Министерство образования и науки Российской Федерации
Балтийский государственный технический университет «Военмех»

В. Л. ФАЙНШМИДТ

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Учебное пособие

Санкт-Петербург
2017

УДК 519.22(075.8)
Ф 17

Файншмидт, В.Л.

Ф 17 Элементы математической статистики: учебное пособие / В.Л.Файншмидт; Балт. гос. техн. ун-т. — СПб., 2017. — 64 с.

ISBN 978-5-9069206-84-3

В пособии, соответствующем программе курса высшей математики, изложены основные сведения по статистике, необходимые будущим инженерам.

Предназначено для студентов инженерных специальностей.

УДК 519.22(075.8)

Рецензент канд-т техн. наук проф. *Э.И. Ульянов*

Утверждено
редакционно-издательским
советом университета

ISBN 978-5-906920-84-3

© В.Л. Файншмидт, 2017
© БГТУ, 2017

1. ЗАДАЧИ СТАТИСТИКИ

Какой бы реальный процесс мы ни рассматривали невозможно указать все влияющие на него факторы и получить точные значения описывающих его величин. Поэтому результаты наблюдений всегда описываются приближенно, т.е. содержат элементы случайности. В соответствии с этим при обработке наблюдений приходится находить (обычно приближенно) параметры и законы распределения случайных величин. Это означает, что обработка невозможна без применения теории вероятностей. Такого рода обработкой занимается математическая статистика. Спрос на статистику в различных областях человеческой деятельности особенно возрос в последние годы. Это объясняется тем, что появились быстродействующие компьютеры, позволяющие за короткое время обрабатывать большие числовые массивы.

Приведем примеры простейших задач, связанных с обработкой наблюдений:

1. Требуется по результатам наблюдений найти функцию и плотность распределения случайной величины.

2. Имеется n значений случайной величины, полученных в результате испытаний. Требуется проверить гипотезу о том, что эта величина описывается заданной функцией распределения $F(x)$.

3. Оценить по результатам серии измерений их среднее значение и величину разброса.

4. Произведена серия наблюдений над парой величин. Выяснить, зависимы ли эти величины, и найти характер этой зависимости.

В этом пособии мы будем рассматривать математическую постановку и решение задач статистики и не будем говорить о том, как используют современные компьютеры для обработки результатов наблюдений. Эти вопросы подробно рассмотрены в большом числе книг. Например, в книге [1] подробно изложена методика таких расчетов с помощью пакетов STATGRAFICS и MATHCAD.

Прежде всего укажем на следующее: в соответствии со сказанным ранее обычно считают, что наблюдения производятся над случайной величиной или системой случайных величин. Множество всех

возможных значений величины или системы величин называют *генеральной совокупностью.* В результате наблюдений получают некоторый набор значений случайной величины или системы случайных величин. Этот набор значений называют *выборкой.* Если набор содержит n элементов, то n называют *объемом выборки.*

Естественно, что наблюдения должны производиться так, чтобы они соответствовали основным характеристикам генеральной совокупности. Для этого элементы выборки стараются получать максимально возможным случайным образом. В таком случае выборка оказывается *репрезентативной (представительной).* Получить такой совершенно случайный набор элементов, к сожалению, удается не всегда.

Отметим сразу же такое существенное обстоятельство: *до проведения испытаний каждый элемент выборки рассматривается как случайная величина (или система случайных величин), имеющая тот же закон распределения, что и генеральная совокупность.* Полученные в результате испытания конкретные значения элементов называют *реализациями* этих случайных величин (систем случайных величин).

Отметим одно обстоятельство, связанное с терминологией. Для решения задач будем, как правило, использовать функции от выборочных данных. Такие функции принято называть *статистиками* (так же как и весь этот раздел науки).

Будем в дальнейшем использовать такие обозначения:

$F(x)$ – функция распределения непрерывной случайной величины,

$f(x)$ – плотность вероятности непрерывной случайной величины,

$M(X)$ – математическое ожидание случайной величины X,

$D(X)$ – дисперсия случайной величины X,

$\sigma(X) = \sqrt{D(X)}$ – среднее квадратическое отклонение случайной величины X,

$m_k = \int\limits_{-\infty}^{+\infty} x^k f(x)dx$ – начальный момент порядка k непрерывной случайной величины с плотностью распределения $f(x)$,

$m_k = \sum_{i=1}^{n} x_i^k p(x_i)$ – начальный момент порядка k дискретной случайной величины.

2. НЕКОТОРЫЕ ПОЛЕЗНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

Напомним читателю некоторые законы распределения, которые изучаются в курсе теории вероятностей и могут понадобиться нам в статистике.

Биномиальное распределение. Пусть X – число появлений события A при n однотипных испытаниях. Тогда X – случайная величина, которая может принимать значения $0, 1, ..., n$. Если p – вероятность появления A при одном испытании, то

$$P(X = k) = C_n^k p^k (1-p)^{n-k} \quad (k = 0, 1, ..., n),$$

причем $M(X) = np$, $D(X) = np(1-p)$.

Отношение числа появлений события A к числу всех испытаний называют частотой события A, т.е. частотой A является случайная величина $p_* = \dfrac{X}{n}$. Очевидно, что

$$P\left(p_* = \frac{k}{n}\right) = C_n^k p^k (1-p)^{n-k} \quad (k = 0, 1, ..., n).$$

Кроме того, $M(p_*) = p$ и $D(p_*) = \dfrac{p(1-p)}{n}$.

Распределение Пуассона. Случайная величина X подчиняется закону распределения Пуассона, если

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, 2, 3, ...),$$

где λ – положительная константа.

Отметим, что $M(X) = \lambda$ и $D(X) = \lambda$.

Равномерное распределение. Случайная величина равномерно распределена на промежутке $[a, b]$, если ее плотность вероятности

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a, b); \\ 0, & x \notin [a, b]. \end{cases}$$

Зная плотность, можно найти функцию распределения этой случайной величины:

$$F(x) = \begin{cases} 0, & x \in (-\infty, a], \\ \dfrac{x-a}{b-a}, & x \in [a, b], \\ 1, & x \in [a, +\infty). \end{cases}$$

Нетрудно показать, что $M(X) = \dfrac{b+a}{2}$ и $D(X) = \dfrac{(b-a)^2}{12}$.

Экспоненциальное распределение. Случайная величина X имеет экспоненциальное распределение, если плотность вероятностей ее

$$f(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x > 0,\ \lambda > 0. \end{cases}$$

Отсюда следует, что

$$F(x) = \begin{cases} 0, & x \leqslant 0, \\ 1 - e^{-\lambda x}, & x \geqslant 0 \end{cases}$$

и $M(X) = \dfrac{1}{\lambda}$, $D(X) = \dfrac{1}{\lambda^2}$.

Заметим, что распределения Пуассона и экспоненциальное оказываются весьма полезными при описании систем массового обслуживания.

Нормальное распределение. Нормальное распределение характеризуется плотностью распределения

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

и, в соответствии с этим, функцией распределения

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}}\, dt.$$

Элементарной заменой функция распределения приводится к виду

$$F(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{t^2}{2}}\, dt = \Phi\left(\frac{x-m}{\sigma}\right),$$

где $\Phi(x)$ – функция Лапласа, т.е.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$

Как известно, $M(X) = m$ и $D(X) = \sigma^2$.

Существуют таблицы функции Лапласа.

Отметим важное для статистики обстоятельство. Имеются таблицы, в которых по значению $\Phi(x)$ указывается соответствующее значение аргумента x. Вообще, значение аргумента x, найденное по значению функции распределения $F(x)$, носит название *квантиль*. Поэтому таблицы называют таблицами квантилей нормального распределения.

Напомним, что нормальное распределение является особенно часто используемым в силу справедливости теоремы Ляпунова. Эта теорема говорит о том, что сумма большого числа независимых случайных величин может приближенно описываться нормальным распределением.

Кроме указанных нам понадобятся еще несколько распределений, которые мы сейчас приведем.

Распределение Стьюдента. Случайная величина T подчиняется закону распределения Стьюдента, если ее плотность вероятности

$$s_k(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

где k – целое положительное число. Заметим, что $\Gamma(\alpha)$ – это гамма-функция Эйлера. Она задается равенством

$$\Gamma(\alpha) = \int\limits_0^{\infty} x^{\alpha-1} e^{-x}\, dx \quad (\alpha > 0)$$

и обычно изучается в основном курсе высшей математики. Ясно, что интегральная функция распределения Стьюдента

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

Параметр k в этом законе называют числом степеней свободы распределения.

В статистических исследованиях часто используются таблицы квантилей распределения $S_k(x)$, в которых для заданных значений $S_k(x)$ приведены соответствующие значения x.

При $k \to \infty$ распределение Стьюдента превращается в нормальное. Поэтому при достаточно больших k можно заменять это распределение нормальным. На практике нормальное распределение используют при $k > 30$.

Распределение Стьюдента оказывается полезным при изучении оценок математического ожидания.

Распределение Пирсона. Случайная величина подчиняется закону распределения Пирсона, если ее плотность вероятностей

$$f_k(x) = \begin{cases} \dfrac{1}{\sqrt{2^k}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & , \quad x > 0; \\ 0 & , \quad x < 0. \end{cases}$$

Случайная величина в этом распределении неотрицательна. Поэтому ее принято обозначать χ^2 (хи в квадрате). Ясно, что

$$P(\chi^2 < \chi_0^2) = F_k(\chi_o^2) = \frac{1}{\sqrt{2^k}\Gamma(\frac{k}{2})} \int\limits_0^{\chi_0^2} t^{\frac{k}{2}-1} e^{-\frac{t}{2}}\, dt.$$

Здесь k также называют числом степеней свободы распределения.

Существуют таблицы квантилей распределения Пирсона.

При $k \to \infty$ распределение χ^2 с n степенями свободы приближается к нормальному.

Можно доказать такое утверждение: если случайные величины $x_1,\ x_2,\ \ldots\ , x_n$ независимы и все подчинены нормальному распределению с параметрами $m = 0$ и $\sigma = 1$, то случайная величина $\sum\limits_{i=1}^{n} x_i^2$ подчиняется закону распределения χ^2 с $n-1$ степенями свободы.

Распределение Пирсона оказывается полезным при построении оценок дисперсии и при изучении закона распределения генеральной совокупности.

Распределение Фишера. Если плотность вероятности случайной величины

$$f_{n,m}(x) = \begin{cases} \dfrac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\sqrt{\dfrac{n^n m^m x^{n-2}}{(m+nx)^{n+m}}} & , \quad x > 0 \\ 0 & , \quad x < 0, \end{cases}$$

то говорят, что величина подчиняется закону распределения Фишера. Это распределение имеет две степени свободы: n и m.

Как и в предыдущем случае, величина, подчиненная закону Фишера, принимает только положительные значения.

Если случайные величины X и Y подчиняются закону распределения Пирсона, причем имеют соответственно n и m степеней свободы, то $F = \dfrac{mX}{nY}$ описывается законом распределения Фишера со степенями свободы n и m.

Имеются таблицы квантилей закона Фишера.

Отметим одно важное обстоятельство: распределения Стьюдента, Пирсона и Фишера используют при обработке результатов наблюдения тогда, когда генеральная совокупность подчиняется нормальному распределению.

Распределение Колмогорова. Случайная величина подчиняется закону распределения Колмогорова, если ее функция распределения

$$K(x) = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2}.$$

Существуют таблицы квантилей этого распределения.

Заметим, что квантили этого распределения нетрудно вычислить приближенно. Действительно, задав близкую к единице вероятность $1-\alpha$, запишем уравнение

$$1 - \alpha = K(x_\alpha).$$

Так как ряд для $K(x)$ быстро сходится, то оставим в нем лишь первый член. Получим

$$1 - \alpha = 1 - 2e^{-2x_\alpha^2},$$

откуда $x_\alpha \approx \sqrt{\frac{1}{2}\ln\frac{2}{\alpha}}$.

3. ПОСТРОЕНИЕ ЭМПИРИЧЕСКОЙ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Предположим, мы произвели выборку объемом n из множества значений некоторой случайной величины X и по результатам выборки хотим найти функцию распределения этой случайной величины. Для этого прежде всего расположим элементы выборки в порядке возрастания. Расположенную таким образом выборку называют *вариационным рядом*. Пусть x_1 – наименьшее, а x_n – наибольшие значения элементов выборки. Разобьем промежуток $[x_1, x_n]$ на m частей. Границы частей обозначим

$$x_1 = \widetilde{x}_0 < \widetilde{x}_1 < .... < \widetilde{x}_{m-1} < \widetilde{x}_m = x_n$$

и подсчитаем число элементов выборки, попадающих в каждый промежуток. Обозначим эти числа соответственно n_1 , n_2 ,..., n_m . Затем найдем отношения $\frac{n_1}{n}$, $\frac{n_2}{n}$,..., $\frac{n_m}{n}$. Заметим, что

$$n_1 + n_2 + ... + n_m = n,$$

а потому $\frac{n_1}{n} + \frac{n_2}{n} + ... + \frac{n_m}{n} = 1$. Ясно, что найденные отношения можно приближенно принять за вероятности попадания значений X в соответствующие промежутки.

Для геометрической иллюстрации результатов над каждым промежутком $(\widetilde{x}_{k-1}, \widetilde{x}_k)$ строят прямоугольник, площадь которого пропорциональна отношению $\frac{n_k}{n}$. В результате получают фигуру, называемую *гистограммой*. На рис. 1 приведен пример такой гистограммы.

Понятно, что гистограмма – это эмпирический график плотности распределения генеральной совокупности.

Нередко используют другое построение. Для этого над серединой каждого из промежутков $(\widetilde{x}_{k-1}, \widetilde{x}_k)$ откладывают отрезок, длина которого равна $\frac{n_k}{n}$ (в соответсвующем масштабе). Затем концы отрезков соединяют. В результате получается *полигон (многоугольник)* эмпирического распределения (рис. 2), который также можно

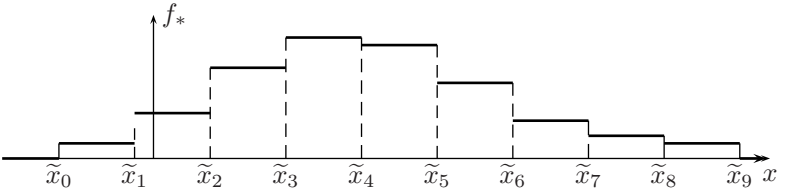

Рис. 1. Гистограмма распределения

рассматривать как график эмпирической (статистической) плотности распределения.

