Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

Томский государственный университет систем управления и радиоэлектроники

(ТУСУР)

ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

2015

Прикладная математическая статистика

Учебное пособие. —Томск: ТУСУР. – 2015. – 86с.

Составитель **А.А. Мицель**

В учебном пособии рассмотрены основные разделы прикладной математической статистики – выборка, точечные и интервальные оценки параметров распределений вероятностей, методы анализа законов распределения вероятностей случайных величин, проверка гипотез о значениях параметров распределений, дисперсионный анализ зависимостей, корреляционный анализ и регрессионный анализ.

Учебное пособие предназначено для магистрантов направления 01.04.02 «Прикладная математика и информатика», обучающихся по магистерской программе «Математическое и программное обеспечение вычислительных комплексов, систем и компьютерных сетей». Представляет интерес для инженеров, аспирантов, преподавателей, ученых, занимающихся вопросами статистической обработки данных.

ОГЛАВЛЕНИЕ

Введение в прикладную статистику	6
Вопросы для самоконтроля	8
Тема 1. Выборка. Эмпирическое распределение	9
1.1. Основные понятия и соотношения	9
1.2. Числовые характеристики выборки	13
Вопросы для самоконтроля	14
Тема 2. Точечные оценки параметров распределений вероятностей	15
2.1. Точечные и интервальные оценки	15
2.2. Вычисление точечных оценок	18
2.2.1. Оценка параметров методом максимального правдоподобия	18
2.2.2. Оценка параметров методом моментов	19
2.2.3. Оценка параметров методом наименьших квадратов	20
2.3. Точечная оценка параметров нормального распределения	21
2.4. Точечная оценка параметров показательного закона распределения	22
2.5. Точечная оценка параметров равномерного закона распределения	23
2.6. Точечная оценка параметров биномиального закона распределения	23
2.7. Планирование экспериментов для оценки параметров распределений	24
2.7.1. Нормальное распределение	24
2.7.2. Экспоненциальное распределение	25
2.7.3 Биномиальное распределение	26
Вопросы для самоконтроля	26
Тема 3. Интервальные оценки параметров распределений	28
3.1 Оценка параметров нормального распределения	28
3.2. Оценка параметров показательного распределения	29
3.3. Оценка параметров биномиального распределения	30
3.4. Примеры интервальных оценок	30
3.5. Интервальные оценки параметров при неизвестном законе распределения	31
3.5.1. Оценки для центра распределения	31

3.5.2 Оценка рассеяния распределения	32
Вопросы для самоконтроля	33
Тема 4. Методы анализа законов распределения вероятностей случайных величин	34
4.1. Общие понятия	34
4.2. Общие критерии согласия	35
4.2.1 Критерии, основанные на сравнении теоретической плотности распределения и эмпирической гистограммы	36
4.2.2 Критерии, основанные на сравнении теоретической и эмпирической функций распределения вероятностей	38
4.3 Критерии нормальности распределения	40
4.3.1 Модифицированный критерий χ^2	40
4.3.2 Критерий типа Колмогорова – Смирнова	41
4.4 Критерий проверки экспоненциальности распределения	43
4.4.1 Критерии типа Колмогорова –Смирнова	43
4.4.2 Критерий Фишера	44
4.5 Критерии согласия для равномерного распределения	44
4.6 Критерий симметрии	45
Вопросы для самоконтроля	46
Тема 5. Проверка гипотез о значениях параметров распределений	47
5.1 Общие сведения	47
5.2 Последовательные методы проверки гипотез о значениях параметров распределений	49
5.3 Проверка гипотезы о параметрах нормального распределения	51
5.3.1 Проверка гипотезы о значении среднего	51
5.3.2 Проверка гипотезы о значении дисперсии	53
5.4 Проверка гипотезы о параметре экспоненциального распределения	55
5.5 Проверка гипотезы о параметре биномиального распределения	56
Вопросы для самоконтроля	58
Тема 6. Дисперсионный анализ зависимостей	59
6.1 Основные положения	59
6.2. Однофакторный анализ	60
6.2.1.0днофакторный дисперсионный анализ	60
6.2.2. Непараметрические методы однофакторного анализа	63
6.3. Двухфакторный анализ	66

6.3.1 Двухфакторный параметрический дисперсионный анализ	67
6.3.2. Двухфакторный непараметрический анализ	68
Вопросы для самоконтроля	69
Тема 7. Корреляционный анализ	70
7.1. Вычисление параметрических коэффициентов корреляции	70
7.2 Вычисление непараметрических коэффициентов корреляции	72
Вопросы для самоконтроля	74
Тема 8. Регрессионный анализ	75
8.1. Построение модели регрессии	77
8.2. Оценка адекватности регрессии	77
8.2.1 Анализ регрессионных остатков	78
8.2.2 Доверительный интервал для уравнения регрессии	79
8.3. Оценка дисперсии коэффициентов регрессии и доверительных интервалов	79
8.4 Пример построения уравнения регрессии	80
Вопросы для самоконтроля	85
Литература	86

Введение в прикладную статистику

Что представляет из себя предмет математической статистки? Можно приводить разные описательные «определения», которые в большей или меньшей степени отражают содержание этого раздела математики. В теории вероятностей выводятся правила, которые позволяют по вероятностям одних случайных событий вычислить вероятности других, которые с ними связаны или по числовым характеристикам и функциям распределения одних случайных величин подсчитывать функции распределения и числовые характеристики других случайных величин. Другими словами, зная состав генеральной совокупности, там изучают распределения для состава случайной выборки. Это типичная *прямая задача* теории вероятностей. Однако часто приходится решать и *обратные задачи,* когда известен состав выборки и по нему требуется определить, какой была генеральная совокупность. Такого рода обратные задачи и составляют, образно говоря, предмет математической статистики.

Несколько уточняя это сравнение, можно сказать так: в теории вероятностей мы, зная природу некоторого явления, выясняем, как будут вести себя (как распределены) те или иные изучаемые нами характеристики, которые можно наблюдать в экспериментах. В математической статистике наоборот — исходными являются экспериментальные данные (как правило, это наблюдения над случайными величинами), а требуется вынести то или иное суждение или решение о природе рассматриваемого явления. Таким образом, мы соприкасаемся здесь с одной из важнейших сторон человеческой деятельности — процессом познания. Тезис о том, что «критерий истины есть практика» имеет самое непосредственное отношение к математической статистике, поскольку именно эта наука изучает методы (в рамках точных математических моделей), которые позволяют отвечать на вопрос, соответствуют ли практика, представленная в виде результатов эксперимента, данному гипотетическому представлению *о* природе явления или нет.

При этом необходимо подчеркнуть, что, как и в теории вероятностей, нас будут интересовать не те эксперименты, которые позволяют делать однозначные, детерминированные выводы о рассматриваемых в природе явлениях, а эксперименты, результатами которых являются случайные события. С развитием науки роль такого рода задач становится все больше и больше, поскольку с увеличением точности экспериментов становится все труднее избежать «случайного фактора», связанного с разного рода помехами и ограниченностью наших измерительных и вычислительных возможностей.

Математическая статистика является частью теории вероятностей в том смысле, что каждая задача математической статистики есть по существу задача (иногда весьма своеобразная) теории вероятностей. Однако сама по себе математическая статистика занимает и самостоятельное положение в табеле о науках. Математическая статистика может рассматриваться как наука о так называемом индуктивном поведении человека (и не только человека), в условиях, когда он должен на основании своего недетерминированного опыта принимать решения с наименьшими для него потерями.

Пример 1. Для многих изделий одним из основных параметров, которым характеризуется качество, является срок службы. Однако срок службы изделия (скажем, электролампы), как правило, случаен, и заранее определить его невозможно. Опыт показывает, что если процесс производства в известном смысле однороден, то сроки службы $\xi_1, \xi_2, \ldots$ соответственно 1-го, 2-го и т.д. изделий можно рассматривать как независимые одинаково распределенные величины. Интересующий нас параметр, определяющий срок службы, естественно

отождествить с числом $\theta = M\xi_i$. Одна из стандартных задач состоит в выяснении, чему равно θ. Для того чтобы определить это значение, берут n готовых изделий и проверяют их. Пусть $x_1, x_2, ..., x_n$ — сроки службы этих проверенных изделий. Мы знаем, что $\frac{1}{n}\sum_{i=1}^{n}\xi_i \underset{п.н.}{\to}\theta$ при $n \to \infty$. Поэтому естественно ожидать, что число $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ при достаточно большом n окажется близким к θ и позволит в какой-то мере ответить на поставленные вопросы. При этом очевидно, что мы заинтересованы в том, чтобы требуемое число наблюдений n было по возможности наименьшим, а наша оценка числа θ по возможности более точной (завышение параметра θ как и его занижение, приведут к материальным потерям).

Пример 2. Радиолокационное устройство в моменты времени $t_1, t_2, ..., t_n$ зондирует заданную часть воздушного пространства с целью обнаружения там некоторого объекта. Обозначим $x_1, x_2, ..., x_n$ значения отраженных сигналов, принятых устройством. Если в заданной части пространства интересующий нас объект отсутствует, то значения x_i можно рассматривать как независимые случайные величины, распределенные так же, как некоторая случайная величина ξ, природа которой обусловлена характером различных помех. Если же в течение всего периода наблюдений объект находился в поле зрения, то x_i будут наряду с помехами содержать «полезный» "сигнал a, и значения x_i будут распределены как $\xi + a$. Таким образом, если в первом случае наблюдения x_i, имели функцию распределения $F(x)$, то во втором случае их функция распределения будет иметь вид $F(x-a)$. По выборке $x_1, x_2, ..., x_n$ требуется решить, какой из этих двух случаев имеет место, т. е. существует в заданном месте интересующий нас объект или нет.

В этой задаче окажется возможным указать в известном смысле «оптимальное решающее правило», которое будет решать поставленную задачу с минимальными ошибками. Сформулированная задача может быть усложнена следующим образом. Сначала объект отсутствует, а затем, начиная с наблюдения неизвестным номером θ, появляется. Требуется по возможности более точно определить момент θ появления объекта. Это так называемая «задача о разладке», имеющая и целый ряд других интерпретаций, важных для приложений.

Пример 3. Некоторый эксперимент производится сначала n_1, раз в условиях A и затем n_2 раз в условиях B. Обозначим $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$ результаты этих экспериментов соответственно в условиях А и В. Спрашивается, сказывается ли изменение условий эксперимента на его результатах? Иными словами, если обозначить через P_A распределение x_i, $1 \le i \le n_1$ и через P_B — распределение y_i, $1 \le i \le n_2$, то вопрос состоит том, выполнено соотношение $P_A = P_B$ или нет.

Например, если нужно установить, влияет ли некоторый препарат на развитие, скажем, растений или животных, то параллельно ставятся две серии экспериментов (с препаратом и без), результаты которых необходимо уметь сравнивать.

Часто возникают и более сложные задачи, когда аналогичный вопрос ставится для многих серий наблюдений, проведенных в различных условиях. Если результаты наблюдений зависят от условий, то бывает необходимым проверить тот

или иной характер этой зависимости (так называемая задача о регрессии).

Пример 3 и названные более сложные проблемы относятся к классу статистических задач *с двумя и более выборками.*

Список примеров типичных статистических задач, разных по сложности и по своему существу, можно было бы продолжить. Однако общими для всех них будут следующие два обстоятельства:

1. Перед нами не было бы никаких проблем, если бы распределения результатов наблюдений, которые фигурируют в задачах, были нам известны.

2. В каждой из этих задач мы должны по результатам экспериментов принимать какое-то решение относительно распределения имеющихся наблюдений (отсюда и название «Теория статистических решений», упоминавшееся выше).

В связи с этими двумя замечаниями принципиальное значение для всего дальнейшего и, в частности, для решения приведенных в качестве примеров задач, приобретает следующий факт. Оказывается, по результатам наблюдений $x_1, x_2, ..., x_n$ над некоторой величиной ξ можно при больших n сколь угодно точно восстановить неизвестное распределение P этой случайной величины. Аналогичное утверждение справедливо и для любого функционала $\theta = \theta(P)$ от этого неизвестного распределения. Этот факт лежит в основе математической статистики.

Вопросы для самоконтроля

1. Что такое математическая статистика?
2. С решением каких задач она связана дисциплина математическая статистика?
3. Привести примеры

Тема 1. Выборка. Эмпирическое распределение

1.1. Основные понятия и соотношения

Множество всех возможных значений случайной величины ξ, распределенной по закону F, называется ***генеральной совокупностью*** F.

Множество $\{x_1, x_2, \ldots, x_n\}$ отдельных значений случайной величины ξ, полученных в серии из n независимых экспериментов (наблюдений), называется ***выборочной совокупностью или выборкой*** объема n из генеральной совокупности.

Выборка $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$, в которой элементы упорядочены по возрастанию, называется ***вариационным рядом***. В вариационном ряду некоторые элементы могут совпадать. Совпадающие элементы объединяют в группы:

$$\underbrace{x_1'}_{1}, \underbrace{x_2', x_3'}_{2} \ldots, \underbrace{x_{n-3}'}_{k-1}, \underbrace{x_{n-2}', x_{n-1}', x_n'}_{k}$$

При большом объеме выборки когда число значений случайной величины X велико или случайная величина является непрерывной строят группированный статистический ряд, что существенно уменьшает вычислительную работу.

Группировка осуществляется следующим образом:

1. Выявляется диапазон выборочных значений от самого меньшего до самого большого.

2. Весь диапазон разбивается на k интервалов или разрядов. Интервалы могут быть равными или неравными между собой. Обычно используют 10-20 интервалов. Шаг интервалов вычисляют по формуле Стерджеса $h = \dfrac{x_{\max} - x_{\min}}{1 + \log_2 n}$, начало первого $x_{нач} = x_{\min} - n/2$.

3. Подыскивается число выборочных значений, попавшее в каждый интервал n_i (i - номер интервала). Сумма $\sum_{i=1}^{m} n_i = n$ - объем выборки. Если выборочное значение попадает точно на границу между интервалами, то заранее нужно договориться куда его относить. Можно, например, прибавлять по $\dfrac{1}{2}$ к числам значений в интервалах справа и слева.

4. Подсчитываются частоты интервалов $\omega_i = n_i / n$, очевидно что $\sum_{i=1}^{m} \omega_i = 1$

5. Находят середины интервалов $\overline{x_i}$, $i = \overline{1, m}$ и составляют таблицу, в которую

заносят середины интервалов и частоты ω_i. Полученный материал называется группировочным рядом относительных частот.

Совокупность пар чисел $(\overline{x}_i, n_i)$, где $\overline{x}_i$, $i=\overline{1,m}$ – наблюдаемые, неповторяющиеся (для непрерывного распределения) в выборке значения, а n_i - число этих значений в выборке, называется ***статистическим рядом абсолютных частот***. Совокупность пар чисел $(\overline{x}_i, \omega_i)$, где $\omega_i = n_i / n$ называется ***статистическим рядом относительных частот***. Совокупность пар чисел $\left(\overline{x}_i, \sum_{k=1}^{i} \omega_k\right)$ называется ***статистическим рядом накопленных частот***. Статистические ряды отображают в виде таблицы:

$\overline{x}_i$	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_m$
n_i	n_1	n_2	…	n_m
ω_i	ω_1	ω_2	…	ω_m
$\sum_{k=1}^{i} \omega_k$	ω_1	$\omega_1 + \omega_2$	…	1

Подобного вида статистический ряд используют обычно для описания выборки из генеральной совокупности с дискретным распределением. В этом случае статистический ряд относительных частот приближенно оценивает ряд распределения дискретной случайной величины.

Ломаная, отрезки которой соединяют точки $(\overline{x}_i, \omega_i)$, называется ***полигоном частот***. Для дискретной случайной величины полигон частот является оценкой многоугольника распределения.

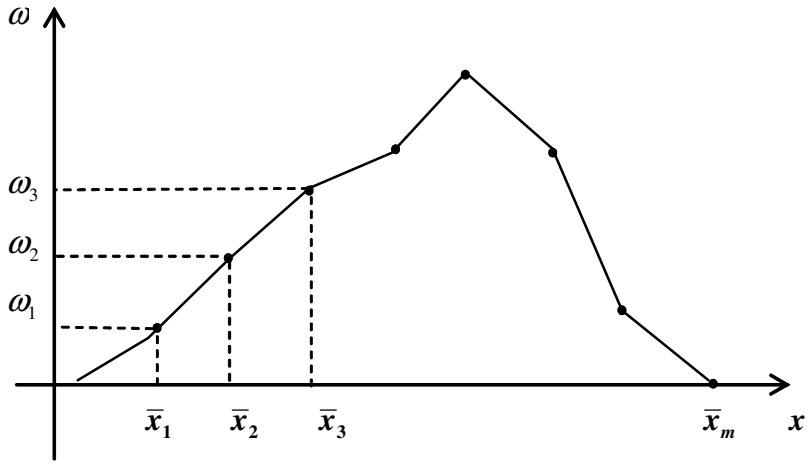

Рис. 1.1. Полигон частот

Пример.

Даны результаты измерений:

178, 160, 154, 183, 155, **153**, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 186, 158, 171, 175, 173, 184, 172

$x_{\min}$=153, $x_{\max}$=186

$h = \dfrac{186-163}{1+\log_2 30} = 5.59$ шаг разбиения выборки.

Примем $h = 6$, $x_{нач} = 153 - \dfrac{6}{2} = 150$.

Исходные данные разбиваем на 6 интервалов:
[150;156), [156,162), [162,168), [168,174),[174,180),[180,186).

Подсчитав число студентов (n_i), попавших в каждый из полученных промежутков, получим интервальный статистический ряд:

$\overline{x}_i$	153	159	165	171	177	183
Интервалы	[150;156)	[156,162)	[162,168)	[168,174)	[174,180)	[180,186)
Частота	4	5	6	7	5	3
Относительная частота ω_i	0.13	0.17	0.2	0.23	0.17	0.1

Для описания выборки из совокупности с непрерывным распределением используют также ***сгруппированные статистические ряды***. Для этого интервал, в котором содержатся все элементы выборки, делится на m равных (или неравных) последовательных, непересекающихся интервалов $\tilde{x}_0 \div \tilde{x}_1, \tilde{x}_1 \div \tilde{x}_2, \ldots, \tilde{x}_{m-1} \div \tilde{x}_m$, и подсчитывают частоты n_i - число элементов выборки, попавших в i-ый интервал. Число интервалов группирования определяют, например, по формуле Стерджесса: $m = 1 + [\log_2 n] \approx 1 + 4 \cdot \lg n$. В результате получаем следующий статистический ряд:

$\overline{x}_i$	$\overline{x}_1$	$\overline{x}_2$	…	$\overline{x}_m$
n_i	n_1	n_2	…	n_m
ω_i	ω_1	ω_2	…	ω_m
ρ_i	ρ_1	ρ_2	…	ρ_m
$\sum \omega_i$	ω_1	$\omega_1 + \omega_2$	…	1

Здесь $\overline{x}_i = \dfrac{\tilde{x}_{i-1} + \tilde{x}_i}{2}$ - середины интервалов группирования, $\rho_i = \dfrac{\omega_i}{\Delta x_i} = \dfrac{\omega_i}{\tilde{x}_i - \tilde{x}_{i-1}}$ - плотность частоты.

В качестве оценки кривой плотности непрерывного распределения используется ***гистограмма частот*** - ступенчатая фигура, состоящая из m прямоугольников, опирающихся на частичные интервалы (см. рисунок). Высота i-го прямоугольника полагается равной плотности частоты ρ_i. Соответственно площадь каждого прямоугольника равна $\rho_i \cdot \Delta x_i = \omega_i$ - относительной частоте.

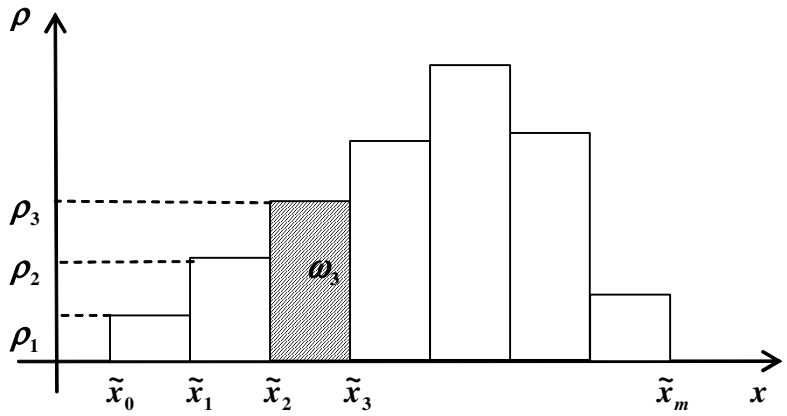

Рис 1.2. Гистограмма частот.

Эмпирической функцией распределения, полученной по выборке $\{x_1, x_2, \ldots, x_n\}$, называется функция, при каждом $x \in R$ равная:

$$F_n^*(x) = \frac{\text{количество } x_i < x}{n}.$$

$F_n^*(x)$ есть ступенчатая функция. Эмпирическая функция распределения является оценкой теоретической функции распределения (рис 3).

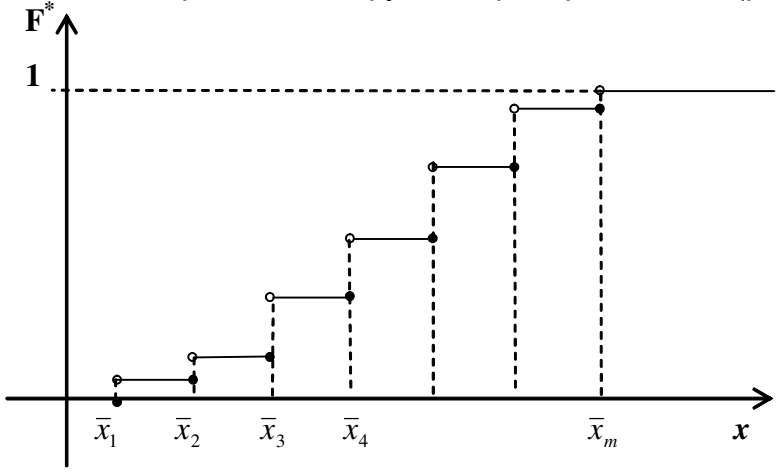

Рис 1.3. Эмпирическая функция распределения.

Теорема [1]. При $n \to \infty$ выборочное распределение $F^*(x)$ стремится к

исходному распределению $F(x)$ почти наверное (с вероятностью 1)

$$F^*(x) \underset{п.н.}{\to} F(x). \blacktriangleleft$$

1.2. Числовые характеристики выборки

В качестве ***числовых характеристик выборки*** используются:

1. Выборочное среднее: $\overline{m} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
2. Выборочная дисперсия $\overline{D} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2$.
3. Несмещенная выборочная дисперсия $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$.
4. Выборочные начальные и центральные моменты

$$\overline{m}_k = \overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} x^k, \qquad \overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(x - \overline{X}\right)^k.$$

5. Выборочная медиана x^* – это среднее значение вариационного ряда

$$x^* = x_m, \ \textit{если} \ n = 2m-1 \ (\textit{нечетно})$$

$$x^* = (x_m + x_{m+1})/2, \ \textit{если} \ n = 2m \ (\textit{четно})$$

Напомним, что медианой ξ непрерывного распределения F называется решение уравнения $F(\xi) = 1/2$

Более общим понятием является понятие квантили ξ_p порядка p. Это число ξ_p, для которого $F(\xi_p) = p$. Так что медиана есть квантиль порядка 1/2. Если F имеет точки разрыва (дискретную компоненту), то это понятие теряет смысл. Поэтому в общем случае мы будем пользоваться следующим определением.

Квантилью ξ_p порядка p распределения F называется число

$$\xi_p = \sup\{x : F(x) \le p\}$$

Как функция от p квантиль ξ_p есть не что иное, как функция $F^{-1}(p)$

6. Выборочная квантиль x_p порядка p равна

$$x_p = x_l, \quad l = [np] + 1.$$

По статистическому ряду значения первых четырех величин могут быть найдены по формулам:

$$\overline{m} = \sum_{i=1}^{m} \overline{x}_i \omega_i, \qquad \overline{D} = \sum_{i=1}^{m} (\overline{x}_i - \overline{m})^2 \omega_i, \qquad s^2 = \frac{n}{n-1}\sum_{i=1}^{m}\left(\overline{x}_i - \overline{m}\right)^2 \omega_i,$$

$$\overline{m}_k = \sum_{i=1}^{m} \overline{x}_i^k \omega_i, \qquad \overline{\mu}_k = \sum_{i=1}^{m}\left(\overline{x}_i - \overline{m}\right)^k \omega_i.$$

Выборочные характеристики являются приближенными значениями соответствующих числовых характеристик случайной величины ξ.

