Учреждение образования
“ГОМЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
ИМЕНИ ФРАНЦИСКА СКОРИНЫ”
Кафедра теоретической физики

Статистические методы обработки данных

Конспект Лекций

Специальность

1-31 04 08 Компьютерная физика

Лекции: 30 часов

Практические занятия: –

Лабораторные занятия: 34 часа

Материал подготовил

Андреев

Виктор Васильевич

доктор физ.-мат. наук, доцент

Гомель, 2018

ОГЛАВЛЕНИЕ

1 Основные понятия теории вероятности и мат. статистики

Для исследования свойств объекта проводят измерения, позволяющие количественно дать характеристики свойств этого объекта. На практике производится ограниченное число измерений n при одинаковых условиях. В результате получаем множество событий (значений) исследуемой величины X: $\{x_1, x_2, \dots x_n\}$, которое представляет собой выборку объема n. Таким образом, возможные исходы некоторого эксперимента (измерения) представляют собой множество точек, которые можно сочетать разными способами. **Такие сочетания называют событиями.**

1.1 Частота попадания случайной величины.

В том случае,(т.е. когда одна и та же характеристика объекта принимает различные значения) говорят о том, что величина X является **случайной величиной**. Случайная величина -это действительное число (или набор действительных чисел), которое заключено между $-\infty$ и $+\infty$, которое сопоставляется каждой возможной точке из числа значений этой характеристики. При соответствующих условиях для каждое событие можно характеризовать **частотой появления** ν_n.

***Определение* 1.1**

Частотой появления ν_n события A называется отношение числа m появления данного события к общему числу проведенных одинаковых испытаний, в каждом из которых могло появиться или не появиться данное событие:

$$\nu_n = \frac{m}{n} \, . \tag{1.1}$$

Определение 1.2

Если число испытаний n велико, то, как правило, частоты появления данного события A в различных сериях измерений отличаются мало друг от друга. Это утверждение записывают следующим образом:

$$\lim_{n\to\infty} \nu_n = p \; . \tag{1.2}$$

Число p называются вероятностью (англ.-probability) случайного события A.

Отметим, что существуют такие события у которых частота появления может сильно отличаться от вероятности, даже при большом числе испытаний.

Как видим из вышеприведенного примера, для описания случайного поведения величины необходима совокупность, содержащая неограниченное число значений измеряемой величины ($n = \infty$). Такая выборка называется генеральной совокупностью. Генеральная совокупность часто используется как важное абстрактное понятие, необходимое в теоретических расчетах, связанных с исследованием поведения физической величины как случайной величины.

Различают два основных типа случайных величин: дискретные случайные величины и непрерывные случайные величины. Если величина X имеет **конечное** число (счетное множество) из последовательности возможных значений $\{x_1, x_2, \ldots x_k, \ldots, \}$ то такая величина называется *дискретной случайной величиной*. Если случайная величина может принимать любое значение из интервала возможных значений, то такая величина называется *непрерывной случайной величиной*.

1.2 Интегральный закон распределения вероятности.

Для характеристики частоты появления различных значений случайной величины X теория вероятностей предлагает пользоваться указанием **закона распределения вероятностей** различных значений этой величины.

При этом различают два вида описания законов распределения:

1. **интегральный закон распределения вероятности**;
2. **дифференциальный закон распределения вероятности**.

***Определение* 1.3**

***Интегральным законом**, или функцией распределения вероятностей $F(x)$ случайной величины X, называют функцию, значения которой представляют вероятность того, что значения x_k случайной величины X меньше некоторого значения x (x – некоторое произвольное число).*

Данное утверждение символически записывается в виде

$$F(x) = \mathrm{Prob}\,(x_k < x)\ , \tag{1.3}$$

где $\mathrm{Prob}\,(x_k < x)$ и представляет собой вероятность события в вышеприведенном определении.

Очевидно, что

$$F\,(a) \le F\,(b)\,, \quad \text{при} \quad a \le b \quad \text{(неубывающая функция)} \tag{1.4}$$

$$F\,(-\infty) = 0\ , \ \ F\,(\infty) = 1\ . \tag{1.5}$$

Для дискретной одномерной случайной величины X закон распределения удобно представить в виде таблицы

$$X = \begin{pmatrix} x_1, x_2, \ldots, x_n \\ p_1, p_2, \ldots, p_n \end{pmatrix} \tag{1.6}$$

где $x_1, x_2, \ldots, x_n$ – значения случайной величины X, а $p_1, p_2, \ldots, p_n$ – вероятности появления этих значений.

Другими словами $\mathrm{Prob}\,(X = x_i) = p_i$. В этом случае интегральный закон вероятности в соответствии с законом распределения вероятности имеет вид:

$$F(x) = \mathrm{Prob}\,(x_k < x) = \sum_{i=1}^{k} p_i\ . \tag{1.7}$$

1.3 Дифференциальный закон распределения вероятности.

Для случайной величины с непрерывной и дифференцируемой функцией распределения $F\,(x)$ можно найти **дифференциальный закон распределения вероятностей**:

$$p(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}\ . \tag{1.8}$$

$p(x)$ называют кривой плотности распределения вероятностей (или просто **плотность вероятности**.)

Свойства $p\,(x)$:

1. $p(x) \geqslant 0$

2. Из (1.8) следует, что

$$F(x) = \int\limits_{-\infty}^{x} p(\xi)\mathrm{d}\xi \ . \tag{1.9}$$

3. Условие нормировки:

$$\int\limits_{-\infty}^{\infty} p\,(x)\,\mathrm{d}x = 1 \ . \tag{1.10}$$

Для дискретной случайной величины из определения (1.6) следует, что

$$p(x) = \sum_{i=1}^{n} \delta\,(x - x_i)\,p_i \ , \tag{1.11}$$

где одномерная δ-функция Дирака определена соотношениями:

$$\delta\,(x) = \begin{cases} +\infty \ , \text{если } x = 0 \\ 0 \ , \text{если } x \neq 0 \end{cases} \ ,$$

$$\int\limits_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1 \ . \tag{1.12}$$

1.4 Квантили распределений

***Определение* 1.4**

Квантилем случайной величины x вероятности q (или уровня значимости $\alpha = 1 - q$), называется величина x_q, для которой имеем:

$$F\,(x_q) = \mathrm{Prob}\,(x < x_q) = \int\limits_{-\infty}^{x_q} p\,(x)\,\mathrm{d}x = q \tag{1.13}$$

или

$$\mathrm{Prob}\,(x > x_q) = \int\limits_{x_q}^{\infty} p\,(x)\,\mathrm{d}x = \alpha = 1 - q \ . \tag{1.14}$$

Отметим, что с помощью законов распределений вероятности можно найти вероятность нахождения случайной величины в некотором интервале $[a,b]$:

$$\mathrm{Prob}\,(a < x < b) = \int_a^b p\,(x)\,\mathrm{d}x = F(b) - F(a)\ . \tag{1.15}$$

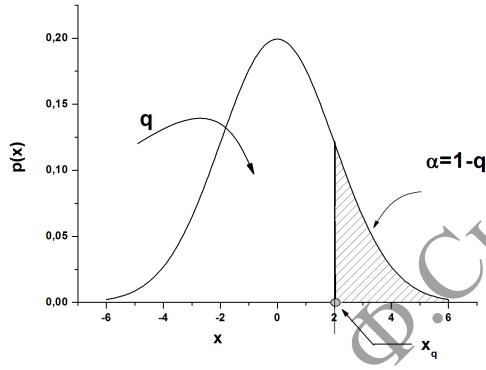

Рисунок 1– Иллюстрация квантиля распределения $p(x)$

1.5 Примеры законов распределений.

Нормальное распределение:

$$p\,(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]\ , \tag{1.16}$$

a и σ -параметры нормального распределения. **Распределение χ^2 с n степенями свободы:**

$$p(\chi^2) = \frac{(\chi^2)^{(\frac{n}{2}-1)} e^{-\frac{\chi^2}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\ , \quad (\chi^2 > 0)\ , \tag{1.17}$$

n – параметры χ^2 распределения.

Равномерное распределение:

$$p(x) = \frac{1}{b-a} \quad ,(b > a)\ , \tag{1.18}$$

a,b - параметры распределения.

Рисунок 2– Нормальное распределение с $a = 0$ и $\sigma = 1$

Рисунок 3– Распределение χ^2 с $n = 3,4,5,6$ степенями свободы

Рисунок 4– Равномерное распределение с $a = 0$ и $b = 1,2,4$

Дискретные распределения:

Распределение Пуассона:

$$p(x) = \frac{e^{-\mu}\mu^k}{k!} k \geq 0 \ , \tag{1.19}$$

μ - параметр распределения, а переменная $k = 0,1,\ldots$.

2 Моменты распределения.

Моменты k-того порядка для непрерывной случайной величины записываются в виде:

$$\mu_k = \int\limits_{-\infty}^{\infty} x^k p(x) dx \ , \tag{2.1}$$

где μ_k – алгебраический момент k-того порядка.

$$\nu_k = \int\limits_{-\infty}^{\infty} (x - \mu_1)^k p(x) dx \ , \tag{2.2}$$

где ν_k – центральный момент k-того порядка.

2.1 Математическое ожидание. Дисперсия.

Первый алгебраический момент называют **математическим ожиданием**:

***Определение* 2.1**

Математическим ожиданием *непрерывной случайной величины с плотностью вероятности* $p(x)$ *называется величина* $\mathrm{E}\{x\}$*, определяемая соотношением:*

$$\mu_1 = \mathrm{E}\{x\} = \int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x \ . \tag{2.3}$$

Для функции случайной величины $g(x)$ соотношение (2.3) обобщается следующим образом:

$$\mathrm{E}\{g(x)\} = \int_{-\infty}^{\infty} g(x) p(x) \, \mathrm{d}x \ . \tag{2.4}$$

Иногда, для сокращения записи, используем следующее обозначение для математического ожидания:

$$\mathrm{E}\{g(x)\} \equiv \langle g(x) \rangle \ . \tag{2.5}$$

Используя (1.11), (2.3) и свойство δ-функции

$$\int\limits_{-\infty}^{\infty} f(x)\delta(x-a)\mathrm{d}x = f(a) \tag{2.6}$$

для дискретной случайной величины получаем, что математическое ожидание вычисляется по формуле:

$$\mathrm{E}\{x\} = \sum_{i=1}^{n} p_i x_i \ . \tag{2.7}$$

Свойства математического ожидания (2.4):

1.
$$\mathrm{E}\{c\} = c \ , \text{если } c = \text{const} \ ; \tag{2.8}$$

2.
$$\mathrm{E}\left\{\sum_{k=1} c_k g_k(x)\right\} = \sum_{k=1}^{n} c_k \mathrm{E}\{g_k(x)\} \ , \text{если } c_k = \text{const} \ ; \tag{2.9}$$

3.
$$\mathrm{E}\left\{\sum_{k=1} c_k \xi_k\right\} = \sum_{k=1} c_k \mathrm{E}\{\xi_k\} \ , \text{если } \xi_k \text{ случайные величины} \ . \tag{2.10}$$

Математическое ожидание характеризует центр распределения $p(x)$. Однако следует отметить, что не для всех распределений существует математическое ожидание.

Наиболее общей характеристикой центра распределения следует считать медиану.

Определение 2.2

Медиана это такое значение случайной величины x_m для которой вероятности появления различных значений случайной величины X $p_1 = \mathrm{Prob}(X < x_m)$ и $p_2 = \mathrm{Prob}(X > x_m)$ равны между собой, т. е. $p_1 = p_2 = 0{,}5$.

Другими словами можно сказать, что медиана это квантиль распределения вероятности $q = 0{,}5$.

Также центр распределения может характеризоваться модой распределения.

Определение 2.3

Мода распределения это такое значение случайной величины x_{mod} для которой плотность вероятности появления значения случайной величины X максимальна, т. е. $p(x_{mod}) = \max p(x)$.

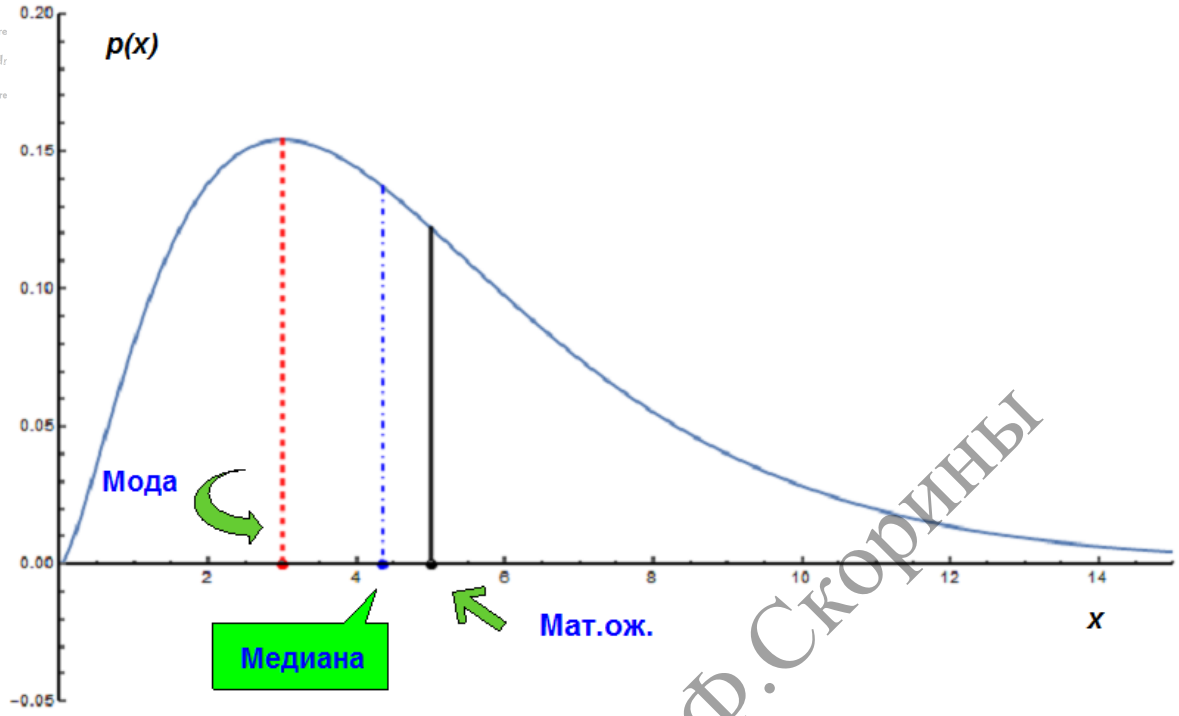

Рисунок 5– Пример распределения

Определение 2.4

Дисперсией непрерывной случайной величины с плотностью вероятности $p(x)$ *называется величина* $\mathrm{D}(x)$*, определяемая соотношением:*

$$\nu_2 = \mathrm{D}\{x\} = \int_{-\infty}^{\infty} (x - \mathrm{E}\{x\})^2 \; p(x)\,\mathrm{d}x \; . \tag{2.11}$$

Из определения (2.4) следует, что

$$\mathrm{D}\{x\} = \mathrm{E}\left\{(x-\mu_1)^2\right\} = \mathrm{E}\{x^2\} - \mathrm{E}^2\{x\} \; . \tag{2.12}$$

Свойства дисперсии (2.12):

1.
$$\mathrm{D}\{c\} = 0 \text{ , если } c = \text{const} \; ; \tag{2.13}$$

2.
$$\mathrm{D}\{cx\} = c^2\mathrm{D}\{x\} \text{ , если } c = \text{const} \; ; \tag{2.14}$$

3.
$$\mathrm{D}\{c + x\} = \mathrm{D}\{x\} \text{ , если } c = \text{const} \; . \tag{2.15}$$

Для дискретной случайной величины, с учетом (1.11) дисперсия вычисляется по формуле:

$$\mathrm{D}\{x\} = \sum_{k=1} (x_k - \mathrm{E}\{x\})^2 p_k \ . \tag{2.16}$$

Дисперсия характеризует рассеяние отдельных значений случайной величины от центра распределения и она определяет форму распределения.Дисперсия имеет размерность квадрата случайной величины и выражает как бы мощность рассеяния, поэтому для более наглядной характеристики используют среднеквадратичное отклонение σ:

$$\sigma_x = +\sqrt{\mathrm{D}\{x\}} \ . \tag{2.17}$$

которое имеет размерность самой случайной величины.