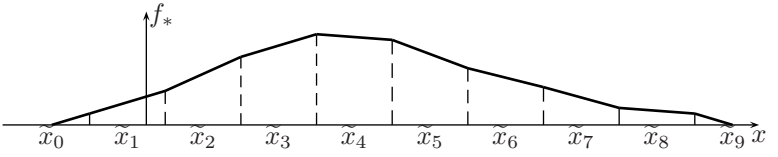

Рис. 2. Полигон распределения

Для того чтобы гистограмма более или менее правдиво отражала характер генеральной совокупности, объем n выборки должен быть достаточно большим (например, измеряться сотнями). При этом количество частей m на практике обычно выбирают в пределах от 12 до 20. Заметим, что существуют различные эмпирические формулы, связывающие m и n.

Наряду со статистической плотностью часто строят статистическую (эмпирическую) функцию распределения. Обозначив через $F_*(x)$ статистическую функцию распределения мы можем принять, что

$$F_*(\widetilde{x}_0) = P(X < \widetilde{x}_1) = 0,$$

$$F_*(\widetilde{x}_1) = \frac{n_1}{n},$$

$$F_*(\widetilde{x}_2) = \frac{n_1}{n} + \frac{n_2}{n},$$

...,

$$F_*(\widetilde{x}_m) = 1.$$

Отложив на графике точки $(\widetilde{x}_k,\ F_*(\widetilde{x}_k)$ и затем соединив их отрезками прямых, получим приближенный вид функции распределения (рис. 3).

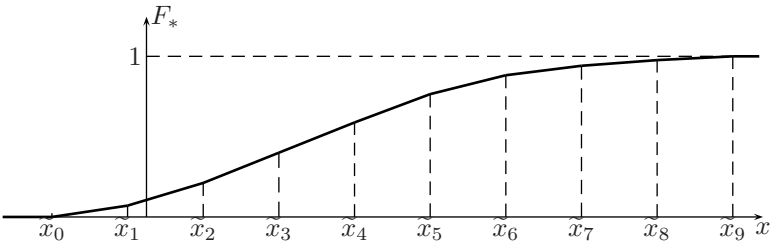

Рис. 3. Статистическая функция распределения

Нередко график эмпирической функции распределения строят иначе, а именно, на высоте, равной $F_*(\widetilde{x}_2) = \dfrac{n_1}{n}$, проводят отрезок от середины первого первого промежутка до середины второго. Затем на высоте $F_*(\widetilde{x}_2) = \dfrac{n_1}{n} + \dfrac{n_2}{n}$ строят отрезок от середины второго промежутка до середины третьего и т.д. В результате получается ступенчатый график статистической функции распределения. Пример такого графика дан на рис. 4.

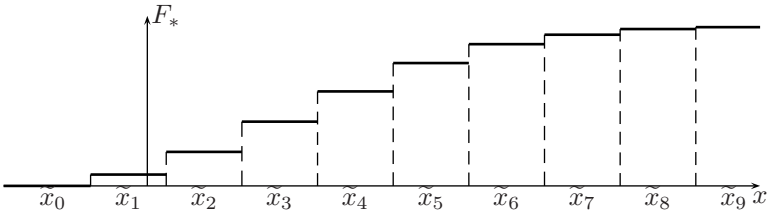

Рис. 4. Ступенчатая статистическая функция распределения

Заметим, что в современной практике длины отрезков, на которые разбивают всю область варьирования, берут, как правило, равными.

Можно доказать, что с ростом объема выборки и числа интервалов статистическая функция распределения приближается к истинной функции распределения генеральной совокупности.

Возможен и другой подход к построению функции распределения и плотности по выборочным данным. Этот подход предполагает, что вид закона распределения заранее известен. В этом случае задача сводится к нахождению, а точнее, к оценке параметров распределения по полученным данным. Например, если известно, что генеральная совокупность подчиняется нормальному распределению, то задача состоит в подборе по итогам эксперимента подходящих значений m и σ.

4. СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Предположим, что закон распределения генеральной совокупности зависит от некоторого параметра ϑ. Мы хотим, произведя выборку из генеральной совокупности, выяснить значение этого параметра. Естественно, что указать точное значение ϑ по выборке невозможно. Можно лишь построить некоторую его оценку. Существуют два подхода к построению такой оценки:

1. Строится некоторая функция g от элементов выборки. В эту функцию подставляют найденные в результате эксперимента значения $x_1, x_2, ..., x_n$ и то, что получится, считается приближенным значением ϑ. Такого рода оценка параметра называется *точечной*.

2. Задается достаточно близкая к единице вероятность P. Эта вероятность называется *доверительной*. Затем с помощью выборки строится интервал, такой, что с вероятностью P истинное значение ϑ находится в этом интервале. Такая оценка называется *интервальной*.

Как правило, точечные оценки строятся при достаточно больших объемах выборки, а интервальные – при малых.

Разумеется, сразу возникает вопрос: каким образом следует выбирать функции для точечной оценки и границ доверительного интервала? Общего ответа на этот вопрос нет, однако существуют некоторые часто используемые способы построения оценок. Ниже мы приведем два наиболее употребительных способа построения точеч-

ных оценок, а сейчас укажем некоторые требования к оценкам.

Начнем с рассмотрения точечной оценки.

Отметим, что g является функцией выборки $x_1, x_2, ..., x_n$. Поэтому до проведения выборки сама величина $g(x_1, x_2, ..., x_n)$ является случайной, поскольку случайными считаются элементы выборки. В соответствии с этим к g обычно предъявляют требования, выраженнные в вероятностных терминах:

1. Желательно, чтобы оценка g была *несмещенной*, т.е. чтобы было $M(g) = \vartheta$.

2. Желательно, чтобы оценка g была *эффективной*, т.е. чтобы $D(g)$ была меньше дисперсий всех других оценок.

Заметим, что, в соответствии с теоремой Рао–Крамера, дисперсия любой оценки не может убывать с ростом n быстрее, чем $\frac{1}{n}$.

3. Желательно, чтобы оценка g была *состоятельной*, т.е. чтобы для любого $\varepsilon > 0$ выполнялось условие

$$\lim_{n\to\infty} P(|\vartheta - g(x_1, x_2, ..., x_n)| < \varepsilon) = 1.$$

Последнее условие означает следующее: если оценка g состоятельная, то при больших объемах выборки можно быть практически уверенным в том, что оценка почти не отличается от истинного значения параметра ϑ.

К сожалению, не всегда удается найти оценку, удовлетворяющую всем трем требованиям.

5. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ ОЦЕНОК ПО МЕТОДУ МОМЕНТОВ

При этом методе считается, что известен вид закона распределения генеральной совокупности. Например, известен вид плотности $f(x, \theta_1, ..., \theta_k)$, но неизвестны значения параметров $\theta_1, ..., \theta_k$. Чтобы оценить эти параметры, вычисляют выборочные моменты

$$\begin{aligned}
&\overline{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n),\\
&\overline{x^2} = \frac{1}{n}(x_1^2 + x_2^2 + ... + x_n^2),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
&\overline{x^k} = \frac{1}{n}(x_1^k + x_2^k + ... + x_n^k).
\end{aligned}$$

. Предполагаем, что выборка осуществляется случайным образом, а потому все ее элементы равноправны. Из-за этого мы при нахождении средних значений сумм делим эти суммы на n.

Эти выборочные моменты затем сравнивают с найденными по $f(x, \theta_1, ..., \theta_k)$ моментами m_1, m_2, ..., m_k. Естественно, мы предполагаем, что моменты существуют. В результате получают систему k уравнений, из которой можно найти оценки $\hat{\theta}_1, ..., \hat{\theta}_k$ неизвестных параметров.

Заметим, что этот метод используют и в случае дискретной генеральной совокупности.

Можно доказать, что полученные этим методом оценки состоятельные и асимптотически несмещенные. Вместе с тем они не всегда оказываются эффективными.

Этот метод был предложен К. Пирсоном.

Приведем несколько примеров применения этого метода.

П р и м е р 1 (оценка параметра экспоненциального распределения). Пусть генеральная совокупность подчинена экспоненциальному распределению. Тогда плотность ее вероятностей должна быть такой (см. разд. 2):

$$f(x) = \begin{cases} 0 & \text{при } x < 0, \\ \lambda e^{-\lambda x} & \text{при } x > 0. \end{cases}$$

Чтобы оценить входящий в плотность параметр λ, найдем выборочный момент $\overline{x}$ и сравним его с моментом m_1, который для этого распределения равен $\frac{1}{\lambda}$. Получим $\overline{x} = \frac{1}{\lambda}$. Отсюда очевидно, что точечная оценка параметра

$$\hat{\lambda} = \frac{1}{\overline{x}}.$$

П р и м е р 2 (оценка параметров нормального распределения). Предположим, что генеральная совокупность имеет нормальное распределение, т.е ее плотность

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

Мы хотим с помощью выборки оценить значения параметров a и σ^2.

Так как неизвестных параметров лишь два, то по выборке найдем только два выборочных момента $\overline{x}$ и $\overline{x^2}$ и сравним их с моментами m_1 и m_2.

Из курса теории вероятности известно, что для нормального распределения $m_1 = a$ и $m_2 = a^2 + \sigma^2$. Поэтому образуется такая система

$$\overline{x} = a, \quad \overline{x^2} = a^2 + \sigma^2,$$

откуда

$$\hat{a} = \overline{x}, \quad \hat{\sigma}^2 = \overline{x^2} - \overline{x}^2.$$

Используя принятое в статистике обозначение

$$\overline{s}^2 = \overline{x^2} - \overline{x}^2,$$

можем записать оценки в виде

$$\hat{a} = \overline{x}, \quad \hat{\sigma}^2 = \overline{s}^2.$$

П р и м е р 3 (оценка параметров Γ-распределения). Мы уже говорили в разделе 2, что случайная величина подчиняется Γ-распределению, если плотность ее вероятностей

$$f(x) = \begin{cases} 0 & \text{при } x < 0, \\ \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad (\lambda > 0) & \text{при } x > 0. \end{cases}$$

Чтобы оценить входящие в закон распределения параметры α и λ, снова находим $\overline{x}$ и $\overline{x^2}$. Моменты Γ-распределения

$$m_1 = \frac{\alpha}{\lambda}, \quad m_2 = \frac{\alpha(\alpha-1)}{\lambda^2}.$$

Это приводит к системе

$$\overline{x} = \frac{\alpha}{\lambda}, \quad \overline{x^2} = \frac{\alpha(\alpha-1)}{\lambda^2},$$

решив которую получим оценки параметров:

$$\hat{\alpha} = \frac{\overline{x}^2}{\overline{x^2} - \overline{x}^2}, \quad \hat{\lambda} = \frac{\overline{x}}{\overline{x^2} - \overline{x}^2}$$

или

$$\hat{\alpha} = \frac{\overline{x}^2}{\overline{s}^2}, \quad \hat{\lambda} = \frac{\overline{x}}{\overline{s}^2}.$$

6. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ ОЦЕНОК ПО МЕТОДУ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Как и в предыдущем разделе, предположим, что известен вид закона распределения генеральной совокупности, но не определены значения входящих в него параметров $\theta_1, ..., \theta_k$.

Если генеральная совокупность подчинена непрерывному распределению, то для оценки этих параметров по выборке $x_1, x_2, ..., x_n$ составляет произведение

$$L = \prod_{i=1}^{n} f(x_i, \theta_1, ..., \theta_k) =$$

$$= f(x_1, \theta_1, ..., \theta_k) f(x_2, \theta_1, ..., \theta_k) ... f(x_n, \theta_1, ..., \theta_k).$$

Если генеральная совокупность подчинена дискретному распределению, то произведение

$$L = \prod_{i=1}^{n} p(x_i, \theta_1, ..., \theta_k).$$

В обоих случаях произведение называется функцией правдоподобия.

В качестве оценок коэффициентов принимают те их значения, при которых функция правдоподобия достигает максимума.

Напомним, что необходимым условием наибольшего значения является равенство нулю частных производных. Это приводит к системе уравнений

$$\frac{\partial L}{\partial \theta_1} = 0, \quad \frac{\partial L}{\partial \theta_2} = 0, ..., \frac{\partial L}{\partial \theta_k} = 0,$$

из которой можно найти оценки параметров. Эти оценки называют оценками наибольшего правдоподобия.

Полученные таким образом оценки всегда состоятельны и асимптотически эффективны, хотя иногда оказываются смещенными.

Этот способ получения оценок был предложен Р. Э. Фишером.

Заметим, что найти производную большого числа сомножителей оказывается весьма трудоемко. Поэтому, учитывая, что $\ln L$ возрастает при возрастании L и убывает при убывании L, обычно находят максимальное значение $\ln L$. Так как

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta_1, ..., \theta_k),$$

то нахождение производных оказывается менее трудоемким. Другими словами, оценки параметров находят из системы уравнений

$$\frac{\partial(\ln L)}{\partial\theta_1} = 0, \quad \frac{\partial(\ln L)}{\partial\theta_2} = 0, ..., \frac{\partial(\ln L)}{\partial\theta_k} = 0.$$

П р и м е р 1 (оценка параметра экспоненциального распределения). Составляем функцию правдоподобия:

$$L = \lambda^n e^{-\lambda x_1} e^{-\lambda x_2} ... e^{-\lambda x_n} = \lambda^n e^{-\lambda(x_1+x_2+...x_n)}.$$

Ясно, что

$$\ln L = n \ln \lambda - \lambda(x_1 + x_2 + ...x_n)$$

и

$$\frac{\partial(\ln L)}{\partial\lambda} = n\frac{1}{\lambda} - (x_1 + x_2 + ...x_n).$$

Приравняем производную к нулю:

$$n\frac{1}{\lambda} - (x_1 + x_2 + ...x_n) = 0,$$

откуда получаем оценку параметра:

$$\hat{\lambda} = \frac{n}{x_1 + x_2 + ...x_n} = \frac{1}{\overline{x}}.$$

Заметим, что результат оказался таким же, как и при методе моментов.