Вопросы для самоконтроля

1. Понятие выборки и формы ее записи.
2. Что такое вариационный ряд?
3. Что такое статистический ряд абсолютных частот?
4. Что такое статистический ряд относительных частот?
5. Что такое статистический ряд накопленных частот.
6. Что такое группированный статистический ряд?
7. Что такое полигон частот, гистограмма?
8. Эмпирическая функция распределения.
9. Числовые характеристики выборки

Тема 2. Точечные оценки параметров распределений вероятностей

2.1 Точечные и интервальные оценки

В реальной жизни практически никогда не бывает так, чтобы исследователь располагал точным знанием закона распределения вероятностей наблюдаемых случайных величин. Ему в общем случае неизвестны как сам закон распределения вероятностей, так и его параметры. В распоряжении исследователя имеется лишь совокупность результатов наблюдений, и. основываясь только на них, он должен сделать выводы о параметрах распределения, если вид закона распределения вероятностей ему известен. Если же нет, то и сам закон распределения вероятностей ему придется выбирать на основании выборочных результатов наблюдений. Здесь мы рассмотрим методы оценки параметров $\theta_1,\theta_2,...,\theta_k$ различных, заранее определенных по форме, распределений вероятностей случайных величин $F(x,\theta_1,\theta_2,...,\theta_k)$. Всякая оценка неизвестного параметра θ по выборке является функцией от выборочных значений $x_1,x_2,...,x_n$, т.е. $\theta^*(x_1,x_2,...,x_n)$. Числовые характеристики $\theta_1^*,\theta_2^*,...,\theta_k^*$, полученные по выборкам, называют статистическими оценками параметров.

Различают два вида оценок параметров *точечные* и *интервальные.* Предположим, оценке подлежит параметр θ некоторого распределения вероятностей по выборочным данным $x_1,x_2,...,x_n$ некоторой случайной величины X. Точечной оценкой параметра θ по выборочным данным является некоторый функционал $\theta^*=\varphi(x_1,x_2,...,x_n)$, позволяющий получить наилучшую оценку в принятых критериях.

В качестве критериев, характеризующих пригодность оценки параметра распределения, используются такие ее свойства, как *несмещенность, состоятельность, эффективность* и *достаточность.*

Оценка θ^* параметра θ называется *несмещенной*, если для любого фиксированного объема выборки n математическое ожидание оценки равно оцениваемому параметру, т.е.

$$M(\theta^*)=\theta. \qquad (2.1)$$

Поясним смысл этого равенства следующим примером. Имеются два алгоритма вычисления оценок для параметра θ. Значения оценок, построенных первым алгоритмом по различным выборкам объема n генеральной совокупности, приведены на рис 2.1,*а*, а с использованием второго алгоритма – на рис 2.1,*б*. Видим, что среднее значение оценок на рис 3.1,*а* совпадает с θ, и, естественно, такие оценки предпочтительнее по сравнению с оценками рис 2.1,*б*, которые концентрируются слева от значения θ и для которых $M(\theta^*)<\theta$, т.е. эти оценки являются смещенными.

Оценка θ^* называется *состоятельной*, если

$$\theta_n^* \xrightarrow{p} \theta,$$

т.е. для любого $\varepsilon>0$ при $n\to\infty$

$$P\left(\left|\theta_n^* - \theta\right| < \varepsilon\right) \to 1 . \qquad (2.2)$$

Поясним смысл этого предельного соотношения. Пусть ε – очень малое положительное число. Тогда (2.2) означает, что чем больше число наблюдений n, тем больше уверенность (вероятность) в незначительном отклонении θ_n^* от неизвестного параметра θ. Очевидно, что «хорошая оценка» должна быть состоятельной, иначе эта оценка не имеет практического смысла, так как увеличение объема исходной информации не будет приближать нас к «истинному» значению θ.

Предположим, что имеются две состоятельные и несмещенные оценки

$$\theta_n^{*(1)} = \varphi_1(x_1, ..., x_n); \quad \theta_n^{*(2)} = \varphi_2(x_1, ..., x_n) \qquad (2.3)$$

одного и того же параметра θ. Как из двух этих оценок выбрать лучшую? Каждая из них является случайной величиной, и мы не можем предсказать индивидуальное значение оценки в каждом частном случае. Однако, рассматривая в качестве меры концентрации распределения оценки θ_n^* около значения параметра θ величину $M(\theta_n^* - \theta)^2$, мы можем теперь точно охарактеризовать сравнительную эффективность оценок $\theta_n^{*(1)}$ и $\theta_n^{*(2)}$. В качестве меры эффективности принимается отношение

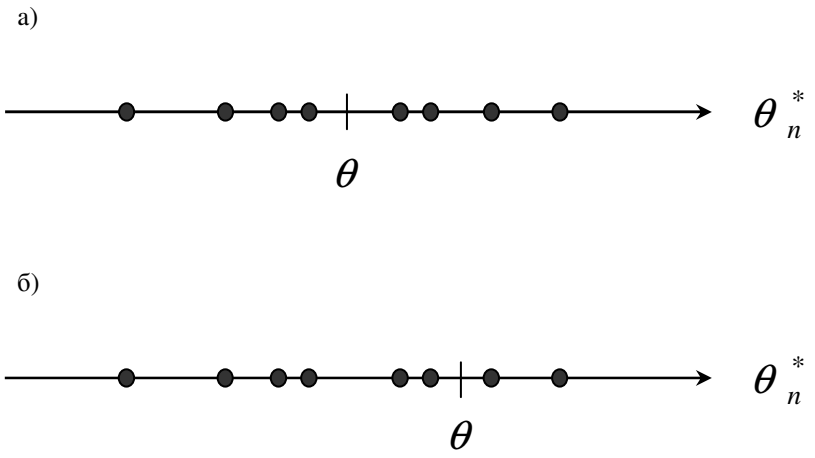

Рис. 2.1. К определению несмещенной оценки

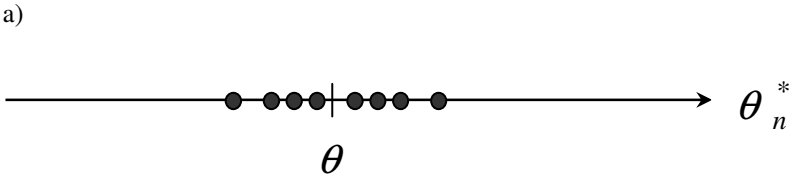

Рис. 2.2. К определению эффективной оценки

$$e = \frac{M(\theta_n^{*(1)} - \theta)^2}{M(\theta_n^{*(2)} - \theta)^2}. \qquad (2.4)$$

Если значение $e > 1$, то оценка $\theta_n^{*(2)}$ более эффективна, чем $\theta_n^{*(1)}$. В случае несмещенных оценок $M(\theta_n^{*(1)}) = \theta, M(\theta^{*(2)}) = \theta$ и поэтому

$$e = \frac{D(\theta_n^{*(1)})}{D(\theta_n^{*(2)})}, \qquad (2.5)$$

где $D(\theta_n^*)$ – дисперсия оценки θ_n^*.

Таким образом, несмещенная оценка θ_n^* параметра θ называется *несмещенной эффективной*, если она среди всех других несмещенных оценок того же параметра обладает *наименьшей дисперсией*.

Приведенная на рис 2.2,*а* оценка θ^* является более эффективной по равнению с оценкой, значения которой нанесены на рис 2.2,*б*.

Оценка θ^* называется достаточной, если оценка θ^* извлекает максимальную информацию из выборки.

Под интервальной оценкой параметра θ понимается интервал, границы которого $a_н^*$ и $a_в^*$ являются функционалами от выборочных значений случайной величины, и который с заданной вероятностью α содержит оцениваемый параметр: $P\{a_н^* < \theta < a_в^*\} = \alpha$. Вероятность α называется *доверительной вероятностью,* а оценки $a_н^*$ и $a_в^*$ – соответственно *нижней* и *верхней доверительными границами.* Интервал $[a_н^*, a_в^*]$ называется *доверительным, интервалом.* Если длина доверительного интервала $l(\alpha) = a_в^* - a_н^* = const$, то для состоятельных и несмещенных оценок $\alpha \to 1$ при $n \to \infty$. При фиксированном объеме выборки n, α будет тем больше, чем больше l.

Различают два вида интервальных оценок: *одно-* и *двусторонние.* При *двусторонней* оценке задаются обе границы доверительного интервала, так что

$$P\{a_н^* < \theta < a_в^*\} = \alpha \text{ и } \quad P\{\theta < a_н^*\} = \alpha'; \; P\{\theta > a_в^*\} = \alpha'',$$

где $\alpha' + \alpha'' = 1 - \alpha$. Если $\alpha' = \alpha'' = \frac{1-\alpha}{2}$, то двусторонний доверительный интервал называется *симметричным.* Для него справедливы соотношения

$$P\{\theta < a_в^*\} = \frac{1+\alpha}{2} \text{ и } P\{a > a_н^*\} = \frac{1+\alpha}{2}.$$

При односторонних доверительных интервалах границы интервалов задаются так, чтобы

$$P\{a < a_в^*\} = \alpha \text{ или } P\{\theta > a_н^*\} = \alpha$$

Величина $q = (1-\alpha)$ — дополнение .доверительной вероятности до единицы называется *уровнем значимости.* Этим термином обозначается вероятность появления события, которую исследователь связывает с *неслучайным (значимым) событием.* Очевидно, что двусторонний интервал для симметричных распределений аналогичен одностороннему при удвоенном уровне значимости.

Наиболее существенной характеристикой оценки параметра распределения является ее *эффективность.* Именно эта характеристика обычно используется для сравнения методов оценки параметров распределения между собой. Как правило, эффективность оценки сравнивается с эффективностью оценки параметра распределения методом максимального правдоподобия (т.е. с наиболее эффективной оценкой). Легко видеть, что применение менее эффективных оценок (требующих, как правило, меньшего объема, вычислений) может быть скомпенсировано соответствующим увеличением объема выборки.

Поясним практический смысл процедуры оценки параметров распределения вероятностей. Так как само распределение наблюдаемых случайных величин является для исследователя той совокупностью данных, которой он располагает относительно наблюдаемого процесса, то и параметры распределения позволяют судить об основных чертах этого процесса. Например, когда мы спрашиваем, какова долговечность прибора, мы по сути, ставим задачу оценки среднего значения (или математического ожидания) наблюдаемого распределения показателей долговечности. Если нас интересует, насколько стабилен наблюдаемый технологический процесс, то ответ на этот вопрос требует оценки разброса (рассеяния) наблюдаемых случайных величии, характеризующих качество технологического процесса.

2.2 Вычисление точечных оценок

Для нахождения вида функции оценивания того или иного параметра используют один из следующих методов:

1) метод максимального правдоподобия;
2) метод моментов;
3) оценивание с помощью метода наименьших квадратов

2.2.1. Оценка параметров методом максимального правдоподобия

Наибольшее распространение получил метод максимального правдоподобия. Суть метода состоит в следующем. Пусть $X = (x_1, x_2, ..., x_n)$ – независимая выборка из распределения F_θ, зависящего от неизвестного параметра $\theta = (\theta_1, \theta_2, ..., \theta_k) \in \Theta \subset R^k$. Функцией правдоподобия $L(\theta, x) = L(\theta_1, ..., \theta_k; x_1, ..., x_n)$ называют функцию

$$L(\theta, x) = \begin{cases} \prod_{j=1}^{n} f(x_j, \theta_1, ..., \theta_k), & \textit{если } x \textit{ непрерывная величина} \\ \prod_{j=1}^{n} p(x_j, \theta_1, ..., \theta_k), & \textit{если } x \textit{ дискретная величина} \end{cases}$$

В качестве оценки параметров $\theta^* = (\theta^*_1, ..., \theta^*_k)$ примем значения этих параметров, при которых функция правдоподобия принимает максимальное значение, т.е. $\theta^* = (\theta^*_1, ..., \theta^*_k) = \arg\left(\max_\theta L(\theta, x)\right)$. Если функция $L(\theta, x) = L(\theta_1, ..., \theta_k; x_1, ..., x_n)$ является дифференцируемой по переменным $\theta_1, \theta_2, ..., \theta_k$, оценки параметров удовлетворяют системе уравнений:

$$\frac{dL(\theta_1, ..., \theta_k; x)}{d\theta_i} = 0, \; i = 1, 2, ..., k.$$

2.2.2. Оценка параметров методом моментов

Идея этого метода заключается в приравнивании теоретических и эмпирических моментов.

Пусть $X = (x_1, x_2, ..., x_n)$ – независимая выборка из распределения F_θ, зависящего от неизвестного параметра $\theta = (\theta_1, \theta_2, ..., \theta_k) \in \Theta \subset R^k$. Моментом i -го порядка называется функция

$$\mu_i(\theta_1, ..., \theta_k) = E[x^i] = \begin{cases} \int x^i f(x, \theta_1, ..., \theta_k) dx, & \textit{если } x \textit{ непрерывная величина} \\ \sum_j x_j^i p(x_j, \theta_1, ..., \theta_k), & \textit{если } x \textit{ дискретная величина} \end{cases}$$

где $f(x, \theta)$ – плотность распределения непрерывной случайной величины x, $p(x_j, \theta)$ – вероятность дискретной случайной величины. Теоретический момент является функцией неизвестных параметров $\mu_i(\theta) = \mu_i(\theta_1, \theta_2, ..., \theta_k)$.

Выборочным (эмпирическим) моментом i -го порядка называется величина

$$m_i(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j^i .$$

Отметим, что по своему определению эмпирические моменты являются функциями от выборки.

Для нахождения неизвестных параметров (будем обозначать их $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)$) составим систему уравнений

$$\mu_1(\hat{\theta}_1, ..., \hat{\theta}_k) = m_1,$$
$$\mu_2(\hat{\theta}_1, ..., \hat{\theta}_k) = m_2,$$
$$.....$$
$$\mu_k(\hat{\theta}_1, ..., \hat{\theta}_k) = m_k.$$

Далее решаем систему относительно параметров $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_k)$. В результате получим

$$\hat{\theta}_1 = \hat{\theta}_1(x_1, ..., x_n),$$
$$\hat{\theta}_2 = \hat{\theta}_2(x_1, ..., x_n),$$
$$.....,$$
$$\hat{\theta}_k = \hat{\theta}_k(x_1, ..., x_n).$$

Найденные параметры зависят от выборки $x = (x_1, x_2, ..., x_n)$.

2.2.3. Оценка параметров методом наименьших квадратов

Пусть дана табличная функция $y(x)$

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

Необходимо аппроксимировать эти данные некоторой параметрической функцией $f(\theta_1, ..., \theta_k; x)$, т.е. заменить функцию $y(x)$ функцией $f(\theta_1, ..., \theta_k; x)$:

$y(x) \approx f(\theta_1, ..., \theta_k; x)$.

Параметры $\theta_1, ..., \theta_k$ будем подбирать таким образом, чтобы расхождение табличной

функции с функцией $f(\theta_1,...,\theta_k;x)$ было минимальным. Для этого построим функционал $F(\theta_1,...,\theta_k)=\sum_{j=1}^{n}\left(y_i-f(\theta_1,...,\theta_k;x_j)\right)^2$ и найдем его минимум.

Необходимое условие минимума имеет вид

$$\frac{dF(\theta_1,...,\theta_k)}{d\theta_i}=0,\ i=1,...,k.$$

Решаем эту систему уравнений и получаем значения параметров $\hat{\theta}=(\hat{\theta}_1,...,\hat{\theta}_k)$.

Пример 2.1. Пусть дана независимая выборка $x=(x_1,x_2,...,x_n)$ из распределения F_θ. Разобьем весь диапазон данных $[x_{\min},x_{\max}]$ на m интервалов и построим гистограмму. Обозначим середины интервалов $\overline{x}_i=\frac{\tilde{x}_{i-1}+\tilde{x}_i}{2},\ i=1,2,...,m$. Тогда плотность частоты $\rho_i=\frac{\omega_i}{\Delta x_i}=\frac{\omega_i}{\tilde{x}_i-\tilde{x}_{i-1}}$ (высоты столбиков гистограммы) будут значениями табличной функции $y_i,\ i=1,2,...,m$. Здесь $\omega_i=n_i/n$ – относительные частоты.

Таким образом, мы получили табличную функцию

$\tilde{x}$	$\tilde{x}_1$	$\tilde{x}_2$	...	$\tilde{x}_n$
ρ	ρ_1	ρ_2	...	ρ_n

Поставим следующую задачу. Подобрать параметры известного закона непрерывного распределения $f(\theta_1,...,\theta_k;x)$ так, чтобы расхождение между гистограммой и функцией $f(\theta,x)$ было минимально. В результате мы приходим к методу наименьших квадратов. ►

2.3. Точечная оценка параметров нормального распределения

Пусть случайная величина X имеет нормальное распределение $N(\mu,\sigma^2)$, где $\sigma>0;\ \mu\in R$. Плотность распределения имеет вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in(-\infty,\infty).$$

Решение. Запишем функцию правдоподобия

$$L(\mu,\sigma^2;x)=\prod_{j=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^{n}\frac{(x_j-\mu)^2}{2\sigma^2}}.$$

Из условия максимума $\frac{dL(\mu,\sigma^2;x)}{d\mu}=0,\ \ \frac{dL(\mu,\sigma^2;x)}{d\sigma}=0$, получим следующие оценки среднего и дисперсии

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2}-\bar{x}^2.$$

Оценка среднего – состоятельная, несмещенная, эффективная, достаточная и распределена как случайная величина тоже нормально со средним $M(\bar{x})=\mu$ и дисперсией $D(\bar{x})=\frac{\sigma^2}{n}$.

Оценка дисперсии – состоятельная, эффективная, достаточная, но смещенная. При $n<30$ рекомендуется использовать несмещенную оценку

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^{\,2}-n\bar{x}^2\right).$$

2.4. Точечная оценка параметров показательного закона распределения

Пусть $X\in\Pi_\lambda$, где Π_λ – показательный закон распределения с параметром λ с плотностью распределения $f(x)=\lambda e^{-\lambda x},\ x\in[0,\infty)$. Найти оценку параметра λ^* методом максимального правдоподобия.

Запишем функцию правдоподобия

$L(\lambda,x)=\prod_{j=1}^{n}\lambda e^{-\alpha x_j}=\lambda^n e^{-\lambda\sum_{j=1}^{n}x_j}$. Из условия максимума $\frac{dL(\lambda;x)}{d\lambda}=0$ получим следующее уравнение: $n\lambda^{n-1}e^{-\lambda\sum_{j=1}^{n}x_j}-\lambda^n\sum_{j=1}^{n}x_j e^{-\lambda\sum_{j=1}^{n}x_j}=0$. Отсюда следует

$$\lambda^*=\frac{1}{\frac{1}{n}\sum_{j=1}^{n}x_j}.$$

Найдем теперь оценку параметра λ^* методом моментов.

$\mu_1(\lambda)=E[x]=\int_0^\infty x\cdot\lambda e^{-\lambda x}dx=\frac{1}{\lambda}$. Приравниваем к $m_1=\frac{1}{n}\sum_{j=1}^{n}x_j$. Отсюда получим:

$\lambda^* = \frac{1}{m_1} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} x_j}$. Таким образом, оценки параметра λ, полученные методом максимального правдоподобия и методом моментов, совпали.

2.5. Точечная оценка параметров равномерного закона распределения

Пусть $X \in U_{\alpha,\beta}$, где $U_{\alpha,\beta}$ – равномерный закон распределения с параметрами α, β. Найдем оценки параметров $\hat{\alpha}$ и $\hat{\beta}$ методом моментов.

$$m_1(\alpha,\beta) = E[x] = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x dx = \frac{\alpha+\beta}{2},\ m_2(\alpha,\beta) = E[x^2] = \frac{1}{\beta-\alpha}\int_{\alpha}^{\beta} x^2 dx = \frac{\beta^2+\alpha\beta+\alpha^2}{3}.$$

Получим систему

$$\frac{\alpha+\beta}{2} = m_1,$$
$$\frac{\beta^2+\alpha\beta+\alpha^2}{3} = m_2.$$

Решение системы: $\hat{\beta} = m_1 + \sqrt{3}\sqrt{m_2 - m_1^2} = m_1 + \sqrt{3}\sigma$, $\hat{\alpha} = m_1 - \sqrt{3}\sigma$. Здесь $\sigma^2 = m_2 - m_1^2$ – дисперсия выборочного распределения. ►

2.6. Точечная оценка параметров биномиального закона распределения

Биномиальное распределение — распределение количества «успехов» в последовательности из n независимых случайных экспериментов, таких что вероятность «успеха» в каждом из них постоянна и равна p.

Определение. Пусть $x_1, x_2, ..., x_n$ — конечная последовательность независимых случайных величин с распределением Бернулли, то есть

$$x_i = \begin{cases} 1 & p \\ 0 & q = 1-p \end{cases},\quad i = 1,2,...,n$$

Построим случайную величину Y:

$$x = \sum_{i=1}^{n} x_i$$

Тогда x – число единиц (успехов) в последовательности $x_1, x_2, ..., x_n$, имеет биномиальное распределение с n степенями свободы и вероятностью «успеха» p. Распределения вероятностей задаётся формулой:

$$f(x;n,p)=\binom{n}{x}\cdot p^x\cdot q^{n-x},\ k=0,1,...,n,$$

где $\binom{n}{x}=\frac{n!}{(n-x)!x!}$ — биномиальный коэффициент

Здесь x – количество появлений события в серии из n испытаний, при условии, что в единичном испытании вероятность его появления равна p.

Функция распределения биномиального распределения может быть записана в виде суммы:

$$F(x;n,p)=\sum_{i=0}^{x}\binom{n}{i}\cdot p^i\cdot q^{n-i},$$

Среднее значение равно $E[x]=np$, а дисперсия – $D[x]=npq$.

Биномиальное распределение зависит от одного параметра p.

Если имеется реализация из n испытаний, в которых событие наблюдалось m раз, то несмещенной точечной оценкой максимального правдоподобия параметра p является величина $p_n=\frac{m}{n}$. Это следует из решения уравнения

$$\left.\frac{df(x;n,p)}{dp}\right|_{x=m}=0.$$

2.7. Планирование экспериментов для оценки параметров распределений

2.7.1. Нормальное распределение

Оценка среднего при известной дисперсии

Объем выборки, необходимый для оценки среднего μ с заданной предельной абсолютной ошибкой ε и доверительной вероятностью α при известной дисперсии σ^2 определяется соотношением

$$n=\left(\frac{u_\alpha\sigma}{\varepsilon}\right)^2.$$

Для α - квантили стандартного нормального распределения можно использовать аппроксимацию $u_\alpha=4,91[\alpha^{0,14}-(1-\alpha)^{0,14}]$. Тогда имеем

$$n=24,108\left\{\frac{\sigma}{\varepsilon}\left[\alpha^{0,14}-(1-\alpha)^{0,14}\right]\right\}^2.$$

Пример 2.2. Напряжение зажигания газоразрядного прибора распределено нормально со стандартным, отклонением $\sigma=50$ В. Найти объем выборки, позволяющий оценить среднее значение напряжения зажигания с предельной абсолютной ошибкой $\varepsilon=20$ В при доверительной вероятности $\alpha=0,95$.

Решение. Имеем $n = 24{,}108\left\{\frac{50}{20}\left[0{,}95^{0{,}14} - (1-0{,}95)^{0{,}14}\right]\right\}^2 = 17$. Следовательно, желаемая точность оценки с вероятностью $\geq$ 0,95 достигается при объеме выборки $n \geq 17$.►

Оценка среднего при неизвестной дисперсии

Необходимый объем выборки определяется из соотношения

$$\delta = \frac{\varepsilon}{\bar{x}} = \frac{t_\alpha(n)}{\sqrt{n}} \cdot \frac{s}{\bar{x}},$$

где $t_\alpha(n)$ – α-квантиль распределения Стьюдента при $v = n$ степенях свободы; s и $\bar{x}$ – выборочные оценки соответственно стандартного отклонения и среднего значения. Здесь δ – относительная погрешность среднего.

Значения $\frac{t_\alpha(n)}{\sqrt{n}}$ приведены в табл. 4 (см. статистические таблицы).

Определение объема выборки происходит в следующей последовательности.

