A_k$ зависят от неизвестного параметра θ и записываются в виде

$$p_j(\theta) = e^{-\theta}\theta^{j-1}/(j-1)!,\ j=1,...,(k-1),\ p_k(\theta) = \sum_{j=k-1}^{\infty} e^{-\theta}\theta^{j-1}/(j-1)!.$$

Для построения модифицированной статистики хи-квадрат требуется построить по информации, содержащейся в выборке $X_1,...,X_n$, оценку $St(n-1)$ параметра θ. Согласно теории (см. теорему Фишера 3.5.19), для этой цели нужно вычислить по частотам $n_1,...,n_k$ мультиномиальную оценку максимального правдоподобия параметра θ, решив систему уравнений вида (3.5.18). В данном примере эта система содержит лишь одно уравнение и его решением при заданной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$ является наблюденное значение $\overline{x}$ выборочного среднего $\overline{X}$. Напомним, что обычной оценкой максимального правдоподобия для параметра θ распределения Пуассона, вычисляемой не по частотам $n_1,...,n_k$, а по исходной выборке $X_1,...,X_n$, является выборочное среднее $\overline{X}$.

Таким образом, в данном примере мультиномиальная оценка максимального правдоподобия параметра θ, вычисляемая по частотам $n_1,...,n_k$, и обычная оценка максимального правдоподобия параметра θ, вычисляемая по исходной выборке $X_1,...,X_n$, совпадают. С учетом этого факта запишем модифицированную статистику хи-квадрат в виде (3.5.16), то есть

$$\chi^2(\hat{\theta}) = \sum_{i=1}^{k}[(n_i - np_i(\hat{\theta}))^2 / np_i(\hat{\theta})],$$

где

$$p_j(\hat{\theta}) = e^{-\hat{\theta}}\hat{\theta}^{j-1}/(j-1)!,\ j=1,...,(k-1),\ p_k(\hat{\theta}) = \sum_{j=k-1}^{\infty} e^{-\hat{\theta}}\hat{\theta}^{j-1}/(j-1)!,\ \hat{\theta} = \overline{x}.$$

Окончательно, при больших объемах выборки, точнее при $n \geq 50$, гипотеза H_0^* отвергается на уровне значимости α, если $T(\vec{X}) = \sqrt{n}(\overline{X} - \mu_0)/\tilde{S}$.

Приведем числовой пример. Отдел технического контроля проверил $n = 200$ партий одинаковых изделий и получил следующее эмпирическое распределение (в первой строке указано количество x_i нестандартных изде-

лий в одной партии; во второй строке указаны частоты n_i, то есть количество партий, содержащих x_i нестандартных изделий):

x_i	0	1	2	3	4
n_i	116	56	22	4	2

Требуется проверить на уровне значимости $\alpha = 0,05$ гипотезу о том, что случайная величина X, характеризующая число нестандартных изделий в партии, описывается распределением Пуассона, то есть $H_0 : L(X) = \Pi y(\theta)$, θ – неизвестный параметр, $H_1 : \sigma^2 \neq \sigma_0^2$. Отметим, что данная параметрическая гипотеза сложная. Оценка максимального правдоподобия неизвестного параметра θ равна

$$\hat{\theta} = n^{-1} \sum n_i x_i = (116 \cdot 0 + 56 \cdot 1 + 22 \cdot 2 + 4 \cdot 3 + 2 \cdot 4) / 200 = 0,6.$$

Вычислим гипотетические частоты $n_i^* = np_i(\hat{\theta}) = n\hat{\theta}^i e^{-\hat{\theta}} / i! = 200 \cdot (0,6)^i e^{-0,6} / i!$, $i = 0,1,2,3$. В данном случае $k = 4$, так как для выполнения неравенства $\min(n_1,...,n_k) \geq 5$ частоты 4 и 2 двух последних состояний объединены. С учетом этого замечания получаем следующие значения гипотетических частот: $n_0^* = 109,76$, $n_1^* = 65,86$, $n_2^* = 19,76$, $n_3^* = 4,56$. Вычисленное значение модифицированной статистики хи-квадрат

$$\chi^2(\hat{\theta}) = \sum_{i=0}^{k-1} [(n_i - np_i(\hat{\theta}))^2 / np_i(\hat{\theta})] = \{(n_0 - n_0^*)^2 / n_0^*\} + ... + \{(n_3 - n_3^*)^2 / n_3^*\} =$$

$$= \{(116 - 109,76)^2 / 109,76\} + \{(56 - 65,86)^2 / 65,86\} + \{(22 - 19,76)^2 / 19,76\} +$$

$$+ \{(6 - 4,56)^2 / 4,56\} = 2,54.$$

С учетом того, что оценивался один параметр θ, число степеней свободы будет равно $k - 1 - 1 = 4 - 2 = 2$. При уровне значимости $\alpha = 0,05$ критическое значение $\chi^2_{1-\alpha,(k-1-1)} = \chi^2_{0,95,(2)} = 5,99$. Так как выполняется неравенство $(\chi^2_{\text{в}} = 2,54) < (\chi^2_{0,95,(2)} = 5,99)$, то нет оснований возражать против высказанной гипотезы $H_0 : L(X) = \Pi y(\theta)$ на уровне значимости $\alpha = 0,05$. Достигнутый уровень значимости или *P-значение* критерия хи-квадрат в данном случае равно $\Pr\{\chi^2(2) \geq 2,5\} \approx 0,28$, что также свидетельствует в пользу гипотезы о том, что случайная величина X имеет распределение Пуассона.

3.9.70. Выборочным методом получены пробы при выплавке стали, которые характеризовались двумерным признаком (X, Y). Измерялся признак X – угар кремния и признак Y – выход стали. Объем выборки $n = 15$. Статистические данные представлены следующей таблицей:

Таблица 3.9.1

X	7,9	0,9	3,7	8,1	6,9	0,8	6,0	7,2	8,2	10,2	11,2	0,5	4,6	9,7	1,0
Y	70,3	85	100	78,1	77,9	98,4	59,2	86,8	70,1	42,2	81,9	97,1	68,2	92,1	91,2

Предполагая, что признак (X,Y) характеризуется двумерным нормальным распределением:

А) проверьте гипотезу независимости $H_0:\rho=0$ против альтернативы $H_1:\rho\neq 0$,

Б) проверьте гипотезу $\tilde{H}_0:\rho=-0,6$ против альтернативы $\tilde{H}_1:\rho\neq -0,6$.

Решение. А) Для проверки гипотез H_0, H_1 нужно выполнить следующие действия.

1. Вычисляем выборочный коэффициент корреляции по формуле (3.7.7):

$$\hat{\rho}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}=-0,547.$$

2. Вычисляем статистику T_n по формуле (3.7.8):

$$T_n=\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}=\frac{-0,547\cdot\sqrt{13}}{\sqrt{1-0,2934}}=-2,32.$$

3. По таблицам квантилей распределения Стьюдента при $\alpha=0,05$ находим $t_{0,975,13}=2,160$.

4. Критерий. Так как выполняется неравенство $(|T_n|=2,32)>(t_{0,975,13}=2,16)$, то гипотезу $H_0:\rho=0$ о независимости признаков X и Y отвергаем на уровне значимости $\alpha=0,05$. Таким образом, выход стали и угар кремния при выплавках стали являются зависимыми случайными величинами, точнее, между ними существует значимая линейная стохастическая связь.

3.9.71. Выборочным методом обследовались пациенты кардиологического центра, которые характеризовались двумерным признаком (X,Y), где X – уровень холестерина в крови, Y – отношение (вес/рост) человека. Объем выборки $n=10$. Статистические данные приведены в табл. 3.9.2.

Таблица 3.9.2

X	254	240	279	284	315	250	298	284	310	337
Y	2,71	2,96	2,62	2,19	2,68	2,64	2,37	2,61	2,12	1,94

Используя критерий отношения правдоподобия для двумерной нормальной модели, основанный на статистике T_n из (3.7.8), проверьте на уровне зна-

чимости $\alpha = 0{,}10$ гипотезу о том, что уровень холестерина в крови коррелирован с отношением (вес/рост) человека.

3.9.72. В данном упражнении принята самостоятельная нумерация формул. Пусть объекты изучаемой совокупности характеризуются двумерным атрибутивным признаком (X, Y), причем признак X принимает k различных уровней $\tilde{X}_1, ..., \tilde{X}_k$, а признак Y принимает m различных уровней $\tilde{Y}_1, ..., \tilde{Y}_m$ с вероятностями $p_{ij} = P\{X = \tilde{X}_i, Y = \tilde{Y}_j\}$, *P-значение*, $j = 1, ..., m$. Отметим, что сумма вероятностей p_{ij}, $i = 1, ..., k$, $j = 1, ..., m,$ равна единице, то есть

$$\sum_{i=1}^{k} \sum_{j=1}^{m} p_{ij} = 1. \tag{1}$$

Определим маргинальные вероятности

$$P\{X = \tilde{X}_i\} = p_{i\bullet} = \sum_{j=1}^{m} p_{ij},\ i = 1, ..., k \ \text{ и } \ P\{Y = \tilde{Y}_j\} = p_{\bullet j} = \sum_{i=1}^{k} p_{ij},\ j = 1, ..., m. \tag{2}$$

Отметим, что

$$\sum_{i=1}^{k} p_{i\bullet} = \sum_{j=1}^{m} p_{\bullet j} = 1. \tag{3}$$

Статистические данные представляются с помощью таблицы сопряженности признаков X и Y (см. разд. 3.7.В), которая содержит частоты n_{ij}, $i = 1, ..., k$, $j = 1, ..., m,$ появления объектов с *i*-м уровнем признака X и с *j*-м уровнем признака Y в серии из n наблюдений, проведенных выборочным методом. Общее число наблюдений n равно сумме частот, то есть

$$n_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} = n, \tag{4}$$

и суммы частот по строкам и по столбцам обозначены соответственно через

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij},\ i = 1, ..., k, \ \ n_{\bullet j} = \sum_{i=1}^{k} n_{ij},\ j = 1, ..., m. \tag{5}$$

В рамках этой параметрической модели, которая характеризуется полиномиальным распределением

$$P\left(\{n_{ij}\}, \{p_{ij}\}\right) = \frac{n!}{n_{11}! \ldots n_{km}!} \prod_{i=1}^{k} \prod_{j=1}^{m} p_{ij}^{n_{ij}}, \tag{6}$$

с неизвестными вероятностями $\{p_{ij}\}$, $p_{i\bullet}$ *и* $p_{\bullet j}$, $i = 1, ..., k$, $j = 1, ..., m$, требуется построить по данным таблицы сопряженности признаков оценки макси-

мального правдоподобия для параметров $p_{i\bullet}$ *и* $p_{\bullet j}$ в предположении, что признаки X и Y независимы, то есть при гипотезе независимости $\hat{p} = r/n$, которая кратко записывается в виде

$$H_0^{**}: p_{ij} = p_{i\bullet} \cdot p_{\bullet j},\ i = 1,...,k,\ j = 1,...,m. \tag{7}$$

Решение. Следуя [9], отметим, что при независимости признаков X и Y, то есть при выполнении равенств (7), выполняются следующие равенства:

$$\prod_{i=1}^{k}\prod_{j=1}^{m} p_{ij}^{n_{ij}} = \left(\prod_{i=1}^{k}\prod_{j=1}^{m} p_{i\bullet}^{n_{ij}}\right)\left(\prod_{i=1}^{k}\prod_{j=1}^{m} p_{\bullet j}^{n_{ij}}\right) = \prod_{i=1}^{k} p_{i\bullet}^{n_{i1}+...+n_{im}} \prod_{j=1}^{m} p_{\bullet j}^{n_{1j}+...+n_{kj}} = \prod_{i=1}^{k} p_{i\bullet}^{n_{i\bullet}} \prod_{j=1}^{m} p_{\bullet j}^{n_{\bullet j}}$$

С учетом этого замечания, согласно (2.3.16), функция правдоподобия $L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**})$ при условии, что верна гипотеза H_0^{**}, запишется в виде

$$L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**}) = \prod_{i=1}^{k} p_{i\bullet}^{n_{i\bullet}} \prod_{j=1}^{m} p_{\bullet j}^{n_{\bullet j}}. \tag{8}$$

Для построения оценок максимального правдоподобия параметров $\hat{p}_{i\bullet}$ *и* $\hat{p}_{\bullet j}$ нужно максимизировать функцию правдоподобия $L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**})$ при наличии двух связей (3) на оцениваемые параметры. Из первой связи следует, что

$$p_{k\bullet} = 1 - \sum_{i=1}^{k-1} p_{i\bullet}. \tag{9}$$

Из второй связи следует, что

$$p_{\bullet m} = 1 - \sum_{j=1}^{m-1} p_{\bullet j}. \tag{10}$$

С учетом этих связей функция правдоподобия $L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**})$, определенная в (8), перепишется в виде

$$L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**}) = \left(1 - \sum_{i=1}^{k-1} p_{i\bullet}\right)^{n_{k\bullet}} \prod_{i=1}^{k-1} p_{i\bullet}^{n_{i\bullet}} \left(1 - \sum_{j=1}^{m-1} p_{\bullet j}\right)^{n_{\bullet m}} \prod_{j=1}^{m-1} p_{\bullet j}^{n_{\bullet j}}. \tag{11}$$

Отсюда логарифм функции правдоподобия

$$\ln L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**}) = n_{k\bullet} \ln\left(1 - \sum_{i=1}^{k-1} p_{i\bullet}\right) + \sum_{i=1}^{k-1} n_{i\bullet} \ln p_{i\bullet} +$$

$$+ n_{\bullet m} \ln\left(1 - \sum_{j=1}^{m-1} p_{\bullet j}\right) + \sum_{j=1}^{m-1} n_{\bullet j} \ln p_{\bullet j}. \tag{12}$$

Оценки максимального правдоподобия $\hat{p}_{i\bullet}$ и $\hat{p}_{\bullet j}$ параметров $p_{i\bullet}$ и $p_{\bullet j}$, $i=1,...,k$, $j=1,...,m$, являются решением системы уравнений

$$\partial \ln L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**}) / \partial\, p_{i\bullet} = 0,\ i=1,...,k; \tag{13}$$

$$\partial \ln L(\{n_{ij}\};\{p_{i\bullet}\},\{p_{\bullet j}\} \mid H_0^{**}) / \partial\, p_{\bullet j} = 0,\ j=1,...,m. \tag{14}$$

Используя (13) для заданной функции правдоподобия и учитывая (12), получаем

$$-\frac{n_{k\bullet}}{\left(1-\sum_{i=1}^{k-1} p_{i\bullet}\right)} + \frac{n_{i\bullet}}{p_{i\bullet}} = 0.$$

Отсюда следует, что $n_{k\bullet} / p_{k\bullet} = n_{i\bullet} / p_{i\bullet}$ или

$$p_{i\bullet} = \frac{p_{k\bullet}}{n_{k\bullet}} \cdot n_{i\bullet}. \tag{15}$$

Суммируя по i обе части в (15), получим $\sum_{i=1}^{k} p_{i\bullet} = \frac{p_{k\bullet}}{n_{k\bullet}} \sum_{i=1}^{k} n_{i\bullet}$. Учитывая далее, что $\sum_{i=1}^{k} p_{i\bullet} = 1$ и $\sum_{i=1}^{k} n_{i\bullet} = n$, из предыдущей формулы получаем $p_{k\bullet} / n_{k\bullet} = 1/n$. Подставив найденное выражение в (15) , окончательно получаем оценки максимального правдоподобия $\hat{p}_{i\bullet}$ параметров $p_{i\bullet}$, $i=1,...,k$, в виде $\hat{p}_{i\bullet} = n_{i\bullet}/n$, $i=1,...,k$. Аналогично из (14) получаем оценки максимального правдоподобия $\hat{p}_{\bullet j}$ параметров $p_{\bullet j}$, $j=1,...,m$, в виде $\hat{p}_{\bullet j} = n_{\bullet j}/n$, $j=1,...,m$.

Итак, оценки максимального правдоподобия теоретических частот np_{ij}, $i=1,...,k$, $j=1,...,m$, при справедливости гипотезы H_0^{**} запишутся в виде $n\hat{p}_{i\bullet} \cdot \hat{p}_{\bullet j} = n_{i\bullet} n_{\bullet j} / n$, $i=1,...,k$, $j=1,...,m$.

3.9.73. Понятие *P-значения* критерия (см. замечание 3.1.34). Пусть изучаемая совокупность состоит из объектов, которые принадлежат двум непересекающимся классам A и $\overline{A}$. Пусть, далее, p – доля объектов, принадлежащих классу A, а $q=1-p$ – доля объектов, принадлежащих классу $\overline{A}$. Статистические задачи, которые возникают при изучении данной совокупности объектов, состоят в построении оценок параметра p, $0<p<1$, и проверки гипотез относительно параметра p. Оценка максимального правдоподобия параметра p записывается в виде относительной частоты $\hat{p}=r/n$, и, следовательно, МП-оценка $\hat{q}$ параметра q равна $\hat{q}=1-\hat{p}=(n-r)/n$. Для проверки гипотезы H_0, состоящей в том, что параметр p равен заданной величине p_0,

против односторонних альтернатив $H_1^+ : p > p_0$, $H_1^- : p < p_0$ либо против двусторонней альтернативы $H_1 : p \neq p_0$ используйте критерий, основанный на статистике $\hat{p} = r/n$ и вычислите *P-значение* этого критерия.

Решение. Используя (3.1.33) и учитывая, что при гипотезе $H_0 : p = p_0$ статистика $\hat{p} = r/n$ имеет биномиальное распределение с параметрами n и p_0, получаем для проверки гипотез $H_0 : p = p_0$, $H_1^+ : p > p_0$ *P-значение* критерия в виде

$$P\text{-значение} = \sum_{i=r}^{n} C_n^r p_0^i (1-p_0)^{n-i}.$$

Далее, для проверки гипотез $H_0 : p = p_0$, $H_1^- : p < p_0$, *P-значение* критерия вычисляется по формуле

$$P\text{-значение} = \sum_{i=0}^{r} C_n^r p_0^i (1-p_0)^{n-i},$$

а для проверки гипотез $H_0 : p = p_0$, $H_1 : p \neq p_0$ – по формуле

$$P\text{-значение} = 2\min\left(\sum_{i=r}^{n} C_n^i p_0^i (1-p_0)^{n-i}, \ \sum_{i=0}^{r} C_n^i p_0^i (1-p_0)^{n-i} \right).$$

Отметим, что при большом числе опытов n и при отсутствии таблиц биномиального распределения для вычисления *P-значений* критерия следует воспользоваться асимптотическими формулами. Для записи этих формул удобно представить случайную величину с биномиальным распределением в виде суммы бернуллиевских случайных величин (см. П.1.9). Итак, пусть событие A состоит в том, что наблюдаемый объект принадлежит классу A. Определим случайную величину X в виде

$$X = \begin{cases} 1, & \text{при осуществлении события } A, \\ 0, & \text{в противном случае, событие } \overline{A}. \end{cases}$$

Результаты опыта представлены выборкой $X_1,...,X_n$, наблюденная реализация которой $x_1,...,x_n$ состоит из единиц и нулей, причем x_i равно единице или нулю, $i = 1,...,n,$ в зависимости от того, наблюденный объект принадлежит или нет классу A. Оценка максимального правдоподобия параметра p запишется в виде $\hat{p} = \overline{X} = (X_1 + ... + X_n)/n$. Далее, согласно локальной теореме Муавра – Лапласа (П.1.15), статистика $T(\vec{X}) = \sqrt{n}(\overline{X} - p_0)/\sqrt{p_0(1-p_0)}$ при гипотезе H_0 имеет асимптотически стандартное нормальное распределение,

то есть при $n \to \infty$ (на практике при $np(1-p) > 9$) справедливо следующее выражение:

$$L\left\{\frac{\sqrt{n}(\bar{X}-p_0)}{\sqrt{p_0(1-p_0)}} \mid H_0\right\} = N(0;1).$$

С учетом этого выражения, для проверки гипотез $H_0 : p = p_0$, $H_1^+ : p > p_0$ асимптотическое *P-значение* критерия, основанного на статистике $T(\vec{X}) = \sqrt{n}(\bar{X} - p_0)/\sqrt{p_0(1-p_0)}$, вычисляется по формуле

$$P\text{-значение} = P\{T(\vec{X}) \geq T(\vec{x}) \mid H_0\} = 1 - \Phi\{(\bar{x} - p_0)/\sqrt{p_0(1-p_0)/n}\},$$

где $\bar{x} = (x_1 + ... + x_n)/n$ – наблюденное значение статистики $\hat{p} = \bar{X} = (X_1 + ... + X_n)/n$.

Для проверки гипотез $H_0 : p = p_0$, $H_1^- : p < p_0$ асимптотическое *P-значение* критерия, основанного на статистике $T(\vec{X}) = \sqrt{n}(\bar{X} - p_0)/\sqrt{p_0(1-p_0)}$, вычисляется по формуле

$$P\text{-значение} = P\{T(\vec{X}) \leq T(\vec{x}) \mid H_0\} = \Phi\{(\bar{x} - p_0)/\sqrt{p_0(1-p_0)/n}\},$$

а для проверки гипотез $H_0 : p = p_0$, $H_1 : p \neq p_0$ асимптотическое *P-значение* критерия, основанного на статистике $T(\vec{X}) = \sqrt{n}(\bar{X} - p_0)/\sqrt{p_0(1-p_0)}$, вычисляется по формуле

$$P\text{-значение} = 2P\{T(\vec{X}) \geq |T(\vec{x})| \mid H_0\} = 2[1 - \Phi\{(\bar{x} - p_0)/\sqrt{p_0(1-p_0)/n}\}].$$

Окончательное решение с использованием *P-значения* принимается следующим образом. Если вычисленное *P-значение* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается.

Приведем числовой пример. Партия изделий принимается, если вероятность того, что изделие окажется бракованным, не превыщает 0,02. Среди случайно отобранных $n = 480$ изделий оказалось $m = 12$ дефектных. Можно ли принять партию?

Решение данной задачи сведем к проверке гипотез $H_0 : p = p_0$, $H_1^+ : p > p_0$, где $p_0 = 0,02$. Наблюденное значение $T(\vec{x})$ статистики $T(\bar{X})$ равно

$$T(\vec{x}) = \sqrt{n}\{(m/n) - p_0\}/\sqrt{p_0(1-p_0)} = \sqrt{480}\{(12/480) - 0,02\}/\sqrt{0,02 \cdot 0,98} = 0,78.$$

Примем уровень значимости $\alpha = 0,05$. Так как произведение $480 \cdot 0,02 \cdot 0,98 = 9,408$, то неравенство $np(1-p) > 9$ выполняется, поэтому воспользуемся нормальной аппроксимацией статистики $T(\bar{X})$. При $\alpha = 0,05$

ее критическое значение $\lambda_{1-\alpha} = \Phi^{-1}(1-\alpha) = \Phi^{-1}(0,95) = 1,645$ и выполняется неравенство $(T(\vec{x}) = 0,78) < (\lambda_{0,95} = 1,645)$, поэтому гипотезу $H_0 : p = 0,02$ принимаем на уровне значимости $\alpha = 0,05$. Далее, асимптотическое $P\text{-значение} = P\{T(\vec{X}) \geq T(x) \mid H_0\} = 1 - \Phi(0,78) = 0,22$, что также позволяет принять партию изделий.

3.9.74. Пусть $L(X) = N(\mu, \sigma^2)$. Рассмотрим стандартные критерии для проверки гипотез о параметрах μ и σ^2 по выборке $X_1, ..., X_n$ из нормального распределения.

А) Вычислите *P-значение* критерия проверки гипотезы $H_0 : \mu = \mu_0$ против односторонних альтернатив $H_1^- : \mu < \mu_0$, $H_1^- : \mu < \mu_0$ либо против двусторонней альтернативы $H_1 : \mu \neq \mu_0$.

Б) Вычислите *P-значение* критерия проверки гипотезы $H_0 : \sigma^2 = \sigma_0^2$ против односторонних альтернатив $H_1^+ : \sigma^2 > \sigma_0^2$, $H_1^- : \sigma^2 < \sigma_0^2$ либо против двусторонней альтернативы $H_1 : \sigma^2 \neq \sigma_0^2$.

Решение. А) Для проверки гипотез пункта (А) воспользуемся критерием, основанным на статистике $T(\vec{X}) = \sqrt{n}(\overline{X} - \mu_0)/\tilde{S}$, где $\overline{X}$ – выборочное среднее и $\tilde{S}$ – положительное значение корня квадратного из подправленной выборочной дисперсии, которая является несмещенной оценкой параметра σ^2. Согласно теореме (2.8.31), статистика $T(\vec{X}) = \sqrt{n}(\overline{X} - \mu_0)/\tilde{S}$ при гипотезе $H_0 : \mu = \mu_0$ имеет распределение Стьюдента с числом степеней свободы $(n-1)$, то есть справедливо следующее выражение:

$$L\left\{\frac{\sqrt{n}(\overline{X} - \mu_0)}{\tilde{S}} \mid H_0\right\} = St(n-1).$$

С учетом этого замечания, *P-значение* критерия, основанного на статистике $T(\vec{X}) = \sqrt{n}(\overline{X} - \mu_0)/\tilde{S}$, для проверки соответствующих гипотез вычисляются по формулам, приведенным в табл. 3.9.3.

Таблица 3.9.3

Нулевая гипотеза	Альтернатива	*P-значение*
$H_0 : \mu = \mu_0$, σ^2 *неизвестно*	$H_1^+ : \mu > \mu_0$	$P = \Pr\{St(n-1) \geq t(\vec{x}) \mid H_0\}$
	$H_1^- : \mu < \mu_0$	$P = \Pr\{St(n-1) \leq t(\vec{x}) \mid H_0\}$
	$H_1 : \mu \neq \mu_0$	$P = 2\min[\Pr\{St(n-1) \geq t(\vec{x}) \mid H_0\}, \Pr\{St(n-1) \leq t(\overline{x}) \mid H_0\}]$

Здесь через $St(n-1)$ обозначена случайная величина с распределением Стьюдента и числом степеней свободы $(n-1)$, а $t(x)=\sqrt{n}(\overline{x}-\mu_0)/\tilde{s}$ – наблюденное значение статистики $T(\vec{X})=\sqrt{n}(\overline{X}-\mu_0)/\tilde{S}$, вычисленное по наблюденной реализации $x_1,...,x_n$ выборки $X_1,...,X_n$. Окончательное решение с использованием *P-значения* принимается следующим образом. Если вычисленное *P-значение* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается.

Б) Для проверки гипотез пункта (Б) воспользуемся критерием, основанным на статистике $\chi^2=(n-1)\tilde{S}^2/\sigma_0^2$, где $\tilde{S}^2$ – подправленная выборочная дисперсия, вычисляемая по выборке $X_1,...,X_n$. Согласно теореме 2.8.31, статистика $\chi^2=(n-1)\tilde{S}^2/\sigma_0^2$ при гипотезе $H_0:\sigma^2=\sigma_0^2$ имеет распределение хи-квадрат с числом степеней свободы $(n-1)$, то есть справедливо следующее выражение:

$$L\left\{\frac{(n-1)\tilde{S}^2}{\sigma_0^2}\mid H_0\right\}=\chi^2(n-1).$$

С учетом этого замечания, *P-значение* критерия, основанного на статистике $\chi^2=(n-1)\tilde{S}^2/\sigma_0^2$, для проверки соответствующих гипотез вычисляются по формулам, приведенным в табл. 3.9.4.