Для описания относительной меры отклонения используют коэффициент вариации

$$V_x = \frac{\sigma_x}{\mathrm{E}\{x\}}. \tag{2.18}$$

2.2 Коэффициент асимметрии. Эксцесс.

Центральные моменты 3,4 порядка дают информацию о виде кривой плотности распределения.

Определение 2.5

Третий центральный момент непрерывной случайной величины с плотностью вероятности $p(x)$, который определяется соотношением:

$$\nu_3 \equiv \int_{-\infty}^{\infty} (x - \mathrm{E}\{x\})^3 \ p(x)\,\mathrm{d}x \ , \tag{2.19}$$

характеризует асимметрию кривой $p(x)$ или скошенность распределения(например,когда один спад - крутой, а другой - пологий).

Для симметричных относительно центра распределений он равен нулю. ν_3 имеет размерность куба случайной величины, поэтому для относительной характеристики используют безразмерный **коэффициент асимметрии**:

$$\gamma_1 = \frac{\nu_3}{\nu_2^{3/2}} \ . \tag{2.20}$$

Четвертый центральный момент

$$\nu_4 \equiv \int_{-\infty}^{\infty} (x - \mathrm{E}\{x\})^4 \; p(x)\, \mathrm{d}x \;, \tag{2.21}$$

характеризует протяженность распределения (островершинность).

Определение 2.6

Безразмерную величину вида

$$\varepsilon = \frac{\nu_4}{\nu_2^2}$$

называют ***эксцессом распределения***.

Область изменения эксцесса: $\varepsilon \in [1, \infty]$. Часто используют **коэффициент эксцесса**

$$\gamma_2 = \varepsilon - 3 \tag{2.23}$$

и **контрэксцесс** $\kappa = 1/\sqrt{\varepsilon}$.

3 Оценки параметров распределений

***Определение* 3.1**

Функция результатов опытов, которая зависит от неизвестных статистических характеристик называют ***статистикой****. Статистика зависит от случайных величин и сама является случайной величиной.*

Для нахождения поведения случайной величины, полученных в результате эксперимента необходима числовая информация о моментах распределения.

Оценкой статистической характеристики $\tilde{\theta}$ называется статистика, которая принимается за неизвестное истинное значение параметра θ. Основное требование к оценке истинного значения состоит в том, чтобы большинство значений статистики сосредоточилось вблизи значений θ и вероятность больших отклонений от этого значения была мала. Также желательно, чтобы с увеличением объема выборки точность оценок также увеличилась.

Если имеется экспериментальная выборка объема n случайной величины $X = \{x_1, x_2, \ldots, x_n\}$, то ее элементы можно рассматривать как n статистически независимых случайных величин. Тогда любая оценка $\tilde{\theta}$ параметра θ должна быть функций элементов выборки, т. е.

$$\tilde{\theta} = \theta\left(x_1, x_2, \ldots, x_n\right) \ . \tag{3.1}$$

При этом закон распределения $\tilde{\theta}$ зависит от закона распределения величины X и от объема выборки.

Существуют два вида оценок параметров распределения: *точечные и интервальные.*

3.1 Точечные оценки

Под точечной оценкой параметра распределения понимают оценку одним числом. К точечным оценкам предъявляют следующие требования:

– **состоятельность**;

– **несмещенность**;

– **эффективность**;

– **устойчивость**;

надежность.

Определение 3.2

Метод оценки параметров называется ***состоятельным****, если оценки, полученные с его помощью, сходятся к истинному значению с увеличением объема выборки* n

Определение 3.3

Если $\tilde{\theta}$ *оценка параметра* θ*, то оценка будет* ***несмещенной****, если смещение*

$$\mathrm{E}\left\{\tilde{\theta}\right\} - \theta \tag{3.2}$$

равно нулю (0)*, т. е.* $\mathrm{E}\left\{\tilde{\theta}\right\} = \theta$.

Определение 3.4

Оценка $\tilde{\theta}_{eff}$ *является эффективной, если она обладает наименьшим разбросом относительно истинного значения параметра* θ*, т. е.*

$$\mathrm{D}\left\{\tilde{\theta}_{eff}\right\} \to \min\left\{\tilde{\theta}_1, \tilde{\theta}_2, \ldots\right\} \ . \tag{3.3}$$

Определение 3.5

Под ***устойчивостью*** *оценки понимают нечувствительность ее к малым отклонениям от точного распределения.*

Точечная оценка для математического ожидания (центра распределения) дается выражением:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \tag{3.4}$$

Для дисперсии $D = \sigma^2$ такая оценка дается выражением:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ , \quad S = \sqrt{\frac{1}{n-1}}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \ . \tag{3.5}$$

Величины $\bar{x}$ и S сами являются случайными величинами и, следовательно, они тоже могут иметь разброс, который характеризуется дисперсией.

Для $\bar{x}$:

$$S_{\bar{x}}^2 = \frac{S^2}{n} \ , \quad S_{\bar{x}} = \frac{S}{\sqrt{n}} \ . \tag{3.6}$$

Для дисперсии:

$$D\left[S^2\right] = \frac{\nu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} \ , \tag{3.7}$$

где ν_4 – четвертый центральный момент, а σ – среднеквадратичное отклонение.

Несмещенными и состоятельными оценками $\tilde{\nu}_3$ и $\tilde{\nu}_4$ для третьего ν_3 и четвертого ν_4 центрального моментов соответственно являются

$$\tilde{\nu}_3 = \frac{n^2}{(n-1)\,(n-2)} m_3 \ , \tag{3.8}$$

$$\begin{aligned} \tilde{\nu}_4 = & \frac{1}{(n-1)\,(n-2)\,(n-3)} \times \\ & \times \left[n\left(n^2-2n+3\right) m_4 - 3n\,(2n-3)\,m_3^2\right] \ , \end{aligned} \tag{3.9}$$

где

$$m_k = \sum_{i=1}^{m} (x_i - \bar{x})^k \ . \tag{3.10}$$

Тогда оценки (знак $\widetilde{\ }$) для коэффициента асимметрии γ_1 и эксцесса ε (смотри соотношения (2.20)и (2.6)) определяются формулами:

$$\begin{aligned} \widetilde{\gamma}_1 &= \frac{\tilde{\nu}_3}{S^3} \ , \\ \widetilde{\varepsilon} &= \frac{\tilde{\nu}_4}{S^4} \ , \end{aligned} \tag{3.11}$$

где S – точечная оценка $\sqrt{D\{x\}}$.

Примечание.

В теории погрешностей, которая составляет основу обработки данных в лабораторных работах вместо обозначения S используют Δx. Эту величину называют **абсолютной погрешностью**, а также **стандартным отклонением**. Коэффициент вариации V_x называют относительной погрешностью и выражают в процентах:

$$V_x \to \epsilon_x = \frac{\Delta x}{\bar{x}} \times 100\% \ . \tag{3.12}$$

3.2 Интервальные оценки

Точечные оценки параметров случайных величин не позволяют судить о степени близости выборочных значений к оцениваемому параметру. Более содержательны процедуры оценивания параметров, связанные с построением интервала с известной степенью доверительности.

Для любой случайной величины X можно определить вероятности попадания в некоторый интервал $[x_{min}, x_{max}]$, если известен закон распределения вероятностей (смотри (1.15)):

$$\text{Prob}\,(x_{min} < x < x_{max}) = \int\limits_{x_{min}}^{x_{max}} p\,(x,\theta)\,\mathrm{d}x = F(x_{max}) - F(x_{min})\ , \qquad (3.13)$$

где в плотность распределения введена величина θ, которая определяет параметры распределения.

Потребуем, чтобы вероятность попадания равнялась некоторому значению $P = 1 - \alpha$, т. е.

$$\text{Prob}\,(x_{min} < x < x_{max}) = P = 1 - \alpha\ . \qquad (3.14)$$

С помощью (3.14) можно "построить" три вида интервалов с заданной вероятностью попадания p:

1. Верхний односторонний (левосторонний) интервал $]-\infty, x_{max}]$;
2. Нижний односторонний (правосторонний) интервал $]x_{min},\infty]$;
3. Двусторонний интервал $[x_{min}, x_{max}]$.

Предполагается, что вероятность попадания в каждый их интервалов равна $P = 1 - \alpha$. Обычно величину P называют доверительной вероятностью (иногда надежностью), а величину α уровнем значимости интервала.

Интервальные оценки для моментов распределения находятся построением некоторой функции случайной величины, куда входят искомые параметры распределений θ и точечные оценки $\tilde{\theta}$. Тогда (3.14) применительно к параметрам можно записать

$$\text{Prob}\left(\tilde{\theta} - \delta_{min} < \theta < \tilde{\theta} + \delta_{max}\right) = P = 1 - \alpha\ , \qquad (3.15)$$

который следует понимать так: вероятность того, чтобы параметр θ находится в интервале $\left[\tilde{\theta} - \delta_{min}, \tilde{\theta} + \delta_{max}\right]$ равна $P = 1 - \alpha$.

Отметим, что задача построения доверительных интервалов для параметров при произвольном объеме "экспериментальной" выборки n разработана только для нормального распределения случайной величины X. Для других распределений имеются только частные случаи.

Например, для любой выборки n из нормальной совокупности с математическим ожиданием ξ функция вида

$$t = \frac{\bar{x} - \xi}{(S/\sqrt{n})} \tag{3.16}$$

имеет t- распределение (распределение Стьюдента) с $n-1$ степенями свободы, плотность которого определяется соотношением:

$$p_{\mathrm{ST}}(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi} n \Gamma\left(\frac{n}{2}\right)} \left[1 + \frac{t^2}{n}\right]^{-\frac{n+1}{2}}. \tag{3.17}$$

Тогда можно найти такое значение $t_{n,P}$ случайной величины с распределением Стьюдента для которой выполняется

$$\mathrm{Prob}\left(\left|\frac{\bar{x} - \xi}{S/\sqrt{n}}\right| < t_{n-1,P}\right) = P = 1 - \alpha \tag{3.18}$$

или

$$\mathrm{Prob}\left(\bar{x} - \frac{t_{n-1,P} S}{\sqrt{n}} < \xi < \bar{x} + \frac{t_{n-1,P} S}{\sqrt{n}}\right) = P = 1 - \alpha\,. \tag{3.19}$$

В итоге имеем

Доверительный интервал для математического ожидания $\mathrm{E}\{x\} = \xi$ (двусторонний), если неизвестна дисперсия распределения:

$$\left[\bar{x} - \frac{t_{n-1,1-\alpha/2} S}{\sqrt{n}} < \xi < \bar{x} + \frac{t_{n-1,1-\alpha/2} S}{\sqrt{n}}\right], \tag{3.20}$$

где $t_{n,P}$ определяется соотношением:

$$\int_{-\infty}^{t_{n,P}} p_{\mathrm{ST}(t)\mathrm{d}t=P}\,.$$

Коэффициенты $t_{n,P}$ носят название коэффициентов Стьюдента.

Для построения доверительного интервала для дисперсии $\mathrm{D}\{x\} = \sigma^2$ используется, то, что функция

$$\frac{(n-1)\,S}{\sigma^2} \tag{3.21}$$

имеет распределение χ^2 с $n-1$ степенями свободы.

Тогда **доверительный интервал для дисперсии** при неизвестном математическом ожидании имеет вид:

$$\left[\frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha/2}} < \mathrm{D}\{x\} < \frac{(n-1)\,S^2}{\chi^2_{n-1,1-\alpha/2}}\right] , \tag{3.22}$$

где $\chi^2_{n,\alpha}$ – квантиль распределения χ^2:

$$\int\limits_{\chi^2_{n,\alpha}}^{\infty} p(\chi^2)\mathrm{d}\chi^2 = \alpha$$

с плотностью

$$p(\chi^2) = \frac{(\chi^2)^{(\frac{n}{2}-1)}e^{-\frac{\chi^2}{2}}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} , \quad (\chi^2 > 0) .$$

Значения $\mathrm{E}\{x\}$ и $\mathrm{D}\{x\}$ лежат в интервале с доверительной вероятностью $P = 1-\alpha$.

3.3 Доверительный интервал для случайной величины

С помощью точечных оценок, рассмотренных нами выше, результат для случайной величины x обычно записывают в виде:

$$(\bar{x} \pm \Delta x) \ . \tag{3.23}$$

При этом согласно правилу приведения результатов:

Правило приведения результатов

Последняя значащая цифра в любом приводимом результате для $\bar{x}$ обычно должна быть того же порядка величины (находиться в той же десятичной позиции), что и абсолютная погрешность Δx.

Однако используемые в расчетах числа должны, как правило, содержать на одну значащую цифру больше, чем это оправдано. Это уменьшит неточности, возникающие при округлении чисел. В конце расчета окончательный ответ следует округлить и избавиться от этой добавочной (и незначащей) цифры).

Придать вероятностный смысл интервалу (3.23) можно, зная лишь функцию распределения вероятности $p(x)$ (плотность вероятности). Если предположить, что наша выборка принадлежит генеральной совокупности с нормальным распределением ($\mathrm{E}\{x\} = \bar{x}$, $\mathrm{D}\{x\} = (\Delta x)^2$), то легко найти, что

$$\begin{aligned}
&\mathrm{Prob}\left(\bar{x} - \Delta x < x < \bar{x} + \Delta x\right) = \\
&= \frac{1}{\Delta x\sqrt{2\pi}} \int\limits_{\bar{x}-\Delta x}^{\bar{x}+\Delta x} \exp\left[-\frac{(x-\bar{x})^2}{2\,(\Delta x)^2}\right] \mathrm{d}x \\
&= \left(z = \frac{x-\bar{x}}{\Delta x}\right) = \\
&= \frac{1}{\sqrt{2\pi}} \int\limits_{-1}^{1} \exp\left[-\frac{z^2}{2}\right] \mathrm{d}z = \\
&= \mathrm{erf}\left(\frac{1}{\sqrt{2}}\right) \approx 0{,}682689\,. \qquad (3.24)
\end{aligned}$$

Поэтому в физических приложениях, часто говорят: вероятность того, что результат измерения окажется в пределах одного стандартного отклонения от истинного результата, составляет 68,3%. Можно сказать и так: вероятность, того измеряемая случайная величина лежит в интервале (3.23) составляет 68,3%.

4 Взвешенное среднее значение

В практических исследованиях часто встречается следующая ситуация: одна и эта же величина измеряется в нескольких экспериментах. Это делается для наиболее надежного определения некоторой величины. При этом собирают измерения различного происхождения, выполненные разными установками (инструментами) и методами. Результаты таких измерений называют часто называют **неравноточными**.

Предположим, что у нас есть n отдельных измерений

$$x_1 \pm \Delta x_1 \ , \quad x_2 \pm \Delta x_2 \ , \ldots , x_n \pm \Delta x_n, \tag{4.1}$$

с соответствующими погрешностями $\Delta x_1 \ , \ldots , \Delta x_n$.

***Определение* 4.1**

Наилучшая оценка величины X, основанная на таких измерениях, равна в средневзвешенному значению $\hat{x}_{sw}$

$$\hat{x}_{sw} = \frac{1}{w} \sum_{i=1}^{n} w_i x_i \ , \tag{4.2}$$

$$w = \sum_{i=1}^{n} w_i \ , \quad w_i = \frac{1}{(\Delta x_i)^2} \ , \tag{4.3}$$

а её среднеквадратичное отклонение $\Delta \hat{x}_{sw}$

$$\Delta \hat{x}_{sw} = \frac{1}{\sqrt{w}} \ . \tag{4.4}$$

Результат такой обработки данных записывается в обычном виде:

$$\hat{x}_{sw} \pm \Delta \hat{x}_{sw} \ . \tag{4.5}$$

5 Ошибка косвенного измеряемой величины

Все, о чем мы говорили выше, относилось непосредственно к измеряемым величинам. А как определить погрешности для косвенно измеряемых величин, т. е. величин измеряемых с помощью физических законов из непосредственно измеряемых? Оказывается, погрешность косвенно измеряемых величин, связана с погрешностями непосредственно измеряемых величин.

Пусть косвенно измеряемая величина Z связана с непосредственно измеряемыми $x_1, \ldots, x_m$ (m величин) величинами функцией $Z = f(x_1, \ldots, x_m)$. Если случайные величины x_i **не коррелированы друг с другом, т. е. независимы**, то абсолютная погрешность величины Z определяется соотношением

$$\sigma_Z = \Delta Z = \sqrt{\sum_{i=1}^{m} \left(\frac{\partial f}{\partial x_i}\right)^2 (\Delta x_i)^2}\,, \tag{5.1}$$

где $\Delta x_1, \ldots, \Delta x_m$ среднеквадратичные отклонения непосредственно измеряемых величин.