Сначала по заданным величинам $\delta = \frac{\varepsilon}{\bar{x}}$ и α и предполагаемому коэффициенту вариации $w = \frac{s}{\bar{x}}$ по табл. 3 находим значение $\frac{t_\alpha(n)}{\sqrt{n}} = \delta / w = \frac{\varepsilon}{s}$ и по нему определяем искомое значение n.

Если для найденного объема выборки n выборочное значение s окажется больше предполагавшегося, то эксперимент должен быть продолжен.

Пример 2.3. Определить необходимый объем выборки для оценки среднего значения с предельной относительной ошибкой $\delta = 0{,}4$ при доверительной вероятности $\alpha = 0{,}95$, если предполагаемое значение коэффициента вариации равно $w = 1$.

Решение. Имеем $\frac{t_{0{,}95}}{\sqrt{n}} = \delta / w = 0{,}4$. Тогда из табл. 3 для $\alpha = 0{,}95$ непосредственно находим $n = 26$.►

2.7.2. Экспоненциальное распределение

Предположим, что в течение некоторого времени t_u испытывается n приборов и при испытаниях обнаруживается r отказов. Необходимо определить значения n и r, обеспечивающие оценку интенсивности отказов λ_0 с заданной относительной предельной ошибкой δ при доверительной вероятности α.

При испытаниях невосстанавливаемых приборов требуемый объем выборки равен

$$n = \frac{r}{\lambda_0 t_u a(r, \alpha)}.$$

Значения коэффициента $a(r, \alpha)$ приведены в табл. 3 [1] (см. табл. 5, статистические таблицы).

Значения r находятся из соотношения $b(r, \alpha) = \frac{1}{1+\delta}$, где $b(r, \alpha)$ –

коэффициент, зависящий от r и α (см. табл. 6, статистические таблицы). По заданным α и δ сначала определяем $b(r,\alpha)$, затем по заданному значению α и вычисленному $b(r,\alpha)$ из табл. 6 находим r. Далее, для найденного значения r и заданного α по табл. 5 определяем значение $a(r,\alpha)$, и по заданному t_u и λ_0 вычисляем требуемый объем выборки n. В случае испытаний восстанавливаемых приборов может быть получена оценка необходимого времени испытаний

$$t_u = \frac{r}{a(r,\alpha)} T_0,$$

где T_0 – ожидаемое время наработки на отказ.

Пример 2.4. Найти требуемый объем испытаний для оценки интенсивности отказов невосстанавливаемого прибора, если заданы время испытаний $t_u = 1000$ ч., предельная относительная ошибка $\delta = 0,2$, предполагаемое значение интенсивности отказов $\lambda_0 = 10^{-3}$, доверительная вероятность $\alpha = 0,95$.

Решение. Находим $\frac{1}{1+\delta} = \frac{1}{1+0,2} = 0,833$. Из табл. 6 для $b(r,\alpha) = 0,833$ и $\alpha = 0,95$ находим $r = 80$. Из табл. 5 для $r = 80$ и $\alpha = 0,95$ находим $a(r,\alpha) = 0,84$. Тогда искомый объем выборки $n = \frac{80}{10^{-3} \cdot 1000 \cdot 0,84} = 95$. ►

2.7.3 Биномиальное распределение

Предположим, что задано некоторое значение параметра биномиального распределения – p_0. Тогда, наименьший объем выборки, необходимый для того, чтобы подтвердить с вероятностью α, что $p \le p_0$ равен

$$n = \frac{\ln(1-\alpha)}{\ln(1-p_0)}.$$

Если среди n испытанных приборов не будет ни одного отказа, то с вероятностью α можно утверждать, что $p \le p_0$.

Пример 2.5. Найти объем выборки, позволяющий с достоверностью $\alpha = 0,90$ установить, что доля дефектных изделий в партии не превышает заданную величину $p_0 = 0,05$.

Решение. Имеем $n = \frac{\ln(1-\alpha)}{\ln(1-p_0)} = \frac{\ln 0,1}{\ln 0,95} = 45$. ►

Вопросы для самоконтроля

1. Что означают понятия точечные и интервальные оценки?
2. Понятие состоятельности, несмещенности и эффективности оценки.
3. Оценка неизвестных параметров закона распределения
4. Функция правдоподобия и оценка максимального правдоподобия.
5. Метод моментов. Оценки математического ожидания и дисперсии случайной величины. Их свойства.
6. Метод наименьших квадратов оценки параметров.
7. Оценки параметров нормального распределения.

8. Оценки параметров экспоненциального распределения.
9. Оценки параметров равномерного распределения.
10. Оценки параметров биномиального распределения.
11. Какой объем выборки необходим для оценки среднего μ при известной дисперсии для нормальной случайной величины?
12. Какой объем выборки необходим для оценки среднего μ нормальной случайной величины при известной дисперсии?
13. Какой объем выборки необходим для оценки среднего μ нормальной случайной величины при неизвестной дисперсии?
14. Какой объем выборки необходим для оценки параметра экспоненциального распределения?
15. Какой объем выборки необходим для оценки параметра биномиального распределения?

Тема 3. Интервальные оценки параметров распределений

3.1 Оценка параметров нормального распределения

Пусть случайная величина x имеет нормальное распределение $N(\mu,\sigma^2)$, где $\sigma>0$, $\mu\in R$ с плотностью $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x\in(-\infty,\infty)$.

Оценка μ при известной дисперсии σ^2

Интервальные оценки с доверительной вероятностью α имеют вид:

$$\mu^{н}(\alpha)=\bar{x}-u_\gamma\frac{\sigma}{\sqrt{n}},\quad \mu^{в}(\alpha)=\bar{x}+u_\gamma\frac{\sigma}{\sqrt{n}},$$

где u_γ – γ-квантиль стандартного нормального распределения; $\gamma=\dfrac{1+\alpha}{2}$ для двусторонней оценки, $\gamma=\alpha$ для односторонней оценки.

В табл. 1 (см. статистические таблицы) приведены квантили стандартного нормального распределения. На практике удобно использовать аппроксимацию квантилей стандартного нормального распределения вида

$$u_p=2{,}0637\left(\ln\frac{1}{1-p}-0{,}16\right)^{0{,}4274}-1{,}5774,\quad 0{,}5\le p\le 0{,}999$$

Оценка μ при неизвестной дисперсии

В этом случае интервальная оценка с доверительной вероятностью α имеет вид:

$$\mu^{н}(\alpha)=\bar{x}-t_\gamma\frac{s}{\sqrt{n}},\quad \mu^{в}(\alpha)=\bar{x}+t_\gamma\frac{s}{\sqrt{n}},$$

где t_γ – γ-квантиль распределения Стьюдента с $v=n-1$ степенями свободы;

$\gamma=\dfrac{1+\alpha}{2}$ для двусторонней оценки, $\gamma=\alpha$ для односторонней оценки.

В работе [1] на стр. 52 приведены критические точки распределения Стьюдента. Аппроксимации для расчетов квантилей имеют вид

$$t_p(v)=\begin{cases}u_p, & v>30;\\ u_p\left\{1+\dfrac{u_p^2+1}{4v}+\dfrac{5u_p^4+16u_p^2+3}{96v^2}+\dfrac{3u_p^6+19u_p^4+17u_p^2+15}{384v^3}\right\}, & v\le 30\end{cases}$$

Оценка дисперсии σ^2

Интервальные оценки при доверительной вероятностью α равны

$$(s^2)^{н}=\frac{1}{\chi^2_\gamma}\sum_{i=1}^{n}(x_i-\bar{x})^2,\quad (s^2)^{в}=\frac{1}{\chi^2_{\gamma'}}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

где χ^2_γ – γ-квантиль распределения χ^2 с $v=n-1$ степенями свободы (если

параметр μ известен, $v = n$); $\gamma' = \dfrac{1+\alpha}{2}$ для двусторонней оценки и $\gamma' = \alpha$ для односторонней оценки; $\gamma'' = \dfrac{1-\alpha}{2}$ для двусторонней оценки и $\gamma'' = 1-\alpha$ для односторонней оценки.

Для аппроксимации можно использовать следующую формулу

$$\chi^2_\gamma(v) = v\left[1 - \frac{2}{9v} + u_\gamma\sqrt{\frac{2}{9v}}\right]^3,$$

u_γ – γ-квантиль стандартного нормального распределения.

Для практического применения для уровней достоверности $\alpha = 0{,}9;\ 0{,}95;\ 0{,}99$ значения u_γ приведены в табл. 3.1

Таблица 3.1

α	Односторонние границы				Двусторонние границы			
	γ'	γ''	$u_{\gamma'}$	$u_{\gamma''}$	γ'	γ''	$u_{\gamma'}$	$u_{\gamma''}$
0,90	0,90	0,10	1,283	-1,283	0,950	0,050	1.645	-1.645
0,95	0,95	0,05	1,645	-1,645	0,975	0,025	1,960	-1,960
0,99	0,99	0,01	2,326	-2,326	0,995	0,005	2.576	- 2.576

Интервальная оценка σ может быть рассчитана также по формулам

$$s^{н} = \sqrt{\frac{n-1}{\chi^2_{\gamma'}}} \cdot s, \quad s^{в} = \sqrt{\frac{n-1}{\chi^2_{\gamma''}}} \cdot s,$$

где $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

В работе [1] приведена таблица значений $\sqrt{\dfrac{n-1}{\chi^2_{\gamma'}}}$ для различных $(n-1)$ от 2 до 100 и трех значений $\alpha = 0{,}9;\ 0{,}95;\ 0{,}99$. Для $\alpha = 0{,}95$ получена следующая эмпирическая формула:

– для нижней границы $\sqrt{\dfrac{1}{\chi^2_{0{,}975}}} = \dfrac{\sqrt{2n-1{,}74}}{1{,}96+\sqrt{2n-2}}$;

– для верхней границы $\sqrt{\dfrac{1}{\chi^2_{0{,}025}}} = \begin{cases} 11{,}54(n-3{,}61)^2 + 1{,}98; & 2 \le n \le 4; \\ \dfrac{\sqrt{2n-0{,}47}}{\sqrt{2n-0{,}8}-1{,}96}; & n \ge 5. \end{cases}$

Критические точки распределения χ^2 приведены в работе [2] на стр. 312.

3.2. Оценка параметров показательного распределения

Пусть $X \in \Pi_\lambda$, где Π_λ – показательный закон распределения с параметром λ с

плотностью распределения $f(x)=\lambda e^{-\lambda x},\ x\in[0,\infty)$.

Интервальная оценка параметра λ при доверительной вероятности α рассчитывается по формулам

$$\lambda^{н}=\frac{\chi^2_{\gamma'}}{2\sum\limits_{i=1}^{n}x_i},\quad \lambda^{в}=\frac{\chi^2_{\gamma''}}{2\sum\limits_{i=1}^{n}x_i},$$

где χ^2_γ – γ-квантиль распределения хи-квадрат с $v=2n$ степенями свободы;

$\gamma'=\frac{1+\alpha}{2},\quad \gamma''=\frac{1-\alpha}{2}$ для двусторонней оценки и $\gamma'=\alpha,\quad \gamma''=1-\alpha$ для односторонней оценки.

3.3. Оценка параметров биномиального распределения

Интервальные оценки параметра p с доверительной вероятностью α являются решениями уравнений Клоппера-Пирсона

$$\sum_{x=m}^{n}C_n^x p_н^x(1-p_н)^{(n-x)}=\frac{1-\alpha}{2},\quad \sum_{x=0}^{m}C_n^x p_в^x(1-p_в)^{(n-x)}=\frac{\alpha}{2}.$$

На практике широко используется аппроксимация нормальным распределением [1]. В этом случае нижняя $p_н$ и верхняя $p_в$ границы равны

$$p_н=\frac{x+\frac{1}{2}u_\gamma^2-u_\gamma\sqrt{\frac{x}{n}(n-x)+\frac{1}{4}u_\gamma^2}}{n+u_\gamma^2},\ p_в=\frac{x+\frac{1}{2}u_\gamma^2+u_\gamma\sqrt{\frac{x}{n}(n-x)+\frac{1}{4}u_\gamma^2}}{n+u_\gamma^2},$$

где u_γ – γ-квантиль стандартного нормального распределения; $\gamma=\frac{1+\alpha}{2}$ для двусторонней оценки, $\gamma=\alpha$ для односторонней оценки.

Эта аппроксимация рекомендуется при $x\ge 4$ и $n-x\ge 4$.

3.4. Примеры интервальных оценок

Пример 3.1. Требуется определить, какое количество книг N по некоторой теме должен иметь продавец, чтобы удовлетворить по возможности всех покупателей, если за четыре дня по этой теме было продано: 18, 12, 13, 15 книг.

Решение. На основании этих данных находим среднее и дисперсию

$\bar{x}=\frac{18+12+13+15}{4}=14{,}5;$

$s^2=\frac{1}{3}\left((18-14{,}4)^2+(12-14{,}4)^2+(13-14{,}4)^2+(15-14{,}4)^2\right)=7{,}0;$

$s=\sqrt{7}=2{,}65;\quad v=n-1=3.$

Примем доверительную вероятность $\alpha=0{,}95$. Тогда для односторонней критической области имеем $\gamma=\alpha$, $q=(1-\alpha)=0{,}05$, $2q=0{,}1$ и по таблице распределения Стьюдента получим $t_{0{,}1}=2{,}35$. Верхняя граница математического

ожидания равна $\mu^{в} = 14,5 + \frac{2,65}{\sqrt{4}} 2,35 = 17,61.$

Следовательно, максимальное возможное количество книг, которое необходимо иметь продавцу, $N = 18$.►

Пример 3.2. В канцелярии офиса работают три секретаря. Время подготовки одного документа каждым секретарем в среднем составляет 16,3; 15,5 и 17,2 мин. Требуется оценить ориентировочное время и возможное отклонение во времени оформления документа, сданного в канцелярию.

Решение. Рассчитываем выборочное среднее значение и выборочную дисперсию $\bar{x} = 16,33;\ \ s^2 = 0,72$.

Приняв доверительную вероятность $\alpha = 0,9$, получим

$\gamma' = \frac{1+\alpha}{2} = 0,95;\ \ \gamma'' = \frac{1-\alpha}{2} = 0,05$. При $v = 2$ по таблицам χ^2-распределения, находим: $\chi^2_{0,95} = 5,99;\ \ \chi^2_{0,05} = 0,103$. Тогда двусторонняя доверительная оценка дисперсии σ^2 равна:

$$(s^2)^{н} = \frac{1}{\chi^2_{\gamma'}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{2 \cdot 0,72}{5,99} = 0,24;\ (s^2)^{в} = \frac{1}{\chi^2_{\gamma''}} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{2 \cdot 0,72}{0,103} = 14,1.$$

В результате получим $0,24 \le \sigma^2 \le 14,1$. После извлечения квадратного корня $0,49 \le \sigma \le 3,61$.►

3.5. Интервальные оценки параметров при неизвестном законе распределения

3.5.1. Оценки для центра распределения

В качестве первичных (достаточно грубых) оценок центра группирования значений случайных величин при неизвестном законе распределения вероятностей могут быть использованы различные предельные неравенства. Рассмотрим неравенства Чебышева. Неравенство Чебышева имеет вид

$$P\left(|x - \mu| \ge k\sigma\right) < \frac{1}{k^2},$$

где μ и σ – соответственно среднее значение и стандартное отклонение.

Из неравенства Чебышева следует, что

$$x - \frac{\sigma}{\sqrt{1-\alpha}} \le \mu \le x + \frac{\sigma}{\sqrt{1-\alpha}},$$

где α – доверительная вероятность. Здесь предполагается, что σ известно.

Если вместо значения случайной величины x используется выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, то имеет место неравенство

$$\bar{x} - \frac{\sigma}{\sqrt{n(1-\alpha)}} \le \mu \le \bar{x} + \frac{\sigma}{\sqrt{n(1-\alpha)}}.$$

Если известно, что распределение симметрично относительно центра μ, то доверительный интервал равен

$$x-\frac{2\sigma}{3\sqrt{(1-\alpha)}}\le\mu\le x+\frac{2\sigma}{3\sqrt{(1-\alpha)}}\ \text{ или }\ \bar{x}-\frac{2\sigma}{3\sqrt{n(1-\alpha)}}\le\mu\le\bar{x}+\frac{2\sigma}{3\sqrt{n(1-\alpha)}}.$$

Отсюда следует, что только знание того факта, что распределение случайной величины симметрично, уже позволяет построить более узкий доверительный интервал для центра распределения.

Пример 3.3. Имеются результаты наблюдений над случайной величиной с неизвестным законом распределения вероятностей (известна только дисперсия $\sigma^2=75$):

$x_i: 1{,}2;\ 3{,}4;\ 6{,}1;\ 8{,}3;\ 12{,}1;\ 13{,}1;\ 14{,}8;\ 16{,}7;\ 21{,}9;\ 23{,}7;\ 24{,}5;\ 28{,}4.$

Найти доверительный интервал для центра распределения при $\alpha=0{,}95$.

Решение. Имеем $\bar{x}=14{,}516;\ \ 14{,}516-\frac{\sqrt{75}}{\sqrt{12(1-0{,}95)}}\le\mu\le14{,}516+\frac{\sqrt{75}}{\sqrt{12(1-0{,}95)}}$

или $3{,}336\le\mu\le25{,}696$.

Если бы располагали информацией о том, что распределение вероятностей случайной величины x симметрично, то имело бы место

$14{,}516-\frac{2\sqrt{75}}{3\sqrt{12(1-0{,}95)}}\le\mu\le14{,}516+\frac{2\sqrt{75}}{3\sqrt{12(1-0{,}95)}}$ или $7{,}062\le\mu\le21{,}96$, т.е. доверительный интервал длины $25{,}696-3{,}336=22{,}36$ уменьшился бы в 1,5 раза до $21{,}96-7{,}062=14{,}898$.►

3.5.2 Оценка рассеяния распределения

Некоторое представление о степени рассеяния непрерывного распределения дают его выборочные квантили. В общем случае доверительные интервал для p-квантили ограничен элементами упорядоченной по возрастанию выборки с номерами r и s, так как доверительная вероятность равна

$$\alpha=I_p(r,n-r+1)-I_p(s,n-s+1)=\sum_{i=r}^{s-1}C_n^i p^i(1-p)^{n-i}=P(x_r\le x_p\le x_s),$$

где $I_p(a,b)$ – функция бэта-распределения с параметрами a и b.

Если $s=n-r+1$ (случай симметричного интервала) то $\alpha=\sum_{i=r}^{n-r}C_n^i p^i(1-p)^{n-i}$.

Разность между $x_{0,75}$ и $x_{0,25}$, называемая интерквартильной широтой, является характеристикой степени рассеяния распределения относительно его центра.

Пример 3.4. В условиях примера 3 найти доверительный интервал для 25%-й квнтили распределения.

Решение. Предположим, что $r=3$ и $s=n-r+1=10$. Тогда доверительная вероятность того, что в интервале $[x_3-x_{10}]$ находится 25%-я квантиль ($p=0{,}25$), равна

$$\alpha=\sum_{i=3}^{9}C_{12}^i 0{,}25^i(1-0{,}25)^{12-i}=0{,}552.\blacktriangleright$$

Вопросы для самоконтроля

1. Оценки среднего нормального распределения при известной дисперсии.
2. Оценки среднего нормального распределения при неизвестной дисперсии.
3. Оценка дисперсии нормального распределения.
4. Интервальная оценка параметров экспоненциального распределения.
5. Интервальная оценка параметров биномиального распределения.
6. Оценки для центра распределения при неизвестном законе распределения
7. Оценка рассеяния распределения при неизвестном законе распределения

Тема 4. Методы анализа законов распределения вероятностей случайных величин

4.1. Общие понятия

Для практического применения методов теории вероятностей и математической статистики знание закона распределения вероятностей чрезвычайно важно. По существу, сама изучаемая случайная величина для исследователя представлена только законом распределения вероятностей реализации ее значений.

Зная закон распределения вероятностей наблюдаемой случайной величины, исследователь или инженер в состоянии решать многие практические задачи, связанные с планированием производства, обеспечением качества продукции, оценкой эффективности и стабильности производства.

Попытка, применить методы анализа результатов наблюдений, разработанные для конкретных законов распределения вероятностей, в условиях, когда реальное распределение отличается от гипотетического, является самой распространенной на практике ошибкой, приводящей к неверным выводам и, в конечном итоге, к существенным материальным потерям и затратам времени.

Именно поэтому любая обработка результатов наблюдений должна неизменно начинаться с ответа на главный вопрос: каково распределение вероятностей обрабатываемого ряда случайных величин? На практике эта проблема обычно формулируется следующим образом. Выдвигается гипотеза — «наблюдаемое распределение случайных величин описывается некоторым конкретным законом (нормальным, экспоненциальным, Вейбулла, ...)». Задача первичного исследования принять или отклонить выдвинутую гипотезу.

Если ни одна из гипотез, связанных с формой закона распределения вероятностей не принимается, то может быть сформулирована более мягкая гипотеза — например, «наблюдаемое распределение симметрично относительно какой-то точки». Даже установление только этого факта дает в руки исследователя более эффективные методы анализа наблюдений, чем полное незнание закона распределения вероятностей. И. наконец, если исследователь не получил достаточных оснований для выбора вида распределения, то возникает задача подбора формы распределения непосредственно по экспериментальным данным. При этом распределение вероятностей должно быть подобрано так, чтобы оно удовлетворительно описывало имеющийся экспериментальный материал.

Мы встречаемся здесь с понятием статистической гипотезы. *Статистической гипотезой* называется предположение, выдвигаемое относительно особенностей распределения вероятностей случайной величины, которое проверяется по результатам наблюдений над ней.

Проверка любой статистической гипотезы сводится к следующему. По выборочным значениям случайной величины подсчитывается некоторая величина – *статистический критерий (статистика критерия).* При допущении, что распределение вероятностей используемой статистики критерия в условиях справедливости проверяемой гипотезы известно, определяется вероятность появления вычисленного значения статистики. На основе так называемого принципа значимости устанавливается *уровень значимости* — наибольшее значение вероятности, несовместимое с признанием случайности экспериментально вычисленного значения статистики критерия. Событие называется *значимым,* (а не случайным), если теоретическая вероятность его случайного появления меньше, чем принятый уровень значимости. Уровнем значимости определяется критическое значение статистики критерия. Как правило, если значение статистики критерия,

вычисленное по экспериментальным данным, больше критического, то гипотеза отклоняется на выбранном уровне значимости. В противном случае она признается не противоречащей результатам наблюдений. Дополнение до единицы уровня значимости называется *уровнем достоверности* (*достоверностью).*

Поскольку статистика критерия для проверки гипотезы вычисляется по выборочным реализациям случайной величины, то и сама она является случайной величиной. Поэтому суждения по гипотезе на основе статистики критерия могут носить только вероятностный характер. При этом различают *ошибки, первого рода,* заключающиеся в отклонении верной гипотезы, и *ошибки второго рода,* заключающиеся в принятии ложной гипотезы. Вероятность ошибки первого рода совпадает (по крайней мере не выше) с уровнем значимости и обозначается в литературе через α. Ошибка второго рода обозначается через β. Эффективность статистического критерия проверки гипотезы оценивается его мощностью $1-\beta$, равной вероятности отклонения ложной гипотезы.

Выбор значений α и β определяется условиями эксперимента и требованиями, предъявляемыми к достоверности суждения по проверяемой гипотезе. Обычно на практике используются значения α, β, равные 0,1; 0,05; 0,01.

Проверяемая гипотеза называется нулевой и обозначается символом H_0. Например, запись $H_0 : F(x) = G(x)$ означает, что проверяется нулевая гипотеза о совпадении функций распределения $F(x)$ и $G(x)$.

В классификации статистических критериев проверки гипотез о законе распределения вероятностей принята определенная терминология. Такие критерии подразделяются на два класса — общие критерии согласия и специальные критерии согласия. *Общие критерии согласия* применимы к самой общей формулировке гипотезы, как гипотезы о согласии наблюдаемых результатов с любым априорно предполагаемым распределением вероятностей. *Специальные критерии согласия* предполагают специальные нулевые гипотезы, формулирующие согласие с определенной формой распределения вероятностей – нормальной, экспоненциальной, Вейбулла и т.д. Такие критерии носят соответствующие названия – критерии нормальности, критерии экспоненциальности и т.п.

4.2. Общие критерии согласия

Нулевая гипотеза при применении общих критериев согласия записывается в форме

$$H_0 : \; F_n(x) = F(x),$$

где $F_n(x)$ – эмпирическая функция распределения вероятностей; $F(x)$ – гипотетическая функция распределения вероятностей.