Таблица 3.9.4

Нулевая гипотеза	Альтернатива	*P-значение*
$H_0:\sigma^2=\sigma_0^2$	$H_1^+:\sigma^2>\sigma_0^2$	$P=\Pr\{\chi^2(n-1)\ge\chi^2_{\text{в}}\mid H_0\}$
	$H_1^-:\sigma^2<\sigma_0^2$	$P=\Pr\{\chi^2(n-1)\le\chi^2_{\text{в}}\mid H_0\}$
	$H_1:\sigma^2\ne\sigma_0^2$	$P=2\min[\Pr\{\chi^2(n-1)\ge\chi^2_{\text{в}}\mid H_0\},$ $\Pr\{\chi^2(n-1)\le\chi^2_{\text{в}}\mid H_0\}]$

Здесь $\chi^2_{\text{в}}=(n-1)\tilde{s}^2/\sigma_0^2$ – наблюденное (вычисленное) значение статистики $\chi^2=(n-1)\tilde{S}^2/\sigma_0^2$. Отметим, что для вычисления *P-значения* по приведенным формулам следует воспользоваться таблицами для интеграла вероятностей χ^2 в [6, с. 140]. Окончательное решение с использованием *P-значения* критерия принимается следующим образом. Если вычисленное *P-значения* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается.

3.9.75. Проверка связанных выборок с помощью парного *t*-критерия. Пусть $L(X_i) = N(\mu_i, \sigma_i^2)$, $i = 1,2$, где μ_i и σ_i^2 – соответственно среднее и дисперсия величины X_i, а $\sigma_{ij} = \sigma_{ji} = COV(X_{ij})$ – ковариация между X_i и X_j, $i, j = 1,2$. Отметим, что $\sigma_{ii} = \sigma_i^2$, $i = 1,2$. Коэффициент корреляции ρ_{ij} между X_i и X_j определяется в виде $\rho_{ij} = \sigma_{ij} / \sigma_i \sigma_j$, $i, j = 1,2$. Из популяции W производим выборку объема n и наблюдаем пару (X_1 и X_2) у каждого индивидуума в выборке. Результатами наблюдений являются $x_{11},...,x_{1n}$ и $x_{21},...,x_{2n}$, так что x_{ik} есть k-е наблюдение случайной величины X_i, $i = 1,2$, $k = 1,...,n$. Оценки максимального правдоподобия параметров μ_i и σ_i^2, σ_{ij} и ρ_{ij} по выборкам $X_{11},...,X_{1n}$ и $X_{21},...,X_{2n}$ соответственно записываются в виде

$$\bar{X}_{i\bullet} = \frac{1}{n}\sum_{k=1}^{n} X_{ik}\ ,\ S_i^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ik} - \bar{X}_{i\bullet})^2\ ,\ S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(X_{ik} - \bar{X}_{i\bullet})(X_{jk} - \bar{X}_{j\bullet})\ ,$$

$$r_{ij} = S_{ij} / S_i S_j\ .$$

Выборочные дисперсии, коэффициенты ковариации и корреляции представляются в матричной форме и соответственно ковариационная и корреляционная матрицы записываются как

$$\begin{bmatrix} S_{11} = S_1^2 & S_{12} \\ S_{21} & S_{22} = S_2^2 \end{bmatrix},\ \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}.$$

Если случайные величины X_1 и X_2 описывают одинаковые или однотипные характерисики *для одного и того же индивидуума* до и после «обработки», то оценка эффекта «обработки» (например, оценка новой технологии) может быть проведена путем сравнения средних значений величин X_1 и X_2. Для осуществления этого требуется проверить гипотезу $H_0 : \mu_1 - \mu_2 = \theta_0$, где θ_0 – заданная постоянная величина, которая может равняться нулю. Проверка гипотезы H_0 осуществляется с помощью *парного t-критерия*, называемого также *t*-критерием для связанных (зависимых) выборок. Статистика критерия (см., например, [3]) имеет вид

$$T(\vec{X}_1, \vec{X}_2) = \frac{\sqrt{n} \cdot [(\bar{X}_{1\bullet} - \bar{X}_{2\bullet}) - \theta_0]}{\sqrt{S_1^2 + S_2^2 - 2S_{12}}},$$

где $\vec{X}_1 = (X_{11},...,X_{1n})$ и $\vec{X}_2 = (X_{21},...,X_{2n})$ – выборки однотипного признака до и после «обработки», а S_D^2 – выборочная диперсия разности $D = (X_1 - X_2)$, вычисляемая по формуле

$$S_D^2 = S_1^2 + S_2^2 - 2S_1S_2\rho_{12} = S_1^2 + S_2^2 - 2S_{12}.$$

При выполнении гипотезы H_0 статистика $T(\vec{X}_1, \vec{X}_2)$ имеет распределение Стьюдента с числом степеней свободы $(n-1)$, то есть справедливо выражение $L\{\sqrt{n}(\overline{D} - \theta_0)/S_D \mid H_0\} = St(n-1)$. *P-значения* критерия зависят от правила принятия решения, то есть от альтернативной гипотезы и приводятся в табл. 3.9.5.

Таблица 3.9.5

Нулевая гипотеза	Альтернатива	*P-значение*
$H_0 : \mu_1 - \mu_2 = \theta_0$	$H_1^+ : \mu_1 - \mu_2 > \theta_0$	$P = \Pr\{St(n-1) \geq t(\vec{x}_1, \vec{x}_2) \mid H_0\}$
	$H_1^- : \mu_1 - \mu_2 < \theta_0$	$P = \Pr\{St(n-1) \leq t(\vec{x}_1, \vec{x}_2) \mid H_0\}$
	$H_1 : \mu_1 - \mu_2 \neq \theta_0$	$P = 2[\Pr\{St(n-1) \geq \mid t(\vec{x}_1, \vec{x}_2) \mid \mid H_0\},$

Окончательное решение с использованием *P-значения* критерия принимается следующим образом. Если вычисленное *P-значение* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается. Пусть для примера объем выборки $n = 108$ и наблюденные значения выборочных оценок следующие: $\overline{x}_{1\bullet} = 7{,}373$, $\overline{x}_{2\bullet} = 7{,}413$, $s_1^2 = 0{,}1253$, $s_2^2 = 0{,}1184$ и $s_{12} = 0{,}1101$. Наблюденные значения выборочных оценок ковариационной и корреляционной матриц сответственно имеют вид

$$\begin{bmatrix} s_{11} = s_1^2 & s_{12} \\ s_{21} & s_{22} = s_2^2 \end{bmatrix} = \begin{bmatrix} 0{,}1253 & 0{,}1101 \\ 0{,}1101 & 0{,}1184 \end{bmatrix} \text{ и } \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1{,}000 & 0{,}904 \\ 0{,}904 & 1{,}000 \end{bmatrix}.$$

Значение выборочного коэффициента корреляции $r_{12} = 0{,}904$ свидетельствует о наличии «тесной» линейной связи между случайными величинами X_1 и X_2. По этим данным проверим гипотезу $H_0 : \mu_1 - \mu_2 = 0$ против односторонней альтернативы $H_1^- : \mu_1 - \mu_2 < 0$, используя *P-значение* критерия, основанного на статистике $T(\vec{X}_1, \vec{X}_2)$. Наблюденное значение $t(\vec{x}_1, \vec{x}_2)$ статистики $T(\vec{X}_1, \vec{X}_2)$ равно

$$t(\vec{x}_1, \vec{x}_2) = \frac{\sqrt{108}(7{,}373 - 7{,}413)}{\sqrt{0{,}1253 + 0{,}1184 - 2 \cdot 0{,}1101}} = -2{,}71.$$

Так как $P\text{-значение} = \Pr\{St(107) \leq -2{,}71\} < 0{,}005$, то гипотезу $H_0 : \mu_1 - \mu_2 = 0$ отвергаем.

Отметим, что статистику $T(\vec{X}_1, \vec{X}_2)$ часто записывают в более удобном для вычислений виде (см. разд. 3.6.Г). Для этого используют выборку $V_i = X_{1i} - X_{2i}$, $i = 1,...,n,$ для разности $V = (X_1 - X_2)$. С учетом этих обозначений другая запись статистики имеет вид $T(\vec{X}_1, \vec{X}_2) = \sqrt{n}(\overline{V} - \theta_0)/\tilde{S}_V$, где $\overline{V}$ – выборочное среднее разности $V = (X_1 - X_2)$ и $\tilde{S}_V^2$ – подправленная выборочная дисперсия этой разности, то есть

$$\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i = \overline{X}_{1\bullet} - \overline{X}_{2\bullet}, \tilde{S}_V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} V_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} V_i\right)^2\right\}.$$

Таким образом, парный *t*-критерий совпадает с *t*-критерием для выборки, состоящей из разностей $V_i = X_{1i} - X_{2i}$, $i = 1,...,n$.

3.9.76. Для сравнения урожайности пшеницы двух сортов (А) и (Б) было выбрано десять участков пашни размерами по две сотки и каждый из них был разбит на два одинаковых участка по одной сотке. Сравнивался сорт (А) – обычная разновидность и сорт (Б) – новый гибрид. На каждом отдельном участке были обеспечены одинаковые условия роста и созревания. Далее, на каждом участке размерами по две сотки случайным образом выбирался участок размером одна сотка и засевался сортом А. Оставшиеся десять участков по одной сотке были засеяны сортом Б. Результаты урожайности приведены в табл. 3.9.6. Требуется по этим данным проверить гипотезу о том, что урожайности сортов (А) и (Б) одинаковые, против альтернативы, утверждающей, что урожайность сорта (Б) выше урожайности сорта (А).

Таблица 3.9.6

№ участка	Сорт (А) $\{x_i\}$	Сорт (Б) $\{y_i\}$	$\{v_i = y_i - x_i\}$
1	36,9	36,8	-0,1
2	35,2	37,1	1,9
3	31,2	31,4	0,2
4	34,1	34,1	0,0
5	36,1	35,9	-0,2
6	34,1	35,2	1,1
7	37,2	37,9	0,7
8	36,8	37,2	0,4
9	29,6	30,2	0.6
10	35,4	36,5	1.1

Решение. Пусть признак X характеризует урожайность сорта (А) и признак Y характеризует урожайность сорта (Б). Пусть далее $\mu_V = M\{(Y - X)\}$ –

истинное среднее значение разности $V=(Y-X)$. Сформулированные гипотезы кратко записываются в виде $H_0:\mu_V=0$, $H_1^+:\mu_V>0$. Для проверки этих гипотез используем парный *t*-критерий, основанный на статистике $T(\vec{X}_1,\vec{X}_2)=\sqrt{n}(\overline{V}-\theta_0)/\tilde{S}_V$. По данным таблицы вычисляем среднюю разность $\overline{v}=(v_1+...+v_n)/n=(-0{,}1+...+1{,}1)/10=0{,}57$ и подправленную выборочную дисперсию разностей $\tilde{s}_V^2$ по формуле

$$\tilde{s}_V^2=\frac{1}{n-1}\sum_{i=1}^{n}(v_i-\overline{v})^2=\{(-0{,}1-0{,}57)^2+...+(1{,}1-0{,}57)^2\}=0{,}4312.$$

Используя эти данные, вычисляем наблюденное значение $t(\vec{x},\vec{y})$ статистики $T(\vec{X},\vec{Y})$

$$t(\vec{x},\vec{y})=\sqrt{n}(\overline{v}-0)/\tilde{s}_V=\sqrt{10}\cdot(0{,}57-0)/\sqrt{0{,}4312}=2{,}7.$$

Далее, используя таблицы функции распределения Стьюдента (см. табл. 3.1.а в [6]), находим $\Pr\{St(9)\le 2{,}7\}=0{,}988$. Таким образом, для критерия проверки гипотез $H_0:\mu_V=0$, $H_1^+:\mu_V>0$ *P-значение* $=\Pr\{St(9)>2{,}7\}=1-0{,}988=0{,}012$ и, следовательно, гипотезу H_0 отвергаем, то есть заключаем, что урожайность сорта (Б) выше урожайности сорта (А).

3.9.77. Проверка равенства двух связанных дисперсий нормальных совокупностей. Оценка эффекта «обработки» при сравнении изменьчивости двух признаков X – до обработки и Y – после обработки по двум *связанным* выбокам $X_1,...,X_n$ и $Y_1,...,Y_n$ из нормальных совокупностей сводится к проверке гипотезы $H_0:\sigma_X^2=\sigma_Y^2$ против двусторонней альтернативы $H_1:\sigma_X^2\ne\sigma_Y^2$. Проверка этих гипотез проводится с помощью критерия, основанного на статистике

$$T(\vec{X},\vec{Y})=\frac{\sqrt{n-2}\cdot(SS_X-SS_Y)}{2\sqrt{SS_X\cdot SS_Y-(SS_{XY})^2}},$$

где SS_X – выборочная сумма квадратов отклонений для признака X, SS_Y – выборочная сумма квадратов отклонений для признака Y, которые вычисляются по формулам

$$SS_X=\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2,\ SS_Y=\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}Y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}Y_i\right)^2,$$

$$SS_{XY}=\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\cdot\sum_{i=1}^{n}X_i\cdot\sum_{i=1}^{n}Y_i.$$

Хорошо известно (см., например, [20]), что при гипотезе $H_0:\sigma_X^2=\sigma_Y^2$ статистика $T(\vec{X},\vec{Y})$ имеет распределение Стьюдента с числом степеней свободы $(n-2)$, то есть справедливо выражение $L\{T(\vec{X},\vec{Y})\,|\,H_0\}=St(n-2)$. С учетом этого замечания, критерий проверки гипотез $H_0:\sigma_X^2=\sigma_Y^2$, $H_1:\sigma_X^2\neq\sigma_Y^2$, основанный на статистике $T(\vec{X},\vec{Y})$, определяется критической областью размера α, которая записывается в виде $\mathrm{X}_{1\alpha}=\{(\vec{x},\vec{y}):\ |t(\vec{x},\vec{y})|>t_{1-\alpha/2,(n-2)}\}$. Приведем числовой пример для следующих данных:

$$\begin{array}{l|llll|l}\{x_i\}&21&18&20&21&\sum x_i=80\\ \hline \{y_i\}&26&33&27&34&\sum y_i=120\end{array}.$$

Вычислив $SS_X=6$, $SS_Y=50$ и $SS_{XY}=(21\cdot26+18\cdot33+20\cdot27+\cdot34)-(80\cdot120)/4=-6$, получаем наблюденное значение статистики

$$|t(\vec{x},\vec{y})|=\frac{\sqrt{4-2}\cdot|6-50|}{2\cdot\sqrt{6\cdot50-(-6)^2}}=1{,}91.$$

Пусть $\alpha=0{,}05$, тогда квантиль уровня $(1-\alpha/2)$ распределения Стьюдента с числом степеней свободы $(n-2)$ равен $t_{1-\alpha/2,(n-2)}=t_{0{,}975,(2)}=4{,}30$. Так как выполняется неравенство $(t(\vec{x},\vec{y})=1{,}91)<(t_{0{,}975,(2)}=4{,}30)$, то гипотезу H_0 принимаем на уровне значимости $\alpha=0{,}05$. Далее, используя таблицы функции распределения Стюдента в работе [6, табл. 3.1а], находим, что $\Pr\{St(2)\le1{,}91\}\approx0{,}90$. Отсюда следует, что *P-значение*$=2P\{St(2)>1{,}91\}\approx0{,}20$ и, следовательно, гипотезу H_0 принимаем. Отметим, что если предположение нормальности изучаемых совокупностей не может быть обосновано, то используют критерий Уилкоксона для разностей пар или знаковый критерий (см. разд. 5.1 – 5.3).

3.9.78. Признак X характеризует размер выпускаемой предприятием детали. Выборочным методом было проведено $n=200$ измерений $X_1,\ldots,X_{200}$ размера детали. В результате этих измерений была получена следующая реализация выборки: $x_1=6{,}75$, $x_2=6{,}80,\ldots,$ $x_{200}=6{,}72$, причем оказалось, что наблюденные значения для 1-й и n-й порядковых статистик соответственно равны $x_{(1)}=6{,}68$ и $x_{(200)}=6{,}84$, то есть наблюденное значение размаха выборки равно $R=6{,}84-6{,}68=0{,}16$. Далее, выбрав первоначально число равных интервалов группироки $k=9$, по наблюденной реализации $x_1=6{,}75$, $x_2=6{,}80,\ldots,$ $x_{200}=6{,}72$ выборки был построен интервальный вариационный ряд частот (см. первые два столбца в табл. 3.9.7). Учитывая, что отношение

$R/k \approx 0,02$, длина равных интервалов группировки $\Delta_1,...,\Delta_9$ была выбрана равной величине $\Delta_i = (y_{i+1} - y_i) = 0,02$, $i = 1,...,9$, причем левая граница первого интервала была выбрана равной $x_{(1)} - \Delta_i / 2 = 6,68 - 0,02/2 = 6,67$, а правая граница последнего (девятого) интервала – $x_{(n)} + \Delta_i / 2 = 6,84 + 0,02/2 = 6,85$. В таблице также приведены накопленные относительные частоты и вычисленные значения $\{np_i\}$ ожидаемых гипотетических частот.

Таблица 3.9.7

Интерв. групп. $\{\Delta_i\}$	Середины интерв. $\{\hat{x}_i\}$	Частоты $\{n_i\}$	Отн. част. $\{n_i/n\}$	Накопл. отн. частоты	$\{p_i\}$	$\{np_i\}$	$\frac{(n_i - np_i)^2}{np_i}$
[6,67, 6,69)	6,68	2	0,010	0,010			
[6,69, 6,71)	6,70	15**(17)**	0,075	0,085	0,057	11,38	2,77
[6,71, 6,73)	6,72	17	0,085	0,170	0,114	22,84	1,49
[6,73, 6,75)	6,74	44	0,220	0,390	0,205	40,94	0,22
[6.75, 6,77)	6,76	52	0,260	0,650	0,240	48,68	0,22
[6,77, 6,79)	6,78	44	0,220	0,870	0,204	40,80	0,25
[6.79, 6,81)	6,80	14	0,070	0,940	0,114	22,80	3,39
[6,81, 6,83)	6,82	11**(12)**	0,055	0,995	0,057	11,40	0,03
[6,83, 6,85)	6,84	1	0,005	1,000			
		$n=200$	1,000				$\chi^2 = 8,37$

Используя хи-квадрат-критерий, проверьте сложную гипотезу о нормальности изучаемого признака X, которая кратко записывается в виде $H_0 : F_X(x) = \Phi\{(x-\mu)/\sigma\}$, параметры μ и σ^2 неизвестны. Постройте график эмпирической функции распределения признака X.

Решение. По данным интервального ряда частот вычисляем оценки неизвестных параметров μ и σ^2:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{k} \hat{x}_i \cdot n_i = \frac{1}{200}(6,68 \cdot 2 + 6,70 \cdot 15 + ... + 6,84 \cdot 1) = 6,76,$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{k} (\hat{x}_i - \hat{\mu})^2 \cdot n_i = \frac{1}{200}\{(6,68 - 6,76)^2 \cdot 2 + ... + (6,84 - 6,76)^2 \cdot 1\} = 0,001,$$

$$\hat{\sigma} = \sqrt{0,001} = 0,032.$$

Отметим, что в данном примере первые два и последние два выделенных в таблице интервалов группировки объединены в один интервал, чтобы выполнить рекомендацию (см. замечание 3.5.13), при которой должно выполняться

неравенство $\min(n_1,...,n_k) \geq 5$. Таким образом, число интервалов группировки становится равным $k = 7$ и, следовательно, число степеней свободы распределения хи-квадрат-статистики критерия, с учетом того, что оценивается два неизвестных параметра μ и σ^2, равно $(k-1-2) = 7-3 = 4$. Для вычисления статистики критерия предварительно вычислим гипотетические вероятности $p_i(\hat{\theta}^*) = P\{X \in \Delta_i \mid H_0^*\}$, $i = 1,...,k$, которые зависят от векторного параметра $\hat{\theta}^* = (\hat{\mu}, \hat{\sigma})$. Для вычисления этих вероятностей используем таблицы стандартной номальной функции распределения $\Phi(x)$. Например, вероятность попадания случайной величины X в первый интервал группировки $\Delta_1 = [6,67, 6,71)$ равна

$$p_1(\hat{\theta}^*) = P\{X \in \Delta_1 \mid H_0^*\} = \Phi\{(6,71-6,76)/0,032\} - \Phi\{(6,67-6,76)/0,032\} =$$
$$= \Phi(-1,56) - \Phi(-2,81) = \Phi(2,81) - \Phi(1,56) = 0,9975 - 0,9406 = 0,0569.$$

Остальные вероятности вычисляются аналогично и занесены в шестом столбце табл. 3.9.7. Используя эти данные, вычисляем наблюденное значение статистики критерия $\chi^2 = 8,37$. Для принятия решения примем уровень значимости $\alpha = 0,05$, следовательно, критическое значение модифицированной статистики хи-квадрат $\chi^2_{1-\alpha,(k-1-2)} = \chi^2_{0,95,(4)} = 9,49$. Так как выполняется неравенство $(\chi^2 = 8,37) < (\chi^2_{0,95,(4)} = 9,49)$, то гипотезу о нормальности признака X принимаем на уровне зачимости $\alpha = 0,05$.

3.9.78а. Во время войны вся территория Лондона была разделена на 576 участков площадью по $0,25$ км^2 и было подсчитано число участков n_k, на которые упало k снарядов. Эти данные представлены в табл. 3.5.2

Таблица 3.5.2

k	0	1	2	3	4	$(k \geq 5)$
n_k	229	211	93	35	7	1

Пусть X – дискретная случайная величина, которая принимает целочисленные значения 0, 1, 2, ... с вероятностями

$$p_k(\theta) = P(X = k) = \frac{\theta^k e^{-\theta}}{k!},\ k = 0, 1, 2, ..., \text{ то есть } L(X) = \Pi y(\theta).$$

Требуется, используя критерий хи-квадрат, проверить сложную гипотезу $H_0 : L(X) = \Pi y(\theta)$, $0 < \theta < \infty$, по данным опыта, представленным в табл. 3.5.2. Принять уровень значимости $\alpha = 0,05$.

3.9.79. Сравнение двух наблюденных частот с применением нормальной аппроксимации. Имеется две достаточно большие совокупности объектов, причем доли объектов, обладающих некоторым признаком (А), (например, А – объект бракованный) в каждой совокупности соответственно равны p_1 и p_2. Параметры p_1 и p_2 *неизвестны*. Выборочным методом в каждой совокупности наблюдается $n_{1\bullet}$ и $n_{2\bullet}$ объектов и наблюденные данные $\{n_{ij}\}$, $i, j = 1,2,$ заносятся в таблицу 2×2 (табл. 3.9.8).

Таблица 3.9.8

Совокупность \ Признак	Бракованные (A)	Хорошие $(\overline{A})$	Всего
1-я совокупность	n_{11}	n_{12}	$n_{1\bullet}$
2-я совокупность	n_{21}	n_{22}	$n_{2\bullet}$
Всего	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Предполагается, что общее число n выбранных объектов из двух совокупностей является заданным числом, а числа $n_{\bullet 1}$ и $n_{\bullet 2}$ являются случайными, так как $n_{\bullet 1} = n_{11} + n_{12}$ и $n_{\bullet 2} = n_{12} + n_{22}$, где n_{12}, n_{21}, n_{12}, n_{22} – случайные величины. По наблюденным данным $\{n_{ij}\}$, $i, j = 1,2,$ требуется проверить гипотезу $H_0 : p_1 = p_2 = p$, где p – некоторое *неизвестное* число, $0 < p < 1$.

Обозначим $\hat{p}_1 = n_{11} / n_{1\bullet}$ и $\hat{p}_2 = n_{21} / n_{2\bullet}$. При справедливости гипотезы $H_0 : p_1 = p_2 = p$ разность $\hat{p}_1 - \hat{p}_2$ является несмещенной оценкой нуля, так как $M\{(\hat{p}_1 - \hat{p}_2) \mid H_0\} = p_1 - p_2 = 0$. Далее, вычислим дисперсию разности $\hat{p}_1 - \hat{p}_2$. При справедливости гипотезы $H_0 : p_1 = p_2 = p$ имеем

$$D\{(\hat{p}_1 - \hat{p}_2) \mid H_0\} = D(\hat{p}_1) + D(\hat{p}_2) = p(1-p)/n_{1\bullet} + p(1-p)/n_{2\bullet} = \\ = p(1-p)\{(1/n_{1\bullet}) + (1/n_{2\bullet})\}.$$

Так как p неизвестное число, то для того, чтобы оценить неизвестную дисперсию разности $(\hat{p}_1 - \hat{p}_2)$, нужно построить оценку $p(1-p) = p - p^2$. Как отмечается в [9, с. 142], несмещенная оценка для $p - p^2$ записывается в виде $\dfrac{n_{\bullet 1}}{n} - \dfrac{n_{\bullet 1}(n_{\bullet 1} - 1)}{n(n-1)} = \dfrac{n_{\bullet 1} n_{\bullet 2}}{n(n-1)}$. Используя эту оценку, получаем несмещенную оценку для дисперсии $D\{(\hat{p}_1 - \hat{p}_2) \mid H_0\}$ в виде

$$\hat{D}\{(\hat{p}_1 - \hat{p}_2) \mid H_0\} = \frac{n_{\bullet 1} n_{\bullet 2}}{n(n-1)} \left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}} \right) = \frac{n_{\bullet 1} n_{\bullet 2}}{n(n-1)} \frac{(n_{2\bullet} + n_{1\bullet})}{n_{1\bullet} n_{2\bullet}} = \frac{n_{\bullet 1} n_{\bullet 2}}{(n-1) n_{1\bullet} n_{2\bullet}}.$$

Далее, так как при гипотезе H_0 выполняется равенство $M(\hat{p}_1) = M(\hat{p}_2)$, то статистика

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{D}\{(\hat{p}_1 - \hat{p}_2) \mid H_0\}}} = \frac{(n_{11}/n_{1\bullet}) - (n_{21}/n_{2\bullet})}{\sqrt{\dfrac{n_{\bullet 1} n_{\bullet 2}}{(n-1) n_{1\bullet} n_{2\bullet}}}} = \frac{n_{11} - \dfrac{n_{1\bullet} n_{\bullet 1}}{n}}{\left[\dfrac{n_{1\bullet} n_{\bullet 1} n_{2\bullet} n_{\bullet 2}}{n^2 (n-1)}\right]^{1/2}}$$

имеет асимптотически стандартное нормальное распределение, то есть справедливо выражение

$$L\left\{\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{D}\{(\hat{p}_1 - \hat{p}_2)\}}} \mid H_0\right\} = N(0;1) \text{ при } n \to \infty .$$

Отметим, что при больших значениях объема выборки n можно заменить число $(n-1)$ на n и, следовательно, статистику критерия можно записать в более удобном для вычислений виде

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{D}\{(\hat{p}_1 - \hat{p}_2) \mid H_0\}}} = \frac{(n_{11}/n_{1\bullet}) - (n_{21}/n_{2\bullet})}{\sqrt{\dfrac{n_{\bullet 1}}{n}\left(1 - \dfrac{n_{\bullet 1}}{n}\right)\left(\dfrac{1}{n_{1\bullet}} + \dfrac{1}{n_{2\bullet}}\right)}} .$$

С использованием поправки на непрерывность, для повышения качества асимптотической аппроксимации распределения статистики, окончательно запишем ее в виде

$$T(\{n_{ij}\}) = [(n_{11}/n_{1\bullet} - 1/2n_{1\bullet}) - (n_{21}/n_{2\bullet} - 1/2n_{2\bullet})] / \sqrt{n_{\bullet 1}(1 - n_{\bullet 1}/n)(1/n_{1\bullet} + 1/n_{2\bullet})/n} .$$

Итак, с учетом нормальной аппроксимации, критерий проверки гипотезы $H_0 : p_1 = p_2 = p$ против односторонней альтернативы $H_1^+ : p_1 > p_2$ определяется критической областью размера α как

$$\mathrm{X}_{1\alpha}^{+} = (\{n_{ij}\} : T(\{n_{ij}\} \geq \lambda_{1-\alpha}),\ \lambda_{1-\alpha} = \Phi^{-1}(1-\alpha) .$$

Критерий проверки гипотезы $H_0 : p_1 = p_2 = p$ против односторонней альтернативы $H_1^- : p_1 < p_2$ определяется критической областью размера α как

$$\mathrm{X}_{1\alpha}^{-} = (\{n_{ij}\} : T(\{n_{ij}\} \leq \lambda_{\alpha}),\ \lambda_{\alpha} = \Phi^{-1}(\alpha) .$$

Критерий проверки гипотезы $H_0 : p_1 = p_2 = p$ против двусторонней альтернативы $H_1 : p_1 \neq p_2$ определяется критической областью размера α в виде

$$\mathrm{X}_{1\alpha} = (\{n_{ij}\} : |T(\{n_{ij}\}| \geq \lambda_{1-\alpha/2}),\ \lambda_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2) .$$

Формулы для *P-значений* критерия, основанного на статистике $T(\{n_{ij}\})$, также зависят от рассматриваемых альтернатив. Для проверки гипотез $H_0: p_1 = p_2 = p$, $H_1^+: p_1 > p_2$ имеем *P-значение* $= 1 - \Phi\{T(\{n_{ij}\})\}$; для проверки гипотез $H_0: p_1 = p_2 = p$, $H_1^-: p_1 < p_2$ – *P-значение* $= \Phi\{T(\{n_{ij}\})\}$ и для проверки гипотез $H_0: p_1 = p_2 = p$, $H_1: p_1 \neq p_2$ – *P-значение* $= 2[1 - \Phi\{T(\{n_{ij}\})\}]$. Окончательное решение с использованием *P-значения* критерия принимается следующим образом. Если вычисленное *P-значение* критерия меньше уровня значимости α, то гипотеза H_0 отвергается, в противном случае H_0 принимается. Отметим (см., например, [9]), что нормальная аппроксимация распределения статистики $T(\{n_{ij}\})$ при нулевой гипотезе считается удовлетворительной, если наименьшая из величин $n_{i\bullet}n_{\bullet j}/n$, $i, j = 1,2$, будет больше 5. Если это правило не выполняется, то следует применять точный критерий Фишера.