Пример. $Z = f(x,y) = x \pm y$

$$\Delta Z^2 = \underbrace{\left(\frac{\partial f}{\partial x}\right)^2}_{=1} (\Delta x)^2 + \underbrace{\left(\frac{\partial f}{\partial y}\right)^2}_{=1} (\Delta y)^2 =$$

$$= (\Delta x)^2 + (\Delta y)^2\,. \tag{5.2}$$

Практическое следствие этого соотношения: для создания оптимальных условий (условий с минимальной погрешностью) основные усилия должны быть направлены не на дальнейшее уточнение тех результатов измерений, которые являются наиболее точными, а на совершенствование наименее точных измерений случайных величин.

6 Графическое представление эмпирических данных

Цель обработки данных заключается в выявлении вида распределений случайных величин и оценки параметров установленного распределения.

Полученные экспериментальные данные представляют, как правило, в виде таблиц. Полученные таблицы удобно представить графически. Используя набор независимых наблюдений $x_1, x_2, \ldots, x_n$ случайной величины X, полезным первым шагом в исследовании поведения случайной величины является организация и представление их таким образом, чтобы их можно было легко интерпретировать и оценивать. Для достаточно большого количества наблюдаемых данных, полигон частот (распределения), гистограмма и кумулятивная линия является отличным графическим представлением данных, что облегчает оценку адекватности предполагаемой модели и оценку параметров распределения.

Гистограмма и полигон распределений являются графическим отображением частот, которые, в свою очередь, представляют собой оценки плотностей вероятностей $p(x)$. Кумулятивная линия - это график накопленных частот, в свою очередь оценивающих интегральную функцию распределения $F(x)$.

Как строить полигон частот

1. Построить вариационный ряд для выборки, т. е. упорядочить значения случайной величины так, чтобы выполнялось условие: $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$.
2. Разбить область на m интервалов (бинов).
3. Построить точки на плоскости x–ν с координатами $\{\tilde{x}_i, \nu_i\}, (i = 1,\ldots,n)$, где

$$\tilde{x}_i = \frac{x_i + x_{i+1}}{2} \tag{6.1}$$

является серединой i-того интервала, ν_i - частота попадания случайной величины X в i-тый интервал.

Это же распределение можно представить в виде гистограммы. Для построения гистограммы необходимо над каждым отрезком оси абсцисс, соответствующим интервалу значений измеряемой величины, построить прямоугольник, площадь которого пропорциональна частоте попадания в этот ин-

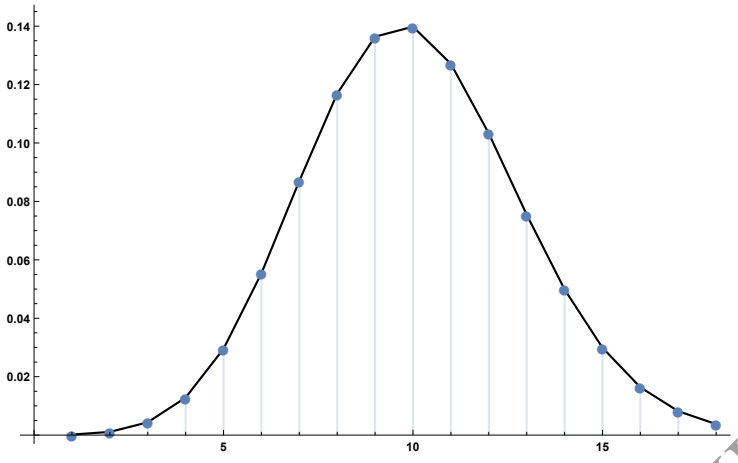

Рисунок 6– Пример полигона частот

тервал. Обычно выбирают интервалы одинаковой ширины, поэтому высота прямоугольников различна.

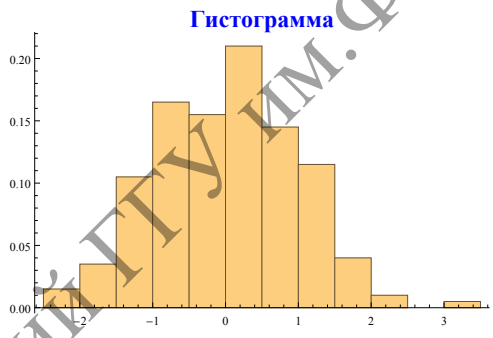

Рисунок 7– Пример гистограммы

Где используются обработка данных для построения гистограммы?

Для определения оценок математического ожидания, с.к.о., эксцесса не требуется какого-либо группирования данных.

Для определения медианы, сгибов, использования критерия согласия Колмогорова-Смирнова или для обнаружения промахов, экспериментальные данные необходимо расположить в порядке возрастания, т.е. построить вариационный ряд (упорядоченную выборку).

Для определения формы распределения, для использования критериев согласия Пирсона и др., для сопоставления гипотез о форме распределения и т.д. простого упорядочения выборки уже недостаточно, а выборка должна быть представлена в виде гистограммы, состоящей из m столбцов с определенной протяженностью d соответствующих им интервалов.

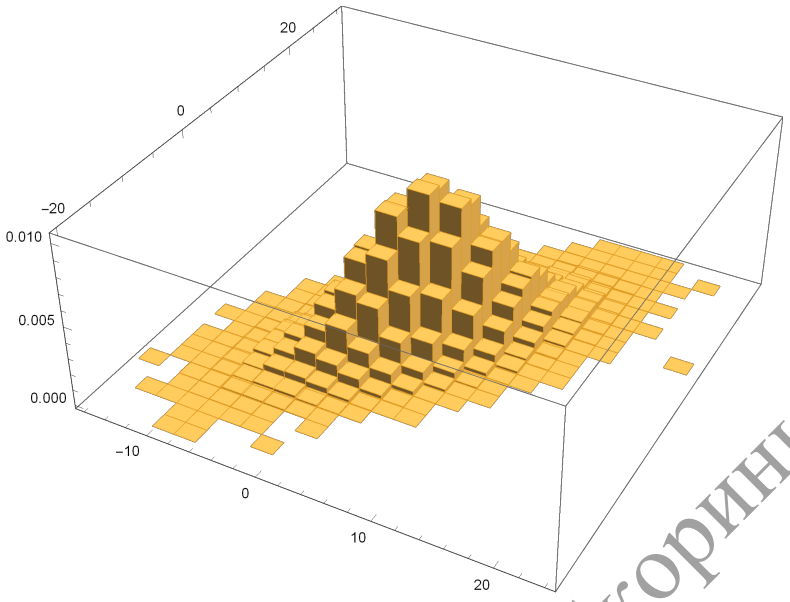

Рисунок 8– Пример 3D-гистограммы

7 Оптимальное число интервалов для получения гистограммы

Как выбрать m и d.

Оптимальное число

Оптимальное число существует!!

Оптимальное число интервалов группирования это такое число, когда ступенчатая огибающая гистограммы наиболее близка к плавной кривой распределения случайной величины.

К примеру, при группировании данных выбор большого числа интервалов ($m > n$), автоматически приведет к тому, что некоторые из них окажутся пустыми или малозначительными. Гистограмма будет отличаться от плавной кривой распределения вследствие изрезанности многими всплесками и провалами.

Тогда можно сформулировать 1-е требование к числу интервалов: Размер интервала (ячейки, бина) должен быть достаточно широким для обеспечения хороших статистических свойств будущей гистограммы (достаточно большая статистика, минимальные корреляции (связи) с соседними ячейками).

При слишком малом числе m интервалов, гистограмма отличается от

действительной кривой распределения вследствие слишком крупной ступенчатости. Из-за чего будут потеряны характерные особенности. Например, если взять $m = 1$, т.е. d равно размаху экспериментальных данных, то любое распределение сводится к равномерному, а если $m = 3$, то любое куполообразное распределение сведется к треугольному. Например, при обработке линейчатых спектров, слишком большая ячейка может привести к потере спектральной линии.

Тогда возникает **2-е требование к числу интервалов**:
Размер ячейки должен быть достаточно узким для того, чтобы прорисовывалась "тонкая структура" исследуемой величины.

Как видим, требования являются противоречивыми. Укрупнение интервалов группирования является методом "фильтрации различных случайных выбросов и провалов", но слишком протяженные интервалы сглаживают особенности искомого закона распределения. Таким образом, задача выбора оптимального числа интервалов при построении гистограммы – это задача оптимальной фильтрации, а оптимальным числом m интервалов является максимальное возможное сглаживание случайных флуктуаций данных, которое сочетается с минимальным искажением от сглаживания самой кривой искомого распределения.

Общепринято делать интервалы `одинаковыми`. Хотя в дальнейшем увидим, что это условие необязательно. Условие равновеликости интервалов удобно с практической точки зрения.

Рекомендации по выбору m.

I группа: эвристические критерии (без доказательства).

Формула Стерджеса

$$m = \log_2 n + 1 \ . \tag{7.1}$$

Формула Брукса и Каррузера

$$m = 5 \lg n \ . \tag{7.2}$$

Формула если $n > 100$

$$m = \sqrt{n} \ . \tag{7.3}$$

Эти три формулы являются наиболее часто встречающимися в литературе по математической статистике.

II группа: с использованием критерия χ^2.

В ней используется рассмотрение интервалов не с равной длинной, а с **равной вероятностью** в соответствии с принимаемой моделью, т. е. предположением о законе распределения. В данном подходе неявно учитывается форма распределения.

Число интервалов с равной вероятностью, которые мы обозначили как K, отличаются от числа m с равной длиной d .

Г. Манн и А. Ваальд установили, что при $n \to \infty$ оптимальное число K равновероятных интервалов задается соотношением:

Критерий Мана-Ваальда

$$K \sim b\sqrt[5]{2}\left(\frac{n}{Z_\alpha}\right)^{2/5}, \tag{7.4}$$

где $b = 2 \div 4$.

Здесь Z_α – квантиль нормального распределения, соответствующий вероятности $P = 1 - \alpha$, α – принятый уровень значимости.

$$Z_\alpha = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_\alpha} e^{-\frac{x^2}{2}} \mathrm{d}x = 1 - \alpha \ .$$

Критерий Мана-Ваальда

На практике часто берут $\alpha = 0.1$, тогда

$$K \simeq 1{,}9\, n^{2/5} \ . \tag{7.5}$$

В итоге приходим к таким рекомендациям при использования равновероятностных бинов K:

1. найти число ячеек, используя (7.5);
2. если окажется, что n/K мало, уменьшить K чтобы выполнялось неравенство $n/K \geqslant 5$;
3. Сформировать K равновероятностных ячеек гистограммы на основе данных. Отметим, что если измеряемая случайная величина многомерная, то существуют различные способы формирования ячеек с одинаковым вероятностным содержанием в K ячейках.

III группа.

Поскольку для K интервалы получаются не равной длины, то это приводит к ряду неудобств при построении гистограмм, но зато при этом мы неявно закладываем при использовании χ^2 выбор K в зависимости от формы распределения.

III группа рекомендаций "устраняет" недостаток II группы возвращаясь к интервалам m с равной длиной d, но при этом и учитывает, в отличие от I группы, форму распределения (форма характеризуется эксцессом ε или контрэксцессом κ).

Примером такого подхода является соотношение:

III группа (формула И.У.Алексеевой)

$$m = \frac{4}{\kappa}\lg\frac{n}{10}\ , \quad \kappa = \frac{1}{\sqrt{\varepsilon}}\ . \tag{7.6}$$

На практике эту формулу в зависимости от n при $\kappa = const$ удобнее аппроксимировать выражением

$$m = \frac{\varepsilon + 1{,}5}{6} n^{2/5} \tag{7.7}$$

Трудность использования III группы состоит в том, что число интервалов часто приходиться выбирать прежде, чем будут найдены оценки $\varepsilon, \bar{x}$, и т.д.

Эту трудность обходят следующим образом: наиболее часто встречаются распределения с ε от 1,8 до 6 (от равномерного до Лапласа, включая нормальное ε=3). Для этих граничных точек имеем

$$m_{min} = 0{,}55\, n^{2/5} \tag{7.8}$$

$$m_{max} = 1{,}25\, n^{2/5} \tag{7.9}$$

Искомое m можно выбрать близким к этому интервалу, при этом m лучше выбрать нечетным, т.к. при четном m для островершинных распределений в центре гистограммы оказывается два столбца равных по высоте и середина распределения принудительно утолщается.

“Практические” рекомендации для построения диаграмм

1. Для практического определения числа интервалов воспользоваться формулами для m_{min} (7.8) и m_{max} (7.9), или сразу формулой (7.6), выбрав при этом m нечетным.

2. Так как крайние точки могут располагаться несимметрично, то ширина d столбца гистограммы определяется по отклонению от центра ΔX_m наиболее удаленной точки:

$$d = \frac{2\Delta X_m}{m} \, . \tag{7.10}$$

   При этом полученное значение d необходимо округлять в большую сторону, чтобы крайняя точка не оказалась за пределами крайнего столбца.

3. Величину d при этом удобно выбирать так, чтобы она делилась на 2 так, чтобы потом центральный столбец можно было бы поделить пополам для уточнения центра распределения.

8 Промахи и методы их исключения

Одним из условий правомерности статистической выборки является требование ее однородности, т.е. принадлежности всех ее членов к одной и той же генеральной совокупности.

Однако на практике это требование очень часто нарушается. И, если скажем, при обработке вручную еще можно вспомнить как (при каких условиях) были получены "подозрительные" данные, то при автоматической обработке данных необходимы методы исключения "чужих" для данной выборки результатов.

***Определение* 8.1**

Отсчёты, резко отклоняющиеся по своим значениям от большинства других отсчетов принято называть ***промахами*** *и исключать их из выборки.*

Важно. Если серия из небольшого числа измерений содержит грубую погрешность — промах, то наличие этого промаха может сильно исказить как среднее значение измеряемой величины, так и границы доверительного интервала. Поэтому из окончательного результата необходимо исключить этот промах.

Обычно промах имеет резко отличающееся от других измерений значение. Однако это отклонение от значений других измерений не дает еще права исключить это измерение как промах, пока не проверено, не является ли это отклонение следствием статистического разброса.

Особую неприятность доставляют отсчеты, которые и не входят в компактную группу отсчетов, но и не удалены от нее на значительное расстояние. Такой отсчет называют предполагаемым промахом. В экспериментальной

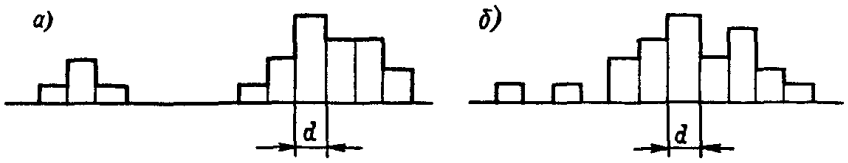

Рисунок 9– Возможные промахи

практике исследователи просто отбрасывали крайние, "слишком удаленные от центра наблюдения". Эта процедура получила название **цензурирование выборки**.

Однако для принятия решения необходимы какие-либо формальные критерии.

Простейший метод заключается в использовании "правила 3σ", когда по выборке с удаленными отсчётами (предполагаемыми промахами) вычисляется оценка σ и граница $|X_{\text{group}}| = 3\sigma$, а все $|x_i| \pm 3\sigma$ отбрасываются.

Правило 3σ обосновано на неравенстве Чебышева:

$$\text{Prob}\{|x_i - \bar{x}| \geqslant a\} \leq \frac{\sigma^2}{a^2}, \quad (a > 0) \tag{8.1}$$

Если все $x_i \geqslant 0$, то

$$\text{Prob}(x \geqslant a) \leqslant \frac{\bar{x}}{a}\,, \quad (a > 0) \tag{8.2}$$

Если x имеет одномодальное распределение (непрерывное), то справедлива более сильная оценка

$$\text{Prob}\{|x_i - \bar{x}| \geqslant a\} \leq \frac{4}{9}\frac{1 + S^2}{\left(\dfrac{a}{\sigma} - |S_{pear}|\right)^2}\,, \tag{8.3}$$

где S_{pear} – пирсоновская мера асимметрии (для распределений симметричных относительно моды $S_{pear} = 0$).

$$S_{pear} = \frac{\overline{x} - \xi}{\sigma}\,,$$

где ξ – максимальная вероятность.

$$S_{pear} = \frac{\gamma_1\,(\gamma_2 + 6)}{2\,(5\gamma_2 - 6\gamma_1^2 + 6)}\,,$$

где

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}\,, \quad \gamma_2 = \frac{m_4}{m_2^2} - 3\,.$$

Значения m_j определяются по формуле:

$$m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^j\,.$$

Если положить $a = 3\sigma$, то имеем

$$P\{|x_i - x| \geq 3\sigma\} \leq \frac{1}{9} \approx 0{,}11$$

или для одномодальных симметричных:

$$F\{|x_i - x| \geq 3\sigma\} \leq \frac{4}{9}\frac{1}{9} = \frac{4}{81} \approx 0{,}05 \ ,$$

т. е. для произвольного распределения вероятность, что $x_i \geqslant \bar{x}$ на 3σ составляет 11%, а для одномодальных симметричных 5%.

