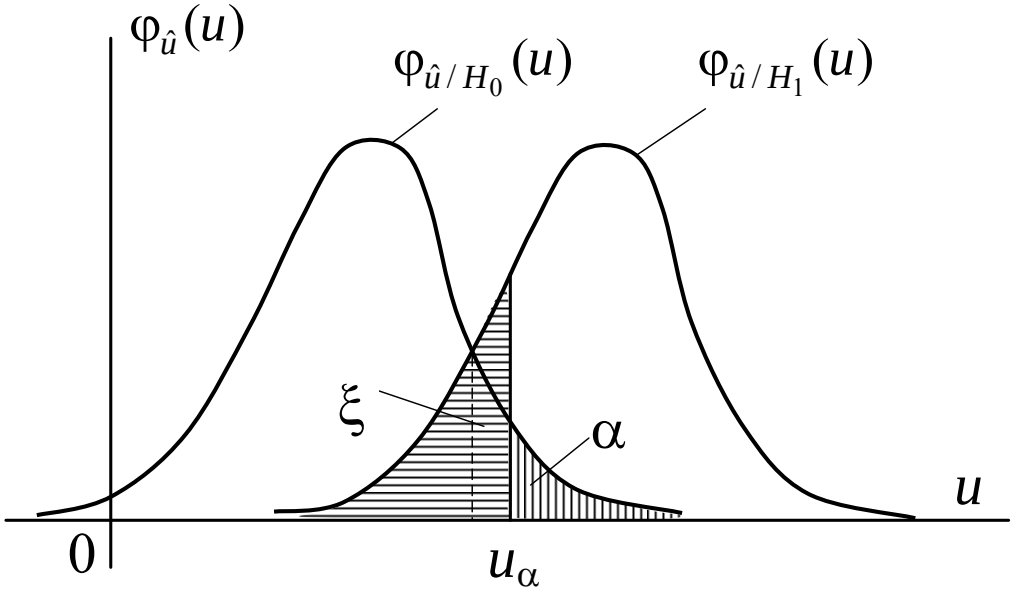
ФЕДЕРАЛЬНОЕ АГЕНСТВО ПО ОБРАЗОВАНИЮ
Государственное образовательное учреждение
Высшего профессионального образования
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ

СЕНЬЧЕНКОВ В. И.

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Учебное пособие



Санкт-Петербург - 2006

Рассматриваются вопросы обработки экспериментальной информации, теоретическую основу которых составляют методы параметрической статистики. Особое внимание уделено статистическому оцениванию законов и параметров распределения случайных объектов, формированию и проверке статистических гипотез. Достаточно подробно изложен один из методов статистического анализа экспериментальных данных - регрессионный анализ. Приводится большое количество примеров, иллюстрирующих практическое применение теоретических положений.

Учебное пособие ориентировано на студентов вечерней и заочной форм обучения, может быть использовано ими для самостоятельной проработки материала по дисциплине “Обработка экспериментальных данных”.

Рецензенты: доктор технических наук профессор В. И. ХИМЕНКО;

заслуженный деятель науки РФ
доктор технических наук профессор А. К. ДМИТРИЕВ;

доктор технических наук профессор В. В. ПОПОВИЧ.

ОГЛАВЛЕНИЕ

1. ПРОБЛЕМА СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1. Наблюдение и измерение

Одним из основных способов изучения окружающего мира, средством познания было и остаётся наблюдение. Под наблюдением принято понимать целенаправленное восприятие свойств процессов и явлений. Опираясь на результаты наблюдения, исследователь строит физическую модель новых процессов и явлений, выдвигает научные гипотезы, создаёт теории, принимает решения. Наблюдение, как правило, сопровождается определённым преобразованием и регистрацией информации о свойствах наблюдаемого объекта.

Для повышения эффективности процесса наблюдения могут использоваться различные технические средства. Наблюдение присуще любому эксперименту. Фактически, эксперимент – это и есть создание условий для наиболее эффективного наблюдения тех или иных свойств изучаемого объекта. Поэтому каждый экспериментатор является в первую очередь наблюдателем. Учитывая большую, во многих случаях решающую роль эксперимента в современных науках, особенно в технических и экономических, можно утверждать, что понятие эксперимента является фундаментальным.

Принятие любого решения невозможно без наличия необходимой информации. Так, в случае управления нужна информация о состоянии управляемого объекта, окружающей среды и цели управления. Такая информация не может быть получена без наблюдения. В то же время от наблюдения до принятия решения информация претерпевает, как правило, существенные преобразования. Наиболее характерными этапами таких преобразований являются измерения и обработка результатов измерений.

В широком смысле измерение состоит в сравнении наблюдаемой величины с эталоном и получении в результате этого её численного значения. Если результаты наблюдений представлены количественно, то имело место измерение. Нередко, кроме измерения, различают подсчёт. Данную операцию можно квалифицировать как завершающий этап наблюдения – регистрацию величин дискретного типа (число студентов в аудитории, количество проросших зёрен на единице площади посева, доход работника фирмы и т.п.). Тогда измерения можно рассматривать как регистрацию величин непрерывного типа (вес, расстояние, скорость и т.д.).

Часто понятия наблюдения и измерения отождествляются. Представляется, что в рамках настоящей брошюры это не принципиально и вполне допустимо. Более того, без особой оговорки, используя как синонимы термины «наблюдение» и «измерение», будем включать в их содержание и подсчёт. Действительно, статистические методы обработки применимы к данным, получаемым как путём измерения, так и подсчёта. В дальнейшем, в силу указанных выше причин, все эти данные будут называться экспериментальными данными.

Говоря об экспериментальных данных, будем предполагать, что речь идёт о числовых величинах, векторах или функциях, т.е. о результатах количественных наблюдений, получаемых путём измерения или подсчёта. Если данные не могут быть представлены количественно, а являются качественными характеристиками, высказываниями, утверждениями, то их обработку следует предоставить специалистам по различным логическим методам.

Задачей обработки данных во многих случаях является принятие решения относительно значений определённых параметров (величин), характеризующих изучаемые явления или процессы. Обозначим эти параметры $a_1, a_2,..., a_m$. При этом возможны два случая. В первом из них непосредственно измеряются указанные величины. Говорят, что имеют место прямые измерения.

Во втором случае величины $a_1, a_2,..., a_m$ непосредственно измерены быть не могут, а измеряются другие переменные $x_1, x_2,..., x_n$, с которыми функционально связаны величины $a_1, a_2,..., a_m$:

$$x_i = f_i(a_1, a_2,..., a_m), \quad i = \overline{1,n}. \tag{1.1.1}$$

Выражение (1.1.1) в векторной форме запишем как

$$X_{<n>} = F_{<n>}(A_{<m>}), \tag{1.1.2}$$

где $X_{<n>} = (x_1, x_2,..., x_n)^{\mathsf{T}}$; $F_{<n>} = (f_1, f_2,..., f_n)^{\mathsf{T}}$; $A_{<m>} = (a_1, a_2,..., a_m)^{\mathsf{T}}$.

Эти измерения называются косвенными. В общем случае функциональные связи (1.1.1) являются нелинейными. Если же схема косвенных измерений линейна, то имеют место соотношения

$$x_i = \sum_{j=1}^{m} u_{ij} a_j, \quad i = \overline{1,n}$$

или в векторно-матричной форме

$$X_{<n>} = U_{[n;m]} A_{<m>}. \tag{1.1.3}$$

Во многих случаях свойства объекта, над которым ставится эксперимент, изменяются с течением времени. Тогда соответственно вместо соотношений (1.1.2) и (1.1.3) при непрерывных измерениях имеем

$$X_{<n>} = F_{<n>}(t; A_{<m>}),$$
$$X_{<n>} = U_{[n;m]}(t) A_{<m>},$$

а при дискретных измерениях

$$X_{<n>_i} = F_{<n>}(t_i; A_{<m>}),$$
$$X_{<n>_i} = U_{[n;m]}(t_i) A_{<m>}.$$

Предполагается, что непрерывные и дискретные измерения проводятся в моменты времени, лежащие внутри интервала [0, T], т.е. $t \in [0, T]$, $t_i \in [0, T]$, $i = \overline{1, N}$.

Результат измерения всегда содержит ошибку (погрешность), представляющую собой отклонение результата измерения от истинного значения. Данное утверждение не относится в полной мере к подсчёту. Подсчитать количество интересующих объектов можно безошибочно. Так, в нормальных условиях можно точно определить количество компьютеров в интернет-классе. Редактор безошибочно может сосчитать число страниц рукописи и т.д. В то же время в некоторых условиях (дефицит времени, несовершенство средства наблюдения и др.) можно допустить ошибки при подсчёте. Например, с вертолёта трудно подсчитать точно количество коров в большом стаде.

Существующая в настоящее время теория ошибок измерений развита в основном применительно к оцениванию непрерывных величин. При этом ошибки принято разделять на систематические и случайные.

Систематические ошибки – это составляющие общей ошибки, вызываемые факторами, действующими одинаковым образом при многократном повторении одних и тех же измерений. Эти ошибки при повторных измерениях остаются неизменными или, если изменяются, то закономерно. Систематические ошибки, как правило, обусловлены погрешностями измерительных приборов (инструментальные ошибки) и несовершенством методов измерений (методические ошибки).

Случайные ошибки – составляющие общей ошибки, изменяющиеся случайным образом при повторных измерениях одной и той же величины. Причиной случайных ошибок являются неконтролируемые факторы, проявление которых неодинаково в каждом измерении и которые заранее не могут быть учтены. Другими словами, случайные ошибки проявляются тогда, когда при измерениях имеют место случайные события. Объективно такие события при наблюдениях и измерениях происходят всегда, как бы ни повышалась «чистота» эксперимента.

Среди случайных ошибок особо следует выделить грубые ошибки (промахи). Их характер и причины существенно отличаются от характера других случайных ошибок измерений. Основная масса случайных ошибок появляется при исправно работающих средствах измерений и правильных действия экспериментатора. Причиной появления грубых ошибок являются неисправность приборов или неточность в работе экспериментатора.

В данной брошюре рассматриваются методы обработки экспериментальных данных, содержащих случайные ошибки (исключая грубые). Обозначим вектор ошибок через $\hat{Z}(t)$. Тогда при их аддитивном учёте и прямых измерениях выражение вектора экспериментальных данных принимает вид

$$\hat{X}_{<m>} = A_{<m>} + \hat{Z}_{<m>}; \qquad (1.1.4)$$

в случае косвенных измерений

$$\hat{X}_{<n>} = F_{<n>}(A_{<m>}) + \hat{Z}_{<n>}. \qquad (1.1.5)$$

При мультипликативном учёте ошибок можно записать

$$\hat{X}_{<m>} = (a_1\hat{z}_1,\ a_2\hat{z}_2,...,a_m\hat{z}_m)^{\mathsf{T}}; \qquad (1.1.6)$$

$$\hat{X}_{<n>} = \left(\hat{z}_1 f_1(A_{<m>}),\ \hat{z}_2 f_2(A_{<m>}),...,\hat{z}_n f_n(A_{<m>})\right)^{\mathsf{T}}. \qquad (1.1.7)$$

1.2. Обобщённая модель обработки экспериментальных данных

Цель обработки данных эксперимента заключается в получении из них сведений о свойствах изучаемого объекта или процесса. На заключительном этапе экспериментатор принимает решение относительно этих свойств. Данное решение может быть связано, например, с оцениванием конкретных значений величин, характеризующих свойства объекта, а также с проверкой предположений о нахождении этих величин в определённых пределах, предположений о возможных законах распределения и т.д. Следует отметить, что процедура оценивания связана с получением оценок характеристик объекта. Под ***оценкой*** будем понимать приближённое значение оцениваемой величины, которое целесообразно принимать за её истинное значение. На рис.1.1 приведена обобщённая модель обработки экспериментальных данных.

Состояние (свойства) исследуемого объекта или процесса абстрагируется пространством ситуаций $\{E\}$. При этом под $E \in \{E\}$ в общем случае понимается абстрактный параметр, характеризующий ситуацию, т.е. состояние объекта (фазовые координаты, структуру и т.д.). Абстрактный параметр может выражаться числом, функцией, функционалом, оператором, отношением, событием и т.п. Пространство ситуаций при оценивании вектора $A_{<m>}$ представляет множество всех возможных его значений. Экспериментатора интересует истинное значение вектора $A_{<m>}$.

В процессе наблюдений регистрируется случайный вектор $\hat{X}$, который связан с вектором A выражениями (1.1.4) – (1.1.7). В общем виде истинную связь обозначим соотношением $X = N(E)$. Случайные воздей-

ствия (помехи) $\hat{Z}$, образующие пространство воздействий $\{\hat{Z}\}$, искажают эту связь и она принимает вид $\hat{X} = N(E; \hat{Z})$.

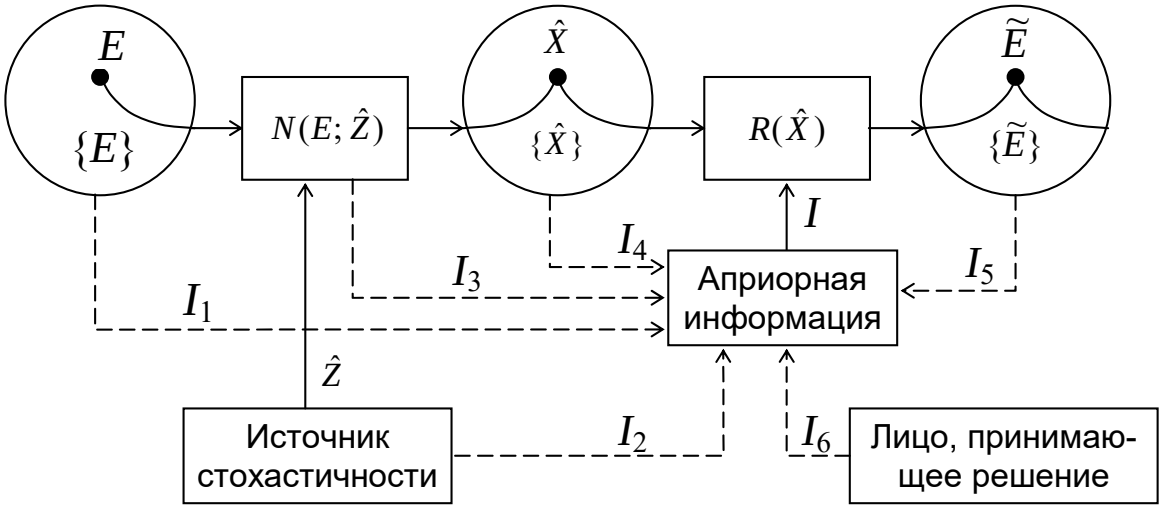


Рис.1.1. Обобщённая модель обработки экспериментальных данных

Помехи $\hat{Z}$ генерируются источником стохастичности. Операция над параметром E и случайной величиной (процессом) $\hat{Z}$, в результате которой формируется вектор $\hat{X}$, обозначена на схеме $N(E; \hat{Z})$.

Множество $\{\hat{X}\}$ всех значений $\hat{X}$, которые могут быть реализованы при всех возможных значениях E и $\hat{Z}$, составляет пространство наблюдений. Информация о параметре E из вектора $\hat{X}$ получается при обработке последнего. Статистическую процедуру преобразования над вектором $\hat{X}$ обозначим через $R(\hat{X})$. Величина I характеризует всю совокупность априорной информации, которая используется в статистической процедуре совместно с результатами наблюдений $\hat{X}$. Следует подчеркнуть, что эта информация известна до начала эксперимента. Результатом преобразования величин $\hat{X}$ и I является принятие решения о параметре E. Поэтому функцию R можно назвать решающей функцией.

Как видно из рис.1.1 (пунктирные линии), априорная информация I_1, I_2, I_3, I_4, I_5, I_6 может отражать определённые свойства пространства ситуаций, источника стохастичности, оператора N, пространства наблюдений, пространства решений и требований лица, принимающего решение.

Множество всех решений $\tilde{E} = R(\hat{X})$ образует пространство решений $\{\tilde{E}\}$. Идеальный исход обработки экспериментальных данных сводится к тождеству $\tilde{E} \equiv E$. Однако на практике это тождество оказывается нереализуемым. Решение $\tilde{E}$ является случайной величиной (вектором, функцией), с помощью которой экспериментатор оценивает интересующую его величину E. При этом степень близости данной оценки к истин-

ному значению E определяется как характером априорной информации об условиях эксперимента и о задаче обработки данных, так и особенностями процедуры принятия решений.

Выше указывалось, что помехи и регистрируемые в процессе наблюдения сигналы представляют собой случайные объекты (величины, векторы, функции). Из теории вероятностей известно, что самой полной характеристикой случайного объекта является закон распределения. Наиболее употребительны формы закона распределения в виде функции распределения $F_{\hat{x}}(x)$ и плотности распределения $\varphi_{\hat{x}}(x)$. В рассматриваемом случае принципиальное влияние на выбор способов обработки и качества обработки имеют законы распределения на пространствах $\{\hat{Z}\}$ и $\{\hat{X}\}$, а также оператор $N(E;\hat{Z})$. Обычно $N(E;\hat{Z})$ предполагается известным. Наиболее распространён случай, когда по отношению к помехе $\hat{Z}$ оператор N является оператором суммирования (помеха аддитивная).

В зависимости от того, известны законы распределений на пространствах $\{\hat{Z}\}$ и $\{\hat{X}\}$ или нет, статистические методы обработки экспериментальных данных подразделяются на параметрические (классические) и непараметрические.

Параметрические методы используются тогда, когда законы распределений на указанных пространствах известны. Однако, на практике они не всегда априори известны. При постановке экспериментов для исследования новых явлений и процессов истинные законы распределений могут быть неизвестны вообще. В других случаях условия эксперимента и характер помех настолько сложны и нестабильны, что трудно говорить о конкретных законах непосредственно в период проведения экспериментальных работ. Так, например, в задачах приёма и обработки телеметрической информации о техническом состоянии бортовых устройств космических аппаратов характер помех меняется в зависимости от времени суток, взаимного расположения пункта приёма и аппарата на орбите, наличия целенаправленных возмущающих воздействий. В настоящее время проблема обработки экспериментальных данных при неизвестных законах распределений решается двумя путями.

Во-первых, используется минимаксный подход. При этом решаемая задача фактически сводится к параметрической, так как данный подход приводит к нахождению закона распределения на $\{\hat{Z}\}$, в некотором смысле «наилучшего среди плохих» (максимин) или «наихудшего среди хороших» (минимакс).

Во-вторых, применяются методы непараметрической статистики.

Во многих научных дисциплинах (исследование операций, теория принятия решений, теория игр и т.д.) по информированности специали-

ста, решающего задачу, различают три уровня – детерминированный, статистический и неопределённый.

Наиболее простым является первый уровень – детерминированный, когда условия решения задачи известны полностью и случайные факторы отсутствуют. Применительно к обобщённой модели на рис.1.1 это означает, что операторы N и R точно известны, Z – неслучайная величина, все измерения проводятся абсолютно точно. Крайний случай этой ситуации – достоверно известно значение параметра E. Говорят, что решение задачи на детерминированном уровне производится в условиях определённости.

На втором, стохастическом уровне известны множество возможных ситуаций и априорное вероятностное распределение случайных факторов и этих ситуаций. В обобщённой модели обработки экспериментальных данных (см. рис.1.1) это касается законов распределений величин $\hat{Z}$ и $\hat{X}$. На стохастическом уровне информированности специалист находится в «условиях риска».

Наконец, на уровне неопределённости известно множество ситуаций, но отсутствует априорная информация о распределениях. Принято говорить, что задача решается в условиях неопределённости.

Последние два уровня обобщаются термином «условия неполной информации». Статистические методы обработки экспериментальных данных используются именно в условиях неполной информации.

При этом параметрические методы в основном используются на стохастическом уровне, а непараметрические – на уроне неопределённости. Утверждение «в основном» подчёркивает, что в принципе и параметрические и непараметрические методы при необходимости могут быть использованы на обоих рассматриваемых уровнях неполной информации.

Многие задачи обработки данных решаются в условиях, когда относительно исходной информации (характеристик исследуемого объекта, ошибок измерений и т.д.) выдвигаются определённые гипотезы. В действительности эти гипотезы могут отличаться от фактических, реально существующих характеристик. Иначе говоря, модель может оказаться неадекватной реальным процессам. В связи с этим необходимо отметить, что различные статистические процедуры обладают различной чувствительностью к такой неадекватности. Чувствительность означает степень влияния экспериментальных данных на результаты их обработки. Поскольку непараметрические методы требуют наименьшего объёма априорной информации, то очевидно, что они и менее чувствительны к искажениям исходных данных.

Выше отмечалось, что экспериментальная информация может моделироваться с помощью случайных величин, векторов или функций. Случайная функция представляет собой наиболее общую модель наблю-

даемого сигнала. Если аргументом случайной функции является время, то для неё используется термин «случайный процесс». Значительная часть результатов экспериментов описывается случайными процессами.

При наиболее распространённой аддитивной ошибке модель непрерывного сигнала имеет вид

$$\hat{X}(t) = \hat{F}(t; A) + \hat{Z}(t),$$

где $\hat{Y} = \hat{F}(t; A)$ – полезная составляющая наблюдаемого сигнала; $A_{<m>}$ - вектор оцениваемых параметров.

Характер полезной составляющей может быть использован как один из признаков для классификации моделей наблюдения. При этом можно выделить следующие типы полезных сигналов: детерминированный сигнал с неизвестными параметрами, случайный сигнал с известной функцией распределения (с точностью до параметров) и случайный сигнал с неизвестной функцией распределения (функция распределения может быть задана только классом распределений).

1.3. Сущность выборочного метода

При рассмотрении обобщённой модели обработки экспериментальных данных отмечалось, что пространство наблюдений $\{\hat{X}\}$ представляет собой множество реализаций вектора $\hat{X}$, или выборочное множество. Остановимся подробнее на понятии выборки, так как оно играет фундаментальную роль в статистических методах обработки информации как параметрических, так и непараметрических.

При экспериментальных исследованиях закономерностей в массовых случайных явлениях предполагается, что опыты могут быть повторены большое число раз при одинаковых условиях. В каждом опыте регистрируется определённый признак изучаемого объекта. Различают общий и основной признаки. ***Общим признаком*** называется свойство, по которому объекты объединяются в однородные совокупности, а ***основным признаком*** – свойство объектов, исследуемое в данном эксперименте. Под однородной будем понимать совокупность, все элементы которой являются реализациями одной и той же случайной величины (функции) с одним и тем же законом распределения.

Если производится исследование веса совокупности однотипных деталей (например, гаек), то тип деталей (гайки) характеризует их общий признак, а вес деталей – основной признак.

Отдельное конкретное значение наблюдаемого основного признака называется его реализацией или вариантом. При статистическом исследовании вероятностных свойств совокупности объектов нет возможности производить опыты над каждым из них. Так, при изучении роста ежеме-

сячных доходов граждан РФ невозможно установить доход каждого гражданина за ограниченное время. Для определения всхожести зерновых перед посевной бессмысленно пытаться обследовать каждое отдельное зерно. Всё же, несмотря на это, существует метод, который позволяет изучить интересующие свойства всей совокупности объектов. Речь идёт о выборочном методе, согласно которому основные признаки совокупности объектов изучаются по некоторой её части, называемой выборкой. Более строго в математической статистике выборкой называют совокупность наблюдаемых реализаций основного признака.

Совокупность всех возможных объектов (вариантов), из которых производится выборка, называется генеральной совокупностью. Следует отметить, что выборка является однородной, если все её элементы извлечены из одной генеральной совокупности.

Обычно предполагается, что выборки формируются при многократной реализации случайного эксперимента, результат которого нельзя заранее точно предсказать. Поэтому такие выборки называются случайными.

Пусть количественно исследуемый основной признак описывается случайной величиной $\hat{x}$. Допустим, что в процессе эксперимента получена последовательность из n значений $x_1,\ x_2,\ldots, x_n$ случайной величины $\hat{x}$. До проведения эксперимента эта последовательность является случайной выборкой и обозначается $\hat{x}_1,\ \hat{x}_2,\ldots,\hat{x}_n$.

Если основой признак наблюдаемого объекта описывается случайной функцией $\hat{x}(t)$, то в процессе эксперимента получаем конечную совокупность реализаций $x_1(t),\ x_2(t),\ldots, x_n(t),\ t \in [0; T]$ случайной функции. Случайная выборка представляет в этом случае последовательность случайных функций $\hat{x}_1(t),\ \hat{x}_2(t),\ldots,\hat{x}_n(t)$. Часто выборку целесообразно описывать с помощью n-мерного случайного вектора $\hat{X}_{<n>}$, компонентами которого являются элементы упорядоченной последовательности $\hat{x}_1$, $\hat{x}_2,\ldots,\ \hat{x}_n$.

Случайный характер выборки выражается в том, что нельзя заранее предсказать возможные значения её элементов, и любые две последовательности из n наблюдаемых значений в общем случае будут различными. В конкретных прикладных задачах элементы выборок представляют собой реализации случайных величин, т.е. детерминированные величины. Таким образом, априорно (до проведения эксперимента) выборка будет случайной, а апостериорно (после проведения опыта) – неслучайной.

Число n элементов выборки является конечным и называется ***объёмом выборки***. Число элементов генеральной совокупности может быть конечным или бесконечным.

В статистике различают повторные и бесповторные выборки. Выборка называется ***повторной***, если отобранный объект после испытания перед отбором следующего объекта снова возвращается в генеральную совокупность. Выборка называется ***бесповторной***, если отобранный объект после испытания не возвращается в генеральную совокупность.

С позиций теории вероятностей элементы случайной выборки рассматриваются как независимые случайные величины с одной и той же функцией распределения $F_{\hat{x}}(x)$ и плотностью распределения $\varphi_{\hat{x}}(x)$. Последнее означает, что для плотности распределения случайного вектора $\hat{X}_{<n>}$ имеет место равенство

$$\varphi_{\hat{X}}(X_{<n>}) = \prod_{i=1}^{n} \varphi_{\hat{x}_i}(x).$$

Далеко не всякая выборка адекватно отражает свойства генеральной совокупности. Убедительный пример: требуется оценить средний рост жителей некоторого города, а в качестве выборки исследователю предлагают городскую баскетбольную команду. Нетрудно понять, насколько будет искажён результат.

Говорят, что выборка должна быть ***представительной*** или ***репрезентативной***. Выборка представительна, если все элементы генеральной совокупности имеют одинаковую вероятность быть выбранными. Чтобы обеспечить это, не имея никаких сведений о генеральной совокупности, можно полагаться только на случайность отбора объектов в выборку. Все прочие способы будут необъективными, носящими следы влияния посторонних факторов. И семена для проверки всхожести, и жителей для оценивания среднего роста – всё нужно отбирать совершенно случайным образом. Иное дело, если экспериментатор заранее знает, что генеральная совокупность состоит из нескольких классов, различных по своим характеристикам. При таких условиях выборку лучше делать из каждого класса в отдельности. Например, изучая рост жителей, целесообразно делать отдельную выборку мужчин, отдельную женщин; при этом можно учесть возраст, профессию.

Из случайного характера выборок неопровержимо следует, что **любое суждение о генеральной совокупности по выборке само является случайным**. Имеется в виду суждение, затрагивающее хотя бы один элемент генеральной совокупности, не попавший в выборку.

А какова же связь между наблюдениями, отбираемыми в состав экспериментальных данных, и выборочным методом? Имеется случайная величина $\hat{x}$ и в результате n независимых испытаний получаются n её допустимых значений. Если все допустимые значения случайной величины $\hat{x}$ считать генеральной совокупностью, то полученные при наблюдениях n значений образуют выборку. По этой выборке изучаются свойства

случайной величины $\hat{x}$. Итак, **производство наблюдений является частным случаем выборочного метода, когда в качестве генеральной совокупности берутся все допустимые значения некоторой случайной величины и исследуются свойства этой величины.**

Если требуется исследовать несколько основных признаков, например рост и вес жителей города, три размера прямоугольных деталей, то экспериментатор наблюдает векторную случайную величину $\hat{X}_{<q>}$. При этом выборка $\hat{X}_{<q>1}, \hat{X}_{<q>2}, \ldots, \hat{X}_{<q>n}$ описывается nq-мерным случайным вектором $\hat{X}_{<nq>}$.

Как уже отмечалось, случайные выборки используются для изучения свойств генеральной совокупности. При этом обычно элементы выборки используются для образования по соответствующим правилам новых случайных величин вида

$$\hat{s}_j = f_j(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n), \quad j = 1,2,\ldots$$

которые называются ***статистиками***. Примерами статистик являются выборочная сумма $\hat{s}_1 = \sum_{i=1}^{n} \hat{x}_i$, выборочное среднее $\hat{s}_2 = \frac{1}{n}\left(\sum_{i=1}^{n} \hat{x}_i \right)$ и т.д.

Все свойства (характеристики) генеральной совокупности, получаемые на основе выборки, называются ***выборочными*** или ***статистическими*** (в отличие от теоретических, изучаемых в теории вероятностей). В дальнейшем все выборочные характеристики будут отмечаться индексом *. Так, $F^*_{\hat{x}}(x)$ обозначает статистическую функцию распределения случайной величины $\hat{x}$; $M^*_{\hat{x}} = \overline{x}^*$ – статистическое математическое ожидание; $D^*_{\hat{x}}$ – статистическую дисперсию и т.д. Все выборочные характеристики представляют собой функции элементов выборки, т.е. статистики.

В выборочном методе большое внимание уделяется изучению законов распределения различных статистик, статистических рядов. Теоретической основой выборочного метода являются предельные теоремы теории вероятностей, которые приведены в виде, адаптированном к рассматриваемой тематике, в § 2.2.

1.4. Задачи и методы обработки экспериментальных данных

Как уже отмечалось, основная цель обработки экспериментальных данных состоит в получении определённых сведений об исследуемом объекте (процессе, явлении). Особенности процесса обработки определяются характером решаемых задач и объёмом информации, получаемой в ходе эксперимента. По указанным признакам обработка может быть первичной и вторичной.

Задачами первичной обработки являются выделение полезного сигнала на фоне помех (шумов), сжатие данных и приведение их к системе измерений, пригодной для дальнейшей обработки или отображения. Наиболее характерной операцией выделения полезного сигнала является устранение грубых ошибок.

Задачи вторичной обработки могут быть распределены в две группы. Первая группа задач сводится к построению математических моделей реальных процессов и явлений, вторая группа – к анализу таких моделей. Математическая модель – это абстрактное информационное отражение реального процесса или явления на языке математики.

В кибернетике (как технической, так и экономической) и теории управления процедура построения математических моделей объектов управления по результатам наблюдения их входных и выходных процессов называется идентификацией. Параллельно с теорией идентификации в кибернетике развивается теория распознавания образов, которая также связана с проблематикой построения математических моделей по результатам обработки экспериментальных данных. Распознавание образов представляет собой задачу обработки данных, в процессе которой делается вывод о принадлежности распознаваемого образа к определённому классу. Этот класс и определяет вид искомой модели.

Наиболее типовыми задачами построения математических моделей на основе статистических методов являются задачи оценивания параметров (например, параметров законов распределения случайных объектов), оценивания неизвестных функциональных зависимостей (например, законов распределения), проверки гипотез, построения уравнений регрессии и распознавания образов.

В задачах анализа моделей производится оценка влияния множества факторов на конечный результат и выбор наиболее важных факторов, а также исследуется структура экспериментальных данных и построенных на их основе математических моделей.

Один из возможных вариантов перечня задач, решаемых при первичной и вторичной обработке показан на рис.1.2.

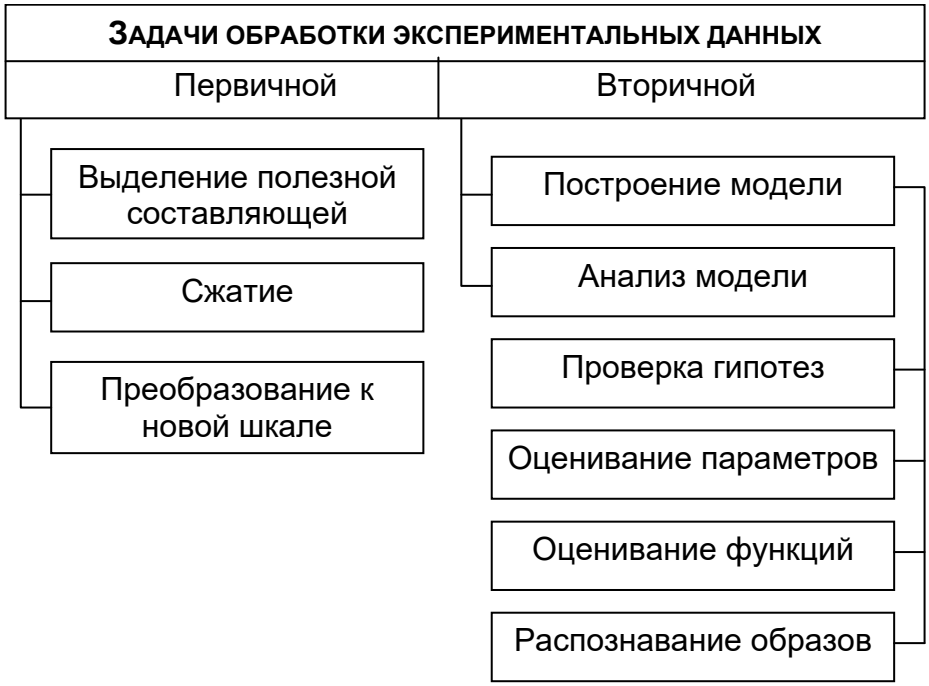


Рис.1.2. Классификация задач обработки экспериментальных данных

2. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ПАРАМЕТРИЧЕСКИХ МЕТОДОВ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

2.1. Основные понятия

Параметрические методы обработки экспериментальных данных опираются на основополагающий факт, в соответствии с которым свойства результатов экспериментальных исследований, рассматриваемых как случайные объекты, описываются некоторым законом распределения. При этом предполагается, что анализ экспериментальных данных позволяет с достаточной степенью точности определить вид и конкретную форму закона распределения или значения его параметров, если нет необходимости в использовании самого закона. Такая информация даёт возможность в полном объёме использовать методы теории вероятностей для решения задач обработки.

Так как действительный закон распределения и значения его параметров неизвестны, то параметрические методы оперируют с их приближениями – статистическими законами распределения и оценками параметров распределения.

Статистическим законом распределения случайной величины $\hat{x}$ называется закон распределения данной величины, установленный с помощью статистических методов обработки данных.

Статистический закон распределения может быть определён в виде статистической функции распределения $F^*_{\hat{x}}(x)$, статистической плотности распределения $\varphi^*_{\hat{x}}(x)$ или статистического ряда распределения $P^*(x_i)$, $i = \overline{1,n}$.

Статистическими оценками параметров закона распределения случайной величины называются приближённые значения данных параметров (статистики), полученные с помощью статистических методов обработки данных.

В дальнейшем статистические оценки для краткости называются просто оценками.

Если некоторый закон распределения характеризуется параметрами a_1, a_2,…, a_m, то их оценки будем обозначать в виде $\tilde{a}_1$, $\tilde{a}_2$,…, $\tilde{a}_m$. Наиболее распространёнными видами параметров законов распределения при обработке экспериментальных данных являются математическое ожидание $M_{\hat{x}}$, дисперсия $D_{\hat{x}}$ или среднее квадратическое отклонение $\sigma_{\hat{x}}$, а для

системы случайных величин – корреляционный момент $K_{\hat{x}_i\hat{x}_j}$ или коэффициент корреляции $r_{\hat{x}_i\hat{x}_j}$. Иногда используются центральные моменты третьего и четвёртого порядков. Соответственно при обработке данных используются их статистические аналоги – оценки математического ожидания, корреляционного момента и т.д.

Таким образом, если имеется совокупность экспериментальных данных $x_1, x_2,\ldots, x_n$, то и статистический закон распределения, например функция $F^*_{\hat{x}}(x)$, и оценки его параметров представляют собой некоторые функции этих данных:

$$F^*_{\hat{x}}(x) = \psi(x_1,\ x_2,...,x_n)\,; \qquad (2.1.1)$$

$$\widetilde{a}_j = f_j(x_1,\ x_2,...,x_n), \quad j = \overline{1,m}\,. \qquad (2.1.2)$$

Вид статистик ψ и f_j определяет качество оценок $F^*_{\hat{x}}(x)$ и $\widetilde{a}_j$. В связи с этим возникает ряд проблем, основной из которых является проблема определения условий, при которых оценки (2.1.1) и (2.1.2) могут с требуемой достоверностью представлять теоретические законы распределения и их параметры. Эти условия формируются ***предельными теоремами*** теории вероятностей. Именно они служат тем фундаментом параметрических методов обработки экспериментальных данных, на основе которого могут быть получены подходящие оценки законов и параметров распределения наблюдаемых характеристик.

Вторая проблема состоит в выборе **достаточной статистики**, т.е. такой статистики, которая позволяет в конкретных условиях получать оценки заданного качества. Так как на основе результатов наблюдений x_1, $x_2,\ldots, x_n$ может быть образован большой спектр статистик (2.1.1) и (2.1.2), данная проблема сводится к выбору из них оптимальной в определённом смысле статистики. Решение проблемы осуществляется методами теории статистических решений.

Как видно из рис.1.1, к проблеме принятия решений при обработке экспериментальных данных сводится не только задача выбора достаточной статистики. Большинство задач обработки данных в разной степени может быть отнесено к задачам принятия решений. В связи с этим фундаментом параметрических методов обработки служат также принципы принятия статистических решений, на основе которых сформированы критерии принятия оптимальных в определённом смысле решений. Особую роль среди данных принципов играет принцип максимального правдоподобия и вытекающий из него для случая нормального закона распределения метод наименьших квадратов.

В настоящей брошюре рассматриваются вопросы параметрической обработки экспериментальных данных.

2.2. Предельные теоремы теории вероятностей

Использование параметрических методов обработки данных предполагает выявление условий, определяющих справедливость априорных предположений о виде закона распределения исследуемой случайной величины и свойствах его параметров. Эти условия формулируются в виде предельных теорем теории вероятностей. Ниже излагаются содержание и сущность теорем без доказательства, а также некоторые рекомендации по их практическому применению.

2.2.1. Теорема Ляпунова

В природе, как известно, широко распространён нормальный закон распределения. Практикой установлено, что этому закону подчиняются ошибки стрельбы и бомбометания, погрешности измерений, погрешности размеров деталей, изготавливаемых промышленными предприятиями, время безотказной работы многих устройств и т.д. Поэтому в процессе обработки экспериментальной информации часто выдвигается предположение о нормальном распределении исследуемой случайной величины. Однако иногда нормальный закон распределения применить нельзя. Ввиду этого необходимо точно знать, когда можно выдвинуть такое предположение и в каких случаях от него следует отказаться. Этому вопросу посвящена центральная предельная теорема и её разновидности (теоремы Ляпунова, Муавра-Лапласа).

Теорема. Если последовательность независимых случайных величин $\hat{x}_1$, $\hat{x}_2$,…, $\hat{x}_n$ удовлетворяет условию Ляпунова

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} M[|\hat{x}_i - M_{\hat{x}_i}|^3]}{\sqrt{\left(\sum_{i=1}^{n} D_{\hat{x}_i}\right)^3}} = 0,$$

где $M[|\hat{x}_i - M_{\hat{x}_i}|^3]$ – третий абсолютный центральный момент, то последовательность случайных величин

$$\vartheta_n = \frac{\sum_{i=1}^{n} (\hat{x}_i - M_{\hat{x}_i})}{\sqrt{\sum_{i=1}^{n} D_{\hat{x}_i}}}$$

сходится по распределению к случайной величине, имеющей нормальное распределение, т.е. существует предел

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}_i})}{\sqrt{\sum_{i=1}^{n} D_{\hat{x}_i}}} < \vartheta \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\vartheta} e^{-\frac{t^2}{2}} dt .$$

На практике часто пользуются случайными величинами, представляющими собой сумму независимых случайных величин:

$$\hat{z}_n = \sum_{i=1}^{n} \hat{x}_i = \hat{\vartheta}_n \sqrt{\sum_{i=1}^{n} D_{\hat{x}_i} + \sum_{i=1}^{n} M_{\hat{x}_i}} .$$

Поскольку случайная величина $\hat{z}_n$ связана со случайной величиной $\hat{\vartheta}_n$ линейной зависимостью, то в пределе она также будет иметь нормальное распределение. Параметры данного распределения можно выразить с помощью теорем о числовых характеристиках:

$$\left.\begin{aligned} M_{\hat{z}_n} &= M\left[\sum_{i=1}^{n} \hat{x}_i\right] = \sum_{i=1}^{n} M_{\hat{x}_i}; \\ D_{\hat{z}_n} &= D\left[\sum_{i=1}^{n} \hat{x}_i\right] = \sum_{i=1}^{n} D_{\hat{x}_i}. \end{aligned}\right\}$$

Условие Ляпунова представляет собой требование малости слагаемых

$$\frac{\hat{x}_i - M_{\hat{x}_i}}{\sqrt{\sum_{i=1}^{n} D_{\hat{x}_i}}}$$

в сумме

$$\frac{\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}_i})}{\sqrt{\sum_{i=1}^{n} D_{\hat{x}_i}}} .$$

Таким образом, сущность центральной предельной теоремы состоит в следующем: закон распределения суммы независимых случайных величин при неограниченном увеличении числа слагаемых приближается к нормальному, если случайные величины, входящие в сумму, имеют дисперсию одного и того же порядка и конечные математические ожидания. Это означает, что удельный вес каждого слагаемого стремится к нулю при увеличении числа слагаемых.

В реальных условиях любое случайное отклонение от закономерного протекания основного явления вызывается бесчисленным множеством случайных факторов, каждый из которых обычно оказывает малое влияние на суммарное воздействие, и часто эти факторы независимы или сла-

бо зависимы. Этим и объясняется широкое распространение нормального закона.

На практике теоремой Ляпунова пользуются и тогда, когда n сравнительно невелико. При суммировании непрерывных случайных величин, имеющих одинаковые симметричные законы распределения с одинаковыми числовыми характеристиками, эту теорему можно применять при $n \geq 8$. Если же суммируются случайные величины с различными несимметричными законами и различными числовыми характеристиками, то теоремой Ляпунова можно пользоваться только при числе слагаемых порядка сотни.

Практическое применение теоремы Ляпунова предполагает использование формул для определения вероятности попадания нормально распределённой случайной величины в интервал $[a; b)$. В данном случае можно воспользоваться следующими формулами:

$$P(a \leq \hat{z}_n < b) \approx \Phi_0\left(\frac{b - M_{\hat{z}_n}}{\sigma_{\hat{z}_n}}\right) - \Phi_0\left(\frac{a - M_{\hat{z}_n}}{\sigma_{\hat{z}_n}}\right); \qquad (2.2.1)$$

$$P(a \leq \hat{z}_n < b) \approx \Phi_1\left(\frac{b - M_{\hat{z}_n}}{\sigma_{\hat{z}_n}}\right) - \Phi_1\left(\frac{a - M_{\hat{z}_n}}{\sigma_{\hat{z}_n}}\right), \qquad (2.2.2)$$

где

$$\Phi_0 = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt, \qquad (2.2.3)$$

$$\Phi_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \qquad (2.2.4)$$

называются функциями нормированного нормального распределения (т. е. распределения с параметрами $M_{\hat{z}} = 0$, $D_{\hat{z}} = 1$) или функциями Лапласа, они являются табличными (см. приложения 2 и 3).

Следует отметить, что формулами (2.2.1) и (2.2.2) можно пользоваться при выполнении условия

$$\left| z - M_{\hat{z}_n} \right| < 3\sigma_{\hat{z}_n}, \qquad (2.2.5)$$

где z принимает значение a или b. Это требование вызвано тем, что за пределами интервала

$$[M_{\hat{z}_n} - 3\sigma_{\hat{z}_n}; M_{\hat{z}_n} + 3\sigma_{\hat{z}_n}]$$

могут быть существенные ошибки.

П р и м е р 2.1. При обработке информации требуется сложить 1000 чисел, каждое из которых округлено с точностью до 0,01. Полагая, что ошибки округления подчинены равномерному закону распределения, найти вероятность того, что суммарная ошибка округления не превысит 0,2.

▼ Обозначим через $\hat{x}_i$, $i = \overline{1,1000}$ ошибку округления i-го числа, а через $\hat{z}_n$ – суммарную ошибку округления (n = 1000).

Далее учитываем, что случайная величина, равномерно распределённая на интервале [a; b] имеет математическое ожидание и дисперсию, которые определяются по формулам

$$M_{\hat{x}} = \frac{a+b}{2}, \quad D_{\hat{x}} = \frac{(b-a)^2}{12}.$$

Ошибки округления в данном случае распределены на интервале

$$[a; b] = [-0{,}005;\ 0{,}005],$$

следовательно,

$$M_{\hat{x}} = \frac{-0{,}005+0{,}005}{2} = 0, \quad D_{\hat{x}} = \frac{(0{,}005-(-0{,}005))^2}{12} = \frac{(0{,}01)^2}{12}.$$

$$M_{\hat{z}_n} = \sum_{i=1}^{1000} M_{\hat{x}_i} = 0; \quad D_{\hat{z}_n} = \sum_{i=1}^{1000} \frac{(0{,}01)^2}{12} = 0{,}00833;$$

$$\sigma_{\hat{z}_n} = \sqrt{D_{\hat{z}_n}} = \sqrt{0{,}00833} = 0{,}0913.$$

Условие (2.2.5) соблюдается, поэтому

$$P(|\hat{z}_n| < 0{,}2) = \Phi_0\left(\frac{0{,}2}{0{,}0913}\right) - \Phi_0\left(\frac{-0{,}2}{0{,}0913}\right) =$$

$$= 2\Phi_0\left(\frac{0{,}2}{0{,}0913}\right) = 2\Phi_0(2{,}19) = 2 \cdot 0{,}4857 = 0{,}971.$$

Последнее равенство вытекает непосредственно из формулы (2.2.1), в нём учтено, что функция $\Phi_0(x)$ является нечётной.

▲

2.2.2. Теорема Муавра-Лапласа

Пусть производится n независимых испытаний, в каждом из которых событие A может появляться с одной и той же вероятностью. Тогда случайная величина $\hat{z}_n$, представляющая собой число появлений события A в n испытаниях, будет иметь биномиальное распределение [4]. Если число испытаний велико, то и случайная величина принимает большое число возможных значений. Пользоваться такой случайной величиной затруднительно из-за сложности вычислений. Поэтому целесообразно применять теорему Муавра-Лапласа, которая доказывает сходимость последовательности случайных величин, имеющих биномиальное распределение, к нормально распределённой случайной величине.

Теорема. Пусть $\hat{z}_n$ число появлений события A в n независимых испытаниях, в каждом из которых вероятность этого события равна p. Тогда при $n \to \infty$ имеет место соотношение

$$\lim_{n\to\infty} P\left(\frac{\hat{z}_n - np}{\sqrt{npq}} < \vartheta\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\vartheta} e^{-\frac{t^2}{2}} dt$$

где $q = 1 - p$; $M_{\hat{z}_n} = np$, $D_{\hat{z}_n} = npq$ – соответственно математическое ожидание и дисперсия биномиально распределённой случайной величины.

Таким образом, нормированная случайная величина

$$\hat{\vartheta}_n = \frac{\hat{z}_n - np}{\sqrt{npq}} \qquad (2.2.6)$$

согласно теореме Муавра-Лапласа в пределе будет подчиняться нормированному нормальному закону распределения. Отсюда вытекает приближённое равенство, справедливое при больших значениях n:

$$P(a \le \hat{\vartheta}_n < b) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{\vartheta^2}{2}} d\vartheta. \qquad (2.2.7)$$

Найдём вероятность попадания случайной величины $\hat{z}_n$ в интервал $[m_1; m_2)$. Для этого подставим граничные точки m_1 и m_2 в формулу (2.2.6):

$$a = \frac{m_1 - np}{\sqrt{npq}}; \qquad b = \frac{m_2 - np}{\sqrt{npq}}.$$

Выражение (2.2.7) принимает вид

$$P(m_1 \le \hat{z}_n < m_2) \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{\vartheta^2}{2}} d\vartheta.$$

Используя табличные функции (2.2.3) и (2.2.4), получаем следующие рабочие формулы:

$$P(m_1 \le \hat{z}_n < m_2) \approx \Phi_0\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{m_1 - np}{\sqrt{npq}}\right); \qquad (2.2.8)$$

$$P(m_1 \le \hat{z}_n < m_2) \approx \Phi_1\left(\frac{m_2 - np}{\sqrt{npq}}\right) - \Phi_1\left(\frac{m_1 - np}{\sqrt{npq}}\right); \qquad (2.2.9)$$

Если n сравнительно мало и разность $|m - np|$ соизмерима с 0,5, то не безразлично, относятся ли граничные точки интервала $[m_1; m_2)$ к числу возможных значений $\hat{z}_n$ или нет. В этом случае вместо z_1 и z_2 следует брать $m_1 - 0{,}5$, $m_2 - 0{,}5$. Тогда соотношения (2.2.8) и (2.2.9) примут вид

$$P(m_1 \le \hat{z}_n < m_2) \approx \Phi_0\left(\frac{m_2 - 0{,}5 - np}{\sqrt{npq}}\right) - \Phi_0\left(\frac{m_1 - 0{,}5 - np}{\sqrt{npq}}\right); \quad (2.2.10)$$

$$P(m_1 \le \hat{z}_n < m_2) \approx \Phi_1\left(\frac{m_2 - 0{,}5 - np}{\sqrt{npq}}\right) - \Phi_1\left(\frac{m_1 - 0{,}5 - np}{\sqrt{npq}}\right). \quad (2.2.11)$$

Следует отметить, что формулы (2.2.10) и (2.2.11) дают более точное приближение, чем (2.2.8) и (2.2.9).

Расчёты по приближённым формулам (2.2.8) – (2.2.11) могут производиться при соблюдении условия

$$|m - np| < 3\sqrt{npq},$$

которое непосредственно вытекает из (2.2.5).

Теорема Муавра-Лапласа описывает поведение биномиального распределения при больших значениях n, что позволяет значительно упростить вычисления. Расчёты по точной формуле

$$P(m_1 \le \hat{z}_n < m_2) = \sum_{m=m_1}^{m_2} C_n^m p^m q^{n-m} \qquad (2.2.12)$$

при больших значениях n очень громоздки.

В выражение (2.2.12) входит число сочетаний из n элементов по m:

$$C_n^m = \frac{n!}{m!(n-m)!},$$

где $n! = 1 \cdot 2 \cdot \ldots \cdot n$, $m! = 1 \cdot 2 \cdot \ldots \cdot m$.

П р и м е р 2.2. По линии связи независимо друг от друга передаётся 90 сообщений, каждое из которых состоит из пяти двоичных чисел. Вероятность искажения хотя бы одного числа в сообщении равна 0,06. Определить вероятность того, что число принятых без искажения сообщений будет находиться в интервале [82; 87].

▼ Обозначим через $\hat{x}_i$ число неискажённых сообщений при i-й передаче. Эта случайная величина принимает значение ноль с вероятностью 0,06 и единица – с вероятностью 0,94. Следовательно

$$p_i = p = 0{,}94; \quad q_i = q = 0{,}06.$$

Обозначим $\hat{z}_n = \sum_{i=1}^{90} \hat{x}_i$ – общее число неискажённых сообщений. Тогда

$$M_{\hat{z}_n} = np = 90 \cdot 0{,}94 = 84{,}6,$$

$$\sigma_{\hat{z}_n} = \sqrt{npq} = \sqrt{90 \cdot 0{,}94 \cdot 0{,}06} = 2{,}25.$$

По формуле (2.2.9) получаем

$$P(82 \le \hat{z}_n \le 87) \approx \Phi_1\left(\frac{87 - 84{,}6}{2{,}25}\right) - \Phi_1\left(\frac{82 - 84{,}6}{2{,}25}\right) =$$

$$= \Phi_1(1{,}067) - \Phi_1(-1{,}156) = 0{,}857 - 0{,}124 = 0{,}733.$$

▲

2.2.3. Неравенство Чебышева

Для любой случайной величины, имеющей конечное математическое ожидание и дисперсию, при каждом $\varepsilon > 0$ имеет место неравенство

$$P(|\hat{x} - M_{\hat{x}}| \geq \varepsilon) \leq \frac{D_{\hat{x}}}{\varepsilon^2}. \qquad (2.2.13)$$

Для противоположного события неравенство Чебышева принимает вид

$$P(|\hat{x} - M_{\hat{x}}| < \varepsilon) \geq 1 - \frac{D_{\hat{x}}}{\varepsilon^2}. \qquad (2.2.14)$$

Неравенства (2.2.13) и (2.2.14) можно использовать для получения оценок вероятностей отклонения случайной величины от своего математического ожидания, если закон распределения случайной величины неизвестен.

П р и м е р 2.3. Найти нижнюю границу вероятности того, что случайная величина $\hat{x}$, имеющая произвольный закон распределения, отклоняется от своего математического ожидания меньше чем на $\pm 3\sigma_{\hat{x}}$.

▼ По формуле (2.2.14) получим

$$P(|\hat{x} - M_{\hat{x}}| < 3\sigma_{\hat{x}}) \geq 1 - \frac{D_{\hat{x}}}{(3\sigma_{\hat{x}})^2} = 1 - \frac{D_{\hat{x}}}{9D_{\hat{x}}} = 1 - \frac{1}{9} = \frac{8}{9}.$$

Известно, что для нормального закона распределения существует так называемое «правило трёх сигм», согласно которому вероятность попадания случайной величины в интервал

$$[M_{\hat{x}} - 3\sigma_{\hat{x}}; M_{\hat{x}} + 3\sigma_{\hat{x}}]$$

близка к единице (≈ 0,997). Подобное правило существует и для случайных величин, имеющих распределение, отличное от нормального, но при этом вероятность указанного события будет не меньше 8/9.

▲

2.2.4. Теоремы Чебышева и Маркова

Частная теорема Чебышева. При неограниченном увеличении независимых испытаний среднее арифметическое полученных при испытаниях значений случайной величины, имеющей конечную дисперсию, сходится по вероятности к её математическому ожиданию:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - M_{\hat{x}}\right| < \varepsilon\right) = 1. \qquad (2.2.15)$$

Из (2.2.15) и (2.2.1) следует, что при ограниченном *n* справедливо приближённое равенство

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - M_{\hat{x}}\right| < \varepsilon\right) \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sigma_{\hat{x}}}\right), \qquad (2.2.16)$$

где $\frac{\sigma_{\hat{x}}}{\sqrt{n}} = \sigma[M_{\hat{x}}]$ – среднее квадратическое отклонение математического ожидания случайной величины $\hat{x}$.

В выражении (2.2.16) учтено, что математическое ожидание случайной величины

$$\left|\frac{1}{n}\sum \hat{x} - M_{\hat{x}}\right|$$

равно нулю.

Применяя неравенство Чебышева (2.2.14) для случайной величины $1/n\sum_{i=1}^{n}\hat{x}_i$, получим

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - M_{\hat{x}}\right| < \varepsilon\right) \geq 1 - \frac{D_{\hat{x}}}{n\varepsilon^2}. \qquad (2.2.17)$$

Формулой (2.2.16) можно пользоваться, когда применима теорема Ляпунова или когда закон распределения каждой случайной величины $\hat{x}_i$, $i = \overline{1,n}$ нормальный. Если же теорема Ляпунова не применима или законы распределения $\hat{x}_i$, $i = \overline{1,n}$ неизвестны, то приходиться определять нижние границы соответствующих вероятностей из соотношения (2.2.17).

При решении практических задач с применением теоремы Чебышева часто возникают трудности, связанные с невозможностью обеспечить независимость испытаний. Теорема Маркова определяет условия, при которых закон больших чисел справедлив и для зависимых испытаний.

Теорема Маркова. Если случайная величина $\hat{x}$ такова, что

$$\lim_{n\to\infty}\frac{D\left[\sum_{i=1}^{n}\hat{x}_i\right]}{n^2} = 0,$$

то для любого числа $\varepsilon > 0$ существует предельное соотношение

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - M_{\hat{x}}\right| < \varepsilon\right) = 1.$$

2.2.5. Теорема Бернулли

Если производится n независимых испытаний и вероятность появления события A в каждом из них равна p, то при любом $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{\hat{z}_n}{n} - p\right| < \varepsilon\right) = 1,$$

где $\hat{z}_n$ – число появления события A в n испытаниях.

При $n \to \infty$ согласно теореме Ляпунова можно считать, что случайная величина

$$P^* = \frac{\hat{z}_n}{n}$$

будет иметь нормальное распределение, поэтому справедливо приближённое равенство

$$P\{|P^* - p| < \varepsilon\} \approx 2\Phi_0\left(\varepsilon\sqrt{\frac{n}{pq}}\right). \qquad (2.2.17)$$

В соотношении (2.2.17) учитывается, что

$$D[|P^* - p|] = D[P^*] = D\left[\frac{\hat{z}_n}{n}\right] = \frac{1}{n^2}D_{\hat{z}_n} = \frac{npq}{n^2} = \frac{pq}{n} \quad \text{или} \quad \sigma[|P^* - p|] = \sqrt{\frac{pq}{n}}.$$

К случайной величине P^* применимо неравенство Чебышева

$$P(|P^* - p| < \varepsilon) \geq 1 - \frac{pq}{n\varepsilon^2}.$$

Теорема Бернулли устанавливает факт устойчивости частоты, когда вероятность появления события от испытания к испытанию не меняется.

2.3. Элементы теории статистических решений

2.3.1. Задачи принятия статистических решений при обработке экспериментальных данных

Конечной целью обработки экспериментальных данных являются некоторые выводы о состоянии или свойствах исследуемого процесса или объекта. Например, это могут быть выводы о законах распределения случайных величин или параметрах законов распределения, о справедливости каких-либо гипотез и, наконец, выводы о наличии и особенностях взаимосвязей различных показателей, характеризующих свойства объекта. В любом случае из множества возможных выводов необходимо выбрать один, оптимальный в каком-либо смысле. Иначе говоря, необходимо принять решение.

Решением называется некоторое заключение, вывод об исследуемом объекте или его свойствах.

Обработка данных всегда осуществляется в условиях неопределённости, обусловленной неполнотой информации об исследуемом объекте, помехами как естественного, так и искусственного характера. В связи с этим принимаемые решения являются статистическими.

Статистическим решением называется некоторое заключение (вывод) об исследуемом объекте или его свойствах, полученное в результате обработки экспериментальных данных.

Основой для принятия решения является ***решающее правило*** (решающая функция), которое служит для выбора из множества возможных решений одного предпочтительного.

Пусть $\{E\}$ - множество возможных состояний исследуемого объекта; $\{X_{<n>}\}$ – множество возможных результатов наблюдений, а $\{\widetilde{E}\}$ – множество возможных решений.

Функция

$$R:\{X_{<n>}\}\to\{\widetilde{E}\},$$

отображающая множество $\{X_{<n>}\}$ результатов наблюдений в множество решений $\{\widetilde{E}\}$, называется решающим правилом (решающей функцией).

Существует много решающих правил, но из этих правил выбирается такое, которое обеспечивает принятие решения требуемого качества, т.е. решения в определённом смысле оптимального.

В связи с этим возникает задача определения оптимального решающего правила. Выбор такого правила определяется рядом факторов:

а) требованиями, которые предъявляются к качеству решения;

б) свойствами экспериментальных данных;

в) условиями, в которых получены данные;

г) дополнительной априорной информацией, которая может быть использована при принятии решения.

Требования к качеству решения определяются потребителем решения (см. рис.1.1) и могут быть сформированы в виде требований минимизации:

– потерь, которые может понести потребитель при неправильном решении;

– риска, связанного с принятием неправильного решения;

– вероятности принятия неправильного решения.

Потерями называются отрицательные последствия, сопровождающие реализацию принятого решения.

Риск – это возможность некоторых потерь со стороны потребителя.

Условия, в которых получены данные, можно разделить на две группы: условия пассивного эксперимента и условия активного эксперимента. В первом случае планирование экспериментальных работ с целью получения данных с необходимыми свойствами отсутствует. Во втором случае эксперимент организуется так, чтобы полученные результаты обладали требуемыми свойствами.

Наиболее важными свойствами совокупности экспериментальных данных, существенно влияющими на качество решающей функции являются объём выборки и её представительность, статистическая устойчивость, однородность, отсутствие аномальных результатов.

Влияние данных свойств на качество решения и вид решающей функции рассмотрено в последующих разделах.

Наиболее важными свойствами априорной информации, оказывающими влияние на формирование решающей функции, являются объём

данной информации и её достоверность. Так, в идеальном случае априорная информация I_1 (см. рис.1.1) позволяет получить ряд распределения вероятностей состояний исследуемого объекта, наиболее полной формой информации I_4 является закон распределения вектора результатов наблюдений.

Описание свойств решений может осуществляться с различных позиций. Чаще всего для этого применяется функция потерь.

Пусть множество решений $\{\widetilde{E}\}$ является дискретным и состоит из l альтернативных решений $\widetilde{E}_j$, $j=\overline{1,l}$. Такими же свойствами пусть обладает и множество $\{E\}$ состояний объекта, мощность этого множества обозначим через m.

Функцией потерь называется функция $\pi(E_i;\ \widetilde{E}_j)$, $i=\overline{1,m}$, $j=\overline{1,l}$, характеризующая последствия принятия решения $\widetilde{E}_j$ при условии, что объект находится в состоянии E_i:

$$\pi(E_i;\ \widetilde{E}_j) = \pi_{ij}. \qquad (2.3.1)$$

Функция (2.3.1) записывается в виде матрицы потерь

$$\Pi_{[m;l]} = \begin{pmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1l} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{i1} & \pi_{i2} & \cdots & \pi_{il} \\ \cdots & \cdots & \cdots & \cdots \\ \pi_{m1} & \pi_{m2} & \cdots & \pi_{ml} \end{pmatrix}.$$

Следует заметить, что для определённого состояния E_i исследуемого объекта величина потерь в общем случае будет величиной случайной. В дальнейшем будем полагать, что информация I_6 (рис.1.1) представляет собой сведения о функции потерь (2.3.1).