Рассмотрим числовой пример, статистические данные для которого взяты из [9, с. 143]. Имеются две группы пациентов: группу (A) образуют пациенты, которым делали прививки шприцами типа (A), и соответственно группу (B) образуют пациенты, которым делали прививки шприцами типа (B). Отметим, что тип шприца выбирается случайно, независимо от пациента. В табл. 3.9.9 приведены данные наблюденных реакций, вызванных у пациентов прививками, которые проводились двумя типами шприцев (A) и (B).

Таблица 3.9.9

Тип шприца	Реакция		
	Положительная	Отрицательная	Всего
1-я группа (A)	27	13	40
2-я группа (B)	5	10	15
Всего	32	23	55

По наблюденным данным требуется проверить гипотезу, согласно которой тип применяемого шприца не приводит к различиям между группами пациентов.

Решение. Сначала убеждаемся, что наименьшая из величин $n_{i\bullet}n_{\bullet j}/n$, $i, j = 1,2$, равна $23 \cdot 15/55 = 6,27$, что больше 5, поэтому нормальная аппроксимация распределения статистики $T(\{n_{ij}\})$ будет удовлетворительной. Далее, вычисляем $n_{\bullet 1}/n = 32/55 = 0,5818$ и статистику

$$T(\{n_{ij}\})=\frac{\left(\frac{n_{11}}{n_{1\bullet}}-\frac{1}{2n_{1\bullet}}\right)-\left(\frac{n_{21}}{n_{2\bullet}}-\frac{1}{2n_{2\bullet}}\right)}{\sqrt{\frac{n_{\bullet 1}}{n}\left(1-\frac{n_{\bullet 1}}{n}\right)\left(\frac{1}{n_{1\bullet}}+\frac{1}{n_{2\bullet}}\right)}}=\frac{\left(\frac{27}{40}-\frac{1}{2\cdot 40}\right)-\left(\frac{5}{15}-\frac{1}{2\cdot 15}\right)}{\sqrt{0{,}5818(1-0{,}5818)\left(\frac{1}{40}+\frac{1}{15}\right)}}=1{,}981.$$

При проверке нулевой гипотезы против двусторонней альтернативы находим

$$\textit{P-значение}=2[1-\Phi\{T(\{n_{ij}\})\}]=2[1-\Phi\{1{,}981\}]=2[1-0{,}9762]=0{,}0476<0{,}05.$$

Таким образом, мы должны сделать заключение, что с уровнем значимости 5 % между первой и второй группами пациентов есть различие, обусловленное различием между типами шприцев.

3.9.А. Задачи и дополнения к разделу 3.8

3.9.80. Пусть объекты изучаемой совокупности (или популяции) W характеризуются некоторым результирующим показателем X. В соответствии с факторным признаком Y вся совокупность разбивается на k групп $W_1,...,W_k$ (или k слоев, или k подпопуляций $W_1,...,W_k$ популяции W), которые соответствуют k значениям (или k уровням) $Y_1,...,Y_k$ фактора Y. Статистическими данными являются наблюденные реализации k выборок $X_{11},...,X_{n_1 1},\ldots,$ $X_{1k},...,X_{n_k k}$ из совокупностей $W_1,...,W_k$ с непрерывными распределениями. Предполагаем, что эти данные $\{X_{ij}\}$, $i=1,...,n_j$, $j=1,...,k,$ представляют собой выборку, полученную в результате n независимых наблюдений над показателем X из k нормальных совокупностей со средними значениями $\mu_1,...,\mu_k$ и с равными, но неизвестными дисперсиями σ^2, то есть модель наблюдений записывается в виде

$$X_{ij}=\mu_j+\varepsilon_{ij},\ i=1,...,n_j,\ j=1,...,k,\ n=(n_1+...+n_k),\ L(\varepsilon_{ij})=N(0;\sigma^2). \quad \textbf{(3.9.81)}$$

В условиях этой модели требуется проверить гипотезу об однородности наблюдаемых данных (или, другими словами, убедиться в том, что изменение фактора Y не влияет на итоговый показатель X). Для гауссовской модели (3.9.81) эта задача на статистическом языке кратко формулируется в виде проверки гипотез

$$H_0:\mu_1=\mu_2=...=\mu_k=\mu,\ H_1:\text{не все }\mu_j\text{ равны},\ j=1,...,k. \quad \textbf{(3.9.82)}$$

Примеры проверки гипотез (3.9.82) с помощью F-критерия Фишера, а также МНК-оценки параметров модели (3.9.81) по конкретным статистическим данным, приводятся в упр. 3.9.94 – 3.9.96. Отметим, что гауссовскую модель

наблюдений (3.9.81) часто удобнее рассматривать в параметризованных вариантах. Используют два варианта параметризации, в зависимости от того, равны или нет объемы выборок $n_1,...,n_k$ из k совокупностей $W_1,...,W_k$. Если объемы выборок $n_1,...,n_k$ различны, то для параметризации модели (3.9.81) определяют средний уровень показателя X по всем совокупностям $W_1,...,W_k$ в виде

$$\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j \,,\ n = (n_1 + ... + n_k)\,. \qquad \textbf{(3.9.83)}$$

Затем вводятся дополнительные параметры $\xi_j = \mu_j - \mu$, которые характеризуют отклонение среднего уровня от j-й обработки, $j = 1,...,k$. Эти параметры называют *дифференциальными эффектами* обработки. Параметризованный вариант модели (3.9.81) записывается в виде

$$X_{ij} = \mu + \xi_j + \varepsilon_{ij}\,,\ i = 1,...,\ n_j\,,\ j = 1,...,k\,,\ n = (n_1 + \cdots + n_k)\,,$$
$$L(\varepsilon_{ij}) = N(0\,;\sigma^2)\,. \qquad \textbf{(3.9.84)}$$

Из (3.9.83) следует, что взвешенная сумма параметров $\xi_1,...\xi_k$ равна нулю, то есть

$$\sum_{j=1}^{k} n_j \xi_j = \sum_{j=1}^{k} n_j (\mu_j - \mu) = \sum_{j=1}^{k} n_j \mu_j - \mu \sum_{j=1}^{k} n_j = 0\,, \qquad \textbf{(3.9.85)}$$

и гипотеза $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ об однородности данных в терминах введенных параметров $\xi_1,...\xi_k$ переписывается в виде

$$H_0 : \xi_1 = \xi_2 = ... = \xi_k = 0\,. \qquad \textbf{(3.9.86)}$$

Отметим, что в модели (3.9.84) имеется число $(k+1)$ параметров, однако общее количество независимых параметров остается равным k, так как взвешенная сумма параметров $\xi_1,...\xi_k$ равна нулю (см. 3.9.85).

Другой вариант параметризации модели (3.9.81) используется, когда объемы выборок $n_1,...,n_k$ из k совокупностей $W_1,...,W_k$ одинаковы. В этом случае средний уровень определяется в виде

$$\mu^* = \frac{1}{k}\sum_{j=1}^{k} \mu_j \qquad \textbf{(3.9.87)}$$

и также вводятся дополнительные параметры $\tau_j = \mu_j - \mu^*$, которые характеризуют отклонение среднего уровня от j-й обработки, $j = 1,...,k$. Параметризованный вариант модели (3.9.81) записывается как

$$X_{ij} = \mu^* + \tau_j + \varepsilon_{ij},\ i = 1,...,n_j,\ j = 1,...,k,\ n = (n_1 + \cdots + n_k),$$
$$L(\varepsilon_{ij}) = N(0;\sigma^2). \qquad \textbf{(3.9.88)}$$

Гипотеза H_0 из (3.9.82) об однородности данных в терминах параметров $\tau_1,...\tau_k$ переписывается в виде $H_0 : \tau_1 = \tau_2 = ... = \tau_k = 0$. Как и в модели (3.9.84), в данной модели имеется число параметров $(k+1)$, однако общее количество независимых параметров остается неизменно равным k, так как $\tau_1 + ... + \tau_k = 0$. Отметим, что если все $n_1,...,n_k$ равны между собой, то оба варианта параметризации модели (3.9.81) идентичны.

(А) Убедитесь, что дифференциальные эффекты *j*-й обработки, $j = 1,...,k$, для моделей (3.9.84) и (3.9.88) связаны между собой соотношениями

$$\xi_j = \tau_j - \frac{1}{n}\sum_{j=1}^{k} n_j \tau_j,\ j = 1,...,k. \qquad \textbf{(3.9.89)}$$

(Б) Убедитесь, что средний квадрат внутри групп $S_W^2 = SS_W/(n-k)$, определенный в (3.8.24), является несмещенной оценкой параметра σ^2 модели (3.9.81), то есть $M(S_W^2) = \sigma^2$.

(В) Убедитесь, что для среднего квадрата между группами $S_B^2 = SS_B/(k=1)$, определенного в (3.8.24), соответственно для моделей (3.9.84) и (3.9.88), справедливы формулы

$$M(S_B^2) = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j \xi_j^2; \qquad \textbf{(3.9.90)}$$

$$M\{S_B^2\} = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j \left(\tau_j - \frac{1}{n}\sum_{j=1}^{k} n_j \tau_j\right)^2. \qquad \textbf{(3.9.91)}$$

Из приведенных формул следует, что средний квадрат между группами $S_B^2 = SS_B/(k-1)$ является несмещенной оценкой параметра σ^2 модели (3.9.81) *только при справедливости гипотезы* H_0, то есть выполняется выражение $M(S_B^2 \mid H_0) = \sigma^2$. Отметим также, что общий средний квадрат $S_T^2 = SS_T/(n-1)$ является наилучшей оценкой параметра σ^2 модели (3.9.81) при справедливости гипотезы H_0, то есть выполняется выражение $M(S_T^2 \mid H_0) = \sigma^2$. Этот факт непосредственно следует из несмещенности подправленной выборочной дисперсии $\tilde{S}^2$, построенной по выборке $X_1,...,X_n$ (см. упр. 2.15.13).

Решение. (А) Отметим, что модели наблюдений (3.9.84) и (3.9.88) связаны между собой соотношениями $\mu_j = \mu + \xi_j = \mu^* + \tau_j$. С учетом этих выражений, формула (3.9.83) запишется в виде

$$\mu = \frac{1}{n}\sum_{j=1}^{k} n_j \mu_j = \frac{1}{n}\sum_{j=1}^{k} n_j (\mu^* + \tau_j) = \mu^* + \frac{1}{n}\sum_{j=1}^{k} n_j \tau_j ,$$

и, следовательно,

$$\xi_j = \mu_j - \mu = (\mu^* + \tau_j) - (\mu^* + n^{-1}\sum n_j \tau_j) = \tau_j - \frac{1}{n}\sum_{j=1}^{k} n_j \tau_j , \ j = 1,...,k .$$

(Б) Для доказательства факта несмещенности оценки S_W^2 в рамках модели (3.9.81) с учетом того, что дисперсии нормальных совокупностей $W_1,...,W_k$ равны, то есть $\sigma_j^2 = \sigma^2$, $\forall j = 1,...,k$, напомним, что выборочная дисперсия S^2, построенная по выборке $X_1,...,X_n$, является смещенной оценкой параметра σ^2 (см. упр. 2.15.13), точнее, ее математическое ожидание равно $M(S^2) = \sigma^2 - \sigma^2/n = \sigma^2(n-1)/n$. Учитывая этот факт, получаем

$$M\{S_W^2\} = M\{SS_W/(n-k)\} = \frac{1}{n-k}\sum_{j=1}^{k} n_j M\left(\frac{1}{n_j}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2\right) =$$

$$= \frac{1}{n-k}\sum_{j=1}^{k} n_j \frac{n_j - 1}{n_j}\sigma_j^2 = \frac{\sigma^2}{n-k}\sum_{j=1}^{k}(n_j - 1) = \sigma^2 .$$

(В) Предполагая, что наблюдения описываются моделью (3.9.84), рассмотрим статистику

$$\tilde{S}_B^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet} - \xi_j)^2 ,$$

которая не зависит от среднего остаточного квадрата S_W^2 и распределена как $\{\sigma^2/(k-1)\}\chi^2(k-1)$ (детали см. в [9, с. 286]). С учетом этого факта получаем, что $M\{\tilde{S}_B^2\} = \{\sigma^2/(k-1)\}\, M\{\chi^2(k-1)\} = \sigma^2$ и, следовательно,

$$(k-1)\sigma^2 = M\{\sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet} - \xi_j)^2\} . \qquad \textbf{(3.9.92)}$$

Учитывая теперь, что

$$(k-1)S_B^2 = \sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 , \ M\{\sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2\} = M\{(k-1)S_B^2\} ,$$

$$M(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) = \xi_j ,$$

из (3.9.92) получаем

$$(k-1)\sigma^2 = M\{\sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet} - \xi_j)^2\} = M\{\sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2\} + \sum_{j=1}^{k} n_j\xi_j^2 -$$

$$-2\sum_{j=1}^{k} n_j\xi_j M(\bar{X}_{\bullet j} - X_{\bullet\bullet}) = M\{\sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2\} - \sum_{j=1}^{k} n_j\xi_j^2 = M\{(k-1)S_B^2\} - \sum_{j=1}^{k} n_j\xi_j^2 .$$

Отсюда следует, что

$$M(S_B^2) = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j\xi_j^2 ,$$

и учитывая (3.9.89), окончательно получаем

$$M\{S_B^2\} = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j\xi_j^2 = \sigma^2 + \frac{1}{k-1}\sum_{j=1}^{k} n_j\left(\tau_j - \frac{1}{n}\sum_{j=1}^{k} n_j\tau_j\right)^2 .$$

Отметим, что если все $\xi_j = 0$ (или все $\tau_j = 0$), то вторые слагаемые в последнем выражении обращаются в ноль, и поэтому $M\{S_B^2 \mid H_0\} = \sigma^2$, то есть S_B^2 – средний квадрат суммы квадратов отклонений SS_B между группами – является несмещенной оценкой дисперсии σ^2 только при справедливости гипотезы H_0 об однородности данных.

Таким образом, несмещенной оценкой параметра σ^2 является средний квадрат $S_W^2 = SS_W/(n-k)$, а оценки $S_B^2 = SS_B/(k-1)$ и $S_T^2 = SS_T/(n-1)$ обладают этим свойством только при справедливости гипотезы H_0 об однородности данных. Отметим, что использование правила сложения (3.8.23) позволяет непосредственно проверить факт несмещенности оценок S_W^2, S_T^2 и S_B^2 параметра σ^2 при гипотезе H_0. В самом деле,

$$M(S_B^2 \mid H_0) = \{M(SS_T \mid H_0) - M(SS_W \mid H_0)\}/(k-1) =$$
$$= \{(n-1)\sigma^2 - (n-k)\sigma^2\}/(k-1) = \sigma^2 .$$

3.9.93. Связь *F*- и *t*-критериев. Пусть число k уровней факторного признака Y (признака классификации) равно двум. Убедитесь, что при этом условии, то есть при $k = 2$, *F*-статистика Фишера, определенная в разд. 3.8, совпадает с квадратом *t*-статистики Стьюдента, определенной формулой (3.6.40).

Решение. Статистика Фишера, определенная в виде отношения двух различных оценок параметра σ^2 гауссовской модели (3.8.3), записывается в виде $F = S_B^2/S_W^2 = \{SS_B/(k-1)\}/\{SS_W/(n-k)\}$, где S_B^2 и S_W^2 определены в (3.8.24) и

S_B^2 называют средним квадратом суммы квадратов SS_B, который вычисляется по наблюдаемым данным $\{X_{ij}\}$, $i=1,...,n_j$, $j=1,...,k$, за итоговым показателем X по формуле

$$S_B^2 = SS_B/(k-1) = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 ,$$

а S_W^2 называют средним квадратом суммы квадратов SS_W, который вычисляется по наблюдаемым данным $\{X_{ij}\}$, $i=1,...,n_j$, $j=1,...,k$, за итоговым показателем X по формуле

$$S_W^2 = SS_W/(n-k) = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 .$$

В этих формулах k – число уровней факторного признака Y, $n=(n_1+...+n_k)$ – общее число наблюдений (объем выборки), $\overline{X}_{\bullet j}$ – групповое выборочное среднее j-й группы, $j=1,...,k$, и $\overline{X}_{\bullet\bullet}$ – общее групповое выборочное среднее, соответственно определенные в (3.8.11) и (3.8.15). При числе уровней $k=2$ факторного признака Y имеем

$$SS_B = n_1(\overline{X}_{\bullet 1} - \overline{X}_{\bullet\bullet})^2 + n_2(\overline{X}_{\bullet 2} - \overline{X}_{\bullet\bullet})^2 = n_1\left[\overline{X}_{\bullet 1} - \frac{n_1\overline{X}_{\bullet 1} + n_2\overline{X}_{\bullet 2}}{n_1+n_2}\right]^2 +$$

$$+n_2\left[\overline{X}_{\bullet 2} - \frac{n_1\overline{X}_{\bullet 1} + n_2\overline{X}_{\bullet 2}}{n_1+n_2}\right]^2 = \frac{n_1n_2}{n_1+n_2}\left(\overline{X}_{\bullet 1} - \overline{X}_{\bullet 2}\right)^2 = \frac{(\overline{X}_{\bullet 1} - \overline{X}_{\bullet 2})^2}{1/n_1 + 1/n_2}.$$

Далее, средний квадрат внутри групп S_W^2 можно привести к виду среднего взвешенного несмещенных оценок групповых дисперсий, то есть к виду

$$S_W^2 = SS_W/(n-k) = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 = \frac{1}{n-k}\sum_{j=1}^{k}(n_j-1)\frac{1}{(n_j-1)}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 =$$

$$= \sum_{j=1}^{k}(n_j-1)\tilde{S}_j^2 \Big/ \sum_{j=1}^{k}(n_j-1) = \frac{(n_1-1)\tilde{S}_1^2 + (n_2-1)\tilde{S}_2^2 + ... + (n_k-1)\tilde{S}_k^2}{n_1+n_2+...+n_k-k},$$

где $\tilde{S}_j^2$ – несмещенная оценка дисперсии σ_j^2, j-й группы, $j=1,...,k$, определяемая в виде

$$\tilde{S}_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 ,\ j=1,...,k .$$

Отсюда при $k=2$ и $n_1=n, n_2=m$ получаем, что внутригрупповой (остаточ-

ный) средний квадрат S_W^2 совпадает с объединенной несмещенной оценкой $\tilde{S}_\Sigma^2$ дисперсии σ^2, построенной по двум выборкам $X_1,...X_n$, $Y_1,...Y_m$, определенной в (3.6.41), то есть

$$S_W^2 = \frac{(n-1)\tilde{S}_1^2 + (m-1)\tilde{S}_2^2}{n+m-2} = \tilde{S}_\Sigma^2 .$$

Учитывая далее, что F-статистика определяется выражением $F=\{SS_B/(k-1)\}/\{SS_W/(n-k)\}$, при $k=2$ имеем $F=\{SS_B\}/\{SS_W/(n_1+n_2-2)\}$, или при $n_1=n$ и $n_2=m$ F-статистика запишется в виде $F=(\overline{X}_{\bullet 1}-\overline{X}_{\bullet 2})^2/\tilde{S}_\Sigma^2(1/n+1/m)$, который совпадает с квадратом t-статистики Стьюдента, определенной формулой (3.6.40).

Отметим, что Ван-дер-Варден [10], используя ссылки на центральную предельную теорему и установленную связь между F- и t-критериями, обосновав робастность F-критерия Фишера при отклонении от гауссовской модели, распространяет свойство робастности к отклонениям от нормальности и для t-критерия Стьюдента (заметим, что это утверждение не согласуется с выводами работ [106, 112]).

Рассмотрим числовой пример. Пусть $L(X)=N(a_1,\sigma^2)$ и $L(Y)=N(a_2,\sigma^2)$. Имеем две независимые выборки $\vec{X}=(X_1,...,X_n)$ и $\vec{Y}=(Y_1,...,Y_m)$ из нормальных совокупностей с равными дисперсиями. Пусть $n=7$ и $m=6$. Наблюденные реализации $\vec{x}=(x_1,...,x_7)$ и $\vec{y}=(y_1,...,y_6)$ этих выборок приведены в таблице.

$\vec{x}$	226,5	224,1	218,6	220,1	228,8	229,6	222,5
$\vec{y}$	221,5	230,2	223,4	224,3	230,8	223,8	----

По этим данным вычислены наблюденные значения статистик: $\overline{x}=224,314$, $\tilde{s}_1^2=17,765$ и $\overline{y}=225,667$, $\tilde{s}_2^2=14,951$. Требуется проверить гипотезу $H_0: a_1-a_2=0$ против двусторонней альтернативы $H_1: a_1-a_2\neq 0$ с помощью t-критерия Стьюдента, основанного на статистике (3.6.40). Согласно (3.6.41), наблюденное значение объединенной выборочной дисперсии

$$\tilde{S}_\Sigma^2 = \frac{\tilde{S}_1^2(n-1)+\tilde{S}_2^2(m-1)}{(n-1)+(m-1)} = \frac{6\cdot 17,765+5\cdot 14,951}{11} = 16,486 .$$

Согласно (3.6.40), наблюденное значение t-статистики вычисляется по формуле

$$T(\vec{x},\vec{y}) = \frac{|\overline{x}-\overline{y}|}{\tilde{s}_\Sigma\sqrt{(1/n+1/m)}} = \frac{|224,314-225,667|}{\sqrt{16,486\cdot(1/7+1/6)}} = 0,599 .$$

Пусть уровень значимости равен $\alpha = 0,05$. В данном примере рассматривается двусторонняя альтернатива, поэтому критическая область статистики $T(\vec{X},\vec{Y})$ определяется выражением (3.6.47а) и ее критическое значение определяется квантилем уровня $(1-\alpha/2)$ распределения Стьюдента с числом степеней свободы $(n+m-2)=11$, то есть $t_{1-\alpha/2,(n+m-2)} = t_{0,975,(11)} = 2,201$. Далее, так как выполняется неравенство $(T(\vec{x},\vec{y}) = 0,599) < (t_{0,975,(11)} = 2,201)$, то гипотезу $H_0 : a_1 = a_2$ принимаем на уровне значимости $\alpha = 0,05$. Для принятия решения также можно вычислить *P-значение* критерия (см. упр. 3.9.74 и 3.9.75). Используя табл. 3.1а функции распределения Стьюдента в [6, с. 174], получаем

$$P\text{-значение } = 2P\{\, T(\vec{X},\vec{Y}) \ge T(\vec{x},\vec{y}) \mid H_0\} = 2P\{St(11) \ge 0,599\,\} = 2\cdot 0,28 = 0,56,$$

что явно свидетельствует в пользу гипотезы $H_0 : a_1 = a_2$.

3.9.94. Пусть объекты изучаемой совокупности W характеризуются итоговым показателем X. В соответствии с четырьмя уровнями Y_1, Y_2, Y_3, Y_4 факторного признака Y совокупность W разбита на четыре нормальные подсовокупности W_1, W_2, W_3, W_4 со средними значениями $\mu_1, \mu_2, \mu_3, \mu_4$, дисперсии которых предполагаются равными неизвестному параметру σ^2. Наблюденные данные $\{x_{ij}\}$, $i = 1,...,4$, $j = 1,..., 4,$ в виде реализаций выборок $X_{1j},...,X_{4j}$, объема $n_j = 4$, $j = 1,..., 4$, из нормальных совокупностей W_1, W_2, W_3, W_4 приведены в табл. 3.9.10.

Таблица 3.9.10

$X \setminus Y$	Y_1	Y_2	Y_3	Y_4
1	140	150	148	150
2	144	149	149	155
3	142	152	146	154
4	145	150	147	152
Сумма Групповое выб. среднее	$X_{\bullet 1} = 571$ $\overline{X}_{\bullet 1} = 142,75$	$X_{\bullet 2} = 601$ $\overline{X}_{\bullet 2} = 150,25$	$X_{\bullet 3} = 590$ $\overline{X}_{\bullet 3} = 147,50$	$X_{\bullet 4} = 611$ $\overline{X}_{\bullet 4} = 152,75$

Предполагаем, что модель наблюдений гауссовская, вида (3.8.3). Требуется по наблюденным данным $\{x_{ij}\}$, $i = 1,...,4$, $j = 1,..., 4,$ проверить гипотезу об однородности совокупностей W_1, W_2, W_3, W_4 в смысле равенства их средних значений, то есть гипотеза будет записываться в виде $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$.

Решение. В данном примере число уровней фактора Y равно $k=4$, объемы выборок из совокупностей W_1, W_2, W_3, W_4 одинаковы, то есть $n_1=n_2=n_3=n_4=4$, общий объем наблюдений $n=n_1+n_2+n_3+n_4=16$. Далее, общая сумма наблюдений

$$X_{\bullet\bullet}=X_{\bullet 1}+X_{\bullet 2}+X_{\bullet 3}+X_{\bullet 4}=571+601+590+611=2373.$$

Межгрупповая сумма квадратов

$$SS_B=\sum_{j=1}^{k}n_j(\bar{X}_{\bullet j}-\bar{X}_{\bullet\bullet})^2=\sum_{j=1}^{k}n_j\bar{X}_{\bullet j}^2+n\bar{X}_{\bullet\bullet}^2-2\bar{X}_{\bullet\bullet}\sum_{j=1}^{k}n_j\bar{X}_{\bullet j}=\sum_{j=1}^{k}X_{\bullet j}^2/n_j-X_{\bullet\bullet}^2/n=$$
$$=(571)^2/4+(601)^2/4+(590)^2/4+(611)^2/4-(2373)^2/16=220{,}19.$$

Общая сумма квадратов

$$SS_T=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet\bullet})^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2-X_{\bullet\bullet}^2/n=$$
$$=(140)^2+(144)^2+\ldots+(152)^2-(2373)^2/16=259{,}46.$$

Остаточная сумма квадратов внутри групп, согласно правилу сложения (3.8.23),

$$SS_W=SS_T-SS_B=259{,}46-220{,}19=39{,}27.$$

Заполненная таблица дисперсионного анализа наблюдений данного примера выглядит следующим образом (табл. 3.9.11).