Чувствительность разных статистических методов к наличию аномальных наблюдений (“промахов”) в экспериментальных данных неодинакова.

Критерий Граббса

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}} \tag{8.4}$$

Тест Граббса определяется для гипотезы:

- H_0: в наборе данных нет выбросов
- H_1: В наборе данных имеется ровно один выброс.

Тест Граббса основан на предположении о нормальности выборки. То есть сначала нужно проверить, что данные могут быть разумно аппроксимированы нормальным распределением перед применением теста Граббса.

Тест Граббса обнаруживает один выброс за раз. Этот выброс исключается из набора данных, и тест повторяется до тех пор, пока не будут обнаружены выбросы. Тем не менее, множественные итерации изменяют вероятности обнаружения, и тест не должен использоваться для размеров выборок в шесть или меньше, поскольку он часто помещает большинство точек в виде выбросов.

Критерий Роснера для обнаружения нескольких выбросов

$$R_i = \max \frac{x_i - \bar{x}}{S} \tag{8.5}$$

или

$$\tau_1 = \max\{\frac{x_1 - \bar{x}}{S}, \frac{x_n - \bar{x}}{S}\} \text{ , если } x_1 < x_2 < \ldots < x_n \ . \tag{8.6}$$

Алгоритм критерия Роснера состоит в следующем. По начальной выборке объема n вычисляются значения $\bar{x}$ и S и статистика τ_1. Затем из выборки удаляется экстремальный член $x_{min}(x_{max})$—в зависимости от того, какое значение более удалено от среднего. Так повторяется k раз.

Полученные значения статистик τ_{ik} $(i = 1, \ldots, k)$ каждый раз сравниваются с критическими значениями

$$\tau^*_{i,n,p} = \frac{n-i}{\sqrt{n-i+1}} \sqrt{\frac{t^2_{p,n-i-1}}{n+i-1+t^2_{p,n-i-1}}}\ ,\ p = 1-\alpha/(2(n-i+1)) \qquad (8.7)$$

для заданных n, k и вероятности p. Превышение критерием τ_{1i} критического значения $\tau^*_{i,n,p}$ для некоторого i, позволяет установить не только наличие выбросов, но и их количество (равное значению i, при котором появляется первая значимая величина критерия τ_{1i}). Вычисление последовательных статистик ведется до тех пор, пока $\tau_{1(i+1)} > \tau_{1i}$.

8.1 Другие критерии для исключения промахов

Однако, правило "3σ" хорошо для нормального распределения. Действительно, при $n = 100$ появление $|x_i \geqslant 3\sigma|$ можно считать промахом, то для равномерного промахом можно считать уже $|x_i| = 1.8\sigma$, а для распределения Лапласа $|x_i| = 3\sigma$ есть отсчет, принадлежащий данной выборке.

На этом примере видно, что граница цензурирования зависит не только от объема выборки n, но также и от формы распределения.

Зависимость от n полуколичественно можно оценить из условия: границы цензурирования должны отсекать в среднем менее одной точки, тогда назначение границ с уровнем значимости $g = 1 - P$, где $P = \frac{n}{n+1}$ дает зависимость от n.

Для расчета количества σ для цензурирования можно использовать аппроксимационные формулы:

Формулы для расчета количества σ ()

$$t_{\text{гр.}} = 1{,}2 + 3{,}6\left(1 - 1/\sqrt{\varepsilon}\right)\lg\left(\frac{n}{10}\right)\ ,$$
$$t_{\text{гр.}} = 1{,}55 + 0{,}8\sqrt{\varepsilon - 1}\lg\left(\frac{n}{10}\right) \qquad (8.8)$$

Тогда интервал цензурирования выглядит следующим образом:

$$\bar{x} \pm t_{\text{гр.}} S\ , \qquad (8.9)$$

где S - точечная оценки среднеквадратичного отклонения (3.5).

В заключение данного раздела отметим, что все оценки $\bar{x}$, S и т.д. должны пересчитываться после цензурирования выборки.

9 Критерии согласия

Постановка задачи.

Одной из задач первичной обработки экспериментальных наблюдений является выбор закона распределения, который наиболее хорошо описывающего случайную величину, выборку которой наблюдают.

Рассмотрим любой эксперимент, в котором измеряется некоторая случайная величина X и мы получаем некоторую выборку объема n. Наша задача описать поведение этой случайной величины. Из теории мы знаем о большом количестве функций распределения вероятностей (нормальное, равномерное распределения, распределение Стьюдента и т.д.).

И у нас возникает мысль: А что если одно из известных теоретических (модельных) распределений подходит для описания поведения измеряемой величины? Как быть? Подойдет ли выбранное нами теоретическое распределение или не подойдет? Такая задача в математической статистике называют проверкой гипотезы.

***Определение* 9.1**

Под статистической гипотезой понимают всякое высказывание о случайной величине (генеральной совокупности), проверяемое по выборке (по результатам наблюдений).

Соответственно, процедура сопоставления высказанной гипотезы с выборочными данными называется **проверкой гипотезы**.

Таким образом, **целью проверки гипотезы о согласии** опытного распределения с теоретическим является стремление удостовериться в том, что данная модель теоретического закона не противоречит наблюдаемым данным и использование ее не приведет к существенным ошибкам при вероятностных расчетах.

Ответ, который должен быть получен в конечном итоге выглядит следующим образом: Данная ("экспериментальная") выборка с вероятностью p соответствует теоретическому распределению с параметрами (например, нормальному распределению с параметрами $a = 3$ и $\sigma = 1$).

Поэтому основная задача на практике состоит в определении вероятности принятия модели (гипотезы) p (или более точно вероятности, что данная выборка не соответствует данной гипотезе $\alpha = 1 - p$).

Что нужно знать о процессе проверки гипотез:

Этап 1.Выбор модели (теоретической функции распределения)

Располагая выборочными данными и руководствуясь конкретными условиями рассматриваемой задачи, формулируют гипотезу H_0, которую называют **основной** или нулевой, и гипотезу H_1 конкурирующую с гипотезой H_0.

Термин "конкурирующая" означает, что являются противоположными следующие два события: по выборке будет принято решение о справедливости для генеральной совокупности гипотезы H_0; и по выборке будет принято решение о справедливости для генеральной совокупности гипотезы H_1. Гипотезу H_1 называют также альтернативной.

Например, если нулевая гипотеза такова: данная выборка имеет нормальное распределение с параметрами $a = 3$ и $\sigma = 1$, то альтернативная гипотеза может быть следующей: данная выборка имеет нормальное распределение, но с параметрами $a < 3$ и $\sigma = 1$ или

$$H_0 : p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right], \text{ где}$$
$$a = 3 \quad \text{и} \quad \sigma = 1\,, \tag{9.1}$$

и

$$H_1 : p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right], \text{ где}$$
$$a < 3 \quad \text{и} \quad \sigma = 1\,. \tag{9.2}$$

Этап 2.Расчет вероятности принятия гипотезы p или вероятности не соответствия данной модели $\alpha = 1 - p$.

Вероятность α часто называют уровнем значимости и если выборка соответствует выбранной нами модели (теоретическому распределению), то уровень значимости относительно мал ($\alpha < 0{,}05$) Поясним ее смысл.

Решение о том, можно ли считать высказывание H_0 справедливым для генеральной совокупности, принимается по выборочным данным, т. е. по ограниченному ряду наблюдений, следовательно, это решение может быть ошибочным.

При этом может иметь место ошибка двух родов: отвергают гипотезу H_0, или, иначе, принимают альтернативную гипотезу H_1, тогда как на самом деле гипотеза H_0 верна; это **ошибка первого рода**; принимают гипотезу H_0, тогда как на самом деле высказывание H_0 неверно, т. е. верной является гипотеза H_1 это **ошибка второго рода**.

Так вот уровень значимости α – это вероятность ошибки первого рода, т. е. вероятность того, что будет принята гипотеза H_1, если на самом деле в

генеральной совокупности верна гипотеза H_0 (или отклонение проверяемой гипотезы H_0 при её справедливости).

Вероятность ошибки второго рода часто обозначают β, т. е. вероятность того, что будет принята гипотеза H_0, если на самом деле верна гипотеза H_1 (или не отклонении H_0 при справедливости H_1).

О мощности критерия.

При использовании критериев согласия, как правило на практике, не задают конкурирующих гипотез: рассматривается принадлежность выборки конкретному закону. А в качестве конкурирующей гипотезы — принадлежность любому другому.

Естественно, что способность критерия отличать закон, соответствующий H_0, от других, близких к закону, соответствующему H_0, и далёких от него, отличаются.

***Определение* 9.2**

Мощностью критерия *по отношению к конкурирующей гипотезе H_1 называется величина $1-\beta$. Критерий тем лучше распознаёт пару конкурирующих гипотез H_0 и H_1, чем выше его мощность.*

Один из возможных вариантов проверки гипотез являются **критерии согласия.**

Критерий χ^2 (критерий Пирсона).

Рассмотрим этапы необходимые для проверки гипотез на примере критерия χ^2.

Процедура проверки гипотез с использованием критериев типа χ^2 предусматривает группирование наблюдений.

1. Выбираем "модель": обычно это теоретическое дифференциальное распределение вероятностей $p(x,\theta)$, где θ – один (или несколько) параметров распределения.

2. Область определения случайной величины разбивают на m непересекающихся интервалов граничными точками $x_0, x_1, \ldots, x_{m-1}, x_m$, где $x_0 < x_1 < \ldots < x_{m-1} < x_m$.

3. В соответствии с заданным разбиением подсчитывают число n_i выборочных значений, попавших в i-й интервал и вероятности попадания в i-й

интервал

$$p_i = \int_{x_{i-1}}^{x_i} p(x,\theta)\mathrm{d}x = \\ = F(x_i) - F(x_{i-1}) \tag{9.3}$$

соответствующие теоретическому закону с интегральной функцией распределения $F(x,\theta)$ для всех m интервалов.

При этом $\sum_{i=1}^{m} n_i = n$ и $\sum_{i=1}^{m} p_i = 1$.

4. Проводим расчет статистики критерия согласия χ^2 Пирсона с помощью соотношения

$$\chi^2_{\text{exp}} = n \sum_{i=1}^{m} \frac{(n_i/n - p_i)^2}{p_i} \tag{9.4}$$

При проверке гипотезы при $n \to \infty$ для которой известны, как вид закона $p(x,\theta)$, так и все его параметры θ (простая гипотеза) функция χ^2_{exp} подчиняется распределению χ^2_r с $r = m - 1$ степенями свободы (доказано в Pearson, Karl (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 50 (302): 157—175.).

5. Далее находим из уравнения величину

$$\int_{\chi^2_{\text{exp}}}^{\infty} p(s)\mathrm{d}s = \\ \int_{\chi^2_{\text{exp}}}^{\infty} \frac{(s)^{(\frac{r}{2}-1)} e^{-\frac{s}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})}\mathrm{d}s = p \ , \text{ или} \\ 1 - F_r(\chi^2_{\text{exp}}) = p \ , \tag{9.5}$$

где $F_r(x)$ – интегральная функция вероятности распределения χ^2_r с r степенями свободы.

6. После вычисления p получаем ответ: **Данная выборка с вероятностью p соответствует теоретическому распределению $p(x,\theta)$ с параметрами θ.**

Или: **Данная выборка с вероятностью $\alpha = 1 - p$ не соответствует теоретическому распределению $p(x,\theta)$ с параметрами θ**

.

Примечания.

★ На практике (например с помощью точечных оценок) удается оценить(рассчитать) все или часть параметров распределения. Тогда статистика χ^2_{exp} при справедливости проверяемой гипотезы подчиняется χ^2_r -распределению с $r = m - k - 1$ степенями свободы, где k количество оцененных по выборке параметров.

★ Некорректное использование критериев согласия (не построен вариационный ряд для выборки, неверно выбрано число интервалов) может приводить к необоснованному принятию (чаще всего) или необоснованному отклонению проверяемой гипотезы.

★ Существуют и другие критерия согласия : критерий Колмогорова-Смирнова, критерий Мизеса и т.д.

Существует несколько более удобный способ понимания критерия χ^2. Введем понятие приведенного значения $\tilde{\chi}^2$ (или $\tilde{\chi}^2$ на одну степень свободы), которое определим как

$$\tilde{\chi}^2 = \frac{\chi^2}{r} \tag{9.6}$$

Тогда, каким бы ни было число степеней свободы, наш критерий можно сформулировать следующим образом: если мы получаем значение $\tilde{\chi}^2$ порядка 1 или меньше, то у нас нет оснований сомневаться в нашем ожидаемом распределении; если мы получаем значение $\tilde{\chi}^2$ много большее, чем единица, то невероятно, чтобы наше ожидаемое распределение было верным. Т.е. если

$$\tilde{\chi}^2 \leqslant 1 \, , \tag{9.7}$$

то наша выборка соответствует теоретическому распределению.

10 Алгоритм предварительной обработки экспериментальных данных

Исходя из изложенного материала можно сформировать необходимые этапы предварительной обработки наблюдений.

1. **Вычисление выборочных характеристик (точечные оценки моментов экспериментального распределения)** $\bar{x}$, S^2, $\widetilde{\nu}_3$ (и соответственно $\widetilde{\gamma}_1$), $\widetilde{\nu}_4$ (и соответственно $\widetilde{\varepsilon}$), а также дополнительные характеристики $S_{\bar{x}}$, и $m_{3,4}$ по формулам:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \ , \\
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \\
m_k &= \sum_{i=1}^{n} (x_i - \bar{x})^k \ , k = 2,3,4 \ , \\
\tilde{\nu}_3 &= \frac{n^2}{(n-1)(n-2)} m_3 \ , \\
\widetilde{\gamma}_1 &= \frac{\tilde{\nu}_3}{S^3} \ , \\
\tilde{\nu}_4 &= \frac{\left(n^2 - 2n + 3\right) m_4 - 3n(2n-3) m_3^2}{(n-1)(n-2)(n-3)} \ , \\
S_{\bar{x}} &= \frac{S}{\sqrt{n}} \ .
\end{aligned} \tag{10.1}$$

(Смотри формулы (3.4), (3.5), (3.8),(3.9)) и другие.

2. Цензурирование выборки (отсев грубых промахов). Последующий пересчёт точечных оценок.

3. Построение гистограмм и кумулятивной кривой. Визуальный анализ на соответствие соответствие теоретической плотности распределения с привлечением точечных оценок.

4. Проверка с помощью критериев согласия на соответствие одной (или нескольких) из теоретических плотностей распределения. Как правило, проверяют соответствие экспериментальной выборки нормальному распределению.

11 Некоторые сведения о двумерных случайных величинах

Переход от одномерных случайных величин к многомерным сопровождается введением новых понятий . Наглядные модельные представления можно построить для двух или трех переменных. Наиболее удобен случай двух переменных, и он чаще других будет использоваться в дальнейшем.

Пусть X и Y две случайные величины, имеющие набор значений x_k и y_k. Индекс k нумерует определенные значения этих величин. Для каждой из одномерных случайных величин можно ввести интегральные $F(x)$, $F(y)$ и дифференциальные функции распределения вероятностей $p(x)$ и $p(y)$. Но можно ввести и совместную функцию распределения вероятностей, описывающие поведение обеих случайных величин, как единый объект (одновременно).