Учитывая изложенное выше, под объёмом априорной информации будем понимать число видов информации, используемой при формировании решающего правила, а под качеством априорной информации – степень её соответствия объективным свойствам исследуемого объекта.

Оптимальным решающим правилом называется правило, обеспечивающее выполнение требований, предъявляемых к качеству решения в конкретных условиях применения данного правила.

Под условиями принятия решения понимается совокупность перечисленных выше факторов, определяющих выбор правила решения. Обозначим вектор условий, в которых применяется решающее правило $R_I(X)$, символом $K_{<4>}$. Компонентами данного вектора являются объём K_{X1} и качество K_{X2} результатов наблюдений, объём K_{I1} и качество K_{I2} априорной информации:

$$K_{<4>} = (K_{X1},\ K_{X2},\ K_{I1},\ K_{I2})^{\intercal}.$$

С учётом этого решающее правило можно представить в виде функции двух переменных – результатов наблюдений и априорной информации, т.е.

$$R_I(X) = R(X_{<n>}; I),$$

а качество (оптимальность) данного правила применительно к условиям $K_{<4>}$ охарактеризовать показателем

$$L = f(R; K_{<4>}),$$

представляющим собой эффект, который достигается в результате использования решающего правила R в условиях $K_{<4>}$. Конкретное выражение показателя L при решении различных задач обработки может быть различным. Поэтому оптимальным решающим правилом R_0 для некоторой совокупности условий обработки $K_{<4>j}$, $j = \overline{1, I}$, целесообразно считать правило, обеспечивающее экстремум показателя L в условиях $K_{<4>j}$:

$$R_0(X_{<n>}; I) = \arg \operatorname*{extr}_{R_i \in \{R\}} \left\{ L(R_i; K_{<4>j}) \right\},$$

где $\{R\}$ – множество решающих правил.

На практике подбор оптимального решающего правила для некоторых условий обработки выполняется на основе принципов принятия решений. В настоящее время сформулирован ряд таких принципов, основными из которых являются принцип максимального правдоподобия и принцип минимальной вероятности ошибки.

Рассмотрим их более подробно.

2.3.2. Принцип максимального правдоподобия

Данный принцип используется в тех случаях, когда известен только условный закон распределения результатов наблюдений относительно состояния E исследуемого объекта.

Пусть случайная величина $\hat{x}$ имеет плотность распределения $\varphi_{\hat{x}/E}(x; E)$, а результаты наблюдения над величиной $\hat{x}$ представляют собой простую (повторную) случайную выборку

$$(\hat{x}_1,\ \hat{x}_2, ..., \hat{x}_n)^{\mathsf{T}} = \hat{X}_{<n>}.$$

Вероятность появления конкретной реализации $X_{<n>}$ пропорциональна элементу вероятности распределения случайного вектора $\hat{X}_{<n>}$:

$$P(\hat{X}_{<n>} = X_{<n>}) = P\left(\bigcap_{i=1}^{n} (x_i \le \hat{x}_i < x_i + dx_i \right) = \varphi_{\hat{X}_{<n>}/E}(X_{<n>}; E) dX_{<n>}. \tag{2.3.2}$$

Плотность распределения для простой выборки представляется выражением

$$\varphi_{\hat{X}_{<n>}/E}(X_{<n>}; E) = \prod_{i=1}^{n} \varphi_{\hat{x}_i/E}(x_i; E) = \prod_{i=1}^{n} \varphi_{\hat{x}/E}(x_i; E).$$

Функция оцениваемого состояния E вида

$$L(E; X_{<n>}) = \varphi_{\hat{X}_{<n>}/E}(X_{<n>}; E) \qquad (2.3.3)$$

называется ***функций правдоподобия***.

Принцип максимального правдоподобия состоит в утверждении, что при данной реализации $X_{<n>}$ вектора $\hat{X}_{<n>}$ наиболее правдоподобным, т.е. наиболее близким к действительному состоянию, является то значение $\widetilde{E}$ оценки состояния исследуемого объекта, при котором вероятность (2.3.2), а следовательно, и функция правдоподобия (2.3.3) максимальны.

Исходя из данного утверждения, для определения наиболее правдоподобного значения состояния E необходимо найти такое значение $\widetilde{E}$, которое обеспечивает экстремум функции правдоподобия. Если оценивается несколько параметров E_1, E_2,…,E_k какого-либо состояния, то решение данной задачи сводится к нахождению корней системы уравнений

$$\frac{\partial L(E_1; E_2; ...; E_k; X_{<n>})}{\partial E_j} = 0, \quad j = \overline{1,k}. \qquad (2.3.4)$$

Выражения (2.3.4) представляют собой необходимое условие экстремума функции многих переменных, которое заключается в равенстве нулю всех частных производных данной функции.

Решение уравнений (2.3.4) даёт оценки исследуемых параметров, которые называются ***оценками максимального правдоподобия***.

Таким образом, принятие решения в соответствии с принципом максимального правдоподобия заключается в следующем. На основе выборки $X_{<n>}$ определяются значения функции правдоподобия для всех возможных состояний E_i, $i = \overline{1,n}$. Среди данных состояний выбирается то, которое обеспечивает максимум функции правдоподобия.

2.3.3. Принцип минимальной вероятности ошибки

Рассматриваемый принцип, так же как и предыдущий, применим в тех случаях, когда известен только условный закон распределения результатов наблюдений. Сущность принципа состоит в том, что минимизируется вероятность принятия неправильного решения. Введём следующие определения.

Ошибкой первого рода называется ошибка, представляющая собой принятие решения о том, что исследуемый объект не находится в предполагаемом состоянии, в то время как в действительности он пребывает именно в этом состоянии.

Ошибкой второго рода называется ошибка, представляющая собой принятие решения о том, что исследуемый объект находится в предполагаемом состоянии, в то время как в действительности он пребывает в другом состоянии.

В общем случае правило решения должно быть таким, чтобы обеспечивалась минимально возможная вероятность принятия ошибочных решений. Если бы была известна функция потерь или функция риска, соответствующая каждому из исходов, то задачу поиска решающего правила, минимизирующего вероятность принятия ошибочного решения, можно было бы переформулировать как задачу нахождения решающего правила, минимизирующего вероятность ошибки либо первого, либо второго рода, в зависимости от того, какая из них связана с б*о*льшими потерями или б*о*льшим риском.

Так как функция потерь (риска) неизвестна, то задача поиска решающего правила формулируется как задача минимизации суммы вероятностей ошибок первого и второго рода. Метод решения данной задачи рассмотрим для случая, когда исследуемый объект может находиться в одном из двух состояний E_1 или E_2. Пусть наблюдаемая переменная $\hat{x}$ является скалярной, а кривые условных законов распределения $\varphi_{\hat{x}/E_i}(x;E_i)$ при условии, что объект может находиться в состоянии E_i, $i=\overline{1,2}$, имеют вид, изображённый на рис.2.1.

Как видно из рисунка, состояниям E_1 и E_2 соответствуют некоторые подмножества значений переменной $\hat{x}$, попадание в которые результата наблюдения с наибольшей вероятностью соответствует тому или иному состоянию объекта. Поэтому, фиксируя попадание наблюдаемого результата в одно или другое подмножество, можно судить с некоторой вероятностью о состоянии, которое принял объект. Пусть такими подмножествами являются $(-\infty; x_\alpha)$ и $[x_\alpha; +\infty)$. Величина x_α является границей данных подмножеств. Тогда при $x \in [x_\alpha; -\infty)$ принимается решение о нахождении объекта в состоянии E_1, а если $x \in [x_\alpha; \infty)$ – о нахождении объекта в состоянии E_2.

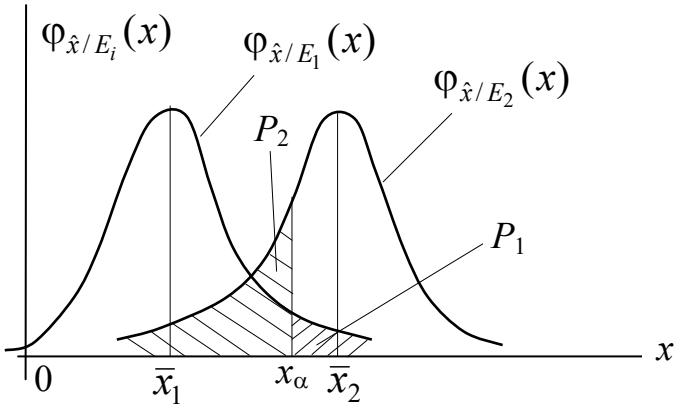


Рис.2.1. Условные законы распределения случайной величины

Поскольку кривые частично перекрываются, существует вероятность ошибки. Так, результат наблюдения с вероятностью

$$P_1 = \int_{x_\alpha}^{\infty} \varphi_{\hat{x}/E_1}(x; E_1)dx \qquad (2.3.5)$$

может попасть в область $[x_\alpha; +\infty)$, если объект находится в состоянии E_1, но при этом будет принято решение, что он пребывает в состоянии E_2. Это означает ошибку первого рода. Наоборот, с вероятностью

$$P_2 = \int_{-\infty}^{x_\alpha} \varphi_{\hat{x}/E_2}(x; E_2)dx \qquad (2.3.6)$$

результат наблюдения может попасть в область $(-\infty; x_\alpha)$, если объект находится в состоянии E_2, но принимается решение о пребывании объекта в состоянии E_1. Будет допущена ошибка второго рода.

Если последствия ошибочных решений оценить невозможно, то очевидно, что решающее правило должно обеспечивать минимум суммы вероятностей (2.3.5) и (2.3.6). Так как данное правило состоит в определении границы x_α, она должна выбираться таким образом, чтобы минимизировать величину

$$u = \int_{-\infty}^{x_\alpha} \varphi_{\hat{x}/E_2}(x; E_2)dx + \int_{x_\alpha}^{\infty} \varphi_{\hat{x}/E_1}(x; E_1)dx\,.$$

Общих правил выбора оптимального значения x_α не существует. На практике чаще всего минимизируется вероятность ошибки первого рода до определённой заранее назначенной величины. На основе этого и выбирается значение критической границы.

В том случае, когда объект может находиться более чем в двух состояниях, применение рассмотренного принципа существенно усложняется. Ввиду этого используется не сам наблюдаемый параметр, а построенная на его основе специальная функция, так называемый показатель согласованности (см. раздел 7).

Следует отметить, что на основе рассмотренных выше принципов может быть сформировано большое число показателей и критериев принятия решений, специфика построения которых определяется особенностями конкретной задачи.

2.4. Элементы теории оценивания

Первичной задачей обработки экспериментальных данных является задача оценивания. При её решении наибольшее распространение получил принцип максимального правдоподобия и вытекающие из него критерии и алгоритмы оценивания.

Пусть схема наблюдения имеет вид

$$\hat{Y} = F(A) + \hat{\mathrm{E}}\,, \qquad (2.4.1)$$

а вектор ошибок измерений $\hat{\mathrm{E}}$ имеет нормальное распределение

$$\varphi_{\hat{\mathrm{E}}}(\mathrm{E}) = C\exp\left(-\frac{1}{2}\mathrm{E}^{\mathrm{T}}K_{\hat{\mathrm{E}}}^{-1}\mathrm{E}\right),$$

где $C = (2\pi)^{n/2} \mid K_{\hat{\mathrm{E}}} \mid^{-1}$ – нормирующий множитель; $K_{\hat{\mathrm{E}}}$ – корреляционная матрица вектора ошибок измерений.

Учитывая, что

$$\hat{\mathrm{E}} = \hat{Y} - F(A),$$

плотность распределения $\varphi_{\hat{\mathrm{E}}}(\mathrm{E})$ можно выразить через $\hat{Y}$ и $F(A)$:

$$\varphi_{\hat{Y}}(Y / A) = C\exp\left(-\frac{1}{2}(Y - F(A)\right)^{\mathrm{T}} K_{\hat{\mathrm{E}}}^{-1}(Y - F(A)).$$

Тогда принцип максимального правдоподобия приводит к следующей функции потерь:

$$V(\hat{Y}; A) = (\hat{Y} - F(A))^{\mathrm{T}} K_{\hat{\mathrm{E}}}^{-1}(\hat{Y} - F(A)). \qquad (2.4.2)$$

Таким образом, при нормальном законе распределения выборки функция потерь является квадратичной. В частном случае, когда все элементы выборки имеют одинаковое распределение с дисперсией σ^2 и независимы, функция потерь (2.4.2) принимает вид

$$V(\hat{Y}; A) = (\hat{Y} - F(A))^{\mathrm{T}}(\hat{Y} - F(A)). \qquad (2.4.3)$$

Метод оценивания, основанный на минимизации квадратичной функции потерь вида (2.4.2) или (2.4.3), называется ***методом наименьших квадратов*** (см. раздел 8). Этот метод является оптимальным и для ряда других распределений ошибок наблюдений.

Если рассматривать схему наблюдения (2.4.1) в предположении, что вектор ошибок измерений имеет распределение Лапласа, то получим функцию потерь в виде суммы модулей ошибок. Метод оценивания вектора параметров, основанный на минимизации функции потерь как суммы модулей ошибок измерений, называется ***методом наименьших модулей*** [8].

В настоящей брошюре он не рассматривается. Следует только отметить, что данный метод является оптимальным и в ряде других задач оценивания.

3. МЕТОДЫ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ

Одной из важнейших задач обработки данных является задача оценивания (экспериментального определения) вероятностных характеристик случайных объектов.

3.1. Постановка задачи оценивания законов и параметров распределения случайных величин

Пусть $\hat{x}$ – случайная величина, характеризующая свойство исследуемого объекта. Требуется на основе случайной выборки $\hat{x}_1,\ \hat{x}_2,\ldots,\hat{x}_n$ выработать объективное суждение о вероятностных свойствах случайной величины $\hat{x}$.

Как было указано в § 1.3, любая функция случайной выборки

$$\hat{s} = s(\hat{x}_1,\ \hat{x}_2,...,\hat{x}_n) \qquad (3.1.1)$$

называется статистикой. Если статистика (3.1.1) используется в качестве приближения неизвестной вероятностной характеристики (закона или параметра распределения) случайной величины, то её значение

$$\widetilde{s} = s(x_1,\ x_2,...,x_n), \qquad (3.1.2)$$

полученное в результате обработки экспериментальных данных по формуле (3.1.2), называется ***оценкой*** этой характеристики.

Известно, что исчерпывающей характеристикой вероятностного поведения случайной величины $\hat{x}$ является закон её распределения $F_{\hat{x}}(x)$ или $\varphi_{\hat{x}}(x)$. Поэтому основной целью в рассматриваемой здесь задаче является построение закона распределения случайной величины $\hat{x}$ по экспериментальным данным, т.е. его представление как функции выборки

$$\widetilde{F}_{\hat{x}}(x) = F_{\hat{x}}(x;\ x_1,\ x_2,...,x_n) = s(x;\ X_{<n>}),$$

которая может служить в качестве оценки функции $F_{\hat{x}}(x)$, обладающей требуемой точностью и надёжностью (достоверностью).

В общем случае функция $F_{\hat{x}}(x)$ зависит как от своего аргумента, так и от параметров распределения:

$$F_{\hat{x}}(x) = F_{\hat{x}}(x;\ a_1,\ a_2,...,a_m) = F_{\hat{x}}(x;\ A_{<m>}).$$

Параметрическая обработка данных опирается на предположение о том, что класс распределений, которому принадлежит функция $F_{\hat{x}}(x)$, априорно известен. Конкретные значения параметров $A_{<m>}$ этого распределения, выделяющие его в рассматриваемом классе, неизвестны. Тогда оценивание функции $F_{\hat{x}}(x)$ сводится к оцениванию её параметров $A_{<m>}$, т.е. к отысканию такой статистики

$$S_{<m>}(X_{<n>}) = \widetilde{A}_{<m>} \approx A_{<m>},$$

которая обеспечивала бы приближённое равенство

$$F_{\hat{x}}(x) = F_{\hat{x}}(x;\ A_{<m>}) \approx F_{\hat{x}}(x;\ \widetilde{A}_{<m>}).$$

Поскольку вся информация об исследуемом объекте содержится в выборке объёма n, то для однозначного решения задачи статистического оценивания m параметров требуется выполнение условия $n > m$.

В качестве критериев оценивания истинных значений характеристик используются соотношения следующего вида:

$$\begin{cases} F_{\hat{x}}(x) \approx s(x;\ X_{<n>}); \\ A_{<m>} \approx S_{<m>}(X_{<n>}) \end{cases}. \qquad (3.1.3)$$

или

$$\begin{cases} s'(x;\ X_{<n>}) \le F_{\hat{x}}(x) \le s''(x;\ X_{<n>}); \\ S'_{<m>}(X_{<n>}) \le A_{<m>} \le S''_{<m>}(X_{<n>}), \end{cases} \qquad (3.1.4)$$

где $s'(x;\ X_{<n>})$, $s''(x;\ X_{<n>})$ – нижняя и верхняя границы интервалов; $S'_{<m>}(X_{<n>})$, $S''_{<m>}(X_{<n>})$ – границы m-мерных областей.

Оценивание вероятностных характеристик в соответствии с критерием (3.1.3) называется ***точечным***, а в соответствии с критерием (3.1.4) – интервальным. Строго говоря оценки всегда являются точечными. Что же касается интервальных оценок, то их назначение - характеризовать качество точечных оценок.

Предположим, что распределение $F_{\hat{x}}(x)$ однопараметрическое, т.е. $A_{<m>} = A_{<1>} = a$. Принятое допущение позволяет существенно повысить наглядность рассуждений, которые без затруднений распространяются на случай многопараметрического распределения. Кроме того, будем считать, что класс распределений, которому принадлежит функция $F_{\hat{x}}(x)$, известно, но неизвестно значение параметра a. В этом случае задача оценивания функции распределения $F_{\hat{x}}(x; a)$ сводится к оцениванию параметра, т.е. к определению соотношения вида

$$F_{\hat{x}}(x) = F_{\hat{x}}(x;\ a) \approx F_{\hat{x}}(x;\ \widetilde{a}) = \widetilde{F}_{\hat{x}}(x),$$

где $\widetilde{a} = S_{<1>}(X_{<n>}) = s(X_{<n>})$ – оценка параметра a.

Поскольку результаты $\hat{X}_{<n>}$ наблюдений над случайной величиной $\hat{x}$ априори являются случайными, то случайной оказывается и оценка $\widetilde{a}$:

$$\widetilde{a} = s(\hat{X}_{<n>}) = \hat{s}.$$

В общем случае $\widetilde{a} \neq a$, следовательно, и после получения оценки $\widetilde{a}$ параметра a его неопределённость для исследователя полностью не снимается. В то же время исследователь даёт вероятностное суждение об истинном значении a согласно результату эксперимента так, чтобы соответствовать ему наилучшим (в некотором смысле) образом. Оценка будет

объективной характеристикой параметра, если она удовлетворяет требованиям несмещённости, состоятельности и эффективности.

Оценка $\widetilde{a}$ параметра a называется ***несмещённой***, если её математическое ожидание равно оцениваемому параметру:

$$M_{\widetilde{a}} = a\,. \qquad (3.1.5)$$

Если $M_{\widetilde{a}} \neq a$, то оценка называется ***смещённой***.

Оценка $\widetilde{a}$ параметра a называется ***состоятельной***, если она сходится по вероятности к оцениваемому параметру:

$$\lim_{n\to\infty} P(|\,\widetilde{a} - a\,| \le \varepsilon) = 1, \quad \forall \varepsilon > 0\,, \qquad (3.1.6)$$

где n – объём выборки.

Очевидно, что состоятельной может быть только несмещённая оценка. Поскольку согласно известному неравенству Чебышева

$$P(|\,\widetilde{a} - a\,| \le \varepsilon) = 1 - \frac{D_{\widetilde{a}}}{\varepsilon^2}\,,$$

то из выражений (3.1.5) и (3.1.6) следует, что

$$\lim_{n\to\infty} D_{\widetilde{a}} = 0\,,$$

т.е. с ростом объёма выборки дисперсия состоятельной оценки стремится к нулю, и наоборот, если с ростом n дисперсия стремится к нулю, то оценка $\widetilde{a}$ состоятельная.

Несмещённая оценка[1] параметра a называется ***эффективной***, если её дисперсия минимальна:

$$D_{\widetilde{a}} = \min_k \{D_{\widetilde{a}_k} \mid k = 1,2,...\}\,, \qquad (3.1.7)$$

где $\widetilde{a}_k = s_k(\hat{X}_{<n>})$ – оценка параметра a с помощью статистики k-го вида.

Если равенство (3.1.7) выполняется только в пределе при $n \to \infty$, то соответствующая оценка называется ***асимптотически эффективной***. Из последнего определения следует, что состоятельная оценка асимптотически эффективна.

Оценки, удовлетворяющие всем трём перечисленным требованиям, называются ***подходящими значениями*** оцениваемых параметров. На практике достичь совместного выполнения всех трёх условий (3.1.5), (3.1.6) и (3.1.7) удаётся не всегда, так как формулы для вычисления эффективной и несмещённой оценки могут оказаться слишком сложными. Поэтому для упрощения расчётов нередко используются незначительно смещённые и не вполне эффективные оценки. Однако, выбор той или иной оценки должен опираться на её критическое рассмотрение со всех указанных выше точек зрения.

Следует отметить, что определение эффективной оценки имеет аналитическое выражение (3.1.7) лишь в случае оценивания единственно-

[1] Для смещённой оценки понятие эффективности не определено.

го параметра распределения $\varphi_{\hat{x}}(x;a)$. Если число m оцениваемых параметров $(a_1, a_2,\ldots, a_m)^{\mathsf{T}} = A_{<m>}$ больше единицы, то в качестве характеристики рассеяния их оценок $(\widetilde{a}_1, \widetilde{a}_2,...,\widetilde{a}_m)^{\mathsf{T}} = \widetilde{A}_{<m>}$ должна использоваться обобщённая дисперсия

$$D_{\widetilde{A}_{<m>}} \overset{d}{=} | K_{\widetilde{A}_{<m>}} |. \qquad (3.1.8)$$

В правой части соотношения (3.1.8) определитель корреляционной матрицы вектора $\widetilde{A}_{<m>}$ оценок параметров $A_{<m>}$. Более полное раскрытие понятия обобщённой дисперсии можно найти, например, в монографии [6].

Ещё раз подчеркнём, что следует различать две фазы оценивания вероятностных характеристик случайных объектов: априорную (доопытную) и апостериорную (послеопытную). На первой фазе оценки рассматриваются как функции случайной выборки и, следовательно, сами случайны. На второй фазе они не случайны, так как представляют собой функции выборки (реализации случайной выборки), элементы которой не случайны. Очевидно, что все требования к оценкам как законов, так и параметров распределения случайной величины предъявляются на первой фазе их оценивания, т.е. априори.

3.2. Качество статистического оценивания

Как отмечалось в § 3.1, при обработке данных принято различать точечное и интервальное оценивание вероятностных характеристик случайных объектов. Однако, строго говоря, собственно оценки $\widetilde{F}_{\hat{x}}(x)$, $\widetilde{A}_{<m>}$ и т.п., представляющие практический интерес, могут быть только точечными и определяются приближёнными равенствами типа (3.1.3). Что касается оценок (3.1.4), то они оценивают не характеристики $F_{\hat{x}}(x)$, $A_{<m>}$, а интервалы, в которых эти характеристики могут находиться, а могут и не находиться. Последнее утверждение обосновывается тем, что статистики $\hat{s}'$, $\hat{s}''$, $\hat{S}'_{<m>}$, $\hat{S}''_{<m>}$ априори случайны (как функции случайных аргументов – элементов случайной выборки $\hat{X}_{<n>}$). Поэтому ясно, что интервальное оценивание сугубо вероятностное и служит для характеристики качества точечного оценивания. Компонентами, которые характеризуют это качество, являются точность и надёжность (достоверность).

Раскроем существо задачи исследования точности и надёжности статистического оценивания. Пусть по результатам n наблюдений случайной величины $\hat{x}$ получена точечная оценка $\widetilde{a}$. Возникает вопрос: насколько эта оценка точна и надёжна.

Точность статистического оценивания характеризуется абсолютной погрешностью (ошибкой)

$$\Delta\widetilde{a} = \widetilde{a} - a\,.$$

Истинную погрешность Δa определить невозможно даже при известной оценке $\widetilde{a}$. Это объясняется тем, что исследователь не знает истинное значение параметра a. Поэтому вводится понятие вероятной погрешности статистической оценки параметра a.

Максимальной вероятной погрешностью статистической оценки называется её максимально возможное отклонение $\varepsilon_\beta > 0$ от оцениваемой характеристики случайного объекта, гарантируемое с вероятностью не менее β. Данная величина (ε_β) имеет также эквивалентное название - ***максимальная с вероятностью*** β ***погрешность оценки*** какой – либо характеристики.

Если оценка несмещённая, то её математическое ожидание равно оцениваемой характеристике, т.е. справедливо равенство (3.1.5). Тогда при симметричном распределении оценки $\widetilde{a}$ относительно математического ожидания имеет место следующее соотношение:

$$\beta = \beta(a) = P(a' \le \widetilde{a} \le a'') = P(a - \varepsilon_\beta \le \widetilde{a} \le a + \varepsilon_\beta) = P(|\,\widetilde{a} - a\,| \le \varepsilon_\beta)\,, \quad (3.2.1)$$

где $\varepsilon_\beta = \varepsilon_\beta(a)$ – максимальная с вероятностью β погрешность оценки $\widetilde{a}$ параметра a.

Соотношение (3.2.1) введено в предположении, что параметр a известен. Но тогда задача его оценивания теряет смысл, её просто не существует. На практике дело обстоит иначе. На основе экспериментальных данных определяется оценка параметра, истинное значение которого остаётся неизвестным. Затем вводится соотношение

$$\beta = P(|\,\widetilde{a} - a\,| \le \varepsilon_\beta) = P(\widetilde{a} - \varepsilon_\beta \le a \le \widetilde{a} + \varepsilon_\beta) = P(\widetilde{a}' \le a \le \widetilde{a}'')\,. \quad (3.2.2)$$

Следует подчеркнуть, что похожие на первый взгляд выражения (3.2.1) и (3.2.2) имеют различный вероятностный смысл. Так, (3.2.1) определяет вероятность того, что случайная величина $\widetilde{a}$ попадает в неслучайный интервал $[a - \varepsilon_\beta;\ a + \varepsilon_\beta]$, а (3.2.2) – вероятность того, что неслучайное (хотя и неизвестное) значение a оцениваемого параметра окажется в пределах случайного интервала $[\widetilde{a} - \varepsilon_\beta;\ \widetilde{a} + \varepsilon_\beta]$. Данный интервал является случайным как по величине, так и по расположению на вещественной оси, т.е. он накрывает точку a.

Интервал $I_\beta = [\widetilde{a} - \varepsilon_\beta;\ \widetilde{a} + \varepsilon_\beta]$ называется ***доверительным интервалом***, соответствующим ***доверительной вероятности*** β, или 100β-процентным доверительным интервалом. Его границы называются доверительными границами для параметра a.

Очевидно, чем у́же доверительный интервал, тем меньше максимальная с вероятностью β погрешность ε_β оценки параметра, тем она

точнее. С другой стороны, чем больше доверительная вероятность, тем более надёжна (достоверна) оценка, тем с бóльшим доверием можно к ней относиться.

Абсолютная достоверность оценивания характеризуется доверительной вероятностью $\beta = 1$. В условиях воздействия случайных факторов такая достоверность не достижима, поэтому реальная доверительная вероятность определяется на основе принципа практической уверенности. Согласно этому принципу события, имеющие вероятности, близкие к единице, считаются практически достоверными, а имеющие вероятности, близкие к нулю, – практически невозможными. Иначе говоря, если вероятность случайного события близка к единице (к нулю), то практически можно быть уверенным, что при однократном проведении опыта это событие произойдёт (не произойдёт).

Вероятность практически достоверного события определяется сущностью решаемой задачи. При анализе качества статистического оценивания обычно принимают $\beta \in [0{,}8;\ 0{,}99]$.

Как было показано в § 3.1 [см. формулу (3.1.4)], доверительные границы представляют собой статистики, т.е. некоторые функции элементов выборки $(x_1,\ x_2,\ldots,\ x_n)^{\mathrm{T}} = X_{<n>}$. В случае одномерного параметра a соотношение (3.1.4) принимает вид

$$s'(X_{<n>}) \le a \le s''(X_{<n>}),$$

где $s'(X_{<n>}) = a'$; $s''(X_{<n>}) = a''$.

Поскольку выборка $\hat{X}_{<n>}$ априори случайна, то и статистики $\hat{s}', \hat{s}''$, а следовательно, и доверительные границы $\hat{a}'$, $\hat{a}''$ априори случайны:

$$\hat{a}' = s'(\hat{X}_{<n>}); \qquad \hat{a}'' = s''(\hat{X}_{<n>}).$$

При анализе качества статистических оценок вся информация об исследуемой переменной содержится в случайной выборке $\hat{X}_{<n>}$. Поэтому не только оценка $\tilde{a}$ параметра, но и её максимальная вероятная погрешность ε_β определяется через выборку:

$$\begin{cases} \hat{a}' = s'(\hat{X}_{<n>}) = \tilde{a}(\hat{X}_{<n>}) - \varepsilon'_\beta(\hat{X}_{<n>}); \\ \hat{a}'' = s''(\hat{X}_{<n>}) = \tilde{a}(\hat{X}_{<n>}) + \varepsilon''_\beta(\hat{X}_{<n>}). \end{cases} \tag{3.2.4}$$

Соотношения (3.2.4) носят общий характер и в явном виде никогда не формируются. На практике погрешность ε_β оценки $\tilde{a}$ выражается через саму оценку:

$$\varepsilon_\beta = \varepsilon_\beta(\tilde{a}).$$

Таким образом, доверительные границы a', a'' для параметра a определяются его оценкой $\tilde{a}$:

$$\begin{cases} \hat{a}' = \tilde{a} - \varepsilon'_\beta(\tilde{a}); \\ \hat{a}'' = \tilde{a} + \varepsilon''_\beta(\tilde{a}), \end{cases} \qquad (3.2.5)$$

где функции $\varepsilon'_\beta(\tilde{a})$ и $\varepsilon''_\beta(\tilde{a})$ в общем случае различные. Соотношения (3.2.5) иллюстрируются рис.3.1.

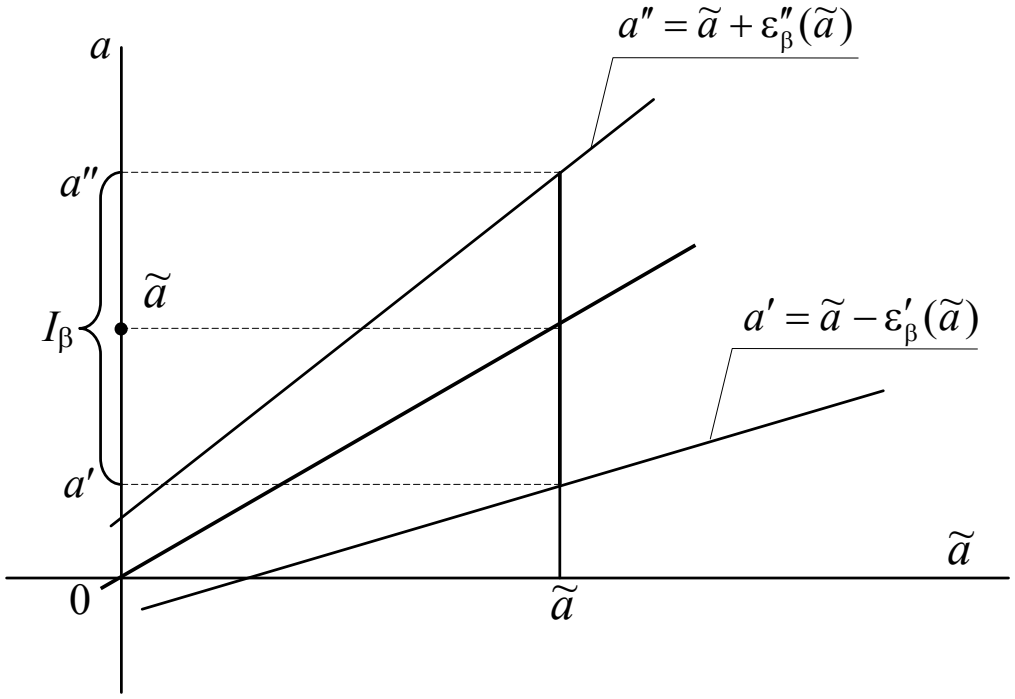


Рис.3.1. Зависимость доверительных границ параметра от его оценки

Выражения (3.2.1) и (3.2.2) были получены в предположении, что оценка $\tilde{a}$ имеет симметричное относительно параметра a распределение. При этом доверительный интервал $I_\beta = [a'; a'']$ оказывается также симметричным. В этом случае должны выполняться условия $\varepsilon'_\beta = \varepsilon''_\beta = \varepsilon_\beta$, которые означают равенство положительной и отрицательной максимальных вероятных погрешностей. В общем случае это требование не обязательно и при определении ε'_β и ε''_β в основу могут быть положены различные соображения. Чаще всего выдвигается требование

$$P(a < \hat{a}') = P(a > \hat{a}'') = \frac{1-\beta}{2}, \qquad (3.2.6)$$

означающее, что истинное значение a оцениваемого параметра может оказаться левее или правее его доверительного интервала с одинаковой вероятностью.

При условии (3.2.6) и симметричном распределении оценки $\tilde{a}$ доверительный интервал для параметра a будет симметричным:

$$I_\beta(a) = [a'; a''] = [\tilde{a} - \varepsilon_\beta; \tilde{a} + \varepsilon_\beta].$$

Следовательно, имеют место равенства

$$\beta = P(a \in I_\beta(a)) = P(\widetilde{a} - \varepsilon_\beta \le a \le \widetilde{a} + \varepsilon_\beta) = P(|\widetilde{a} - a| \le \varepsilon_\beta) = \beta_\varepsilon(a). \qquad (3.2.7)$$

Если распределение оценки $\widetilde{a}$ не симметрично относительно оцениваемого параметра a, то при условии (3.2.6) доверительный интервал не симметричен, и соотношение (3.2.7) принимает вид

$$\beta = P(a \in I_\beta(a)) = P(\widetilde{a} - \varepsilon'_\beta \le a \le \widetilde{a} + \varepsilon''_\beta),$$

где ε'_β и ε''_β – соответственно абсолютные значения максимальных с вероятностью β отрицательной и положительной погрешностей оценки $\widetilde{a}$ параметра a.

Итак, доверительный интервал (его составляющие ε'_β и ε''_β) характеризует точность, доверительная вероятность β – надёжность (достоверность) оценки $\widetilde{a}$, а вместе они определяют качество оценивания параметра a.

В § 3.1 отмечалось, что с ростом объёма n выборки оценка $\widetilde{a}$ сходится по вероятности к оцениваемому параметру a (закон больших чисел) и при этом её дисперсия стремится к нулю. Это значит, что с увеличением n растёт как точность, так и надёжность оценивания. В результате оказываются связанными между собой три характеристики качества статического оценивания:

– доверительный интервал $I_{\beta,n}(a)$;

– доверительная вероятность $\beta_{\varepsilon,n}(a)$;

– объём $n_{\beta,\varepsilon}(a)$ выборки, потребный для оценивания параметра a с заданной точностью и надёжностью.

Указанные характеристики связаны соотношениями, позволяющими управлять качеством статистического оценивания[1]

$$\left.\begin{aligned} n\uparrow &\Rightarrow \begin{cases} \varepsilon = \text{const} \Rightarrow \beta\uparrow \\ \beta = \text{const} \Rightarrow \varepsilon\downarrow \end{cases} \\ \beta\uparrow &\Rightarrow \begin{cases} n = \text{const} \Rightarrow \varepsilon\uparrow \\ \varepsilon = \text{const} \Rightarrow n\uparrow \end{cases} \\ \varepsilon\downarrow &\Rightarrow \begin{cases} n = \text{const} \Rightarrow \beta\downarrow \\ \beta = \text{const} \Rightarrow n\uparrow \end{cases} \end{aligned}\right\} \qquad (3.2.8)$$

Таким образом, при исследовании качества статистического оценивания решается одна из трёх основных задач:

– определение доверительного интервала $I_{\beta,n}$ (или половины его длины $\varepsilon_{\beta,n}$) для параметра a при заданной доверительной вероятности β и фиксированном объёме n выборки;

---

[1] Символы $\uparrow$, $\downarrow$ означают соответственно возрастание и убывание.

– определение доверительной вероятности $\beta_{I,n}$ (или $\beta_{\varepsilon,n}$) при заданном доверительном интервале I (или максимальной вероятной погрешности ε) и фиксированном объёме n выборки;

– определение объёма $n_{\beta,I}$ (или $n_{\beta,\varepsilon}$) выборки, потребного для оценивания параметра a с требуемой надёжностью β и точностью I (или ε).

3.3. Оценивание вероятности случайного события

В результате реализации определённого комплекса условий может произойти некоторое случайное событие A, вероятность $P(A) = p$ появления которого неизвестна. Требуется по результатам наблюдений данного события в некотором эксперименте оценить вероятность p.

Для решения поставленной задачи проводится серия n независимых и однородных испытаний – схема Бернулли, т.е. осуществляется n независимых реализаций одного и того же комплекса условий. Подсчитывается число $m(A) = m$ испытаний, в которых событие A появилось. Отношение

$$\frac{m}{n} = P^*(A) = p^*(n) \qquad (3.3.1)$$

называется *частотой* события A в серии n испытаний или его *статистической вероятностью*. Проанализируем свойства частоты p^* как оценки вероятности p.

1. Поскольку число $\hat{m}$ появлений события A в n независимых и однородных испытаниях подчинено биномиальному закону распределения, то

$$M_{p^*} = \frac{1}{n} M_{\hat{m}} = \frac{np}{n} = p\,. \qquad (3.3.2)$$

Из выражения (3.3.2) следует, что частота (3.3.1) является несмещённой оценкой вероятности p.

2. Согласно теореме Бернулли

$$\lim_{n\to\infty} P(|\, p^* - p \,| < \varepsilon) = 1, \quad \forall \varepsilon > 0\,,$$

т.е. частота p^* сходится по вероятности к вероятности p. Следовательно, рассматриваемая частота – это состоятельная оценка вероятности p.

3. Дисперсия частоты

$$D_{p^*} = \frac{1}{n^2} D_{\hat{m}} = \frac{npq}{n^2} = \frac{pq}{n}\,, \qquad (3.3.3)$$

где $q = 1 - p$. Из соотношения (3.3.3) вытекает, что при $n \to \infty$ дисперсия $D_{p^*} \to 0$. Это означает асимптотическую эффективность указанной оцен-

ки. Можно показать, что при любом n дисперсия частоты - минимально возможная величина, следовательно, p^* является эффективной оценкой p.

Таким образом, частота p^* события A в серии n независимых однородных испытаний есть подходящее значение его вероятности, т.е. наилучшая точечная оценка.

Исследуем качество оценивания вероятности p по его частоте p^*. Итак, полагаем, что

$$p \approx \tilde{p} = p^* = \frac{m}{n}.$$

Априори число $\hat{m}$ случайно и подчинено биномиальному закону распределения с параметрами n, p. Согласно теореме Муавра-Лапласа при достаточно больших n (практически при $np(1-p) > 9$) биномиальное распределение может быть с достаточной точностью аппроксимировано нормальным распределением с параметрами $M_{\hat{z}} = np$, $\sigma_{\hat{z}} = \sqrt{np(1-p)}$. В этом случае справедливо соотношение

$$F_{\hat{m}}(z) = F^{\text{б}}_{\hat{m}}(z; n; p) \approx F^{\text{н}}_{\hat{z}}\left(z; np; \sqrt{np(1-p)}\right) = \Phi_1\left(\frac{z-np}{\sqrt{np(1-p)}}\right).$$

Поскольку оценка $\tilde{p} = \hat{m}/n$ связана с $\hat{m}$ линейной зависимостью, она будет распределена приближённо нормально с параметрами

$$M_{\tilde{p}} = p; \quad \sigma_{\tilde{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Тогда справедливо

$$F_{\tilde{p}}(\vartheta) \approx F_{\tilde{\vartheta}}\left(\vartheta; p; \sqrt{\frac{p(1-p)}{n}}\right) = \Phi_1\left(\frac{(\vartheta-p)\sqrt{n}}{\sqrt{p(1-p)}}\right). \qquad (3.3.4)$$

Так как закон распределения (3.3.4) оценки $\tilde{p}$ симметричен относительно оцениваемой вероятности p, доверительный интервал $I_{\beta,n}(p)$ будет симметричен относительно оценки $\tilde{p}$. Для определения данного интервала достаточно знать половину его длины, которая равна максимальной с доверительной вероятностью $\beta(p)$ абсолютной погрешности $\varepsilon(p)$:

$$\varepsilon'_{\beta,n} = \varepsilon''_{\beta,n} = \varepsilon_{\beta,n} = \varepsilon.$$

В результате доверительная вероятность для p будет определяться следующим равенством:

$$\begin{aligned} \beta = \beta_{I,n} = \beta_{\varepsilon,n} = P(|\tilde{p} - p| \le \varepsilon) = 2\Phi_0\left(\frac{\varepsilon}{\sigma_{\tilde{p}}}\right) = \\ = 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\tilde{p}(1-\tilde{p})}}\right). \end{aligned} \qquad (3.3.5)$$

Разрешив уравнение (3.3.5) относительно ε, получим

$$\varepsilon = \varepsilon_{\beta,n} = t_\beta \sqrt{\frac{p(1-p)}{n}} \approx t_\beta \sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n}} = \widetilde{\varepsilon}\,, \qquad (3.3.6)$$

откуда

$$I = I_{\beta,n} = [\,p';\, p''\,] = [\,\widetilde{p} - \varepsilon;\ \widetilde{p} + \varepsilon\,] \approx [\,\widetilde{p} - \widetilde{\varepsilon};\ \widetilde{p} + \widetilde{\varepsilon}\,]. \qquad (3.3.7)$$

В выражении (3.3.6) величина t_β – квантиль нормированного нормального распределения:

$$t_\beta = \Phi_0^{-1}\left(\frac{\beta}{2}\right) \quad \text{или} \quad t_\beta = \Phi_1^{-1}\left(\frac{1+\beta}{2}\right).$$

Значения функции t_β приведены в приложении 4.

Если необходимые точность ε и надёжность β заданы, то потребное для их обеспечения число $n_{\beta,n}$ испытаний находится из уравнения (3.3.6):

$$n = n_{\beta,n} \geq \frac{p(1-p)}{\varepsilon^2} t_\beta^2 \approx \frac{\widetilde{p}(1-\widetilde{p})}{\varepsilon^2} t_\beta^2\,, \qquad (3.3.8)$$

Формулы (3.3.5) – (3.3.8) определяют решения трёх основных задач исследования качества статистического оценивания (см. § 3.2) применительно к оценке вероятности случайного события по его частоте в серии n независимых однородных испытаний.

Из соотношения (3.3.8) видно, что потребный объём выборки обратно пропорционален квадрату максимальной вероятной погрешности ε оценки $\widetilde{p}$ и пропорционален квадрату функции t_β, который растёт быстрее, чем β. Поэтому для оценивания вероятности случайного события по его частоте с достаточной точностью и надёжностью требуется проведение довольно длинной серии испытаний. Сказанное иллюстрируется табл.3.1, в которой приведены потребные числа $n_{0,95;\varepsilon}$ испытаний, обеспечивающие с доверительной вероятностью $\beta = 0{,}95$ необходимую точность ε оценивания различных значений вероятности p.

Таблица 3.1

Зависимость числа испытаний от требуемой доверительной вероятности

ε	p				
	0,9	0,8	0,7	0,6	0,5
0,05	139	246	323	369	385
0,01	3458	6147	8068	9220	9604

Из табл.3.1 видно, что потребное число $n_{\beta;\varepsilon}$ испытаний растёт не только с увеличением необходимой точности оценивания, но и с приближением истинного значения p оцениваемой вероятности к 0,5. Это объяснимо, поскольку при $p = 0{,}5$ дисперсия оценки $\widetilde{p} = p^*$ достигает максимального значения, равного $0{,}25/n$ [см. формулу (3.3.3)]. Указанный факт используется для определения верхней границы потребного числа испы-

таний. Так, полагая $p = 0{,}5$, $\beta = 0{,}95$, имеем значение $t_\beta = t_{0{,}95} = 1{,}96 \approx 2$ (см. приложение 4). В соответствии с выражением (3.3.8) получаем

$$n = n_{0{,}95;n} \geq \frac{0{,}5 \cdot 0{,}5}{\varepsilon^2} 1{,}96^2 \approx \frac{1}{\varepsilon^2}. \qquad (3.3.9)$$

Пример 3.1. В процессе эксперимента выполнено 200 опытов, частота события A оказалась $p^* = 0{,}34$.

1. Построить 85%-й доверительный интервал для вероятности события A.

2. Найти доверительную вероятность β для вероятности события A, если максимальная вероятная погрешность $\varepsilon_\beta = 0{,}1$.

▼ 1) Для $\beta = 0{,}85$ в приложении 4 находим $t_\beta = 1{,}439$. Тогда по формуле (3.3.6) оценка максимальной вероятной ошибки составит

$$\widetilde{\varepsilon} = 1{,}439\sqrt{\frac{0{,}34(1-0{,}34)}{200}} = 0{,}048.$$

Находим доверительный интервал из соотношения (3.3.7)

$$I_{0{,}85;\,200} \approx [0{,}34 - 0{,}048;\ 0{,}34 + 0{,}048] = [0{,}292;\ 0{,}388].$$

2) По формуле (3.3.5) находим доверительную вероятность

$$\beta_{0{,}1;200} \approx 2\Phi_0\left(\frac{0{,}1\sqrt{200}}{\sqrt{0{,}34(1-0{,}34)}}\right) = 2\Phi_0\left(\frac{1{,}414}{0{,}474}\right) = 2\Phi_0(2{,}98) = 2 \cdot 0{,}4986 = 0{,}9972.$$

Значение функции $\Phi_0(x)$ взято из приложения 2.

▲

Пример 3.2. В процессе эксперимента выполняются опыты, частота события составляет $p^* = 0{,}7$.

1. Определить потребный объём выборки, чтобы максимальная вероятная погрешность оценки p^* составляла $\varepsilon \leq 0{,}05$ при доверительной вероятности $\beta = 0{,}9$.

2. Найти верхнюю границу потребного числа опытов при любой частоте события .

▼ 1) По заданному β находим $t_\beta = 1{,}643$. Тогда в соответствии с формулой (3.3.8) потребный объём выборки составит

$$n \geq \frac{0{,}7(1-0{,}7)}{(0{,}05)^2} \cdot 1{,}643 = 138.$$

2) Из выражения (3.3.9) имеем

$$n \geq \frac{1}{0{,}0025} = 400.$$

▲

4. ОЦЕНИВАНИЕ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Оценками законов распределения случайных величин являются ***статистические законы распределения***. Их построение осуществляется на основе информации, содержащейся в выборке $\hat{X}_{<n>}$.

В дальнейшем, в рамках раздела 4 предполагается, что выборка $\hat{X}_{<n>}$ простая, т.е. повторная выборка из распределения

$$\varphi_{\hat{x}}(x) = \varphi_{\hat{x}}(x;\ A_{<m>}),$$

где $A_{<m>}$ - вектор параметров распределения.

4.1. Статистические ряды распределения

При проведении серии испытаний экспериментальные данные (выборка) представляются в виде табл.4.1, которая называется простым статистическим рядом.

Таблица 4.1

Простой статистический ряд

Номера испытаний	1	2	3	…	i	…	n
Варианты признака $\hat{x}$	x_1	x_2	x_3	…	x_i	…	x_n

Если элементы случайной выборки упорядочены по возрастанию, т.е.

$$x_1 \le x_2 \le x_3 \le \ldots \le x_i \le \ldots \le x_n,$$

получаемая таблица называется ***вариационным рядом***. Разность $w_n = x_n - x_1$ между наибольшим и наименьшим элементами вариационного ряда называется *размахом выборки*.

Вариационный ряд является простейшей формой статистического закона распределения, который определяется в результате эксперимента.

Если наблюдаемый случайный признак является дискретным, или точность измерений ограничена, или результаты наблюдений округляются, значения некоторых вариантов признака в выборке могут совпадать. Множества совпадающих вариантов наблюдаемого признака называются *связками*. Из этого следует, что различные варианты могут появляться с разной частотой.

Таблица 4.2

Вариационный ряд с вариантами различной частоты

Варианты признака $\hat{x}$	x_1	x_2	x_3	…	x_l	…	x_n
Частота вариантов	m_1/n	m_2/n	m_3/n	…	m_l/n	…	m_n/n

В табл. 4.2 представлен вариационный ряд, в котором m_l – число появлений в выборке варианта x_l; $P_l^* = m_l / n$ – частота варианта x_l, $l = \overline{1, n}$. Такой ряд можно рассматривать как ряд распределения некоторой дискретной случайной величины.

Пример 4.1. Проведены испытания 12 однотипных микросхем и с точностью до 1 часа зарегистрировано время безотказной работы каждой из них. Результаты испытаний сведены в простой статистический ряд.

Таблица 4.3

Простой статистический ряд (к примеру 4.1)

i	1	2	3	4	5	6	7	8	9	10	11	12
τ_i, ч	30	108	36	69	117	161	143	500	108	135	89	36

На основе простого статистического ряда, табл.4.3, построен вариационный ряд.

Таблица 4.4

Вариационный ряд (к примеру 4.1)

l	1	2	3	4	5	6	7	8	9	10	11	12
τ_l, ч	30	36	36	69	89	108	108	117	135	143	161	500

Массив данных в табл. 4.4 может быть представлен вариационным рядом с частотами вариантов.

Таблица 4.5

Вариационный ряд с частотами вариантов (к примеру 4.1)

τ_l, ч	30	36	69	89	108	117	135	143	161	500
P_l^*	1/12	2/12	1/12	1/12	2/12	1/12	1/12	1/12	1/12	1/12

Из табл.4.5 видно, что в выборке содержатся две связки, в каждую из которых входит два варианта.

По распределению величины $\hat{\tau}$ – времени безотказной работы микросхем, можно судить о распределении генеральной совокупности, тем точнее и надёжнее, чем больше объём n выборки.

При большом объёме выборки из генеральной совокупности, наблюдаемый признак которой имеет непрерывное распределение, вариационный ряд (табл. 4.2) становится громоздким. В этом случае диапазон наблюдаемых вариантов x_i разбивают на интервалы, называемые разрядами, вычисляют частоты P_l^* попадания вариантов случайной величины $\hat{x}$ в эти разряды и представляют результаты в виде табл. 4.6. Эта таблица называется ***интервальным вариационным*** или ***статистическим рядом*** случайной величины $\hat{x}$.

Таблица 4.6

Интервальный вариационный ряд

J_l	$x_1; x_2$	$x_2; x_3$	…	$x_l; x_{l+1}$	…	$x_r; x_{r+1}$
m_l	m_1	m_2	…	m_l	…	m_r
P_l^*	P_1^*	P_2^*	…	P_l^*	…	P_r^*

В представленной таблице $J_l = [x_l; x_{l+1})$ – l-й разряд; m_l – число значений (вариантов) случайной величины $\hat{x}$, попавших в l-й разряд; x_l; x_{l+1} – границы l-го разряда; r – число разрядов; $P_l^* = m_l/n = P^*(x_l \le x_i < x_{l+1})$, $l = \overline{1, r}$.

Число r разрядов, на которые разбивается случайная выборка, не должно быть слишком большим. В этом случае частоты подвергаются незакономерным колебаниям и статистический ряд становится невыразительным. Указанное число не должно быть и слишком малым, так как описание распределения случайной величины становится грубым. Чем однороднее статистический материал и чем больше его объём, тем большее число разрядов можно выбирать. Для ориентировочного определения числа разрядов можно пользоваться соотношениями

$$r \approx 5 \lg n \quad \text{или} \quad r \approx \sqrt{n}\,.$$

При этом целесообразно, чтобы выполнялись условия $5 < r < 25$, $m_l \ge 5$, $l = \overline{1, r}$. Длины разрядов можно брать как одинаковыми, так и различными. В последнем случае статистическая обработка экспериментальных данных несколько усложняется, однако при значительной неравномерности распределения наблюдаемой случайной величины в областях её наибольшей изменчивости разряды целесообразно делать более короткими.

Примечание. Если наблюдаемый признак $\hat{x}$ представляет собой дискретную случайную величину, то статистический ряд её распределения может иметь лишь форму табл.4.2. Для такого признака интервальный статистический ряд лишён физического смысла.

Пример 4.2. Проведены испытания 100 однотипных микросхем и зарегистрировано время безотказной работы каждой из них в часах. Результаты испытаний сведены в табл.4.7.

Таблица 4.7

Результаты испытаний микросхем (к примеру 4.2)

i	τ_i	i	τ_i	i	τ_i	i	τ_i	i	τ_i
1	151,5	21	151,5	41	2,0	61	105,0	81	21,1
2	190,0	22	38,5	42	56,2	62	30,0	82	107,8
3	67,2	23	190,0	43	211,4	63	73,3	83	111,0
4	38,5	24	46,0	44	151,5	64	138,8	84	123,8
5	281,0	25	301,7	45	40,2	65	84,3	85	301,7
6	156,4	26	261,2	46	114,1	66	33,0	86	13,1
7	38,5	27	40,2	47	3,0	67	3,9	87	8,5
8	18,6	28	120,3	48	212,0	68	190,0	88	105,0
9	34,4	29	26,3	49	114,1	69	96,7	89	241,4
10	58,2	20	63,3	50	51,3	60	267,0	90	18,6
11	18,5	31	105,0	51	177,4	71	494,8	91	75,3
12	49,5	32	30,0	52	30,0	72	105,0	92	24,5
13	204,3	33	156,2	53	73,2	73	33,0	93	13,1
14	65,2	34	67,2	54	107,8	74	46,0	94	67,2
15	86,7	35	89,1	55	71,2	75	27,8	95	123,8
16	107,8	36	386,2	56	102,2	76	9,5	96	86,7
17	43,2	37	58,2	57	40,2	77	54,2	97	230,3
18	21,1	38	34,4	58	120,3	78	5,8	98	167,1
19	107,8	39	75,3	59	26,3	79	40,2	99	34,4
20	7,2	40	221,1	60	63,3	80	22,3	100	117,2

На основе статистического материала табл.4.7 построен интервальный вариационный ряд, который приведён в табл.4.8.

Таблица 4.8

Интервальный вариационный ряд (к примеру 4.2)

J_l	0; 50	50; 100	100; 150	150; 200	200; 250
m_l	38	21	18	10	6
P_l^*	0,38	0,21	0,18	0,1	0,06
J_l	250; 300	300; 350	350; 400	400; 450	450; 500
m_l	3	2	1	0	1
P_l^*	0,03	0,02	0,01	0,00	0,01

Для большей наглядности ряд может оформляться графически в виде полигона. При его построении определяются «представители» разрядов ряда распределения, т.е. их средние точки с абсциссами

$$\bar{x}_l = 0{,}5(x_l + x_{l+1}), \quad l = \overline{1, r}.$$

Из этих точек восстанавливаются перпендикуляры, длины которых равны частотам P_l^*. Описанный переход от вариационного ряда к статистическому можно интерпретировать как переход к новой дискретной случайной величине $\hat{\bar{x}}$, принимающей значения $\bar{x}_l$ с «вероятностями» P_l^*.

Совокупность точек A_1, A_2,…, A_r, которые лежат на верхних концах перпендикуляров, называется ***огивой распределения*** случайной величины $\hat{\tilde{x}}$ [9]. Соединив смежные точки огивы отрезками прямых, получим полигон распределения случайной величины $\hat{\tilde{x}}$, который является полной аналогией многоугольника распределения. На рис.4.1 изображён полигон, который построен на основании статистического материала примера 4.2.

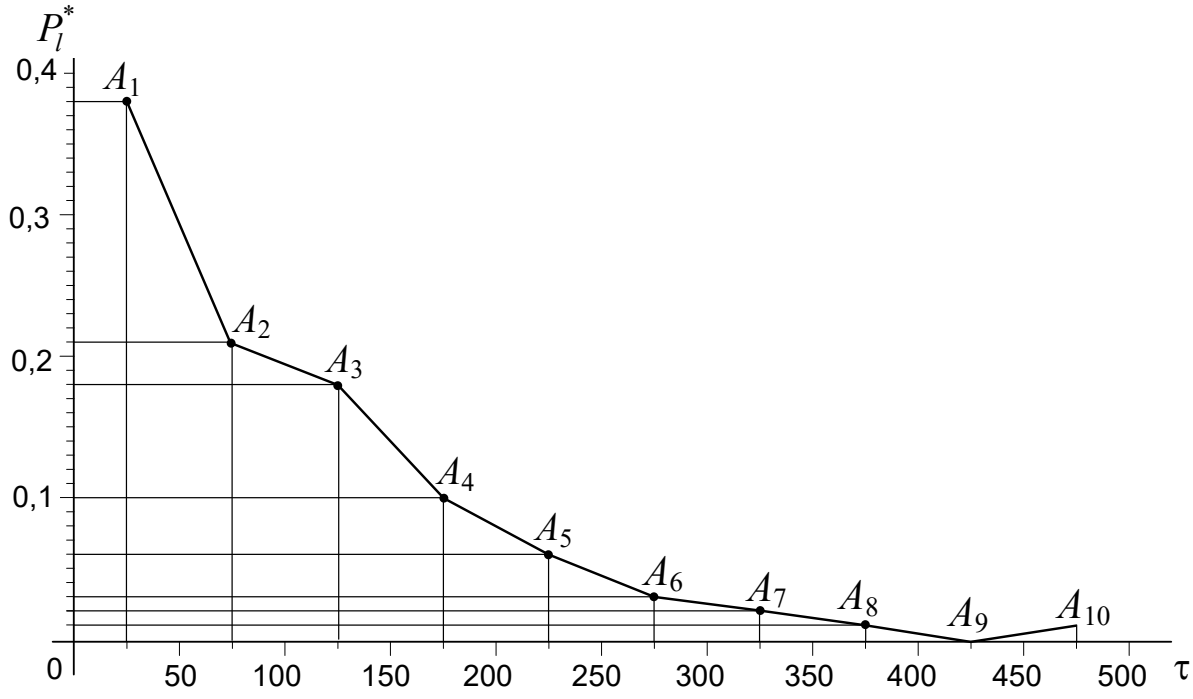


Рис.4.1. Полигон распределения (к примеру 4.2)

4.2. Статистические плотности распределения

Если экспериментальные данные представлены случайной величиной непрерывного типа, необходимо искать и более наглядную, чем интервальный вариационный ряд, форму статистического закона её распределения. Для этого частоту P_l^* попадания наблюдаемой случайной величины в соответствующий разряд статистического ряда необходимо распределить по всем её значениям из данного разряда. Существуют различные способы выполнения такой операции, из которых наиболее употребительны способы **полигона** и **гистограммы**.

4.2.1. Нормированный полигон распределения

При способе полигона предполагается, что разряды статистического ряда имеют длину

$$h_l = x_{l+1} - x_l$$

и что частоты попадания в разряды наблюдаемых значений случайной величины $\hat{x}$ плавно изменяются в пределах разрядов по линейному закону. Тогда, выполняя деление ординаты полигона на h_l, получим нормированный полигон распределения случайной величины $\hat{x}_r$. При этом крайние точки A'_1 и A'_r нормированной огивы следует соединить горизонтальными отрезками прямых с точками A'_0 и A'_{r+1} соответственно. Для примера 4.2 нормированный полигон распределения показан на рис.4.2. Пунктиром показана кривая показательного распределения, которому подчиняется время безотказной работы микросхем.

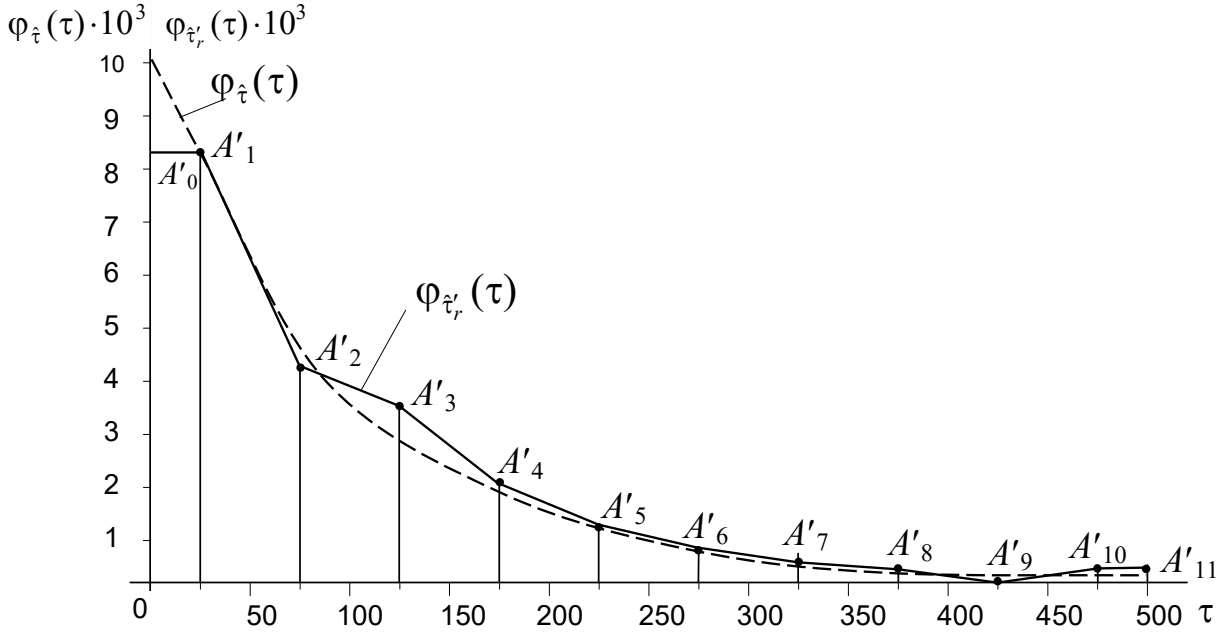


Рис.4.2. Нормированный полигон распределения (к примеру 4.2)

Нетрудно видеть, что построенный многоугольник A'_0, $A'_1,\ldots,A'_{r+1}$ представляет собой кривую распределения некоторой непрерывной случайной величины $\hat{x}'_r$, а описывающая её зависимость

$$y = \varphi_{\hat{x}'_r}(x)$$

обладает всеми свойствами плотности распределения, в частности

$$\int_{x_1}^{x_{r+1}} \varphi_{\hat{x}'_r}(x)dx = 1.$$

4.2.2. Гистограмма распределения

При способе гистограммы предполагается, что в пределах l-го разряда статистического ряда плотность распределения непрерывной случайной величины $\hat{x}''_r$ постоянна и равна

$$\varphi_{\hat{x}''_r}(\overline{x}_l) = \frac{P_l^*}{h_l},$$

где $h_l = x_{l+1} - x_l$ – длина l-го разряда.

