Министерство образования Российской Федерации

Международный образовательный консорциум «Открытое образование»

Московский государственный университет экономики, статистики и информатики

АНО «Евразийский открытый институт»

Трошин Л. И.

Математическая статистика

Учебно-практическое пособие

Москва, 2003

УДК 519.22
ББК 22.172
Т 766

Трошин Л.И. **Математическая статистика:** Учебно-практическое пособие / Московский университет экономики, статистики и информатики. – М., 2003. – с.144

ISBN 5–7764–0283–4

ПРЕДМЕТ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика занимается обработкой результатов случайного эксперимента. В отличие от теории вероятностей, в ней математическая модель эксперимента не известна (либо частично, либо полностью). Следовательно, возникает задача: по имеющимся наблюдаемым данным эксперимента (их всегда конечное, может быть и достаточно большое число) установить или восстановить неизвестное распределение вероятностей или объективно оценить параметры распределения.

Можно утверждать, что математическая статистика занимается задачами, обратными задачам теории вероятностей. Для пояснения сказанного рассмотрим следующий пример.

В контейнере содержится N деталей, среди которых M бракованных и, следовательно, N - M стандартных. Из контейнера извлекается без возвращения n деталей (бесповторный отбор). Если бы мы знали количество бракованных и стандартных деталей в контейнере, то по соответствующей модели можно было бы получить распределение вероятностей числа m бракованных деталей в выборке объема n (гипергеометрическое распределение). Рассмотренная задача относится к компетенции теории вероятностей.

На практике, в большинстве подобных задач выборок из генеральной совокупности элементов с альтернативным признаком, не известны параметры генеральной совокупности (N, M). Следовательно, здесь возникает обратная задача, в которой по анализу выборки требуется сделать заключение о составе генеральной совокупности, например о генеральной доле M/N бракованных деталей или о числе M бракованных деталей в контейнере при известном общем числе N деталей. Таким образом, имеем типичную задачу математической статистики.

Приведенный простой пример указывает на тесную связь между теорией вероятностей и математической статистикой, и одну из этих дисциплин можно считать частью другой.

Для отыскания подходящей вероятностной модели наблюдаемого случайного явления применяются математико-статистические методы оценки параметров и проверки гипотез. Рассмотрим пример, иллюстрирующий это положение.

Будем предполагать, что рождение мальчика и рождение девочки - в каждом отдельном наблюдении равновероятные события (то есть с вероятностной точки зрения наблюдения рождения ребенка и бросания монеты аналогичны), вероятность которых равна 0,5. Возникает вопрос о том, насколько эта модель (эксперимент с двумя равновероятными исходами) соответствует действительности, то есть реальному опыту. По данным большого объема наблюдений, проводившихся в Швейцарии с 1871 по 1900 годы, родилось 1359671 мальчик и 1285086 девочек.[1] Следовательно, для n = 2644757 получено m =1359671 и m/n = 0,5141. В математической статистике доказывается, что этот результат не согласуется с выбранной моделью равновероятности рождения мальчика и девочки. На самом деле вероятность рождения мальчика больше 0,5, а наилучшей оценкой вероятности рождения мальчика, полученной по этим выборочным данным, является относительная частота p = 0,5141.

Отметим, что в каждой из приведенных задач математической статистики по результатам эксперимента, выборки принимается определенное решение относительно распределения имеющихся наблюдений. Следовательно, математическую статистику можно назвать теорией статистических решений.

Теория вероятностей и математическая статистика возникли из решения задач, связанных с азартными играми. Свое дальнейшее развитие и совершенствование эти науки

[1] Б.Л. ван дер Варден. Математическая статистика. И. Л. М., 1960. с. 43.

получили благодаря запросам естественных наук, таких как теория ошибок наблюдений, статистика народного хозяйства, в частности, социально-демографическая статистика, и другие. В настоящее время можно утверждать, что теоретико-вероятностные концепции и математико-статистические методы применяются повсюду, где возникает потребность в обработке экспериментальных или наблюдаемых данных вероятностной природы.

1. ОПИСАТЕЛЬНАЯ СТАТИСТИКА: ВЫБОРОЧНЫЕ АНАЛОГИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

1.1. Генеральная совокупность и выборка

Математическая статистика изучает генеральные совокупности с помощью выборки.

Под генеральной совокупностью объектов, элементов, обладающих некоторыми признаками, понимается подходящая случайная величина. Если речь идет о многомерной совокупности, то рассматривается несколько признаков – случайных компонент случайного вектора. Множество элементов генеральной совокупности соответствует множеству возможных значений случайной величины, то есть пространству выборок. Случайно взятый элемент выборочной совокупности или зарегистрированное значение случайной величины, ее реализация, позволяет при достаточно большом объеме таких элементов судить о вероятностном распределении признака в генеральной совокупности. Конечно, выбранная теоретико-вероятностная модель – модель случайной величины – генеральной совокупности является идеализацией реального положения вещей, реальной совокупности, однако, если удается подобрать подходящую теоретико-вероятностную модель реальной совокупности, то оказывается возможным получить результаты, объясняющие и прогнозирующие развитие процесса и анализа реальных закономерностей.

Выборка определяется в математической статистике двояко.

Под выборкой объема n из генеральной совокупности понимается n наблюдаемых результатов испытания, например, n измерений (числовых или нечисловых) некоторого признака или нескольких признаков. В таком определении выборка записывается как n символов малыми латинскими буквами $x_1, x_2, \ldots, x_n$.

Под выборкой во втором смысле этого термина понимается система n независимых случайных величин (записывается большими латинскими буквами) $X_1, X_2, \ldots, X_n$, распределенных одинаково и так же, как генеральная совокупность или случайная величина X. Разумеется, речь может идти о многомерной случайной величине, тогда выборка представляет собой n k-мерных случайных векторов (или их реализаций) и записывается в виде матрицы порядка $n \times k$ или $k \times n$.

Выборка в первом смысле также представляет собой идеализацию реального положения вещей и может сильно отклоняться от условия случайности или, по-другому, представительности (репрезентативности) генеральной совокупности.

Таким образом, чтобы эффективно применять теоретико-вероятностные или математико-статистические методы анализа информации с целью принятия по возможности правильных решений, мы должны быть уверены в том, что генеральная совокупность соответствует реальной изучаемой и что представленная для обработки информация отражает в большей или меньшей степени основные особенности распределения признаков.

1.2. Вариационный ряд. Группировка

Пусть имеется выборка объема n из одномерной генеральной совокупности X с функцией распределения $F(x)$ $x_1, x_2, \ldots, x_n$.

Будем использовать малые буквы, понимая, когда это необходимо для теоретических утверждений, выборку во втором смысле.

Расположим наблюдаемые значения признака X в порядке неубывания, получим ряд $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, который называется вариационным рядом.

Для сжатия и дальнейшей обработки информации оказывается полезным сгруппировать элементы вариационного ряда либо естественным образом, вытекающим из дискретности распределения признака в генеральной совокупности, например, группировка семей работников по числу имеющихся детей, либо путем разбиения значений непрерывной случайной величины на некоторое число интервалов с общими смежными границами и, для удобства сравнения, одинаковой длины, за исключением может быть начального и конечного, которые могут заменяться даже лучами.

Сгруппированный вариационный ряд представляет собой последовательность пар (вариант признака, численность элементов, соответствующих варианту), вместо вариантов могут выступать интервалы, группы или некоторые значения признака X, представляющие группы, например, середины интервалов. Вместо численностей групп-частот признака X – употребляются относительные частоты-частости, выборочные доли изучаемого признака. Сгруппированный вариационный ряд изображается в виде ряда распределения частот или частостей, оформляемого таблицей, например,

X	x_1	x_2	…	x_r
m	m_1	m_2	…	m_r

или

$(a; b]$	$(a_1; b_1]$	$(a_2; b_2]$	…	$(a_r; b_r]$
w	w_1	w_2	…	w_r

$$\sum_{j=1}^{r} m_j = n$$

$$w_j = \frac{m_j}{n}, \sum_{j=1}^{r} w_j = 1,$$

где x_j – отдельные варианты-значения признака, представляющие группы (или интервалы), $(a_j; b_j]$ – интервалы группировки, для которых $b_j = a_{j+1}$, m_j – частота варианта x_j или интервала $(a_j; b_j]$, w_j – частость или выборочная доля j-го, соответствующего первого элемента пары.

При сопоставлении выборочных рядов распределения с частостями с рядами распределения вероятностей случайной величины, которые в случае непрерывности X являются аппроксимациями плотностей распределения, очевидна аналогия между выборочными долями и вероятностями. Эта аналогия логически оправдана в случае действия закона больших чисел, заключающегося в устойчивости частостей и при достаточно большом объеме выборки мало отличающимися от вероятностей получения варианта признака или попадания варианта в интервал.

Рассмотрим далее по аналогии с рядами распределения дискретных случайных величин или непрерывных случайных величин, аппроксимированных интервальными (по существу дискретными) рядами распределения, понятия, относящиеся к выборочным распределениям.

Для вариационного ряда аналогом функции распределения признака является выборочная функция распределения, определяемая по формуле

$$F_n(x) = \sum_{x_i \le x} w_j, \quad w_i = \frac{1}{n}, i \in \{1,2,...,n\}.$$

Для сгруппированного ряда или выборочного ряда распределения долей интегральная выборочная функция распределения может быть выражена формулой

$$F_n(x) = \sum_{x_j \le x} w_j, \quad w_j = \frac{m_j}{n}, j \in \{1,2,...,r\}.$$

Тогда $F_n(x)$ можно записать в виде ряда по накопленным частостям, обозначенным через f_j:

X	x_1	x_2	…	x_r
f	w_1	$w_1 + w_2$	…	1

$$f_j = \sum_{j'=1}^{j} w_{j'}.$$

Можно получить различные графические изображения рядов выборочного распределения (полигон, гистограмма) и выборочной функции распределения (кумулята).

1.3. Характеристики выборочных распределений

Характеристики выборочных распределений или выборочные характеристики генеральной совокупности, признака или случайной величины X вычисляются по формулам, аналогичным формулам определения и вычисления характеристик, параметров генеральной совокупности (генеральных параметров), с заменой вероятностей на частости, выборочные доли.

Характеристики положения.

1. Средняя арифметическая выборки определяется по формуле

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (простая)},$$

или для сгруппированного ряда

$$\overline{X} = \frac{\sum x_j m_j}{\sum m_j} \text{ (взвешенная)},$$

$$\sum m_j = \sum_{j=1}^{r} m_j = n, \; x_j = \frac{1}{2}\left(a_j + b_j\right)$$

2. Выборочная медиана. Для вариационного ряда медиана определяется по формуле

$$M_e = \begin{cases} x_{(l)} \text{ при } n = 2l - 1, \\ \frac{1}{2}\left(x_{(l)} + xl + 1\right) \text{ при } n = 2l. \end{cases}$$

Для интервального вариационного ряда распределения

$$M_e = a_{M_e} + h\frac{l - f_{M_e-1}}{m_{M_e}},$$

где a_{M_e} – нижняя граница, начало интервала, содержащего медиану, медианного интервала; h – длина интервала; $f_{M_e\text{-}1}$ – накопленная частота интервала, предшествующего медианному интервалу; l удовлетворяет условиям $n = 2l$ или $n = 2l - 1$, n – объем выборки.

3. Выборочная мода. Эта характеристика выборки определяется в предположении одномодального распределения признака в генеральной совокупности и для сгруппированного вариационного ряда:

$$M_{\text{o}} = a_{M_{\text{o}}} + h\frac{d_1}{d_1 + d_2},$$

где

$a_{M\text{o}}$ – начало модального интервала, то есть интервала, имеющего наибольшую частоту,

h – длина интервала,

$d_1 = m_{M\text{o}}\text{-}m_{M\text{o}\text{-}1}$ – разность частот модального и предшествующего модальному интервалов,

$d_2 = m_{M\text{o}}\text{-}m_{M\text{o}+1}$ – разность частот модального и следующего за модальным интервалов.

Как видно, средняя арифметическая и другие выборочные характеристики обладают свойствами, которыми обладают аналогичные генеральные характеристики. Как аналог математического ожидания средняя арифметическая функция выборки определяется по формуле

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n}\varphi\left(x_i\right) = \frac{\sum y_j m_j}{\sum m_j},$$

$$y_i = \varphi\left(x_i\right),$$

$$y_j = \varphi\left(x_j\right).$$

Характеристики рассеивания.

1. Выборочная дисперсия определяется формулой

$$S^2 = \overline{\left(x - \overline{x}\right)^2} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{\sum\left(x_j - \overline{x}\right)^2 m_j}{\sum m_j},$$

$$j \in \{1,2,...,k\}, \sum m_j = n\,.$$

2. Выборочное среднее квадратическое отклонение

$$S = \sqrt{S^2}$$

3. Выборочный коэффициент вариации

$$V = \frac{S}{\overline{X}}(100\%)$$

4. Размах выборки

$$R = x_{\max} - x_{\min} = x_{(n)} - x_{(1)}.$$

Выборочные моменты.

1. Начальный момент *l*-го порядка определяется по формуле

$$\nu_l^* = \overline{x^l} = \frac{1}{n}\sum_{i=1}^{n} x_i^l = \frac{\sum x_j^l m_j}{\sum m_j},\ \ \sum m_j = n,\ \ j \in \{1,2,...,k\}$$

В частности:

$$\nu_0^* = 1,\ \nu_1^* = \overline{x},\ \nu_2^* = \overline{x^2},\ \nu_3^* = \overline{x^3},\ \nu_4^* = \overline{x^4}$$

2. Центральные выборочные моменты определяются и вычисляются по формулам

$$\mu_l^* = \overline{\left(x - \overline{x}\right)^l} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^l = \frac{\sum\left(x_j - \overline{x}\right)^l m_j}{\sum m_j},$$

$$\sum m_j = n,\ \ j \in \{1,2,..., k\}.$$

В частности:

$$\mu_0^* = 1,\ \mu_1^* = 0,\ \mu_2^* = \overline{\left(x - \overline{x}\right)^2} = \nu_2^* - \left(\nu_1^*\right)^2 = \overline{x^2} - \left(\overline{x}\right)^2 = S^2,$$

$$\mu_3^* = \overline{\left(x - \overline{x}\right)^3} = \nu_3^* - 3\nu_2^* \cdot \nu_1^* + 2\left(\nu_1^*\right)^3,$$

$$\mu_4^* = \overline{\left(x - \overline{x}\right)^4} = \nu_4^* - 4\nu_3^* \cdot \nu_1^* + 6\nu_2^* \cdot \left(\nu_1^*\right)^2 - 3\left(\nu_1^*\right)^4.$$

Выборочные характеристики асимметрии и эксцесса.

1. Выборочный коэффициент асимметрии определяется формулой

$$A_c = \frac{\mu_3^*}{\left(\sqrt{\mu_{2*}}\right)^3}.$$

2. Выборочный коэффициент эксцесса определяется по формуле

$$E_k = \frac{\mu_4^*}{\left(\mu_2^*\right)^2} - 3$$

1.4. Двумерный ряд распределения выборки и его характеристики

Пусть имеется выборка объемом n из двумерной генеральной совокупности

$$(X,Y)^T : (x_1, y_1), (x_2, y_2), ..., (x_n, y_n).$$

Сгруппируем наблюдения в виде корреляционной таблицы, имеющей вид (табл.1.4.1).

Таблица 1.4.1

Корреляционная таблица частот

Y / X	$(c_1; d_1]$	. . .	$(c_j; d_j]$	. . .	$(c_S; d_S]$	n_x
$(a_1; b_1]$	n_{11}	. . .	n_{1j}	. . .	n_{1S}	n_{1*}
⋮	⋮		⋮		⋮	⋮
$(a_i; b_i]$	n_{i1}	. . .	n_{ij}	. . .	n_{iS}	n_{i*}
⋮	⋮		⋮		⋮	⋮
$(a_r; b_r]$	n_{r1}	. . .	n_{rj}	. . .	n_{rS}	n_{r*}
n_y	n_{*1}	. . .	n_{*2}	. . .	n_{*S}	$n_{**} = n$

Корреляционная таблица частот или частостей является аналогом дискретного двумерного ряда распределения случайной величины $(X,Y)^T$ с заменой в клетках (i, j) вероятностей p_{ij} на частоты n_{ij} или выборочные доли $W_{ij} = \frac{n_{ij}}{n}$. Таким образом, n_{ij} – число двумерных точек, попавших в клетку (i, j), прямоугольник с вершинами (a_i, c_j), (a_i, d_j), (b_i, d_j), (b_i, c_j) и длинами сторон $h_x = b_i - a_i$ и $h_y = d_j - b_j$. В итоговых строке и столбце указаны частоты одномерных рядов распределения выборки или частных выборочных распределений признаков X и Y, при этом $n_{i*} = \sum_j n_{ij}$, $j \in \{1,2,...,S\}$ есть частота варианта $x_i = \frac{a_i + b_i}{2}$ или интервала $(a_i; b_i]$ признака X, $n_{*j} = \sum_i n_{ij}$, $i \in \{1,2,...,r\}$ – частота j-й группы по признаку Y, $n_{**} = \sum_i \sum_j n_{ij} = \sum_i n_{i*} = \sum_j n_{*j} = n$.

Аналогично расчетам вектора средних и ковариационной матрицы двумерного случайного вектора (дискретное распределение) рассчитываем в дополнительных строках и столбцах выборочной корреляционной таблицы основные элементы формул моментов до второго порядка включительно. Будем иметь следующую схему расчетной таблицы (табл. 1.4.2).

Таблица 1.4.2

Расчетная таблица основных выборочных характеристик двумерной генеральной совокупности

Y / X	y_1	...	y_j	...	y_S	n_x	xn_x	x^2n_x	$\sum yn_{xy}$	$x\sum yn_{xy}$
x_1	n_{11}	...	n_{1j}	...	n_{1S}	n_{1*}	x_1n_{1*}	$x_1^2n_{1*}$	$\sum y_jn_{1j}$	$x_1\sum y_jn_{1j}$
⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮
x_i	n_{i1}	...	n_{ij}	...	n_{iS}	n_{i*}	x_in_{i*}	$x_i^2n_{i*}$	$\sum y_jn_{ij}$	$x_i\sum y_jn_{ij}$
⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮
x_r	n_{r1}	...	n_{rj}	...	n_{rS}	n_{r*}	x_rn_{r*}	$x_r^2n_{r*}$	$\sum y_jn_{rj}$	$x_r\sum y_jn_{rj}$
n_y	n_{*1}	...	n_{*j}	...	n_{*S}	$n_{**}=n$	$\sum x_in_{i*}$	$\sum x_i^2n_{i*}$	$\sum y_jn_{*j}$	$\sum\sum x_iy_jn_{ij}$
yn_y	y_1n_*	...	y_jn_{*j}	...	y_Sn_{*S}	$\sum y_jn_{*j}$				
y^2n_y	$y_1^2n_{*}$	...	$y_j^2n_{*j}$	...	$y_S^2n_{*S}$	$\sum y_j^2n_{*j}$				

Основные выборочные характеристики вычисляются по следующим формулам:

$$\overline{X}=\frac{1}{n}\sum x_in_{i*},\ \ S_x^2=\frac{1}{n}\sum x_i^2n_{i*}-\left(\overline{x}\right)^2,\ \ S_x=\sqrt{S_x^2},$$

$$\overline{Y}=\frac{1}{n}\sum y_jn_{*j},\ \ S_y^2=\frac{1}{n}\sum y_j^2n_{*j}-\left(\overline{y}\right)^2,\ \ S_y=\sqrt{S_y^2},$$

$$cov^*\left(x,y\right)=\overline{xy}-\overline{x}\cdot\overline{y}=\frac{1}{n}\sum_i\sum_j x_iy_jn_{ij}-\overline{x}\cdot\overline{y},$$

$$r=r_{xy}=\frac{\frac{1}{n}\sum\sum x_iy_jn_{ij}-\overline{x}\cdot\overline{y}}{\sqrt{S_x^2\cdot S_y^2}},\ [S]=\begin{pmatrix}S_x^2 & rS_xS_y\\ rS_xS_y & S_y^2\end{pmatrix}.$$

Дополнительные характеристики выборки вычисляются по следующим формулам:

$$b_{yx}=r\frac{S_y}{S_x},\qquad \overline{y}/x=\overline{y}+b_{yx}\left(x-\overline{x}\right),\qquad S_{y\ ост.}^2=S_y^2\left(1-r^2\right),$$

$$b_{xy}=r\frac{S_x}{S_y},\qquad \overline{x}/y=\overline{x}+b_{xy}\left(y-\overline{y}\right),\qquad S_{x\ ост.}^2=S_x^2\left(1-r^2\right).$$

Для подстановки в формулы используются результаты расчетной таблицы, помещенные в части итоговой строки и части итогового столбца, выделенные жирными линиями. Названия выборочных характеристик следующие: $(\bar{x}, \bar{y})$ - центр выборки, $\bar{x}$ - выборочная средняя (арифметическая) признака *X*, $\bar{y}$ - выборочная средняя признака *Y*, S_x^2, S_y^2 - выборочные дисперсии признаков *X* и *Y* - соответственно, S_x, S_y - выборочные средние квадратические отклонения признаков *X* и *Y*, $r = r_{xy}$ - выборочный коэффициент корреляции между *X* и *Y*, [*S*] - ковариационная матрица выборки, b_{yx}, b_{xy} - коэффициенты линейной регрессии выборки, $\overline{y/x_l}$ - условная средняя линейной регрессии *Y* на *X* выборки, $\overline{x/y_l}$ - условная средняя выборочной регрессии *X* на *Y* (линейной) – левые части выборочных линейных уравнений регрессии, $S^2_{y\ ост.}, S^2_{x\ ост.}$ - остаточные выборочные дисперсии линейных уравнений регрессии *Y* на *X* и *X* на *Y* выборки соответственно.

Из корреляционной таблицы можно получить выборочные аналоги условных рядов распределений и их характеристик. Рассмотрим выборочные характеристики зависимости *Y* от *X*.

Выборочный условный ряд частотного распределения признака *Y* при фиксированном $X = x_i$ имеет вид:

y	y_1	y_2	...	$y \cdot S$
$n_{y/x}$	n_{i1}	n_{ir}	...	n_{iS}

,

$$\sum_{j=1}^{S} n_{ij} = n_{i*},$$

или для выборочных условных долей-частостей:

y	y_1	y_2	...	y_j	y_S
$w_{y/x}$	$w_{1/i}$	$w_{2/i}$	...	$w_{j/i}$	$w_{S/i}$

,

$$w_{j/i} = \frac{n_{ij}}{n_{i*}}, \quad \sum_j w_{j/i} = 1.$$

Условная выборочная средняя признака *Y* при фиксированном признаке $X = x_i$ вычисляется по формуле:

$$\overline{y/x_i} = \sum_{j=1}^{S} y_j \cdot w_{j/i} = \frac{\sum_j y_j \cdot n_{ij}}{\sum_j n_{ij}} = \frac{\sum y_j \cdot n_{ij}}{n_{i*}}.$$

Множество точек $(x_i, \bar{y}/x_i)$, $i \in \{1,2,...,r\}$ образуют эмпирическую (выборочную) регрессию *Y* на *X*.

Условная выборочная дисперсия признака *Y* при фиксированном $X = x_i$ вычисляется по формуле:

$$S^2_{y/x_i} = \sum_{j=1}^{S} y_j^2 w_{j/i} - \left(\overline{y/x_i}\right)^2 = \frac{\sum y_j^2 n_{ij}}{n_{i*}} - \left(\overline{y/x_i}\right)^2$$

Выборочная остаточная дисперсия вычисляется по формуле:

$$S^2_{y\ ост.} = \overline{S^2_{y/x_i}} = \sum_{i=1}^{r} S^2_{y/x_i} \frac{n_{i*}}{n} = \frac{\sum y_j^2 n_{*j}}{n} - \frac{\sum \left(\overline{y/x_i}\right)^2 n_{i*}}{n}$$

Тогда выборочную дисперсию регрессии Y на X можно вычислить по формуле разложения дисперсий

$$S^2_{y\ регр.} = S^2_y - S^2_{y\ ост.}$$

Выборочные корреляционное и дисперсионное отношения получаются по формулам:

$$\eta^*_{y/x} = \sqrt{\frac{S^2_{y\ регр.}}{S^2_y}}, \quad \eta^{*2}_{y/x} = \frac{S^2_{y\ регр.}}{S^2_y}, \qquad 1 - \eta^{*2}_{y/x} = \frac{S^2_{y\ ост.}}{S^2_y}$$

Справедливо свойство $r^2 \le \eta^2$, где $\eta^2 = \eta^2_{y/x}$ или $\eta^2 = \eta^2_i$.

Аналогичные формулы можно получить при изучении зависимости X от Y в выборке.

Очевидно, методика получения множества точек регрессии выборки применима тогда, когда хотя бы при одном и том же x_i имеется $n_{ij} > 1$ точек y_j.

Если элементы выборки не сгруппированы и, например, все x_i различны, то множество точек эмпирической регрессии совпадает с самой выборкой (x_i, y_i), $i \in \{1,2,\ldots,n\}$. Следовательно, выборочные условные дисперсии равны нулю и $S^2_{Y\ общ.} = S^2_{Y\ регр.}$. Для получения аналогов коэффициентов связи следует ввести в качестве регрессии Y на X некоторую функцию, называющуюся выравнивающей кривой (прямой) регрессии.

Аппроксимация множества точек регрессии в выборке с помощью прямой линии имеет смысл, когда множество точек регрессии, например, $(x_i, \overline{y}/x_i)$ или $(\overline{x}_i, y_i)$, $i \in \{1,2,\ldots r\}$, колеблется около этой прямой линии.

Таким образом, в качестве аналогов условных математических ожиданий выступают значения $\overline{y}/x_i$, получаемые при подстановке x_i в формулу выравнивающей кривой (прямой) регрессии, как это делается в регрессионном анализе. Если выравнивающая линия является прямой, то в качестве аналогов связи берутся выборочные коэффициенты корреляции и регрессии. Кроме того, здесь можно вычислить аналоги генеральных остаточной дисперсии и дисперсии регрессии Y на X; выборочное дисперсионное отношение здесь будет равно выборочному коэффициенту детерминации или $\eta^*_Y = |r|$.

1.5. Пояснения, примеры и решения задач

1. Рассмотрим примерную схему обработки исходной информации, которая предполагается выборкой из генеральной совокупности X с некоторым законом распределения $F(x)$.

Пусть имеются данные о числе детей в семье работников, полученные на основании обследования выборки, объемом в 100 семей. Информация представлена в виде сгруппированного ряда:

Число детей в семье (X)	0	1	2	3	4	5	Итого
Число семей (m) с числом детей (X)	20	43	28	5	3	1	100

Запишем этот ряд, указав частости каждого варианта признака X - числа детей в семье и накопленные частости:

Число детей в семье (X)	0	1	2	3	4	5	Итого
w	0,20	0,43	0,28	0,05	0,03	0,01	1,00
f	0,20	0,63	0,91	0,96	0,99	1,00	–

Вычислим выборочные характеристики положения признака X.

Средняя арифметическая:

$$\overline{X} = 0 \cdot 0{,}2 + 1 \cdot 0{,}43 + 2 \cdot 0{,}28 + 3 \cdot 0{,}05 + 4 \cdot 0{,}03 + 5 \cdot 0{,}01 = 1{,}31 \ чел.$$

Выборочная медиана:

$$M_e = 1 \ чел.,$$

так как в (ранжированном) вариационном ряде семьи с номерами 50 и 51 попадают в группу семей с одним ребенком.

Выборочная мода:

$$M_0 = 1 \ чел.,$$

т.к. имеется вариант с наибольшей частотой (частостью 0,43), он равен 1.

Вычислим характеристики вариации выборки.

Размах выборки:

$$R = X_{\max} - X_{\min} = 5 - 0 = 5 \ \ чел.$$

Выборочная дисперсия:

$$S^2 = \overline{X^2} - (\overline{X})^2,$$

$$\overline{X^2} = 0^2 \cdot 0{,}2 + 1^2 \cdot 0{,}43 + 2^2 \cdot 0{,}28 + 3^2 \cdot 0{,}05 + 4^2 \cdot 0{,}03 + 5^2 \cdot 0{,}01 = 2{,}73,$$

$$S^2 = 2{,}73 - 1{,}31^2 = 1{,}0139 \ \ чел.$$

Среднее квадратическое отклонение выборки:

$$S = \sqrt{S^2} = 1{,}006926 \ \ чел.$$

Выборочный коэффициент вариации (по среднему квадратическому отклонению):

$$V = \frac{S}{\overline{X}} \cdot 100\% = 76{,}86\%$$

Выборочные моменты:

начальные

$\nu_0^* = 1;$

$\nu_1^* = \overline{X} = 1{,}31;$

$\nu_2^* = \overline{X^2} = 2{,}43;$

$\nu_3^* = 0^3 \cdot 0{,}2 + 1^3 \cdot 0{,}43 + 2^3 \cdot 0{,}28 + 3^3 \cdot 0{,}05 + 4^3 \cdot 0{,}03 + 5^3 \cdot 0{,}01 = 7{,}19;$

$\nu_4^* = 0^4 \cdot 0{,}2 + 1^4 \cdot 0{,}43 + 2^4 \cdot 0{,}28 + 3^4 \cdot 0{,}05 + 4^4 \cdot 0{,}03 + 5^4 \cdot 0{,}01 = 22{,}89;$

центральные

$\mu_0^* = 1;$

$\mu_1^* = 0;$

$\mu_2^* = S_x^2 = 1{,}0139;$

$\mu_3^* = \nu_3^* - 3\nu_2^* \cdot \nu_1^* + 2\nu_1^{*3} = 2{,}1363;$

$\mu_4^* = \nu_4^* - 4\nu_3^* \cdot \nu_1^* + 6\nu_2^* \cdot \nu_1^{*2} - 3\nu_1^{*4} = 1{,}4001.$

Выборочные коэффициенты асимметрии и эксцесса:

$$A_c = \frac{\mu_3^*}{(\sqrt{\mu_2^*})^3} = 2{,}0925;$$

$$E_k = \frac{\mu_4^*}{\mu_2^{*2}} - 3 = -1{,}6380.$$

2. Рассмотрим построение интервального ряда для выборки из непрерывной генеральной совокупности (или из совокупности с непрерывным распределением признака).

В таблице приведены значения темпа роста прибыли 100 малых предприятий по сравнению с базисным периодом в процентах.

105,3	101,8	101,2	102,7	105,7	101,1	101,9	101,5	102,8	105,2
102,8	106,7	102,0	105,4	102,3	100,8	101,8	101,3	108,3	104,3
105,5	103,9	105,1	109,6	102,8	102,0	103,5	103,7	103,7	104,9
102,6	103,6	103,3	102,8	102,3	103,9	103,3	101,8	100,8	104,2
101,9	104,6	106,2	103,2	102,2	101,8	104,8	105,3	104,6	103,8
102,3	104,3	105,6	102,5	101,6	102,0	108,4	106,9	101,2	103,6
105,3	103,1	103,4	104,7	106,2	104,7	107,8	107,1	103,9	101,5
103,3	103,9	106,2	104,9	105,3	102,8	103,5	102,9	102,7	101,5
101,0	100,2	100,5	105,9	104,2	102,6	101,2	104,4	105,2	103,8
106,4	104,6	102,7	101,2	105,3	104,3	103,1	103,1	104,9	105,2

Для построения сгруппированного ряда распределения выборки следует определить число интервалов, величину интервала (шаг) и численность (частоту) единиц, попавших в каждый интервал. Число интервалов обычно берут в пределах от 5 до 15 в связи с меньшим или большим объемом выборки. При выборке числа интервалов руководствуются теми соображениями, что оно не должно быть, с одной стороны, слишком большим, так как в каждом интервале окажется мало единиц, что ведет к трудностям выявления формы распределения выборки, с другой стороны число интервалов не должно быть чрезмерно малым, так как тогда могут быть упущены существенные черты распределения.

Более или менее удовлетворительный подход к процедуре определения числа интервалов дает следующая эмпирическая формула:

$$h = \frac{R}{\log_2 n} = \frac{x_{\max} - x_{\min}}{1{,}4427 \cdot \ln n} = \frac{x_{\max} - x_{\min}}{3{,}322 \cdot \lg n},$$

где h – ширина интервала, R – размах выборки.

В нашем примере подчеркнуты в исходной таблице данных x_{max} и x_{min}, поэтому

$$h = \frac{109{,}6 - 100{,}2}{6{,}644} = \frac{9{,}4}{6{,}644} = 1{,}41$$

чтобы исходные значения признака меньше попадали на границы между интервалами, h имеет на один десятичный знак больше, чем данные значения.

Чтобы x_{min} попало внутрь группировки, отступаем от x_{min} влево не более чем на полшага и получаем левый конец a_1 первого интервала системы интервалов, затем, откладывая от a_1 длину интервала h, получаем правый конец b_1 первого интервала, совпадающий с a_2 - началом второго интервала. Повторяем процедуру откладывания шага h до тех пор, когда наблюдение x_{max} окажется внутри последнего интервала.

Таким образом будет получена система, состоящая из $r \cong \frac{R}{\log_2 n} + 1$ примыкающих последовательно друг к другу интервалов шириной h, содержащая внутри все n элементов выборки.

В нашем примере это будет система с началом в начале первого интервала $a_1 = 100{,}2 - \frac{1}{2} \cdot 1{,}41 = 99{,}50$ и концом в конце восьмого интервала $b_8 = 109{,}77$:

(99,50; 100,91], (100,91; 102,32], (102,32; 103,73], (103,73; 105,14], (105,14; 106,55], (106,55; 107,96], (107,96; 109,37], (109,37; 110,78].

В таблице 1.5.1 представлен интервальный и дискретный ряд распределения значений темпов роста прибыли 100 малых предприятий с различными частотами и частостями. Значения x_j являются серединами интервалов (a_j, b_j), $j \in \{1,2,...,8\}$.

Вычислим выборочную моду полученного ряда распределения:

$$M_o = a_{Mo} + h \frac{m_{Mo} - m_{Mo-1}}{2m_{Mo} - m_{Mo+1} - m_{Mo-1}}.$$

В таблице находим модальный интервал – третий, следовательно, $a_{Mo} = 102{,}32$, $m_{Mo} = 26$, $m_{Mo-1} = 24$, $m_{Mo+1} = 22$. Имея $n = 100$ и $h = 1{,}41$, получим

$$M_o = 102{,}32 + 1{,}41 \frac{26 - 24}{52 - 24 - 22} = 102{,}79\%.$$

Для нахождения медианы определяем медианный интервал, используя, например, столбец f накопленных частот: $f_{Me} = 54$. Тогда $f_{Me-1} = 28$, $m_{Me} = 26$, $a_{Me} = 102{,}32$. Далее $l = 50$, $h = 1{,}41$. Значение выборочной медианы получаем по формуле:

$$M_e = a_{Me} - h \frac{l - f_{Me-1}}{m_{Me}} = 102{,}32 - 1{,}41 \frac{50 - 28}{26} = 103{,}51\%.$$

Для отыскания выборочных моментов результаты вычислений целесообразно оформить в виде таблицы расчета моментов по упрощающей вычисления схеме (табл.1.5.2), естественно, при расчете вручную с калькулятором.

Введем вспомогательную переменную $x'_j = \dfrac{x_j - x_o}{h}$, где в качестве x_o берется середина серединного интервала; так как таких интервалов два (четвертый и пятый), выбираем из них тот, который имеет наибольшую частоту (или любой), таким образом, выбираем пятый интервал, следовательно, $x_o = 104{,}435$.

Таблица 1.5.1

Ряд распределения темпов роста прибыли малых предприятий (%)

(*a; b*]	*x*	*m*	*w*	*f*	$F_n(x)$
(99,50; 100,91]	100,205	4	0,04	4	0,04
(100,91; 102,32]	101,615	24	0,24	28	0,28
(102,32; 103,73]	103,025	26	0,26	54	0,54
(103,73; 105,14]	104,435	22	0,22	76	0,76
(105,14; 106,55]	105,845	17	0,17	93	0,93
(106,55; 107,96]	107,255	4	0,04	97	0,97
(107,96; 109,37]	108,665	2	0,02	99	0,99
(109,37; 110,78]	110,075	1	0,01	100	1,00
Итого	–	100	1,00	–	–

Таблица 1.5.2

Вычисление моментов

x'	*m*	$x'm$	x'^2m	x'^3m	x'^4m
-3	4	-12	36	-108	324
-2	24	-48	96	-192	384
-1	26	-26	26	-26	26
0	22	-86	0	-326	0
1	17	17	17	17	17
2	4	8	16	32	64
3	2	6	18	54	162
4	1	4	16	64	256
		35		167	
Итого	100	-51	225	-159	1233
ν'_k		-0,51	2,5	-1,59	12,33

В результате вместо исходных чисел x_j получаем числа целые, наименее уклоняющиеся от нуля, что значительно облегчает набор и вычисления. Для новой переменной x'_j вычисляем начальные моменты ν'_k по известным формулам (числа в рамках – положительные и отрицательные промежуточные суммы при расчете моментов нечетного порядка). Для получения выборочных характеристик используются формулы, выражающие их через моменты ν'_k, а именно:

$\overline{X} = \nu'_1 \cdot h + x_0 = -0{,}51 \cdot 1{,}41 + 104{,}435 = 103{,}72\%$

$S^2 = \mu'_2 \cdot h^2 = (\nu'_2 - \nu'^2_1) \cdot h^2 = (2{,}25 - (-0{,}51)^2) \cdot 1{,}41^2 = 3{,}95612(\%)$;

$\mu_2^{'} = 1{,}9899$

$S = \sqrt{S^2} = 1{,}989\%; \quad \sqrt{\mu_2^{'}} = 1{,}4106$

$\mu_3^{'} = \nu_3^{'} - 3\nu_2^{'} \cdot \nu_1^{'} + 2\nu_1^{'} = -1{,}59 + 3 \cdot 2{,}25 \cdot 0{,}51 - 2 \cdot 0{,}51^3 = 1{,}7198$

$\mu_4^{'} = \nu_4^{'} - 4\nu_3^{'} \cdot \nu_1^{'} + 6\nu_2^{'} \cdot \nu_1^{'2} - 3\nu_1^{'4} = 12{,}33 - 4 \cdot 1{,}59 \cdot 0{,}51 + 6 \cdot 2{,}25 \cdot 0{,}51^2 -$

$- 3 \cdot 0{,}51^4 = 12{,}3948$

$$A_c = \frac{\mu_3^*}{\mu_2^{*3/2}} = \frac{\mu_3^{'}}{\mu_2^{'3/2}} = \frac{1{,}7198}{1{,}4106^3} = 0{,}613$$

$$E_k = \frac{\mu_4^{'}}{\mu_2^{'2}} - 3 = 0{,}130$$

Справедливость приведенных формул легко доказать, заменив $x_j^{'}$ по обратной формуле $x_j^{'} = x_j h + x_0$.

Выборочный коэффициент вариации

$$V = \frac{S}{\overline{X}} \cdot 100\% = 1{,}92\% .$$

На рисунке 1.5.1 дано графическое изображение полигона и гистограммы и показан порядок расположения характеристик M_0, M_e и $\bar{x}$ на оси x-ов.

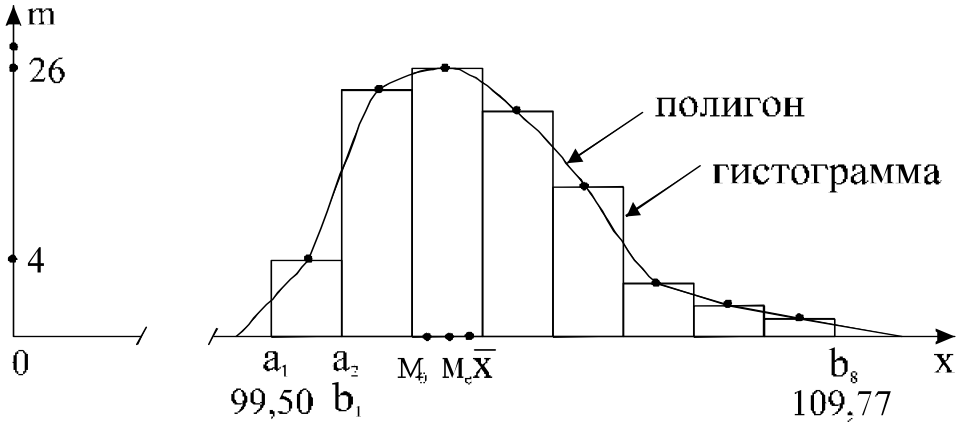

Рис. 1.5.1. Графики гистограммы и полигона распределения

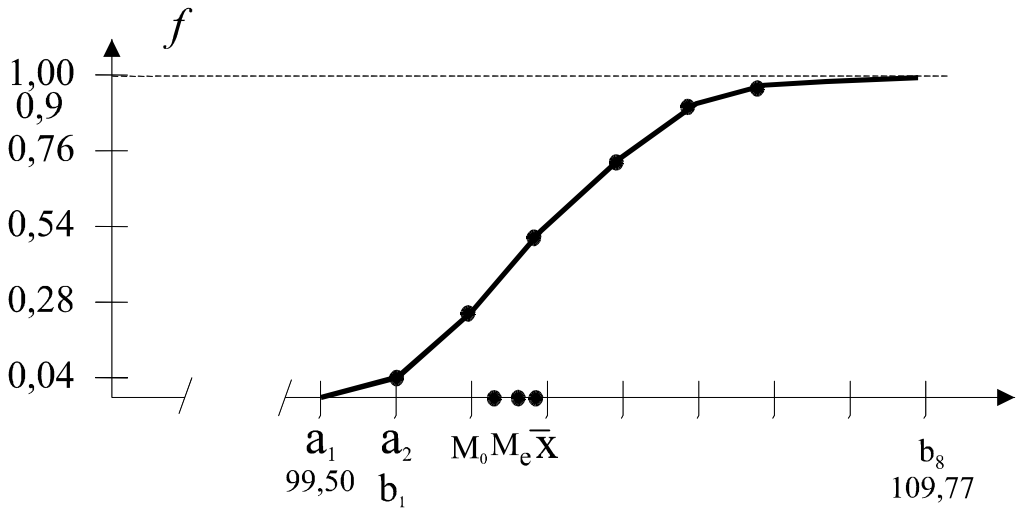

Рис. 1.5.2

Рис. 1.5.2 изображает кумуляту выборочного распределения частостей

3. Построение ряда сгруппированных данных и графическое его изображение в некоторых случаях помогают подобрать подходящую модель распределения признака в генеральной совокупности. Кроме этого, как правило, должно выполняться приблизительное равенство соотношений между характеристиками выборки и характеристиками совокупности.