***Определение* 11.1**

Совместная функция распределения $F(x,y)$ - это вероятность, того что значения двумерной величины приписанные выборочному множеству k, удовлетворяют одновременно неравенствам $x(k) \leqslant x$ и $y(k) \leqslant y$ или

$$F(x,y) = \text{Prob}(x_k \leqslant x \text{ и } y_k \leqslant y) \ . \tag{11.1}$$

Легко найти, что

1.
$$F(-\infty,y) = 0 \ , \tag{11.2}$$

2.
$$F(x,-\infty) = 0 \ , \tag{11.3}$$

3.
$$F(\infty,-\infty) = 1 \ . \tag{11.4}$$

Совместный дифференциальный закон распределения $p(x,y)$ определяется соотношением:

$$p(x,y) = \frac{\partial^2}{\partial y \partial x}\left[F(x,y)\right] \tag{11.5}$$

Очевидно что

1.
$$p(x,y) \geqslant 0 \ , \tag{11.6}$$

2.

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} p(\xi,\eta)\,\mathrm{d}\xi\mathrm{d}\eta\,, \tag{11.7}$$

Плотность вероятности для случайных величин X и Y в отдельности, выражаются через совместную функцию распределений с помощью соотношений:

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,\mathrm{d}y\,, \tag{11.8}$$

$$p(y) = \int_{-\infty}^{\infty} p(x,y)\,\mathrm{d}x\,. \tag{11.9}$$

Функции (11.8) и (11.9) называют безусловными или маргиальными плотностями распределения случайных величин X и Y.

Определение 11.2

Две случайные величины X и Y являются статистически независимыми (или взаимно независимыми), если

$$p(x,y) = p(x)\,p(y)\,. \tag{11.10}$$

Математическое ожидание

Если имеется функция случайных величин $G(x,y)$, то ее математическое ожидание

$$E\{G(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(x,y)p(x,y)\,\mathrm{d}x\mathrm{d}y\,. \tag{11.11}$$

Дисперсия

Дисперсия $D\{G(x,y)\}$ функция случайных величин $G(x,y)$ выражается следующим образом:

$$D\{G(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (G(x,y) - E\{G(x,y)\})^2\,p(x,y)\,\mathrm{d}x\mathrm{d}y\,. \tag{11.12}$$

11.1 Алгебраические и центральные моменты

Алгебраические имеют соответственно следующий вид (и частном случае для (11.11), когда $G = x$ или $G = y$):

$$E\{x\} = \mu_x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)\,\mathrm{d}x\mathrm{d}y\,, \tag{11.13}$$

$$E\{y\} = \mu_y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp(x,y)\,\mathrm{d}x\mathrm{d}y\,. \tag{11.14}$$

My title

С учетом определений (11.8) и (11.9) для безусловных функций распределения имеем, что

$$E\{x\} = \mu_x = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x\,, \tag{11.15}$$

$$E\{y\} = \mu_y = \int_{-\infty}^{\infty} yp(y)\,\mathrm{d}y\,. \tag{11.16}$$

Центральные моменты, по аналогии с одномерной случайной величиной, могут быть записаны в виде

$$D\{x\} = \sigma_x^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E\{x\})^2\, p(x,y)\,\mathrm{d}x\mathrm{d}y\,, \tag{11.17}$$

$$D\{y\} = \sigma_y^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - E\{y\})^2\, p(x,y)\,\mathrm{d}x\mathrm{d}y\,. \tag{11.18}$$

Но кроме этих простейших моментов второго порядка (11.17) и (11.18), имеющих уже хорошо известную форму дисперсий, появляются новые возможности для образования моментов второго порядка.

Ковариация

Так можно определить дисперсию $D\{x,y\}$ посредством соотношения:

$$D\{x,y\} = \operatorname{cov}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E\{x\})(y - E\{y\})\, p(x,y)\, \mathrm{d}x\mathrm{d}y\ . \tag{11.19}$$

***Определение* 11.3**

Центральный момент (11.19) называются ***ковариацией*** *между x и y.*

Если x и y статистически независимы (см. (11.10)), т. е. $p(x,y) = p(x)p(y)$, тогда ковариация обращается в нуль, Действительно, после очевидных преобразований, имеем, что

$$\operatorname{cov}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E\{x\})(y - E\{y\})\, p(x)\, p(y)\, \mathrm{d}x\mathrm{d}y =$$

$$= \int_{-\infty}^{\infty} p(x)(x - E\{x\})\, \mathrm{d}x \times \int_{-\infty}^{\infty} (y - E\{y\})\, p(y)\, \mathrm{d}y = 0\ . \tag{11.20}$$

Во всех других случаях смешанный момент второго порядка не равен нулю. Обычно вводится безразмерная переменная - **коэффициент корреляции** $\rho(x,y)$, определяемый как

$$\rho(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}\ . \tag{11.21}$$

Можно, показать, что

$$-1 \le \rho(x,y) \le 1\ . \tag{11.22}$$

На практике бывает необходимо оценить коэффициент корреляции $\rho(x,y)$ для конкретной выборки объема n. Выборочная ковариация (несмещенная точечная оценка) $r(x,y)$ можно определить по формуле

$$r(x,y) = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_x}\frac{(y - \bar{y})}{S_y} p_i\ . \tag{11.23}$$

Здесь $\bar{x}$ и $\bar{y}$ - средние значения случайных переменных; p_i - частота попадания переменных x_i и y_i в i-ый интервал значений, а $S_{x,y}$ - точечные оценки среднеквадратичных отклонений $\sigma_{x,y}$.

Определение 11.4

Ковариация (11.19) имеет смешанный характер и вместе с остальными моментами образует матрицу (11.17) и (11.18), которую называют ***ковариационной матрицей****, или матрицей вторых моментов, или дисперсионной матрицей, или матрицей ошибок.*

Ковариационная матрица

$$\mathbf{V} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} D\{x\} & D\{x,y\} \\ D\{y,x\} & D\{y\} \end{pmatrix}. \tag{11.24}$$

Поскольку $D\{x,y\} = D\{y,x\}$, то матрица корреляции симметрична, диагональные элементы которой представляют собой дисперсии.

Каждый элемент ковариационной матрицы $\mathbf{V}_{ij}$ можно трактовать как математическое ожидание ij-элемента произведения вектор-столбца $(\mathbf{x}-\boldsymbol{\mu})$ на вектор-строку $(\mathbf{x}-\boldsymbol{\mu})^T$, где

$$(\mathbf{x}-\boldsymbol{\mu}) = \begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}$$
$$(\mathbf{x}-\boldsymbol{\mu})^T = \begin{pmatrix} x-\mu_x & y-\mu_y \end{pmatrix} \tag{11.25}$$

Теперь в матричной записи (11.24) имеет вид

$$\mathbf{V} = E\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\}. \tag{11.26}$$

Корреляционные связи имеют важное значение в исследованиях. Удобство работы с корреляциями связано с тем, что корреляция - безразмерная величина в отличие от ковариации. Из коэффициентов корреляции можно также построить корреляционную матрицу $\mathbf{R}$. В двумерном случае, в следствие, того, что

$$\rho(x,x) = \rho(y,y) = 1 \tag{11.27}$$

имеем, что

$$\mathbf{R} = \begin{pmatrix} \rho(x,x) & \rho(x,y) \\ \rho(x,y) & \rho(y,y) \end{pmatrix} = \\ = \begin{pmatrix} 1 & \dfrac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} \\ \dfrac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} & 1 \end{pmatrix} . \tag{11.28}$$

Важно, что если x и y Образуют двумерную случайную величину, то дисперсия линейной функции $z = x + y$ находится с помощью соотношения:

$$D\{x+y\} = D\{x\} + D\{y\} + 2\operatorname{cov}(x,y) . \tag{11.29}$$

11.2 Примеры многомерных функций распределения вероятностей

Пусть имеем многомерную случайную величину $\{X_1, X_2, \ldots, X_k\}$ размерности k. Для центральных и алгебраических моментов распределения, а также ковариационной матрицы введем следующие обозначения:

$$\begin{aligned} &\mathrm{E}\{X_i\} = \mu_i \ , i = 1, \ldots, k \ , \\ &\mathrm{D}\{X_i\} = \sigma_i \ , \\ &\operatorname{cov}(X_i, X_j) = \rho^{ij} \sigma_i \sigma_j \ . \end{aligned} \tag{11.30}$$

Введем дополнительные k-мерные векторы: $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$ и $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_k\}$ и матрицу ошибок

$$\mathbf{V}_{ij} = \begin{cases} \sigma_i^2, & \text{если } i = j \\ \rho_{ij} \sigma_i \sigma_j & \text{если } i \neq j \end{cases} . \tag{11.31}$$

Многомерное нормальное распределение

Тогда функция распределения

$$p(\mathbf{X}) = \frac{1}{(2\pi)^{k/2}\,|V|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right] \quad ,(11.32)$$

где $|V|$ -детерминант ковариационной матрицы $\mathbf{V}$ определяет k-мерную плотность нормального распределения.

Наиболее наглядным вариантом является двумерное нормальное распределение, которое определяется параметрами $\mu_1, \mu_2, \sigma_1, \sigma_2$ и ρ_{12}:

$$p_N(x,y) = \frac{1}{2\pi\sqrt{(1-\rho_{12}^2)\,\sigma_1^2\sigma_2^2}} \times$$
$$\times \exp\left\{\frac{\sigma_2^2(x-\mu_1)^2 - 2\rho_{12}\sigma_2\sigma_1(x-\mu_1)(y-\mu_2) + \sigma_1^2(y-\mu_2)^2}{2(\rho_{12}^2-1)\,\sigma_1^2\sigma_2^2}\right\} \quad (11.33)$$

Если $\mu_1 = \mu_2 = 0$ и $\sigma_1 = \sigma_2 = 1$, а $\rho_{12} = \rho$, то из (11.33) получим стандартизованное двумерное нормальное распределение:

$$p_N(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{\frac{x^2 - 2\rho xy + y^2}{2(\rho^2-1)}\right\} . \quad (11.34)$$

График стандартизированного двухмерного нормального распределения с $\rho = 0$ (отсутствуют корреляции между x и y), представлен на рисунке 11.2.

Естественно, существуют и другие теоретические функции распределения многомерных случайных величин.

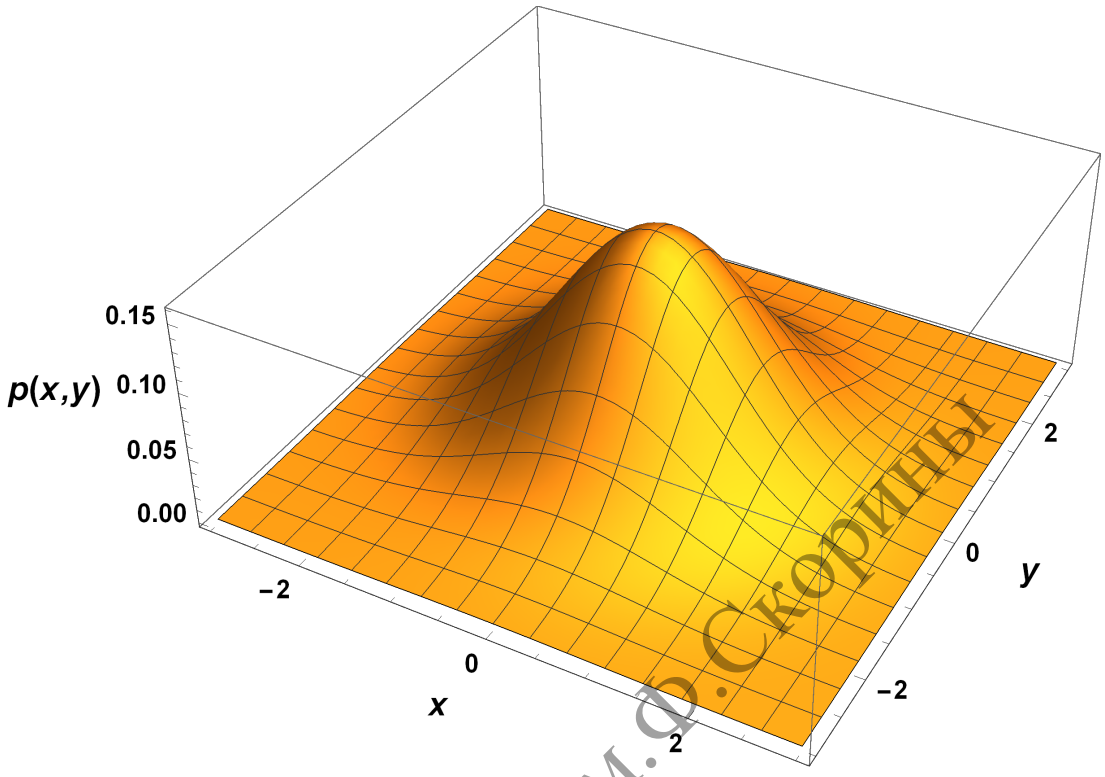

Рисунок 10– Двумерное нормальное распределение с $\mu_1 = \mu_2 = 0$ и $\sigma_1 = \sigma_2 = 1$, а $\rho = 0$

12 Корреляционный и регрессионный анализ

Статистические связи между случайными величинами можно изучать методами корреляционного и регрессионного анализа. Основной целью регрессионного анализа является установление формы и изучение зависимости между переменными.

Целью корреляционного анализа является определить степень взаимосвязи между исследуемыми случайными величинами. В случае, если исследуется связь двух переменных, корреляционный анализ называют парным; если число переменных более двух — множественным.

Статистическая зависимость между переменными, при которой каждому значению одной или нескольких случайных величин соответствует определенное среднее значение другой, называется **корреляционной**.

Корреляционный и регрессионный анализ решает две основные задачи:

1. Определение уравнения (или системы уравнений) связи, т.е. в установлении математической формы, в которой выражается данная связь. Это очень важно, так как от правильного выбора формы связи зависит конечный результат изучения взаимосвязи между признаками.

2. Вычисление степени взаимосвязи, т.е. меры связи между признаками с целью установить степень влияния данных случайных величин (факторов) на конечную случайную величину (результат). Эта задача решается путем определения параметров корреляционного уравнения различными математическими методами.

Далее проводятся оценка и анализ полученных результатов при помощи различных показателей регрессионно-корреляционного метода (коэффициентов детерминации, линейной и множественной корреляции и т.д.), а также проверка существенности связи между изучаемыми признаками.

В результате использование корреляционного и регрессионного анализа можно решить такие задачи, как

– Взаимосвязь. Есть ли взаимосвязь между случайными величинами, и насколько велико их влияние на конечную величину.

– Прогнозирование. Если известно поведение одного параметра, то можно предсказать поведение другого параметра, коррелирующего с первым в области, где нет экспериментальной информации.

– Обнаружение неизвестных причинных связей.

Более часто используемым показателем степени тесноты корреляционной связи является линейный коэффициент корреляции. При расчете этого показателя учитываются не только отклонения индивидуальных значений случайной величины от среднего значения, но и сама величина этих отклонений.

В самом общем случае, при большом числе наблюдений одно и то же значение случайной величины X может встретиться n_x раз, одно и то же значение случайной величины Y может встретиться n_y раз, а одна и та же пара чисел $\{x, y\}$ может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т.е. подсчитывают частоты n_x, n_y и n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют корреляционной.

Таблица 1– Пример корреляционной таблицы

y\x	1	2	3	4	5	n_y
0	–	–	–	6	4	10
1	–	–	1	4	6	11
2	–	5	9	5	–	19
3	3	7	–	–	–	10
n_x	3	12	10	15	10	50

В первой строке указаны наблюдаемые значения $\{1,2,3,4,5\}$ случайной величины X, а в первом столбце таблицы – наблюдаемые значения $\{0,1,2,3\}$ случайной величины Y.

На пересечении строк и столбцов находятся частоты n_{xy} наблюдаемых пар значений случайных величин X и Y. Например, частота 7 указывает, что пара чисел $\{2,3\}$ наблюдалась 7 раз. Все частоты помещены в прямоугольнике. В последнем столбце записаны суммы частот строк. В последней строке записаны суммы частот столбцов. Общее число наблюдений $n = 50$.