Проводя через точки A'_1, $A'_2, \ldots, A'_r$ нормированной огивы горизонтальные отрезки прямых, получают семейство прямоугольников, называемое ***гистограммой распределения*** случайной величины $\hat{x}$. На рис.4.3 приведена гистограмма по данным примера 4.2. Для сравнения там же пунктиром изображена теоретическая кривая показательного распределения. Легко заметить, что площади прямоугольников, составляющих гистограмму, равны соответствующим частотам, а площадь всей гистограммы равна единице:

$$\int_{x_1}^{x_{r+1}} \varphi_{\hat{x}''_r}(x)dx = 1.$$

Следовательно, огибающая гистограммы обладает свойствами кривой распределения, а описывающая её зависимость

$$y = \varphi_{\hat{x}''_r}(x)$$

имеет свойства плотности распределения.

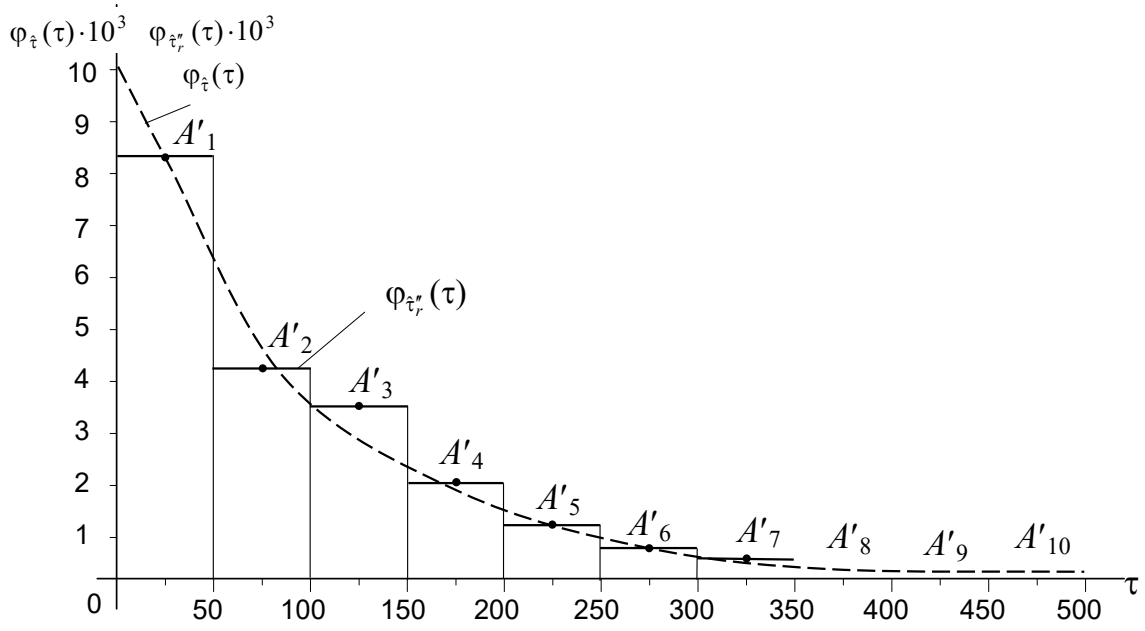


Рис.4.3. Гистограмма распределения (к примеру 4.2)

С увеличением объёма n выборки и, следовательно, числа разрядов статистического ряда огибающие полигона $\varphi_{\hat{x}'_r}(x)$ и гистограммы $\varphi_{\hat{x}''_r}(x)$ всё более приближаются к кривой распределения $\varphi_{\hat{x}}(x)$ случайной величины $\hat{x}$. Таким образом, они могут использоваться для приближённого описания плотности её распределения:

$$\varphi_{\hat{x}}(x) \approx \varphi_{\hat{x}'_r}(x) = \widetilde{\varphi}_{\hat{x}}(x)$$

или

$$\varphi_{\hat{x}}(x) \approx \varphi_{\hat{x}_r''}(x) = \widetilde{\varphi}_{\hat{x}}(x).$$

4.3. Статистические функции распределения

4.3.1. Выборочная функция распределения

По вариационному ряду табл.4.2 можно построить ***статистическую*** или ***выборочную функцию распределения*** $F^*_{\hat{x}}(x)$ случайной величины $\hat{x}$.

По определению

$$F^*_{\hat{x}}(x) = F^*_{\hat{x}}(x;n) = \sum_{x_k < x} P^*(x_i = x_k).$$

Следовательно, в явном виде статистическая функция распределения примет вид

$$F^*_{\hat{x}}(x) = \begin{cases} 0, & x \le x_1 \\ \dfrac{m_1}{n}, & x_1 < x \le x_2 \\ \cdots\cdots\cdots\cdots \\ \dfrac{1}{n}\sum\limits_{i=1}^{k} m_i, & x_k < x \le x_{k+1} \\ \cdots\cdots\cdots\cdots \\ 1, & x > x_n \end{cases}. \qquad (4.3.1)$$

График функции (4.3.1) в условиях примера 4.1 показан на рис.4.4. Обоснованием применимости функции $F^*_{\hat{x}}(x;n)$ для оценивания истинной функции распределения $F_{\hat{x}}(x)$ случайной величины $\hat{x}$ служит предельная теорема В.И. Гливенко, которая формулируется следующим образом.

При увеличении объёма n выборки статистическая функция распределения $F^*_{\hat{x}}(x;n)$ неограниченно приближается (сходится по вероятности) к истинной функции распределения $F_{\hat{x}}(x)$ случайной величины $\hat{x}$:

$$\lim_{n\to\infty} P\left(\max_x \left\{ \left| F^*_{\hat{x}}(x;n) - F_{\hat{x}}(x) \right| \right\} < \varepsilon\right) = 1, \quad \forall\, \varepsilon > 0. \qquad (4.3.2)$$

Таким образом, $F^*_{\hat{x}}(x)$ – состоятельная оценка $F_{\hat{x}}(x)$. Известно также, что функция $F^*_{\hat{x}}(x)$ является несмещённой оценкой для $F_{\hat{x}}(x)$:

$$M[F^*_{\hat{x}}(x)] = F_{\hat{x}}(x).$$

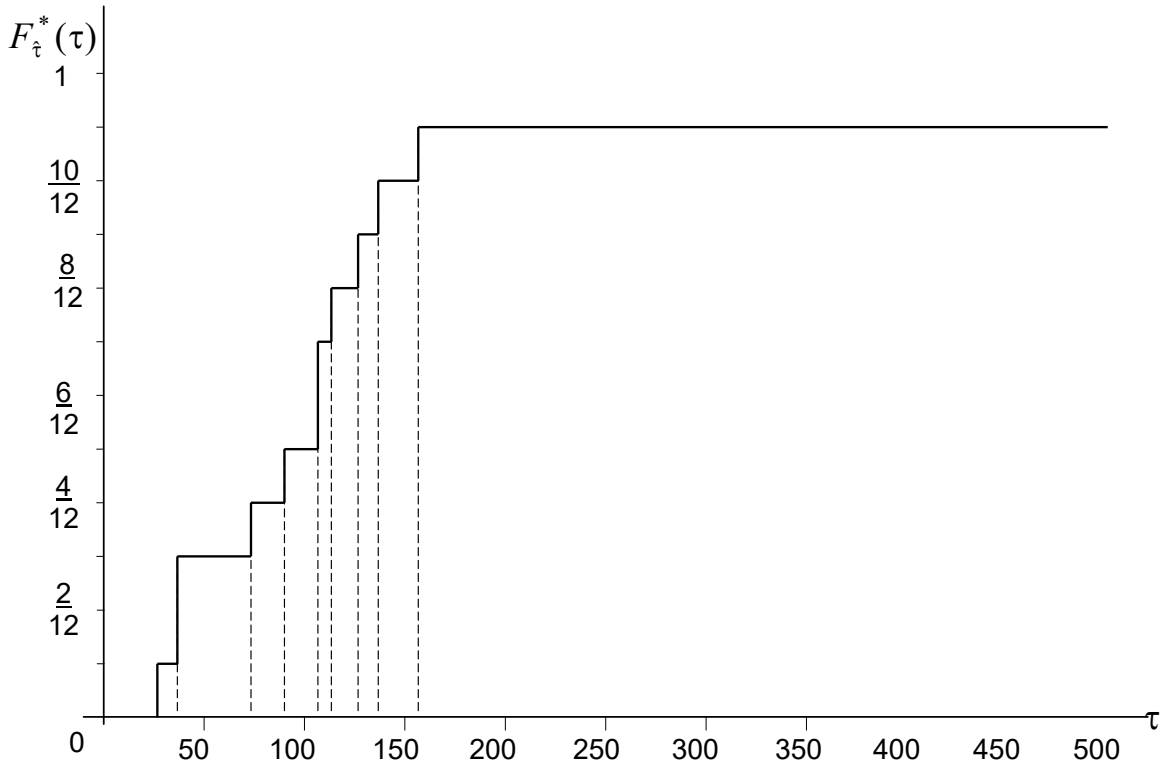


Рис.4.4. Статистическая функция распределения (к примеру 4.1)

В силу соотношения (4.3.2) дисперсия $D[F^*_{\hat{x}}(x)]$ с ростом n уменьшается, следовательно, эта оценка является асимптотически эффективной. Отсюда вытекает, что при достаточно большом массиве экспериментальных данных функцию распределения изучаемой случайной величины можно приближённо заменять её выборочной функцией распределения.

4.3.2. Кумулята распределения

Используя полигон или гистограмму, можно построить статистическую функцию распределения (так называемую кумуляту распределения случайной величины) путём интегрирования функции $\widetilde{\varphi}_{\hat{x}}(x)$:

$$F_{\hat{x}}(x) \approx F^*_{\hat{x}}(x) = \widetilde{F}_{\hat{x}}(x) = \int_{-\infty}^{x} \widetilde{\varphi}_{\hat{x}}(z)dz\,.$$

При численном интегрировании получим

$$\widetilde{F}_{\hat{x}}(x)=\begin{cases} 0, & x \le x_1; \\ P_1^*\dfrac{\Delta x_1}{h_1}, & x_1 < x \le x_2; \\ P_1^* + P_2^*\dfrac{\Delta x_2}{h_2}, & x_2 < x \le x_3; \\ \cdots\cdots & \\ \sum\limits_{i=1}^{l-1} P_i^* + P_l^*\dfrac{\Delta x_l}{h_l}, & x_l < x \le x_{l+1}; \\ \cdots\cdots & \\ 1, & x > x_r, \end{cases} \qquad (4.3.3)$$

где $h_l = x_{l+1} - x_l$, $\Delta x_l = x - x_l$.

График функции (4.3.3) в условиях примера 4.2 показан на рис.4.5, на котором пунктиром изображён график теоретической функции показательного закона распределения.

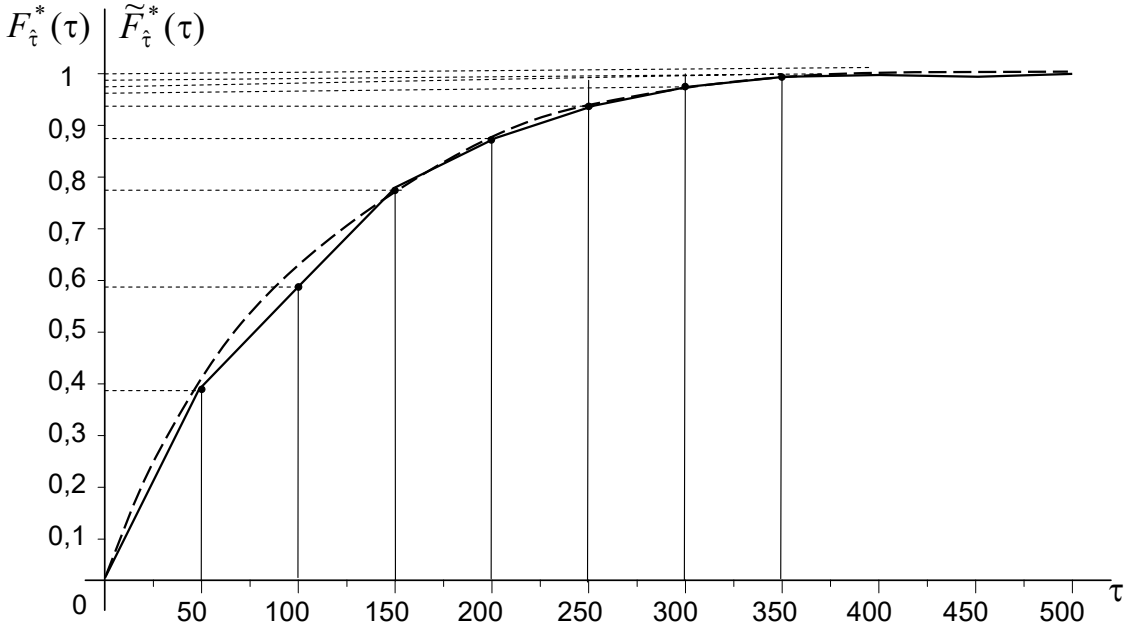


Рис.4.5. Кумулята распределения (к примеру 4.2)

Примечание. Поскольку функция распределения дискретной случайной величины ступенчатая, то её оценка может иметь лишь форму рис.4.4. Другими словами для дискретной случайной величины кумуляты не существует.

4.3.4. Качество оценивания функций распределения

По определению функция распределения $F_{\hat{x}}(x)$ случайной величины $\hat{x}$ есть вероятность случайного события:

$$F_{\hat{x}}(x) \overset{d}{=} P(\hat{x} < x).$$

Пусть в качестве её оценки $\widetilde{F}_{\hat{x}}(x)$ используется частота статистического аналога $x_i < x$ события $\hat{x} < x$ в серии n испытаний:

$$\widetilde{F}_{\hat{x}}(x) = F^*_{\hat{x}}(x;n) = P^*(x_i < x). \qquad (4.3.4)$$

Поскольку (4.3.4) является оценкой вероятности случайного события $\hat{x} < x$, то её качество должно исследоваться методами, рассмотренными в § 3.3.

Если объём выборки достаточно велик, то закон распределения оценки функции $F_{\hat{x}}(x)$ может быть аппроксимирован нормальным законом аналогично (3.3.4):

$$F_{\widetilde{F}_{\hat{x}}(x)}(\vartheta) \approx F_{\widetilde{F}_{\hat{x}}(x)}(\vartheta; M_{\widetilde{F}_{\hat{x}}(x)}; \sigma_{\widetilde{F}_{\hat{x}}(x)}) = F_{\widetilde{F}_{\hat{x}}(x)}\left(\vartheta; F_{\hat{x}}(x); \sqrt{\frac{F_{\hat{x}}(x)(1-F_{\hat{x}}(x))}{n}}\right) =$$

$$= \Phi_1\left(\frac{(\vartheta - F_{\hat{x}}(x))\sqrt{n}}{\sqrt{F_{\hat{x}}(x)(1-F_{\hat{x}}(x))}}\right). \qquad (4.3.5)$$

С учётом (4.3.5) по аналогии с формулами (3.3.5) – (3.3.7) получим

$$\forall x: \ \beta_x = \beta_{I,n} = \beta_{\varepsilon,n} = \left(P \mid \widetilde{F}_{\hat{x}}(x) - F_{\hat{x}}(x) \mid \le \varepsilon\right) =$$

$$= 2\Phi_0\left(\frac{\varepsilon}{\sigma_{\widetilde{F}_{\hat{x}}(x)}}\right) = 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{F_{\hat{x}}(x)(1-F_{\hat{x}}(x))}}\right); \qquad (4.3.6)$$

$$\forall x: \ \ \varepsilon_x = \varepsilon_{\beta,n} = t_\beta \sqrt{\frac{F_{\hat{x}}(x)(1-F_{\hat{x}}(x))}{n}}; \qquad (4.3.7)$$

$$\forall x: \ I_x = I_{\beta,n} = [F'_{\hat{x}}(x); \ F''_{\hat{x}}(x)] = [\widetilde{F}_{\hat{x}}(x) - \varepsilon_x; \ \widetilde{F}_{\hat{x}}(x) + \varepsilon_x]. \qquad (4.3.8)$$

При исследовании качества оценки $\widetilde{F}_{\hat{x}}(x)$ возникает проблема, связанная с незнанием истинной функции распределения $F_{\hat{x}}(x)$. При большом объёме выборки эта функция заменяется её оценкой $\widetilde{F}_{\hat{x}}(x)$ и формулы (4.3.6) – (4.3.8) приобретают вид:

$$\beta_x = \beta_{I,n} = \beta_{\varepsilon,n} \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{\widetilde{F}_{\hat{x}}(x)(1-\widetilde{F}_{\hat{x}}(x))}}\right); \qquad (4.3.9)$$

$$\varepsilon_x = \varepsilon_{\beta,n} \approx t_\beta \sqrt{\frac{\widetilde{F}_{\hat{x}}(x)(1-\widetilde{F}_{\hat{x}}(x))}{n}}; \qquad (4.3.10)$$

$$I = I_{\beta,n} \approx [\widetilde{F}_{\hat{x}}(x) - \varepsilon_x; \ \widetilde{F}_{\hat{x}}(x) + \varepsilon_x]. \qquad (4.3.11)$$

Из соотношений (4.3.6) – (4.3.11) видно, что все показатели качества оценивания функции распределения $F_{\hat{x}}(x)$ зависят от её аргумента x.

Так, при фиксированных доверительной вероятности β = const и объёме выборки n = const доверительные границы для $F_{\hat{x}}(x)$ будут функциями:

$$F'_{\hat{x}}(x) = f_1(x); \quad F''_{\hat{x}}(x) = f_2(x).$$

П р и м е р 4.4. Пусть признак $\hat{x}$ массива экспериментальных данных распределён нормально, т.е.

$$F_{\hat{x}}(x) = F^{\text{н}}_{\hat{x}}(x;\ M_{\hat{x}};\ \sigma_{\hat{x}}) = F^{\text{н}}_{\hat{x}}(x; 0; 1) = \Phi_1(x).$$

По результатам ста наблюдений (n = 100) построена статистическая функция распределения

$$\widetilde{F}_{\hat{x}}(x) = F^{*}_{\hat{x}}(x; 100) = \widetilde{\Phi}_1(x).$$

Требуется построить для функции распределения 95-процентный доверительный интервал (β = 0,95).

▼ По формулам (4.3.7) и (4.3.8) получим

$$\varepsilon_{0,95;100} = t_\beta \sqrt{\frac{\Phi_1(x)(1-\Phi_1(x))}{n}} = 0{,}196\sqrt{\Phi_1(x)(1-\Phi_1(x))};$$

$$I_{0,95;100} = [\widetilde{\Phi}_1(x) - 0{,}196\sqrt{\Phi_1(x)(1-\Phi_1(x))};\ \widetilde{\Phi}_1(x) + 0{,}196\sqrt{\Phi_1(x)(1-\Phi_1(x))}].$$

На рис.4.6 изображена доверительная область для функции $\Phi_1(x)$, границы которой даны пунктиром. График функции $\Phi_1(x)$ – сплошная линия. ▲

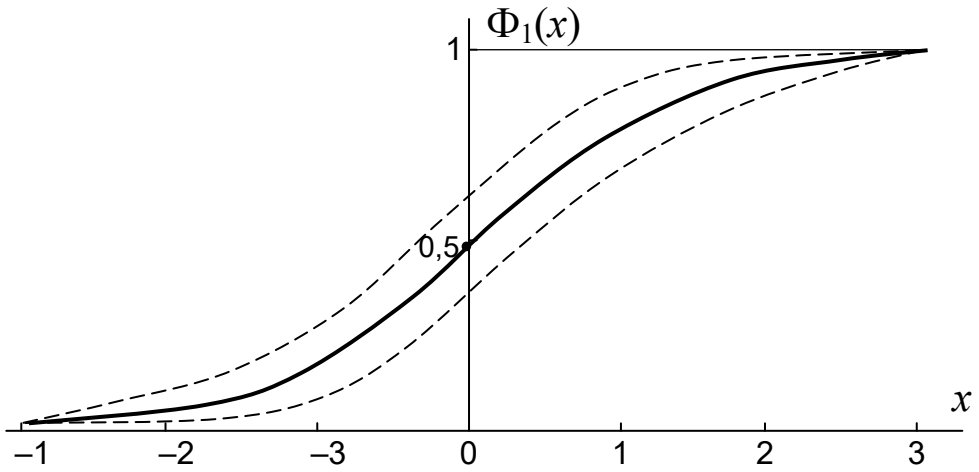


Рис.4.6. Доверительная область функции распределения (к примеру 4.4)

4.3.5. Потребный объём экспериментальных данных

Поскольку $\widetilde{F}_{\hat{x}}(x)$ является оценкой вероятности случайного события $\hat{x} < x$, то объём n выборки, потребный для оценивания функции распределения $F_{\hat{x}}(x)$ с необходимыми точностью ε и надёжностью β, определяется выражением, аналогичным (3.3.8):

$$n = n_{\beta,\varepsilon} \geq \frac{F_{\hat{x}}(x)(1-F_{\hat{x}}(x))}{\varepsilon^2} t_\beta^2 \approx \frac{\widetilde{F}_{\hat{x}}(x)(1-\widetilde{F}_{\hat{x}}(x))}{\varepsilon^2} t_\beta^2 . \qquad (4.3.12)$$

Как видно из рис.4.6, доверительная область для $F_{\hat{x}}(x)$ зависит от x и имеет наибольшую ширину при $x = 0$. Это означает, что наибольшее число наблюдений потребуется для оценивания значения $F_{\hat{x}}(0)$ (см. табл.3.1). Согласно формуле (4.3.12) и табл.3.1. при $x \to \infty$ потребное число n экспериментальных точек снижается, однако при этом уменьшается правомерность предположения о нормальном распределении оценки $\widetilde{F}_{\hat{x}}(x)$. Поэтому при оценивании функции распределения объём выборки берётся максимальным, обеспечивающим требуемые точность и надёжность оценки $\widetilde{F}_{\hat{x}}(x)$ при всех $x \in \{x\}$. Указанный объём определяется соотношением

$$n = n_{\beta,\varepsilon} \geq \frac{t_\beta^2}{4\varepsilon^2} . \qquad (4.3.13)$$

5. ОЦЕНИВАНИЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК И ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ОБЪЕКТОВ

Закон распределения в любой его форме является исчерпывающей характеристикой вероятностного поведения случайного объекта (величины, вектора, функции). Поскольку задача его определения достаточно сложна, на практике часто определяются лишь числовые характеристики, основными из которых являются математическое ожидание и дисперсия. Для случайных векторов характеристикой рассеяния является корреляционная матрица.

Следует отметить, что решение такой более упрощённой задачи имеет большую практическую ценность, так как во многих случаях знать закон распределения не требуется. Кроме того, часто на основе каких-то априорных предположений вид закона распределения исследуемого случайного объекта известен и для его полного описания достаточно определить лишь параметры.

Как уже отмечалось, задача оценивания параметров распределения, в частности числовых характеристик, сводится к отысканию таких статистик (функций случайной выборки), которые могут служить наилучшими в каком-то смысле оценками истинных значений параметров. Все рассматриваемые оценки должны проверяться на наличие у них основных необходимых свойств: несмещённости, состоятельности и эффективности.

5.1. Оценивание математического ожидания случайной величины

Пусть имеется случайная величина $\hat{x}$, математическое ожидание которой $M_{\hat{x}} = \bar{x}$ неизвестно. Над случайной величиной проведено n независимых опытов (наблюдений). По их результатам $x_1,\ x_2,\ldots, x_n$ требуется найти состоятельную, несмещённую и эффективную оценку $\widetilde{M}_{\hat{x}}$ параметра $M_{\hat{x}}$, т.е. найти функцию

$$\widetilde{M}_{\hat{x}} = M_{\hat{x}}(x_1,\ x_2,...,x_n) = \widetilde{M}_{\hat{x}}(X_{<n>}).$$

В качестве оценки математического ожидания (среднего значения) случайной величины $\hat{x}$ могут приниматься различные характеристики случайной выборки. Все общие методы статистического оценивания дают в качестве наилучшей точечной оценки математического ожидания случайной величины её статистическое среднее. Однако при нахождении оценки математического ожидания, удовлетворяющей требованиям со-

стоятельности, несмещённости и эффективности, следует различать равноточные и неравноточные наблюдения (однородные и неоднородные опыты).

5.1.1. Равноточные наблюдения

Статистическое (выборочное) среднее или ***статистическое математическое ожидание*** случайной величины находится по формуле

$$\widetilde{M}_{\hat{x}} = \widetilde{M}_{\hat{x}}(X_{<n>}) = M^*_{\hat{x}} = \bar{x}^* = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i \,. \qquad (5.1.1)$$

Поскольку наблюдения равноточны, то случайные величины $\hat{x}_1,\ \hat{x}_2,...,\hat{x}_n$ по существу представляют собой «экземпляры» одной и той же случайной величины $\hat{x}$ и, следовательно, имеют один и тот же закон распределения с числовыми характеристиками:

$$M_{\hat{x}_i} = M_{\hat{x}} = \bar{x}\,; \quad D_{\hat{x}_i} = D_{\hat{x}}; \quad \sigma_{\hat{x}_i} = \sigma_{\hat{x}}\,, \quad i = \overline{1,n}\,.$$

Покажем, что оценка $M^*_{\hat{x}}$ удовлетворяет всем трём общим требованиям.

1. Из выражения (5.1.1) следует, что

$$M[M^*_{\hat{x}}] = M\left[\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\right] = \frac{1}{n}\sum_{i=1}^{n}M_{\hat{x}_i} = \frac{nM_{\hat{x}}}{n} = M_{\hat{x}}\,.$$

Таким образом, $M^*_{\hat{x}}$ является несмещённой оценкой параметра $M_{\hat{x}}$.

2. Согласно теореме Чебышева среднее арифметическое наблюдаемых значений случайной величины сходится по вероятности к её математическому ожиданию

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - M_{\hat{x}_i}\right| < \varepsilon\right) = 1.$$

Следовательно, статистическое среднее $M^*_{\hat{x}}$ есть состоятельная оценка параметра $M_{\hat{x}}$.

3. Согласно выражению (5.1.1) дисперсия статистического среднего

$$D[M^*_{\hat{x}}] = D\left[\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}D_{\hat{x}_i} = \frac{nD_{\hat{x}}}{n^2} = \frac{D_{\hat{x}}}{n} \qquad (5.1.2)$$

с ростом объёма *n* выборки неограниченно убывает и, следовательно, $M^*_{\hat{x}}$ асимптотически эффективная оценка $M_{\hat{x}}$. Доказано, что если случайная величина $\hat{x}$ подчинена нормальному закону распределения, то при любых *n* дисперсия (5.1.2) будет минимально возможной. В таком случае $M^*_{\hat{x}}$ является эффективной оценкой математического ожидания $M_{\hat{x}}$.

Следовательно, $M^*_{\hat{x}}$ – это подходящее значение $M_{\hat{x}}$:

$$M_{\hat{x}} \approx \widetilde{M}_{\hat{x}} = M^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \,. \qquad (5.1.3)$$

П р и м е р 5.1. В условиях примера 4.1 найти оценку математического ожидания случайной величины $\tilde{\tau}$.

▼ Согласно формуле (5.1.3)

$$\widetilde{M}_{\hat{\tau}} = M^*_{\hat{\tau}} = \frac{1}{12}\sum_{i=1}^{12} \tau_i = 127{,}6 \text{ ч.}$$

Случайная выборка, приведённая в табл.4.3, была получена при наблюдении случайной величины $\tilde{\tau}$, математическое ожидание которой $M_{\hat{\tau}} = \bar{\tau} = 100$ ч. Сравнительно невысокая точность полученной оценки обусловлена малым объёмом выборки, но ни в коей мере не способом её вычисления. ▲

Если объём *n* выборки достаточно велик, то вычисления по формуле (5.1.3) оказываются громоздкими. Задачу можно упростить, если использовать данные интервального вариационного ряда, т.е. полагать

$$\widetilde{\widetilde{M}}_{\hat{x}} = \sum_{l=1}^{r} \bar{x}_l P_l^* \,, \qquad (5.1.4)$$

где $\bar{x}_l = 0{,}5(x_l + x_{l+1})$ – представитель (середина) *l*-го разряда вариационного ряда; P_l^* – частота попадания вариантов x_i случайной величины $\hat{x}$ в *l*-й разряд.

Значение оценки $\widetilde{\widetilde{M}}_{\hat{x}}$, определяемое по формуле (5.1.4), оказывается приближённым, однако с ростом *n* (а, следовательно, и *r*) точность данной формулы возрастает.

П р и м е р 5.2. В условиях примера 4.2 найти приближённое значение оценки $\widetilde{\widetilde{M}}_{\hat{\tau}}$ математического ожидания $M_{\hat{\tau}}$ случайной величины $\hat{\tau}$.

▼ Используем данные табл.4.7. По формуле (5.1.4) получим

$$\widetilde{M}_{\hat{\tau}} \approx \widetilde{\widetilde{M}}_{\hat{\tau}} = \sum_{l=1}^{100} \bar{\tau}_l P_l^* = 102 \text{ ч.},$$

▲

Следует обратить внимание на то, что выборка, приведённая в табл.4.7, принадлежит той же генеральной совокупности, что и в табл.4.3. Однако, как видно из решений примеров 5.1 и 5.2, в последнем случае даже приближённое значение оценки $\widetilde{M}_{\hat{\tau}}$ меньше отличается от истинного значения параметра $M_{\hat{\tau}} = 100$ ч. Это объясняется бóльшим объёмом *n* и, следовательно, большей информативностью выборки, приведённой в табл.4.7.

5.1.2. Неравноточные наблюдения

Пусть характеристики точности наблюдений от опыта к опыту изменяются так, что наблюдаемая в *i*-м опыте случайная величина $\hat{x}_i$ имеет дисперсию

$$D_{\hat{x}_i} = \sigma_i^2, \quad i = \overline{1,n}.$$

При этом среднее значение случайной величины $\hat{x}$ от опыта к опыту не изменяется:

$$M_{\hat{x}_i} = \bar{x} = M_{\hat{x}}, \quad i = \overline{1,n}.$$

В данном случае оценка $\widetilde{M}_{\hat{x}}$ математического ожидания случайной величины $\hat{x}$ по-прежнему будет являться функцией случайной выборки:

$$\widetilde{M}_{\hat{x}} = \widetilde{M}_{\hat{x}}(X_{<n>}).$$

Необходимо так выбрать вид этой зависимости, чтобы оценка имела простое аналитическое выражение и была несмещённой, состоятельной и эффективной.

Так как наиболее простой функциональной зависимостью является линейная, то будем искать оценку $\widetilde{M}_{\hat{x}}$ в классе линейных функций:

$$\widetilde{M}_{\hat{x}} = \sum_{i=1}^{n} c_i x_i . \qquad (5.1.5)$$

Очевидно, что теперь решение поставленной задачи состоит в отыскании значений коэффициентов c_i, $i = \overline{1,n}$ линейной формы (5.1.5), при которых оценка $\widetilde{M}_{\hat{x}}$ будет удовлетворять всем трём указанным выше требованиям.

1. Чтобы оценка была несмещённой, должно выполняться равенство

$$M[\widetilde{M}_{\hat{x}}] = M_{\hat{x}}.$$

Поскольку в этом случае

$$M[\widetilde{M}_{\hat{x}}] = M\left[\sum_{i=1}^{n} c_i x_i\right] = \sum_{i=1}^{n} c_i M_{\hat{x}_i} = M_{\hat{x}} \sum_{i=1}^{n} c_i = M_{\hat{x}},$$

то коэффициенты c_i должны удовлетворять условию

$$\sum_{i=1}^{n} c_i = 1.$$

2. Для того чтобы оценка $\widetilde{M}_{\hat{x}}$ была эффективной, её дисперсия

$$D[\widetilde{M}_{\hat{x}}] = D\left[\sum_{i=1}^{n} c_i \hat{x}_i\right] = \sum_{i=1}^{n} c_i^2 D_{\hat{x}_i} = \sum_{i=1}^{n} c_i^2 D_i \qquad (5.1.6)$$

должна быть минимальной при условии, что

$$1-\sum_{i=1}^{n} c_i = 0. \qquad (5.1.7)$$

Условный экстремум (минимум) функции (5.1.6) с переменными $c_1,\ c_2,\ldots,\ c_n$ отыскиваем методом неопределённых множителей Лагранжа. При этом учитываем, что должно выполняться равенство (5.1.7). Следовательно, исследуем на минимум вспомогательную функцию

$$D[\widetilde{M}_{\hat{x}}] = \sum_{i=1}^{n} c_i^2 D_i + 2\lambda\left(1-\sum_{i=1}^{n} c_i\right) = Q(c_1,\ c_2,...,c_n),$$

где λ – неопределённый множитель Лагранжа.

Решаем систему n уравнений

$$\frac{\partial Q}{\partial c_i} = 2c_i D_i - 2\lambda = 0, \quad i = \overline{1,n}$$

относительно переменных $c_1,\ c_2,\ldots,\ c_n$ и получаем

$$c_i = \frac{\lambda}{D_i}.$$

Таким образом, вес c_i, с которым должен входить результат i-го наблюдения в формулу для оценки $\widetilde{M}_{\hat{x}}$, должен быть обратно пропорционален его дисперсии. Иными словами, чем точнее наблюдение, тем с бóльшим весом необходимо учитывать его результат. Вывод, полученный формально, полностью согласуется с вербальными рассуждениями: чем точнее наблюдение, тем больше ему следует доверять.

Поскольку $\sum_{i=1}^{n} c_i = 1$, то $\lambda\sum_{i=1}^{n}\frac{1}{D_i} = 1$ и, следовательно,

$$\lambda = \frac{1}{\sum_{i=1}^{n}\frac{1}{D_i}}. \qquad (5.1.8)$$

Обозначим $1/D_i = d_i$, тогда (5.1.8) представляется как $\lambda = 1\Big/\sum_{i=1}^{n} d_i$ и

$$c_i = \frac{d_i}{\sum_{i=1}^{n} d_i}, \quad i = \overline{1,n}. \qquad (5.1.9)$$

Таким образом, выражение для оценки (5.1.5) будет иметь вид

$$\widetilde{M}_{\hat{x}} = \frac{\sum_{i=1}^{n} d_i x_i}{\sum_{i=1}^{n} d_i}. \qquad (5.1.10)$$

Оценка вида (5.1.10) является эффективной, так как получена на основе требования минимума дисперсии.

3. Минимальная дисперсия несмещённой оценки $\widetilde{M}_{\hat{x}}$

$$D[\widetilde{M}_{\hat{x}}] = \sum_{i=1}^{n} c_i^2 D_i = \frac{1}{\left(\sum_{i=1}^{n} d_i\right)^2} \sum_{i=1}^{n} d_i^2 \frac{1}{d_i} = \frac{1}{\sum_{i=1}^{n} d_i} = \lambda, \qquad (5.1.11)$$

а её среднее квадратическое отклонение

$$\sigma[\widetilde{M}_{\hat{x}}] = \sqrt{\lambda} = \frac{1}{\sqrt{\sum_{i=1}^{n} d_i}}.$$

Поскольку $d_i = 1/D_i = \text{const}$, $i = \overline{1,n}$, то из выражения (5.1.11) вытекает, что при неограниченном возрастании количества наблюдений $\lambda \to 0$. Следовательно, $\widetilde{M}_{\hat{x}}$ сходится по вероятности к $M_{\hat{x}}$, т.е. является состоятельной оценкой математического ожидания $M_{\hat{x}}$.

Частный случай. Предположим, что все наблюдения равноточны. Это означает, что $D_i = D_{\hat{x}}$, $d_i = 1/D_{\hat{x}} = d$, $c_i = 1/n$, $i = \overline{1,n}$ и тогда

$$\widetilde{M}_{\hat{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i = M_{\hat{x}}^*$$

Получим результат как и в пп.5.1.1 – оценкой математического ожидания случайной величины $\hat{x}$ является её статистическое среднее $M_{\hat{x}}^*$.

П р и м е р 5.3. Дальность $\hat{x}$ до центра масс ракеты измеряется тремя методами, точность которых характеризуется средними квадратическими отклонениями $\sigma_{\hat{x}_1} = 0{,}2$ км, $\sigma_{\hat{x}_2} = 0{,}5$ км, $\sigma_{\hat{x}_3} = 1$ км. Измерения дальности $\hat{x}$ этими методами дали следующие результаты: $x_1 = 10{,}0$ км; $x_2 = 9{,}5$ км; $x_3 = 10{,}8$ км.

Найти оценку $\widetilde{M}_{\hat{x}}$ математического ожидания $M_{\hat{x}}$ дальности $\hat{x}$ и среднее квадратическое отклонение $\sigma[\widetilde{M}_{\hat{x}}]$ этой оценки.

▼ По условию задачи

$$d_1 = \frac{1}{\sigma_{\hat{x}_1}^2} = \frac{1}{0{,}04} = 25, \quad d_2 = \frac{1}{\sigma_{\hat{x}_2}^2} = \frac{1}{0{,}25} = 4, \quad d_3 = \frac{1}{\sigma_{\hat{x}_3}^2} = \frac{1}{1} = 1.$$

Далее согласно равенствам (5.1.9)

$$c_1 = \frac{25}{30}, \quad c_2 = \frac{4}{30}, \quad c_3 = \frac{1}{30}.$$

По формуле (5.1.5) получаем

$$\widetilde{M}_{\hat{x}} = \frac{1}{30}(25 \cdot 10 + 4 \cdot 9{,}5 + 1 \cdot 10{,}8) = 9{,}9 \text{ км}.$$

В соответствии с выражением (5.1.11)

$$D[\widetilde{M}_{\hat{x}}] = \frac{1}{\sum\limits_{i=1}^{3} d_i} = \frac{1}{30} \text{ км}^2, \quad \sigma[\widetilde{M}_{\hat{x}}] = \sqrt{\frac{1}{30}} = 0{,}183 \text{ км}.$$

▲

5.1.3. Качество оценивания математического ожидания

Качество статистического оценивания математического ожидания при заданном объёме n выборки определяется доверительным интервалом

$$I_{\beta,n} = [M'_{\hat{x}};\ M''_{\hat{x}}] \tag{5.1.12}$$

и доверительной вероятностью

$$\beta_{I,n} = \beta_{\varepsilon,n} = P[M'_{\hat{x}} \le M_{\hat{x}} \le M''_{\hat{x}}]. \tag{5.1.13}$$

Как указывалось в § 3.2, процедура построения доверительного интервала зависит, с одной стороны, от характера распределения наблюдаемого признака $\hat{x}$ и, как следствие, от распределения оценки $\widetilde{M}_{\hat{x}}$, а с другой – от объёма n случайной выборки $\hat{X}_{<n>}$. Кроме того, и в первую очередь, она зависит от типа статистики $s(X_{<n>})$, используемой в качестве оценки $\widetilde{M}_{\hat{x}}$ математического ожидания. В п.п. 5.1.1 и 5.1.2 применялись линейные статистики в виде средневзвешенного элементов выборки.

В случае равноточных независимых наблюдений наилучшей (по трём критериям, см. § 3.1) оценкой математического ожидания является статистическое математическое ожидание (5.1.1). При этом, если наблюдаемая случайная величина $\hat{x}$ подчинена нормальному закону распределения, то при любом n оценка $\widetilde{M}_{\hat{x}}$ будет иметь нормальное распределение с параметрами

$$\left.\begin{aligned} M[\widetilde{M}_{\hat{x}}] &= M_{\hat{x}}; \\ D[\widetilde{M}_{\hat{x}}] &= \frac{D_{\hat{x}}}{n}; \\ \sigma[\widetilde{M}_{\hat{x}}] &= \frac{\sigma_{\hat{x}}}{\sqrt{n}} \end{aligned}\right\}. \tag{5.1.14}$$

Наряду с этим оценка $\widetilde{M}_{\hat{x}} = M^*_{\hat{x}}$ асимптотически нормальна, т.е. при $n \to \infty$ для любого распределения признака $\hat{x}$ распределение оценки $\widetilde{M}_{\hat{x}}$ приближается к нормальному с параметрами, определяемыми равенствами (5.1.14). Данное утверждение вытекает из предельной теоремы Ляпунова.

Тогда доверительная вероятность (5.1.13) будет определяться отношением

$$\beta = \beta_{I,n} = \beta_{\varepsilon,n} = P(|\widetilde{M}_{\hat{x}} - M_{\hat{x}}| \le \varepsilon) \approx 2\Phi_0\left(\frac{\varepsilon}{\sigma[M^*_{\hat{x}}]}\right) = \\ = 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sigma_{\hat{x}}}\right) \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\widetilde{\sigma}_{\hat{x}}}\right). \tag{5.1.15}$$

Первое приближённое равенство в (5.1.15) обусловлено отличием закона распределения признака $\hat{x}$ от нормального, а второе – заменой неизвестного $\sigma_{\hat{x}}$ его оценкой $\widetilde{\sigma}_{\hat{x}}$. При нормальном распределении $\hat{x}$ и известном $\sigma_{\hat{x}}$ соотношение (5.1.15) становится точным.

Разрешив (5.1.15) относительно ε, получим

$$\varepsilon = \varepsilon_{\beta,n} = \frac{\sigma_{\hat{x}}}{\sqrt{n}}\Phi_0^{-1}\left(\frac{\beta}{2}\right) = \frac{\sigma_{\hat{x}}}{\sqrt{n}}t_\beta \approx \frac{\widetilde{\sigma}_{\hat{x}}}{\sqrt{n}}t_\beta, \tag{5.1.16}$$

откуда находим интервал (5.1.12):

$$I_{\beta,n} = [M'_{\hat{x}};\ M''_{\hat{x}}] = [\widetilde{M}_{\hat{x}} - \varepsilon_{\beta,n};\ \widetilde{M}_{\hat{x}} + \varepsilon_{\beta,n}] = \\ = \left[\widetilde{M}_{\hat{x}} - \frac{\sigma_{\hat{x}}}{\sqrt{n}}t_\beta;\ \widetilde{M}_{\hat{x}} + \frac{\sigma_{\hat{x}}}{\sqrt{n}}t_\beta\right] \approx \left[\widetilde{M}_{\hat{x}} - \frac{\widetilde{\sigma}_{\hat{x}}}{\sqrt{n}}t_\beta;\ \widetilde{M}_{\hat{x}} + \frac{\widetilde{\sigma}_{\hat{x}}}{\sqrt{n}}t_\beta\right]. \tag{5.1.17}$$

Из соотношения (5.1.17) видно, что при большом объёме выборки доверительный интервал для $M_{\hat{x}}$ симметричен и полностью определяется его оценкой и максимальной с вероятностью β абсолютной погрешностью $\varepsilon_{\beta,n}$. На рис.5.1 дана геометрическая интерпретация соотношения (5.1.17).

Из уравнения (5.1.16) выражаем *n*, при этом будем иметь

$$n = n_{\beta,I} = n_{\beta,\varepsilon} \ge \left(\frac{\sigma_{\hat{x}}}{\varepsilon}t_\beta\right)^2 \approx \left(\frac{\widetilde{\sigma}_{\hat{x}}}{\varepsilon}t_\beta\right)^2. \tag{5.1.18}$$

Формулой (5.1.18) определяется потребный объём выборки для оценивания математического ожидания случайной величины $\hat{x}$.

5.2. Оценивание дисперсии и среднего квадратического отклонения случайной величины

Над случайной величиной $\hat{x}$ производится *n* независимых равноточных наблюдений. Требуется по результатам эксперимента определить состоятельные и несмещённые оценки $\widetilde{D}_{\hat{x}}$ и $\widetilde{\sigma}_{\hat{x}}$ характеристик рассеяния $D_{\hat{x}}$ и $\sigma_{\hat{x}}$ случайной величины $\hat{x}$, т.е. найти

$$\widetilde{D}_{\hat{x}} = \widetilde{D}_{\hat{x}}(X_{<n>}) \ \text{ и } \ \widetilde{\sigma}_{\hat{x}} = \widetilde{\sigma}_{\hat{x}}(X_{<n>}).$$

Ограничимся рассмотрением наиболее важного для практики случая, когда случайная величина $\hat{x}$ подчинена нормальному закону распределения с параметрами $M_{\hat{x}}$ и $\sigma_{\hat{x}}$.

При решении поставленной задачи следует различать два случая – когда параметр $M_{\hat{x}}$ известен и когда он неизвестен.

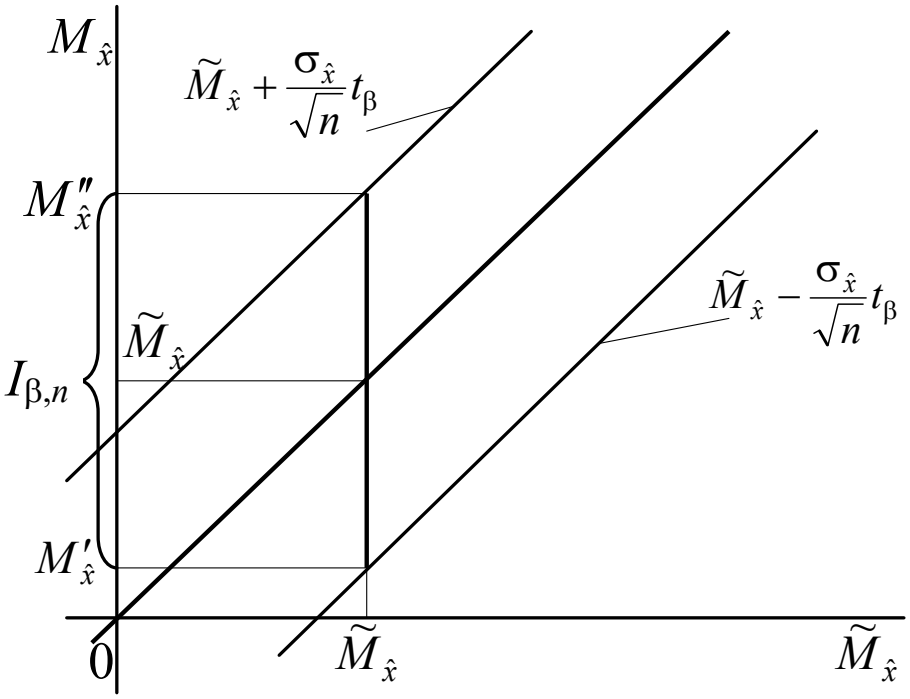


Рис.5.1. Доверительный интервал для математического ожидания

5.2.1. Оценивание дисперсии и среднего квадратического отклонения при известном математическом ожидании

Вводим случайную величину

$$\widetilde{D}_{\hat{x}} = D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}})^2 , \qquad (5.2.1)$$

которая называется ***дисперсией случайной выборки*** или ***статистической, выборочной дисперсией***. Установим некоторые из свойств случайной величины (5.2.1).

1. Преобразуем $D^*_{\hat{x}}$ к виду

$$D^*_{\hat{x}} = \frac{1}{n}\sigma^2_{\hat{x}}\sum_{i=1}^{n}\frac{(\hat{x}_i - M_{\hat{x}})^2}{\sigma^2_{\hat{x}}} = \frac{\sigma^2_{\hat{x}}}{n}\hat{\chi}^2_n,$$

т.е. $D^*_{\hat{x}}$ является линейной функцией от случайной величины $\hat{\chi}^2_n$, подчинённой хи-квадрат распределению (распределению К. Пирсона) с n степенями свободы. Следовательно

$$M[D_{\hat{x}}^*] = \frac{\sigma_{\hat{x}}^2}{n} M[\hat{\chi}_n^2] = \frac{\sigma_{\hat{x}}^2}{n} n = \sigma_{\hat{x}}^2 = D_{\hat{x}}. \qquad (5.2.2)$$

Таким образом, $D_{\hat{x}}^*$ – несмещённая оценка $D_{\hat{x}}$.

2. Поскольку

$$D[D_{\hat{x}}^*] = \frac{\sigma_{\hat{x}}^4}{n^2} D[\hat{\chi}_n^2] = \frac{\sigma_{\hat{x}}^4}{n^2} 2n = \frac{2}{n} D_{\hat{x}}^2; \quad \sigma[D_{\hat{x}}^*] = \sqrt{\frac{2}{n}} \sigma_{\hat{x}}^2, \qquad (5.2.3)$$

то при $n \to \infty$ имеет место $D[D_{\hat{x}}^*] \to 0$. Иначе, дисперсия $D_{\hat{x}}^*$ случайной выборки асимптотически эффективная оценка $D_{\hat{x}}$.

3. Как следует из (5.2.2) и (5.2.3), случайная величина $D_{\hat{x}}^*$ имеет числовые характеристики

$$M[D_{\hat{x}}^*] = D_{\hat{x}}; \quad D[D_{\hat{x}}^*] = \frac{2}{n} D_{\hat{x}}^2; \quad \sigma[D_{\hat{x}}^*] = \sqrt{\frac{2}{n}} \sigma_{\hat{x}}^2.$$

Поскольку

$$\lim_{n \to \infty} D[D_{\hat{x}}^*] = 0,$$

то оценка $D_{\hat{x}}^*$ является состоятельной.

Итак, при $n \to \infty$ дисперсия случайной выборки (5.2.1) представляет собой подходящее значение дисперсии $D_{\hat{x}}$ случайной величины $\hat{x}$. При малых n она в общем случае не вполне эффективна.

П р и м е р 5.4. Полагая $M_{\hat{\tau}}$ = 100 ч, в условиях примера 4.1 найти оценку $\widetilde{D}_{\hat{\tau}}$ дисперсии $D_{\hat{\tau}}$ случайной величины $\hat{\tau}$.

▼ Используем данные табл.4.3 и по формуле (5.2.1) получаем

$$\widetilde{D}_{\hat{\tau}} = D_{\hat{\tau}}^* = \frac{1}{12} \sum_{i=1}^{12} (\tau_i - 100)^2 = 15115 \text{ ч}^2.$$

▲

Если объём n выборки достаточно велик, то для вычисления оценки $\widetilde{D}_{\hat{x}}$ можно пользоваться приближённой формулой

$$\widetilde{D}_{\hat{x}} \approx \widetilde{\widetilde{D}}_{\hat{x}} = \sum_{l=1}^{r} (\bar{x}_l - M_{\hat{x}})^2 P_l^*, \qquad (5.2.4)$$

где $\bar{x}_l$ и P_l^* имеют тот же смысл, что и в формуле (5.1.4).

П р и м е р 5.5. Полагая $M_{\hat{\tau}}$ = 100 ч, в условиях примера 4.2 найти приближённое значение оценки $\widetilde{D}_{\hat{\tau}}$ дисперсии случайной величины $\hat{\tau}$.

▼ Используя табл.4.7, по формуле (5.2.4) получаем

$$\widetilde{D}_{\hat{\tau}} \approx \widetilde{\widetilde{D}}_{\hat{\tau}} = \sum_{l=1}^{10} (\bar{\tau}_i - 100)^2 P_l^* = 7975 \text{ ч}^2.$$

▲

Теперь найдем оценку среднего квадратического отклонения $\sigma_{\hat{x}}$. Формула для определения статистического среднего квадратического отклонения имеет вид

$$\sigma^*_{\hat{x}} = \sqrt{D^*_{\hat{x}}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}})^2}\,.$$

Требуется выявить основные свойства $\sigma^*_{\hat{x}}$.

1. Из вышеизложенного следует, что $\sigma^*_{\hat{x}}$ является состоятельной и асимптотически эффективной оценкой $\sigma_{\hat{x}}$.

2. Поскольку

$$\sigma^*_{\hat{x}} = \sqrt{\frac{1}{n}\sigma^2_{\hat{x}}\hat{\chi}^2_n} = \frac{\sigma_{\hat{x}}}{\sqrt{n}}\hat{\chi}_n,$$

то

$$M[\sigma^*_{\hat{x}}] = \frac{\sigma_{\hat{x}}}{\sqrt{n}}M[\hat{\chi}_n] = \sigma_{\hat{x}}\sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \neq \sigma_{\hat{x}},$$

где $\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt$ - гамма-функция (интеграл Эйлера 2 рода).

Полученное соотношение указывает на смещённость оценки среднего квадратического отклонения.

Если величину $\sigma^*_{\hat{x}}$ исправить, умножив её на коэффициент

$$k_n = \sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)},$$

то полученная в результате функция случайной выборки

$$\tilde{\sigma}_{\hat{x}} = k_n\sigma^*_{\hat{x}} = k_n\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_{\hat{x}})^2} \qquad (5.2.5)$$

будет состоятельной, несмещённой и асимптотически эффективной оценкой среднего квадратического отклонения $\sigma_{\hat{x}}$ случайной величины $\hat{x}$. В табл.5.1 приведены значения коэффициента k_n для некоторых n. Эти значения используются при вычислении оценки (5.2.5).

Таблица 5.1

Значения коэффициента k_n

n	2	3	4	5	6	9	12	18	24
k_n	1,128	1,085	1,064	1,051	1,042	1,028	1,021	1,014	1,010

Из таблицы видно, что необходимость в исправлении оценки возникает лишь при малых объёмах выборки, так как с их увеличением коэффициент k_n достаточно быстро приближается к единице.

5.2.2. Оценивание дисперсии и среднего квадратического отклонения при неизвестном математическом ожидании

Для отыскания оценки дисперсии случайной величины необходимо знать её математическое ожидание $M_{\hat{x}}$. В случае, если данный параметр неизвестен, используют его оценку $\widetilde{M}_{\hat{x}} = M^*_{\hat{x}}$.

Рассмотрим функцию случайной выборки в виде статистической дисперсии

$$D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - M^*_{\hat{x}})^2, \qquad (5.2.6)$$

и исследуем её свойства.

1. По аналогии со случаем, когда математическое ожидание известно, можно показать состоятельность оценки (5.2.6).

2. Преобразуем выражение (5.2.6)

$$D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}\Big((\hat{x}_i - M_{\hat{x}}) - (M^*_{\hat{x}} - M_{\hat{x}})\Big)^2 =$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}})^2 - 2(M^*_{\hat{x}} - M_{\hat{x}})\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}}) + \sum_{i=1}^{n}(M^*_{\hat{x}} - M_{\hat{x}})^2\right) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - M_{\hat{x}})^2 - (M^*_{\hat{x}} - M_{\hat{x}})^2 = \frac{\sigma^2_{\hat{x}}}{n}\left(\sum_{i=1}^{n}\left(\frac{\hat{x}_i - M_{\hat{x}}}{\sigma_{\hat{x}}}\right)^2 - \left(\frac{M^*_{\hat{x}} - M_{\hat{x}}}{\sigma_{\hat{x}}/\sqrt{n}}\right)^2\right) =$$

$$= \frac{\sigma^2_{\hat{x}}}{n}(\hat{\chi}^2_n - \hat{\chi}^2_1).$$

Поскольку случайные величины $\hat{\chi}^2_n$ и $\hat{\chi}^2_1$ являются функциями одной и той же выборки, то они зависимы. Причём их зависимость такова, что разность этих случайных величин оказывается подчинённой закону распределения хи-квадрат с $n-1$ степенями свободы. Таким образом

$$D^*_{\hat{x}} = \frac{\sigma^2_{\hat{x}}}{n}\hat{\chi}^2_{n-1}, \qquad (5.2.7)$$

откуда

$$M[D^*_{\hat{x}}] = \frac{\sigma^2_{\hat{x}}}{n}M[\hat{\chi}^2_{n-1}] = \frac{n-1}{n}\sigma^2_{\hat{x}} \neq D_{\hat{x}}.$$

Следовательно, статистическая дисперсия $D^*_{\hat{x}}$ оказывается смещённой оценкой параметра $D_{\hat{x}}$. Для исправления оценки $D^*_{\hat{x}}$ её достаточно

умножить на коэффициент $n(n–1)$. С ростом объёма n выборки указанный коэффициент стремится к единице, поэтому при достаточно больших n смещённостью оценки $D^*_{\hat{x}}$ можно пренебречь.

3. Поскольку

$$D[D^*_{\hat{x}}] = \frac{\sigma^4_{\hat{x}}}{n^2} D[\hat{\chi}^2_{n-1}] = \frac{2(n-1)}{n^2} D^2_{\hat{x}}, \quad \sigma[D^*_{\hat{x}}] = \frac{\sqrt{2(n-1)}}{n} \sigma^2_{\hat{x}}, \qquad (5.2.8)$$

то при $n \to \infty$ имеет место $D[D^*_{\hat{x}}] \to 0$. Результат, полученный на основе анализа выражения (5.2.8), свидетельствует об асимптотической эффективности оценки $D^*_{\hat{x}}$.

Итак, при $n \to \infty$ исправленная статистическая дисперсия

$$\widetilde{D}_{\hat{x}} = \frac{n}{n-1} D^*_{\hat{x}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widetilde{M}_{\hat{x}})^2 \qquad (5.2.9)$$

является подходящим значением дисперсии $D_{\hat{x}}$ случайной величины $\hat{x}$. С уменьшением объёма n выборки эффективность этой оценки несколько падает.

Оценка (5.2.9) имеет следующие числовые характеристики:

$$M[\widetilde{D}_{\hat{x}}] = D_{\hat{x}}; \quad D[\widetilde{D}_{\hat{x}}] = \frac{2}{n-1} D^2_{\hat{x}}; \quad \sigma[\widetilde{D}_{\hat{x}}] = \sqrt{\frac{2}{n-1}} \sigma^2_{\hat{x}}. \qquad (5.2.10)$$

Вычисление дисперсии $D[\widetilde{D}_{\hat{x}}]$ связано со сложными выкладками, поэтому её выражение приведено без вывода.

При большом объёме n выборки приближённое значение оценки $\widetilde{D}_{\hat{x}}$ можно вычислять по формуле

$$\widetilde{D}_{\hat{x}} = \widetilde{\widetilde{D}}_{\hat{x}} = \sum_{l=1}^{r} (\overline{x}_l - \widetilde{M}_{\hat{x}})^2 P^*_l. \qquad (5.2.11)$$

Перейдём к отысканию оценки для среднего квадратического отклонения $\sigma_{\hat{x}}$ случайной величины $\hat{x}$ в случае неизвестного математического ожидания. Указанная оценка определяется по формуле

$$\widetilde{\sigma}'_{\hat{x}} = \sqrt{\widetilde{D}_{\hat{x}}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widetilde{M}_{\hat{x}})^2}. \qquad (5.2.12)$$

Проанализируем свойства оценки (5.2.12).

1. Из вышеизложенного следует, что данная оценка состоятельна и асимптотически эффективна.

2. Согласно выражениям (5.2.7) и (5.2.9) имеем

$$\widetilde{\sigma}'_{\hat{x}} = \frac{\sigma_{\hat{x}}}{\sqrt{n-1}} \hat{\chi}_{n-1}.$$

Поэтому

$$M[\widetilde{\sigma}'_{\hat{x}}] = \frac{\sigma_{\hat{x}}}{\sqrt{n-1}} M[\hat{\chi}_{n-1}] = \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma_{\hat{x}} \neq \sigma_{\hat{x}}$$

и, следовательно, (5.2.12) является смещённой оценкой $\sigma_{\hat{x}}$. Для исправления данной оценки её достаточно умножить на коэффициент

$$k_{n-1} = \sqrt{\frac{n-1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}.$$

Полученная при этом функция случайной выборки

$$\widetilde{\sigma}_{\hat{x}} = k_{n-1}\widetilde{\sigma}'_{\hat{x}} = k_{n-1}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^2} \qquad (5.2.13)$$

будет состоятельной, несмещённой и асимптотически эффективной оценкой среднего квадратического отклонения $\sigma_{\hat{x}}$ случайной величины $\hat{x}$.

Подходящее значение $\widetilde{\sigma}_{\hat{x}}$ можно получить и непосредственно, используя статистическую дисперсию (5.2.6). Для исправления получаемой при этом оценки $\sigma^*_{\hat{x}} = \sqrt{D^*_{\hat{x}}}$ её необходимо умножить на коэффициент

$$k'_{n-1} = \sqrt{\frac{n}{n-1}} k_{n-1},$$

т.е. величина

$$\widetilde{\sigma}_{\hat{x}} = k'_{n-1}\sigma^*_{\hat{x}} = \sqrt{\frac{n}{n-1}} k_{n-1} \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^2}$$

является состоятельной, несмещённой и асимптотически эффективной оценкой.