Так, в первом примере можно предположить, что распределение семей подчиняется биномиальному закону или закону Пуассона, так как $\bar{x} = S^2 = \lambda *$. Во втором примере выборочные характеристики положения (M_0, M_e , $\bar{x}$) близки друг к другу, а асимметрия и эксцесс достаточно малы. Можно сделать предположение о нормальном законе распределения темпа роста прибыли предприятий в генеральной совокупности, откуда взята выборка.

4. Пусть имеется выборка объемом в 10 единиц из генеральной совокупности (X,Y) с распределением F(x,y):

(1;9), (3;10), (4;11), (5;11), (6;10), (7;15), (7;16), (8;13), (9;17), (10;17),

где x_i, y_i – расходы на рекламу и объем продаж за определенный период времени фирм, в условных единицах стоимости (у.е.с.). Рассчитаем следующие выборочные характеристики распределения:

средние, дисперсии, средние квадратические отношения

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60}{10} = 6 y.e.c., \quad \overline{x^2} = \frac{\sum x_i^2}{n} = \frac{430}{10} = 43, \quad S_x^2 = \overline{x^2} - \left|\bar{x}\right|^2 = 43 - 36 = 7;$$

$$S_x = \sqrt{7} = 2{,}6 y.e.c.,$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{134}{10} = 13{,}4, \quad \overline{y^2} = \frac{1876}{10} = 187{,}6, \quad S_y^2 = 187{,}6 - 13{,}4^2 = 8{,}04;$$

$$S_y = 2{,}84 y.e.c.,$$

коэффициенты корреляции, детерминации, регрессии

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{872}{10} = 87{,}2, \quad r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{S_x^2 \cdot S_y^2}} = \frac{87{,}2 - 6 \cdot 13{,}4}{\sqrt{7 \cdot 8{,}04}} = \frac{6{,}8}{7{,}502} = 0{,}9064249,$$

$$r^2 = 0{,}8216062; \quad 1 - r^2 = 0{,}1783938, \quad b_{yx} = r\frac{S_y}{S_x} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x^2} = \frac{87{,}2 - 6 \cdot 13{,}4}{7} = \frac{6{,}8}{7} =$$

$$= 0{,}971; \quad b_{xy} = r\frac{S_x}{S_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_y^2} = \frac{6{,}8}{8{,}04} = 0{,}846, \quad \textit{контроль}: r^2 = b_{yx} \cdot b_{xy} = \frac{6{,}8}{7} \cdot \frac{6{,}8}{8{,}04} =$$

$$= 0{,}8216062,$$

уравнения линейной регрессии

$$\bar{y}/x = \bar{y} + b_{yx}(x - \bar{x}), \; \bar{y}/x = 13{,}4 + 0{,}971(x - 6), \; \bar{x}/y = 6 + 0{,}846(y - 13{,}4)$$

остаточные дисперсии (равные условным дисперсиям), дисперсии регрессии

$$S^2_{y\,ост.} = S^2_{y/x} = S_y^2(1 - r^2) = 8{,}04 \cdot 0{,}1784 = 1{,}43, \quad S^2_{x\,ост.} = S^2_{x/y} = S_x^2(1 - r^2) =$$

$$= 7 \cdot 0{,}1784 = 1{,}25, \quad S^2_{y\,регр.} = S_y^2 \cdot r^2 = 8{,}04 \cdot 0{,}8216 = 6{,}61, \quad S^2_{x\,регр.} = S_x^2 \cdot r^2 = 7 \cdot 0{,}8216 =$$

$$= 5{,}75.$$

Построим множество выборочных точек (поле корреляции) и обе линии регрессии в системе координат $x0y$ (рис.1.5.3.). Для построения каждой прямой достаточно знать координаты двух ее точек. Одна точка - $(\bar{x}, \bar{y})$, через которую проходят обе линии. В качестве другой точки прямой можно выбрать любую, не выходящую за пределы рисунка или поля корреляции, например, конец отрезка, отсекаемого от оси абсцисс $Oy(x = 0, \bar{x}/y = 0)$.

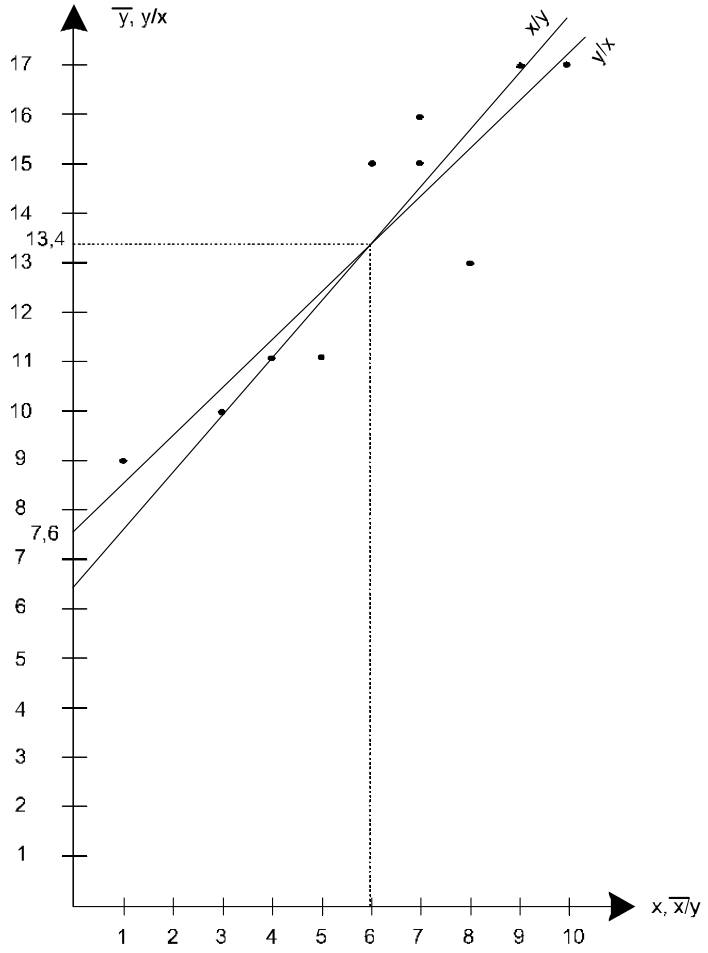

Рис 1.5.3

1.6. Упражнения

По данным выборкам построить сгруппированные (интервальные) ряды и вычислить основные характеристики.

1. Расходы на горючее при доставке товара от поставщика к потребителю составили для 50 автомашин в у.е.:

327, 664, 596, 654, 517, 103, 588, 491, 591, 388,
438, 401, 510, 218, 417, 523, 772, 561, 784, 572,
503, 724, 671, 589, 599, 572, 784, 561, 772, 523,
618, 825, 645, 754, 631, 737, 849, 698, 712, 680,
607, 441, 785, 998, 866, 715, 523, 581, 431, 359.

2. Затраты времени при поездке из дома на работу составили в минутах для 50 рабочих дней:

24	21	22	20	25	26	27	28	29	21
21	21	19	24	23	27	21	19	18	17
19	21	22	25	24	26	20	23	22	19
20	21	20	22	22	21	19	21	19	20
19	19	18	21	21	23	24	25	25	24

3. При контроле распродаваемых комплектов из 10 электрических батареек оказалось, что в выборке 50 комплектов число неисправных батареек в комплекте было следующим:

1 0 3 4 2 2 1 2 1 1
1 1 0 3 5 2 0 1 0 1
1 2 1 1 1 4 2 0 1 2
0 1 2 1 1 1 1 1 3 1
2 1 0 1 1 2 1 2 2 1

4. Даны следующие числа бутылок с трещинами в 50 выбранных тарных ящиках по 20 бутылок в каждом ящике.

3, 1, 0, 3, 1, 2, 2, 2, 2, 0, 1, 2, 2, 2, 3, 1, 2, 2, 2, 3, 2, 3, 1, 3, 1, 2, 1, 0, 0, 0, 5, 0, 1, 2, 3, 6, 0, 0, 0, 1, 0, 1, 1, 3, 4, 4, 4, 5, 2, 2.

5. По данным таблицы 1.6.1 найти выборочные характеристики (средние, дисперсии, средние квадратические отклонения, коэффициенты корреляции, детерминации линейной регрессии).

6. По данным таблицы 1.6.1 найти корреляционные и дисперсионные отношения.

Таблица 1.6.1

Распределение 100 рабочих по числу фактически отработанных часов и величине выручки за выполненную работу (в усл. ед. стоимости) в течение недели

Выручка Y	Фактически отработанное время - X				Итого рабочих
	30	35	40	45	
1000	12	10	-	-	22
5000	2	22	15	8	47
9000	1	14	10	6	31
Итого рабочих	15	46	25	14	100

7. По выборке объема n = 50 составлена корреляционная таблица

Y / X	10	15	20	25	n_x
20	-	4	5	1	10
30	6	10	12	1	29
40	3	8	-	-	11
n_y	9	22	17	2	50

Вычислить выборочные коэффициенты корреляции, детерминации, корреляционные и дисперсионные отношения.

8. По данным задач 1-4 построить графические изображения выборочных рядов распределения.

9. По данным таблицы 1.6.1. и задачи 7 построить поле корреляции, множества точек эмпирической регрессии и прямые линии регрессии в системе координат x0y, при этом n_{ij} точек выборки располагаются равномерно внутри клетки (прямоугольника) (i,j).

10. Выполнить вычисления всех характеристик связи согласно с решением задачи 4 пункта 1.5 на основе данных таблицы 1.6.2.

Таблица 1.6.2

Заказы на выпуск продукции по промышленным предприятиям и организациям региона в 1999 году

Дата	Объем имеющихся заказов на выпуск продукции в последующие периоды (млн руб.) - X	Обеспеченность производства заказами (месяцев) - Y
на 01.01	8480,6	1,7
на 01.02	17666,3	4,3
на 01.03	13672,9	2,9
на 01.04	22129,2	3,7
на 01.05	23542,0	3,8
на 01.06	20679,1	3,6
на 01.07	19956,2	3,3
на 01.08	20180,4	3,8
на 01.09	17544,2	2,6
на 01.10	16642,0	2,4
на 01.11	15298,1	2,0
на 01.12	14464,3	1,9
на 01.01 2000 г.	15842,0	1,8

11. На основе следующих десяти выборочных данных о расстоянии (X, км) до места вызова и времени (Y, мин.) поездки ремонтной бригады получить линейное уравнение регрессии y на x, подсчитать остаточную дисперсию и построить график регрессии и поле корреляции: (1,0;7), (1,2;10), (1,4;10), (2,5;14), (2,9;15), (4,1;15), (4,5;19), (4,8;20), (4,8;17), (5,0;20).

2. СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Рассмотрим выборку из одномерной генеральной совокупности X объемом n

$$X_1, X_2, ..., X_n$$

как систему взаимно независимых случайных величин, имеющих такое же распределение, как и сама генеральная совокупность X. Распределение выборки можно рассматривать как распределение n-мерного вектора с независимыми случайными компонентами, плотность вероятностей которого можно записать в виде (если речь идет о непрерывной генеральной совокупности):

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i).$$

Аналогично, для дискретной случайной величины также можно определить распределение выборки или выборочное распределение.

Всякая функция выборки является случайной величиной (если аргументы рассматриваются как случайные величины) и называется статистикой. Одни статистики могут включать в качестве аргументов параметры распределения генеральных совокупностей, другие могут не включать их. Обычно, если речь идет о статистике, то предполагается, что входящие в нее в качестве аргументов параметры считаются известными, то есть для статистики мы должны иметь конкретное (числовое) значение, если имеется выборка во втором смысле этого слова.

Плотность распределения можно рассматривать как статистику, если значения всех параметров, от которых она зависит, известны. Если же, наоборот, считать известными величины, образующие выборку, то плотность выборки, рассматриваемая как функция неизвестных параметров генеральной совокупности, называется функцией правдоподобия выборки и записывается как

$$p(x_1, ..., x_n, \theta),$$

где θ – неизвестный параметр или вектор параметров.

К задачам статистического вывода в первую очередь относится задача оценивания параметров распределения и самих законов распределения. Рассмотрим оценивание неизвестных параметров генеральной совокупности. Оценки могут быть точечными и интервальными (в случае векторного параметра аналогом интервальной оценки не обязательно является прямоугольник или многомерный брус).

2.1. Точечные оценки и некоторые их свойства

Точечной оценкой неизвестного параметра θ называется некоторая статистика, значение которой может быть принято в качестве приближенного значения этого параметра. Так, статистики, рассмотренные ранее как аналоги характеристик генеральной совокупно-

сти, являются точечными оценками соответствующих параметров, характеристик, генеральной совокупности.

Очевидно, можно предположить не одну статистику в качестве оценки некоторого параметра. Например, для симметричного одномодального распределения признака можно указать три оценки математического ожидания – выборочные среднюю арифметическую моду, медиану. Для генеральной дисперсии оценкой являются статистики

$$S^2,\ \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ S_*^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - MX)^2$$

(последняя статистика будет являться оценкой, то есть может быть вычислена, когда параметр MX – известен). Вообще существует бесчисленное множество оценок (например, оценкой математического ожидания можно считать статистику вида $\bar{x} + \frac{a}{n}$, где a – любое число), поэтому выбирают наиболее предпочтительные оценки, обладающие некоторыми полезными свойствами или удовлетворяющие некоторым требованиям. Будем обозначать оценку параметра θ в виде

$$T = T_n(x_1, x_2, ..., x_n).$$

1. Оценка T (точечная) называется несмещенной, если $MT = \theta$.

Покажем, что $\bar{x}$ является несмещенной оценкой для MX. Действительно,

$$M\overline{X} = M\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}MX_i = \frac{1}{n}\sum_{i=1}^{n}MX = \frac{1}{n}nMX = MX.$$

В этой цепочке равенств использованы известные свойства математического ожидания и определение выборки во втором смысле: каждое X_i независимо от остальных и имеет то же распределение, что и признак X, то есть $MX_i = MX$ для $i \in \{1,2,\ldots n\}$.

Выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ является смещенной оценкой для DX.

Действительно,

$$M(S^2) = M(\overline{X^2} - (\overline{X})^2) = M\overline{X^2} - M(\overline{X})^2 = M\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - M\left[\frac{(\sum X_i)^2}{n^2}\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}MX_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}MX_i^2 + \sum_{i\neq j}MX_iMX_j\right) = \frac{1}{n}n\nu_2 - \frac{1}{n^2}\left(n\nu_2 - n(n-1)\nu_1^2\right) =$$

$$= \nu_2 - \frac{1}{n}\nu_2 - \frac{n-1}{n}\nu_1^2 = \frac{n-1}{n}\nu_2 - \frac{n-1}{n}\nu_1^2 = \frac{n-1}{n}(\nu_2 - \nu_1^2) = \frac{n-1}{n}DX =$$

$$= DX - \frac{1}{n}DX.$$

Величина $\frac{1}{n}DX$ называется смещением выборочной дисперсии. При достаточно больших n смещением можно пренебречь, то есть считать S^2 несмещенной оценкой генеральной дисперсии. В этом случае также говорят, что оценка S^2 является асимптотически несмещенной оценкой DX.

Очевидно, можно "исправить " S^2 и получить несмещенную оценку DX

$$\hat{S}^2 = \frac{n}{n-1}S^2, \text{ так как } M\hat{S}^2 = DX .$$

Таким образом, если выборка мала, исправленная дисперсия $\hat{S}^2$ является более предпочтительной, чем выборочная дисперсия S^2.

Несмещенная оценка в измерении параметра гарантирует от систематических ошибок. При смещенности оценки наблюдаемые значения статистики группируются не около истинного значения параметра, а около математического ожидания статистики, поэтому могут получиться измерения «с недостатком» или, наоборот, измерения оценивают параметр «с избытком».

2. Оценка T параметра θ называется состоятельной, если она сходится по вероятности к оцениваемому параметру, то есть

$$\lim_{n\to\infty}\left(\left|T_n(x_1,...,x_n)-\theta\right|<\varepsilon\right)=1, \varepsilon>0 .$$

Следовательно, если оценка имеет математическое ожидание, равное параметру, или смещение, стремящееся к нулю при $n \to \infty$, и, кроме того, дисперсию, стремящуюся к нулю при $n \to \infty$, то такая оценка, согласно теореме Чебышева, является состоятельной.

Так как $D\overline{X} = \frac{1}{n^2}D(X_1 + ...X_n) = \frac{1}{n^2}\sum_{i=1}^{n}DX_i = \frac{DX}{n}$ и $M\overline{X} = MX$, то средняя арифметическая является состоятельной оценкой математического ожидания, в частности, частость W является состоятельной оценкой вероятности p.

Важность свойства состоятельности состоит в том, что при достаточно большом объеме выборки практически достоверна замена неизвестного параметра на значение статистики, являющейся его точечной оценкой.

Выборочная и исправленная дисперсия также являются состоятельными оценками DX, так как первое условие (асимптотической несмещенности) для них, как показано, выполняется и выполняется условие стремления к нулю дисперсии этих оценок (последнее можно доказать, если вычислить эту дисперсию).

3. Пусть дано несколько несмещенных оценок параметра θ. Оценка, имеющая среди всех наименьшую дисперсию при одном и том же объеме выборки, является эффективнее, чем остальные. Если это свойство выполняется по отношению к любой другой статистике, оценивающей тот же параметр, то оценка называется эффективной. Доказывается, что эффективными являются $\overline{X}$ и S^2, если речь идет о выборке из нормально распределенной генеральной совокупности с параметрами μ и σ^2.

2.2. Законы распределения некоторых статистик

При решении практических задач, особенно когда приходится иметь дело с малыми выборками, важно знать точные законы распределения выборочных характеристик.

При больших выборках, когда значение n велико, можно пользоваться асимптотическими законами. Однако следует помнить, что асимптотические распределения иногда резко расходятся с истинными распределениями, особенно при малых n.

Предположим, что имеется выборка объемом n из одномерной генеральной совокупности с функцией распределения $F(x)$ и требуется определить закон распределения статистики $T_n(x_1, x_2, ..., x_n)$. Эта задача сводится к отысканию закона распределения функции $T_n(x_1, x_2, ..., x_n)$ от n независимых случайных величин $X_1, X_2, ..., X_n$ с одной и той же функцией распределения $F(x)$.

Доказано, что если заданы функции F и T_n, то всегда существует единственное решение.

Точное, или относительно точное решение удается получить лишь в сравнительно редких случаях. Кроме случая среднего арифметического получено мало общих результатов. Лишь в одном частном случае, когда выборка берется из нормальной генеральной совокупности, получаются достаточно полные результаты. Именно этот случай мы и будем рассматривать в дальнейшем.

Распределение $\overline{X}$

Пусть из генеральной совокупности X, имеющей нормальный закон распределения $N(\mu, \sigma)$ с математическим ожиданием μ и средним квадратическим отклонением σ, взята выборка объемом n и определена средняя арифметическая

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

где X_i - результат i-го наблюдения.

Здесь имеется в виду выборка во втором смысле, то есть выборка рассматривается как система $X_1, X_2, ..., X_n$ независимых, одинаково распределенных нормально случайных величин с параметрами распределения μ и σ.

Можно показать, что $\overline{X}$ имеет нормальный закон распределения $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ с математическим ожиданием μ и средним квадратическим отклонением $\frac{\sigma}{\sqrt{n}}$ или дисперсией $\frac{\sigma^2}{n}$, $M\overline{X} = \mu$, $D\overline{X} = \frac{\sigma^2}{n}$, например, используя правила получения закона распределения при умножении случайной величины на число и композицию (свертку) законов распределения для случая независимых слагаемых величин.

Распределение $\left(\overline{X} - \overline{Y}\right)$

Пусть из генеральной совокупности X, имеющей нормальный закон распределения $N(\mu_x, \sigma_1)$, взята выборка объемом n_1, а из генеральной совокупности Y с распределением $N(\mu_y, \sigma_2)$ – выборка объемом n_2, причем выборки независимы.

Аналогично предыдущему можно показать, что случайная величина $Z = \overline{X} - \overline{Y}$, где

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \; ; \; \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \; ,$$

имеет нормальный закон распределения с параметрами

$$MZ = \mu_x - \mu_y \; ; \; \sigma_z = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \; .$$

Рассмотрим некоторые свойства распределения χ^2.

1. Если $U_1^2, U_2^2, ..., U_l^2$ – независимые случайные величины, имеющие распределение χ^2, с соответственно $\nu_1, \nu_2, ..., \nu_l$ числом степеней свободы, то случайная величина

$$U^2 = \sum_{i=1}^{l} U_i^2$$

также имеет распределение χ^2 с $\nu = \sum_{i=1}^{l} \nu_i$ числом степеней свободы.

2. Если случайная величина U^2 имеет распределение χ^2 с ν степенями свободы, причем $U^2 = \sum_{i=1}^{l} U_i^2$, где U_i^2 имеет также распределение χ^2 с ν_i степенями свободы ($i \in \{1,2,\ldots,l\}$), то необходимым и достаточным условием независимости случайных величин U_i^2 является равенство $\nu = \sum_{i=1}^{l} \nu_i$ (теорема Кохрана).

Последнее свойство широко используется в дисперсионном и регрессионном анализе при оценке существенности влияния различных факторов на результативный признак.

3. Если случайная величина U^2 имеет распределение χ^2 с ν степенями свободы, то при $\nu \to \infty$ закон распределения случайной величины $\sqrt{2U^2}$ сходится к нормальному закону распределения $N\left(\sqrt{2\nu - 1};\ 1\right)$.

Эта сходимость настолько быстра, что уже при $\nu \geq 30$ можно пользоваться нормальным законом.

Отсюда следует, что случайная величина $Z = \sqrt{2U^2} - \sqrt{2\nu - 1}$ имеет нормированное нормальное распределение $N(0,1)$.

Статистики, имеющие распределение χ^2

Закон распределения χ^2 тесно связан с распределением точечных оценок генеральной дисперсии в том случае, когда выборка берется из нормальной генеральной совокупности X.

Пусть из генеральной совокупности X, имеющей нормальный закон распределения $N(\mu, \sigma)$ с известным математическим ожиданием μ, взята выборка объемом n .

Тогда статистика $\chi^2_* = \frac{nS^2_*}{\sigma^2}$, где $S^2_* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, имеет распределение χ^2 с n степенями свободы.

Докажем это положение. Подставив выражение S^2_* в правую часть χ^2_*, получим $\chi^2_* = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \mathrm{Y}_i^2$.

Из второго определения выборки следует, что $\mathrm{Y}_i = \frac{x_i - \mu}{\sigma}$ – есть независимые нормированные нормальные случайные величины $N(0,1)$. Тогда согласно определению случайная величина χ^2_* имеет распределение χ^2 с n степенями свободы, и доказательство закончено.

Теперь докажем, что статистика $U^2 = \chi^2 = \frac{nS^2}{\sigma^2}$ имеет распределение χ^2 с (n-1) степенями свободы.

Известно, что если каждое значение X_i ($i \in \{1,2,\ldots,n\}$) заменить на $X_i^{'} = X_i - \mu$, то дисперсия выборки S^2 не изменится. Поэтому без ограничения общности можно предположить, что X_i — независимые случайные величины с нулевым математическим ожиданием $\mu = 0$ и нормальным законом распределения $N(0,\sigma)$.

Перейдем от X_i к новой системе случайных величин Y_j, $i, j \in \{1,2,\ldots, n\}$, связанных с X_i соотношениями

$$Y_j = \frac{1}{\sqrt{j(j+1)}}\left(\sum_{i=1}^{j} X_i - j \cdot X_{j+1}\right) \text{ для } \mathrm{j} \in \{1,2,\ldots, \mathrm{n\text{-}1}\}$$

$$\text{и } Y_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$$

Тогда

$$Y_1 = \frac{1}{\sqrt{1 \cdot 2}}(X_1 - X_2)$$

$$Y_2 = \frac{1}{\sqrt{2 \cdot 3}}(X_1 + X_2 - 2X_3)$$

……..

……..

$$Y_{n-1} = \frac{1}{\sqrt{(n-1)n}}\left[X_1 + X_2 + \ldots + X_{n-1} - (n-1)X_n\right]$$

$$Y_n = \frac{1}{\sqrt{n}}(X_1 + X_2 + \ldots + X_n).$$

Составим матрицу A коэффициентов преобразования системы $X_1, X_2, \ldots, X_n$:

$$A = \begin{bmatrix} \frac{1}{\sqrt{1\cdot 2}} & -\frac{1}{\sqrt{1\cdot 2}} & 0 & 0 & \ldots\ldots & 0 & 0 \\ \frac{1}{\sqrt{2\cdot 3}} & \frac{1}{\sqrt{2\cdot 3}} & -\frac{2}{\sqrt{2\cdot 3}} & 0 & \ldots\ldots & 0 & 0 \\ \frac{1}{\sqrt{3\cdot 4}} & \frac{1}{\sqrt{3\cdot 4}} & \frac{1}{\sqrt{3\cdot 4}} & -\frac{3}{\sqrt{3\cdot 4}} & \ldots\ldots & 0 & 0 \\ & \ldots\ldots & & & & & \\ & \ldots\ldots & & & & & \\ \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \frac{1}{\sqrt{(n-1)n}} & \ldots\ldots & \frac{1}{\sqrt{(n-1)n}} & -\frac{n-1}{\sqrt{(n-1)n}} \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \ldots\ldots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{bmatrix}$$

Предварительно покажем, что A удовлетворяет так называемому условию ортогональности:

- сумма квадратов элементов одной строки (одного столбца) равна единице

$$\sum_{i=1}^{n} a_{ji}^2 = 1, \quad \sum_{j=1}^{n} a_{ji}^2 = 1;$$

- сумма произведений соответствующих элементов двух параллельных строк (столбцов) равна нулю

$$\sum_{i=1}^{n} a_{ji} a_{ki} = 0 \quad (j \neq k);$$

$$\sum_{j=1}^{n} a_{ji} a_{jk} = 0 \quad (i \neq k).$$

Легко проверить непосредственно, что матрица А удовлетворяет условиям ортогональности. Например, сумма произведений элементов второго и третьего столбцов равна

$$0 - \frac{2}{2\cdot 3} + \frac{1}{3\cdot 4} + \frac{1}{4\cdot 5} + \ldots + \frac{1}{(n-1)n} + \frac{1}{n} = -\frac{1}{3} + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{4} - \frac{1}{5}\right) + \ldots +$$

$$+ \left(\frac{1}{n-1} - \frac{1}{n}\right) + \frac{1}{n} = 0.$$

Ортогональным называют линейное преобразование n-мерного евклидова пространства, сохраняющее длину каждого вектора. Оно связано с вращением n-мерного пространства вокруг начала координат. При таком вращении коэффициенты линейного преобразования связаны соотношениями ортогональности и представляют косинус углов наклона новых осей относительно старых.

Ввиду того, что длина векторов остается неизменной,

$$\sum_{i=1}^{n} X_i^2 = \sum_{j=1}^{n} Y_j^2, \quad \sum_{i=1}^{n} X_i = \sum_{j=1}^{n} Y_j .$$

что легко проверить, возведя в квадрат и складывая уравнения преобразований с учетом условий ортогональности.

Из этих же уравнений и условий $MX_i = 0$ и $DX_i = \sigma^2$ следует, что $MY_j = 0$, $DY_j = \sigma^2$ и $M(Y_k Y_j) = 0$ при $k \neq j$.

Так как Y_j есть линейные функции независимых нормальных случайных величин X_i, то Y_j также имеют нормальный закон распределения $N(0,\sigma)$.

Из того, что Y_j взаимно некоррелированы, $M(Y_k Y_j) = 0$ при $k \neq j$, и имеют в совокупности нормальный закон распределения, следует, что они попарно независимы.

Согласно последнему уравнению преобразования имеем $Y_n = \overline{X}\sqrt{n}$.

Тогда с учетом сохранения расстояния

$$nS^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 = \sum_{i=1}^{n} Y_i^2 - Y_n^2 = \sum_{i=1}^{n-1} Y_i^2.$$

Отсюда следует, что $\chi^2 = \dfrac{nS^2}{\sigma^2} = \sum_{i=1}^{n-1}\left(\dfrac{Y_i}{\sigma}\right)^2$,

где $\dfrac{Y_i}{\sigma}$ - независимые случайные величины, имеющие нормированный нормальный закон распределения $N(0,1)$.

Тогда статистика χ^2 представляет сумму квадратов (n - 1) независимых случайных величин с распределением $N(0,1)$ и согласно определению распределения имеет распределение χ^2 с (n-1) степенями свободы, что и требовалось доказать.

Докажем теперь, что в случае выборки из нормально распределенной генеральной совокупности средняя арифметическая $\overline{X}$ и дисперсия S^2 независимы.

В самом деле, мы доказали, что $Y_1, Y_2, ..., Y_{n-1}$ и $Y_n = \overline{X}\sqrt{n}$ независимы, но тогда величины $\overline{X} = \dfrac{Y_n}{\sqrt{n}}$ и $S^2 = \dfrac{1}{n}\sum_{i=1}^{n-1} Y_i^2$ также независимы.

Таким образом, мы доказали, что в случае выборки из нормальной совокупности $N(\mu, \sigma)$ средняя арифметическая $\overline{X}$ и дисперсия S^2 взаимно независимы, причем $\overline{X}$ имеет распределение $N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$, а $\dfrac{nS^2}{\sigma^2}$ – распределение χ^2 с (n-1) степенями свободы.

Определим теперь другим способом математическое ожидание MS^2 и дисперсию DS^2 статистики S^2. Для случайной величины U^2, имеющей распределение $\chi^2 : MU^2 = \nu, \quad DU^2 = 2\nu$, поэтому

$$M\left(\frac{nS^2}{\sigma^2}\right) = n - 1; \quad D\left(\frac{nS^2}{\sigma^2}\right) = 2(n-1).$$

Тогда, учитывая свойства математического ожидания и дисперсии, будем иметь

$$\frac{n}{\sigma^2} MS^2 = n-1 \text{ и } MS^2 = \frac{n-1}{n}\sigma^2,$$

$$D\left(\frac{nS^2}{\sigma^2}\right) = \frac{n^2}{\sigma^4} DS^2 = 2(n-1) \text{ и } DS^2 = \frac{2(n-1)}{n^2} \cdot \sigma^4.$$

Из полученных формул следует состоятельность оценки S^2.

Статистики, имеющие t – распределение (*Стьюдента*)

Пусть из генеральной совокупности X с нормальным законом распределения $N(\mu, \sigma)$ взята выборка объемом n. Докажем, что статистика $T = \frac{\overline{X} - \mu}{S}\sqrt{n-1}$ имеет распределение Стьюдента с n-1 степенями свободы.

Было доказано, что $\overline{X}$ имеет нормальный закон распределения $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, откуда следует, что статистика $Z = \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ имеет стандартный нормальный закон распределения $N(0,1)$.

Доказано также, что выборочная характеристика $U^2 = \frac{nS^2}{\sigma^2}$ имеет распределение χ^2 с n-1 степенями свободы, причем $\overline{X}$ и S независимы. Но тогда независимы и величины Z и U и, согласно определению распределения, статистика $T = \frac{Z}{U}\sqrt{n-1} = \frac{\overline{X} - \mu}{S}\sqrt{n-1}$ имеет распределение Стьюдента с n-1 степенями свободы.

Рассмотрим еще одну выборочную характеристику, имеющую распределение Стьюдента.

Пусть из генеральной совокупности X с нормальным законом распределения $N(\mu_x, \sigma_1)$ взята выборка объемом n_1, а из генеральной совокупности Y с распределением $N(\mu_y, \sigma_2)$ – выборка объемом n_2. Покажем, что при независимости выборок и равенстве дисперсий $\sigma_1^2 = \sigma_2^2 = \sigma^2$ статистика

$$T = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

имеет распределение Стьюдента с n_1+n_2 - 2 степенями свободы.

Доказано, что случайная величина $\left(\overline{X} - \overline{Y}\right)$ имеет нормальный закон распределения $N\left(\mu_x - \mu_y;\ \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$.

Перейдем к нормированной случайной величине Z с распределением $N(0,1)$

$$Z = \frac{X - Y - (\mu_x - \mu_y)}{\sigma \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}}.$$

Учитывая, что величины $U_1^2 = \frac{n_1 S_1^2}{\sigma^2}$ и $U_2^2 = \frac{n_2 S_2^2}{\sigma^2}$ независимы и имеют распределение χ^2 с числом степеней свободы соответственно $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$, их композиция $U^2 = U_1^2 + U_2^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}$ также имеет распределение χ^2 с $\nu = \nu_1 + \nu_2 = n_1 + n_2 - 2$ числом степеней свободы.

Тогда выборочная характеристика

$$T = \frac{Z}{U}\sqrt{\nu} = \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

имеет распределение Стьюдента с $(n_1 + n_2 - 2)$ числом степеней свободы, что и требовалось доказать.

Этим результатом мы воспользуемся в дальнейшем при проверке гипотезы о равенстве средних μ_x и μ_y генеральных совокупностей X и Y.

Распределение отношения исправленных дисперсий

Докажем, что если $\hat{S}_1^2$ и $\hat{S}_2^2$ - исправленные выборочные дисперсии двух независимых выборок объемом n_1 и n_2 из нормальных генеральных совокупностей X и Y с равными средними квадратическими отклонениями σ, то статистика

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

имеет распределение Фишера-Снедекора с n_1-1 и n_2-1 степенями свободы.

Так как выборочные характеристики $U_1^2 = \frac{(n_1 - 1)\hat{S}_1^2}{\sigma^2}$ и $U_2^2 = \frac{(n_2 - 1)\hat{S}_2^2}{\sigma^2}$ имеют распределение χ^2 с соответственно n_1 - 1 и n_2 - 1 степенями свободы и по условию выборок U_1^2 и U_2^2 независимы, то согласно определению статистика

$$F = \frac{U_1^2}{U_2^2} \cdot \frac{n_2 - 1}{n_1 - 1} = \frac{\hat{S}_1^2}{\hat{S}_2^2}$$

имеет F - распределение с числом степеней свободы n_1 - 1 и n_2 - 1 .

2.3. Методы получения точечных оценок

Рассмотрим методы получения точечных оценок.

Наиболее простым методом получения точечных оценок является метод моментов. Этот метод заключается в приравнивании нескольких первых генеральных моментов к выборочным и решении полученной системы относительно неизвестных параметров, являющихся аргументами приравненных моментов генеральной совокупности. Например, имеем следующую систему двух уравнений

$$\begin{cases} \nu_1 = \nu_1^* \\ \nu_2 = \nu_2^* \end{cases} \text{ или } \begin{cases} \nu_1(MX) = \nu_1^* \\ \nu_2(MX, DX) = \nu_2^* \end{cases}$$

с двумя неизвестными параметрами *MX* и *DX*. Решая такую систему относительно *MX* и *DX,* получим

$$MX = \nu_1^*,$$

$$DX = \nu_2^* - \nu_1^{*2}.$$

Метод, дающий асимптотически нормальные и эффективные оценки, носящий название метода максимального правдоподобия, заключается в следующем.

Находится функция правдоподобия выборки, то есть плотность распределения непрерывной, случайной величины или аналитическая формула ряда распределения дискретной случайной величины, рассматриваемые как известные функции параметров. В силу независимости выборочных наблюдений эти функции представляются как произведения *n* однотипных сомножителей. Задача заключается в том, чтобы найти максимум функции правдоподобия, а именно значения аргументов – параметров функции правдоподобия, сообщающие этой функции максимум. Так как отыскание точек максимума функции связано с дифференцированием и приравниванием к нулю градиента скалярной функции, то вместо максимума исходной функции правдоподобия находят максимум ее логарифма (легче брать производные от суммы, чем от произведения).

К недостаткам метода следует отнести трудности решения системы уравнений.

Применяется также метод наименьших квадратов, который рассматривается в регрессионном анализе.

2.4. Интервальные оценки и их свойства

Рассмотрим вопрос о получении интервальных оценок неизвестного параметра, то есть построение так называемого доверительного интервала, внутри которого находится истинное значение параметра с заданной доверительной вероятностью или надежностью. Метод построения интервальной оценки сводится к отысканию доверительных границ.

Таким образом, интервальной оценкой называют интервал $\left[\underline{\theta}; \overline{\theta}\right]$, границами $\underline{\theta}$ – (нижняя) и $\overline{\theta}$ – (верхняя) которого является пара статистик. Интервальная оценка должна удовлетворять следующим двум требованиям.

1. Оценка должна быть достаточно точной. Точность интервальной оценки есть –

$$\Delta_\theta = \frac{\overline{\theta} - \underline{\theta}}{2},$$

таким образом, длина интервала должна быть достаточно малой.

2. Оценка должна быть достаточно надежной. Надежностью интервальной оценки называют вероятность

$$\gamma = P\left\{\underline{\theta} \le \theta \le \overline{\theta}\right\}$$

накрытия доверительным интервалом неизвестного параметра θ. Такая доверительная вероятность должна быть близкой к единице, чтобы считать событие $\left\{\underline{\theta} \le \theta \le \overline{\theta}\right\}$ практически достоверным.

Найдем нижнюю $\underline{\theta}$ и верхнюю $\overline{\theta}$ границы интервала для параметра θ. Построим две статистики $t_1(\theta)$ и $t_2(\theta)$ такие, чтобы для любого возможного значения параметра θ выполнялось неравенство

$$t_1(\theta) \le T \le t_2(\theta),$$

где T – статистика, оценивающая параметр θ, причем при каждом фиксированном значении параметра θ закон распределения (например, плотность распределения) T известен. Возьмем две вероятности P_1 и P_2, удовлетворяющие условию

$$0 \le P_1 < P_2 \le 1,$$

и найдем для любого значения параметра θ два числа $t_1(\theta)$ и $t_2(\theta)$, отвечающих равенствам

$$P(T<t_1(\theta)) = P_1 \text{ и } P(T \le t_2(\theta)) = P_2,$$

используя, например, квантили таблиц распределения.

Следовательно, статистики $t_1(\theta)$ и $t_2(\theta)$ удовлетворяют условию

$$P\left\{t_1(\theta) \le T \le t_2(\theta)\right\} = P_2 - P_1 = \gamma,$$

и зависят от θ и от γ (или от P_1 и P_2).

Будем считать, что функции t_1 и t_2 параметра θ монотонно возрастают. Возьмем далее наблюдаемое значение статистики T, полученное по выборке. Тогда границы $\underline{\theta}$ и $\overline{\theta}$ будут получаться как единственные решения уравнений

$$t_2(\theta) = T \text{ и } t_1(\theta) = T,$$

что можно записать как

$$\underline{\theta} = t_2^{-1}(T) \text{ и } \overline{\theta} = t_1^{-1}(T).$$

Можно проверить, что неравенства

$$t_1(\theta) \le T \le t_2(\theta) \text{ и } t_2^{-1}(T) \le \theta \le t_1^{-1}(T)$$

равносильны при оговоренных условиях монотонности t_1 и t_2. Следовательно, вероятность второго неравенства такая же, как первого, то есть

$$\mathrm{P}\{\underline{\theta} \le \theta \le \overline{\theta}\} = \mathrm{P}\{\mathrm{t}_1(\theta) \le \mathrm{T} \le \mathrm{t}_2(\theta)\} = \gamma.$$

Таким образом, интервал $[\underline{\theta}, \overline{\theta}]$ является доверительным для параметра θ с надежностью γ и точностью $\Delta_\theta = \frac{1}{2}\left[\overline{\theta} - \underline{\theta}\right]$.

Следует отметить, что из-за зависимости статистик t_1 и t_2 от γ выбор интервала не однозначен. Интервал желательно выбирать таким образом, чтобы точность его была наибольшей, то есть длина наименьшей. Однако, ввиду трудностей такого выбора, на практике используют интервал, для которого $t_1(\theta)$ и $t_2(\theta)$ находятся из условия

$$P\left(T \le t_1(\theta)\right) = P\left(T \ge t_2(\theta)\right) = \frac{\alpha}{2} = \frac{1-\gamma}{2},$$

где $\alpha = 1 - \gamma$ - вероятность ошибки, заключающейся в том, что полученная указанным способом интервальная оценка на самом деле не накрывает неизвестный параметр θ.

Во многих случаях удобно применить более простой метод нахождения интервальных оценок, который называют центральным. Его можно применить тогда, когда мы можем найти некоторую статистику, зависящую от наблюдений и неизвестного параметра, однако распределение которой не зависит от оцениваемого и других параметров. Такая статистика называется центральной, и по ее распределению можно получить неравенство вида

$$\mathrm{t}_1 \le \mathrm{T} \le \mathrm{t}_2,$$

которое затем преобразовывается в неравенство для интервальной оценки.

Интервальная оценка генеральной средней нормально распределенной совокупности при известной дисперсии

Выбираем в качестве статистики T

$$T = Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

предполагая известной генеральную дисперсию σ^2.

Известно, что T распределена нормально по стандартному закону. Следовательно, статистика является центральной, поэтому

$$P\{-t \le T \le t\} = \frac{1}{2}\left[\Phi(t) - \Phi(-t)\right] = \Phi(t) = \gamma.$$

Отсюда

$$\overline{x} - \Phi^{-1}(\gamma)\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{x} + \Phi^{-1}(\gamma)\frac{\sigma}{\sqrt{n}}.$$

Формула

$$\Delta_\mu = \Phi^{-1}(\gamma) \cdot \frac{\sigma}{\sqrt{n}}$$

используется для решения различных задач в зависимости от того, какие буквы в ней считаются заданными и что надо найти. Например, можно найти объем выборки по формуле

$$n = \frac{t_\gamma^2 \cdot \sigma^2}{\Delta_\mu^2},$$

где $t_\gamma = \Phi^{-1}(\gamma)$ является решением уравнения $\Phi(t) = \gamma$ и находится приближенно с помощью таблиц стандартного нормального закона. Следовательно, в этой задаче предполагаются известными $\gamma, \Delta_\mu, \sigma_\mu^2$.