Если имеем k значений случайной величины X и m значений случайной величины Y в корреляционной таблице с числом попаданий n_i, n_j и n_{ij} для x_i , y_i и пар $\{x_i,y_j\}$ соответственно, то оценку коэффициент корреляции можно в данном случае можно найти из выражения:

$$r_{xy} = \frac{n\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{m} x_i y_j n_{ij} - \left(\sum\limits_{i=1}^{k} x_i n_i\right)\left(\sum\limits_{j=1}^{m} y_j n_j\right)}{\sqrt{n\sum\limits_{i=1}^{k} x_i^2 n_i - \left(\sum\limits_{i=1}^{k} x_i n_i\right)^2}\sqrt{n\sum\limits_{j=1}^{m} y_j^2 n_j - \left(\sum\limits_{j=1}^{m} y_j n_j\right)^2}}\,. \tag{12.1}$$

В упрощенном варианте, когда имеется n значений пар $\{x_i, y_i\}$ двумерной случайной величины, выражение коэффициент корреляции (12.1) упрощается и приобретает вид:

$$r_{xy} = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \left(\sum\limits_{i=1}^{n} x_i\right)\left(\sum\limits_{j=1}^{n} y_j\right)}{\sqrt{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{k} x_i\right)^2}\sqrt{n\sum\limits_{j=1}^{n} y_j^2 - \left(\sum\limits_{j=1}^{n} y_j\right)^2}}\,. \tag{12.2}$$

Последний вариант часто встречается в физических исследованиях.

Иногда одному значению x_i соответствует m значений $\{\tilde{y}_{i1},\tilde{y}_{i2},\ldots,\tilde{y}_{i,k}\}$ величины Y. Тогда, после нахождения среднего значения и среднеквадратичного отклонения по известным формулам:

$$y_i = \frac{1}{m}\sum_{j=1}^{m}\tilde{y}_{ij}\ ,$$

$$s_{y_i} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(y_i - \tilde{y}_{ij})^2}\ , \tag{12.3}$$

наборам пар $\{x_i, y_i\}$ с дополнительным набором ошибок $\{s_{y_1}, \ldots, s_{y_n}\}$. Для определения степени связи используют обычно таблицу Чеглока (смотри таблицу 2).

Таблица 2– Пример корреляционной таблицы

Значение $\lvert r_{xy}\rvert$	Степень связи
0,1 – 0,3	слабая
0,3 – 0,5	умеренная
0,5 – 0,7	заметная
0,7 – 0,9	высокая
0,9 – 0,99	весьма высокая связь

13 Линейный регрессионный анализ

Корреляционный анализ позволяет установить степень взаимосвязи двух и более случайных величин. Однако наряду с этим желательно иметь модель этой связи, которая дала бы возможность предсказывать значения одной случайной величины по конкретным значениям другой. Методы решения подобных задач составляют раздела математической статистики “регрессионный анализ”.

Для упрощения изложения рассмотрим случай двух случайных величин x и y. Линейная связь между двумя случайными величинами означает, что прогноз значения y по данному значению x имеет вид:

$$\hat{y} = A + Bx \tag{13.1}$$

Если данные связаны идеальной линейной зависимостью $|r_{xy}| = 1$, то предсказанное значение $\tilde{y}_i$ будет соответствовать наблюдаемому значению y_i при любом данном x_i. На практике идеальная зависимость отсутствует (за счет случайных разбросов данных, за счет нелинейных эффектов).

Однако, если предположить наличие линейной связи, то можно подобрать A и B, которые дадут возможность предсказывать ожидаемое значение y_i для любого данного x_i (при этом предсказанное $\tilde{y}_i$ не совпадает с наблюдаемыми y_i, однако оно будет равно среднему значению всех таких наблюденных значений).

Наиболее общепринятая процедура определения коэффициентов A и B состоит в таком выборе, который минимизирует сумму квадратов отклонений наблюденных значений от предсказанного значения y_i.

Этот метод называют методом наименьших квадратов (`МНК`). Он разработан в 1795-1805 гг. Лежандром и Гауссом.

13.1 Метод наименьших квадратов

Рассчитывается отклонение $y_i - \hat{y}_i = \Delta i$, а затем находятся коэффициенты так, что

$$Q = \sum_{i=1}^{n} \Delta_i^2 \to \min \tag{13.2}$$

Для этого надо взять частные производные функции Q (13.2) по параметрам A и B и приравнять их к нулю.

В итоге имеем, что

$$\begin{cases} \dfrac{\partial Q}{\partial A} = 0, \\ \dfrac{\partial Q}{\partial B} = 0 \end{cases} . \tag{13.3}$$

Из этой системы получаем оценки пар A и B.

Проделав цепочку преобразований

$$\frac{\partial Q}{\partial A} = -2\sum_{i=1}^{n}(y_i - A - Bx_i) = 0 \, . \tag{13.4}$$

$$An + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \, ,$$

$$\frac{\partial Q}{\partial B} = -2\sum_{i=1}^{n}(y_i - A - Bx_i)\,x_i = 0 \, ,$$

$$-\sum_{i=1}^{n} y_i x_i + A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 = 0 \, ,$$

$$An + B\left(\sum_{i=1}^{n} x_i\right) = \left(\sum_{i=1}^{n} y_i\right) \, ,$$

$$\left(\sum_{i=1}^{n} x_i\right) A + B\left(\sum_{i=1}^{n} x_i^2\right) = \left(\sum_{i=1}^{n} x_i y_i\right)$$

в итоге имеем:

$$\begin{cases} A = \dfrac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}, \\ B = \dfrac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} \end{cases} . \tag{13.5}$$

Эти оценки можно использовать для модели $\hat{y} = A + Bx$, которую называют **прямой линейной регрессии y на x**.

13.2 Свойства метода наименьших квадратов

Из первого уравнения имеем:

$$A + B\sum_{i=1}^{n}\frac{x_i}{n} = \sum_{i=1}^{n}\frac{y_i}{n} = A + B\overline{x} = \overline{y}\ , \tag{13.6}$$

т.е. кривая регрессии проходит через центр тяжести экспериментальных точек.

Если теперь до проведения МНК все исходные данные отцентрировать, т.е. $\overline{x}$ и $\overline{y}$ перенести в начало координат $\overline{x}$=$\overline{y}$=0, то первое уравнение превратится в тождество, т.е. система уравнений сокращается, что особенно важно, когда имеет место ситуация с большим числом коэффициентов.

Уравнение прямой можно записать:

$$\hat{y} = A + Bx = \overline{y} + B(x - \overline{x})\ . \tag{13.7}$$

Вторая особенность МНК состоит в том, что полученные этим методом оценки необратимы, т.е. если имеется модель регрессии y на x: $y = A + Bx$, то регрессию x на y нельзя обратить:

$$x = \frac{(y - A)}{B} = \frac{1}{b}y - A\ . \tag{13.8}$$

Коэффициент наклона при y обозначим как B'. Тогда имеем, что

$$x = B'y - A\ . \tag{13.9}$$

Можно найти по аналогии с вышеизложенным, что

$$B' = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum\limits_{i=1}^{n} y_i^2 - n\overline{y}^2}\ . \tag{13.10}$$

Коэффициент (13.10) связан с B через коэффициент корреляции r_{xy} соотношением

$$\sqrt{B'B} = r_{xy}\ . \tag{13.11}$$

13.3 Точность оценок A и B.

Пусть $E\{A\} = a$, где a – "истинное" значение. Тогда доверительные интервалы для коэффициентов линейной регрессии A и B запишутся в виде:

$$a - A = \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right)^{\frac{1}{2}} S_{y/x} t_{n-2,\alpha/2} \ ,$$

$$b - B = \left(\frac{1}{\sum\limits_{i=1} \left(x_i - \overline{x}^2\right)} \right)^{\frac{1}{2}} S_{y/x} t_{n-2,\alpha/2} \ , \tag{13.12}$$

где

$$S_{y/x} = \left[\frac{\sum\limits_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \right]^{\frac{1}{2}} = \left[\left(\frac{n-1}{n-2} \right) S_y^2 \left(1 - r_{xy}^2\right) \right]^{\frac{1}{2}} \ , \tag{13.13}$$

а $t_{n-2,\alpha/2}$ - квантили распределения Стьюдента.

13.4 Редукция некоторых задач нелинейного регрессионного анализа к линейному

Как сводить некоторые задачи регрессионного анализа к линейному регрессионному анализу.

Например, модель типа

$$\frac{1}{y} = a_1 + b_1 x$$

путем замены переменных $\frac{1}{y} = y_1$ имеем $y_1 = a_1 + b_1 x$.

Аналогично

$$y = a_1 + \frac{b_1}{x}\frac{1}{x} \to x \ ,$$

$$\frac{y}{x} = a_1 + b_1 x \ ,$$

$$y = ax^b \lg y = \lg a + b \lg x \ ,$$

$$\lg y \to y \lg x \to x \ ,$$

$$y = ab^x \lg y = \lg a + x \lg b \ .$$

Такие преобразования, строго говоря, допустимы, если исходные величины измерены точно.Например: $y = Ax^B$измерен с Δy: $y = Ax^B + \Delta y$.

Тогда линеаризированная форма будет отличаться от исходного за счет неясности преобразования Δy. Однако в первичном анализе Δy можно пренебречь, если это необоснованно, то вводим переменную y' путем

$$y' = y - \Delta y = Ax^B \ .$$

Метод МНК достаточно чувствителен к неоднородностям статистики. Наличие неоднородной статистики приводит иногда к абсурдным результатам, поэтому окончательное решение об МНК должно приниматься по однородной, очищенной статистике.

14 Элементы нелинейного регрессионного анализа

Пусть одному значению x_i из выборки объема n соответствует m значений $\{\tilde{y}_{i1},\tilde{y}_{i2},\ldots,\tilde{y}_{i,k}\}$ величины Y. Тогда, после нахождения среднего значения и среднеквадратичного отклонения, получим набор пар $\{x_i, y_i\}$ с дополнительным набором ошибок $\{\sigma_{y_1},\ldots,\sigma_{y_n}\}$. Такая ситуация возникает, например, при построении гистограмм.

Цель регрессионного анализа найти уравнение взаимосвязи случайных величин Y и X (в данном случае, парная регрессионная модель) вида: $y = \phi(x;\boldsymbol{\theta})$. В силу воздействия неучтенных случайных факторов отдельные значения y_i будут в большей или меньшей мере отклоняться от функции регрессии $\phi(x;\boldsymbol{\theta})$. В этом случае уравнение взаимосвязи двух переменных (парная регрессионная модель) может быть представлено в виде:

$$y = \phi(x;\boldsymbol{\theta}) + \epsilon\,, \tag{14.1}$$

где $\epsilon = \{\epsilon_1,\ldots,\epsilon_n\}$ случайная величина, характеризующая отклонение от функции регрессии. Эту переменную часто называют возмущением; величина $\boldsymbol{\theta} = \{\theta_1,\theta_2,\ldots,\theta_k\}$ определяет набор параметров регрессионной модели, которые необходимо определить.

Основные положения регрессионного анализа:

Основные положения регрессионного анализа

1. Одним из основных требований регрессионного анализа является, условие равенства нулю математического ожидания "возмущения" ϵ:
$$\mathrm{E}\{\epsilon_i\} = 0\,. \tag{14.2}$$
2. Также предполагают, что случайные величины ϵ_i подчиняются одномерному нормальному распределению, если все y_i и y_j не коррелируют с другом или многомерному нормальному распределению, если такая корреляция имеет место.
3. Дисперсия "возмущения" ϵ_i (или зависимой переменной y для любого i) задается соотношением:
$$\mathrm{D}\{\epsilon_i\} = \sigma_i^2\,. \tag{14.3}$$
Наиболее простой вариант (14.3) состоит в требовании того, чтобы все дисперсии отклонений были одинаковые.

В случае постоянства дисперсии и отсутствия корреляции, оценки параметров регрессии полученные, например, с помощью метода наименьших квадратов, обладают важными свойствами, а именно:

– несмещенность;

– состоятельность;

– эффективность.

В случае нелинейной зависимости, свойства постоянства дисперсии и отсутствия корреляции могут не выполняться, тогда полученные оценки параметров регрессии не будут обладать указанными характеристиками. Или связь между переменными x и y линейна, но на исследуемый показатель воздействует фактор, не включенный в модель.

Для решения этой задачи в случае линейной регрессии использовался метод наименьших квадратов. Рассмотрим некоторые методы поиска оптимальной оценки $\widehat{\boldsymbol{\theta}}$ в случае нелинейной регрессии. Естественно, эти методы могут быть использованы и в линейном варианте функциональной зависимости.

14.1 Метод максимального правдоподобия

Пусть имеется выборка из n независимых результатов наблюдений $x = \{x_1,x_2,\ldots,x_n\}$ с с плотностью вероятности $p\,(x,\boldsymbol{\theta})$ ($\boldsymbol{\theta}$ - набор параметров).

***Определение* 14.1**

Совместная функция вероятности

$$L\,(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}) = p\,(x_1,\boldsymbol{\theta})\,p\,(x_2,\boldsymbol{\theta})\cdots p\,(x_n,\boldsymbol{\theta}) = \prod_{i=1}^{n} p\,(x_i,\boldsymbol{\theta}) \qquad (14.4)$$

называют ***функцией максимального правдоподобия.***

Совместная функция вероятности $L\,(x_1,x_2,\ldots,x_n,\boldsymbol{\theta})$ может быть представлена в виде (14.4) в силу независимости результатов каждого измерения. Две функции правдоподобия являются равными, если одна есть произведение второй на некоторую скалярную величину.

$L\,(x_1,x_2,\ldots,x_n,\boldsymbol{\theta})$ рассматривают как функцию $\boldsymbol{\theta}$ при фиксированных, полученных в измерениях x. Чем больше значение $L\,(x_1,x_2,\ldots,x_n,\boldsymbol{\theta})$, тем

более вероятна или правдоподобна, выборка значений $\{x_1,x_2,\ldots,x_n\}$ при заданном значении $\boldsymbol{\theta}$. Поэтому $L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)$ и называют функцией правдоподобия.

Метод максимального правдоподобия или метод наибольшего правдоподобия (ММП, ML, MLE — англ. maximum likelihood estimation) в математической статистике — это метод оценивания неизвестного параметра путём максимизации функции правдоподобия. Основан на предположении о том, что вся информация о статистической выборке содержится в функции правдоподобия.

Метод максимального правдоподобия состоит в поиске максимума функции $L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)$, при этом $\{x_1,x_2,\ldots,x_n\}$ считаются постоянными, а параметр $\boldsymbol{\theta}$ - переменным. Далее находят такое значение $\boldsymbol{\theta}$, при котором функция $L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)$ становится наиболее правдоподобной, т.е. **принимает максимальное значение**.

Для эффективного использования ММП требуются достаточно большие объемы выборок, точное знание анализируемого закона распределения, достаточно устойчивые распределения и не очень большое число неизвестных параметров.

Доказано, что вторая производная от функции правдоподобия $L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)$ меньше нуля и, таким образом, равенство нулю первой производной дает действительно максимальное значение $L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)$. Максимум определяют по стандартной методике, однако вместо самой функции для удобства вычислений берут логарифм функции и ищут его максимум. Поскольку максимумы функции правдоподобия и логарифмической функции совпадают, то для удобства вычислений берут логарифм функции и ищут его максимум.

Введем обозначение

$$\ell\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)=\ln L\left(x_1,x_2,\ldots,x_n,\boldsymbol{\theta}\right)=\sum_{i=1}^{n}\ln p\left(x_i,\boldsymbol{\theta}\right)\ . \qquad (14.5)$$

Система уравнений для получения оценок параметров $\boldsymbol{\theta}=\{\theta_1,\theta_2,\ldots,\theta_k\}$

можно записать в виде

$$\begin{cases} \dfrac{\partial \ell\left(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}\right)}{\partial \theta_1} = 0 \ , \\ \dfrac{\partial \ell\left(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}\right)}{\partial \theta_2} = 0 \ , \\ \cdots \\ \dfrac{\partial \ell\left(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}\right)}{\partial \theta_k} = 0 \ , \end{cases} \tag{14.6}$$

Решение системы уравнений (14.6) дает набор оптимальных значений параметров $\widetilde{\boldsymbol{\theta}} = \left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$.

Наибольшие трудности возникают не при оценке параметров с помощью решения системы (14.6), а при выявлении погрешностей, с которыми эти оценки сделаны. Возможные (неявные и естественные) связи между параметрами $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ приводят к необходимости учитывать корреляции (ковариации) между оценками различных параметров $\widetilde{\boldsymbol{\theta}} = \left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$.