Таблица 3.9.11

Сумма кв. отклонений	Степени свободы	Средний квадрат	Наблюденное значение F-статистики
Межгрупповая $SS_B=220{,}19$	$(k-1)=3$	$SS_B/(k-1)=74{,}3$	$F=\dfrac{SS_B/(k-1)}{SS_W/(n-k)}=\dfrac{73{,}4}{3{,}27}=22{,}45$
Внутригрупповая $SS_W=SS_T-SS_B=$ $=39{,}27$	$(n-k)=12$	$SS_W/(n-k)=3{,}27$	
Общая $SS_T=259{,}46$	$(n-1)=15$	$SS_T/(n-1)=17{,}30$	

Пусть уровень значимости $\alpha=0{,}05$. По таблицам квантилей F-распределения Фишера [6, с. 208], находим квантиль уровня 0,95 F-распределения $F_{1-\alpha,(k-1;n-k)}=F_{0{,}95,(3,12)}=3{,}49$. Так как выполняется неравенство $(F=22{,}45)>(F_{0{,}95,(3,12)}=3{,}49)$, то гипотезу $H_0:\mu_1=\mu_2=\mu_3=\mu_4$ отклоня-

ем на уровне значимости $\alpha = 0,05$. Используя понятие достигнутого уровня значимости (или *P-значения* критерия) отмечаем, что в данном примере для критерия, основанного на *F*-статистике, получаем *P-значение* $< 0,001$, поэтому гипотезу $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ следует отклонить. Далее, выборочный коэффициент детерминации равен

$$\eta^2 = (SS_B / n)/(SS_T / n) = SS_B / SS_T = 220,19 / 259,46 = 0,849 .$$

Это значение коэффициента детерминации говорит о том, что 84,9 % общей вариации наблюдаемого показателя *X* обусловлено меняющимся фактором *Y*.

В данном примере гипотеза $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ оказалась отвергнутой, поэтому следует выяснить, какие именно из совокупностей W_1, W_2, W_3, W_4 различаются значимо. Для ответа на этот вопрос воспользуемся упорядоченной процедурой парных сравнений Шеффе (см. разд. 3.8.В) в виде последовательной проверки гипотез $H_0^* : \mu_i - \mu_j = 0$, $1 \le i \ne j \le 4$, с использованием алгоритма, описанного в замечании 3.8.48. Согласно этому алгоритму, сначала следует расположить совокупности W_1, W_2, W_3, W_4 в порядке возрастания их групповых выборочных средних. В данном примере получаем $\hat{\mu}_1 < \hat{\mu}_3 < \hat{\mu}_2 < \hat{\mu}_4$. Общее возможное число парных сравнений совокупностей с помощью процедуры Шеффе будет равно $C_4^2 = 6$. Для последовательной серии парных сравнений совокупностей при проверке гипотез $H_0^* : \mu_i - \mu_j = 0$, $i \ne j$, $H_1^* : \mu_i - \mu_j \ne 0$, используем критерий, статистика которого определена в (3.8.49). Эта статистика при $c_i = 1$, $c_j = -1$ и всех остальных из $c_1, ..., c_k$, равных нулю, записывается в виде

$$Z = \{\overline{X}_{\bullet i} - \overline{X}_{\bullet j}\}^2 / \{(k-1) \cdot S_W^2 \cdot (1/n_i + 1/n_j)\},\ 1 \le i \ne j \le k .$$

Учитывая, что в данном примере $\overline{X}_{\bullet 1} = 142,72$, $\overline{X}_{\bullet 4} = 152,75$, $k = 4$, $n_j = 4$, $S_W^2 = 3,27$ и критическое значение $F_{1-\alpha,(k-1;n-k)} = F_{0,95,(3,12)} = 3,49$, при проверке гипотез $H_0^* : \mu_1 - \mu_4 = 0$, $H_1^* : \mu_1 - \mu_4 \ne 0$, статистика Z будет равна

$$Z = \{\overline{X}_{\bullet 1} - \overline{X}_{\bullet 4}\}^2 / \{(k-1) \cdot S_W^2 \cdot (1/n_i + 1/n_j)\} =$$
$$= \{142,75 - 152,75\}^2 / \{3 \cdot 3,27 \cdot (1/4 + 1/4)\} = 20,51 > 3,39 .$$

Из последнего неравенства заключаем, что гипотеза $H_0^* : \mu_1 - \mu_4$ отвергается на уровне значимости $\alpha = 0,05$, то есть $\mu_1 \ne \mu_4$ и, следовательно, совокупности W_1 и W_4 значимо различаются своими средними значениями. Аналогично

реализуется упорядоченная процедура парных сравнений путем последовательной проверки следующих гипотез:

$$H_0^*: \mu_1 - \mu_2, \text{ статистика } Z = 11{,}56 > 3{,}49, \text{ то есть } \mu_1 \neq \mu_2,$$
$$H_0^*: \mu_1 - \mu_3, \text{ статистика } Z = 4{,}65 > 3{,}49, \text{ то есть } \mu_1 \neq \mu_3,$$
$$H_0^*: \mu_3 - \mu_4, \text{ статистика } Z = 5{,}61 > 3{,}49, \text{ то есть } \mu_3 \neq \mu_4,$$
$$H_0^*: \mu_3 - \mu_2, \text{ статистика } Z = 2{,}15 < 3{,}49, \text{ то есть } \mu_3 = \mu_2,$$
$$H_0^*: \mu_2 - \mu_4, \text{ статистика } Z = 1{,}27 < 3{,}49, \text{ то есть } \mu_2 = \mu_4.$$

Таким образом, в данном примере из общего числа $C_4^2 = 6$ пар совокупностей значимо отличаются все совокупности, кроме пар (W_2, W_3) и (W_2, W_4), средние значения которых не отличаются при уровне значимости $\alpha = 0{,}05$.

Отметим, что при решении данного примера предполагалось, что наблюдаемые данные описываются гауссовской моделью (3.8.3), при формулировке которой также предполагается равенство дисперсий совокупностей W_1, W_2, W_3, W_4 – см. (3.8.3г). Для проверки гипотезы $H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ Бартлетт (см. разд. 3.6) предложил модифицированный вариант статистики отношения правдоподобия (3.6.20), в котором n_j заменяется на $n_j - 1$ и оценки максимального правдоподобия $\hat{\sigma}_j^2$, $j = 1,...,k$, и $\hat{\sigma}^2$ заменяются несмещенными оценками $\tilde{S}_j^2$ и S_W^2. В результате такой модификации статистика (3.6.20) запишется в виде

$$B\{\vec{X}_1,...,\vec{X}_k\} = -\sum_{j=1}^{k}(n_j - 1)\ln(\tilde{S}_j^2 / S_W^2) = (n-k)\ln S_W^2 - \{(n_1 - 1)\ln \tilde{S}_1^2 + \cdots + (n_k - 1)\ln \tilde{S}_k^2\},$$

где $\tilde{S}_j^2$ – подправленная выборочная дисперсия *j*-й группы, вычисляемая по формуле

$$\tilde{S}_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2 = \frac{1}{n_j - 1}\left\{\sum_{i=1}^{n_j} X_{ij}^2 - (X_{\bullet j})^2 / n_j\right\}, \; j = 1,...,k,$$

и S_W^2 – остаточный средний квадрат

$$S_W^2 = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_{\bullet j})^2 = (SS_T - SS_B)/(n-k).$$

Для повышения качества аппроксимации распределения статистики $B\{\vec{X}_1,...,\vec{X}_k\}$ при нулевой гипотезе распределением хи-квадрат Бартлетт предложил нормировать эту статистику на коэффициент C^*, вычисляемый по

формуле

$$C^* = 1 + \{1/(n_1 - 1) + \cdots + 1/(n_k - 1) - 1/(n-k)\}/\{3(k-1)\}.$$

С учетом этого замечания статистика критерия Бартлетта записывается в виде

$$B^*\{\vec{X}_1,...,\vec{X}_k\} = B\{\vec{X}_1,...,\vec{X}_k\}/C^*.$$

При больших объемах выборки для нее справедливо выражение $L(B^*(\vec{X}_1,...,\vec{X}_n)\,|\,H_0) = \chi^2(k-1)$. Воспользуемся этим критерием для проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ по данным, приведенным в табл. 3.9.10. Подправленные групповые выборочные дисперсии соответственно равны

$$\tilde{S}_1^2 = \{(140)^2 + (144)^2 + (142)^2 + (145)^2 - (571)^2/4\}/3 = 4{,}92,$$
$$\tilde{S}_2^2 = \{(150)^2 + (149)^2 + (152)^2 + (150)^2 - (601)^2/4\}/3 = 1{,}58,$$
$$\tilde{S}_3^2 = \{(148)^2 + (149)^2 + (146)^2 + (147)^2 - (590)^2/4\}/3 = 1{,}67,$$
$$\tilde{S}_4^2 = \{(150)^2 = (155)^2 + (154)^2 + (152)^2 - (611)^2/4\}/3 = 4{,}92.$$

Далее, для вычисленных оценок групповых дисперсий, с учетом того, что $k = 4$, $n_j = 4$ и $S_W^2 = 3{,}27$, получаем

$$B(\vec{X}_1,...,X_k) = -\sum_{j=1}^{k}(n_j - 1)\ln(\tilde{S}_j^2/S_W^2) =$$
$$= 3\{\ln(3{,}27/4{,}92) + \ln(3{,}27/1{,}58) + \ln(3{,}27/1{,}67) + \ln(3{,}27/4{,}92)\} = 1{,}75$$

и нормировочный коэффициент C^* равен

$$C^* = 1 + \{1/(n_1 - 1) + \cdots + 1/(n_k - 1) - 1/(n-k)\}/\{3(k-1)\} = 1 + \{4/3 - 1/12\}/9 = 1{,}139.$$

Окончательно получаем, что наблюденное значение статистики $B^*\{\vec{X}_1,...,\vec{X}_k\}$ критерия Бартлетта равно

$$B^*\{\vec{x}_1,...,\vec{x}_k\} = B\{\vec{x}_1,...,\vec{x}_k\}/C^* = 1{,}750/1{,}139 = 1{,}54.$$

Для уровня значимости $\alpha = 0{,}05$ критическое значение статистики $B^*\{\vec{X}_1,...,\vec{X}_k\}$ равно квантилю уровня $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы $(k-1)$, то есть $\chi^2_{1-\alpha,(k-1)} = \chi^2_{0{,}95,(3)} = 7{,}81$.

Так как выполняется неравенство $(B\{\vec{x}_1,...,\vec{x}_k\}/C^* = 1{,}54) < (\chi^2_{0{,}95,(3)} = 7{,}81)$, то гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$ не противоречит наблюденным данным и, следовательно, не отклоняется на уровне значимости $\alpha = 0{,}05$. Используя таблицу интеграла вероятностей χ^2 в [6, с.140], находим, что в дан-

ном примере для критерия Бартлетта *P-значение* $=0,67$, что явно свидетельствует в пользу принятия гипотезы о равенстве дисперсий совокупностей W_1, W_2, W_3, W_4.

3.9.95. Пусть объекты изучаемой совокупности W характеризуются итоговым показателем X. В соответствии с четырьмя уровнями Y_1, Y_2, Y_3, Y_4 факторного признака Y совокупность W разбита на четыре нормальные подсовокупности W_1, W_2, W_3, W_4 со средними значениями $\mu_1, \mu_2, \mu_3, \mu_4$, дисперсии которых предполагаются равными неизвестному параметру σ^2. Наблюденные данные $\{x_{ij}\}$, $i=1,...,9$, $j=1,...,4$, в виде реализаций выборок $X_{1j},...,X_{9j}$ объема $n_j=9$, $j=1,...,4$, из нормальных совокупностей W_1, W_2, W_3, W_4 приведены в табл. 3.9.12.

Итак, в рамках гауссовской модели (3.8.3) требуется по наблюденным данным $\{x_{ij}\}$, $i=1,...,9$, $j=1,...,4$, проверить гипотезу $H_0: \mu_1=\mu_2=\mu_3=\mu_4=\mu$ об однородности данных совокупностей W_1, W_2, W_3, W_4 в смысле равенства их средних значений.

Решение. В данном примере число уровней фактора Y равно $k=4$, объемы выборок из совокупностей W_1, W_2, W_3, W_4 одинаковы, то есть $n_1=n_2=n_3=n_4=9$, общий объем наблюдений $n=n_1+n_2+n_3+n_4=36$. Далее, общая сумма наблюдений

$$X_{\bullet\bullet}=X_{\bullet 1}+X_{\bullet 2}+X_{\bullet 3}+X_{\bullet 4}=36,867+41,627+42,729+45,398=166,621$$

и общее групповое среднее $\overline{X}_{\bullet\bullet}=X_{\bullet\bullet}/n=166,621/36=4,628$.

Таблица 3.9.12

$X \setminus Y$	Y_1	Y_2	Y_3	Y_4
1	4,079	4,368	4,169	4,928
2	4,859	5,668	5,709	5,608
3	3,540	3,752	4,416	4,940
4	5,047	5,848	5,666	5,291
5	3,298	3,802	4,123	4,674
6	4,679	4,844	5,059	5,038
7	2,870	3,578	4,403	4,905
8	4,648	5,393	4,496	5,208
9	3,847	4,374	4,688	4,806
Сумма	$X_{\bullet 1}=36,867$	$X_{\bullet 2}=41,627$	$X_{\bullet 3}=42,729$	$X_{\bullet 4}=45,398$
Гр. среднее	$\overline{X}_{\bullet 1}=4,096$	$\overline{X}_{\bullet 2}=4,625$	$\overline{X}_{\bullet 3}=4,748$	$\overline{X}_{\bullet 4}=5,044$

Межгрупповая сумма квадратов

$$SS_B = \sum_{j=1}^{k} n_j (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} n_j \bar{X}_{\bullet j}^2 + n\bar{X}_{\bullet\bullet}^2 - 2\bar{X}_{\bullet\bullet} \sum_{j=1}^{k} n_j \bar{X}_{\bullet j} = \sum_{j=1}^{k} X_{\bullet j}^2 / n_j - X_{\bullet\bullet}^2 / n =$$

$$= (36{,}867)^2/9 + (41{,}627)^2/9 + (42{,}729)^2/9 + (45{,}398)^2/9 - (166{,}621)^2/36 = 4{,}232 .$$

Общая сумма квадратов

$$SS_T = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - X_{\bullet\bullet}^2 / n =$$

$$= (4{,}079)^2 + (4{,}859)^2 + \ldots + (4{,}806)^2 - (166{,}621)^2 / 36 = 18{,}289 .$$

Остаточная сумма квадратов внутри групп, согласно (3.8.12),

$$SS_W = SS_T - SS_B = 18{,}289 - 4{,}232 = 14{,}057 .$$

Для удобства и наглядности вычисления наблюденного значения F-статистики Фишера по исходным данным табл. 3.9.12, заполним таблицу дисперсионного анализа, которая выглядит следующим образом (табл. 3.9.13):

Таблица 3.9.13

Сумма кв. отклонений	Степени свободы	Средний квадрат	Наблюденное значение F-статистики
Межгрупповая $SS_B = 4{,}232$	$(k-1)=3$	$SS_B/(k-1)=1{,}411$	$F = \dfrac{SS_B/(k-1)}{SS_W/(n-k)} = \dfrac{1{,}411}{0{,}439} = 3{,}21$
Внутригрупповая $SS_W = SS_T - SS_B = 14{,}057$	$(n-k)=32$	$SS_W/(n-k)=0{,}439$	
Общая $SS_T = 18{,}289$	$(n-1)=35$	$SS_T/(n-1)=0{,}523$	

Пусть уровень значимости $\alpha = 0{,}05$. По таблицам квантилей F-распределения Фишера (см. [6, с. 208]) находим $F_{1-\alpha,(k-1;n-k)} = F_{0{,}95,(3,32)} = 2{,}90$. Так как $(F = 3{,}21) > (F_{0{,}95,(3,32)} = 2{,}90)$, то гипотезу $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ отклоняем на уровне значимости $\alpha = 0{,}05$. Используя понятие достигнутого уровня значимости (или *P-значения* критерия), отмечаем, что в данном примере для критерия, основанного на F-статистике, получаем *P-значение* $< 0{,}05$, поэтому гипотезу $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ следует отклонить и воспользоваться методом множественных сравнений, чтобы решить, какие именно уровни фактора Y различаются. Выборочный коэффициент детерминации в этом случае

$$\eta^2 = (SS_B/n)/(SS_T/n) = SS_B/SS_T = 4{,}232/18{,}289 = 0{,}231 .$$

Это значение коэффициента детерминации говорит о том, что 23,1 % общей вариации наблюдаемого показателя X обусловлено меняющимся фактором Y.

Оценки средних значений совокупностей W_1, W_2, W_3, W_4

$$\hat{\mu}_1 = \overline{X}_{\bullet 1} = 4{,}096\,,\ \hat{\mu}_2 = \overline{X}_{\bullet 2} = 4{,}625\,,\ \hat{\mu}_3 = \overline{X}_{\bullet 3} = 4{,}748\,,\ \hat{\mu}_4 = \overline{X}_{\bullet 4} = 5{,}044\,.$$

Оценка среднего значения всей совокупности W равна общему групповому выборочному среднему $\hat{\mu} = \overline{X}_{\bullet\bullet} = X_{\bullet\bullet}/n = 166.621/36 = 4{,}628$.

Далее, представив модель наблюдений (3.8.3) в параметризованном виде (3.8.7), получим оценки дифференциальных эффектов каждого уровня факторного признака Y. Эти оценки следующие:

$$\hat{\xi}_1 = \hat{\mu}_1 - \hat{\mu} = 4{,}096 - 4{,}628 = -0{,}532\,,\ \hat{\xi}_2 = \hat{\mu}_2 - \hat{\mu} = 4{,}625 - 4{,}628 = -0{,}003\,,$$
$$\hat{\xi}_3 = \hat{\mu}_3 - \hat{\mu} = 4{,}748 - 4{,}628 = 0{,}120\,,\ \hat{\xi}_4 = \hat{\mu}_4 - \hat{\mu} = 5{,}044 - 4{,}628 = 0{,}416\,.$$

Так как гипотеза $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ отклонена, то для оценки дисперсии σ^2 всей совокупности W следует использовать средний квадрат внутри групп, то есть $S_W^2 = SS_W/(n-k) = 0{,}439$.

В данном примере гипотеза $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ оказалась отвергнутой, поэтому следует выяснить, какие именно из совокупностей W_1, W_2, W_3, W_4 различаются значимо. Для ответа на этот вопрос воспользуемся процедурой упорядоченных парных сравнений Шеффе (см. разд. 3.8.B) в виде последовательной проверки гипотез $H_0^* : \mu_i - \mu_j = 0$, $1 \le i \ne j \le 4$, с использованием алгоритма, описанного в замечании 3.8.48. Согласно этому алгоритму, сначала следует расположить совокупности W_1, W_2, W_3, W_4 в порядке возрастания их групповых выборочных средних. В данном примере получаем $\hat{\mu}_1 < \hat{\mu}_2 < \hat{\mu}_3 < \hat{\mu}_4$. Общее возможное число парных сравнений совокупностей с помощью процедуры Шеффе $C_4^2 = 6$. Для последовательной серии парных сравнений совокупностей при проверке гипотез $H_0^* : \mu_i - \mu_j = 0$, $i \ne j$, $H_1^* : \mu_i - \mu_j \ne 0$ построим доверительные интервалы (3.8.45), которые при $c_i = 1$, $c_j = -1$ и всех остальных из $c_1, \ldots, c_k$, равных нулю, записываются в виде

$$(\overline{X}_{\bullet i} - \overline{X}_{\bullet j}) \pm \{(k-1) \cdot S_W^2 \cdot F_{1-\alpha}(k-1, n-k) \cdot (1/n_i + 1/n_j)\}^{1/2},\ 1 \le i \ne j \le k\,.$$

Пусть $\alpha = 0{,}05$. Учитывая, что $F_{1-\alpha,(k-1;n-k)} = F_{0{,}95,(3,32)} = 2{,}90$, $k-1 = 3$, $S_W^2 = 0{,}439$, $n_j = 9$, $j = 1,2,3,4$, получаем, что 95 %-й доверительный интервал для проверки гипотезы $H_0^* : \mu_1 - \mu_4 = 0$ против альтернативы $H_1^* : \mu_1 - \mu_4 \ne 0$ записывается в виде

$$(\bar{X}_{\bullet 1}-\bar{X}_{\bullet 4})\pm\{(k-1)\cdot S_W^2\cdot F_{1-\alpha}(k-1,n-k)\cdot(1/n_1+1/n_4)\}^{1/2}=$$
$$=(4{,}096-5{,}044)\pm\{3\cdot 0{,}439\cdot 2{,}90\cdot(1/9+1/9)\}^{1/2}=-0{,}948\pm 0{,}921=(-1{,}869\,;-0{,}027).$$

Так как вычисленный интервал $(-1{,}869\ ;\ -0{,}027)$ не накрывает ноль, то гипотеза $H_0^*:\mu_1-\mu_4=0$ отвергается на уровне значимости $\alpha=0{,}05$, то есть значимо выполняется неравенство $\mu_1\neq\mu_4$. Далее, при проверке гипотез $H_0^*:\mu_1-\mu_3=0$, $H_1^*:\mu_1-\mu_3\neq 0$ аналогично получаем 95 %-й доверительный интервал $(-1{,}573\ ;\ 0{,}269)$. Этот интервал накрывает ноль, поэтому гипотеза $H_0^*:\mu_1-\mu_3=0$ принимается на уровне значимости $\alpha=0{,}05$. Отметим также, что при проверке гипотез $H_0^*:\mu_1-\mu_3=0$, $H_1^*:\mu_1-\mu_3\neq 0$ наблюденное значение статистики Z вида (3.8.49) равно

$$Z=\{\bar{X}_{\bullet 1}-\bar{X}_{\bullet 3}\}^2/\{(k-1)\cdot S_W^2\cdot(1/n_1+1/n_3)\}=$$
$$=\{4{,}096-4{,}747\}^2/\{3\cdot 0{,}439\cdot(1/9\ +\ 1/9)\}=1{,}45.$$

Так как выполняется неравенство $(Z=1{,}45)<(F_{0{,}95,\ (3,32)}=2{,}90)$, то гипотеза $H_0^*:\mu_1-\mu_3=0$ сохраняется на уровне значимости $\alpha=0{,}05$. Далее проверяем гипотезу $H_0^*:\mu_2-\mu_4=0$ против альтернативы $H_1^*:\mu_2-\mu_4\neq 0$. В этом случае статистика $Z=0{,}59<2{,}90$, поэтому гипотеза $H_0^*:\mu_2-\mu_4=0$ сохраняется на уровне значимости $\alpha=0{,}05$, то есть μ_2 и μ_4 значимо не различаются. Аналогично, при проверке гипотез $H_0^*:\mu_3-\mu_4=0$, $H_1^*:\mu_3-\mu_4\neq 0$ получаем статистику $Z=0{,}25<2{,}90$, поэтому гипотеза $H_0^*:\mu_3-\mu_4=0$ сохраняется на уровне значимости $\alpha=0{,}05$, то есть μ_3 и μ_4 значимо не различаются.

Подведем итог множественного парного сравнения совокупностей с помощью упорядоченной процедуры Шеффе. Не различаются значимо совокупности W_1,W_2,W_3, W_2,W_3,W_4, W_3,W_4, то есть $\mu_1=\mu_2=\mu_3\neq\mu_4$, $\mu_1\neq\mu_2=\mu_3=\mu_4$ и $\mu_3=\mu_4$. Отсюда заключаем, что при общем доверительном уровне, равном 95 %, значимым является неравенство $\mu_1\neq\mu_4$, то есть средние значения совокупностей W_1 и W_4 значимо различаются.

3.9.96. Статистические данные этого упражнения взяты из примера (2.4.2) в [3, с. 99]. В эксперименте с крысами изучалось влияние 21 лекарства на количество X (в заданных единицах) соляной кислоты, выделяемой в желудке крысы. Каждое лекарство давали определенной случайной выборке крыс, а первая выборка служила для контроля. Следовательно, в данном примере число популяций (или число различных значений факторного признака Y) равно $k=22$, то есть вся популяция крыс W разбита на 22 подпопуляции

$W_1,...,W_{22}$, причем W_j – популяция крыс, получивших j-е лекарство $j=2,...,22$, а W_1 – популяция крыс, не получивших лекарство, из которой извлекалась контрольная выборка. Результаты наблюдений за итоговым показателем X обозначим через $\{x_{ij}\}$, $i=1,...,n_j$, $j=1,...,22$. По этим данным были вычислены реализации групповых выборочных средних $\overline{X}_{\bullet j}$, которые приведены в табл. 3.9.14 в порядке их возрастания с указанием номера выборки и соответствующих объемов выборок n_j, $j=1,...,22$. По приведенным статистическим данным требуется проверить гипотезу в рамках модели (3.8.3) о том, что среднее выделение соляной кислоты одинаково для всех 22 популяций. Полагая, что $x_{1j},...,x_{n_j j}$ – наблюденные реализации выборок $X_{1j},...,X_{n_j j}$ из нормального распределения $N(\mu_j,\sigma_j^2)$, $j=1,...,22$, при условии, что $\sigma_1^2=\sigma_2^2\cdots=\sigma_{22}^2=\sigma^2$, требуется проверить гипотезу $H_0:\mu_1=\mu_2=\cdots=\mu_{22}$ против альтернативы $H_1:$ *не все* μ_j *равны*, $j=1,...,22$.

Таблица 3.9.14

Номер выборки $j=1,...,22$	Объем выборки n_j	Групповое выб. среднее $\hat{\mu}_j=\overline{X}_{\bullet j}$	Номер выборки $j=1,...,22$	Объем выборки n_j	Групповое выб. среднее $\hat{\mu}_j=\overline{X}_{\bullet j}$
7	22	73,73	8	14	333,29
15	25	146,32	11	27	341,30
6	13	147,92	5	32	374,06
14	18	165,61	4	8	412,13
3	8	191,13	1(контр.)	71	417,32
13	17	213,47	21	16	459,81
18	17	224,41	10	19	460,37
9	14	263,86	22	19	477,53
19	14	303,14	17	18	484,61
12	15	313,20	20	18	507,56
2	6	329,83	16	19	566,37

Решение. Вычисления дисперсионного анализа наблюдений данного примера приведены в табл. 3.9.15. Пусть уровень значимости $\alpha=0,05$. По таблицам квантилей F-распределения Фишера (см. [6, с. 208]) находим $F_{1-\alpha,(k-1;n-k)}=F_{0,95,(21,408)}=1,57$. Так как выполняется неравенство

$$(F=6,49)\ >\ (F_{0,95,(21,408)}=1,57),$$

то гипотезу $H_0:\mu_1=\mu_2=\cdots=\mu_{22}$ отклоняем на уровне значимости $\alpha=0,05$.

Используя понятие достигнутого уровня значимости (или *P-значения* критерия) отмечаем, что в данном примере для критерия, основанного на *F*-статистике, получаем *P-значение* $< 0,001$, поэтому гипотезу $H_0: \mu_1 = \mu_2 = \cdots = \mu_{22}$ следует отклонить и воспользоваться методом множественных сравнений, чтобы решить, какие именно уровни фактора Y различаются значимо.

Таким образом, проверяемые лекарства значимо различаются по своему влиянию на секрецию соляной кислоты в желудке крысы.