Приведем еще ряд формул, использующих известные распределения статистики.

Интервальная оценка для генеральной средней μ при неизвестной генеральной дисперсии

Оценка имеет вид

$$\bar{x} - \Delta_\mu \le \mu \le \bar{x} + \Delta_\mu,$$

причем

$$\Delta_\mu = \frac{St^{-1}(\alpha, n-1)}{\sqrt{n-1}} S,\ \alpha = 1 - \gamma,$$

где $St^{-1}(\alpha, n\text{-}1)$ находится по таблицам распределения Стьюдента.

Формулы для нахождения интервальной оценки генеральной доли p

При достаточно больших объемах n выборки из генеральной совокупности, распределенной по биномиальному закону, аппроксимируемому нормальным (интегральная теорема Муавра-Лапласа), используется формула

$$P\left(\left|\frac{m}{n} - p\right| \le \Delta_p\right) = \Phi(t_\gamma),$$

где

$$\Delta_p = t_\gamma \sqrt{\frac{p(1-p)}{n}},\quad t_\gamma = \Phi^{-1}(\gamma).$$

Для получения интервальной оценки решается квадратное неравенство

$$\left|\frac{m}{n} - p\right|^2 \le t_\gamma^2 \frac{p(1-p)}{n}.$$

Решение этого неравенства

$$\underline{p} \le p \le \overline{p}$$

дает интервальную оценку с надежностью γ ($\underline{p}$ и $\overline{p}$ - корни квадратного трехчлена).

При сравнительно больших n ($n > 100$) иногда используется интервальная оценка

$$\frac{m}{n} - t_\gamma \sqrt{\frac{w(1-w)}{n}} \le p \le \frac{m}{n} + t_\gamma \sqrt{\frac{w(1-w)}{n}}, \text{ где } w = \frac{m}{n}.$$

При $m = 0$ и $m = n$ интервальные оценки имеют вид соответственно:

В других случаях используется так называемая неполная бета-функция или F–распределение. Однако, информации, имеющейся в таблицах F – распределения, может не

$$0 \le p \le 1 - \sqrt[n]{\alpha}, \quad \sqrt[n]{\alpha} \le p \le 1.$$

хватить для точного нахождения интервальной оценки.

Интервальные оценки генеральной дисперсии σ^2 и среднего квадратического отклонения σ

Пусть из генеральной совокупности X, распределенной по нормальному закону $N(\mu, \sigma)$, взята выборка объемом n и получена выборочная дисперсия S^2. Требуется определить с надежностью γ интервальные оценки для генеральной дисперсии σ^2 и среднего квадратичного отклонения σ.

При построении доверительного интервала используют статистику $\frac{nS^2}{\sigma^2}$, которая имеет распределение χ^2 с $n - 1$ степенями свободы. По таблице распределения χ^2 всегда можно найти два числа $\chi_1^2 = U_1^2$ и $\chi_2^2 = U_2^2$, для которых выполняется соотношение

$$P\left(\chi_1^2 < \frac{nS^2}{\delta^2} < \chi_2^2\right) = \gamma.$$

Существует бесчисленное множество пар чисел, удовлетворяющих этому условию. Как отмечалось, используют доверительный интервал, для которого

$$P\left(\chi^2 < \chi_1^2\right) = P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2} = \frac{\alpha}{2}.$$

После преобразований получим окончательные выражения для интервальной оценки σ^2:

$$\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}, \text{ причем } P\left(\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}\right) = \gamma,$$

и интервальной оценки σ

$$\sqrt{\frac{nS^2}{\chi_2^2}} \le \sigma \le \sqrt{\frac{nS^2}{\chi_1^2}}, \text{ причем } P\left(\frac{\sqrt{n}S}{\chi_2} \le \sigma \le \frac{\sqrt{n}S}{\chi_1}\right) = \gamma.$$

Так как таблицы распределения Пирсона χ^2 содержат значения $\chi_\text{т}^2$, удовлетворяющие условию $P(\chi^2 > \chi_\text{т}^2) = \alpha$, то граничные значения χ_1^2 и χ_2^2 определяют для числа степеней свободы $\nu = n - 1$ из соотношений

$$P\left(\chi^2 > \chi_1^2\right) = \frac{1+\gamma}{2};$$

$$P\left(\chi^2 > \chi_2^2\right) = \frac{1-\gamma}{2}.$$

Следовательно, $\chi_1^2 = P_i^{-1}\left(1-\frac{\alpha}{2}, n-1\right), \chi_2^2 = P_i^{-1}\left(\frac{\alpha}{2}, n-1\right).$

Как уже отмечалось, при достаточно больших объемах выборки (n-1> 30) можно считать, что случайная величина $\sqrt{2\chi^2}$ имеет нормальное распределение $N\left(\sqrt{2\nu-1}; 1\right).$

Поэтому для построения доверительного интервала σ (или σ^2 с границами – квадратами границ σ) при n-1> 30 используют статистику

$$T = \sqrt{2\frac{nS^2}{\sigma^2}} - \sqrt{2n-3},$$

которая имеет асимптотическое стандартное нормальное распределение $N(0,1)$. Тогда, по таблице Φ(t) можно определить значение t_γ, для которого

$$\text{P}(\mid \text{T} \mid \le \text{t}_\gamma) = \Phi(\text{t}_\gamma) = \gamma.$$

Раскрыв неравенство, получим

$$-t_\gamma \le \frac{S}{\sigma}\sqrt{2n} - \sqrt{2n-3} \le t_\gamma.$$

Отсюда после преобразования найдем доверительный интервал для среднего квадратического отклонения:

$$\frac{\sqrt{2n}}{\sqrt{2n-3}+t_\gamma} \cdot S \le \sigma \le \frac{\sqrt{2n}}{\sqrt{2n-3}-t_\gamma} \cdot S.$$

В заключение заметим, что если θ_n^* – оценка максимального правдоподобия параметра θ, то при достаточно больших n интервальная оценка параметра θ определяется соотношением

$$P\left(\theta_n^* - t_\gamma\sqrt{D\theta_n^*} \le \theta \le \theta_n^* + t_\gamma\sqrt{D\theta_n^*}\right) = \gamma,$$

где γ – надежность, доверительная вероятность оценки, $t_\gamma = \Phi^{-1}(t_\gamma)$, точность $\Delta_\theta = t_\gamma\sqrt{D\theta_n^*}$, и, таким образом, имеет вид:

$$\theta_n^* - \Delta_\theta \le \theta \le \theta_n^* + \Delta_\theta.$$

2.5. Пояснения, примеры и решения задач

1. Покажем, что частость $\frac{m}{n}$ появления события A является несмещенной оценкой вероятности p появления события A в отдельном испытании. Действительно, пусть имеется выборка объемом n

$$X_1, X_2, ..., X_n,$$

причем $X_1, X_2, ..., X_n$ – независимые случайные величины, значит появление события A в каждом испытании не зависит от появления события A в других испытаниях. Если предполагается, что событие A в каждом испытании появляется с одной и той же вероятностью p, то $MX_i = p$. Кроме того, частота появления события A, то есть $m = \sum_{i=1}^{n} X_i$, а $\frac{m}{n} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. По доказанному ранее средняя арифметическая, полученная по выборке из одного и того же распределения, является несмещенной оценкой математического ожидания, то есть $M\left(\frac{m}{n}\right) = \frac{np}{n} = p$.

Кроме того, так как $D\left(\frac{m}{n}\right) = \frac{pq}{n}$, частость $\frac{m}{n}$ является состоятельной оценкой вероятности.

2. Найдем максимально правдоподобную оценку $T_n(X_1, X_2, ..., X_n)$ параметра λ закона Пуассона. Составим функцию правдоподобия (или закон распределения выборки)

$$p(x_1, x_2, ..., x_n, \lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!}e^{-\lambda} ... \frac{\lambda^{x_n}}{x_n!}e^{-\lambda}.$$

Логарифмируя p, получим

$$L = \ln p(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\ln\lambda - \ln\prod_{i=1}^{n} x_i!.$$

Продифференцируем это уравнение по параметру λ и приравняем производную к нулю, получим

$$\frac{\partial L}{\partial \lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0.$$

Отсюда

$$\lambda^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Следовательно, средняя арифметическая есть максимально-правдоподобная оценка параметра λ закона Пуассона.

3. Найдем оценки максимального правдоподобия для μ и σ^2 нормального закона распределения.

Имеем функцию правдоподобия и её натуральный логарифм (*ln*):

$$p(x_1,...,x_n,\mu,\sigma^2) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$L = \ln p(\mu,\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Найдем частные производные от L по μ и по σ^2, приравняем их к нулю, получим следующую систему уравнений

$$\begin{cases} \dfrac{\partial L}{\partial \mu} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0, \\ \dfrac{\partial L}{\partial \sigma^2} = \dfrac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 - \dfrac{n}{2\sigma^2} = 0 \end{cases}.$$

Решением являются следующие оценки максимального правдоподобия:

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$(\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = S^2.$$

4. Приведем типичные примеры вычисления интервальных оценок.

Пример 1. По результатам $n = 10$ измерений установлено, что средняя длина между опорами ЛЭП равна $\bar{x} = 85{,}5$ м. В предположении, что ошибки измерения распределены по нормальному закону со средним квадратическим отклонением $\sigma = 0{,}5м$, определить с надежностью $\gamma = 0{,}95$ интервальную оценку для генеральной средней μ.

Решение. Поскольку параметр σ нам известен, интервальную оценку будем искать по формуле, включающей σ.

По таблице интегральной функции Лапласа из условия $\Phi(t_\gamma) = 0{,}95$ найдем $t_\gamma = 1{,}96$.

Тогда точность оценки равна

$$\Delta_\mu = t_\gamma \frac{\sigma}{\sqrt{n}} = 1{,}96 \frac{0{,}5}{\sqrt{10}} = 0{,}31.$$

Откуда доверительный интервал имеет вид

$$85{,}5\text{-}0{,}31 \le \mu \le 85{,}5\text{+}0{,}31$$

и окончательно

$$85{,}19 \le \mu \le 85{,}81 \text{ (м)}.$$

Пример 2. Решим пример 1 при условии, что параметр σ неизвестен, а выборочное среднее квадратическое отклонение $S = 0{,}5$м.

Решение. Так как σ нам неизвестно, то интервальную оценку генеральной средней μ будем искать по формуле, включающей S. Из таблиц t – распределения для числа степеней свободы $\nu = n\text{-}1 = 9$ и $\alpha = 1\text{-}\gamma = 0{,}05$ найдем $St^{-1}(\alpha = 0{,}05, \nu = 10\text{-}1) = 2{,}262$.

Тогда точность оценки равна

$$\Delta_\mu = 2{,}26 \frac{0{,}5}{\sqrt{9}} = 0{,}38.$$

Откуда доверительный интервал имеет вид:

$$85{,}5\text{-}0{,}38 \le \mu \le 85{,}5\text{+}0{,}38,$$

и окончательно

$$85{,}12 \le \mu \le 85{,}88\text{(м)}.$$

Пример 3. По результатам контроля $n = 10$ деталей вычислено выборочное среднее квадратическое отклонение $S = 10$ мк. В предположении, что ошибка изготовления распределена нормально, определить с надежностью $\gamma = 0{,}95$ доверительный интервал для параметра σ.

Решение. Имеем $P\left(\chi^2 \ge \chi_1^2\right) = 0{,}975;\ P\left(\chi^2 \ge \chi_2^2\right) = 0{,}025.$

По таблице распределения χ^2 для числа степеней свободы $n\text{-}1 = 9$ и найденных вероятностей определим $\chi_1^2 = 2{,}700$ и $\chi_2^2 = 19{,}023$.

Искомый доверительный интервал есть

$$\frac{\sqrt{10} \cdot 10}{4{,}36} \le \sigma \le \frac{\sqrt{10} \cdot 10}{1{,}64}$$

или

$$7{,}25 \le \sigma \le 19{,}25 \text{ (мк)}.$$

Пример 4. При испытании зерна на всхожесть из $n = 300$ зерен проросло $m = 225$. С надежностью $\gamma = 0{,}95$ определить доверительный интервал для генеральной доли p проросших зерен.

Решение. По таблице интегральной функции Лапласа из условия $\Phi(t_\gamma) = 0{,}95$ определим $t_\gamma = 1{,}96$.

Учитывая, что $\frac{m}{n} = 0{,}75$, определим точность оценки

$$\Delta_p = 1{,}96\sqrt{\frac{0{,}75 \cdot 0{,}25}{300}} = 0{,}049.$$

Тогда доверительный интервал имеет вид

$$0{,}75 - 0{,}049 \le p \le 0{,}75 + 0{,}049,$$

$$0{,}701 \le p \le 0{,}799.$$

5. Определим минимальный объем выборки n, при котором можно найти интервальную оценку генеральной средней μ с известным средним квадратическим отклонением σ и с заданными доверительной вероятностью γ и точностью Δ_μ. Решение получается по формуле:

$$n = \frac{t_\gamma^2 \sigma^2}{\Delta_\mu^2}.$$

Пусть $\sigma = 2$, $\gamma = 0{,}95$, $\Delta_\mu = 0{,}5$. Тогда $t_\gamma = \Phi^{-1}(0{,}9500) = 1{,}96$ и

$$n = \frac{1{,}96^2 \cdot 2^2}{0{,}5^2} = 61{,}4656 = 62.$$

Ответ целесообразно записывать в виде ближайшего целого числа, превосходящего полученное нецелое.

Вычислим надежность γ интервальной оценки μ при $\Delta_\mu = 1$, $n = 50$, $\sigma = 2$. Используем формулу

$$t_\gamma = \frac{\Delta_\mu \sqrt{n}}{\sigma}$$

и получаем

$$t_\gamma = \frac{1 \cdot \sqrt{50}}{2} = 3{,}5355 = 3{,}54.$$

Отсюда

$$\gamma = \Phi(t_\gamma) = \Phi(3{,}54) = 0{,}9996.$$

6. Найдем интервальную оценку с надежностью $\gamma = 0{,}90$ для генерального среднего квадратического отклонения σ, используя точную и приближенную формулы, при исходных данных: $n = 31$, $S^2 = 10$.

Точное решение получим, используя табличные значения распределения хи-квадрат Пирсона.

$$\chi_2^2 = P_i^{-1}(1 - \frac{\alpha}{2}; n-1) = P_i^{-1}(0{,}95;30) = 43{,}773,$$

$$\chi_1^2 = P_i^{-1}(\frac{\alpha}{2}; n-1) = P_i^{-1}(0{,}05;30) = 18{,}493.$$

Тогда интервальная оценка получается по формуле

$$\frac{nS^2}{\chi_2^2} \le \sigma^2 \le \frac{nS^2}{\chi_1^2}.$$

После подстановки значений всех символов будем иметь

$$\frac{31 \cdot 10}{43{,}773} \le \sigma^2 \le \frac{31 \cdot 10}{18{,}493},$$

или

$$7{,}0820 \le \sigma^2 \le 16{,}7631,$$

откуда

$$2{,}661 \le \sigma \le 4{,}094.$$

Таким образом, с доверительной вероятностью (надежностью) 0,90 можно утверждать, что генеральное среднее квадратическое отклонение признака, распределенного нормально, лежит в границах от 2,661 до 4,094 единиц его измерения.

Теперь воспользуемся приближенной формулой интервальной оценки σ^2:

$$\frac{2nS^2}{(\sqrt{2n-3} + t_\gamma)^2} \le \sigma^2 \le \frac{2nS^2}{(\sqrt{2n-3} - t_\gamma)^2}.$$

Найдем $t_\gamma = \Phi^{-1}(0{,}9000) = \Phi^{-1}(0{,}8910) = 1{,}64$ по таблицам $\Phi(t)$ стандартного нормального распределения, подставим численные значения в формулу интервальной оценки:

$$\frac{2 \cdot 31 \cdot 10}{(\sqrt{2 \cdot 31 - 3} + 1{,}64)^2} \le \sigma^2 \le \frac{2 \cdot 31 \cdot 10}{(\sqrt{2 \cdot 31 - 3} - 1{,}64)^2}$$

и после вычислений получим

$$7{,}135971 \le \sigma^2 \le 16{,}988423,$$

откуда интервальная оценка для σ будет иметь вид:

$$2{,}671 \le \sigma \le 4{,}122.$$

Таким образом, эта оценка менее точна, чем предыдущая.

7. Решим обратную задачу об интервальной оценке неизвестной генеральной дисперсии σ^2 для $X \sim N(\mu, \sigma^2)$.

Пусть по выборке объёма n = 31 найдена выборочная дисперсия $S^2 = 10$. Для заданной интервальной оценки $\sigma^2 \in [9; 16]$ найти надежность γ. Для решения воспользуемся равенствами

$$P_1 = P\left(\chi^2 > \frac{nS^2}{\underline{\sigma}^2}\right) = P_i\left(\frac{nS^2}{\underline{\sigma}^2}; n-1\right) = P_i\left(\frac{310}{9}; 31-1\right) =$$

$$= P_i(34{,}444; 30) = P_i(34{,}800) = 0{,}25\,,$$

$$P_2 = P\left(\chi^2 > \frac{nS^2}{\overline{\sigma}^2}\right) = P\left(\chi^2 > \frac{310}{16}\right) = P_i(18{,}493) = 0{,}95\cdot$$

При расчете используем табличные значения распределения хи-квадрат, наиболее близкие к вычисленным. В результате получаем доверительную вероятность

$$\gamma = P_2 - P_1 = 0{,}95 - 0{,}25 = 0{,}70.$$

Теперь определим надежность γ, считая объём выборки $n = 31$ большим, то есть используя асимптотическое распределение:

$$T = \sqrt{\frac{2nS^2}{\sigma^2}} - \sqrt{2n-3} \sim N(0;1)\,.$$

Получаем

$$t_1 = \sqrt{\frac{2nS^2}{\overline{\sigma}^2}} - \sqrt{2n-3} = \sqrt{\frac{2\cdot 31\cdot 10}{16}} - \sqrt{2\cdot 31-3} = -1{,}46,$$

$$t_2 = \sqrt{\frac{2nS^2}{\underline{\sigma}^2}} - \sqrt{2n-3} = \sqrt{\frac{2\cdot 31\cdot 10}{9}} - \sqrt{2\cdot 31-3} = 062.$$

Используя таблицу Φ(t) стандартного нормального распределения, будем иметь

$$\gamma = \frac{1}{2}\left[\Phi(t_2) + \Phi(t_1)\right] = \frac{1}{2}\left[\Phi(0{,}62) + \Phi(1{,}46)\right] =$$

$$= \frac{1}{2}\left[0{,}4647 + 0{,}8557\right] = 0{,}6602.$$

Таким образом, ответы по точной формуле с использованием ближайшего табличного значения хи-квадрат и приближенной формуле совпадают с точностью до одного знака после запятой.

8. Применение метода получения формул вычисления нижней и верхней доверительных границ неизвестного параметра генеральной совокупности в случае генеральной доли p биномиального закона распределения приводит к решению уравнений:

$$\sum_{j=0}^{m} C_n^j \overline{p}^j (1-\overline{p})^{n-j} = \frac{\alpha}{2}, \sum_{j=m}^{n} C_n^j \underline{p}^j (1-\underline{p})^{n-j} = \frac{\alpha}{2},$$

где $P(\underline{p} \le p \le \overline{p}) = \gamma = 1-\alpha.$

Известно, что левые части уравнений выражаются в терминах F- распределения Снедекора – Фишера. В результате применения табличных значений F- распределения получаются следующие формулы, по которым вычисляются точные границы генеральной доли:

$$\underline{p} = \frac{mF_{1-\frac{\alpha}{2}}(2m,2(n-m+1))}{n-m+1+mF_{1-\frac{\alpha}{2}}(2m,2(n-m+1))},$$

$$\overline{p} = \frac{(m+1)F_{\frac{\alpha}{2}}(2(m+1),2(n-m))}{n-m+(m+1)F_{\frac{\alpha}{2}}(2(m+1),2(n-m))},$$

где $F_p(\nu_1,\nu_2)$ – табличное значение статистики F, удовлетворяющее условию:

$$P(F>F_p(\nu_1,\nu_2)) = F_i(F_p(\nu_1,\nu_2)),$$

ν_1 – число степеней свободы числителя, ν_2 - число степеней свободы знаменателя. Напомним, что

$$F_p(\nu_1,\nu_2) = \frac{1}{F_{1-p}(\nu_2,\nu_1)}.$$

Найдем доверительные границы для неизвестной вероятности $\rho = P(A)$ события А, если при $n = 30$ независимых испытаниях оно появилось $m = 10$ раз, с доверительной вероятностью $\gamma = 0{,}90$, используя точные и приближенные формулы.

Границы, получаемые по точным формулам с применением таблицы распределения Фишера – Снедекора, следующие:

$$\underline{p} = \frac{10 \cdot F_{0,95}(20,42)}{21+10 \cdot F_{0,95}(20,42)} = \frac{10}{21 \cdot F_{0,05}(42,20)+10} =$$

$$= \frac{10}{21 \cdot F_{0,05}(24,20)+10} = \frac{10}{21 \cdot 2{,}08+10} = 0{,}186$$

$$\overline{p} = \frac{11 \cdot F_{0,05}(22,40)}{20 + 11 \cdot F_{0,05}(22,40)} = \frac{11 \cdot F_{0,05}(24,30)}{20 + 11 \cdot F_{0,05}(24,30)} = \frac{11 \cdot 1{,}89}{20 + 11 \cdot 1{,}89} = 0{,}510.$$

При вычислении использовались ближайшие табличные значения. Значения $F_{0,05}(42,20) = 1{,}99$ и $F_{0,05}(22,40) = 1{,}82$, получаемые интерполяцией значений более подробных таблиц F- распределения, дают точные границы $\underline{p} = 0{,}193$ и $\overline{p} = 0{,}500$, таким образом, оценка [0,186;0,510] менее точна, чем [0,193;0,500].

Применим далее приближенную формулу интервальной оценки как решения квадратного неравенства

$$\left|\frac{10}{30} - p\right|^2 \le 1{,}64^2 \cdot \frac{p(1-p)}{30},$$

где $\Phi^{-1}(0{,}9000) = \Phi^{-1}(0{,}8990) = 1{,}64$.

После преобразований получим неравенство с коэффициентами, округленными до трех знаков после запятой

$$1{,}090p^2 - 0{,}756p + 0{,}111 \le 0.$$

Решение этого неравенства

$$0{,}211 \le p \le 0{,}483.$$

Полученный интервал [0,211; 0,483] совпадает с интервалом [0,193; 0,500], вычисленным по точным формулам с точностью до 0,1.

Наконец, найдем интервальную оценку с помощью простейшей приближенной формулы

$$\mathrm{w} - \Delta_p^* \le p \le \mathrm{w} + \Delta_p^*,$$

где $\mathrm{w} = m/n$, $\Delta_p^* = \Phi^{-1}(\gamma)\sqrt{\dfrac{\mathrm{w}(1-\mathrm{w})}{n}}$

Получим $\mathrm{w} = 10/30 = 0{,}333$, $\Delta_p^* = 1{,}64 \cdot \sqrt{\dfrac{2}{270}} = 0{,}012$ и $\underline{p} = 0{,}333 - 0{,}012 = 0{,}321$, $\overline{p} = 0{,}333 + 0{,}012 = 0{,}345$. Интервал не сравним с точным: использование простой формулы непригодно ввиду недостаточно большого объема выборки $n = 30$.

9. Иногда требуется найти только одну доверительную границу, так как другая не представляет интереса и (или) очевидна. В таком случае полагают вероятности $P_1 = 0$ или $P_2 = 1$, то есть сосредотачивают α целиком на соответствующем хвосте распределения. Например, из партии электроламп была взята на контроль 121 штука. По результатам контроля оказалось, что средняя продолжительность горения электролампы составляет 750 часов, а среднее квадратическое отклонение – 17 часов. Найдем с доверительной вероятностью 0,95 минимальное значение среднего срока службы электролампы в партии, считая, что продолжительность горения электролампы подчиняется нормальному закону распределения вероятностей. Для решения используем распределение Стьюдента, исходя из

данных: $n = 121$, $\overline{x} = 750$, $s = 17$. Найдем для одной границы $St^{-1}(2\alpha,\ n-1) = St^{-1}(0{,}1;\ 120) = 1{,}658$, тогда

$$\underline{\mu} = \overline{X} - St^{-1}(2\alpha, n-1)\frac{s}{\sqrt{n-1}} = 750 - 1{,}658\frac{17}{\sqrt{120}} = 750 - 2{,}573 = 747(час).$$

Можно воспользоваться нормальным законом распределения статистики $\overline{X}$, полагая S = σ. Получим

$$\underline{\mu} = \overline{X} - \Phi^{-1}(1-2\alpha)\frac{\sigma}{\sqrt{n}} = 750 - 1{,}64 \cdot \frac{17}{11} = 750 - 2{,}53 = 747(час).$$

Как было указано в 2.4, для отыскания интервальной оценки генеральной доли при возникновении ситуации, когда статистика m принимает крайние значения 0 или n, вычисляется только одна доверительная граница. Например, если $n = 10$ и $\gamma = 0{,}9$, то в случае $m = 0$ получим $0 \le p \le 1 - \sqrt[10]{0{,}1}$, $0 \le p \le 1 - 0{,}7943$, а в случае $m = 10$ будем иметь $0{,}7943 \le p \le 1$.

10. Формула

$$\Delta_p = t_\gamma \sqrt{\frac{p(1-p)}{n}} = t_\gamma \cdot \sqrt{D\left(\frac{m}{n}\right)},\ t_\gamma = \Phi^{-1}(\gamma)$$

часто используется для определения (минимально необходимого) объема выборки n при исследовании единицы совокупности, каждая из которых обладает некоторым свойством с неизвестной вероятностью p. Если по предварительным исследованиям или экспертным оценкам генеральная доля p равна p_0, то применяется формула

$$n = \frac{t_\gamma^2 p_0(1-p_0)}{\Delta_p^2}.$$

Если же оценка p_0 отсутствует, то используется равенство $\min D\left(\frac{m}{n}\right) = \frac{1}{4n}$ и формула объема выборки имеет вид

$$n = \frac{t_\gamma^2}{4\Delta_p^2}.$$

Например, если $\gamma = 0{,}9973$, то $t_\gamma = 3{,}00$ и при $\Delta_p = 0{,}01$ получим

$$n = \frac{9}{4 \cdot 0{,}0001} = 22500,$$

если к тому же $p_0 = 0{,}1$, то получаем значительно меньший объем выборки

$$n = \frac{9 \cdot 0{,}1 \cdot 0{,}9}{0{,}0001} = 8100.$$

Интервальная оценка может быть интерпретирована следующим образом: с доверительной вероятностью (надежностью) γ справедливо утверждение о том, что генеральный параметр θ находится в границах (пределах) от $\underline{\theta}$ до $\overline{\theta}$. При этом вероятность ошибки этого утверждения равна $\alpha = 1 - \gamma$ или, по-другому, вероятность события, состоящего в том, что доверительный интервал не накроет оцениваемый параметр θ, равна α. Далее, надежность γ означает, что в среднем (в целом, приблизительно) в $\gamma \cdot 100$ случаев из 100 указанная интервальная оценка содержит параметр θ, а в $\alpha \cdot 100$ случаев - не содержит.

2.6. Упражнения

1. Доказать, что выборочные начальные моменты *k*-го порядка являются несмещенными оценками соответствующих генеральных начальных моментов.

2. Доказать, что выборочные начальные моменты являются состоятельными оценками соответствующих генеральных моментов.

3. Найти значение параметра *p*, при котором дисперсия частости события *A* принимает максимальное значение, и само это максимальное значение.

4. В результате независимых наблюдений над случайной величиной *X* получена выборка объемом *n*. Найти оценку максимального правдоподобия неизвестного параметра *X* в предположении, что случайная величина *X* имеет показательный закон распределения с плотностью $p(x) = \alpha e^{-\alpha x}$ при $x \geq 0$ и $p(x) = 0$ при $x < 0$.

5. Данные об урожайности ржи на 10 опытных участках одинакового размера приведены в таблице.

Урожайность (ц/га) x	16,8	16,0	18,9	15,7	19,1	16,7
Число участков *m*	2	1	2	2	2	1

Найти интервальные оценки для неизвестных параметров μ и σ нормальной генеральной совокупности *X* с надежностью

$\gamma_1 = 0{,}99;\ \ \gamma_2 = 0{,}9;\ \ \gamma_3 = 0{,}95.$

6. Средний размер основных фондов у 39 малых предприятий составляет $\overline{X} = 0{,}85\ млн.руб.$, а $S = 0{,}09\ млн.руб.$ В предположении о нормальном распределении определить с надежностью 0,98 доверительный интервал для σ.

7. С надежностью γ определить границы генеральной доли малых предприятий региона, нарушающих инструкции по отчетности, если среди 100 проверенных предприятий 40 оказались нарушителями инструкции, $\gamma = 0{,}95$.

8. Средний вес мешка цемента, полученный по выборке 10 мешков из железнодорожного вагона, составил 70 *кг*, а среднее квадратическое отклонение – S = 1,5 *кг*. Найти доверительную границу среднего веса мешка в вагоне с надежностью 0,95. Найти доверительную вероятность того, что среднее квадратическое отклонение веса мешка в вагоне не превзойдет 4 кг. Вес мешка цемента в вагоне распределен по нормальному закону.

9. Найти интервальную оценку генеральной доли p, если частота m при n = 8 равна 1) 8 и 2) 0 с доверительной вероятностью $\gamma = 0{,}95$.

10. С надежностью γ найти нижнюю и верхнюю границы среднего процента успеваемости учащихся в колледже, если из 10 наудачу выбранных учащихся 7 успешно ответили на вопросы контрольного теста. Использовать точную $\gamma = 0{,}90$ и приближенную $\gamma = 0{,}9545$ формулы интервальной оценки параметра биномиального закона распределения.

3. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

3.1. Понятия статистической гипотезы и статистического критерия

В математической статистике существует класс задач, состоящих в определении правила, по которому любому результату наблюдений (выборки объема n) соответствует одно и только одно принимаемое решение. Задачи с двумя решениями обычно относятся к классу задач статистической проверки статистической гипотезы, которая должна быть принята или отвергнута.

Для простоты изложения и обозначений будем рассматривать выборку из генеральной совокупности с одним признаком – одномерной случайной величиной X, заданной функцией распределения $F(x)$. Множество выборок $\vec{x} = (x_1, x_2, ..., x_i, ..., x_n)^T$ образует пространство выборок R_n.

Будем предполагать, что пространство выборок R_n совладеет с n-мерным евклидовым пространством (т.е. прямой R_1 при $n = 1$, плоскостью R_2 при $n = 2$, "физическим" пространством R_3 при $n = 3$ и т.д.). Такой подход не нарушит общности рассуждений, если положить вероятность события $\vec{X} = \vec{x}$ или плотность $p(\vec{x})$ в любой точке $\vec{x} = (x_1, ..., x_n)^T$ пространства выборок, для которой указанное событие невозможно, равным нулю.

Построение и развитие любой научной теории сводится к выдвижению и практической проверке гипотез, утверждений, отражающих некоторые свойства изучаемых объектов или отношения между ними. Например, к научным гипотезам относятся гипотеза о существовании внеземной цивилизации или утверждение о смене одного общественно-политического строя другим, более прогрессивным. Гипотезы, рассматриваемые в математической статистике, отличаются от вышеприведенных.

Статистическая гипотеза является частным случаем общенаучной гипотезы и представляет собой утверждение, высказываемое о свойствах распределения признаков в генеральной совокупности или при анализе выборок из генеральной совокупности, т.е. касающееся поведения наблюдаемых случайных величин. Дадим более точную формулировку. Пусть имеется выборка $\vec{x} = (x_1, x_2, ..., x_i, ..., x_n)^T$. Всякое непротиворечивое множество предположений о виде распределения вероятностей (вероятностной меры) системы n случайных величин $(x_1, x_2, ..., x_i, ..., x_n)$ (которые могут быть и векторами одной и той же размерности K) называется статистической гипотезой.

Примерами статистических гипотез являются следующие утверждения:

1. Нормальное распределение признака X с известной дисперсией генеральной совокупности имеет генеральную среднюю равной μ_0.

2. Генеральная доля альтернативного признака X удовлетворяет неравенству $p \le p_0$, где p_0 - заданная доля (стандарт).

3. Нормальное распределение признака X с неизвестной средней генеральной совокупности имеет генеральную дисперсию равной σ_0^2.

4. Распределение признака X в генеральной совокупности является нормальным с заданными параметрами $MX = \mu_0$ и $DX = \sigma_0^2$.

5. Две независимые выборки $(x_1, x_2, ..., x_{n_1})$ и $(x_1', x_2', ..., x_{n_2}')$ взяты из одной и той же генеральной совокупности распределения признака X с непрерывной и неизвестной функцией распределения $F(x)$.

Статистическая гипотеза называется простой, если в соответствующем ей утверждении распределение признака (признаков) в генеральной совокупности определено однозначно. Например, гипотезы (1) и (4) являются простыми в силу того, что нормальный закон распределения вероятностей случайной величины X (одномерной) определяется заданием двух параметров, например, MX и DX, которые в утверждениях гипотез имеют конкретные значения. В противном случае гипотеза называется сложной. Сложную статистическую гипотезу можно определить и как непротиворечивое множество простых гипотез, содержащее более одной простой гипотезы. Сложными гипотезами будут гипотезы (2), (3) и (5). Гипотеза (2) является сложной из-за множества различных значений параметра p, определяющего биномиальный закон распределения выборочной доли $\left(\frac{m}{n}\right)$ или частоты (m) единиц в выборке, обладающих изучаемым признаком, и удовлетворяющего неравенству, содержащемуся в гипотезе. Гипотеза (3) является сложной из-за того, что согласно ей один из определяющих нормальный закон параметров (MX) может быть равен, вообще говоря, любому действительному числу. Наконец, гипотеза (5) - сложная из-за неопределенности функции распределения $F(x)$, непосредственно сформулированной в утверждении.

Проверяемая гипотеза часто называется нулевой и обозначается через H_0.

Для уточнения проверяемой гипотезы и самого правила проверки, наряду с гипотезой H_0, рассматривают гипотезу, являющуюся логическим отрицанием H_0. Эта гипотеза называется альтернативной или конкурирующей гипотезой и обозначается через H_1. Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез.

Правило проверки гипотезы называют критерием гипотезы. Оно состоит в следующем.

Все пространство выборок R_n делится на две взаимно-дополняющие области W_n и $R_n\overline{W_n}$. Область W_n называется критической областью или областью отвержения гипотезы; область $R_n\overline{W_n}$ – областью принятия гипотезы. Производится выборка и проверяется, в какую область эта выборка попадает. Если выборка попала в критическую область, проверяемая гипотеза (H_0) отвергается, т.е. принимается логическое отрицание гипотезы (H_1). Если данная выборка попала в дополнительную область $R_n\overline{W_n}$, проверяемая гипотеза (H_0) принимается.

Иногда критическую область (как и само правило) называют критерием гипотезы.

Требования, предъявляемые к критерию

Всякая простая гипотеза $H^{\text{П}}$ полностью определяет распределение выборок, конкретнее, функцию правдоподобия $f(\vec{x}/H^{\text{П}})$. Всякой сложной гипотезе H соответствует класс функций правдоподобия $f(\vec{x}/H)$ так, что любой простой гипотезе $H^{\text{П}}$ из семейства H, образующего сложную гипотезу, отвечает конкретная функция правдоподобия из класса таких функций.

Зная функцию правдоподобия, можно в принципе определить вероятность попадания выборки в любую область пространства.

Естественно, что критическая область W_n строится исходя из проверяемой гипотезы H_0.

Известно, что события, вероятность которых достаточно мала, в соответствии с принципом практической невозможности считаются невозможными. Пусть α – определенная заранее верхняя граница вероятности редких событий (т.е. событие A считается практически невозможным, если $P(A) \le \alpha$).

Область W_n строится таким образом, чтобы вероятность попадания в нее выборки $\vec{X}$ не превосходила α при справедливости проверяемой гипотезы, т.е. событие, состоящее в попадании выборки $\vec{X}$ в критическую область, в случае истинности проверяемой гипотезы (ошибка первого рода) должно быть практически невозможным.

Вероятность α – вероятность ошибки первого рода – называют уровнем значимости критерия. Вероятность попадания выборки в заданную критическую область в случае сложной проверяемой гипотезы будет, вообще говоря, разной (не превосходящей α) для различных простых гипотез, составляющих сложную. Верхняя граница этих вероятностей обычно называется размером критической области или критерия. Следовательно, критическая область должна выбираться так, чтобы ее размер не превосходил уровня значимости α.

Если бы критическая область определялась только уровнем значимости критерия, то даже в случае простой проверяемой гипотезы выбор этой области был бы очень произвольным. Следует обратить внимание на тот факт, что правило проверки не исключает и ошибки второго рода: принять гипотезу H_0, когда она неверна. Вероятность ошибки второго рода, состоящей в том, что выборка $\vec{X}$ попадает в область принятия гипотезы при условии, что эта гипотеза неверна, обозначается через β. Эта вероятность может быть в принципе вычислена, когда альтернативная гипотеза является простой, т.е. вполне определена функция правдоподобия альтернативной гипотезы. Вероятность 1-β называется мощностью критерия относительно простой альтернативной гипотезы $H_1^{\text{п}}$. Если среди всех областей W_n с размерами, не превосходящими заданного уровня значимости α, имеется область с наибольшей мощностью 1-β, то, очевидно, следует взять именно эту область, т.е. выбрать наиболее мощный критерий относительно простой альтернативной гипотезы.

При сложной альтернативной гипотезе мощность критерия будет зависеть от простой гипотезы, пробегающей семейство гипотез, составляющих эту альтернативную гипотезу. Предпочтительнее выбирать в таком случае так называемый равномерно наиболее мощный критерий - наиболее мощный критерий относительно всех простых гипотез, составляющих сложную альтернативную гипотезу.

Критерий для проверки простой гипотезы $H_0^{\text{п}}$ называется несмещенным, если вероятность ошибки первого рода не превосходит мощности критерия относительно любой простой гипотезы, входящей в семейство, определяющее альтернативную гипотезу, т.е.

$$\mathrm{P}\left(\vec{X} \in W_n / H_0^{\text{П}}\right) \le \mathrm{P}\left(\vec{X} \in W_n / H_1\right)$$

для любой $H_1^{\text{П}} \in H_1$. В противном случае критерий называется смещенным. Очевидно, предпочтительнее выбирать критерий, являющийся одновременно и несмещенным и равномерно наиболее мощным (или наиболее мощный несмещенный критерий в случае, когда имеется простая альтернативная гипотеза).

Если проверяемая гипотеза является сложной, то к критерию предъявляется требование, состоящее в том, чтобы для любой простой гипотезы, составляющей проверяемую, вероятность попадания $\vec{X}$ в W_n в точности равнялась заданному уровню

значимости α. В этом случае область W_n называется подобной пространству выборок или подобной критической областью, а сам критерий – подобным.

Наконец, если $\lim\limits_{n\to\infty} P\left(\vec{X} \in W_n / H_1\right) = 1,$ то критерий, определяемый областями W_n размера, не превосходящего заданного α, называется состоятельным критерием уровня значимости α гипотезы H_0.

Множество перечисленных требований, предъявляемых к критериям, связано в большей степени с тем, что не всегда наиболее предпочтительные критерии (удовлетворяющие всем или большинству перечисленных требований) существуют. Для проверки простой гипотезы против простой альтернативы существует достаточно общее решение (лемма Неймана-Пирсона). Для сложных гипотез более предпочтительные критерии получены при некоторых специальных условиях. При достаточно больших объемах выборки существуют равномерно наиболее мощные несмещенные критерии (критерии отношения правдоподобия) для некоторых сложных гипотез, в частности при проверке гипотез о параметрах распределения при условии асимптотической нормальности оценок максимального правдоподобия параметров.

Как правило, проверка гипотез проводится на основе некоторой статистики $T(\vec{X})$, которая называется статистикой критерия или тестовой статистикой. Чаще всего тестовая статистика строится на основе отношения функций правдоподобия, так как при этом получаются более предпочтительные критерии.

В предыдущем изложении предполагалось, что уровень значимости α критерия задан заранее. Выбор α может диктоваться следующими соображениями.

Известно, что уменьшение ошибки одного рода при фиксированном объеме выборки ведет к увеличению ошибки другого рода. Если известна "стоимость" ошибок первого, второго рода и объема выборки

$$C = C_1(\alpha) + C_2(\beta) + C_3(n),$$

то задача в принципе может быть решена при проверке простой гипотезы против простой альтернативы путем нахождения таких β, α и n, которые минимизируют общую стоимость, разумеется, при известных функциях $C_1(\alpha)$, $C_2(\beta)$, $C_3(n)$ и $\beta(\alpha, n)$. Если же такую "стоимость" можно оценить лишь качественно, то следует предпочесть меньший уровень значимости, когда мы уверены в справедливости проверяемой гипотезы в большей степени, чем альтернативной, за счет увеличения β, естественно, при фиксированном объеме выборки. Следует увеличить риск ошибки первого рода α в случае, когда риск, связанный с ошибкой второго рода, приводит к качественно большим потерям (например, установление закономерности, где ее на самом деле нет), и за счет увеличения α следует уменьшить риск ошибки второго рода β.

Обычно α берут равным 0,001; 0,01; 0,05 и 0,1 (или 1 - α = 0,999; 0,99; 0,95; 0,9). Для перечисленных значений строятся таблицы квантилей соответствующих тестовых статистик.