Введем обозначения для сокращения записи соотношений:

$$\begin{aligned} &\ell\left(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}\right) = \ell\left(\boldsymbol{\theta}\right) \ , \\ &L\left(x_1, x_2, \ldots, x_n, \widetilde{\boldsymbol{\theta}}\right) = L_{max} \ . \end{aligned} \tag{14.7}$$

Для оценки погрешностей параметров $\widetilde{\boldsymbol{\theta}} = \left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$ разложим логарифмическую функцию правдоподобия в ряд Тейлора в окрестностях точек $\left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$:

$$\begin{aligned} \ell\left(\theta\right) = \ell\left(\widetilde{\boldsymbol{\theta}}\right) + \sum_{i=1}^{k} \left(\frac{\partial \ell}{\partial \theta_i}\right)_{\widetilde{\theta}} \left(\theta_i - \widetilde{\theta}_i\right) + \\ + \frac{1}{2} \sum_{p=1}^{k} \sum_{m=1}^{k} \left(\frac{\partial^2 \ell}{\partial \theta_p \theta_m}\right)_{\widetilde{\theta}} \left(\theta_i - \widetilde{\theta}_i\right) \left(\theta_m - \widetilde{\theta}_m\right) + \ldots \end{aligned} \tag{14.8}$$

Величины $\left(\theta_i - \widetilde{\theta}_i\right)$ трактуются как дисперсии оценок максимального правдоподобия $\widetilde{\theta}_i$.

Из определения оценок $\widetilde{\boldsymbol{\theta}}$ следует, что второй член в разложении равен нулю для всех k. Пренебрегая членами разложения выше квадратичных, логарифмическую функцию правдоподобия можно записать в матричном виде:

$$\ell\left(\boldsymbol{\theta}\right) = \ln L_{max} + \frac{1}{2}\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{T}\right)\mathbf{A}\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right) \ , \tag{14.9}$$

где

$$\mathbf{A} = \left(\begin{array}{cccc} \dfrac{\partial^2 \ell}{\partial \theta_1^2} & \dfrac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \cdots & \dfrac{\partial^2 \ell}{\partial \theta_1 \theta_k} \\ \dfrac{\partial^2 \ell}{\partial \theta_1 \theta_2} & \dfrac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 \ell}{\partial \theta_2 \theta_2} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 \ell}{\partial \theta_1 \theta_k} & \dfrac{\partial^2 \ell}{\partial \theta_2 \theta_k} & \cdots & \dfrac{\partial^2 \ell}{\partial \theta_k^2} \end{array} \right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \ . \tag{14.10}$$

В асимптотическом пределе $n \to \infty$, элементы матрицы практически не зависят от конкретной выборки $\{x_1, x_2, \ldots, x_n\}$, и их можно заменить математическими ожиданиями:

$$\mathbf{B} = \mathrm{E}\left\{\mathbf{A}\right\} = \left(\begin{array}{cccc} \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_1^2}\right\} & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_1 \theta_2}\right\} & \cdots & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_1 \theta_k}\right\} \\ \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_1 \theta_2}\right\} & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_2^2}\right\} & \cdots & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_2 \theta_2}\right\} \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_1 \theta_k}\right\} & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_2 \theta_k}\right\} & \cdots & \mathrm{E}\left\{\dfrac{\partial^2 \ell}{\partial \theta_k^2}\right\} \end{array} \right)_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \ . \tag{14.11}$$

После потенцирования (14.9) следует, что функция правдоподобия имеет вид

$$L\left(\boldsymbol{\theta}\right) = L_{max} \exp\left\{\frac{1}{2}\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^{T}\right)\mathbf{B}\left(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right)\right\} \ . \tag{14.12}$$

Сопоставление с известными теоретическими распределениями приводит к выводу, что (14.12) представляет собой k-мерное нормальное распределение со средними оценками $\left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$ и ковариационной матрицей

$$\mathbf{V} = -\mathbf{B}^{-1} \ . \tag{14.13}$$

Диагональные элементы ковариационной матрицы (14.13) - являются дисперсиями $c_{11} = \sigma_{\widetilde{\theta}_1}, \ldots, c_{kk} = \sigma_{\widetilde{\theta}_k}$ оценок максимального правдоподобия $\left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$, а недиагональные элементы представляют собой ковариации между всевозможными парами оценок $c_{ij} = \mathrm{cov}\left(\widetilde{\theta}_i, \widetilde{\theta}_j\right)$.

Коэффициент корреляции между оценками определяется формулой

$$\rho\left(\widetilde{\theta}_i,\widetilde{\theta}_j\right)=\frac{\operatorname{cov}\left(\widetilde{\theta}_i,\widetilde{\theta}_j\right)}{\sigma_{\widetilde{\theta}_i}\sigma_{\widetilde{\theta}_j}}\ ,\ \ i\neq j\ . \tag{14.14}$$

Поскольку $L\left(\boldsymbol{\theta}\right)$ в окрестности $\widetilde{\boldsymbol{\theta}}$ представляет собой k-мерное нормальное распределение, то квадратичная форма

$$Q\left(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}\right)=\left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right)^T\mathbf{V}^{-1}\left(\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\right) \tag{14.15}$$

распределена по закону χ^2 с r степенями свободы. Поэтому можно определить вероятностное утверждение:

$$\operatorname{Prob}[Q\left(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}\right)\leqslant Q_\alpha]=\alpha\ , \tag{14.16}$$

где квантиль распределения χ^2 с r степенями свободы Q_α уровня α задается соотношением

$$\int_0^{Q_\alpha}\frac{2^{-r/2}}{\Gamma\left(r/2\right)}\exp\left[-x/2\right]x^{r/2-1}d\mathrm{x}=1-\alpha\ . \tag{14.17}$$

Гиперэллипсоид, определяемый уравнением

$$Q\left(\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}\right)=Q_\alpha \tag{14.18}$$

будет задавать доверительную область в пространстве переменных $\widetilde{\boldsymbol{\theta}}$ с вероятностным содержанием (доверительной вероятностью) $P=1-\alpha$.

Для приближенного построения гиперэллипсоида (14.18), который определяет возможные значения оптимальных параметров с доверительной вероятностью P, можно использовать уравнение вида (смотри (14.9)):

$$\ln L\left(\boldsymbol{\theta}\right)\approx\ln L_{max}-\Delta(\ln L\left(\boldsymbol{\theta}\right)) \tag{14.19}$$

Таким образом, как следует из (14.17) и (14.18), значение $\Delta(\ln L\left(\boldsymbol{\theta}\right))$ определяется задаваемым уровнем достоверности (C.L.) и числом параметров, входящих в набор $\boldsymbol{\theta}$.

С помощью (14.19) для доверительной вероятности $P=95\%$ можно найти для числа параметров 1, 2 и 3, числовое значение $\Delta(\ln L\left(\boldsymbol{\theta}\right)\approx 3{,}84,\ 5{,}99$

и 7,82 соответственно. Аналогичные значения для стандартного отклонения с $P = 68{,}27\%$ приводят к значениям $\Delta(\ln L(\boldsymbol{\theta}) \approx 1{,}00$, 2,30 и 3,53.

В этом случае, можно построить контурные графики функции

$$\ln L(\boldsymbol{\theta}) - \ln L_{max} = \Delta(\ln L(\boldsymbol{\theta})) \tag{14.20}$$

для различных пар параметров и найти среднеквадратичные отклонения $\sigma_{\widetilde{\theta}_1}, \ldots, \sigma_{\widetilde{\theta}_k}$ оценок максимального правдоподобия $\left\{\widetilde{\theta}_1, \widetilde{\theta}_2, \ldots, \widetilde{\theta}_k\right\}$.

14.2 Метод наименьших квадратов

Метод наименьших квадратов также используется для получения "наилучших" (оптимальных) значений модельных параметров при описании экспериментальных значений.

Метод наименьших квадратов совпадает с методом максимального правдоподобия в следующем частном случае. Рассмотрим набор из n независимых измерений y_i в известных точках x_i.

Из y_i из $\mathbf{y} = \{y_1 \pm \sigma_1, \ldots y_n \pm \sigma_k\}$ подчиняются подчиняются нормальному распределению с математическим ожиданием $\phi(x_i, \boldsymbol{\theta})$ и известной дисперсиейσ_i^2. Таким образом предполагается, что описание значений y_i с помощью модельной регрессионной кривой $\phi(x_i, \boldsymbol{\theta})$ должно быть оптимальным. Цель состоит в том, чтобы построить оценки для неизвестных параметров $\boldsymbol{\theta}$.

Для получения оптимальных значений $\widetilde{\boldsymbol{\theta}}$ набора модельных параметров $\boldsymbol{\theta}$ используют функцию

$$\chi^2(\boldsymbol{\theta}) = -2 \ln L(\boldsymbol{\theta}) + \text{const} = \sum_{i=1}^{n} \left[\frac{y_i - \phi(x_i, \boldsymbol{\theta})}{\sigma_i}\right]^2 . \tag{14.21}$$

Соотношение (14.21) можно получить из определения функции максимального правдоподобия (14.1), используя плотность нормального распределения

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i - \phi(x_i, \boldsymbol{\theta}))}{2\sigma_i^2}\right] . \tag{14.22}$$

Исходя из требования минимального значения функции $\chi^2(\boldsymbol{\theta})$ при $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}$, т. е.

$$\chi^2\left(\widetilde{\boldsymbol{\theta}}\right) = \chi^2_{min} \tag{14.23}$$

получаем систему уравнений

$$\begin{cases} \dfrac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_1} = 0 , \\ \dfrac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_2} = 0 , \\ \cdots \\ \dfrac{\partial \chi^2(\boldsymbol{\theta})}{\partial \theta_k} = 0 , \end{cases} \tag{14.24}$$

для определения "оптимальных" значений $\widetilde{\boldsymbol{\theta}}$. Процедура аналогична линейному регрессионному анализу, с той лишь разницей, что поиск 'оптимальных" значений $\widetilde{\boldsymbol{\theta}}$, как правило, находится численно, а не аналитически.

Значение χ^2_{min} является мерой уровня согласия между экспериментальными данными и подогнанной кривой (fit):

$$\chi^2_{min} = \sum_{i=1}^{n} \left[\frac{y_i - \phi\left(x_i, \widetilde{\boldsymbol{\theta}}\right)}{\sigma_i} \right]^2 . \tag{14.25}$$

Поэтому χ^2_{min} можно использовать как статистику соответствия предполагаемую функциональную форму $\phi\left(x_i, \widetilde{\boldsymbol{\theta}}\right)$. Известно, что (14.25) имеет распределение χ^2 с $n_d = n - k$ степенями свободы.

В этом случае, в соответствии с разделом проверки гипотез, если $\chi^2_{min}/n_d \leq 1$, то совпадение будет "хорошим". Или можно найти вероятность соответствия модели $\phi\left(x_i, \widetilde{\boldsymbol{\theta}}\right)$ экспериментальным данным P (p-value) из соотношения:

$$P = \frac{1}{2^{n_d} \int\limits_{\chi^2_{min}}^{\infty} \Gamma\{n_d/2\}} t^{n_d/2-1} \exp(-t/2) . \tag{14.26}$$

Следующим шагом решения данной задачи состоит в нахождении доверительных интервалов (ошибок $\sigma_{\widetilde{\theta}_i}$) для оптимальных значений $\widetilde{\boldsymbol{\theta}} = \left\{\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k\right\}$. Эта задача решается так же как и для метода максимального правдоподобия, если предположить, что оценки $\widetilde{\theta}_i, i = 1, \ldots, k$ подчиняются k-мерному нормальному распределению (14.12).

Отличием, состоит в том, что приближенного построения гиперэллипсоида (14.18), который определяет возможные значения оптимальных парамет-

ров с вероятностным содержанием C.L., используется уравнение вида:

$$\chi^2(\boldsymbol{\theta}) = \chi^2_{min} + \Delta\chi^2_{\text{crit}} \ . \tag{14.27}$$

Значение $\Delta\chi^2_{\text{crit}}$ определяется задаваемым уровнем достоверности p и числом параметров, входящих в набор $\boldsymbol{\theta}$ и совпадает с $\Delta(\ln L(\boldsymbol{\theta}))$.

Замечание.

Если y_i не являются независимыми, и имеют ковариационную матрицу $\mathbf{V}_{ij} = \operatorname{cov}(y_i, y_j)$, то наилучшие оценки параметров $\widetilde{\boldsymbol{\theta}} = \left\{\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k\right\}$ путем поиска минимума функционала

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - \phi(x_i, \boldsymbol{\theta}))(\mathbf{V})^{-1}_{ij}(y_j - \phi(x_j, \boldsymbol{\theta})) \ . \tag{14.28}$$

15 Сравнение моделей

При поиске зависимости между измеряемыми переменными по выборочным данным наиболее важной задачей является поиск модели регрессии. Построение эмпирического уравнения регрессии - начальный этап анализа. При этом, на первом этапе не всегда удается получить “оптимальную” модель регрессии. Тогда для описания зависимости между случайными величинами может использоваться несколько моделей (уравнений регрессии). Поэтому возникает задача об оценке “качества” полученных моделей.

Качество модели оценивается по следующим направлениям:

1. Содержательная оценка качества модели.
2. Статистическая оценка качества модели.

Содержательный анализ подразумевает анализ физического (экономического и т.д.) смысла модели. Результат данной работы должен быть ответы на вопросы вида:

✓ *Являются ли те факторы, которые используются в модели, значимыми для описания данного физического явления или процесса?*

✓ *Какой физический смысл параметров регрессионного уравнения (модели)?*

Статистическая оценка качества модели включает следующие этапы:

✓ Проверка статистической значимости параметров уравнения регрессии при помощи различных критериев.

✓ Проверку общего качества уравнения регрессии.

✓ Проверку свойств данных, выполнение которых предполагалось при оценивании уравнения.

15.1 Проверка статистической значимости параметров уравнения регрессии

После того, как найдено уравнение регрессии (например,линейной регрессии (13.1)), проводится оценка значимости как уравнения в целом, так и отдель-

ных его параметров.

Оценка значимости уравнения регрессии в целом дается с помощью F-критерия Фишера. При этом выдвигается нулевая гипотеза о том, что коэффициент регрессии B равен нулю, т. е. $H_0 : B = 0$. Следовательно, фактор случайной величины x не оказывает влияния на результат y.

Перед расчетом критерия проводятся анализ дисперсии. Общая сумма квадратов отклонений (СКО) y от среднего значения $\bar{y}$ раскладывается на две части - "объясненную" и "необъясненную":

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}_{\text{Общая СКО}} = \overbrace{\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y}_i)^2}_{\text{Факторная СКО}}}^{\text{(Объясненная)}} + \overbrace{\underbrace{\sum_{i=1}^{n}(y_i-\hat{y})^2}_{\text{Остаточная СКО}}}^{\text{(Необъясненная)}} . \qquad (15.1)$$

Проиллюстрируем соотношение (15.1) на примере двух крайних вариантов. Если общая СКО в точности равна остаточной, то случайный фактор x не оказывает влияния на результат, вся дисперсия y обусловлена воздействием прочих факторов. При этом прямая для линейной регрессии параллельна оси OX.

Когда общая СКО равна факторной, то никакие другие (прочие) факторы не влияют на результат. Величина y связана с x с вероятностью 100%, и остаточная СКО равна нулю.

Однако на практике в правой части (15.1) присутствуют оба слагаемых. Пригодность линии регрессии для прогноза зависит от того, какая часть общей вариации y приходится на объясненную вариацию. Если объясненная СКО будет больше остаточной СКО, то уравнение регрессии статистически значимо и фактор x оказывает существенное воздействие на результат y. Это равносильно тому, что коэффициент корреляции будет приближаться к единице.

Для использования критерия Фишера для оценки значимости регрессии необходимо ввести понятие число степей свободы для различных величин.

***Определение* 15.1**

Число степеней свободы (df-degrees of freedom) - это минимально необходимое число значений зависимой переменной, которых достаточно для получения искомой характеристики выборки и которые могут свободно изменяться с учетом того, что для этой выборки известны все другие величины, используемые для расчета искомой характеристики. Или другими словами-это ***число независимо варьируемых значений признака****.*

Уравнение для определения F-статистики в случае многомерной регрессии имеет вид:

$$F_{\text{ф}} = \left(\frac{\sum\limits_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2}{m} \right) \left\{ 1 / \frac{\left(\sum\limits_{i=1}^{n} \left(y_i - \hat{y}\right)^2 \right)}{(n - m - 1)} \right\} , \tag{15.2}$$

где n- объем выборки; m- число независимых переменных в факторной части СКО.