В заключение отметим, что поскольку известное соотношение для дисперсии

$$D_{\hat{x}} = \nu_2[\hat{x}] - \nu_1^2[\hat{x}]$$

справедливо и для её оценки, т.е.

$$\widetilde{D}_{\hat{x}} = \widetilde{\nu}_2[\hat{x}] - \widetilde{\nu}_1^2[\hat{x}],$$

то формулам (5.2.1), (5.2.4) и (5.2.9), (5.2.11) соответственно можно придать более удобный для практического использования вид:

$$\left.\begin{aligned} \widetilde{D}_{\hat{x}} &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - M_{\hat{x}}^2 \\ \widetilde{D}_{\hat{x}} &= \widetilde{\widetilde{D}}_{\hat{x}} = \sum_{l=1}^{r} \bar{x}_l^2 P_l^* - M_{\hat{x}}^2 \end{aligned}\right\}$$

$$\left.\begin{aligned} \widetilde{D}_{\hat{x}} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \widetilde{M}_{\hat{x}}^2\right)\frac{n}{n-1} \\ \widetilde{D}_{\hat{x}} &\approx \widetilde{\widetilde{D}}_{\hat{x}} = \sum_{l=1}^{r} \overline{x}_l^2 P_l^* - \widetilde{M}_{\hat{x}}^2 \end{aligned}\right\} \qquad (5.2.14)$$

П р и м е р 5.6. Производятся измерения одного из габаритных размеров $\hat{x}$ однотипных деталей. Данные измерений сведены в табл.5.2.

Т а б л и ц а 5.2

Результаты измерений размера деталей

i	x_i	i	x_i	i	x_i	i	x_i	i	x_i
1	10,5	5	10,4	9	10,3	13	10,5	17	10,8
2	10,8	6	10,6	10	10,8	14	10,7	18	10,7
3	11,2	7	10,9	11	10,6	15	10,8	19	10,9
4	10,9	8	11,0	12	11,3	16	10,9	20	11,0

1. Найти оценку $\widetilde{M}_{\hat{x}}$ математического ожидания величины $\hat{x}$ и построить доверительный интервал, соответствующий доверительной вероятности $\beta = 0{,}8$.

2. Определить доверительную вероятность β для математического ожидания случайной величины $\hat{x}$, если максимальная вероятная погрешность $\varepsilon_\beta = 0{,}07$.

▼ 1. В соответствии с выражением (5.1.1) находим оценку математического ожидания

$$\widetilde{M}_{\hat{x}} = \frac{1}{20}\sum_{i=1}^{20} x_i = 10{,}78.$$

По первой формуле (5.2.14) находим оценку дисперсии

$$\widetilde{D}_{\hat{x}} = \left(\frac{1}{20}\sum_{i=1}^{20} x_i^2 - (10{,}78)^2\right)\frac{20}{19} = 0{,}064.$$

Оценка среднего квадратического отклонения

$$\widetilde{\sigma}_{\hat{x}} = \sqrt{\widetilde{D}_{\hat{x}}} = \sqrt{0{,}064} = 0{,}253.$$

При заданном $\beta = 0{,}8$ величина $t_\beta = 1{,}282$ (см. приложение 4), тогда максимальное вероятное отклонение математического ожидания найдём по формуле (5.1.16), что составит

$$\varepsilon = \varepsilon_{0{,}8;\,20} = \frac{0{,}253}{\sqrt{20}} \cdot 1{,}282 = 0{,}072.$$

Выражение (5.1.17) даёт следующий результат:

$$I_{0{,}8;\,20} = [10{,}78 - 0{,}072;\ \ 10{,}78 + 0{,}072] = [10{,}71;\ 10{,}85].$$

2. В соответствии с соотношением (5.1.15) находим

$$\beta = \beta_{0{,}07;\,20} = 2\Phi_0\left(\frac{0{,}07\sqrt{20}}{0{,}253}\right) = 2\Phi_0(1{,}237) = 2 \cdot 0{,}3912 = 0{,}7824.$$

Значение функции $\Phi_0(x)$ найдены в приложении 2.

▲

П р и м е р 5.7. Габаритный размер $\hat{x}$ деталей измеряется методом, который характеризуется дисперсий $D_{\hat{x}} = 0{,}064$. Определить потребный объём выборки, чтобы максимальная вероятная погрешность ε оценки среднего размера деталей не превосходила 0,06 при доверительной вероятности $\beta = 0{,}93$.

▼ Для $\beta = 0{,}93$ в приложении 4 находим $t_\beta = 1{,}810$. Тогда из выражения (5.1.18) получаем

$$n \geq \left(\frac{\sqrt{0{,}064}}{0{,}06} 1{,}810 \right)^2 = 58 .$$

▲

5.2.3. Качество оценивания дисперсии

Качество оценивания дисперсии $D_{\hat{x}}$ характеризуется доверительным интервалом

$$I_{\beta,n} = [D'_{\hat{x}}; D''_{\hat{x}}]$$

и доверительной вероятностью

$$\beta_{I,n} = P(D'_{\hat{x}} \leq D_{\hat{x}} \leq D''_{\hat{x}}].$$

При оценивании дисперсии различают случаи известного и неизвестного математического ожидания (см. пп. 5.2.1, 5.2.2). Рассмотрим второй, наиболее распространённый случай. При неизвестном $M_{\hat{x}}$ состоятельная, несмещённая и асимптотически эффективная оценка дисперсии определяется равенством (5.2.9).

При достаточно большом объёме выборки распределение оценки (5.2.9) будет близким к нормальному с параметрами (5.2.10). Следовательно, доверительная вероятность для дисперсии будет приближённо определяться соотношением

$$\beta = \beta_{I,n} = \beta_{\varepsilon,n} = P(|\widetilde{D}_{\hat{x}} - D_{\hat{x}}| \leq \varepsilon) \approx 2\Phi_0\left(\frac{\varepsilon}{\sigma[D_{\hat{x}}]} \right) = 2\Phi_0\left(\frac{\varepsilon\sqrt{n-1}}{D_{\hat{x}}\sqrt{2}} \right), \quad (5.2.15)$$

где $\varepsilon = \varepsilon_{\beta,n}$ – максимальная с вероятностью β абсолютная погрешность оценки $\widetilde{D}_{\hat{x}}$ дисперсии $D_{\hat{x}}$.

Выражаем ε из уравнения (5.2.15) и в результате получим

$$\varepsilon = \varepsilon_{\beta,n} \approx D_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta , \quad (5.2.16)$$

откуда доверительный интервал

$$I_{\beta,n} = [D'_{\hat{x}}; D''_{\hat{x}}] \approx \left[\widetilde{D}_{\hat{x}} - D_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta;\ \ \widetilde{D}_{\hat{x}} + D_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta \right]. \qquad (5.2.17)$$

Выражения (5.2.15) – (5.2.17) были получены в предположении, что дисперсия $D_{\hat{x}}$ известна. В действительности известна лишь её оценка $\widetilde{D}_{\hat{x}}$, которая только и может фигурировать в этих формулах. С учётом сказанного будем иметь

$$\beta_{\varepsilon,n} \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n-1}}{\widetilde{D}_{\hat{x}}\sqrt{2}} \right); \qquad (5.2.18)$$

$$\varepsilon_{\beta,n} \approx \widetilde{D}_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta; \qquad (5.2.19)$$

$$I_{\beta,n} \approx \left[\widetilde{D}_{\hat{x}} - \widetilde{D}_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta;\ \ \widetilde{D}_{\hat{x}} + \widetilde{D}_{\hat{x}} \sqrt{\frac{2}{n-1}} t_\beta \right]. \qquad (5.2.20)$$

Как видно из выражения (5.2.20), доверительный интервал для дисперсии $D_{\hat{x}}$ оказывается симметричным относительно оценки $\widetilde{D}_{\hat{x}}$. При этом его ширина зависит от значения $\widetilde{D}_{\hat{x}}$ (рис.5.2).

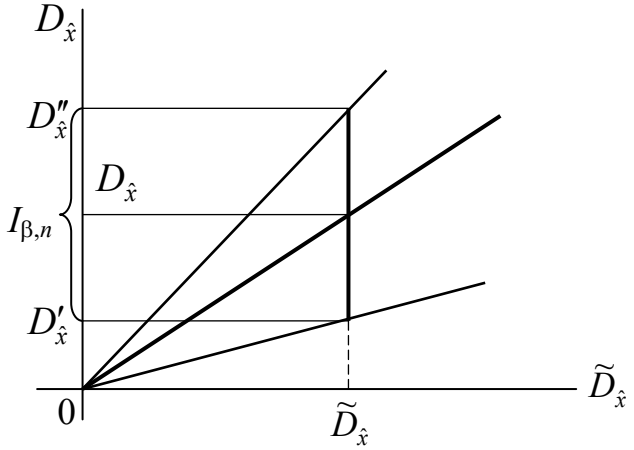


Рис.5.2. Доверительный интервал для дисперсии

Потребный объём экспериментальных данных для оценивания дисперсии $D_{\hat{x}}$ с заданными точностью и надёжностью получим, если выразить n из уравнения (5.2.15):

$$n = n_{\beta,\varepsilon} \geq 2\left(\frac{\widetilde{D}_{\hat{x}} t_\beta}{\varepsilon} \right)^2 + 1, \qquad (5.2.21)$$

Следует отметить, что n определяется, когда дисперсия ещё только подлежит оцениванию. Но значения и вероятностные характеристики оценки $\widetilde{D}_{\hat{x}}$, входящей в выражение (5.2.21), зависят от объёма выборки. Поэтому объём n определяется методом последовательных приближений.

В первом приближении задаются некоторым ориентировочным значением n_0, при котором вычисляется приближение $\widetilde{D}_{\hat{x}}^{0}$ оценки $\widetilde{D}_{\hat{x}}$. Затем оно уточняется в последующих циклах вычислений. Очевидно, что если n_0 превышает найденное по формуле (5.2.21) значение n, то принимается $n = n_0$. В данном случае оценка $\widetilde{D}_{\hat{x}}^{0} = \widetilde{D}_{\hat{x}}$ уже удовлетворяет требованиям по точности и надёжности.

П р и м е р 5.8. В условиях примера 5.6:

1) найти приближённое значение числовых характеристик дисперсии случайной величины $\hat{x}$;

2) построить 80-процентный доверительный интервал для дисперсии;

3) определить доверительную вероятность β для дисперсии, если максимальная с вероятностью β погрешность $\varepsilon_\beta = 0{,}02$.

▼ 1. По формулам (5.2.10) получаем:

$$M[\widetilde{D}_{\hat{x}}] \approx \widetilde{D}_{\hat{x}} = 0{,}064; \quad D[\widetilde{D}_{\hat{x}}] \approx \frac{2}{20-1}(0{,}064)^2 = 0{,}0004;$$

$$\sigma[\widetilde{D}_{\hat{x}}] \approx \sqrt{\frac{2}{20-1}}\,0{,}064 = 0{,}20.$$

2. Используем выражение (5.2.19):

$$\varepsilon = \varepsilon_{0{,}8;\,20} = 0{,}064\sqrt{\frac{2}{20-1}}\,1{,}282 = 0{,}027.$$

Доверительный интервал для дисперсии в соответствии с (5.2.20):

$$I = I_{0{,}8;\,20} \approx [0{,}064 - 0{,}027;\ 0{,}064 + 0{,}027] = [0{,}037;\ 0{,}091].$$

3. По формуле (5.2.18)

$$\beta = \beta_{0{,}02;\,20} \approx 2\Phi_0\left(\frac{0{,}02\sqrt{20-1}}{0{,}064\sqrt{2}}\right) = 2\Phi_0(0{,}97) = 2 \cdot 0{,}3337 = 0{,}6674.$$

Значение $\Phi_0(x)$ найдено в приложении 2.

▲

5.3. Оценивание числовых характеристик и параметров распределения случайных векторов

5.3.1. Двумерный случайный вектор

В разделе 4 были рассмотрены методы оценивания закона распределения случайной величины во всех его возможных формах – ряда, функции и плотности распределения. Аналогичная задача возникает и при обработке экспериментальных данных в виде случайных векторов,

т.е. систем стохастически связанных между собой случайных величин. Такие системы характеризуются многомерными законами распределения.

В данном параграфе более подробно остановимся на оценивании параметров распределения случайных векторов. При этом начнём с рассмотрения частного случая – двумерного вектора, т.е. системы двух случайных величин.

Над системой двух случайных величин $(\hat{x}; \hat{y})$ произведено n независимых равноточных наблюдений, в результате которых получена последовательность пар чисел $(x_i; y_i)$, $i = \overline{1,n}$ (табл.5.3), которые можно интерпретировать как координаты точек $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$ плоскости.

Таблица 5.3

Двумерный массив экспериментальных данных

i	1	2	…	i	…	n
x_i	x_1	x_2	…	x_i	…	x_n
y_i	y_1	y_2	…	y_i	…	y_n

Требуется по результатам наблюдений определить состоятельные, несмещённые и эффективные (асимптотически эффективные) оценки числовых характеристик $M_{\hat{x}}$, $M_{\hat{y}}$, $D_{\hat{x}}$, $D_{\hat{y}}$, $K_{\hat{x}\hat{y}}$ системы случайных величин $(\hat{x}; \hat{y})$. В данном случае $K_{\hat{x}\hat{y}}$ – корреляционный момент $\hat{x}$ и $\hat{y}$.

Задача определения точечных оценок параметров двумерного распределения решается так же, как и для одной случайной величины. При этом оценки координат $M_{\hat{x}}$, $M_{\hat{y}}$ центра рассеяния системы $(\hat{x}; \hat{y})$ находятся по формулам

$$\widetilde{M}_{\hat{x}} = M^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \widetilde{M}_{\hat{y}} = M^*_{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{5.3.1}$$

а оценки элементов её корреляционной матрицы $K_{\hat{x}\hat{y}}$ определяются выражениями

$$\left.\begin{aligned} \widetilde{D}_{\hat{x}} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^2; \\ \widetilde{D}_{\hat{y}} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \widetilde{M}_{\hat{y}})^2; \\ \widetilde{K}_{\hat{x}\hat{y}} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})(y_i - \widetilde{M}_{\hat{y}}). \end{aligned}\right\} \tag{5.3.2}$$

Если математические ожидания $M_{\hat{x}}$, $M_{\hat{y}}$ известны, то элементы корреляционной матрицы определяются выборочными дисперсиями и корреляционным моментом (см. п.п. 5.2.1):

$$\left.\begin{aligned}
\widetilde{D}_{\hat{x}} &= D^*_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^2; \\
\widetilde{D}_{\hat{y}} &= D^*_{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widetilde{M}_{\hat{y}})^2; \\
\widetilde{K}_{\hat{x}\hat{y}} &= K^*_{\hat{x}\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})(y_i - \widetilde{M}_{\hat{y}}).
\end{aligned}\right\}$$

При вычислении оценок $\widetilde{D}_{\hat{x}}$, $\widetilde{D}_{\hat{y}}$, $\widetilde{K}_{\hat{x}\hat{y}}$ целесообразно воспользоваться известной связью между центральными и начальными моментами, которая имеет место и для их статистических аналогов

$$\left.\begin{aligned}
D^*_{\hat{x}} &= \mu^*_2[\hat{x}] = \nu^*_2[\hat{x}] - (\nu^*_1[\hat{x}])^2; \\
\widetilde{D}_{\hat{y}} &= \mu^*_2[\hat{y}] = \nu^*_2[\hat{y}] - (\nu^*_1[\hat{y}])^2; \\
\widetilde{K}_{\hat{x}\hat{y}} &= \mu^*_{11}[\hat{x};\,\hat{y}] = \nu^*_{11}[\hat{x};\,\hat{y}] - \nu^*_1[\hat{x}]\nu^*_1[\hat{y}].
\end{aligned}\right\} \qquad (5.3.3)$$

С учётом (5.3.3) выражения (5.3.2) принимают вид, который обычно используется на практике:

$$\left.\begin{aligned}
\widetilde{D}_{\hat{x}} &= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \widetilde{M}^2_{\hat{x}}\right)\frac{n}{n-1}; \\
\widetilde{D}_{\hat{y}} &= \left(\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \widetilde{M}^2_{\hat{y}}\right)\frac{n}{n-1}; \\
\widetilde{K}_{\hat{x}\hat{y}} &= \left(\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)\right)\frac{n}{n-1}.
\end{aligned}\right\} \qquad (5.3.4)$$

Очевидно, что оценка $\widetilde{r}_{\hat{x}\hat{y}}$ коэффициента корреляции $r_{\hat{x}\hat{y}}$ найдётся по формуле

$$\widetilde{r}_{\hat{x}\hat{y}} = \frac{\widetilde{K}_{\hat{x}\hat{y}}}{\widetilde{\sigma}_{\hat{x}}\widetilde{\sigma}_{\hat{y}}}. \qquad (5.3.5)$$

П р и м е р 5.9. Пусть $\hat{x}$ и $\hat{y}$ – координаты пробоины в мишени после выстрела (в сантиметрах). По мишени произведено 10 независимых выстрелов, результаты которых сведены в табл.5.4, где i – номер выстрела. Найти оценки числовых характеристик $M_{\hat{x}}$, $M_{\hat{y}}$, $\sigma_{\hat{x}}$, $\sigma_{\hat{y}}$, $r_{\hat{x}\hat{y}}$, системы случайных величин $(\hat{x};\,\hat{y})$.

Т а б л и ц а 5.4

Координаты пробоин в мишени

i	1	2	3	4	5	6	7	8	9	10
x_i	3	1	2,5	1,5	4	3,5	2	2,5	1,5	1
y_i	5	2	4	3,5	1,5	5,5	2,5	4,5	2	0

▼ По формулам (5.3.1) получим

$$\widetilde{M}_{\hat{x}} = \frac{1}{10}\sum_{i=1}^{10} x_i = 2{,}55 \text{ см}, \quad \widetilde{M}_{\hat{y}} = \frac{1}{10}\sum_{i=1}^{10} y_i = 2{,}95 \text{ см}.$$

Используя соотношения (5.3.4), имеем:

$$\widetilde{D}_{\hat{x}} = \left(\frac{1}{10}\sum_{i=1}^{10} x_i^2 - (2{,}55)^2\right)\frac{10}{10-1} = 1{,}068 \text{ см}^2;$$

$$\widetilde{D}_{\hat{y}} = \left(\frac{1}{10}\sum_{i=1}^{10} y_i^2 - (2{,}95)^2\right)\frac{10}{10-1} = 3{,}359 \text{ см}^2;$$

$$\widetilde{K}_{\hat{x}\hat{y}} = \left(\frac{1}{10}\sum_{i=1}^{10} x_i y_i - 2{,}55 \cdot 2{,}95\right)\frac{10}{10-1} = 0{,}986 \text{ см}^2.$$

Наконец, используя табл.5.1, по формулам (5.2.13) и (5.3.5), получим:

$$\widetilde{\sigma}_{\hat{x}} = k_9\sqrt{\widetilde{D}_{\hat{x}}} = 1{,}028\sqrt{1{,}068} = 1{,}064\text{см};$$

$$\widetilde{\sigma}_{\hat{y}} = 1{,}028\sqrt{3{,}359} = 1{,}885 \text{ см};$$

$$\widetilde{r}_{\hat{x}\hat{y}} = \frac{0{,}986}{1{,}064 \cdot 1{,}885} = 0{,}491.$$

▲

5.3.2. Многомерный случайный вектор

Аналогично решается задача оценивания числовых характеристик системы произвольного числа случайных величин.

Пусть имеется m случайных величин $(\hat{x}_1, \hat{x}_2, ..., \hat{x}_m)^{\mathsf{T}} = \hat{X}_{<m>}$. Над системой произведено n независимых равноточных наблюдений, результаты которых оформлены в виде табл.5.5. Указанная таблица носит название ***простой статистической матрицы***.

Таблица 5.5

Простая статистическая матрица

i	x_{ij}					
	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{im}
1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1m}
2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2m}
...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{im}
...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	⋯	x_{nj}	⋯	x_{nm}

В табл.5.5 величины x_{ij}, $i = \overline{1,n}$, $j = \overline{1,m}$ – это значения, принятые случайной величиной $\hat{x}_j$ в i-м опыте.

Требуется найти оценки числовых характеристик системы случайных величин, т.е. оценки математических ожиданий $(M_{\hat{x}_1},\, M_{\hat{x}_2},...,M_{\hat{x}_m})^{\mathsf{T}} = M_{\hat{X}_{<m>}}$ и элементов корреляционной матрицы

$$K_{\hat{X}_{<m>}} = \| K_{il} \|_m^m = \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1m} \\ & K_{22} & \cdots & K_{2m} \\ & & \cdots & \cdots \\ & & & K_{mm} \end{pmatrix},$$

где $K_{jl} = K_{\hat{x}_j \hat{x}_l}$, $K_{jj} = K_{\hat{x}_j \hat{x}_j} = D_{\hat{x}_j}$, $j,l = \overline{1,m}$.

Выведенные ранее формулы для вычисления состоятельных, несмещённых и эффективных (асимптотически эффективных) оценок числовых характеристик в общем случае системы m случайных величин приобретают следующий вид:

$$\left.\begin{aligned} &\widetilde{M}_j = \widetilde{M}_{\hat{x}_j} = M^*_{\hat{x}_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}; \\ &\widetilde{D}_j = \widetilde{D}_{\hat{x}_j} = \frac{n}{n-1} D^*_{\hat{x}_j} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \widetilde{M}_{\hat{x}_j})^2; \\ &\widetilde{K}_{jl} = \widetilde{K}_{\hat{x}_j \hat{x}_l} = \frac{n}{n-1} K^*_{\hat{x}_j \hat{x}_l} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ij} - \widetilde{M}_{\hat{x}_j})(x_{il} - \widetilde{M}_{\hat{x}_l}). \end{aligned}\right\} \qquad (5.3.6)$$

Для вычисления оценок (5.3.6) могут быть использованы формулы типа (5.3.4). Оценки средних квадратических отклонений:

$$\widetilde{\sigma}_j = \widetilde{\sigma}_{\hat{x}_j} = \sqrt{\widetilde{D}_{\hat{x}_j}}, \quad j = \overline{1,m}$$

Зная $\widetilde{K}_{jl}$ и $\widetilde{\sigma}_j$, нетрудно найти оценки элементов нормированной корреляционной матрицы по формулам

$$\widetilde{r}_{jl} = \widetilde{r}_{\hat{x}_j \hat{x}_l} = \frac{\widetilde{K}_{jl}}{\widetilde{\sigma}_j \widetilde{\sigma}_l}, \quad j,l = \overline{1,m}. \qquad (5.3.7)$$

5.3.3. Качество оценивания числовых характеристик случайных векторов

Основные вероятностные свойства m-мерного случайного вектора

$$\hat{X}_{<m>} = (\hat{x}_1,\, \hat{x}_2,...,\hat{x}_m)^{\mathsf{T}}$$

описываются m-мерным вектором математических ожиданий его компонент

$$M_{\hat{X}_{<m>}} = (M_{\hat{x}_1},\, M_{\hat{x}_2},...,M_{\hat{x}_m})^{\mathsf{T}} = (\overline{x}_1,\, \overline{x}_2,...,\overline{x}_m)^{\mathsf{T}} = \overline{X}_{<m>}.$$

и корреляционной матрицей m-го порядка

$$K_{\hat{X}_{<m>}} = \| K_{\hat{x}_j \hat{x}_l} \|_m^m = \begin{pmatrix} D_{\hat{x}_1} & K_{\hat{x}_1\hat{x}_2} & \cdots & K_{\hat{x}_1\hat{x}_m} \\ & D_{\hat{x}_2} & \cdots & K_{\hat{x}_2\hat{x}_m} \\ & & \cdots & \cdots \\ & & & D_{\hat{x}_m} \end{pmatrix} = \overline{\dot{X}_{<m>} \dot{X}^{\mathsf{T}}_{<m>}},$$

т.е. всего $m + m^2$ числовыми характеристиками. Поскольку корреляционная матрица симметрична, то число различных её элементов равно $0{,}5(m + m^2)$. Таким образом, при оценивании числовых характеристик m-мерного случайного вектора $\hat{X}_{<m>}$ необходимо вычислить $0{,}5(m^2 + 3m)$ оценок, из которых m оценок $\widetilde{M}_{\hat{x}_j}$ математических ожиданий $M_{\hat{x}_j}$, $j = \overline{1,m}$, такое же количество оценок $\widetilde{D}_{\hat{x}_j}$ дисперсий $D_{\hat{x}_j}$, $j = \overline{1,m}$ и $\widetilde{K}_{\hat{x}_j\hat{x}_l}$ оценок корреляционных моментов $K_{\hat{x}_j\hat{x}_l}$, $i, j = \overline{1,m}$.

Задача анализа качества оценивания числовых характеристик случайного вектора $\hat{X}_{<m>}$ заключается в построении для них $0{,}5(m^2+3m)$ доверительных интервалов

$$I_{\beta_1,n} = [M'_{\hat{x}_j}; M''_{\hat{x}_j}], \quad j = \overline{1,m},$$

$$I_{\beta_2,n} = [D'_{\hat{x}_j}; D''_{\hat{x}_j}], \quad j = \overline{1,m},$$

$$I_{\beta_3,n} = [K'_{\hat{x}_j\hat{x}_l}; K''_{\hat{x}_j\hat{x}_l}], \quad j,l = \overline{1,m},\ j < l$$

и вычислении такого же числа доверительных вероятностей:

$$\beta_{I_1,n} = P(M'_{\hat{x}_j} \le M_{\hat{x}_j} \le M''_{\hat{x}_j}), \quad j = \overline{1,m},$$

$$\beta_{I_2,n} = P(D'_{\hat{x}_j} \le D_{\hat{x}_j} \le D''_{\hat{x}_j}), \quad j = \overline{1,m},$$

$$\beta_{I_3,n} = P(K'_{\hat{x}_j\hat{x}_l} \le K_{\hat{x}_j\hat{x}_l} \le K''_{\hat{x}_j\hat{x}_l}), \quad j,l = \overline{1,m},\ j < l.$$

Методики анализа точности и надёжности оценивания числовых характеристик положения (математических ожиданий и рассеяния (дисперсий и средних квадратических отклонений) были подробно рассмотрены в § 5.2. Поэтому здесь основное внимание будет уделено анализу качества оценивания числовых характеристик связи (корреляционных моментов и коэффициентов корреляции) компонент случайного вектора $\hat{X}_{<m>}$.

При $k_{n-1} \approx 1$ точечные оценки $\widetilde{K}_{\hat{x}_j\hat{x}_l}$ и $\widetilde{r}_{\hat{x}_j\hat{x}_l}$ вычисляются соответственно по третьей формуле (5.3.6) и (5.3.7). Известно, что корреляционный момент кроме связи компонент случайного вектора $\hat{X}_{<m>}$ характеризует и их рассеяние. Поэтому в качестве основной характеристики связи чаще всего используется коэффициент корреляции $r_{\hat{x}_j\hat{x}_l} = r_{jl} = r$.

При вычислении доверительной вероятности $\beta = \beta_{I,n}$ и построении доверительного интервала $I = I_{\beta,n}$ для коэффициента корреляции необходимо знать закон распределения его оценки $\tilde{r}$. Оказывается, что независимо от распределения случайного вектора $\hat{X}^{il}_{<2>} = (\hat{x}_j; \hat{x}_l)^{\mathsf{T}}$ при достаточно большом объёме n выборки (практически при $n > 30$) закон распределения оценки $\tilde{r}$ близок к нормальному [1] с параметрами

$$M_{\tilde{r}} \approx r, \quad \sigma_{\tilde{r}} \approx \frac{\sqrt{1-r^2}}{\sqrt{n}},$$

иначе

$$\varphi_{\tilde{r}} = \varphi^{\text{н}}_{\tilde{r}}(x;\, M_{\tilde{r}};\, \sigma_{\tilde{r}}) = \varphi^{\text{н}}_{\tilde{r}}\left(x;\, r;\, \frac{\sqrt{1-r^2}}{\sqrt{n}}\right).$$

Поэтому доверительная вероятность

$$\begin{aligned} \beta = \beta_{I,n} = \beta_{\varepsilon,n} = P(|\tilde{r} - r| \le \varepsilon) = 2\Phi_0\left(\frac{\varepsilon}{\sigma_{\tilde{r}}}\right) = \\ = 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{1-r^2}}\right) \approx 2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sqrt{1-\tilde{r}^2}}\right). \end{aligned} \tag{5.3.8}$$

Выражая из уравнения (5.3.8) величину ε, будем иметь

$$\varepsilon = \varepsilon_{\beta,n} = \sigma_{\tilde{r}} t_\beta = \frac{\sqrt{1-r^2}}{\sqrt{n}} t_\beta \approx \frac{\sqrt{1-\tilde{r}^2}}{\sqrt{n}} t_\beta, \tag{5.3.9}$$

откуда

$$\begin{aligned} I = I_{\beta,n} = [r'; r''] = \left[\tilde{r} - \frac{\sqrt{1-r^2}}{\sqrt{n}} t_\beta;\ \tilde{r} + \frac{\sqrt{1-r^2}}{\sqrt{n}} t_\beta\right] \approx \\ \approx \left[\tilde{r} - \frac{\sqrt{1-\tilde{r}^2}}{\sqrt{n}} t_\beta;\ \tilde{r} + \frac{\sqrt{1-\tilde{r}^2}}{\sqrt{n}} t_\beta\right]. \end{aligned} \tag{5.3.10}$$

Соотношением (5.3.10) описывается 100β-процентной доверительный интервал для коэффициента корреляции r.

Разрешив уравнение (5.3.9) относительно n, получим выражение для объёма выборки, потребного при оценивании r с точностью ε (с абсолютной ошибкой, не превосходящей ε) и надёжностью β (доверительной вероятностью β):

$$n = n_{\beta,\varepsilon} \ge \frac{1-r^2}{\varepsilon^2} t_\beta^2 \approx \frac{1-\tilde{r}^2}{\varepsilon^2} t_\beta. \tag{5.3.11}$$

Следует обратить внимание на то, что во всех выражениях (5.3.8)–(5.3.11) реальными являются лишь приближённые равенства, так как на этапе оценивания (как точечного, так и интервального) истинное значе-

ние коэффициента корреляции неизвестно. В связи с этим потребный объём выборки может быть определён лишь методом последовательных приближений. Сущность этого метода была раскрыта в п.п.5.2.3.

П р и м е р 5.10. Пусть $\hat{x}$, $\hat{y}$ – координаты пробоины в мишени. Произведено 40 выстрелов. Коэффициент корреляции случайных величин $\hat{x}$ и $\hat{y}$ составил $r = 0{,}605$. Найти:

1) доверительную вероятность β для r, если максимальная с вероятностью β абсолютная погрешность оценки $\tilde{r}$ должна быть не более 0,1;

2) доверительный интервал для r при доверительной вероятности $\beta = 0{,}85$.

▼ 1. По формуле (5.3.8) находим

$$\beta = \beta_{0{,}1;\,40} = 2\Phi_0\left(\frac{0{,}1\sqrt{40}}{\sqrt{1-(0{,}605)^2}}\right) = 2\Phi_0(0{,}79) = 0{,}5704.$$

Значение функции $\Phi_0(x)$ – в приложении 2.

2. Для $\beta = 0{,}85$ в приложении 4 находим $t_\beta = t_{0{,}85} = 1{,}439$.

Для вычисления погрешности ε используем выражение (5.3.9):

$$\varepsilon = \varepsilon_{0{,}85;\,40} = \frac{\sqrt{1-(0{,}65)^2}}{\sqrt{40}}1{,}439 = 0{,}181.$$

В соответствии с выражением (5.3.10) доверительный интервал

$I = I_{0{,}85;\,40} = [0{,}605–0{,}181;\quad 0{,}605+0{,}181] = [0{,}424;\quad 0{,}786]$.

▲

5.4. Оценивание числовых характеристик случайных функций

Известно, что случайная функция может рассматриваться как обобщение понятия случайного вектора (системы случайных величин) на бесконечное множество составляющих его компонентов (сечений случайной функции). Исчерпывающего вероятностного описания такого случайного объекта не существует, поэтому на практике используются лишь законы распределения и числовые характеристики систем конечного числа сечений случайных функций. При этом из-за сложности построения статистических законов распределения многомерных случайных векторов наиболее широкое применение получила корреляционная теория, в рамках которой изучаются лишь первые и вторые моменты распределений случайных функций, т.е. их математические ожидания, дисперсии и корреляционные функции.

5.4.1. Нестационарные случайные функции

Реализации $x_i(t)$, $i=\overline{1,n}$ случайной функции $\hat{x}(t)$ представляют собой неслучайные функции, значения которых $x_i(t_j)$ в фиксированных точках t_j, $j = 1,2,\ldots$ являются реализациями x_{ij} случайных величин $\hat{x}_j = \hat{x}(t_j)$.

Пусть над случайной функцией $\hat{x}(t)$ произведено n независимых равноточных наблюдений (опытов), в результате которых получено n её реализаций $x_i(t)$, $i=\overline{1,n}$ (рис.5.3).

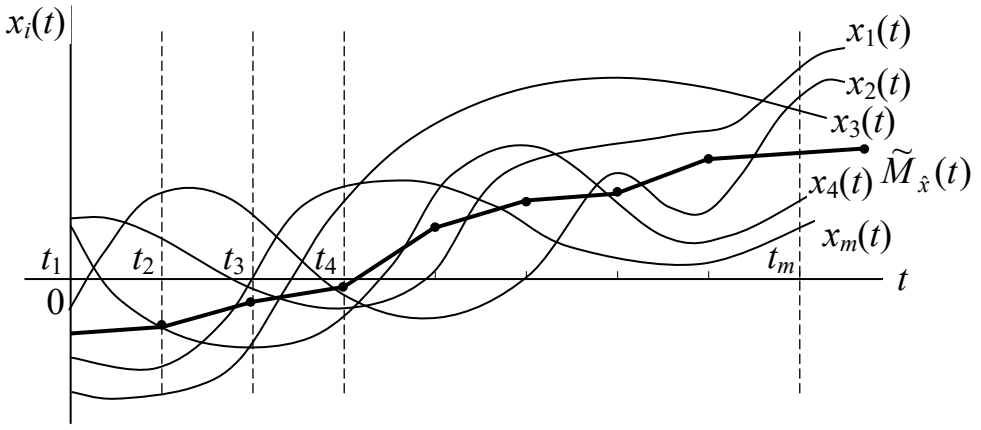


Рис.5.3. Реализации случайной функции

Требуется найти оценки числовых характеристик случайной функции: математического ожидания $M_{\hat{x}}(t)=\overline{x}(t)$, дисперсии $D_{\hat{x}}(t)=\overline{\dot{x}^2}(t)$ и корреляционной функции $K_{\hat{x}}(t';t'')=\overline{\dot{x}(t')\dot{x}(t'')}$, удовлетворяющие требованиям состоятельности, несмещённости и эффективности (асимптотической эффективности).

В ряде сечений случайной функции, соответствующих моментам времени t_1, $t_2,\ldots$, $t_j,\ldots,t_m$, фиксируются значения, принятые реализациями $x_i(t)$ функции $\hat{x}(t)$ в эти моменты. Поскольку наблюдалось n реализаций, то каждому из моментов t_j, $j=\overline{1,m}$ будут соответствовать n значений, принятых случайной величиной $\hat{x}_j = \hat{x}(t_j)$. Указанная случайная величина является j-м сечением случайной функции $\hat{x}(t)$. Расстояния

$$h = h_j = t_{j+1} - t_j$$

между фиксируемыми сечениями $\hat{x}(t_j)$ случайной функции $\hat{x}(t)$ обычно берутся одинаковыми и назначаются так, чтобы последовательность $x_i(t_j)$, $j=\overline{1,m}$ позволяла восстановить основной характер зависимости $x_i(t)$. Нередко в основу выбора кладётся теорема В.А. Котельникова, согласно которой для точного восстановления непрерывной функции достаточно её наблюдать в равноотстоящих дискретных точках с частотой, в два раза превышающей максимум её частотного спектра [2,13]. Бывает, что при-

ведённые соображения являются излишними и расстояние задаётся темпом работы регистрирующей аппаратуры.

Для удобства последующей статистической обработки зарегистрированные данные сводятся в таблицу, строки которой соответствуют реализациям, а столбцы – сечениям случайной функции (табл.5.6).

Таблица 5.6

Зарегистрированные значения случайной функции

$x_i(t)$	t							
	t_1	t_2	...	t_j	...	t_l	...	t_m
$x_1(t)$	$x_1(t_1)$	$x_1(t_2)$	...	$x_1(t_j)$	...	$x_1(t_l)$	...	$x_1(t_m)$
$x_2(t)$	$x_2(t_1)$	$x_2(t_2)$	...	$x_2(t_j)$	...	$x_2(l_j)$	...	$x_2(t_m)$
...	...	...	...	...	...	...	...	...
$x_i(t)$	$x_i(t_1)$	$x_i(t_2)$	...	$x_i(t_j)$	...	$x_i(t_l)$	...	$x_i(t_m)$
...	...	...	...	...	...	...	...	...
$x_n(t)$	$x_n(t_1)$	$x_n(t_2)$	...	$x_n(t_j)$	...	$x_n(t_l)$	...	$x_n(t_m)$

Приведённая в табл.5.6 совокупность значений случайной функции $\hat{x}(t)$ представляет собой результаты n наблюдений m-мерного случайного вектора

$$\hat{X}_{<m>} = \left(\hat{x}(t_1),\ \hat{x}(t_2),...,\hat{x}(t_m)\right)^{\mathsf{T}}$$

и обрабатывается по методике § 5.3.

Так, оценки математических ожиданий сечений $\hat{x}(t_j)$ случайной функции $\hat{x}(t)$ находятся по формулам

$$\widetilde{M}_{\hat{x}}(t_j) = M^*_{\hat{x}}(t_j) = \frac{1}{n}\sum_{i=1}^{n} x_i(t_j),\quad j=\overline{1,m}. \qquad (5.4.1)$$

Соединяя точки $\left(t_j;\widetilde{M}_{\hat{x}}(t_j)\right)$ отрезками прямых, можно построить приближённый график (рис.5.3) оценки $\widetilde{M}_{\hat{x}}(t)$ математического ожидания $M_{\hat{x}}(t)$ случайной функции $\hat{x}(t)$. Очевидно, что возможны и другие виды интерполяции, например, квадратичная.

Несмещенные оценки дисперсий и корреляционных моментов сечений определяются соответственно следующими соотношениями:

$$\widetilde{D}_{\hat{x}}(t_j) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t_j)-\widetilde{M}_{\hat{x}}(t_j)\right)^2,\quad j=\overline{1,m}. \qquad (5.4.2)$$

$$\widetilde{K}_{\hat{x}}(t_j;t_l) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i(t_j)-\widetilde{M}_{\hat{x}}(t_l)\right)\left(x_i(t_l)-\widetilde{M}_{\hat{x}}(t_l)\right),\quad j,l=\overline{1,m}. \qquad (5.4.3)$$

Легко заметить, что формула (5.4.2) может быть получена и из выражения (5.4.3) при $l=j$, поскольку $D_{\hat{x}}(t_j) = K_{\hat{x}}(t_j;t_j)$.

В вычислительном отношении более удобны формулы, основанные на связи начальных и центральных моментов, т.е.

$$\widetilde{D}_{\hat{x}}(t_j)=\left(\frac{1}{n}\sum_{i=1}^{n}(x_i^2(t_j)-\widetilde{M}_{\hat{x}}^2(t_j)\right)\frac{n}{n-1},\quad j=\overline{1,m}. \qquad (5.4.4)$$

$$\widetilde{K}_{\hat{x}}(t_j;t_l)=\left(\frac{1}{n}\sum_{i=1}^{n}x_i(t_j)x_i(t_l)-\widetilde{M}_{\hat{x}}(t_l)\widetilde{M}_{\hat{x}}(t_l)\right)\frac{n}{n-1},\quad j,l=\overline{1,m}. \qquad (5.4.5)$$

При практическом использовании формул (5.4.4) и (5.4.5) рекомендуется начало отсчёта значений случайной функции перенести ближе к её математическому ожиданию. Это позволит избежать вычислений разности близких чисел.

П р и м е р 5.11. Результаты наблюдения 11 реализаций случайной функции $\hat{x}(t)$ в момент времени $t_j=\overline{0,10}$ с приведены в табл.5.7.

Т а б л и ц а 5.7

Реализации случайной функции

$x_i(t)$	t										
	0	1	2	3	4	5	6	7	8	9	10
$x_1(t)$	0,7	1,2	2,0	3,2	4,7	6,0	6,4	6,6	6,3	5,6	5,0
$x_2(t)$	1,2	2,0	3,6	4,6	5,1	5,5	6,0	6,2	6,2	6,0	6,0
$x_3(t)$	2,0	3,3	4,1	4,4	4,5	4,5	4,8	5,5	6,0	6,3	6,2
$x_4(t)$	2,5	2,9	3,0	3,2	3,8	4,7	5,4	5,5	5,4	5,7	6,2
$x_5(t)$	2,7	3,8	4,7	5,1	5,3	5,2	5,0	4,9	5,1	5,7	6,6
$x_6(t)$	3,2	3,9	4,1	4,1	4,0	4,2	5,0	6,0	6,3	6,1	5,7
$x_7(t)$	3,8	4,5	5,0	5,4	5,5	5,6	5,5	5,5	5,2	5,0	4,9
$x_8(t)$	4,1	3,8	3,6	3,9	4,8	5,8	6,2	6,1	5,7	5,4	5,4
$x_9(t)$	4,2	4,9	5,0	4,6	4,3	4,0	4,2	4,7	5,7	6,6	6,8
$x_{10}(t)$	5,4	4,3	3,2	2,9	3,1	3,8	4,5	5,5	6,4	7,0	7,2
$x_{11}(t)$	5,8	5,6	5,4	5,2	4,8	4,5	4,3	4,3	4,4	4,5	4,8

Требуется определить оценки числовых характеристик случайной функции $\hat{x}(t)$.

▼ По формуле (5.4.1) вычисляются оценки $\widetilde{M}_{\hat{x}}(t_j)$ и результаты сводятся в табл.5.8.

Т а б л и ц а 5.8

Оценки математических ожиданий сечений случайной функции

t_j	0	1	2	3	4	5	6	7	8	9	10
$\widetilde{M}_{\hat{x}}(t_j)$	3,2	3,7	4,0	4,2	4,5	4,9	5,2	5,5	5,7	5,8	5,9

По формуле (5.4.3) или (5.4.5) вычисляются оценки $\widetilde{K}_{\hat{x}}(t_j;t_l)$ и результаты сводятся в табл.5.9, диагональные элементы $\widetilde{K}_{\hat{x}}(t_j;t_j)$ которой представляют собой оценки $\widetilde{D}_{\hat{x}}(t_j)$ дисперсий $D_{\hat{x}}(t_j)$ в сечениях $\hat{x}(t_j)$ случайной функции $\hat{x}(t)$.

В рассмотренном примере пришлось 11 производить вычисления по формуле (5.4.1) и 66 раз – по формуле (5.4.3) или (5.4.5). Это свиде-

тельствует о большой трудоёмкости задачи оценивания вероятностных характеристик нестационарных случайных функций.

Таблица 5.9

Корреляционные моменты случайной функции

t_l	t_j										
	0	1	2	3	4	5	6	7	8	9	10
0	2,6	1,9	0,9	0,2	–0,3	–0,7	–0,8	–0,7	–0,4	–0,1	–0,1
1		1,6	1,1	0,5	–0,07	–0,5	–0,7	–0,7	–0,5	–0,1	0,07
2			1,0	0,7	0,3	–0,06	–0,5	–0,6	–0,4	–0,2	–0,05
3				0,7	0,7	0,09	–0,2	–0,3	–0,3	–0,3	–0,2
4					0,5	0,4	0,2	–0,03	–0,2	–0,3	–0,3
5						0,6	0,5	0,3	0	–0,3	–0,4
6							0,6	0,4	0,1	–0,1	–0,3
7								0,5	0,3	–0,2	–0,4
8									0,4	0,3	0,2
9										0,5	0,5
10											0,7

▲

5.4.2. Стационарные случайные функции

По определению, случайная функция $\hat{u}(t)$ является стационарной (в широком смысле), если её математическое ожидание и дисперсия постоянны, а корреляционная функция зависит лишь от расстояния между сечениями случайной функции:

$$M_{\hat{u}(t)} = \overline{u(t)} = M_{\hat{u}} = \text{const};$$

$$D_{\hat{u}}(t) = \overline{\dot{u}^2(t)} = D_{\hat{u}} = \text{const};$$

$$K_{\hat{u}}(t';t'') = \overline{\dot{u}(t')\dot{u}(t'')} = K_{\hat{u}}(t';t'+\tau) = K_{\hat{u}}(\tau).$$

Класс стационарных случайных функций достаточно многообразен. Однако в практическом отношении наибольший интерес представляют стационарные случайные функции, обладающие эргодическим свойством, для которых одна реализация достаточно большой продолжительности содержит о случайной функции столько же информации, сколько её содержит и множество реализаций той же суммарной продолжительности. Другими словами, каждая из реализаций эргодической стационарной случайной функции является представителем всего их ансамбля. Как отмечено в работе [12], следует различать эргодические свойства случайных функций по отношению к моментам их распределения различных порядков. При этом под эргодичными обычно понимаются случайные функции, обладающие такими свойствами по отношению к моментам первого и второго порядков, т.е. к математическому ожиданию и корре-

ляционной функции (следовательно, и к дисперсии). Далее рассматриваются только такие случайные функции.

Оценки числовых характеристик эргодичных случайных функций могут быть приближённо определены не как средние по множеству реализаций, а как средние по времени T наблюдения одной реализации по следующим формулам:

$$\widetilde{M}_{\hat{u}} = \widetilde{M}[\hat{u}(t)] \approx \frac{1}{T}\int_0^T u(z)dz, \quad t \in [0;\, T]; \tag{5.4.6}$$

$$\widetilde{K}_{\hat{u}}(\tau) = \widetilde{M}[\dot{u}(t)\dot{u}(t+\tau)] \approx \frac{1}{T-\tau}\int_0^{T-\tau} \dot{u}(z)\dot{u}(z+\tau)dz, \quad t \in [0;\, T\text{–}\tau]; \tag{5.4.7}$$

$$\widetilde{D}_{\hat{u}} = \widetilde{D}[\hat{u}(t)] = \widetilde{K}_{\hat{u}}(0) \approx \frac{1}{T}\int_0^T \dot{u}^2(z)dz, \quad t \in [0;\, T], \tag{5.4.8}$$

где $\dot{u}(t) = u(t) - \widetilde{M}_{\hat{u}}$.

Обоснованием применимости формул (5.4.6) – (5.4.8) служит тот факт, что для эргодичных стационарных случайных функций средние во времени оценки сходятся по вероятности к оцениваемым ими характеристикам $M_{\hat{u}}$, $K_{\hat{u}}(\tau)$, $D_{\hat{u}}$.

Из выражений (5.4.6) – (5.4.8) видно, что для их практического применения требуется интегрировать ряд функций от реализации $u(t)$ случайной функции $\hat{u}(t)$. Чаще всего на практике для нахождения оценок (5.4.6) – (5.4.8) используется следующая методика.

Пусть на интервале времени [0; T] наблюдалась реализация эргодичной стационарной случайной функции $\hat{u}(t)$, значения которой $u(t_j)$ в ряде равноотстоящих опорных моментов времени $t_j = 0{,}5h + kh$, $j = \overline{1,m}$, $k = \overline{0,m-1}$ зарегистрированы и сведены в табл.5.10.

Таблица 5.10

Значения реализации эргодичной стационарной функции в выбранные моменты времени

t	t_1	t_2	…	t_j	…	t_m
$u(t)$	$u(t_1)$	$u(t_2)$	…	$u(t_j)$	…	$u(t_m)$

Требуется по данным этой таблицы определить оценки $\widetilde{M}_{\hat{u}}$, $\widetilde{K}_{\hat{u}}(\tau)$, $\widetilde{D}_{\hat{u}}$ числовых характеристик $M_{\hat{u}}$, $K_{\hat{u}}(\tau)$, $D_{\hat{u}}$ случайной функции $\hat{u}(t)$

Интервал [0; T] наблюдения случайной функции $\hat{u}(t)$ разбивается на m равных подынтервалов длиной $h = T/m$, расположенных симметрично относительно опорных моментов времени t_1, t_2,…, t_m (рис.5.4).

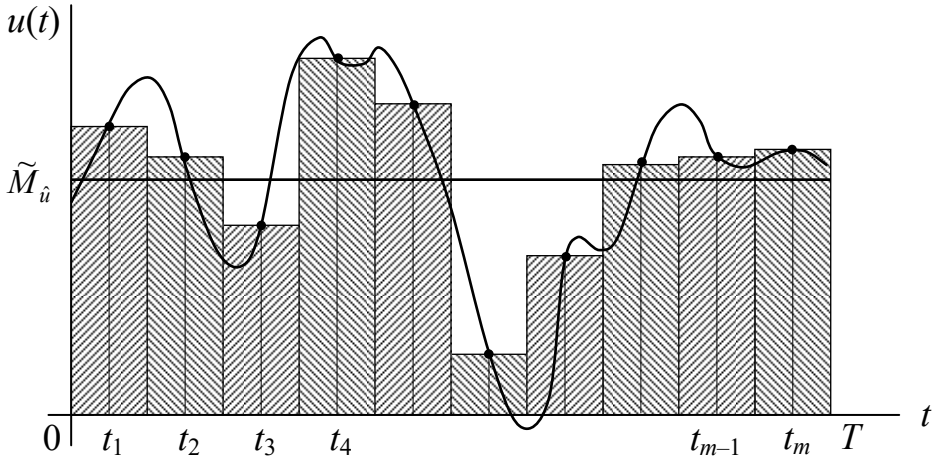


Рис.5.4. Реализация эргодичной стационарной случайной функции

Далее предполагается, что в пределах подынтервала $[t_j–0{,}5h;\ t_j+0{,}5h]$ функция, описывающая реализацию случайной функции, постоянна:

$$u(t) = u(t_j) = \text{const}, \quad t \in [t_j - 0{,}5h;\ t_j + 0{,}5h], \quad j = \overline{1,m},$$

Если h достаточно мало, то можно приближённо полагать, что

$$\int_{t_j-0{,}5h}^{t_j+0{,}5h} u(t)dt = hu(t_j), \quad j = \overline{1,m}. \qquad (5.4.9)$$

Суммирование результатов (5.4.9) по j даёт

$$M_{\hat{u}} \approx \frac{1}{T}\int_0^T u(t)dt = \frac{h}{T}\sum_{j=1}^{m} u(t_j) = \frac{1}{m}\sum_{j=1}^{m} u(t_j). \qquad (5.4.10)$$

Несложно заметить, что выражение (5.4.10) реализует процедуру численного интегрирования по формуле прямоугольников.

Аналогично вычисляется и оценка корреляционной функции для значений аргумента

$$\tau_l = lh = \frac{lT}{m}.$$

Поскольку в выражении (5.4.7) длина интервала интегрирования

$$T - \tau_l = T - \frac{lT}{m} = \frac{m-l}{m}T,$$

то, поделив его на $m - l$ равных участков и вынося на каждом из них за знак интеграла среднее значение функции $\dot{u}(t)\dot{u}(t+\tau)$, получим

$$\widetilde{K}_{\hat{u}}(\tau_l) = \widetilde{K}_{\hat{u}}\left(\frac{lT}{m}\right) \approx \frac{1}{m-l}\sum_{j=1}^{m-l}\dot{u}(t_j)\dot{u}(t_{j+l}), \quad l = \overline{0,L}, \qquad (5.4.11)$$

где $L = m/4$ [2].

При $\tau = 0$ формула (5.4.11) даёт

$$\widetilde{D}_{\hat{u}} = \widetilde{K}_{\hat{u}}(0) \approx \frac{1}{m}\sum_{j=1}^{m} \dot{u}^2(t_j). \qquad (5.4.12)$$

Вычисления по формуле (5.4.11) ведутся последовательно для $l = 0, 1, 2,\ldots$ вплоть до таких значений l_k, при которых функция $\widetilde{K}_{\hat{u}}(lT/m)$ становится практически равной нулю или начинает совершать незначительные колебания около нуля. По полученным точкам может быть построен приближённый график корреляционной функции $K_{\hat{u}}(\tau)$ (рис.5.5).

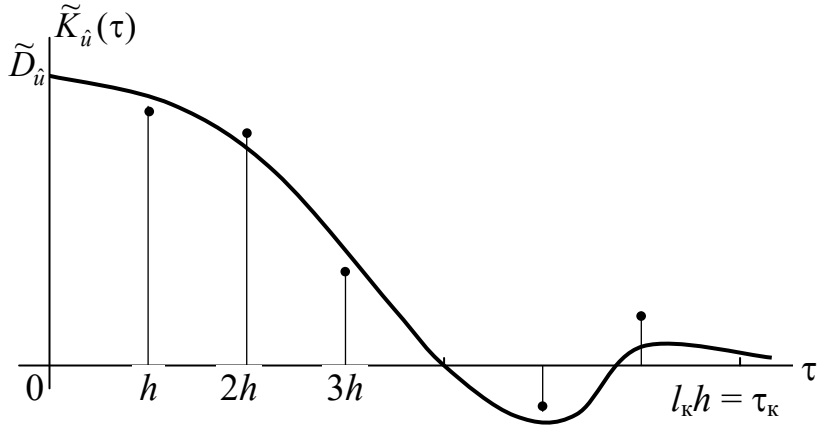


Рис.5.5. Приближённый график корреляционной функции

На представленном рисунке через $\tau_к$ обозначена длина интервала корреляции случайной функции $\hat{u}(t)$, т.е. наименьшее расстояние между сечениями случайной функции, на котором корреляция между ними практически отсутствует ($\tau_к = l_к h$).

П р и м е р 5.12. В табл.5.11 приведены результаты наблюдения эргодичной стационарной случайной функции $\hat{u}(t)$ на интервале времени продолжительностью $T = 28$ с с периодичностью $h = 1$ с в моменты времени t_j, $j = \overline{1, 28}$.

Т а б л и ц а 5.11

Реализация эргодичной стационарной случайной функции

t_j, с	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
$u(t_j)$	4,0	4.1	3,8	2,0	1,0	–0,3	–0,2	1,0	3,5	2,5	2,6	2,0	0,6	2,5
t_j, с	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}	t_{21}	t_{22}	t_{23}	t_{24}	t_{25}	t_{26}	t_{27}	t_{28}
$u(t_j)$	3,3	3,8	1,2	0,5	–0,5	0,5	3,2	4,0	2,3	0,5	2,4	3,2	2,0	0,5

Требуется определить оценки числовых характеристик функции $\hat{u}(t)$.

▼ Необходимые вычисления производятся по формулам (5.4.10–(5.4.12). Результаты расчётов оформлены в виде табл.5.12, наглядно иллюстрирующей все этапы решения задачи.

▲

Решение рассмотренной задачи потребовало однократного использования формулы (5.4.10) и одиннадцатикратного – формулы (5.4.11). Если интервал корреляции $\tau_к$ случайной функции соизмерим с интервалом T её наблюдения, то аргумент τ корреляционной функции должен варьироваться от 0 до T. В этом случае параметр l в формуле (5.4.11) пробегает значения от 0 до $m-1$ и, следовательно, формула (5.4.11) реализуется m раз.

Расчёты показывают, что по сравнению с нестационарными случайными функциями трудоёмкость оценивания числовых характеристик стационарных эргодичных случайных функций существенно снижается.

5.4.3. Качество оценивания числовых характеристик случайных функций

Известно, что сечения случайной функции $\hat{x}(t)$ представляют собой обычные случайные величины, а совокупности сечений – случайные векторы, вероятностные характеристики которых (законы и параметры распределения) зависят от значений t_j, $j=\overline{1,m}$ параметра t.

Выше отмечалось, что формально случайная функция может интерпретироваться как обобщение понятия случайного вектора на случай бесконечного множества его компонентов. Однако на практике используются законы распределения случайных величин ограниченного и, как правило, невысокого порядка. Такая постановка эквивалентна моделированию случайной функции $\hat{x}(t)$ случайным вектором

$$\hat{X}_{<m>}(T_{<m>}) = \left(\hat{x}(t_1),\, \hat{x}(t_2),...,\hat{x}(t_m)\right)^{\mathsf{T}} = (\hat{x}_1,\, \hat{x}_2,...,\hat{x}_m)^{\mathsf{T}} = \hat{X}_{<m>}. \quad (5.4.13)$$

Таким образом, в основе подходов к оцениванию числовых характеристик случайных функций лежат методы оценивания числовых характеристик случайных величин и векторов, рассмотренные ранее. Поэтому ниже рассматриваются лишь некоторые особенности, присущие только случайным функциям.

Таблица 5.12

Расчётная таблица оценок числовых характеристик стационарной эргодичной случайной функции

j	$u(t_j)$	$u(t_j) - \widetilde{M}_{\hat{u}}$	$\dot{u}(t_j)\dot{u}(t_j + \tau_l) = (u(t_j) - \widetilde{M}_{\hat{u}})(u(t_j + \tau_l) - \widetilde{M}_{\hat{u}})$										
			l										
			0	1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	4,0	2,0	4,00	4,20	3,60	0,00	–2,00	–4,60	–4,40	–2,00	3,00	1,00	1,20
2	4,1	2,1	4,41	3,78	0,00	–2,10	–4,83	–4,62	–2,10	3,15	1,05	1,26	0,00
3	3,8	1,8	3,24	0,00	–1,80	–4,14	–3,96	–1,80	2,70	0,90	1,08	0,00	–2,52
4	2,0	0,0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
5	1,0	–1,0	1,00	2,30	2,20	1,00	–1,50	–0,50	–0,60	0,00	1,40	–0,50	–1,30
6	–0,3	–2,3	5,29	5,06	2,30	–3,45	–1,15	–1,38	0,00	3,22	–1,15	–2,99	–4,14
7	–0,2	–2,2	4,84	2,20	–3,30	–1,10	–1,32	0,00	3,08	–1,10	–2,86	–3,96	1,76
8	1,0	–1,0	1,00	–1,50	–0,50	–0,60	0,00	1,40	–0,50	–1,30	–1,80	0,80	1,50
9	3,5	1,5	2,25	0,75	0,90	0,00	–2,10	0,75	1,95	2,70	–1,20	–2,25	–3,75
10	2,5	0,5	0,25	0,30	0,00	–0,70	0,25	0,65	0,90	–0,40	–0,75	–1,25	–0,75
11	2,6	0,6	0,36	0,00	–0,84	0,30	0,78	1,08	–0,48	–0,90	–1,50	–0,90	0,72
12	2,0	0,0	0,00	0,00	0,00	0,00	0,00	1,00	0,00	0,00	0,00	0,00	0,00
13	0,6	–1.4	1,96	–0,70	–1,82	–2,52	1,12	2,10	3,50	2,10	–1,68	–2,80	–0,42
14	2,5	0,5	0,25	0,65	0,90	–0,40	-0,75	–1,25	–0,75	0,60	1,00	0,15	–0,75
15	3,3	1,3	1,69	2,34	–1,04	–1,95	–3,25	–1,95	1,56	2,60	0,39	–1,95	0,52
16	3,8	1,8	3,24	–1,44	–2,70	–4,50	–2,70	2,16	3,60	0,54	–2,70	0,72	2,16
17	1,2	–0,8	0,64	1,20	2,00	1,20	–0,96	–1,60	–0,24	1,20	–0,32	–0,96	0,00
18	0,5	–1,5	2,25	3,75	2,25	–1,80	–3,00	–0,45	2,25	–0,60	–1,80	0,00	2,25
19	–0,5	–2,5	6,25	3,75	–3,00	–5,00	–0,75	3,75	–1,00	–3,00	0,00	3,75	–
20	0,5	–1,5	2,25	–1,80	–3,00	–0,45	2,25	–0,60	–1,80	0,00	2,25	–	–
21	3,2	1,2	1,44	2,40	0,36	–1,80	0,48	1,44	0,00	–1,80	–	–	–
22	4,0	2,0	4,00	0,60	–3,00	0,80	2,40	0,00	–3,00	–	–	–	–
23	2,3	0,3	0,09	–0,45	0,12	0,36	0,00	–0,045	–	–	–	–	–
24	0,5	–1,5	2,25	–0,60	–1,80	0,00	2,25	–	–	–	–	–	–

j	$u(t_j)$	$u(t_j)-\widetilde{M}_{\hat{u}}$	$\dot{u}(t_j)\dot{u}(t_j+\tau_l)=(u(t_j)-\widetilde{M}_{\hat{u}})(u(t_j+\tau_l)-\widetilde{M}_{\hat{u}})$										
			l										
			0	1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10	11	12	13	14
25	2,4	0,4	0,16	0,48	0,00	–0,60	–	–	–	–	–	–	–
26	3,2	1,2	1,44	0,00	–1,80	–	–	–	–	–	–	–	–
27	2,0	0,0	0,00	0,00	–	–	–	–	–	–	–	–	–
28	0,5	–1,5	2,25	–	–	–	–	–	–	–	–	–	–
Σ	56	0,0	56,80	27,27	–9,97	–27,45	–18,74	–4,87	4,67	5,91	–5,59	–9,88	–3,52
$\widetilde{M}_{\hat{u}}=\frac{1}{28}\cdot 56=2$		$m-l$	28	27	26	25	24	23	22	21	20	19	18
		$\widetilde{K}_{\hat{u}}(\tau_l)$	2,03	1,01	–0,38	–1,10	–0,78	–0,21	0,21	0,28	–0,28	–0,52	–0,19

Числовые характеристики $M_{\hat{x}}(t)$, $D_{\hat{x}}(t)$, $K_{\hat{x}}(t_1;t_2)$ случайной функции $\hat{x}(t)$ зависят от параметра *t*, следовательно, от него зависят как оценки $\widetilde{M}_{\hat{x}}(t)$, $\widetilde{D}_{\hat{x}}(t)$, $\widetilde{K}_{\hat{x}}(t_1;t_2)$, так и характеристики их качества – доверительные интервалы:

$$I_{\beta_1,n} = [M'_{\hat{x}}(t); M''_{\hat{x}}(t)]; \quad (5.4.14)$$

$$I_{\beta_2,n} = [D'_{\hat{x}}(t); D''_{\hat{x}}(t)]; \quad (5.4.15)$$

$$I_{\beta_3,n} = [K'_{\hat{x}}(t_1;t_2); K''_{\hat{x}}(t_1;t_2)], \quad (5.4.16)$$

а также доверительные вероятности:

$$\beta_{I_1,n} = P(M'_{\hat{x}}(t) \le M_{\hat{x}}(t) \le M''_{\hat{x}}(t)); \quad (5.4.17)$$

$$\beta_{I_2,n} = P(D'_{\hat{x}}(t) \le D_{\hat{x}}(t) \le D''_{\hat{x}}(t)); \quad (5.4.18)$$

$$\beta_{I_3,n} = P(K'_{\hat{x}}(t_1;t_2) \le K_{\hat{x}}(t_1;t_2) \le K''_{\hat{x}}(t_1;t_2)), \quad (5.4.19)$$

Во всех приведённых выражениях *n* – это число наблюдаемых реализаций случайной функции $\hat{x}(t)$. Если случайная функция $\hat{u}(t)$ стационарна, то выражения (5.4.14)–(5.4.19) упрощаются и принимают следующий вид:

$$I_{\beta_1,T} = [M'_{\hat{u}}; M''_{\hat{u}}];$$

$$I_{\beta_2,T} = [D'_{\hat{u}}; D''_{\hat{u}}];$$

$$I_{\beta_3,T} = [K'_{\hat{u}}(\tau); K''_{\hat{u}}(\tau)],$$

$$\beta_{I_1,T} = P(M'_{\hat{u}} \le M_{\hat{u}} \le M''_{\hat{u}});$$

$$\beta_{I_2,T} = P(D'_{\hat{u}} \le D_{\hat{u}} \le D''_{\hat{u}});$$

$$\beta_{I_3,T} = P(K'_{\hat{u}}(\tau) \le K_{\hat{u}}(\tau) \le K''_{\hat{u}}(\tau)),$$

где *T* – продолжительность наблюдения случайной функции $\hat{u}(t)$.