3.2. Лемма Неймана-Пирсона

Общая проблема проверки гипотезы ставится следующим образом.

При заданном уровне значимости α найти такую критическую область W_n , для которой вероятность ошибки 1-го рода не превосходит α, а мощность критерия достигает максимума.

Эта проблема решается полностью в случае, когда обе гипотезы, нулевая и альтернативная, являются простыми, с помощью фундаментальной леммы Неймана-Пирсона.

Пусть имеются две функции правдоподобия $f(\vec{x}/H_0^{\Pi})$ и $f(\vec{x}/H_1^{\Pi})$ полностью определенные, когда справедливы соответствующие гипотезы. Для проверки гипотезы H_0^{Π} против конкурирующей H_1^{Π} найдется такая константа $c(\alpha)$, зависящая от α, что область W_n , состоящая из всех точек $\vec{x}$, для которых выполняются соотношения:

$$f(\vec{x}/H_0^{\Pi}) = 0,$$

$$\frac{f(\vec{x}/H_1^{\Pi})}{f(\vec{x}/H_0^{\Pi})} \geq c(\alpha) \text{ при } f(\vec{x}/H_0^{\Pi}) \neq 0,$$

является решением задачи:

(1) $P\left(\vec{X} \in W_n / H_0^{\Pi}\right) \leq \alpha,$

(2) $P\left(\vec{X} \in W_n / H_1^{\Pi}\right) \geq P\left(\vec{X} \in S_n / H_1^{\Pi}\right)$ для любой области S_n, удовлетворяющей условию $P\left(\vec{X} \in S_n / H_0^{\Pi}\right) = \alpha.$

Докажем сформулированную лемму в предположении непрерывности распределения $\vec{X}$ при гипотезах H_0^{Π} и H_1^{Π}, т.е. когда $f(\vec{x}/H_0^{\Pi})$ и $f(\vec{x}/H_1^{\Pi})$ являются плотностями распределения $p(\vec{x}/H_0^{\Pi})$ и $p(\vec{x}/H_1^{\Pi})$ при гипотезах H_0^{Π} и H_1^{Π} соответственно.

Пусть $\vec{x} \in R_n$ – любая точка, для которой $p(\vec{x}/H_0^{\Pi}) = 0$, тогда можно включить $\vec{x}$ в W_n, так как это не нарушает условия (1) задачи. Введем функцию случайного аргумента $\vec{X}$

$$C = \frac{f(\vec{X}/H_1^{\Pi})}{f(\vec{X}/H_0^{\Pi})}$$

при любом $\vec{X}$, для которого плотность $p(\vec{x}/H_0^{\Pi}) \neq 0$. Рассмотрим функцию распределения случайной величины C при гипотезе H_0^{Π}:

$$F(c) = P(C < c) = P\left(\frac{f(\vec{X}/H_1^{\Pi})}{f(\vec{X}/H_0^{\Pi})} < c \,\middle/ H_0^{\Pi}\right) = P(f(\vec{X}/H_1^{\Pi}) < cf(\vec{X}/H_0^{\Pi})/H_0^{\Pi}),$$

или функцию аргумента c:

$$\alpha(c) = 1 - F(c) = P(f(\vec{X}/H_1^{\Pi}) \geq cf(\vec{X}/H_0^{\Pi})/H_0^{\Pi}).$$

Заметим, что $\alpha(c)$ не возрастает и непрерывна слева. Пусть сначала для заданного $\alpha(c)$ существует единственное решение уравнения

$$\alpha(c) = P(f(\vec{X}/H_1^{\Pi}) \geq cf(\vec{X}/H_0^{\Pi})/H_0^{\Pi}).$$

Тогда область W_n, определяемая соотношениями

$$\begin{cases} f(\vec{X}/H_0^{\Pi}) = 0, \\ f(\vec{X}/H_1^{\Pi}) \geq c(\alpha) f(\vec{X}/H_0^{\Pi}), \end{cases}$$

является искомой. Действительно, по построению

$$P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = P\left(f\left(\vec{X}/H_1^{\Pi}\right) = \alpha(c) \geq c\,(\alpha) f(\vec{X}/H_0^{\Pi})/H_0^{\Pi}\right),$$

т.е. условие (1) задачи выполнено. Далее, пусть S_n- другая критическая область, удовлетворяющая условию $P\left(\vec{X} \in S_n / H_0^{\Pi}\right) = \alpha(c)$.

Рассмотрим разбиения областей W_n и S_n вида:

$$W_n = W_n \overline{S}_n + W_n S_n,\ S_n = S_n W_n + S_n \overline{W}_n$$

Имеем, согласно доказанному,

$$P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = P\left(\vec{X} \in W_n \overline{S}_n / H_0^{\Pi}\right) + P(\vec{X} \in W_n S_n / H_0^{\Pi}) = \alpha,$$

и, согласно предположению,

$$P\left(\vec{X} \in S_n / H_0^{\Pi}\right) = P\left(\vec{X} \in S_n W_n / H_0^{\Pi}\right) + P(\vec{X} \in S_n \overline{W}_n / H_0^{\Pi}) = \alpha.$$

Тогда $P\left(\vec{X} \in W_n \overline{S}_n / H_0^{\Pi}\right) = P(\vec{X} \in S_n \overline{W}_n / H_0^{\Pi})$.

Так как $S_n \overline{W}_n \in \overline{W}_n$, то для всех точек области $S_n \overline{W}_n$ выполняется неравенство $f(\vec{X}/H_1^{\Pi}) < c(\alpha) f(\vec{X}/H_0^{\Pi})$, аналогично, для всех точек области $W_n \overline{S}_n \in W_n$ имеет место неравенство $f(\vec{X}/H_1^{\Pi}) \geq c(\alpha) f(\vec{X}/H_0^{\Pi})$.

Тогда получаем

$$P(\vec{X} \in W_n / H_1^{\Pi}) = P(\vec{X} \in W_n S_n / H_1^{\Pi}) + P(\vec{X} \in \overline{S}_n W_n / H_1^{\Pi}) =$$

$$= P(\vec{X} \in W_n S_n / H_1^{\Pi}) + \int_{\overline{S}_n W_n} f(\vec{X}/H_1^{\Pi}) d\vec{x} \geq P(\vec{X} \in W_n S_n / H_1^{\Pi}) +$$

$$+ c(\alpha) \int_{\overline{S}_n W_n} f(\vec{X}/H_0^{\Pi}) d\vec{x} = P(\vec{X} \in W_n S_n / H_1^{\Pi}) + c(\alpha) P(\vec{X} \in \overline{S}_n W_n / H_0^{\Pi}) =$$

$$= P(\vec{X} \in W_n S_n / H_1^{\Pi}) + c(\alpha) P(\vec{X} \in \overline{W}_n S_n / H_0^{\Pi}) = P(\vec{X} \in W_n S_n / H_1^{\Pi}) +$$

$$+ c(\alpha) \int_{\overline{W}_n S_n} f(\vec{X}/H_0^{\Pi}) d\vec{x} \geq P(\vec{X} \in W_n S_n / H_1^{\Pi}) + \int_{\overline{W}_n S_n} f(\vec{X}/H_1^{\Pi}) d\vec{x} =$$

$$= P(\vec{X} \in W_n S_n / H_1^{\Pi}) + P(\vec{X} \in \overline{W}_n S_n / H_1^{\Pi}) = P(\vec{X} \in S_n / H_1^{\Pi}),$$

что и требовалось доказать.

Пусть уравнение $\alpha(c) = P(f(\vec{X}/H_1^{\Pi}) \geq cf(\vec{X}/H_0^{\Pi})/H_0^{\Pi})$ имеет не единственное решение c (α), а целый интервал решений $c_1 < c(\alpha) \leq c_2$. Обозначим через V_n область R_n, соответствующую интервалу решений $(c_1, c_2]$. Но

$$P(\vec{X} \in V_n / H_0^{\Pi}) = \alpha(c_1 + 0) - \alpha(c_2) = 0 \text{ и } f(\vec{X}/H_0^{\Pi}) > 0,$$

следовательно, на множество V_n можно не обращать внимания, так как в силу непрерывности $\vec{X}$ это множество имеет меру нуль и тогда $P(\vec{X} \in V_n / H_1^{\Pi}) = 0$ также. Таким образом, можно взять любое из множества решений $(c_1, c_2]$.

Наконец, пусть рассматриваемое выше уравнение не имеет решений для заданного α. Тогда существует единственное $c = c_0$ такое, что $\alpha(c_0 + 0) < \alpha < \alpha(c_0)$.

По определению $\alpha(c)$:

$$\alpha(c_0) - \alpha(c_0 + 0) = P\left(f\left(\vec{X}/H_1^{\Pi}\right) = c_0 f\left(\vec{X}/H_0^{\Pi}\right)/H_0^{\Pi}\right) = \Delta > 0.$$

Построим область W_n следующим образом. Сначала возьмем всю область с $f\left(\vec{X}/H_1^{\Pi}\right) > c_0 f\left(\vec{X}/H_0^{\Pi}\right)$ и затем добавим к ней любую такую часть множества с $f\left(\vec{X}/H_1^{\Pi}\right) = c_0 f\left(\vec{X}/H_0^{\Pi}\right)$, чтобы $P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = \alpha$. Доказательство того, что полученная таким образом область определяет наиболее мощный критерий, ничем не отличается от предыдущего, так как замена неравенства $f\left(\vec{X}/H_1^{\Pi}\right) < c(\alpha) f\left(\vec{X}/H_0^{\Pi}\right)$ для всех точек области $S_n \overline{W}_n$ на неравенство $f\left(\vec{X}/H_1^{\Pi}\right) \leq c_0 f\left(\vec{X}/H_0^{\Pi}\right)$ не меняет цепочки равенств и неравенств предыдущего доказательства.

Докажем теперь лемму Неймана-Пирсона в предположении дискретности распределения $\vec{X}$ при гипотезах H_0^{Π} и H_1^{Π}, т.е. когда $f\left(\vec{X}/H_0^{\Pi}\right)$ и $f\left(\vec{X}/H_1^{\Pi}\right)$ являются вероятностями события $\vec{X} = x$ при гипотезах H_0^{Π} и H_1^{Π} соответственно. Множество возможных точек $\vec{X}$, для которых $f\left(\vec{X}/H_0^{\Pi}\right) > 0$ и (или) $f\left(\vec{X}/H_1^{\Pi}\right) > 0$, является либо конечным, либо счетным. Сначала включим точки $\vec{X}$, для которых $f\left(\vec{X}/H_0^{\Pi}\right) = 0$ и $f\left(\vec{X}/H_1^{\Pi}\right) > 0$, в критическое множество W_n. Оставшиеся возможные точки упорядочим соответственно неубыванию величины отношения $f\left(\vec{X}/H_1^{\Pi}\right)/f\left(\vec{X}/H_0^{\Pi}\right)$. Включим в множество W_n точки $\vec{X}$, соответствующие максимальному числу членов указанной последовательности, согласующемуся с условием (1) задачи. Тогда область W_n, определяемая соотношениями

$$f\left(\vec{X}/H_0^{\Pi}\right) = 0,$$

$$f\left(\vec{X}/H_1^{\Pi}\right) \geq c(\alpha) f\left(\vec{X}/H_0^{\Pi}\right) \text{ при } f\left(\vec{X}/H_0^{\Pi}\right) \neq 0,$$

где $c(\alpha)$ есть решение уравнения

$$P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = \sum_{\vec{x}: f\left(\vec{x}/H_1^{\Pi}\right) \geq cf\left(\vec{x}/H_0^{\Pi}\right)} f\left(\vec{X}/H_0^{\Pi}\right) = \alpha,$$

является искомой. Доказательство этого утверждения отличается от случая непрерывного распределения $\vec{X}$ только формальной заменой знака интеграла знаком суммы. Трудность возникает в ситуации, когда уравнение $P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = \alpha$ не имеет решения $c(\alpha)$, а именно, когда включив в W_n очередную точку $\vec{X}_{l-1}$, мы не достигаем еще уровня α, включив же следующую: $\vec{X}_l$ - превзойдем его. Теоретически эта трудность может быть преодолена путем "расщепления" последней точки $\vec{X}_l$ на две $\vec{X}_l^{'}$ и $\vec{X}_l^{''}$. Пусть $P\left(\vec{X} \in W_n + \{\vec{X}_l\} / H_0^{\Pi}\right) = \alpha + \delta,\ \delta > 0$, т.е. область $W_n + \{X_l\}$ имеет слишком большой размер, а $P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = \alpha - \varepsilon,\ \varepsilon > 0$. Припишем точкам $\vec{X}_l^{'}$ и $\vec{X}_l^{''}$ вероятности $P\left(\vec{X}_l^{'}\right) = \varepsilon$ и $P\left(\vec{X}_l^{''}\right) = \delta$ и включим точку $\vec{X}_l^{'}$ в область W_n.

Для осуществления расщепления применим прием рандомизации, состоящий в устройстве, например, лотереи с вероятностью выигрыша $\frac{\varepsilon}{\varepsilon + \delta}$, не зависящей от выборки. Если возникла ситуация, при которой выборочная точка $\vec{X}_l$ подлежит расщеплению, то производится лотерея. Если в результате получен выигрыш, то гипотеза H_0^{Π} отвергается, в противном случае - принимается. Указанный критерий имеет уровень значимости, равный α. Действительно, событие $\vec{X} = \vec{x}_l$ имеет вероятность при гипотезе H_0^{Π}, равную $\varepsilon + \delta$. Событие $\vec{X} = \vec{x}_l^{'}$ осуществляется тогда и только тогда, когда происходит событие $\vec{X} = \vec{x}_l$ и независимое от этого событие: "выигрыш". Тогда

$$P\left(\vec{X} = \vec{x}_l^{'} / H_0^{\Pi}\right) = P\left(\vec{X} = \vec{x}_l / H_0^{\Pi}\right) \cdot P\ (\text{"выигрыш"}) = (\varepsilon + \delta) \cdot \frac{\varepsilon}{\varepsilon + \delta} = \varepsilon,$$

$$\text{и } P\left(\vec{X} \in W_n / H_0^{\Pi}\right) + P\left(\vec{X} = \vec{x}_l / H_0^{\Pi}\right) = \alpha - \varepsilon + \varepsilon = \alpha$$

есть вероятность отвергнуть гипотезу H_0^{Π}, что и требовалось доказать.

На практике необязательно следует пользоваться описанным приемом рандомизации. Достаточно к области, соответствующей неравенству $f\left(\vec{x}/H_1^{\Pi}\right) > cf\left(\vec{x}/H_0^{\Pi}\right)$, добавить столько точек $\vec{x}$, удовлетворяющих $f\left(\vec{x}/H_1^{\Pi}\right) = cf\left(\vec{x}/H_0^{\Pi}\right)$, чтобы вероятность выборке попасть в построенную область была по возможности близка к α (но не превосходила α). Полученная критическая область W_n является решением новой экстремальной задачи с уровнем значимости, несколько меньшим исходного α (и большей ошибкой второго рода по сравнению с критерием уровня значимости α, построенным приемом рандомизации).

Заметим, наконец, что лемма, очевидно, справедлива, когда $\alpha = 0$ и $\alpha = 1$. Достаточно тогда положить $C(0) = +\infty$, $C(0) = -\infty$ и $0 \cdot \infty = 0$ для определенности.

Из леммы Неймана-Пирсона можно получить следствие, состоящее в том, что построенный в лемме наиболее мощный критерий уровня значимости α является

несмещенным, т.е. является наиболее мощным несмещенным критерием относительно данной альтернативной гипотезы H_1^{Π} .

Пусть критерий леммы определяет область W_n такую, что $P\left(\vec{X} \in W_n / H_0^{\Pi}\right) = \alpha$ и $P\left(\vec{X} \in \overline{W}_n R_n / H_1^{\Pi}\right) = \beta$. Надо доказать, что 1-$\beta \geq \alpha$ для построенного критерия.

Если $c \geq 1$, то $f\left(\vec{x} / H_1^{\Pi}\right) \geq f\left(\vec{x} / H_0^{\Pi}\right)$ и

$\int\limits_{W_n} f\left(\vec{x} / H_1^{\Pi}\right) d\vec{x} \geq \int\limits_{W_n} f\left(\vec{x} / H_0^{\Pi}\right) d\vec{x}$, т.е. 1-$\beta \geq \alpha$,

что и требовалось доказать. Пусть теперь $c<1$. Заметим, что для любой точки $\vec{x} \in W_n$ $f\left(\vec{x} / H_1^{\Pi}\right) \geq cf\left(\vec{x} / H_0^{\Pi}\right)$, и для любой точки $\vec{x} \in \overline{W}_n R_n$ $f\left(\vec{x} / H_1^{\Pi}\right) \leq cf\left(\vec{x} / H_0^{\Pi}\right)$, а в силу $c<1$, для любой точки $\vec{x} \in \overline{W}_n R_n$ $f\left(\vec{x} / H_1^{\Pi}\right) < f\left(\vec{x} / H_0^{\Pi}\right)$.

Выделим в W_n подмножество S_n всех точек $\vec{x}$ таких, что $f\left(\vec{x} / H_1^{\Pi}\right) \geq f\left(\vec{x} / H_0^{\Pi}\right)$ для любой из этих точек. Покажем, что такое $S_n \in W_n$ существует. Действительно, в противном случае для всех $\vec{x} \in W_n$ выполняется неравенство $f\left(\vec{x} / H_1^{\Pi}\right) < f\left(\vec{x} / H_0^{\Pi}\right)$, и, согласно сделанному выше замечанию, такое неравенство выполняется для любой точки $\vec{x} \in R_n$. Однако это противоречит свойству плотности или ряда распределения, согласно которому

$$\int\limits_{R_n} f\left(\vec{x} / H_1^{\Pi}\right) d\vec{x} = \int\limits_{R_n} f\left(\vec{x} / H_0^{\Pi}\right) d\vec{x} = 1.$$

Таким образом, для любой точки $\vec{x} \in S_n$

$$f\left(\vec{x} / H_1^{\Pi}\right) \geq f\left(\vec{x} / H_0^{\Pi}\right),$$

а для любой точки $\vec{x} \in \overline{S}_n R_n$

$$f\left(\vec{x} / H_1^{\Pi}\right) < f\left(\vec{x} / H_0^{\Pi}\right).$$

Тогда

$$P\left(\vec{X} \in S_n / H_1^{\Pi}\right) - P\left(\vec{X} \in S_n / H_0^{\Pi}\right) =$$

$$= 1 - P\left(\vec{X} \in \overline{S}_n R_n / H_1^{\Pi}\right) - 1 + P\left(\vec{X} \in \overline{S}_n R_n / H_0^{\Pi}\right) =$$

$$= P\left(\vec{X} \in \overline{S}_n R_n / H_0^{\Pi}\right) - P\left(\vec{X} \in \overline{S}_n R_n / H_1^{\Pi}\right) =$$

$$= \int\limits_{\overline{S}_n R_n} \left[f\left(\vec{x} / H_0^{\Pi}\right) - f\left(\vec{x} / H_1^{\Pi}\right) \right] d\vec{x}.$$

Подынтегральная функция в области $\overline{S}_n R_n$ положительна в силу выбора подмножества S_n, следовательно, при $\overline{S}_n W_n \in \overline{S}_n R_n$

$$\int\limits_{\overline{S}_n R_n} \left[f\left(\vec{x} / H_0^{\Pi}\right) - f\left(\vec{x} / H_1^{\Pi}\right) \right] d\vec{x} \geq \int\limits_{\overline{S}_n W_n} \left[f\left(\vec{x} / H_0^{\Pi}\right) - f\left(\vec{x} / H_1^{\Pi}\right) \right] d\vec{x} =$$

$$= P\left(\vec{X} \in \overline{S}_n W_n / H_0^{\Pi}\right) - P\left(\vec{X} \in \overline{S}_n W_n / H_1^{\Pi}\right).$$

Таким образом

$$P\left(\vec{X} \in S_n / H_1^{\Pi}\right) - P\left(\vec{X} \in S_n / H_0^{\Pi}\right) \geq P\left(\vec{X} \in S_n W_n / H_0^{\Pi}\right) - P\left(\vec{X} \in S_n W_n / H_1^{\Pi}\right).$$

Отсюда

$$P\left(\vec{X} \in W_n / H_1^{\Pi}\right) \geq P\left(\vec{X} \in W_n / H_0^{\Pi}\right),$$

т.е. 1-$\beta \geq \alpha$, что и требовалось доказать.

3.3. Примеры построения наиболее предпочтительных критериев

Приведем примеры, иллюстрирующие построение критерия по лемме Неймана-Пирсона.

Пример 1. Проверить гипотезу о прямоугольном (равномерном) законе распределения против гипотезы о треугольном законе распределения случайной величины. Конкретнее, пусть

$$H_0^{\Pi} : p(x) = \begin{cases} \dfrac{1}{2} & при \ \ |x| \leq 1 \\ 0 & при \ \ |x| > 1 \end{cases}; \ H_1^{\Pi} : p(x) = \begin{cases} 1 - |x| & при \ \ |x| \leq 1 \\ 0 & при \ \ |x| > 1 \end{cases}$$

Для простоты будем считать объем выборки *n* равным 1.

Согласно лемме, критическая область $W_n = W_1$, отвечающая более предпочтительному критерию уровня значимости α, удовлетворяет соотношениям

$$|x| > 1,$$

$$\frac{f\left(x / H_1^{\Pi}\right)}{f\left(x / H_0^{\Pi}\right)} = \frac{1 - |x|}{\dfrac{1}{2}} \geq C \ \ при \ \ |x| \leq 1,$$

или

$$\begin{cases} |x| \leq 1 - \dfrac{1}{2} C, \\ |x| > 1. \end{cases}$$

Тогда, считая выборку, удовлетворяющую условию $|x| > 1$, невозможным событием, будем иметь:

при $C = 2$ область W_1 состоит из единственной выборочной точки $x = 0$;

при $C > 2$ область W_1 является пустым множеством, т.е. гипотеза H_0^{Π} не отвергается при любой возможной выборке;

при $0 < C < 2$ область W_1 имеет вид $|x| \le 1 - \frac{1}{2}C$, т.е. является окрестностью нуля с радиусом, не превосходящим единицы;

наконец, при $C = 0$ область W_1 совпадает с областью возможных значений X, т.е. при любой выборке гипотеза $H_0^{\text{П}}$ будет отвергаться.

Выясним, как по заданному уровню значимости α определить критическое значение C, равное $C(\alpha)$. Имеем

$$1 - F(c) = P\left(X \in W_1 / H_0^{\text{П}}\right) = P\left(|X| \le 1 - \frac{1}{2}C / H_0^{\text{П}}\right) = \int \frac{1}{2} dz = \alpha$$

$$|x| \le \frac{1}{2}C .$$

Решая уравнение

$$2\int_0^{1-\frac{1}{2}C} \frac{1}{2} dz = \alpha$$

относительно C, получим

$$C(\alpha) = 2(1 - \alpha)$$

и, следовательно, критическая область, соответствующая наиболее мощному критерию, имеет вид

$$|x| \le \alpha .$$

Вычислим мощность критерия

$$1 - \beta = \int_{W_1} f\left(x / H_1^{\text{П}}\right) dx = 2\int_0^{\alpha} (1 - x) dx = 2\alpha - \alpha^2 .$$

Покажем, что критерий является несмещенным. Действительно, $1 - \beta = 2\alpha - \alpha^2 = \alpha[1 + (1 - \alpha)]$ что, очевидно, не меньше α.

Итак, правило проверки гипотезы $H_0^{\text{П}}$ будет состоять в следующем: если выборочное значение x случайной величины X удовлетворяет неравенству $|x| \le \alpha$, то гипотеза о прямоугольном распределении X отвергается с вероятностью ошибки первого рода, равной α, т.е. предпочтение оказывают треугольному распределению; если же наблюдаемое значение X попало в область $|x| > \alpha$, то гипотеза о прямоугольном распределении X не отвергается, при этом вероятность ошибки второго рода равна

$$\beta = 1 - 2\alpha + \alpha^2 = (1 - \alpha)^2 .$$

Пример 2. Проверить гипотезу $H_0^{\text{П}}$ о том, что нормально распределенная генеральная совокупность X с известной дисперсией σ^2 имеет генеральную среднюю, равную μ_0, против альтернативы $H_1^{\text{П}} : MX = \mu_1 > \mu_0$.

Рассмотрим выборку (X_1, X_2, …, X_n) объема n из данной генеральной совокупности. Тогда функция правдоподобия при различных гипотезах будет иметь вид:

$$f\left(\vec{x}/H_0^{\Pi}\right)=\left(2\pi\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2\right\}$$

$$f\left(\vec{x}/H_1^{\Pi}\right)=\left(2\pi\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu_1\right)^2\right\},$$

и определена для любой точки $\vec{x}\in R_n$.

Отношение правдоподобия равно

$$C=\exp\left\{-\frac{1}{2\sigma^2}\left[2n\left(\mu_1-\mu_0\right)\bar{x}-n\left(\mu_1^2-\mu_0^2\right)^2\right]\right\}.$$

Так как C является монотонной функцией средней арифметической $\bar{x}$ выборки, то критическая область W_n определяется неравенством относительно статистики $\bar{x}$ вида $\bar{x}>\bar{x}'_{кр.}\left(\alpha\right)$,

где $\bar{x}'_{кр.}\left(\alpha\right)$ - критическое значение статистики $\bar{x}$, отвечающее заданному уровню α, т.е. $\int\limits_{\bar{x}>\bar{x}'_{кр.}(\alpha)} P\left(\vec{x}/H_0^{\Pi}\right)d\vec{x}=\alpha$.

Известно, что статистика $\bar{x}$ распределена нормально с параметрами при гипотезе $H_0^{\Pi}: M\bar{x}=\mu_0$ и $D\bar{x}=\dfrac{\sigma^2}{n}$.

Следовательно, гипотеза H_0^{Π} отвергается в случае $\bar{x}>\bar{x}'_{кр.}\left(\alpha\right)$,

где $\bar{x}'_{кр.}\left(\alpha\right)=\dfrac{\sigma}{\sqrt{n}}\Phi_*^{-1}\left(1-\alpha\right)+\mu_0$; Φ_*^{-1} - функция, обратная функции нормального распределения.

Справедливы следующие формулы, связывающие функцию $\Phi_*(t)$ и функцию $\Phi(t)$, которая табулирована (см. Приложение):

$$\Phi_*\left(t\right)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{t}e^{-\frac{1}{2}y^2}\,dy,$$

$$\Phi_*\left(t\right)=\frac{1}{2}\left[\Phi\left(t\right)+1\right];$$

$$\Phi\left(t\right)=2\Phi_*\left(t\right)-1,\ \Phi\left(t\right)=\frac{2}{\sqrt{2\pi}}\int\limits_{0}^{t}e^{-\frac{1}{2}y^2}\,dy,\ \Phi\left(-t\right)=-\Phi\left(t\right).$$

Рассчитаем мощность полученного критерия:

$$1-\beta=\int\limits_{\bar{x}>\bar{x}'_{кр.}(\alpha)} f\left(\vec{x}/H_1^{\text{П}}\right)d\vec{x}=1-\int\limits_{\bar{x}\le\bar{x}'_{кр.}(\alpha)} f\left(\vec{x}/H_1^{\text{П}}\right)d\vec{x}=1-\Phi_*\left(\frac{\bar{x}_{кр.}(\alpha)-\mu_1}{\sigma/\sqrt{n}}\right).$$

Так как критерий гипотезы $H_0^{\text{П}}:\mu=\mu_0$, задаваемый неравенством $\bar{x}>\bar{x}'_{кр.}(\alpha)$, не зависит от $\mu_1>\mu_0$, то он является равномерно наиболее мощным относительно всех простых гипотез $\mu_1>\mu_0$, составляющих сложную альтернативную гипотезу $H_1:\mu>\mu_0$. Этот же критерий является несмещенным, так как

$$1-\beta=1-\Phi_*\left(\frac{\bar{x}'_{кр.}(\alpha)-\mu_1}{\sigma/\sqrt{n}}\right)>1-\Phi_*\left(\frac{\bar{x}'_{кр.}(\alpha)-\mu_0}{\sigma/\sqrt{n}}\right)=\alpha$$

при всех $\mu_1>\mu_0$.

Заметим, что по аналогии с гипотезой $H_1:\mu>\mu_0$, критерий гипотезы $H_0^{\text{П}}:\mu=\mu_0$ против альтернативной сложной гипотезы $H_1:\mu<\mu_0$ имеет вид

$$\bar{x}<\bar{x}''_{кр.}(\alpha),$$

где $\bar{x}''_{кр.}(\alpha)=\mu_0+\frac{\sigma}{\sqrt{n}}\Phi_*^{-1}(\alpha)$. Критерий также является равномерно наиболее мощным и несмещенным относительно $H_1:\mu<\mu_0$.

Пример 3. Проверить гипотезу $H_0^{\text{П}}$ о том, что нормально распределенная генеральная совокупность X с известной дисперсией σ^2 имеет генеральную среднюю μ равной μ_0, против альтернативы $H_1:\mu\neq\mu_0$.

Согласно примеру 2, при проверке гипотезы $H_1:\mu=\mu_0$ против $H_1:\mu<\mu_0$ критическая область имеет вид $\bar{x}<\bar{x}''_{кр.}(\alpha)$; если же альтернативой является гипотеза $H_1:\mu>\mu_0$, то критическая область имеет вид $\bar{x}>\bar{x}'_{кр.}(\alpha)$. Заметим, что ни одна из этих двух областей не является предпочтительной при альтернативе $H_1:\mu\neq\mu_0$, так как любая из областей дает смещенный критерий против этой альтернативы. Например, мощность критерия, область W_n которого определяется неравенством $\bar{x}>\bar{x}'_{кр.}(\alpha)$, имеет вид $1-\Phi_*\left(\frac{\bar{x}'_{кр.}(\alpha)-\mu}{\sigma/\sqrt{n}}\right)$ и будет, очевидно, меньше $\alpha=1-\Phi_*\left(\frac{\bar{x}'_{кр.}(\alpha)-\mu_0}{\sigma/\sqrt{n}}\right)$ при $\mu<\mu_0$.

Для отыскания более предпочтительного критерия рассмотрим ту же статистику $\bar{x}$, а в качестве критической области возьмем «симметричную» область: $\bar{x}<\bar{x}''_{кр.}\left(\frac{\alpha}{2}\right)$, $\bar{x}>\bar{x}'_{кр.}\left(\frac{\alpha}{2}\right)$. Следовательно, если наблюдаемое значение статистики $\bar{x}$ удовлетворяет неравенству

$$\left|\bar{x}_{набл.} - \mu_0\right| > \frac{\sigma}{\sqrt{n}}\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right),$$

то гипотеза $H_1 : \mu = \mu_0$ отвергается с вероятностью ошибки такого решения, равной α.

Покажем, что найденный критерий является не смещенным. Функция мощности для любого μ имеет вид

$$\Phi_*\left\{\frac{\sqrt{n}}{\sigma}(\mu-\mu_0)-\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right)\right\}+\Phi_*\left\{\frac{\sqrt{n}}{\sigma}(\mu_0-\mu)-\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right)\right\}.$$

Нетрудно заметить, что она имеет минимум в точке $\mu = \mu_0$, равный α, и поэтому найденный критерий не смещен.

С помощью модификации леммы Неймана-Пирсона можно показать, что полученный критерий является равномерно наиболее мощным несмещенным критерием.

3.4. Односторонние и двусторонние критические области

При сравнении критериев, построенных в примерах 2 и 3 предыдущего параграфа, можно заметить, что критические области для одной и той же тестовой статистики $\bar{x}$ отличаются друг от друга в зависимости от вида конкурирующей гипотезы. Гипотезе $H_1 : \mu > \mu_0 (\mu < \mu_0)$ соответствует критическая область вида $\bar{x} > \bar{x}'_{кр.}(\alpha)$ $\left(\bar{x} < \bar{x}'''_{кр.}(\alpha)\right)$, называемая односторонней или, точнее, правосторонней (левосторонней) критической областью. Соответствующий критерий называется односторонним. Гипотезе $H_1 : \mu \neq \mu_0$ соответствует двусторонний критерий, так как критическая область имеет вид: $\bar{x} > \bar{x}_{кр.}\left(\frac{\alpha}{2}\right)$ или $\bar{x} < \bar{x}'''_{кр.}\left(\frac{\alpha}{2}\right)$,и множество простых гипотез $H_1^{\text{П}}$, составляющих альтернативу H_1, определяет множество значений параметра μ, расположенных по обе стороны от μ_0.

При проверке простой гипотезы $H_0^{\text{П}}$ против сложной альтернативы желательно иметь альтернативу, приводящую к односторонней критической области тестовой статистики в силу того, что односторонний критерий оказывается предпочтительней двустороннего критерия, основанного на той же статистике. Информация об указанном виде альтернативной гипотезы иногда встречается на практике; например, когда речь идет о допустимой доле брака $p = p_0$; альтернативой здесь служит гипотеза о том, что доля брака превышает стандарт: $p > p_0$. Двусторонние критерии вынужденно применяются тогда, когда множество простых гипотез невозможно упорядочить, таковыми являются критерии согласия, когда альтернативой служит множество всевозможных теоретических распределений, отличных от предполагаемого распределения.

При дальнейшем изложении полезно ввести понятие параметрических и непараметрических статистических гипотез. Статистическая гипотеза называется параметрической, если в соответствующем ей утверждении форма распределения

признаков в генеральной совокупности известна с точностью до параметров, характеристик, относительно которых и идет речь в гипотезе. При заранее неизвестном виде распределения гипотеза называется непараметрической.

Примерами параметрических гипотез служат гипотезы (1), (2) и (3). Гипотезы (4) и (5) являются непараметрическими.

Наиболее полно разработана теория проверки параметрических статистических гипотез. Теория проверки непараметрических статистических гипотез развита недостаточно и для знакомства с ней приводятся лишь некоторые, получившие широкое распространение, задачи.

Параметрические гипотезы, определяемые параметром θ, вида $\theta = \theta_0$, $\theta \geq \theta_0$ или $\theta \leq \theta_0$, где θ_0 - известное значение, называются обычно гипотезами сравнения со стандартом (θ_0). Гипотезы вида $\theta^1 = \theta^2$ или $\theta^1 = \theta^2 = \ldots = \theta^k$ называются гипотезами сравнения двух или k совокупностей. Иногда удобно гипотезу $\theta^1 = \theta^2$ записать в виде $\theta^1 - \theta^2 = 0$, т.е. привести ее к гипотезе сравнения со стандартом (0). Различные выборки, применяемые для проверки гипотезы сравнения, независимы.

Относительно параметрических гипотез можно заметить, что оценка параметров и проверка гипотез типа сравнения со стандартом обычно производятся на основе одних и тех же статистик, однако для оценки и для проверки применять одну и ту же выборку теоретически нельзя, так как это может привести к неправильным заключениям.

3.5. Гипотезы о генеральных долях

3.5.1. Сравнение генеральной доли со стандартом

Рассмотрим генеральную совокупность объема N с альтернативным признаком: M единиц совокупности обладают изучаемым признаком и N - M единиц - не обладают этим признаком. Генеральная доля обозначается через $p = \frac{M}{N}$. Гипотезы о генеральной доле проверяются с помощью выборки объема n, которая может быть получена при возвращении случайно отобранной единицы в генеральную совокупность.

Проверим гипотезу о том, что генеральная доля p равна стандартному значению p_0, против альтернативы, состоящей в том, что генеральная доля p равна p_1, причем $p_1 > p_0$. Здесь обе гипотезы являются простыми, и на основе леммы Неймана-Пирсона проверка может быть произведена с помощью тестовой статистики m - частоты числа единиц, обладающих изучаемым признаком, выборки объема n единиц. Статистика m распределена по биномиальному закону с параметрами n и p. Наиболее мощная критическая область размера, не превышающего α, задается неравенством $m > m_{кр.}$, где $m_{кр.}$ является решением уравнения

$$\sum_{m=0}^{m_{кр.}} C_n^m p_0^m (1 - p_0)^{n-m} = 1 - \alpha. \qquad (1)$$

Заметим, что из-за дискретности случайной величины m это уравнение может не иметь решения, поэтому, если не прибегать к рандомизации, следует выбрать, как было указано в 3.2, такое $m_{кр.}$, чтобы левая часть уравнения отличалась от правой как можно меньше, но была бы не меньше, чем правая часть. После того, как такое $m_{кр.}$ будет

найдено, можно уменьшить уровень значимости α так, чтобы левая часть при найденном $m_{кр.}$ в точности равнялась правой части уравнения при новом уровне значимости.

Напомним, что левая часть рассмотренного уравнения может быть аппроксимирована в соответствии с законом Пуассона при достаточно большом объеме выборки n и при малой (или большой) стандартной доле p_0, или в соответствии с нормальным законом, когда n велико и p_0 не намного отличается от 0,5. При безвозвратной выборке можно заменить левую часть уравнения в соответствии с гипергеометрическим законом.

Мощность указанного критерия вычисляется по формуле

$$1-\beta = 1-\sum_{m=0}^{m_{кр.}} C_n^m p_1^m \left(1-p_1\right)^{n-m} = \sum_{m_{кр.}+1}^{n} C_n^m p_1^m \left(1-p_1\right)^{n-m}. \qquad (2)$$

Если вычисленная мощность оказывается недостаточно большой, то можно, кроме задания уровня значимости α, задать удовлетворительную вероятность ошибки 2-го рода β, однако в этом случае объем выборки нельзя фиксировать заранее, и придется вычислять одновременно с $m_{кр.}$, воспользовавшись системой

$$\begin{cases} \sum_{m=0}^{m_{кр.}} C_n^m p_0^m \left(1-p_0\right)^{n-m_0} = 1-\alpha, \\ \sum_{m=0}^{m_{кр.}} C_n^m p_1^m \left(1-p_1\right)^{n-m} = \beta, \end{cases} \qquad (3)$$

или системой с аппроксимированными левыми частями уравнений, как было указано выше.

Пример 1. Из партии изделий извлекается выборка объемом в 100 единиц с целью проверки гипотезы $H_0^{\text{П}}$ о том, что доля брака в партии составляет 5%, против гипотезы $H_1^{\text{П}}$ о том, что эта доля равна 10%. Найти наиболее предпочтительный критерий уровня значимости, не превышающего $\alpha = 0{,}05$, и вычислить его мощность.

Решение. Согласно уравнению (1) имеем

$$\sum_{m=0}^{m_{кр.}} C_{100}^m \cdot 0{,}05^m \cdot 0{,}95^{100-m} = 0{,}95.$$

Так как доля брака $p_0 = 0{,}05$ достаточно мала, а объем выборки $n = 100$ достаточно велик, левую часть уравнения можно аппроксимировать в соответствии с законом Пуассона с параметром $\lambda_0 = 100 \cdot 0{,}05 = 5$. Тогда P ($m \le m_{кр} = 8$) = 0,9319, а P ($m \le m_{кр} = 9$) = 0,9682. Следовательно, $m_{кр} = 9$ и $\alpha' = 0{,}0318 = 0{,}03$. Таким образом, критическая область размера $\alpha' = 0{,}03$ задается неравенством: $m > 9$, т.е. если в выборке объемом в 100 изделий окажется более 9 бракованных, то гипотезу о доле брака в партии, равной 10%, следует предпочесть гипотезе о доле брака, равной 5%. Вычислим мощность полученного критерия, взяв $\lambda_1 = 100 \cdot 0{,}1 = 10$, и воспользовавшись формулой (2), с аппроксимацией пуассоновским распределением:

$$1-\beta = 1-\sum_{m=0}^{9} e^{-10} \cdot \frac{10}{m!} = 1-0{,}4580 = 0{,}5420.$$

Таким образом, ошибка 2-го рода, равная 0,46, может оказаться чрезмерно большой.

Пример 2. В условиях примера 1 вместо заданного объема выборки задана ошибка 2-го рода, равная 0,1. Найти предпочтительный критерий и соответствующий минимальный объем выборки.

Решение. Выпишем систему (3), приняв аппроксимацию левых частей в соответствии с нормальным законом (интегральная теорема Муавра-Лапласа) с параметрами $\mu_0 = np_0$, $\sigma_0^2 = np_0(1-p_0)$, когда справедлива гипотеза $H_0^{\text{П}}: p = p_0$, и $\mu_1 = np_1$, $\sigma_1^2 = np_1(1-p_1)$ при справедливости гипотезы $H_1^{\text{П}}: p = p_1$:

$$P\left(0 \le m \le m_{кр.} / H_0^{\text{П}}\right) = \Phi_*\left(\frac{m_{кр.} - p_0 n}{\sqrt{p_0(1-p_0)n}}\right) - \Phi_*\left(\frac{0 - p_0 n}{\sqrt{p_0(1-p_0)n}}\right) = 1-\alpha,$$

$$P\left(0 \le m \le m_{кр.} / H_1^{\text{П}}\right) = \Phi_*\left(\frac{m_{кр.} - p_1 n}{\sqrt{p_1(1-p_1)n}}\right) - \Phi_*\left(\frac{0 - p_1 n}{\sqrt{p_1(1-p_1)n}}\right) = \beta$$

Так как n достаточно велико,

$$\Phi_*\left(\frac{-p_0 n}{\sqrt{p_0(1-p_0)n}}\right) \text{ и } \Phi_*\left(\frac{-p_1 n}{\sqrt{p_1(1-p_1)n}}\right)$$

достаточно близки к нулю, и тогда

$$\frac{m_{кр.} - p_0 n}{\sqrt{p_0(1-p_0)n}} = \Phi_*^{-1}(1-\alpha), \quad \frac{m_{кр.} - p_1 n}{\sqrt{p_1(1-p_1)n}} = \Phi_*^{-1}(\beta)$$

Отсюда

$$n = \frac{\left[\Phi_*^{-1}(1-\alpha)\sqrt{p_0(1-p_0)} - \Phi_*^{-1}(\beta)\sqrt{p_1(1-p_1)}\right]^2}{(p_1-p_0)^2},$$

$$m_{кр.} = \frac{p_1\Phi_*^{-1}(1-\alpha)\sqrt{p_0(1-p_0)} - p_0\Phi_*^{-1}(\beta)\sqrt{p_1(1-p_1)}}{\Phi_*^{-1}(1-\alpha)\sqrt{p_0(1-p_0)} - \Phi_*^{-1}(\beta)\sqrt{p_1(1-p_1)}} \cdot n,$$

Подставив числовые значения в последние формулы, получим

$$n = \frac{\left[1{,}64\sqrt{0{,}05 \cdot 0{,}95} - (-1{,}28)\cdot\sqrt{0{,}1 \cdot 0{,}9}\right]^2}{(0{,}1-0{,}05)^2} = 220,$$

$$m_{кр} = \frac{0{,}1 \cdot 1{,}64\sqrt{0{,}05 \cdot 0{,}95} - 0{,}05 \cdot (-1{,}28)\cdot\sqrt{0{,}1 \cdot 0{,}9}}{1{,}64 \cdot \sqrt{0{,}05 \cdot 0{,}95} - (-1{,}28)\sqrt{0{,}1 \cdot 0{,}9}} \cdot 220 = 16.$$

Таким образом, если в выборке объемом 220 наблюдаемое число бракованных изделий $m_{набл.}$ будет больше 16, то гипотеза о доле брака в партии, равной 5%, отвергается с вероятностью ошибки 0,05 (риск производства), если же $m_{набл.}$ будет не больше 16, то

гипотеза о доле брака в 5% принимается с вероятностью ошибки, равной 0,1 (риск потребления).