Оценка дисперсии σ^2 наблюдаемого признака X (где X – количество в заданных единицах соляной кислоты, выделяемой в желудке крысы) всей совокупности крыс W равна среднему квадрату внутри групп, то есть $S_W^2 = SS_W/(n-k) = 55299$. Для среднего μ_1 контрольной группы, согласно (3.8.36), 95 %-й доверительный интервал

$$\overline{X}_{\bullet 1} \pm t_{1-\alpha/2,(n-k)}\sqrt{S_W^2/n_1} = 417,32 \pm t_{0,975,(408)}\sqrt{55299/71} =$$
$$= 417,32 \pm 1,96(27,9) = (362,6\ ;\ 472,0).$$

Далее, согласно (3.8.37), 95 %-й доверительный интервал для разности между средним μ_1 контрольной популяции W_1 и средним μ_7 седьмой популяции W_7

$$(\overline{X}_{\bullet 1} - \overline{X}_{\bullet 7}) \pm t_{1-\alpha/2,(n-k)}\sqrt{S_W^2(1/n_1 - 1/n_7)} =$$
$$= (417,32 - 73,73) \pm 1,96 \cdot \sqrt{55299 \cdot (1/71 + 1/22)} =$$
$$= 343,59 \pm 1,96 \cdot (57,38) = (231,13\ ;\ 456,05).$$

Таблица 3.9.15

Сумма кв. отклонений	Степени свободы	Средний квадрат	Наблюденное значение *F*-статистики
Межгрупповая $SS_B = 7536412$	$(k-1) = 21$	$SS_B/(k-1) = 358877$	$F = \frac{SS_B/(k-1)}{SS_W/(n-k)} = \frac{358877}{55299} = 6,49$
Внутригрупповая $SS_W = SS_T - SS_B = 22561794$	$(n-k) = 408$	$SS_W/(n-k) = 55299$	
Общая $SS_T = 30098206$	$(n-1) = 429$	$SS_T/(n-1) = 70159$	

В данном примере *F*-критерий отклоняет гипотезу $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, и это означает, что имеется, по крайней мере, одна пара совокупностей W_i и W_j $(i \neq j)$, для которой $\mu_i \neq \mu_j$. Но, зная лишь значение *F*-статистики, мы не можем указать номера совокупностей i и j, для которых $\mu_i \neq \mu_j$. Для ответа на

этот вопрос нужно проводить дополнительный анализ, чтобы узнать, средние значения каких совокупностей отличаются значимо. Конечно, можно было бы попытаться произвести серию парных сравнений совокупностей при помощи *t*-критерия Стьюдента (см. разд. 3.6 и 3.8.42). Однако такой подход ошибочен [47, с. 182], так как при заданном уровне значимости α для *t*-критерия Стьюдента *при проведении серий парных сравнений вероятность отклонить верную гипотезу становится значительно больше, чем* α, то есть не обеспечивается заданный уровень значимости. Для решения возникающей проблемы разработаны различные процедуры (см., например, критерий Дункана в упр. 3.9.100). Далее, метод множественного сравнения Шеффе дает возможность провести серию парных сравнений так, что вероятность хотя бы одного неверного заключения в точности равна α. Воспользуемся процедурой упорядоченных парных сравнений Шефе (см. разд. 3.8.B) в виде последовательной проверки гипотез $H_0^*: \mu_i - \mu_j = 0$, $1 \le i \ne j \le 4$, с использованием алгоритма, описанного в замечании 3.8.48. Согласно этому алгоритму, сначала следует расположить совокупности $W_1,...,W_{22}$ в порядке возрастания их групповых выборочных средних. В данном примере получаем

$$\hat{\mu}_7 < \hat{\mu}_{15} < \hat{\mu}_6 < \hat{\mu}_{14} < \hat{\mu}_3 < \hat{\mu}_{13} < \hat{\mu}_{18} < \hat{\mu}_9 < \hat{\mu}_{19} < \hat{\mu}_{12} < \hat{\mu}_2 < \hat{\mu}_8 < \hat{\mu}_{11} < \hat{\mu}_5 < \hat{\mu}_4 < \\ < \hat{\mu}_1 < \hat{\mu}_{21} < \hat{\mu}_{10} < \hat{\mu}_{22} < \hat{\mu}_{17} < \hat{\mu}_{20} < \hat{\mu}_{16}. \qquad \textbf{(3.9.97)}$$

Общее возможное число парных сравнений совокупностей с помощью процедуры Шеффе будет равно $C_{22}^2 = 22!/(2! \cdot 20!) = 231$. Для последовательной серии парных сравнений совокупностей при проверке гипотез $H_0^*: \mu_i - \mu_j = 0$, $i \ne j$, $H_1^*: \mu_i - \mu_j \ne 0$, используем критерий, статистика которого определена в (3.8.49). Эта статистика при $c_i = 1$, $c_j = -1$ и всех остальных из $c_1,...,c_k$, равных нулю, записывается в виде

$$Z = \{\overline{X}_{\bullet i} - \overline{X}_{\bullet j}\}^2 / \{(k-1) \cdot S_W^2 \cdot (1/n_i + 1/n_j)\},\ 1 \le i \ne j \le k. \qquad \textbf{(3.9.98)}$$

Эквивалентная процедура для проверки гипотез $H_0^*: \mu_i - \mu_j = 0$, $i \ne j$, $H_1^*: \mu_i - \mu_j \ne 0$ основана на построении доверительных интервалов вида

$$(\overline{X}_{\bullet i} - \overline{X}_{\bullet j}) \pm \{(k-1) \cdot S_W^2 \cdot F_{1-\alpha}(k-1, n-k) \cdot (1/n_i + 1/n_j)\}^{1/2},\ 1 \le i \ne j \le k. \qquad \textbf{(3.9.99)}$$

Если ноль накрывается этим интервалом, то нет оснований возражать против гипотезы H_0^*, в противном случае гипотеза H_0^* отвергается на уровне значимости α.

При проведении серии парных сравнений с помощью упорядоченной процедуры Шеффе (см замечание 3.8.48) поступаем следующим образом. По-

сле того как все опыты (или рассматриваемые совокупности $W_1,...,W_{22}$) расположены в порядке возрастания их выборочных групповых средних (см. (3.9.97)), проводится серия парных сравнений совокупностей в определенном порядке. Сначала проводится парное сравнение совокупности, обладающей наименьшим выборочным средним, с совокупностью, обладающей наибольшим выборочным средним. Затем совокупность с наименьшим выборочным средним сравнивается со всеми последующими совокупностями в порядке убывания их выборочных средних, и этот процесс продолжается до принятия гипотезы $H_0^*: \mu_i - \mu_j = 0$, $i \neq j$. Далее повторяется описанная процедура для совокупности со вторым по величине выборочным средним, то есть эта совокупность сравнивается со всеми последующими совокупностями в порядке убывания их выборочных средних, и опять этот процесс продолжается до принятия гипотезы $H_0^*: \mu_i - \mu_j = 0$, $i \neq j$. Затем повторяется этот процесс для совокупности, имеющей третье по величине выборочное среднее, и т.д. Для данного примера этот алгоритм реализуется в виде следующей последовательности парных сравнений методом Шеффе. С учетом (3.9.97) процесс начинается с проверки гипотезы $H_0^*: \mu_7 - \mu_{16} = 0$ против альтернативы $H_1^*: \mu_7 - \mu_{16} \neq 0$. Согласно (3.9.99), вычисление $(1-\alpha)$ %-го доверительного интервала при $\alpha = 0{,}05$ с учетом того, что $k-1=21$, $S_W^2 = 55299$ и $F_{1-\alpha}(k-1, n-k) = F_{0,95}(21, 408) = 1{,}57$, дает следующий результат:

$$(\bar{X}_{\bullet 7} - \bar{X}_{\bullet 16}) \pm \{(k-1)\cdot S_W^2 \cdot F_{1-\alpha}(k-1,\ n-k)\cdot(1/n_7 + 1/n_{16})\}^{1/2} =$$
$$= (73{,}73 - 566{,}37) \pm \{21\cdot 55299\cdot 1{,}57\cdot(1/22 + 1/19)\}^{1/2} =$$
$$= -492{,}64 \pm 422{,}70 = (-915{,}34\ ;\ -69{,}94).$$

Данный интервал не накрывает ноль, поэтому гипотеза $H_0^*: \mu_7 - \mu_{16} = 0$ отвергается, то есть средние значения совокупностей W_7 и W_{16} значимо различаются. Далее проверяется гипотеза $H_0^*: \mu_7 - \mu_{20} = 0$ против альтернативы $H_1^*: \mu_7 - \mu_{20} \neq 0$. Аналогично из (3.9.99) получаем 95 %-й доверительный интервал в виде

$$(\bar{X}_{\bullet 7} - \bar{X}_{\bullet 20}) \pm \{(k-1)\cdot S_W^2 \cdot F_{1-\alpha}(k-1,\ n-k)\cdot(1/n_7 + 1/n_{20})\}^{1/2} =$$
$$= (73{,}73 - 507{,}56) \pm \{21\cdot 55299\cdot 1{,}57\cdot(1/22 + 1/18)\}^{1/2} =$$
$$= -433{,}83 \pm 429{,}12 = (-862{,}95\ ;\ -4{,}71).$$

Данный интервал также не накрывает ноль, поэтому гипотеза $H_0^*: \mu_7 - \mu_{20} = 0$ отвергается. Продолжая сравнивать совокупность W_7 последовательно с совокупностями $W_{17}, W_{22}, W_{10}, W_{21}, W_1$, обнаруживаем, что гипотеза об однород-

ности их средних значений также отвергается, а начиная с совокупности W_4, и для всех последующих с меньшим групповыми выборочными средними, гипотеза об однородности их средних значений принимается. Например, при проверке гипотез $H_0^*: \mu_7 - \mu_1 = 0$, $H_1^*: \mu_7 - \mu_1 \neq 0$ получаем 95 %-й доверительный интервал в виде

$$(\bar{X}_{\bullet 7} - \bar{X}_{\bullet 1}) \pm \{(k-1) \cdot S_W^2 \cdot F_{1-\alpha}(k-1,\ n-k) \cdot (1/n_7 + 1/n_1)\}^{1/2} =$$
$$= (73{,}73 - 417{,}32) \pm \{21 \cdot 55299 \cdot 1{,}57 \cdot (1/22 + 1/71)\}^{1/2} = (-676{,}2\ ;\ -11{,}0).$$

Этот интервал не включает ноль, поэтому гипотезу $H_0^*: \mu_7 - \mu_1 = 0$ отвергаем на уровне значимости $\alpha = 0{,}05$.

Далее, при проверке гипотез $H_0^*: \mu_7 - \mu_4 = 0$, $H_1^*: \mu_7 - \mu_4 \neq 0$, 95 %-й доверительный интервал записывается в виде

$$(\bar{X}_{\bullet 7} - \bar{X}_{\bullet 4}) \pm \{(k-1) \cdot S_W^2 \cdot F_{1-\alpha}(k-1,\ n-k) \cdot (1/n_7 + 1/n_4)\}^{1/2} =$$
$$= (73{,}73 - 412{,}13) \pm \{21 \cdot 55299 \cdot 1{,}57 \cdot (1/22 + 1/8)\}^{1/2} =$$
$$= -388{,}40 \pm 566{,}73 = (-955{,}13\ ;\ 178{,}33).$$

Этот интервал накрывает ноль, следовательно, гипотеза $H_0^*: \mu_7 - \mu_4 = 0$ принимается на уровне значимости $\alpha = 0{,}05$. В качестве примера приведем также проверку гипотез $H_0^*: \mu_7 - \mu_5 = 0$, $H_1^*: \mu_7 - \mu_5 \neq 0$. В данном случае 95 %-й доверительный интервал записывается в виде

$$(\bar{X}_{\bullet 7} - \bar{X}_{\bullet 5}) \pm \{(k-1) \cdot S_W^2 \cdot F_{1-\alpha}(k-1,\ n-k) \cdot (1/n_7 + 1/n_5)\}^{1/2} =$$
$$= (73{,}73 - 374{,}06) \pm \{21 \cdot 55299 \cdot 1{,}57 \cdot (1/22 + 1/32)\}^{1/2} =$$
$$= -300{,}33 \pm 377{,}52 = (-677{,}85\ ;\ 77{,}19).$$

Так как этот интервал включает ноль, то гипотезу $H_0^*: \mu_7 - \mu_5 = 0$ принимаем на уровне значимости $\alpha = 0{,}05$.

В качестве другого примера приведем результат проверки гипотез $H_0^*: \mu_7 - \mu_{15} = 0$, $H_1^*: \mu_7 - \mu_{15} \neq 0$, В этом случае получаем 95 %-й доверительный интервал $(-467{,}41\ ;\ 322{,}23)$, который накрывает ноль, поэтому гипотеза $H_0^*: \mu_7 - \mu_{15} = 0$ принимается на уровне значимости $\alpha = 0{,}05$. Согласно описанному выше алгоритму, далее проверяем гипотезы $H_0^*: \mu_{15} - \mu_{16} = 0$, $H_0^*: \mu_{15} - \mu_{20} = 0$ и т.д.

Множественные сравнения 22 выборочных групповых средних секреции соляной кислоты в желудке крысы, представленных в табл. 3.9.14, проведенные с помощью описанной процедуры парных сравнений Шеффе при уровне

значимости $\alpha = 0{,}05$, приведены в табл. 3.9.16. В этой таблице жирным шрифтом выделены множества номеров опытов (или номеров популяций $W_1,...,W_{22}$), для которых гипотезы $H_0^*: \mu_i - \mu_j = 0$, $i \neq j$, не отвергаются на уровне значимости $\alpha = 0{,}05$, то есть для этих номеров популяций $W_1,...,W_{22}$ результаты эксперимента не противоречат предположению об однородности средних значений выделенных популяций. Например, из первой строки табл. 3.9.16 следует, что гипотеза

$$H_0: \mu_7 = \mu_{15} = \mu_6 = \mu_{14} = \mu_3 = \mu_{13} = \mu_{18} = \mu_9 = \mu_{19} = \mu_{12} = \mu_2 = \mu_8 = \mu_{11} = \mu_5 = \mu_4$$

приемлема на уровне значимости $\alpha = 0{,}05$.

Итак, внутри каждого из выделенных подмножеств популяций ни одна из пар средних не различается значимо, но средние значения из разных подмножеств могут различаться значимо.

Таблица 3.9.16

Номер опыта

7	**15**	**6**	**14**	**3**	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	1	21	10	22	17	20	16
7	**15**	**6**	**14**	**3**	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	21	10	22	17	20	16
7	15	**6**	**14**	**3**	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	**21**	**10**	**22**	17	20	16
7	15	6	**14**	**3**	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	**21**	**10**	**22**	**17**	20	16
7	15	6	14	**3**	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	**21**	**10**	**22**	**17**	20	16
7	15	6	14	3	**13**	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	**21**	**10**	**22**	**17**	**20**	16
7	15	6	14	3	13	**18**	**9**	**19**	**12**	**2**	**8**	**11**	**5**	**4**	**1**	**21**	**10**	**22**	**17**	**20**	**16**

3.9.100. Критерий Дункана. Если F-критерий отклоняет гипотезу $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$, то это означает, что имеется, по крайней мере, одна пара совокупностей W_i и W_j $(i \neq j)$, для которой $\mu_i \neq \mu_j$. Но, зная лишь значение F-статистики, мы не можем указать номера совокупностей i и j, для которых $\mu_i \neq \mu_j$. Для ответа на этот вопрос нужно проводить дополнительный анализ, чтобы узнать, средние значения каких совокупностей отличаются значимо. Для проверки гипотезы $H_0: \mu_{j'} = \mu_{j''}$, $j', j'' = 1,...,k$, $j' \neq j''$, о равенстве двух, произвольно выбранных, средних значений $\mu_{j'}$ и $\mu_{j''}$ используют критерий Дункана, который основан на статистике вида

$$D = \frac{|\hat{\mu}_{j'} - \hat{\mu}_{j''}|}{\left(\dfrac{[(1/n_{j'}) + (1/n_{j''})]}{(n-k)} \displaystyle\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_{\bullet j})^2 \right)^{1/2}}.$$

Гипотеза $H_0 : \mu_{j'} = \mu_{j''}$, $j', j'' = 1,...,k$, $j' \neq j''$, отвергается, если наблюденное значение этой статистики превышает ее критическое значение. Критические значения находятся по табл. 9 в работе [44, с. 159].

3.9.101. Случайные эффекты, модель II. Пусть признак X характеризует качество объектов, которые хранятся на складе предприятия в большом количестве контейнеров. С целью определения качества всех объектов, хранящихся на складе, случайно выбирается заданное число k контейнеров и в каждом j-м контейнере, $j = 1,...,k$, берется выборка $X_{1j},...,X_{n_j j}$ объема n_j, $j = 1,...,k$. Математической моделью наблюдений над признаком X в условиях данного эксперимента является так называемая модель II дисперсионного анализа со случайными эффектами (*или модель компонентного анализа*). Эта модель формулируется следующим образом. Предполагается, что k групп (подпопуляций) случайно выбираются из бесконечной совокупности всех возможных подпопуляций. Каждой группе присваивается номер от 1 до k, и j-я группа считается соответствующей j-му уровню фактора Y, который характеризует свойства рассматриваемых подпопуляций. Далее, из каждой группы случайно выбирается n_j объектов и рассматриваются значения $x_{1j},...,x_{n_j j}$, $j = 1,...,k$, итогового показателя X. Предполагается, что наблюдаемые данные $X_{1j},...,X_{n_j j}$ распределены нормально со средним значением a_j и дисперсией σ^2, не зависящей от уровня j фактора Y, $j = 1,...,k$, то есть для i-го элемента выборки в группе с номером j выполняется выражение $L(X_{ij} \mid j) = N(a_j, \sigma^2)$. Также предполагается, что $a_1,...,a_k$ представляют случайную выборку из совокупности объектов, признак X которых имеет нормальное распределение со средним значением μ и дисперсией σ_a^2, то есть справедливо выражение $L(a_j) = N(\mu, \sigma_a^2)$, $j = 1,...,k$. Для удобства параметризации этой модели, так же как и для модели I вида (3.8.3), определяют j-й дифференциальный (главный) эффект фактора Y равенством $\alpha_j = a_j - \mu$. Отметим, что в отличие от *постоянных* дифференциальных эффектов $\xi_j = \mu_j - \mu$ в модели I с фиксированными эффектами (см. (3.8.7)), эффекты $\alpha_j = a_j - \mu$, $j = 1,...,k$, рассматриваемой модели II, являются *случайными величинами*, распределенными нормально с нулевым средним и дисперсией σ_a^2, то есть справедливы выражения $L(\alpha_j) = N(0, \sigma_a^2)$, $j = 1,...,k$.

Итак, *модель однофакторного дисперсионного анализа со случайными эффектами* (коротко, модель II) описывается выражением (3.8.50), согласно

которому наблюдения записываются в виде

$$X_{ij} = a_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \qquad \textbf{(3.9.102)}$$

где все α_j и ε_{ij} – взаимно независимые случайные величины с нормальными распределениями, то есть выполняются выражения $L(\alpha_j) = N(0, \sigma_a^2)$ и $L(\varepsilon_{ij}) = N(0, \sigma^2)$, $i = 1,...,n_j$, $j = 1,...,k$.

Убедимся, что в рамках модели II однофакторного дисперсионного анализа со случайными эффектами, для которой статистическая модель наблюдений описывается выражением (3.8.50), математическое ожидание $M(S_B^2)$ среднего квадрата между группами $S_B^2 = SS_B/(k-1)$, где SS_B определено в (3.8.20) и имеет вид

$$SS_B = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \sum_{j=1}^{k} X_{\bullet j}^2/n_j - X_{\bullet\bullet}^2/n,$$

вычисляется по формуле

$$M(S_B^2) = \frac{1}{k-1}\sum_{j=1}^{k} n_j M(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2 = \frac{1}{k-1}\left(n - \frac{1}{n}\sum_{j=1}^{k} n_j^2\right)\sigma_a^2 + \sigma^2.$$

Также убедимся, что в частном случае, когда объемы выборок в группах равны, то есть при $n_1 + n_2 + ... + n_k = m$ и $n = km$, эта формула принимает более простой вид

$$M(S_B^2) = \frac{1}{k-1}\left(km - \frac{km^2}{km}\right)\sigma_a^2 + \sigma^2 = m\sigma_a^2 + \sigma^2.$$

Решение. Используя выражение (3.9.102), запишем формулу для общего выборочного среднего $\bar{X}_{\bullet\bullet}$ в виде

$$\bar{X}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mu + \alpha_j + \varepsilon_{ij}) = \mu + \bar{\alpha}_\bullet + \bar{\varepsilon}_{\bullet\bullet},$$

где приняты следующие обозначения:

$$\bar{\varepsilon}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\varepsilon_{ij} = \frac{1}{n}\sum_{j=1}^{k} n_j\bar{\varepsilon}_{\bullet j},\ \bar{\varepsilon}_{\bullet j} = (\varepsilon_{1j} + ... + \varepsilon_{n_j j})/n_j,\ \bar{\alpha}_\bullet = \frac{1}{n}\sum_{j=1}^{k} n_j\alpha_j.$$

Далее, групповое выборочное среднее j-й группы $\bar{X}_{\bullet j}$ запишется в виде

$$\bar{X}_{\bullet j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = \frac{1}{n_j}\sum_{i=1}^{n_j}(\mu + \alpha_j + \varepsilon_{ij}) = \mu + \alpha_j + \bar{\varepsilon}_{\bullet j},\ j = 1,...,k,$$

где $\overline{\varepsilon}_{\bullet j} = (\varepsilon_{1j} + ... + \varepsilon_{n_j j})/n_j$, $j = 1,...,k$. С учетом этих выражений разность $\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet}$ запишется в виде

$$\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet} = (\alpha_j - \overline{\alpha}_\bullet) + (\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet}),\ j = 1,...,k. \qquad \textbf{(3.9.103)}$$

Далее, так как все α_j и ε_{ij} – взаимно независимые случайные величины, то $M\{(\alpha_j - \overline{\alpha}_\bullet)(\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet})\} = 0$, поэтому из (3.9.103) имеем

$$M(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 = M(\alpha_j - \overline{\alpha}_\bullet)^2 + M(\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet})^2. \qquad \textbf{(3.9.104)}$$

Рассмотрим теперь разности

$$\alpha_j - \overline{\alpha}_\bullet = \alpha_j - \frac{1}{n}\sum_{j=1}^{k} n_j \alpha_j = (1 - n_j/n)\alpha_j - \frac{1}{n}\sum_{l \neq j} n_l \alpha_l,$$

$$\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet} = \overline{\varepsilon}_{\bullet j} - \frac{1}{n}\sum_{j=1}^{k} n_j \overline{\varepsilon}_{\bullet j} = (1 - n_j/n)\overline{\varepsilon}_{\bullet j} - \frac{1}{n}\sum_{l \neq j} n_l \overline{\varepsilon}_{\bullet l}.$$

В этих формулах слагаемые независимы, причем $M(\alpha_j) = 0$, $M(\overline{\varepsilon}_{\bullet j}) = 0$ и $M(\alpha_j^2) = \sigma_a^2$, $M[(\overline{\varepsilon}_{\bullet j})^2] = \sigma^2/n_j$, $j = 1,...,k$. Поэтому

$$M(\alpha_j - \overline{\alpha}_\bullet)^2 = \left[\left(1 - \frac{n_j}{n}\right)^2 + \frac{1}{n^2}\sum_{l \neq j} n_l^2\right] \cdot \sigma_a^2 = \left[1 - \frac{2n_j}{n} + \frac{1}{n^2}\sum_{j=1}^{k} n_j\right] \cdot \sigma_a^2,$$

$$M(\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet})^2 = \left(1 - \frac{n_j}{n}\right)^2 \cdot \frac{\sigma^2}{n_j} + \frac{1}{n^2}\sum_{l \neq j} (\sigma^2/n_l) = (1/n_j + 1/n) \cdot \sigma^2.$$

Таким образом, используя данные выражения и формулу (3.9.104), убеждаемся, что математическое ожидание среднего квадрата между группами $S_B^2 = SS_B/(k-1)$ вычисляется по формуле

$$M(S_B^2) = \frac{1}{k-1}\sum_{j=1}^{k} n_j M(\overline{X}_{\bullet j} - \overline{X}_{\bullet\bullet})^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j [M(\alpha_j - \overline{\alpha}_\bullet)^2 + M(\overline{\varepsilon}_{\bullet j} - \overline{\varepsilon}_{\bullet\bullet})^2] =$$

$$= \frac{1}{k-1}\sum_{j=1}^{k} n_j \left[\left(1 - \frac{2n_j}{n} + \frac{1}{n^2}\sum_{j=1}^{k} n_j^2\right) \cdot \sigma^2 + \left(\frac{1}{n_j} - \frac{1}{n}\right) \cdot \sigma^2\right] =$$

$$= \frac{1}{k-1}\left[\sigma_a^2 \sum_{j=1}^{k} n_j - \frac{2\sigma_a^2}{n}\sum_{j=1}^{k} n_j^2 + \frac{n\sigma_a^2}{n^2}\sum_{j=1}^{k} n_j^2 + (k-1) \cdot \sigma^2\right] = \frac{1}{k-1}\left(n - \frac{1}{n}\sum_{j=1}^{k} n_j^2\right) \cdot \sigma_a^2 + \sigma^2.$$

Отсюда, при $n_1 = n_2 = ... = n_k = m$ и $n = km$ имеем

$$M(S_B^2) = \frac{1}{k-1}\left(km - \frac{km^2}{km} \right) \cdot \sigma_a^2 + \sigma^2 = m\sigma_a^2 + \sigma^2 .$$

3.9.105. Оптимальное число наблюдений для модели II дисперсионного анализа. Пусть, как и в предыдущем упражнении, признак X характеризует качество объектов, которые хранятся на складе предприятия в большом количестве контейнеров. С целью определения качества всех объектов, хранящихся на складе, случайно выбирается заданное число k контейнеров и в каждом j-м контейнере, $j = 1,...,k$, берется выборка $X_{1j},...,X_{n_j j}$ объема n_j, $j = 1,...,k$. Математической моделью наблюдений над признаком X в условиях данного эксперимента является модель II дисперсионного анализа со случайными эффектами, которая определена в (3.9.102). Пусть далее стоимость наблюдений, связанных со случайным выбором контейнера, равна c_1, а стоимость каждой пробы в выбранном контейнере равна c_2. Таким образом, получение m проб $n_1 = n_2 = ... = n_k = m$ в каждом контейнере будет стоить $c_1 + mc_2$, а если наблюдается k контейнеров, то общая стоимость такого эксперимента будет равна $k(c_1 + mc_2)$. Предположим, что средства, отпущенные на проведение эксперимента, ограничены и равны величине C. Требуется выбрать такие m и k, удовлетворяющие этому ограничению C, которые бы минимизировали дисперсию оцененного среднего значения признака X, то есть нужно минимизировать дисперсию $D(\overline{X}_{\bullet\bullet}) = (\sigma^2 + m\sigma_a^2)/km$ (см. формулу (3.8.73)) при условии, что m и k удовлетворяют условию $k(c_1 + mc_2) = C$.

Решение. Учитывая, что $k(c_1 + mc_2) = C$, следовательно, $kmc_2 = C - kc_1$, перепишем формулу для дисперсии $\overline{X}_{\bullet\bullet}$ в виде

$$D(\overline{X}_{\bullet\bullet}) = (\sigma^2 + m\sigma_a^2)/km = \sigma_a^2/k + \sigma^2 c_2/(C - kc_1) .$$

Дифференцируя по k и приравнивая результат нулю, получим

$$\partial D(\overline{X}_{\bullet\bullet})/\partial k = -\sigma_a^2/k^2 + \sigma^2 c_1 c_2/(C - c_1 k)^2 = 0 .$$

Отсюда, с учетом равенства $kmc_2 = C - kc_1$, следует оптимальное значение m в виде

$$m_0 = \frac{\sigma}{\sigma_a}\sqrt{\frac{c_1}{c_2}} . \; m_0 = (\sigma/\sigma_a) \cdot (c_1/c_2)^{1/2} .$$

Подставив найденное значение m_0 в $k(c_1 + mc_2) = C$, получим $k_0 = C/(c_1 + m_0 c_2)$. Отметим, что вычисленные значения k_0 и m_0 могут оказаться дробными ве-

личинами. В таких случаях следует использовать наибольшую целую часть вычисленных k_0 и m_0. Отметим также, что для применения полученных формул нужно предварительно (в пробных опытах) получить оценки параметров σ, σ_a (см. формулы (3.8.60) и (3.8.62)) и задать стоимости c_1, c_2 проведения опытов.

3.9.106. Проверяется нулевая гипотеза $H_0 : \sigma_a^2 = 0$ с помощью *F*-критерия, определенного в (3.8.59а), по статистическим данным, которые описываются моделью II однофакторного дисперсионного анализа со случайными эффектами. Для этой модели наблюдения записываются в виде $X_{ij} = a_j + \varepsilon_{ij} = \mu + \alpha_j + \varepsilon_{ij}$, где все α_j и ε_{ij} – взаимно независимые случайные величины с нормальными распределениями, то есть для них выполняются выражения $L(\alpha_j) = N(0, \sigma_a^2)$ и $L(\varepsilon_{ij}) = N(0, \sigma^2)$, $i = 1,...,n_j$, $j = 1,...,k$. Требуется вычислить мощность *F*-критерия при различных значениях отношения σ_a^2 / σ^2.