П р и м е р 2 (оценка параметров нормального распределения). В этом случае функция правдоподобия

$$L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{(x_1-a)^2}{\sigma^2} - \frac{(x_2-a)^2}{\sigma^2} - ... - \frac{(x_n-a)^2}{\sigma^2}}.$$

Находим последовательно

$$\ln L = -n\ln(\sqrt{2\pi}) - n\ln\sigma -$$
$$-\frac{(x_1-a)^2 + (x_2-a)^2 + ... + (x_n-a)^2}{2\sigma^2}.$$
$$\frac{\partial(\ln L)}{\partial a} = -\frac{2(x_1-a) + 2(x_2-a) + ... + 2(x_n-a)}{\sigma^2},$$
$$\frac{\partial(\ln L)}{\partial \sigma} = -\frac{n}{\sigma} + 2\frac{(x_1-a)^2 + (x_2-a)^2 + ... + (x_n-a)^2}{\sigma^3}.$$

Приравняем эти производные к нулю:

$$-\frac{2(x_1-a) + 2(x_2-a) + ... + 2(x_n-a)}{\sigma^2} = 0,$$
$$-\frac{n}{\sigma} + \frac{(x_1-a)^2 + (x_2-a)^2 + ... + (x_n-a)^2}{\sigma^3} = 0.$$

Несложные преобразования приводят эти уравнения к виду

$$a = \frac{x_1 + x_2 + ...x_n}{n},$$
$$\sigma^2 = \frac{(x_1-a)^2 + (x_2-a)^2 + ... + (x_n-a)^2}{n},$$

откуда получаются оценки

$$\hat{a} = \overline{x}, \quad \hat{\sigma}^2 = \overline{s}^2.$$

Как и в предыдущем примере, эти оценки совпали с теми, что были найдены по методу моментов.

П р и м е р 3 (оценка параметра распределения Пуассона). Как уже говорилось в разд. 2, случайная величина X подчиняется распределению Пуассона, если она принимает целые неотрицательные значения, причем

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad (k = 0, 1, 2, ...).$$

Чтобы оценить параметр λ по выборке $x_1, x_2, ..., x_n$, составляем функцию правдоподобия

$$L = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\ \frac{\lambda^{x_2}}{x_2!}e^{-\lambda}\ ...\quad \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{\lambda^{x_1+x_2+...+x_n}}{x_1!x_2!...x_n!}e^{-n\lambda}.$$

Отсюда

$$\ln L = (x_1 + x_2 + ... + x_n)\ln\lambda - \ln(x_1!x_2!...x_n!) - n\lambda$$

и

$$\frac{\partial(\ln L)}{\partial\lambda} = (x_1 + x_2 + ... + x_n)\frac{1}{\lambda} - n.$$

Приравняв частную производную к нулю, получим, как нетрудно видеть, оценку параметра:

$$\hat{\lambda} = \overline{x}.$$

7. ТОЧЕЧНАЯ ОЦЕНКА СРЕДНЕГО ЗНАЧЕНИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Пусть имеется некоторая случайная величина (генеральная совокупность). Мы хотим по выборке оценить неизвестное нам ее математическое ожидание m.

Полагая, что выборка $x_1, x_2, ..., x_n$ состоит из независимых элементов, примем в качестве точечной оценки для m величину

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n},$$

т.е. среднее арифметическое выборочных результатов.

Покажем, что такая оценка является "приятной во всех отношениях": несмещена, эффективна и состоятельна.

Для этого напомним, что до испытаний каждый элемент выборки – случайная величина, имеющая тот же закон распределения, что и генеральная совокупность. Значит, если математическое ожидание генеральной совокупности равно m, а дисперсия – σ^2, то для каждого x_k должны выполняться равенства $M(x_k) = m$ и $D(x_k) = \sigma^2$. Отсюда

$$M(\overline{x}) = M\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) =$$

$$= \frac{1}{n}(M(x_1) + M(x_2) + ... + M(x_n)) = \frac{1}{n}mn = m.$$

Значит, оценка несмещенная.

Так как элементы выборки независимы,то

$$D(\overline{x}) = D\left(\frac{x_1 + x_2 + ... + x_n}{n}\right) =$$

$$= \frac{1}{n^2}(D(x_1) + D(x_2) + ... + D(x_n)) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Такое значение дисперсии оценки, пропорциональное $\frac{1}{n}$, означает, что она эффективна.

Теперь напомним теорему Чебышева, которая обычно рассматривается в курсе теории вероятностей:

если имеется последовательность независимых случайных величин x_k $(k = 1, 2, 3, ...)$ с математическими ожиданиями m_k и дисперсиями σ_k^2, причем все дисперсии не превышают некоторой величины, то для любого фиксированного числа $\varepsilon > 0$ выполняется равенство

$$\lim_{n\to\infty} P\left(\left|\frac{x_1 + x_2 + ... + x_n}{n} - \frac{m_1 + m_2 + ... + m_n}{n}\right| < \varepsilon\right) = 1.$$

В частности, если все члены последовательности имеют одинаковые законы распределения с математическими ожиданиями m, то оказывается

$$\lim_{n\to\infty} P\left(\left|\frac{x_1 + x_2 + ...x_n}{n} - m\right| < \varepsilon\right) = 1.$$

Так как элементы выборки независимы и имеют одинаковые законы распределения, то к ним применимо последнее утверждение, т.е. для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P(|\overline{x} - m| < \varepsilon) = \lim_{n\to\infty} P\left(\left|\frac{x_1 + x_2 + ...x_n}{n} - m\right| < \varepsilon\right) = 1.$$

Значит, оценка $\overline{x}$ состоятельна.

Состоятельность этой оценки означает, что при большом объеме выборки можно быть практически уверенным в выполнении равенства $m = \overline{x}$.

8. ТОЧЕЧНАЯ ОЦЕНКА ДИСПЕРСИИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Примем в качестве оценки дисперсии величину

$$\overline{s}^2 = \frac{1}{n}\sum_{k=1}^{n}(x - \overline{x})^2.$$

Прежде чем оценивать качество этой оценки, преобразуем выражение для нее:

$$\overline{s}^2 = \frac{1}{n}\sum_{k=1}^{n}(x - \overline{x})^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k^2 - 2x_k\overline{x} + \overline{x}^2) =$$

$$= \frac{1}{n}\left(\sum_{k=1}^{n}x_k^2 - 2\overline{x}\sum_{k=1}^{n}x_k + n\overline{x}^2\right) = \frac{1}{n}\left(\sum_{k=1}^{n}x_k^2 - 2n\overline{x}^2 + n\overline{x}^2\right).$$

Отсюда

$$\overline{s}^2 = \frac{1}{n}\sum_{k=1}^{n}x_k^2 - \overline{x}^2$$

или

$$\overline{s}^2 = \overline{x^2} - \overline{x}^2.$$

Теперь найдем математическое ожидание $\overline{s}^2$. Для этого вначале вычислим $M(x_k^2)$. Так как

$$D(x_k) = M(x_k^2) - (M(x_k))^2,$$

то

$$M(x_k^2) = D(x_k) + (M(x_k))^2 = \sigma^2 + m^2.$$

Далее, как было показано, $M(\overline{x}) = m$ и $D(\overline{x}) = \dfrac{\sigma^2}{n}$. Поэтому

$$M(\overline{x}^2) = D(\overline{x}) + (M(\overline{x}))^2 = \frac{\sigma^2}{n} + m^2.$$

Следовательно,

$$M(\overline{s}^2) = \frac{1}{n}\sum_{k=1}^{n} M(x_k^2) - M(\overline{x}^2) =$$

$$= \frac{1}{n} n(\sigma^2 + m^2) - \left(\frac{\sigma^2}{n} + m^2\right) = \frac{n-1}{n}\sigma^2.$$

Таким образом, $M(\overline{s}^2) \neq \sigma^2$, т.е. оценка $M(\overline{s}^2)$ оказывается смещенной. Если умножить $\overline{s}^2$ на $\dfrac{n}{n-1}$, то получится несмещенная оценка дисперсии:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2.$$

Видно, что при больших значениях n оценки практически одинаковы; при $n > 30$ их, как правило, не различают.

Можно доказать, что обе оценки дисперсии состоятельны и оценка $\overline{s}^2$ эффективна.

Величина

$$s = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

обычно принимается в качестве точечной оценки для σ_x – среднего квадратического отклонения генеральной совокупности.

П р и м е р. В результате испытаний получены такие значения случайной величины: 6,2; 4,8; 4,9; 5,6; 5,7; 6,3; 4,9; 5,3; 6,5; 5,5. Нетрудно подсчитать, что в этом случае

$$\overline{x} = 5,57,\ \overline{s}^2 = 0,338,\ s^2 = 0,376,\ s = 0,61.$$

9. ТОЧЕЧНАЯ ОЦЕНКА МЕДИАНЫ

Напомним, что медианой случайной величины X называют число ϑ, для которого $P(x < \vartheta) = P(x > \vartheta) = \dfrac{1}{2}$.

Чтобы оценить медиану, вначале расположим элементы выборки в порядке возрастания, т.е. построим вариационный ряд. Если вариационный ряд содержит нечетное число элементов, то за оценку $\hat{\vartheta}$ медианы принимают срединный элемент ряда. Если же в вариационном ряде четное число элементов, то в качестве оценки $\hat{\vartheta}$ берут полусумму двух срединных элементов ряда.

П р и м е р. Найдем оценку медианы для результатов испытания из предыдущего примера. Для этого расположим их в порядке возрастания: 4,8; 4,9; 4,9; 5,3; 5,5; 5,6; 5,7; 6,2; 6,3; 6,5. В выборке 10 элементов. Поэтому $\overline{\vartheta} = \dfrac{5,5+5,6}{2} = 5,55$.

10. ТОЧЕЧНАЯ ОЦЕНКА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Из курса теории вероятностей известно, что корреляционным моментом пары случайных величин X и Y называют величину

$$K_{xy} = M(XY) - M(X)M(Y),$$

а коэффициентом корреляции –

$$k_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y},$$

где σ_x и σ_y – средние квадратические отклонения величин X и Y.

Напомним, что корреляционный момент характеризует линейную связь случайных величин. Именно, чем ближе $|k_{xy}|$ к единице, тем ближе зависимость между X и Y к линейной

Предположим, что генеральной совокупностью является система случайных величин (X, Y). В результате испытаний получено n пар значений этих величин (x_k, y_k). Мы хотим оценить корреляционный момент K_{xy} и коэффициент корреляции k_{xy} этих величин. Для этого вначале находим

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad \text{и} \quad \overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k,$$

а затем в качестве оценки корреляционного момента принимаем величину

$$\overline{R}_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}).$$

Для упрощения практических вычислений это выражение нетрудно преобразовать к виду

$$\overline{R}_{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{x}\ \overline{y}.$$

Можно показать, что $M(\overline{R}_{xy}) = \dfrac{n-1}{n}K_{xy}$, т.е. наша оценка смещенная. Несмещенной оценкой корреляционного момента, очевидно, будет

$$R_{xy} = \frac{n}{n-1}\overline{R}_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) =$$

$$= \frac{1}{n-1}\sum_{k=1}^{n} x_k y_k - \frac{n}{n-1}\overline{x}\ \overline{y}.$$

Обе оценки состоятельны и при больших значениях практически не различаются.

В соответствии со сказанным в качестве оценки коэффициента корреляции принимают величину

$$r_{xy} = \frac{R_{xy}}{s_x s_y} = \frac{\overline{R}_{xy}}{\overline{s}_x \overline{s}_y}.$$

П р и м е р. В результате выборки из генеральной совокупности (X, Y) получены такие пары значений (0, 1); (1, 8); (2, 16); (3, 26); (4, 34); (5, 38); (6, 48); (7, 62); (8, 74). Найдем точечную оценку коэффициента корреляции.

Находим последовательно

$$\overline{x} = 4; \quad \overline{y} = 34,1;$$

$$\overline{R}_{xy} = \frac{1}{9}\sum_{k=1}^{9} x_k y_k - \overline{x}\ \overline{y} = 58,9.$$

Затем вычисляем

$$\overline{s}_x^2 = \frac{1}{9}\sum_{k=1}^{9} x_k^2 - \overline{x}^2 = 6,7; \quad \overline{s}_x = 2,6;$$

$$\overline{s}_y^2 = \frac{1}{9}\sum_{k=1}^{9} y_k^2 - \overline{y}^2 = 528,4; \quad \overline{s}_y = 23,0.$$

Наконец, получаем

$$r_{xy} = \frac{\overline{R}_{xy}}{\overline{s}_x \overline{s}_y} = \frac{58,9}{2,6 \cdot 23,0} = 0,985.$$

Такая величина коэффициента корреляции говорит о почти линейной зависимости случайных величин.

11. ИНТЕРВАЛЬНАЯ ОЦЕНКА ВЕРОЯТНОСТИ СОБЫТИЯ ПО ЕГО ЧАСТОТЕ

Предположим, что произведено n одинаковых испытаний, в каждом из которых могло осуществиться событие A с вероятностью p. В результате испытаний оказалось, что событие A осуществилось k раз, так что частота его оказалась такой: $p_* = \dfrac{k}{n}$. Требуется, зная частоту p_*, оценить значение p.

Прежде чем строить интервальную оценку вероятности p, заметим, что p_* можно считать точечной оценкой вероятности p. Другими словами, можно принять $p \approx p_*$.