Все известные общие критерии согласия можно разбить на три основные группы:

- критерии, основанные на изучении разницы между теоретической плотностью распределения и эмпирической гистограммой:
- критерии, основанные на расстоянии между теоретической и эмпирической функциями распределения вероятностей;
- корреляционно-регрессионные критерии, основанные на изучении корреляционных и регрессионных связей между эмпирическими и теоретическими порядковыми статистиками.

4.2.1 Критерии, основанные на сравнении теоретической плотности распределения и эмпирической гистограммы

Критерий χ^2 (Пирсона) для простой гипотезы

Пусть дана выборка $\{x_1, x_2, \ldots, x_n\}$ из генеральной совокупности F. Проверяется гипотеза $H_0 : F_n(x) = F(x)$ против альтернативы $H_1 : F_n(x) \neq F(x)$.

Представим выборку в виде группированного ряда, разбив предполагаемую область значений случайной величины на m интервалов. Пусть n_i - число элементов выборки попавших в i-ый интервал, а $p_i = F(x_{i+1}) - F(x_i)$ - теоретическая вероятность попадания случайной величины в i-й интервал при условии истинности H_0. Составим статистику $\rho(x) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$, которая характеризует сумму квадратов отклонения наблюдаемых значений n_i от ожидаемых np_i по всем интервалам группирования.

Теорема Пирсона. Если H_0 верна, то при фиксированном m и $n \to \infty$

$$\rho(x) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} \Rightarrow \chi^2_\alpha(m-1). \tag{4.1}$$

Таким образом, статистику $\rho(x)$ можно использовать в качестве статистики критерия согласия для проверки гипотезы о виде закона распределения, который будет иметь вид:

$$F_n(x) = \begin{cases} H_0, & \rho(x) < \chi^2_\alpha \\ H_1, & \rho(x) \geq \chi^2_\alpha \end{cases}, \ \rho(x) = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{4.2}$$

где χ^2_α-квантиль распределения $\chi^2(m-1)$ с $(m-1)$ степенями свободы.

Данный критерий называется *критерием χ^2 или критерием согласия Пирсона.* Дисперсия статистики $\rho(x)$ равна $D(\rho) = 2(m-1) + \frac{1}{n}\left(\sum_{i=1}^{m} \frac{1}{p_i} - m^2 - 2m + 2\right)$. Если $\sum_{i=1}^{m} \frac{1}{p_i} \ll n$ и $m \ll n$, то $D(\rho) = 2(m-1)$, т.е. совпадает с дисперсией случайной величины, имеющей χ^2 - распределение.

На мощность статистического критерия χ^2 сильное влияние оказывает число интервалов разбиения гистограммы m и порядок её разбиения (т.е. выбор длин интервалов внутри диапазона изменения значений случайной величины). На практике принято считать, что статистику χ^2 можно использовать при $np_i \geq 5$.

Показано, что такое приближение допустимо и тогда, когда не более, чем в 20% интервалов имеет место $1 \le np_i \le 5$. При $n \ge 200$ рекомендуется выбирать m из условия $m = 4(n-1)^{2/5}$, но не превышающее $m = n/5$.
При $n < 200$ значение m можно выбирать из условия $m = 1 + 3{,}32 \cdot \lg n \approx 1 + 4 \cdot \lg n$. Считается, что оптимальное значение $m = 10$.

Правило проверки гипотезы просто: если $\rho(x) > \chi^2_\alpha(v)$, то на уровне значимости α, т.е. с достоверностью $(1-\alpha)$ гипотеза H_0 отвергается.

Критерий χ^2 (Пирсона) для сложной гипотезы

Пусть $\{x_1, x_2, \ldots, x_n\}$ выборка из генеральной совокупности F. Проверяется сложная гипотеза $H_0 : F_n(x) = F_\theta(x)$, где θ - неизвестный параметр распределения F (или вектор параметров), против альтернативы $H_1 : F_n(x) \ne F_\theta(x)$.

Пусть выборка по прежнему представлена в виде группированного ряда и n_i - число элементов выборки попавших в i-ый интервал, $i \in \{1, 2, \ldots, m\}$. Статистику (1) мы не можем в этом случае использовать для построения критерия Пирсона, так как не можем вычислить теоретические значения вероятностей p_i, которые зависят от неизвестного параметра θ. Пусть θ^* - оценка параметра θ, а $p_i^*(\theta^*)$ - соответствующие ей оценки вероятностей p_i. Составим статистику

$$\rho(x) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}.$$

Теорема Пирсона. Если H_0 верна, и l - число компонент вектора θ (число неизвестных параметров распределения), то при фиксированном m и $n \to \infty$

$$\rho(x) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*} \Rightarrow \chi^2_\alpha(m-l-1). \qquad (4.3)$$

Таким образом, критерий Пирсона для параметрической гипотезы будет иметь вид:

$$F_n(x) = \begin{cases} H_0, & \rho(x) \le \chi^2_\alpha \\ H_1, & \rho(x) > \chi^2_\alpha \end{cases}, \qquad \rho(x) = \sum_{i=1}^{m} \frac{\left(n_i - np_i^*\right)^2}{np_i^*}, \qquad (4.4)$$

где $\chi^2_\alpha(m-l-1)$ - квантиль распределения χ^2 с $m-l-1$ степенями свободы.

Замечание. Вообще говоря, оценки θ^*, используемые для построения статистики критерия хи-квадрат, должны быть определены из условия минимума статистики $\rho(x)$.

Поэтому желательно уточнить оценки θ^*, найденные другим способом (методом максимального правдоподобия или методом моментов) путем минимизации $\rho(x)$.

4.2.2 Критерии, основанные на сравнении теоретической и эмпирической функций распределения вероятностей

Пусть дана выборка $x_1 \le x_2 \le .... \le x_n$. Обозначим через $F_n(x)$ эмпирическую функцию распределения вероятностей. а через $F(x)$ — теоретическую функцию распределения ($x_i = F^{-1}\left(\frac{i-0,5}{n}\right)$).

Расстояние между эмпирической и теоретической функциями распределения вероятностей является весьма эффективной статистикой для проверки гипотез о виде закона распределения вероятностей случайной величины.

Среди известных критериев согласия такого типа отметим серию критериев, использующих различные варианты анализа расстояния между $F_n(x)$ и $F(x)$: критерий Джини, критерий Крамера-фон Мизеса, критерий Колмогорова-Смирнова, критерий Смирнова-Крамера-фон Мизеса и др.

Критерий Колмогорова-Смирнова

Пусть $F_n(x)$ – эмпирическая функция распределения случайной величины x, представленной выборкой $x_1 \le x_2 \le .... \le x_n$:

$$F_n(x) = \begin{cases} 0, & x < x_1; \\ \dfrac{i}{n}, & x_i \le x \le x_{i+1},\ 1 \le i \le n-1; \\ 1, & x \ge x_n. \end{cases}$$

Для проверки нулевой гипотезы $H_0 : F_n(x) = F(x)$, где $F(x)$ – полностью определенная (с точностью до параметров) теоретическая функция распределения, рассматривается расстояние между эмпирической и теоретической функциями распределения

$$D_n = \sup_{|x|<\infty} \left|F_n(x) - F(x)\right|;\ D_n^+ = \sup_{|x|<\infty} \left(F_n(x) - F(x)\right);\ D_n^- = -\inf_{|x|<\infty} \left(F_n(x) - F(x)\right).$$

Здесь $\sup$, $\inf$ – точные верхняя и нижняя границы соответствующих разностей.

Для практического применения используются формулы

$$D_n^+ = \max_{1\le i\le n}\left(\frac{i}{n} - F(x_i)\right);\ D_n^- = \max_{1\le i\le n}\left(F(x_i) - \frac{i-1}{n}\right);\ D_n = \max\left(D_n^+, D_n^-\right).$$

Критические значения разностей рассчитываются по приближенным формулам

$$D_n(\alpha) = \left\{\frac{1}{2}\ln\frac{2}{1-\alpha}\right\}^{-1};\ D_n^{+(-)}(\alpha) = \left\{\frac{1}{2}\ln\frac{1}{1-\alpha}\right\}^{-1}$$

Если $D_n > D_n(\alpha)$, то гипотеза согласия H_0 отклоняется на уровне значимости α.

При $n \ge 20$ полезна аппроксимация

$$\chi^2 = \frac{1}{9n}\left(6nD_n^{+(-)} + 1\right)^2,$$

распределение которой описывается распределением χ^2 с $v = 2$ степенями свободы.

При $n \geq 10$ необходимо использовать более точное приближение

$$D_n^{+(-)}(\alpha) = \left\{\frac{1}{2n}\left(y - \frac{2y^2 - 4y - 1}{18n}\right)\right\}^{1/2} - \frac{1}{6n} \approx \left(\frac{y}{2n}\right)^{1/2} - \frac{1}{6n},$$

где $y = -\ln\alpha$ для $D_n^{+(-)}(\alpha)$ и $y = -\ln(\alpha/2)$ для D_n, при $0{,}01 \leq \alpha \leq 0{,}2$ и $0{,}005 \leq \alpha$.

Стефенс предложил следующие преобразования статистик $D_n^{+(-)}$, D_n

$\widetilde{D}_n = D_n\left(\sqrt{n} + 0{,}275 - \frac{0{,}04}{\sqrt{n}}\right)$ — для нижней процентной точки;

$\widetilde{D}_n = D_n\left(\sqrt{n} + 0{,}12 + \frac{0{,}11}{\sqrt{n}}\right)$ — для верхней процентной точки;

$\widetilde{D}_n^{+(-)} = D_n\left(\sqrt{n} + 0{,}12 + \frac{0{,}11}{\sqrt{n}}\right).$

Критические значения статистик Стефенса приведены в табл. 4.1.

Таблица 4.1. Процентные точки статистик $\widetilde{D}_n$ и $\widetilde{D}_n^{+(-)}$

α	0,150	0,100	0,050	0,025	0,010
$\widetilde{D}_n$	0,973	1,073	1,224	1,358	1,518
$\widetilde{D}_n^{+(-)}$	1,138	1,224	1,358	1,480	1,628

Критерий Колмогорова-Смирнова применяется при $n \geq 50$.

Критерий Крамера-фон Мизеса

Статистика критерия имеет вид

$$w^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i) - \frac{2i-1}{2n}\right\}^2,$$

где $F(x)$ – теоретическая функция распределения.

Необходимо помнить, что теоретическая функция распределения должна быть известна с точностью до параметров. Распространенная ошибка — использование в качестве $F(x)$ функции распределения с параметрами, оцениваемыми по выборке приводит к уменьшению величины критического значения статистики, т.е. к увеличению количества ошибок второго рода.

При объеме выборки $n > 40$ можно использовать приведенные в табл. 5.2 квантили распределения w^2, которые следуют из его предельного распределения (α – уровень значимости, принятый для проверки H_0).

Таблица 4.2. Квантили распределения w^2

α	0,900	0,950	0,990	0,995	0,999
$w^2(\alpha)$	0,3473	0,4614	0,7435	0,8694	1,1679

При $n < 40$ можно использовать аппроксимацию

$$(w^2)' = \left(w^2 - \frac{0,4}{n} + \frac{0,6}{n^2} \right) \cdot \left(1 + \frac{1}{n} \right).$$

4.3 Критерии нормальности распределения

Нормальный закон распределения вероятностей получил наибольшее распространение в практических задачах обработки экспериментальных данных. Большинство прикладных методов математической статистики исходит из предположения нормальности распределения вероятностей изучаемых случайных величин.

Широкое распространение этого распределения вызвало необходимость разработки специальных критериев согласия эмпирических распределений с нормальным законом. Рассмотрим два критерия.

4.3.1 Модифицированный критерий χ^2

Пусть дана выборка $x_1, x_2, ..., x_n$ данных из распределения $F(x)$. После оценки параметров $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ распределения совокупность выборочных данных разбивается на m равновероятных интервалов ($p_i = \frac{1}{m} = const$) и статистика критерия подсчитывается по формуле

$$\chi^2 = \frac{m}{n}\sum_{i=1}^{m} n_i^2 - n,$$

где n – объем выборки; n_i – количество членов выборки, попавшие в i-й интервал.

Границы интервалов определяются как

$$a_i = \bar{x} + c_i s \ \ (i = 0,1,...,m).$$

Значения коэффициентов c_i приведены в табл. 5.3. Следует отметить, $c_0 = -\infty$ и $c_m = \infty$. Так как c_i симметричны относительно нуля, то недостающие значения c_i можно найти из соотношений

$c_{\frac{1}{2}(m-1)+i} = -c_{\frac{1}{2}(m-1)-i}$, $(i = 1,..., \frac{m-1}{2})$ – для нечетных m;

$c_{\frac{1}{2}m+i} = -c_{\frac{1}{2}m-i}$, $(i = 1,..., \frac{m-2}{2})$ – для четных m.

Таблица 4.3. Значения коэффициентов c_i модифицированного

χ^2-критерия нормальности для $m \in [3;15]$

m	c_1	c_2	c_3	c_4	c_5	c_6	c_7
3	-0,4307						
4	-0,6745	0					
5	-0,8416	-0,2533					
6	-0,9074	-0,4307	0				
7	-1,0676	-0,5659	-0,1800				
8	-1,1503	-0,6745	-0,3180	0			
9	-1, 2206	-0,7647	-0,4307	-0,1397			
10	-1,2816	-0,8416	-0,5244	-0,2533	0		
11	-1,3352	-0,9085	-0,6040	-0,3488	-0,1142		
12	-1,3830	-0,9674	-0,6745	-0,4307	-0,2194	0	
13	-1,4201	-1,0201	-0,7303	-0,5024	-0,2934	-0,0966	
14	-1,4652	-1,0676	-0,7916	-0,5660	-0,3661	-0,1800	0
15	-1,5011	-1,1108	-0,8416	-0,6229	-0,4307	-0,2533	-0,0837

Если $\chi^2 > d_m(\alpha)$, где $d_m(\alpha)$ – критическое значение статистики критерия на уровне значимости α, то гипотеза нормальности отклоняется. Критические значения $d_m(\alpha)$ приведены в табл. 5.4

Таблица 4.4. Критические значения $d_m(\alpha)$ модифицированного χ^2 - критерия нормальности

m	α			m	α		
	0,10	0,05	0,01		0,10	0,05	0,01
3	2,371	3,248	5,418	10	12, 384	14,438	18,852
4	3,928	5,107	7,917	11	13,694	15,843	20,431
5	5,442	6,844	10,075	12	14,988	17,226	21,977
6	6,905	8,479	12,021	13	16,267	19,589	23,495
7	8,322	10,038	13,837	14	17,535	19,937	24,990
8	9,703	11,543	15,567	15	18,792	21,270	26,464
9	11,055	13,007	17,234				

4.3.2 Критерий типа Колмогорова – Смирнова

Рассмотрим применение критерия Колмогорова-Смирнова (см. раздел 5.2.2) для проверки нормальности распределения в ситуации, когда оба его параметра оцениваются по выборке. Алгоритм проверки нулевой гипотезы H_0 для этого случая сохраняется. При этом используется модифицированная статистика

$$D_n^{и} = D_n\left(\sqrt{n} - 0{,}01 + \frac{0{,}85}{\sqrt{n}}\right).$$

Критические значения $D_n^{и}(\alpha)$ (α – уровень значимости) приведены в табл. 4.5

Таблица 4.5. Критические значения статистики Колмогорова – Смирнова, модифицированной для проверки нормальности распределения

α	0,15	0,10	0,05	0,03	0,01
$D_n^{и}(\alpha)$	0,775	0,819	0,895	0,955	1,035

Применим критерий согласия w^2 (см. раздел 5.2.2) для задачи проверки гипотезы нормальности распределения вероятностей случайных величин. Алгоритм вычисления статистики критерия в этом случае не меняется — меняются только критические значения статистики проверки гипотезы. Для различных ситуаций. когда параметры гипотетического распределения оцениваются непосредственно по самой выборке, критические значения статистики w^2 приведены в табл. 4.6.

Таблица 4.6. Критические значения статистики w^2 *для* проверки нормальности распределения ($1-\alpha$ – уровень значимости)

Исходные условия	α				
	0,90	0,95	0,99	0,995	0,999
Параметры (μ и σ) известны заранее	0,3473	0,4614	0,7435	0,8694	1,1679
Параметр σ известен, а параметр μ оценивается по выборке $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	0,1344	0,1653	0,2380	0,2698	0,3443
Параметр μ известен, а параметр σ оценивается по выборке $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$	0,2370	0,4418	0,7245	0,8506	1,1490
Параметры (μ и σ) оцениваются по выборке	0,1035	0,1260	0,1788	0,2018	0,2559

4.4 Критерий проверки экспоненциальности распределения

4.4.1 Критерии типа Колмогорова –Смирнова

Предположим, имеет место гипотетически закон распределения вероятностей $F(x)=1-\exp\left(-\frac{(x-\mu)}{v}\right)$, где μ и v – неизвестные параметры, оценки которых по выборке могут быть найдены по формулам (отметим, что выборка упорядочена, т.е. $x_1 \le x_2 \le .... \le x_n$)

$$\hat{v}=\frac{n(\bar{x}-x_1)}{n-1};\ \hat{\mu}=x_1-\frac{\hat{v}}{n}.$$

Обозначим $z_i=\frac{x_i-\hat{\mu}}{\hat{v}}$ и перейдем к нормированному экспоненциальному распределению $F(z_i)=1-\exp(-z_i)$, для которого можно применить следующие критерии согласия

- Критерий Колмогорова-Смирнова

$$D_n^+=\max\left[\frac{1}{n}-F(z_i)\right];\ D_n^-=\max\left[F(z_i)_i-\frac{i-1}{n}\right];\ D_n \max\left(D_n^+, D_n^-\right);$$

- Критерий Смирнова-Крамера-фон Мизеса $w^2=\sum_{i=1}^{n}\left(F(z_i)-\frac{2i-1}{2n}\right)^2+\frac{1}{12n}$.

Для случая проверки экспоненциальности распределения с неизвестными параметрами критические значения для различных уровней значимости приведены в табл. 4.7.

Таблица 4.7. Критические значения статистик критериев согласия типа Колмогорова-Смирнова для проверки экспоненциальности распределения с неизвестными параметрами

n	Уровень значимости α (верхние процентные точки)					
	0,25	0,15	0,10	0,05	0,025	0,01
	Статистика $\sqrt{n}D_n$					
5	0,683	0,749	0,793	0,865	0,921	0,992
10	0,753	0,833	0,889	0,977	1,048	1,119
15	0,771	0,865	0,912	1,002	1,079	1,163
20	0,786	0,872	0,927	1,021	1,099	1,198
25	0,792	0,878	0,936	1,033	1,115	1,215
50	0,813	0,879	0,960	1,061	1,149	1,257
100	0,824	0,911	0,972	1,072	1,171	1,278
∞	0,840	0,927	0,995	1,094	1,184	1,298
	Статистика w^2					

5	0,083	0,102	0,117	0,141	0,166	0,197
10	0,097	0,122	0,142	0,176	0,211	0,259
15	0,103	0,130	0,151	0,188	0,229	0,281
20	0,106	0,133	0,157	0,195	0,237	0,293
25	0,107	0,135	0,160	0,199	0,247	0,301
50	0,111	0,141	0,166	0,209	0,256	0,319
100	0,113	0,144	0,170	0,215	0,263	0,328
∞	0,116	0,148	0,175	0,222	0,271	0,338

4.4.2 Критерий Фишера

Критерий Фишера имеет вид $F = \dfrac{\sum_{i=1}^{n} x_i}{(n-1)x_1}$.

Эта статистика имеет F-распределение с $v_1 = 2n-2$ и $v_2 = 2$ степенями свободы.

Если $\dfrac{\sum_{i=1}^{n} x_i}{(n-1)x_1} > F_\alpha(2n-2,2)$, то нулевая гипотеза отклоняется. Здесь $F_\alpha(v_1, v_2)$ – α - критическое значение F-статистики с v_1 и v_2 степенями свободы.

Критические значения F-статистики при уровне значимости $\alpha = 0,05$ приведены в таблице (см. табл. 7 статистические таблицы).

4.5 Критерии согласия для равномерного распределения

Критерии типа Колмогорова-Смирнова

Приведем модифицированные формы критериев Колмогорова-Смирнова для задачи проверки равномерности распределения порядковой статистики $U_1 < U_2 < ... < U_n$.

$$D^+ = \max_i \left(U_i - \frac{i}{n+1} \right);\ D^- = \max_i \left(\frac{i}{n+1} - U_i \right);\ D = \max\left(D^+, D^- \right).$$

Распределения указанных статистик быстро сходятся к предельному, если использовать их модификации:

$$\widetilde{D}^+ = \left(D^+ + \frac{0,4}{n} \right) \cdot \left(\sqrt{n} + 0,2 + \frac{0,68}{\sqrt{n}} \right);\ \widetilde{D}^- = \left(D^- + \frac{0,4}{n} \right) \cdot \left(\sqrt{n} + 0,2 + \frac{0,68}{\sqrt{n}} \right);$$

$$\widetilde{D} = \left(D + \frac{0,4}{n} \right) \cdot \left(\sqrt{n} + 0,2 + \frac{0,68}{\sqrt{n}} \right).$$

Критические значения для модифицированных статистик приведены в табл. 4.8.

Таблица 4.8. Критические значения $\widetilde{D}^+$, $\widetilde{D}^-$, $\widetilde{D}$ критериев равномерности

n	Уровень значимости α				
	0,01	0,025	0,05	0,1	0,15

$\widetilde{D}^{+}$	1,518	1,358	1,224	1,073	0,973
$\widetilde{D}^{-}$	1,518	1,358	1,224	1,073	0,973
$\widetilde{D}$	1,628	1,480	1,358	1,224	1,138

4.6 Критерий симметрии

Если отсутствуют предпосылки для проверки согласия эмпирического распределения с каким-либо теоретическим, то выявление даже самых общих свойств эмпирического распределения дает некоторую информацию для выбора приемов и методов обработки экспериментального материала.

Одним из таких практически важных свойств распределения является его симметричность относительно центра группирования значений случайной величины.

Рассмотрим один из критериев проверки на симметричность – критерий Кенуя. Для этого нам потребуется ввести понятие *порядковой статистики*.

Как только любому члену наблюдаемого выборочного ряда ставится в соответствие его номер в упорядоченном по возрастанию ряду выборочных значений — этот член выборки становится *порядковой статистикой*. Для полного координирования порядковой статистики необходимо указать объем выборки и номер статистики. Для того, чтобы отличить просто член выборки x_i от порядковой статистики, будем применять для ее обозначения символ $x_{(i,n)}$ – т.е. i *-я* порядковая статистика в выборке объема n. Или $x_{(i)}$, если по умолчанию предполагается известным объем выборки.

Статистика критерия строится следующим образом. Выборочные значения упорядочиваются по возрастанию: $x_1 \le x_2 \le ... \le x_n$, определяются порядковые статистики с номерами $\frac{n}{16}, \frac{15n}{16}, \frac{n}{2}$, т.е. $x_{\left(\frac{n}{16}\right)}, x_{\left(\frac{15n}{16}\right)}, x_{\left(\frac{n}{2}\right)}$. По ним вычисляется мера асимметрии $A = x_{\left(\frac{15n}{16}\right)} - 2x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{16}\right)}$ и статистика критерия

$$\widetilde{A} = A\frac{\sqrt{n}}{3s},$$

где $s = \frac{1}{3}\left(x_{\left(\frac{15n}{2}\right)} - x_{\left(\frac{n}{2}\right)} \right)$, или $A = \sqrt{n}\,\frac{x_{\left(\frac{15n}{16}\right)} - 2x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{16}\right)}}{x_{\left(\frac{15n}{2}\right)} - x_{\left(\frac{n}{2}\right)}}$.

При $n \ge 50$ распределение статистики может быть аппроксимировано стандартным нормальным распределением. Тогда гипотеза симметрии отклоняется с достоверностью α, если $\left|\widetilde{A}\right| > u_{\frac{1+\alpha}{2}}$, где u_γ – γ-квантиль стандартного нормального распределения.

Пример 4. 7. В выборке объёма $n = 64$ имеются порядковые статистики $x_{\left(\frac{n}{16}\right)} = x_{(4)} = 17;\ x_{\left(\frac{n}{2}\right)} = x_{(32)} = 44;\ x_{\left(\frac{15n}{16}\right)} = x_{(60)} = 127$. Проверить гипотезу

симметричности распределения случайной величины критерием Кенуя при доверительной вероятности $\alpha = 0,95$.