5.4.4. Потребный объём наблюдений

При оценивании числовых характеристик нестационарной случайной функции задача состоит в определении потребного числа *n* её реализаций $x_i(t)$, $i = \overline{1, n}$. Поскольку все числовые характеристики случайной функции, а также характеристики (5.4.14) – (5.4.19) качества оценивания зависят от параметра *t*, то от него будет зависеть и потребный объём *n* выборки реализаций:

$$n_1 = n_{\beta_1,I_1} = n_1(t); \quad (5.4.20)$$

$$n_2 = n_{\beta_2,I_2} = n_2(t); \quad (5.4.21)$$

$$n_3 = n_{\beta_3,I_3} = n_3(t_1;t_2). \quad (5.4.22)$$

Выражениями (5.4.20)–(5.4.22) определяется потребный объём реализаций случайной функции для оценки соответственно математического

ожидания, дисперсии и корреляционной функции. Из данных выражений следует, что оценивание числовых характеристик случайной функции для различных её сечений может потребовать различного числа реализаций. Такое требование практически неосуществимо. Поэтому требуемый объём n_k, $k = 1, 2, 3$ определяется как $\max\limits_{t}\{n_k(t)\}$, соответствующий минимуму дисперсии оценки k-й числовой характеристики случайной функции:

$$n_1 = \max_{t\in[0;T]}\{n_1(t)\};$$

$$n_2 = \max_{t\in[0;T]}\{n_2(t)\};$$

$$n_3 = \max_{t_1,t_2\in[0;T]}\{n_3(t_1;t_2)\}.$$

При оценивании числовых характеристик стационарной случайной функции $\hat{u}(t)$ наблюдается всего одна её реализация $u(t)$. Поэтому здесь возникает вопрос о длительности наблюдения T, потребной для оценивания характеристик случайной функции с заданной точностью I или ε и надёжностью β.

Поскольку случайная функция $\hat{u}(t)$ стационарна, то все её сечения распределены одинаково. По этой причине последовательность значений $u(t_j)$, $j = \overline{1,m}$ наблюдаемой реализации $u(t)$ случайной функции $\hat{u}(t)$ может интерпретироваться как однородная выборка $u(t_1)$, $u(t_2)$,…, $u(t_m)$, элементы которой принадлежат одной и той же генеральной совокупности. Теперь задача состоит в определении потребного объёма n этой выборки.

По методикам, изложенным в §§ 5.1 – 5.3, получим соотношения:

$$n_1 = n_{\beta_1,I_1};$$

$$n_2 = n_{\beta_2,I_2};$$

$$n_3 = n_{\beta_3,I_3} = \max_{\tau\in[0;T]}\{n_{\beta_3,I_3}(\tau)\}.$$

Так как наблюдается всего одна реализация случайной функции, то потребное число n её измерений определяется следующим соотношением:

$$n = \max\{n_1; n_2; n_3\}. \qquad (5.4.23)$$

Поскольку регистрация значений реализации $u(t)$ стационарной случайной функции $\hat{u}(t)$ обычно производится через равные промежутки времени длительностью h, то потребная длительность наблюдения реализации определяется равенством

$$T = nh,$$

где n находится из соотношения (5.4.23).

6. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ В ЗАДАЧАХ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

6.1. Понятие статистической гипотезы. Виды гипотез

Гипотезой принято называть предположение о некоторых свойствах изучаемых явлений. При обработке экспериментальных данных рассматриваются гипотезы о свойствах генеральной совокупности, например о виде закона распределения исследуемого признака, о параметрах закона распределения. Эти гипотезы проверяются путём обработки случайной выборки и в дальнейшем называются ***статистическими***.

Наряду с принятой гипотезой рассматривают и противоречащую ей гипотезу, которая может быть принята в том случае, если первая не подтвердится. Для отличия эти гипотез друг от друга выдвинутую гипотезу принято называть **нулевой** или **основной** и обозначать символом H_0, а гипотезу, противоречащую нулевой, – **конкурирующей** или **альтернативной** и обозначать символом H_1.

Для краткости записи гипотез используют специальное обозначение. Пусть нулевая гипотеза состоит в предположении, что математические ожидания двух нормально распределённых случайных величин $\hat{x}$ и $\hat{y}$ равны, а конкурирующая гипотеза состоит в том, что они не равны. Эти гипотезы записываются следующим образом:

$$H_0 : M_{\hat{x}} = M_{\hat{y}}; \quad H_1 : M_{\hat{x}} \neq M_{\hat{y}}.$$

Гипотезы принято подразделять на простые и сложные. Простой называют гипотезу, содержащую только одно предположение, например $H_0 : M_{\hat{x}} = M_{\hat{y}}$. Сложной называют гипотезу, содержащую конечное или бесконечное число предположений. Например, гипотеза $H_0 : M_{\hat{x}} > 10$ состоит из бесконечного множества простых гипотез вида $H_i : M_{\hat{x}} = b_i$, где b_i – любое число, превосходящее 10.

Задачи проверки гипотез можно разделить на несколько классов, отличающихся друг от друга как по форме, так и по методом решения. Прежде всего, эти задачи делятся на ***параметрические***, когда вид закона распределения известен, и ***непараметрические***, когда закон распределения неизвестен. В свою очередь, каждый из данных классов содержит следующие подклассы.

1. **Задачи согласия**. Данные задачи сводятся к проверке согласия (соответствия) вида закона распределения или значений параметров распределения, выдвинутых в качестве предполагаемых, с законом распределения или параметрами закона распределения исследуемой случайной величины.

Формулировка данных задач имеет следующий вид. Нулевая гипотеза:

$$H_0 : Q_{\hat{x}} = S_{\hat{x}}.$$

Конкурирующие гипотезы:

$$H_1^{(1)} : Q_{\hat{x}} < S_{\hat{x}}; \qquad H_1^{(2)} : Q_{\hat{x}} > S_{\hat{x}}; \qquad H_1^{(3)} : Q_{\hat{x}} \neq S_{\hat{x}}.$$

Здесь $Q_{\hat{x}}$ и $S_{\hat{x}}$ – символы сравниваемых случайных объектов. Так, если речь идёт о проверке согласия закона распределения, то $Q_{\hat{x}} = F_{\hat{x}}$, $S_{\hat{x}} = F_{\hat{x}\text{г}}$, где $F_{\hat{x}}$ – закон распределения исследуемой случайной величины; $F_{\hat{x}\text{г}}$ – гипотетический закон распределения. Для данного случая получим:

$$H_0 : F_{\hat{x}} = F_{\hat{x}\text{г}}; \qquad H_1 : F_{\hat{x}} \neq F_{\hat{x}\text{г}}.$$

При проверке согласия параметров распределения альтернативные гипотезы могут выдвигаться в форме гипотез, содержащих бесчисленное множество предположений.

2. **Задачи независимости**. Эти задачи возникают в тех случаях, когда необходимо проверить, являются ли компоненты некоторого случайного вектора независимыми. Очевидно, что если компоненты вектора

$$\hat{X}_{<n>} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)^{\mathsf{T}}$$

независимы, то

$$F_{\hat{X}_{<n>}}(X_{<n>}) = \prod_{i=1}^{n} F_{\hat{x}_i}(x_i),$$

где $F_{\hat{X}_{<n>}}(X_{<n>})$ – закон распределения случайного вектора $X_{<n>}$; $F_{\hat{x}_i}(x_i)$, $i = \overline{1,n}$ – закон распределения *i*–го компонента вектора $X_{<n>}$.

Поэтому задачу проверки гипотезы о независимости можно сформулировать следующим образом:

$$H_0 : F_{\hat{X}_{<n>}}(X_{<n>}) = \prod_{i=1}^{n} F_{\hat{x}_i}(x_i);$$

$$H_1 : F_{\hat{X}_{<n>}}(X_{<n>}) \neq \prod_{i=1}^{n} F_{\hat{x}_i}(x_i).$$

3. **Задачи проверки выборки**. Данные задачи появляются в случае необходимости проверки того факта, что полученная выборка является простой, т.е. варианты выборки подчинены одному и тому же закону распределения.

Задачи такого типа формулируются в виде соотношений:

$$H_0 : F_{\hat{X}_{<n>}}(x_1, x_2, .., x_n) = \prod_{i=1}^{n} F_{\hat{x}}(x_i);$$

$$H_1:\ F_{\hat{X}_{<n>}}(x_1,\, x_2,..,x_n) \neq \prod_{i=1}^{n} F_{\hat{x}}(x_i)\,.$$

6.2. Общий подход к проверке гипотез

Подход к решению задачи проверки гипотез рассмотрим на следующих двух примерах.

П р и м е р 6.1. На склад готовой продукции микросхемы одного типа поступают с двух заводов, выпускающих продукцию разного качества, и такими же партиями микросхемы отпускаются со склада потребителю. Качество продукции заводов характеризуется вероятностью p того, что случайным образом выбранная микросхема является бракованной. Для одного завода $p = p_0$, для другого $p = p_1$ ($p_0 < p_1$). Потребитель произвольно выбирает одну партию микросхем. Необходимо на основании результатов контроля решить, на каком заводе изготовлена выбранная партия микросхем.

▼ Введём нулевую гипотезу H_0, состоящую в том, что выбранная партия микросхем изготовлена на одном заводе (вероятность брака равна p_0), и конкурирующую гипотезу H_1 о том, что партия микросхем изготовлена на другом заводе (вероятность брака равна p_1).

Отберём из партии случайным образом n изделий. Обозначим число бракованных микросхем среди отобранных символом $\hat{x}$. Очевидно, что $\hat{x}$ – дискретная случайная величина, множество значений которой

$$X_{\{n\}} = \{0,\, 1,\, 2,..,n\}.$$

Назовём ***решающим правилом*** или ***критерием проверки гипотезы*** совокупность условий, при которых нулевая гипотеза принимается или отвергается.

В рассматриваемом примере решающее правило будет состоять в некотором разбиении множества $X_{\{n\}}$ на два подмножества X_0 и X_1 ($X_0 \cup X_1 = X,\ \ X_0 \cap X_1 = \varnothing$) таких, что при попадании возможного значения случайной величины $\hat{x}$ в множество X_0 гипотеза H_0 принимается, а в множество X_1 – отвергается. ▲

Разбиение множества X на подмножества X_0 и X_1 можно осуществлять различным образом, поэтому прежде чем решать поставленную задачу, необходимо определить, какое из возможных разбиений множества X на подмножества X_0 и X_1 следует выбрать.

П р и м е р 6.2. На вход приёмного устройства в некоторый момент времени поступает случайный сигнал $\hat{y}$, который представляет сумму известного сигнала x и случайной помехи $\hat{z}$, либо одну помеху $\hat{z}$. Измеряется величина $\hat{y}$, и на основании полученного числового значения y

необходимо установить, присутствовал ли на входе сигнал, т.е. выбрать одну из возможностей: $\hat{y} = x + \hat{z}$ или $\hat{y} = \hat{z}$.

▼ Введём нулевую гипотезу H_0, состоящую в том, что сигнал присутствует ($H_0 : \ \hat{y} = x + \hat{z}$) и конкурирующую гипотезу о том, что сигнал на входе отсутствует ($H_1 : \ \hat{y} = \hat{z}$).

Множество Y возможных значений случайной величины $\hat{y}$ представляет собой всю числовую ось. Решающее правило в данном случае будет состоять в разбиении множества Y на две части Y_0 и Y_1, такие, что при попадании возможного значения случайной величины $\hat{y}$ в множество Y_0 гипотеза H_0 принимается, а при попадании возможного значения $\hat{y}$ в множество Y_1 эта гипотеза отвергается. Как и в предыдущей задаче, необходимо решить вопрос о таком разбиении. ▲

Из приведённых примеров видно, что при наличии способов разбиения множества возможных значений наблюдаемой величины $\hat{x}$ или $\hat{y}$ на подмножества, соответствующие приёму и отклонению гипотезы H_0, общий подход к решению задачи проверки гипотез включает в себя следующие этапы.

1. Выдвигается нулевая и конкурирующая гипотезы.

2. Выбирается некоторая величина, которая представляет собой функцию элементов выборки, связана с нулевой и конкурирующей гипотезами и зависит от условий проведения эксперимента. В дальнейшем эту величину будем называть ***показателем согласованности гипотезы***.

3. Выбирается критерий проверки (критерий согласия, критерий соответствия), т.е. совокупность правил, указывающих, при каких значениях показателя согласованности гипотеза отвергается, а при каких не отвергается.

4. Множество возможных значений показателя согласованности в соответствии с принятым критерием разбивается на два подмножества таким образом, что попадание возможного значения данного показателя в одно из этих подмножеств означает принятие гипотезы H_0, а в другое – отклонение указанной гипотезы.

5. Проводится эксперимент, вычисляется величина показателя и определяется, к какому из подмножеств относится эта величина, на основании чего принимается решение о приёме или отклонении гипотезы H_0.

Из описанного выше общего подхода к решению задачи проверки гипотез следует, что это решение связано с предварительным выбором показателя согласованности и критерия проверки гипотез, которые должны обладать определёнными свойствами.

6.3. Показатель согласованности и его свойства

Показателем согласованности или ***статистической характеристикой*** гипотезы называется случайная величина $\hat{u}$, являющаяся функцией гипотетических данных и результатов наблюдений, предназначенная для проверки нулевой гипотезы.

Конкретный вид показателя согласованности для различных гипотез может быть различным. Так, при проверке гипотезы о законе распределения показатель согласованности может задаваться следующими способами:

– в виде зависимости от гипотетической функции распределения $F_{\hat{x}\text{г}}(x)$, т.е. функции распределения, выдвинутой в качестве нулевой гипотезы, и статистической функции распределения $F^*_{\hat{x}}(x)$, полученной экспериментально:

$$\hat{u} = f_1\left(F_{\hat{x}\text{г}}(x);\, F^*_{\hat{x}}(x)\right); \qquad (6.3.1)$$

– в виде зависимости от гипотетической вероятности p и частоты p^*, полученной в результате проведения эксперимента:

$$\hat{u} = f_2\left(p;\, p^*\right).$$

При проверке гипотезы о равенстве математических ожиданий двух независимых случайных величин $\hat{x}$ и $\hat{y}$ показатель согласованности может выбираться в виде различного рода зависимостей от начальных и центральных моментов первого и второго порядков от случайных величин $\hat{x}$ и $\hat{y}$:

$$\hat{u} = f_3\left(\nu^*_1(\hat{x});\, \nu^*_2(\hat{y});\, \mu^*_2(\hat{x});\, \mu_2(\hat{y})\right).$$

Применяются также и другие виды зависимостей. Однако, несмотря на такое разнообразие, в любом случае показатель согласованности должен удовлетворять ряду требований. Поскольку это величина случайная, то и требования формулируются применительно к закону распределения показателя согласованности. Состоят они в следующем.

1. Показатель согласованности должен определяться нулевой и конкурирующей гипотезами, а также условиями проведения эксперимента. Так, в показателе согласованности, определяемом выражением (6.3.1), эта зависимость представлена наличием как гипотетической, так и статистической функций распределения в качестве аргументов функции f_1.

2. Показатель согласованности должен представлять собой случайную величину, точное или приближённое распределение которой известно. В настоящее время наиболее распространён выбор показателей согласованности, распределённых по нормальному закону, законам хи-квадрат, Стьюдента, Фишера. Причём показатели согласованности с различными законами распределения обозначаются разными символами.

Так, показатели, распределённые по нормальному закону, обозначают через u или z, по закону хи-квадрат – через χ^2, по закону Стьюдента – через t, по закону Фишера – через F.

3. Закон распределения показателя согласованности должен быть инвариантен к виду закона распределения исследуемой случайной величины. Именно данное обстоятельство и определило широкое распространение показателей согласованности, имеющих указанные выше законы распределения.

4. Для построения закона распределения показателя согласованности должен быть востребован минимум априорных сведений, так как возможность получения достоверных сведений до опыта существенно ограничена.

5. Закон распределения показателя согласованности должен быть критичен по отношению к проверяемой гипотезе. Указанное требование означает, что условные плотности распределения $\varphi_{\hat{u}/H_0}(u)$ и $\varphi_{\hat{u}/H_1}(u)$ должны существенно отличаться друг от друга.

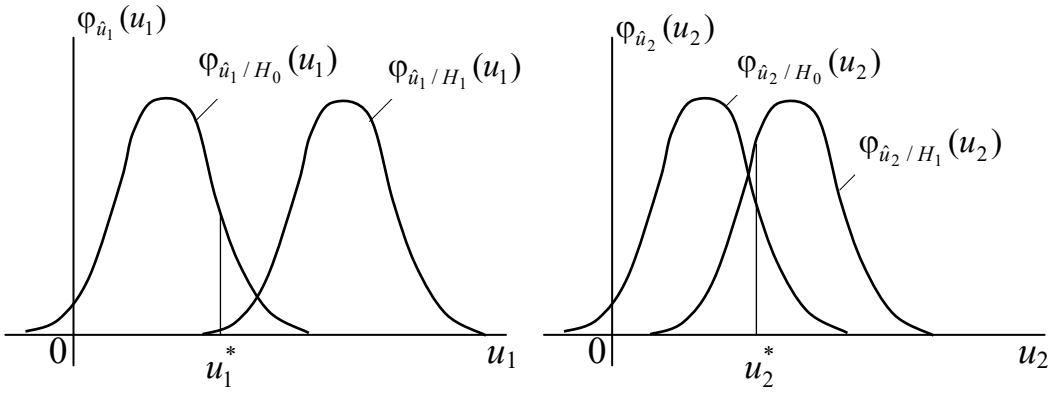


Рис.6.1. Условные плотности распределения показателей согласованности

На рис.6.1 изображены кривые условных плотностей распределения двух различных показателей согласованности $\hat{u}_1$ и $\hat{u}_2$ при нулевой и конкурирующей гипотезах. Из сравнения кривых видно, что применение показателя $\hat{u}_1$ предпочтительнее, так как он обеспечивает более высокую степень уверенности различения гипотез H_0 и H_1, чем показатель $\hat{u}_2$. Действительно, при одном и том же значении $u = u_1^* = u_2^*$, т.е. при наступлении одного и того же события $\hat{u} = u$, вероятность отнесения его к нулевой гипотезе значительно выше, когда используется показатель $\hat{u}_1$.

В заключение следует отметить, что для проверки гипотезы по данным выборки вычисляют частные значения входящих в показатель согласованности величин и, таким образом, получают частное значение показателя согласованности гипотезы. Это значение, вычисленное по

данным выборки, в дальнейшем будем называть ***наблюдаемым значением*** показателя согласованности $\hat{u}$ и обозначать через u.

6.4. Методы задания критической области

Как отмечалось выше, проверка гипотез требует задания решающего правила, т.е. метода разбиения множества U возможных значений показателя согласованности $\hat{u}$ на два подмножества: подмножество U_0, при попадании в которое наблюдаемого значения u показателя $\hat{u}$ нулевая гипотеза принимается, и подмножество U_1, при попадании в которое наблюдаемого значения u нулевая гипотеза отвергается. В дальнейшем область, соответствующую U_0, будем называть ***областью допустимых значений*** (областью принятия гипотезы H_0) и обозначать символом D, а область, соответствующую подмножеству U_1 – ***критической областью*** показателя $\hat{u}$ и обозначать символом Q.

Поскольку $\hat{u}$ – одномерная случайная величина, то все её возможные значения принадлежат некоторому интервалу. Поэтому области Q и D являются интервалами, следовательно, существуют разделяющие их точки. Они называются ***критическими точками*** (***границами***). Задание критической области и сводится к заданию критических точек.

Сущность задания критических точек состоит в следующем. Рассмотрим события:

A – верна гипотеза H_0;

$\overline{A}$ – верна гипотеза H_1;

B – наблюдаемое значение u показателя согласованности $\hat{u}$ попало в область D;

$\overline{B}$ – наблюдаемое значение u попало в область Q.

Тогда в процессе принятия решения возможен один из следующих исходов:

$A \cap B$ – верна гипотеза H_0 и принято решение о её справедливости;

$A \cap \overline{B}$ – верна гипотеза H_0, а принято решение о справедливости гипотезы H_1;

$\overline{A} \cap B$ – верна гипотеза H_1, а принято решение о справедливости гипотезы H_0;

$\overline{A} \cap \overline{B}$ – верна гипотеза H_1 и принято решение о её справедливости.

Очевидно, что исходы $A \cap \overline{B}$ и $\overline{A} \cap B$ являются ошибочными, первому из них соответствует ошибка первого рода, а второму – ошибка второго рода.

Таким образом, под ошибкой первого рода понимается принятие решения об отклонении нулевой гипотезы в случае, если в действительности она является правильной, а под ошибкой второго рода – решение о принятии нулевой гипотезы, если в действительности она не верна.

Поскольку рассмотренные события являются случайными, то им могут быть поставлены в соответствие вероятности наступления данных событий, а именно:

p_{11} – вероятность наступления события $A \cap B$;

p_{12} – вероятность наступления события $A \cap \overline{B}$;

p_{21} – вероятность наступления события $\overline{A} \cap B$;

p_{22} – вероятность наступления события $\overline{A} \cap \overline{B}$.

Значения p_{11}, p_{12}, p_{21}, p_{22} можно вычислить как вероятности попадания случайной величины $\hat{u}$ в области D и Q следующим образом.

Пусть законы распределения показателя согласованности $\hat{u}$ при условии, что справедлива нулевая или конкурирующая гипотеза заданы в форме плотности распределения $\varphi_{\hat{u}/H_0}(u)$ и $\varphi_{\hat{u}/H_1}(u)$, а границей данных областей является точка u^*. Предположим, что взаимное расположение кривых распределения $\varphi_{\hat{u}/H_0}(u)$ и $\varphi_{\hat{u}/H_1}(u)$, а также областей D и Q имеет вид, изображённый на рис.6.2.

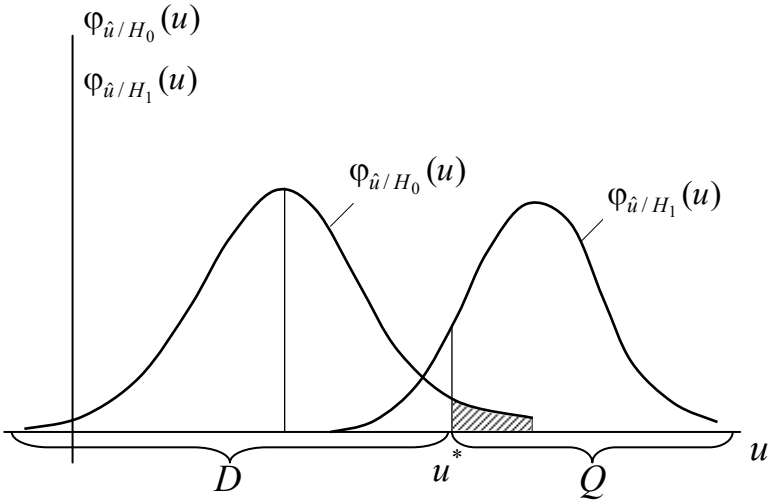


Рис.6.2. Кривые распределения показателя согласованности при различных гипотезах

В этом случае вероятности p_{11}, p_{12}, p_{21}, p_{22} определяются следующими соотношениями:

$$p_{11} = P(A/B) = \int_D \varphi_{\hat{u}/H_0}(u)du; \qquad (6.4.1)$$

$$p_{12} = P(A/\overline{B}) = \int_Q \varphi_{\hat{u}/H_0}(u)du; \qquad (6.4.2)$$

$$p_{21} = P(\overline{A}/B) = \int_D \varphi_{\hat{u}/H_1}(u)du; \qquad (6.4.3)$$

$$p_{22} = P(\overline{A}/\overline{B}) = \int_Q \varphi_{\hat{u}/H_1}(u)du. \qquad (6.4.4)$$

Из выражений (6.4.1) – (6.4.4) следует, что значения p_{11}, p_{12}, p_{21}, p_{22} зависят от размеров и расположения области D допустимых значений и критической области Q. Поэтому, предъявляя соответствующие требования к вероятностям p_{11}, p_{12}, p_{21}, p_{22}, можно определить расположение и размеры данных областей, т.е. критические границы.

Так как проверка гипотезы связана с принятием решения о справедливости или несправедливости выдвинутой гипотезы, то при нахождении областей D и Q целесообразно опираться на результаты теории статистических решений (см. § 2.3). Практическое применение данных результатов зависит от объёма априорной информации, которая может быть использована при проверке гипотез. В связи с этим методы задания критической области принято делить на две группы [5]:

– методы, опирающиеся на оценки потерь от неправильного решения;

– методы, опирающиеся на оценки вероятностей ошибок при принятии решений.

Наибольшее практическое распространение получили методы второй группы, так как их применение требует минимальной априорной информации при проверке гипотез, однако, достоверность принятия решений с помощью данных методов ниже, чем для методов первой группы.

6.5. Проверка гипотез как задача принятия решений

Чтобы формализовать задачу проверки статистических гипотез в виде задачи принятия решения, опишем эту задачу в терминах теории статистических решений.

В качестве объекта наблюдения здесь выступает гипотеза H. Будем полагать, что два возможных варианта данной гипотезы – нулевая H_0 и конкурирующая H_1 представляют собой простые гипотезы.

В качестве статистической характеристики гипотезы используется показатель согласованности $\hat{u}$, являющийся некоторой функцией результатов наблюдения.

Множество решений включает в себя: решение $\widetilde{E}_1$, состоящее в принятии гипотезы H_0, и решение $\widetilde{E}_2$, состоящее в отклонении гипотезы H_0 (т.е. принятии гипотезы H_1).

Объём априорной информации в процессе принятия гипотез может быть различен. Так, при минимальной неопределённости она включает в себя (см. § 2.3):

– вероятности наступления гипотез H_0 и H_1, данные вероятности запишем как $P(H = H_0)$ и $P(H = H_1)$;

– законы распределения показателя согласованности $\hat{u}$ при условии справедливости нулевой и конкурирующей гипотез, т.е. условные плотности распределения $\varphi_{\hat{u}/H_0}(u)$ и $\varphi_{\hat{u}/H_1}(u)$;

– функцию потерь π, задаваемую в виде матрицы потерь:

$$\pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix}.$$

Решающее правило для данного случая состоит в разделении множества возможных значений показателя согласованности на два подмножества D и Q. Попадание наблюдаемого значения в первое из них означает принятие решения о справедливости гипотезы H_0, а второе - принятие решения об отклонении H_0. Таким образом, решающее правило определяет выбор критической границы или границ, если их несколько, и, таким образом задаёт критическую область Q.

Решающее правило может быть сформулировано на основе принципов принятия статистических решений. При проверке гипотез используется четыре вида правил, причём применение того или иного вида зависит от полноты априорных данных.

Если задача проверки гипотез сформулирована как задача выбора решений и матрица потерь π определена, то оптимальное решение может быть получено на основе байесовского или минимаксного правила.

Если функция потерь π не определена, то для однозначного выбора решения при проверке гипотез можно использовать два подхода. Применительно к задачам с известным априорным распределением гипотез наиболее полной характеристикой степени соответствия каждой из них результатам произведённого испытания является апостериорная вероятность этой гипотезы. При этом истинной считается апостериорная вероятность. Указанное правило называется ***правилом апостериорной вероятности.***

При отсутствии данных об априорном распределении гипотез единственной характеристикой степени соответствия той или иной гипотезы результатам наблюдения является функция правдоподобия.

Поэтому в таких случаях выбор решений производится на основе ***правила максимума правдоподобия***, в соответствии с которым истинной считается гипотеза с наибольшим значением функции правдоподобия.

6.6. Проверка гипотез классическим методом

Применение рассмотренных выше методов, как уже отмечалось, связано с использованием априорной информации, которая далеко не всегда имеется в распоряжении исследователя при обработке экспериментальных данных. В связи с этим на практике наибольшее распростра-

нение получили методы проверки гипотез, опирающиеся при назначении критических границ только на информацию о вероятностях ошибок первого и второго рода.

Сущность данных методов состоит в задании вероятности p_{12} ошибки первого рода. Зная эту величину и зависимость, связывающую вероятность p_{12} с показателем согласованности $\hat{u}$, можно определить границы и расположение критической области Q.

В связи с тем, что вероятность p_{12} ошибки первого рода играет в данных методах ведущую роль, она имеет специальное название – ***уровень значимости критерия проверки гипотезы*** и специальное обозначение α. Таким образом, в соответствии с выражением (6.4.2)

$$\alpha = p_{12} = P(A/\overline{B}) = \int_{Q} \varphi_{\hat{u}/H_0}(u)du\,. \qquad (6.6.1)$$

Если известны $\varphi_{\hat{u}/H_0}(u)$ и α, то выражение (6.6.1) позволяет определить критические границы, т.е. границы области Q. Данные границы в дальнейшем будем обозначать символами $u_{\alpha 1}$, $u_{\alpha 2}$,…, если область Q состоит из нескольких подобластей, или символом u_α, если она представляет собой одну область, т.е. возможные значения показателя согласованности $\hat{u}$ делятся границей u_α на две полупрямые.

При конкретном выборе критических границ u_α необходимо учитывать два дополнительных обстоятельства, а именно:

– соотношение между условными законами распределения показателя согласованности $\hat{u}$, соответствующими нулевой и альтернативной гипотезам;

– взаимозависимость ошибок первого и второго рода.

Поясним эти обстоятельства более подробно. Соотношение между условными законами распределения показателя согласованности $\hat{u}$, соответствующими нулевой и конкурирующей гипотезам, выражается в виде взаимного расположения их кривых распределения на оси абсцисс. Особенности характеристики $\hat{u}$, а также нулевой и конкурирующей гипотез приводят к тому, что кривые распределения $\varphi_{\hat{u}/H_0}(u)$, $\varphi_{\hat{u}/H_1}(u)$ могут располагаться друг относительно друга тремя различными способами:

– известно, что кривая распределения $\varphi_{\hat{u}/H_1}(u)$ сдвинута относительно $\varphi_{\hat{u}/H_0}(u)$ вправо (рис.6.3,а);

– известно, что кривая распределения $\varphi_{\hat{u}/H_1}(u)$ сдвинута относительно $\varphi_{\hat{u}/H_0}(u)$ влево (рис.6.3,б);

– сдвиг кривой распределения $\varphi_{\hat{u}/H_1}(u)$ неизвестен, т.е. она может быть сдвинута относительно $\varphi_{\hat{u}/H_0}(u)$ как вправо, так и влево (рис.6.3,в).

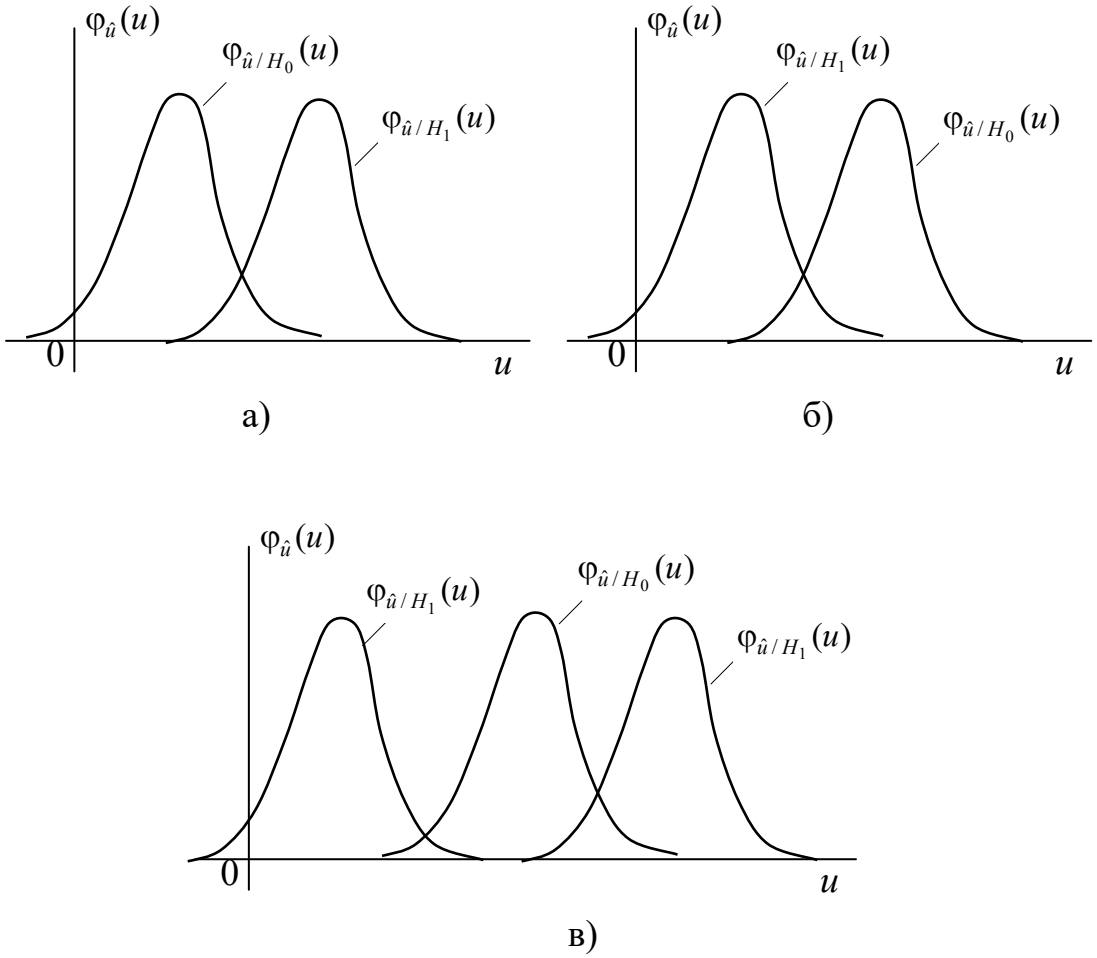


Рис.6.3.Варианты расположения условных законов распределения показателя согласованности

Очевидно, что вид критической области для каждого способа должен быть различен, а именно в первом случае должна быть выбрана правосторонняя критическая область, во втором – левосторонняя и в третьем – двусторонняя. Правосторонней называют критическую область, определяемую неравенством $u > u_\alpha$, левосторонней – неравенством $u < u_\alpha$, двусторонней – неравенствами $u < u_{\alpha 1}$, $u > u_{\alpha 2}$, где $u_{\alpha 2} > u_{\alpha 1}$.

При отыскании критической области достаточно найти критические точки. Методика их отыскания состоит в следующем. Задаются уровнем значимости α и ищут критическую точку u_α, исходя из требования, чтобы при условии справедливости нулевой гипотезы вероятность попадания показателя согласованности $\hat{u}$ в критическую область была равна принятому уровню значимости.

На основании (6.6.1) для правосторонней критической области это условие имеет вид

$$P(\hat{u} \geq u_\alpha) = \int_{u_\alpha}^{\infty} \varphi_{\hat{u}/H_0}(u)du = \alpha,$$

для левосторонней

$$P(\hat{u} < u_\alpha) = \int_{-\infty}^{u_\alpha} \varphi_{\hat{u}/H_0}(u)du = \alpha,$$

для двусторонней

$$P(\hat{u} < u_{\alpha 1}) + P(\hat{u} \geq u_{\alpha 2}) = \int_{-\infty}^{u_\alpha} \varphi_{\hat{u}/H_0}(u)du + \int_{u_\alpha}^{\infty} \varphi_{\hat{u}/H_0}(u)du = \alpha.$$

В последнем случае чаще всего выбирают симметрично расположенные критические точки:

$$P(\hat{u} < u_{\alpha 1}) = P(\hat{u} \geq u_{\alpha 2}) = \frac{\alpha}{2}.$$

Критические точки, удовлетворяющие приведённым выше условиям, находят по соответствующим таблицам (см., например, приложения 5 и 6).

Из вышеизложенного следует, что при выборе критических областей необходимо учитывать не только свойства нулевой, но и свойства конкурирующей гипотезы.

Для пояснения взаимозависимости ошибок первого и второго рода рассмотрим условные плотности $\varphi_{\hat{u}/H_0}(u)$, $\varphi_{\hat{u}/H_1}(u)$ и правостороннюю критическую область с критической границей u_α, см. рис.6.4.

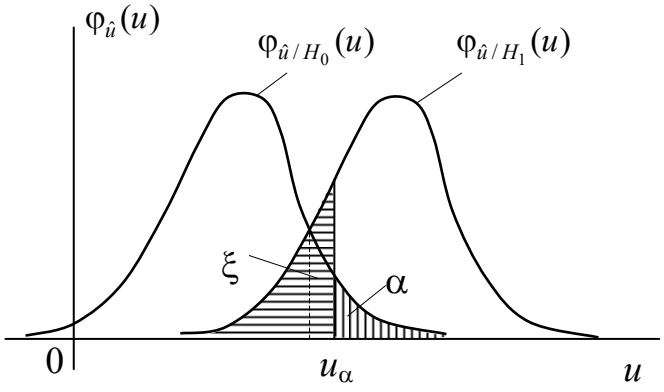


Рис.6.4. Взаимозависимость ошибок первого и второго рода

Следует напомнить, что вероятность ошибки первого рода равняется вероятности α попадания показателя $\hat{u}$ в критическую область. Чем меньше уровень значимости α, тем реже будет допускаться ошибка первого рода, а, следовательно, отвергаться правильная нулевая гипотеза H_0. Однако было бы неверно на основании этого делать вывод о том, что значение вероятности α должен быть выбрано как можно меньшим.

Из рис. 6.4 видно, чем меньше α, тем больше вероятность

$$\xi = \int_{-\infty}^{u_\alpha} \varphi_{\hat{u}/H_1}(u)du \qquad (6.6.2)$$

ошибки второго рода.

Указанная взаимосвязь ошибок первого и второго рода позволяет сделать следующий важный вывод: для уменьшения вероятности ошибки при принятии гипотезы критическую границу необходимо выбирать таким образом, чтобы сумма вероятностей ошибок первого и второго рода была минимальной (см. п.п.2.3.3). Если показатель согласованности подчинён нормальному закону, то минимум суммы вероятностей ошибок первого и второго рода достигается при выборе критической границы в абсциссе точки пересечения кривых распределения $\varphi_{\hat{u}/H_0}(u)$ и $\varphi_{\hat{u}/H_1}(u)$. На рис.6.4. указанное деление показано пунктиром.

Вместе с тем следует заметить, что не во всех случаях подход к выбору u_α с учётом минимума суммы вероятностей ошибок первого и второго рода целесообразен. На практике выбор целесообразной величины α зависит от «тяжести» последствий ошибок первого и второго рода для каждой конкретной задачи. Например, если ошибка первого рода повлечёт большие потери, а второго рода – малые, то целесообразно принять возможно меньшее α. Вернёмся к примеру 6.1 и рассмотрим, какую величину уровня значимости критерия проверки целесообразно выбрать с точки зрения потребителя микросхем. Заметим, что в задачах подобного типа вероятность приёма бракованной партии изделий (в нашем примере – микросхем), т.е. вероятность ошибки первого рода, принято называть *риском потребителя*, а вероятность признать бракованной партию качественных изделий (вероятность ошибки второго рода) – *риском производителя*. С точки зрения потребителя желательно уменьшать вероятность приёма бракованной партии микросхем, т.е. уменьшать вероятность ошибки первого рода, в связи с чем величину α целесообразно выбирать возможно меньшей.

Поскольку вероятность ошибки второго рода также играет важную роль при выборе критической области, то критерий проверки принято характеризовать так называемой мощностью показателя согласованности.

Мощностью γ показателя согласованности называют вероятность попадания показателя согласованности $\hat{u}$ в критическую область при условии, что справедлива конкурирующая гипотеза.

На основе определения можно записать

$$\gamma = \int_{u_\alpha}^{\infty} \varphi_{\hat{u}/H_1}(u)du\,. \qquad (6.6.3)$$

С учётом выражения (6.6.2) равенство (6.6.3) примет вид

$$\gamma = 1 - \int_{-\infty}^{u_\alpha} \varphi_{\hat{u}/H_1}(u)du = 1 - \xi\,.$$

Таким образом, мощность показателя согласованности – вероятность того, что не будет допущена ошибка второго рода. Поэтому для уменьшения ошибки второго рода критическую область необходимо строить так, чтобы мощность показателя согласованности при заданном уровне значимости была максимальной.

Мощность показателя согласованности позволяет обоснованно подойти к выбору односторонних критических областей. Предположим, что в качестве критической выбрана правосторонняя область (рис.6.4), но кривая условной плотности распределения для конкурирующей гипотезы $\varphi_{\hat{u}/H_1}(u)$ смещена относительно кривой $\varphi_{\hat{u}/H_0}(u)$ влево (рис.6.5).

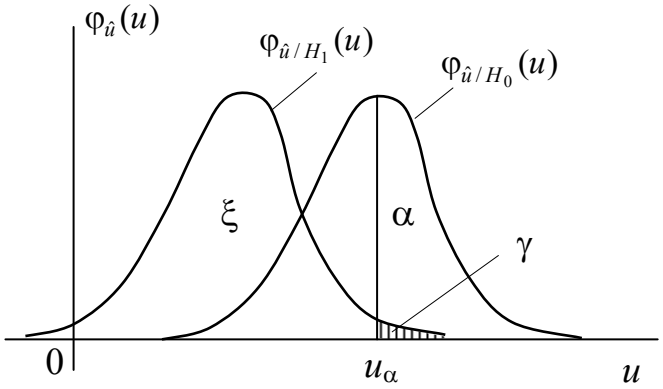


Рис.6.5. Условные распределения показателя согласованности гипотез

Найденные в этих условиях вероятности

$$\xi = \int_{-\infty}^{u_\alpha} \varphi_{\hat{u}/H_1}(u)du \approx 1; \qquad \gamma = 1 - \xi \approx 0$$

показывают, что при таком выборе критической области показатель согласованности $\hat{u}$ становится непригодным для статистической проверки гипотезы H_0, так как его мощность близка к нулю, а ошибки второго рода становятся практически достоверными. Очевидно, что для увеличения мощности показателя согласованности необходимо выбрать левостороннюю критическую область.

Если характер конкурирующей гипотезы неясен, то в качестве критической целесообразно выбирать двустороннюю симметричную область. Но следует отметить, что единственный способ одновременного уменьшения вероятностей ошибок первого и второго рода состоит в увеличении объёма выборки.

6.7. Проверка гипотез об аномальности результатов наблюдений

При обработке экспериментальных данных существенное значение имеет процесс предварительной обработки, одним из этапов которого является исключение результатов, содержащих грубые ошибки, т.е. аномальных результатов.

В любой выборке сомнительными являются, как правило, наибольший и наименьший элементы, которые и подлежат проверке. Обозначим через $\hat{x}_1$ и $\hat{x}_n$ наименьший и наибольший элементы случайной выборки $\hat{x}_1$, $\hat{x}_2$,…, $\hat{x}_n$, а через x_1 и x_n – реализации этих элементов в данном эксперименте.

Предположим, что сомнительным является наибольший элемент x_n случайной выборки. При этом будем полагать, что наблюдаемая величина $\hat{x}$ подчинена нормальному закону распределения с известными числовыми характеристиками $M_{\hat{x}}$ и $\sigma_{\hat{x}}$. Уровень значимости примем равным некоторой достаточно малой вероятности α. Для выборки, состоящей из одного элемента x, можно утверждать, что он является следствием грубой ошибки, если

$$x > M_{\hat{x}} + \sigma_{\hat{x}} t_{1-2\alpha} = x_\alpha \qquad (6.7.1)$$

или

$$\frac{x - M_{\hat{x}}}{\sigma_{\hat{x}}} > t_{1-2\alpha}.$$

Следовательно, при $n = 1$ в качестве показателя согласованности гипотезы целесообразно использовать случайную величину

$$\hat{u} = \frac{\hat{x} - M_{\hat{x}}}{\sigma_{\hat{x}}}, \qquad (6.7.2)$$

а критическую границу определять на основе выражения

$$u_\alpha = t_{1-2\alpha}. \qquad (6.7.3)$$

Однако при проверке аномальности крайнего элемента случайной выборки объёма $n > 1$ использование показателя согласованности вида (6.7.2) может привести к грубым ошибкам. Покажем это на примере.

П р и м е р 6.3. Пусть $n = 1$ и $\alpha = 0{,}05$.

▼ В приложении 4 находим

$$u_\alpha = t_{1-2\alpha} = t_{0,9} = 1{,}64. \qquad (6.7.4)$$

Подставив найденное значение u_α в формулу (6.7.1), получим

$$x > x_\alpha = M_{\hat{x}} + 1{,}64\,\sigma_{\hat{x}}. \qquad (6.7.5)$$

Таким образом, при одном испытании будем констатировать факт грубой ошибки, если наблюдаемое значение случайной величины удовлетворяет неравенству (6.7.5).

Увеличим теперь число испытаний до 20 и найдём вероятность того, что наибольший член $\hat{x}_{20}$ выборки превзойдёт величину x_α:

$$P(\hat{x}_{20} > x_\alpha) = 1 - P(\hat{x}_{20} < x_\alpha) = 1 - P\left(\bigcap_{i=1}^{20}(\hat{x}_i < x_\alpha)\right). \qquad (6.7.6)$$

Предполагая, что испытания независимы, и учитывая, что

$$P(\hat{x}_i < x_\alpha) = 1 - P(\hat{x}_i > x_\alpha) = 1 - \alpha, \quad i = \overline{1,n},$$

из выражения (6.7.6) получим

$$P(\hat{x}_{20} > x_\alpha) = 1 - (1-\alpha)^n = 1 - (0{,}95)^{20} \approx 0{,}65.$$

Таким образом, при использовании критической границы, определяемой выражением (6.7.4), 65% нормальных наибольших элементов выборки следует признать аномальными. Иначе, вероятность ошибки первого рода при 20 испытаниях увеличится до 0,65.

▲

С целью устранения указанного недостатка необходимо по мере увеличения объёма выборки сдвигать критическую границу вправо относительно значения u_α, определяемого равенством (6.7.4).

Для построения критической области, удовлетворяющей указанному требованию, рассмотрим функцию распределения $F_{\hat{x}_n}(x)$. Примем во внимание то обстоятельство, что для наступления события $(\hat{x}_n < x)$ необходимо, чтобы все элементы выборки были меньше x:

$$\hat{x}_n < x = \bigcap_{i=1}^{n}(\hat{x}_i < x). \qquad (6.7.7)$$

Далее учитываем, что

$$P(\hat{x}_i < x) = F_{\hat{x}_i}(x) = F_{\hat{x}}(x) = \Phi_1\left(\frac{x - M_{\hat{x}}}{\sigma_{\hat{x}}}\right). \qquad (6.7.8)$$

К правой части выражения (6.7.7) применяем теорему умножения вероятностей независимых событий $(\hat{x}_i < x)$, $i = \overline{1,n}$ и принимаем во внимание (6.7.8). В результате получаем

$$F_{\hat{x}_n}(x) = P(\hat{x}_n < x) = \prod_{i=1}^{n} P(\hat{x}_i < x) = \left(F_{\hat{x}}(x)\right)^n = \left(\Phi_1\left(\frac{x - M_{\hat{x}}}{\sigma_{\hat{x}}}\right)\right)^n. \qquad (6.7.9)$$

Границу u_α критической области, отвечающей уровню значимости α, можно найти как квантиль случайной величины $\hat{x}_n$ при аргументе $1-\alpha$. Зависимость между u_α и α определяется равенствами:

$$u_\alpha = F_{\hat{x}_n}^{-1}(1-\alpha);$$
$$F_{\hat{x}_n}(u_\alpha) = 1 - \alpha. \qquad (6.7.10)$$

Подставляем $x = u_\alpha$ в формулу (6.7.9) и на основании равенства (6.7.10) получим уравнение

$$\left(\Phi_1\left(\frac{u_\alpha - M_{\hat{x}}}{\sigma_{\hat{x}}}\right)\right)^n = 1-\alpha. \qquad (6.7.11)$$

Решаем (6.7.11) относительно u_α и находим

$$u_\alpha = M_{\hat{x}} + \sigma_{\hat{x}} t_{2(1-\alpha)^{1/n}-1}. \qquad (6.7.12)$$

Так как вероятность α обычно мала, то

$$(1-\alpha)^{\frac{1}{n}} \approx 1-\frac{\alpha}{n}$$

и, следовательно,

$$t_{2(1-\alpha)^{1/n}-1} \approx t_{1-\frac{2\alpha}{n}}. \qquad (6.7.13)$$

Подставляя соотношение (6.7.13) в (6.7.12), получим

$$u_\alpha = M_{\hat{x}} + \sigma_{\hat{x}} t_{1-\frac{2\alpha}{n}}. \qquad (6.7.14)$$

Выражение (6.7.12) и применяется для определения критической границы в случае, если $n > 1$.

Порядок проверки гипотезы об аномальном значении наибольшего элемента выборки состоит в следующем.

1. Элемент, относительно которого выдвигается гипотеза, исключается из выборки, т.е. её объём уменьшается на единицу.

2. Назначается уровень значимости α и по приложению 4 определяется значение $t_{1-\frac{2\alpha}{n}}$.

3. По формуле (6.7.14) определяется критическая граница. При отсутствии априорных значений $M_{\hat{x}}$ и $\sigma_{\hat{x}}$ в данной формуле используются оценки $\widetilde{M}_{\hat{x}}$ и $\widetilde{\sigma}_{\hat{x}}$. При их вычислении предполагаемый аномальный результат из выборки исключается.

4. Наблюдаемое значение показателя согласованности гипотезы определяется по формуле

$$u = \frac{x_n - M_{\hat{x}}}{\sigma_{\hat{x}}}. \qquad (6.7.15)$$

5. Проверяется условие $u > u_\alpha$. Если оно выполняется, то наибольший элемент выборки, по отношению к которому выдвигалось предположение о наличии грубой ошибки, отбрасывается. При $u \le u_\alpha$ этот элемент сохраняется в выборке, поскольку данные эксперимента не подтверждают гипотезы о наличии грубой ошибки. Для подтверждения полученного вывода необходимо повторить проверку по пунктам 1–5, но с включением сомнительного элемента в выборку.

При рассмотрении в качестве аномального наименьшего элемента $\hat{x}_1$ случайной выборки, порядок проверки гипотезы сохраняется, но в ка-

честве показателя согласованности последней используется случайная величина

$$\hat{u} = \frac{M_{\hat{x}} - \hat{x}_1}{\sigma_{\hat{x}}}.$$

При этом проверяемый элемент отбрасывается, если $u > u_\alpha$, и сохраняется в противном случае.

П р и м е р 6.4. С помощью радиодальномера производятся 20 измерений дальности $\hat{z}$ до объекта. Точность радиодальномера характеризуется среднеквадратическим отклонением $\sigma_{\hat{x}} = 50$ м. Имеются ли основания полагать, что наибольшее отклонение

$$\hat{z}_{20} - M_{\hat{z}} = 180 \text{ м},$$

зафиксированное в данной серии наблюдений, содержит грубую ошибку? Уровень значимости критерия проверки гипотезы принять равным 0,05.

▼ По условию задачи $n = 20$, $\alpha = 0{,}05$. В соответствии с выражением (6.7.15) значение показателя согласованности

$$u = \frac{z_{20} - M_{\hat{z}}}{\sigma_{\hat{z}}} = \frac{180}{50} = 3{,}6.$$

Границу критической области находим в приложении 4:

$$u_\alpha = t_{1-\frac{2\alpha}{n}} = t_{1-\frac{0{,}1}{20}} = t_{0{,}995} = 2{,}82.$$

Так как $u > u_\alpha$, то наибольшее отклонение содержит грубую ошибку и его следует из дальнейшего рассмотрения исключить.

▲

7. МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ О ЗАКОНАХ РАСПРЕДЕЛЕНИЯ И ПАРАМЕТРАХ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ

7.1. Проверка гипотез о законах распределения

7.1.1. Выравнивание статистических рядов

При обработке экспериментальных данных одним из основных вопросов является обоснование закона распределения исследуемой случайной величины.

Пусть экспериментальным путём получена случайная выборка $\hat{x}_1$, $\hat{x}_2,\ldots,\hat{x}_n$. В связи с её ограниченностью при обработке статистического материала приходится решать две задачи.

1. Подобрать для полученного статистического ряда теоретическую кривую распределения, выражающую лишь существенные черты статистического материала, но не случайности, связанные с недостаточным объёмом экспериментальных данных (задача выравнивания или сглаживания статистических рядов).

2. Определить, чем объясняются неизбежные расхождения между подобранной теоретической кривой распределения и статистическим распределением: случайными обстоятельствами или тем, что подобранная кривая неудовлетворительно выравнивает данное статистическое распределение (задача проверки гипотезы о законах распределения).

Процедура выравнивания заключается в том, чтобы подобрать теоретическую кривую распределения, с той или иной точки зрения наилучшим образом описывающую данное статистическое распределение.

Задача о наилучшем выравнивании статистических рядов, как и вообще задача о наилучшем аналитическом представлении эмпирических функций, является в значительной мере неопределённой, и решение её зависит от того, что условиться считать «наилучшим». Например, при сглаживании эмпирических зависимостей часто используют метод наименьших квадратов (раздел 8), согласно которому наилучшим приближением к эмпирической зависимости в данном классе функций является такое, при котором сумма квадратов отклонений обращается в минимум. При этом вопрос о том, в каком именно классе функций следует искать наилучшее приближение, решается уже не из математических соображений. Вид функции, выражающей исследуемую зависимость часто известен заранее. Из опыта требуется получить лишь некоторые числен-

ные параметры, входящие в выражение функции. Именно эти параметры подбираются с помощью метода наименьших квадратов.

Аналогично обстоит дело и с задачей выравнивания статистических рядов. Как правило, вид теоретической кривой выбирается заранее из соображений, связанных с существом задачи, а в некоторых случаях просто с внешним видом статистического распределения. Аналитическое выражение выбранной кривой распределения зависит от некоторых параметров. Поэтому задача выравнивания статистического ряда переходит в задачу рационального выбора тех значений параметров, при которых соответствие между статистическим и теоретическим распределениями оказывается наилучшим.

Следует иметь в виду, что любая аналитическая функция $f(x)$, с помощью которой выравнивается статистическое распределение, должна обладать основными свойствами плотности распределения:

$$f(x) \geq 0; \quad \int_{-\infty}^{\infty} f(x)dx = 1. \qquad (7.7.1)$$

Предположим, что исходя из тех или иных соображений выбрана теоретическая кривая распределения $\varphi_{\hat{x}}(x)$, удовлетворяющая условиям (7.1.1). С помощью данной кривой требуется выровнять данное статистическое распределение. В выражение функции $\varphi_{\hat{x}}(x)$ входят параметры $(a_1, a_2,\ldots,a_m)^{\mathrm{T}} = A_{<m>}$, т.е.

$$\varphi_{\hat{x}}(x) = \varphi_{\hat{x}}(x; A_{<m>}). \qquad (7.1.2)$$

Необходимо подобрать эти параметры так, чтобы кривая (7.1.2) наилучшим образом описывала данный статистический материал. Один из методов, применяемых для решения данной задачи – **метод моментов**.

Согласно методу моментов параметры выбираются с таким расчётом, чтобы несколько важнейших числовых характеристик (моментов) теоретического распределения были равны соответствующим статистическим характеристикам. Например, если теоретическая кривая зависит только от двух параметров:

$$\varphi_{\hat{x}}(x) = \varphi_{\hat{x}}(x; A_{<2>}),$$

эти параметры выбираются так, чтобы математическое ожидание $M_{\hat{x}}$ и дисперсия $D_{\hat{x}}$ теоретического распределения совпадали с соответствующими статистическими характеристиками $M^*_{\hat{x}}$ и $D^*_{\hat{x}}$. Если кривая $\varphi_{\hat{x}}(x)$ зависит от трёх параметров, можно подобрать их так, чтобы совпали первые три момента, и т.д.

П р и м е р 7.1. Произведено 500 измерений отклонения по вертикали при стрельбе в мишень. Результаты измерений сведены в статистический ряд, табл.7.1. Требуется выровнять данное распределение с помощью нормального закона.

Таблица 7.1

Интервальный статистический ряд (к примеру 7.1)

J_l	–4; –3	–3; –2	–2; –1	–1; 0	0; 1	1; 2	2; 3	3; 4
m_l	6	25	72	133	120	88	46	10
P_l^*	0,012	0,050	0,144	0,266	0,240	0,176	0,092	0,020

▼ Нормальный закон распределения

$$\varphi_{\hat{x}}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

зависит от двух параметров: $M_{\hat{x}} = m$ и $\sigma_{\hat{x}} = \sigma$. Подберём эти параметры так, чтобы сохранить первые два момента – математическое ожидание и дисперсию статистического распределения. Оценку математического ожидания вычисляем по формуле (5.1.4):

$$\widetilde{M}_{\hat{x}} = \sum_{l=1}^{8} \overline{x}_l P_l^* = -3{,}5 \cdot 0{,}012 - 2{,}5 \cdot 0{,}050 - 1{,}5 \cdot 0{,}144 - 0{,}5 \cdot 0{,}266 + 0{,}5 \cdot 240 +$$

$$+ 1{,}5 \cdot 0{,}176 + 2{,}5 \cdot 0{,}092 + 3{,}5 \cdot 0{,}020 = 0{,}168.$$

Оценку дисперсии вычисляем по второй формуле (5.2.14). Для этого находим оценку второго начального момента

$$\widetilde{\nu}_2 = \sum_{l=1}^{8} \overline{x}_l^2 P_l^* = 2{,}126.$$

В итоге получаем

$$\widetilde{D}_{\hat{x}} = \sum_{l=1}^{8} \overline{x}_l^2 P_l^* - \widetilde{M}_{\hat{x}}^2 = 2{,}126 - 0{,}028 = 2{,}098.$$

В соответствии с методом моментов должны выполняться условия

$$m = \widetilde{M}_{\hat{x}}\,, \quad \sigma^2 = \widetilde{D}_{\hat{x}}\,.$$

Это означает, что

$$m = 0{,}168, \qquad \sigma = \sqrt{2{,}098} = 1{,}448.$$

Выражение нормального закона распределения принимает вид

$$\varphi_{\hat{x}}(x) = \frac{1}{1{,}448\sqrt{2\pi}} e^{-\frac{(x-0{,}168)^2}{2 \cdot 1{,}448^2}}\,. \qquad (7.1.3)$$

Вычисляем значения функции (7.1.3) на границах разрядов, результаты сводим в табл.7.2.

Таблица 7.2

Значения плотности распределения нормального закона (к примеру 7.1)

x	–4	–3	–2	–1	0	1	2	3	4
$\varphi_{\hat{x}}(x)$	0,004	0,025	0,090	0,199	0,274	0,234	0,124	0,041	0,008

На одном графике (рис.7.1) строим гистограмму и выравнивающую её кривую распределения.

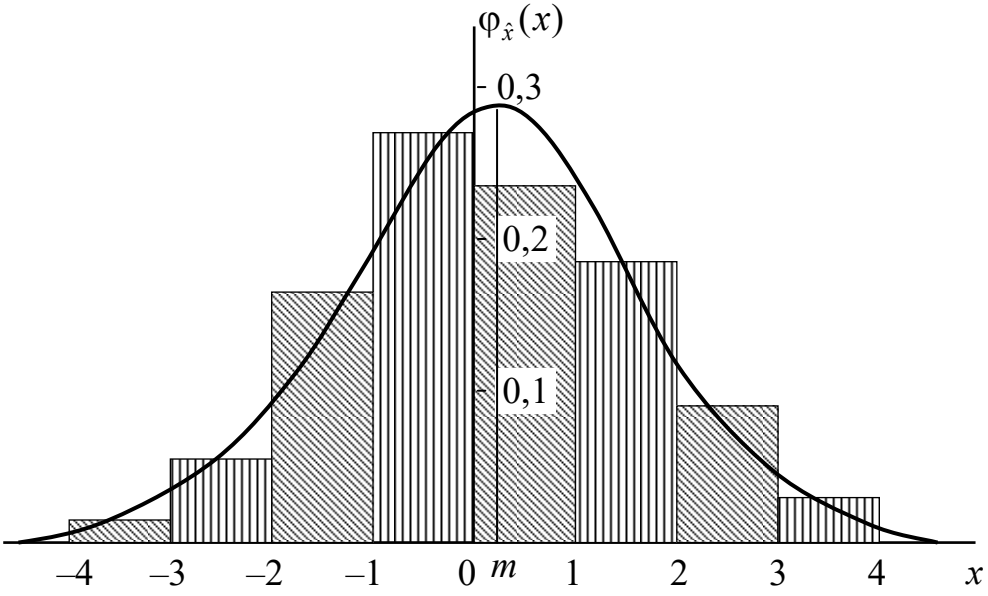


Рис.7.1. Гистограмма и теоретическая кривая распределения (к примеру 7.1)

Из графика видно, что теоретическая кривая распределения $\varphi_{\hat{x}}(x)$ сохраняет в основном существенные особенности статистического распределения. Но при этом она свободна от случайных неправильностей хода гистограммы, которые, по-видимому, могут быть отнесены за счёт случайных причин. Для более серьёзного обоснования последнего суждения необходимо выполнить проверку гипотезы о принятом законе распределения. ▲

Выравнивание статистического ряда теоретической кривой распределения может рассматриваться как выдвижение нулевой гипотезы о виде распределения. Но эта задача может быть решена и аналитически.