Проверим гипотезу о том, что генеральная доля p не больше стандартного значения p_0, против конкурирующей гипотезы, заключающейся в том, что эта генеральная доля превышает стандартное значение p_0. Обе гипотезы $H_0 : p \le p_0$ и $H_1 : p > p_0$ являются сложными. Наиболее предпочтительным служит здесь только что рассмотренный критерий с критической областью, определяемой неравенством

$$m > m_{\text{кр.}},$$

где $m_{кр.}$, является решением уравнения (1), либо того же уравнения с измененной левой частью в соответствии с пуассоновым, нормальным или гипергеометрическим законом. Можно доказать, что рассмотренная критическая область определяет равномерно наиболее мощный (и несмещенный) критерий, который является, очевидно, односторонним (правосторонним).

Для проверки гипотезы $H_0 : p \ge p_0$ против альтернативы $H_0 : p_1 < p_0$ по аналогии с вышеизложенным наиболее предпочтительный критерий задается с помощью тестовой статистики m в виде следующего неравенства, определяющего критическую область :

$$m < m_{\text{кр.}},$$

где $m_{кр.}$ есть решение уравнения

$$\sum_{m=m_{кр.}}^{n} C_n^m p_0^m \left(1 - p_0\right)^{n-m} = 1 - \alpha. \qquad (4)$$

Этот критерий также является односторонним (левосторонним) и равномерно наиболее мощным (несмещенным).

Заметим, что гипотезы: $H_0^{\text{П}} : p = p_0$ против $H_1 : p > p_0$ или против $H_1 : p < p_0$ являются частными случаями гипотез, рассмотренных выше, поэтому для них соответствующие односторонние критерии оказываются также наиболее предпочтительными.

Пример 3. Из партии в 10 изделий извлекается наудачу 3 изделия для контроля, приводящего к уничтожению контролируемых изделий. Найти наиболее предпочтительный критерий проверки партии уровня $\alpha = 0{,}1$, если условие принятия партии (т.е. 7 оставшихся изделий) состоит в том, что в партии из данных 10 изделий дефектными оказываются не более 2 изделий, в противном случае партия не принимается.

Решение. Требуется проверить гипотезу $H_0 : M \le 2$ против альтернативы $H_1 : M > 2$, где M - число дефектных изделий в партии, на основе выборки без возвращения объемом $n = 3$ единицы. Для нахождения $m_{кр.}$ (критического числа дефектных изделий в выборке) воспользуемся измененным уравнением (3), в котором P_m вычисляется по формуле, соответствующей гипергеометрическому закону распределения

$$P_m = P\left(X = m\right) = \frac{C_{M_0}^m C_{N-M_0}^{n-m}}{C_N^n}$$

при справедливости гипотезы $H_0^{\text{П}} : M = M_0$. Рассчитаем ряд этого распределения, положив $N = 10,\ M_0 = 2,\ n = 3$:

$$X = \begin{Bmatrix} m: & 0 & 1 & 2 \\ P_m: & \frac{7}{15} & \frac{7}{15} & \frac{1}{15} \end{Bmatrix}.$$

Если взять в качестве критического числа дефектных изделий в выборке $m_{кр.} = 1$, то левая часть уравнения, равная $\frac{7}{15} + \frac{7}{15} = \frac{14}{15}$, окажется больше правой части, равной 1-0,1 = 0,9 и наиболее близкой к 0,9. Поэтому уровень значимости можно уменьшить: $1 - \frac{14}{15} = \frac{1}{15} = 0,07$. Итак, критерий состоит в следующем. Если число дефектных изделий в выборке окажется больше одного, то партия не принимается с вероятностью ошибки $\alpha' = 0,07$. Если же $m_{набл.} \le 1$, то партия принимается.

Пример 4. В условиях примера 3 вычислить максимальное и минимальное значения вероятности ошибки 2-го рода для критической области размера, не превосходящего 0,01.

Решение. Если не прибегать к рандомизации, то критическая область определяется неравенствам $m_{набл.} > 2$, причем вероятность попадания выборки в эту область равна нулю при справедливости гипотезы $H_0 : M \le 2$. Последнее становится более ясным, если обратить внимание на тот факт, что при числе дефектных изделий в выборке, большем 2, партия по определению отвергается, т.е. не совершается при этом ошибки. Максимальная вероятность ошибки 2-го рода получается при гипотезе $H_1^{\text{П}} : M_1 = 3$, т.е. при самом близком к $M_0 = 2$ числе дефектных изделий или минимальном числе дефектных изделий в партии, которая не принимается. Эта вероятность вычисляется по формуле

$$\beta = \sum_{m=0}^{m_{кр.}} P_m = \sum_{m=0}^{2} P_m,$$

где $P_m = P\left(X = m / M_1 = 3\right) = \dfrac{C_3^m C_7^{3-m}}{C_{10}^3}$.

Имеем следующий гипергеометрический ряд распределения при справедливости альтернативы $M_1 = 3$:

$$X = \begin{Bmatrix} m: & 0 & 1 & 2 & 3 \\ P_m: & \frac{7}{24} & \frac{21}{40} & \frac{7}{40} & \frac{1}{120} \end{Bmatrix}.$$

Отсюда $\beta = \frac{7}{24} + \frac{21}{40} + \frac{7}{40} = 1 - \frac{1}{120} = \frac{119}{120}$,

т.е. вероятность ошибки 2-го рода при $M_1 = 3$ достаточно велика. Минимальное значение вероятности ошибки 2-го рода получится при $M_1 = 9$

$$\beta = 0 + 0 + 0,3 = 0,3,$$

исходя из следующего ряда распределения:

$$X/\left(M_1=9\right)=\left\{\begin{matrix} m: & 0 & 1 & 2 & 3 \\ P_m: & 0 & 0 & 0,3 & 0,7 \end{matrix}\right\}.$$

Таким образом, для более достоверного различения гипотез значения параметра, их определяющего, должны быть достаточно удалены друг от друга, либо можно прибегнуть к увеличению объема выборки. Первое от математической статистики не зависит, второе же может оказаться бессмысленным ввиду малочисленности партии и уничтожения контролируемых изделий. Отметим, что если проверяется гипотеза $M \le 2$ против альтернативы $M = 10$ с критической областью $m_{\text{набл.}} > 2$, то задача становится не вероятностной, так как при $M \le 2$ любая партия будет приниматься, а при $M = 10$ - отвергаться по определению и без ошибок, т.е. при справедливости $M \le 2$ любая выборка с $n = 3$ не будет попадать в критическую область, а при справедливости $M = 10$ будет попадать в критическую область.

Очевидно, примеры 3 и 4 можно дать в эквивалентной формулировке проверки гипотез о генеральной доле $\frac{M}{N}$ вместо M.

Задача оценки генеральной доли часто приводит к необходимости решить, согласуются ли данные с гипотезой $p = p_0$ или нет. Следовательно, требуется проверить гипотезу $H_0^{\text{П}}: p = p_0$, против альтернативы $H_1: p \ne p_0$. Равномерно наиболее мощный несмещенный критерий уровня, не превосходящего α, определяет доверительный интервал вида

$$\underline{m}_{кр} \le m_{набл} \le \overline{m}_{кр}$$

или двустороннюю критическую область вида

$$m_{набл} < \underline{m}_{кр}, \quad m_{набл} > \overline{m}_{кр}.$$

Точные границы $\underline{m}_{кр}$ и $\overline{m}_{кр}$ в общем случае находятся путем громоздких расчетов. Практически при достаточно больших объемах выборки пользуются аппроксимацией биномиального или гипергеометрического законов нормальным, когда p_0 не слишком близко к 0 или 1.

В результате критическая область является симметричной и задается в виде неравенства

$$\left|m_{набл} - np_0\right| > \sqrt{np_0(1-p_0)}\,\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = \sqrt{np_0(1-p_0)}\,\Phi_*^{-1}(1-\alpha) \qquad (5)$$

Если же n достаточно велико и p_0 близко к 0 или 1, то целесообразно воспользоваться аппроксимацией законом Пуассона и, исходя из логической сущности задачи, соответствующим односторонним критерием, заменив соответственно альтернативную гипотезу $H_1: p \ne p_0$ на $H_1: p > p_0$ или на $H_1: p < p_0$.

Пример 5. На основе выборки объема в 100 единиц из большой партии изделий, изготовленных на двух станках, найти наиболее предпочтительный критерий для проверки гипотезы "в партии количества изделий, изготовленных на разных станках, одинаковы".

Решение. Требуется проверить гипотезу $H_0^{\text{П}} : p = p_0 = 0{,}5$, где p_0 - доля изделий в партии, изготовленных на первом станке, против альтернативы $H_1 : p \neq p_0 = 0{,}5$. Взяв $\alpha = 0{,}05$, согласно (5), получим

$$\overline{m}_{кр} = np_0 + \sqrt{np_0(1-p_0)}\,\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) =$$

$$= 100 \cdot 0{,}5 + \sqrt{100 \cdot 0{,}5 \cdot 0{,}5} \cdot 1{,}96 = 40{,}2$$

$$\underline{m}_{кр} = np_0 - \sqrt{np_0(1-p_0)}\,\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = 59{,}8$$

Критическая область задается неравенствами

$$m_{набл} < 41, \qquad m_{набл} > 59$$

и имеет несколько меньший размер, чем $\alpha = 0{,}05$.

Если в результате проверки гипотеза $H_0^{\text{П}} : p = p_0 = 0{,}5$ отвергается, то можно проверять ту же гипотезу на том же уровне значимости с помощью одностороннего критерия соответственно против гипотезы $H_1 : p > p_0 = 0{,}5$ или $H_1 : p < p_0 = 0{,}5$, конечно, на основе другой выборки. Проверка гипотезы с помощью одностороннего критерия на основе старой выборки подтвердила бы альтернативу.

3.5.2. Сравнение нескольких долей

Пусть изучаются две генеральные совокупности относительно одного и того же альтернативного признака, распределение которого полностью определяется генеральной долей p_1 для совокупности №1 и p_2 - для совокупности №2. Тогда сравнение этих совокупностей с математико-статистической точки зрения сводится к сравнению их генеральных долей. Проверяемая гипотеза имеет вид

$$H_0^{\text{П}} : p_1 = p_2 = p_0$$

или, что то же самое,

$$H_0^{\text{П}} : p_1 - p_2 = 0\,,$$

так что само значение p_0 может быть и неизвестным.

Альтернативной является гипотеза

$$H_1 : p_1 \neq p_2\,.$$

Проверка указанной гипотезы производится на основе двух независимых выборок, взятых соответственно из каждой совокупности.

Если объемы n_1 выборки из совокупности №1 и n_2 из №2 достаточно велики, то можно воспользоваться асимптотически нормальными распределениями для выборочных долей

$$W_1 = \frac{m_1}{n_1} \text{ и } W_2 = \frac{m_2}{n_2},$$

где m_1 и m_2 - численности единиц выборок из совокупностей №1 и №2, обладающих изучаемым признаком.

В случае справедливости гипотезы $p_1 = p_2 = p_0$ статистика

$$T = \frac{W_1 - W_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

распределена нормально с $MT = 0$ и с $DT = 1$.

Разумеется при достаточно большом $n = n_1+n_2$ можно вместо неизвестной p_0 воспользоваться ее оценкой $\hat{p}_0$ (при справедливости гипотезы $p_1 = p_2 = p_0$):

$$\hat{p}_0 = \frac{m_1 + m_2}{n_1 + n_2} = \frac{W_1 \cdot n_1 + W_2 n_2}{n_1 + n_2}.$$

Критическая область симметрична и определяется неравенством

$$|t_{набл}| > \Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}(1-\alpha).$$

Пример 6. Сравнить две партии изделий с точки зрения доли брака

Партия	Объем выборки	Число дефектных изделий в выборке
№1	200	5
№2	300	10

Решение. Пусть $\alpha = 0{,}05$, тогда $\Phi_*^{-1}(0{,}975) = 1{,}96$ и критическая область задается неравенством

$$|t_{набл}| > 1{,}96.$$

Оценим $p_0 : \hat{p}_0 = \frac{15}{500} = 0{,}03$; вычислим значение статистики T в условиях задачи

Так как 0,54 < 1,96, то гипотеза об идентичности партий с точки зрения доли брака не отвергается (партии можно считать одинаковыми или незначимо различными, если

$$t_{набл} = \frac{(5/200)-(10/300)}{\sqrt{0{,}03 \cdot 0{,}97\,(1/200+1/300)}} = -0{,}54.$$

вспомнить, что речь идет об однократной выборке, на основе которой принимается одно из двух решений: принять иди отвергнуть гипотезу).

Рассмотрим теперь случай, когда объемы выборок не очень велики. Запишем данные выборок в таблицу

Совокупность	Выборка		
	Объем	число наблюдений	
		A	$\overline{A}$
№1	n_{1*}	n_{11}	n_{12}
№2	n_{2*}	n_{21}	n_{22}
Всего	$n_{**} = n_{1*}+n_{2*} = n_{*1}+n_{*2}$	n_{*1}	n_{*2}

где в столбце A помещены численности единиц, обладающих изучаемым признаком A, а в столбце $\overline{A}$ - не обладающих A; * означает суммирование по первому или второму индексу.

При справедливости гипотезы $H_0^{\text{П}} : p_1 = p_2$ оценка величины p_0 имеет вид

$$\hat{p}_0 = \frac{n_{11} + n_{21}}{n_{1*} + n_2} = \frac{n_{*1}}{n_{**}}.$$

Тогда в соответствии с этой оценкой теоретическое распределение частот единиц, обладающих и не обладающих изучаемым признаком A, будет выглядеть следующим образом:

Совокупность	Выборка		
	Объем	число наблюдений	
		A	$\overline{A}$
№1	n_{1*}	$n'_{11} = n_{1*}\dfrac{n_{*1}}{n_{**}}$	$n'_{12} = n_{1*}\left(1 - \dfrac{n_{*1}}{n_{**}}\right)$
№2	n_{2*}	$n'_{21} = n_{2*}\dfrac{n_{*1}}{n_{**}}$	$n'_{22} = n_{2*}\left(1 - \dfrac{n_{*1}}{n_{**}}\right)$
Всего	n_{**}	n_{*1}	n_{*2}

Теперь проверка сводится к установлению существенности разницы между векторами эмпирических частот $(n_{11}, n_{12}, n_{21}, n_{22})$ и теоретических частот $(n'_{11}, n'_{12}, n'_{21}, n'_{22})$. Проверку можно осуществить с помощью критерия согласия χ^2 при числе степеней свободы $\nu = 1$, так как только одна из четырех величин $n_{11}, n_{12}, n_{21}, n_{22}$ является независимой.

Статистика критерия имеет вид:

$$\chi^2_{набл} = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}},$$

а критическая область определяется неравенством

$$\chi^2_{набл} > \chi^2(\alpha,\nu) = \mathrm{P}_i^{-1}(\alpha,\nu), \qquad (6)$$

где $\chi^2(\alpha,\nu)$ находится по таблицам распределения на пересечении столбца α и строки $\nu = 1$.

Следует отметить, что применение критерия χ^2, вообще говоря, требует, чтобы частоты $n'_{11}, n'_{12}, n'_{21}, n'_{22}$ были не меньше 5.

Пример 7. Сравнить два участка цеха, выпускающих одну и ту же продукцию, с точки зрения долей рабочих, не выполнивших сменного задания.

Участок	Выборка		
	Объем	число наблюдений	
		не выполнившие задание	выполнившие задание
№1	50	15	35
№2	70	20	50
Всего	120	35	85

Решение. Так как оценка доли p рабочих, не выполнивших сменного задания, при гипотезе $p_1 = p_2 = p$ равна $\hat{p} = \frac{35}{120}$, то в соответствии с вышеприведенными формулами получаем следующую таблицу теоретических данных:

Участок	Выборка		
	Объем	число наблюдений	
		не выполнившие задание	выполнившие задание
№1	50	$50 \cdot \frac{35}{120} = 14{,}6$	$50 \cdot \frac{85}{120} = 35{,}4$
№2	70	$70 \cdot \frac{35}{120} = 20{,}4$	$70 \cdot \frac{85}{120} = 49{,}6$
Всего	120	35	85

Находим наблюдаемое значение статистики χ^2

$$\chi^2_{набл} = \frac{0{,}4^2}{14{,}6} + \frac{0{,}4^2}{35{,}4} + \frac{0{,}4^2}{20{,}4} + \frac{0{,}4^2}{49{,}6} = 0{,}03.$$

Взяв уровень значимости $\alpha = 0{,}01$, находим по таблицам $\chi^2(0{,}01;1) = 2{,}71$. Таким образом, гипотеза об однородности обоих участков цеха с рассматриваемой точки зрения не отвергается.

Рассмотрим случай сравнения генеральных долей, когда одна или более величин $n'_{11}, n'_{12}, n'_{21}, n'_{22}$ меньше 5. Здесь можно воспользоваться расчетом условного распределения одной из величин n_{ij}, например, n_{11} при справедливости $H_0^{\text{П}}: p_1 = p_2 = p$, так как лишь одна из них независима.

Если совокупности одинаковы, то наиболее предпочтительной оценкой неизвестной общей генеральной доли является $\hat{p} = \frac{n_{*1}}{n_{**}}$. Образуем новую генеральную совокупность, объединив обе выборки объемом в $N = n_{**}$ единиц, из которых $M = n_{*1}$ единиц обладают изучаемым признаком A, а остальные $N\text{-}M = n_{*2}$ единиц этим признаком не обладают. Из этой совокупности производится безвозвратная выборка объемом в $n = n_{1*}$ единиц. Тогда число единиц, обладающих признаком A, в этой выборке $m = n_{11}$ распределено по

гипергеометрическому закону. Таким образом рассчитывается ряд распределения тестовой статистики n_{11}

$$X = \begin{Bmatrix} n_{11} & 0 & 1 & \dots n_{1*} \\ P(n_{11}) & P(0) P(1) \dots P(n_{1*}) \end{Bmatrix},$$

где

$$P(n_{11}) = \frac{n_{*1}! n_{*2}! n_{1*}! n_{2*}!}{n_{**}! n_{11}! n_{12}! n_{21}! n_{22}!},$$

при условии, что проверяемая гипотеза типа сравнения со стандартом $H_0^{\text{П}} : p = p_0$ справедлива.

Если n_{1*} достаточно мало, то можно использовать правосторонний критерий, т.е. в качестве альтернативы взять гипотезу $H_1 : p > p_0$. Тогда критическая область определяется неравенством

$$n_{11} > n_{11кр},$$

где $n_{11кр}$ находится из решения уравнения

$$\sum_{n_{11}=0}^{n_{11кр}} P(n_{11}) = 1 - \alpha. \quad (7)$$

Полученный критерий носит название точного критерия Фишера.

Пример 8. Сравнить работу двух станков с точки зрения выпуска дефектных изделий за некоторый период.

Станок	Выборка		
	Объем	Дефектные	Годные
№1	30	3	27
№2	50	1	49
Всего	80	4	76

Решение. Построим ряд распределения случайной величины n_{11}-числа дефектных изделий в первой выборке – при условии, что доля дефектных изделий, изготовленных на станке №1, равна доле дефектных изделий, изготовленных на станке №2:

$$P(n_{11} = 0) = \frac{30!\,50!\,4!\,76!}{80!\,0!\,30!\,4!\,46!} = 0{,}146,$$

$$P(n_{11} = 1) = \frac{30!\,50!\,4!\,76!}{80!\,1!\,29!\,3!\,47!} = 0{,}372,$$

$$P(n_{11} = 2) = \frac{30!\,50!\,4!\,76!}{80!\,2!\,28!\,2!\,48!} = 0{,}337,$$

$$P(n_{11}=3)=\frac{30!\,50!\,4!\,76!}{80!\,3!\,27!\,1!\,49!}=0,128,$$

$$P(n_{11}=4)=\frac{30!\,50!\,4!\,76!}{80!\,4!\,26!\,0!\,50!}=0,017.$$

В качестве приближенного решения уравнения (7) возьмем при $\alpha = 0,05$ $n_{11кр} = 3$. Тогда критическая область, определяемая неравенством $n_{11} > 3$, будет иметь уровень значимости $\alpha' = 0,017 < \alpha$.

Так как $n_{11набл} = 3$ не попадает в критическую область, то гипотеза об одинаковой работе станков с точки зрения выпуска дефектных изделий за некоторый период не отвергается. Здесь уместно отметить, что принимать гипотезу нецелесообразно, так как $n_{11набл}$ слишком близко к критическому значению, если есть возможность повторить проверку.

Критерий χ^2 применяется и при проверке гипотезы о сравнении долей более чем двух генеральных совокупностей с несколькими признаками. Проверяемая гипотеза утверждает идентичность совокупностей и имеет вид:

$$H_0^{\Pi}: p_{1j}=p_{2j}=...=p_{ij}=p_j=p_{rj}=\frac{1}{r}\sum_{i=1}^{r}p_{ij};\quad j\in\{1,2,...,s\},$$

где r - число сравниваемых генеральных совокупностей,

s – число характеристик, по которым различаются эти совокупности. Так, в рассмотренном выше случае двух генеральных совокупностей с одним альтернативным признаком $r = 2$ и $s = 2$, так как сравниваемых характеристик было две: $p_{11}=p_0$ и $1 - p_{11}=p_{12}=1-p_0$.

Для расчета χ^2 эмпирические данные (данные выборок) удобно изобразить в виде таблицы

Характеристики / Совокупности	1	2	…	j	…	s	Объем выборки
№ 1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	n_{1*}
№ 2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	n_{2*}
…	…	…	…	…	…	…	…
№ i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	n_{i*}
…	…	…	…	…	…	…	…
№ r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	n_{r*}
Всего	n_{*1}	n_{*2}	…	n_{*j}	…	n_{*s}	n_{**}

В этой таблице n_{ij} - наблюдаемые частоты единиц, обладающих признаком j, в выборке из совокупности № i.

Теоретические n'_{ij} частоты вычисляются при справедливости проверяемой гипотезы по формулам

$$n'_{ij}=\frac{n_{i*}\cdot n_{*j}}{n_{**}},\quad i\in\{1,2,...,r\},\quad j\in\{1,2,...,s\}. \qquad (8)$$

Наблюдаемое значение χ^2 определяется выражением

$$\chi^2_{набл} = \sum_{i=1}^{r}\sum_{j=1}^{S}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}.$$

Для получения числа степеней свободы имеем $r \cdot s$ случайных величин n_{ij}, которые связаны r линейными уравнениями

$$\sum_{j=1}^{S} n_{ij} = n_{i*}, \quad i \in \{1,2,...,r\},$$

и s линейными уравнениями

$$\sum_{i=1}^{r} n_{ij} = n_{*j}, \quad j \in \{1,2,...,s\}.$$

Система этих $r+s$ уравнений линейно зависима, так как

$$\sum_{i=1}^{r} n_{i*} = \sum_{j=1}^{S} n_{*j} = n_{**},$$

поэтому ранг системы равен $r+s-1$ и число степеней свободы есть

$$\nu = r \cdot s - (r + s - 1) = (r - 1) \cdot (s - 1).$$

Критическая область определяется неравенством (6).

Пример 9. Сравнить предприятия по качественному составу годной продукции, классифицированной по сортам, на основе следующих данных

Номер предприятия	Сорт			Число годных изделий
	1-й	2-й	3-й	
	число изделий			
1	147	109	39	295
2	184	113	57	354
3	120	192	33	345
4	282	139	49	470
Всего	733	553	178	1464

Решение. Проверим сначала гипотезу о том, что генеральная доля продукции любого сорта не зависит от предприятия, на котором она изготовлена (гипотезу однородности), т.е.

$$H_0 : p_{1j} = p_{2j} = p_{3j} = p_{4j} = \frac{1}{4}\sum_{i=1}^{4} p_{ij}, \quad j \in \{1,2,3\}.$$

Теоретические частоты n'_{ij} получим, используя формулы (8):

$$n'_{11} = \frac{295 \cdot 733}{1464} = 147,7 \qquad n'_{12} = \frac{295 \cdot 552}{1464} = 111,4 \qquad n'_{13} = 35,9$$

$$n'_{21} = \frac{354 \cdot 733}{1464} = 177,2 \qquad n'_{22} = 133,7 \qquad n'_{23} = 43,1$$

$$n'_{31} = 172,7 \qquad n'_{32} = 130,3 \qquad n'_{33} = 42,0$$

$$n'_{41} = 235,3 \qquad n'_{42} = 177,5 \qquad n'_{43} = 57,2$$

Отсюда

$$\chi^2_{набл} = \left(\frac{0,7^2}{147,7} + \frac{2,4^2}{111,4} + \frac{3,1^2}{35,9}\right) + \left(\frac{6,8^2}{177,2} + \frac{20,7^2}{133,7} + \frac{3,9^2}{43,1}\right) +$$
$$+ \left(\frac{52,7^2}{172,7} + \frac{60,3^2}{130,3} + \frac{9^2}{42,0}\right) + \left(\frac{46,7^2}{235,3} + \frac{38,5^2}{177,5} + \frac{8,2^2}{57,2}\right) = 75,4$$

Так как наблюдаемое значение статистики χ^2 больше критического значения $\chi^2_{кр}(0,05;6) = 12,6$, то гипотеза однородности отвергается с вероятностью ошибки 0,05. Таким образом, качественная структура годной продукции существенно зависит от предприятия, на котором она производится.

Выделим теперь предприятия, продукция которых по качественному составу существенно отличается от продукции остальных предприятий. Для решения этого вопроса исключим из задачи сравнения прежде всего предприятие № 3, для которого слагаемое

$$\sum_{j=1}^{3} \frac{\left(n_{3j} - n'_{3j}\right)^2}{n'_{3j}} = \frac{52,7^2}{172,7} + \frac{60,3^2}{130,3} + \frac{9^2}{33} = 48,0$$

вносит самый большой вклад в наблюдаемое значение χ^2. Для оставшихся трех предприятий, совершенно аналогично тому, как мы это делали для четырех предприятий, проверяем гипотезу однородности. Имеем следующие исходные данные

Номер предприятия	Сорт			Число годных изделий
	1-й	2-й	3-й	
	число изделий			
1	147	109	39	295
2	184	113	57	354
4	282	139	49	470
Всего	613	361	145	1119

Отсюда

$$\chi^2 = \left(\frac{14,6^2}{161,6} + \frac{13,8^2}{95,2} + \frac{0,8^2}{38,2}\right) + \left(\frac{9,9^2}{193,9} + \frac{1,2^2}{114,2} + \frac{11,1^2}{45,9}\right) +$$
$$+ \left(\frac{24,5^2}{257,5} + \frac{12,6^2}{151,6} + \frac{11,9^2}{60,9}\right) = 12,24$$

Так как $\chi^2_{набл}$ больше $\chi^2_{кр}\ (0,05;4)=9,49$, то гипотеза однородности отвергается с вероятностью ошибки 0,05. Исключим из задачи сравнения предприятие № 4, дающее наиболее слагаемое (5,71) в $\chi^2_{набл}$. Для оставшихся предприятий № 1 и № 2 имеем эмпирические и теоретические частоты:

$$n_{11}=147;\quad n_{12}=109;\quad n_{13}=39;\quad n_{21}=184;$$
$$n_{22}=113;\quad n_{23}=57;$$

$$n'_{11}=150,5;\quad n'_{12}=100,9;\quad n'_{13}=43,6;\quad n'_{21}=180,5;$$
$$n'_{22}=121,1;\quad n'_{23}=52,4;$$

откуда

$$\chi^2=\left(\frac{3,5^2}{150,5}+\frac{8,1^2}{100,9}+\frac{4,6^2}{43,6}\right)+\left(\frac{3,5^2}{180,5}+\frac{8,1^2}{121,1}+\frac{4,6^2}{52,4}\right)=2,24.$$

Так как $\chi^2_{набл}=2,24$ меньше $\chi^2_{кр}\ (0,05;2)=5,99$, то гипотеза однородности не отвергается; примем эту гипотезу, следовательно, предприятия №1 и №2 идентичны по качественному составу годной продукции, но отличаются в этом смысле от предприятий № 3 и № 4. Проверим теперь, можно ли отнести предприятия № 3 и №4 в одну группу. Имеем следующие эмпирические и теоретические частоты (сохраняя старые индексы):

$$n_{31}=120;\quad n_{32}=192;\quad n_{33}=33;\quad n_{41}=282;$$
$$n_{42}=139;\quad n_{43}=49;$$

$$n'_{31}=170,2;\quad n'_{32}=140,1;\quad n'_{33}=34,7;\quad n'_{41}=231,8;$$
$$n'_{42}=190,9;\quad n'_{43}=47,3.$$

Тогда

$$\chi^2_{набл}=\left(\frac{50,2^2}{170,2}+\frac{51,9^2}{140,1}+\frac{1,7^2}{34,7}\right)+\left(\frac{50,2^2}{231,8}+\frac{51,9^2}{190,9}+\frac{1,7^2}{47,3}\right)=59,2.$$

Так как $\chi^2_{набл}$ больше $\chi^2_{кр}\ (0,05;2)=5,99$, то предприятия № 3 и № 4 различны в смысле качества выпускаемой продукции. Таким образом, имеем разбиение четырех предприятий на три группы по качественному составу годной продукции, классифицированной по сортам.

3.5.3. Гипотеза о нормальном распределении

Расчет теоретических частот $m'_j = m_j^T$ *в предположении о нормальном законе распределения*

При расчете теоретических частот за оценку параметров μ и σ нормального закона распределения принимают значения соответствующих выборочных характеристик, т.е.

$$\mu = \overline{X} = 4{,}002 \text{ и } \sigma = S = 0{,}1378 .$$

Вероятности p_j попадания в j-й интервал и теоретические частоты m_j^T определяются по уравнениям (9).

Результаты расчетов приводятся в табл.3.5.1. Значения интегральной функции Лапласа $\Phi(t)$ находились по данным таблицы стандартного нормального закона, приведенной в Приложении:

$$p_j = P\{a_j < X \le b_j\} = \frac{1}{2}\left[\Phi\left(t_{2j}\right) - \Phi\left(t_{1j}\right)\right], m_j^{\mathrm{T}} = np_j . \qquad (9)$$

Таблица 3.5.1

Интервалы	m_j	t_1	t_2	$\frac{1}{2}\Phi(t_1)$	$\frac{1}{2}\Phi(t_2)$	p_j	$n \cdot p_j$	m_j^T
3,65-3,75	1	-∞	-1,83	-0,5000	-0,4664	0,0336	1,680	2
3,75-3,85	6	-1	-1,10	-0,4664	-0,3645	0,1019	5,095	5
3,85-3,95	11	-1,10	-0,38	-0,3645	-0,1480	0,2165	10,825	11
3,95-4,05	15	-0,38	0,35	-0,1480	0,1368	0,2848	14,240	14
4,05-4,15	9	0,35	1,07	0,1368	0,3577	0,2209	11,045	11
4,15-4,25	6	1,07	1,79	0,3577	0,4635	0,1058	5,290	5
4,25-4,35	2	1,79	∞	0,4635	0,5000	0,0365	1,825	2
Итого	50	-	-	-	-	1,0000	50	50

Проверка гипотезы о нормальном законе распределения

Вычисление фактического (наблюдаемого) значения функции $\chi^2_{набл.} = \sum_{j=1}^{l}\left(m_j - m_j^{\mathrm{T}}\right)^2 / m_j^{\mathrm{T}}$ приведено в табл. 3.5.2 по данным табл. 3.5.1.

Как уже отмечалось, следует учитывать, что при использовании критерия согласия Пирсона должно быть достаточно большим общее число наблюдений $n \ge 50$ и достаточно заполненные частотами интервалы. Если отдельные теоретические частоты окажутся слишком малыми (меньше 5), то при вычислении критерия рекомендуется объединять такие интервалы, а частоты складывать. В случае нормального закона распределения оцениваются два параметра μ и σ, поэтому $r = 2$ и число степеней свободы $\nu = l - 3$, l – общее число интервалов после объединения.

Таблица 3.5.2

m_j	m_j^T	$\left(m_j - m_j^T\right)^2$	$\left(m_j - m_j^T\right)^2 / m_j^T$
1	2	3	4
1 6	2 5	0	0
11	11	0	0
15	14	1	0,071
9	11	4	0,363
6 2	5 2	1	0,143
50	50	-	$\chi^2_{набл.} = 0{,}577$

Из табл. 3.5.2 (графа 4) получаем $\chi^2_{набл.} = 0{,}577$. Табличное (критическое) значение $\chi^2_{кр.}$ определим по таблице 3 Приложения для числа степеней свободы $\nu = 5\text{-}3 = 2$ и уровня значимости $\alpha = 0{,}05 - \chi^2_{кр.}(2;\ 0{,}05) = 5{,}991$.

Так как наблюдаемое значение статистики Пирсона меньше табличного $\left(\chi^2_{набл.} < \chi^2_{кр.}\right)$, то согласно критерию Пирсона выдвинутая гипотеза о нормальном законе распределения не противоречит данным наблюдений.

3.6. Гипотезы о дисперсиях нормально распределенных генеральных совокупностей

3.6.1. Сравнение дисперсии со стандартом

Пусть на основе выборки объема *n* из генеральной совокупности, значение признака в которой распределено по нормальному закону с параметрами $MX = \mu$ и $DX = \sigma^2$, требуется проверить гипотезу

$$H_0 : \sigma^2 = \sigma_0^2.$$

Для проверки используется статистика (выборочная дисперсия)

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

Известно, что статистика $\dfrac{nS^2}{\sigma_0^2}$ при справедливости гипотезы H_0 распределена как χ^2 с $r = n - 1$ степенями свободы.

Критерий, основанный на этой статистике, при альтернативной гипотезе $H_1 : \sigma^2 > \sigma_0^2$ является равномерно наиболее мощным для проверки гипотезы $H_0 : \sigma^2 \le \sigma_0^2$ и задается неравенством

$$\chi^2_{\cdot} = \frac{nS^2}{\sigma_0^2} > \mathrm{P}_i^{-1}(\alpha, n-1),$$

определяющим правосторонюю критическую область.

Критерий для проверки гипотезы $H_0 : \sigma^2 \ge \sigma_0^2$ против альтернативы $H_1 : \sigma^2 < \sigma_0^2$ определяется посредством неравенства, выражающего левосторонюю критическую область

$$\chi^2 = \frac{nS^2}{\sigma_0^2} < \mathrm{P}_i^{-1}(1-\alpha, n-1),$$

и является равномерно наиболее мощным среди всех несмещенных критериев уровня α.

При проверке гипотезы $H_0 : \sigma^2 = \sigma_0^2$ против альтернативы $H_1 : \sigma^2 \ne \sigma_0^2$ равномерно наиболее мощный критерий среди всех несмещенных критериев уровня α задается неравенствами

$$\frac{nS^2}{\sigma_0^2} < C_{1\,\text{кр.}},\ \frac{nS^2}{\sigma_0^2} > C_{2\,\text{кр.}},\ C_{1\,\text{кр.}} < C_{2\,\text{кр.}},$$

где константы $C_{1\,\text{кр.}}$ и $C_{2\,\text{кр.}}$ определяются из уравнений

$$\int_{C_{1\,\text{кр.}}}^{C_{2\,\text{кр.}}} f_{n-1}(y)dy = 1-\alpha,$$

$$\int_{C_{1\,\text{кр.}}}^{C_{2\,\text{кр.}}} f_{n+1}(y)dy = 1-\alpha,$$

где $f_{n-1}(y)$, $f_{n+1}(y)$ – плотности распределения χ^2 соответственно с $\nu = n-1$ и $\nu = n+1$ степенями свободы. Достаточной для практических применений этого критерия служит аппроксимация его критерием с "равными хвостами" с критической областью, задаваемой неравенствами

$$\frac{nS^2}{\sigma_0^2} < \mathrm{P}_i^{-1}\left(1-\frac{\alpha}{2}, n-1\right),\ \frac{nS^2}{\sigma_0^2} > \mathrm{P}_i^{-1}\left(\frac{\alpha}{2}, n-1\right).$$

Такая аппроксимация справедлива при не очень малых n и σ^2, не очень близких к 0 или ∞.

Если генеральное среднее μ является известной величиной, то во всех критериях следует заменить статистику $\dfrac{nS^2}{\sigma_0^2}$ на $\sum_{i=1}^{n}(x_i-\mu)^2 \big/ \sigma_0^2$ и использовать χ^2 с измененным числом степеней свободы $n-1$ на n, однако на практике этот случай встречается редко.

Мощность приведенных критериев (при неизвестном μ) определяется следующими формулами:

$$1-\beta = P\left(\chi^2 > \frac{\sigma_0^2}{\sigma_1^2}\,\mathrm{P}_i^{-1}(\alpha, n-1)\right)$$ для $\mathrm{H}_0 : \sigma^2 \le \sigma_0^2$ против $\mathrm{H}_1 : \sigma_1^2 > \sigma_0^2$,

$$1-\beta = P\left(\chi^2 < \frac{\sigma_0^2}{\sigma_1^2}\,\mathrm{P}_i^{-1}(1-\alpha, n-1)\right)$$ для $\mathrm{H}_0 : \sigma^2 \ge \sigma_0^2$ против $\mathrm{H}_1 : \sigma_1^2 > \sigma_0^2$,

$$1-\beta = P\left(\chi^2 < \frac{\sigma_0^2}{\sigma_1^2}\,\mathrm{P}_i^{-1}\left(1-\frac{\alpha}{2}, n-1\right)\right) + P\left(\chi^2 > \frac{\sigma_0^2}{\sigma_1^2}\,\mathrm{P}_i^{-1}\left(1-\frac{\alpha}{2}, n-1\right)\right)$$

для $H_0 : \sigma^2 = \sigma_0^2$ против $H_1 : \sigma_1^2 \ne \sigma_0^2$,

Пример 1. Точность некоторого измеряемого параметра технологического процесса характеризуется стандартной величиной $\sigma^2 = 400$. Исходя из факта подчинения измеряющихся значений этого параметра нормальному закону, проверить на основе данных контрольной выборки: $n = 25$, $S^2 = 625$, нарушается ли технологический процесс с точки зрения точности параметра.

Решение. Здесь требуется проверить гипотезу $H_0 : \sigma^2 \le 400$ против альтернативы $H_1 : \sigma^2 > 400$. Возьмем $\alpha = 0{,}01$, тогда $\chi^2_{\text{кр.}} = \mathrm{P_i}^{-1}(0{,}01;\ 24) = 43{,}0$. Так как $\chi^2_{\text{набл.}} = \dfrac{25 \cdot 625}{400} = 39$ меньше $\chi^2_{\text{кр.}}$, то нет оснований считать, что процесс протекает с нарушениями.

Пример 2. Исходя из данных примера 1, вычислить мощность критерия при альтернативе $H_1 : \sigma^2 = 625$.

Решение. Согласно сказанному выше

$$1-\beta = P\left(\chi^2 > \frac{\sigma_0^2}{\sigma_1^2}\,\mathrm{P}_i^{-1}(\alpha, n-1)\right) = P\left(\chi^2 > \frac{400}{625} \cdot 43\right) = P\left(\chi^2 > 27{,}52\right)$$

Используя таблицу распределения χ^2 в строке $\nu = 24$, будем иметь

$$\mathrm{P_i}^{-1}(0{,}30;\ 24) = 27{,}1,\ \mathrm{P_i}^{-1}(0{,}20;\ 24) = 29{,}6.$$

Отсюда приблизительно $1 - \beta = 0{,}25$.

Пример 3. Найти критическую область для проверки гипотезы $H_0 : \sigma^2 = 50$ с уровнем значимости $\alpha = 0{,}05$ на основе выборки $n = 50$ для статистики S^2.

Решение. Так как альтернативой здесь служит противоположная гипотеза $\sigma^2 \ne 50$, то будем иметь двустороннюю критическую область. Таблица χ^2 обрывается при $\nu = n - 1 = 30$. Поэтому воспользуемся (как делалось ранее) статистикой

Так как область принятия гипотезы имеет вид для статистики T

$$T = \sqrt{\frac{2S^2}{\sigma_0^2}}\,n - \sqrt{2(n-1)-1} = \sqrt{2S}\ - \sqrt{97}.$$

$$\left|t_{\text{набл.}}\right| \le \Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}(1-\alpha)$$

или

$$\left|\sqrt{2S^2} - \sqrt{97}\right| \le 1{,}96,$$

то

$$31{,}12 \le S^2 \le 69{,}74.$$

Таким образом, критическую область можно записать в виде:

$$S^2 < 31,\ S^2 > 70,$$

и она имеет уровень значимости несколько меньший 0,05.