Решение. Находим $A = \sqrt{64}\,\dfrac{x_{(60)} - 2x_{(32)} + x_{(4)}}{x_{(60)} - x_{(4)}} = 4,07$.

Для $\alpha = 0,95$ из таблицы квантилей нормального распределения (см. статистические таблицы) $u_{0,975} = 1,96$. Так как $\left|\widetilde{A}\right| = 4,07 > 1,96$, то гипотеза симметрии отклоняется.►

Вопросы для самоконтроля

1. Задачи статистической проверки гипотез. Понятие гипотезы.
2. Уровень значимости, уровень достоверности.
3. Критерии, основанные на сравнении теоретической плотности распределения и эмпирической гистограммы. Критерий χ^2 (Пирсона) для простой гипотезы.
4. Критерий χ^2 (Пирсона) для сложной гипотезы.
5. Критерии, основанные на сравнении теоретической и эмпирической функций распределения вероятностей. *Критерий Колмогорова-Смирнова.*
6. *Критерий Крамера-фон Мизеса.*
7. Критерии нормальности распределения. Модифицированный критерий χ^2.
8. Критерий типа Колмогорова – Смирнова проверки нормальности распределения.
9. Критерий проверки экспоненциальности распределения. Критерии типа Колмогорова –Смирнова.
10. Критерий Фишера проверки экспоненциальности распределения
11. Критерии согласия для равномерного распределения. Критерии типа Колмогорова-Смирнова.
12. Критерий симметрии.

Тема 5. Проверка гипотез о значениях параметров распределений

5.1 Общие сведения

На бытовом языке слово *гипотеза* означает *предположение.* В математической статистике это предположение относится к распределению вероятностей на выборочном пространстве. Предположения могут быть как о конкретном законе распределения, так и о значениях его параметров. Таким образом, статистическая гипотеза — это предположение о распределении вероятностей, которое мы хотим проверить по имеющимся данным.

На практике часто встречаются задачи сравнения двух выборочных совокупностей. Например, нас может интересовать сравнение двух методов обработки, т.е. двух различающихся действий, направленных к одной цели: двух методик обучения, двух технологий управления, двух способов получения информации и т.д. Для формирования статистической гипотезы необходимо черты, присущие конкретной проблеме, выразить в терминах, относящихся к распределению вероятностей. Этот процесс является творческим, и его невозможно формализовать. При этом следует помнить, что для ряда типовых случаев математическая теория разработана очень подробно, и стараться по возможности свести дело к одной из типовых статистических задач.

Статистические гипотезы служат инструментом проверки выдвигаемых теоретических предположений. Предположения могут быть сделаны относительно параметров статистического распределения вероятностей (например, в случае нормального закона — относительно математического ожидания m или дисперсии σ^2). Тогда гипотезу называют *параметрической.* Предположения могут быть сделаны относительно самого распределения случайной величины (подчинение закону Бернулли, Пуассона, нормальному и т.д.) в этом случае проверяемую гипотезу *называют непараметрической.*

Процедура обоснованного сопоставления высказанного предположения с имеющимися в нашем распоряжении выборочными данными $x_1, x_2, ..., x_n$, сопровождающаяся количественной оценкой степени достоверности полученного вывода, осуществляется с помощью *статистического критерия* и называется *статистической проверкой гипотез.*

При проверке гипотез используется понятие нулевой (прямой) и альтернативной (обратной) гипотез. *Прямая гипотеза* обозначается как H_0 и формулируется в предположении отсутствия различий между выборочной и генеральной совокупностями. Исследовав выборку, мы принимаем решение: согласуется она с этой гипотезой или нет. *Альтернативная гипотеза* H_1, является конкурирующей по отношению к нулевой и принимается после того, как отвергается нулевая. Рассмотрим гипотезу относительно математического ожидания нормально распределенной случайной величины.

Нулевая гипотеза:

$$H_0 : m = m_0 .$$

Альтернативная гипотеза:

$$H_1 : m \neq m_0$$

или, что то же самое,

$$(m < m_0) \cup (m > m_0) .$$

Простая гипотеза содержит только одно предположение относительно оцениваемого параметра, например $H_0 : a = 4$. *Сложная гипотеза* состоит из конечного или бесконечного числа простых гипотез. Например, $H_0 : a > 4$ означает, что гипотеза состоит из бесконечного набора вида:

$$H_0 : a = a_i ,$$

где $a_i = \{4,5;\ 5,8;\ ....;\ \infty\}$.

Доказать справедливость простой гипотезы значительно легче, поэтому при формулировании исходного предположения стараются в качестве нулевой использовать простую гипотезу. При этом может быть применим принцип доказательства «от противного».

Пример 5.1

Пусть мы хотим доказать, что студенты пятого курса в среднем выше ростом, чем студенты первого.

Решение. В качестве нулевой гипотезы выдвинем предположение о равенстве их среднего роста:

$$H_0 : \overline{L}_5 = \overline{L}_1$$

где $\overline{L}_5$ — средний рост студентов 5-го курса;

$\overline{L}_1$ —- средний рост студентов 1-го курса.

В качестве альтернативы могла бы быть сформулирована гипотеза, утверждающая существенные отличия их роста:

$$H_1 : \overline{L}_5 \neq \overline{L}_1 .$$

Если в результате статистической проверки гипотеза H_0 будет отвергнута, то тем самым будет доказана возможность принятия гипотезы H_1 Если следовать здравому смыслу, то рост студентов в конце обучения уменьшиться не может, следовательно, он действительно стал больше. Таким образом, альтернативную гипотезу в данном примере можно сформулировать как одностороннюю:

$$H_1 : \overline{L}_5 > \overline{L}_1 \blacktriangleright$$

Суждение об истинности или логичности статистической гипотезы строится на основе критериальной (тестовой) проверки. Статистические критерии отражают результаты сравнений значений, определенных по выборочным данным, с критическими, установленными теоретическим путем. Если наблюдаемое значение не превышает критического, то теоретически отсутствуют основания, чтобы отвергнуть нулевую гипотезу H_0. В противном случае принимается справедливость альтернативной гипотезы H_1.

Критериальная проверка гипотез допускает определенную вероятность ошибки в выводах. При этом различают вероятность ошибки первого рода (q) — отвергнуть нулевую гипотезу, когда она справедлива, и второго рода — принять нулевую гипотезу, когда она ложна (β).

Для примера воспользуемся ситуацией, описанной И. Ильфом и Е. Петровым в знаменитом романе «Золотой теленок». О. Бендер из разговора с Шурой Балагановым узнает, что в г. Черноморске живет богатый человек по фамилии Корейко. Все дальнейшие события романа связаны с выдвижением гипотезы H_0 о том, что *гражданин Корейко = подпольный миллионер.* Рискованные операции, на

которые пускался Бендер, основывались на его уверенности, что у Корейко есть миллион, с вероятностью $P = 1 - q$. Конечно, можно было бы не поверить Шуре Балаганову и предположить (т.е. выдвинуть гипотезу), что

H_0: Корейко = честный бухгалтер, тогда вероятность получения миллиона составила бы $P = 1 - \beta$. Но Бендеру очень хотелось иметь миллион, и возможность пропустить этот шанс для него значительно важнее, чем, возможно, напрасные хлопоты по изъятию этого миллиона. И вообще следование пессимистическим прогнозам не в характере великого комбинатора.

Область с малой вероятностью q попадания критериальных значений (рис. 5.1) характеризуется как критическая (пороговая), а область $1 - q$ — допустимая. Гипотеза H_0 принимается только при попадании критерия в область допустимых значений.

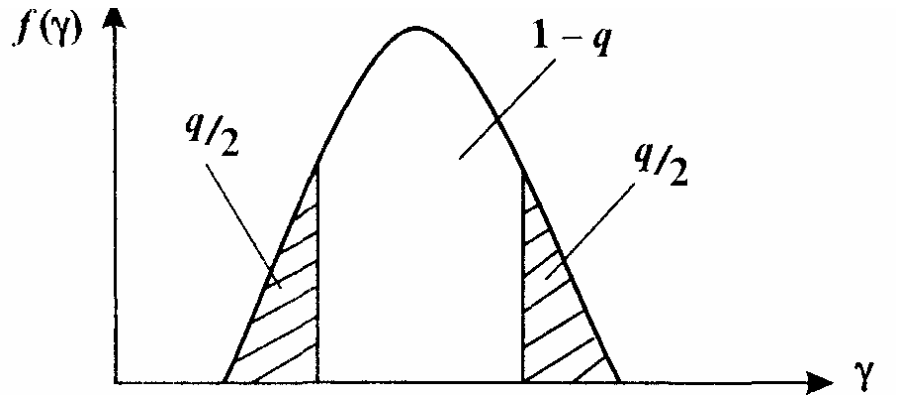

Рис. 5.1 График плотности распределения критической статистики γ.

Для нулевой гипотезы $H_0 : \bar{\theta}_1 = \bar{\theta}_2$ о равенстве статистических параметров распределения вероятностей альтернативная гипотеза, утверждающая неравенство параметров, распадается на две:

$$H_1 : \bar{\theta}_1 > \bar{\theta}_2 \text{ и } H_1 : \bar{\theta}_1 < \theta_2 .$$

В этом случае для проверки H_0 применяют двусторонний критерий со значениями $K_{1-q/2}$ и $K_{q/2}$. Если одно из неравенств практически невозможно (например, заведомо известно, что $P(\bar{\theta}_1 > \bar{\theta}_2) = 0$ то используют односторонний критерий K_q^1. Наиболее распространенными в практике значениями q являются: 0,01, 0,05; 0,1, что указывает на вероятность получения достоверного вывода ($1 - q$), равную 0,99; 0,95; 0,9.

5.2 Последовательные методы проверки гипотез о значениях параметров распределений

Большинство методов проверки статистических гипотез используют фиксированный объем выборки. Вальд предложил теорию последовательной проверки гипотез (последовательный анализ), существенным отличием которой является то, что число наблюдений, необходимое для принятия решения по гипотезе, зависит от исходов испытаний и является случайной, не фиксированной

заранее величиной.

Метод последовательной проверки гипотезы предполагает на каждой стадии наблюдений (эксперимента) принятие одного из возможных решений: принять гипотезу, отклонить ее или продолжить наблюдения. Обычно при последовательном анализе нулевая гипотеза относительно значения параметра θ формулируется в форме предположения об одном из двух его возможных значений θ_0 или θ_1:

$$H_0 : \theta = \theta_0; \;\; H_1 : \theta = \theta_1 .$$

Задача последовательного анализа в ходе эксперимента выбрать одну из гипотез. Вальд показал, что для проверки гипотез методами последовательного анализа требуется в среднем в два раза меньше наблюдений, чем при проверке классическими методами, основанными на заранее фиксированном числе наблюдений. Впоследствии было показано, что при определенных условиях выигрыш от применения последовательной процедуры по сравнению с классической теоретически неограничен.

Перед планированием процедуры последовательного анализа назначают приемлемые величины вероятностей допустимых ошибок: α – вероятность принятия гипотезы H_1, когда верна гипотеза H_0 (ошибка первого рода) и β – вероятность принятия гипотезы H_0, когда верна гипотеза H_1 (ошибка второго рода).

Наибольший выигрыш последовательный анализ дает при $\alpha \gg \beta$ или $\alpha \ll \beta$ т. е. когда α и β являются величинами разного порядка малости.

Так как число наблюдений n в последовательном анализе является величиной случайной, то необходимо знать либо его функцию распределения вероятностей, либо параметры этого распределения (например, среднее количество необходимых наблюдений). Среднее значение зависит только от истинного значения параметра, относительно которого проверяется гипотеза.

Функция $\bar{n}(\theta)$, определяющая зависимость $\bar{n}$ от θ, называется функцией среднего числа наблюдений.

На практике обычно находят средние значения числа наблюдений $\bar{n}(\theta_0)$ и $\bar{n}(\theta_1)$, соответствующие гипотетическим значениям параметра θ_0 и θ_1, между которыми осуществляется выбор, и максимальное среднее значение числа наблюдений $\bar{n}_{\max}$, необходимое для окончания последовательной процедуры проверки гипотезы.

Если знания только среднего числа наблюдений недостаточно и требуется определить либо вероятность того, что для завершения последовательной процедуры понадобится не более некоторого, наперед заданного, числа наблюдений, либо число наблюдений, соответствующее заданной вероятности завершения последовательной процедуры, используются таблицы распределения Вальда с функцией

$$P(x < a) = W_c(a) = \sqrt{\frac{c}{2\pi}} \int_0^a x^{-3/2} \exp\left\{ -\frac{c}{2}\left(x + \frac{1}{x} - 2 \right) \right\} dx$$

где $x = \frac{n}{=}{n}$ — отношение числа наблюдений к его среднему значению; c – параметр распределения, определяемый видом распределения исследуемой случайной величины и гипотетическим значением параметра θ (θ_0 или θ_1).

Вероятность γ завершения процедуры последовательного анализа и число испытаний $n(\theta)$ при некотором значении θ (θ_0 или θ_1) связаны соотношением $\gamma = W_{c(\theta)}\left(\frac{n(\theta)}{\bar{n}(\theta)}\right)$, из которого можно определить либо γ, соответствующее заданному $n(\theta)$, либо $n(\theta)$, соответствующее заданному значению γ.

Значения $W_c\left(x = \frac{n}{\bar{n}}\right)$ приведены в табл. 10 (см. статистические таблицы)

5.3 Проверка гипотезы о параметрах нормального распределения

5.3.1 Проверка гипотезы о значении среднего

Проверяется нулевая гипотеза $H_0 : \mu = \mu_0$ против альтернативы $H_0 : \mu = \mu_1$ $(\mu_0 < \mu_1)$. Полагается, что дисперсии σ_1^2 и σ_2^2 известны заранее, причем $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Введем обозначения

$$A = \frac{\sigma^2}{\mu_1 - \mu_0}\ln\frac{\beta}{1-\alpha} + n\frac{\mu_1 + \mu_0}{2};\ B = \frac{\sigma^2}{\mu_1 - \mu_0}\ln\frac{1-\beta}{\alpha} + n\frac{\mu_1 + \mu_0}{2}.$$

Если:

$\sum_{i=1}^{n} x_i \le A$, то принимается гипотеза H_0;

$\sum_{i=1}^{n} x_i \ge B$, то принимается гипотеза H_1;

$A < \sum_{i=1}^{n} x_i < B$, то наблюдения продолжаются.

Средние объемы выборок, необходимые для завершения процедуры последовательного анализа, равны

$$\bar{n}(\mu_0) = 2\sigma^2\frac{(1-\alpha)\ln\frac{1-\alpha}{\beta} + \alpha\ln\frac{\alpha}{1-\beta}}{(\mu_1 - \mu_0)^2};\ \bar{n}(\mu_1) = 2\sigma^2\frac{(1-\beta)\ln\frac{1-\beta}{\alpha} + \beta\ln\frac{\beta}{1-\alpha}}{(\mu_1 - \mu_0)^2};$$

$$\bar{n}_{\max} = -\sigma^2\frac{\ln\frac{\beta}{1-\alpha}\cdot\ln\frac{1-\beta}{\alpha}}{(\mu_1 - \mu_0)^2}.$$

Здесь $\bar{n}_{\max}$ – максимальное среднее значение числа наблюдений, необходимое для окончания последовательной процедуры проверки гипотезы.

Параметр c распределения Вальда находится по формуле

$$c = K\frac{\left|\mu - \frac{\mu_0+\mu_1}{2}\right|}{\mu_1-\mu_0}, \text{ где } K = \begin{cases} \ln\frac{1-\alpha}{\beta}, \ \textit{если } \mu=\mu_0, \ \alpha \ll \beta; \\ \ln\frac{1-\beta}{\alpha}, \ \textit{если } \mu=\mu_1, \ \beta \ll \alpha. \end{cases}.$$

Пример 5.2. Предположим, что параметр прибора распределен нормально со стандартным отклонением $\sigma = 200$. Необходимо проверить при $\alpha = 0{,}1$ и $\beta = 0{,}01$ гипотезу о том, что параметр прибора равен $\mu = \mu_0 = 1800$, против альтернативы $\mu = \mu_1 = 2000$.

Найти контрольные границы и средние объемы выборок для последовательной проверки гипотезы. Определить объем выборки $n_{0,95}$, для которого с вероятностью не менее 0,95 процедура последовательного анализа закончится принятием решения по гипотезе.

Решение. Находим

$$A = \frac{200^2}{2000-1800}\ln\frac{0{,}01}{0{,}9} + n\frac{2000+1800}{2} = -899{,}96 + 1900\cdot n;$$

$$B = \frac{200^2}{2000-1800}\ln\frac{0{,}99}{0{,}1} + n\frac{2000+1800}{2} = 458{,}5 + 1900\cdot n.$$

Далее вычисляем средние объемы выборок

$$\bar{n}(1800) = 2\cdot 200^2\frac{0{,}9\cdot\ln\frac{0{,}9}{0{,}01} + 0{,}1\cdot\ln\frac{0{,}1}{0{,}99}}{(2000-1800)^2} = 8;$$

$$\bar{n}(2000) = 2\cdot 200^2\frac{0{,}01\cdot\ln\frac{0{,}01}{0{,}9} + 0{,}99\cdot\ln\frac{0{,}99}{0{,}1}}{(2000-1800)^2} = 5;$$

$$\bar{n}_{\max} = -200^2\frac{\ln\frac{0{,}01}{0{,}9}\cdot\ln\frac{0{,}99}{0{,}1}}{(2000-1800)^2} = 10.$$

Так как $\beta \ll \alpha$, то $K = \ln\frac{0{,}99}{0{,}1} = 2{,}2925$. Вычисляем параметр c для гипотезы H_1:

$$c = 2{,}2925\cdot\frac{\left|2000 - \frac{2000+1800}{2}\right|}{2000-1800} = 1{,}147.$$

Из таблицы $W_c(x)$ (см. стр. 440, а также статистические таблицы) находим для $c = 1{,}1462$ значение $x = \frac{n_{0,95}}{\bar{n}}$, соответствующее условию $W_c(x) = 0{,}95$. Путем интерполяции получим $x = 2{,}845$. Тогда для гипотезы H_1 получим: $n_{0,95} = 2{,}845\cdot\bar{n}(2000) = 2{,}845\cdot 5 = 15$. Таким образом, с вероятностью 0,95 для принятия решения по гипотезе потребуется не более 15 испытаний.

Итак, принимаем гипотезу H_0, если $\sum_{i=1}^{n} x_i \le -899,96 + 1900 \cdot n$, и принимаем гипотезу H_1, если $\sum_{i=1}^{n} x_i \ge 458,5 + 1900 \cdot n$. В любом ином случае испытания необходимо продолжить. При H_0 в среднем понадобится 8 испытаний, а при H_1 – 5 испытаний. Максимальное среднее число испытаний не превысит 10.►

5.3.2 Проверка гипотезы о значении дисперсии

Проверяется гипотеза $H_0: \sigma^2 = \sigma_0^2$ против альтернативы $H_1: \sigma^2 = \sigma_1^2$ $(\sigma_1 > \sigma_2)$ при известном среднем μ.

Гипотеза H_0 принимается, если $\sum_{i=1}^{n} x_i^2 \le A$; если $\sum_{i=1}^{n} x_i^2 \ge B$, принимается гипотеза H_1. Если $A < \sum_{i=1}^{n} x_i^2 < B$, то испытания продолжаются.

Здесь $A = \frac{2 \cdot \sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln\left[\frac{\beta}{1-\alpha}\left(\frac{\sigma_1}{\sigma_0}\right)^n\right]$; $B = \frac{2 \cdot \sigma_0^2 \sigma_1^2}{\sigma_1^2 - \sigma_0^2} \ln\left[\frac{1-\beta}{\alpha}\left(\frac{\sigma_1}{\sigma_0}\right)^n\right]$.

Средние объемы выборок, необходимые для завершения последовательной проверки, равны

$$\bar{n}(\sigma_0^2) = 2\frac{(1-\alpha)\cdot \ln\frac{1-\alpha}{\beta} + \alpha \cdot \ln\frac{\alpha}{1-\beta}}{\left(\frac{\sigma_0}{\sigma_1}\right)^2 + 2 \cdot \ln\frac{\sigma_1}{\sigma_0} - 1}; \quad \bar{n}(\sigma_1^2) = 2\frac{\beta \cdot \ln\frac{\beta}{1-\alpha} + (1-\beta)\cdot \ln\frac{1-\beta}{\alpha}}{\left(\frac{\sigma_1}{\sigma_0}\right)^2 - 1 - 2 \cdot \ln\frac{\sigma_1}{\sigma_0}};$$

$$\bar{n}_{\max} = -\frac{\ln\frac{\beta}{1-\alpha} \cdot \ln\frac{1-\beta}{\alpha}}{\left(\ln\left(\frac{\sigma_1}{\sigma_0}\right)^2\right)^2}.$$

Вероятностные оценки необходимого числа испытаний могут быть найдены по аналогии с критерием для проверки гипотезы о среднем значении, с помощью таблицы распределения Вальда $W_c(x)$ ([1] стр. 440) при $c = K\frac{\left|\sigma^2 - \frac{\sigma_0^2 + \sigma_1^2}{2}\right|}{\sigma_1^2 - \sigma_0^2}$, где

$$K = \begin{cases} \ln\frac{1-\alpha}{\beta}, \ \textit{если } \sigma^2 = \sigma_0^2, \ \ \alpha \ll \beta; \\ \ln\frac{1-\beta}{\alpha}, \ \textit{если } \sigma^2 = \sigma_1^2, \ \ \beta \ll \alpha. \end{cases}$$

Если значение среднего μ неизвестно, то все приведенные соотношения сохраняются при замене n на $n-1$.

Пример 5.3. Предположим, что параметр прибора распределен нормально с

известным средним μ. Необходимо проверить при $\alpha = 0,01$ и $\beta = 0,1$ гипотезу о том, что дисперсия значений параметра прибора равна $\sigma^2 = \sigma_0^2 = 25$, против альтернативы $\sigma^2 = \sigma_1 = 49$.

Найти контрольные границы и средние объемы выборок для последовательной проверки гипотез. Определить объем выборки $n_{0,9}$, для которого с вероятностью не менее $\gamma = 0,9$ процедура последовательного анализа закончится принятием решения по гипотезе. Определить вероятность того, что последовательная процедура потребует не более 20 испытаний.

Решение.

Находим $A = \frac{2 \cdot 25 \cdot 49}{49 - 25} \ln\left[\frac{0,1}{0,99} \cdot \left(\frac{7}{5}\right)^n\right] = -234,029 + 34,348 \cdot n$;

$B = \frac{2 \cdot 25 \cdot 49}{49 - 25} \ln\left[\frac{0,9}{0,01} \cdot \left(\frac{7}{5}\right)^n\right] = 459,354 + 34,348 \cdot n$.

Если $\sum_{i=1}^{n} x_i^2 \le -234,029 + 34,348 \cdot n$, то принимается гипотеза о том, что $\sigma^2 = \sigma_0^2 = 25$; если $\sum_{i=1}^{n} x_i^2 \ge 459,354 + 34,348 \cdot n$, то принимается гипотеза о том, что $\sigma^2 = \sigma_1^2 = 49$. В случае $-234,029 + 34,348 \cdot n < \sum_{i=1}^{n} x_i^2 < 459,354 + 34,348 \cdot n$ испытания продолжаются.

Далее вычисляем средние объемы выборок

$$\bar{n}(25) = 2 \cdot \frac{0,99 \cdot \ln\frac{0,99}{0,1} + 0,01 \cdot \ln\frac{0,01}{0,9}}{\left(\frac{5}{7}\right)^2 + 2 \cdot \ln\frac{7}{5} - 1} = 25;$$

$$\bar{n}(49) = 2 \cdot \frac{0,1 \cdot \ln\frac{0,1}{0,99} + 0,9 \cdot \ln\frac{0,9}{0,01}}{\left(\frac{7}{5}\right)^2 - 1 - 2 \cdot \ln\frac{7}{5}} = 27;$$

$$\bar{n}_{\max} = -\frac{\ln\frac{0,1}{0,99} \cdot \ln\frac{0,9}{0,01}}{\left(\ln\left(\frac{7}{5}\right)^2\right)^2} = 46.$$

В нашем случае $\alpha \ll \beta$, тогда находим $K = \ln\frac{0,99}{0,1} = 2,292$ и

$$c = 2,292 \cdot \frac{\left|25 - \frac{25 + 49}{2}\right|}{49 - 25} = 1,146.$$

Из таблицы распределения Вальда имеем $W_{1,146}(x)=0,9$ при $x=\frac{n_{0,9}}{\bar{n}}\approx 2,12$, тогда $n_{0,9}=\bar{n}(25)\cdot 2.12=53$.