Решение. Напомним, что при проверке нулевой гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ с помощью *F*-критерия (см. упр. 3.9.61) функция мощности от отношения дисперсий $\tau = \sigma_1^2 / \sigma_2^2$ записывается в виде

$$W(\tau) = P\{F(\nu_1, \nu_2) > \tau^{-1} \cdot F_{1-\alpha}(\nu_1, \nu_2)\},$$

где $F(\nu_1, \nu_2)$ – случайная величина, имеющая *F*-распределения Фишера с числами степеней свободы ν_1 и ν_2, а $F_{1-\alpha}(\nu_1, \nu_2)$ – квантиль уровня $(1-\alpha)$ этого распределения.

Используем этот результат для вычисления мощности *F*-критерия в данном случае. В условиях модели II мы имеем дело со статистиками S_B^2 и S_W^2, для которых математические ожидания (см. упр. 3.9.101) вычисляются по формулам $M(S_B^2) = m\sigma_a^2 + \sigma^2$, $M(S_W^2) = \sigma^2$. Пусть отношение этих ожидаемых значений равно $\tau = (m\sigma_a^2 + \sigma^2)/\sigma^2 = 1 + m \cdot (\sigma_a^2 / \sigma^2)$. Тогда формула для мощности *F*-критерия в условиях модели II запишется в виде

$$W\{(\sigma_a^2 / \sigma^2)\} = P\{F(\nu_1, \nu_2) > [1 + m(\sigma_a^2 / \sigma^2)]^{-1} \cdot F_{1-\alpha}(\nu_1, \nu_2)\}.$$

Для иллюстрации численного расчета пусть параметры, характеризующие условия проведения опыта, следующие: число случайно выбираемых групп $k = 22$ и в каждой группе объем выборки $m = 4$, а общее число наблюдений $n = km = 88$. Для заданных значений параметров числа степеней свободы *F*-распределения Фишера равны $\nu_1 = k - 1 = 21$ и $\nu_2 = n - k = 88 - 22 = 66$. Далее, выберем уровень значимости $\alpha = 0{,}05$ и рассмотрим конкретные значения отношения σ_a^2 / σ^2.

а) Пусть $\sigma_a^2/\sigma^2=1/4$ и $\tau=1+m\cdot(\sigma_a^2/\sigma^2)=2$, а $F_{1-\alpha}(\nu_1,\nu_2)=F_{0,95}(21,66)=1,72$. Тогда мощность F-критерия будет приближенно равна 0,5 , то есть

$$W(\tau=1/4)=P\{\ F(21,66)>0,5\cdot 1,72\ \}=P\{\ F(21,66)>0,86\ \}=$$
$$=P\{\ F(66,21)<1,163\ \}\approx 0,5,$$

так как $P\{\ F(66,21)<1,02\ \}=0,5$.

б) Пусть $\sigma_a^2/\sigma^2=3/4$, тогда $\tau=1+m\cdot(\sigma_a^2/\sigma^2)=4$, $F_{0,95}(21,66)=1,72$ и

$$W(\tau=3/4)=P\{\ F(21,66)>0,25\cdot 1,72\ \}=P\{\ F(21,66)>0,43\ \}=$$
$$P\{\ F(66,21)<2,326\ \}>0,975,$$

так как $P\{\ F(66,21)<2,173\ \}=0,975$.

3.9.107. Двухфакторные повторяемые планы с фиксированными и случайными эффектами, смешанные модели. В данном упражнении, как и в разд. 3.8, чтобы удобнее было различать варианты фиксированных и случайных эффектов, будем обозначать греческими буквами α_j, β_i *фиксированные* главные эффекты факторов A и B, а *случайные* эффекты этих факторов будем обозначать латинскими буквами a_j, b_i, $j=1,...,k$, $i=1,...,r$.

Фиксированные эффекты. Рассмотрим сначала случай, когда оба фактора A (с k уровнями) и B (с r уровнями) характеризуются фиксированными эффектами. В этом случае параметризованный вариант нормальной модели (3.8.81) двухфакторного дисперсионного анализа с *фиксированными эффектами и повторениями* (с одинаковым числом наблюдений n в каждой (i,j)-ячейке табл. 3.8.6) записывается виде

$$X_{ijl}=\mu+\alpha_j+\beta_i+(\alpha\beta)_{ij}+\varepsilon_{ijl},\ i=1,...,r,\ j=1,...,k,\ l=1,...,n, \quad \textbf{(3.9.108)}$$

где ε_{ijl} – независимые в совокупности случайные величины, которые имеют нормальное распределение с нулевым средним и одинаковой дисперсией σ^2, то есть выполняется выражение $L(\varepsilon_{ijl})=N(0,\sigma^2)$ для всех $i=1,...,r$, $j=1,...,k$, $l=1,...,n$. Параметры μ, α_j, β_i этой модели определяются следующим образом. Общее среднее μ изучаемой совокупности определяется в виде $\mu=\dfrac{1}{rk}\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{k}\mu_{ij}$, параметр α_j обозначает j-й главный эффект (или дифференциальный эффект) фактора A, который определяется в виде $\alpha_j=\sum\limits_{i=1}^{r}(\mu_{ij}/r)-\mu$, $j=1,...,k$, параметр β_i обозначает i-й главный эффект фактора B, который определяется в виде $\beta_i=\sum\limits_{j=1}^{k}(\mu_{ij}/k)-\mu$, $i=1,...,r$. Вели-

чина $(\alpha\beta)_{ij}$ называется двухфакторным взаимодействием *j*-го уровня фактора A и *i*-го уровня фактора B и определяется в виде

$$(\alpha\beta)_{ij} = \mu_{ij} + \mu - \left(\sum_{i=1}^{r} (\mu_{ij}/r) + \sum_{j=1}^{k} (\mu_{ij}/k) \right),\ i = 1,...,r,\ j = 1,...,k.$$

Отметим, что модель, в которой взаимодействия $(\alpha\beta)_{ij} = 0$ при всех $i = 1,...,r$, $j = 1,...,k$, называется аддитивной. Отметим также, что двухфакторный дисперсионный анализ обычно начинают с проверки нулевой гипотезы об аддитивности модели, которая записывается в виде H_{AB}: *все* $(\alpha\beta)_{ij} = 0$. Гипотезы об отсутствии главных эффектов факторов A и B соответственно записываются в виде $H_A : \alpha_1 = ... = \alpha_k = 0$ и $H_B : \beta_1 = ... = \beta_k = 0$.

Двухфакторные повторяемые планы со случайными эффектами. Если оба фактора A и B соответствуют модели II со случайными эффектами, то модель называют двухфакторным повторяемым планом со случайными эффектами, для которой наблюдения записываются в виде

$$X_{ijl} = \mu + a_j + b_i + (ab)_{ij} + \varepsilon_{ijl},\ i = 1,...,r,\ j = 1,...,k,\ l = 1,...,n, \qquad \textbf{(3.9.109)}$$

где μ – генеральное среднее, a_j – независимые случайные величины с нормальным распределением $N(0, \sigma_a^2)$, b_j – независимые случайные величины с нормальным распределением $N(0, \sigma_b^2)$, $(ab)_{ij}$ – независимые случайные величины с нормальным распределением $N(0, \sigma_{ab}^2)$, ε_{ijl} – независимые случайные величины с нормальным распределением $N(0, \sigma^2)$. Считается, что все случайные величины a_j, $b_i, (ab)_{ij}$, ε_{ijl} независимы в совокупности. В этой модели выделяют четыре компоненты дисперсии σ_a^2, $\sigma_b^2, \sigma_{ab}^2$ и остаточную дисперсию σ^2, которые связаны соответственно с факторами A, B, (AB) и ошибками измерений $\{\varepsilon_{ij}\}$.

Смешанная модель. Повторяемые планы с рандомизированными блоками. Модель, описывающая повторяемый план с рандомизированными блоками, называется *двухфакторным смешанным повторяемым планом*, для которого наблюдения записываются в виде

$$X_{ijl} = \mu + \alpha_j + b_i + (\alpha b)_{ij} + \varepsilon_{ijl},\ i = 1,...,r,\ j = 1,...,k,\ l = 1,...,n, \qquad \textbf{(3.9.110)}$$

где μ – генеральное среднее, α_j обозначает *j*-й дифференциальный эффект фактора A (с фиксированными эффектами). Считается, что все случайные величины $b_i, (ab)_{ij}$, ε_{ijl} независимы в совокупности. Случайные величины b_j

независимы и имеют нормальное распределение $N(0,\sigma_b^2)$, случайные величины $(ab)_{ij}$ независимы и имеют нормальное распределение $N(0,\sigma_{ab}^2)$, а ошибки ε_{ijl} независимы и имеют нормальное распределение $N(0,\sigma^2)$.

В данном разделе мы обсудим лишь случай фиксированных эффектов факторов A и B. Обсуждение смешанных моделей и варианта случайных эффектов может быть найдено в [3]. Итак, обсудим двухфакторный дисперсионный анализ с повторениями, для которого модель наблюдений определяется выражением (3.9.108). Обычно этот анализ начинают с проверки нулевой гипотезы об аддитивности модели, которая записывается в виде H_{AB}: все $(\alpha\beta)_{ij}=0$. Гипотезы об отсутствии главных эффектов факторов A и B соответственно записываются в виде H_A:$\alpha_1=...=\alpha_k=0$ и H_B:$\beta_1=...=\beta_k=0$. Если в результате проверки сформулированные гипотезы не являются значимыми, то есть не отвергаются на заданном уровне значимости, то дальнейшей целью дисперсионного анализа является построение оценок параметров μ, α_j, β_i, $(\alpha\beta)_{ij}$ по наблюденным данным $x_{ij1},...,x_{ijn}$ выборки $X_{ij1},...,X_{ijn}$, $i=1,...,r$, $j=1,...,k$, $l=1,...,n$. Для этих параметров, как обычно, выполняются следующие условия:

$$\sum_{j=1}^{k}\alpha_j=0,\ \sum_{i=1}^{r}\beta_i=0,\ \sum_{i=1}^{r}(\alpha\beta)_{ij}=\sum_{j=1}^{k}(\alpha\beta)_{ij}=0. \qquad \textbf{(3.9.111)}$$

При их выполнении однозначные МНК-оценки параметров μ, $\{\alpha_j\}$, $\{\beta_i\}$, $(\alpha\beta)_{ij}$, $i=1,...,r$, $j=1,...,k$, находятся путем минимизации суммы

$$S=\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl}-\mu-\alpha_j-\beta_i-(\alpha\beta)_{ij})^2.$$

Решая систему нормальных уравнений, в частности, из уравнения

$$\partial S/\partial\mu=-2\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl}-\mu-\alpha_j-\beta_i-(\alpha\beta)_{ij})=0,$$

с учетом (3.9.111) , получаем МНК-оценку параметра μ в виде

$$\hat{\mu}=\overline{X}_{\bullet\bullet\bullet}=\frac{1}{rkn}\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n}X_{ijl}. \qquad \textbf{(3.9.112)}$$

Аналогично находим МНК-оценки параметров $\{\alpha_j\}$, $\{\beta_i\}$, $(\alpha\beta)_{ij}$, $j=1,...,k$, $i=1,...,r$, в виде

$$\hat{\alpha}_j = \bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet} = \frac{1}{nr}\sum_{i=1}^{r}\sum_{l=1}^{n} X_{ijl} - \frac{1}{rkn}\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n} X_{ijl},\ j=1,...,k,$$

$$\hat{\beta}_i = \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet\bullet\bullet} = \frac{1}{nk}\sum_{j=1}^{k}\sum_{l=1}^{n} X_{ijl} - \frac{1}{rkn}\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n} X_{ijl},\ i=1,...,r,$$

$$(\alpha\hat{\beta})_{ij} = \bar{X}_{ij\bullet} - \bar{X}_{\bullet j\bullet} - \bar{X}_{i\bullet\bullet} + \bar{X}_{\bullet\bullet\bullet},\ i=1,...,r,\ j=1,...,k, \qquad \textbf{(3.9.113)}$$

где $$\bar{X}_{\bullet\bullet\bullet} = \frac{1}{rkn}\sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n} X_{ijl} = \frac{1}{k}\sum_{j=1}^{k}\bar{X}_{\bullet j\bullet} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_{i\bullet\bullet} = \frac{1}{kr}\sum_{j=1}^{k}\sum_{i=1}^{r}\bar{X}_{ij\bullet},$$

$$\bar{X}_{i\bullet\bullet} = \frac{1}{nk}\sum_{j=1}^{k}\sum_{l=1}^{n} X_{ijl},\ \bar{X}_{\bullet j\bullet} = \frac{1}{nr}\sum_{i=1}^{r}\sum_{l=1}^{n} X_{ijl},\ \bar{X}_{ij\bullet} = \frac{1}{n}\sum_{l=1}^{n} X_{ijl}. \qquad \textbf{(3.9.114)}$$

Далее для наглядности и удобства дальнейших ссылок соответствующие суммы квадратов, используемые для проверки гипотез

$$H_A: \alpha_1 = ... = \alpha_k = 0,\ H_B: \beta_1 = ... = \beta_r = 0,$$
$$H_{AB}: (\alpha\beta)_{ij} = 0,\ i=1,...,r,\ j=1,...,k, \qquad \textbf{(3.9.115)}$$

приведены в табл. 3.9.17.

Таблица 3.9.17

Источник рассеивания	Сумма квадратов	Степени свободы	Средний квадрат	$M(S^2_{(\cdot)})$
Фактор A	$SS_A = rn\sum_{j=1}^{k}(\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet})^2$	$\nu_A = k-1$	SS_A/ν_A	$\sigma^2 + rn\sigma^2_A$
Фактор B	$SS_B = kn\sum_{i=1}^{r}(\bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet\bullet\bullet})^2$	$\nu_B = r-1$	SS_B/ν_B	$\sigma^2 + kn\sigma^2_B$
Взаимодействие AB	$SS_{AB} = n\sum_{i=1}^{r}\sum_{j=1}^{k}(\bar{X}_{ij\bullet} - \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet j\bullet} + \bar{X}_{\bullet\bullet\bullet})^2$	$\nu_{AB} = (r-1)\times(k-1)$	SS_{AB}/ν_{AB}	$\sigma^2 + n\sigma^2_{AB}$
Остаток (ошибки)	$SS_R = \sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl} - \bar{X}_{ij\bullet})^2$	$\nu_R = rk(n-1)$	SS_R/ν_R	σ^2
Полная сумма кв.	$SS_T = \sum_{i=1}^{r}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl} - \bar{X}_{\bullet\bullet\bullet})^2$	$\nu_T = rkn-1$	------	------

Как обычно, средние квадраты определяются путем деления соответствующих сумм квадратов на число степеней свободы и записываются в виде

$$S_A^2 = SS_A / \nu_A,\ S_B^2 = SS_B / \nu_B,\ S_{AB}^2 = SS_{AB} / \nu_{AB},\ S_R^2 = SS_R / \nu_R. \qquad \textbf{(3.9.116)}$$

В правом столбце табл. 3.9.17 также приведены формулы для математических ожиданий соответствующих средних квадратов. В этих формулах использованы следующие обозначения:

$$\sigma_A^2 = \frac{1}{k-1}\sum_{j=1}^{k}\alpha_j^2,\ \sigma_B^2 = \frac{1}{r-1}\sum_{i=1}^{r}\beta_i^2,\ \sigma_{AB}^2 = \frac{1}{(r-1)(k-1)}\sum_{i=1}^{r}\sum_{j=1}^{k}(\alpha\beta)_{ij}^2. \qquad \textbf{(3.9.117)}$$

Следуя работе [3], приведем критерии проверки гипотез для повторяемого двухфакторного плана с фиксированными эффектами в табл. 3.9.18.

Таблица 3.9.18

Отсутствие эффектов взаимодействия $H_{AB}:(\alpha\beta)_{ij}=0$, $i=1,...,r,\ j=1,...,k$	Отсутствие главных эффектов фактора A $H_A:\alpha_1=...=\alpha_k=0$	Отсутствие главных эффектов фактора B $H_B:\beta_1=...=\beta_k=0$
$F_{AB}=\frac{SS_{AB}/\nu_{AB}}{SS_R/\nu_R}$, $L(F_{AB}\|H_{AB})=F(\nu_1,\nu_2)$, $\nu_1=(k-1)(r-1)$, $\nu_2=rk(n-1)$. Если $F_{AB}>F_{1-\alpha}(\nu_1,\nu_2)$, то гипотеза H_{AB} отвергается на уровне значимости α	$F_A=\frac{SS_A/\nu_A}{SS_R/\nu_R}$, $L(F_A\|H_A)=F(\nu_1,\nu_2)$, $\nu_1=k-1,\ \nu_2=rk(n-1)$. Если $F_A>F_{1-\alpha}(\nu_1,\nu_2)$, то гипотеза H_A отвергается на уровне значимости α	$F_B=\frac{SS_B/\nu_B}{SS_R/\nu_R}$, $L(F_B\|H_B)=F(\nu_1,\nu_2)$, $\nu_1=r-1,\ \nu_2=rk(n-1)$. Если $F_B>F_{1-\alpha}(\nu_1,\nu_2)$, то гипотеза H_B отвергается на уровне значимости α

Отметим, что *P-значение* критериев проверки гипотез (3.9.115), которые основаны на использовании статистик F_{AB}, F_A, F_B (см. табл. 3.9.18), определяется как площадь справа от наблюденного значения $F_{(\cdot)}$-статистики под кривой плотности распределения Фишера $F(\nu_1,\nu_2)$ с числами степеней свободы ν_1 и ν_2. Если *P-значение* критерия меньше уровня значимости α, то соответствующая гипотеза отвергается.

Замечание 3.9.118. Мы уже отмечали (см. замечание 3.8.80), что в ходе решения задач дисперсионного анализа часто пересматривают модель наблюдений в результате проведенных исследований. Например, переходят от одно-

факторной модели к двухфакторной. При использовании двухфакторной модели (3.9.108) обычно сначала проверяется гипотеза об аддитивности модели, которая записывается в виде $H_{AB}:(\alpha\beta)_{ij}=0$, $i=1,...,r$, $j=1,...,k$. Если в результате проверки гипотеза об аддитивности модели не оказалась значимой (то есть гипотеза не отвергается), то при анализе главных эффектов факторов A и B поступают различным образом (см. [3]). Один из вариантов является обычным. Он игнорирует факт отсутствия взаимодействия факторов A и B и состоит в обычной проверке гипотез $H_A:\alpha_1=...=\alpha_k=0$, $H_B:\beta_1=...=\beta_r=0$ с помощью *F*-критериев, указанных в табл. 3.9.18. Другой вариант учитывает факт того, что эффект взаимодействия факторов A и B не является значимым и основан на объединении остаточной суммы квадратов SS_R с суммой квадратов SS_{AB}, которая характеризует взаимодействие факторов A и B . Такое объединение позволяет построить новую оценку параметра σ^2 . В результате этого объединения образуется новая сумма квадратов $SS_P=SS_R+SS_{AB}$ с числом степеней свободы $\nu_P=\nu_R+\nu_{AB}$, и новой оценкой дисперсии σ^2 служит средний квадрат SS_P/ν_P . Далее гипотеза $H_A:\alpha_1=...=\alpha_k=0$ проверяется с помощью критерия, основанного на статистике $F_A=\{SS_A/\nu_A\}/\{SS_P/\nu_P\}$, а гипотеза $H_B:\beta_1=...=\beta_r=0$ – с помощью критерия, основанного на статистике $F_B=\{SS_B/\nu_B\}/\{SS_P/\nu_P\}$. *P-значение* критериев проверки этих гипотез определяется как площадь справа от наблюденного значения $F_{(\cdot)}$-статистики под кривой плотности распределения Фишера $F(\nu_1,\nu_P)$ с числами степеней свободы ν_1 и ν_P (см. табл. 3.9.18).

Отметим, что в работе [3, с. 249] всех пользователей дисперсионного анализа делят на три группы: «*часто объединяющие*, *редко объединяющие и необъединяющие*». Представители группы «*необъединяющие*» не изменяют оценку σ^2 независимо от результата проверки гипотезы H_{AB} :*все* $(\alpha\beta)_{ij}=0$, и они действуют в соответствии с критериями табл. 3.9.18. Представители другой группы «ч*асто объединяющие*» пересчитывают оценку для параметра σ^2 , если взаимодействие факторов A и B оказалось несущественно, то есть гипотеза об аддитивности модели не отвергнута. Представители следующей группы «*редко объединяющие*» переходят к объединению, только если *P-значение* критерия, полученное при проверке гипотезы аддитивности, достаточно велико (например, если *P-значение* $>0{,}5$). Итак, никаких точных рекомендаций при решении данного вопроса не существует, и поэтому каждый исследователь сам определяет свой «статус объединителя», исходя из своих пристрастий, интуиции и опыта. Отметим также, что вопрос об объединении намного существеннее для двухфакторной модели II со случайными эффектами, чем для модели I двухфакторного дисперсионного анализа с фиксиро-

ванными эффектами. Так как для модели II объединение может привести к существенному увеличению числа степеней свободы знаменателя F-статистики, что, в свою очередь, приводит к увеличению мощности F-критерия при проверке гипотез относительно главных эффектов факторов A и B.

Пример 3.9.119. Приведем числовой пример, статистические данные для которого взяты из [3]. Изучаемым показателем X является количество выдыхаемого азота случайно выбранного пациента. Фактор A характеризует диету питания (рассматривается четыре варианта диеты: A_1 – не содержит белок, A_2 – 23 % белков, A_3 – 32 % белков, A_4 – 67 % белков; априорные представления состоят в том, что количество выдыхаемого азота в среднем возрастает при возрастании потребления белков). Фактор B характеризует пол пациента и имеет два уровня: B_1 – мужчины, B_2 – женщины. С целью изучения влияния диеты и пола пациента на количество выдыхаемого азота X проводился эксперимент, в котором каждая комбинация пола и диеты была повторена три раза. Итак, адекватной моделью для описания условий данного эксперимента является модель двухфакторного дисперсионного анализа с повторениями (то есть с равным числом наблюдений в ячейках) и с фиксированными эффектами факторов A и B, которая, согласно (3.9.108), записывается в виде

$$X_{ijl} = \mu + \alpha_j + \beta_i + (\alpha\beta)_{ij} + \varepsilon_{ijl},\ i = 1,...,r,\ j = 1,...,k,\ l = 1,...,n.$$

В данном эксперименте фактор B – пол – имеет два уровня, то есть $r = 2$, фактор A – диета – четыре уровня, то есть $k = 4$ и число повторений эксперимента в каждой (i, j)-ячейке равно $n = 3$. Наблюденные данные приведены в табл. 3.9.19. Для удобства в этой таблице также приведены вычисленные по формулам (3.9.114) частные средние значения по строкам $\overline{X}_{i\bullet\bullet}$, $i = 1,2$, и частные средние значения по столбцам $\overline{X}_{\bullet j\bullet}$, $j = 1,...,4$, а также общее среднее $\overline{X}_{\bullet\bullet\bullet}$.

Таблица 3.9.19

$B_i \setminus A_j$	A_1	A_2	A_3	A_4	Среднее по строкам
B_1	4,079 4,859 3,540	4,368 5,668 3,752	4,169 5,709 4,416	4,928 5,608 4,940	$\overline{X}_{2\bullet\bullet} = 4{,}6697$
B_2	2,870 4,648 3,847	3,578 5,393 4,374	4,403 4,496 4,688	4,905 5,208 4,806	$\overline{X}_{2\bullet\bullet} = 4{,}4347$
Среднее по столбцам	$\overline{X}_{\bullet 1\bullet} = 3{,}9738$	$\overline{X}_{\bullet 2\bullet} = 4{,}5222$	$\overline{X}_{\bullet 3\bullet} = 4{,}6468$	$\overline{X}_{\bullet 4\bullet} = 5{,}0658$	$\overline{X}_{\bullet\bullet\bullet} = 4{,}5522$

Оценки наименьших квадратов параметров μ, $\{\alpha_j\}$, $\{\beta_i\}$, $(\alpha\beta)_{ij}$, $i=1,...,r$, $j=1,...,k$, вычисляются по формулам (3.9.112) и (3.9.113). В частности, $\hat{\mu}=\bar{X}_{\bullet\bullet\bullet}=4{,}5522$. Главные эффекты фактора B (пол пациента) соответственно равны:

для мужчин

$$\hat{\beta}_1=\bar{X}_{1\bullet\bullet}-\bar{X}_{\bullet\bullet\bullet}=4{,}4669-4{,}5522=-0{,}0853$$

и для женщин

$$\hat{\beta}_2=\bar{X}_{2\bullet\bullet}-\bar{X}_{\bullet\bullet\bullet}=4{,}4347-4{,}5522=-0{,}1175.$$

Главные эффекты фактора A (диета) соответственно равны:

для диеты A_1 – $\hat{\alpha}_1=\bar{X}_{\bullet1\bullet}-\bar{X}_{\bullet\bullet\bullet}=3{,}9738-4{,}5522=-0{,}5784$,

для диеты A_2 – $\hat{\alpha}_2=\bar{X}_{\bullet2\bullet}-\bar{X}_{\bullet\bullet\bullet}=4{,}5222-4{,}5522=-0{,}030$,

для диеты A_3 – $\hat{\alpha}_3=\bar{X}_{\bullet3\bullet}-\bar{X}_{\bullet\bullet\bullet}=4{,}6468-4{,}5522=0{,}0940$,

для диеты A_4 – $\hat{\alpha}_4=\bar{X}_{\bullet4\bullet}-\bar{X}_{\bullet\bullet\bullet}=5{,}0658-4{,}5522=0{,}5136$.

МНК-оценки взаимодействия различных уровней факторов A и B вычисляются по формуле (3.9.113). В частности, величина взаимодействия диеты A_1 и мужчин B_1 равна $(\alpha\hat{\beta})_{11}=\bar{X}_{11\bullet}-\bar{X}_{\bullet1\bullet}-\bar{X}_{1\bullet\bullet}+\bar{X}_{\bullet\bullet\bullet}$ = (4,079+4,859+3,540)/3 – – 3,9738 – 4,6697 + 4,5522 = 0,0680. Аналогично получаем другие значения величин взаимодействия диеты и пола, например взаимодействие диеты A_4 и женщин B_2 равно

$$(\alpha\hat{\beta})_{24}=\bar{X}_{24\bullet}-\bar{X}_{\bullet4\bullet}-\bar{X}_{2\bullet\bullet}+\bar{X}_{\bullet\bullet\bullet}=$$
$$=(4{,}905+5{,}208+4{,}806)/3-5{,}0658-4{,}4347+4{,}5522=0{,}0247.$$

Используя вычисленные оценки параметров μ, $\{\alpha_j\}$, $\{\beta_i\}$, $(\alpha\beta)_{ij}$, $i=1,...,r$, $j=1,...,k$, и формулы для соответствующих сумм квадратов отклонений (см. табл. 3.9.17), заполним табл. 3.9.20 дисперсионного анализа.