7.1.2. Выбор нулевой гипотезы аналитическим способом

Для аналитического выбора нулевой гипотезы может быть использована следующая методика. По данным эксперимента определяются статистические оценки коэффициента асимметрии $\widetilde{a}_{\hat{x}}$ и коэффициента эксцесса $\widetilde{e}_{\hat{x}}$:

$$\widetilde{a}_{\hat{x}} = \frac{\widetilde{\mu}_3}{\widetilde{\sigma}_{\hat{x}}^3}; \qquad \widetilde{e}_{\hat{x}} = \frac{\widetilde{\mu}_4}{\widetilde{\sigma}_{\hat{x}}^4} - 3, \tag{7.1.4}$$

где $$\widetilde{\sigma}_{\hat{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^2}{n-1}} = \sqrt{\frac{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}{n-1}};$$

$$\widetilde{\mu}_3 = \frac{\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^3}{n}; \qquad \widetilde{\mu}_4 = \frac{\sum_{i=1}^{n}(x_i - \widetilde{M}_{\hat{x}})^4}{n}.$$

При большом объёме выборки оценки центральных моментов третьего и четвёртого порядков могут вычисляться по формулам, аналогичным (5.2.11) для дисперсии:

$$\tilde{\mu}_3 = \sum_{l=1}^{r} (\bar{x}_l - \tilde{M}_{\hat{x}})^3 P_l^* ; \qquad \tilde{\mu}_4 = \sum_{l=1}^{r} (\bar{x}_l - \tilde{M}_{\hat{x}})^4 P_l^* . \qquad (7.1.5)$$

В теории распределений [12] доказано, что каждому закону свойственно определённое соотношение между коэффициентами асимметрии и эксцесса, т.е. может быть построена диаграмма, изображённая на рис.7.2.

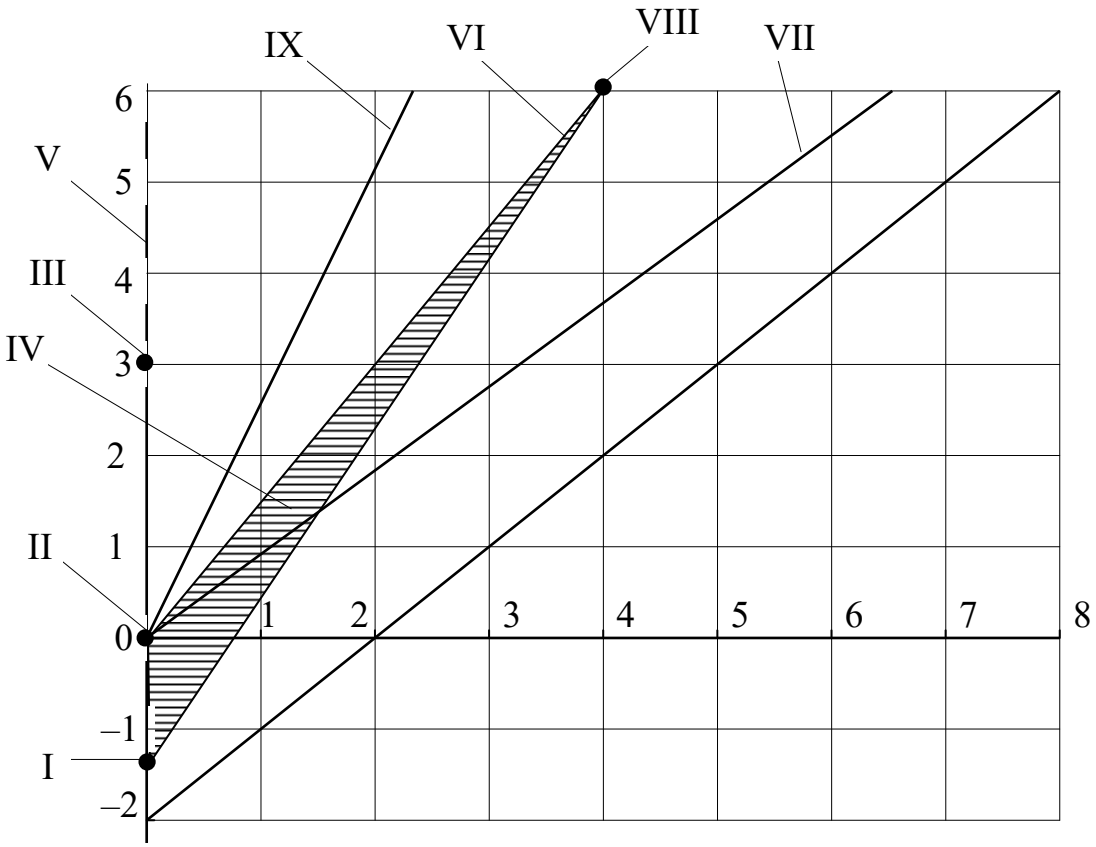


Рис.7.2. Диаграмма соотношений между коэффициентами асимметрии и эксцесса

На представленной диаграмме выделены следующие характерные точки, прямые и области. Точки (0; –1,2), (0; 0), (0; 3), (4; 6) отвечают соответственно равномерному и нормальному распределениям, распределению Лапласа и показательному распределению. Так, для любого нормального закона $a_{\hat{x}} = 0$, $e_{\hat{x}} = 0$, что и определяет координаты точки (0; 0). Гамма-распределение, логарифмически нормальное распределение, распределение Стьюдента и Пуассона показаны на диаграмме прямыми, а бета-распределение представлено областью. При этом обозначения следующие: I – равномерный закон, II – нормальный закон; III – закон Лапласа; IV – бета-распределение; V – закон Стьюдента (прямая, совпа-

дающая с осью ординат); VI – гамма-распределение; VII – закон Пуассона; VIII – показательный закон; IX – логарифмически нормальное распределение.

Знание оценок коэффициентов асимметрии и эксцесса позволяет приближённо определить гипотетический закон распределения. Для этого по полученным значениям оценок на диаграмму наносится точка $(\widetilde{a}_{\hat{x}}^2;\widetilde{e}_{\hat{x}})$. Если она окажется вблизи точки, прямой или области, соответствующих одному из распределений, то последнее и следует выдвинуть в качестве гипотезы.

При попадании точки в области диаграммы, для которых не определён закон распределения, выдвижение гипотетического закона должно осуществляться на основании каких-либо дополнительных априорных соображений.

П р и м е р 7.2. В условиях примера 7.1 выбрать нулевую гипотезу аналитическим способом.

▼ По формулам (7.1.5) вычисляем оценки центральных моментов третьего и четвёртого порядков:

$$\widetilde{\mu}_3 = \sum_{l=1}^{8}(\overline{x}_l - \widetilde{M}_{\hat{x}})^3 P_l^* = (-3{,}5-0{,}168)^3\,0{,}012 + (-2{,}5-0{,}168^3 0{,}050 +$$

$$+(-1{,}5-0{,}168)^3\,0{,}144 + (-0{,}5-0{,}168)^3\,0{,}266 + (0{,}5-0{,}168)^3\,0{,}24 + \quad ;$$

$$+(1{,}5-0{,}168)^3\,0{,}176 + (2{,}5-0{,}168)^3\,0{,}092 + (3{,}5-0{,}168)^3\,0{,}02 =$$

$$= -0{,}592 - 0{,}950 - 0{,}668 - 0{,}079 + 0{,}009 + 0{,}416 + 1{,}167 + 0{,}740 = 0{,}043;$$

$$\widetilde{\mu}_4 = \sum_{l=1}^{8}(\overline{x}_l - \widetilde{M}_{\hat{x}})^4 P_l^* = 11{,}64\,.$$

По формулам (7.1.4) вычисляем оценки коэффициента асимметрии и коэффициента эксцесса:

$$\widetilde{a}_{\hat{x}} = \frac{\widetilde{\mu}_3}{\widetilde{\sigma}_{\hat{x}}^3} = \frac{0{,}043}{(1{,}448)^3} = 0{,}014\,;$$

$$\widetilde{e}_{\hat{x}} = \frac{\widetilde{\mu}_4}{\widetilde{\sigma}_{\hat{x}}^4} - 3 = \frac{11{,}64}{(1{,}448)^3} - 3 = -0{,}353\,.$$

Точку $(\widetilde{a}_{\hat{x}}^2;\ \widetilde{e}_{\hat{x}}) = (0{,}0001;-0{,}353)$ наносим на диаграмму, рис.7.2. Данная точка находится в непосредственной близости от точки (0; 0). Следовательно, принимается нулевая гипотеза о нормальном распределении отклонений по вертикали при стрельбе в мишень.

▲

Проверка гипотезы о виде закона распределения выполняется после решения предыдущей задачи, т.е. выбора теоретического распределения.

7.1.3. Проверка гипотез о законах распределения по методу К.Пирсона

Задача проверки гипотезы о виде закона распределения формулируется следующим образом.

Пусть в результате эксперимента получена случайная выборка $\hat{x}_1$, $\hat{x}_2,\ldots,\hat{x}_n$ и для неё выбран теоретический закон распределения, характеризуемый функцией распределения $F_{\hat{x}}(x)$ или плотностью распределения $\varphi_{\hat{x}}(x)$.

Необходимо на основании обработки и анализа полученной выборки проверить гипотезу H_0 о том, что исследуемая случайная величина подчинена выбранному закону распределения.

В настоящее время существует ряд методов решения данной задачи, однако наибольшее распространение получил метод К. Пирсона. Достаточно употребляемыми являются также методы А.Н. Колмогорова и Н.В. Смирнова [4, 6, 12]. Указанные методы отличаются друг от друга видом меры рассогласования между статистическим и гипотетическим законами распределения. Так, в методах А.Н. Колмогорова и Н.В. Смирнова такой мерой является функция разности между статистической функцией распределения $F^*_{\hat{x}}(x)$ и функцией распределения $F_{\hat{x}}(x)$ гипотетического закона:

$$d = f(F^*_{\hat{x}}(x) - F_{\hat{x}}(x)).$$

В методе К. Пирсона в качестве таковой используется функция разности между частотой и вероятностью попадания случайной величины в заданные интервалы:

$$d = f(p^*_j - p_j), \qquad (7.1.6)$$

где j – номер интервала.

Рассмотрим метод К. Пирсона более подробно. Мера расхождения (7.1.6) в явном виде представляется суммой квадратов разностей между частотой и вероятностью попадания случайной величины $\hat{x}$ в интервалы, на которые разбивается множество возможных значений этой величины:

$$\hat{u} = \sum_{l=1}^{r} c_l (p^*_l - p_l)^2, \qquad (7.1.7)$$

где r – число интервалов; l – номер интервала.

Коэффициенты c_l введены в выражение (7.1.7) для учёта того, что абсолютные значения разностей $p^*_l - p_l$ неравнозначны при различных значениях p_l. Действительно, одно и то же значение разности $p^*_l - p_l$ является малозначимым при большой величине p_l и представляет собой заметную величину, если вероятность p_l мала.

К. Пирсон показал, что коэффициенты c_l целесообразно брать обратно пропорциональными вероятностям p_l. При этом, если данные коэффициенты определять на основе выражения

$$c_l = \frac{n}{p_l}, \quad l = \overline{1,r},$$

то при больших значениях n закон распределения случайной величины

$$\hat{u} = \sum_{l=1}^{r} \frac{n(p_l^* - p_l)^2}{p_l} \tag{7.1.8}$$

не зависит от вида распределения случайной величины $\hat{x}$ и объёма выборки n, а зависит только от числа интервалов r. Кроме этого при увеличении n закон распределения случайной величины (7.1.8) приближается к распределению хи-квадрат [6].

Докажем это утверждение.

Рассмотрим случайную величину $\hat{m}_l$ – число попаданий случайной величины $\hat{x}$ в l-й интервал. Эта случайная величина распределена по биномиальному закону с характеристиками

$$M_{\hat{m}_l} = np_l, \quad \sigma_{\hat{m}_l} = \sqrt{np_l(1-p_l)}.$$

Однако при достаточно большом n величину $\hat{m}_l$ на основании теоремы Муавра-Лапласа можно считать распределённой по нормальному закону с теми же характеристиками. Выполняя нормирование случайной величины $\hat{m}_l$, получим

$$\hat{z}_l = \frac{\hat{m}_l - np_l}{\sqrt{np_l(1-p_l)}}.$$

Нормированные случайные величины $\hat{z}_l$ связаны между собой линейным соотношением

$$\sum_{l=1}^{r} \hat{z}_l \sqrt{np_l(1-p_l)} = \sum_{l=1}^{r} \hat{m}_l - n\sum_{l=1}^{r} p_l = n - n = 0.$$

На основании этого утверждаем, что случайная величина $\sum_{l=1}^{r} z_l^2$ будет приближённо следовать хи-квадрат (χ^2) распределению. Если эту случайную величину принять за показатель согласованности гипотезы, то получим равенство

$$\hat{u} = \hat{\chi}^2 = \sum_{l=1}^{r} \frac{(\hat{m}_l - np_l)^2}{np_l(1-p_l)}. \tag{7.1.9}$$

Преобразуем выражение (7.1.9), учитывая, что

$$\sum_{l=1}^{r} \hat{m}_l = n$$

и $1 - p_l \approx 1$ при больших значениях n:

$$\hat{u} = \sum_{l=1}^{r} \frac{n^2\left(\frac{\hat{m}_l}{n} - p_l\right)^2}{np_l(1-p_l)} = \sum_{l=1}^{r} \frac{n(p_l^* - p_l)^2}{p_l} \qquad (7.1.10)$$

или

$$\hat{u} = \sum_{l=1}^{r} \frac{\hat{m}_l^2 + 2\hat{m}_l np_l + n^2 p_l^2}{np_l} = \sum_{l=1}^{r} \frac{\hat{m}_l^2}{np_l} - 2\sum_{l=1}^{r} \hat{m}_l + n\sum_{l=1}^{r} p_l = \\ = \sum_{l=1}^{r} \frac{\hat{m}_l^2}{np_l} - 2n + n = \sum_{l=1}^{r} \frac{\hat{m}_l^2}{np_l} - n. \qquad (7.1.11)$$

Выражения (7.1.10) или (7.1.11) используются в зависимости от формы представления результатов наблюдения, т.е. в зависимости от того, являются ли исходными данными p_l^* или m_l.

Как известно, распределение χ^2 зависит от числа степеней свободы $f = r - s$, равного числу интервалов r минус число независимых условий (связей), наложенных на частоты p_l^*. В формуле (7.1.10) предполагается наличие только одного условия

$$\sum_{l=1}^{r} p_l^* = 1, \qquad (7.1.12)$$

которое накладывается всегда. Тогда принимаем $s = 1$ и число степеней свободы $f = r - 1$. Равенство (7.1.12) есть сумма вероятностей несовместных событий, образующих полную группу.

В случае, когда теоретическое распределение подбирается так, чтобы совпадали его математическое ожидание и оценка математического ожидания, полученная по результатам наблюдения, т.е.

$$\sum_{l=1}^{r} \bar{x}_l p_l^* = M_{\hat{x}},$$

число связей увеличивается на единицу. Следовательно, $s = 2$ и число степеней свободы $f = r - 2$. Если условие совпадения параметров теоретического и статистического распределения распространяется и на дисперсию

$$\sum_{l=1}^{r} (\bar{x}_l - M_{\hat{x}})^2 p_l^* = D_{\hat{x}},$$

то $s = 3$, $f = r - 3$ и т.д.

Таким образом, число степеней свободы распределения χ^2 зависит при проверке гипотез от условий проведения проверки, что необходимо учитывать, используя показатель согласованности гипотезы (7.1.13) или (7.1.14).

Можно показать, что при невыполнении гипотезы H_0 по мере возрастания n значение показателя согласованности $\hat{u}$ будет неограниченно

увеличиваться, т.е. кривая распределения $\varphi_{\hat{u}/H_1}(u)$ сдвинута относительно кривой $\varphi_{\hat{u}/H_0}(u)$ вправо. Поэтому в соответствии с рекомендациями предыдущего раздела в качестве критической целесообразно выбрать правостороннюю критическую область, рис.7.3.

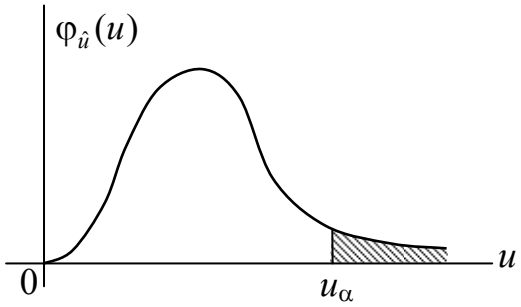


Рис.7.3. Правосторонняя критическая область

В этом случае для определения критической границы u_α можно использовать приложение 7, в котором даны критические точки распределения χ^2 в зависимости от уровня значимости α и числа степеней свободы f.

Порядок проверки гипотезы о виде закона распределения состоит в следующем.

1. Назначается уровень значимости α, и по таблице критических точек распределения χ^2 (приложение 7) определяется критическая граница u_α. Входами в таблицу служат уровень значимости α и число степеней свободы f.

2. Результаты эксперимента представляются в виде интервального статистического (вариационного) ряда (табл.4.5), в котором m_l и p_l^* – число и частота попаданий исследуемой величины $\hat{x}$ в l-й интервал ($l = \overline{1,r}$) соответственно.

3. Вычисляются вероятности p_l попадания случайной величины $\hat{x}$, которая подчиняется гипотетическому закону распределения, в l-й разряд:

$$p_l = P(x_l < \hat{x} < x_{l+1}) = \int_{x_l}^{x_{l+1}} \varphi_{\hat{x}}(x)dx,$$

где $\varphi_{\hat{x}}(x)$ – плотность распределения гипотетического закона. Очевидно, что должно выполняться условие

$$\sum_{l=1}^{r} p_l = 1.$$

4. Рассчитывается значение u показателя согласованности гипотезы по формуле (7.1.10) или (7.1.11).

5. Проверяется условие $u \le u_\alpha$. Если оно выполняется, то расхождение между экспериментальными данными и гипотезой H_0 полагается незначительным. В противном случае нулевая гипотеза отвергается.

Существенное достоинство метода К. Пирсона состоит в возможности его применения тогда, когда априорно известен лишь вид гипотетического распределения, но не известны его параметры. В этом случае параметры распределения заменяются оценками, которые используются в дальнейшем для вычисления вероятностей p_l, а число степеней свободы уменьшается на число заменяемых параметров. Метод К. Пирсона имеет следующие недостатки:

а) он применим только при большой выборке ($n \ge 100$), так как показатель согласованности подчиняется распределению хи-квадрат лишь при достаточно большом n;

б) результаты проверки в значительной степени зависят от способа разбиения выборки на интервалы, причём их число целесообразно делать не менее 8–10, а количество попаданий случайной величины $\hat{x}$ в любой из интервалов должно быть не менее 5.

П р и м е р 7.3. В условиях примера 7.1 проверить согласованность теоретического и статистического распределений.

▼ Назначаем уровень значимости $\alpha = 0{,}05$. Число степеней свободы $f = 8 - 3 = 5$. По таблице приложения 7 определяем критическую границу $u_{0,05} = 11{,}1$.

Пользуясь теоретическим нормальным законом распределения с параметрами $m = 0{,}168$ и $\sigma = 1{,}448$, находим вероятности попадания в разряды по формуле

$$p_l = \Phi_1\left(\frac{x_{l+1} - m}{\sigma}\right) - \Phi_1\left(\frac{x_l - m}{\sigma}\right),$$

где x_l, x_{l+1} – границы l-го разряда. Значения функции Φ_1 находим в таблице приложения 3. Затем составляем расчётную таблицу 7.3.

Т а б л и ц а 7.3

Расчётные данные (к примеру 7.3)

J_l	–4; –3	–3; –2	–2; –1	–1; 0	0; 1	1; 2	2; 3	3; 4
p_l^*	0,012	0,050	0,144	0,266	0,240	0,176	0,092	0,020
p_l	0,012	0,052	0,142	0,244	0,264	0,181	0,076	0,021
$p_l^* - p_l$	0	–0,002	0,002	0,022	–0,024	–0,005	0,012	–0,001
$(p_l^* - p_l)^2$	0	$4\cdot10^{-6}$	$4\cdot10^{-6}$	$484\cdot10^{-6}$	$576\cdot10^{-6}$	$25\cdot10^{-6}$	$144\cdot10^{-6}$	10^{-6}
$\frac{n(p_l^* - p_l)^2}{p_l}$	0	0,038	0,014	0,992	1,091	0,069	0,947	0,024

По формуле (7.1.10) находим значение показателя согласованности гипотезы

$$u = \sum_{l=1}^{8} \frac{500(p_l^* - p_l)^2}{p_l} = 3{,}18.$$

Поскольку $u = 3{,}18$, $u_{0,05} = 11{,}1$, то $u < u_{0,05}$ – гипотеза о нормальном распределении отклонений по вертикали при стрельбе в мишень принимается. ▲

7.2. Проверка гипотез о параметрах законов распределения

7.2.1. Проверка гипотез о равенстве математических ожиданий

Пусть имеются две независимые случайные величины $\hat{x}$ и $\hat{y}$, распределённые по нормальному закону. Эксперимент состоит в том, что над случайными величинами $\hat{x}$ и $\hat{y}$ осуществляется соответственно n и m независимых испытаний, в результате которых получаются случайные выборки $\hat{x}_1$, $\hat{x}_2, \ldots, \hat{x}_n$ и $\hat{y}_1$, $\hat{y}_2, \ldots, \hat{y}_m$. По этим выборкам определяются оценки математических ожиданий

$$\widetilde{M}_{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i, \qquad \widetilde{M}_{\hat{y}} = \frac{1}{m}\sum_{i=1}^{m} \hat{y}_j.$$

Требуется по полученным оценкам проверить гипотезу о равенстве математических ожиданий $M_{\hat{x}}$ и $M_{\hat{y}}$.

Такая задача ставится потому, что, как правило, оценки математических ожиданий оказываются различными. Причина этого может быть двоякой: либо действительно отличны и оценки и математические ожидания, либо $M_{\hat{x}}$ и $M_{\hat{y}}$ одинаковы, а отличие оценок вызвано случайными причинами, в частности, случайным отбором вариантов выборки. Если окажется, что нулевая гипотеза справедлива ($M_{\hat{x}} = M_{\hat{y}}$), то различие в оценках $\widetilde{M}_{\hat{x}}$ и $\widetilde{M}_{\hat{y}}$ обусловлено случайными причинами, иначе различными являются математические ожидания. При решении данной задачи остановимся на том случае, когда дисперсии $D_{\hat{x}}$ и $D_{\hat{y}}$ известны.

В качестве показателя согласованности гипотезы выберем случайную величину

$$\hat{u} = \frac{\widetilde{M}_{\hat{x}} - \widetilde{M}_{\hat{y}}}{\sigma[\widetilde{M}_{\hat{x}} - \widetilde{M}_{\hat{y}}]}. \qquad (7.2.1)$$

Целесообразность выбора показателя согласованности вида (7.2.1) определяется следующими соображениями.

Введём в рассмотрение случайную величину

$$\hat{z} = \widetilde{M}_{\hat{x}} - \widetilde{M}_{\hat{y}}, \qquad (7.2.2)$$

которая, очевидно, распределена по нормальному закону и имеет числовые характеристики:

$$M_{\hat{z}} = M_{\hat{x}} - M_{\hat{y}}; \quad D_{\hat{z}} = \frac{\sigma^2_{\hat{x}}}{n} + \frac{\sigma^2_{\hat{y}}}{m}; \quad \sigma_{\hat{z}} = \sqrt{\frac{\sigma^2_{\hat{x}}}{n} + \frac{\sigma^2_{\hat{y}}}{m}}.$$

Нормируем случайную величину (7.2.2) и получаем

$$\hat{u} = \frac{\hat{z}}{\sigma_z} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - \frac{1}{m}\sum_{j=1}^{m}\hat{y}_j}{\sqrt{\frac{\sigma^2_{\hat{x}}}{n} + \frac{\sigma^2_{\hat{y}}}{m}}}. \qquad (7.2.3)$$

Случайная величина (7.2.3) подчинена нормальному закону распределения, параметры которого известны: $M_{\hat{u}} = 0$, $\sigma_{\hat{u}} = 1$, что существенно упрощает процедуру проверки нулевой гипотезы. Действительно, если гипотеза H_0 справедлива ($M_{\hat{x}} = M_{\hat{y}}$), то случайная величина центрирована, откуда следует, что $M_{\hat{u}} = 0$. Так как выборки независимые, то $\sigma_{\hat{u}} = 1$.

Критическая область строится в зависимости от вида конкурирующей гипотезы, которая может быть сформулирована тремя различными способами:

$$M_{\hat{x}} \neq M_{\hat{y}}; \; M_{\hat{x}} > M_{\hat{y}}; \; M_{\hat{x}} < M_{\hat{y}}.$$

Рассмотрим методику проверки гипотезы H_0 для каждого из приведённых способов формулировки конкурирующей гипотезы.

1. H_0: $M_{\hat{x}} = M_{\hat{y}}$; H_1: $M_{\hat{x}} \neq M_{\hat{y}}$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания в неё показателя согласованности в предположении о справедливости нулевой гипотезы была равна принятому уровню значимости α.

Наибольшая мощность критерия достигается тогда, когда левая и правая критические точки $u_{\alpha 1}$, $u_{\alpha 2}$ выбраны так, что вероятность попадания показателя согласованности $\hat{u}$ в каждый из двух интервалов критической области равна $\alpha/2$:

$$P(\hat{u} < u_{\alpha 1}) = \frac{\alpha}{2}; \quad P(\hat{u} \geq u_{\alpha 2}) = \frac{\alpha}{2}.$$

Поскольку $\hat{u}$ – нормированная нормально распределённая случайная величина и её распределение симметрично относительно нуля, то критические точки также симметричны относительно нуля:

$$|u_{\alpha 1}| = |u_{\alpha 2}| = |u_{\alpha}|.$$

Используя функцию нормированного нормального распределения (функцию Лапласа), вероятность попадания показателя согласованности в критическую область можно определить выражением

$$1 - P(|\hat{u}| < u_\alpha) = 1 - 2\Phi_0(u_\alpha) = \alpha,$$

откуда

$$u_\alpha = \Phi_0^{-1}\left(\frac{1-\alpha}{2}\right) = t_{1-\alpha}. \qquad (7.2.4)$$

Двусторонняя критическая область будет определяться неравенствами $u < -u_\alpha$, $u > u_\alpha$. Таким образом, правило проверки гипотезы H_0 для рассматриваемого случая состоит в следующем.

а). Назначается уровень значимости α и в соответствии с формулой (7.2.4) по таблице приложения 4 определяются границы критической области $u_{\alpha 1} = -u_\alpha$, $u_{\alpha 2} = u_\alpha$.

б). На основе случайных выборок вычисляется наблюдаемое значение показателя u по формуле

$$u = \frac{\widetilde{M}_{\hat{x}} - \widetilde{M}_{\hat{y}}}{\sqrt{\frac{\sigma_{\hat{x}}^2}{n} + \frac{\sigma_{\hat{y}}^2}{m}}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i - \frac{1}{m}\sum_{j=1}^{m}\hat{y}_j}{\sqrt{\frac{\sigma_{\hat{x}}^2}{n} + \frac{\sigma_{\hat{y}}^2}{m}}}. \qquad (7.2.5)$$

в). Проверяется условие $|u| > u_\alpha$. Если оно выполняется, то гипотеза H_0 отвергается. В противном случае данные эксперимента не противоречат нулевой гипотезе.

П р и м е р 7.4. Производится контрольный отстрел двух партий снарядов, причём из первой партии проверяется 10 снарядов, а из второй – 15. В результате отстрела получены следующие оценки математических ожиданий отклонения точек попадания снарядов от точки прицеливания по дальности: для первой партии отклонение равно –0,8 км, для второй +0,4 км. Среднеквадратические отклонения по дальности для снарядов первой и второй партий известны и равны соответственно 2 и 1,5 км. Необходимо проверить гипотезу о совпадении проекций центров рассеивания на ось дальности в обеих партиях.

▼ Пусть $\hat{x}$ и $\hat{y}$ – отклонение точек попадания снарядов от точки прицеливания по дальности соответственно для первой и второй партий.

По условию задачи $n = 10$, $m = 15$, $\widetilde{M}_{\hat{x}} = -0{,}8$ км, $\widetilde{M}_{\hat{y}} = 0{,}4$ км, $\sigma_{\hat{x}} = 2$ км, $\sigma_{\hat{y}} = 1{,}5$ км.

Задаёмся уровнем значимости $\alpha = 0{,}05$ и в приложении 4 находим

$$u_\alpha = t_{1-\alpha} = t_\gamma = t_{0{,}95} = 1{,}96.$$

Используя формулу (7.2.5), вычисляем абсолютное значение показателя согласованности:

$$|u| = \left| \frac{-0{,}8 - 0{,}4}{\sqrt{\frac{4}{10} + \frac{2{,}25}{15}}} \right| = 1{,}62 .$$

Так как $|u| < u_\alpha$, нулевая гипотеза $M_{\hat{x}} = M_{\hat{y}}$ не противоречит данным контрольного отстрела.

2. H_0: $M_{\hat{x}} = M_{\hat{y}}$; H_1: $M_{\hat{x}} > M_{\hat{y}}$.

Такой случай возможен, если априорные сведения позволяют предположить, что $M_{\hat{x}} > M_{\hat{y}}$. В этом случае строят такую правостороннюю критическую область, чтобы вероятность попадания в неё показателя согласованности в предположении о справедливости нулевой гипотезы была равна α:

$$P(\hat{u} \geq u_\alpha) = \alpha . \qquad (7.2.6)$$

Для того чтобы критическую точку найти с помощью функции Лапласа, перепишем выражение (7.2.6) в виде

$$P(\hat{u} \geq u_\alpha) = P(u_\alpha \leq \hat{u} < \infty) = 1 - \Phi_1(u_\alpha) = \alpha .$$

Из предыдущего выражения получим

$$\Phi_1(u_\alpha) = 1 - \alpha$$

и, следовательно,

$$u_\alpha = \Phi_1^{-1}(1 - \alpha) = t_{1-2\alpha} \qquad (7.2.7)$$

Правило проверки гипотезы для рассматриваемого случая.

а). Назначается уровень значимости α и в соответствии с (7.2.7) по таблице приложения 4 определяется величина u_α. При этом в таблицу следует входить со значением $1–2\alpha$,

б). Определяется величина u по формуле (7.2.5).

в). Проверяется условие $u > u_\alpha$. Если оно выполняется, гипотеза H_0 отвергается, в противном случае принимается.

3. H_0: $M_{\hat{x}} = M_{\hat{y}}$; H_1: $M_{\hat{x}} < M_{\hat{y}}$.

При указанной формулировке конкурирующей гипотезы левостороннюю критическую область строят так, чтобы вероятность попадания в неё показателя согласованности в предположении о справедливости нулевой гипотезы была равна принятому уровню значимости: $P(\hat{u} < u_\alpha) = \alpha$.

Учитывая, что показатель $\hat{u}$ имеет симметричное распределение относительно нуля, заключаем, что точка $u_{\alpha 1}$ симметрична такой точке $u_\alpha > 0$, для которой $P(\hat{u} \geq u_\alpha) = \alpha$, это значит $u_{\alpha 1} = –u_\alpha$. Следовательно, методика определения $u_{\alpha 1}$ полностью совпадает с методикой предыдущего случая, только полученное значение берётся с отрицательным знаком.

Правило проверки гипотезы также аналогично рассмотренному выше правилу, за исключением последнего пункта, а именно, если $u < u_{\alpha 1}$, нулевая гипотеза отвергается, в противном случае – принимается.

Выше предполагалось, что случайные величины $\hat{x}$ и $\hat{y}$ распределены нормально, а их дисперсии известны. При этих предположениях показатель согласованности гипотезы распределён по нормальному закону с параметрами $M_{\hat{u}} = 0,\ \sigma_{\hat{u}} = 1$. Если хотя бы одно из предположений не выполняется, описанный метод проверки гипотезы о равенстве математических ожиданий неприменим. Однако при больших объёмах независимых выборок (≥ 30 вариантов каждая) оценки математических ожиданий и дисперсий распределены приближённо нормально и закон распределения $\hat{u}$ можно считать близким к нормальному. В этом случае проверку гипотезы можно проводить по описанной выше методике, подставляя в формулу (7.2.5) оценки дисперсий, но к полученным результатам следует относиться с осторожностью.

7.2.2. Проверка гипотез о равенстве дисперсий

Проверка гипотез о равенстве дисперсий – одна из важнейших задач статистической обработки экспериментальных данных. На практике задача сравнения дисперсий возникает, если требуется сравнить погрешности показаний приборов, точность методов измерений и т.д.

Сформулируем задачу проверки гипотезы о равенстве дисперсий. Пусть имеются две случайные величины $\hat{x}$ и $\hat{y}$, каждая из которых подчиняется нормальному закону распределения с дисперсиями $D_{\hat{x}}$ и $D_{\hat{y}}$. По независимым выборкам $x_1, x_2,\ldots,x_n$ и $y_1, y_2,\ldots,y_m$ найдены оценки дисперсий:

$$\left.\begin{aligned} \widetilde{D}_{\hat{x}} &= \frac{1}{n-1}\sum_{i=1}^{n}(\hat{x}_i - \widetilde{M}_{\hat{x}})^2 = \frac{D_{\hat{x}}}{n-1}\hat{\chi}^2_{n-1}; \\ \widetilde{D}_{\hat{y}} &= \frac{1}{m-1}\sum_{j=1}^{m}(\hat{y}_j - \widetilde{M}_{\hat{y}})^2 = \frac{D_{\hat{y}}}{m-1}\hat{\chi}^2_{m-1}. \end{aligned}\right\}, \qquad (7.2.8)$$

где $\hat{\chi}^2_{n-1}$, $\hat{\chi}^2_{m-1}$ - хи- квадрат распределения с $n-1$ и $m-1$ степенями свободы соответственно.

Обычно полученные оценки различны, в связи с чем возникает вопрос, можно ли на основе обработки экспериментальных данных полагать, что $D_{\hat{x}} = D_{\hat{y}}$ (нулевая гипотеза).

Если нулевая гипотеза справедлива, то это означает, что выборочные дисперсии (7.2.8) представляют собой оценки одной и той же характеристики рассеивания генеральной совокупности и их различие определяется случайными причинами. В противном случае различие оценок су-

щественно и является следствием того, что дисперсии генеральных совокупностей различны.

В качестве показателя согласованности гипотезы о равенстве дисперсий примем отношение большей оценки дисперсии к меньшей. Для определённости будет полагать $D_{\hat{x}} > D_{\hat{y}}$, тогда

$$\hat{u} = \frac{D_{\hat{x}}}{D_{\hat{y}}}. \qquad (7.2.9)$$

Учитывая оценки (7.2.8) при условии, что нулевая гипотеза справедлива, на основе отношения (7.2.9) получаем следующее выражение показателя согласованности:

$$\hat{u} = \frac{\hat{\chi}^2_{n-1}(m-1)}{\hat{\chi}^2_{m-1}(n-1)} = \mathsf{F}_{(n-1;\,m-1)}.$$

Таким образом, показатель согласованности представляет собой случайную величину, подчинённую закону распределения Фишера со степенями свободы $f_1 = n - 1$ и $f_2 = m - 1$. Как известно, распределение Фишера зависит только от значений степеней свободы и уровня значимости, а от других параметров не зависит.

Критическая область в зависимости от вида конкурирующей гипотезы строится по-разному. Как и ранее, рассмотрим три вида конкурирующей гипотезы:

$$D_{\hat{x}} \neq D_{\hat{y}}; \quad D_{\hat{x}} > D_{\hat{y}}; \quad D_{\hat{x}} < D_{\hat{y}}.$$

Построение критических областей для каждого из этих видов осуществляется следующим образом.

1. $H_0: D_{\hat{x}} = D_{\hat{y}}; \quad H_1: D_{\hat{x}} \neq D_{\hat{y}}$.

В этом случае строят двустороннюю критическую область, исходя из того, чтобы вероятность попадания в неё показателя согласованности в предположении о справедливости нулевой гипотезы была равна уровню значимости α. При этом достигается наибольшая мощность критерия проверки, когда вероятности попадания показателя согласованности в каждый из двух интервалов критической области будут одинаковы и равны $\alpha/2$. Таким образом, при построении критической области должны выполняться следующие условия (рис.7.4):

$$\left.\begin{aligned} P(\hat{u} < u_{\alpha 1}) = \alpha/2; \\ P(\hat{u} \geq u_{\alpha 2}) = \alpha/2. \end{aligned}\right\}$$

Правая критическая точка $u_{\alpha 2}$ может быть найдена непосредственно по таблице критических точек распределения Фишера (приложение 5). При этом входами в таблицу будут величины $\alpha/2$, $f_1 = n - 1$, $f_2 = m - 1$. В результате имеем

$$u_{\alpha 2} = \mathsf{F}_{\left(\frac{\alpha}{2};\, n-1;\, m-1\right)} = \mathsf{F}_{\alpha 2}\,.$$

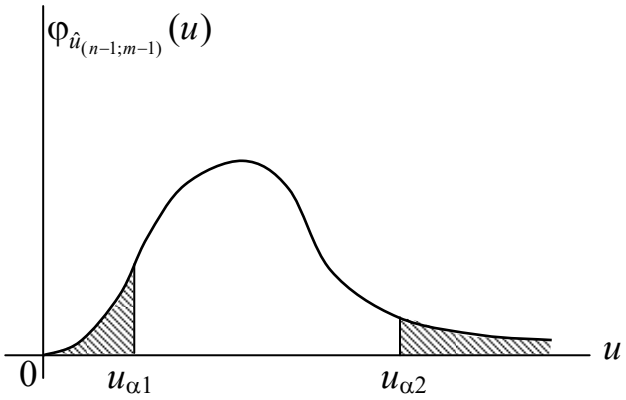


Рис.7.4. Двусторонняя критическая область

Однако левых критических точек данная таблица не содержит и найти непосредственно $u_{\alpha 1}$ невозможно. В связи с этим для нахождения левой критической границы $u_{\alpha 1}$ необходимо использовать следующий приём.

Рассмотрим события

$$\mathsf{F}_{(n-1;\, m-1)} < \mathsf{F}_{\alpha 1} \quad \text{и} \quad \frac{1}{\mathsf{F}_{(n-1;\, m-1)}} \geq \frac{1}{\mathsf{F}_{\alpha 1}}\,.$$

Так как эти события эквивалентны, то их вероятности равны:

$$\frac{\alpha}{2} = P(\hat{\mathsf{F}}_{(n-1;\, m-1)} < \mathsf{F}_{\alpha 1}) = P\left(\frac{1}{\hat{\mathsf{F}}_{(n-1;\, m-1)}} \geq \frac{1}{\mathsf{F}_{\alpha 1}}\right).$$

Как известно [1], случайная величина $1\big/\hat{\mathsf{F}}_{(n-1;\, m-1)}$ также подчиняется закону распределению Фишера со степенями свободы $f_1 = m - 1, f_2 = n - 1$. Поэтому значение $1/\mathsf{F}_{\alpha 1}$ может быть найдено как верхний $100(\alpha/2)$-процентный предел этого закона распределения:

$$\frac{1}{\mathsf{F}_{\alpha 1}} = \mathsf{F}_{\left(\frac{\alpha}{2};\, m-1;\, n-1\right)}.$$

Таким образом, для определения $1/\mathsf{F}_{\alpha 1}$ необходимо войти в таблицу критических точек распределения Фишера с аргументами $\alpha/2$, $f_1 = m - 1$, $f_2 = n - 1$. Значение левой критической границы определяется как величина, обратная значению, найденному по таблице.

Учитывая изложенное выше, правило проверки гипотезы о равенстве дисперсий можно сформулировать в следующем виде.

а). Назначается уровень значимости α и по таблице критических точек распределения Фишера находятся критические границы $u_{\alpha 1}$ и $u_{\alpha 2}$. При нахождении критической границы $u_{\alpha 2}$ в таблицу следует входить с

аргументами $\alpha/2$, $f_1 = n - 1$, $f_2 = m - 1$, а при определении критической границы $u_{\alpha 1}$ – с аргументами $\alpha/2$, $f_1 = m - 1$, $f_2 = n - 1$. В последнем случае табличное значение $\mathsf{F}_{\left(\frac{\alpha}{2}; m-1; n-1\right)}$ используется для определения критической границы $u_{\alpha 1}$ из выражения

$$u_{\alpha 1} = \frac{1}{\mathsf{F}_{\left(\frac{\alpha}{2}; m-1; n-1\right)}}. \qquad (7.2.10)$$

б). Вычисляется значение показателя согласованности

$$u = \frac{\widetilde{D}_{\hat{x}}}{\widetilde{D}_{\hat{y}}} = \frac{\widetilde{\sigma}^2_{\hat{x}}}{\widetilde{\sigma}^2_{\hat{y}}}. \qquad (7.2.11)$$

в). Проверяется неравенство

$$u_{\alpha 1} < u < u_{\alpha 2}.$$

Если оно выполняется, то наблюдаемое значение показателя согласованности попадает в область допустимых значений. В этом случае делается вывод об отсутствии существенного различия между сравниваемыми дисперсиями и гипотеза H_0 принимается. Если $u < u_{\alpha 1}$ или $u > u_{\alpha 2}$, то нулевая гипотеза отвергается.

П р и м е р 7.5. При исследовании стабилизатора напряжения проведено семь испытаний и получена оценка дисперсии выходного напряжения, равная 0,06 B^2. После доработки стабилизатора проведено ещё 13 испытаний, в результате чего оценка дисперсии выходного напряжения стала равна 0,10 B^2. Есть ли основание полагать, что в результате доработки точность стабилизатора не изменилась?

▼ Обозначим $\widetilde{D}_{\hat{x}} = 0{,}10\ B^2$, $\widetilde{D}_{\hat{y}} = 0{,}06 B^2$. Тогда $n = 13$, $m = 7$. Задаёмся уровнем значимости $\alpha = 0{,}10$ и в приложении 5 находим $u_{\alpha 2}$ для $\alpha/2 = 0{,}05$, $f_1 = n - 1 = 12$, $f_2 = m - 1 = 6$. Также находим $u_{\alpha 1}$ для $\alpha/2 = 0{,}05$, $f_1 = m - 1 = 6$, $f_2 = n - 1 = 12$. Получаем $u_{\alpha 2} = 4$, $\mathsf{F}_{(0{,}05;6;12)} = 3$ и, следовательно, $u_{\alpha 1} = 0{,}33$.

Значение показателя согласованности по формуле (7.2.11):

$$u = \frac{\widetilde{D}_{\hat{x}}}{\widetilde{D}_{\hat{y}}} = \frac{0{,}10}{0{,}06} = 1{,}67.$$

Так как $u_{\alpha 1} < u < u_{\alpha 2}$, то гипотеза H_0 о том, что доработка не повлияла на точность стабилизатора напряжения, принимается. ▲

2. $H_0: D_{\hat{x}} = D_{\hat{y}}$; $H_1: D_{\hat{x}} > D_{\hat{y}}$.

В этом случае строят правостороннюю критическую область таким образом, чтобы вероятность попадания в эту область показателя согласованности в предположении о справедливости нулевой гипотезы была равна принятому уровню значимости:

$$P(\hat{u} > u_{\alpha}) = \alpha.$$

Критическую точку $u_\alpha = \mathsf{F}_{(\alpha; f_1; f_2)}$ находят по таблице критических точек распределения Фишера, используя в качестве аргументов α, $f_1 = n - 1$, $f_2 = m - 1$. Наблюдаемое значение показателя согласованности определяется по формуле (7.2.11). Если $u < u_\alpha$, то нет оснований отвергнуть нулевую гипотезу, в противном случае она отвергается.

3. H_0: $D_{\hat{x}} = D_{\hat{y}}$; H_1: $D_{\hat{x}} < D_{\hat{y}}$.

В данном случае строят левостороннюю критическую область таким образом, чтобы

$$P(\hat{u} < u_\alpha) = \alpha .$$

Критическая точка находится по таблице критических точек распределения Фишера на основе отношения

$$u_\alpha = \frac{1}{\mathsf{F}_{(\alpha; m-1; n-1)}} . \qquad (7.2.12)$$

В знаменателе (7.2.12) – табличное значение, найденное при аргументах α, $f_1 = m - 1$, $f_2 = n - 1$.

Наблюдаемое значение показателя согласованности определяется по формуле (7.2.11). Если $u > u_\alpha$, то нулевая гипотеза принимается, в противном случае она должна быть отвергнута.

В заключение следует отметить, что показатель согласованности гипотезы (7.2.9) можно использовать для сравнения дисперсий и в том случае, когда для одной из дисперсий найдена не оценка, а её точное значение. В этом случае число степеней свободы закона распределения Фишера в числителе или знаменателе выражения (7.2.9) следует устремить к бесконечности, в остальном методика проверки гипотезы остаётся прежней.

П р и м е р 7.6. Из партии снарядов с известной характеристикой рассеивания по дальности $\sigma_{\hat{x}_1} = 20$ м испытываются 10 снарядов, хранившихся без специальной тары. Есть ли основание полагать, что по причине такого хранения рассеивание снарядов по дальности возросло, если в результате испытаний получена оценка $\tilde{\sigma}_{\hat{x}_1} = 27$ м?

▼ В данном примере кривая распределения характеристики $\hat{u}$ при конкурирующей гипотезе смещена влево, поэтому в качестве критической выбираем левостороннюю область.

Пусть $\alpha = 0{,}05$, тогда для определения u_α входим в таблицу приложения 5 со значениями $\alpha = 0{,}05$, $f_1 = m - 1 = 9$, $f_2 = \infty$. Получим $\mathsf{F}_{(0{,}05;9;\infty)} = 1{,}88$, следовательно,

$$u_\alpha = \frac{1}{1{,}88} = 0{,}53 .$$

Вычисляем значение показателя согласованности

$$u = \frac{\widetilde{\sigma}^2_{\hat{x}}}{\widetilde{\sigma}^2_{\hat{y}}} = \left(\frac{27}{20}\right)^2 = 1{,}82\,.$$

Так как $u > u_\alpha$ и значение показателя согласованности попало в область допустимых значений, то нет оснований утверждать, что в результате хранения без специальной тары рассеивание снарядов по дальности возросло.

▲

8. СТАТИСТИЧЕСКИЙ АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

8.1. Статистический анализ и обработка данных

В предыдущих разделах были рассмотрены методы определения характеристик, описывающих свойства случайных объектов (величин, векторов, функций). Однако цель обработки экспериментальных данных в конечном счёте состоит в выявлении причинно-следственных связей, определяющих состояние и развитие изучаемого явления. Установление этих связей позволяет не только глубоко анализировать различные процессы, но и определять оптимальные пути управления ими.

Решение указанной задачи осуществляется с помощью ряда методов, объединяемых единым названием – **методы статистического анализа** экспериментальных данных. В число этих методов входят методы дисперсионного, корреляционного, регрессионного, компонентного и факторного анализов, метод наименьших квадратов.

Все эти методы целесообразно разделить на две группы:

– методы статического статистического анализа (ССА), в которых фактор времени в явном виде не учитывается;

– методы динамического статистического анализа (ДСА), в которых экспериментальные данные представляются в форме динамических или временных рядов.

Из всего многообразия методов статистического анализа ниже будут изложены широко распространённые методы – наименьших квадратов и регрессионного анализа. Они рассматриваются как методы ССА.

В методах ССА признак, характеризующий причины, принято называть ***факторным признаком*** или для краткости – ***фактором***. Признак, характеризующий следствия, принято называть ***результативным признаком*** или для краткости – ***результатом*** (***результатом наблюдений***).

При получении и обработке данных предполагается, что результат наблюдения y зависит от одного или нескольких факторов x_1, x_2,…, x_m, и фиксируется по отношению к данным факторам. В процессе обработки решается ряд вопросов.

1. Справедливо ли предположение о зависимости результата y от факторов x_1, x_2,…,x_m?

2. Как оценить степень этой зависимости?

3. Как выделить среди факторов наиболее существенные?

4. Нельзя ли сократить число факторов, используемых при анализе?

5. Какой вид имеет причинно-следственная зависимость между факторами и результатом?

Прежде чем приступить к рассмотрению данных вопросов, остановимся на процедуре формального представления причинно-следственных связей между результатом y и факторами $x_1, x_2,\ldots,x_m$. Указанная процедура сводится к определению зависимости

$$y = f(x_1, x_2,\ldots,x_m).$$

Подход к решению данной задачи различен в зависимости от свойств факторов $X_{<m>}$, функции f и, наконец, свойств результатов наблюдений. По этой причине постановка задачи анализа и методы её решения могут быть существенно различными. Для описания данной задачи будем обозначать символом f функциональную (детерминированную), а символом $\hat{f}$ – стохастическую зависимость между $X_{<m>}$ и y.

Напомним, что детерминированная – это зависимость величины y или её некоторой характеристики, например математического ожидания, от факторов $X_{<m>}$. Зависимость закона распределения результата y от факторов $X_{<m>}$ является стохастической.

Тогда могут иметь место следующие виды зависимостей результата y от факторов $X_{<m>}$.

1. Функциональная зависимость от неслучайных факторов

$$\hat{y} = f(x_1,\ x_2,...,x_m) + \hat{\varepsilon}, \qquad (8.1.1)$$

при которой случайный характер результата y обусловливается только ошибками $\hat{\varepsilon}$ при наблюдении данного результата.

2. Стохастическая зависимость от неслучайных факторов

$$\hat{y} = \hat{f}(x_1,\ x_2,...,x_m), \qquad (8.1.2)$$

при которой случайный характер результата y обусловливается стохастическим характером зависимости $\hat{f}$.

3. Функциональная зависимость от случайных факторов

$$\hat{y} = f(\hat{x}_1,\ \hat{x}_2,...,\hat{x}_m), \qquad (8.1.3)$$

при которой случайный характер результата y обусловливается случайным характером факторов.

4. Полная стохастическая зависимость

$$\hat{y} = \hat{f}(\hat{x}_1,\ \hat{x}_2,...,\hat{x}_m). \qquad (8.1.4)$$

Решение задачи анализа для зависимостей типа (8.1.1) и (8.1.3) опирается на метод наименьших квадратов, а для (8.1.2) и (8.1.4) – на методы регрессионного и корреляционного анализов.

8.2. Сущность метода наименьших квадратов

Метод наименьших квадратов (МНК) получил широкое распространение при обработке экспериментальных данных в целях исследования различных функциональных зависимостей, определения параметров распределений и т.д.

Существует широкий класс задач, в которых МНК является оптимальным методом обработки данных. В других классах задач использование МНК часто оправдывается алгоритмической простотой его реализации ценой небольших потерь в оптимальности получаемого результата. Для нелинейных задач статистического анализа данных зачастую невозможно использование каких-либо других методов, кроме МНК.

Эти и другие причины объясняют широкое распространение МНК при статистическом анализе экспериментальных данных, в частности, при выявлении функциональных зависимостей. Исторически МНК возник значительно раньше других методов обработки данных. Вероятностное обоснование МНК дано К. Гауссом в начале XIX в. и А.А. Марковым в начале XX в.

Предположим, что требуется определить компоненты вектора

$$A_{<k>} = (a_1, a_2,\ldots, a_k)^{\mathrm{T}},$$

который в общем случае не поддаётся непосредственному наблюдению. Однако можно наблюдать вектор

$$Y_{<n>} = (y_1, y_2,\ldots, y_n)^{\mathrm{T}},$$

функционально связанный с искомым вектором $A_{<k>}$:

$$Y_{<n>} = F_{<n>}(t; A_{<k>}). \qquad (8.2.1)$$

При этом соотношение размерностей векторов A и Y может быть произвольным. В частном случае A может быть скалярной величиной, а Y – вектором, и наоборот.

В общем случае вектор-функция F является нелинейной. Схема оценивания, в которой по наблюдениям в некоторые моменты времени t_i, $i = \overline{1, N}$ одного набора параметров (в данном случае компонентов вектора Y) необходимо оценить компоненты другого набора параметров (компоненты вектора A), функционально связанного с первым, называется схемой косвенных наблюдений.

Процесс наблюдения всегда сопровождается ошибками. Наблюдаемое значение функции (8.2.1) в момент времени t_i отклоняется от теоретического вследствие случайных факторов. Следовательно, результат наблюдения всегда представляет собой реализацию случайной величины. В общем случае ошибка наблюдения нелинейным образом связана с наблюдаемой функцией.

На практике часто удаётся путём линеаризации уравнений модели (8.2.1) относительно случайных ошибок свести уравнения к форме, когда

случайные ошибки входят аддитивно или мультипликативно (см. §1.1). Однако наиболее простым и самым распространённым типом связи ошибок наблюдения и наблюдаемых величин является линейная аддитивная связь, когда модель наблюдения может быть представлена уравнениями

$$\hat{Y}_{<n>_i} = F_{<n>_i}(t_i; A_{<k>}) + \hat{\mathrm{E}}_{<n>_i}, \quad i = \overline{1, N}, \tag{8.2.2}$$

где $\hat{\mathrm{E}}_i$ – вектор аддитивной ошибки в *i*-й момент наблюдения.

В дальнейшем будем рассматривать эту схему наблюдения. Уравнения типа (8.2.2) называются ***уравнениями наблюдения***. Поскольку компоненты вектора $\hat{\mathrm{E}}_i$ являются случайными неизвестными наблюдателю величинами, то для поиска оценок вектора $A_{<k>}$ используется уравнение вида

$$\hat{Y}_{<n>_i} = F_{<n>_i}(t_i; A_{<k>}), \tag{8.2.3}$$

которое может оказаться и несовместным, поскольку отражает наблюдаемый процесс приближённо. Поэтому уравнение (8.2.3) принято называть ***условным***.

Если моменты времени t_i, $i = \overline{1, N}$ представляют собой известные и в данной задаче фиксированные величины, то фактически вектор-функция *F* является функцией только вектора *A*. Поэтому в дальнейшем в число аргументов будем включать моменты времени t_i тогда, когда они либо неизвестны, либо известны с ошибкой. С учётом сказанного уравнение наблюдения запишется в виде

$$\hat{Y}_{<n>} = F_{<n>}(A_{<k>}) + \hat{\mathrm{E}}_{<n>}. \tag{8.2.4}$$

В соответствии с методом наименьших квадратов оценки компонентов вектора *A* отыскиваются на основе минимизации суммы квадратов отклонений между *Y* и *F*:

$$\hat{V}(\widetilde{A}) = \min_{A \in \mathbf{R}^k} \{(\hat{Y} - F(A))^{\mathsf{T}} (\hat{Y} - F(A))\}, \tag{8.2.5}$$

где $\widetilde{A}$ – вектор, представляющий собой решение задачи (8.2.5); $\mathbf{R}^k$ – *k*-мерное вещественное пространство.

Часто вместо минимизации квадратичной функции (8.2.5) для оценивания вектора *A* используют минимизацию квадратичной функции более общего вида:

$$\hat{V}(\widetilde{A}) = \min_{A \in \mathbf{R}^k} \{(\hat{Y} - F(A))^{\mathsf{T}} Q_{[n]} (\hat{Y} - F(A))\}, \tag{8.2.6}$$

где $Q_{[n]}$ – неотрицательно определённая симметричная матрица, которая называется весовой.

Очевидно, что задача (8.2.5) является частным случаем задачи (8.2.6), если в качестве весовой выбрать единичную матрицу.

Показатель качества оценивания (8.2.5) в скалярной форме имеет вид

$$\hat{V}(\widetilde{A}) = \min_{A \in \mathbf{R}^k} \left\{ \left(\sum_{i=1}^{n} (\hat{y}_i - f_i(A))^2 \right) \right\}, \qquad (8.2.7)$$

где $\hat{y}_i$ и f_i, $i = \overline{1,n}$ – компоненты вектора Y и вектор-функции F соответственно.

Скалярная форма показателя (8.2.6):

$$\hat{V}(\widetilde{A}) = \min_{A \in \mathbf{R}^k} \left\{ \left(\sum_{i,j=1}^{n} (\hat{y}_i - f_i(A)) q_{ij} (\hat{y}_j - f_j(A)) \right) \right\}, \qquad (8.2.8)$$

Для сокращения записей, а также упрощения некоторых выкладок в последующем будем использовать в основном векторно-матричную запись показателя качества оценивания вектора A.

Необходимое условие минимума функции (8.2.5) или (8.2.6) состоит в том, что её частные производные по всем компонентам a_i вектора A должны быть равны нулю:

$$\frac{\partial \hat{V}}{\partial a_i} = 0, \quad i = \overline{1,k}. \qquad (8.2.9)$$

Развёрнутый вид условий (8.2.9) в скалярной форме представляется следующей системой уравнений:

$$\begin{cases} \dfrac{\partial \hat{V}}{\partial a_1} = \sum\limits_{i=1}^{n} (\hat{y}_i - f_i(a_1, a_2, ..., a_k)) \dfrac{\partial f_i(a_1, a_2, ..., a_k)}{\partial a_1} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial \hat{V}}{\partial a_j} = \sum\limits_{i=1}^{n} (\hat{y}_i - f_i(a_1, a_2, ..., a_k)) \dfrac{\partial f_i(a_1, a_2, ..., a_k)}{\partial a_j} = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial \hat{V}}{\partial a_k} = \sum\limits_{i=1}^{n} (\hat{y}_i - f_i(a_1, a_2, ..., a_k)) \dfrac{\partial f_i(a_1, a_2, ..., a_k)}{\partial a_n} = 0 \end{cases} \qquad (8.2.10)$$

Система уравнений (8.2.10) в теории МНК называется системой ***нормальных уравнений***. Её решение, т.е. вектор $\widetilde{A}$ будем в дальнейшем называть МНК-оценкой.

П р и м е р 8.1. Требуется оценить скалярную величину a, для которой уравнение наблюдения имеет вид

$$\hat{y}_i = a + \hat{\varepsilon}_i, \quad i = \overline{1,n}.$$

▼ Предположим, что оценку параметра a необходимо отыскивать минимизацией суммы квадратов отклонений

$$\hat{V}(a) = \sum_{i=1}^{n} (\hat{y}_i - a)^2.$$

Тогда необходимое условие минимума этой суммы квадратов в соответствии с (8.2.9) запишется в виде

$$\frac{\partial \hat{V}}{\partial a} = -2\sum_{i=1}^{n}(\hat{y}_i - a) = \sum_{i=1}^{n}(y_i - a) = 0\,.$$

Оценка искомой величины

$$\widetilde{a} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\,.$$

Следует заметить, что при решении данной задачи обоснование выбора функции $V(a)$ отсутствовало, хотя ранее в § 2.4 указывалось, что оптимальный выбор данной функции диктуется условиями задачи.

Замечание приведено в связи с тем, что в дальнейшем будет даваться и иное решение этой же задачи. ▲

В общем случае система нормальных уравнений нелинейная относительно искомых параметров, а искомые параметры – компоненты вектора A – выражаются нелинейным образом через компоненты вектора наблюдения $\hat{Y}$.

8.3. Метод наименьших квадратов при линейной связи наблюдаемых и оцениваемых параметров

8.3.1. Линейная модель наблюдения

Рассматриваемый ниже метод наименьших квадратов при линейной модели наблюдения получил название схемы Гаусса–Маркова.

Пусть наблюдаемые параметры и параметры искомой функциональной зависимости связаны линейным уравнением, а ошибки наблюдения аддитивны, причём имеют равные нулю математические ожидания:

$$\hat{Y}_{<n>} = X_{[n;k]}A_{<k>} + \hat{\mathrm{E}}_{<n>}, \qquad (8.3.1)$$

$$M[\hat{\mathrm{E}}_{<n>}] = 0\,, \qquad (8.3.2)$$

где $X_{[n;k]}$ – прямоугольная матрица, называемая матрицей наблюдения; $\hat{Y}_{<n>}$ и $\hat{\mathrm{E}}_{<n>}$, как и ранее, соответственно случайные векторы наблюдения и ошибок наблюдения.

На основании равенства (8.3.2) можно записать

$$M[\hat{Y}_{<n>}] = X_{[n;k]}A_{<k>}\,. \qquad (8.3.3)$$

В модели Гаусса–Маркова предполагается, что относительно вектора $\hat{\mathrm{E}}$ известна некоторая дополнительная информация. Рассмотрим возможные варианты использования данной информации.

Один из наиболее простых случаев тот, когда известно, что наблюдения некоррелированы и равноточны. В этом варианте корреляционная

матрица вектора $\hat{\text{Ê}}$ или, что то же самое, вектора $\hat{Y}$ выражается формулой

$$K_{\hat{\text{Ê}}[n]} = \sigma^2 E_{[n]}, \qquad (8.3.4)$$

где σ^2 – дисперсия наблюдения; $E_{[n]}$ – единичная матрица.

Дисперсия наблюдения может быть и неизвестной, тогда она подлежит оценке наряду с компонентами вектора A. При неизвестной дисперсии σ^2 не представляется возможным получить какие-либо характеристики точности оценивания.

Более общим является случай, когда наблюдения коррелированны и равноточны, однако для них известна только нормированная корреляционная матрица $G_{[n]}$, а корреляционная матрица вектора $\hat{\text{Ê}}$ имеет вид

$$K_{\hat{\text{Ê}}[n]} = \sigma^2 G_{[n]}, \qquad (8.3.5)$$

причём σ^2 – в общем случае неизвестная дисперсия наблюдения.

