

1. Основные понятия математической статистики

Математическая (или теоретическая) статистика опирается на методы и понятия теории вероятностей, но решает в каком-то смысле обратные задачи.

В теории вероятностей рассматриваются случайные величины с **заданным** распределением или случайные эксперименты, свойства которых **целиком известны**. Предмет теории вероятностей — свойства и взаимосвязи этих величин (распределений).

Но часто эксперимент представляет собой черный ящик, выдающий лишь некие результаты, по которым требуется сделать вывод о свойствах самого эксперимента. Наблюдатель имеет набор числовых (или их можно сделать числовыми) результатов, полученных повторением одного и того же случайного эксперимента в одинаковых условиях.

При этом возникают, например, следующие вопросы: Если мы наблюдаем одну случайную величину — как по набору ее значений в нескольких опытах сделать как можно более точный вывод о ее распределении?

Если мы наблюдаем одновременно проявление двух (или более) признаков, т.е. имеем набор значений нескольких случайных величин — что можно сказать об их зависимости? Есть она или нет? А если есть, то какова эта зависимость?

Часто бывает возможно высказать некие предположения о распределении, спрятанном в «черном ящике», или о его свойствах. В этом случае по опытным данным требуется подтвердить или опровергнуть эти предположения («гипотезы»). При этом надо помнить, что ответ «да» или «нет» может быть дан лишь с определенной степенью достоверности, и чем дольше мы можем продолжать эксперимент, тем точнее могут быть выводы. Наиболее благоприятной для исследования оказывается ситуация, когда можно уверенно утверждать о некоторых свойствах наблюдаемого эксперимента — например, о наличии функциональной зависимости между наблюдаемыми величинами, о нормальности распределения, о его симметричности, о наличии у распределения плотности или о его дискретном характере, и т.д.

Итак, о (математической) статистике имеет смысл вспоминать, если

- имеется случайный эксперимент, свойства которого частично или полностью неизвестны,

- мы умеем воспроизводить этот эксперимент в одних и тех же условиях некоторое (а лучше — какое угодно) число раз.

1.1. Основные понятия выборочного метода

Пусть $\xi : \Omega \rightarrow \mathbf{R}$ — случайная величина, наблюдаемая в случайном эксперименте. Предполагается, что вероятностное пространство задано (и не будет нас интересовать).

Будем считать, что проведя n раз этот эксперимент в одинаковых условиях, мы получили числа x_1, x_2, \dots, x_n — значения этой случайной величины в первом, втором, и т.д. экспериментах. Случайная величина ξ имеет некоторое распределение \mathcal{F} , которое нам **частично или полностью неизвестно**.

Рассмотрим подробнее набор $\mathbf{X} = (X_1, \dots, X_n)$, называемый **выборкой**.

В серии уже произведенных экспериментов выборка — это набор чисел. Но если эту серию экспериментов повторить еще раз, то вместо этого набора мы получим **новый** набор чисел. Вместо числа x_1 появится другое число — одно из значений случайной величины ξ . То есть x_1 (и x_2 , и x_3 , и т.д.) — **переменная величина**, которая может принимать те же значения, что и случайная величина ξ , и так же часто (с теми же вероятностями). Поэтому до опыта x_1 — случайная величина, одинаково распределенная с ξ , а после опыта — число, которое мы наблюдаем в данном первом эксперименте, т.е. одно из возможных значений **случайной величины** x_1 .

Выборка $\mathbf{X} = (X_1, \dots, X_n)$ объема n — это набор из n **независимых и одинаково распределенных** случайных величин («копий ξ »), имеющих, как и ξ , распределение \mathcal{F} .

Что значит «по выборке сделать вывод о распределении»? Распределение характеризуется функцией распределения, плотностью или таблицей, набором числовых характеристик — $E\xi$, $D\xi$, $E\xi^k$ и т.д. По выборке нужно уметь строить приближения для всех этих характеристик.

1.2. Эмпирическая функция распределения, гистограмма

Поскольку неизвестное распределение \mathcal{F} можно описать, например, его функцией распределения $F(u) = P(X_1 < u)$, построим по выборке «оценку» для этой функции.

Определение 1.