3.6.2. Сравнение нескольких генеральных дисперсий

При сравнении двух (или более) нормально распределенных генеральных совокупностей обычно в первую очередь решается задача сравнения их дисперсий (или средних квадратических отклонений). Это связано с тем, что наиболее эффективно решается задача сравнения средних в случае равенства генеральных дисперсий.

Рассмотрим задачу проверки гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ против альтернативы $H_1 : \sigma_1^2 \neq \sigma_2^2$, где σ_1^2 и σ_2^2 есть генеральные дисперсии нормально распределенных совокупностей № 1 и № 2, на основе соответствующих (независимых) выборок объемом n_1 и объемом n_2.

Рассмотрим сначала выборки достаточно большого объема. Так как среднее квадратическое отклонение S выборки, взятой из нормальной генеральной совокупности с дисперсией σ^2, распределено асимптотически нормально с параметрами MS = σ и $DS = \sigma^2/2n$, то при справедливости гипотезы H_0 статистики S_1 и S_2 распределены нормально с параметрами

$MS_1 = MS_2 = \sigma$, $DS_1 = \sigma^2/2n_1$, $DS_2 = \sigma^2/2n_2$.

Тогда статистика

$$T = \frac{S_1 - S_2}{\sigma\sqrt{1|(2n_1) + 1|(2n_2)}}$$

распределена по нормированному нормальному закону и может служить в качестве тестовой статистики. Разумеется, что в рассматриваемых условиях σ^2 можно заменить оценкой (несмещенной):

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

Критическая область размера α определяется как обычно:

$$\left|t_{\text{набл.}}\right| > \Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}(1-\alpha).$$

Пример 4. Проверить идентичность настройки двух станков, выпускающих одну и ту же деталь, на основе следующих выборочных данных измерения основного параметра деталей: $n_1 = 100$; $S_1 = 3$ мм; $n_2 = 200$; $S_2 = 8$ мм.

Решение. Оценим σ^2:

$$\hat{S}^2 = \frac{100\cdot 3^2 + 200\cdot 8^2}{100+200-2} = \frac{13700}{298}$$

Вычислим

$$t_{\text{набл.}} = \frac{3-8}{\sqrt{\frac{13700}{298}\left(\frac{1}{200}+\frac{1}{400}\right)}} = -\frac{5}{0{,}59} = -8{,}47.$$

Так как при $\alpha = 0{,}05$ имеем $\left|t_{\text{набл.}}\right| > 1{,}96$, то настройка станков считается различной (с вероятностью ошибки 0,05).

Пусть теперь объемы n_1 и n_2 выборок невелики. Так как при справедливости гипотезы $n_1S_1^2/\sigma^2$ имеет распределение χ^2 с $\nu_1 = n_1 - 1$ степенями свободы, а $n_2S_2^2/\sigma^2$ имеет распределение χ^2 с $\nu_2 = n_2 - 1$ степенями свободы, то статистика (дисперсионное отношение)

$$F = \frac{n_1S_1^2/(n_1-1)}{n_2S_2^2/(n_2-1)}$$

имеет распределение Фишера-Снедекора (F – распределение) с $\nu_1 = n_1 - 1$ степенями свободы в числителе и с $\nu_2 = n_2 - 1$ степенями свободы в знаменателе и является тестовой. Если в качестве альтернативы берется $H_1 : \sigma_1^2 \neq \sigma_2^2$, то практически достаточной является критическая область $F > F^{-1}_{i\alpha/2}(\nu_1, \nu_2)$. Иногда вместо альтернативы $H_1 : \sigma_1^2 \neq \sigma_2^2$ берут альтернативу $H_1 : \sigma_1^2 > \sigma_2^2$ и используют, следовательно, правостороннюю критическую область

$$F_{набл.} > F_{кр.} = F_i^{-1}(\alpha, \nu_1, \nu_2),$$

причем совокупности нумеруются так, чтобы числитель дисперсионного отношения был не меньше знаменателя.

Пример 5. Проверить гипотезу $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ на основе выборочных данных $n_1 = 15$; $S_1^2 = 0{,}1$; $n_2 = 20$; $S_2^2 = 0{,}9$.

Решение. Находим несмещенные оценки дисперсии σ^2 по данным выборок

$$\frac{n_1S_1^2}{n_1-1} = \frac{15\cdot 0{,}1}{14} = 0{,}107; \quad \frac{n_2S_2^2}{n_2-1} = \frac{25\cdot 0{,}9}{24} = 0{,}938.$$

Находим дисперсионное отношение

$$F_{\text{набл.}} = \frac{0{,}938}{0{,}107} = 8{,}77.$$

По таблицам F распределения для $\alpha = 0{,}05$ на пересечении столбца $25 - 1 = 24$ и строки $15 - 1 = 14$ находим $F_{\text{кр.}} = P_i^{-1}(0{,}05;\ 24;\ 14) = 2{,}35$.
Так как $F_{\text{набл.}}$ больше $F_{\text{кр.}}$, то гипотеза отвергается с вероятностью ошибки 0,05.

Для сравнения более двух генеральных дисперсий приведем лишь два наиболее часто употребляемых критерия.

Критерий Кохрана применяется при проверке гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2 = \sigma^2$ по выборкам одинакового объема n, взятым соответственно из нормальных совокупностей № 1, № 2, …, № r. Тестовой статистикой является

$$G = \frac{\hat{S}_{max}^2}{\sum_{i=1}^{r} \hat{S}_i^2},$$

где $\hat{S}_i^2 = \dfrac{nS_i^2}{n-1}$ есть несмещенная оценка генеральной дисперсии σ_i^2, $\hat{S}_{max}^2$ - наибольшая из r оценок $\hat{S}_i^2$. При справедливости гипотезы H_0 статистика G имеет распределение Кохрана с $\nu = n - 1$ степенями свободы и количеством выборок r. В качестве критической области берется правосторонняя область, задаваемая неравенством:

$$G_{\text{набл.}} > G_{\text{кр.}}(\alpha,\ \nu,\ r).$$

Пример 6. По четырем выборкам из нормальных генеральных совокупностей объемом n = 10 получены исправленные дисперсии $\hat{S}_1^2 = 0{,}52$; $\hat{S}_2^2 = 0{,}64$; $\hat{S}_3^2 = 0{,}25$; $\hat{S}_4^2 = 0{,}48$. Проверить гипотезу равенства генеральных дисперсий.

Решение. Находим

$$G_{\text{набл.}} = \frac{0{,}65}{0{,}52 + 0{,}65 + 0{,}35 + 0{,}48} = \frac{0{,}65}{2} = 0{,}325$$

и сравниваем $G_{\text{набл.}}$ с $G_{\text{кр.}}(0{,}05;\ 9;\ 4) = 0{,}502$.
Так как $G_{\text{набл.}}$ меньше $G_{\text{кр.}}$, то гипотеза о равенстве генеральных дисперсий не отвергается.

При разных объемах выборок применяется критерий Бартлета, основанный на том, что статистика

$$\chi^2 = \left\{ (n_* - r)\, ln \left[\frac{\sum_{i=1}^{r} (n_i - 1)\hat{S}_i^2}{n_* - r} \right] - \sum_{i=1}^{r} (n_i - 1)\, ln\, \hat{S}_i^2 \right\} \cdot C,$$

где $C = \left[1 + \frac{\sum_{i=1}^{r} \frac{1}{n_i - 1} - \frac{1}{\sum n_i - 2}}{3(r-1)}\right]^{-1}$, $n_* = \sum_{i=1}^{r} n_i$ имеет приблизительно распределение χ^2 с $\nu = r - 1$ степенями свободы. Критическая область является правосторонней:

$$\chi^2_{\text{набл.}} > \chi^2_{\text{кр.}} = \text{P}_\text{i}^{-1}(\alpha, \text{r} - 1).$$

Пример 7. На основе данных пяти выборок из нормальных генеральных совокупностей

№№	объем	S_i^2
1	15	4,5
2	16	9,2
3	10	7,8
4	9	8,5
5	20	4,9

Проверить гипотезу о равенстве генеральных дисперсий.

Решение. Вычисления удобно производить в таблице

i	$n_i - 1$	lnS_i^2	$(n_i - 1)S_i^2$	$(n_i - 1)lnS_i^2$
1	14	1,50	63,0	21,00
2	15	2,22	138,0	33,30
3	9	2,05	70,2	18,45
4	8	2,14	68,0	17,12
5	19	1,59	93,1	30,21
Итого	70 – 5	–	432,3	120,08

Отсюда

$$\chi^2 = [(70 - 5) \cdot ln\frac{432,3}{70 - 5} - 120,08] \cdot$$

$$\cdot \left[1 + \frac{\frac{1}{14} + \frac{1}{15} + \frac{1}{9} + \frac{1}{8} + \frac{1}{19} - \frac{1}{70 - 5}}{3(5-1)}\right]^{-1} =$$

$$= (122,85\text{-}120,08) \cdot 0,97.$$

Так как $\chi^2_{\text{набл.}} = 2,77 \cdot 0,97$ меньше $\chi^2_{\text{кр.}} = \text{P}_\text{i}^{-1}(0,05;\ 4) = 9,49$, то гипотеза о равенстве генеральных дисперсий не отвергается. Следует отметить, что в данном случае множитель С можно не считать, так как С<1, а 2,77<9,49.

3.7. Гипотезы о генеральных средних нормально распределенных совокупностей

3.7.1. Сравнение генеральной средней со стандартом

При проверке гипотез о сравнении среднего значения μ признака в нормально распределенной генеральной совокупности при известной генеральной дисперсии σ^2 со стандартом μ_0 употребляется тестовая статистика

$$T = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}},$$

распределение которой подчиняется нормальному закону.

Выпишем неравенства, соответствующие наиболее предпочтительным критериям при различных гипотезах.

1) Гипотезе $H_0 : \mu \le \mu_0$ против гипотезы $H_1 : \mu > \mu_0$ соответствует правосторонняя критическая область: $t > \Phi_*^{-1}(1 - \alpha) = \Phi^{-1}(1-2\alpha)$.

2) Гипотезе $H_0 : \mu \ge \mu_0$ против гипотезы $H_1 : \mu < \mu_0$ соответствует левосторонняя критическая область: $t_{\text{набл.}} < \Phi_*^{-1}(\alpha) = -\Phi^{-1}(1-2\alpha)$.

Таким образом, для (1) и (2) критическая область записывается как $|t| > \Phi^{-1}(1-2\alpha)$.

Критерии являются равномерно наиболее мощными. Мощность определяется по формуле

$$1-\beta = \Phi_*\left(\frac{|\mu-\mu_0|}{\sigma/\sqrt{n}} + \Phi_*^{-1}(\alpha)\right) = \frac{1}{2}\left[1 + \Phi\left(\frac{|\mu_1-\mu_0|}{\sigma/\sqrt{n}} - \Phi^{-1}(1-2\alpha)\right)\right].$$

3) Гипотезе $H_0^{\text{п}} : \mu = \mu_0$ против гипотезы $H_1 : \mu \ne \mu_0$ соответствует симметричная критическая область $|t\ | > \Phi_*^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}(1-\alpha) = \Phi^{-1}(\gamma)$.

Критерий является равномерно наиболее мощным среди всех несмещенных критериев уровня α. Мощность определяется по формуле:

$$1-\beta = \Phi_*\left(\Phi_*^{-1}\left(\frac{\alpha}{2}\right) - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) + \Phi_*\left(\Phi_*^{-1}\left(\frac{\alpha}{2}\right) + \frac{\mu-\mu_0}{\sigma/\sqrt{n}}\right) =$$

$$= 1 - \frac{1}{2}\left[\Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + \Phi^{-1}(1-\alpha)\right) - \Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - \Phi^{-1}(1-\alpha)\right)\right].$$

Если проверяется гипотеза $H_0^{\text{п}} : \mu = \mu_0$ против гипотезы $H_1^{\text{п}} : \mu = \mu_1 > \mu_0$, то для заданных вероятностей ошибок α и β легко найти критическую границу $\overline{x}_{\text{кр.}}$ критической области $\overline{X} > \overline{X}_{\text{кр.}}$ и объем выборки:

$$\overline{X}_{\text{кр.}} = \frac{\mu_1 \Phi_*^{-1}(1-\alpha) - \mu_0 \Phi_*^{-1}(\beta)}{\Phi_*^{-1}(1-\alpha) - \Phi_*^{-1}(\beta)} = \frac{\mu_1 \Phi^{-1}(1-2\alpha) + \mu_0 \Phi^{-1}(1-2\beta)}{\Phi^{-1}(1-2\alpha) + \Phi^{-1}(1-2\beta)}.$$

$$n = \frac{\left[\Phi_*^{-1}(1-\alpha) - \Phi_*^{-1}(\beta)\right]^2 \sigma^2}{(\mu_1 - \mu_0)^2} = \frac{\left[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-2\beta)\right]^2 \cdot \sigma^2}{(\mu_1 - \mu_0)^2}.$$

Пример 8. Найти критическую область и объем выборки для проверки гипотезы $H_0^{\text{п}}$: $\mu = 100$ против альтернативы $H_1^{\text{п}}$: $\mu = 120$, если известна $\sigma^2 = 625$ и даны $\alpha = 0{,}01$; $\beta = 0{,}05$.

Решение. $\overline{x}_{\text{кр.}} = \dfrac{120 \cdot 2{,}33 - 100 \cdot (-1{,}64)}{2{,}33 - (-1{,}64)} \approx 112,$

следовательно, критическая область определяется неравенством

$$\overline{x} > 112.$$

Объем выборки равен $n = \dfrac{3{,}97^2 \cdot 625}{400} = 25.$

При проверке гипотез о сравнении среднего значения μ признака в нормально распределенной генеральной совокупности при неизвестной генеральной дисперсии σ^2 со стандартом μ_0 употребляется тестовая статистика

$$T = \frac{\overline{X} - \mu_0}{\hat{S}/\sqrt{n}} = \frac{\overline{X} - \mu_0}{S/\sqrt{n-1}},$$

имеющая распределение Стьюдента с $\nu = n - 1$ степенями свободы (t – распределение).

Выпишем неравенства, соответствующие наиболее предпочтительным критериям при различных гипотезах.

4) Гипотезе $H_0 : \mu \le \mu_0$ против гипотезы $H_1 : \mu > \mu_0$ соответствует правосторонняя критическая область: $t > St^{-1}(2\alpha, n-1)$.

5) Гипотезе $H_0 : \mu \ge \mu_0$ против гипотезы $H_1 : \mu < \mu_0$ соответствует левосторонняя критическая область: $t < -St^{-1}(2\alpha, n-1)$.

6) Гипотезе $H_0 : \mu = \mu_0$ против гипотезы $H_1 : \mu \ne \mu_0$ соответствует симметричная критическая область: $|t| > St^{-1}(\alpha, n-1)$.

Критерии являются равномерно наиболее мощными среди всех несмещенных критериев. Расчет мощности производится аналогично случаю с известной генеральной дисперсией.

Пример 9. Проверить гипотезу $H_0 : \mu = 30$ против $H_1 : \mu \ne 30$ с уровнем $\alpha = 0{,}05$ на основе данных выборки: $n = 16$; $\overline{x} = 25$; $S^2 = 25$. Вычислить мощность критерия при $\mu_1 = 25$.

Решение. Находим по таблице распределения Стьюдента $St^{-1}(\alpha;n-1)=St^{-1}(0{,}05;15)=2{,}131$ и сравним с $t_{\text{набл.}} = \dfrac{25-30}{5/\sqrt{16}} = -4.$ Так как $|t_{\text{набл.}}|$ больше $St^{-1}(\alpha, n-1)$, то гипотеза $H_0 : \mu = 30$ отвергается с вероятностью ошибки 0,05. Для расчета мощности применим формулы:

$$1-\beta=\mathrm{P}\left(\bar{x}<\bar{x}''_{кр}/H_1\right)+\mathrm{P}\left(\bar{x}>\bar{x}'_{кр}/H_1\right),$$

$$\bar{x}'_{кр.}=\mu_0+\frac{\hat{S}}{\sqrt{n}}St^{-1}(\alpha,n-1)=32{,}664,$$

$$\bar{x}''_{кр.}=\mu_0-\frac{\hat{S}}{\sqrt{n}}St^{-1}(\alpha,n-1)=27{,}336.$$

Используя ближайшее табличное значение распределения Стьюдента, получим:

$$1-\beta=P\left(t<\frac{\bar{x}''_{кр}-\mu_1}{\hat{S}^2/\sqrt{n}}\right)+P\left(t>\frac{\bar{x}_{кр}}{\hat{S}^2/\sqrt{n}}\right)=$$

$$=P(1<1{,}869)+P(t>6{,}131)\cong 1-\frac{0{,}1}{2}+0=0{,}95.$$

3.7.2. Сравнение нескольких генеральных средних

Рассмотрим две нормально распределенные генеральные совокупности с известными дисперсиями σ_1^2 и σ_2^2 и неизвестными μ_1 и μ_2.

На основе выборок n_1 и n_2 из этих совокупностей требуется проверить гипотезу $H_0^{\text{п}}$: $\mu_1=\mu_2=\mu$.

Тестовой статистикой является величина

$$T=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}},$$

распределенная при истинности гипотезы H_0 по нормированному нормальному закону.

В зависимости от конкурирующей гипотезы $H_1:\ \bar{x}_1>\bar{x}_2$, $H_1:\ \bar{x}_1<\bar{x}_2$ или $H_1:\ \bar{x}_1\neq\bar{x}_2$ строится критическая область соответственно $t_{\text{набл.}}>\Phi_*^{-1}(1-\alpha)$,

$$t_{\text{набл.}}<\Phi_*^{-1}(\alpha)\ \text{или}\ \left|t_{\text{набл.}}\right|>\Phi_*^{-1}\left(1-\frac{\alpha}{2}\right).$$

При неизвестных генеральных дисперсиях либо требуется достаточно большой объем выборки для их надежной и точной оценки, либо требуется, чтобы эти дисперсии были пропорциональны с известным коэффициентом пропорциональности $\lambda:\sigma_1^2=\lambda\cdot\sigma_2^2$. В других случаях известные приближенные критерии с практической точки зрения не всегда эффективны (проблема Беренса-Фишера).

Если генеральные дисперсии одинаковы ($\lambda = 1$), то для проверки гипотезы H_0 : $\mu_1 = \mu_2 = \mu$ используется тестовая статистика

$$T = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{n_1 S_1^2 + n_{.2} S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}},$$

имеющая распределение Стьюдента с $\nu = n_1 + n_2 - 2$ степенями свободы. Вид критической области зависит, как обычно, от конкурирующей гипотезы.

Сравнение более двух генеральных средних при одинаковых генеральных неизвестных дисперсиях производится в дисперсионном анализе.

3.8. Пояснения, примеры и решения задач.

1. Рассмотрим графическое изображение основных понятий статистической проверки гипотез.

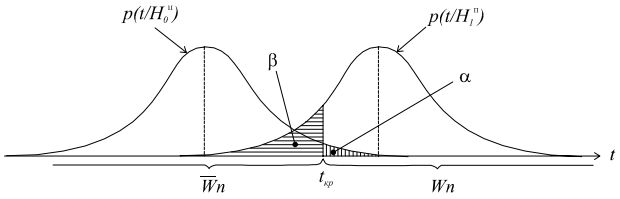

Рис. 3.8.1. Правосторонняя критическая область *W*n

На рисунке 3.8.1 изображена правосторонняя критическая область $t > t_{кр}$, вероятность попадания в которую наблюдаемого значения статистики $T = t_{набл.}$ равна уровню значимости критерия α при условии, что проверяемая гипотеза H_0 истинна:

$$\alpha = P\left(T > t_{кр} \middle| H_0^{п}\right) = \int_{t_{кр}}^{\infty} p\left(t \middle| H_0\right) dt,$$

где α - площадь области, изображенной вертикальной штриховкой.

Область принятия гипотезы $t \le t_{кр}$, вероятность попадания в которую при условии что $H_0^{п}$ ложна или, следовательно, $H_1^{п}$ истинна:

$$\beta = P\left(T \le t_{кр} \middle| H_1\right) = \int_{-\infty}^{t_{кр.}} p\left(t \middle| H_1\right) dt,$$

β - площадь области, заштрихованной горизонтально. Площадь области, расположенной правее прямой $t = t_{кр}$ и под плотностью $p(t|H_1)$ над критической областью, есть мощность критерия

$$1 - \beta = \int_{t_{кр}}^{\infty} p\left(t \middle| H_1^{п}\right) dt .$$

Площадь над областью принятия гипотезы и под плотностью $p(t|H_0)$ есть доверительная вероятность принятия гипотезы H_0, когда она истинна

$$\gamma = 1 - \alpha = \int_{-\infty}^{t_{кр}} p(t|H_0)dt .$$

t – (тестовая) статистика критерия, $t_{кр}$ – критическая точка, граница между W_n и $\overline{W}_n$. Из рисунка 3.8.1 следует, что построенная критическая область дает наиболее мощный несмещенный при $0 < \alpha < 0{,}5$ критерий проверки простой гипотезы против простой альтернативы.

2. Из 250 столовых комплектов, состоящих из шести одинаковых предметов, оказалось следующее распределение числа m комплектов по числу x нестандартных предметов в комплекте

x	0	1	2	3	4	5	6
m	21	29	83	77	30	7	3

Проверим с уровнем значимости $\alpha = 0{,}05$ гипотезу о биномиальном законе распределения x с параметром p – генеральной доле нестандартных предметов. Результаты расчетов поместим в таблицу

x	m	xm	$P_n(x)$	$m^T = P_n(x) \cdot N$	$\frac{(m - m^T)^2}{m^T}$
0	13	0	0,0467	12	0,083
1	44	44	0,1866	47	0,191
2	76	152	0,3110	78	0,051
3	77	231	0,2765	69	0,928
4	30	120	0,1382	34*	0,471
5	7 }10	35	0,0369	9 }10	0
6	3	18	0,0041	1	
Итого	250	600	1,0000	250	1,724

Вычисляем среднюю арифметическую $\bar{x} = \frac{600}{250} = 2{,}4 шт$. Приравниваем среднюю арифметическую математическому ожиданию частоты np биномиального закона. Так как n по условию задачи равно 6 предметам, то оценкой генеральной доли будет $p = 2{,}4 : 6 = 0{,}4$ и $q = 1 - 0{,}4 = 0{,}6$. Теперь можно рассчитать вероятности частот по формуле Бернулли. Получим $P_{10}(x = 0) = C_{10}^0 \cdot 0{,}4^0 \cdot 0{,}6^6 = 0{,}0467$, $P_{10}(x = 1) = C_{10}^1 \cdot 0{,}4^1 \cdot 0{,}6^5 = 0{,}1866$ и т. д. Вероятности рассчитываем с точностью до четырех верных цифр после запятой. Далее получаем теоретические частоты $m^T = P_{10}(x) \cdot N$. Так как по условию задачи имеем $N = 250$ наблюдений, то $m^T(X = 0) = 0{,}0467 \cdot 250 = 12$, $m^T(X = 1) = 0{,}1866 \cdot 250 = 47$ и т. д. Если в результате округления сумма теоретических частот не совпадает с $N = 250$, то добавляем или вычитаем 1 из тех теоретических частот, которые имеют наибольшую ошибку округления. Так $m^T = 34^* = 35 - 1$. Для проверки гипотезы применяем критерий Пирсона хи - квадрат. Число степеней свободы $\nu = k - l - 1$, где k – число значений признака x после объединения малочисленных ($m^T < 5$) значений в группу (для одномодальных распределений такая процедура может быть осуществлена на хвостах распределения). Так как $m^T(X = 6) = 1 < 5$, то объединяем последнее и предпоследнее значения и получаем объединенную группу значений с теоретической частотой, равной $1 + 9 = 10$, и эмпирической частотой, равной соответственно

сумме наблюдаемых частот $m = m(X = 6) + m(X = 5) = 3 + 7 = 10$. Таким образом, $k = 6$. Далее l – число параметров предполагаемого (теоретического) распределения, оцениваемого по выборке. Здесь оценивался один параметр, а именно, $p = \overline{x}/n$, следовательно, $l = 1$, а число степеней свободы равно: $\nu = 6 - 1 - 1 = 4$. По таблицам распределения хи–квадрат Пирсона находим $\chi_{кр}^2 = P_i^{-1}(0{,}05;4) = 9{,}488$. Критическая область имеет вид $\chi^2 > 9{,}488$.

Вычисляем $\chi^2_{набл.}$ по формуле

$$\chi^2_{набл} = \sum_{j=1}^{к} \frac{\left(m_j - m_j^T\right)^2}{m_j^T},$$

предварительно подсчитав каждое слагаемое суммы в последнем столбце таблицы. В итоге получим

$$\chi^2_{набл.} = 0{,}083 + 0{,}191 + \cdots + 0 = 1{,}724.$$

Так как $\chi^2_{набл.}$ не попало в критическую область, то проверяемая гипотеза не отвергается. Примем её. Следовательно, распределение числа нестандартных предметов в комплекте подчиняется биномиальному закону.

3. Так как статистика $\sqrt{\frac{2nS^2}{\sigma^2}}$ распределена асимптотически нормально с параметрами $M\sqrt{\frac{2nS^2}{\sigma^2}} = \sqrt{2n-3}$ и $D\sqrt{\frac{2nS^2}{\sigma^2}} = 1$, то $\frac{\sigma}{\sqrt{2n}} \cdot \sqrt{\frac{2nS^2}{\sigma^2}} = S$ распределена асимптотически также нормально с параметрами

$$M\left(\frac{\sigma}{\sqrt{2n}} \cdot \sqrt{\frac{2nS^2}{\sigma^2}}\right) = \frac{\sigma}{\sqrt{2n}}\sqrt{2n-3} \underset{n\to\infty}{=} \sigma,$$

$$D\left(\frac{\sigma}{\sqrt{2n}} \cdot \sqrt{\frac{2nS^2}{\sigma^2}}\right) = \frac{\sigma^2}{2n} \cdot 1 = \frac{\sigma^2}{2n}.$$

Следовательно, для $n > 31$ можно пользоваться тестовой статистикой стандартного нормального закона. Так, например, для проверки гипотезы H_0: $\sigma = \sigma_0$ против альтернативы H_1: $\sigma = \sigma_1 > \sigma_0$ или $\sigma = \sigma_1 < \sigma_0$ критическая область уровня значимости α имеет вид

$$|t| > t_{кр} = \Phi^{-1}(1 - 2\alpha),$$

а при H_1: $\sigma = \sigma_1 \neq \sigma_0$

$$|t| > t_{кр} = \Phi^{-1}(1 - \alpha),$$

при этом

$$t_{набл} = \frac{S - \sigma_0}{\sigma_0}\sqrt{2n}.$$

Для вычисления мощности критерия рассмотрим, например, правостороннюю критическую область

$$S > S_{кр} = \sigma_0 + \Phi^{-1}(1-2\alpha)\frac{\sigma_0}{\sqrt{2n}},$$

получаемую преобразованием неравенства $t > t_{кр} = \Phi^{-1}(1 - 2\alpha)$. Тогда мощность критерия по определению вычисляется по формуле

$$1-\beta = P(S > S_{кр}|H_1) = P\left(\frac{S-\sigma_1}{\sigma_1}\sqrt{2n} > \frac{S_{кр}-\sigma_1}{\sigma_1}\sqrt{2n}\right) = \frac{1}{2}\left[\Phi(\infty) - \Phi\left(\frac{S_{кр}-\sigma_1}{\sigma_1}\sqrt{2n}\right)\right] =$$

$$= \frac{1}{2}\left[1 - \Phi\left(\frac{S_{кр}-\sigma_1}{\sigma_1}\sqrt{2n}\right)\right] = \frac{1}{2}\left[1 + \Phi\left(\frac{\sigma_1 - S_{кр1}}{\sigma_1}\sqrt{2n}\right)\right], \left(S_{кр} = \sigma_0 + \Phi^{-1}(1-2\alpha)\frac{\sigma_0}{\sqrt{2n}}\right) .$$

Аналогично, при гипотезе $H_1 : \sigma = \sigma_1 < \sigma_0$, получим левостороннюю критическую область

$$S < S_{кр} = \sigma_0 - \Phi^{-1}(1-2\alpha)\frac{\sigma_0}{\sqrt{2n}}$$

и мощность

$$1-\beta = \frac{1}{2}\left[1 + \Phi\left(\frac{S_{кр}-\sigma_1}{\sigma_1}\sqrt{2n}\right)\right].$$

При двусторонней критической области, когда H_1: $\sigma \neq \sigma_1$, мощность критерия вычисляется по формуле

$$1-\beta = 1 - \frac{1}{2}\left[\Phi\left(\frac{S_{кр.1}-\sigma_1}{\sigma_1}\sqrt{2n}\right) - \Phi\left(\frac{S_{кр.2}-\sigma_1}{\sigma_1}\sqrt{2n}\right)\right],$$

где

$$S_{кр.1} = \sigma_0 + \Phi^{-1}(1-\alpha)\frac{\sigma_0}{\sqrt{2n}} \; ; \; S_{кр.2} = \sigma_0 - \Phi^{-1}(1-\alpha)\frac{\sigma_0}{\sqrt{2n}} .$$

4. Приведем формулы расчёта мощности критерия проверки гипотезы $H_0^{п}$: $\mu = \mu_0$ против различных альтернатив $H_1^{1п}$: $\mu = \mu_1 > \mu_0$, $H_1^{2п}$: $\mu = \mu_1 < \mu_0$ и H_1^{3}: $\mu = \mu_1 \neq \mu_0$, когда σ - неизвестная величина:

$$1-\beta = 1 - \frac{1}{2}St\left(\frac{|\mu_1-\mu_0|}{S}\sqrt{n-1} - St^{-1}(2\alpha, n-1), n-1\right)$$

при $H_1^{1п}$ и $H_1^{2п}$,

$$1-\beta = 1 - \frac{1}{2}\left[St\left(\frac{\mu_0-\mu_1}{S}\sqrt{n-1} - St^{-1}(\alpha, n-1), n-1\right) - St\left(\frac{\mu_0-\mu_1}{S}\sqrt{n-1} + St^{-1}(\alpha, n-1), n-1\right)\right]$$

при H_1^{3}. Напомним, что для таблиц распределения Стьюдента, имеющихся в Приложении,

$$P(t_1 \le T \le t_2) = \frac{1}{2}\left[St(t_1, n-1) - St(t_2, n-1)\right]$$

и

$$St(-t, n-1) = 2 - St(t, n-1), St(-\infty) = 2; St(0) = 1, St(\infty) = 0 .$$

5. Проверку однородности двух нормально распределенных совокупностей $X \sim N(\mu_x, \sigma_x^2)$ и $Y \sim N(\mu_y, \sigma_y^2)$ одного и того же признака по имеющимся результатам независимых выборок из них объёмами n_x и n_y соответственно, можно осуществить следующим образом. Так как нормальный закон определяется двумя параметрами μ и σ^2, то можно получить оценки этих параметров: $\bar{x}, \bar{y}, S_x^2, S_y^2$.. Сначала проверяется гипотеза о равенстве генеральных дисперсий, например, с помощью критерия Фишера – Снедекора. Если гипотеза отвергается, то задача решена – совокупности не однородны из-за различных дисперсий. В противном случае гипотеза о равенстве дисперсий принимается и проверяется при этом условии гипотеза о равенстве генеральных средних по критерию Стьюдента. Если гипотеза отвергается, то получаем ответ: генеральные совокупности распределены неодинаково из-за различия в средних, если же гипотеза не отвергается, то можно её принять. Следует отметить, что рассматриваемые критерии чувствительны к отклонениям от нормальности распределений и различию в объемах выборок, поэтому предлагаемую проверку нельзя считать окончательной в случае принятия гипотезы.

3.9. Упражнения

1. Проверить гипотезу о том, что генеральная доля учащихся колледжа, успешно сдавших экзамен, составляет 90%, если при выборке 10 учащихся оказалось, что 4 из них экзамен не сдали. Уровень значимости равен 0,05. Проверить мощность критерия полученной двусторонней критической области при гипотезе H_1: $p_1 = 50\%$.

2. При контроле качества изделий оказалось, что среди 100 проконтролированных изделий партии 70 изделий не имели ни одного дефекта. С уровнем значимости $\alpha = 0{,}1$ проверить гипотезу о том, что генеральная доля изделий в партии составляет 0,8 против 0,6. Вычислить мощность критерия.

3. По данным задачи 1 из 1.6 проверить гипотезу о нормальном законе распределения по критерию хи-квадрат Пирсона при $\alpha = 0{,}1$.

4. По данным задачи 3 из 1.6 проверить гипотезу о пуассоновом законе распределения с $\alpha = 0{,}01$ и проверить гипотезу о биномиальном законе распределения. Взять уровень значимости $\alpha = 0{,}05$ и применить критерий Пирсона хи-квадрат.

5. На уровне значимости $\alpha = 0{,}05$ сравнить успеваемость 3-х групп учащихся колледжа по результатам контрольного теста по данным, приведенным в таблице

NN групп	Оценки ответов				Кол-во
	отлично	хорошо	удовлетвор.	неудовлетвор.	уч-ся
1	2	6	19	3	30
2	1	8	10	1	20
3	4	10	15	1	25

Использовать критерий Пирсона хи-квадрат.

6. По данным задачи 5 из 2.6 проверить гипотезу о том, что генеральная средняя урожайность ржи составляет 16 ц/га против альтернативы – 18 ц/га на уровне значимости

0,01 (использовать нормальное распределение урожайности пшеницы на участках). Вычислить мощность критерия.

7. Исходя из условий задачи 6 (2.6) проверить:

а) H_0: $\mu = 0{,}90$ млн руб. против H_1: $\mu = 0{,}70$ млн руб., вычислить мощность критерия, $\alpha = 0{,}05$;

б) H_0: $\sigma = 0{,}05$ млн руб. против H_1: $\sigma = 0{,}12$ млн руб., вычислить мощность критерия. Применить критерий на основе асимптотически нормального распределения статистики S, $\alpha = 0{,}01$.

8. На основе данных выборки из нормально распределенной генеральной совокупности объемом $n = 30$ получено $S = 10$. С уровнем значимости $\alpha = 0{,}01$ проверить гипотезуH_0: $\sigma = 20$. Вычислить мощность критерия для полученной двусторонней критической области при гипотезе H_1: $\mu = 9$.

9. Дано: $n = 20$, $\bar{x} = 15$, $\sigma = 5$. На уровне значимости $\alpha = 0{,}1$ проверить гипотезу H_0: $\mu = 20$ против H_1: $\mu = 9$. Вычислить мощность критерия.

10. С двух хлебозаводов на контроль были взяты некоторые количества изделий. В результате проверки были получены следующие данные:

изделия хлебозавода №1: 15 изделий, $\bar{x}_1 = 1{,}006$ кг, $S_{x_1} = 0{,}012$ кг,

изделия хлебозавода №2: 20 изделий, $\bar{x}_2 = 0{,}997$ кг, $S_{x_2} = 0{,}015$ кг.

Проверить гипотезу однородности распределений веса изделий на уровне значимости 0,1, принять гипотезу о нормальном законе распределения веса изделий.

11. Для сравнения четырех нормально распределенных совокупностей были вычислены четыре выборочные дисперсии
$S_1^2 = 5{,}21$, $S_2^2 = 6{,}25$, $S_3^2 = 5{,}34$, $S_4^2 = 7{,}97$
по независимым выборкам объема $n_1 = n_2 = n_3 = n_4 = 9$. На уровне значимости $d = 0{,}01$ проверить гипотезу однородности (равенства) генеральных дисперсий, применив критерий Кохрана.

12. Для сравнения трех нормально распределенных генеральных совокупностей на основе выборок объемом $n_1 = 5$, $n_2 = 10$, $n_3 = 7$ были получены выборочные дисперсии $S_1^2 = 7{,}21$, $S_2^2 = 8{,}45$, $S_3^2 = 6{,}50$. Проверить гипотезу однородности дисперсий по критерию Бартлета на уровне значимости $\alpha = 0{,}05$.

4. СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ СВЯЗИ

Рассмотрим некоторые простейшие методы и модели статистического исследования связи между признаками.

Несмотря на то, что идея связи вызывает представление о строгой (функциональной) зависимости, при исследовании массовых наблюдаемых явлений такая зависимость отсутствует. Причиной этого служат многие, не поддающиеся учету факторы, являющиеся источником размытости наблюдаемой зависимости, рассеяния значений наблюдаемых признаков, вызванному изменением уровней неконтролируемых, "мешающих" факторов. Задачей статистического анализа является обнаружение и измерение существующей связи, несмотря на помехи неконтролируемых факторов, между изучаемыми признаками.

Следует отметить, что получаемые результаты статистического анализа зависимостей недостаточны для надежности вывода о причинной связи, нужны аргументы, относящиеся к физической природе явлений.

4.1. Корреляционный анализ

Корреляционный анализ является одним из методов статистического анализа взаимозависимости нескольких признаков.

В настоящее время он определяется как метод, применяемый тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа состоит в оценке корреляционной матрицы генеральной совокупности по выборке и определении на ее основе оценок других коэффициентов связи.

Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) состоит в оценке уравнений регрессии.

Для компактности изложения и удобства восприятия материала приводятся некоторые формулы и схемы расчетов, рассматривавшиеся в 1.4 и 1.5.

4.1.1. Двумерная модель

Рассмотрим генеральную совокупность с двумя признаками X и Y, совместное распределение которых задано плотностью двумерного нормального закона:

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} - \exp\{R_2(x, y)\},$$

где $R,(x, y) = -\frac{1}{2\sqrt{1-\rho^2}}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\cdot\frac{y-\mu_y}{\sigma_y} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]$,

определяемого пятью параметрами:

$$M[x] = \mu_x,\ D_x = \sigma_x^2,\ M[y] = \mu_y,\ D_y = \sigma_y^2,\quad M\left[\frac{X-\mu_x}{\sigma_x}\cdot\frac{Y-\mu_y}{\sigma_y}\right] = \rho\ \left(\rho^2 \neq 1\right).$$

Имея эти параметры, можно получить уравнения линий регрессии, показывающих изменение условных математических ожиданий в зависимости от изменения соответствующих значений случайных аргументов:

$MY/X - MY = \beta_{yx}(X - MX)$ – прямая регрессии Y по X;
$MY/Y - MX = \beta_{xy}(Y - MY)$ – прямая регрессии X по Y;

$$\beta_{yx} = \rho\frac{\sigma_y}{\delta_x}$$ – коэффициент регрессии Y по X;

$$\beta_{xy} = \rho\frac{\sigma_x}{\delta_y}$$ – коэффициент регрессии X по Y.

Полезно вспомнить, что квадрат коэффициента корреляции ρ^2, то есть коэффициент детерминации, в рассматриваемой модели указывает долю дисперсии одной случайной величины, обусловленную вариацией другой.

Задача двумерного корреляционного анализа состоит, прежде всего, в оценке пяти параметров, определяющих генеральную совокупность.

Рассмотрим выборку из генеральной совокупности (X, Y) объема n:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n).$$

Образуем двумерный вариационный ряд в виде таблицы, называемой корреляционной:

x / y	$\ldots x_k \ldots$	m_y
.	.	.
.	.	.
.	.	.
y_l	$\ldots m_{kl} \ldots$	m_{*l}
.	.	.
.	.	.
.	.	.
m_x	$\ldots m_{k*} \ldots$	n

В строке x в возрастающем порядке расположены варианты x_i, а в столбце y варианты y_i. На пересечении столбца x_k и строки y_l находится частота m_{kl}, означающая число точек выборки, равных точке (x_k, y_l).

В столбце m_y помещены частоты одномерного вариационного ряда y, в строке m_x – частоты x, полученные путем суммирования соответствующих частот m_{kl}. Наконец, n равно сумме частот любого из одномерных рядов x или y.

В качестве точечных оценок неизвестных начальных моментов первого и второго порядков генеральной совокупности берутся соответствующие выборочные моменты.

Точечные же оценки неизвестных пяти параметров получают с помощью формул, аналогичных формулам вычисления самих параметров, через генеральные начальные моменты. Таким образом, будем иметь:

$$\bar{x} = \frac{\sum x_k m_{k^*}}{n}$$ – оценка для μ_x,

$$\bar{y} = \frac{\sum y_l m_{*l}}{n}$$ – оценка для μ_y,

$$\overline{x^2} = \frac{\sum x_k^2 m_{k^*}}{n}$$ – оценка для $M(X^2)$,

$$\overline{y^2} = \frac{\sum y_l^2 m_{*l}}{n}$$ – оценка для $M(Y^2)$,

$$\overline{xy} = \frac{\sum\sum x_k y_l m_{kl}}{n}$$ – оценка для $M(X \cdot Y)$,

откуда

$$S_x^2 = \overline{x^2} - (\bar{x})^2$$ – оценка для σ_x^2,

$$S_y^2 = \overline{y^2} - (\bar{y})^2$$ – оценка для σ_y^2,

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{Sx \cdot Sy}$$ – оценка для ρ.

Оценки генеральных коэффициентов регрессии b_{yx} и b_{xy} получаются соответственно по формулам:

$$b_{yx} = r\frac{Sy}{Sx},\ b_{xy} = r\frac{Sx}{Sy},$$

откуда оценки уравнений регрессии имеют вид:

$$\overline{y/x} - \bar{y} = b_{yx}(x - \bar{x}),\ \ \overline{x/y} - \bar{x} = b_{xy}(y - \bar{y}).$$

При этом $\overline{y/x}$ и $\overline{x/y}$ – обозначения оценок для условных математических ожиданий *MY*/*X* и *MX*/*Y* генеральной совокупности.

Следует отметить, что вышеприведенные точечные оценки являются состоятельными, а $\bar{x}$ и $\bar{y}$ еще и несмещенными и эффективными; кроме того, в корреляционной модели распределение выборочных средних ($\bar{x}$, $\bar{y}$) не зависит от распределения (S_x^2, S_y^2, r), наконец, выборочный коэффициент корреляции r по абсолютной величине не превосходит единицы.