В этой таблице сумма квадратов отклонений, соответствующая фактору A, была вычислена по формуле

$$SS_A=rn\sum_{j=1}^{k}(\bar{X}_{\bullet j\bullet}-\bar{X}_{\bullet\bullet\bullet})^2=rn\sum_{j=1}^{k}(\hat{\alpha}_j)^2=$$
$$=2\cdot3\cdot[(-0{,}5784)^2+(-0{,}03)^2+(0{,}094)^2+(0{,}5136)^2]=3{,}65.$$

Аналогично, сумма квадратов отклонений, соответствующая фактору B, равна

$$SS_B=kn\sum_{i=1}^{r}(\bar{X}_{i\bullet\bullet}-\bar{X}_{\bullet\bullet\bullet})^2=kn\sum_{i=1}^{r}(\hat{\beta}_i)^2=4\cdot3\cdot[(-0{,}0853)^2+(-0{,}1175)^2]=0{,}25.$$

Таблица 3.9.20

Источник вариации	Сумма кв. отклонений	Степени свободы	Средний квадрат	F-статистика
Между столбцами, фактор A (диета)	$SS_A=3{,}65$	$\nu_A=k-1=3$	$SS_A/\nu_A=$ $=1{,}217$	$F_A=\dfrac{SS_A/\nu_A}{SS_R/\nu_R}=2{,}48$
Между строками, фактор B (пол)	$SS_B=0{,}25$	$\nu_B=r-1=1$	$SS_B/\nu_B=$ $=0{,}250$	$F_B=\dfrac{SS_B/\nu_B}{SS_R/\nu_R}=0{,}51$
Взаимодействие факторов A и B	$SS_{AB}=0{,}04$	$\nu_{AB}=(k-1)\times$ $\times(r-1)=3$	$SS_{AB}/\nu_{AB}=$ $=0{,}013$	$F_{AB}=\dfrac{SS_{AB}/\nu_{AB}}{SS_R/\nu_R}=0{,}03$
Ошибки, остаток R	$SS_R=7{,}84$	$\nu_R=rk\times$ $\times(n-1)=16$	$SS_R/\nu_R=$ $=0{,}490$	
Общая вариация	$SS_T=11{,}78$	$\nu_T=rkn-$ $-1=23$		

Другие суммы квадратов отклонений SS_{AB}, SS_R, SS_T были вычислены по формулам табл. 3.9.17 с учетом того, что $SS_{AB}=SS_C-SS_A-SS_B$, $SS_R=SS_T-SS_C$, где $SS_C=n\sum\limits_{i,j}(\bar{X}_{ij\bullet}-\bar{X}_{\bullet\bullet\bullet})^2$.

Используя F-критерии проверки гипотез, приведенные в табл. 3.9.18, и вычисленные значения F-статистик в правом столбце табл. 3.9.20, проверим гипотезу о взаимодействии факторов A, B и гипотезы относительно равенства нулю их главных эффектов.

Проверим гипотезу $H_{AB}:\text{все}\,(\alpha\beta)_{ij}=0$ о взаимодействии факторов «диета и пол». Пусть уровень значимости равен $\alpha=0{,}05$. В данном примере числа степеней свободы ν_1 и ν_2 соответственно $\nu_1=(k-1)(r-1)=3$, $\nu_2=rk(n-1)=16$ и квантиль уровня $(1-\alpha)$ для F-распределения Фишера $F_{1-\alpha}(\nu_1,\nu_2)=F_{0{,}95}(3{,}16)=3{,}239$. Далее, так как $\{F_{AB}=0{,}03\}<\{F_{0{,}95}(3{,}16)=3{,}239\}$, то гипотеза H_{AB} о взаимодействии факторов «диета и пол» не является значимой, то есть не отвергается на уровне значимости $\alpha=0{,}05$.

Проверим гипотезу $H_A:\alpha_1=\ldots=\alpha_k=0$ об отсутствии дифференциальных эффектов фактора «диета». В данном случае $\nu_1=k-1=3$, $\nu_2=rk(n-1)=16$. Так как $\{F_A=2{,}48\}<\{F_{0{,}95}(3,16)=3{,}239\}$, то гипотеза H_A об отсутствии дифференциальных эффектов фактора «диета» не является значимой, то есть не отвергается на уровне значимости $\alpha=0{,}05$.

Проверка гипотезы $H_B : \beta_1 = ... = \beta_r = 0$ об отсутствии дифференциальных эффектов фактора «пол пациента». В данном случае числа степеней свободы ν_1 и ν_2 соответственно равны $\nu_1 = r - 1 = 1$, $\nu_2 = rk(n-1) = 16$ и $F_{0,95}(1, 16) = 4,494$. Так как $\{F_B = 0,51\} < \{F_{0,95}(1,16) = 4,494\}$, то гипотеза H_B об отсутствии дифференциальных эффектов фактора «пол пациента» не является значимой, то есть не отвергается на уровне значимости $\alpha = 0,05$.

Итак, все три проверяемые гипотезы не являются значимыми, то есть не противоречат данным эксперимента, приведенным в таблице (3.9.19), в рамках модели I двухфакторного дисперсионного анализа с повторениями и с фиксированными эффектами фактора *A* – «диета» и фактора *B* – «пол пациента».

Поскольку эффект взаимодействия факторов «диета и пол» оказался незначимым, то в соответствии с замечанием 3.9.118, можно вычислить объединенную сумму квадратов $SS_P = SS_R + SS_{AB} = 0,04 + 7,84 = 7,88$ с числом степеней свободы $\nu_P = \nu_R + \nu_{AB} = 3 + 16 = 19$ и средний квадрат $SS_P / \nu_P = 7,88/19 = 0,42$. В результате такого объединения новые значения *F*-статистик будут соответственно равны $F_A^* = \{SS_A / \nu_A\} / \{SS_P / \nu_P\} = 1,217/0,42 = 2,90$ и $F_B^* = \{SS_B / \nu_B\} / \{SS_P / \nu_P\} = 0,25/0,42 = 0,60$.

Сравнивая вычисленные значения статистик F_A^* и F_B^* соответственно с квантилями $F_{0,95}(3, 19) = 3,127$ и $F_{0,95}(1, 19) = 4,381$, заключаем, что и после объединения главные эффекты факторов «диета» и «пол пациента» значимо не отличаются от нуля при $\alpha = 0,05$. Отметим также, что оценка дисперсии σ^2 равна $\hat{\sigma}^2 = SS_R / \nu_R = 0,49$ (без объединения) и $\hat{\sigma}_*^2 = SS_P / \nu_P = 0,42$ (с объединением).

ЛИТЕРАТУРА

1. *Абрамовиц М.*, *Стиган И.* Справочник по специальным функциям с формулами, графиками и математическими таблицами: пер. с англ. под ред. В.А. Диткина и Л.Н. Кармазинной. – М.: Наука, 1979. – 830 с.
2. *Аптон Г.* Анализ таблиц сопряженности. – М.: Финансы и статистика, 1982. – 144 с.
3. *Афифи А.*, *Эйзен С.* Статистический анализ: Подход с использованием ЭВМ: пер. с англ. – М.: Мир, 1982. – 488 с.
4. *Бабкин Н.В.*, *Макшанов А.В.*, *Мусаев А.А.* Робастные методы статистического анализа навигационной информации. – Центр. НИИ «Румб», 1985.
5. *Бикел П.*, *Доксум К.* Математическая статистика. – М.: Финансы и статистика, 1983. – Вып. 1. – 280 с.; Вып. 2. – 197 с.

5а. *Болдин М.В.*, *Симонова Г.И.*, *Тюрин Ю.Н.* Знаковый статистический анализ линейных моделей. – М.: Наука; Физматлит, 1997. – 288 с.

6. *Большев Л.Н.*, *Смирнов Н.В.* Таблицы математической статистики. – М.: Наука, 1983. – 416 с.
7. *Боровков А.А.* Математическая статистика. – М.: Наука, 1984. – 472 с.
8. *Боровков А.А.* Теория вероятностей. – М.: Наука, 1986. – 431 с.
9. *Браунли К. А.* Статистическая теория и методология в науке и технике. – М.: Наука, 1977. – 407 с.
10. *Ван-дер-Варден Б.Л.* Математическая статистика. – М.: ИЛ, 1960. – 434 с.
11. *Ватутин В.А.*, *Ивченко Г.И.*, *Медведев Ю.И.*, *Чистяков В.П.* Теория вероятностей и математическая статистика в задачах: учеб. пособие для вузов. – 3-е изд., испр. – М.: Дрофа, 2005. – 315 с.

11а. *Воинов В.Г.*, *Никулин М.С.* Несмещенные оценки и их применения. – М.: Наука, 1989. – 437 с.

12. *Гаек Я.*, *Шидак З.* Теория ранговых критериев. – М.: Наука, 1971. – 376 с.
13. *Гихман И.И.*, *Скороход А.В.*, *Ядренко М.И.* Теория вероятностей и математическая статистика. – 2-е изд., перераб. и доп. – Киев: Выща шк.; Головное изд-во, 1988. – 439с.
14. *Гнеденко Б.В.* Курс теории вероятностей. – М.: Наука, 1988. – 446 с.
15. *Градштейн И.С.*, *Рыжик И.М.* Таблицы интегралов, сумм, рядов и произведений. – М.: ГИФМЛ, 1963. – 1100 с.
16. *Гусаров В.М.* Статистика. – М.: ЮНИТИ, 2003. – 463 с.
17. *Ершов А.А.* Стабильные методы оценки параметров (обзор) // Автоматика и телемеханика. – 1978. – № 8. – С. 66–100.
18. *Дэйвид Г.* Порядковые статистики. – М.: Наука, 1979. – 336 с.

19. *Ефимов А.В.* Сборник задач по математике для ВТУЗОВ. Специальные курсы / под ред. А.В. Ефимова. – М.: Наука, 1984. – 606 с.
20. *Закс Ш.* Теория статистических выводов. – М.: Мир, 1974. – 776 с.
21. *Закс Л.* Статистическое оценивание. – М.: Статистика, 1976. – 598 с.
22. *Ивченко Г.И.*, *Медведев Ю.И.* Математическая статистика. – М.: Высшая школа, 1984. – 248 с.
23. *Кендэлл М.*, *Стьюарт А.* Теория распределений. – М.: Наука, 1966. – 587 с.
24. *Кендэлл М.*, *Стьюарт А.* Статистические выводы и связи. – М.: Наука, 1973. – 899 с.
25. *Кендэлл М.* Ранговые корреляции. – М.: Статистика, 1975. – 212 с.
26. *Климов Г.П.* Теория вероятностей и математическая статистика. – М.: Изд-во МГУ, 1983. – 328 с.
27. *Коваленко И.Н.*, *Филиппова А.А.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1982. – 256 с.
28. *Козлов М.В.*, *Прохоров А.В.* Введение в математическую статистику. – М.: Изд-во МГУ, 1987. – 264 с.
29. *Кокрен У.* Методы выборочного исследования: пер. с англ. И.М. Сонина / под ред. А.Г. Волкова; Вступит. статья Н.К. Дружинина. – М.: Статистика, 1976. – 440 с.
30. *Кокс Д.*, *Хинкли Д.* Теоретическая статистика. – М.: Мир, 1978. – 560 с.
31. *Кокс Д.*, *Льюис П.* Статистический анализ последовательностей событий. – М.: Мир, 1969. – 312 с.
32. *Колемаев В.А.*, *Калинина В.Н.* Теория вероятностей и математическая статистика / под ред. В.А. Колемаева. – М.: ИНФРА-М, 1997. – 302 с.
33. *Корн Г.*, *Корн Т.* Справочник по математике для научных работников и инженеров. – М.: Наука, 1984. – 831 с.
34. *Королюк В.С.*, *Бобровских Ю.В.* Теория U-статистик. – Киев: Наукова думка, 1989. – 383 с.
35. *Королюк В.С.*, *Портенко Н.И.*, *Скороход А.В.*, *Турбин А.Ф.* Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.
36. *Крамер Г.* Математические методы статистики. – М.: Мир, 1974. – 648 с.
37. *Леман Э.* Проверка статистических гипотез. – М.: Наука, 1964. – 498 с.
38. *Леман Э.* Теория точечного оценивания. – М.: Наука, 1991. – 448с.
39. *Ликеш И.*, *Ляга И.* Основные таблицы математической статистики. – М.: Финансы и статистика, 1985. – 356 с.
40. *Лонер Р.Л.*, *Уилкинсон Г.Н.* Устойчивые статистические методы оценки данных: пер. с англ. под ред. Н.Г. Волкова. – М.: «Машиностроение», 1984. – 229 с.
41. *Маршал А.*, *Олкин И.* Неравенства. Теория мажоризации и её приложения: пер. с англ. под ред. В.М. Круглова. – М.: Мир, 1983. – 480 с.

42. *Мельник М.* Основы прикладной статистики: пер с англ. – М.: Энергоатомиздат, 1983. – 416 с.
43. *Михок Г.*, *Урсяну В.* Выборочный метод и статистическое оценивание: пер. с рум. В.М. Остиану / под ред. В.Ф. Матвеева. – М.: Финансы и статистика, 1982. – 245 с.
44. *Мюллер П.*, *Нойман П.*, *Шторм Р.* Таблицы по математической статистике. – М.: Финансы и статистика, 1982. – 278 с.
45. *Натансон Н.П.* Теория функций вещественной переменной. – 2-е изд. – М.: Гостехиздат, 1957. – 480 с.
46. *Оуэн Д.Б.* Сборник статистических таблиц: 2-е изд., испр. – М.: ВЦ АН СССР, 1973. – 586 с.
47. *Поллард Дж.* Справочник по вычислительным методам статистики. – М.: Финансы и статистика, 1982. – 344 с.
47а. *Прудников А.П.*, *Брычков Ю.А.*, *Маричев О.И.* Интегралы и ряды. – М.: Наука, 1981. – 798 с.
48. *Пугачев В.С.* Теория вероятностей и математическая статистика. – М.: Наука, 1979. – 496 с.
49. *Рао С.Р.* Линейные статистические методы и их применение. – М.: Наука, 1968. – 548 с.
50. *Романовский В.* Математическая статистика. – М., 1938; Л.: ГОНТИ НКТП СССР, 1938. – 527 с.
51. *Рунион Р.* Справочник по непараметрической статистике. – М.: Финансы и статистика, 1982. – 198 с.
52. *Сархан А.Е.*, *Гринберг В.* Введение в теорию порядковых статистик: пер. с англ. под ред. Ф.Я. Боярского. – М.: Статистика, 1970. – 414 с.
53. *Смирнов Н.В.* Теория вероятностей и математическая статистика. Избранные труды. – М.: Наука, 1970. – 290 с.
54. *Смирнов Н.В.*, *Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. – 2-е изд., испр. и доп. – М.: Наука, 1965. – 511с.
55. *Сошникова Л.А. и др.* Многомерный статистический анализ в экономике / под ред. В.Н. Тамашевича.. – М.: ЮНИТИ, 1999. – 598 с.
56. *Справочник* по прикладной статистике / под ред. Э. Ллойда, У. Ледермана. – М.: Финансы и статистика, 1989. – Т. 1. – 509 с.; 1990. – Т. 2. – 525 с.
57. *Справочник* по специальным функциям с формулами, графиками и рисунками / под ред. К. Эйнслейна, Э. Рэлстона, Г.С. Уолфа. – М.: Наука, 1979. – 830 с.
58. *Статистические* методы в экспериментальной физике: пер. с англ. под ред. А.А. Тяпкина. – М.: Атомиздат, 1976. – 335 с.
59. *Тарасенко Ф.П.* Непараметрическая статистика. – Томск: Изд-во Том. ун-та, 1976. – 292 с.

60. *Тарасенко Ф.П.*, *Шуленин В.П.* О статистической связи между наблюдением и его рангом // Тр. СФТИ при Томском ун-те. – 1971. – Вып. 60. – С. 220–228.
61. *Тарасенко Ф.П.*, *Шуленин В.П.* О некоторых свойствах одной непараметрической процедуры обнаружения сигналов в шумах // Радиотехника. – 1971. – № 4.
62. *Тарасенко Ф.П.*, *Шуленин В.П.* Критерии согласия // Тр. СФТИ при Томском ун-те. – 1974. – Вып. 62. – 67 с.
63. *Тарасенко Ф.П.*, *Шуленин В.П.* Основные идеи устойчивости статистических процедур // Тр. 10-й Всесоюзной школы-семин. по статистической гидроакустике. – Москва – Сухуми, 1978. – С. 114–119.
64. *Тарасенко Ф.П.*, *Шуленин В.П.* Об одном классе оценок параметра положения // VI Международный симпозиум по теории информации: тез. докл. Ч. I. – Москва – Ташкент, 1984. – С. 171–173.
65. *Тюрин Ю.Н.*, *Макаров А.А.* Статистический анализ данных на компьютере / под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528 с.
65а. *Тюрин Ю.Н.* Непараметрические методы статистики // Знание. Сер. Математика, кибернетика. – 1978. – № 4. – С. 1–64.
66. *Уилкс С.* Математическая статистика. – М.: Наука, 1967. – 632 с.
67. *Феллер В.* Введение в теорию вероятностей и ее приложения. – М.: Мир, 1967. – Т. 1. – 498 с.; Т. 2. – 752 с.
68. *Хаджи П.И.* Функция вероятности. – Кишинев: Ред.-изд. отд. АН Молд. ССР, 1971. – 398 с.
69. *Хальд А.* Математическая статистика с техническими приложениями. – М.: ИЛ, 1956. – 664 с.
70. *Хампель Ф.*, *Рончетти Э.*, *Рауссей П.*, *Штаэль В.* Робастность в статистике. Подход на основе функций влияния. – М.: Мир, 1989. – 512 с.
71. *Хан Г.*, *Шапиро С.* Статистические модели в инженерных задачах. – М.: Статистика, 1980. – 444 с.
72. *Харин Ю.С.*, *Степанова М.Д.* Практикум на ЭВМ по математической статистике. – Минск: Изд-во «Университетское», 1987.
73. *Хастингс Н.*, *Пикок Дж.* Справочник по статистическим распределениям. – М.: Статистика, 1980. – 95 с.
74. *Хеттсманспергер Т.* Статистические выводы, основанные на рангах. – М.: Финансы и статистика, 1987. – 334 с.
75. *Химмельблау Д.* Анализ процессов статистическими методами: пер. с англ. В.Д. Скаржинского / под ред. В.Г. Горского. – М.: Мир, 1973. – 957 с.
76. *Холлендер М.*, *Вулф Д.* Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.
77. *Хьюбер П.* Робастность в статистке. – М.: Мир, 1984. – 304 с.
78. *Четыркин Е. М.*, *Калихман И.Л.* Вероятность и статистика. – М.: Финансы и статистика, 1982. – 319 с.

79. *Чибисов Д.М.* Некоторые критерии типа хи-квадрат для непрерывных распределений // Теория вероятностей и ее применения. – 1971. – Т. XVI. – Вып. 1.
80. *Шеффе Г.* Дисперсионный анализ. – М.: Физматгиз, 1963. – 625 с.
81. *Ширяев А.Н.* Вероятность. – М.: Наука, 1980. – 574 с.
82. *Шуленин В.П.* Об одном свойстве выборочных интервалов из равномерного распределения // Тр. СФТИ при Томском ун-те. – 1971. – Вып. 60.
83. *Шуленин В.П.* Об устойчивости класса оценок Ходжеса – Лемана // Тр. VI Всесоюзной конференции по теории кодирования и передачи информации. – Москва – Вильнюс, 1978. – С. 174–151.
84. *Шуленин В.П.* О некоторых свойствах устойчивости α-урезанной оценки Ходжеса – Лемана // Стохастические системы управления. – Новосибирск: Наука, 1979. – С. 54–62.
85. *Шуленин В.П.* Устойчивые оценки параметра положения, основанные на ранговых статистиках, вычисляемых по α-урезанной выборке // Математическая статистика и её приложения. – Томск. Изд-во Том. ун-та, 1978. – Вып. 5. – С. 23 – 37.
86. *Шуленин В.П.* Минимаксные оценки и их числовые характеристики устойчивости // Математическая статистика и её приложения. – Томск. Изд-во Том. ун-та, 1980. – Вып. 6. – С. 32 – 47.
87. *Шуленин В.П.* Использование устойчивых оценок параметра положения для построения оценок симметричных распределений // Математическая статистика и её приложения. – Томск: Изд-во Том. ун-та, 1981. – С. 188–198.
88. *Шуленин В.П.* Устойчивые меры масштабного параметра, их оценивание и сравнение // Математическая статистика и её приложения. – Томск: Изд-во Том. ун-та, 1981. – С. 199–215.
89. *Шуленин В.П.* Минимаксные R_α-оценки и их числовые характеристики при конечных выборках // Адаптация и обучение в системах управления и принятия решений. – Новосибирск: Наука, 1982. – С. 110–122.
90. *Шуленин В.П.* Адаптивная оценка урезанной средней разности Джини // Методы и программное обеспечение обработки информации и прикладного статистического анализа данных на ЭВМ. – Минск, 1985. – С. 113–114.
91. *Шуленин В.П.* Исследование устойчивости и асимптотических свойств урезанной средней разности Джини // Тр. IV Международной конференции по теории вероятности и математической статистике. – Вильнюс, 1985. – С. 330–332.
92. *Шуленин В.П.* Исследование свойств оценки медианы абсолютных разностей // V совещание-семинар по непараметрическим и робастным методам статистики в кибернетике. – Томск: Изд-во Том. ун-та, 1987. – Ч. II. – С. 460–467.

93. *Шуленин В.П.* Асимптотические свойства обобщенных оценок Ходжеса – Лемана // IX Всесоюзная конференция по теории кодирования и передачи информации. – Одесса, 1988. – Ч. II. – С. 111–113.
94. *Шуленин В.П.* Асимптотические свойства и робастность обобщенных L-оценок // Тр. V Международной конференции по теории вероятности и математической статистике. – Вильнюс, 1989. – Т. 4. – С. 377–378.
95. *Шуленин В.П.* Асимптотические свойства обобщенных L-оценок, вычисляемых по урезанным выборкам // Непараметрические и робастные статистические методы в кибернетике и информатике. – Томск: Изд-во Том. ун-та, 1990. – С. 564–570.
96. *Шуленин В.П.* Числовые характеристики устойчивости обобщенных L-оценок // Проблемы компьютерного анализа данных и моделирование. – Минск, 1991. – С. 209–214.
97. *Шуленин В.П.* Асимптотические свойства и робастность MD-оценок // Теория вероятностей и её применение. – 1992. – Т. 37. – Вып. 4. – С. 816–818.
98. *Шуленин В.П.* Связи между классами робастных оценок параметра положения // Компьютерный анализ данных и моделирование. – Минск, 1992. – С. 61.
99. *Шуленин В.П.* Асимптотические свойства и робастность MD-оценок и их связи с M,L,R-оценками // Современные проблемы компьютерного анализа данных и моделирование: сб. научных статей. – Минск: Белгосуниверситет, 1993. – С. 118–124.
100. *Шуленин В.П.* Введение в робастную статистику. – Томск: Изд-во Том. ун-та, 1993. – 227 с.
101. *Шуленин В.П.*, *Серых А.П.* Робастные и непараметрические алгоритмы обработки данных физических экспериментов // Изв. вузов. Физика. – 1993. – № 10. – С. 128–136.
102. *Шуленин В.П.* Границы эффективности оценок, построенных методом минимума расстояния Крамера – Мизеса // Изв. вузов. Физика. – 1995. – № 9. – С. 84–89.
103. *Шуленин В.П.* О робастности оценок класса функционалов от распределения вероятностей // Второй Сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-96): тез. докл. – Новосибирск, 1996.
104. *Шуленин В.П.* Границы эффективностей оценок, построенных методом минимума расстояния Крамера – Мизеса // СИБКОНВЕРС'95. Международная конференция по использованию результатов конверсии науки в вузах Сибири для международного сотрудничества. Томск, 2 – 4 октября 1995 г.: труды конф. Т. 1. – Томск, 1996. – С. 124–125.
105. *Шуленин В.П.*, *Деева Т.А.* Свойства ранговых критериев при конечных объемах выборки для λ-распределений Тьюки // Международная конфе-

ренция Всесибирские чтения по математике и механике 17 – 20 июня 1997 г., Томск. Избранные доклады. Т. 1. – Томск: Изд-во Том. ун-та, 1997. – С. 104 – 109.

106. *Шуленин В.П.*, *Усачева Н.А.* О робастности F- и H-критериев проверки гипотез к отклонениям от гауссовской модели // Новые технологии и комплексные решения: наука, образование, производство: материалы Всероссийской научно-практической конференции 19 окт. 2001 г. Ч. II (Математика). – Анжеро-Судженск, 2001. – С. 80–82.
107. *Шуленин В.П.* Асимптотические свойства GL- и U-статистик на урезанных выборках // Новые технологии и комплексные решения: наука, образование, производство: материалы Всероссийской научно-практической конференции 19 окт. 2001 г. Ч. II (Математика). – Анжеро-Судженск, 2001. – С. 82–85.
108. *Шуленин В.П.*, *Вертышева И.О.* Приближенные формулы вычисления моментов порядковых статистик // Международная конференция по математике и механике. 16 – 18 сентября 2003 г., Томск: тез. докл. – Томск, 2003. – С. 81–82.
109. *Шуленин В.П.*, *Тарасенко Ф.П.* Функции регрессии наблюдений и их рангов // Вестник Томского государственного университета. – 2003. – № 280. Серия « Математика. Кибернетика. Информатика». – С. 213–216.
110. *Шуленин В.П.* Асимптотические свойства GL- и U-статистик // Вестник Томского государственного университета. Приложение. – 2004. – № 9 (11). – С. 184–190.
111. *Шуленин В.П.* Робастные статистические процедуры // Труды 10-го юбилейного симпозиума по непараметрическим и робастным статистическим методам в кибернетике. – Томск: Изд-во НТЛ, 2004. – С. 159 – 172.
112. *Шуленин В.П.*, *Табольжин.* Изучение свойств ранговых аналогов F-критерия Фишера при отклонениях от гауссовской модели дисперсионного анализа // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2008. – № 1(2). – С. 87–96.
113. *Шуленин В.П.* Эффективные и робастные MD-оценки Крамера – Мизеса // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2010. – № 3(12). – С. 107–121.
114. *Шуленин В.П.* Свойства адаптивных оценок Ходжеса – Лемана в асимптотике и при конечных объемах выборки. Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2010. – № 2(11). – С. 96–112.
115. *Shulenin V.P.* Asymptotic properties and robustness a generalized Hodges – Lehmann estimate // 11th Prague Conference 1990 on Information Theory, Statistical Decision Functions and Random Processes: Abstracts. – Prague, August 27 – 31, 1990.

116. *Shulenin V.P.*, *Tarasenko F.P.* Connection of MD-estimates with classes of robust estimates of location parameter // 12-th PRAGUE Conf. Inf. Theory, Stat. Decision Functions, Random Processes. – Prague, 1994. – P. 220–223.
117. *Shulenin V.P.*, *Deeva T.A.* Asymptotic Efficiency for the Generalized Hodges-Lehmann Estimator under the Normal Scale Mixture Distributions // CADM, MINSK. June 8 – 12, 1998. – V. 2. – P. 107 – 112.
118. *Shulenin V.P.*, *Deeva T.A.* The numerical characteristics of robustness of the class of the Hodges-Lehmann estimators // Proc. of the Third Russian-Korean International Symposium on Science and Technology. KORUS'99. – V. 2. – P. 510–514.
119. *Shulenin V.P.* Asymptotic Properties of the Trimmed GL-and U-Statistics // 6^{th} Prague Symposium on Asymptotic Statistics. Prague, 1998, August 23 – 28. PRAGUE STOCHATICS'98: Abstracts. – P. 84.
120. *Shulenin V.P.* Asymptotic properties of the robust estimators of functionals for measures variability // 14-th PRAGUE Conference on Information Theory, Stat. Decision Functions, Random Processes. Prague, August 19 – 23, 2002.
121. *Ali M.M.* Stochastic Ordering and Kurtosis Measure // J. Amer. Statist. Assoc. – 1974. – V. 69. – P. 543–545.
122. *Andrews D.F.*, *Bickel P.J.*, *Hampel F.R.*, *et al.* Robust estimation of location: survey and advances. – Princeton. N.Y.: Princeton Univ. Press, 1972. – 375 p.
123. *Anscomb F.Z.* Topics in the investigation of linear relation fitted by the method of least squares // J. Roy. Stat. Soc., ser. B. – V. 29. – P. 1–52.
124. *Azencott R.* Robustass des R-estimateurs // Asterisque. – 1977. – V. 43-44. – P. 189–202.
125. *Barlow R.E.*, *Proschan F.* Statistical Theory of Reliability and Life Testing: Probability Models. – Holt, N.Y., 1975.
126. *Bennett C.A.* Asymptotic properties of ideal linear estimator: Ph. D. Thesis. – Univ. Michigan, 1952.
127. *Beran R.* Jackknife approximation to bootstrap estimates // Ann. Statist. – 1984. – V. 12. – No. 1. – P. 101–118.
128. *Bickel P.J.*, *Lehmann E.L.* Descriptive Statistics for nonparametric models. I. Introduction. II. Location // Ann. Statist. – 1975. – V. 3. – P. 1038–1044, 1045–1069.
129. *Bickel P.J.* Another look at robustness: a review of reviews and some new development // Scand. J. Statist. Theory and Appl. – 1976. – V. 3. – P. 145–168.
130. *Bickel P.J.*, *Lehmann E.L.* Measures of location and scale // Proc. Prague Symp. Asymptotic Statist. 1973. V. 1. – Prague: Prague Charles Univ., 1974. – P. 25–36.
131. *Bickel P.J.*, *Lehmann E.L.* Descriptive statistics for nonparametric models. III. Dispersion // Ann. Statist. – 1976. – V. 4. – No. 6. – P. 1139–1158.