Эмпирической функцией распределения, построенной по выборке $X = (X_1, \dots, X_n)$ объема n , называется случайная функция $F_n^* : \mathbb{R} \times \Omega \rightarrow [0, 1]$, при каждом $y \in \mathbb{R}$ равная

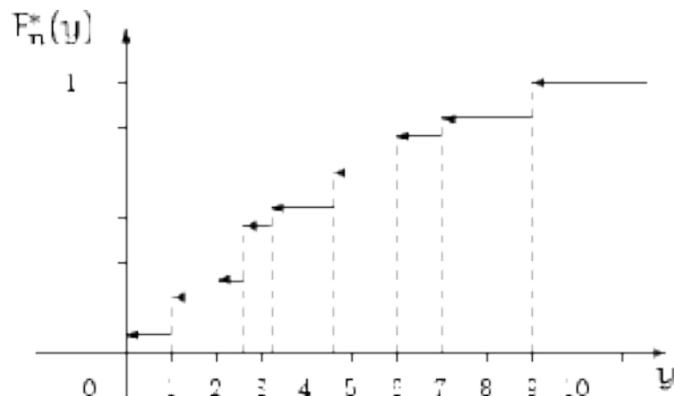
$$F_n^*(y) = \frac{\text{количество } X_i \in (-\infty, y]}{n} = \frac{1}{n} \sum_{i=1}^n I(X_i < y).$$

Пример 1.

Выборка: $X = (0; 2; 1; 2,6; 3,1; 4,6; 1; 4,6; 6; 2,6; 6; 7; 9; 9; 2,6)$.

Вариационный ряд: $(0; 1; 1; 2; 2,6; 2,6; 2,6; 3,1; 4,6; 4,6; 6; 6; 7; 9; 9)$.

Рис. 1. Пример 1



Эмпирическая функция распределения имеет скачки в точках выборки, величина скачка в точке x_i равна m_i/n , где m_i — количество элементов выборки, совпадающих с x_i .

Другой характеристикой распределения является таблица (для дискретных распределений) или плотность (для абсолютно непрерывных). Эмпирическим, или выборочным аналогом таблицы или плотности является так называемая **гистограмма**.

Гистограмма строится по **группированным** данным. Предполагаемую область значений случайной величины ξ (или область выборочных данных) делят **независимо от выборки** на некоторое количество интервалов (не обязательно одинаковых). Пусть A_1, \dots, A_k — интервалы на прямой, называемые интервалами группировки. Обозначим для $j = 1, \dots, k$ через n_j число элементов выборки, попавших в интервал A_j :

$$v_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^n I(X_i \in A_j), \quad \text{здесь} \quad \sum_{j=1}^k v_j = n. \quad (1)$$

На каждом из интервалов A_j строят прямоугольник, площадь которого пропорциональна v_j . Общая площадь всех прямоугольников должна равняться единице. Пусть l_j — длина интервала A_j . Высота f_j прямоугольника над A_j равна

$$f_j = \frac{v_j}{n l_j}.$$

Полученная фигура называется гистограммой.

Пример 2.

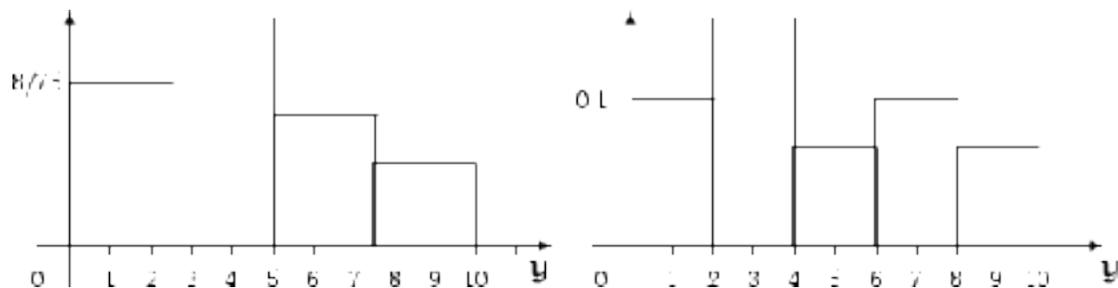
Имеется вариационный ряд (см. пример 1):

$$(0; 1; 1; 2; 2,6; 2,6; 2,6; 3,1; 4,6; 4,6; 6; 6; 7; 9; 9).$$

Разобьем отрезок $[0, 10]$ на 4 равных отрезка. В отрезок $A_1 = [0; 2,5]$ попали 4 элемента выборки, в $A_2 = [2,5; 5]$ — 6, в $A_3 = [5; 7,5]$ — 3, и в отрезок $A_4 = [7,5; 10]$ попали 2 элемента выборки. Строим гистограмму (рис. 2). На рис. 3 — тоже гистограмма для той же выборки, но при разбиении области на 5 равных отрезков.

Рис. 2. Пример 2

Рис. 3. Пример 2



1.3. Выборочные моменты

Знание моментов распределения также многое может сказать о его виде и свойствах. Введем выборочные аналоги неизвестных истинных моментов распределения.