4.1.2. Приемы вычисления выборочных характеристик

Если объем выборки n невелик, то наблюдаемые точки располагают в порядке их регистрации (не образуя вариационного ряда) и обрабатывают по схеме:

x	y	x^2	y^2	xy
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_j	y_j	x_j^2	y_j^2	$x_j y_j$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$\sum x_j$	$\sum y_j$	$\sum x_j^2$	$\sum y_j^2$	$\sum x_j y_j$

последовательно заполняя столбцы таблицы результатами операций, указанных сверху. В последней строке вычисляются соответствующие суммы элементов столбцов. Далее используют формулы:

$$\bar{x} = \frac{\sum x_j}{n}, \bar{y} = \frac{\sum y_j}{n}; S_x^2 = \frac{\sum x_j^2}{n} - (\bar{x})^2, \mathrm{S}_\mathrm{y}^2 = \frac{\sum y_j^2}{n} - (\bar{y})^2,$$

$$r = \frac{\sum x_j y_j - \left(\sum x_j \cdot \sum y_j\right)/n}{\left\{\left[\sum x_j^2 - \left(\sum x_j\right)^2/n\right]\left[\sum y_j^2 - \left(\sum y_j\right)^2/n\right]\right\}^{\frac{1}{2}}},$$

$$b_{yx} = \frac{\sum x_j y_j - \left(\sum x_j \cdot \sum y_j\right)/n}{\sum x_j^2 - \left(\sum x_j\right)^2/n}, \; b_{xy} = \frac{\sum x_j y_j - \left(\sum x_j \cdot \sum y_j\right)/n}{\sum y_j^2 - \left(\sum y_j\right)^2/n}.$$

Если выборка многочисленна, то данные группируют путем построения двумерного ряда, корреляционная таблица для которого имеет вид:

x / y	$\ldots (a_k - b_k] \ldots$	m_y
.	.	.
.	.	.
.	.	.
$(c_l - d_l]$	$\ldots m_{kl} \ldots$	m_{*l}
.	.	.
.	.	.
.	.	,
m_x	$\ldots m_{k*} \ldots$	n

где m_{kl} - частота прямоугольника, в основании которого лежит полуинтервал $(a_k - b_k]$, а в высоте - $(c_l - d_l]$, то есть число точек выборки, попавших внутрь или на часть границы прямоугольника, задаваемой полуинтервалами; длины интервалов по x одинаковы и равны h_x, то же самое относится и к y (с одинаковой длиной h_y).

Для вычисления характеристик интервального ряда переходим к условному дискретному вариационному ряду с условными вариантами:

$$x_k' = \frac{x_k - x_0}{h_x}, \quad y_l' = \frac{y_l - y_0}{h_y},$$

где x_0, y_0 – рабочие средние, выбираются обычно равными центрам интервалов, лежащих в середине соответствующих одномерных рядов, x_k и y_l - центры интервалов; таким образом, условные варианты – целые числа, наименее уклоняющиеся от нуля по абсолютной величине.

Вычисления удобно производить по схеме, последовательно заполняя строки, лежащие ниже таблицы двумерного ряда условных вариантов (1÷4), и столбцы, лежащие справа от этой таблицы (1÷2):

	x' / y'	$\dots x_k \dots$	m_y	1 $y' m_y$	2 $(y')^2 m_y$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	y'_l	$\dots m_{kl} \dots$	m_{*l}	$y'_l m_{*l}$	$(y'_l)^2 m_{*l}$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	m_x	$\dots m_{k*} \dots$	n	$\sum y'_l m_{*l}$	$\sum (y'_l)^2 m_{*l}$
1	$x' m_x$	$\dots x'_k m_{k*} \dots$	$\sum x'_k m_{k*}$		
2	$(x')^2 m_x$	$\dots (x'_k)^2 m_{k*} \dots$	$\sum (x'_k) m_{k*}$		
3	$\sum y' m_{xy}$	$\dots \sum y'_l m_{kl} \dots$	$\sum y'_l m_{*l}$		
4	$x' \sum y' m_{xy}$	$\dots x'_k \sum y'_l m_{kl} \dots$	$\sum\sum x'_k y'_l m_{kl}$		

Заметим, что для контроля вычислений можно использовать равенство $\sum\sum y'_l m_{kl} = \sum y'_l m_{*l}$, то есть равенство чисел в конце строки 3 и столбца 1.

Далее используются формулы:

$$\bar{x}_{\text{гр.}} = \frac{\sum x_k' m_{k*}}{n} \cdot h_x + x_0, \quad \bar{y}_{\text{гр.}} = \frac{\sum y_l' m_{*l}}{n} \cdot h_y + y_0,$$

$$S^2_{x_{\text{гр.}}} = \left[\frac{\sum \left(x_k'\right)^2 m_{k*}}{n} - \left(\frac{\sum x_k' m_{k*}}{n} \right)^2 \right] h_x^2, \; S^2_{y_{\text{гр.}}} = \left[\frac{\sum \left(y_l'\right)^2 m_{*l}}{n} - \left(\frac{\sum y_l' m_{*l}}{n} \right)^2 \right] h_y^2,$$

$$r_{\text{гр.}} = \frac{n \sum\sum x_k' y_l' m_{kl} - \sum x_k' m_{k*} \cdot \sum y_l' \cdot m_{*l}}{\left\{ \left[n \sum \left(x_k'\right)^2 m_{k*} - \left(\sum x_k' m_{k*} \right)^2 \right] \cdot \left[n \sum \left(y_l'\right)^2 m_{*l} - \left(\sum y_l' \cdot m_{*l} \right)^2 \right] \right\}^{\frac{1}{2}}},$$

$$b_{yx_{\text{гр.}}} = \frac{n \sum\sum x_k' \cdot y_l' m_{kl} - \sum x_k' m_{k*} \cdot \sum y_l' m_{*l}}{n \sum \left(x_k'\right)^2 m_{k*} - \left(\sum x_k' m_{k*} \right)^2} \cdot \frac{h_y}{h_x},$$

$$b_{xy_{\text{гр.}}} = \frac{n\sum\sum x_k' \cdot y_l' m_{kl} - \sum x_k' m_{k*} \cdot \sum y_l' m_{*l}}{n\sum \left(y_l'\right)^2 m_{*l} - \left(\sum y_l' m_{*l}\right)^2} \cdot \frac{h_x}{h_y}.$$

При группировке вычисленные характеристики могут сильно отличаться от выборочных. Оценки по группированным данным центральных моментов второго порядка S_x^2 и S_y^2 можно улучшить поправками Шеппарда:

$$S_x^2 \cong S_{x_{\text{гр.}}}^2 - \frac{1}{12}h_x^2, \quad S_y^2 \cong S_{y_{\text{гр.}}}^2 - \frac{1}{12}h_y^2.$$

Эти поправки часто сглаживают ошибки, возникающие от группировки, если длина интервала (h) не превосходит восьмой части размаха соответствующего признака.

4.1.3. Проверка значимости параметров связи

В двумерной модели параметрами связи являются коэффициент корреляции ρ (или его квадрат, называемый коэффициентом детерминации) и коэффициенты регрессии β_{yx} и β_{xy}.

Назовем параметр связи значимо отличающимся от нуля, если гипотеза о равенстве параметра нулю отвергается с заданным уровнем значимости α. Если же эта гипотеза принимается, то параметр связи в генеральной совокупности называется незначимым.

Заметим, что в двумерной модели достаточно проверить значимость только коэффициента корреляции. Если коэффициент корреляции незначим, то X и Y считаются независимыми в генеральной совокупности.

Статистика r, вычисляемая для выборки из двумерной нормально распределенной совокупности с $\rho = 0$, связана со статистикой t, имеющей распределение Стьюдента с $\nu = n - 2$ степенями свободы, формулой

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}.$$

Зная границы для t, соответствующие обычным уровням значимости ($\alpha = 10\,\%$, $5\,\%$, $2\,\%$, $1\,\%$), можно получить границы для r, воспользовавшись этой формулой. Границы для r табулированы. Таким образом, для проверки гипотезы $H_0 : \rho = 0$ по данным α и $\nu = n - 2$ находим $r_{\text{табл.}}$ Если $|r_{\text{набл.}}| > r_{\text{табл.}}$, то гипотеза H_0 отвергается с вероятностью ошибки α, если же $|r_{\text{набл.}}| \le r_{\text{табл.}}$, то гипотеза не отвергается. При $\nu > 100$ для проверки $H_0 : \rho = 0$ можно пользоваться нормированным нормальным законом распределения статистики

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}$$

или статистики

$$r\sqrt{n - 1}.$$

Если наблюдаемая величина (или $r\sqrt{n-1}$) расположена в доверительном интервале $[-t_{1-\alpha}, t_{1-\alpha}]$, то гипотеза H_0 не отвергается, в противном случае H_0 отвергается с вероятностью ошибки α, $t_{1-\alpha} = \Phi^{-1}(1 - \alpha)$.

4.1.4. Интервальные оценки параметров связи

Для значимых параметров связи имеет смысл найти интервальные оценки.

При нахождении доверительного интервала для коэффициента корреляции ρ используют статистику, введенную Фишером:

$$Z_r = \frac{1}{2} ln \frac{1+r}{1-r},$$

которая при $n > 10$ распределена приблизительно нормально с генеральным средним $MZ_r \approx \frac{1}{2} ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$ и дисперсией $DZ_r \approx \frac{1}{n-3}$.

Тогда доверительный интервал, оценивающий MZ_r с надежностью $\gamma = 1 - \alpha$, имеет вид

$$Z_r - t_\gamma \sqrt{\frac{1}{n-3}} \le M Z_r \le Z_r + t_\gamma \sqrt{\frac{1}{n-3}},$$

где $t_\gamma = \Phi^{-1}(1-\alpha) = \Phi^{-1}(\gamma)$ находится по таблицам интеграла Лапласа для данного γ (или $\alpha = 1 - \gamma$). Для перехода от Z к ρ имеется таблица, после использования которой получаем интервальную оценку с надежностью γ вида

$$r_{min} \le \rho \le r_{max},$$

где r_{min} и r_{max} выбираются с учетом того, что Z_r – функция нечетная. При этом поправочным членом $\frac{\rho}{2(n-1)}$ у MZ_r пренебрегают.

Если коэффициент корреляции значим, то коэффициенты регрессии также значимо отличаются от нуля (с тем же уровнем α). Интервальные оценки для них получаются по формулам:

$$|t| \le St^{-1}(\alpha, \nu), t = (b_{yx} - \beta_{yx}) \frac{Sx\sqrt{n-2}}{Sy\sqrt{1-r^2}},$$

$$t = (b_{xy} - \beta_{xy}) \frac{Sy\sqrt{n-2}}{Sx\sqrt{1-r^2}},$$

где t имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы (доказывается в регрессионном анализе).

Переход от неравенств $|t| \le St^{-1}(\alpha, \nu)$ к интервальным оценкам для коэффициентов регрессии осуществляется с помощью тождественных алгебраических преобразований.

Для значимого коэффициента корреляции ρ некоторые авторы рекомендуют более предпочтительную оценку, чем r:

$$r\left(1 + \frac{1-r^2}{n-4}\right);$$

предпочтительной оценкой ρ^2 является выражение

$$\frac{(n-1)r^2-1}{n-2}.$$

Этими точечными оценками следует пользоваться при небольших объемах n выборки.

4.1.5. Задачи, решаемые с помощью статистики Фишера

Кроме нахождения интервальной оценки для ρ с помощью преобразования

$$Z_r = \frac{1}{2}\ln\frac{r+1}{r-1}$$

можно решить следующие задачи:

1 . Проверить, согласуется ли выборочный коэффициент корреляции r с предполагаемым значением генерального коэффициента корреляции ρ_0. Для этого, взяв уровень значимости α, проверяем, попадает ли абсолютная величина разности $\left|Z_r - Z_{\rho_0}\right|$ в интервал $\left[\Phi^{-1}_{(1-\alpha)}/\sqrt{n-3}\right]$. Если попадает, то гипотеза $H_0: \rho = \rho_0$ не отвергается; в противном случае отвергается с вероятностью ошибки α.

2. Проверить гипотезу однородности коэффициентов корреляции.

Пусть $r_1, r_2, \ldots, r_k$ - коэффициенты корреляции, полученные из k нормально распределенных совокупностей по независимым выборкам с объемами $n_1, n_2, \ldots, n_k$. Проверяется гипотеза

$H_0: \rho_1 = \rho_2 = \ldots \rho_k = \rho.$

Статистика

$$\sum_{i=1}^{k}\frac{(Z_i - Z_\rho)^2}{1/(n_i-3)}$$

имеет тогда распределение χ^2 с k степенями свободы. Если заменить Z_ρ на среднюю арифметическую

$$\overline{Z} = \frac{\sum Z_i}{\sum n_i},$$

то получим, что

$$\sum_{i=1}^{k}\frac{(Z_{ri} - \overline{Z}_r)^2}{1/(n_i-3)}$$

распределена по закону χ^2 с $\nu = k$-1 степенями свободы.

Если теперь для заданных α и $\nu = k$-1

$$\chi^2_{табл.} > \sum_{i=1}^{k}\frac{(Z_{ri} - \overline{Z}_r)^2}{1/(n_i-3)},$$

то гипотеза однородности отвергается с вероятностью ошибки α, в противном случае гипотеза не отвергается.

В случае принятия гипотезы однородности предпочтительной точечной оценкой ρ является значение r, полученное обратным преобразованием из $\overline{Z}$.

4.2. Регрессионный анализ

Регрессионный анализ – метод исследования зависимости результативного признака Y (случайной величины) от нескольких случайных величин $x_1, x_2, \ldots, x_k$, называемых факторами или регрессорами. При этом имеется в виду математическая, функциональная зависимость между числовой характеристикой случайной величины, как правило, математическим ожиданием, и соответствующими значениями факторных признаков. В качестве формы зависимости выбирается определенный класс функций, зависящих от неизвестных параметров. Задачей регрессивного анализа является оценка параметров по ряду независимых наблюдений и проверка гипотез относительно таких неизвестных параметров. Выбор класса функций осуществляется экспертным путем, исходя из соображений, касающихся изучаемой зависимости (экономических, социальных и др.). В случае неизвестной формы зависимости выбирают такую функцию, которая в некотором определенном смысле давала бы значения результативного признака, близкие к полученным реализациям случайной величины (случайных величин) при наблюдаемых или опытных значениях регрессоров. Кроме того, можно использовать графическое изображение наблюдаемых переменных, а также руководствоваться по возможности простой формой зависимости. Когда класс функций регрессии, отражающих зависимость математических ожиданий результативного признака от значений регрессоров – независимых аргументов выбран, то задачей регрессионного анализа становится оценка неизвестных параметров. Самым распространенным методом оценки параметров функции регрессии (регрессионной модели) является метод наименьших квадратов, дающий при определенных условиях несмещенные оценки с наименьшей дисперсией. Для интервального оценивания и проверки гипотез о параметрах регрессионной модели требуется нормальность распределения наблюдаемых случайных величин, характеризующих результативный признак.

4.2.1. Простая линейная регрессионная модель

Пусть имеется ряд наблюдений $(x_1, y_1),\ldots, (x_n, y_n)$. Регрессионная модель называется простой линейной, если она задается следующей системой уравнений

$$Y_1 = \beta_0+\beta_1 x_1+\varepsilon_1$$
$$Y_2 = \beta_0+\beta_1 x_2+\varepsilon_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_n = \beta_0+\beta_1 x_n+\varepsilon_n$$

В этих уравнениях $x_1, x_2,\ldots, x_n$ являются известными, точно измеренными, неслучайными величинами или фиксированными значениями одного и того же факторного признака x. Далее β_0 и β_1 - неизвестные параметры – коэффициенты регрессии линейной модели – неслучайные величины; ε_i – случайные независимые величины, называемые остатками, ошибками регрессии, по их величине можно судить о качестве выбранной линейной модели, при этом $M\varepsilon_i = 0$ и $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2$ – остаточная дисперсия – постоянная величина – неизвестный параметр модели. Следовательно, Y_i – независимые случайные величины с $MY_i = \beta_0+\beta_1 x_i$, $DY_i = D\varepsilon_i = \sigma^2$ и $M[(Y_i - MY_i)(Y_{i'} - MY_{i'})] = M(\varepsilon_i \cdot \varepsilon_{i'}) = 0$ для $i \neq i'$, $i, i' \in \{1, 2,\ldots, n\}$. Если (x_i, Y_i) взяты из двумерной генеральной совокупности, то

$$MY_i = M(Y/x = x_i) = \beta_0 + \beta_1 x_1$$

являются условными математическими ожиданиями, уравнение

$$\tilde{y} = M(Y/x) = \beta_0 + \beta_1 x$$

называется уравнением линейной регрессии.

4.2.2. Метод наименьших квадратов оценивания параметров модели

Метод заключается в минимизации суммы квадратов остатков (МНК):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \tilde{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2 .$$

Для получения МНК-оценок неизвестных параметров модели найдем частные производные функции $Q(\beta_0,\beta_1)$ по неизвестным β_0 и β_1 и приравняем их к нулю. После очевидных преобразований получим следующую систему нормальных уравнений, опуская для простоты записи нижние индексы:

$$n\beta_0 + \left(\sum x\right)\beta_1 = \sum y ,$$

$$\left(\sum x\right)\beta_0 + \left(\sum x^2\right)\beta_1 = \sum xy .$$

Обозначим решение этой системы через (b_0,b_1), тогда будем иметь

$$b_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{n\sum x^2 - \left(\sum x\right)^2},$$

$$b_1 = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2}.$$

Предполагается, что $n\Sigma x^2 - \left(\Sigma x\right)^2 \neq 0$, и в силу положительности $\sum_{i=1}^{n} \varepsilon_i^2$, за исключением единственного значения: $\varepsilon_i = ... = \varepsilon_n = 0$, точечные МНК оценки коэффициентов регрессии составляют минимум $Q(\beta_0,\beta_1)$.

Нетрудно доказать, что эти оценки совпадают с соответствующими оценками, получаемыми в корреляционном анализе

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = r\frac{S_y}{S_x} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left|\bar{x}\right|^2} = b_{yx} .$$

Точечной оценкой уравнения регрессии является

$$\hat{y} = \overline{y/x} = b_0 + b_1 x .$$

Таким образом, оценка ожидаемого значения или прогнозным значением результативного признака будет при $x = x_i$

$$\hat{y}_i = \bar{y}/x_i = b_0 + b_1 x_i ,$$

а наблюдаемым значением остатка будет

$$e_i = Y_i - \hat{Y}_i .$$

Найдем математические ожидания полученных оценок коэффициентов регрессии:

$$Mb_1 = M\frac{n\sum xy - \sum x\cdot\sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{n\sum x\cdot My - \sum x\cdot\sum My}{n\sum x^2 - \left(\sum x\right)^2} =$$

$$= \frac{n\sum x(\beta_0 + \beta_1 x) - \sum x\cdot\sum(\beta_0 + \beta_1 x)}{n\sum x^2 - \left(\sum x\right)^2} =$$

$$= \frac{n\beta_0\sum x + n\beta_1\sum x^2 - n\beta_0\sum x - \beta_1\sum x\cdot\sum x}{n\sum x^2 - \left(\sum x\right)^2} =$$

$$= \frac{n\sum x^2 - \left(\sum x\right)^2}{n\sum x^2 - \left(\sum x\right)^2}\beta_1 = \beta_1,$$

$$Mb_0 = M\left(\overline{Y} - b_1\bar{x}\right) = M\left(\beta_0 + \beta_1\bar{x} + \bar{\varepsilon} - b_1\bar{x}\right) = \beta_0.$$

Следовательно, МНК оценки коэффициентов регрессии являются несмещенными. Найдем дисперсии этих оценок, имея в виду независимость случайных величин Y_i:

$$Db_1 = D\frac{n\sum xy - \sum x\cdot\sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{D\sum\left(nx - \sum x\right)y}{\left[n\sum x^2 - \left(\sum x\right)^2\right]^2} =$$

$$= \frac{\sum D\left(nx - \sum x\right)y}{\left[n\sum x^2 - \left(\sum x\right)^2\right]^2} = \frac{\sum\left(nx - \sum x\right)^2\sigma^2}{\left[n\sum x^2 - \left(\sum x\right)^2\right]^2} =$$

$$= \frac{n\left[\sum x^2 - \left(\sum x\right)^2\right]\sigma^2}{\left[\sum x^2 - \left(\sum x\right)^2\right]^2} = \frac{\sigma^2}{\sum(x-\bar{x})^2},$$

$$Db_0 = D\frac{\sum y\sum x^2 - \sum x\sum xy}{n\sum x^2 - \left(\sum x\right)^2} = \frac{\sum\left[\sum x^2 - \left(\sum x\right)x\right]^2\sigma^2}{\left[n\sum x^2 - \left(\sum x\right)^2\right]^2} =$$

$$= \frac{\sum x^2\left[n\sum x^2 - \left(\sum x\right)^2\right]\sigma^2}{\left[n\sum x^2 - \left(\sum x\right)^2\right]^2} = \frac{\sum x^2\sigma^2}{n\sum x^2 - \left(\sum x\right)^2} = \frac{\sum x^2\sigma^2}{n\sum(x-\bar{x})^2}.$$

Докажем полезное для дальнейшего изложения тождество

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-y)^2 \ .$$

Имеем очевидное тождество

$$y - \bar{y} = \left(y - \hat{y}\right) + (\hat{y} - \bar{y}).$$

Возведем обе части этого тождества в квадрат и просуммируем их, получим

$$A = \sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2 + 2\sum (y - \hat{y})(\hat{y} - \bar{y}).$$

Осталось доказать, что удвоенная сумма равна нулю. Действительно

$$\sum (y - \bar{y})(\hat{y} - \bar{y}) = \sum (y - \hat{y})(b_0 + b_1 x - b_0 - b_1 \bar{x}) = b_1 \sum (y - \bar{y})(x - \bar{x}) =$$
$$= b_1 \left[\sum (y - \hat{y})x - \bar{x}\sum (y - \hat{y})\right].$$

Далее используем первое нормальное уравнение, согласно которому $\sum \hat{y}_i = \sum y$, поэтому $A = b_1\left[\sum (y - \hat{y})x\right] = b_1\left[\sum yx - \sum \hat{y}x\right] = b_1 \cdot 0 = 0$, согласно второму уравнению системы нормальных уравнений для МНК оценок коэффициентов регрессии.

Введем следующие обозначения и определения:

$Q_{общ.} = \sum (y - \bar{y})^2$ – общая сумма квадратов отклонения результативного признака Y от средней арифметической;

$Q_{регр.} = \sum (\hat{y} - \bar{y})^2$ – сумма квадратов отклонений, обусловленных регрессией;

$Q_{ост.} = \sum (y - \hat{y})^2$ – остаточная сумма квадратов, показывающая качество «подгонки» исследуемой регрессии к эмпирическим точкам. Таким образом, имеет место формула разложения суммы квадратов

$$Q_{общ.} = Q_{регр.} + Q_{ост.}.$$

Найдем математические ожидания сумм квадратов

$$MQ_{общ.} = M\sum (y - \bar{y})^2 = M\sum (\beta_0 + \beta_1 x + \varepsilon - \beta_0 - \beta_1 \bar{x} - \bar{\varepsilon})^2 =$$
$$= M\sum \left[\beta_1 (x - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})\right]^2 = \beta_1^2 \sum (x_i - \bar{x})^2 + \sum M(\varepsilon_i - \bar{\varepsilon})^2,$$

(математическое ожидание удвоенной суммы произведений равно нулю в силу условия $M\varepsilon = 0$). Далее

$$M(\varepsilon - \bar{\varepsilon})^2 = M(\varepsilon^2 - 2\varepsilon\bar{\varepsilon} + \bar{\varepsilon}^2) = \sigma^2 - \frac{2\sigma^2}{n} + \frac{n\sigma^2}{n^2} = \sigma^2 - \frac{\sigma^2}{n}.$$

(используются условия $M\varepsilon_i^2 = 0$, $M(\varepsilon_i \cdot \varepsilon_{i'}) = 0$
при $i \neq i'$, $i, i' \in \{1, 2, \ldots, n\}$). Следовательно,

$$MQ_{общ.} = \beta_1^2 \sum (x_i - \bar{x})^2 + (n-1)\sigma^2.$$

Найдем $MQ_{\text{регр.}}$

$$MQ_{регр.} = M\sum (\hat{y} - \bar{y})^2 = M\sum (b_0 + b_1 x - b_0 - b_1 \bar{x})^2 =$$
$$= \sum (x - \bar{x})^2 Mb_1^2 = \sum (x - \bar{x})^2 (Db_1 + \beta_1^2) = \sum (x - \bar{x})^2 \left(\frac{\sigma^2}{\sum (x - \bar{x})^2} + \beta_1^2\right) =$$
$$= \beta_1^2 \sum (x - \bar{x})^2 + \sigma^2.$$

Из тождества разложения суммы квадратов получим

$$MQ_{ост.} = MQ_{общ.} - MQ_{регр.} = \beta_1 \sum \left(x_i - \bar{x}\right)^2 + (n-1)\sigma^2 - \beta_1 \sum \left(x_i - \bar{x}\right)^2 - \sigma^2 =$$
$$= (n-2)\sigma^2.$$

Из этой формулы, очевидно, можно получить следующую несмещенную оценку остаточной дисперсии σ^2:

$$\hat{S}^2 = \frac{Q_{ост.}}{n-2} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n-2}.$$

Статистика $\hat{Y} = b_0 + b_1 x_0$ является несмещенной оценкой для условного математического ожидания $M(Y/x_0) = \beta_0 + \beta_1 x_0$, так как

$$M\,\hat{Y} = M\left(b_0 + b_1 x_0\right) = \beta_0 + \beta_1 x_0,$$

и ее дисперсия равна

$$D\,\hat{y} = \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum \left(x - \bar{x}\right)^2} \right).$$

Если предполагается сделать следующее $n + 1$ наблюдение y_{n+1} при заданном значении регрессора x_0, то несмещенной оценкой математического ожидания y_{n+1} является

$$\hat{y} = b_0 + b_1 x_0,$$

а дисперсия отклонения $y_{n+1} - \hat{y}$ увеличивается на величину $Dy_{n+1} = \sigma^2$, то есть

$$D\left(y_{n+1} - \hat{y} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)}{\sum \left(x - \bar{x}\right)^2} \right).$$

4.2.3. Проверка гипотез о параметрах регрессионной модели

Теперь мы должны предположить, что результативный признак распределен нормально с параметрами $MY_i = \mu_i = \beta_0 + \beta_1 x_i$ или $M(Y/x_i) = \beta_0 + \beta_1 x_i$ и постоянной остаточной дисперсией $DY_i = D(Y/x_i) = \sigma^2$, следовательно, остатки ε_i являются независимыми случайными величинами, распределенными по одному и тому же нормальному закону с параметрами $(0, \sigma^2)$, причем $M(\varepsilon_i, \varepsilon_{i'}) = 0$
для $i \neq i'$, $i, i \in \{1, 2, ..., n\}$.

Проверим гипотезу о значимости уравнения регрессии, которая в нашем случае простейшей линейной модели сводится к проверке значимости (существенного отличия от нуля) коэффициента регрессии β_1 при меняющемся регрессоре x. На уровне значимости α требуется проверить гипотезу

$$H_0 : \beta_1 = 0$$

против альтернативы

$$H_1 : \beta_1 \neq 0$$

по данным наблюдениям (X_i, Y_i), $i \in \{1, 2, ..., n\}$.

В качестве критерия рассмотрим критерий Фишера – Снедекора, статистика которого имеет вид

$$F = \frac{\mathrm{Q}_{регр.}/1}{\mathrm{Q}_{ост.}/(n-2)}.$$

Критическая область задается неравенством

$$F > F_i(\alpha/2,\ \nu_1 = 1,\ \nu_2 = n - 2).$$

Обоснованием этого критерия является независимость статистик $Q_{\text{регр.}}$ и $Q_{\text{ост.}}/(n-2)$, так как они являются несмещенными ($Q_{\text{регр.}}$ при $\beta_1 = 0$, а $Q_{\text{ост.}}/(n\text{-}2)$ независимо от величины β_1) оценками одной и той же дисперсии σ^2. Согласно теореме Кохрана статистики $Q_{\text{общ.}}$, $Q_{\text{регр.}}$, $Q_{\text{ост.}}$ удовлетворяют ее условиям, то есть являются при гипотезе H_0: $\beta_1 = 0$ распределенными (после деления на σ^2) по Пирсону (хи-квадрат) и независимыми, так как число степеней свободы у $Q_{\text{общ.}}$, равное n-1, составляет сумму чисел степеней свободы у слагаемых $\nu_1 = 1$ и $\nu_2 = n$ -2. Следует отметить, что часто берут одностороннюю критическую область по аналогии с проверкой гипотезы о равенстве нулю коэффициента детерминации.

4.2.4. Интервальные оценки

Для построения доверительных интервалов коэффициентов регрессии, условного математического ожидания и интервала предсказания нового наблюдения используется известное положение, по которому статистика, являющаяся несмещенной оценкой параметра θ, линейно зависящая от наблюдаемых случайных величин, распределенных нормально и независимых, также нормально распределена. Так как дисперсия этой оценки неизвестна, то она заменяется несмещенной оценкой (σ^2 заменяется на $\hat{S}^2$) и используется для построения доверительных интервалов (и проверки гипотез) распределение Стьюдента с числом степеней свободы $\nu_2 = n - 2$.

Приведем формулы интервальных оценок с доверительной вероятностью $\gamma = 1\text{-}\alpha$.

1. Интервальная оценка для β_0

$$b_0 - \Delta\beta_0 \le \beta_0 \le b_0 + \Delta\beta_0,\ \Delta\beta_0 = St^{-1}(\alpha, n-2)\sqrt{\hat{S}^2 \frac{\sum x^2}{n\sum\left(x - \bar{x}\right)^2}}.$$

2. Интервальная оценка для β_1

$$b_1 + \Delta\beta_1 \le \beta_1 \le b_1 + \Delta\beta_1,\ \Delta\beta_1 = St^{-1}(\alpha, n-2)\sqrt{\hat{S}^2 \frac{1}{\sum\left(x - \bar{x}\right)^2}}.$$

3. Интервальная оценка для $M(Y/x_0)$

$$\hat{Y} - \Delta\mu \le M(Y / x_0) \le \hat{y} + \Delta\mu,\ \hat{y} = b_0 + b_1 x_0,\ \Delta\mu = St^{-1}(\alpha, n-2)\sqrt{\hat{S}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}\right)}.$$

4. Интервал предсказания y_{n+1}

$$\hat{Y} - \Delta y_{n+1} \le y_{n+1} \le \hat{Y} + y_{n+1},$$

$$\hat{Y} = b_0 + b_1 x_0,\ \Delta y_{n+1} = St^{-1}(\alpha, n-2)\sqrt{\hat{S}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x-\bar{x})^2}\right)}.$$

4.3. Пояснения, примеры и решения задач

1. На основании выборочных данных (табл. 4.3.1) провести корреляционный анализ зависимости между основными фондами (X) и объемом валовой продукции (Y) предприятий мебельной промышленности.

Таблица 4.3.1

Валовая продукция в млн. руб. (Y)	Основные фонды предприятий (X) мебельной промышленности, в млн. руб.					Всего предприятий, m_y
	0 – 1,4	1,4 – 2,8	2,8 – 4,2	4,2 – 5,6	5,6 – 7,0	
0,0 – 0,8	1	3				4
0,8 –1,6		4	6			10
1,6 – 2,4			10	8		18
2,4 – 3,2			5	9	1	15
3,2 – 4,0					3	3
Всего предприятий, m_x	1	7	21	17	4	50

Решение. Первоначально полезно получить эмпирическое подтверждение того, что выборка взята из генеральной совокупности (*X*, *Y*), имеющей двумерный нормальный закон распределения. С этой целью (согласно п. 1.10) построим поле корреляции (рис. 4.3.1). Так как на рисунке "облако" точек имеет вытянутую форму, причем точки группируются около некоторой прямой, то предположение нормальности не отвергается. Будем считать, что совокупность (*X*, *Y*) подчиняется двумерному нормальному закону распределения.

Определим оценки параметров этого распределения.

Для удобства вычисления перейдем к условным вариантам, которые определим по формулам:

$$x_j' = \frac{x_j - x_0}{h_x} \quad \text{и} \quad y_\nu' = \frac{y_\nu - y_0}{h_y},$$

где x_j и y_ν – центры соответствующих интервалов

$$j = 1,2, \ldots к;$$

$$\nu = 1, 2, \ldots, l$$

h_x и h_y – ширина интервалов;

x_0 и y_0 – центры серединных интервалов.

В нашем примере:

$$x_0 = 3,5;\ y_0 = 2;$$

$$h_x = 1,4;\ h_y = 0,8.$$

Для расчетов удобно использовать вспомогательную таблицу (табл.4.3.2).

При заполнении вспомогательной таблицы частные суммы вычисляются по формулам:

$$m_x = m_{j*} = \sum_{\nu=1}^{l} m_{i\nu};\quad m_y = m_{*\nu} = \sum_{j=1}^{k} m_{j\nu}$$

Предпоследняя строка таблицы $\left(\sum_{\nu=1}^{l} y_\nu' m_{j\nu}\right)$ получена путем суммирования произведений y_ν' на соответствующую частоту $m_{j\nu}$.

$(-2) \cdot 1 = -2;\ (-2) \cdot 3 + (-1) \cdot 4 = -10$ и т.д.

В конце строки записывают сумму всех этих величин:

$$\sum_{j=1}^{k} \sum_{\nu=1}^{l} y_\nu' m_{j\nu} = -2 - 10 - 1 + 9 + 7 = 3.$$

Контрольные суммы: $\sum_j m_{j*} = \sum_{\nu=1}^{l} m_{*\nu} = 50$

И $\sum_{\nu=1}^{l} y_\nu' m_{*\nu} = \sum_{j=1}^{k} \sum_{\nu=1}^{l} y_\nu' m_{j\nu} = 3.$

Таблица 4.3.2

x_j' / y_ν'	-2	-1	0	1	2	$m_{*\nu}$	$y_\nu' m_{*\nu}$	$(y_\nu')^2 m_{*\nu}$
-2	1	3				4	-8	16
-1		4	6			10	-10	10
0			10	8		18	0	0
1			5	9	1	15	15	15
2					3	3	6	12
m_{j*}	1	7	21	17	4	50	$\sum_\nu y_\nu' m_{*\nu} = 3$	$\sum_\nu (y_\nu')^2 m_{*\nu} = 53$
$x_j' m_{j*}$	-2	-7	0	17	8	$\sum_j x_j' m_{j*} = 16$		
$(x_j')^2 m_{j*}$	4		0	17	16	$\sum_j (x_j')^2 m_{j*} = 44$		
$\sum_\nu y_\nu' m_{j\nu}$	-2	-10	-1	9	7	$\sum_j \sum_\nu y_\nu' m_{j\nu} = 3$		
$x_j' \sum_\nu y_\nu' m_{j\nu}$	4	10	0	9	14	$\sum_j x_j' \sum_\nu y_\nu' m_{j\nu} = 37$		

Используя свойства средней арифметической и дисперсии, можно записать:

$$\bar{x} = x_0 + \frac{h_x}{n}\sum_j x_j\, m_{j*} = x_0 + h_x \overline{x'}; \quad \bar{y} = y_0 + h_y \overline{y'};$$

$$S_x^2 = h_x^2 \left[\frac{\sum_j (x_j')^2 m_{j*}}{n} - \left(\frac{\sum_j x_j' m_{j*}}{n} \right)^2 \right] = h_x^2 \cdot S_{x'}^2;$$

$$S_y^2 = h_y^2 \left[\frac{\sum_\nu (y_\nu')^2 m_{*\nu}}{n} - \left(\frac{\sum_\nu y_\nu' m_{*\nu}}{n} \right)^2 \right] = h_y^2 \cdot S_{y'}^2;$$

$$\overline{x'y'} = \frac{1}{n}\sum_{j=1}^{k}\sum_{\nu=1}^{l} x_j' y_\nu' m_{j\nu};$$

тогда

$$r = \frac{\overline{x'y'} - \overline{x'}\,\overline{y'}}{S_{x'}\, S_{y'}}.$$

В нашем примере по данным таблицы 4.3.2 имеем:

$$\bar{x} = 3{,}5 + \frac{16}{50} \cdot 1{,}4 = 3{,}948;\ \bar{y} = 2 + \frac{3}{50} \cdot 0{,}8 = 2{,}048$$

Откуда $\overline{x'} = 0{,}32;\ \overline{y'} = 0{,}06;$

$$S_x^2 = 1{,}4^2 \cdot \left[\frac{44}{50} - 0{,}32^2 \right] = 1{,}524;\ S_x = 1{,}235;$$

$$S_y^2 = 0{,}8^2 \cdot \left[\frac{53}{50} - 0{,}06^2 \right] = 0{,}676;\ S_y = 0{,}822.$$

Тогда $S_{x'}^2 = 0{,}7776,\ S_y^2 = 1{,}0564.$

Учитывая, что $\overline{X'Y'} = \frac{1}{50} \cdot 37 = 0{,}74$, получим

$$\mathrm{r} = \frac{0{,}74 - 0{,}32 \cdot 0{,}06}{\sqrt{0{,}7776 \cdot 1{,}0564}} = 0{,}795.$$

Проверим значимость коэффициента корреляции при $\alpha = 0{,}05$, т.е. проверим гипотезу $H_0 : \rho = 0$ на уровне значимости $\alpha = 0{,}05$. Для проверки гипотезы воспользуемся таблицей Фишера-Иейтса (таблица 5 Приложения) и найдем при $\alpha = 0{,}05$ и числе степеней свободы $\nu = n - 2 = 48$ табличное значение $r_{\text{кр.}} = 0{,}273$. Так как наблюдаемое значение $|\mathrm{r}|$ больше табличного, т.е. $|\mathrm{r}| > r_{\text{кр.}}$, то гипотеза отвергается. Следовательно, величины X и Y зависимые.

Теперь определим с надежностью $\gamma = 0{,}95$ интервальную оценку для ρ. Воспользуемся Z – преобразованием Фишера и для $r = 0{,}7953 \approx 0{,}8$ по таблице 6 Приложения найдем $Z_r = 1{,}0986$. Предварительно определим доверительный интервал для MZ_r, который имеет вид

$$Z_r - t_\gamma \sqrt{\frac{1}{n-3}} \le MZ_r \le Z_2 + t_\gamma \sqrt{\frac{1}{n-3}},$$

где $t_\gamma = 1{,}96$ находим по таблице 1 Приложения из условия

$$\Phi(\mathrm{t}_\gamma) = 0{,}95.$$

Тогда $1{,}0986 - 1{,}96 \sqrt{\frac{1}{47}} \le MZ_r \le 1{,}0986 + 1{,}96 \sqrt{\frac{1}{47}}$ и

$0{,}8127 \le MZ \le 1{,}3845.$

Снова воспользуемся таблицей Z – преобразования (таблица 6 Приложения) для перехода от Z к ρ и получим интервальную оценку для ρ с надежностью γ: $0{,}67 \le \rho \le 0{,}88$.

Найдем точечные оценки b_{yx} и b_{xy} генеральных коэффициентов регрессии β_{yx} и β_{xy}:

$$b_{yx} = r\frac{S_y}{S_x} = 0{,}795 \cdot \sqrt{\frac{0{,}676}{1{,}524}} = 0{,}530 \text{ и}$$

$$b_{xy} = r\frac{S_x}{S_y} = 0{,}795 \cdot \sqrt{\frac{1{,}524}{0{,}676}} = 1{,}194.$$

Оценка уравнения регрессии Y по X имеет вид

$$\hat{y} = \overline{y/x} = 2{,}048 + 0{,}5297(x - 3{,}948);$$

и окончательно $\hat{y} = 0{,}530\,x - 0{,}043$.

Оценка уравнения регрессии X по Y имеет вид

$$\hat{x} = 3{,}948 + 1{,}194(y - 2{,}048)$$

и окончательно

$$\hat{x} = 1{,}502 + 1{,}194y.$$

Линии регрессии y по x и x по y изображены на рис. 4.3.1.

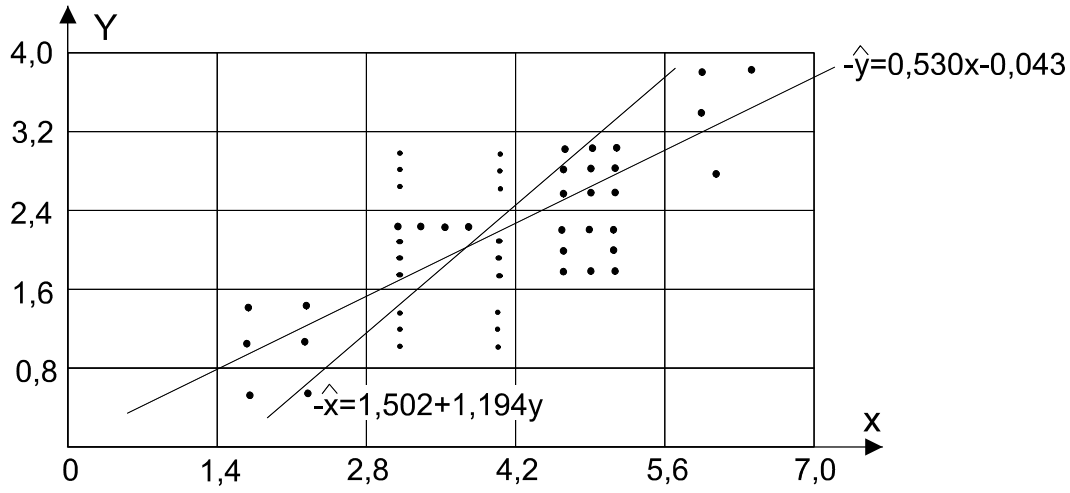

Рис. 4.3.1. Поле корреляции и линии регрессии

Определим интервальные оценки коэффициентов регрессии β_{yx} и β_{xy} с надежностью $\gamma = 0{,}95$.