Покажем, что модель наблюдения (8.3.3), (8.3.4) легко сводится к модели (8.3.3), (8.3.5). Из линейной алгебры известно, что любая симметричная положительно-определённая матрица (а матрица G является таковой) может быть представлена в виде

$$G_{[n]} = D_{[n]} D_{[n]}^{\text{т}}, \qquad (8.3.6)$$

где $D_{[n]}$ – невырожденная матрица.

Произведём замену переменных по формуле

$$\hat{Z} = D^{-1}\hat{Y}, \qquad (8.3.7)$$

тогда

$$\hat{Y} = D\hat{Z}.$$

Корреляционная матрица вектора $\hat{Z}$ вычисляется следующим образом:

$$K_{\hat{Z}[n]} = M\left((\hat{Z} - M_{\hat{Z}})(\hat{Z} - M_{\hat{Z}})^{\text{т}}\right) =$$
$$= M\left(D^{-1}(\hat{Y} - M_{\hat{Y}})(\hat{Y} - M_{\hat{Y}})^{\text{т}} D^{-1\text{т}}\right) =$$
$$= D^{-1} K_{\hat{\text{Ê}}} D^{-1\text{т}} = \sigma^2 D^{-1} D D^{\text{т}} D^{-1\text{т}} = \sigma^2 E_{[n]}.$$

Получили выражение, аналогичное выражению (8.3.4).

Условное уравнение (8.3.3) с учётом (8.3.7) приобретает вид

$$M_{\hat{Z}} = D^{-1} X A = \overline{X} A.$$

Таким образом, модель наблюдения (8.3.3), (8.3.5) с помощью невырожденного линейного преобразования сводится к соотношениям

$$\hat{Z} = \overline{X} A + \hat{\text{Ê}}, \quad M[\hat{\text{Ê}}] = 0, \quad K_{\hat{\text{Ê}}} = \sigma^2 E.$$

Это и есть исходная модель наблюдения.

Если наблюдения коррелированы и неравноточны, то корреляционная матрица ошибок наблюдения

$$K_{\hat{E}[n]} = R_{[n]}, \tag{8.3.8}$$

где $R_{[n]}$ – известная симметричная положительно-определённая матрица, которая, как и матрица G в формуле (8.3.5), может быть представлена в виде произведения двух невырожденных квадратных матриц аналогично равенству (8.3.6). Это означает, что преобразованием, аналогичным преобразованию (8.3.7), модель наблюдения (8.3.3), (8.3.8) сводится к исходной модели.

По этим причинам в данном подразделе детально рассматривается только наиболее простая модель наблюдения (8.3.3), (8.3.4), а в конце его с помощью преобразования типа (8.3.6) получаются аналогичные результаты для схем наблюдения с корреляционными матрицами ошибок (8.3.5) и (8.3.8).

8.3.2. Нормальные уравнения и оценки наименьших квадратов

Для линейной модели наблюдения (8.3.1) квадратичная функция (8.2.6), при минимизации которой отыскиваются оценки компонентов вектора A, будет иметь вид

$$\hat{V} = (\hat{Y} - XA)^{\mathsf{T}} Q(\hat{Y} - XA), \tag{8.3.9}$$

а нормальные уравнения (8.2.9) – вид

$$X^{\mathsf{T}}QXA - X^{\mathsf{T}}QY = 0. \tag{8.3.10}$$

Можно показать, что система нормальных уравнений (8.3.10) всегда совместна [10]. Будем считать, что матрица X имеет ранг k (предполагается, что $n \geq k$), а матрица весов Q – невырожденная. Тогда матрица $X^{\mathsf{T}}QX$ будет невырожденной, а потому из равенства (8.3.10) можно получить выражение для оценки вектора A:

$$\widetilde{A} = (X^{\mathsf{T}}QX)^{-1} X^{\mathsf{T}}Q\hat{Y}. \tag{8.3.11}$$

Если весовая матрица единичная ($Q = E$), то вместо соотношения (8.3.11) получим равенство

$$\widetilde{A} = (X^{\mathsf{T}}X)^{-1} X^{\mathsf{T}}\hat{Y}. \tag{8.3.12}$$

Если относительно вектора ошибок ничего не известно, то ничего нельзя сказать и о свойствах оценок (8.3.11) или (8.3.12). Если же соотношение (8.3.2) выполняется, то оценки (8.3.11), (8.3.12) являются несмещёнными. Действительно

$$M_{\widetilde{A}} = M[(X^{\mathsf{T}}QX)^{-1} X^{\mathsf{T}}Q\hat{Y}] = M[(X^{\mathsf{T}}QX)^{-1} X^{\mathsf{T}}Q(XA + \hat{E})] =$$
$$= M[(X^{\mathsf{T}}QX)^{-1}(X^{\mathsf{T}}QX)A] + M[(X^{\mathsf{T}}QX)^{-1} X^{\mathsf{T}}Q\hat{E})] =$$
$$= M_A + (X^{\mathsf{T}}QX)^{-1} X^{\mathsf{T}}QM_{\hat{E}} = M_A = A,$$

поскольку $M_{\hat{E}} = 0$.

Пусть корреляционная матрица вектора ошибок наблюдений имеет вид (8.3.4). Вычислим корреляционную матрицу вектора оценок $\widetilde{A}$. Обозначим

$$X^{\mathsf{T}}QX = S_{[n]}; \quad (X^{\mathsf{T}}QX)^{-1} = C_{[n]}.$$

В этих обозначениях формула для МНК-оценки (8.3.11) перепишется в виде

$$\widetilde{A} = CX^{\mathsf{T}}Q\hat{Y} = C\hat{Z},$$

где $\hat{Z} = X^{\mathsf{T}}Q\hat{Y}$.

Тогда корреляционная матрица вектора оценок $\widetilde{A}$ вычисляется по формуле

$$\begin{aligned} K_{\widetilde{A}[k]} &= M[(CX^{\mathsf{T}}Q\hat{Y} - CX^{\mathsf{T}}QM_{\hat{Y}})(CX^{\mathsf{T}}Q\hat{Y} - CX^{\mathsf{T}}QM_{\hat{Y}})^{\mathsf{T}}] = \\ &= CX^{\mathsf{T}}QM[(\hat{Y} - M_{\hat{Y}})(\hat{Y} - M_{\hat{Y}})^{\mathsf{T}}]Q^{\mathsf{T}}XC^{\mathsf{T}} = \\ &= CX^{\mathsf{T}}QK_{\hat{Y}[n]}Q^{\mathsf{T}}XC^{\mathsf{T}} = \sigma^2 CX^{\mathsf{T}}QQ^{\mathsf{T}}XC^{\mathsf{T}}. \end{aligned} \qquad (8.3.13)$$

В процессе преобразований в выражении (8.3.13) учтено, что

$$(CX^{\mathsf{T}}Q)^{\mathsf{T}} = Q^{\mathsf{T}}XC^{\mathsf{T}}, \quad K_{\hat{Y}[n]} = \sigma^2 E_{[n]}.$$

Если $Q = E$, то

$$K_{\widetilde{A}[k]} = \sigma^2 (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1} = \sigma^2 (X^{\mathsf{T}}X)^{-1} = \sigma^2 C. \qquad (8.3.14)$$

На практике наиболее часто используется вариант МНК-оценивания, при котором в качестве весовой матрицы Q выбирается матрица $K_{\hat{Y}}^{-1}$, поскольку в таком случае МНК-оценка получается эффективной. Доказательство этого факта можно найти, например, в [10]. Для такого варианта корреляционная матрица оценки вычисляется по формуле

$$K_{\widetilde{A}} = (X^{\mathsf{T}}K_{\hat{Y}}^{-1}X)^{-1},$$

которая при $K_{\hat{Y}} = \sigma^2 E$ совпадает с выражением (8.3.14).

Когда величина σ^2 известна, формулы (8.3.13), (8.3.14) позволяют отыскать корреляционную матрицу вектора $\widetilde{A}$. Из данных формул следует, что даже при некоррелированных равноточных наблюдениях компоненты вектора оценок оказываются коррелированными. Для их некоррелированности необходима ещё ортогональность столбцов матрицы наблюдений X (при $Q = E$) или столбцов матрицы $CX^{\mathsf{T}}Q$ в общем случае.

Если дисперсия наблюдений σ^2 неизвестна, её необходимо оценить наряду с компонентами вектора A, иначе невозможно отыскать корреляционную матрицу вектора $\widetilde{A}$, которая даёт характеристики точности оценивания.

Покажем, каким образом можно получить оценку величины σ^2. В математической статистике доказывается, что остаточная сумма квадратов

$$\hat{V}(\tilde{A}) = (\hat{Y} - X\tilde{A})^{\mathrm{T}}(\hat{Y} - X\tilde{A}) \qquad (8.3.15)$$

($\tilde{A}$ – МНК-оценка при линейной модели наблюдения) имеет закон распределения $\sigma^2\chi^2_{n-k}$, если вектор ошибок наблюдения $\hat{\mathrm{E}}$ характеризуется корреляционной матрицей $\sigma^2 E$ (n – размерность вектора $\hat{Y}$, k – число оцениваемых параметров).

В скалярной форме выражение (8.3.15) имеет вид

$$V(\tilde{a}_1,\ \tilde{a}_2,...,\tilde{a}_k) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} x_{ij}\tilde{a}_j \right)^2 .$$

На основании свойств $\sigma^2\chi^2_{n-k}$ - распределения получается, что

$$M[V(\tilde{A})] = (n-k)\sigma^2,$$

а потому оценку для величины σ^2 можно приближённо вычислять по формуле

$$\tilde{\sigma}^2 = \frac{V(\tilde{A})}{n-k}. \qquad (8.3.16)$$

Отметим, что это оценка несмещённая.

Без доказательства укажем, что МНК-оценки не всегда получаются эффективными. Ранее уже отмечалось со ссылкой на работу [10], что свойством эффективности обладают МНК-оценки для моделей наблюдения (8.3.3), (8.3.4) при $Q = E$, (8.3.3), (8.3.5) при $Q = G^{-1}$ и (8.3.3), (8.3.8) при $Q = R^{-1}$. Это следует, в частности, для нормального распределения вектора ошибок $\hat{\mathrm{E}}$ из эффективности оценок, получаемых по методу максимального правдоподобия, поскольку МНК и ММП-оценки совпадают при указанном способе выбора весовых матриц. При всех остальных способах выбора матрицы весов МНК-оценки неэффективны. Однако это не означает, что варианты выбора матрицы весов Q, приводящие к неэффективным оценкам, нецелесообразны. Различные соображения могут привести к выбору матрицы весов в другой форме.

П р и м е р 8.2. Известно, что величина a – постоянная, схема оценивания имеет вид

$$\hat{y}_i = a + \hat{\varepsilon}_i, \quad i = \overline{1, n},$$

где $\hat{y}_i$ – наблюдаемая величина, $\hat{\varepsilon}_i$ – ошибка наблюдения. Требуется по результатам наблюдения определить оценку величины a с использованием МНК, а также получить характеристики точности оценивания при известных моментных характеристиках ошибки $\hat{\varepsilon}_i$:

$$M[\hat{\varepsilon}_i] = 0, \ \ M[\hat{\varepsilon}_i^2] = \sigma^2, \quad i = \overline{1, n}.$$

Данная задача совпадает с задачей из примера 8.1, однако там ничего не говорилось о статистических свойствах ошибок измерения $\hat{\varepsilon}_i$.

▼ Случайную величину $\hat{y}$ представим в виде вектора

$$\hat{Y} = (\hat{y}_1,\ \hat{y}_2,...,\hat{y}_n)^{\mathsf{T}}.$$

Роль вектора оцениваемых параметров играет скалярная величина a, т.е. $k = 1$. Поэтому матрица X, имеющая размерность $n \times k$, представляет собой вектор-столбец, состоящий из единиц:

$$X = (1,\ 1,...,1)^{\mathsf{T}}$$

В качестве матрицы весов Q выбираем единичную матрицу $E_{[n]}$. МНК-оценку скалярной величины a вычисляем по формуле (8.3.12). Предварительно найдём матрицы $(X^{\mathsf{T}}X)^{-1}$ и $X^{\mathsf{T}}Y$:

$$X^{\mathsf{T}}X = (1,\ 1,...,1)(1,\ 1,...,1)^{\mathsf{T}} = n;$$

$$(X^{\mathsf{T}}X)^{-1} = \frac{1}{n};$$

$$X^{\mathsf{T}}Y = (1,\ 1,...,1)(\hat{y}_1,\ \hat{y}_2,...,\hat{y}_n)^{\mathsf{T}} = \sum_{i=1}^{n}\hat{y}_i.$$

Окончательно получим

$$\tilde{a} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i.$$

Заметим, что решение совпадает с полученным в примере 8.1.

Эта оценка является несмещённой и эффективной. Определим точность оценки $\tilde{a}$. На основании формулы (8.3.14) получим

$$C = (X^{\mathsf{T}}X)^{-1} = \frac{1}{n},$$

$$K_{\tilde{a}} = \frac{\sigma^2}{n}. \qquad (8.3.17)$$

Если бы величина σ^2 не была априори известной, то её оценку можно было бы получить на основании формулы (8.3.16):

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \tilde{a})^2}{n-1}.$$

Затем можно найти дисперсию оценки $K_{\tilde{a}}$ по формуле (8.3.17) с заменой σ^2 на $\tilde{\sigma}^2$.

Данный пример фактически указывает, каким образом следует оценивать математическое ожидание случайной величины и каковы при этом точечные характеристики этих оценок. Напомним, что ранее тот же результат для оценки математического ожидания был получен с помощью предельных теорем.

▲

П р и м е р 8.3. Рассмотрим задачу, аналогичную приведённой в примере 8.2, с тем отличием, что независимые измерения величины a производятся с различной от эксперимента к эксперименту точностью, которая характеризуется дисперсией σ_i^2, $i=\overline{1,n}$.

▼ Итак, имеем схему неравноточных наблюдений:

$$\hat{y}_i = a + \hat{\varepsilon}_i; \quad \hat{Y}_{<n>} = (\hat{y}_1,\ \hat{y}_2,...,\hat{y}_n)^{\intercal}; \quad diagK_{\hat{Y}[n]} = (\sigma_1^2,\ \sigma_2^2,...,\sigma_n^2).$$

Корреляционная матрица $K_{\hat{Y}}$ взята в диагональной форму в силу независимости измерений.

Данную задачу сводим к предыдущей, используя преобразования (8.3.6), (8.3.7). При этом

$$K_{\hat{Y}[n]} = D_{[n]}D_{[n]}^{\intercal},$$

где D – диагональная матрица с диагональю

$$diagD_{[n]} = (\sigma_1,\ \sigma_2,...,\sigma_n).$$

Тогда формула (8.3.7) приводит к новому вектору

$$\hat{Z} = D^{-1}\hat{Y} = \left(\frac{\hat{y}_1}{\sigma_1}, \frac{\hat{y}_2}{\sigma_2},...,\frac{\hat{y}_n}{\sigma_n}\right)^{\intercal}.$$

Легко убедиться, что корреляционная матрица вектора $\hat{Z}$ получается при этом единичной, а задача сводится к предыдущей. Уравнение (8.3.3) записывается в виде

$$\hat{Z} = D^{-1}XA = \overline{X}A,$$

где $\overline{X} = D^{-1}X = \left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2},...,\frac{1}{\sigma_n}\right)^{\intercal}$.

Далее находим, как и в примере 8.2:

$$\overline{X}^{\intercal}\overline{X} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}; \quad (\overline{X}^{\intercal}\overline{X})^{-1} = \frac{1}{\sum\limits_{i=1}^{n}\frac{1}{\sigma_i^2}}; \quad \overline{X}^{\intercal}\hat{Z} = \overline{X}^{\intercal}D^{-1}\hat{Y} = \sum_{i=1}^{n}\frac{\hat{y}_i}{\sigma_i^2};$$

$$\tilde{a} = (\overline{X}^{\intercal}\overline{X})^{-1}\overline{X}^{\intercal}\hat{Z} = \frac{1}{\sum\limits_{i=1}^{n}\frac{1}{\sigma_i^2}}\sum_{j=1}^{n}\frac{\hat{y}_j}{\sigma_j^2}.$$

Данная оценка является несмещённой и эффективной, причём

$$K_{\tilde{a}} = \frac{1}{\sum\limits_{i=1}^{n}\frac{1}{\sigma_i^2}}.$$

▲

Сравнивая полученное решение с решением задачи 8.1 или 8.2, видим, что учёт неравноточности измерений приводит к иному решению. Возвращаясь к замечанию, указанному в примере 8.1, можно опять же подчеркнуть, что учёт статистических свойств ошибок измерений приводит к необходимости иного выбора функции $V(a)$, чем это было сделано в примере 8.1 или 8.2.

9. МЕТОДЫ РЕГРЕССИОННОГО АНАЛИЗА

9.1. Сущность и задачи регрессионного анализа

Регрессионный анализ – совокупность статистических методов обработки экспериментальных данных, позволяющих в условии стохастической зависимости исследуемой величины от неслучайных или случайных переменных определять данную зависимость.

В дальнейшем будем рассматривать две модели регрессионного анализа (РА).

М о д е л ь 1. В данной модели зависимая переменная $\hat{y}$ – случайная величина, а независимые переменные x_j, $j=\overline{1,k}$ – неслучайные, точно заданные переменные. Таким образом, модель 1 регрессионного анализа имеет вид (8.1.2).

М о д е л ь 2. В данной модели как зависимая переменная, так и независимые переменные являются случайными величинами. Следовательно, модель 2 регрессионного анализа имеет вид (8.1.4).

В дальнейшем регрессионный анализ на основе модели 1 будем называть РА-1, а на основе модели 2 – РА-2. В некоторых источниках РА-2 принято объединять с корреляционным анализом. В данной брошюре РА-2 рассматривается как самостоятельный вид регрессионного анализа, при выполнении которого привлекаются методы корреляционного анализа. Так как РА-1 и РА-2 имеют много общего, то основное внимание уделяется методам РА-1, а для РА-2 показывается лишь специфика соответствующих методов анализа.

Сущность регрессионного анализа состоит в замене стохастической зависимости между переменными $\hat{y}$ и $\hat{x}_j$, $j=\overline{1,k}$ некоторой детерминированной зависимостью *f*, достаточно хорошо аппроксимирующей основные свойства исходной стохастической зависимости. В дальнейшем переменные $\hat{x}_j$, $j=\overline{1,k}$ будем обозначать также вектором $\hat{X}_{<k>}$. Иначе говоря, в процессе регрессионного анализа устанавливается аналитическая зависимость между некоторой характеристикой случайной величины $\hat{y}$ и независимыми переменными $\hat{X}_{<k>}$. Очевидно, что в данном случае возникает проблема выбора соответствующей характеристики случайной величины $\hat{y}$. В регрессионном анализе в качестве такой характеристики используется условное математическое ожидание

$$M[\hat{y} \mid x_1,\ x_2,...,x_k] = M[\hat{y} \mid \hat{x}_1 = x_1,\ \hat{x}_2 = x_2,...,\hat{x}_k = x_k]$$

случайной величины $\hat{y}$ при условии, что независимые переменные $\hat{X}_{<k>}$ приняли определённые значения $X_{<k>}$. Таким образом, сущность регрессионного анализа состоит в замене зависимостей вида (8.1.2) или (8.1.4) зависимостью вида

$$M[\hat{y} \mid x_1,\ x_2,...,x_k] = f(x_1, x_2,..., x_k). \qquad (9.1.1)$$

Выражение (9.1.1) называется ***регрессией***, именно это название и определило наименование методов, объединённых в регрессионном анализе.

Замена стохастической зависимости регрессионной определяет и ограниченность методов регрессионного анализа. Она состоит в том, что данные методы позволяют провести не всестороннее исследование того, как зависит $\hat{y}$ от $\hat{X}_{<k>}$, а лишь один аспект этой стохастической зависимости. Всесторонний анализ имел место, если бы, например, устанавливалась зависимость между законом распределения случайной величины $\hat{y}$ и переменными $\hat{X}_{<k>}$. Тем не менее, с практической точки зрения этот единственный аспект в большинстве случаев является наиболее существенным.

Можно провести классификацию видов регрессионного анализа.

По виду функции f в выражении (9.1.1) регрессионный анализ принято делить на ***линейный***, в котором указанная функция является линейной относительно оцениваемых параметров, т.е.

$$f(x_1, x_2,..., x_k) = \sum_{j=1}^{p} a_j \varphi_j\ , \qquad (9.1.2)$$

и ***нелинейный***, в котором она нелинейная относительно параметров a_j. В выражении (9.1.2) функции φ_j могут определяться одной, несколькими или всеми независимыми переменными.

По числу независимых переменных регрессионный анализ принято подразделять на ***однофакторный***, если имеет место только одна такая переменная, и ***многофакторный***, если число независимых переменных более одной.

Очевидно, что для установления зависимости (9.1.1) необходимо решить ряд задач, которые и составляют собственно регрессионный анализ. К их числу относятся:

1) выбор класса функций, в рамках которого определяется взаимосвязь между $\hat{y}$ и $\hat{X}_{<k>}$;

2) определение подходящих значений параметров a_j, определяющих конкретный вид функции;

3) оценка точности аппроксимации зависимости (8.1.2) или (8.1.4) функцией (9.1.1).

Необходимо отметить, что первая из перечисленных задач формально не решается методами регрессионного анализа. Иначе говоря, класс функции Ψ определяется на основе соображений, которые находятся вне рамок данных методов. Регрессионный анализ позволяет только оценить, насколько удачен этот выбор. При этом наилучшей оценкой зависимости $\hat{y}$ от $X_{<k>}$ в заданном классе Ψ является функция, реализующая минимум математического ожидания квадрата ошибки, т.е. величины

$$\varepsilon(X_{<k>}) = M\left[(\hat{y} - \tilde{y}(X_{<k>}))^2\right]. \qquad (9.1.3)$$

Оценка $\tilde{y}$ случайной величины $\hat{y}$, принадлежащая определённому классу функций Ψ и минимизирующая ошибку (9.1.3), называется средней квадратической регрессией $\hat{y}$ на $X_{<k>}$ класса Ψ.

Вместе с тем некоторые рекомендации по выбору класса функций Ψ могут быть сделаны на основе анализа совокупности результатов наблюдений, в частности, при построении выборочной кривой регрессии. Это можно сделать, по крайней мере, на качественном уровне.

9.2. Однофакторный регрессионный анализ

9.2.1. Модели однофакторного регрессионного комплекса

В однофакторном регрессионном анализе предполагается, что переменная $\hat{y}$ определяется только одной независимой переменной (одним фактором), следовательно, модели РА-1 и РА-2 имеют вид

$$\hat{y} = \hat{f}(x), \quad \hat{y} = \hat{f}(\hat{x})$$

соответственно.

Данные модели могут быть представлены в несколько иной форме, а именно:

$$\hat{y} = f(x) + \hat{\varepsilon}, \qquad (9.2.1)$$

$$\hat{y} = f(\hat{x} + \hat{\delta}) + \hat{\varepsilon}, \qquad (9.2.2)$$

где $\hat{\varepsilon}$ – ошибка результата наблюдения; $\hat{\delta}$ – ошибка наблюдаемого значения фактора.

В ряде источников модели однофакторного регрессионного анализа именуются моделями парной регрессии.

Регрессионный комплекс, соответствующий модели (9.2.1), описывается следующим образом. Пусть проводится исследование некоторой системы, при этом выполняется n опытов. В результате фиксируется n значений параметра x, характеризующего воздействие среды на систему. При каждом значении x_i, $i = \overline{1, n}$ данного параметра фиксируется m значе-

ний наблюдаемого признака $\hat{y}$, который характеризует воздействие системы на среду (рис.9.1). Результаты наблюдений могут быть представлены в виде табл.9.1.

Рис.9.1. Взаимодействие системы и среды

Таблица 9.1

Представление результатов однофакторного эксперимента (модель РА-1)

Значения фактора	Наблюдаемые значения результата						$\bar{y}^*$
	1	2	···	j	···	m	
x_1	y_{11}	y_{12}	···	y_{1j}	···	y_{1m}	$\bar{y}_1^*$
x_2	y_{21}	y_{22}	···	y_{2j}	···	y_{2m}	$\bar{y}_2^*$
...	...	...	...	...	...	...	...
x_i	y_{i1}	y_{i2}	···	y_{ij}	···	y_{im}	$\bar{y}_i^*$
...	...	...	...	...	...	...	...
x_n	y_{n1}	y_{n2}	···	y_{nj}	···	y_{nm}	$\bar{y}_n^*$

В данной таблице кроме результатов наблюдений приведены оценки $\bar{y}_i^*$ условных математических ожиданий результатов наблюдений для различных значений фактора x.

Графически результаты наблюдений могут быть изображены в виде поля корреляции, показанного на рис.9.2.

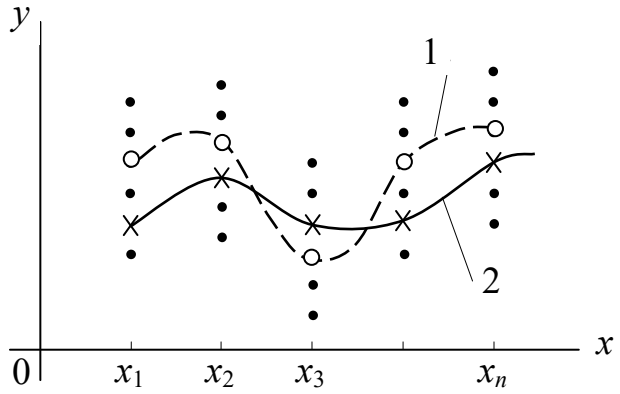


Рис.9.2. Поле корреляции и функции регрессии

На данном поле зачернённые точки соответствуют результатам наблюдений, крестики – значениям оценок математических ожиданий

$$\widetilde{M}[\hat{y} \mid \hat{x} = x_i] = \bar{y}_i^*, \quad i = \overline{1, n},$$

а кружки – значениям математических ожиданий

$$M[\hat{y} \mid \hat{x} = x_i] = \bar{y}_i, \quad i = \overline{1, n}.$$

Предположим, что между $\hat{y}$ и x существует зависимость, которая может быть описана в виде

$$M[\hat{y} \mid \hat{x} = x_i] = f(x_i), \quad i = \overline{1, n}.$$

Тогда кривая, проведённая через значения $M_{\hat{y}}(x_i)$ представляет собой *функцию регрессии генеральной совокупности* (кривая 1, рис.9.2). Так как каждая оценка $\widetilde{M}[\hat{y} \mid \hat{x} = x_i]$ условного математического ожидания $M[\hat{y} \mid \hat{x} = x_i]$ представляет собой несмещённую оценку, то кривая 2, рис.9.2, проходящая через значения $\widetilde{M}[\hat{y} \mid \hat{x} = x_i]$, будет одной из приемлемых оценок функции регрессии генеральной совокупности. Данная кривая называется *выборочной функцией регрессии* или *оценкой функции регрессии*.

Задачей регрессионного анализа является определение выборочной функции регрессии, наилучшим образом (в каком-либо смысле) соответствующей функции регрессии генеральной совокупности.

Для того чтобы данная задача была конструктивной, т.е. допускала решение, вводится ряд предположений, в рамках которых справедливо применение регрессионного анализа. Эти предположения состоят в следующем.

1. Величина x является неслучайной, т.е. задаётся или измеряется без ошибок.

2. Результаты наблюдений получены таким образом, что

$$M[\hat{y}_{ij}] = M_{\hat{y}_i}, \quad i = \overline{1, n}, \; j = \overline{1, m}. \qquad (9.2.3)$$

3. Для каждого x_i распределение величины $\hat{y}$ имеет постоянную дисперсию:

$$D_{\hat{y}_i} = D_{\hat{\varepsilon}_i} = \sigma_i^2 = \sigma^2, \quad i = \overline{1, n}. \qquad (9.2.4)$$

Учитывая (9.2.4), для любого y_i можно записать

$$D[\widetilde{M}_{\hat{y}_i}] = \frac{\sigma^2}{m}, \quad i = \overline{1, n}. \qquad (9.2.5)$$

Так как $\widetilde{M}_{\hat{y}_i}$ является несмещённой оценкой $M_{\hat{y}_i}$, то ошибка

$$\hat{\varepsilon}_i = \widetilde{M}_{\hat{y}_i} - M_{\hat{y}_i}, \quad i = \overline{1, n} \qquad (9.2.6)$$

представляет собой случайную величину с математическим ожиданием

$$M_{\hat{\varepsilon}_i} = M[\widetilde{M}_{\hat{y}_i}] - M_{\hat{y}_i} = M_{\hat{y}_i} - M_{\hat{y}_i} = 0$$

и дисперсией

$$D_{\hat{\varepsilon}_i} = D[\widetilde{M}_{\hat{y}_i}] = \frac{\sigma^2}{m}.$$

4. Величины $\hat{\varepsilon}_i$ и x_i являются стохастически независимыми, так как x_i являются детерминированными и, следовательно, справедливо равенство

$$K_{\hat{\varepsilon}_i x_i} = 0, \quad i = \overline{1, n},$$

где $K_{\hat{\varepsilon}_i x_i}$ - корреляционный момент $\hat{\varepsilon}_i$ и x_i.

5. Результаты наблюдений являются независимыми:

$$K_{\hat{y}_i \hat{y}_l} = K_{\hat{\varepsilon}_i \hat{\varepsilon}_l} = 0, \quad i \neq l, \quad i = \overline{1, n}, \quad l = \overline{1, n}.$$

6. Величины $\hat{y}_i$ и, следовательно, ошибки $\hat{\varepsilon}_i$ распределены по нормальному закону. Необходимо заметить, что отклонения от нормального закона встречаются часто, однако имеют существенное значение только в том случае, если они велики.

В большинстве практических случаев сбор данных или весьма затруднён, или связан с большими затратами. Поэтому нередко каждому значению фактора x соответствует только одно значение результата и тогда $m = 1$, $y_{ij} = y_i$. В связи с тем, что это значение извлекается случайным образом из генеральной совокупности, величина y_i является несмещённой оценкой величины $M[\hat{y} \mid \hat{x} = x_i]$. Учитывая данное обстоятельство, имеем

$$M_{\hat{\varepsilon}_i} = 0, \quad D_{\hat{y}_i} = D_{\hat{\varepsilon}_i} = \sigma^2, \quad i = \overline{1, n}.$$

Регрессионный комплекс, соответствующий модели (9.2.2), значительно отличается от рассмотренного и имеет следующие особенности.

Пусть проводится исследование некоторой системы, при этом выполняется n опытов. В каждом опыте может быть зарегистрировано m значений фактора $\hat{x}$, характеризующего воздействие среды на систему, и столько же соответствующих значений результата $\hat{y}$, который является характеристикой воздействия системы на среду. Поскольку любая точка $(x_{ij}; y_{ij})$ случайным образом извлекается из генеральной совокупности, то её можно рассматривать как результат i-го опыта:

$$(x_{ij}; y_{ij}) = (x_i; y_i), \quad i = \overline{1, n}, \quad i = \overline{1, m}.$$

В этом случае результаты наблюдений могут быть представлены в виде табл.9.2.

Таблица 9.2

Представление результатов однофакторного эксперимента (модель РА-2)

Значения фактора	x_1	x_1	···	x_1	···	x_1
Значения результата	y_1	y_1	···	y_1	···	y_1

Очевидно, что нижняя строка табл.9.2 представляет собой одновременно и оценки условных математических ожиданий:

$$\widetilde{M}[\hat{y} \mid \hat{x} = x_i] = \bar{y}^* = y_i, \quad i = \overline{1, n}.$$

Как и в модели РА-1 графически результаты наблюдений могут быть представлены в виде поля корреляции, рис.9.3.

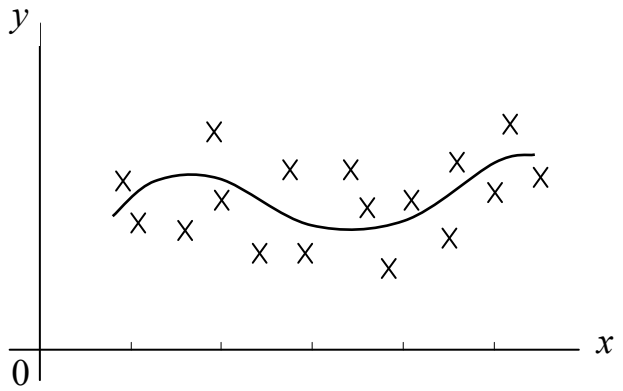


Рис.9.3. Поле корреляции и кривая регрессии

Из сравнения рис.9.2 и 9.3 вида разница между моделями РА-1 и РА-2. Если в первой модели результаты наблюдений рассеивались при определённых значениях фактора x, то во второй результаты располагаются произвольно по всему полю корреляции. Случайный характер значений фактора $\hat{x}$ определяет и особенности РА-2. Эти особенности проявляются как в системе предположений, в рамках которых выполняется РА-2, так и в методах, которые используются при его проведении. Для модели РА-2 справедливы предположения 2, 3, 5, 6.

Так как фактор является случайным, то в общем случае

$$K_{\hat{y}_i\hat{x}_i} = K_{\hat{\varepsilon}_i\hat{x}_i} \neq 0,$$

$$K_{\hat{y}_i\hat{x}_i} = M[(\hat{y}_i - M_{\hat{y}_i})(\hat{x}_i - M_{\hat{x}_i})], \quad i = \overline{1, n}.$$

Особенности методов, используемых в РА-2, рассмотрены ниже.

9.2.2. Построение уравнения регрессии

Пусть из каких-либо соображений выбран класс функций Ψ, которому принадлежит функция регрессии

$$y = f(x). \qquad (9.2.6)$$

Эта функция определяется также и вектором числовых параметров

$$(a_0,\ a_1,\ldots,\ a_k)^{\mathrm{T}} = A_{<k+1>}. \qquad (9.2.7)$$

Поэтому выражение (9.2.6) можно представить в виде

$$y = f(x;\ a_0,\ a_1,\ldots,\ a_k) = f(x;\ A_{<k+1>}). \qquad (9.2.8)$$

Построение уравнения регрессии сводится к решению задачи оценивания параметров (9.2.7), которые называются коэффициентами регрессии. Эта задача может быть решена на основе ряда принципов, являющихся базовыми для статистических методов обработки данных. В практике исследований наиболее широкое применение имеет подход,

опирающийся на принцип максимального правдоподобия и, в частности, подход, использующий метод наименьших квадратов.

В соответствии с данным методом задача сводится к получению подходящей оценки $\widetilde{A}_{<k+1>}$ вектора $A_{<k+1>}$, минимизирующей сумму квадратов отклонений (невязок) наблюдаемых значений результата от выборочной функции регрессии. Указанные невязки представляются выражением

$$e_i = y_i - \widetilde{M}_{\hat{y}_i}, \quad i = \overline{1, n}. \tag{9.2.9}$$

Следовательно, необходимо найти минимальное значение величины

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widetilde{M}_{\hat{y}_i})^2. \tag{9.2.10}$$

Рассмотрим только случай, когда функция регрессии (9.2.8) является линейной относительно оцениваемых параметров:

$$y = f(x;\, a_0,\, a_1,...,a_k) = a_0 f_0(x) + a_1 f_1(x) + ... + a_k f_k(x) = \sum_{j=0}^{k} a_j f_j(x). \tag{9.2.11}$$

Тогда оценки математических ожиданий результата определяются из выражения

$$\widetilde{M}_{\hat{y}} = \widetilde{y} = \sum_{j=0}^{k} \widetilde{a}_j f_j(x). \tag{9.2.12}$$

Принимая во внимание (9.2.12), соотношение (9.2.10) можно записать в виде

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \widetilde{a}_j f_j(x) \right)^2.$$

Следовательно, оценки $\widetilde{a}_0,\, \widetilde{a}_1 ...,\widetilde{a}_k$ должны быть таковы, чтобы выполнялось условие

$$V = \min_{\widetilde{a}_j \in \mathbf{R}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \widetilde{a}_j f_j(x) \right)^2 \right\}. \tag{9.2.13}$$

Из раздела 8 следует, что для нахождения минимума суммы квадратов невязок (9.2.13) необходимо составить систему нормальных уравнений вида (8.2.10). При условии, что используется функция регрессии (9.2.11), указанная система записывается следующим образом:

$$\begin{cases} \sum_{i=1}^{n}\left(y_i-(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i))\right)f_0(x_i)=0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n}\left(y_i-(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i))\right)f_j(x_i)=0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n}\left(y_i-(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i))\right)f_k(x_i)=0 \end{cases} \qquad (9.2.14)$$

В уравнениях (9.2.14) учтено, что

$$\frac{\partial\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{a}_j f_j(x_i)\right)\right)^2}{\partial a_j}=2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{a}_j f_j(x_i)\right)\frac{\partial\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{a}_j f_j(x_i)\right)\right)}{\partial a_j}=$$

$$=2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{a}_j f_j(x_i)\right)(-f_j(x_i)=-2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{a}_j f_j(x_i)\right)f_j(x_i)=$$

$$=-2\sum_{i=1}^{n}\left(y_i-\left(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i)\right)\right)f_j(x_i).$$

(9.2.15)

В выражении (9.2.15) использованы правила дифференцирования сложной функции многих переменных. Поскольку частная производная (9.2.15) приравнивается к нулю, имеем

$$-2\sum_{i=1}^{n}\left(y_i-(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i))\right)f_j(x_i)=0. \qquad (9.2.16)$$

Обе части уравнения (9.2.16) умножаем на –2 и, таким образом, получаем *j*-е уравнение системы (9.2.14):

$$\sum_{i=1}^{n}\left(y_i-\left(\widetilde{a}_0 f_0(x_i)+\widetilde{a}_1 f_1(x_i)+\ldots+\widetilde{a}_k f_k(x_i)\right)\right)f_j(x_i)=0.$$

Выполняем почленное суммирование в уравнениях (9.2.14), слагаемые, содержащие y_i переносим в правую часть, затем умножаем на –1 обе части каждого уравнения.

В результате получаем систему

$$\begin{cases} \widetilde{a}_0\sum\limits_{i=1}^{n} f_0^2(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{n} f_1(x_i)f_0(x_i)+\ldots+\widetilde{a}_k\sum\limits_{i=1}^{n} f_k(x_i)f_0(x_i)=\sum\limits_{i=1}^{n} y_i f_0(x_i) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \widetilde{a}_0\sum\limits_{i=1}^{n} f_0(x_i)f_j(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{n} f_1(x_i)f_j(x_i)+\ldots+\widetilde{a}_k\sum\limits_{i=1}^{n} f_k(x_i)f_j(x_i)=\sum\limits_{i=1}^{n} y_i f_j(x_i) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \widetilde{a}_0\sum\limits_{i=1}^{n} f_0(x_i)f_k(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{n} f_1(x_i)f_k(x_i)+\ldots+\widetilde{a}_k\sum\limits_{i=1}^{n} f_k^2(x_i)=\sum\limits_{i=1}^{n} y_i f_k(x_i) \end{cases} \tag{9.2.17}$$

Уравнения (9.2.17) представляют собой систему линейных уравнений относительно оценок $\widetilde{a}_0$, $\widetilde{a}_1 ..., \widetilde{a}_k$. Следовательно, она решается любым из методов решения систем таких уравнений.

Например, оценки коэффициентов регрессии могут быть найдены по формулам Крамера:

$$\widetilde{a}_0=\frac{|A_0|}{|A|};...;\widetilde{a}_j=\frac{|A_j|}{|A|};...;\widetilde{a}_k=\frac{|A_k|}{|A|}, \tag{9.2.18}$$

где $|A|$ – определитель коэффициентов при неизвестных системы (9.2.17); $|A_j|$, $j=\overline{0,k}$ – определители, которые формируются на основе определителя $|A|$ путём замены j-го столбца столбцом свободных членов.

Таким образом, развёрнутый вид данных определителей будет следующим:

$$|A|=\begin{vmatrix} \sum\limits_{i=1}^{n} f_0^2(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_0(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k(x_i)f_0(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} f_0(x_i)f_j(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_j(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k(x_i)f_j(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} f_0(x_i)f_k(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_k(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k^2(x_i) \end{vmatrix};$$

$$|A_0|=\begin{vmatrix} \sum\limits_{i=1}^{n} y_i f_0(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_0(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k(x_i)f_0(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} y_i f_j(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_j(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k(x_i)f_j(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} y_i f_k(x_i) & \sum\limits_{i=1}^{n} f_1(x_i)f_k(x_i) & \cdots & \sum\limits_{i=1}^{n} f_k^2(x_i) \end{vmatrix};$$

$$|A_j| = \begin{vmatrix} \sum_{i=1}^{n} f_0^2(x_i) & \cdots & \sum_{i=1}^{n} y_i f_0(x_i) & \cdots & \sum_{i=1}^{n} f_k(x_i) f_0(x_i) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} f_0(x_i) f_j(x_i) & \cdots & \sum_{i=1}^{n} y_i f_j(x_i) & \cdots & \sum_{i=1}^{n} f_k(x_i) f_j(x_i) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} f_0(x_i) f_k(x_i) & \cdots & \sum_{i=1}^{n} y_i f_k(x_i) & \cdots & \sum_{i=1}^{n} f_k^2(x_i) \end{vmatrix};$$

$$|A_k| = \begin{vmatrix} \sum_{i=1}^{n} f_0^2(x_i) & \sum_{i=1}^{n} f_1(x_i) f_0(x_i) & \cdots & \sum_{i=1}^{n} y_i f_0(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} f_0(x_i) f_j(x_i) & \sum_{i=1}^{n} f_1(x_i) f_j(x_i) & \cdots & \sum_{i=1}^{n} y_i f_j(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} f_0(x_i) f_k(x_i) & \sum_{i=1}^{n} f_1(x_i) f_k(x_i) & \cdots & \sum_{i=1}^{n} y_i f_k(x_i) \end{vmatrix}.$$

Запишем систему уравнений (9.2.17) в матричной форме:

$$(F^{\mathsf{T}}_{[k+1;n]} F_{[n;k+1]}) \widetilde{A}_{<k+1>} = F^{\mathsf{T}}_{[k+1;n]} Y_{<n>}, \qquad (9.2.19)$$

где

$$F = \begin{pmatrix} f_0(x_1) & f_1(x_1) & \cdots & f_k(x_1) \\ \cdots & \cdots & \cdots & \cdots \\ f_0(x_i) & f_1(x_i) & \cdots & f_k(x_i) \\ \cdots & \cdots & \cdots & \cdots \\ f_0(x_n) & f_1(x_n) & \cdots & f_k(x_n) \end{pmatrix};$$

$$\widetilde{A} = (\widetilde{a}_0, \widetilde{a}_1, ..., \widetilde{a}_{k+1})^{\mathsf{T}}; \quad Y = (y_1, y_2, ..., y_n)^{\mathsf{T}}.$$

Умножим слева обе части матричного уравнения (9.2.19) на квадратную матрицу $(F^{\mathsf{T}}F)^{-1}_{[k+1]}$:

$$(F^{\mathsf{T}}F)^{-1}(F^{\mathsf{T}}F)\widetilde{A} = (F^{\mathsf{T}}F)^{-1} F^{\mathsf{T}} Y.$$

Далее учитываем, что

$$(F^{\mathsf{T}}F)^{-1}(F^{\mathsf{T}}F) = E, \quad E\widetilde{A} = \widetilde{A},$$

где $E = E_{[k+1]}$ – единичная матрица порядка k+1.

Окончательно получаем выражение для вычисления оценок искомых параметров:

$$\widetilde{A}_{<k+1>} = (F^{\mathsf{T}}_{[k+1;n]} F_{[n;k+1]})^{-1} F^{\mathsf{T}}_{[k+1;n]} Y_{<n>}. \qquad (9.2.20)$$

Рассмотренный выше метод определения оценок коэффициентов регрессии без каких-либо изменений применим и в модели РА-2. Однако при этом необходимо учитывать то обстоятельство, что в модели РА-1

зависимость между $\hat{y}$ и x является односторонней. При рассмотрении же модели РА-2 исследователь сталкивается со взаимностью зависимости между $\hat{y}$ и $\hat{x}$. Поэтому в последнем случае правомерна формулировка задач двух типов.

1. Исследование зависимости $\hat{y}$ от $\hat{x}$ и построение уравнения регрессии $\hat{x}$ на $\hat{y}$:

$$\widetilde{y} = \sum_{j=0}^{k} \widetilde{a}_j f_j(x).$$

2. Исследование зависимости $\hat{x}$ на $\hat{y}$ и построение уравнения регрессии $\hat{y}$ на $\hat{x}$:

$$\widetilde{x} = \sum_{j=0}^{k} \widetilde{c}_j \varphi_j(y),$$

где $\widetilde{c}_j$ – оценки коэффициентов регрессии.

9.2.3. Проверка адекватности уравнения регрессии

Под ***адекватностью*** уравнения регрессии понимается соответствие данного уравнения экспериментальным данным.

Адекватность уравнения определяется, прежде всего, правильным выбором класса Ψ функций регрессии.

Для проверки соответствия выбранного класса функций регрессии опытным данным рассмотрим очевидное тождество

$$\hat{y}_i - \bar{y}^* = (\hat{y}_i - \widetilde{y}_i) - (\bar{y}^* - \widetilde{y}_i),\ i = \overline{1, n}. \qquad (9.2.21)$$

Возведём обе части (9.2.21) в квадрат и найдём для них математические ожидания:

$$M[(\hat{y}_i - \bar{y}^*)^2] = M[((\hat{y}_i - \widetilde{y}_i) - (\bar{y}^* - \widetilde{y}_i))^2] =$$
$$= M[(\hat{y}_i - \widetilde{y}_i)^2 - 2(\hat{y}_i - \widetilde{y}_i)(\bar{y}^* - \widetilde{y}_i) + (\bar{y}^* - \widetilde{y}_i)^2] =$$
$$= M[(\hat{y}_i - \widetilde{y}_i)^2] - 2M[(\hat{y}_i - \widetilde{y}_i)(\bar{y}^* - \widetilde{y}_i)] + M[(\bar{y}^* - \widetilde{y}_i)^2] =$$
$$= M[(\hat{y}_i - \widetilde{y}_i)^2] + M[(\bar{y}^* - \widetilde{y}_i)^2],$$

так как

$$M[(\hat{y}_i - \widetilde{y}_i)(\bar{y}^* - \widetilde{y}_i)] = M[\hat{y}_i - \widetilde{y}_i]M[\bar{y}^* - \widetilde{y}_i] = 0.$$

В результате получим

$$M[(\hat{y}_i - \bar{y}^*)^2] = M[(\hat{y}_i - \widetilde{y}_i)^2] + M[(\widetilde{y}_i - \bar{y}^*)^2]$$

или

$$\sigma^2 = \sigma_1^2 + \sigma_2^2, \qquad (9.2.22)$$

где σ^2 – общая дисперсия результатов наблюдений (дисперсия выходной переменной), σ_1^2 – дисперсия, характеризующая рассеивание результатов наблюдений относительно регрессионной зависимости (остаточная дисперсия); σ_2^2 – дисперсия, характеризующая отклонение регрессионной зависимости относительно истинной.

Чем меньше рассеивание результатов относительно регрессионной зависимости, тем лучше последняя аппроксимирует истинную (но неизвестную) зависимость (удовлетворяет экспериментальным данным). Таким образом, чем больше отношение σ^2/σ_1^2, тем более адекватно уравнение регрессии. Поэтому выражение

$$\hat{F} = \frac{\tilde{\sigma}^2}{\tilde{\sigma}_1^2}, \qquad (9.2.23)$$

принимается в качестве показателя согласованности при проверке нулевой гипотезы H_0 о соответствии выбранного класса функций регрессии экспериментальным данным. В связи с тем, что случайная величина (9.2.23) подчиняется закону распределения Фишера [1], критерием правильности гипотезы H_0 является выполнение неравенства

$$F = \frac{\tilde{\sigma}^2}{\tilde{\sigma}_1^2} > F_{(\alpha;n\text{-}1;n-k\text{-}1)}, \qquad (9.2.24)$$

где F – наблюдаемое (вычисленное) значение показателя согласованности гипотезы H_0; $F_{(\alpha;n\text{-}1;n-k\text{-}1)}$ – критическое значение данного показателя при уровне значимости α и степенях свободы $f_1 = n - 1$, $f_2 = n - k - 1$. Критическое значение берётся по таблице критических точек распределения Фишера (приложение 5). Входами в таблицу являются величины α, f_1, f_2.

Оценки дисперсий $\tilde{\sigma}^2$ и $\tilde{\sigma}_1^2$ находятся по формулам:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}; \qquad (9.2.25)$$

$$\tilde{\sigma}_1^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-k-1}. \qquad (9.2.26)$$

В случае невыполнения условия (9.2.24) принимается конкурирующая гипотеза H_1, т.е. возникает проблема выбора другого класса функций регрессии, более точно соответствующего экспериментальным данным. Проверка неравенства (9.2.24) называется проверкой адекватности уравнения регрессии по критерию Фишера.

9.2.4. Проверка значимости коэффициентов регрессии

Важным аспектом качества регрессионной зависимости является значимость коэффициентов регрессии a_j, $j = \overline{0, k}$. Поскольку оценки этих коэффициентов получены на основе случайной выборки, возникает задача проверки того, что их величина определяется именно видом зависимости между $\hat{y}$ и x, а не случайным характером выборки. Эта задача сводится к проверке гипотез о неравенстве нулю коэффициентов a_j:

$$H_{0j}: a_j \neq 0, \quad j = \overline{0, k}$$

при соответствующих конкурирующих гипотезах H_{1j}: $a_j = 0$.

В качестве показателя согласованности при проверке гипотез H_{0j} используется выражение

$$\hat{t}_j = \frac{|\widetilde{a}_j|}{\widetilde{\sigma}_{\widetilde{a}_j}}, \quad j = \overline{0, k}, \tag{9.2.27}$$

где $|\widetilde{a}_j|$ – модуль величины $\widetilde{a}_j$; $\widetilde{\sigma}_{\widetilde{a}_j}$ – оценка среднего квадратического отклонения коэффициента $\widetilde{a}_j$.

Получение оценок $\widetilde{\sigma}_{\widetilde{a}_j}$ осуществляется так же, как это было описано в п.п.8.3.2. Здесь учтём специфику рассматриваемого вида функций регрессии (9.2.11) и уточним данную процедуру.

В соответствии с выражением (8.3.14) можно написать:

$$K_{\widetilde{A}[k+1]} = \widetilde{\sigma}_1^2 (F^{\mathsf{T}}_{[k+1;\,n]} F_{[n;\,k+1]})^{-1}. \tag{9.2.28}$$

На главной диагонали корреляционной матрицы (9.2.28) вектора $\widetilde{A}_{<k+1>}$ будут дисперсии оценок коэффициентов регрессии:

$$diagK_{\widetilde{A}} = \left(\widetilde{\sigma}^2_{\widetilde{a}_0}, \widetilde{\sigma}^2_{\widetilde{a}_1}, ..., \widetilde{\sigma}^2_{\widetilde{a}_k}\right). \tag{9.2.29}$$

Случайная величина (9.2.27) подчинена закону распределения Стьюдента с $f = n - k - 1$ степенями свободы [1, 4]. Поэтому критерием правильности гипотез H_{0j} является выполнение неравенств

$$t_j = \frac{|\widetilde{a}_j|}{\widetilde{\sigma}_{\widetilde{a}_j}} > t_{(\alpha;\,n-k-1)}, \; j = \overline{0, k}, \tag{9.2.30}$$

где t_j – наблюдаемое значение показателя согласованности гипотезы H_{0j}; $t_{(\alpha;\,n-k-1)}$ – критическое значение данного показателя при уровне значимости α и f степенях свободы.

Критическое значение $t_{(\alpha;\,n-k-1)}$ берётся из таблицы критических точек распределения Стьюдента (приложение 6). Очевидно, что входами в таблицу являются α и f.

Коэффициенты регрессии, для которых не выполняется условие (9.2.30), принимаются равными нулю. Проверка неравенств (9.2.30) назы-

вается проверкой значимости коэффициентов регрессии по критерию Стьюдента.

Следует отметить, что при решении практических задач встречаются случаи, когда после исключения незначимых коэффициентов в соответствии с критерием Стьюдента регрессионная зависимость становится неадекватной. Такие случаи вызывают сомнения в универсальности этого широко применяемого в настоящее время критерия. Поэтому после исключения незначимых коэффициентов проверку адекватности уравнения регрессии по критерию Фишера целесообразно повторить.

9.2.5. Примеры однофакторного регрессионного анализа

Из вышеизложенного следует, что однофакторный регрессионный анализ проводится в следующей последовательности.

1. Выбирается вид функции регрессии.
2. Составляется система нормальных уравнений.
3. Находится решение системы нормальных уравнений (определяются оценки коэффициентов регрессии).
4. Проверяется адекватность построенного уравнения регрессии экспериментальным данным.
5. Проверяется значимость коэффициентов регрессии.
6. Повторно проверяется адекватность уравнения регрессии после исключения незначимых коэффициентов.

П р и м е р 9.1. Для исследования зависимости выходного сигнала y системы от входного воздействия x проведены испытания, результаты которых сведены в табл.9.3.

Необходимо построить уравнение регрессии $y = f(x)$ в предположении, что оно является алгебраическим полиномом третьей степени. Расчёты произвести в скалярной форме.

Т а б л и ц а 9.3

Массив экспериментальных данных

x	–2	–1	0	1	3
y	5	4	5	2	–39

▼ 1. Класс функций Ψ задан в условии задачи – это полиномы третьей степени

$$y = a_0x^3 + a_1x^2 + a_2x + a_3. \qquad (9.2.31)$$

Они являются функциями вида (9.2.11). Для данного случая выражение (9.2.11) представляется как

$$y = a_0f_0(x) + a_1f_1(x) + a_2f_2(x) + a_3f_3(x), \qquad (9.2.32)$$

где $$f_0(x) = x^3, \quad f_1(x) = x^2, \quad f_2(x) = x, \quad f_3(x) = 1. \qquad (9.2.33)$$

2. Для функции (9.2.32) и заданного количества опытов система нормальных уравнений (9.2.17) принимает вид

$$\begin{cases} \widetilde{a}_0\sum\limits_{i=1}^{5} f_0^2(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{5} f_1(x_i)f_0(x_i)+\widetilde{a}_2\sum\limits_{i=1}^{5} f_2(x_i)f_0(x_i)+\widetilde{a}_3\sum\limits_{i=1}^{5} f_3(x_i)f_0(x_i)=\sum\limits_{i=1}^{5} y_i f_0(x_i) \\ \widetilde{a}_0\sum\limits_{i=1}^{5} f_0(x_i)f_1(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{5} f_1^2(x_i)+\widetilde{a}_2\sum\limits_{i=1}^{5} f_2(x_i)f_1(x_i)+\widetilde{a}_3\sum\limits_{i=1}^{5} f_3(x_i)f_1(x_i)=\sum\limits_{i=1}^{5} y_i f_1(x_i) \\ \widetilde{a}_0\sum\limits_{i=1}^{5} f_0(x_i)f_2(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{5} f_1(x_i)f_2(x_i)+\widetilde{a}_2\sum\limits_{i=1}^{5} f_2^2(x_i)+\widetilde{a}_3\sum\limits_{i=1}^{n} f_3(x_i)f_2(x_i)=\sum\limits_{i=1}^{5} y_i f_2(x_i) \\ \widetilde{a}_0\sum\limits_{i=1}^{5} f_0(x_i)f_3(x_i)+\widetilde{a}_1\sum\limits_{i=1}^{5} f_1(x_i)f_3(x_i)+\widetilde{a}_2\sum\limits_{i=1}^{5} f_2(x_i)f_3(x_i)+\widetilde{a}_3\sum\limits_{i=1}^{5} f_3^2(x_i)=\sum\limits_{i=1}^{5} y_i f_3(x_i) \end{cases} \tag{9.2.34}$$

С учётом (9.2.33) система уравнений (9.2.34) преобразуется следующим образом:

$$\begin{cases} \widetilde{a}_0\sum\limits_{i=1}^{5} x_i^6+\widetilde{a}_1\sum\limits_{i=1}^{5} x_i^5+\widetilde{a}_2\sum\limits_{i=1}^{5} x_i^4+\widetilde{a}_3\sum\limits_{i=1}^{5} x_i^3=\sum\limits_{i=1}^{5} y_i x_i^3 \\ \widetilde{a}_0\sum\limits_{i=1}^{5} x_i^5+\widetilde{a}_1\sum\limits_{i=1}^{5} x_i^4+\widetilde{a}_2\sum\limits_{i=1}^{5} x_i^3+\widetilde{a}_3\sum\limits_{i=1}^{5} x_i^2=\sum\limits_{i=1}^{5} y_i x_i^2 \\ \widetilde{a}_0\sum\limits_{i=1}^{5} x_i^4+\widetilde{a}_1\sum\limits_{i=1}^{5} x_i^3+\widetilde{a}_2\sum\limits_{i=1}^{5} x_i^2+\widetilde{a}_3\sum\limits_{i=1}^{5} x_i=\sum\limits_{i=1}^{5} y_i x_i \\ \widetilde{a}_0\sum\limits_{i=1}^{5} x_i^3+\widetilde{a}_1\sum\limits_{i=1}^{5} x_i^2+\widetilde{a}_2\sum\limits_{i=1}^{5} x_i+\widetilde{a}_3 n=\sum\limits_{i=1}^{5} y_i \end{cases} \tag{9.2.35}$$

3. Оценки коэффициентов уравнения регрессии (т.е. решение системы линейных уравнений (9.2.35)) находим по формулам (9.2.18):

$$\widetilde{a}_0=\frac{|A_0|}{|A|};\quad \widetilde{a}_1=\frac{|A_1|}{|A|};\quad \widetilde{a}_2=\frac{|A_2|}{|A|};\quad \widetilde{a}_3=\frac{|A_3|}{|A|}. \tag{9.2.36}$$

Развёрнутый вид определителей в соотношениях (9.2.36):

$$|A|=\begin{vmatrix} \sum\limits_{i=1}^{5} x_i^6 & \sum\limits_{i=1}^{5} x_i^5 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 \\ \sum\limits_{i=1}^{5} x_i^5 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 \\ \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i \\ \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i & n \end{vmatrix};\qquad |A_0|=\begin{vmatrix} \sum\limits_{i=1}^{5} y_i x_i^3 & \sum\limits_{i=1}^{5} x_i^5 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 \\ \sum\limits_{i=1}^{5} y_i x_i^2 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 \\ \sum\limits_{i=1}^{5} y_i x_i & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i \\ \sum\limits_{i=1}^{5} y_i & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i & n \end{vmatrix};$$

$$|A_1|=\begin{vmatrix} \sum_{i=1}^{5}x_i^6 & \sum_{i=1}^{5}y_ix_i^3 & \sum_{i=1}^{5}x_i^4 & \sum_{i=1}^{5}x_i^3 \\ \sum_{i=1}^{5}x_i^5 & \sum_{i=1}^{5}y_ix_i^2 & \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}x_i^2 \\ \sum_{i=1}^{5}x_i^{4`} & \sum_{i=1}^{5}y_ix_i & \sum_{i=1}^{5}x_i^2 & \sum_{i=1}^{5}x_i \\ \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}y_i & \sum_{i=1}^{5}x_i & n \end{vmatrix};\quad |A_2|=\begin{vmatrix} \sum_{i=1}^{5}x_i^6 & \sum_{i=1}^{5}x_i^5 & \sum_{i=1}^{5}y_ix_i^3 & \sum_{i=1}^{5}x_i^3 \\ \sum_{i=1}^{5}x_i^5 & \sum_{i=1}^{5}x_i^4 & \sum_{i=1}^{5}y_ix_i^2 & \sum_{i=1}^{5}x_i^2 \\ \sum_{i=1}^{5}x_i^{4`} & \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}y_ix_i & \sum_{i=1}^{5}x_i \\ \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}x_i^2 & \sum_{i=1}^{5}y_i & n \end{vmatrix};$$

$$|A_3|=\begin{vmatrix} \sum_{i=1}^{5}x_i^6 & \sum_{i=1}^{5}x_i^5 & \sum_{i=1}^{5}x_i^4 & \sum_{i=1}^{5}y_ix_i^3 \\ \sum_{i=1}^{5}x_i^5 & \sum_{i=1}^{5}x_i^4 & \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}y_ix_i^2 \\ \sum_{i=1}^{5}x_i^{4`} & \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}x_i^2 & \sum_{i=1}^{5}y_ix_i \\ \sum_{i=1}^{5}x_i^3 & \sum_{i=1}^{5}x_i^2 & \sum_{i=1}^{5}x_i & \sum_{i=1}^{5}y_i \end{vmatrix}.$$

Составляем расчётную таблицу 9.4 для вычисления коэффициентов системы линейных уравнений (9.2.35).