Из курса теории вероятностей известно, что при n одинаковых испытаниях частота p_* подчиняется биномиальному закону распределения, т.е.

$$P\left(p_* = \frac{k}{n}\right) = C_n^k p^k (1-p)^{n-k}.$$

В соответствии с теоремой Муавра–Лапласа при достаточно больших n биномиальный закон можно заменить нормальным с параметрами $m = p$ и $\sigma = \sqrt{\dfrac{p(1-p)}{n}}$. Иначе говоря, при таких значе-

ниях n частота описывается функцией Лапласа, так что

$$P(p_* < x) = \Phi\left(\frac{(x-p)\sqrt{n}}{\sqrt{p(1-p)}}\right).$$

Отсюда

$$P(|p_* - p| < \varepsilon) = P(p - \varepsilon < p_* < p + \varepsilon) =$$

$$= \Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) =$$

$$= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1.$$

Для построения доверительного интервала зададим близкую к единице вероятность $1-\alpha$. Эту вероятность называют доверительной. Затем найдем такое ε, для которого

$$P(|p_* - p| < \varepsilon) = 1 - \alpha.$$

Это равенство запишем в виде

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 = 1 - \alpha$$

или

$$\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) = 1 - \frac{\alpha}{2}.$$

С помощью таблицы квантилей нормального распределения находим $\dfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} = x_\alpha$, откуда $\varepsilon = x_\alpha\sqrt{\dfrac{p(1-p)}{n}}$. Заменяя в последнем выражении неизвестное нам p на его приближенное значение p_*, получаем $\varepsilon = x_\alpha\sqrt{\dfrac{p_*(1-p_*)}{n}}$. Значит, с доверительной вероятностью $P = 1 - \alpha$ можно утверждать, что истинное значение p находится в интервале

$$\left(p_* - x_\alpha\sqrt{\frac{p_*(1-p_*)}{n}}, p_* + x_\alpha\sqrt{\frac{p_*(1-p_*)}{n}}\right).$$

П р и м е р. При 35 однотипных испытаниях событие A осуществилось 24 раза. Найдем с доверительной вероятностью $P = 0,9$ вероятность p появления A при одном испытании.

Очевидно, что $p_* = \dfrac{24}{35} \approx 0,69$.

Далее, так как $\alpha = 0,1$, по таблице квантилей находим $x_{0,1} = 1,64$. Поэтому $\varepsilon = 1,64\sqrt{\dfrac{0,69 \cdot 0,31}{35}} \approx 0,13$.

Таким образом, с вероятностью 0,9 можно утверждать, что значение p находится в интервале $(0,56;\ 0,82)$.

12. ИНТЕРВАЛЬНАЯ ОЦЕНКА СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Пусть генеральная совокупность имеет нормальное распределение. Получив выборку $x_1,\ x_2,\ \ldots,x_n$, мы хотим построить доверительный интервал для ее среднего значения (математического ожидания) m. Для этого введем случайную величину

$$T = \sqrt{n-1}\ \frac{\overline{x} - m}{\overline{s}},$$

где, как и выше,

$$\overline{x} = \frac{x_1 + x_2 = \ldots + x_n}{n} \quad \text{и} \quad \overline{s} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x - \overline{x})^2}.$$

Эта случайная величина подчиняется закону распределения Стьюдента с $n-1$ степенью свободы, т.е. описывается функцией распределения $S_{n-1}(x)$.

Зададим доверительную вероятность $P = 1 - \alpha$, где α мало, и по таблице квантилей распределения Стьюдента найдем квантиль t_α, для которого $P(|T| < t_\alpha) = 1 - \alpha$. Следовательно, можно утверждать, что с вероятностью $P = 1 - \alpha$ выполняется неравенство

$$\sqrt{n-1}\frac{|\overline{x} - m|}{\overline{s}} < t_\alpha$$

или, что то же,

$$\overline{x} - t_\alpha \frac{\overline{s}}{\sqrt{n-1}} < m < \overline{x} + t_\alpha \frac{\overline{s}}{\sqrt{n-1}}.$$

Итак, доверительный интервал для среднего значения генеральной совокупности имеет вид

$$\left(\overline{x} - t_\alpha \frac{\overline{s}}{\sqrt{n-1}},\ \overline{x} + t_\alpha \frac{\overline{s}}{\sqrt{n-1}} \right).$$

Таким образом, истинное среднее значение нашей генеральной совокупности является одним из чисел построенного интервала с вероятностью $P = 1 - \alpha$.

Из приведенной методики построения доверительного интервала нетрудно понять, что при увеличении объема выборки n – уменьшается пропорционально $\dfrac{1}{\sqrt{n}}$, а при увеличении доверительной вероятности его длина увеличивается .

П р и м е р. В примере из разд. 7 были заданы 10 выборочных значений из генеральной совокупности. По этим значениям мы нашли точечную оценку среднего $\overline{x} = 5,57$ и дисперсии $\overline{s}^2 = 0,338$. Теперь зададим доверительную вероятность $P = 0,9$ и построим доверительный интервал для этого среднего значения. В нашем случае $\alpha = 0,1$ и $n - 1 = 9$. По таблице квантилей находим $t_{0,1} = 1,833$, откуда $t_\alpha \dfrac{\overline{s}}{\sqrt{n-1}} = 1,833 \dfrac{\sqrt{0,338}}{3} = 0,36$, а потому доверительный интервал для среднего значения имеет вид $(5,21;\ 5,93)$.

Таким образом, с доверительной вероятностью 0,9 среднее значение генеральной совокупности является одним из чисел интервала $(5,21;\ 5,93)$.

Если бы мы увеличили доверительную вероятность до $P = 0,95$, то оказалось бы, что $t_{0,05} = 2,262$ и потому доверительный интервал стал бы таким: $(5,13;\ 6,01)$.

К сказанному следует добавить два замечания.

Во-первых, как уже говорилось, при $n > 30$ вместо распределения Стьюдента следует использовать нормальное распределение.

Во-вторых, если заранее известна дисперсия σ^2 генеральной совокупности, то вместо подчиняющейся закону Стьюдента случайной величины T следует рассматривать случайную величину $\sqrt{n}\dfrac{\overline{x} - m}{\sigma}$,

подчиняющуюся нормальному распределению с параметрами 0 и 1. Естественно, что в этом случае длина доверительного интервала будет при той же доверительной вероятности меньше, поскольку у нас имеется более полная информация о генеральной совокупности.

П р и м е р. Как и в предыдущем примере, считаем, что $n = 10$ и $\overline{x} = 5,57$, но известна дисперсия генеральной совокупности $\sigma^2 = 0,338$. В таком случае $\sqrt{10}\dfrac{\overline{x} - 5,57}{\sqrt{0,338}}$, имеет нормальное распределение с параметрами 0 и 1. Задав $P = 0,9$, находим по таблице квантилей нормального распределения $t_{0,1} = 1,64$. Поэтому доверительный интервал для среднего значения оказывается таким $(5,27; 5,87)$.

В заключение отметим два очевидных свойства любого доверительного интервала:

1)увеличение доверительной вероятности влечет за собой удлинение доверительного интервала;

2) увеличение объема выборки приводит к уменьшению доверительного интервала, которое пропорционально $\dfrac{1}{\sqrt{n}}$, где n – объем выборки.

13. ИНТЕРВАЛЬНАЯ ОЦЕНКА ДИСПЕРСИИ НОРМАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Как и выше, предположим. что генеральная совокупность имеет нормальное распределение. Исходя из выборки, строим статистику

$$\chi^2 = n\,\frac{\overline{s}^2}{\sigma^2}.$$

Эта статистика, очевидно, является случайной величиной. Можно доказать, что она подчиняется закону распределения Пирсона с $n-1$ степенью свободы. Задав доверительную вероятность $P = 1 - \alpha$, ищем такой интервал $(\chi_1^2,\ \chi_2^2)$, чтобы $P(\chi_1^2 < \chi^2 < \chi_2^2) = 1 - \alpha$. Это значит, что должно быть $F_{n-1}(\chi_2^2) - F_{n-1}(\chi_1^2) = 1 - \alpha$, где $F_{n-1}(\chi_2)$ – функция распределения Пирсона. Обычно, используя таблицу квантилей распределения Пирсона, выбирают границы ин-

тервала так, чтобы было

$$F_{n-1}(\chi_2^2) = 1 - \frac{\alpha}{2}, \quad F_{n-1}(\chi_1^2) = \frac{\alpha}{2}.$$

Значит, с доверительной вероятностью выполняется неравенство

$$\chi_1^2 < n\,\frac{\overline{s}^2}{\sigma^2} < \chi_2^2$$

или

$$n\,\frac{\overline{s}^2}{\chi_2^2} < \sigma^2 < n\,\frac{\overline{s}^2}{\chi_1^2}.$$

Таким образом, доверительный интервал для дисперсии

$$\left(n\,\frac{\overline{s}^2}{\chi_2^2},\ n\,\frac{\overline{s}^2}{\chi_1^2}\right).$$

Ясно, что доверительный интервал для среднего квадратичного отклонения

$$\left(\sqrt{n}\,\frac{\overline{s}}{\chi_2},\ \sqrt{n}\,\frac{\overline{s}}{\chi_1}\right).$$

П р и м е р. В разд. 7 у нас было $n = 10$ и $\overline{s}^2 = 0,338$. Задав доверительную вероятность $P = 0,9$, по таблице квантилей находим χ_1^2 и χ_2^2 так, чтобы было $F_9(\chi_2^2) = 0,95$ и $F_9(\chi_1^2) = 0,05$. Оказывается, $\chi_2^2 = 16,9$ и $\chi_1^2 = 3,32$. Отсюда

$$n\,\frac{\overline{s}^2}{\chi_2^2} = 10\frac{0,338}{16,9} = 0,20; \quad n\,\frac{\overline{s}^2}{\chi_1^2} = 10\frac{0,338}{3,32} = 1,02.$$

Таким образом, доверительный интервал для дисперсии имеет вид $(0,20;\ 1,02)$.

Легко понять, что доверительный интервал для среднего квадратического отклонения такой: $(0,45;\ 1,01)$.

14. ИНТЕРВАЛЬНАЯ ОЦЕНКА ПАРАМЕТРА ЭКСПОНЕНЦИАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

Пусть генеральная совокупность X имеет экспоненциальное распределение с неизвестным параметром λ, который мы хотим оценить с помощью выборки. Для этого, получив выборку $x_1, x_2, ..., x_n$,

найдем $\overline{x}$ и составим случайную величину

$$\chi^2 = 2\lambda n\overline{x}.$$

Оказывается, что эта случайная величина подчиняется закону распределения Пирсона с $2n$ степенями свободы. Поэтому, задав доверительную вероятность $P = 1 - \alpha$, найдем два таких значения χ_1^2 и χ_2^2, для которых выполняются условия

$$F_{2n}(\chi_2^2) = 1 - \frac{\alpha}{2}, \quad F_{2n}(\chi_1^2) = \frac{\alpha}{2}.$$

Теперь можно утверждать, что с заданной доверительной вероятностью выполняется неравенство

$$\chi_1^2 < 2\lambda n\overline{x} < \chi_2^2.$$

Это значит, что доверительный интервал для λ

$$\left(\frac{\chi_1^2}{2n\overline{x}},\ \frac{\chi_2^2}{2n\overline{x}}\right).$$

Полезно отметить следующее обстоятельство. Для экспоненциально распределенной с параметром λ случайной величины $MX = \dfrac{1}{\lambda}$. Поэтому доверительный интервал для ее среднего значения

$$\left(\frac{2n\overline{x}}{\chi_2^2},\ \frac{2n\overline{x}}{\chi_1^2}\right).$$

П р и м е р. Пусть $\overline{x} = 2,6$ и $n = 12$.

Зададим $P = 0,9$ и из уравнений $F_{24}(\chi_1^2) = 0,05$ и $F_{24}(\chi_2^2) = 0,95$ с помощью таблиц находим $\chi_1^2 = 13,8$ и $\chi_2^2 = 36,4$. Отсюда

$$\frac{\chi_1^2}{2n\overline{x}} = \frac{13,8}{24 \cdot 2,6} = 0,22, \quad \frac{\chi_2^2}{2n\overline{x}} = \frac{36,4}{24 \cdot 2,6} = 0,58,$$

т.е. найден доверительный интервал для параметра λ: $(0,22;\ 0,58)$.

Заметим, что доверительный интервал для среднего значения, случайной величины то есть для $\dfrac{1}{\lambda}$, будет таким: $(1,71;\ 4,52)$.

15. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

Гипотезой называют предположение о каком-либо свойстве генеральной совокупности. Статистическая проверка заключается в

принятии или отклонении этой гипотезы по выборочным данным.

Обычно используется следующая методика. Формулируют проверяемую гипотезу. Назовем ее H_0. Затем строят некоторую функцию от выборочных данных $g(x_1, x_2, ..., x_n)$, которую называют критерием проверки.

Предположив, что гипотеза H_0 справедлива, находят закон распределения величины $g(x_1, x_2, ..., x_n)$. После этого, задав доверительную вероятность $P = 1-\alpha$, делят область возможных значений величины g на две части G и $\overline{G}$ такие, что $P(g \in G|H_0) = 1-\alpha$ и $P(g \in \overline{G}|H_0) = \alpha$.

Область G называют допустимой для гипотезы H_0, а область $\overline{G}$ – критической.

Если при подстановке в критерий проверки g полученных в результате испытаний значений окажется $g(x_1, x_2, ..., x_n) \in \overline{G}$, то гипотеза H_0 отвергается. Если же $g(x_1, x_2, ..., x_n) \in G$, то это означает, что результаты испытания не противоречат гипотезе H_0, т.е. она правдоподобна.

На рис. 5, где отрезок AB представляет собой область возможных значений g, указаны типичные виды областей G и $\overline{G}$.

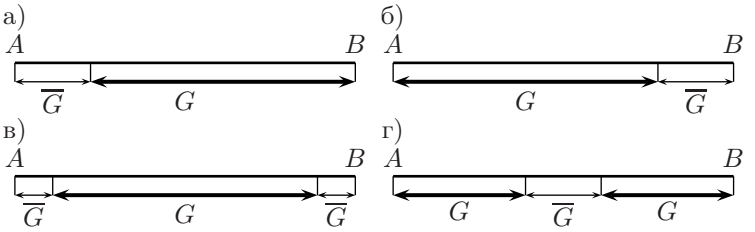

Рис. 5. Виды допустимых и критических областей

Случаи а) и б) соответствуют так называемым односторонним критериям, а в) и г) – двусторонним.