Согласно п.4.1.4 интервальные оценки определяются по формулам:

$$b_{yx} - t\frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}} \le \beta_{yx} \le b_{yx} + t\frac{S_y\sqrt{1-r^2}}{S_x\sqrt{n-2}}$$

и

$$b_{xy} - t\frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}} \le \beta_{xy} \le b_{xy} + t\frac{S_x\sqrt{1-r^2}}{S_y\sqrt{n-2}},$$

где t определяют по таблице распределения Стьюдента при $\alpha = 1 - \gamma$ и $\nu = n - 2$. По таблице 2 Приложения для $\alpha = 1 - 0{,}95 = 0{,}05$ и $\gamma = 50 - 2 = 48$ найдем $t = 2{,}021$.

Тогда

$$0{,}530 - 2{,}021 \cdot \frac{0{,}822 \cdot 0{,}606}{1{,}235 \cdot 6{,}928} \le \beta_{yx} \le 0{,}530 + 2{,}021 \cdot \frac{0{,}822 \cdot 0{,}606}{1{,}235 \cdot 6{,}928},$$

$$0{,}412 \le \beta_{yx} \le 0{,}648.$$

Аналогично рассуждая, получим

$$0{,}928 \le \beta_{xy} \le 1{,}460.$$

2. На основании данных о производительности труда (X) и себестоимости продукции (Y), полученных с $n = 12$ однотипных предприятий за месяц (табл.4.3.3), проверить при $\alpha = 0{,}05$ значимость коэффициента корреляции ρ и с надежностью $\gamma = 0{,}9$ найти интервальную оценку для ρ. Предполагается, что совместное распределение признаков (X,Y) подчиняется двумерному закону распределения.

Таблица 4.3.3

№№ п/п	X	Y	X^2	Y^2	XY
1	112	64	12544	4096	7168
2	74	41	5476	1681	3034
3	100	47	10000	2209	4700
4	111	61	12321	3721	6771
5	94	51	8836	2601	4794
6	92	45	8464	2025	4140
7	100	59	11664	3481	6372
8	104	57	10816	3249	5928
9	8	43	7056	1849	3612
10	87	46	7569	2126	4002
11	64	28	4096	784	1792
12	70	34	4900	1156	2380
Итого:	1100	576	103742	28968	54694

Точечная оценка коэффициента корреляции находится по формуле

$$r = \frac{\overline{xy} - \overline{x}\,\overline{y}}{S_x S_y},$$

где

$$\overline{X} = \frac{1}{n}\sum_j x_j = \frac{1100}{12} = 91{,}67; \quad \overline{Y} = \frac{576}{12} = 48{,}00;$$

$$S_x^2 = \frac{1}{n}\sum_j x_j^2 - (\overline{x})^2 = \frac{103742}{12} - (91{,}67)^2 = 241{,}78; \quad S_x = 15{,}55;$$

$$S_y^2 = \frac{1}{n}\sum_j y_j^2 - (\overline{y})^2 = \frac{28968}{12} - (48{,}00)^2 = 110{,}00; \quad S_y = 10{,}49;$$

$$\overline{xy} = \frac{1}{n}\sum_j x_j y_j = \frac{54694}{12} = 4557,83.$$

Тогда

$$r = \frac{4557,83 - 91,67 \cdot 48,00}{15,55 \cdot 10,49} = 0,97.$$

Проверим значимость коэффициента корреляции при $\alpha = 0,05$, т.е. гипотезу $H_0 : \rho = 0$. Воспользуемся критерием Стьюдента и определим

$$t_{\text{набл.}} = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2} = 12,62.$$

По таблице распределения Стьюдента (таблица 2 Приложения) для $\alpha = 0,05$ и числа степеней свободы $\nu = n - 2 = 10$ найдем $t_{\text{кр.}} = 2,228$. Так как $|t_{\text{набл.}}| > t_{\text{кр.}}$, то гипотеза $H_0 : \rho = 0$ отвергается с вероятностью ошибки $\alpha = 0,05$ и можно сделать вывод, что показатели производительности труда (X) и себестоимости продукции (Y) зависимы.

Определим интервальную оценку коэффициента корреляции ρ с надежностью $\gamma = 0,9$.

По таблице 6 Приложения для $r = 0,97$ найдем $Z_r = 2,0923$, а по таблице 1 Приложения из условия $\Phi(t_\gamma) = 0,9$ найдем $t_\gamma = 1,64$. Тогда интервальная оценка для MZ_r определяется как

$$2,0923 - 1,64\sqrt{\frac{1}{9}} \le MZ_r \le 2,0923 + 1,64\sqrt{\frac{1}{9}},$$

$$1,545 \le MZ_r \le 2,639.$$

Снова воспользуемся таблицей 6 Приложения и перейдем от Z к ρ и найдем интервальную оценку для ρ с надежностью $\gamma = 0,9$:

$$0,91 \le \rho \le 0,99.$$

3. По данным задачи 2 п. 4.3 на уровне значимости $\alpha = 0,05$ проверить гипотезу $H_0 : \rho = 0,8$ против H_1: $\rho > 0,8$. Критическая область является правосторонней и задается неравенством

$$Z_r > Z_{0,8} + \Delta Z = 1,0986 + \frac{\Phi^{-1}(1 - 2\alpha)}{\sqrt{n - 3}} = 1,0986 + \frac{1,64}{3} = 1,6463.$$

Так как $Z_{r\ набл.} = Z_{0,97} = 2,0923$ попадает в критическую область, гипотеза $H_0 : \rho = 0,8$ отвергается с вероятностью ошибки, равной 0,05.

4. В линейном регрессионном анализе рассматриваются модели, в которых в качестве преобразованного результативного признака может выступать некоторая известная функция исходного результативного признака, то же самое относится и к факторным признакам, однако коэффициенты регрессии входят в модель линейно, например,

$$\ln Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i,$$

примером нелинейной регрессии в указанном смысле может служить модель

$$Y_i = \beta_0^2 + x^{\beta_0} + \beta_1 x^2.$$

5. Проведем регрессионный анализ простой линейной модели на примере следующих выборочных данных

x	10	25	10	31	15	50	70	75	90	100
y	4,5	10	4,3	12	6,2	24,4	30,5	31	40	51
$\hat{y}$	3,4	10,6	3,4	13,5	5,8	22,5	33,1	34,4	41,6	46,4

где x – расстояние от пункта производства до пункта продажи товара, y – затраты на перевозку. Предполагается нормальность распределения y при каждом x. Найдем нужные характеристики выборки.

$$\sum x = 476;\ \sum x^2 = 33136;$$
$$\sum y = 213{,}9;\ \sum xy = 15183, n = 10;\ \sum (x - \bar{x})^2 = 10478{,}4.$$

Точечные оценки коэффициентов регрессии:

$$b_1 = \frac{10 \cdot 15183 - 476 \cdot 213{,}9}{10 \cdot 33136 - 476^2} = 0{,}4773018$$

$$b_0 = 21{,}39 - 0{,}4773018 \cdot 47{,}6 = -1{,}3295693$$

Точечная оценка уравнения регрессии:

$$\hat{y} = -1{,}33 + 0{,}477x.$$

Найдем модельные (прогнозные, расчетные) значения результативного признака – оценки условного математического ожидания и запишем их в третью строку таблицы исходных данных. Вычислим сумму квадратов остатков и регрессии с помощью второй и третьей строк в таблице исходных данных, а также общую сумму для контроля вычислений:

$Q_{ост.} = 46{,}24;\ Q_{регр.} = 2388{,}55;\ Q_{общ.} = 2433{,}47.$

Видимо, имеется ошибка округления, приблизительно равная единице, что не повлияет на дальнейшие результаты.

Точечная несмещенная оценка остаточной дисперсии

$$\hat{S}^2 = 46{,}24 / 8 = 4{,}49 = 4{,}5.$$

На уровне значимости $\alpha = 0{,}1$ критическая область для двустороннего критерия имеет вид

$$F > F_{\text{i}}(0{,}05;1;8) = 5{,}32.$$

Найдем $F_{набл.} = \dfrac{Q_{регр.}/1}{\hat{S}^2} = \dfrac{2388}{4{,}5} = 530{,}67.$

Так как $F_{\text{набл.}}$ попало в критическую область, гипотеза $H_1 : \beta_1 = 0$ отвергается с вероятностью ошибки 0,1. Можно сделать вывод о том, что модель (уравнение) регрессии значима; то же самое касается и коэффициента β_1 регрессии, он значимо отличается от нуля.

Найдем интервальные оценки коэффициентов регрессии с надежностью $\gamma = 0{,}95$. Из таблицы распределения Стьюдента находим $St^{-1}(\alpha = 0{,}05, \nu = 8) = 2{,}306$, тогда получаются следующие результаты:

– интервальная оценка коэффициента регрессии β_0 при

$$\Delta\beta_0 = 2{,}306\sqrt{4{,}5\cdot\frac{33136}{10\cdot 10478{,}4}} = 2{,}75$$

и $b_0 = -1{,}33$:

$$-4{,}08 \le \beta_0 \le 1{,}42,$$

– интервальная оценка коэффициента регрессии β_1 при

$$\Delta\beta_1 = 2{,}306\sqrt{4{,}5\cdot\frac{1}{10478{,}4}} = 0{,}048$$

и $b_1 = 0{,}477$:

$$0{,}431 \le \beta_1 \le 0{,}523.$$

По поводу полученных интервальных оценок можно утверждать, что интервал для β_0 включает нулевое значение генерального коэффициента β_0 и, следовательно, является незначимым. Доверительный интервал для коэффициента регрессии β_1, наоборот, значим, так как он не содержит нулевого значения.

Найдем интервальную оценку для условного математического ожидания при условии $x_0 = 60$. Получим

$$\Delta\mu = 2{,}306\cdot\sqrt{4{,}5\left(\frac{1}{10}+\frac{(60-47{,}6)^2}{10418{,}4}\right)} = 1{,}7;\quad \hat{y} = -1{,}33 + 0{,}477\cdot 60 = 27{,}3,$$

следовательно,

$$25{,}6 \le M(Y/x = 60) \le 29{,}0.$$

Интервал предсказания для будущего наблюдения при значении регрессора $x = 60$ имеет при точности $\Delta y_{n+1} = 2{,}306\sqrt{4{,}5\left(1{,}1+\frac{(60-47{,}6)^2}{10418{,}4}\right)} = 5{,}2$ следующий вид

$$27{,}3 - 5{,}2 \le y_{n+1} \le 27{,}3 + 5{,}2$$

или окончательно

$$22{,}1 \le y_{n+1} \le 32{,}5.$$

4.4. Упражнения

Во всех задачах предполагается условие нормальности распределения.

1. Провести корреляционный анализ по данным задачи 6 пункта 1.6, взяв $\alpha = 0{,}1$.

2. Провести корреляционный анализ по данным задачи 7 пункта 1.6, взяв $\alpha = 0{,}05$.

3. Провести корреляционный анализ по данным задачи 10 пункта 1.6, с уровнем значимости $\alpha = 0{,}01$.

4. Провести корреляционный анализ по данным задачи 11 пункта 1.6, с $\alpha = 0{,}02$.

5. По результатам выборки объема $n = 25$ получены характеристики: $S_x^2 = 31{,}16$, $S_y^2 = 9{,}21$, $b_{yx} = -0{,}38$. Проверить значимость и найти интервальную оценку коэффициента ρ_{xy} с уровнем значимости $\alpha = 0.02$.

6. По результатам четырех независимых выборок получены данные: $r_1 = 0{,}07$, $n_1 = 40$; $r_2 = 0{,}85$, $n_2 = 20$; $r_3 = 0{,}6$, $n_3 = 45$; $r_4 = 0{,}65$, $n_4 = 25$. На уровне значимости $\alpha = 0{,}05$ проверить гипотезу $H_0 : \rho = \rho_1 = \rho_2 = \rho_3 = \rho_4$.

7. По условиям задачи 9 найти наиболее предпочтительную оценку неизвестного коэффициента корреляции.

8. На основе выборки объема $n = 40$ найден выборочный коэффициент корреляции $r = 0{,}7$. На уровне значимости $\alpha = 0{,}03$ проверить гипотезу $\rho = 0{,}5$.

9. Предполагая возможность изучения зависимости цены холодильника (условн. ден. единицы) от его возраста (годы) по простой линейной модели, на основе следующих данных:

x	0	1	1	2	2	3	4	5
y	15	12	14	10	8	10	8	5

,

найти точечные оценки параметров модели и построить графическое изображение поля корреляции и линии регрессии.

10. По данным предыдущей задачи проверить гипотезу о значимости уравнения регрессии и коэффициентов регрессии на уровне значимости $\alpha = 0{,}1$.

11. По данным предыдущей задачи найти интервальные оценки коэффициентов регрессии с надежностью $\gamma = 0,\ 98$.

12. Найти интервальную оценку условного математического ожидания цены холодильника и интервал предсказания (прогноза) его цены для возраста $x_0 = 3$ года и $x_0 = 6$ лет. Сравнить точность интервальных оценок ($\alpha = 0{,}95$) по указанным выше данным.

Приложение 1

Нормальный закон распределения

Целые и десятые доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0,0797	0,0876	0,0955	0,1034	0,1113	0,1192	0,1271	0,1350	0,1428	0,1507
0,2	0,1585	0,1663	0,1741	0,1819	0,1897	0,1974	0,2051	0,2128	0,2205	0,2282
0,3	0,2358	0,2434	0,2510	0,2586	0,2661	0,2737	0,2812	0,2886	0,2960	0,3035
0,4	0,3108	0,3182	0,3255	0,3328	0,3401	0,3473	0,3545	0,3616	0,3688	0,3759
0,5	0,3829	0,3899	0,3969	0,4039	0,4108	0,4177	0,4245	0,4313	0,4381	0,4448
0,6	0,4515	0,4581	0,4647	0,4713	0,4778	0,4843	0,4907	0,4971	0,5035	0,5098
0,7	0,5161	0,5223	0,5285	0,5346	0,5407	0,5467	0,5527	0,5587	0,5646	0,5705
0,8	0,5763	0,5821	0,5878	0,5935	0,5991	0,6047	0,6102	0,6157	0,6211	0,6265
0,9	0,6319	0,6372	0,6424	0,6476	0,6528	0,6579	0,6629	0,6679	0,6729	0,6778
1,0	0,6827	0,6875	0,6923	0,6970	0,7017	0,7063	0,7109	0,7154	0,7199	0,7243
1,1	0,7287	0,7330	0,7373	0,7415	0,7457	0,7499	0,7540	0,7580	0,7620	0,7660
1,2	0,7699	0,7737	0,7750	0,7813	0,7850	0,7887	0,7923	0,7959	0,7994	0,8029
1,3	0,8064	0,8098	0,8132	0,8165	0,8198	0,8230	0,8262	0,8293	0,8324	0,8355
1,4	0,8385	0,8415	0,8444	0,8473	0,8501	0,8529	0,8557	0,8584	0,8611	0,8638
1,5	0,8664	0,8690	0,8715	0,8740	0,8764	0,8789	0,8812	0,8836	0,8859	0,8882
1,6	0,8904	0,8926	0,8948	0,8969	0,8990	0,9011	0,9031	0,9051	0,9070	0,9090
1,7	0,9109	0,9127	0,9146	0,9164	0,9181	0,9199	0,9216	0,9233	0,9249	0,9265
1,8	0,9281	0,9297	0,9312	0,9327	0,9342	0,9357	0,9371	0,9385	0,9399	0,9412
1,9	0,9426	0,9439	0,9451	0,9464	0,9476	0,9488	0,9500	0,9512	0,9523	0,9534
2,0	0,9545	0,9556	0,9566	0,9576	0,9586	0,9596	0,9606	0,9616	0,9625	0,9634
2,1	0,9643	0,9651	0,9660	0,9668	0,9676	0,9684	0,9692	0,9700	0,9707	0,9715
2,2	0,9722	0,9729	0,9736	0,9743	0,9749	0,9756	0,9762	0,9768	0,9774	0,9780
2,3	0,9786	0,9791	0,9797	0,9802	0,9807	0,9812	0,9817	0,9822	0,9827	0,9832
2,4	0,9836	0,9841	0,9845	0,9849	0,9853	0,9857	0,9861	0,9865	0,9869	0,9872
2,5	0,9876	0,9879	0,9883	0,9886	0,9899	0,9892	0,9895	0,9898	0,9901	0,9904
2,6	0,9907	0,9910	0,9912	0,9915	0,9917	0,9920	0,9922	0,9924	0,9926	0,9928
2,7	0,9931	0,9933	0,9935	0,9937	0,9939	0,9940	0,9942	0,9944	0,9946	0,9947
2,8	0,9949	0,9951	0,9952	0,9953	0,9955	0,9956	0,9958	0,9959	0,9960	0,9961
2,9	0,9963	0,9964	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972
3,0	0,9973	0,9974	0,9975	0,9976	0,9976	0,9977	0,9978	0,9979	0,9979	0,9980
3,1	0,9981	0,9981	0,9982	0,9983	0,9983	0,9984	0,9994	0,9985	0,9985	0,9986
3,5	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997	0,9997
3,6	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998	0,9998
3,7	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,0	0,999936	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
4,5	0,999994	-	-	-	-	-	-	-	-	-
5,0	0,99999994	-	-	-	-	-	-	-	-	-

Приложение 2

Распределение Стьюдента (t - распределение)

V	Вероятность $\alpha = St(t) = P(\|T\| > t_{табл})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0.158	0,325	0,510	0,727	1,000	1,376	1,963	3.078	6,314	12.706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0.137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,043	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,953	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2.896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,327	0,542	0,700	0,879	1.093	1,372	1,812	2,228	2,764	3,169	4,583
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0.395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3.012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0.393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4.073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	1,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,833

V	Вероятность $\alpha = St(t) = P(\|T\| > t_{табл})$												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,868	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0.256	0.390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,402	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0.256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3.690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0.255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Приложение 3

Распределение Пирсона (χ^2 - распределение)

Значения $\chi^2_{табл.}$ для вероятностей $P(\chi^2 > \chi^2_{табл.})$

V	Вероятность										
	0,999	0,995	0,99	0,98	0,975	0,95	0.90	0,80	0,75	0,70	0,50
1	0,05157	0,04393	0,03157	0,03628	0,03982	0,00393	0,0158	0,0642	0,102	0,148	0,455
2	0,00200	0,0100	0,0201	0,0404	0,0506	0,103	0,211	0,446	0,575	0,713	1,386
3	0,0243	0,0717	0,115	0,185	0,216	0,352	0,584	1,005	1,213	1,424	2,366
4	0,0908	0,207	0,297	0,429	0,484	0,711	1,064	1,649	1,923	2,195	3,357
5	0,210	0,412	0,554	0,752	0,831	1,145	1,610	2,343	2,675	3,000	4,351
6	0,381	0,676	0,872	1,134	1,237	1,635	2,204	3,070	3,455	3,828	5,348
7	0,598	0,989	1,239	1,564	1,690	2,167	2,833	3,822	4,255	4,671	6,346
8	0,857	1,344	1,646	2,032	2,180	2,733	3,490	4,594	5,071	5,527	7,344
9	1,152	1,735	2,088	2,532	2,700	3,325	4,168	5,380	5,899	6,393	8,343
10	1,479	2,156	2,558	3,059	3,247	3,240	4,865	6,179	6,737	7,267	9,342
11	1,834	2,603	3,053	3,609	3,816	4,575	5,578	6,989	7,584	8,148	10,341
12	2,214	3,074	3,571	4,178	4,404	5,226	6,304	7,807	8,438	9,034	11,340
13	2,617	3,565	4,107	4,765	5,009	5,892	7,042	8,634	9,299	9,926	12,340
14	3,041	4.075	4,660	5,368	5,629	6,571	7,790	9,467	10,165	10,821	13,339
15	3,483	4,601	5,229	5,985	6,262	7,261	8,547	10,307	11,036	11,721	14,339
16	3.942	5,142	5,812	6,614	6,908	7,962	9,312	11,152	11,912	12,624	15,338
17	4,416	5,697	6,408	7,255	7,564	8,672	10,085	12,002	12,892	13,531	16,338

V	Вероятность										
	0,999	0,995	0,99	0,98	0,975	0,95	0.90	0,80	0,75	0,70	0,50
18	4,905	6,265	7,015	7,906	8,231	9,390	10,865	12,857	13,675	14,440	17,338
19	5,407	6,844	7,633	8,567	8,907	10,117	11,651	13,716	14,562	15,352	18,338
20	5,921	7,434	8,260	9,237	9,591	10,871	12,443	14,578	15,452	16,266	19,337
21	6,447	8,034	8,897	9,915	10,283	11.591	13,240	15,445	16,344	17,182	20,337
22	6,983	8,643	9,542	10,600	10,982	12,338	14,041	16,314	17,240	18,101	21,337
23	7,529	9,260	10,196	11,293	11,688	13,091	14,848	17,187	18,137	19,021	22,337
24	8,035	9,886	10,856	11,992	12,401	13,848	15,659	18,062	19,037	19,943	23,337
25	8,649	10,520	11.524	12,697	13,120	14,611	16,173	18,940	19,939	20,887	24,337
26	9,222	11,160	12,198	13,409	13,844	15,379	17,292	19,820	20,843	21,792	25,336
27	9,803	11,808	12,879	14,125	14,573	16,151	18,114	20,703	21,749	22,719	26.136
28	10,391	12,461	13,565	14,847	15,308	16,928	18,937	21,588	22,657	23,617	27,336
29	10,986	13,121	14,256	15,574	16,047	17,708	19,768	22,475	23,567	24,577	28,336
30	11 588	13,787	14,953	16,306	16,791	18,493	20,599	23,364	24,478	25,508	29,336

Вероятность										V
0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001	
1,074	1,323	1,642	2,706	3,841	5,024	5,412	6,635	7,879	10,827	1
2,408	2,773	3,219	4,605	5.991	7,378	7,824	9,210	10,597	13,815	2
3,665	4,108	4,642	6,251	7,815	9,348	9,837	11,345	12,838	16,268	3
4,878	5,385	5,989	7,779	9,488	11,143	11,668	13,277	14,860	18,465	4
6,064	6,626	7,289	9,236	11,070	12,839	13,388	15,086	16,750	20,517	5
7,231	7,841	8,558	10,645	12,592	14,449	15,033	16,812	18,548	22,457	6
8,383	9,037	9,803	12,017	14,067	16,013	16,622	18,475	20,278	24,322	7
9,524	10,219	11,030	13,362	15,507	17,535	18,168	20,090	21,955	26,125	8
10,656	11,389	12,242	14,684	16,919	19,023	19,679	21,666	23,589	27,877	9
11,781	12,549	13,412	15,987	18,307	20,483	21,161	23,209 .	25,188	29,588	10
12,899	13,701	14,631	17,275	19,675	21,920	22,618	24,725	26,757	31,264	11
14,011	14,845	15,812	18,549	21,026	23,337	24,054	26,217	28,300	32,909	12
15,119	15,984	16,985	19,812	22,362	24,736	25,472	27,688	29,819	34,528	13
16,222	17,117	18,151	21,064	23,685	26,119	26,873	29,141	31,319	36,123	14
17,322	18,245	19,311	22,307	24,996	27,488	28,259	30,578	32,801	37,697	15
18,418	19,369	20.465	23,542	26,296	28,845	29,633	32,000	34,267	39,252	16
19,511	20,489	21,615	24,769	27,587	30,191	30,995	33,409	35,718	40,790	17
20,601	21,605	22,760	25,989	28,869	31,526	32,346	34,805	37,156	42,312	18
21,689	22,718	23,900	27,204	30,144	32,852	33,687	36,191	38,582	43,820	19
22,775	23,828	25,038	28,412	31,410	34,170	35,020	37,566	39,997	45,315	20
23,858	24,935	26,171	29,615	32,671	35,479	36,343	38,932	41,401	46,797	21
24,939	26,039	27,301	30,813	33,924	36,781	37,659	40,289	42,796	48,268	22

Вероятность										V
0,30	0,25	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,001	
26,018	27,141	28,429	32,007	35,172	38,076	38,968	41,638	44,181	49,728	23
27,096	28,241	29,553	33,196	36.415	39,364	40,270	42,980	45,558	51,170	24
28,172	29,339	30,675	34,382	37,652	40,046	41,566	44,314	46,928	52,620	25
29,246	30,434	31,795	35,563	38,885	41,923	42,856	45,642	48,290	54,052	26
30,319	31,528	32,912	36,741	40,113	43,194	44,140	46,963	49,645	55,476	27
31,391	32,620	34,027	37,916	41,337	44,461	45,419	48,278	50,993	56,893	28
32,461	33,711	35,139	39,087	42,557	45,722	46,693	49,588	52,336	58,302	29
33,530	34,800	36,250	40,256	43.773	46,979	47.962	50,892	53,672	59,703	30

Приложение 4

Распределение Фишера-Снедекора (F-распределение).

Значения $F_{табл}$, удовлетворяющие условию $P(F > F_{табл})$. Первое значение соответствует вероятности 0,05; второе - вероятности 0,01 и третье-вероятности 0,001; v_1 - число степеней свободы числителя; v_2 - знаменателя.

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞	t
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	253,3	12,71
	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366	63,66
	406523	500016	536700	562527	576449	585953	598149	610598	623432	636535	636,2
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19.45	19,50	4,30
	98,49	99,01	00,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50	9,92
	998,46	999,00	999,20	999,20	999,20	999,20	999,40	999,60	999,40	999,40	31,00
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53	3,18
	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12	5,84
	67,47	148,51	141,10	137,10	134,60	132,90	130,60	128,30	125,90	123,50	12,94
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63	2,78
	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46	4,60
	74,13	61,24	56,18	53,43	51,71	50,52	49,00	47,41	45,77	44,05	8,61
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36	2,57
	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02	4,03
	47,04	36,61	33,20	31,09	20,75	28,83	27,64	26,42	25,14	23,78	6,86
6	5,99	5,14	4,76	4.53	4,39	4,28	4,15	4,00	3,84	3,67	2,45
	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88	3,71
	35,51	26,99	23,70	21,90	20,81	20,03	19,03	17,99	16,89	15,75	5,96

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞	*t*
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23	2,36
	12,25	9,55	8,45	7,85	7,46	7,19	6.84	6,47	6,07	5,65	3,50
	29,22	21,69	18,77	17,19	16,21	15,52	14,63	13,71	12,73	11,70	5,40
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,99	2,31
	11,26	8,65	7,59	7,10	6,63	6,37	6,03	5,67	5,28	4,86	3,36
	25,42	18,49	. 15,83	14,39	13,49	12,86	12,04	11,19	10,30	9,35	5,04
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3.07	2,90	2,71	2,26
	10,56	8,02	6,99	6,42	6,06	5.SO	5,47	5,11	4,73	4,31	3,25
	22,86	16,39	13,90	12,56	11,71	11,13	10,37	9,57	8,72	7,81	4,78
10	4.96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54	2,23
	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91	3,17
	21,04	14,91	12,55	11,28	10,48	9,92	9,20	8,45	7,64	6,77	4,59
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40	2,20
	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60	3,11
	19,69	13,81	11,56	10,35	9,58	9,05	8,35	7,62	6,85	6,00	4,49
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30	2,18
	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36	3,06
	18,64	12,98	10,81	9,63	8,89	8,38	7,71	7,00	6,25	5,42	4,32
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21	2,16
	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16	3,01
	17,81	12,31	10,21	9,07	8,35	7,86	7,21	6,52	5,78	4,97	4,12
14	4,60	3,74	3,34	3.11	2,96	2,85	2,70	2,53	2,35	2.13	2,14
	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00	2,98

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞	*t*
	17,14	11,78	9,73	8,62	7,92	7,44	6,80	6,13	5,41	4,60	4,14
15	4,45	3.68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07	2,13
	8.68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87	2,95
	16,59	11,34	9,34	8,25	7,57	7,09	6,47	5,81	5,10	4,31	4,07
16	4,41	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01	2,12
	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3.18	2,75	2,92
	16,12	10,97	9,01	7,94	7,27	6,80	6,20	5,55	4,85	4,06	4,02
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2.19	1,96	2,11
	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65	2,90
	15,72	10,66	. 8,73	7,68	7,02	6,56	5,96	5,32	4,63	3,85	3,96
18	4.41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92	2,10
	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,01	2,57	2,88
	15,38	10,39	8,49	7,46	6,81	6,35	5,76	5,13	4,45	3,67	3,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88	2,09
	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49	2,86
	15,08	10,16	8,28	7,26	6,61	6,18	5,59	4,97	4,29	3,52	3,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84	2,09
	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42	2,84
	14,82	9,95	8,10	7,10	6,46	6,02	5,44	4,82	4,15	3,38	3,85
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,82	2,08
	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36	2,83
	14,62	9,77	7,94	6,95	6,32	5,88	5,31	4,70	4,03	3,26	3,82
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2.23	2,03	1,78	2,07

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞	*t*
	7,94	5,72	4,82	4,31	3,99	3,75	3,45	3,12	2.75	2,30	2,82
	14,38	9,61	7,80	6,81	6,19	5,76	5,19	4,58	3,92	3,15	3,79
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2.20	2,00	1,76	2,07
	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2.70	2,26	2,81
	14,19	9,46	7,67	6,70	6,08	5.56	5,09	4,48	3,82	3,05	3,77
24	4,26	3,40	3,01	2,78	2,62	2,51	2.36	2,18	1,98	1,73	2,06
	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21	2,80
	14,03	9,34	7,55	6,59	5,98	5,55	4,99	4,39	3,74	2,97	3,75
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1.71	2,06
	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17	2,79
	13,88	9,22	7,45	6,49	5,89	5,46	4,91	4,31	3,66	2,89	3,72
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69	2,06
	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13	2,78
	13,74	9,12	7,36	6,41	5,80	5,38	4,83	4,24	3,59	2,82	3,71
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67	2,05
	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10	2,77
	13,61	9,02	7,27	6,33	5,73	5,31	4,76	4,17	3,52	2,76	3,69
28	4,19	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65	2,05
	7,64	5,54	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06	2,76
	13,50	8,93	7,18	6,25	5,66	5,24	4,69	4,11	3,46	2,70	3,67
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64	2,05
	7,60	5,42	4,54	4,04	3.73	3,50	3,20	2,87	2,49	2,03	2,76
	13,39	8,85	7,12	6,19	5,59	5,18	4,65	4,05	3,41	2,64	3,66

v_2 \ v_1	1	2	3	4	5	6	8	12	24	∞	*t*
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62	2,04
	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01	2,75
	13,29	8,77	7,05	6,12	5,53	5,12	4,58	4,00	3,36	2,59	3,64
60	4,00	3,15	2,76	2,52	2,37	2,15	2,10	1,92	1,70	1,39	2,00
	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2.12	1,60	2,66
	11,97	7,76	6,17	5,31	4,76	4,37	3,87	3,31	2,76	1,90	3,36
∞	3,84	2.99	2,60	2,37	2,21	2,09	1,94	1.75	1,52	1.03	1,96
	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,04	2,58
	10,83	6,91	5,42	4,62	4,10	3,74	3,27	2,74	2,13	1,05	3,29

Приложение 5

Таблица Фишера-Иейтса

Зачения $r_{кр}$, найденные для уровня значимости α и чисел степеней свободы $v = n - 2$ в случае парной корреляции и $v = n - l - 2$, где l - число исключенных величин в случае частной корреляции

V	Двусторонние границы				V	Двусторонние границы			
	0,05	0,02	0,01	0,001		0,05	0,02	0,01	0,001
1	0,997	1,000	1,000	1,000	16	0,468	0,543	0,590	0,708
2	0,950	0,980	0,990	0,999	17	0,456	0,529	0,575	0,693
3	0,878	0,934	0,959	0,991	18	0,444	0,516	0,561	0,679
4	0,811	0,882	0,917	0,974	19	0,433	0,503	0,549	0,665
5	0,754	0,833	0,875	0,951	20	0,423	0,492	0,537	0,652
6	0,707	0,789	0,834	0,925	25	0,381	0,445	0,487	0,597
7	0,666	0,750	0,798	0,898	30	0,349	0,409	0,449	0,554
8	0,632	0,715	0,765	0,872	35	0,325	0,381	0,418	0,519
9	0,602	0,685	0,735	0,847	40	0,304	0,358	0,393	0,490
10	0,576	0,658	0,708	0,823	45	0,288	0,338	0,372	0,465
11	0,553	0,634	0,684	0,801	50	0,273	0,322	0,354	0,443
12	0,532	0,612	0,661	0,780	60	0,250	0,295	0,325	0,408
13	0,514	0,592	0,641	0,760	70	0,232	0,274	0,302	0,380
14	0,497	0,574	0,623	0,742	80	0,217	0,257	0,283	0,338
15	0,482	0,558	0,606	0,725	90	0,205	0,242	0,267	0,338
					100	0,195	0,230	0,254	0,321
V	0,025	0,01	0,005	0,0005	V	0,025	0,01	0,005	0,0005
	Односторонние границы					Односторонние границы			

Приложение 6

Таблица Z – преобразования Фишера

$$Z = \frac{1}{2}\{in(1+r) - in(1-r)\}$$

r	0	1	2		4	5	6	7	8	9
0,0	0,0000	0,0101	0,0200	0,0300	0,0400	0,0501	0,0601	0,0601	0,0701	0,0902
1	0,1003	0,1104	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
3	0,3095	0,3205	0,3316	0,3428	0,3541	0,3654	0,3767	0,3884	0,4001	0,4118
4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
5	0,5493	0,5627	0,5764	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,6777
6	0,6932	0,7089	0,7250	0,7414	0,7582	0,7753	0,7928	0,8107	0,8291	0,8480
7	0,8673	0,8872	0,9077	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
9	1,4722	1,5275	1,5275	1,6584	1,7381	1,8318	1,9459	2,0923	2,2976	2,6467
0,99	2,6466	2,6996	2,7587	2,8257	2,9031	2,9945	3,1063	3,2504	3,4534	3,8002

Приложение 7

Значение плотности $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ значение для нормированного нормального закона распределения $f(-t) = f(t)$

Целые и десятые доли t	Сотые доли t									
	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0.3982	0,3980	0,3977	0,3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0,5	3525	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3-56	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2631	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0.2203
1,1	2179	2155	2131	2107	2083	2059	2036	3012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804

Целые и десятые доли *t*	Сотые доли *t*									
	0	1	2	3	4	5	6	7	8	9
1,8	0790	0775	0762	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,478	0,0568	0,0459	0,0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2.8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0,0001	0.0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001

Приложение 8

Значение функции Пуассона $P\left(X=m\right)=\frac{\lambda^m}{m!}e^{-\lambda}$

m \ λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679
1	0,0905	0,1637	0.2223	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679
2	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1216	0,1438	0,1547	0,1839
3	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613
4	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153
5	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0007	0,0012	0,0020	0,0031
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001

m \ λ	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
0	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0001
1	0,2707	0,1494	0,0733	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005
2	0.2707	0,2240	0,1465	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023
3	0,1805	0,2240	0,1954	0,1404	0,892	0,0521	-,0286	0,0150	0,0076
4	0,0902	0,1681	0,1954	0,1755	0,1339	0,0912	0,0572	0,0337	0,0189
5	0,0361	0,1008	0,1563	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0120	0,0504	0,1042	0,1462	0,1606	0,1490	0,1221	0,0911	0,0631
7	0,0034	0,0216	0,0595	0,1045	0,1377	0,1490	0,1396	0,1171	0,0901
8	0,0009	0,0081	0,0298	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9	0,0002	0,0027	0,0132	0,363	0,0689	0,1014	0,1241	0,1318	0,1251
10	0,0000	0,0008	0,0053	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11	0,0000	0,0002	0,0019	0,0082	0,0225	0,0452	0,0722	0,970	0,1137

m \ λ	2,0	3,0	4,0	5,0	6,0	7,0	8,0	9,0	10,0
12	0,0000	0,0001	0,0006	0,0034	0,0113	0,0264	0,0481	0,728	0,0948
13	0,0000	0,0000	0,0002	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14	0,0000	0,0000	0,0001	0,0005	0,0022	0,0071	0,0169	0,0324	0,0521
15	0,0000	0,0000	0,0000	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16	0,0000	0,0000	0,0000	0,0000	0,0003	0,0015	0,0045	0,0109	0,0217
17	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0021	0,0058	0,0128
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0029	0,0071
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0037

Приложение 9

G – распределение.

Пяти- и однопроцентное пределы для отношения G наибольшей выборочной дисперсии к сумме L выборочных дисперсий, полученных из L независимых выборок объемом n. Первое значение соответствует уровню значимости $\alpha = 0{,}05$, а второе - $\alpha = 0{,}01$

L \ n-1	1	2	3	4	5	6	7	8	9	10	16	36	144	∞
	0,998	0,975	0,939	0,906	0,877	0,853	0,838	0,816	0,801	0,788	0,734	0,660	0,518	0,500
2	0,999	0,995	0,979	0,959	0,937	0,917	0,809	0,882	0,867	0,854	0,795	0,700	0,606	0,500
	0,967	0,871	0,798	0,746	0,707	0,677	0,653	0,633	0,617	0,603	0,547	0.475	0,403	0,333
3	0,993	0,942	0,883	0,834	0,903	0,761	0,734	0,711	0,691	0,674	0,606	0,515	0,423	0,333
	0,906	0,768	0,684	0,629	0,590	0,560	0,537	0,518	0,502	0,488	0,437	0,372	0,309	0,250
4	0,968	0,864	0,781	0,721	0,676	0,641	0,613	0,590	0,570	0,554	0,488	0,406	0,325	0,250
	0,841	0,684	0.598	0,544	0,507	0,478	0,456	0,439	0,424	0,412	0,365	0,307	0,251	0,200
5	0,928	0,789	0,696	0,633	0,588	0,553	0,526	0,504	0,485	0,470	0,409	0,335	0,254	0,200
	0,781	0,616	0,532	0,480	0,445	0,418	0,398	0,382	0,368	0,357	0,314	0,261	0,212	0,167
6	0,883	0,722	0,626	0.564	0,520	0,487	0,461	0,440	0,423	0,408	0,353	0,286	0,223	0,167
	0,727	0,561	0,480	0,431	0,397	0,373	0,354	0,338	0,326	0,315	0,276	0,228	0,183	0,143
7	0,838	0,664	0,569	0,508	0.466	0,435	0,411	0,391	0,375	0,362	0,311	0,249	0,193	0,143
	0,680	0,516	0,438	0,391	0,360	0,336	0,319	0,304	0.293	0,283	0,246	0,202	0,162	0,15
8	0,795		0,521	0,463	0,423	0,393	0,370	0,352	0,337	0,325	0,278	0,221	0.170	0,125
	0,639	0,478	0,403	0,358	0,329	0,307	0,290	0,277	0,266	0,257	0.223	0,182	0,145	0,111
9	0,754	0,573	0,481	0,425	0,387	0,359	0,338	0,321	0,307	0,295	0,251	0,199	0,152	0,111

Приложение 10

Значение функции e^{-x}

x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}
0,00	1,0000	0,40	0,6703	0,80	0,4493	1,20	0,3012	1,60	0,2019
01	9900	41	6637	81	4449	21	2982	61	1999
02	9802	42	6570	82	4404	22	2952	62	1979
03	9704	43	6505	83	4360	23	2923	63	1959
04	9608	44	6440	84	4317	24	2894	64	1940
05	9512	45	6376	85	4274	25	2865	65	1920
06	9418	46	6313	86	4232	26	2837	66	1901
07	9324	47	6250	87	4190	27	2808	67	1882
08	9231	48	6188	88	4148	28	2780	68	1864
09	9139	49	6126	89	4107	29	2753	69	1845
10	9048	50	6065	90	4066	30	2725	70	1827
II	8958	51	6005	91	4025	31	2698	71	1809
12	8869	52	5945	92	3985	32	2671	72	1791
13	8781	53	5886	93	3916	33	2645	73	1773
14	8694	54	5827	94	3906	34	2618	74	1755
15	8607	55	5769	95	3867	35	2592	75	1738
16	8521	56	5712	96	3829	36	2567	76	1720
17	8437	57	5655	97	3791	37	2541	77	1703
18	8353	58	5599	98	3753	38	2516	78	1686
19	8270	59	5543	99	3716	39	2491	79	1670
20	8187	60	5488	1,00	3679	40	2466	80	1653

x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}	x	e^{-x}
21	8106	61	5434	01	3642	41	2441	81	1637
22	8025	62	5379	02	3606	42	2417	82	1620
23	7945	63	5326	03	3570	43	2393	83	1504
24	7866	64	5273	04	3535	44	2369	84	1588
25	7788	65	5220	05	3499	45	2346	85	1572
26	7711	66	5169	06	3465	46	2322	86	1557
27	7634	67	5117	07	3430	47	2299	87	1541
28	7558	68	5066	08	3396	48	2254	88	1526
29	7483	69	5616	09	3362	49	2254	89	I5II
30	7408	70	4966	10	3329	50	2231	90	1496
31	7334	71	4916	II	3296	51	2209	91	1481
32	7261	72	4868	12	3263	52	2187	92	1466
33	7189	73	4819	13	3230	53	2165	93	1451
34	7118	74	4771	14	3198	54	2144	94	1437
35	7047	75	4724	15	3166	55	2122	95	1424
36	6977	76	4677	16	3135	56	2101	96	1409
37	6907	77	4630	17	3104	57	2080	97	1395
38	6839	78	4584	18	3073	58	2060	98	1381
39	6771	79	4538	19	3042	59	2039	99	1367
								2,00	1353

Литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. –1022 с.
2. Боровков А.А. Теория вероятностей. – М.: Наука, 1986.
3. Виленкин Н.Я. Популярная комбинаторика. – М.: Наука, 1975.
4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1998 – 400 с.
5. Дубров А.М., Мхитарян В.С., Трошин Л.И., Маслен-ченко Н.В. Многомерные статистические методы. – М.: "Финансы и статистика", 1991.
6. Калинина В.Н., Панкин В.Ф. Математическая статистика. – М.: Высшая школа, 1994 – 336 с.
7. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1991 – 400 с.
8. Прохоров Ю.В., Разанов Ю.А. Теория вероятностей. – М.: Наука, 1987.

Оглавление