Пусть $E\xi = EX_1 = a$, $D\xi = DX_1 = \sigma^2$, $E\xi^k = EX_1^k = m_k$ — теоретические среднее, дисперсия, k -й момент. Мы уже знакомы с соответствующими характеристиками выборочного распределения $E\xi^* = \bar{X}$, $D\xi^* = S^2$, $E(\xi^*)^k = \bar{X}^k$.

Теоретические характеристики	Эмпирические характеристики
$E\xi = EX_1 = a$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ — выборочное среднее
$D\xi = DX_1 = \sigma^2$	$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ — выборочная дисперсия или $S_0^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ — несмещенная выборочная дисперсия
$E\xi^k = EX_1^k = m_k$	$\bar{X}^k = \frac{1}{n} \sum_{i=1}^n X_i^k$ — выборочный k -й момент

Список числовых характеристик и их оценок можно продолжать, рассмотрев, например, центральные, абсолютные и т.п. моменты. В общем случае

момент $E g(\xi)$ будем оценивать величиной $\overline{g(X)} = \frac{1}{n} \sum_{i=1}^n g(X_i)$.

1.4. Свойства эмпирической функции распределения

Теорема 1.

Пусть $X = (X_1, \dots, X_n)$ — выборка объема n из неизвестного распределения \mathcal{F} с функцией распределения F . Пусть F_n^* — эмпирическая функция распределения, построенная по этой выборке. Тогда для любого $y \in \mathbb{R}$

$$F_n^*(y) \xrightarrow{P} F(y) \quad \text{при} \quad n \rightarrow \infty.$$

Замечание 1.

$F_n^*(y)$ — случайная величина, так как она является функцией от случайных величин X_1, \dots, X_n . То же самое можно сказать про гистограмму и выборочные моменты.

2.1. Параметрические семейства распределений

Предположим, что имеется выборка объема n , элементы которой X_1, \dots, X_n независимы, одинаково распределены и имеют распределение \mathcal{F}_θ , известным образом зависящее от неизвестного параметра θ .

Здесь \mathcal{F}_θ — некий класс распределений, целиком определяющихся значением скалярного или векторного параметра θ . Параметр θ принимает значения из некоторого множества Θ .

Например, для всех $i = 1, \dots, n$

- X_i имеют распределение Пуассона Π_λ , где $\lambda > 0$ — неизвестный параметр; здесь $\mathcal{F}_\theta = \Pi_\lambda$, $\theta = \lambda$, $\Theta = (0, \infty)$;

- X_i имеют распределение Бернулли B_p , где $p \in (0, 1)$ — неизвестный параметр; здесь $\mathcal{F}_\theta = B_p$, $\theta = p$, $\Theta = (0, 1)$;
- X_i имеют равномерное распределение $U_{a,b}$, где $a < b$ — неизвестные параметры; здесь $\mathcal{F}_\theta = U_{a,b}$, $\theta = (a, b)$, $\Theta = \{(a, b) : a < b\}$;
- X_i имеют равномерное распределение $U_{0,\theta}$, где $\theta > 0$ — неизвестный параметр; здесь $\mathcal{F}_\theta = U_{0,\theta}$, $\Theta = (0, \infty)$;
- X_i имеют нормальное распределение N_{μ,σ^2} , где $\mu \in \mathbb{R}$, $\sigma > 0$ — неизвестные параметры; здесь $\mathcal{F}_\theta = N_{\mu,\sigma^2}$, $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times (0, \infty)$;
- X_i имеют нормальное распределение $N_{\mu,\sigma}$, где $\mu \in \mathbb{R}$ — неизвестный параметр; здесь $\mathcal{F}_\theta = N_{\mu,\sigma}$, $\theta = \mu$, $\Theta = \mathbb{R}$.

Такая постановка имеет смысл, поскольку редко о проводимом эксперименте совсем ничего нельзя сказать. Обычно тип распределения ясен заранее, и требуется лишь указать значения параметров этого распределения.

Так, в широких предположениях рост юношей одного возраста имеет нормальное распределение (с неизвестными средним и дисперсией), а число покупателей в магазине в течение часа (не часа пик) — распределение Пуассона, и опять-таки с неизвестной «интенсивностью» λ .

2.2. Точечные оценки. Несмещенность, состоятельность оценок

Итак, пусть X_1, \dots, X_n — выборка объема n из параметрического семейства распределений \mathcal{F}_θ , $\theta \in \Theta$.