Теперь определим вероятность окончания последовательной процедуры при $n\le 20$. Имеем $x=\frac{n}{\bar{n}}=\frac{20}{25}=0,8$. Из таблицы распределения Вальда для $c=1,146$ и $x=0,8$ получим $W_{1,146}(0,8)=0,56$. Следовательно, с вероятностью 0,56 для окончания последовательной процедуры потребуется не более 20 испытаний, при условии, что $\sigma^2=\sigma_0^2=25$.►

5.4 Проверка гипотезы о параметре экспоненциального распределения

Проверяется гипотеза $H_0:\lambda=\lambda_0$ (напомним, что $\lambda=1/v$ – средняя наработка до отказа, v – интенсивность отказов), против альтернативы $H_1:\lambda=\lambda_1>\lambda_0$.

Введем обозначения

$$A=\frac{n}{\lambda_1-\lambda_0}\cdot\ln\left[\frac{\lambda_1}{\lambda_0}\cdot\left(\frac{\alpha}{1-\beta}\right)^{1/n}\right];\ B=\frac{n}{\lambda_1-\lambda_0}\cdot\ln\left[\frac{\lambda_1}{\lambda_0}\cdot\left(\frac{1-\alpha}{\beta}\right)^{1/n}\right].$$

Если $\sum_{i=1}^{n}x_i\le A$, принимается гипотеза H_0; если $\sum_{i=1}^{n}x_i\ge B$, принимается гипотеза H_1; если $A<\sum_{i=1}^{n}x_i<B$, то испытания продолжаются.

Средние объемы выборок, необходимые для завершения процедуры последовательного анализа, вычисляются по формулам

$$\bar{n}(\lambda_0)=\frac{(1-\alpha)\cdot\ln\frac{1-\alpha}{\beta}+\alpha\cdot\ln\frac{\alpha}{1-\beta}}{\frac{\lambda_1}{\lambda_0}-1-\ln\frac{\lambda_1}{\lambda_0}};\ \bar{n}(\lambda_1)=2\sigma^2\frac{(1-\beta)\cdot\ln\frac{1-\beta}{\alpha}+\beta\cdot\ln\frac{\beta}{1-\alpha}}{\frac{\lambda_0}{\lambda_1}-1+\ln\frac{\lambda_1}{\lambda_0}};$$

$$\bar{n}_{\max}=-\frac{\ln\frac{1-\alpha}{\beta}\cdot\ln\frac{\alpha}{1-\beta}}{\left(\ln\left(\frac{\lambda_1}{\lambda_0}\right)\right)^2}.$$

Параметр распределения Вальда вычисляются по формулам

$$c=K\frac{\left|\ln\frac{\lambda_1}{\lambda_0}-\frac{\lambda_1-\lambda_0}{2}\right|}{\left(\frac{\lambda_1-\lambda_0}{2}\right)^2},\ \text{где}\ K=\begin{cases}\ln\frac{1-\alpha}{\beta},\ \textit{если}\ \lambda=\lambda_0,\ \alpha\ll\beta;\\ \ln\frac{1-\beta}{\alpha},\ \textit{если}\ \lambda=\lambda_1,\ \beta\ll\alpha.\end{cases}.$$

Пример 5.4. Найти параметры плана последовательных испытаний при $\alpha = 0,2$ и $\beta = 0,05$ для проверки гипотезы H_0 о том, что средняя наработка на отказ электронного прибора равна $\lambda = \lambda_0 = 100$ ч. против альтернативы $H_1: \lambda = \lambda_1 = 150$ ч. Вычислить средние объемы выборок, необходимых для окончания последовательного анализа. Найти вероятность того, что последовательная процедура позволит принять решение уже после $n = 20$ испытаний.

Решение. Находим

$$A = \frac{n}{150-100} \cdot \ln\left[\frac{150}{100} \cdot \left(\frac{0,2}{0,95}\right)^{1/n}\right] = 0,00811 \cdot n - 0,0312;$$

$$B = \frac{n}{150-100} \cdot \ln\left[\frac{150}{100} \cdot \left(\frac{0,8}{0,05}\right)^{1/n}\right] = 0,00811 \cdot n - 0,0554.$$

Далее вычисляем

$$\bar{n}(100) = \frac{0,8 \cdot \ln\frac{0,8}{0,05} + 0,2 \cdot \ln\frac{0,2}{0,95}}{\frac{150}{100} - 1 - \ln\frac{150}{100}} = 21; \quad \bar{n}(150) = \frac{0,05 \cdot \ln\frac{0,05}{0,8} + 0,95 \cdot \ln\frac{0,95}{0,2}}{\frac{100}{150} - 1 + \ln\frac{150}{100}} = 19;$$

$$\bar{n}_{\max} = -\frac{\ln\frac{0,8}{0,05} \cdot \ln\frac{0,2}{0,95}}{\left(\ln\left(\frac{150}{100}\right)\right)^2} = 27.$$

Теперь найдем, учитывая, что $\beta \ll \alpha$ и $\lambda = \lambda_1 = 150$, $K = \ln\left(\frac{0,95}{0,2}\right) = 1,558$ и

$$c = 1,558\frac{\left|\ln\frac{150}{100} - \frac{150-100}{2}\right|}{\left(\frac{150-100}{2}\right)^2} = 0,061.$$

Имеем $x = \frac{n}{\bar{n}(\lambda_1)} = \frac{20}{19} \approx 1$ и из табл. Вальда для $x = 1$ и $c = 0,061$ имеем $W_c(x) = 0,85$. Таким образом, с вероятностью 0,85 процедура последовательного анализа закончится при $n = 20$, если $\lambda = \lambda_1 = 150$.►

5.5 Проверка гипотезы о параметре биномиального распределения

Проверяется гипотеза $H_0: p = p_0$ против альтернативы $H_1: p > p_1$.

Введем обозначения

$$A = \frac{\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right) + \ln\left(\frac{1-p_0}{1-p_1}\right)} + n \cdot \frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right) + \ln\left(\frac{1-p_0}{1-p_1}\right)};$$

$$B=\frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right)+\ln\left(\frac{1-p_0}{1-p_1}\right)}+n\cdot\frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln\left(\frac{p_1}{p_0}\right)+\ln\left(\frac{1-p_0}{1-p_1}\right)}.$$

Пусть x – число поступлений наблюдаемого события (например, количество дефектных приборов в партии). Если $x\le A$, то принимается гипотеза H_0; если $x\ge B$, принимается гипотеза H_1; в случае $A<x<B$ испытания продолжаются. Средние объемы выборок равны

$$\bar{n}(p_0)=\frac{(1-\alpha)\cdot\ln\left(\frac{\beta}{1-\alpha}\right)+\alpha\cdot\ln\left(\frac{1-\beta}{\alpha}\right)}{p_0\cdot\ln\left(\frac{p_1}{p_0}\right)-(1-p_0)\cdot\ln\left(\frac{1-p_0}{1-p_1}\right)};\ \bar{n}(p_1)=\frac{\beta\cdot\ln\left(\frac{\beta}{1-\alpha}\right)+(1-\beta)\cdot\ln\left(\frac{1-\beta}{\alpha}\right)}{p_1\cdot\ln\left(\frac{p_1}{p_0}\right)-(1-p_1)\cdot\ln\left(\frac{1-p_0}{1-p_1}\right)};$$

$$\bar{n}_{\max}=-\frac{\ln\left(\frac{1-\beta}{\alpha}\right)\cdot\ln\left(\frac{\beta}{1-\alpha}\right)}{\ln\left(\frac{p_1}{p_0}\right)\cdot\ln\left(\frac{1-p_0}{1-p_1}\right)}.$$

Параметр распределения Вальда находится по формуле

$$c=K\cdot\frac{\left|p_1\cdot\ln\left(\frac{p_1}{p_0}\right)-(1-p_1)\cdot\ln\left(\frac{1-p_0}{1-p_1}\right)\right|}{p_1\cdot(1-p_1)\cdot\left(\ln\left(\frac{p_1}{p_0}\right)+\ln\left(\frac{1-p_0}{1-p_1}\right)\right)},\ \text{где}\ K=\begin{cases}\ln\frac{1-\alpha}{\beta},\ \textit{если}\ p=p_0,\ \alpha\ll\beta;\\ \ln\frac{1-\beta}{\alpha},\ \textit{если}\ p=p_1,\ \beta\ll\alpha.\end{cases}.$$

Пример 5.5. Необходимо найти параметры плана последовательных испытаний для проверки гипотезы H_0 о том, что доля дефектных изделий в партии $p=p_0=0,01$ против альтернативы H_1: $p=p_1=0,02$. Определить количество испытаний, для которого с вероятностью $\gamma=0,95$ последовательная процедура закончится принятием решения. Заданы ошибка первого рода $\alpha=0,1$ и ошибка второго рода $\beta=0,01$.

Решение. Находим

$$A=\frac{\ln\left(\frac{0,01}{0,9}\right)}{\ln\left(\frac{0,02}{0,01}\right)+\ln\left(\frac{0,99}{0,98}\right)}+n\cdot\frac{\ln\left(\frac{0,99}{0,98}\right)}{\ln\left(\frac{0,02}{0,01}\right)+\ln\left(\frac{0,99}{0,98}\right)}=-6,399+0,0144\cdot n;$$

$$B=\frac{\ln\left(\frac{0,99}{0,1}\right)}{\ln\left(\frac{0,02}{0,01}\right)+\ln\left(\frac{0,99}{0,98}\right)}+n\cdot\frac{\ln\left(\frac{0,99}{0,98}\right)}{\ln\left(\frac{0,02}{0,01}\right)+\ln\left(\frac{0,99}{0,98}\right)}=3,26+0,0144\cdot n.$$

Далее

$$\bar{n}(0,01)=\frac{0,9\cdot\ln\left(\frac{0,01}{0,9}\right)+0,1\cdot\ln\left(\frac{0,99}{0,1}\right)}{0,01\cdot\ln\left(\frac{0,01}{0,9}\right)-0,99\cdot\ln\left(\frac{0,99}{0,98}\right)}=1225;$$

$$\bar{n}(0,02)=\frac{0,01\cdot\ln\left(\frac{0,01}{0,9}\right)+0,99\cdot\ln\left(\frac{0,99}{0,1}\right)}{0,02\cdot\ln\left(\frac{0,02}{0,01}\right)-0,98\cdot\ln\left(\frac{0,99}{0,98}\right)}=569;$$

$$\bar{n}_{\max}=-\frac{\ln\left(\frac{0,99}{0,1}\right)\cdot\ln\left(\frac{0,01}{0,9}\right)}{\ln\left(\frac{0,02}{0,01}\right)\cdot\ln\left(\frac{0,99}{0,98}\right)}=1466.$$

Вычисляем параметр c (при $\beta \ll \alpha$ и $p=p_1=0,02$):

$$K=\ln\left(\frac{1-\beta}{\alpha}\right)=\ln\left(\frac{0,99}{0,1}\right)=2,2925,$$

$$c=2,2925\cdot\frac{\left|0,02\cdot\ln\left(\frac{0,02}{0,01}\right)-0,98\cdot\ln\left(\frac{0,99}{0,98}\right)\right|}{0,02\cdot0,98\cdot\left(\ln\left(\frac{0,02}{0,01}\right)+\ln\left(\frac{0,99}{0,98}\right)\right)}=0,65.$$

Из таблицы Вальда находим $x=\frac{n_{0,95}}{\bar{n}}$, соответствующее условию $W_c\ (x)=0,95$.

Имеем $W_{0,65}(x)=0,95$ при $x\approx 3,5$. Отсюда, $n_{0,95}=569\cdot 3,5=1992$. Следовательно, с вероятностью 0,95 при $p=p_1=0,02$ процедура последовательного анализа потребует не более 1992 испытаний.►

Вопросы для самоконтроля

1. Проверка гипотез о значениях параметров распределений.
2. Последовательные методы проверки гипотез о значениях параметров распределений
3. Проверка гипотезы о числовом значении математического ожидания нормального распределения при известной дисперсии (случаи равных дисперсий).
4. Проверка гипотезы о числовом значении дисперсии нормального распределения
5. Проверка гипотезы о числовом значении параметра экспоненциального распределения
6. Проверка гипотезы о числовом значении параметра биномиального распределения

Тема 6. Дисперсионный анализ зависимостей

6.1 Основные положения

В предыдущих главах были рассмотрены различные методы и приемы математической статистики, позволяющие оценить параметры статистических совокупностей, сравнивать их между собой. При этом, как правило, предполагалась взаимная независимость сравниваемых совокупностей. В настоящей главе рассматриваются вопросы оценки связей между статистическими совокупностями.

Для оценки связей между статистическими совокупностями используются методы дисперсионного, корреляционного и регрессионного анализов, являющиеся последовательными ступенями при исследовании связей между случайными величинами.

Методами дисперсионного анализа устанавливается наличие влияния заданного фактора на изучаемый процесс, отображаемый наблюдаемой статистической совокупностью выборочных данных. Корреляционный анализ позволяет оценить силу такой связи, а методами регрессионного анализа можно выбрать конкретную математическую модель и оценить адекватность отражения ею установленной взаимосвязи случайных величин.

В последние годы стремительно развивается самостоятельное прикладное направление математической статистики – математическая теория активного эксперимента. Базируясь на комбинации методов дисперсионного и регрессионного анализов, методы математического планирования эксперимента дополняют их.

Для установления самого факта наличия (или отсутствия) статистически значимой связи между результирующими показателями Y и объясняющими переменными X могут быть использованы параметрический и непараметрический подходы.

Предполагают, что все наблюдения принадлежат некоторому семейству распределений (т.е. изменяется только m_x при изменении уровня фактора). Если в качестве такого распределения рассматривается нормальное, то для обработки данных могут применяться *методы дисперсионного анализа.* Если предположение о нормальности не является правомерным, тогда используют различные *методы непараметрического анализа.*

Если проводить наблюдения над каким-либо явлением, характеризуемым случайной величиной Y, то значение этой наблюдаемой величины может изменяться от каких-то определенных факторов A, качественных или количественных, а также от совокупности случайных воздействий, делающих фактически эту величину случайной. Как правило, влияние случайных факторов не приводит к смещению среднего значения наблюдаемой величины (изменяется только дисперсия), влияние же определенных факторов отмечается таким смещением. Исследование влияния факторов на изменчивость средних значений является *основной задачей дисперсионного анализа.* Для оценки этого влияния используется свойство аддитивности дисперсии изучаемой случайной величины, обусловленной действием *независимых* факторов. В 1938 г. Р.Н. Фишер впервые определил дисперсионный анализ как отделение дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам. Исходя из этого определения, в зависимости от числа источников дисперсии различают однофакторный и двухфакторный дисперсионный анализ.

Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость изучаемой величины. Для этого производится разложение суммарной выборочной дисперсии на составляющие, обусловленные независимыми факторами. Каждая из этих составляющих представляет собой

оценку генеральной дисперсии. Для определения значимости влияния того или иного определенного фактора дается оценка отношения выборочной дисперсии, соответствующей этому фактору, к дисперсии, обусловленной случайными факторами (дисперсия воспроизводимости). Оценка осуществляется по критерию Фишера.

При этом делаются следующие допущения:

- случайные ошибки наблюдения распределены нормально;
- в эксперименте используются равноточные методы измерения;
- факторы влияют только на изменение средних значений, а дисперсия наблюдаемой величины считается постоянной.

Понятие «определенный фактор» обусловлено следующим:

- это исследуемый фактор, т.е. априори установленный;
- этот фактор имеет количественную или качественную шкалу измерения;
- эта шкала разделена на уровни.

Например, исследуется влияние климатических условий на урожайность. Климатические условия включают в себя три фактора: световой, температурный и влажностный. Первый фактор определяется тремя значениями: солнечной, переменной и пасмурной погодой; второй — средней температурой в пределах: 5-10 °С, 10-15 °С, 15-20 °С, 20-25 °С, 25 30 °С; третий также имеет три уровня, соответствующих понятиям: дождливая погода, нормальная погода, сухая погода. Первый и третий имеют качественную оценку уровней, а второй — количественную.

6.2. Однофакторный анализ

6.2.1.0днофакторный дисперсионный анализ

Рассмотрим влияние фактора A на исследуемый процесс X, принимающего k различных значений — уровней фактора. На каждом i-м уровне производится n_i наблюдений, результаты которых занесены в таблицу 6.1.

Результат каждого наблюдения может быть представлен в виде модели:

$$x_{ji} = \mu + \alpha_i + e_{ji},\ i = 1,...,n, \qquad (6.1)$$

где μ – суммарный эффект во всех опытах; α_i – эффект фактора A на i-м уровне; e_{ji} – ошибка определения x_{ji} на i-м уровне.

Таблица 6.1. Форма представления экспериментальных данных однофакторной модели

Номер Наблюдения	Уровни фактора A			
	A_1	A_2	A_i	A_k
1	x_{11}	x_{12}	x_{1i}	x_{1k}
2	x_{21}	x_{22}	x_{2i}	x_{2k}
….	….	….	….	….
j	x_{j1}	x_{j2}	x_{ji}	x_{jk}

….	….	…..	…..	…..
n	$x_{n_1 1}$	$x_{n_2 2}$	$x_{n_i i}$	$x_{n_k k}$
Средние значения по уровням	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_i$	$\bar{x}_k$

Предположим, что наблюдения на i-м фиксированном уровне фактора нормально распределены относительно среднего значения $(\mu + \alpha_i)$ с общей дисперсией σ^2.

Общее число опытов

$$N = \sum_{i=1}^{k} n_i \,. \tag{6.2}$$

Следует установить, оказывает ли влияние фактор A на исследуемый процесс X. Сформулируем гипотезу H_0 о том, что расхождение наблюдений в сериях опытов для различных уровней факторов можно объяснить только случайными причинами. На статистическом языке это предположение означает, что все данные таблицы x_{ij} принадлежат одному и тому же распределению.

Осуществим проверку нулевой гипотезы равенства средних значений на различных уровнях фактора A:

$$H_0 : m_1 = m_2 = ... = m_k = m \,.$$

Наиболее часто расчет проводится при равном числе опытов на каждом уровне A, т.е. $n_1 = n_2 = ... = n_k = n$. При этом общее число наблюдений $N = k \times n$.

Среднее значение результатов наблюдений на i-м уровне:

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ji} \,. \tag{6.3}$$

Общее среднее значение для всей выборки из N наблюдений:

$$\bar{x} = \frac{1}{k} \sum_{i=1}^{k} \bar{x}_i = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ji} \,. \tag{6.4}$$

Выборочная дисперсия на каждом уровне:

$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)^2 = \frac{1}{n-1} \left[\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{n} \left(\sum_{j=1}^{n} x_{ji} \right)^2 \right]. \tag{6.5}$$

При выполнении условия $s_i^2 = const$, т.е. между дисперсиями s_i^2 нет значимых различий (однородность дисперсий s_i^2 определяется по критерию Кохрена), находим оценку дисперсии, характеризующей рассеяние значений x_{ji} вне влияния фактора A (т.е. дисперсии, характеризующей фактор случайности) по формуле

$$s_{сл}^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2 = \\ = \frac{1}{k(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ji}\right)^2\right]. \qquad (6.6)$$

Легко видеть, что если при оценке s_i^2 мы имеем $(n-1)$ степеней свободы, то оценка $s_{сл}^2$ имеет $v = k(n-1) = N - k$ степеней свободы.

Общая выборочная дисперсия всех наблюдений равна:

$$s_0^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ji} - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}\right)^2\right]. \qquad (6.7)$$

Исходя из свойства аддитивности дисперсии можно общую дисперсию σ_0^2 представить как сумму составляющих этой дисперсии: σ_A^2 — характеризующий вклад фактора A и $\sigma_{сл}^2$ — характеризующий фактор случайности (совокупность случайных факторов):

$$\sigma_0^2 = \sigma_A^2 + \sigma_{сл}^2 .$$

Тогда с учетом выборочных оценок дисперсии

$$s_0^2 = s_A^2 + s_{сл}^2 . \qquad (6.8)$$

Введем теперь оценку дисперсии s_A^2, характеризующей изменение средних $\bar{x}_i$, связанное с влиянием фактора A

$$s_A^2 = \sum_{j=1}^{n}\left(\frac{1}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2\right) = \frac{n}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 \qquad (6.9)$$

с числом степеней свободы $v = k - 1$.

Теперь проверка влияния фактора A на изменение средних может быть сведена к сравнению дисперсий s_A^2 и $s_{сл}^2$. Если дисперсия s_A^2 значимо отличается от $s_{сл}^2$ то нулевая гипотеза $H_0 : m_1 = m_2 = ... = m_k = m$ отвергается и влияние фактора A считается существенным. Для проверки гипотезы применяется односторонний критерий Фишера: влияние A считается значимым, если

$$F = \frac{s_A^2}{s_{сл}^2} > F_q(v_1, v_2),\ v_1 = k-1;\ v_2 = N - k = k(n-1) . \qquad (6.10)$$

Если отношение $\frac{s_A^2}{s_{сл}^2} \le F_q(v_1, v_2)$, то влияние фактора A следует считать незначимым. При этом общая дисперсия s_0^2 будет связана только с фактором случайности и ее можно считать однородной.

При значимости влияния фактора A, т.е. при значимости различия между $m_1 = m_2 = ... = m_k = m$, можно выяснить, какие именно средние m_i различны. Для этого используют критерии Стьюдента или ранговый критерий Дункана.

6.2.2. Непараметрические методы однофакторного анализа

Если предположение о нормальности распределения выборки не является правомерным или наблюдаются качественные показатели, следует использовать различные *методы непараметрического анализа,* среди которых наиболее развиты ранговые методы.

Рассмотрим таблицу наблюдений (см. табл. 6.1) (однофакторный дисперсионный анализ). Сформулируем гипотезу H_0 о том, что расхождение наблюдений в сериях опытов для различных уровней факторов можно объяснить только случайными причинами. На статистическом языке это предположение означает, что все данные таблицы x_{ji} принадлежат одному и тому же распределению. Рассмотрим ранговый однофакторный анализ.

Если мы ничего не знаем о распределении x_{ji}, то в этом случае проще опираться в своих выводах на отношения «больше — меньше» между наблюдениями, так как они не зависят от распределения наблюдений. При этом вся информация содержится в тех рангах r_{ji}, что получают числа x_{ji} при упорядочении всей их совокупности. Ранговые критерии применяются и тогда, когда измерения сделаны в *порядковой шкале,* например, представлены текстовыми баллами или экспертными суммами. Здесь значения x_{ji} вообще являются условностью, а содержательный смысл имеют лишь отношения «больше — меньше» между ними (например, оценки 2, 3, 4, 5). Будем рассматривать наиболее простой случай, когда среди чисел x_{ji} нет совпадающих. В этом случае преобразование таблицы 5.1 в таблицу 5.2 будет однозначным.

Для проверки гипотезы H_0 необходимо сконструировать некоторую статистику, т.е. функцию от рангов r_{ji}, которая легла бы в основу критерия проверки гипотезы. Основные требования к этой статистике следующие: ее значение при гипотезе H_0 должно заметно отличаться от ее значений при альтернативах. Рассмотрим два метода.

Таблица 6.2. Форма представления преобразованных данных

Ранги результатов наблюдений	Уровни фактора A			
	A_1	A_2	A_i	A_k
1	r_{11}	r_{12}	r_{1i}	r_{1k}
2	r_{21}	r_{22}	r_{2i}	r_{2k}
….	….	….	….	….
j	r_{j1}	r_{j2}	r_{ji}	r_{jk}
….	….	…..	…..	…..
n	$r_{n_1 1}$	$r_{n_2 2}$	$r_{n_i i}$	$r_{n_k k}$

Однофакторный непараметрический анализ на основе критерия Краскела-Уоллеса (произвольные альтернативы)

Этот метод используется, когда невозможно сказать что-либо определенное об альтернативах H_0, так как он свободен от распределения. Заменим наблюдения x_{ji} их рангами r_{ji}, упорядочивая всю совокупность $\{x_{ji}\}$ в порядке возрастания. Затем для каждой обработки i (уровня фактора, столбца таблицы) надо вычислить суммарный и средний ранги:

$$R_i = \sum_{j=1}^{n_i} r_{ji} \text{ и } \overline{R}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ji} . \quad (6.11)$$

Если между столбцами нет систематических различий, то средние ранги $\overline{R}_i$, не должны значительно отличаться от среднего, рассчитанного по всей совокупности $\{r_{ji}\}$. Значение последнего $\widehat{R} = \frac{N+1}{2}$. Здесь N — общее число наблюдений.

$$N = \sum_{i=1}^{k} n_i .$$

Вычислим величины $(\overline{R}_i - \widehat{R})^2$ для каждого уровня фактора $\left(\overline{R}_i - \frac{N+1}{2}\right)^2, ..., \left(\overline{R}_k - \frac{N+1}{2}\right)^2$.