132. *Birnbaum A.*, *Laska E.* Optimal robustness: a general method with application to linear estimators of location // J. Amer. Statist. Assoc. – 1976. – V. 62. – P. 1230–1240.
133. *Boos D.D.* A differential for L-statistics // Ann. Statist. – 1979. – V. 7. – P. 955–959.
134. *Boos D.D.*, *Serfling R.J.* A note on differentials and CLT and LIL for statistical functions, with application to M-estimates // Ann. Statist. – 1980. – V. 8. – P. 618–624.
135. *Boos D.D.* Minimum distance estimators for location and goodness of fit // J. Amer. Statist. Assoc. – 1981. – V. 76. – No. 375. – P. 663–670.
136. *Box G.E.P.* Non-normality and tests on variances // Biometrika. – 1953. – V. 40. – P. 318–335.
137. *Box G.E.P. and Anderson S.L.* Permutation Theory in the Derivation of Robust Criteria and the Study of the Departures from Assumption // J. Royal Statistical Society, B (Methodological). – 1955. – V. 17. – P. 1–26.
138. *Chernoff H.*, *Gastwirth J.L.*, *Johns M.V.* Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation // Ann. Math. Statist. – 1976. – V. 38. – P. 52–72.
139. *Costa V.*, *Deshayes J.* Comparison des R-L-M-estimatuers // Asterisque. – 1977. – No. 43-44. – P. 209–238.
140. *David F.N.* Tables of the Correlation Coefficient. – Cambridge University Press, 1938. Reprinted in: Biometrical Tables for Statisticians. – 1966. – V. I. – 3rd edition.
141. *David H.A.* Gini's mean difference rediscovered // Biometrika. – 1968. – V. 55. – P. 573–575.
141a. *Denker M.* Asymptotic Distribution Theory in Nonparametric Statistics. – Braunschweig/Wiesbaden: Vieweg, 1985. – P. 1–204.
142. *Doksum K.A.* Measures of location asymmetry // Scand. J. Statist. and Appl. – 1975. – V. 2. – No. 1. – P. 11–22.
143. *Doksum K.A.* Starshaped transformations and the power of rank tests // Ann. Math. Statist. – 1969. – V. 40. – P. 1167–1176.
144. *Ferenholz L.T.* Von Mises Calculus for Statistical Functionals // Lecture Notes in Statistics. – 1983. – V. 19. – No. 8. – P. 1–124.
145. *Fraser D.A.S.* Nonparametric Methods in Statistics. – N.Y.: Wiley, 1957. – 229 p.
146. *Gastwirth J.L.*, *Rubin H.* The behavior of robust estimators and dependent data // Ann. Statist. – 1975. – V. 3. – P. 1070–1100.
147. *Gastwirth J.L.* On Asymptotic Relative Efficencies of a class of Rank Tests // J. Roy. Statist. Soc. – 1970. – P. 227–232.
148. *Gastwirth J.L.* On robust rank tests. Nonparametric Techniques in Stat. Inference. – Cambrige: Univ. Press., 1970. – P. 89–109.

149. *Ghosh J.K.* A new proof of the Bahadur representation of quantiles and application // Ann. Math. Statist. – 1971. – V. 42. – No. 6. – P. 1957–1961.
149a. *Gibbons J.D.* Nonparametric Statistical Inference. – N.Y.: McGraw-Hill, 1971.
150. *Hajek J.* Course in Nonparametric Statistics. – Academic Press, 1970. – 184 p.
151. *Hajek J.* Asymptotic normality of simple linear rank statistics under alternatives // Ann. Math. Statist. – 1968. – V. 39. – P. 325–346.
152. *Hampel F.R.* Contribution to the theory of robust estimation: Ph. D. diss. – Berkeley: Univ. California, 1968. – 103 p.
153. *Hampel F.R.* A general qualitative definition of robustness // Ann. Math. Statist. – 1971. – V. 42. – P. 1887–1896.
154. *Hampel F.R.* The influence curve and its role in robust estimation. J. Amer. Statist. Assoc. – 1974. – V. 69. – No. 346. – P. 383–393.
155. *Hettmansperger T.R.*, *Keenan M.A.* Tail weight statistical inference and families of distribution a brief survey // Mod. Course Statist. Distrib. Sci. Work. – Dordrecht – Boston, 1975. – P. 161–172.
156. *Hodges J.L.*, *Lehmann E.L.* Estimation of location based on rank tests // Ann. Math. Statist. – 1963. – V. 34. – P. 598–611.
157. *Hodges J.L.* Efficiency in normal samples and tolerance of extreme values for some estimates of location // Proc. 5th Berkeley Symp. Math. Statist. and Prob. 1965 – 1966. – V. I. – Berkeley – Los Angeles, 1967. – P. 163–186.
158. *Hoefding W.* A class of statistics with asymptotically normal distribution // Ann. Math. Statist. – 1948. – V. 19. – P. 292–325.
159. *Hogg R.V.* Adaptive robust procedures: A partial review and some suggestions for future applications and theory // J. Amer. Statist. Assoc. – 1974. – V. 69. – P. 909–923.
160. *Huber P.J.* Robust estimation of location parameter // Ann. Math. Statist. – 1964. – V. 35. – No. 1. – P. 73–101.
161. *Huber P.J.* Robust statistics: a review // Ann. Math. Statist. – 1972. – V. 43. – P. 1041–1067.
162. *Jaeckel L.A.* Robust estimation of location: symmetry and asymmetric contamination // Ann. Math. Statist. – 1971. – V. 42. – No. 3. – P. 1020–1034.
163. *Jaeckel L.A.* Some flexible estimates of location // Ann. Math. Statist. – 1971. – V. 42. – P. 1540–1552.
164. *Janssen P.*, *Serfling R.*, *Veraverbeke M.* Asymptotic normality for a general class of statistical functions and applications to measures of spread // Ann. Statist. – 1984. – V. 12. – No. 4. – P. 1369–1379.
165. *Janssen P.*, *Serfling R.*, *Veraverbeke M.* Asymptotic normality of U-statistics based on trimmed samples // J. Statist. Planning and Inference. – 1987. – V. 16. – P. 63–74.
166. *Jureckova J.* M-L-R-estimators. Handbook of Statistics. V. 4 / eds. P.R. Krishnaiah, P.K. Sen. – Elsevier Science Publishers, 1984. – P. 463–485.

167. *Jureckova J.* Asymptotic relations of M-estimates and R-estimates in linear regression model // Ann. Statist. – 1977. – V. 5. – P. 464–472.
168. *Jureckova J.* Tail-behaviour of location estimators // Ann. Statist. – 1981. – V. 9. – P. 578–585.
169. *Krishnaiah P.R.*, *Sen P.K.*, eds. Handbook of Statistics. V. 4. – Elsevier Science Publishers, 1984. – 967 p.
170. *Lawrence M.J.* Inequalities of S-ordered distributions // Ann. Statist. – 1975. – V. 3. – P. 413–428.
171. *Lehmann E. L.* Nonparametric: Statistical Methods Based on Ranks. Holden – Day. – San Francisco, 1975. – 326 p.
172. *Loh Wei-Yin.* Bounds on ARE's for restricted classes of distribution defined via tail-orderings // Ann. Statist. – 1984. – V. 12. – No. 2. – P. 685–701.
173. *Loh Wei-Yin.* Random quotients and robust estimation // Commun. Statist. Theory Meth. – 1984. – V. 13(22). – P. 2757–2769.
174. *Maritz J.S.*, *Wu M.*, *Staudte R.C. Jr.* A location estimator based on a U-statistic // Ann. Statist. – 1977. – V. 5. – No. 4. – P. 779–786.
175. *Von Mises R.* On the asymptotic distributions of differentiable statistical functions // Ann. Math. Statist. – 1947. – V. 18. – P. 309–348.
175a. *Noether G.E.* On a theorem of Pitman // Ann. Math. Statist. – 1955. – V. 26. – P. 64–68.
176. *Parr W.C.* Minimum distance estimation: a bibliography // Comm. Statist. – 1981. – V. A10. – P. 1205–1224.
177. *Parr W.C.* A note on Hajek protections an the influence curve // Statistics and Probability Letters. – 1983. – V. 1. – P. 177–179.
178. *Parr W.C.*, *De Wet.* On minimum weighted Cramer-von Mists statistical estimation // Comm. Statist. – 1981. – V. A10(12). – P. 1149–1166.
178a. *Parr W.S. and Schucany W.R.* Minimum distance and robust estimation // J. Amer. Statist. Assoc. – 1980. – V. 75. – No. 371. – P. 616–624.
179. *Portnoy S.L.* Robust estimation in dependent situations // Ann. Statist. – 1977. – V. 5. – P. 22–43.
180. *Puri M.L.*, *Sen P.K.* Nonparametric methods in Multivariate Analysis. – N.Y.: John Wiley, 1970. – 440 p.
181. *Ramberg J.S.*, *Schmeiser B.W.* An approximative method for generating symmetric random variables // Commun ACM. – 1972. – V. 15. – P. 987–990.
181a. *Randles R.H. and Wolfe D.A.* Introduction to the theory of nonparametric statistics. – N.Y.: Wiley, 1979.
182. *Reeds J.A.* On the definition of von Mises functionals: Ph. D. Thesis. Dept. of Statistics. – Harvard, 1976. – 178 p.
183. *Serfling R.J.* Approximation Theorems of Mathematical Statistics. – N.Y.: Wiley, 1980. – 371 p.
184. *Serfling R.J.* Generalized L-M-R-statistics // Ann. Statist. – 1984. – V. 12. – P. 76–86.

184а. *Sholz F.W.* A comparison of efficient location estimators // Ann. Statist. – 1974. – V. 2. – No. 6. – P. 1323–1326.
185. *Stigler S.M.* Simon Newcomb, Percy Daniel and history of robust estimations // J. Amer. Statist. Assoc. – 1973. – V. 68. – P. 872–879.
186. *Stigler S.M.* Linear function of order statistics with smooth weight functions // Ann. Statist. – 1974. – V. 2. – P. 676–693.
187. *Tate J.W.* On a double inequality of the normal distribution // Ann. Math. Statist. – 1953. – V. 24. – P. 132–144.
188. *Tukey J.W.* A survey of sampling from contaminated distributions // Contributions to Prob. and Statist. (Ingram Olkin ed.) – Stanford Univ. Press, 1960. – P. 448–485.
189. *Tukey J.W.* Bias and confidence in not-quite large samples. (Abstract) // Ann. Math. Statist. – 1958. – V. 29. – P. 614.
189а. *Tukey J.W.* «Data Analysis, Computation and Mathematics», Quarterly of Applied Mathematics, V. XXX, April 1972. – No. I, Special Issue: «Symposium on the Future of Applied Mathematics». – P. 51 – 65. (см. также: Современные проблемы математики: сборник: пер. с англ. – М.: «Знание», серия математика, кибернетика. – 1977. – № 12. – С. 41–64).
189б. *Tukey J.W.* Exploratory data analysis. – Addison Wesley, Reading. Mass. 1970 – 71.
190. *Van Zwet W.R.* Convex transformations of Random Variables // Math. Centrum. Amsterdam. – 1964.
191. *Wald A.* Note on the consistency of the maximum likelihood estimate // Ann. Math. Statist. – 1949. – V. 20. – P. 595–601.
192. *Wolfowitz J.* The minimum distance method // Ann. Math. Statist. – 1957. – V. 28. – P. 75–88.

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

$P\{A\}$ – вероятность события A

$F_X(x)$ – функция распределения (ф.р.) случайной величины (с.в.) X

$f_X(x)$ – плотность распределения вероятностей (п.р.в.) случайной величины X

$\vec{X} = (X_1,...,X_n)$ – случайный вектор, компоненты которого являются независимыми и одинаково распределенными (н.о.р.) случайными величинами ($X_1,...,X_n$ – выборка объема n)

$F_{\vec{X}}(x_1,...,x_n)$ – функция распределения случайного вектора $\vec{X}=(X_1,...,X_n)$

$f_{\vec{X}}(x_1,...,x_n)$ – плотность распределения вероятностей случайного вектора $\vec{X} = (X_1,...,X_n)$

$\vec{x} = (x_1,...,x_n)$ – наблюденная реализация наблюдаемого случайного вектора $\vec{X} = (X_1,...,X_n)$

$X_{(1)},...,X_{(n)}$ – порядковые статистики выборки $X_1,...,X_n$

$x_{(1)},...,x_{(n)}$ – вариационный ряд наблюденной реализации $\vec{x} = (x_1,...,x_n)$ случайного вектора $\vec{X} = (X_1,...,X_n)$

$\vec{R} = (R_1,...,R_n)$ – ранговый вектор выборки $X_1,...,X_n$

$\vec{r} = (r_1,...,r_n)$ – наблюденная реализация рангового вектора $\vec{R} = (R_1,...,R_n)$

$M(X)$ – математическое ожидание случайной величины X

$D(X)$ – дисперсия случайной величины X

$P\{A \mid B\}$ – условная вероятность события A при условии B

$M(X \mid Y = y)$ – математическое ожидание случайной величины X при условии, что случайная величина Y приняла значение y

$COV(X,Y)$ – ковариация случайных величин X и Y

ρ_{XY} – коэффициент корреляции случайных величин X и Y

$L(X) = N(a, \sigma^2)$ – закон распределения вероятностей с.в. X нормальный с параметрами a и σ^2

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{1}{2}x^2} dx$ – стандартная нормальная функция распределения вероятностей

$\Phi\{(x-a)/\sigma\}$ – функция распределения нормальной с. в. X с параметрами a и σ^2

$\Phi^{-1}(t)$ – обратная (квантильная) функция для нормальной функции распределения $\Phi(x)$

$\lambda_\alpha = \Phi^{-1}(\alpha)$ – квантиль заданного уровня α стандартного нормального распределения

$L(X) = St(n)$ – случайная величина X имеет распределение Стьюдента с числом степеней свободы n

$t_{\alpha,(n)}$ – квантиль уровня α распределения Стьюдента с числом степеней свободы n

$L(X) = \chi^2(n)$ – случайная величина X имеет хи-квадрат-распределение с числом степеней свободы n

$\chi^2_{\alpha,(n)}$ – квантиль уровня α хи-квадрат-распределения с числом степеней свободы n

$L(X) = F(\nu_1, \nu_2)$ – случайная величина X имеет F-распределение Снедекора – Фишера с числами степеней свободы ν_1 *и* ν_2

$F_{\alpha,(\nu_1,\nu_2)}$ – квантиль уровня α для F-распределения с числами степеней свободы ν_1 и ν_2

$L(X) = Bi(n, p)$ – случайная величина X имеет биномиальное распределение с параметрами n и p

ОГЛАВЛЕНИЕ

Часть 1

ПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

ПРЕДИСЛОВИЕ 3

Глава 1. **Основные понятия, определения и обозначения** 6

1.1. Понятие выборки и статистической модели, суть статистических задач 7
1.2. Типы статистических моделей 12
1.2.А. Параметрические модели 12
1.2.Б. Непараметрические модели 13
1.2.В. Модели робастной статистики. Понятие супермодели 13
1.3. Вероятностное описание случайных величин 15
1.3.А. Непрерывный тип 16
1.3.Б. Дискретный тип 20
1.3.В. Двумерный вариант, дискретный тип 23
1.3.Г. Двумерный вариант, непрерывный тип 24
1.3.Д. Двумерный вариант, смешанный тип 31
1.3.Е. Многомерный вариант, непрерывный тип 33
1.4. Упражнения, задачи и дополнения 36

Глава 2. **Характеристика подходов и методов теории оценок параметров** 73

2.1. Основные понятия теории оценок параметров 74
2.2. Свойства точечных оценок параметров 76
2.3. Функция правдоподобия и количество информации Фишера 79
2.4. Неравенство Рао – Крамера 88
2.4.А. Критерий эффективности 90
2.5. Достаточные статистики 99
2.5.А. Критерий факторизации 104

2.5.Б. Эффективные оценки для экспоненциального семейства распределений 111
2.6. Асимптотическая эффективность оценок параметров . 112
2.7. Методы построения статистических оценок 115
2.7.А. Эмпирическая функция распределения (непрерывный случай) 116
2.7.Б. Дискретный случай 120
2.7.В. Многомерный вариант 123
2.8. Методы построения точечных оценок параметров 123
2.8.А. Метод подстановки 124
2.9. Метод моментов 130
2.10. Метод максимального правдоподобия 136
2.10.А. Общие свойства МП-оценок 138
2.10.Б. Метод накопления для приближенного вычисления МП-оценок 144
2.10.В. Асимптотические свойства МП-оценок 148
2.10.Г. Состоятельность оценок максимального правдоподобия 148
2.10.Д. Эффективность и асимптотическая нормальность МП-оценок 151
2.11. Метод минимума расстояний 160
2.12. Интервальные оценки параметров. Доверительные интервалы 163
2.13. Асимптотический подход к построению доверительных интервалов 175
2.14. Доверительные интервалы, основанные на распределениях точечных оценок параметров 180
2.15. Упражнения, задачи и дополнения 186

Глава 3. **Проверка статистических гипотез** 236

3.1. Основные понятия теории проверки статистических гипотез 236
3.1.А. Понятие статистического критерия и критической области 239
3.1.Б. Вероятности ошибок и мощность критерия 245
3.1.В. Общие свойства критериев и их сравнение 252

3.2. Примеры формулировок статистических гипотез 260
3.2.А. Параметрические гипотезы 260
3.2.Б. Непараметрические гипотезы 263
3.3. Параметрические гипотезы. Подход Неймана – Пирсона 265
3.3.А. Непрерывный случай 266
3.3.Б. Дискретный случай 274
3.4. Параметрические сложные гипотезы 280
3.4.А. Статистические модели с монотонным отношением правдоподобия 281
3.4.Б. Критерий отношения правдоподобия 286
3.4.В. Асимптотические свойства критерия отношения правдоподобия 298
3.4.Г. О связи задач проверки гипотез и интервальных оценок параметров 304
3.4.Д. Общие выводы относительно свойств одновыборочных критериев проверки параметрических гипотез 310
3.5. Критерии согласия 315
3.5.Б. χ^2-критерий согласия в непрерывном случае 321
3.5.В. χ^2-критерий согласия для сложной гипотезы 323
3.5.Г. Критерий квантилей 331
3.6. Критерии однородности 334
3.6.А. Двухвыборочный критерий отношения дисперсий 335
3.6.Б. Критерий равенства дисперсий 339
3.6.В. Двухвыборочные критерии проверки гипотезы о равенстве математических ожиданий нормальных распределений 345
3.6.Г. Парный *t*-критерий Стьюдента 354
3.6.Д. Критерий однородности χ^2 359
3.7. Критерии независимости и случайности 365
3.7.А. Проверка гипотезы независимости в гауссовском случае 366
3.7.Б. Проверка гипотезы о «силе» связи между случайными величинами X и Y в гауссовском случае 369

3.7.В. Статистический анализ таблиц сопряженности атрибутивных признаков 372
3.7.Г. Задача согласия 374
3.7.Д. Проверка гипотезы независимости атрибутивных признаков 375
3.7.Е. Проверка гипотезы случайности 381
3.8. Основные понятия дисперсионного анализа 391
3.8.А. Однофакторный дисперсионный анализ: модель I 394
3.8.Б. Доверительные интервалы 407
3.8.В. Метод множественных сравнений Шеффе 409
3.8.Г. Однофакторный дисперсионный анализ: модель II 413
3.8.Д. Двухфакторный дисперсионный анализ 425
3.9. Упражнения, задачи и дополнения 438
3.9.А. Задачи и дополнения к разделу (3.8) 481

ЛИТЕРАТУРА 518

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ 530

Часть 2

НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

Глава 4. **Статистические свойства упорядоченной статистики и ранговых векторов**

4.1. Математическое описание статистических данных
4.1.А. Упорядоченная статистика
4.1.Б. Понятие ранга и антиранга
4.2. Статистические свойства порядковых статистик
4.2.А. Плотность распределения вероятностей упорядоченной статистики
4.2.Б. Функция распределения вероятностей *r*-й порядковой статистики
4.2.В. Плотность распределения вероятностей *r*-й порядковой статистики
4.2.Г. Совместная плотность распределения вероятностей *r*-й и *s*-й порядковых статистик
4.3. Числовые характеристики порядковых статистик

4.3.А. Вероятностное интегральное преобразование
4.3.Б. Моменты порядковых статистик для равномерного распределения в интервале [0,1]
4.3.В. Приближенные формулы для вычисления моментов порядковых статистик
4.4. Распределение вероятностей выборочного квантиля
4.4.А. Асимптотическая нормальность оценки квантиля
4.4.Б. Состоятельность выборочного квантиля
4.5. Вероятностное описание рангов наблюдений
4.5.А. Различные варианты определения рангов наблюдений
4.6. О статистической связи между наблюдением и его рангом
4.6.А. Функции регрессии наблюдений и их рангов
4.6.Б. Коэффициент корреляции между наблюдениями и их рангами
4.7. Непараметрические критерии согласия и однородности, основанные на порядковых статистиках
4.7.А. Критерии Колмогорова – Смирнова
4.7.Б. Критерии Крамера – Мизеса – Смирнова и Андерсона – Дарлинга
4.7.В. Критерии Колмогорова и Крамера – Мизеса для сложной гипотезы
4.8. Критерии нормальности
4.9. Непараметрические доверительные интервалы и толерантные пределы, основанные на порядковых статистиках
4.9.А. Непараметрические доверительные интервалы для квантиля заданного уровня
4.9.Б. Непараметрические толерантные пределы
4.10. Критерии отбраковки выбросов, основанные на порядковых статистиках
4.10.А. Статистические модели для описания выбросов
4.10.Б. Критерии Граббса (параметры μ и σ^2 известны)
4.10.В. Критерии Томпсона (параметры μ и σ^2 неизвестны)
4.10.Г. Критерии Диксона
4.10.Д. Критерии Титьена – Мура
4.11. Упражнения, задачи и дополнения

Глава 5. **Некоторые непараметрические критерии, основанные на рангах и знаках наблюдений**

5.1. Критерий знаков
5.1.А. Одновыборочный случай
5.1.Б. Парные сравнения
5.2. Критерий знаковых рангов Уилкоксона

5.3. Двухвыборочный ранговый критерий Уилкоксона
5.4. Ранговые критерии с метками общего вида
5.4.А. Статистики ранговых меток в одновыборочном случае
5.4.Б. Эффективность одновыборочных критериев, основанных на статистиках с метками общего вида
5.4.В. Двухвыборочный вариант с альтернативой сдвига
5.4.Г. Эффективность двухвыборочных критериев, основанных на статистиках с метками общего вида
5.5. Ранговые критерии для линейной регрессии и гипотезы независимости
5.5.А. Линейная регрессия
5.5.Б. Ранговые критерии независимости
5.6. Ранговый критерий Краскела – Уоллиса
5.7. Ранговые критерии Фридмана и Пейджа
5.7.А. Критерий Фридмана
5.7.Б. Критерий Пейджа
5.8. Упражнения, задачи и дополнения

ПРИЛОЖЕНИЯ

П.1. Дискретные распределения
П.2. Непрерывные распределения
П.2.А. Таблицы распределений

ЛИТЕРАТУРА

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

Ч а с т ь 3

РОБАСТНАЯ СТАТИСТИКА

Г л а в а 6. **Основные понятия теории робастного оценивания**

6.1. Что такое – робастная статистика? Об устойчивости статистических выводов, смысл термина «робастность»
6.2. О робастности и «неробастности» стандартных статистических процедур
6.3. Различные варианты задания супермоделей
6.4. Подходы к определению робастных процедур
6.4.А. Качественный подход к определению робастных процедур
6.4.Б. Количественный подход к определению робастных процедур
6.5. Функция влияния и числовые характеристики устойчивости оценок
6.5.А. Функция влияния статистик критериев
6.6. Связи между различными понятиями робастности оценок

Глава 7. **Методы анализа асимптотических распределений статистик**

7.1. Дифференциальный подход Мизеса к анализу асимптотических свойств статистик
7.2. Примеры анализа асимптотических свойств конкретных статистик методом Мизеса
7.3. Метод проекций. Связь функции влияния Хампеля с функцией, определяющей проекцию Гаека
7.4. Некоторые понятия и результаты теории U-статистик

Глава 8. **Основные типы робастных оценок функционалов**

8.1. Оценки типа максимального правдоподобия (M-оценки)
8.2. L-оценки в виде линейной комбинации порядковых статистик
8.3. R-оценки, основанные на использовании ранговых критериев
8.3.А. R-оценки
8.3.Б. R_α-оценки
8.3.В. Предел устойчивости R_α-оценок
8.4. Обобщенные L-оценки
8.5. Обобщенные L-оценки, основанные на урезанных выборках ($GL_{\alpha\beta}$-оценки)
8.6. U-статистики, основанные на урезанных выборках ($U_{\alpha\beta}$-оценки)
8.7. Оценки параметров, построенные методом минимума расстояний
8.8. MD-оценки, основанные на урезанных выборках (MD_α-оценки)
8.9. Упражнения, задачи и дополнения

Глава 9. **Оценки параметра положения и их сравнение**

9.1. Определение параметра положения
9.2. Связи между M-, L- и R-оценками параметра положения
9.3. Связи между MD-оценками и M-, L- и R-оценками параметра положения
9.4. Сравнение асимптотически эффективных оценок параметра положения
9.5. Сравнение R- и R_α-оценок параметра положения
9.6. Относительные эффективности оценок и их границы для супермоделей с упорядоченными распределениями
9.7. Адаптивные оценки параметра положения
9.8. Упражнения, задачи и дополнения

Глава 10. **Оценки масштабного параметра**

10.1. Различные меры масштабного параметра
10.2. Сравнение оценок масштабного параметра
10.3. Урезанные варианты стандартного отклонения и среднего абсолютных отклонений

10.4. Интер-α-квантильные размахи
10.5. Медиана абсолютных разностей
10.6. Средняя разность Джини и её урезанный вариант
10.7. Упражнения, задачи и дополнения

ПРИЛОЖЕНИЯ

П.3. Некоторые методы и теоремы
П.3.А. Виды сходимости последовательностей случайных величин
П.3.Б. Неравенства
П.3.В. Предельные теоремы
П.3.Г. Теорема Кокрена о разбиении суммы квадратов нормальных случайных величин
П.3.Д. Метод проекций
П.3.Е. Асимптотическая эффективность Питмена
П.3.Ж. Процедура Ходжеса – Лемана

ЛИТЕРАТУРА

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ

Шуленин Валерий Петрович

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Часть 1. Параметрическая статистика

Учебник

Редактор *Т.С. Портнова*
Дизайн, верстка *Д.В. Фортеса*

Изд. лиц. ИД № 04000 от 12.02.2001. Подписано к печати 23.12.2011.
Формат 60 × 84 $^{1}/_{16}$. Бумага офсетная. Печать офсетная. Гарнитура «Таймс».
Усл. п. л. 31,39. Уч.-изд. л. 35,15. Тираж 500 экз. Заказ № 3.

ООО «Издательство научно-технической литературы»
634050, Томск, пл. Новособорная, 1, тел. (3822) 533-335

Отпечатано в типографии ЗАО «М-Принт», г. Томск, ул. Пролетарская, 38/1