Таблица 9.4

Расчётная таблица

x_i	y_i	x_i^2	x_i^3	x_i^4	x_i^5	x_i^6	y_ix_i	$y_ix_i^2$	$y_ix_i^3$
1	2	3	4	5	6	7	8	9	10
–2	5	4	–8	16	–32	64	–10	20	–40
–1	4	1	–1	1	–1	1	–4	4	–4
0	5	0	0	0	0	0	0	0	0
1	2	1	1	1	1	1	2	2	2
3	–39	9	27	81	243	729	–117	–351	–1053
$\sum_{i=1}^{5}x_i=$ $=1$	$\sum_{i=1}^{5}y_i=$ $=-23$	$\sum_{i=1}^{5}x_i^2=$ $=15$	$\sum_{i=1}^{5}x_i^3=$ $=19$	$\sum_{i=1}^{5}x_i^4=$ $=99$	$\sum_{i=1}^{5}x_i^5=$ $=211$	$\sum_{i=1}^{5}x_i^6=$ $=795$	$\sum_{i=1}^{5}y_ix_i=$ $=-129$	$\sum_{i=1}^{5}y_ix_i^2=$ $=-325$	$\sum_{i=1}^{5}y_ix_i^3=$ $=-1095$

Получаем систему уравнений (9.2.35) с числовыми значениями коэффициентов при неизвестных:

$$\begin{cases} 795\widetilde{a}_0+211\widetilde{a}_1+99\widetilde{a}_2+19\widetilde{a}_3=-1095 \\ 211\widetilde{a}_0+99\widetilde{a}_1+19\widetilde{a}_2+15\widetilde{a}_3=-325 \\ 99\widetilde{a}_0+19\widetilde{a}_1+15\widetilde{a}_2+\widetilde{a}_3=-19 \\ 19\widetilde{a}_0+15\widetilde{a}_1+\widetilde{a}_2+5\widetilde{a}_3=-23. \end{cases}$$

Определитель $|A|$ четвёртого порядка вычисляем разложением по первой строке:

$$|A|=\begin{vmatrix}795 & 211 & 99 & 19\\ 211 & 99 & 19 & 15\\ 99 & 19 & 15 & 1\\ 19 & 15 & 1 & 5\end{vmatrix}=795(-1)^2\begin{vmatrix}99 & 19 & 15\\ 19 & 15 & 1\\ 15 & 1 & 5\end{vmatrix}+$$

$$+211(-1)^3\begin{vmatrix}211 & 19 & 15\\ 99 & 15 & 1\\ 19 & 1 & 5\end{vmatrix}+99(-1)^4\begin{vmatrix}211 & 99 & 15\\ 99 & 19 & 1\\ 19 & 15 & 5\end{vmatrix}+19(-1)^5\begin{vmatrix}211 & 99 & 19\\ 99 & 19 & 15\\ 19 & 15 & 1\end{vmatrix}=$$

$$=795\cdot 2716-211\cdot 3780+99(-13384)-19(-3696)=106848.$$

При разложении определителя $|A|$ получены четыре определителя третьего порядка, которые вычисляем также разложением по первой строке:

$$\begin{vmatrix}99 & 19 & 15\\ 19 & 15 & 1\\ 15 & 1 & 5\end{vmatrix}=99(-1)^2\begin{vmatrix}15 & 1\\ 1 & 5\end{vmatrix}+19(-1)^3\begin{vmatrix}19 & 1\\ 15 & 5\end{vmatrix}+15(-1)^4\begin{vmatrix}19 & 15\\ 15 & 1\end{vmatrix}=$$

$$=99\cdot 74-19\cdot 80+15(-206)=2716;$$

$$\begin{vmatrix}211 & 19 & 15\\ 99 & 15 & 1\\ 19 & 1 & 5\end{vmatrix}=211(-1)^2\begin{vmatrix}15 & 1\\ 1 & 5\end{vmatrix}+19(-1)^3\begin{vmatrix}99 & 1\\ 19 & 5\end{vmatrix}+15(-1)^4\begin{vmatrix}99 & 15\\ 19 & 1\end{vmatrix}=$$

$$=211\cdot 74-19\cdot 476+15(-186)=3780;$$

$$\begin{vmatrix}211 & 99 & 15\\ 99 & 19 & 1\\ 19 & 15 & 5\end{vmatrix}=211(-1)^2\begin{vmatrix}19 & 1\\ 15 & 5\end{vmatrix}+99(-1)^3\begin{vmatrix}99 & 1\\ 19 & 5\end{vmatrix}+15(-1)^4\begin{vmatrix}99 & 19\\ 19 & 15\end{vmatrix}=$$

$$=211\cdot 80-99\cdot 476+15\cdot 1124=-13384;$$

$$\begin{vmatrix}211 & 99 & 19\\ 99 & 19 & 15\\ 19 & 15 & 1\end{vmatrix}=211(-1)^2\begin{vmatrix}19 & 15\\ 15 & 1\end{vmatrix}-99(-1)^3\begin{vmatrix}99 & 15\\ 19 & 1\end{vmatrix}+19(-1)^4\begin{vmatrix}99 & 19\\ 19 & 15\end{vmatrix}=$$

$$=211(-206)+99(-186)+19\cdot 1124=-3696.$$

Аналогично находим

$$|A_0|=\begin{vmatrix}-1095 & 211 & 99 & 19\\ -325 & 99 & 19 & 15\\ -129 & 19 & 15 & 1\\ -23 & 15 & 1 & 5\end{vmatrix}=-103992;$$

$$|A_1| = \begin{vmatrix} 795 & -1095 & 99 & 19 \\ 211 & -325 & 19 & 15 \\ 99 & -129 & 15 & 1 \\ 19 & -23 & 1 & 5 \end{vmatrix} = -209016;$$

$$|A_2| = \begin{vmatrix} 795 & 211 & -1095 & 19 \\ 211 & 99 & -325 & 15 \\ 99 & 19 & -129 & 1 \\ 19 & 15 & -23 & 5 \end{vmatrix} = -3216;$$

$$|A_3| = \begin{vmatrix} 795 & 211 & 99 & -1095 \\ 211 & 99 & 19 & -325 \\ 99 & 19 & 15 & -129 \\ 19 & 15 & 1 & -23 \end{vmatrix} = 531360.$$

По формулам (9.2.36) вычисляем оценки коэффициентов регрессии:

$$\widetilde{a}_0 = \frac{-103992}{106848} = -0{,}97; \quad \widetilde{a}_1 = \frac{-209016}{106848} = -1{,}96;$$

$$\widetilde{a}_2 = \frac{-3216}{106848} = -0{,}03; \quad \widetilde{a}_3 = \frac{531360}{106848} = 4{,}97.$$

Получили уравнение регрессии

$$\widetilde{y} = -0{,}97x^3 - 1{,}96x^2 - 0{,}03x + 4{,}97.$$

4. Проверяем адекватность регрессионной зависимости экспериментальным данным. Для этого необходимо вычислить оценки дисперсий (9.2.25) и (9.2.26). Составляем табл.9.5.

Таблица 9.5

Расчётная таблица

x_i	y_i	$y_i - \bar{y}^*$	$(y_i - \bar{y}^*)^2$	$\widetilde{y}_i$	$y_i - \widetilde{y}_i$	$(y_i - \widetilde{y}_i)^2$
–2	5	9,6	92,16	4,95	0,05	0,0025
–1	4	8,6	73,96	4,01	–0,01	0,0001
0	5	9,6	92,16	4,97	0,03	0,0009
1	2	6,6	43,56	2,01	–0,01	0,0001
3	–39	–34,4	1183	–38,95	–0,05	0,0025
$\sum_{i=1}^{5}(y_i - \bar{y}^*)^2 = 1485$				$\sum_{i=1}^{5}(y_i - \widetilde{y}_i)^2 = 0{,}0061$		

Оценка математического ожидания выходной переменной *y*, которая используется при расчётах в табл.9.5, найдена по формуле (5.1.1):

$$\bar{y}^* = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5}(5+4+5+2-39) = -4{,}6\,.$$

Необходимые данные для вычисления оценок дисперсий σ^2 и σ_1^2 берём из табл.9.5:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}^*)^2}{n-1} = \frac{1485}{4} = 371{,}25\,;$$

$$\tilde{\sigma}_1^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-k-1} = \frac{0{,}0061}{5-4} = 0{,}0061\,.$$

Наблюдаемое значение показателя согласованности (9.2.23):

$$\mathsf{F} = \frac{\tilde{\sigma}^2}{\tilde{\sigma}_1^2} = \frac{371{,}25}{0{,}0061} = 60860\,.$$

Критическое значение показателя согласованности при уровне значимости $\alpha = 0{,}01$ и степенях свободы $f_1 = n - 1 = 4$, $f_2 = n - k - 1 = 1$, находим в приложении 5:

$$\mathsf{F}_{(0{,}01;\,4;\,1)} = 5625.$$

Очевидно, что неравенство (9.2.24) выполняется, т.е. $\mathsf{F} > \mathsf{F}_{(0{,}01;\,4;\,1)}$. Следовательно, нулевая гипотеза H_0 о соответствии функции регрессии вида (9.2.31) экспериментальным данным принимается.

5. Проверяем значимость коэффициентов уравнения регрессии.

Для вычисления наблюдаемых значений показателя согласованности (9.2.30) необходимо найти диагональные элементы корреляционной матрицы (9.2.28). В рассматриваемом примере матрица F представляется следующим образом:

$$F = \begin{pmatrix} f_0(x_1) & f_1(x_1) & f_2(x_1) & f_3(x_1) \\ f_0(x_2) & f_1(x_2) & f_2(x_2) & f_3(x_2) \\ f_0(x_3) & f_1(x_3) & f_2(x_3) & f_3(x_3) \\ f_0(x_4) & f_1(x_4) & f_2(x_4) & f_3(x_4) \\ f_0(x_5) & f_1(x_5) & f_2(x_5) & f_3(x_5) \end{pmatrix} = \begin{pmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \\ x_4^3 & x_4^2 & x_4 & 1 \\ x_5^3 & x_5^2 & x_5 & 1 \end{pmatrix}. \quad (9.2.37)$$

Находим произведение транспонированной матрицы F на исходную:

$$F^{\mathsf{T}}_{[5;4]} F_{[4;5]} = \begin{pmatrix} x_1^3 & x_2^3 & x_3^3 & x_4^3 & x_5^3 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & x_5^2 \\ x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ x_3^3 & x_3^2 & x_3 & 1 \\ x_4^3 & x_4^2 & x_4 & 1 \\ x_5^3 & x_5^2 & x_5 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} \sum\limits_{i=1}^{5} x_i^6 & \sum\limits_{i=1}^{5} x_i^5 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 \\ \sum\limits_{i=1}^{5} x_i^5 & \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 \\ \sum\limits_{i=1}^{5} x_i^4 & \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i \\ \sum\limits_{i=1}^{5} x_i^3 & \sum\limits_{i=1}^{5} x_i^2 & \sum\limits_{i=1}^{5} x_i & 5 \end{pmatrix} = \begin{pmatrix} 795 & 211 & 99 & 19 \\ 211 & 99 & 19 & 15 \\ 99 & 19 & 15 & 1 \\ 19 & 15 & 1 & 5 \end{pmatrix}. \quad (9.2.38)$$

Далее требуется найти элементы главной диагонали матрицы $(F^{\intercal}F)^{-1}$:

$$diag(F^{\intercal}_{[5;4]}F_{[4;5]})^{-1} = \left(\frac{A_{11}}{|F^{\intercal}F|}, \frac{A_{22}}{|F^{\intercal}F|}, \frac{A_{33}}{|F^{\intercal}F|}, \frac{A_{44}}{|F^{\intercal}F|} \right),$$

где $|F^{\intercal}F|$ – определитель матрицы $F^{\intercal}F$; A_{ii}, $i = \overline{1,4}$ – алгебраические дополнения элементов главной диагонали этой же матрицы.

Вычисляем определитель и алгебраические дополнения:

$$|F^{\intercal}F| = \begin{vmatrix} 795 & 211 & 99 & 19 \\ 211 & 99 & 19 & 15 \\ 99 & 19 & 15 & 1 \\ 19 & 15 & 1 & 5 \end{vmatrix} = 106848;$$

$$A_{11} = (-1)^2 \begin{vmatrix} 99 & 19 & 15 \\ 19 & 15 & 1 \\ 15 & 1 & 5 \end{vmatrix} = 2716; \quad A_{22} = (-1)^4 \begin{vmatrix} 795 & 99 & 19 \\ 99 & 15 & 1 \\ 19 & 1 & 5 \end{vmatrix} = 8172;$$

$$A_{33} = (-1)^6 \begin{vmatrix} 795 & 211 & 19 \\ 211 & 99 & 1 \\ 19 & 15 & 5 \end{vmatrix} = 76576; \quad A_{44} = (-1)^8 \begin{vmatrix} 795 & 211 & 99 \\ 211 & 99 & 19 \\ 99 & 19 & 15 \end{vmatrix} = 49248.$$

Таким образом, имеем

$$diag(F^{\intercal}F)^{-1} = \left(\frac{2716}{106848}; \frac{8172}{106848}; \frac{76576}{106848}; \frac{49248}{106848} \right) =$$

$$= (0{,}025;\ 0{,}076;\ 0{,}717;\ 9{,}461).$$

Главная диагональ (9.2.29) корреляционной матрицы вектора оценок коэффициентов регрессии:

$$diagK_{\widetilde{A}[4]} = \left(\widetilde{\sigma}^2_{\widetilde{a}_0}, \widetilde{\sigma}^2_{\widetilde{a}_1}, \widetilde{\sigma}^2_{\widetilde{a}_2}, \widetilde{\sigma}^2_{\widetilde{a}_3} \right) = (0{,}0061 \cdot 0{,}025;\ 0{,}061 \cdot 0{,}076;$$

$$0{,}0061 \cdot 0{,}717;\ 0{,}0061 \cdot 0{,}461) = (1{,}5 \cdot 10^{-4};\ 4{,}6 \cdot 10^{-4};\ 4{,}3 \cdot 10^{-4};\ 28{,}1 \cdot 10^{-4}).$$

Вычисляем оценки средних квадратических отклонений коэффициентов $\widetilde{a}_j$, $j = \overline{0,3}$:

$$\sigma_{\tilde{a}_0} = \sqrt{1{,}5 \cdot 10^{-4}} = 0{,}012; \quad \sigma_{\tilde{a}_1} = \sqrt{4{,}6 \cdot 10^{-4}} = 0{,}021;$$

$$\sigma_{\tilde{a}_2} = \sqrt{43{,}7 \cdot 10^{-4}} = 0{,}066; \quad \sigma_{\tilde{a}_3} = \sqrt{28{,}1 \cdot 10^{-4}} = 0{,}053.$$

Находим наблюдаемые значения показателя (9.2.27):

$$t_0 = \frac{0{,}97}{0{,}012} = 80{,}8; \quad t_1 = \frac{1{,}96}{0{,}021} = 93{,}3; \quad t_2 = \frac{0{,}03}{0{,}066} = 0{,}45; \quad t_3 = \frac{4{,}97}{0{,}053} = 93{,}8.$$

Критическое значение показателя согласованности при уровне значимости $\alpha = 0{,}01$ и одной степени свободы $f = 1$ находим в приложении 6: $t_{(0{,}01;\ 1)} = 63{,}7$.

Проверяем условие (9.2.30) и получаем:

$$t_0 > t_{(0{,}01;\ 1)}; \quad t_1 > t_{(0{,}01;\ 1)}; \quad t_2 < t_{(0{,}01;\ 1)}; \quad t_3 > t_{(0{,}01;\ 1)}.$$

На основе приведённых неравенств делаем вывод, что коэффициенты a_0, a_1 и a_3 являются значимыми, а коэффициент a_2 принимаем равным нулю.

Окончательный вид уравнения регрессии

$$\tilde{y} = -0{,}97x^3 - 1{,}96x^2 + 4{,}97.$$

6. Проверяем адекватность последнего уравнения по критерию Фишера. Составляем табл.9.6.

Таблица 9.6

Расчётная таблица

x_i	y_i	$\tilde{y}_i$	$y_i - \tilde{y}_i$	$(y_i - \tilde{y}_i)^2$
–2	5	4,89	0,11	0,0121
–1	4	3,98	0,02	0,0004
0	5	4,97	0,03	0,0009
1	2	2,04	–0,04	0,0016
3	–39	–38,86	–0,14	0,0196
$\sum_{i=1}^{5} (y_i - \tilde{y}_i)^2 = 0{,}0346$				

Вычисляем оценку остаточной дисперсии с учётом результата, полученного в табл.9.6:

$$\tilde{\sigma}_1^2 = \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n - k - 1} = \frac{0{,}0346}{1} = 0{,}0346.$$

Общая дисперсия остаётся прежней. Поэтому наблюдаемое значение показателя согласованности (9.2.33) будет следующим:

$$\mathsf{F} = \frac{\tilde{\sigma}^2}{\tilde{\sigma}_1^2} = \frac{371{,}25}{0{,}0346} = 10730.$$

Очевидно, что при критическом значении показателя согласованности $\mathsf{F}_{(0{,}01;\ 4;\ 1)} = 5625$ неравенство (9.2.24) выполняется. Таким образом,

повторная проверка адекватности по критерию Фишера подтверждает справедливость гипотезы о соответствии функции регрессии вида (9.2.31) экспериментальным данным.

▲

П р и м е р 9.2. В условиях примера 9.1 построить уравнение регрессии, но расчёты произвести в матричной форме.

▼ Матричное уравнение (9.2.19) для рассматриваемого примера имеет вид

$$(F_{[4;5]}^{\intercal}F_{[5;4]})\widetilde{A}_{<4>} = F_{[4;5]}^{\intercal}Y_{<5>}F^{\intercal}F\,, \qquad (9.2.39)$$

где матрицы F и $F^{\intercal}F$ представлены выражениями (9.2.37), (9.2.38) соответственно;

$\widetilde{A}_{<4>} = (\widetilde{a}_0,\ \widetilde{a}_1,\ \widetilde{a}_2,\ \widetilde{a}_3)^{\intercal}$; $\quad Y_{<5>} = (y_1,\ y_2,\ y_3,\ y_4,\ y_5)^{\intercal} = (5;\ 4;\ 5;\ 2;\ -39)^{\intercal}$.

Выражение (9.2.20) для вычисления оценок коэффициентов регрессии представляется в виде

$$\widetilde{A}_{<4>} = (F_{[4;5]}^{\intercal}F_{[5;4]})^{-1}F_{[4;5]}^{\intercal}Y_{<5>}\,. \qquad (9.2.40)$$

Вычисляем обратную матрицу $(F^{\intercal}F)^{-1}$ для матрицы

$$F_{[4;5]}^{\intercal}F_{[5;4]} = \begin{pmatrix} 795 & 211 & 99 & 19 \\ 211 & 99 & 19 & 15 \\ 99 & 19 & 15 & 1 \\ 19 & 15 & 1 & 5 \end{pmatrix}.$$

Алгебраические дополнения элементов матрицы $F^{\intercal}F$:

$A_{11} = 2716$; $A_{21} = -3780$; $A_{31} = -13384$; $A_{41} = 3696$;
$A_{12} = -3780$; $A_{22} = 8172$; $A_{32} = 15480$; $A_{42} = -13248$;
$A_{13} = -13384$; $A_{23} = 126304$; $A_{33} = 76576$; $A_{43} = -10896$;
$A_{14} = 3696$; $A_{24} = -13248$; $A_{34} = -10896$; $A_{44} = 49248$.

Определитель указанной матрицы $|F^{\intercal}F| = 106848$.

В результате получили обратную матрицу

$$(F^{\intercal}F)^{-1} = \frac{1}{|F^{\intercal}F|} = \begin{pmatrix} A_{11} & A_{21} & A_{31} & A_{41} \\ A_{12} & A_{22} & A_{32} & A_{42} \\ A_{13} & A_{23} & A_{33} & A_{43} \\ A_{14} & A_{24} & A_{34} & A_{44} \end{pmatrix} =$$

$$= \frac{1}{106848}\begin{pmatrix} 2716 & -3780 & -13384 & 3696 \\ -3780 & 8172 & 15480 & -13248 \\ -13384 & 15480 & 76576 & -10896 \\ 3696 & -13248 & -10896 & 49248 \end{pmatrix}.$$

Далее вычисляем матрицу в правой части уравнения (9.2.39):

$$F^{\mathsf{T}}_{[5;4]}Y_{<5>}=\begin{pmatrix} x_1^3 & x_2^3 & x_3^3 & x_4^3 & x_5^3 \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & x_5^2 \\ x_1 & x_2 & x_3 & x_4 & x_5 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}=$$

$$=\begin{pmatrix} -8 & -1 & 0 & 1 & 27 \\ 4 & 1 & 0 & 1 & 9 \\ -2 & -1 & 0 & 1 & 3 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 4 \\ 5 \\ 2 \\ -39 \end{pmatrix}=\begin{pmatrix} -1095 \\ -325 \\ -129 \\ -23 \end{pmatrix}.$$

По формуле (9.2.40) находим оценки коэффициентов регрессии

$$\widetilde{A}_{<4>}=\begin{pmatrix} \widetilde{a}_0 \\ \widetilde{a}_1 \\ \widetilde{a}_2 \\ \widetilde{a}_3 \end{pmatrix}=\frac{1}{106848}\begin{pmatrix} 2716 & -3780 & -13384 & 3696 \\ -3780 & 8172 & 15480 & -13248 \\ -13384 & 15480 & 76576 & -10896 \\ 3696 & -13248 & -10896 & 49248 \end{pmatrix}\begin{pmatrix} -1095 \\ -325 \\ -129 \\ -23 \end{pmatrix}=$$

$$=\frac{1}{106848}\begin{pmatrix} -103992 \\ -209016 \\ -3216 \\ 531360 \end{pmatrix}=\begin{pmatrix} -0{,}97 \\ -1{,}96 \\ -0{,}03 \\ 4{,}97 \end{pmatrix}.$$

Так же, как и в примере 9.1 при расчётах в скалярной форме, получили уравнение регрессии

$$\widetilde{y}=-0{,}97x^3-1{,}96x^2-0{,}03x+4{,}97\,.$$

Проверка адекватности уравнения экспериментальным данным и значимость коэффициентов регрессии выполняется аналогично тому, как это сделано в примере 9.1.

▲

9.3. Многофакторный линейный регрессионный анализ

9.3.1. Модели многофакторного линейного регрессионного анализа

В § 9.2 рассматривались модели однофакторного регрессионного анализа, линейные относительно коэффициентов регрессии. В то же время они могут быть нелинейными относительно независимой переменной

(фактора), в примерах 9.1 и 9.2 рассматривалась именно такая нелинейная модель.

В настоящем подпараграфе рассмотрим модели многофакторного (множественного) регрессионного анализа, являющиеся линейными как относительно коэффициентов регрессии, так и относительно факторов.

Модель РА-1 определяется выражением

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k , \qquad (9.3.1)$$

а модель РА-2 – выражением

$$\hat{y} = b_0 + b_1 \hat{x}_1 + b_2 \hat{x}_2 + ... + b_k \hat{x}_k . \qquad (9.3.2)$$

Условное математическое ожидание (9.1.1) результата наблюдения представляется как

$$M[\hat{y} \mid X_{<k>}] = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k .$$

Для модели РА-1 экспериментальные данные могут быть представлены табл.9.7.

Таблица 9.7

Представление результатов многофакторного эксперимента (модель РА-1)

Опыты	Факторы						Результаты наблюдений						Средние значения результатов наблюдений $\bar{y}^*$
	x_1	x_2	···	x_j	···	x_k	y_1	y_2	···	y_s	···	y_m	
1	x_{11}	x_{12}	···	x_{1j}	···	x_{1k}	y_{11}	y_{12}	···	y_{1s}	···	y_{1m}	$\bar{y}_1^*$
2	x_{21}	x_{22}	···	x_{2j}	···	x_{2k}	y_{21}	y_{22}	···	y_{2s}	···	y_{2m}	$\bar{y}_2^*$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	···	x_{ij}	···	x_{ik}	y_{i1}	y_{i2}	···	y_{is}	···	y_{im}	$\bar{y}_i^*$
...	...	...	...	...	...	...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	···	x_{nj}	···	x_{nk}	y_{n1}	y_{n2}	···	y_{ns}	···	y_{nm}	$\bar{y}_n^*$

В данной таблице значения факторов $X_{<k>i}$, $i = \overline{1, n}$ являются фиксированными, что даёт возможность получить при данных значениях m результатов наблюдений. В частном случае может быть $m = 1$. При построении регрессионных зависимостей используется среднее значение результатов.

Для модели РА-2 экспериментальные данные представляются табл.9.8. В данной таблице имеют место системы случайных величин $(\hat{y}_i; \hat{X}_{<k>i})$, $i = \overline{1, n}$, т.е. каждому сочетанию значений факторов $\hat{X}_{<k>}$ соответствует одно значение результата $\hat{y}$.

В рассматриваемом случае линейного регрессионного анализа методы построения моделей вида (9.3.1) и (9.3.2) одинаковы. Это связано с

тем, что математические ожидания ошибок наблюдений равны нулю. Предположения, рассмотренные в п.п.9.2.1, остаются в силе.

Таблица 9.8

Представление результатов многофакторного эксперимента (модель РА-2)

Опыты	Факторы						Результаты наблюдений
	x_1	x_2	⋯	x_j	⋯	x_k	
1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1k}	y_1
2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2k}	y_2
...	...	...	...	...	...	...	...
i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{ik}	y_i
...	...	...	...	...	...	...	...
n	x_{n1}	x_{n2}	⋯	x_{nj}	⋯	x_{nk}	y_n

9.3.2. Построение уравнения множественной регрессии

Задача построения уравнения регрессии сводится к оцениванию коэффициентов

$$(b_0,\ b_1,\ b_2,...,b_k)^{\mathsf{T}} = B_{<k+1>} \tag{9.3.3}$$

в выражении

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n = b_0 + \sum_{j=1}^{k} b_j x_j\ . \tag{9.3.4}$$

Поскольку значения факторов могут иметь различный порядок, то для упрощения вычислений целесообразно использовать их центрированные значения

$$\dot{x}_{ij} = x_{ij} - \bar{x}_j^{\,*}, \quad i = \overline{1, n},\ \ j = \overline{1, k}, \tag{9.3.5}$$

где $\bar{x}_j^{\,*} = \dfrac{1}{n}\sum_{i=1}^{n} x_{ij}$.

Кроме этого вводится n-мерный единичный вектор

$$\dot{X}_{<n>} = (1,\ 1,...,1)^{\mathsf{T}},$$

что необходимо для оценки свободного члена уравнения регрессии. Матрица центрированных значений факторов при этом имеет вид

$$\dot{X}_{[n;\ k+1]} = \begin{pmatrix} 1 & \dot{x}_{11} & \dot{x}_{12} & \cdots & \dot{x}_{1j} & \cdots & \dot{x}_{1k} \\ 1 & \dot{x}_{21} & \dot{x}_{22} & \cdots & \dot{x}_{2j} & \cdots & \dot{x}_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & \dot{x}_{i1} & \dot{x}_{i2} & \cdots & \dot{x}_{ij} & \cdots & \dot{x}_{ik} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & \dot{x}_{n1} & \dot{x}_{n2} & \cdots & \dot{x}_{nj} & \cdots & \dot{x}_{nk} \end{pmatrix}. \tag{9.3.6}$$

Из вышеизложенного следует, что коэффициенты регрессии оцениваются в уравнении

$$y = b_0 x_0 + b_1 \dot{x}_1 + b_2 \dot{x}_2 + ... + b_n \dot{x}_n = \sum_{j=0}^{k} b_j \dot{x}_j . \qquad (9.3.7)$$

После вычисления данных оценок возврат к уравнению вида (9.3.4) осуществляется подстановкой (9.3.5). При этом значение коэффициента b_0 изменяется.

Подходящую оценку $\widetilde{B}_{<k+1>}$ вектора $B_{<k+1>} = (b_0, b_1, b_2,\ldots,b_k)^{\mathrm{T}}$ находим методом наименьших квадратов. Так же, как и в случае однофакторной регрессии, необходимо минимизировать сумму квадратов невязок (9.2.10). В данной формуле оценки $\widetilde{M}_{\hat{y}_i}$ математических ожиданий результата находятся из соотношения

$$\widetilde{y} = \widetilde{M}_{\hat{y}} = \sum_{j=0}^{k} \widetilde{b}_j \dot{x}_j . \qquad (9.3.8)$$

Учитывая (9.3.8), выражение (9.2.10) принимает вид

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \widetilde{b}_j \dot{x}_j \right)^2 . \qquad (9.3.9)$$

Таким образом, найденные оценки $\widetilde{b}_0, \widetilde{b}_1,...,\widetilde{b}_k$ должны удовлетворять условию

$$V = \min_{\widetilde{b}_j \in \mathbf{R}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} \widetilde{b}_j \dot{x}_j \right)^2 \right\}. \qquad (9.3.10)$$

Для определения минимума (9.3.10) функции (9.3.9) составляем систему нормальных уравнений вида (8.2.10):

$$\begin{cases} \sum_{i=1}^{n} \left(y_i - (\widetilde{b}_0 \dot{x}_0 + \widetilde{b}_1 \dot{x}_1 + ... + \widetilde{b}_k \dot{x}_k) \right) \dot{x}_0 = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n} \left(y_i - (\widetilde{b}_0 \dot{x}_0 + \widetilde{b}_1 \dot{x}_1 + ... + \widetilde{b}_k \dot{x}_k) \right) \dot{x}_j = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n} \left(y_i - (\widetilde{b}_0 \dot{x}_0 + \widetilde{b}_1 \dot{x}_1 + ... + \widetilde{b}_k \dot{x}_k) \right) \dot{x}_k = 0 \end{cases} . \qquad (9.3.11)$$

Уравнения (9.3.11) получены исходя из того, что частная производная квадратичной функции (9.3.9) по переменной $\widetilde{b}_j$ находится следующим образом:

$$\frac{\partial\left(\sum\limits_{i=1}^{n}\left(y_i-\sum\limits_{j=0}^{k}\widetilde{b}_j\dot{x}_j\right)\right)^2}{\partial b_j}=2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}\widetilde{b}_j\dot{x}_j\right)\frac{\partial\left(\sum\limits_{i=1}^{n}\left(y_i-\sum\limits_{j=0}^{k}\widetilde{b}_j\dot{x}_j\right)\right)}{\partial b_j}=$$

$$=2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}b_j\dot{x}_j\right)(-\dot{x}_j)=-2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{k}b_j\dot{x}_j\right)\dot{x}_j= \qquad (9.3.12)$$

$$=-2\sum_{i=1}^{n}(y_i-(b_0\dot{x}_0+b_1\dot{x}_1+b_2\dot{x}_2+...+b_k\dot{x}_k))\dot{x}_j.$$

Частную производную (9.3.12) приравниваем к нулю и обе части полученного уравнения умножаем на –2. В результате имеем *j*-е уравнение системы (9.3.11). Далее выполняем почленное суммирование в уравнениях рассматриваемой системы, переносим в правую часть слагаемые, содержащие y_i, а затем умножаем на –1 обе части каждого уравнения. Указанная последовательность операции приводит к эквивалентной системе уравнений

$$\begin{cases}\widetilde{b}_0\sum\limits_{i=1}^{n}\dot{x}_{i0}^2+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}\dot{x}_{i0}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}\dot{x}_{i0}=\sum\limits_{i=1}^{n}y_i\dot{x}_{i0}\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \widetilde{b}_0\sum\limits_{i=1}^{n}\dot{x}_{i0}\dot{x}_{ij}+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}\dot{x}_{ij}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}\dot{x}_{ij}=\sum\limits_{i=1}^{n}y_i\dot{x}_{ij}\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \widetilde{b}_0\sum\limits_{i=1}^{n}\dot{x}_{i0}\dot{x}_{ik}+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}\dot{x}_{ik}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}^2=\sum\limits_{i=1}^{n}y_i\dot{x}_{ik}.\end{cases}$$

Учитывая, что $\dot{x}_{i0}=1$, окончательно получаем

$$\begin{cases}\widetilde{b}_0 n+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}=\sum\limits_{i=1}^{n}y_i\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \widetilde{b}_0\sum\limits_{i=1}^{n}\dot{x}_{ij}+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}\dot{x}_{ij}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}\dot{x}_{ij}=\sum\limits_{i=1}^{n}y_i\dot{x}_{ij}\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \widetilde{b}_0\sum\limits_{i=1}^{n}\dot{x}_{ik}+\widetilde{b}_1\sum\limits_{i=1}^{n}\dot{x}_{i1}\dot{x}_{ik}+...+\widetilde{b}_k\sum\limits_{i=1}^{n}\dot{x}_{ik}^2=\sum\limits_{i=1}^{n}y_i\dot{x}_{ik}.\end{cases} \qquad (9.3.13)$$

Система (9.3.13) является системой линейных уравнений относительно оценок коэффициентов регрессии $\widetilde{b}_0, \widetilde{b}_1,...,\widetilde{b}_k$. Данные оценки находим по формулам Крамера:

$$\widetilde{b}_0 = \frac{|B_0|}{|B|}, ..., \widetilde{b}_j = \frac{|B_j|}{|B|}, ..., \widetilde{b}_k = \frac{|B_k|}{|B|}, \qquad (9.3.14)$$

где $|B|$ – определитель коэффициентов при неизвестных системы уравнений (9.3.13); $|B_j|$, $j = \overline{0, k}$ – определители, полученные из определителя $|B|$ заменой j-го столбца столбцом свободных членов.

Развёрнутый вид данных определителей:

$$|B| = \begin{vmatrix} n & \sum\limits_{i=1}^{n} \dot{x}_{i1} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ij} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ij} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}\dot{x}_{ij} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ik} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ik} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}^2 \end{vmatrix}; \quad |B_0| = \begin{vmatrix} \sum\limits_{i=1}^{n} y_i & \sum\limits_{i=1}^{n} \dot{x}_{i1} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} y_i\dot{x}_{ij} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ij} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}\dot{x}_{ij} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} y_i\dot{x}_{ik} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ik} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}^2 \end{vmatrix};$$

$$|B_j| = \begin{vmatrix} n & \cdots & \sum\limits_{i=1}^{n} y_i & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ij} & \cdots & \sum\limits_{i=1}^{n} y_i\dot{x}_{ij} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}\dot{x}_{ij} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ik} & \cdots & \sum\limits_{i=1}^{n} y_i\dot{x}_{ik} & \cdots & \sum\limits_{i=1}^{n} \dot{x}_{ik}^2 \end{vmatrix}; \quad |B_k| = \begin{vmatrix} n & \sum\limits_{i=1}^{n} \dot{x}_{i1} & \cdots & \sum\limits_{i=1}^{n} y_i \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ij} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ij} & \cdots & \sum\limits_{i=1}^{n} y_i\dot{x}_{ij} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{n} \dot{x}_{ik} & \sum\limits_{i=1}^{n} \dot{x}_{i1}\dot{x}_{ik} & \cdots & \sum\limits_{i=1}^{n} y_i\dot{x}_{ik} \end{vmatrix}.$$

Система уравнений (9.2.13) может быть записана в виде матричного уравнения

$$\left(\dot{X}^{\intercal}_{[k+1;\, n]}\dot{X}_{[n;\, k+1]}\right)\widetilde{B}_{<k+1>} = \dot{X}^{\intercal}_{[k+1;\, n]}Y_{<n>}, \qquad (9.3.15)$$

где

$$\dot{X} = \begin{pmatrix} 1 & \dot{x}_{11} & \cdots & \dot{x}_{1k} \\ \cdots & \vdots & \cdots & \cdots \\ 1 & \dot{x}_{i1} & \cdots & \dot{x}_{ik} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \dot{x}_{n1} & \cdots & \dot{x}_{nk} \end{pmatrix},$$

$$\widetilde{B} = (\widetilde{b}_0,\, \widetilde{b}_1, ..., \widetilde{b}_k)^{\intercal}; \quad Y = (y_1,\, y_2, ..., y_n)^{\intercal}.$$

Для получения вектора $\widetilde{B}$ умножаем обе части матричного уравнения (9.3.15) на $(\dot{X}^{\intercal}\dot{X})^{-1}$ слева:

$$(\dot{X}^{\intercal}\dot{X})^{-1}(\dot{X}^{\intercal}\dot{X})\widetilde{B} = (\dot{X}^{\intercal}\dot{X})^{-1}\dot{X}^{\intercal}Y,$$

отсюда

$$E\widetilde{B} = (\dot{X}^{\intercal}\dot{X})^{-1}\dot{X}^{\intercal}Y$$

или

$$\widetilde{B}_{<k+1>} = \left(\dot{X}^{\mathsf{T}}_{[k+1;\,n]}\dot{X}_{[n;\,k+1]}\right)^{-1}\dot{X}^{\mathsf{T}}_{[k+1;\,n]}Y_{<n>}, \qquad (9.3.16)$$

где $E = E_{[k+1]}$ – единичная матрица порядка $(k+1)$.

Рассмотренный метод построения уравнения регрессии применим как для модели РА-1, так и РА-2.

9.3.3. Проверка адекватности уравнения множественной регрессии

Подобно однофакторному регрессионному анализу, проверка адекватности экспериментальным данным уравнения множественной регрессии производится на основании анализа отношения дисперсий. В качестве показателя согласованности нулевой гипотезы H_0 об адекватности уравнения принимается выражение, аналогичное выражению (9.2.23):

$$\hat{F} = \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}_1^2}, \qquad (9.3.17)$$

где $\widetilde{\sigma}^2$ – оценка общей дисперсии наблюдаемой переменной; $\widetilde{\sigma}_1^2$ – оценка остаточной дисперсии.

Оценки дисперсий, входящих в соотношение (9.3.17), определяются по формулам

$$\widetilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y}^*)^2}{n-1}, \qquad (9.3.18)$$

$$\widetilde{\sigma}_1^2 = \frac{\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k}\widetilde{b}_j\dot{x}_{ij}\right)^2}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \widetilde{y}_i)^2}{n-k-1}, \qquad (9.3.19)$$

где $\bar{y}^*$ – оценка математического ожидания наблюдаемой переменной; $n-1$ – число степеней свободы дисперсии σ^2; $n-k-1$ – число степеней свободы дисперсии σ_1^2.

Случайная величина (9.3.17) подчинена закону распределения Фишера [1]. Следовательно, для принятия нулевой гипотезы необходимо и достаточно, чтобы выполнялось условие

$$\mathsf{F} = \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}_1^2} > \mathsf{F}_{(\alpha;\,n\text{-}1;\,n\text{-}k\text{-}1)}, \qquad (9.3.20)$$

где F –наблюдаемое значение показателя согласованности гипотезы H_0; $\mathsf{F}_{(\alpha;\,n\text{-}1;\,n\text{-}k\text{-}1)}$ – критическое значение данного показателя при уровне значимости α и степенях свободы $f_1 = n-1$, $f_2 = n-k-1$. Критическое значе-

ние показателя берётся по таблице критических точек распределения Фишера (приложение 5).

При невыполнении условия (9.3.20) принимается конкурирующая гипотеза H_1 о том, что линейная модель (9.3.1) или (9.3.2) неадекватна экспериментальным данным.

9.3.4. Селекция факторов

Уравнение регрессии связывает наблюдаемую переменную с совокупностью k факторов. Можно предположить (выдвинуть гипотезу), что какая-то часть этих факторов не оказывает существенного влияния на величину переменной y. Такие факторы без ущерба для точности могут быть исключены из уравнения. Возникает задача выявления таких факторов. Значимость любого фактора определяется, прежде всего, величиной коэффициента регрессии при данном факторе. Следовательно, задача селекции факторов сводится к проверке значимости коэффициентов регрессии. Такая задача решалась в однофакторном регрессионного анализе (пп. 9.2.4).

Следуя представленной в указанном подпараграфе схеме, необходимо проверить нулевые гипотезы о неравенстве нулю коэффициентов b_j:

$$H_{0j}: b_j \neq 0, \quad j = \overline{0, k}.$$

Конкурирующие гипотезы состоят в предположении о равенстве нулю коэффициентов:

$$H_{1j}: b_j = 0.$$

В качестве показателя согласованности при проверке гипотез H_{0j} используется выражение

$$\hat{t}_j = \frac{|\widetilde{b}_j|}{\widetilde{\sigma}_{\widetilde{b}_j}}, \quad j = \overline{0, k}, \tag{9.3.21}$$

где $|\widetilde{b}_j|$ – модуль величины $\widetilde{b}_j$; $\widetilde{\sigma}_{\widetilde{b}_j}$ – оценка среднего квадратического отклонения коэффициента $\widetilde{b}_j$.

Известно [1], что случайная величина (9.3.21) подчинена закону распределения Стьюдента.

Для того, чтобы выражение (9.3.21) можно было использовать практически, необходимо иметь методику вычисления величин $\widetilde{\sigma}_{\widetilde{b}_j}$. Она аналогична представленной в пп.9.2.4.

Рассматриваем корреляционную матрицу вектора $\widetilde{B}_{<k+1>}$:

$$K_{\widetilde{B}[k+1]} = \widetilde{\sigma}_1^2 (\dot{X}^{\intercal}_{[k+1;\,n]} \dot{X})^{-1}. \tag{9.3.22}$$

Элементы главной диагонали матрицы (9.3.22) и являются дисперсиями оценок коэффициентов регрессии:

$$diagK_{\tilde{B}} = (\tilde{\sigma}^2_{\tilde{b}_0},\ \tilde{\sigma}^2_{\tilde{b}_1},...,\tilde{\sigma}^2_{\tilde{b}_k}).$$

Проверяются условия

$$t_j = \frac{|\tilde{b}_j|}{\tilde{\sigma}_{\tilde{b}_j}} > t_{(\alpha;n-k-1)}, \quad j = \overline{0,k}, \tag{9.3.23}$$

где t_j – вычисленное значение показателя согласованности гипотезы H_{0j}; $t_{(\alpha;n-k-1)}$ – критическое значение данного показателя при уровне значимости α и $f_1 = n-k-1$ степенях свободы. Критическое значение $t_{(\alpha;n-k-1)}$ берётся из таблицы критических точек распределения Стьюдента (приложение 6).

Коэффициенты регрессии, для которых условие (9.3.23) не выполняется, принимаются равными нулю. Следовательно, соответствующие им факторы являются незначимыми. Проверка условий (9.3.23) называется селекцией факторов по критерию Стьюдента.

9.3.5. Пример многофакторного линейного регрессионного анализа

Из вышеизложенного очевидно, что многофакторный линейный регрессионный анализ проводится в следующей последовательности.

1. Определяется количество слагаемых в линейной функции регрессии вида (9.3.1) или (9.3.2) в зависимости от числа факторов.
2. Выполняется центрирование факторов.
3. Составляется система нормальных уравнений.
4. Находится решение системы нормальных уравнений (определяются оценки коэффициентов регрессии).
5. Проверяется адекватность построенного уравнения регрессии экспериментальным данным.
6. Производится селекция факторов.
7. Повторно проверяется адекватность уравнения регрессии после исключения незначимых факторов.

П р и м е р 9.3. Для исследования зависимости выходного сигнала y системы от входного воздействия $X_{<2>} = (x_1, x_2)^{\mathrm{T}}$, проведены испытания, результаты которых сведены в табл.9.9.

Необходимо построить регрессионную зависимость y от $X_{<2>}$ в предположении, что она является линейным алгебраическим полиномом.

Т а б л и ц а 9.9

Массив экспериментальных данных

x_1	–0,5	0	0,8	0,4	0,5	0,6
x_2	–3	–1	2	0,5	1,5	6
y	–15,1	–1	19,9	9,5	16,5	47,9

▼ 1. В рассматриваемой задаче функция регрессии представляет собой линейный алгебраический полином от двух независимых переменных

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

2. Составляем таблицу экспериментальных данных с центрированными значениями факторов.

Предварительно вычисляем средние значения факторов:

$$\bar{x}_1^{\,*} = \frac{1}{6}\sum_{i=1}^{6} x_{i1} = \frac{1}{6}(-0{,}5 + 0 + 0{,}8 + 0{,}4 + 0{,}5 + 0{,}6) = 0{,}3;$$

$$\bar{x}_2^{\,*} = \frac{1}{6}\sum_{i=1}^{6} x_{i2} = 1;$$

Используя формулы (9.3.5), производим центрирование факторов, результаты заносятся в таб.9.10.

Таблица 9.10

Массив экспериментальных данных с центрированными значениями факторов

$\dot{x}_1$	–0,8	–0,3	0,5	0,1	0,2	0,3
$\dot{x}_2$	–4	–2	1	–0,5	0,5	5
y	–15,1	–1	19,9	9,5	16,5	47,9

Таким образом, вначале коэффициенты регрессии оцениваем в выражении функции (9.3.7), которая для данной задачи принимает вид

$$y = b_0 + b_1 \dot{x}_1 + b_2 \dot{x}_2 \qquad (9.3.24)$$

3. Система нормальных уравнений представляется следующим образом:

$$\begin{cases} 6\tilde{b}_0 + \tilde{b}_1 \sum_{i=1}^{6} \dot{x}_{i1} + \tilde{b}_2 \sum_{i=1}^{6} \dot{x}_{i2} = \sum_{i=1}^{2} y_1 \\ \tilde{b}_0 \sum_{i=1}^{6} \dot{x}_{i1} + \tilde{b}_1 \sum_{i=1}^{6} \dot{x}_{i1}^2 + \tilde{b}_2 \sum_{i=1}^{6} \dot{x}_{i1}\dot{x}_{i2} = \sum_{i=1}^{6} y_1 \dot{x}_{i1} \\ \tilde{b}_0 \sum_{i=1}^{6} \dot{x}_{i2} + \tilde{b}_1 \sum_{i=1}^{6} \dot{x}_{i1}\dot{x}_{i2} + \tilde{b}_2 \sum_{i=1}^{6} \dot{x}_{i2}^2 = \sum_{i=1}^{6} y_1 \dot{x}_{i2}. \end{cases} \qquad (9.3.25)$$

4. Представим и решим систему (9.3.25) в матричной форме.

Матричное уравнение, эквивалентное данной системе, принимает вид

$$\left(\dot{X}^{\mathsf{T}}_{[3;6]} \dot{X}_{[6;3]}\right) \tilde{B}_{<3>} = \dot{X}^{\mathsf{T}}_{[3;6]} Y_{<6>}, \qquad (9.3.26)$$

где $$\dot{X} = \begin{pmatrix} 1 & \dot{x}_{11} & \dot{x}_{12} \\ 1 & \dot{x}_{21} & \dot{x}_{22} \\ 1 & \dot{x}_{31} & \dot{x}_{23} \\ 1 & \dot{x}_{41} & \dot{x}_{24} \\ 1 & \dot{x}_{51} & \dot{x}_{25} \\ 1 & \dot{x}_{61} & \dot{x}_{26} \end{pmatrix} = \begin{pmatrix} 1 & -0{,}8 & -4 \\ 1 & -0{,}3 & -2 \\ 1 & 0{,}5 & 1 \\ 1 & 0{,}1 & -0{,}5 \\ 1 & 0{,}2 & 0{,}5 \\ 1 & 0{,}3 & 5 \end{pmatrix};$$

$$\widetilde{B} = (\widetilde{b}_0, \widetilde{b}_1, \widetilde{b}_2)^{\mathsf{T}};$$

$$Y = (y_1, y_2, y_3, y_4, y_5, y_6)^{\mathsf{T}} = (-15{,}1;\ -1;\ 19{,}9;\ 9{,}5;\ 16{,}5;\ 47{,}9)^{\mathsf{T}}.$$

Выражение (9.3.16) для вычисления оценок коэффициентов регрессии представляется равенством

$$\widetilde{B}_{<3>} = \left(\dot{X}^{\mathsf{T}}_{[3;6]}\dot{X}_{[6;3]}\right)^{-1}\dot{X}^{\mathsf{T}}_{[3;6]}Y_{<6>}. \qquad (9.3.27)$$

Вычисляем матрицу $\dot{X}^{\mathsf{T}}\dot{X}$:

$$\dot{X}^{\mathsf{T}}\dot{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -0{,}8 & -0{,}3 & 0{,}5 & 0{,}1 & 0{,}2 & 0{,}3 \\ -4 & -2 & 1 & -0{,}5 & 0{,}5 & 5 \end{pmatrix} \begin{pmatrix} 1 & -0{,}8 & -4 \\ 1 & -0{,}3 & -2 \\ 1 & 0{,}5 & 1 \\ 1 & 0{,}1 & -0{,}5 \\ 1 & 0{,}2 & 0{,}5 \\ 1 & 0{,}3 & 5 \end{pmatrix} =$$

$$= \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1{,}12 & 5{,}85 \\ 0 & 5{,}85 & 46{,}5 \end{pmatrix}.$$

Для полученной матрицы находим обратную матрицу:

$$\left(\dot{X}^{\mathsf{T}}\dot{X}\right)^{-1} = \frac{1}{107}\begin{pmatrix} 17{,}9 & 0 & 0 \\ 0 & 279 & -35{,}1 \\ 0 & -35{,}1 & 6{,}72 \end{pmatrix}. \qquad (9.3.28)$$

Далее находим матрицу в правой части уравнения (9.3.26):

$$\dot{X}^{\mathsf{T}}Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -0{,}8 & -0{,}3 & 0{,}5 & 0{,}1 & 0{,}2 & 0{,}3 \\ -4 & -2 & 1 & -0{,}5 & 0{,}5 & 5 \end{pmatrix} \begin{pmatrix} -15{,}1 \\ -1 \\ 19{,}9 \\ 9{,}5 \\ 16{,}5 \\ 47{,}9 \end{pmatrix} = \begin{pmatrix} 77{,}7 \\ 40{,}9 \\ 321 \end{pmatrix}.$$

По формуле (9.3.27) вычисляем оценку вектора коэффициентов регрессии

$$\widetilde{B} = \frac{1}{107}\begin{pmatrix} 17{,}8 & 0 & 0 \\ 0 & 279 & -35{,}1 \\ 0 & -35{,}1 & 6{,}72 \end{pmatrix}\begin{pmatrix} 77{,}7 \\ 40{,}9 \\ 321 \end{pmatrix} = \begin{pmatrix} 12{,}9 \\ 1{,}35 \\ 6{,}74 \end{pmatrix}.$$

Получим следующее уравнение регрессии:

$$\widetilde{y} = 12{,}9 + 1{,}35\dot{x}_1 + 6{,}74\dot{x}_2 . \qquad (9.3.29)$$

5. Проверяем адекватность уравнения (9.3.29) экспериментальным данным.

Предварительно вычисляем оценки дисперсий (9.3.18) и (9.3.19). Для этого составляем табл.9.11.

Таблица 9.11

Расчётная таблица

$\dot{x}_1$	$\dot{x}_2$	y_i	$y_i - \bar{y}^*$	$(y_i - \bar{y}^*)^2$	$\widetilde{y}_i$	$y_i - \widetilde{y}_i$	$(y_i - \widetilde{y}_i)^2$
–0,8	–4	–15,1	–28	784	–15,2	0,01	0,0001
–0,3	–2	–1	–13,9	193	–0,98	–0,02	0,0004
0,5	1	19,9	7	49	20,3	–0,40	0,16
0,1	–0,5	9,5	–3,4	11,6	9,66	0,16	0,0256
0,2	0,5	16,5	3,6	13	16,6	–0,10	0,01
0,3	5	47,9	35	1225	47	0,90	0,81
$\sum_{i=1}^{6}(y_i - \bar{y}^*)^2 = 2716$					$\sum_{i=1}^{6}(y_i - \widetilde{y}_i)^2 = 1{,}01$		

Оценка математического ожидания выходной переменной *y*, используемая при расчётах в табл.9.11, найдена по формуле (5.1.1):

$$\bar{y}^* = \frac{1}{6}\sum_{i=1}^{6} y_i = \frac{1}{6}(-15{,}1 - 1 + 19{,}9 + 9{,}5 + 16{,}5 + 47{,}9) = 12{,}9 .$$

Необходимые данные для вычисления оценок дисперсий берём из табл.9.11:

$$\widetilde{\sigma}^2 = \frac{\sum_{i=1}^{6}(y_i - \bar{y})^2}{6-1} = \frac{2716}{5} \approx 543 ,$$

$$\widetilde{\sigma}_1^2 = \frac{\sum_{i=1}^{6}(y_i - \widetilde{y}_i)^2}{6-2-1} = \frac{1{,}01}{3} = 0{,}33 , \qquad (9.3.30)$$

Наблюдаемое значение показателя согласованности (9.3.17):

$$F = \frac{\widetilde{\sigma}^2}{\widetilde{\sigma}_1^2} = \frac{543}{0{,}33} = 1645$$

Для отыскания критического значения показателя согласованности при уровне значимости $\alpha = 0{,}01$ и степенях свободы $f_1 = n - 1 = 5$, $f_2 = n - k - 1 = 3$ используем приложение 5 и получаем $F_{(0{,}01;5;3)} = 28{,}24$.

Поскольку неравенство (9.3.20) выполняется ($F > F_{(0{,}01;5;3)}$), нулевую гипотезу об адекватности функции регрессии вида (9.3.24) экспериментальным данным принимаем.

6. Выполняем селекцию факторов. Для этого находим элементы главной диагонали корреляционной матрицы (9.3.22). Учитывая выражения (9.3.28) и (9.3.30), имеем

$$diagK_{\tilde{B}} = (\tilde{\sigma}^2_{\tilde{b}_0}, \tilde{\sigma}^2_{\tilde{b}_1}, \tilde{\sigma}^2_{\tilde{b}_2}) = (0{,}078;\ 1{,}224;\ 0{,}030)\,.$$

Оценки средних квадратических отклонений коэффициентов $\tilde{b}_j$, $j = \overline{0,2}$ принимают значения:

$$\tilde{\sigma}_{\tilde{b}_0} = 0{,}279; \qquad \tilde{\sigma}_{\tilde{b}_1} = 1{,}106; \qquad \tilde{\sigma}_{\tilde{b}_2} = 0{,}173\,.$$

Для каждого фактора находим наблюдаемое значение показателя согласованности (9.3.21):

$$t_0 = \frac{|\tilde{b}_0|}{\tilde{\sigma}_{\tilde{b}_0}} = \frac{12{,}9}{0{,}279} = 46{,}2; \quad t_1 = \frac{|\tilde{b}_1|}{\tilde{\sigma}_{\tilde{b}_1}} = \frac{1{,}35}{1{,}106} = 1{,}22; \quad t_2 = \frac{|\tilde{b}_2|}{\tilde{\sigma}_{\tilde{b}_2}} = \frac{6{,}74}{0{,}173} = 38{,}9\,.$$

Для числа степеней свободы $f = n - k - 1 = 3$ и уровня значимости $\alpha = 0{,}01$ критическое значение показателя согласованности $t_{(0{,}01;3)} = 5{,}84$. Следовательно,

$$t_0 > t_{(0{,}01;3)}, \qquad t_1 < t_{(0{,}01;3)}, \qquad t_3 > t_{(0{,}01;3)}.$$

В отношении фактора $\dot{x}_1$ принимаем конкурирующую гипотезу о его незначимости. Тогда в правой части выражения (9.3.24) второе слагаемое приравниваем к нулю. Поскольку в исходной матрице $\dot{X}$ исключается второй столбец, оценки коэффициентов регрессии b_0 и b_2 необходимо пересчитать.

Пересчёт выполняем в том же порядке, который приведён выше (сохраняем ту же нумерацию пунктов).

1. Функция регрессии в данном случае представляет собой линейный алгебраический полином от одной независимой переменной

$$y = b_0 + b_2 x_2. \qquad (9.3.31)$$

2. Составляем табл.9.12 экспериментальных данных с центрированными значениями фактора x_2.

Таблица 9.12

Массив экспериментальных данных с центрированными значениями фактора

$\dot{x}_2$	–4	–2	1	–0,5	0,5	5
y	–15,1	–1	19,9	9,5	16,5	47,9

Коэффициенты регрессии предварительно оцениваем в уравнении

$$y = b_0 + b_2 \dot{x}_2 . \qquad (9.3.32)$$

3. Система нормальных уравнений принимает вид

$$\begin{cases} 6\widetilde{b}_0 + \widetilde{b}_2 \sum\limits_{i=1}^{6} \dot{x}_{i2} = \sum\limits_{i=1}^{6} y_1 \\ \widetilde{b}_0 \sum\limits_{i=1}^{6} \dot{x}_{i2} + \widetilde{b}_2 \sum\limits_{i=1}^{6} \dot{x}_{i2}^2 = \sum\limits_{i=1}^{6} y_1 \dot{x}_{i2}. \end{cases} \qquad (9.3.33)$$

4. Матричное уравнение, эквивалентное системе (9.3.33), представляется как

$$\left(\dot{X}^{\mathsf{T}}_{[2;6]} \dot{X}_{[6;2]}\right) \widetilde{B}_{<2>} = \dot{X}^{\mathsf{T}}_{[2;6]} Y_{<6>}, \qquad (9.3.34)$$

где

$$\dot{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \dot{x}_{12} & \dot{x}_{22} & \dot{x}_{32} & \dot{x}_{42} & \dot{x}_{52} & \dot{x}_{62} \end{pmatrix}^{\mathsf{T}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -4 & -2 & 1 & -0{,}5 & 0{,}5 & 5 \end{pmatrix}^{\mathsf{T}};$$

$$\widetilde{B} = (\widetilde{b}_0,\ \widetilde{b}_2)^{\mathsf{T}};$$

$$Y = (y_1,\ y_2,\ y_3,\ y_4,\ y_5,\ y_6)^{\mathsf{T}} = (-15{,}1;\ -1;\ 19{,}9;\ 9{,}5;\ 16{,}5;\ 47{,}9)^{\mathsf{T}}.$$

Оценки коэффициентов регрессии определяются равенством

$$\widetilde{B}_{<2>} = \left(\dot{X}^{\mathsf{T}}_{[2;6]} \dot{X}_{[6;2]}\right)^{-1} \dot{X}^{\mathsf{T}}_{[2;6]} Y_{<6>} . \qquad (9.3.35)$$

Вычисляем матрицу $\dot{X}^{\mathsf{T}} \dot{X}$:

$$\dot{X}^{\mathsf{T}} \dot{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -4 & -2 & 1 & -0{,}5 & 0{,}5 & 5 \end{pmatrix} \begin{pmatrix} 1 & -4 \\ 1 & -2 \\ 1 & 1 \\ 1 & -0{,}5 \\ 1 & 0{,}5 \\ 1 & 5 \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 0 & 46{,}5 \end{pmatrix}.$$

Находим обратную матрицу:

$$\left(\dot{X}^{\mathsf{T}} \dot{X}\right)^{-1} = \frac{1}{279} \begin{pmatrix} 45{,}5 & 0 \\ 0 & 6 \end{pmatrix}..$$

Матрица в правой части уравнения (9.3.33) есть не что иное, как вектор-столбец с двумя компонентами:

$$\dot{X}^{\mathsf{T}}Y=\begin{pmatrix}1 & 1 & 1 & 1 & 1 & 1\\ -4 & -2 & 1 & -0{,}5 & 0{,}5 & 5\end{pmatrix}\begin{pmatrix}-15{,}1\\ -1\\ 19{,}9\\ 9{,}5\\ 16{,}5\\ 47{,}5\end{pmatrix}=\begin{pmatrix}77{,}7\\ 321\end{pmatrix}.$$

Оценку вектора коэффициентов регрессии находим по формуле (9.3.35):

$$\widetilde{B}=\frac{1}{279}\begin{pmatrix}46{,}5 & 0\\ 0 & 6\end{pmatrix}\begin{pmatrix}77{,}7\\ 321\end{pmatrix}=\begin{pmatrix}12{,}9\\ 6{,}90\end{pmatrix}.$$

Получим уравнение регрессии с одним фактором

$$\widetilde{y}=12{,}9+6{,}9\dot{x}_2. \qquad (9.3.36)$$

5. Проверяем адекватность уравнения (9.3.36) экспериментальным данным.

Для вычисления оценки дисперсии (9.3.19) составляем расчётную табл.9.13. Оценка дисперсии (9.3.18) остаётся прежней.

Таблица 9.13

Расчётная таблица

$\dot{x}_{i2}$	y_i	$\widetilde{y}_i$	$y_i-\widetilde{y}_i$	$(y_i-\widetilde{y}_i)^2$
–4	–15,1	–14,7	–0,4	0,16
–2	–1	–0,9	–0,1	0,01
1	19,9	19,8	0,1	0,01
–0,5	9,5	9,4	0,1	0,01
0,5	16,5	16,3	0,2	0,04
5	47,9	47,4	0,5	0,25
$\sum_{i=1}^{6}(y_i-\widetilde{y}_i)^2=0{,}48$				

Получаем оценку остаточной дисперсии:

$$\widetilde{\sigma}_1^2=\frac{\sum_{i=1}^{6}(y_i-\widetilde{y}_i)^2}{6-1-1}=\frac{0{,}48}{4}=0{,}12.$$

Показатель согласованности (9.3.17) принимает значение

$$\mathsf{F}=\frac{\widetilde{\sigma}^2}{\widetilde{\sigma}_1^2}=\frac{543}{0{,}12}=4525.$$

Критическое значение данного показателя при $\alpha=0{,}01$, $f_1=n-1=5$, $f_2=n-k-1=4$ составляет $\mathsf{F}_{(0{,}01;5;4)}=15{,}52$. Поскольку имеет место неравенство $\mathsf{F}>\mathsf{F}_{(0{,}01;5;4)}$, нулевая гипотеза об адекватности функции регрессии (9.3.31) экспериментальным данным принимается.

6. Выполняем селекцию факторов. Главная диагональ корреляционной матрицы (9.3.22) с учётом выражений (9.3.28) и (9.3.30) принимает вид

$$diagK_{\widetilde{B}} = (\widetilde{\sigma}^2_{\widetilde{b}_0},\ \widetilde{\sigma}^2_{\widetilde{b}_2}) = (0{,}020;\ 0{,}003)\,.$$

Из полученного результата следует, что

$$\widetilde{\sigma}_{\widetilde{b}_0} = 0{,}14; \qquad \widetilde{\sigma}_{\widetilde{b}_2} = 0{,}05\,.$$

Наблюдаемые значения показателя согласованности (9.3.21) для факторов $\dot{x}_0$ и $\dot{x}_2$:

$$t_0 = \frac{|\widetilde{b}_0|}{\widetilde{\sigma}_{\widetilde{b}_0}} = \frac{12{,}9}{0{,}14} = 92{,}1; \qquad t_2 = \frac{|\widetilde{b}_2|}{\widetilde{\sigma}_{\widetilde{b}_2}} = \frac{6{,}9}{0{,}05} = 138\,.$$

Находим критическое значение данного показателя в приложении 6, оно составляет $t_{(0{,}01;4)} = 4{,}6$. Таким образом,

$$t_0 > t_{(0{,}01;\,4)}, \qquad t_2 > t_{(0{,}01;4)}.$$

Принимаем нулевую гипотезу о значимости факторов $\dot{x}_0$ и $\dot{x}_2$ в уравнении (9.3.36).

Переходим к уравнению вида (9.3.31) с нецентрированными факторами:

$$\widetilde{y} = \widetilde{b}_0 + \widetilde{b}_2(x_2 - \bar{x})\,,$$

т.е.

$$\widetilde{y} = 12{,}9 + 6{,}9(x_2 - 1) = 12{,}9 + 6{,}9x_2 - 6{,}9 = 6 + 6{,}9x_2$$

или, окончательно

$$\widetilde{y} = 6 + 6{,}9x_2\,.$$

▲