Ясно, что критерий проверки является случайной величиной, поскольку он зависит от элементов выборки.

Поэтому, каким бы ни был критерий проверки гипотезы, всегда можно совершить одну из двух ошибок:

1) *ошибка первого рода* – гипотеза отвергается, хотя она справедлива;

2) *ошибка второго рода* – гипотеза принимается, хотя она не справедлива.

Вероятность ошибки первого рода обычно обозначают через α. Нетрудно понять, что $\alpha = P(g \in \overline{G}|H_0)$. Вероятность α называют *уровнем значимости*. Вероятность ошибки второго рода принято обозначать через β. Обозначив через $\overline{H}_0$ – альтернативную гипотезу, т.е. отрицание гипотезы H_0, можем написать, что $\beta = P(g \in G|\overline{H}_0)$. Величину $1 - \beta$ называют *мощностью* критерия проверки.

Естественно, что для проверки одной и той же гипотезы можно использовать различные критерии, причем лучшетот, для которого вероятности обеих ошибок малы. Однако уменьшение вероятности одной ошибки ведет к росту вероятности другой. В связи с этим нередко поступают так: задав значение β, находят такой критерий проверки, у которого α минимально по сравнению с другими.

В заключение повторим порядок проверки гипотез:

1) формулируются проверяемая гипотеза H_0 и альтернативная $\overline{H}_0$;

2) выбирается критерий проверки $g(x_1, x_2, ..., x_n)$;

3) находится закон распределения критерия g при условии, что справедлива гипотеза H_0;

4) задается уровень значимости α и по нему строятся допустимая G и критическая $\overline{G}$ области значений критерия g;

5) производится выборка $x_1, x_2, ..., x_n$ из генеральной совокупности;

6) полученная выборка подставляется в критерий g.

Если получившееся значение критерия попадает в область G, то это означает, что выборка не противоречит гипотезе H_0. В противном случае гипотеза H_0 отвергается.

16. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ СРЕДНИХ ЗНАЧЕНИЙ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

1. Предположим, что имеются две генеральные совокупности X и Y, подчиняющиеся нормальному распределению с одинаковыми, но неизвестными дисперсиями. Обозначим их математические ожи-

дания соответственно a_1 и a_2. Мы хотим провести статистическую проверку гипотезы $H_0: a_1 = a_2$.

Для проверки сделаем независимые друг от друга выборки из обеих генеральных совокупностей: $x_1, x_2, ..., x_n$ и $y_1, y_2, ..., y_m$. По этим выборкам вначале найдем $\overline{x}$, $\overline{y}$, $\overline{s}_x^2$, $\overline{s}_y^2$.

В качестве критерия проверки нашей гипотезы примем величину

$$T = \sqrt{\frac{nm(n+m-2)}{n+m}}\ \frac{\overline{x}-\overline{y}}{\sqrt{n\overline{s}_x^2 + m\overline{s}_y^2}}.$$

Если гипотеза о равенстве средних значений справедлива, то T подчиняется закону распределения Стьюдента с $n+m-2$ степенями свободы. В соответствии с этим, задав доверительную вероятность $P = 1-\alpha$, найдем по таблицам квантилей распределения Стьюдента такое t_α, что $P(|T| < t_\alpha) = 1-\alpha$. Значит, допустимая область значений T для нашей гипотезы является интервалом $(-t_\alpha,\ t_\alpha)$, а критической оказывается $(-\infty,\ -t_\alpha) + (t_\alpha, +\infty)$.

Итак, если при подстановке в T выборочных данных окажется $T \in (-t_\alpha,\ t_\alpha)$, то гипотеза H_0 о равенстве математических ожиданий обеих совокупностей не отвергается.

П р и м е р. Пусть получены выборочные данные, у которых $n = 12$, $m = 10$, $\overline{x} = 93$, $\overline{y} = 96$, $\overline{s}_x^2 = 2,54$, $\overline{s}_y^2 = 2,60$.

Здесь критерий T подчиняется закону распределения Стьюдента с 20 степенями свободы. Задав $P = 0,9$, найдем по таблице квантилей $t_{0,1} = 1,73$. Значит, допустимой областью для T является интервал $(-1,73;\ 1,73)$.

Подставив в T выборочные данные, получим

$$T = \sqrt{\frac{12\cdot 10\cdot 20}{22}}\ \frac{-3}{\sqrt{12\cdot 2,54 + 10\cdot 2,60}} = -4,17.$$

Очевидно, что T не попало в допустимую область. Следовательно, гипотеза о равенстве средних значений обеих совокупностей должна быть отвергнута.

2. Предположим опять, что имеются две генеральные совокупности X и Y, подчиняющиеся нормальному распределению, с одинаковыми дисперсиями. Однако теперь будем считать, что дисперсии этих совокупностей нам известны и равны σ^2.

В этом случае, если считать гипотезу H_0 справедливой, разность $\overline{x} - \overline{y}$ подчиняется нормальному распределению с нулевым средним и дисперсией $\dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m} = \dfrac{n+m}{nm}\sigma^2$. Поэтому величина

$$T = \sqrt{\frac{nm}{n+m}}\frac{\overline{x} - \overline{y}}{\sigma}$$

имеет нормальное распределение с нулевым средним и единичной дисперсией. Значит,

$$P(|T| < t) = 2\Phi(t) - 1.$$

Теперь, задав доверительную вероятность $P = 1 - \alpha$, найдем по таблице квантилей нормального распределения такое t_α, для которого $P(|T| < t_\alpha) = 1 - \alpha$.

Если для сделанных выборок окажется $|T| < t_\alpha$, то гипотеза H_0 не отвергается.

П р и м е р. Пусть, как и выше, $n = 12$, $m = 10$, $\overline{x} = 93$, $\overline{y} = 96$. Кроме того считаем, что нам известно $\sigma^2 = 2,6$.

Задав доверительную вероятность $P = 0,9$, по таблице квантилей нормального распределения найдем, что $t_{0,1} = 1,64$, откуда

$$T = \sqrt{\frac{12 \cdot 10}{12 + 10}}\frac{(-3)}{\sqrt{2,6}} = -4,34.$$

Так как $4,34 > 1,64$, то гипотеза о равенстве средних значений отвергается.

17. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ ДИСПЕРСИЙ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

Как и раньше, предположим, что имеются две генеральные совокупности, подчиненные нормальному распределению. Проверяем гипотезу H_0: обе генеральные совокупности имеют одну и ту же дисперсию, т.е. $\sigma_1^2 = \sigma_2^2$. При этом альтернативной считаем гипотезу $\sigma_1^2 \neq \sigma_2^2$.

Для проверки сделаем выборки из этих генеральных совокупностей объемом m и n элементов. По выборкам построим оценки дисперсий, которые обозначим соответственно s_1^2 и s_2^2. Будем считать, что $s_2^2 > s_1^2$ и составим отношение $F = \dfrac{s_2^2}{s_1^2}$. Это отношение подчиняется распределению Фишера с $n-1$ и $m-1$ степенями свободы. Теперь зададим доверительную вероятность $P = 1-\alpha$ и найдем такие два числа F_1 и F_2, для которых $P(F < F_1) = P(F > F_2) = \dfrac{\alpha}{2}$. Тогда у нас будет

$$P\left(F_1 < F < F_2\right) = 1 - \alpha.$$

Используя таблицу квантилей распределения Фишера по параметрам $n-1$, $m-1$ и $\dfrac{\alpha}{2}$, найдем такое значение F_2, для которого $P(F > F_2) = \dfrac{\alpha}{2}$. Далее заметим, что величина $\dfrac{1}{F}$ тоже подчиняется закону распределения Фишера, но с числом степеней свободы $m-1$ и $n-1$. В соответствии с этим найдем по параметрам $m-1$, $n-1$ и $\dfrac{\alpha}{2}$ такое F', для которого $P(\dfrac{1}{F} > F') = \dfrac{\alpha}{2}$. Последнее равенство означает, что $P(F < \dfrac{1}{F'}) = \dfrac{\alpha}{2}$. Значит, должно быть $F_1 = \dfrac{1}{F'}$.

Таким образом, если величина F попадает в допустимую область (F_1, F_2), то гипотеза о равенстве дисперсий не отвергается.

П р и м е р. Пусть $\overline{s}_x^2 = 4.76$, $n = 20$, $\overline{s}_y^2 = 3,50$, $m = 16$.

Проверим гипотезу о равенстве дисперсий, задав доверительную вероятность $P = 0,9$. Так как уровень значимости $\alpha = 0,1$ и числа степеней свободы равны 19 и 15, то по таблице квантилей распределения Фишера найдем $F_2 = F_{0,05;19;15} = 2,3$. Далее по той же таблице найдем $F' = F(0,05;15;19) = 2,2$. Значит, $F_1 = \dfrac{1}{2,2} = 0,45$. Таким образом, допустимой областью для нашей гипотезы является интервал $(0,45;\ 2,20)$.

Отношение $\dfrac{s_x^2}{s_y^2} = \dfrac{4,76}{3,50} = 1,36$ лежит в этом интервале, поэтому с доверительной вероятностью 0,9 гипотеза о равенстве дисперсий не отвергается

18. ПРОВЕРКА ГИПОТЕЗЫ ОБ ОДНОРОДНОСТИ НАБЛЮДЕНИЙ (ОТСУТСТВИИ ГРУБЫХ ОШИБОК)

Грубые ошибки обычно возникают при неправильном поведении прибора, при неправильном снятии его показаний, при неверных вычислениях и т.д.

Будем считать, что генеральная совокупность описывается функцией распределения $F(x)$ и что из нее сделана выборка, состоящая из n элементов $x_1, x_2, ..., x_n$.

Пусть гипотеза H_0 такая: наибольший из элементов выборки $x_{\max}$ принадлежит генеральной совокупности (не является грубой ошибкой).

При этой гипотезе событие $x_{\max} < x$ равносильно тому, что все n элементов выборки меньше x. Поэтому

$$P(x_{\max} < x) = P(x_1 < x)P(x_2 < x)...P(x_n < x) = F^n(x).$$

Задав для проверки гипотезы уровень значимости α, получим

$$P(x_{\max} < x) = 1 - \alpha$$

или

$$F^n(x) = 1 - \alpha,$$

откуда

$$F(x) = (1 - \alpha)^{\frac{1}{n}}.$$

Найдя из последнего уравнения квантиль x, сравним с ним $x_{\max}$. Если окажется $x_{\max} > x$, то гипотеза H_0 отвергается.

Аналогичным образом проверяется гипотеза о том, что наименьший из элементов выборки $x_{\min}$ является грубой ошибкой. В этом случае, задав уровень значимости α, получим

$$P(x_{\min} > x) = (1 - F(x))^n = 1 - \alpha$$

или

$$F(x) = 1 - (1 - \alpha)^{\frac{1}{n}}.$$

Из этого уравнения найдем квантиль x. Если $x_{\min} < x$, то гипотеза об отсутствии грубых ошибок отвергается.

П р и м е р. Пусть генеральная совокупность подчинена нормальному распределению с параметрами $m = 10,3$ и $\sigma = 1,8$, Из совокупности сделана выборка, состоящая из 10 элементов. Наименьший из этих элементов $x_{\min} = 6$. Проверим гипотезу о том, что этот элемент не является грубой ошибкой. В соответствии с условием задачи функция распределения имеет вид $\Phi\left(\dfrac{x-10,3}{1,8}\right)$. Задав уровень значимости $\alpha = 0,2$ получим уравнение

$$\Phi\left(\frac{x-10,3}{1,8}\right) = 1 - 0,8^{\frac{1}{10}}.$$

В силу свойства функции Лапласа перепишем его:

$$\Phi\left(\frac{10,3-x}{1,8}\right) = 0,8^{\frac{1}{10}} \approx 0,98.$$

По таблице квантилей нормального распределения из уравнения

$$\Phi(t) = 0,98$$

найдем $t = 2,05$. Отсюда получим $\dfrac{10,3-x}{1,8} = 2,05$, так что

$$x_\alpha = 10,3 - 1,8 \cdot 2,05 = 6,6.$$

Поскольку $6,6 > 6 = x_{\min}$, то значение следует считать грубой ошибкой.

19. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ВЫБОРКИ ЗАДАННОМУ РАСПРЕДЕЛЕНИЮ ПО КРИТЕРИЮ ПИРСОНА

Проверяется гипотеза H_0: выборка принадлежит генеральной совокупности с функцией распределения $F(x)$.

Пусть объем выборки равен n. Для проверки разобъем размах выборки на l разрядов (интервалов): $\Delta_1, \Delta_2, ..., \Delta_l$. Обозначим через n_i – количество элементов выборки, попавших в интервал Δ_i. Заметим, что рекомендуется, чтобы в каждом интервале было не

менее 5 элементов. Пусть p_i – найденные по $F(x)$ вероятности попадания в интервал Δ_i. Составим статистику

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}.$$

Можно доказать, что при $n \to \infty$ закон распределения этой статистики стремится к закону распределения Пирсона χ^2 с $l-1$ степенью свободы (см. разд. 7). Это означает, что при достаточно больших значениях n можно считать, что построенная статистика подчиняется закону распределения Пирсона с $l-1$ степенью свободы, т.е.

$$P(\chi^2 < \chi_0^2) = F_{l-1}(\chi_0^2).$$

На практике критерий Пирсона используют при $n \geqslant 50$, $l \geqslant 5$ и $n_i \geqslant 5$.

В соответствии со сказанным проверка гипотезы производится так: по таблице квантилей распределения Пирсона по заданному уровню значимости α и числу степеней свободы $l-1$ найдем такой квантиль $\chi^2_{\alpha,l-1}$, для которого $F_{l-1}(\chi^2_{\alpha,l-1}) = 1-\alpha$. Затем сравним нашу статистику с этим квантилем. Если окажется, что $\chi^2 > \chi^2_{\alpha,l-1}$, то гипотеза H_0 отвергается.