Эти значения при H_0 в совокупности должны быть небольшими. Составляя общую характеристику, разумно учесть различия в числе наблюдений для разных обработок (уровней факторов) и взять в качестве меры отступления от чистой случайности величину

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{n} n_i \left(\overline{R}_i - \frac{N+1}{2}\right)^2 . \quad (6.12)$$

Эта величина называется статистикой *Краскела-Уоллеса*

Множитель $\frac{12}{N(N+1)}$ присутствует в качестве нормировочного для обеспечения сходимости распределения статистики H к χ^2 с числом степеней свободы $v = k-1$. Если $H > \chi_q^2(v)$, то фактор считается значимым и гипотеза H_0 отвергается при уровне значимости q.

Если среди x_{ji} есть совпадающие значения, то при ранжировании и переходе к r_{ji} надо использовать средние ранги (например, если 2 значения (5 и 5) занимают ранги 11, 12, то средний ранг (11,5) надо присвоить им обоим). Если совпадений много, рекомендуется использовать модифицированную форму статистики H':

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{m} T_j /(N^3 - N)\right)}, \quad (6.13)$$

где m — число групп совпадающих наблюдений; $T_j = r_j^3 - t_j$ (t_j — число совпадающих наблюдений в группе j).

Однофакторный непараметрический анализ на основе критерия Джонкхиера (альтернативы с упорядочением)

Нередко исследователю заранее известно, что имеющиеся группы результатов упорядочены по возрастанию влияния фактора. Пусть первый столбец таблицы $\{x_{ji}\}$ соответствует наименьшему уровню, а последний — наибольшему. В таких случаях критерий Джонкхиера более чувствителен (более мощный) в сравнении с упорядоченным влиянием фактора.

Рассмотрим сначала случай, когда сравниваются только 2 способа обработки (2 уровня фактора). Фактически речь идет тогда об однородности двух выборок. Для проверки этой гипотезы рассмотрим статистику Манна-Уитни.

Пусть имеем 2 выборки: $x_1, x_2, ..., x_m$ и $y_1, y_2, ..., y_n$. Положим

$$\varphi(x_i, y_j) = \begin{cases} 0, & если\ x_i > y_j; \\ \dfrac{1}{2}, & если\ x_i = y_j; \\ 1, & если\ x_i < y_j. \end{cases} \qquad (6.14)$$

Статистика Манна-Уитни:

$$U = \sum_{i=1}^{m} \sum_{i=1}^{n} \varphi(x_i, y_j). \qquad (6.15)$$

Обратившись теперь к общему случаю, когда сравниваются k способов обработки (k уровней), поступим следующим способом. Для каждой пары уровней u и v, где $1 \le u < v \le k$, составим по выборкам с номерами u и v статистики Манна-Уитни:

$$U(u,v) = \sum_{i=1}^{m} \sum_{i=1}^{n} \varphi(x_i, y_j). \qquad (6.16)$$

Получим $U(1,2), U(1,3), ..., U(1,k), U(2,3), ..., U(2,k), ..., U(k-1,k)$.

Определим статистику Джонкхиера I как $I = \sum U(u,v)$ для $1 \le u < v \le k$. Свидетельством против H_0 (в пользу альтернативы H_1) служат большие значения статистики I, полученные в эксперименте. Для больших объемов выборок в отношении статистики I действует нормальное распределение $I \sim N(MI, DI)$, с математическим ожиданием и дисперсией

$$MI = \frac{1}{4}\left(N^2 - \sum_{i=1}^{k} n_j \right),$$

$$DI = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3) \right], \qquad (6.17)$$

где n_j – количество наблюдений в каждом уровне;

N — общий объем наблюдений;

Рассмотрим нормированную статистику $I^* = \frac{I - MI}{\sqrt{DI}}$. Тогда, свидетельством против H_0 (в пользу альтернативы H_1) служат большие значения статистики $I^* = \frac{I - MI}{\sqrt{DI}}$, полученные в эксперименте, в сравнении с α-процентными точками нормального распределения $\Phi(I^*) = \alpha$ (табличные значения нормированной функции Лапласа). Тогда $q - 1 - \alpha$ — уровень значимости принятия гипотезы H_0.

6.3. Двухфакторный анализ

Иногда в однофакторной модели влияние интересующего нас фактора не проявляется, хотя такое влияние должно быть. Причиной этого может быть большой внутригрупповой разброс, на фоне которого действие фактора остаётся незаметным. Очень часто этот разброс вызван не только случайными причинами, но и действиями еще одного фактора. Если мы в состоянии указать такой фактор, то можно попытаться включить его в модель, чтобы уменьшить статистическую неоднородность наблюдений. Конечно, не всегда удается поправить дело введением *мешающего* фактора и переходом к двухфакторной схеме. Иногда приходится рассматривать трехфакторные и более сложные модели. Замысел во всех этих случаях остается прежним.

Назовём фактор A (рис. 1) главным, B — мешающим. Пусть фактор A принимает k значений, а мешающий — n значений. Фактор B разбивает все группы наблюдений (столбцы таблицы $\{x_{ji}\}$) на блоки.

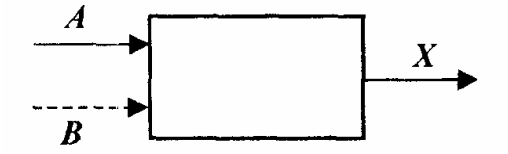

Рис. 6.1. Двухфакторная модель

Каждый блок соответствует определенному уровню фактора B. В частном случае таблица содержит $n \times k$ наблюдений (по одному в клетке). Отличие этой таблицы от однофакторной в том, что наблюдения в любом столбце не являются однородными, если влияние мешающего фактора значимо (табл. 6.3).

Таблица 6.3
Форма представления экспериментальных данных двухфакторной модели

Блоки фактора B	Уровни основного фактора A			
	A_1	A_2	A_i	A_k
B_1	x_{11}	x_{12}	x_{1i}	x_{1k}
B_2	x_{21}	x_{22}	x_{2i}	x_{2k}

….	….	….	….	….
B_j	x_{j1}	x_{j2}	x_{ji}	x_{jk}
….	….	…..	…..	…..
B_n	x_{n1}	x_{n2}	x_{ni}	x_{nk}

Для описания двухфакторного эксперимента обычно применяют аддитивную модель. Она предполагает, что значения отклика x_{ji} являются суммой вкладов соответствующих уровней факторов A и B и независимых случайных факторов: $x_{ji} = a_i + b_j + e_{ji}$. В этой модели величины вкладов A и B не могут быть восстановлены однозначно. Действительно, при одновременном увеличении всех a_i на одну и ту же константу и при уменьшении всех b_j на ту же константу значения x_{ji} не изменяются.

Для однозначности вкладов удобно перейти к представлению в виде:

$$x_{ji} = \eta + \alpha_i + \beta_j, \ при \ \sum_i \alpha_i = 0, \ \sum_j \beta_j = 0.$$

Параметр η интерпретируется как среднее всех x_{ji} (т.е. $\bar{x}$), а α_i и β_j — отклонения от η в результате действия факторов A и B.

6.3.1 Двухфакторный параметрический дисперсионный анализ

Если есть основания предполагать, что случайные величины e_{ji} имеют нормальное распределение с нулевым средним и одинаковой при всех i и j дисперсией σ^2, можно использовать метод, аналогичный однофакторному дисперсионному анализу. Предположим, что влияние фактора A отсутствует. Сформулируем гипотезу $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$. Проверим гипотезу H_0, также основываясь на сравнении двух независимых оценок дисперсии: σ_A^2 и $\sigma_{сл}^2$.

Дисперсионный анализ для двухфакторных таблиц проводится в следующей последовательности. Вычисляются суммы (полагаем $n_1 = n_2 = ... = n_k = n$)

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}^2 \ ; \ Q_2 = \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ji}\right)^2 \ ; \ Q_3 = \frac{1}{k}\sum_{j=1}^{n}\left(\sum_{i=1}^{k} x_{ji}\right)^2 \ ; \ Q_4 = \frac{1}{nk}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ji}\right)^2 .$$

Далее находятся оценки дисперсий

$$s_{сл}^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(n-1)} \ ; \ s_A^2 = \frac{Q_2 - Q_4}{k-1} \ ; \ s_B^2 = \frac{Q_3 - Q_4}{k-1} .$$

Если выполняется неравенство

$\frac{s_A^2}{s_{сл}^2} > F_q\left(v_1 = k-1, v_2 = (n-1)(k-1)\right)$, то H_0 отвергается и влияние фактора A

считается существенным.

Аналогично значимым признается влияние фактора B , если

$$\frac{s_B^2}{s_{сл}^2} > F_q\left(v_1 = n-1,\, v_2 = (n-1)(k-1)\right).$$

6.3.2. Двухфакторный непараметрический анализ

Двухфакторный непараметрический анализ используется при проверке гипотезы H_0, если о распределении случайной величины e_{ji} известно только то, что она непрерывна и независима. Рассмотрим решение задачи с использованием критерия Фридмана, который не предъявляет требований к упорядочению уровней факторов.

Двухфакторный непараметрический анализ по критерию Фридмана (произвольные альтернативы)

Критерий основан на идее перехода от значений x_{ji} к их рангам r_{ji}. В отличие от однофакторного анализа ранжирование происходит не по всей таблице $\{x_{ji}\}$, а по блокам, т.е. рассматривается каждая отдельная строка таблицы. При фиксированном j осуществляется ранжирование величин x_{ji} при $i = 1,2,...,k$. Тем самым устраняется влияние мешающего фактора B, значение которого для каждой строки постоянно. Обозначим полученные ранги величин x_{ji} через r_{ji}. Ясно, что r_{ji}. изменяются от 1 до k, а каждая строка представляет перестановку чисел $1,2,...,k$ (при совпадении x_{ji} надо использовать средние ранги). При гипотезе $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$ все $k!$ перестановок равновероятны. Введем величину $\bar{r}_i = \frac{1}{n}\sum_{j=1}^{n} r_{ji}$ — среднее значение ранга по столбцу. При H_0 значение для каждого столбца не должно сильно отличаться от $\bar{r} = \frac{k+1}{2}$ — среднего ранга всех элементов таблицы.

Статистика *Фридмана* имеет следующий вид

$$S = \frac{12n}{k(k+1)}\sum_{i=1}^{k}\left(\bar{r}_i - \bar{r}\right)^2. \qquad (6.18)$$

Гипотеза H_0 отвергается в пользу альтернативы о наличии эффектов в обработке, если $S \geq \chi_q^2(v = k-1)$. Для небольших значений n, k величина критерия Фридмана $S(q,n,k)$ может быть найдена по специальным статистическим таблицам.

Двухфакторный непараметрический анализ по критерию Пейджа (альтернативы с упорядочением)

Если уровни факторов в таблице $\{x_{ji}\}$ упорядочены, то для проверки гипотезы $H_0: \alpha_1 = \alpha_2 = ... = \alpha_k$ против альтернативы $H_1: \alpha_1 < \alpha_2 < ... < \alpha_k$ используется статистика *Пейджа.*

Введем величину $R_i = \sum_{j=1}^{n} r_{ji}$.

Вычисляется статистика

$$L = \sum_{i=1}^{k} i \cdot R_i = 1 \cdot R_1 + 2 \cdot R_2 + 3 \cdot R_3 + ... + k \cdot R_k . \quad (6.19)$$

H_0 отклоняется в пользу H_1, если $L_{набл} > L_{\alpha}(k,n)$, где значение $L_{\alpha}(k,n)$ найдено по специальным статистическим таблицам.

Критические значения $L_{\alpha}(k,n)$ приведены в табл. 9 (см. статистические таблицы). Для $n > 10$ справедлива аппроксимация статистики Пеиджа

$$L^* = \frac{L - M(L)}{\sqrt{D(L)}}, \text{ где } M(L) = \frac{nk(k+1)^2}{4}, \; D(L) = \frac{n(k^3 - k)^2}{144(k-1)}.$$

При $L^* > u_{\alpha}$ нулевая гипотеза отклоняется (u_{α} – квантиль стандартного нормального распределения.

Вопросы для самоконтроля

1. Что такое дисперсионный анализ зависимостей. Приведите основные понятия.
2. Однофакторный параметрический дисперсионный анализ.
3. Однофакторный непараметрический анализ на основе критерия Краскела-Уоллеса (произвольные альтернативы).
4. Однофакторный непараметрический анализ на основе критерия Джонкхиера (альтернативы с упорядочением)
5. Двухфакторный дисперсионный анализ. Двухфакторный параметрический дисперсионный анализ.
6. Двухфакторный непараметрический анализ по критерию Фридмана (произвольные альтернативы). Двухфакторный непараметрический анализ по критерию Пейджа (альтернативы с упорядочением).

Тема 7. Корреляционный анализ

Корреляционный анализ имеет своей задачей количественное определение тесноты связи между изучаемыми явлениями. Корреляционная связь — эта связь, при которой определенному значению факторного признака соответствует среднее значение результирующего признака. Изучение тесноты и направления связи является важной задачей математической статистики. Оценка тесноты связи между признаками предполагает определение меры соответствия вариации результирующего признака от одного (при изучении парных зависимостей) или нескольких (множественных) факторных признаков.

7.1. Вычисление параметрических коэффициентов корреляции

Выборочный коэффициент корреляции между двумя случайными величинами x и y был впервые введен Пирсоном, поэтому его часто *называют коэффициентом корреляции Пирсона.* В теории разработаны и на практике применяются различные модификации формулы расчета данного коэффициента. Приведем некоторые из них:

$$r = \frac{\overline{(x-\bar{x})(y-\bar{y})}}{s_x s_y} = \frac{\overline{xy} - \bar{x}\bar{y}}{s_x s_y}.$$

Используя преобразование, получают

$$r = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{n \cdot s_x s_y} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}. \qquad (7.1)$$

При $n<15$ необходимо скорректировать коэффициент корреляции

$$r^* = r\left(1+\frac{1-r^2}{2(n-3)}\right). \qquad (7.2)$$

При большом объеме выборки выборочный коэффициент корреляции будет приближаться к корреляционному моменту генеральной совокупности ρ, который определяется как

$$\rho = \frac{M\left[(x-m_x)(y-m_y)\right]}{\sigma_x \sigma_y}. \qquad (7.3)$$

Если $r=0$, то величины x и y независимы, а при $r=1$ зависимость между x и y является функциональной. Коэффициент корреляции тесно связан с коэффициентом регрессии:

$$r = b\frac{s_x}{s_y}, \qquad (7.4)$$

где b — коэффициент регрессии в уравнении вида $y_i = a + bx_i$;

s_x — среднеквадратичное отклонение x;

s_y — среднеквадратичное отклонение y. Значимость коэффициента корреляции

проверяется на основе критерия Стьюдента. При проверке этой гипотезы вычисляется t-статистика:

$$t_{pac} = \sqrt{\frac{r^2(n-2)}{1-r^2}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}\,. \tag{7.5}$$

Расчетное значение сравнивается с табличным значением $t_q(v = n-2)$. Если расчетное значение больше табличного, это свидетельствует о значимости коэффициента корреляции, а, следовательно, и о статистической существенности зависимости между x и y.

При большом числе наблюдений ($n > 100$) используется следующая формула t-статистики:

$$t_{pac} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n}\,. \tag{7.6}$$

Множественный коэффициент корреляции рассчитывается при наличии линейной связи между результирующим признаком Y и несколькими факторными признаками, а также между парой факторных признаков.

В случае оценки связи между результирующим признаком Y и двумя факторными признаками x_1 и x_2 множественный коэффициент корреляции можно определить по формуле:

$$R_{y/x_1,x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r_{x_1x_2}^2}}\,, \tag{7.7}$$

где r_{ab} — парные коэффициенты корреляции между признаками.

В общем случае коэффициент множественной корреляции между результирующим признаком Y и m факторными признаками $x_1, x_1, ..., x_m$ определяется по формуле

$$R_{y/x_1,x_2,...,x_m} = \sqrt{1-\frac{|\rho|}{|\rho_1|}}\,, \tag{7.8}$$

где $|\rho|$ – определитель матрицы парной корреляции

$$\rho = \begin{pmatrix} 1 & \rho_{yx_1} & \rho_{yx_2} & ... & \rho_{yx_m} \\ \rho_{x_1y} & 1 & \rho_{x_1x_2} & ... & \rho_{x_1x_m} \\ \rho_{x_2y} & \rho_{x_2x_1} & 1 & ... & \rho_{x_2x_m} \\ ... & ... & ... & ... & ... \\ \rho_{x_my} & \rho_{x_mx_2} & \rho_{x_mx_3} & ... & 1 \end{pmatrix}; \tag{7.9}$$

$|\rho_1|$ – алгебраическое дополнение элемента ρ_{11}.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1. Приближение R к единице свидетельствует о сильной зависимости между

признаками. При небольшом числе наблюдений ($\frac{n-k}{k}<20$) величина множественного коэффициента корреляции, как правило, завышается. В этом случае множественный коэффициент корреляции корректируется:

$$\widetilde{R}_{y/x_1,x_2,...,x_m}=\sqrt{1-(1-R^2)\frac{n-1}{n-k-1}}, \quad (7.10)$$

где $\widetilde{R}$ — скорректированное значение;
n — число наблюдений;
k — число факторных признаков.

Корректировка R не производится при условии, если $\frac{n-k}{k}\geq 20$.

Проверка значимости множественного коэффициента корреляции осуществляется по критерию Фишера:

$$F_{pac}=\frac{\frac{1}{2}R^2_{y/x_1,x_2}}{\frac{1}{3}\left(1-R^2_{y/x_1,x_2}\right)}. \quad (7.11)$$

Множественный коэффициент корреляции считается значимым, если $F_{pac}>F_q(v_1=2,\ v_2=n-3)$.

На основе приведенных выше формул (7.1)-(7.9) произведем вычисление коэффициентов корреляции и проверим их значимость (вынести на практику).

7.2 Вычисление непараметрических коэффициентов корреляции

Если признаки подчиняются различным (отличным от нормального) законам распределения, то можно рассчитать непараметрические коэффициенты корреляции. Эти коэффициенты могут быть использованы для определения тесноты связи как между количественными, так и между качественными признаками при условии, если их значения упорядочить или проранжировать по степени убывания или возрастания признака.

Среди непараметрических методов оценки тесноты связи наибольшее распространение получили следующие:

1. *Коэффициент ранговой корреляции Спирмана:*

$$\rho_{x/y}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}, \quad (7.12)$$

где d_i^2 – квадрат разности рангов;
n — число наблюдений (число пар рангов).

Коэффициент Спирмана принимает значения от -1 до 1. Значимость коэффициента проверяется на основе t-критерия Стьюдента. Расчетное значение t-критерия определяется по формуле:

$$t_{рас} = \rho_{x/y}\sqrt{\frac{n-2}{1-\rho^2_{x/y}}}.$$

Значение коэффициента корреляции считается статистически существенным, если $t_{рас} > t_{\alpha}(v = n-2)$.

2. Коэффициент ранговой корреляции Кендалла

$$\tau = \frac{2(P-Q)}{n(n-1)} = \frac{2S}{n(n-1)}, \qquad (7.13)$$

где n — число наблюдений;

S — разность между числом последовательностей и числом инверсий по второму признаку. Расчет коэффициента Кендалла производится в такой последовательности:

- значения X ранжируются в порядке возрастания или убывания;
- значения Y располагаются в порядке, соответствующем значениям X;
- для каждого ранга Y определяется число следующих за ним значений рангов, превышающих его по величине. Суммируя эти числа, определяем величину P (число последовательностей) — меру соответствия последовательностей рангов X и Y;
- для каждого ранга Y определяется число следующих за ним рангов, меньших его величины. Суммируя величины, получаем величину Q (число инверсий);
- определяется разность по всем членам ряда $S = P - Q$ и вычисляется τ.

Связь между признаками можно признать статистически значимой, если значение коэффициента корреляции $|\tau| > \tau_{\alpha} = u_{\alpha}\sqrt{\frac{2(2n+5)}{9n(n-1)}}$ (на практике часто полагают связь между признаками значимой, если $\tau > 0{,}5$).

3. *Коэффициент конкордации.* Для определения тесноты связи между несколькими ранжированными признаками применяют *множественный коэффициент ранговой корреляции — коэффициент конкордации,* который вычисляется по формуле:

$$W = \frac{12S_W}{m^2(n^3-n)}, \qquad (7.14)$$

где m —- количество факторов;

n — число наблюдений;

S_W — отклонение суммы квадратов рангов от среднего квадрата суммы рангов

$$S_W = \left(\widetilde{R^2} - \frac{\widehat{R}^2}{n}\right), \qquad (7.15)$$

где $\widetilde{R} = \sum_{j=1}^{10} R_j$, $\widetilde{R^2} = \sum_{j=1}^{10} R_j^2$.

Критическое значение коэффициента конкордации равно $W_{\alpha} = \frac{1}{m(n-1)}\chi^2_{\alpha}(n-1)$.

Если $W > W_\alpha$, то с вероятностью α корреляция между изучаемыми признаками признается значимой.

Если среди последовательностей рангов есть совпадения, то коэффициент конкордации следует вычислять по формуле

$$W = \frac{12S_W}{m^2(n^2-1) - m\sum_{j=1}^{m} T_j}, \qquad (7.16)$$

где $T_j = t_j^3 - t_j$, t_j – количество совпавших рангов в j-й последовательности.

Совпавшим рангам присваиваются средние ранги.

На основе приведенных выше формул (7.10) – (7.12) произведем вычисление непараметрических коэффициентов корреляции и проверим их значимости. Для непараметрических методов необходимы дополнительные преобразования, которые покажем на примерах (вынесено на практику).

Вопросы для самоконтроля

1. Для чего используется корреляционный анализ?
2. Как вычисляются параметрические коэффициенты корреляции?
3. Как определяется значимость коэффициента корреляции?
4. Что такое множественный коэффициент корреляции?
5. Как проверяется значимость множественного коэффициента корреляции?
6. Вычисление непараметрических коэффициентов корреляции. Коэффициент ранговой корреляции Спирмана.
7. Коэффициент ранговой корреляции Кендалла.
8. Коэффициент конкордации.

Тема 8. Регрессионный анализ

Рассмотренные ранее методы дисперсионного и корреляционного анализа позволяют выявить наличие связи между случайными величинами и оценить силу этой связи.

Следующей ступенью является выявление конкретного функционального вида связи между случайными величинами.

При наличии корреляционной связи между x и y имеет место соотношение $F(y)=F(x,y)$ т.е. функция распределения случайной величины y зависит от значения случайной величины x. Любая функция распределения полностью определяется своими параметрами. Изменение функции распределения случайной величины y от x можно задать зависимостями

$$\mu_{1y}=f(x;\beta);\ \mu_{2y}=f_2(x;\beta);\ \mu_{3y}=f_3(x;\beta);\ \mu_{4y}=f_4(x;\beta),$$

называемыми соответственно – *регрессионной, скедастической, клитической и синагической*. Здесь $\beta=(\beta_1,...,\beta_k)$ – параметры модели. На практике обычно предполагается, что дисперсия и моменты высших порядков распределения y не зависят от значения x. Наибольший практический интерес представляет определение зависимости $\mu_{1y}=f(x;\beta)$, описывающей истинную зависимость между y и x. Зависимость средних значений $\mu_{1y}=f(x;\beta)$ называется регрессией y по x, а методы нахождения таких зависимостей и оценки их статистических свойств составляют содержание регрессионного анализа.

По выборочным данным можно найти только оценку истинной регрессии, содержащую ошибку, связанную со случайностью выборки.