Поясним на примере линейной регрессии. Факторную СКО в этом случае можно выразить так:

$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2 = \sum_{i=1}^{n} \left([A + Bx_i] - [A + B\bar{x}] \right)^2 =$$
$$= \sum_{i=1}^{n} \left(Bx_i - B\bar{x}\right)^2 = B^2 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 \ . \tag{15.3}$$

СКО (15.3) зависит только от одного параметра B. Следовательно, факторная СКО в случае линейной регрессии имеет одну степень свободы, т. е. $m = 1$.

Можно и рассчитать число m и другим способом. Для того вспомним выражение (13.7) для регрессионной прямой

$$\hat{y} = A + Bx = \overline{y} + B\left(x - \overline{x}\right) \ . \tag{15.4}$$

Как видно из (15.4) прямая регрессии определяется только одним параметром.

Разделив каждую СКО на свое число степеней свободы, получим среднее квадратичные отклонения (или дисперсии на одну степень свободы):

$$D_{\text{общ.}} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \ ,$$

$$D_{\text{факт.}} = \frac{1}{m}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 \ ,$$

$$D_{\text{ост.}} = \frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y})^2 \ . \qquad (15.5)$$

С учетом (15.5) F-критерий (15.2) можно записать в виде:

$$F_{\text{ф}} = \frac{D_{\text{факт.}}}{D_{\text{ост.}}} \ . \qquad (15.6)$$

В теории вероятности доказано, что для выборки из генеральной совокупности у которой отсутствует связь между зависимой y и независимой x (или $x_1, x_2, \ldots, x_k$) переменной имеет распределение Фишера:

$$p_{F_R}(x, f_1, f_2) = \frac{f_1^{f_1/2} f_2^{f_2/2} x^{f_1/2-1} (f_1 x + f_2)^{-\frac{1}{2}(f_1+f_2)}}{B\left(\frac{f_1}{2}, \frac{f_2}{2}\right)} \ , \quad x \geq 0 \ ,$$

$$F_R(x, f_1, f_2) = I_{\frac{f_1 x}{f_1 x + f_2}}\left(\frac{f_1}{2}, \frac{f_2}{2}\right) \ . \qquad (15.7)$$

где $f_{1,2}$ - параметры распределения; функция $I_n\,(a,b)$ - регуляризованная неполная бэта функция.

Благодаря этому для осуществления статистической проверки значимости уравнения регрессии формулируется нулевая гипотеза об отсутствии связи между переменными (все коэффициенты при переменных равны нулю) и выбирается уровень значимости α.

Уровень значимости α - это вероятность совершить ошибку первого рода, т. е. отвергнуть в результате проверки верную нулевую гипотезу. В рассматриваемом случае совершить ошибку первого рода означает признать по выборке наличие связи между переменными в генеральной совокупности, **когда на самом деле ее там нет**.

Вычисленное значение $F_{\text{ф}}$ признается достоверным (отличным от единицы), если оно больше табличного квантиля распределения Фишера F_α

$$\int_{F_\alpha)}^{\infty} p_{F_R}(x, m, n-m-1)\mathrm{d}x = \alpha \ , \tag{15.8}$$

т.е.

$$F_{\text{ф}} \geq F_\alpha \ . \tag{15.9}$$

В этом случае нулевая гипотеза об отсутствии связи между переменными отклоняется и делается вывод о существенности превышения $D_{\text{факт.}}$ над $D_{\text{ост.}}$. Или другими словами делается вывод о существенной статистической зависимости между зависимой y и независимыми величинами $x = x_1, \ldots, x_k$.

Если $F_{\text{ф}} \leq F_\alpha$, то вероятность нулевой гипотезы выше заданного уровня α (например, 0,05 ÷ 0,10), и эта гипотеза не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи между y и x. Уравнение регрессии считается статистически незначимым, гипотеза H_0 не отклоняется.

С помощью компьютерных вычислений, практически всегда можно найти α путем решения уравнения:

$$\int_{F_{\text{ф}}}^{\infty} p_{F_R}(x, m, n-m-1)\mathrm{d}x = 1 - F_R(F_{\text{ф}}, m, n-m-1) = \alpha \ . \tag{15.10}$$

В если α относительно велико ($\alpha > 0{,}1$), то о наличии корреляции говорит сложно (или с большими оговорками).

Как и в случае парной регрессии, статистическая значимость коэффициентов множественной линейной регрессии с m объясняющими переменными проверяется на основе t-критерия Стьюдента. Величина стандартной ошибки совместно с t- распределением Стьюдента при $n-m$ степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительных интервалов. Для этого для каждого параметра регрессии a_i, полученного в результате вычислений рассчитывается величина

$$t_{\text{exp}} = \frac{a_i}{\sigma_{a_i}} \ , \tag{15.11}$$

т.е. величина параметров регрессии сравнивается с его стандартной ошибкой.

Значение (15.11) сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы. По сути проверяется нулевая

гипотеза в виде H_0: $a_i = 0$. Если $t_{\text{exp}} > t_{n-m,\alpha/2}$, то гипотеза H_0: $a_i = 0$ должна быть отклонена, а статистическая связь y с x считается установленной. В случае $t_{\text{exp}} > t_{n-m,\alpha/2}$ нулевая гипотеза не может быть отклонена, и влияние x на y признается несущественным.

15.2 Проверка общего качества уравнения регрессии

Оценить общее качество уравнения регрессии означает, установить соответствует ли математическая модель, выражающая зависимость между переменными экспериментальным данным и достаточно ли включенных в модель переменных, объясняющих поведение предсказанной величины (y). Оценить общие качества модели = оценить надежность модели = оценить достоверность уравнения регрессии.

15.3 Коэффициент детерминации

Для оценки качества различных моделей используют коэффициент детерминации.

***Определение* 15.2**

Коэффициент детерминации (R^2 — R-квадрат) — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными.

Более точно — это единица минус доля необъяснённой дисперсии (дисперсии случайной ошибки модели, или условной по факторам дисперсии зависимой переменной) в дисперсии зависимой переменной. Его рассматривают как универсальную меру зависимости одной случайной величины от множества других. В частном случае линейной зависимости R^2 является квадратом так называемого множественного коэффициента корреляции между зависимой переменной и объясняющими переменными.
В частности, для модели парной линейной регрессии коэффициент детерминации равен квадрату обычного коэффициента корреляции между y и x.

Истинный коэффициент детерминации модели зависимости случайной величины y от факторов x определяется следующим образом:

$$R^2 = 1 - \frac{D\{y|x\}}{D\{y\}} = 1 - \frac{\sigma^2}{\sigma_y^2}, \tag{15.12}$$

где $D\{y|x\} = \sigma^2$ — условная (по факторам x) дисперсия зависимой переменной (дисперсия случайной ошибки модели).

В данном определении используются истинные параметры, характеризующие распределение случайных величин. Если использовать выборочную оценку значений соответствующих дисперсий, то получим формулу для выборочного коэффициента детерминации (который обычно и подразумевается под коэффициентом детерминации):

Коэффициент детерминации

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{SS_{res}/n}{SS_{tot}/n} = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{15.13}$$

где

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{15.14}$$

сумма квадратов остатков регрессии, y_i, $\hat{y}_i$ — фактические ("экспериментальные") и расчётные значения объясняемой переменной, а величина

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n\hat{\sigma}_y^2 \tag{15.15}$$

так называемая **общая сумма квадратов**, а

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ .$$

В случае линейной регрессии с константой $SS_{tot} = SS_{reg} + SS_{res}$, где

$$SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{15.16}$$

объяснённая сумма квадратов, поэтому получаем более простое определение в этом случае — коэффициент детерминации — это доля объяснённой суммы квадратов в общей:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \ . \tag{15.17}$$

Необходимо подчеркнуть, что формула (15.17) справедлива только для модели с константой, в общем случае необходимо использовать предыдущую формулу.

Интерпретация

1. Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50% (в этом случае коэффициент множественной корреляции превышает по модулю 70%). Модели с коэффициентом детерминации выше 80% можно признать достаточно хорошими (коэффициент корреляции превышает 90%). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

Интерпретация

2. При отсутствии статистической связи между объясняемой переменной и факторами, статистика nR^2 для линейной регрессии имеет

асимптотическое распределение $\chi^2(k-1)$, где $k-1$ — количество факторов модели. В случае линейной регрессии с нормально распределёнными случайными ошибками статистика $F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$ имеет точное (для выборок любого объёма) распределение Фишера $F(k-1, n-k)$ (так называемый F-тест). Информация о распределении этих величин позволяет проверить статистическую значимость регрессионной модели исходя из значения коэффициента детерминации. Фактически в этих тестах проверяется гипотеза о равенстве истинного коэффициента детерминации нулю.

3. В общем случае коэффициент детерминации может быть и отрицательным, это говорит о крайней неадекватности модели: простое среднее приближает лучше.

15.4 Недостаток R^2 и альтернативные показатели

Основная проблема применения (выборочного) R^2 заключается в том, что его значение увеличивается (не уменьшается) от добавления в модель новых переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют! Поэтому сравнение моделей с разным количеством факторов с помощью коэффициента детерминации, вообще говоря, некорректно. Для этих целей можно использовать альтернативные показатели.

Скорректированный (adjusted) R^2

Для того, чтобы была возможность сравнивать модели с разным числом факторов так, чтобы число факторов не влияло на статистику R^2 обычно используется скорректированный коэффициент детерминации, в котором используются несмещённые оценки дисперсий:

$$R^2_{adj} = 1 - \frac{s^2}{s^2_y} = 1 - \frac{SS_{res}/(n-k)}{SS_{tot}/(n-1)} = 1 - (1-R^2)\frac{(n-1)}{(n-k)} \leqslant R^2 \quad (15.18)$$

который даёт “штраф” за дополнительно включённые факторы, где n — количество наблюдений, а k — количество параметров.

Данный показатель всегда меньше единицы, но теоретически может быть и меньше нуля (только при очень маленьком значении обычного коэффициента детерминации и большом количестве факторов). Поэтому теряется интерпретация показателя как “доли”. Тем не менее, применение показателя в сравнении вполне обоснованно.

Для моделей с одинаковой зависимой переменной и одинаковым объёмом выборки сравнение моделей с помощью скорректированного коэффициента детерминации эквивалентно их сравнению с помощью остаточной дисперсии $s^2 = SS_{res}/(n-k)$ или стандартной ошибки модели s. Разница только в том, что последние критерии чем меньше, тем лучше.

Для оценки качества различных моделей при описании данных также используют информационные критерии.

***Определение* 15.3**

Информационный критерий — применяемая в статистике мера относительного качества статистических моделей, учитывающая степень “подгонки” модели под данные с корректировкой (“штрафом”) на используемое количество оцениваемых параметров. То есть критерии основаны на неком компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс.

Информационный характер критериев связан с концепцией информационной энтропии и расстоянием Кульбака-Лейблера, на основе которой был разработан исторически первый критерий — критерий Акаике (AIC), предложенный в 1974 году Хиротсугу Акаике (H. Akaike, “A new look at the statistical model identification” IEEE Transactions on Automatic Control, vol. 19, № . 6, pp. 716-723, 1974.)

Информационные критерии используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Они не позволяют тестировать модели в смысле проверки статистических гипотез. Обычно чем меньше значения критериев, тем выше относительное качество модели.

Информационный критерий Акаике (AIC)

Предложен Хиротугу Акаике в 1971 году, описан и исследован им же в 1973, 1974, 1983 годах. Первоначально аббревиатура AIC, предложенная автором, расшифровывалась как an information criterion ("некий информационный критерий"), однако последующие авторы называли его Akaike information criterion. Исходная расчетная формула критерия имеет вид:

$$\mathrm{AIC} = 2k - 2L \tag{15.19}$$

где L - значение логарифмической функции правдоподобия построенной модели, k-количество использованных (оцененных) параметров.

Многие современные авторы, а также во многих программных продуктах применяется несколько иная формула, предполагающая деление на объем выборки n, по которой строилась модель:

$$\mathrm{AIC} = \frac{2k - 2L}{n}\,. \tag{15.20}$$

Данный подход позволяет сравнивать модели, оцененные по выборках разного объема.

Чем меньше значение критерия, тем лучше модель.

Многие другие критерии являются модификациями AIC.

Байесовский информационный критерий (BIC) или критерий Шварца (SC)

Байесовский информационный критерий (Bayesian information criterion — BIC) предложен Шварцем в 1978 году, поэтому часто он называется также критерием Шварца (Schwarz criterion — SC). Он разработан исходя из байесовского подхода и является наиболее часто используемой модификацией AIC:

$$\mathrm{BIC} = \mathrm{SC} = k \ln n - 2L\,. \tag{15.21}$$

Как видно из формулы, данный критерий налагает больший штраф на увеличение количества параметров по сравнению с `AIC`, так как $\ln n$ больше 2 уже при количестве 8 наблюдений.

Прочие информационные критерии

Состоятельный критерий Акаике (`Consistent AIC` — `CAIC`) предложенный в 1987 году Бозdoганом:

$$\mathrm{CAIC} = (1 + \ln n)k - 2L \; . \tag{15.22}$$

Данный критерий асимптотически эквивалентен `BIC`. Тот же автор в 1994 году предложил модификации, увеличивающие коэффициент при количестве параметров (вместо 2-3 или 4 для AIC_3 и AIC_4).

Скорректированный критерий Акаике (`Corrected AIC` – AIC_c), который рекомендуется применять на малых выборках (предложен в 1978 году Sugiura):

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1} \; . \tag{15.23}$$

Данный критерий, наряду с `AIC` и `BIC` выдается в результатах оценки моделей в `Wolfram Mathematica` при использовании оператора `NonlinearModelFit`.

Критерий Ханнана-Куинна (Hannan-Quinn, HQ) предложен авторами в 1979 году

$$\mathrm{HQ} = 2k \ln \ln n/n - 2L/n \; . \tag{15.24}$$

Имеются также модификации `AIC`, использующие более сложные штрафные функции, зависящие от различных характеристик.

Замечание

Высокие значения коэффициента детерминации, вообще говоря, не свидетельствуют о наличии причинно-следственной зависимости между переменными (также как и в случае обычного коэффициента корреляции). Например, если объясняемая переменная и факторы, на самом деле не связанные с объясняемой переменой, имеют возрастающую динамику, то коэффициент детерминации будет достаточно высок. **Поэтому логическая и смысловая адекватность модели имеют первостепенную важность**. Кроме того, необходимо использовать критерии для всестороннего анализа качества модели.

16 Рекомендуемая литература

Рекомендуемая литература

1. Лавренчик, В.Н. Постановка физического эксперимента и статистическая обработка его результатов/ В.Н. Лавренчик.. – М.: Энергоатомиздат, 1986. - 272 с.

2. Бендат Дж., Пирсол А. Прикладной анализ случайных данных/ Дж.Бендат, А.Пирсол. – М.: Мир,1989. -504 с.

3. Новицкий, П.В. Оценка погрешностей результатов измерений/ П.В.Новицкий, И.А..Зограф - Л.: Энергоатомиздат, 1991. - 304 с

4. Тейлор Дж. Введение в теорию ошибок/ Дж.Тейлор. - М.: Мир, 1985. - 45 с.

5. Гмурман, В. Е. Теория вероятностей и математическая статистика: Учеб. пособие для вузов/В. Е. Гмурман. - 9-е изд., стер. - М.: Высш. шк., 2003. - 479 с.

Дополнительная

Рекомендуемая литература

6. Дьяконов, В.П. Mathematica 4: учебный курс/ В.П. Дьяконов. - СПб.: Питер, 2001 . - 654 с.

7. Воробьёв Е.М. Введение в систему Mathematica./ Е.М.Воробьёв. - М.: Финансы и статистика, 1998. - 345 с.

8. Львовский, Б.Н. Статистические методы построения эмпирических формул: Учеб. пособие для втузов/ Б.Н.Львовский. – М.: Высш. шк., 1988 – 239 с.

9. Кобзарь А. И. Прикладная математическая статистика. Для инженеров и научных работников/ А. И. Кобзарь. – М.: Физматлит, 2006. – 816 с.

10. Боровков Л. Л. Математическая статистика/ Л. Л.Боровков-Учебник.- М.: Наука. Главная редакция физико-математической литературы, 1984. – 472 с.

11. Кассандрова, О. Н. Обработка результатов наблюдений/ О. Н. Кассандрова, В. В. Лебедев - М.:Наука, Главная редакция физ.-мат. литературы, 1970 г. – 104 с.
12. Ивченко, Г. И. Математическая статистика/ Г.И.Ивченко, Ю. И. Медведев Учеб. пособие для втузов. - М.: Высш. шк., 1984. – 248 с.