Заметим, что выражение для χ^2 нетрудно преобразовать к виду

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{l} \frac{n_i^2}{p_i} - n.$$

Рекомендуем читателю показать это.

П р и м е р. Проверить гипотезу о том, что генеральная совокупность подчиняется равномерному распределению на промежутке $[0, 6]$, если результаты испытания таковы:

Δ_i	$0-1$	$1-2$	$2-3$	$3-4$	$4-5$	$5-6$
n_i	14	16	17	15	13	16

Зададим уровень значимости $\alpha = 0,1$. Так как $l-1 = 5$, то по таблице квантилей распределения Пирсона найдем $\chi^2_{0,1;5} = 7,3$. Теперь вычислим статистику χ^2. В соответствии с гипотезой очевидно, что для всех Δ_i должно быть $p_i = \dfrac{1}{6}$. Кроме того, $n = 91$.

Поэтому

$$\chi^2 = \frac{6}{91}(14^2 + 2 \cdot 16^2 + 17^2 + 15^2 + 13^2) - 91 = 0,7.$$

Так как $0,7 < 7,3$, то наша выборка не противоречит сделанной гипотезе.

20. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ВЫБОРКИ ЗАДАННОМУ РАСПРЕДЕЛЕНИЮ ПО КРИТЕРИЮ КОЛМОГОРОВА

Снова проверяется гипотеза H_0 : сделанная выборка принадлежит генеральной совокупности с функцией распределения $F(x)$.

Обозначим через $F_n(x)$ статистическую функцию распределения и положим

$$D_n = \max|F_n(x) - F(x)|.$$

Можно доказать, что если справедлива гипотеза H_0, то при достаточно больших значениях n величина $\sqrt{n}D_n$ практически подчиняется закону распределения Колмогорова (см. разд. 7), так что оказывается

$$P(\sqrt{n}D_n < x) \approx K(x).$$

Отсюда следует способ проверки нашей гипотезы. Задав уровень значимости α, найдем по таблице квантилей распределения Колмогорова такое x_α, для которого $K(x_\alpha) = 1 - \alpha$. Иначе говоря, найдем x_α такое, что

$$P(\sqrt{n}D_n > x_\alpha) = \alpha.$$

Если для сделанной выборки окажется $\sqrt{n}D_n > x_\alpha$, то гипотеза H_0 отвергается. В противном случае выборка не противоречит этой гипотезе.

Критерий Колмогорова обычно используют при больших объемах выборок. В таком случае, как уже говорилось, весь размах выборки разбивают на некоторое количество разрядов, а величины $|F_n(x)-F(x)|$ вычисляют лишь в серединах этих разрядов. При этом срединное значение учитывается столько раз, сколько элементов выборки попало в разряд. Заметим, что величины разрядов должны быть не очень большими. Специалисты рекомендуют при объемах

выборки больше сотни брать от 8 до 20 разрядов. При этом рекомендуется брать α равным $0,2$ или $0,3$.

Естественно, что разбиение на разряды приводит к погрешностям в нахождении D_n. Однако эти погрешности, как правило, невелики.

Этот критерий можно использовать только для непрерывных случайных величин. Кроме того, параметры закона распределения $F(x)$ должны быть известны заранее, а не из выборки.

П р и м е р. Результаты испытания приведены в таблице, в верхней строке которой указаны разряды, а в нижней – число элементов выборки в разряде.

$0-3$	$3-6$	$6-9$	$9-12$	$12-15$
2	14	26	58	78

$15-18$	$18-21$	$21-24$	$24-27$
54	28	16	4

Требуется проверить гипотезу о том, что генеральная совокупность подчиняется нормальному распределению с параметрами $m = 13$ и $\sigma = 5$.

Вначале найдем значение статистической функции распределения $F_{280}(x)$ в серединах разрядов. Так как $n = 280$, то

$$F_{280}(1,5) = \frac{2}{280} = 0,007; \quad F_{280}(4,5) = \frac{16}{280} = 0,057;$$
$$F_{280}(7,5) = \frac{42}{280} = 0,150; \quad F_{280}(10,5) = \frac{100}{280} = 0,357;$$
$$F_{280}(13,5) = \frac{178}{280} = 0,636; \quad F_{280}(16,5) = \frac{232}{280} = 0,829;$$
$$F_{280}(19,5) = \frac{260}{280} = 0,929; \quad F_{280}(22,5) = \frac{276}{280} = 0,986;$$
$$F_{280}(25,5) = 1.$$

Из формулировки гипотезы следует, что $F(x) = \Phi\left(\frac{x-13}{5}\right)$, где Φ – функция Лапласа. Поэтому

$$F(1,5) = \Phi(-2.3) = 0,011; \quad F(4,5) = \Phi(-1,7) = 0,045;$$
$$F(7,5) = \Phi(-1,1) = 0,136; \quad F(10,5) = \Phi(-0,5) = 0,308;$$
$$F(13,5) = \Phi(0,1) = 0,540; \quad F(16,5) = \Phi(0,7) = 0,758;$$
$$F(19,5) = \Phi(1,3) = 0,908; \quad F(22,5) = \Phi(1.9) = 0,971;$$
$$F(25,5) = \Phi(2,5) = 0,994.$$

В соответствии с найденными величинами оказывается, что

$$D_{280} = \max|F_{280}(x) - F(x)| = |F_{280}(13,5) - F(13,5)| =$$

$$= |0,636 - 0,540| = 0,096.$$

Отсюда

$$\sqrt{n}D_n = \sqrt{280} \cdot 0,096 = 1,61.$$

Зададим уровень значимости $\alpha = 0,2$. По таблице квантилей распределения Колмогорова найдем $x_{0,2} = 1,07$.

Так как $1,67 > 1,07$, то наша гипотеза должна быть отвергнута.

21. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ДВУХ ВЫБОРОК ОДНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ (РАНГОВЫЙ КРИТЕРИЙ УИЛКОКСОНА)

Имеются две случайные выборки $x_1.x_2, ..., x_m$ и $y_1, y_2, ..., y_n$, сделанные из генеральных совокупностей X и Y. Мы хотим проверить гипотезу H_0 о том, что обе выборки принадлежат одной генеральной совокупности, т.е. $X = Y$. При этом альтернативная гипотеза $\overline{H}_0$ состоит в том, что выборки принадлежат разным генеральным совокупностям, причем генеральная совокупность Y сдвинута в положительную сторону по отношению к X. В частности, это предположение означает, что $M(Y) > M(X)$.

Для этого расположим все элементы обеих выборки в порядке возрастания, т.е. составим из них единый вариационный ряд. Затем элементам построенного ряда присвоим номера. При этом если несколько рядом стоящих элементов окажутся одинаковыми, то просуммируем их номера, найдем среднее арифметическое суммы и каждому из них присвоим номер, равный среднему арифметическому. Например, если четыре элемента с номерами 7, 8, 9, 10 оказались одинаковыми, то каждому из них присваивается номер $\frac{7+8+9+10}{4} = 8,5$. Эти присвоенные номераназывают рангами элементов.

Очевидно, что ранг каждого элемента является случайной величиной, как и сам элемент. Если гипотеза H_0 справедлива, то элементы обеих выборок имеют одинаковые распределения. Поэтому для всех рангов вероятности должны быть одинаковыми. Так как номера принимают значения от 1 до $m+n$, то вероятность каждого ранга равна $\dfrac{1}{m+n}$. Теперь обозначим через r_i ранг элемента y_i и рассмотрим сумму

$$W = \sum_{i=1}^{n} r_i.$$

При справедливости гипотезы H_0 статистика W является случайной величиной, подчиненной закону распределения Уилкоксона. Таблицы этого распределения имеются в различных изданиях,в частности, в книге [4]. Заметим, что в этих таблицах приведены минимальные уровни значимости, т.е. наибольшие доверительные вероятности, соответствующие значениям m и n. Зная это, зададим уровень значимости α и найдем по таблице значение $W(\alpha,\ m,\ n)$. Если окажется $W < W(\alpha,\ m,\ n)$, то гипотеза H_0 отвергается.

Если альтернативной является гипотеза о том, что Y сдвинуто в отрицательную сторону по сравнению с X, то H_0 отвергается, если $W > n(m+n+1) - W(\alpha,\ m,\ n))$.

Когда альтернативная гипотеза заключается лишь в том, что X и Y имеют различные законы распределения , то задав α, по таблице найдем $W(\alpha/2,\ m,\ n)$. Затем строим интервал

$$(W(\alpha/2,\ m,\ n), n(n+m+1) - W(\alpha/2,\ m,\ n)).$$

Если найденное по выборке W не попадает в этот интервал, то гипотеза H_0 отвергается.

П р и м е р. Имеются две выборки:

$$x_i:\ 51,\ 39,\ 39,\ 59,\ 36,\ 58,\ 57,\ 45,\ 58$$

и

$$y_i:\ 35,\ 33,\ 54,\ 24,\ 42,\ 58,\ 47,\ 54.$$

Проверим гипотезу о принадлежности выборок одной генеральной совокупности при альтернативной гипотезе о том, что Y имеет отрицательный сдвиг относительно X.

Очевидно, что $m = 9$, $n = 8$. Для проверки гипотезы о том, что обе выборки принадлежат одной генеральной совокупности, зададим уровень значимости $\alpha = 0,1$ и по таблице распределения Уилкоксона найдем с учетом параметров выборки $W(0,1;9;8) = 86$. Отсюда

$$n(m+n+1) - W(\alpha,\ m,\ n) = 8 \cdot 17 - 86 = 50.$$

Теперь составим из всех выбранных элементов вариационный ряд, в котором для удобства элементы y_i отметим чертой сверху:

$$\overline{24},\ \overline{33},\ \overline{35},\ 36,\ 39,\ 39,\ \overline{42},\ 45,\ \overline{47},\ 51,\ \overline{54},\ \overline{54},\ 57,$$

$$\overline{58},\ 58,\ 58,\ 59.$$

Найдем сумму рангов для второй выборки (для элементов y_i):

$$W = 1 + 2 + 3 + 7 + 9 + 11,5 + 11,5 + 14 = 59.$$

Так как $59 > 50$, то гипотеза H_0 отвергается. Другими словами, с вероятностью $0,9$ можно утверждать, что генеральная совокупность Y имеет отрицательный сдвиг относительно X.

При $m, n > 25$ распределение Уилкоксона можно с достаточной точностью заменить нормальным распределением с

$$M(W) = \frac{n(m+n+1)}{2} \quad \text{и} \quad D(W) = \frac{mn(m+n+1)}{12}.$$

В заключение обратим внимание на то, что критерий Уилкоксона не требует знания законов распределения исследуемых генеральных совокупностей. Критерии такого типа называют непараметрическими.

22. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ МЕДИАН ДВУХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ (КРИТЕРИЙ ЗНАКОВЫХ РАНГОВ УИЛКОКСОНА)

Имеются две генеральные совокупности: X и Y. Обозначим их медианы соответственно $\vartheta(X)$ и $\vartheta(Y)$. Гипотеза $H_0 : \vartheta(X) = \vartheta(Y)$.

Для проверки этой гипотезы сделаем парные выборки из генеральных совокупностей (x_i, y_i), где $i = 1, ..., n$ и составим разности

$z_i = y_i - x_i$. Найдем модули этих разностей $|z_i|$. Эти модули расположим в порядке возрастания и в соответствии с этим каждому присвоим ранг, так же как в предыдущем разделе. Обозначим через R_i ранг $|z_i|$. Затем положим

$$\psi_i = \begin{cases} 1 \text{ при } z_i > 0, \\ 0 \text{ при } z_i < 0 \end{cases}$$

и построим статистику

$$T = \sum_{i=1}^{n} \psi_i R_i.$$

Случайная величина подчиняется закону распределения, введенному Уилкоксоном. Таблицы этого распределения имеются, например, в [5] (таблица А4).

Используя таблицу, по заданному уровню значимости α найдем $t(\alpha, n)$, для которого $P(T < t(\alpha, n)) = 1 - \alpha$.

Если альтернативная гипотеза состоит в том, что $\vartheta(Y) > \vartheta(X)$, то при $T > t(\alpha, n)$ гипотеза H_0 отвергается.

При альтернативной гипотезе $\vartheta(Y) < \vartheta(X)$ нужно отвергнуть H_0 , если $T < \dfrac{n(n+1)}{2} - t(\alpha, n)$.

Когда альтернативной является гипотеза $\vartheta(Y) \neq \vartheta(X)$ находим $t(\alpha/2, n)$. В этом случае гипотеза H_0 отвергается, если $T > t(\alpha/2, n)$ или $T < \dfrac{n(n+1)}{2} - t(\alpha/2, n)$.

П р и м е р. Испытание дало такие результаты

X	24	33	35	32	25	40	47	36
Y	29	35	36	38	24	38	48	40

Проверим гипотезу $\vartheta(X) = \vartheta(Y)$ при альтернативной гипотезе $\vartheta(Y) > \vartheta(X)$.

Для проверки зададим уровень значимости $\alpha = 0,1$ и найдем $t(0,1; 8) = 28$. Мы использовали таблицу А4 из книги [4].

Составим разности:

$z_1 = 5$, $z_2 = 2$, $z_3 = 1$, $z_4 = 6$, $z_5 = -1$, $z_6 = -2$, $z_7 = 1$, $z_8 = 4$.

Так как $|z_3| = |z_5| = |z_7| = 1$, $|z_2| = |z_6| = 2$, $|z_8| = 4, |z_1| = = 5, |z_4| = 6|$, то $R_3 = R_5 = R_7 = 2$; $R_2 = R_6 = 4,5;$, $R_8 = 6$, $R_1 = 7$, $R_4 = 8$.

В нашем случае $\psi_1 = \psi_2 = \psi_3 = \psi_4 = \psi_7 = \psi_8 = 1$, а $\psi_5 = \psi_6 = = 0$. Поэтому

$$T = \sum_{i=1}^{8} \psi_i R_i = 7 + 4,5 + 2 + 8 + 2 + 6 = 29,5.$$

Так как $29,5 > 28$, то гипотезу о равенстве медиан следует отвергнуть.

23. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

В приложениях нередко возникает такой вопрос: влияет ли некоторый фактор x на величину y. Например, влияет ли на твердость сплава добавление в него некоторого элемента. Для ответа на поставленный вопрос с помощью статистики задается несколько различных уровней фактора x и при каждом из этих уровней производится несколько измерений значения y. Естественно, что измерения не производятся абсолютно точно. Поэтому их можно считать реализациями некоторой случайной величины Y, для которой $y = M(Y)$. При этом, если фактор x не влияет на y испытания проводятся в одинаковых условиях, то случайная величина Y при всех его уровнях не изменяется. Мы приведем один из возможных способов обработки наблюдений, позволяющий ответить на поставленный вопрос. Этот способ называют **дисперсионным анализом**.

Пусть гипотеза H_0 утверждает, что фактор x не влияет на значение величины y. В этом случае случайная величина Y не зависит от уровня x. В частности от x не зависят ее математическое ожидание $M(Y) = a$ и дисперсия $D(Y) = \sigma^2$.

Обозначим через $x_1, ..., x_m$ заданные уровни фактора x. На каждом уровне получено несколько значений y. Результаты приведены в таблице.

	Значения y
x_1	$y_{11},\ y_{21},\ ...,\ y_{n_1 1}$
x_2	$y_{12},\ y_{22},\ ...,\ y_{n_2 2}$
...	.....
x_m	$y_{1m},\ y_{2m},\ ...,\ y_{n_m m}$

Введем такие обозначения:

$$n = n_1 + n_2 + ... + n_m,$$

$$\overline{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \quad (j = 1, 2, ...m),$$

$$\overline{y} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} y_{ij}.$$

Найдем полную сумму квадратов отклонений

$$Q = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2.$$

Преобразуем эту сумму:

$$Q = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j + \overline{y}_j - \overline{y})^2 =$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)^2 + \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\overline{y}_j - \overline{y})^2 + 2 \sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)(\overline{y}_j - \overline{y}).$$

Последняя двойная сумма равна нулю. Действительно,

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j)(\overline{y}_j - \overline{y}) = \sum_{j=1}^{m} (\overline{y}_j - \overline{y}) \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_j) =$$

$$= \sum_{j=1}^{m} (\overline{y}_j - \overline{y}) \left(\sum_{i=1}^{n_j} y_{ij} - n_j \overline{y}_j \right).$$

Из определения величины $\overline{y}_j$ видно, что

$$\sum_{i=1}^{n_j} y_{ij} - n_j \overline{y}_j = 0.$$

Значит, вся рассматриваемая двойная сумма равна нулю.

Следовательно,

$$Q = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})^2.$$

Положим

$$Q_1 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2, \quad Q_2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(\overline{y}_j - \overline{y})^2.$$

Величина Q_1 представляет собой сумму квадратов отклонений внутри групп. Ее называют *остаточным рассеиванием.*

Величина Q_2 есть сумма квадратов отклонений между группами. Ее называют *рассеиванием по уровням фактора.*

Суммы Q_1 и Q_2 нетрудно преобразовать так:

$$Q_1 = \sum_{j=1}^{m}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{m} n_j \overline{y}_j^2,$$

$$Q_2 = \sum_{j=1}^{m} n_j \overline{y}_j^2 - n\overline{y}^2.$$

Рекомендуем читателю сделать преобразования самостоятельно.

Теперь найдем $M(Q_1)$ и $M(Q_2)$, предположив, что гипотеза H_0 справедлива, т.е. фактор X не влияет на значения y. В этом случае должно быть

$$M(y_{ij}) = a, \quad D(y_{ij}) = \sigma^2, \quad M(y_{ij}^2) = \sigma^2 + a^2,$$

$$M(\overline{y}_j) = a, \quad D(\overline{y}_j) = \frac{\sigma^2}{n_j}, \quad M(\overline{y}_j^2) = \frac{\sigma^2}{n_j} + a^2,$$

$$M(\overline{y}) = a, \quad D(\overline{y}) = \frac{\sigma^2}{n}, \quad M(\overline{y}^2) = \frac{\sigma^2}{n} + a^2.$$

Следовательно,

$$M(Q_1) = \sum_{j=1}^{m}\sum_{i=1}^{n_j} M(y_{ij}^2) - \sum_{j=1}^{m} n_j M(\overline{y}_j^2) =$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{n_j} (\sigma^2 + a^2) - \sum_{j=1}^{m} n_j \left(\frac{\sigma^2}{n_j} + a^2 \right) =$$

$$= n(\sigma^2 + a^2) - (m\sigma^2 + na^2) = (n-m)\sigma^2,$$

$$M(Q_2) = \sum_{j=1}^{m} n_j M(\overline{y}_j^2) - nM(\overline{x}^2) =$$

$$= \sum_{j=1}^{m} n_j \left(\frac{\sigma^2}{n_j} + a^2 \right) - n \left(\frac{\sigma^2}{n} + a^2 \right) =$$

$$= m\sigma^2 + na^2 - n \left(\frac{\sigma^2}{n} + a^2 \right) = (m-1)\sigma^2.$$

Из полученного следует:

$$M \left(\frac{Q_1}{n-m} \right) = M \left(\frac{Q_2}{m-1} \right) = \sigma^2.$$

Таким образом, если справедлива гипотеза H_0, то величины $\dfrac{Q_1}{n-m}$ и $\dfrac{Q_2}{m-1}$ можно рассматривать как две оценки одной и той же дисперсии. Поэтому дробь $F = \dfrac{(n-m)Q_2}{(m-1)Q_1}$ должна подчиняться закону распределения Фишера со степенями свободы $n-m$ и $m-1$.

Из сказанного вытекает правило проверки гипотезы о независимости y от уровней фактора x: задав уровень значимости α, зная $m-1$ и $n-m$, по таблице квантилей распределения Фишера находим такое $F_\alpha(m-1, n-m)$, для которого

$$P(F < F_\alpha(m-1, n-m)) = 1 - \alpha.$$

Если окажется $F = \dfrac{(n-m)Q_2}{(m-1)Q_1} > F_\alpha(m-1, n-m)$, то гипотеза о независимости отвергается. Другими словами, в этом случае считают, что фактор x значимо влияет на y.

Заметим, что наши выводы опирались на два предположения:

1) на всех уровнях фактора случайная величина Y не меняется;

2) случайная величина Y имеет нормальное распределение (мы использовали распределение Фишера, основанное на нормальном).

Укажем на важное обстоятельство. Если $\dfrac{Q_1}{n-m} > \dfrac{Q_2}{m-1}$, т.е. разброс между уровнями меньше, чем разброс внутри уровней, то это означает, что фактор не влияет значимо на y. Следовательно, подробный дисперсионный анализ проводить не нужно.

П р и м е р. При четырех различных значениях фактора x получены значения y, приведенные ниже в таблице.

	значения y
x_1	60, 61, 65, 68, 70, 72, 80
x_2	58, 64, 64, 70, 75
x_3	46, 54, 60, 62, 64, 66, 74, 82
x_4	51, 52, 53, 55, 60, 68

Проверим, зависит ли y от фактора x.
Ясно, что в этом случае

$$m = 4,\ n_1 = 7,\ n_2 = 5,\ n_3 = 8,\ n_4 = 6,\ n = 26.$$

Значит, $m - 1 = 3$ и $n - m = 22$. Пусть уровень значимости $\alpha = 0,05$. По таблице квантилей распределения Фишера найдем $F_{0,1}(3, 22) = 3,05$. Теперь последовательно получаем

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{4}\sum_{i=1}^{n_j} y_{ij} = \frac{60 + 61 + ... + 68}{26} = \frac{1654}{26} = 63,6;$$

$$\sum_{j=1}^{4}\sum_{i=1}^{n_j} y_{ij}^2 = 60^2 + 61^2 + ... + 68^2 = 107226;$$

$$\overline{y}_1 = 68,\ \overline{y}_2 = 66,2,\ \overline{y}_3 = 63,5,\ \overline{y}_4 = 56,5;$$

$$\sum_{j=1}^{4} n_j \overline{y}_j^2 = 7 \cdot 68^2 + 5 \cdot 66,2^2 + 8 \cdot 63,5^2 + 6 \cdot 56,5^2 =$$

$$= 105691,7.$$

$$Q_1 = \sum_{j=1}^{4}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{4} n_j \overline{y}_j^2 = 107226 - 105691,7 = 1534,3;$$

$$Q_2 = \sum_{j=1}^{4}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{4} n_j \overline{y}_j^2 = 105691,7 - 26 \cdot 63,6^2 = 522,7$$

$$F = \frac{(n-m)Q_2}{(m-1)Q_1} = \frac{22 \cdot 522,7}{3 \cdot 1534,3} = 2,50.$$

Так как $2,50 < 3,05$, то с вероятностью 0,95 можно утверждать, что наши данные не противоречат гипотезе о независимости y от x. При этом величину $\overline{y} = 63,6$ можно принять за оценку истинного значения y.

Обратим внимание на такой факт. Если бы мы задали уровень значимости $\alpha = 0,1$, то по таблице нашли бы $F_{0,1}(3,22) = 2,35$. Так как $2,50 > 2,35$, то на этом уровне значимости ответ на поставленный вопрос оказался бы противоположным.

Этот пример весьма поучителен. Он говорит о том, что к статистическим выводам нельзя относиться как к достоверным. Любой такой вывод справедлив лишь с некоторой вероятностью.

Аналогичным, но гораздо более громоздким, способом проверяется зависимость величины от нескольких факторов.

24. ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Требуется выяснить влияние двух факторов x и y на величину z. Для этого придаем факторам значения x_1, $x_2, ..., x_n$ и y_1, $y_2, ..., y_m$ и для каждой пары значений (x_i, y_j) проводим испытания, в результате которых получаем такие значения z_{ij1} ,..., $z_{ijk_{ij}}$.

Введем такие обозначения:

$$N = \sum_{i=1}^{n} \sum_{j=1}^{m} k_{ij},$$

$$\overline{z} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{\nu=1}^{k_{ij}} z_{ij\nu}, \qquad \overline{z}_{ij} = \sum_{\nu=1}^{k_{ij}} z_{ij\nu},$$

$$\overline{z}_{\cdot j} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{\nu=1}^{k_{ij}} z_{ij\nu}}{\sum\limits_{i=1}^{n} k_{ij}}, \qquad \overline{z}_{i\cdot} = \frac{\sum\limits_{j=1}^{m} \sum\limits_{\nu=1}^{k_{ij}} z_{ij\nu}}{\sum\limits_{j=1}^{m} k_{ij}}.$$

Будем опять рассматривать полную сумму квадратов

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\nu=1}^{k_{ij}}(z_{ij\nu} - \overline{z})^2.$$

Аналогично тому, как было сделано выше, эту сумму можно преобразовать к такому виду

$$Q = Q_1 + Q_2 + Q_3 + Q_4,$$

где

$$Q_1 = \sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}(\overline{z}_{ij} - \overline{z}_{\cdot j} - \overline{z}_{i\cdot} + \overline{z})^2,$$

$$Q_2 = \sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}(\overline{z}_{i\cdot} - \overline{z})^2, \qquad Q_3 = \sum_{i=1}^{n}\sum_{j=1}^{m} k_{ij}(\overline{z}_{\cdot j} - \overline{z})^2,$$

$$Q_4 = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\nu=1}^{k_{ij}}(z_{ij\nu} - \overline{z}_{ij})^2.$$

Здесь Q_1 характеризует рассеяние от взаимодействия факторов x и y, Q_2 и Q_3 – рассеяние от влияния факторов x и y, Q_4 — рассеяние внутри комбинаций уровней.

Если справедлива гипотеза о том, что факторы не влияют на z, то можно показать, что величины

$$\frac{Q_1}{(n-1)(m-1)}, \quad \frac{Q_4}{N-nm}, \quad \frac{Q_2}{n-1}, \quad \frac{Q_3}{m-1}$$

являются несмещенными оценками генеральной дисперсии σ^2 при справедливости гипотезы об отсутствии влияния факторов на z.

В таком случае отношение

$$F_0 = \frac{(N-nm)Q_1}{(n-1)(m-1)Q_4}$$

должно подчиняться распределению Фишера с числом степеней свободы $(n-1)(m-1)$ и $N-nm$.

Распределению Фишера, очевидно, должны также подчиняться отношения

$$F_1 = \frac{(N-nm)Q_2}{(n-1)Q_4}, \quad F_2 = \frac{(N-nm)Q_3}{(m-1)Q_4},$$

причем F_1 имеет $n-1$ и $N-nm$ степеней свободы, а F_2 – $m-1$ и $N-nm$.

В соответствии со сказанным проверку гипотезы производят так: задав малый уровень значимости α, находят по таблицам такое значение $F_{\alpha,\ (n-1)(m-1),\ N-nm}$, для которого

$$P(F_0 < F_{\alpha,\ (n-1)(m-1),\ N-nm}) = 1-\alpha.$$

Если окажется $F_0 < F_{\alpha,\ (n-1)(m-1),\ N-nm}$, то это это означает, что взаимодействие факторов незначимо влияет на значение величины z. Другими словами, влиянием взаимодействия факторов можно пренебречь.

В этом случае далее следует проверить влияние каждого из факторов x и y в отдельности по критерию Фишера. Для этого, задав α, находим по таблицам $F_{\alpha,\ n-1,\ N-nm}$ и $F_{\alpha,\ m-1,\ N-nm}$ и сравниваем с ними F_1 и F_2 соответственно. Если при этом окажется, например, что $F_1 > F_{\alpha,\ n-1,\ N-nm}$, то делается вывод о значимом влиянии фактора x на z и т.д.

Если же $F_0 > F_{\alpha,\ (n-1)(m-1),\ N-nm}$, то делается вывод о значимом влиянии взаимодействия факторов на z. В таком случае обычными методами дисперсионного анализа невозможно выяснить, каково влияние на z каждого фактора в отдельности. Нужно либо модифицировать дисперсионный анализ, либо менять модель исследования.

25. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Одной из существенных задач статистики является нахождение функциональных связей между величинами. Математически эта задача ставится так:

В результате испытаний установлено (например, с помощью дисперсионного анализа), что величина y зависит от факторов x_1, x_2 ,..., x_m. Требуется по результатам испытания подобрать функцию

$y = \psi(x_1, x_2, ..., x_m)$, которая лучше других описывает связь y с факторами.

Обычно считается, что значения факторов x_1, x_2,..., x_m измеряются с ошибкой, которая мала по сравнению с ошибкой при измерении y. Иначе говоря, значения факторов считают практически не случайными, а найденные значения функции представляются реализациями случайной величины Y, для которой $y = M(Y)$. Поэтому задача сводится к подбору функции $y = M(Y) = \psi(x_1, x_2, ..., x_m)$. Такая функция называется *функцией регрессии*.

Как правило, вид функции ψ заранее не известен, его подбирают, анализируя набор полученных результатов. Чаще всего функцию регрессии ищут в виде многочлена от аргументов x_1, x_2,..., x_m. В этом случае получаются наиболее простые уравнения для нахождения параметров функции. Они являются коэффициентами многочлена. Найденные коэффициенты многочлена называют коэффициентами регрессии. Эти коэффициенты обычно находят методом наименьших квадратов.

Начнем изложение этого метода с самого простого случая. Будем считать, что y зависит лишь от одного фактора x и что функция регрессии является многочленом первой степени.

Предположим, что для значений $x_1, x_2, ..., x_n$ фактора x были найдены соответствующие значения $y_1, y_2, ..., y_n$. При этом на графике пары точек (x_k, y_k) $(k = 1, ..., n)$ оказываются близкими к прямой линии. Поэтому мы ищем зависимость между величинами в виде линейной регрессии

$$y = M(Y) = b_0 + b_1 x.$$

Наша задача найти такие b_0 и b_1, которые наилучшим образом описывают зависимость.

За меру отличия полученных значений y_k $(k = 1, ..., n)$ от вычисленных $b_0 + b_1 x_k$ примем сумму квадратов разностей

$$\delta = \sum_{k=1}^{n} (b_0 + b_1 x_k - y_k)^2.$$

При такой мере отличия наиболее точно зависимость будет описывать прямая с такими значениями b_0 и b_1, для которых δ принимает наименьшее значение.

Из дифференциального исчисления мы знаем, что наименьшее значение δ достигается при необходимых условиях:

$$\frac{\partial \delta}{\partial b_0} = 0, \quad \frac{\partial \delta}{\partial b_1} = 0.$$

Последние равенства в развернутом виде оказываются такими:

$$2\sum_{k=1}^{n}(b_0 + b_1 x_k - y_k) = 0,$$
$$2\sum_{k=1}^{n}(b_0 + b_1 x_k - y_k)x_k = 0.$$

Переписав эти уравнения:

$$b_0 n + b_1 \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k,$$
$$b_0 \sum_{k=1}^{n} x_k + b_1 \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k,$$

найдем из них неизвестные нам коэффициенты. Естественно, что найденные значения являются лишь статистическими оценками истинных значений коэффициентов линейной зависимости. Обозначив эти оценки соответственно $\hat{b}_0$ и $\hat{b}_1$, получим из уравнений

$$\hat{b}_0 = \frac{\sum_{k=1}^{n} y_k \sum_{k=1}^{n} x_k^2 - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2},$$
$$\hat{b}_1 = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2}.$$

Для сокращения записи введем обозначения:

$$[x] = \sum_{k=1}^{n} x_k, \quad [y] = \sum_{k=1}^{n} y_k, \quad [xy] = \sum_{k=1}^{n} x_k y_k.$$

Аналогичные обозначения будем применять и для других сумм такого же типа. В этих обозначениях формулы для найденных коэффициентов принимают вид

$$\hat{b}_0 = \frac{[y][x^2] - [xy][x]}{n[x^2] - [x]^2}, \quad \hat{b}_1 = \frac{n[xy] - [x][y]}{n[x^2] - [x]^2}.$$

Можно формально доказать, что найденные коэффициенты доставляют именно минимум (а не максимум) величине δ. Не будем этого делать.

Таким образом, мы получили приближенное уравнение, связывающее y и x:

$$y = \frac{[y][x^2] - [xy][x]}{n[x^2] - [x]^2} + \frac{n[xy] - [x][y]}{n[x^2] - [x]^2}\ x.$$

Если при значении фактора x_i получено несколько различных значений y_{i1}, y_{12},..., y_{im}, то для упрощения вычислений в δ введем их среднее арифметическое

$$\overline{y}_i = \frac{y_{i1} + y_{i2} + ... + y_{im}}{m}.$$

П р и м е р. Результаты испытания:

x_k	9	15	20	26	30	35	40
y_k	6,3	7,5	9,8	10,6	12,0	14,3	15,1

Построим на графике полученные точки. Чтобы график не занимал слишком много места, возьмем разные масштабы на осях координат (рис.6). Видим, что построенные точки близки к некоторой прямой. Поэтому ищем коэффициенты линейной регрессии.

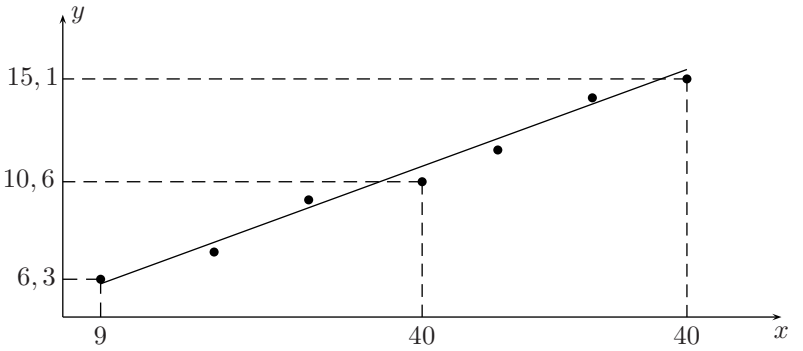

Рис. 6. Линейная регрессия

Нетрудно подсчитать, что

$$n = 7; \quad [x] = 175; \quad [y] = 75,6$$

$$[x^2] = 5107; \quad [xy] = 2106,3.$$

Следовательно, в соответствии с написанными выше формулами,

$$\hat{b}_0 = \frac{75,6 \cdot 5107 - 2106,3 \cdot 175}{7 \cdot 5107 - 175^2} = 3,41;$$

$$\hat{b}_1 = \frac{7 \cdot 2106,3 - 175 \cdot 75,6}{7 \cdot 5107 - 175^2} = 0,29.$$

Таким образом,

$$y = 3,41 + 0,30x$$

или, если округлить значения коэффициентов,

$$y = 3,4 + 0,3x.$$

Заметим, что такое округление имеет смысл, поскольку результат получен по небольшому числу данных. Соответствующая этому уравнению прямая приведена на рис. 6.

Для уточнения зависимостей между величинами можно использовать многочлены более высоких степеней. В частности, можно использовать параболическую регрессию, при которой зависимость описывается квадратичной функцией:

$$y = c_0 + c_1 x + c_2 x^2.$$

В этом случае

$$\delta = \sum_{k=1}^{n} (y_k - c_0 - c_1 x_k - c_2 x_k^2)^2.$$

Коэффициенты c_0, c_1 и c_2, минимизирующие δ, находятся из условий

$$\frac{\partial \delta}{\partial c_0} = 0, \quad \frac{\partial \delta}{\partial c_1} = 0, \quad \frac{\partial c_2}{\partial c_2} = 0.$$

Эти условия, как нетрудно показать, приводят к системе линейных уравнений:

$$\begin{aligned} &nc_0 + [x]c_1 + [x^2]c_2 = [y], \\ &[x]c_0 + [x^2]c_1 + [x^3]c_2 = [yx], \\ &[x^2]c_0 + [x^3]c_1 + [x^4]c_2 = [yx^2]. \end{aligned}$$

Аналогичным образом строится функции регрессии и тогда, когда Y зависит от нескольких факторов. Действительно, например, мы хотим описать зависимость вида $z = C_0 + c_1 x + c_2 y$. В этом случае, имея выборку вида (z_k, x_k, y_k), где $k = 1, 2, ..., n$, ищем минимум величины

$$\delta = \sum_{k=1}^{n} (z_k - b_0 - b_1 x - b_2 y)^2.$$

В результате получим систему

$$\begin{aligned} &nc_0 + [x]c_1 + [y]c_2 = [z], \\ &[x]c_0 + [x^2]c_1 + [xy]c_2 = [xz], \\ &[y]c_0 + [yx]c_1 + [y^2]c_2 = [yz], \end{aligned}$$

из которой найдем оценки коэффициентов.

Заметим, что более сложные зависимости между y и x нередко можно свести к линейным. Например, в случае $y = ae^{bx}$, логарифмируя, получаем линейную зависимость между $\ln y$ и x: $\ln y = \ln a + bx$.

Добавим к сказанному следующее. После того,как система решена, т.е. найдены оценки коэффициентов регрессии $\hat{b}_0$, $\hat{b}_1$,..., производится, если это возможно, статистический анализ результатов. Этот анализ состоит в следующем:

1. Находятся оценки дисперсий величин $\hat{b}_i$. Это позволяет оценить ошибки при определении коэффициентов.

2. Находятся корреляции найденных коэффициентов друг с другом, описывающие зависимость коэффициентов между собой.

3. Проверяется статистическая значимость коэффициентов, т.е. с некоторой доверительной вероятностью устанавливается, что найденные коэффициенты меньше или больше, чем ошибки в их определении.

4. Проверяется адекватность уравнения регрессии реальному процессу.

К сожалению, единой методики проведения такого анализа не существует.

Рассмотренный метод построения регрессионного уравнения имеет ряд существенных недостатков.

Прежде всего, метод весьма трудоемок, особенно,при большом количестве факторов. Далее, оценки коэффициентов оказываются статистически зависимыми (их коэффициенты корреляции обычно отличны от нуля). Наконец, значение каждой оценки существенно зависит от числа членов функции регрессии. Действительно, добавление одного слагаемого в функцию регрессии приводит к существенному изменению системы уравнений.

Библиографический список

1. *Шапорев С.Д.* Прикладная статистика. Учебное пособие. С.-Пб.: БГТУ, 2003.
2. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики для технических приложений. М.: Наука, 1969.
3. *Вентцель Е.С., Овчаров Л.А.* Теория вероятностей и ее инженерные приложения. М.: Наука, 1988.
4. *Пустыльник Е.И.* Статистичекские методы анализа и обработки наблюдений. М.: Наука, 1968.
5. *Холландер М., Вулф Д.* Непараметрические методы статистики. М.,Финансы и статистика, 1983.
6. *Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: Высшая школа, 2002.

О Г Л А В Л Е Н И Е

1. ЗАДАЧИ СТАТИСТИКИ 3

2. НЕКОТОРЫЕ ПОЛЕЗНЫЕ ЗАКОНЫ РАСПРЕДЕЛЕНИЯ 5

3. ПОСТРОЕНИЕ ЭМПИРИЧЕСКОЙ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ 10

4. СТАТИСТИЧЕСКАЯ ОЦЕНКА ПАРАМЕТРОВ СЛУЧАЙНОЙ ВЕЛИЧИНЫ 13

5. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ ОЦЕНОК ПО МЕТОДУ МОМЕНТОВ 14

6. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ ОЦЕНОК ПО МЕТОДУ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ 17

7. ТОЧЕЧНАЯ ОЦЕНКА СРЕДНЕГО ЗНАЧЕНИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ 20

8. ТОЧЕЧНАЯ ОЦЕНКА ДИСПЕРСИИ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ 22

9. ТОЧЕЧНАЯ ОЦЕНКА МЕДИАНЫ 23

10. ТОЧЕЧНАЯ ОЦЕНКА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ 24

11. ИНТЕРВАЛЬНАЯ ОЦЕНКА ВЕРОЯТНОСТИ СОБЫТИЯ ПО ЕГО ЧАСТОТЕ 26

12. ИНТЕРВАЛЬНАЯ ОЦЕНКА СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ 28

13. ИНТЕРВАЛЬНАЯ ОЦЕНКА ДИСПЕРСИИ НОРМАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ 30

14. ИНТЕРВАЛЬНАЯ ОЦЕНКА ПАРАМЕТРА ЭКСПОНЕНЦИАЛЬНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ 31

15. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ 32

16. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ СРЕДНИХ ЗНАЧЕНИЙ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ 34

17. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ ДИСПЕРСИЙ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ 36

18. ПРОВЕРКА ГИПОТЕЗЫ ОБ ОДНОРОДНОСТИ НАБЛЮДЕНИЙ (ОТСУТСТВИИ ГРУБЫХ ОШИБОК) 38

19. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ВЫБОРКИ ЗАДАННОМУ РАСПРЕДЕЛЕНИЮ ПО КРИТЕРИЮ ПИРСОНА 39

20. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ВЫБОРКИ ЗАДАННОМУ РАСПРЕДЕЛЕНИЮ ПО КРИТЕРИЮ КОЛМОГОРОВА 41

21. ПРОВЕРКА ГИПОТЕЗЫ О ПРИНАДЛЕЖНОСТИ ДВУХ ВЫБОРОК ОДНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ (РАНГОВЫЙ КРИТЕРИЙ УИЛКОКСОНА) 43

22. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ МЕДИАН ДВУХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ (КРИТЕРИЙ ЗНАКОВЫХ РАНГОВ УИЛКОКСОНА) 45

23. ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ 47

24. ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ 52

25. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ 54

Библиографический список 61

Файншмидт Виктор Лейбович

Элементы математической статистики

Редактор *Г.М. Звягина*
Корректор *Л.А. Петрова*
Компьютерный набор и верстка *В.Л. Файншмидта* и *Н.В. Тарасовой*
Подписано в печать 21.11.2017. Формат 60x84/16. Бумага документная.
Печать трафаретная. Усл. печ. л. 3.725. Тираж 350 экз. Заказ № 170
Балтийский государственный технический университет
Типография БГТУ
190005, С.-Петербург, 1-я Красноармейская ул., д. 1