В основе регрессионного анализа лежит принцип наименьших квадратов, в соответствии с которым в качестве уравнения регрессии

$$y=f(x;\beta)+\varepsilon,$$

выбирается функция, доставляющая минимум сумме квадратов разностей $s=\sum_{i=1}^{n}\left(y_i-f(x_i;\beta)\right)^2$. Здесь ε – случайная погрешность уравнения. Как правило, вид функции $f(x;\beta)$ определяется заранее, а методом наименьших квадратов определяются ее коэффициенты β_j, минимизирующие s. Количественной мерой рассеяния значений y_i вокруг регрессии $f(x;\beta)$ является дисперсия

$$D=\frac{1}{n-k}\sum_{i=1}^{n}\left(y_i-f(x_i;\beta)\right)^2,$$

где k – число коэффициентов, входящих в аналитическое выражение регрессии (например, если $f(x;\beta)$ – многочлен степени m, то $k=m+1$).

В зависимости от вида уравнения регрессии $y=f(x;\beta)+\varepsilon$ различают линейную ($f(x;\beta)$ – многочлен первой степени) и нелинейную ($f(x;\beta)$ – многочлен степени ≥ 2) регрессии.

Вид функции $f(x;\beta)$ выбирается исходя из особенностей исследуемого явления (процесса), а также из общего графического анализа зависимости между y и x.

Чаще всего ограничиваются рассмотрением линейной регрессионной модели, а

при нелинейной зависимости $y = f(x;\beta) + \varepsilon$ используют различные линеаризующие преобразования переменных y и x. Наиболее распространенные из этих преобразований приведены в табл. 8.1 (для случая двух параметров).

Таблица 8.1

Линеаризующие функциональные преобразования

$(y^* = a^* + b^* x^*)$

Исходная зависимость $y = f(x;\beta)$	Преобразование переменных		Преобразование коэффициентов	
	y^*	x^*	a^*	b^*
$y = a + \frac{b}{x}$	y	$\frac{1}{x}$	a	b
$y = \frac{a}{b+x}$	$\frac{1}{y}$	x	$\frac{b}{a}$	$\frac{1}{a}$
$y = \frac{ax}{b+x}$	$\frac{1}{y}$	$\frac{1}{x}$	$\frac{b}{a}$	$\frac{1}{a}$
$y = \frac{x}{a+bx}$	$\frac{x}{y}$	x	a	b
$y = ab^x$	$\lg y$	x	$\lg a$	$\lg b$
$y = ax^b$	$\lg y$	$\lg x$	$\lg a$	b
$y = ae^{bx}$	$\ln y$	x	$\ln a$	b
$y = ae^{b/x}$	$\ln y$	$\frac{1}{x}$	$\ln a$	b
$y = a + bx^n$	y	x^n	a	b

Схема регрессионного анализа включает в себя последовательное решение следующих задач: нахождение выборочной оценки истинной регрессии (т.е. нахождение коэффициентов регрессии); оценки статистической значимости выборочной регрессии в сравнении с безусловным разбросом значений y_i, характеризующимся дисперсией σ_y^2; определение доверительных областей, с заданной вероятностью включающих в себя истинную регрессию.

Среди дополнительных задач, позволяющих получить полную статистическую картину изучаемой регрессии, отметим: анализ так называемых регрессионных остатков (разница между выборочной регрессией и выборочными значениями функции); анализ наличия грубых отклонений от регрессии (выбросов); построение толерантных границ для регрессии.

Разработанный в настоящее время аппарат регрессионного анализа предполагает, что значения y_i взаимно независимы и нормально распределены. Выполнение этих условий должно быть предварительно проверено с помощью критериев нормальности (см. раздел 3.3) и критериев сравнения дисперсий (см.

раздел 4.3).

8.1. Построение модели регрессии

Рассмотрим линейную по коэффициентам модель регрессии:

$$y = f\left(x;\beta\right)+\varepsilon=\beta_0+\beta_1 f_1(x)+\beta_2 f_2(x)+\ldots+\beta_k f_k(x)+\varepsilon, \quad (8.1)$$

где ε - случайная величина с математическим ожиданием равным нулю и дисперсией σ^2; $f_j(x),\ j=1,...,k$ – некоторые заданные функции.

Полагая, $x_j = f_j(x), \quad j=\overline{1,k}$ перейдем к модели множественной линейной регрессии:

$$y = f\left(x;\beta\right)+\varepsilon=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k+\varepsilon. \quad (8.2)$$

Пусть для оценки неизвестных параметров $\beta_j,\ j=\overline{0,k}$ уравнения регрессии (2) взята выборка объемом n из значений величин $(Y,X_1,X_2,\ldots,X_k)$. Тогда

$$Y = X\beta+\varepsilon,$$

где $Y=\left(\ y_1, y_2 ..., y_n\right)^T$ - вектор значений переменной y;

$\beta=\left(\beta_0,\beta_1,\ldots,\beta_k\right)^T$ - вектор параметров модели;

$\varepsilon=\left(\varepsilon_1,\ldots,\varepsilon_n\right)^T$ – вектор ошибок, где $\varepsilon_i\in \mathrm{N}\left(0,\sigma^2\right)$ и независимы;

X - матрица исходных данных переменных X_j размерами $n\times(k+1)$. Первый столбец матрицы X содержит единицы (значения фиктивной переменной x_0), остальные столбцы значения переменных $x_1, x_2, ..., x_k$:

$$X=\begin{pmatrix} 1 & x_1^1 & \cdots & x_k^1 \\ 1 & x_1^2 & \cdots & x_k^2 \\ & & \cdots & \\ 1 & x_1^n & \cdots & x_k^n \end{pmatrix}.$$

Для нахождения оценки β^* вектора параметров $\beta=\left(\beta_0,\beta_1,\ldots,\beta_k\right)^T$ используем метод наименьших квадратов, согласно которому в качестве оценок $\beta_0^*, \beta_1^*,\ldots, \beta_k^*$ берутся такие, которые минимизируют сумму квадратов Q отклонений значений y_i от $f(\vec{x}_i)$:

$$Q=\sum_{i=1}^{n}\left(y_i-f(\vec{x}_i)\right)^2=\sum_{i=1}^{n}\varepsilon_i^2=\varepsilon^T\varepsilon=\left(Y-X\beta\right)^T\left(Y-X\beta\right). \quad (8.3)$$

Оценка β^* метода наименьших квадратов имеет вид:

$$\beta^*=\left(X^T X\right)^{-1}X^T Y. \quad (8.4)$$

8.2. Оценка адекватности регрессии

Количественной мерой адекватности является отношение дисперсии S^2, определяемой рассеянием значений y_i вокруг линии регрессии, к дисперсии S_y^2

естественного рассеяния значений y_i вокруг своего среднего $\overline{y}$. Другими словами это можно сформулировать так: ошибки, обусловленные заменой истинной зависимости на выборочную регрессию, находятся на уровне естественного разброса, наблюдаемых случайных величин.

Оценим сначала величину дисперсии модели S^2:

$$S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\varepsilon^2 = \frac{1}{n-k-1}\varepsilon^T\varepsilon,$$

где $\hat{y}_i = \beta_0^* + \beta_1^* x_i + \ldots \beta_k^* x_k$.

Оценка дисперсии разброса случайных чисел y_i вокруг своего среднего значения равна $S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$.

Если $\frac{S^2}{S_y^2} > F_\alpha(v_1 = n-k-1,\, v_2 = n-1)$, где $F_\alpha(v_1, v_2)$ – квантиль распределения Фишера с v_1 и v_2 степенями свободы, то ошибка в определении регрессии с доверительной вероятностью α признается статистически значимой, а модель неадекватной. Если $\frac{S^2}{S_y^2} < F_\alpha(v_1, v_2)$, то модель можно признать адекватной.

Примечание. *В пакете Excel используется статистика* $F = \dfrac{\frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. *Если* $F > F_\alpha(v_1 = n-k-1,\, v_2 = k)$, *то модель признается адекватной (т.е. отбрасывается гипотеза о том, что коэффициенты модели* $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ *равны нулю) (см. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Том 1. – М.: Финансы и статистика, 1986. – 366с. [3], на стр.53).*

Качество модели также можно оценить с использованием оценки коэффициента детерминации: $R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$. Чем ближе значения R^2 к 1, тем большую долю дисперсии величины Y объясняет модель регрессии.

8.2.1 Анализ регрессионных остатков

Определенную информацию об адекватности уравнения регрессии дает исследование остатков вида $\varepsilon_i = \hat{y}_i - y_i$. Если выборочная регрессия $\hat{y}$ удовлетворительно описывает истинную зависимость между y и x, остатки ε_i

должны быть независимыми нормально распределенными случайными величинами с нулевым средним и в значениях e_i должен отсутствовать тренд. Нормальность распределения остатков ε_i может быть установлена одним из критериев согласия (см. раздел 3.3).

Гипотезу о равенстве $M(\varepsilon)=0$ можно проверить любым параметрическим или непараметрическим критерием сравнения среднего с заданным значением (в нашем случае с нулем, см. раздел 4.4).

Независимость в последовательности значений ε_i $(i=1,...,n)$ может быть проверена с помощью сериального коэффициента корреляции Дарбина-Ватсона. Статистика сериального коэффициента корреляции Дарбина-Ватсона имеет вид

$$D=\frac{\sum_{i=2}^{n}\left(\varepsilon_i-\varepsilon_{i-1}\right)^2}{\sum_{i=1}^{n}\varepsilon_i^2}.$$

Если $D>D_1(\alpha)$ или $D>4-D_1(\alpha)$, то с достоверностью α принимается гипотеза о наличии соответственно отрицательной или положительной корреляции остатков.
Если $D_2(\alpha)>D>D_1(\alpha)$ или $4-D_1(\alpha)>D>4-D_2(\alpha)$, то критерий не позволяет принять решение по гипотезе о наличии или отсутствии корреляции остатков. Если $D_2(\alpha)<D<4-D_2(\alpha)$, то гипотеза корреляции остатков отклоняется. Критические значения $D_1(\alpha)$ и $D_2(\alpha)$ для различных α и числа k коэффициентов в регрессии приведены в табл. 11 (см. статистические таблицы).

8.2.2 Доверительный интервал для уравнения регрессии

Доверительный интервал для условного среднего $\hat{y}=M(Y\mid X=x)$ в многомерной точке $X_0=(1,x_1^0,\ldots,x_k^0)^T$ определяется по формуле:
$\left[\left(X_0^T\beta^*\right)\pm t_{1-\alpha/2}S\sqrt{\left(X_0^T\left[(X^TX)^{-1}\right]X_0\right)}\right]$, где t_α – квантиль распределение Стьюдента с $n-k-1$ степенью свободы. Соответственно доверительный интервал для значений y в точке $X_0=(1,x_1^0,\ldots,x_k^0)^T$ будет иметь вид:
$\left[\left(X_0^T\beta^*\right)\pm t_{1-\alpha/2}S\left(1+\sqrt{X_0^T\left[(X^TX)^{-1}\right]X_0}\right)\right]$, так как погрешность модели $y=f(x)+\varepsilon$ будет определяться двумя источниками: погрешностью $(\Delta f)^2=S^2\left(X_0^T\left[(X^TX)^{-1}\right]X_0\right)$, связанной с погрешностями параметров модели, и погрешностью собственно модели $\varepsilon^2=S^2$.

8.3. Оценка дисперсии коэффициентов регрессии и доверительных интервалов

Оценка дисперсии коэффициента β_j находится по формуле: $s_j^2=S^2\left[(X^TX)^{-1}\right]_{jj}$, где $\left[(X^TX)^{-1}\right]_{jj}$ соответствующий диагональный элемент матрицы $(X^TX)^{-1}$.

Доверительные интервал для σ^2 находится с использованием статистики $\chi^2 = (n-k-1)S^2/\sigma^2$, которая при нормальном распределении ε_i имеет распределение хи-квадрат с $(n-k-1)$ степенью свободы.

Для проверки значимости коэффициентов уравнения регрессии используем статистику $t_j = \dfrac{\beta_j^*}{\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}}$, которая при истинности гипотезы $H_0: \beta_j = 0$, имеет распределение Стьюдента с $(n-k-1)$ степенью свободы. Если для заданного уровня значимости α значение $|t_j|$ больше критического $t_{крит} = t_{1-\alpha/2}(n-k-1)$, то нулевая гипотеза отвергается и коэффициент признается значимым. В противном случае коэффициент признается незначимым, и соответствующее слагаемое исключается из модели.

В пакете Excel рассчитывается также уровень значимости α статистики $|t_j|$, т.е. вероятность $P\left(x > \left|t_j\right|\right)$. Степень значимости параметров распределения качественно определяется по уровню значимости: не значимые ($\alpha \geq 0{,}100$), слабо значимые ($0{,}100 > \alpha \geq 0{,}050$), статистически значимые ($0{,}050 > \alpha \geq 0{,}010$), сильно значимые ($0{,}010 > \alpha \geq 0{,}001$), высоко значимые ($0{,}001 > \alpha$).

Для уровня значимости α доверительный интервал рассчитывается по формуле $\beta_j^* \pm t_\alpha\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}$, где t_α – квантиль распределение Стьюдента с $n-k-1$ степенью свободы.

8.4 Пример построения уравнения регрессии

Имеется выборка значений совместно наблюдаемых величин X и Y (см. табл. 8.2).

Таблица 8.2

X	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
Y	2,96	0,61	4,63	2,44	2,23	4,89	4,98	3,89	6,74	8,07
X	5,5	6	6,5	7	7,5	8	8,5	9	9,5	10
Y	8,34	9,56	9,30	12,35	11,46	11,09	7,91	8,16	6,54	7,88

Требуется подобрать подходящую модель регрессии, характеризующую зависимость Y от X, если известно, что ошибка $\sigma^2 = 1{,}3$.

Нанесем точки (X, Y) на координатную плоскость – построим корреляционное поле, соответствующее нашей выборке (рис. 8.1)

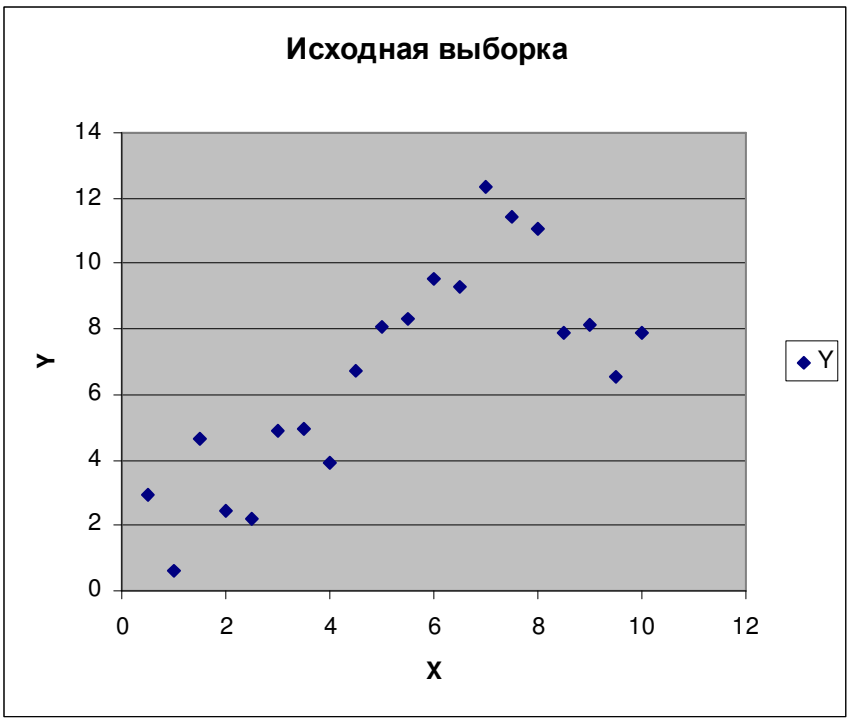

Рис. 8.1. Исходные данные

Видим, что существует зависимость, между значениями Y и X, причем зависимость явно нелинейная. Попробуем аппроксимировать эту зависимость для начала полиномами различных порядков. Возьмем в качестве уравнения регрессии квадратное уравнение:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Чтобы воспользоваться МНК для оценки коэффициентов, проведем линеаризацию модели, положив $x_1 = x,\ x_2 = x^2$, получим

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Тогда оценку вектора параметров, согласно МНК, найдем как

$$\beta^* = \left(X^T X\right)^{-1} X^T Y$$

Здесь X - матрица, первый столбец которой содержит единицы, а второй и последующий значения x_1 и x_2.

Для облегчения подбора модели можно воспользоваться встроенными функциями пакета EXCEL (для выбранной модели все равно потом потребуется провести все вычисления вручную, чтобы построить доверительные интервалы). В пакете анализа необходимо выбрать функцию “регрессия”, задать столбец значений Y и матрицу, соответствующую X (единичный столбец в этом случае задавать не надо). Если выбрать вывод остатков, то помимо регрессионной статистики, будут

выведены и предсказанные значения Y, т.е.

$$y^* = \beta_0^* + \beta_1^* x + \beta_2^* x^2$$

Для нашей модели, регрессионная статистика, полученная пакетом Fxcel будет иметь следующий вид:

Таблица 8.3.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,852379622
R-квадрат	0,72655102
Нормированный R-квадрат	0,694380552
Стандартная ошибка	1,820336831
Наблюдения	20

Таблица 8.4.

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	149,6702188	74,83510939	22,58750495	1,63336E-05
Остаток	17	56,32303623	3,313119778		
Итого	19	205,993255	10,841750	0,305589	

Таблица 8.5.

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-0,963028418	1,354297293	-0,711090853	0,486670901
Переменная X1	2,604940094	0,594036759	4,385149663	0,000403854
Переменная X2	-0,167559332	0,054953956	-3,049085893	0,007253372

(продолжение таблицы 8.3)

Нижние 95%	*Верхние 95%*
-3,82010793	1,894090386
1,351577249	3,858001699
-0,283477711	-0,051610008

Здесь в первой таблице 8.3:

- Множественный R – корень квадратный из коэффициента детерминации $\sqrt{R^2}$;

- R-квадрат – коэффициент детерминации $R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$;
- Нормированный R-квадрат – это скорректированная величина коэффициента детерминации, вычисляемая по формуле $\hat{R}^2 = 1 - (1 - R^2)\dfrac{n-1}{n-k-1}$;
- Стандартная ошибка – значение $S = \sqrt{S^2}$, где $S^2 = \dfrac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ - оценка дисперсии предсказания σ^2;
- Наблюдения – объем выборки n.

Во второй таблице 8.4:

- df – степени свободы v;
- SS – сумма квадратов разностей:

  1) между модельными значениями и средним

  $$SS_{регрессия} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = k \cdot S_R^2;$$

  2) остатки $SS_{остаток} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S^2 \cdot (n-k-1)$;

  3) между исходными данными и средним

  $$SS_{итого} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1) \cdot S_y^2;$$

- $MS = SS / df$;
- F – статистика $F_{Excel} = \dfrac{\frac{1}{k}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \dfrac{S_R^2}{S^2}$.

Если рассчитать $F = \dfrac{S^2}{S_y^2}$, то получим $F = \dfrac{3{,}313119}{10{,}841750} = 0{,}305589$.

- Значимость F – значение вероятности $F(x,k,n-k-1)$ при $x=\frac{S_R^2}{S^2}$, т.е. это уровень значимости принятия нулевой гипотезы H_0.

В третьей таблице 8.5:

- Коэффициенты – значения оценок коэффициентов $\beta_0^*, \beta_1^*, \beta_2^*$;
- Стандартная ошибка – значения оценок среднеквадратичных отклонений коэффициентов $s_j=\sqrt{S^2\cdot\left[(X^TX)^{-1}\right]_{jj}}$;
- t-статистика – наблюдаемые значения статистик критерия проверки значимости коэффициентов соответственно $t_j=\frac{\beta_j^*}{\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}}$;
- P-значения – достигнутые значения уровня значимости $P\left(x>\left|t_j\right|\right)$.
- Нижние и верхние границы 95%-го доверительного интервала $\beta_j^*\pm t_{0,05}(v)\sqrt{S^2\left[(X^TX)^{-1}\right]_{jj}}$.

Соответствующий график предсказанных значений в сравнении с исходными данными имеет вид (рис. 8.2):

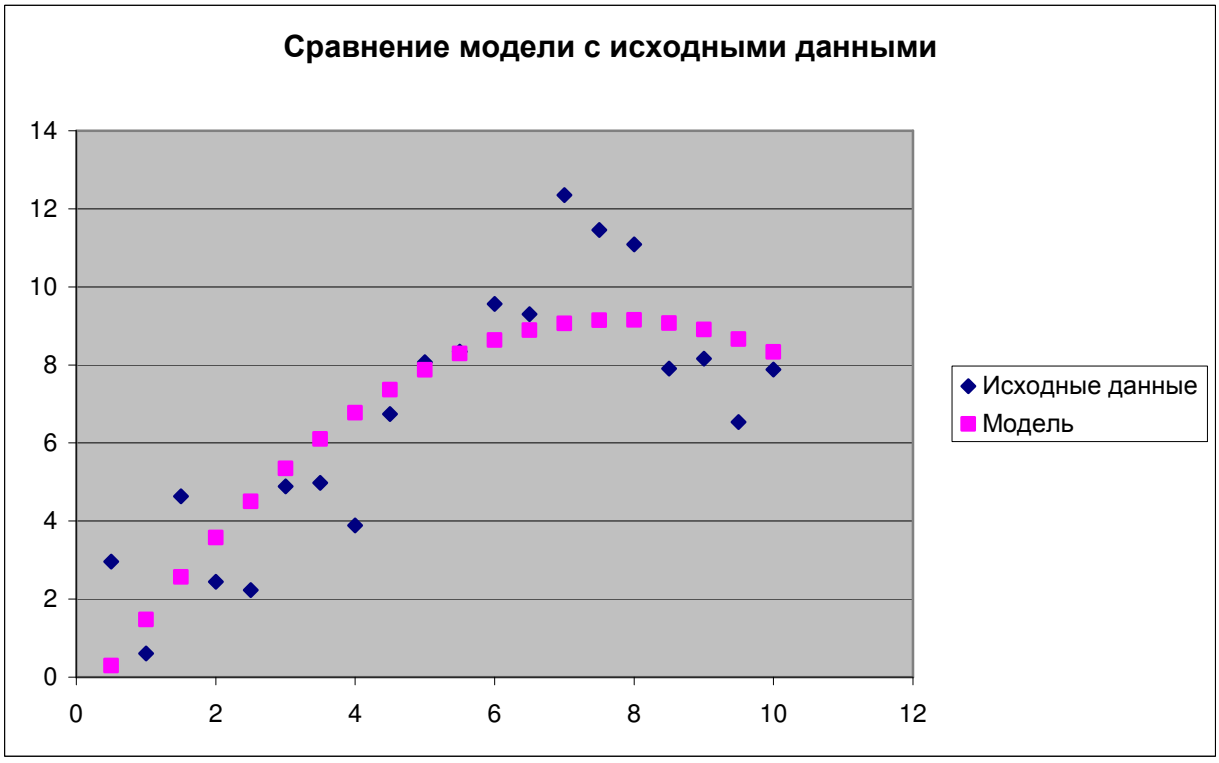

Рис. 8.2. Сравнение модели $y=\beta_0+\beta_1x+\beta_2x^2$ с исходными данными

Отметим, что полученная оценка значения $\sigma = \sqrt{1,3} = 1,14$ велика: $S = 1,82$. Что касается коэффициентов модели, то, кроме β_0, все они значимо отличаются от нуля (достигнутый уровень значимости достаточно мал, поэтому можно отвергнуть гипотезу о равенстве коэффициентов нулю).

Вопросы для самоконтроля

1. Что такое регрессионный анализ?
2. Регрессионная, скедастическая, клитическая и синагическая зависимости изменения функции распределения случайной величины y от x.
3. Построение модели регрессии.
4. Оценка адекватности регрессии. Доверительный интервал для уравнения регрессии.
5. Оценка дисперсии коэффициентов регрессии и доверительных интервалов.

Литература

- Кобзарь А.И. Прикладная математическая статистика. – М.: ФИЗМАТЛИТ, 2012. – 813с.
- Свешников А.А. Прикладные методы теории вероятностей. — Санкт-Петербург: Лань, 2012. – 480 с. [Электронный ресурс]. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_cid=25&pl1_id=3184
- Туганбаев А.А., Крупин В.Г. Теория вероятностей и математическая статистика — Санкт-Петербург: Лань, 2011. – 320 с. [Электронный ресурс]. – Режим доступа: http://e.lanbook.com/books/element.php?pl1_cid=25&pl1_id=652
- Белов А.А., Баллод Б.А., Елизарова Н.Н. Теория вероятностей и математическая статистика: учебник для вузов. – Ростов н/Д: Феникс, 2008. – 318 с. (3 экз. в библиотеке ТУСУР)