Заметим, что все характеристики случайных величин X_1, \dots, X_n зависят от параметра θ . Так, например, для X_i с распределением Пуассона Π_λ

$$E X_1 = \lambda, \quad P(X_1 = 2) = \frac{\lambda^2}{2} e^{-\lambda}, \quad D X_1 = \lambda \quad \text{и т. д.}$$

Чтобы отразить эту зависимость, будем писать $E_{\theta} X_1$ вместо $E X_1$ и т.д. Так, $D_{\theta_1} X_1$ означает дисперсию, вычисленную в предположении $\theta = \theta_1$.

Во многих случаях эта условность необходима. Предположим, что X_i имеют распределение Пуассона Π_{λ} . В предположении, что $\lambda = 1$, имеем $E X_1 = 1$, тогда как при $\lambda = 7$ имеем $E X_1 = 7$. Таким образом, запись $E X_1$, без указания на распределение X_1 , оказывается просто бессмысленной.

Определение 2.

Статистикой называется произвольная функция $\theta^* = \theta^*(X_1, \dots, X_n)$ от элементов выборки.

Замечание 2.

Статистика есть функция от эмпирических данных, но никак не от параметра θ . Статистика, как правило, предназначена именно для оценивания неизвестного параметра θ (поэтому ее иначе называют «оценкой»), и уже поэтому от него зависеть не может.

Конечно, статистика есть не «любая», а «измеримая» функция от выборки (борелевская, для которой прообраз любого борелевского множества из \mathbb{R} есть снова борелевское множество в \mathbb{R}^n), но мы никогда не встретимся с иными функциями, и более на это обращать внимание не будем.

Определение 3.

Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется несмещенной оценкой параметра θ , если для любого $\theta \in \Theta$ выполнено равенство

$$E_{\theta} \theta^* = \theta.$$

Определение 4.

Статистика $\theta^* = \theta^*(X_1, \dots, X_n)$ называется состоятельной оценкой параметра θ , если для любого $\theta \in \Theta$ имеет место сходимость

$$\theta^* \xrightarrow{P} \theta \quad \text{при } n \rightarrow \infty.$$

Несмещенность — свойство оценок при фиксированном n . Означает это свойство отсутствие ошибки «в среднем», т.е. при систематическом использовании данной оценки.

Свойство состоятельности означает, что последовательность оценок приближается к неизвестному параметру при увеличении количества данных. Понятно, что при отсутствии этого свойства оценка совершенно «несостоятельна» как оценка.

2.3. Методы нахождения оценок: метод моментов

Метод моментов заключается в следующем: любой момент случайной величины X_1 (например, k -й) зависит, часто функционально, от параметра θ . Но тогда и параметр θ может оказаться функцией от теоретического k -го момента. Подставив в эту функцию вместо неизвестного теоретического k -го момента его выборочный аналог, получим вместо параметра θ оценку θ^* . Пусть X_1, \dots, X_n — выборка объема n из параметрического семейства распределений \mathcal{F}_θ , где $\theta \in \Theta$. Выберем некоторую функцию $g(y)$ так, чтобы существовал момент

$$E_{\theta} g(X_1) = h(\theta), \quad (1)$$

и функция h была обратима в области Θ . Тогда в качестве оценки θ^* для θ возьмем решение уравнения

$$\overline{g(X)} = h(\theta^*).$$

Или (что то же самое), сначала решаем уравнение (1) относительно θ , а затем вместо истинного момента берем выборочный:

$$\theta = h^{-1}(E_{\theta} g(X_1)), \quad \theta^* = h^{-1}(\overline{g(X)}) = h^{-1}\left(\frac{1}{n} \sum_{i=1}^n g(X_i)\right).$$

Чаще всего в качестве функции $g(y)$ берут $g(y) = y^k$. В этом случае

$$E_{\theta} X_1^k = h(\theta),$$

и, если функция h обратима в области Θ , то

$$\theta = h^{-1}\left(\mathbb{E}_\theta X_1^k\right), \quad \theta^* = h^{-1}\left(\overline{X^k}\right) = h^{-1}\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right).$$

Можно сказать, что мы берем в качестве оценки такое (случайное) значение параметра θ , при котором истинный момент совпадает с выборочным.

2.4. Состоятельность оценок метода моментов

Теорема 2.

Пусть $\theta^* = h^{-1}\left(\overline{g(X)}\right)$ — оценка параметра θ , полученная по методу моментов, причем функция h^{-1} непрерывна. Тогда θ^* состоятельна.

Замечание 3.

Может случиться так, что $\theta^* = h^{-1}\left(\overline{g(X)}\right) \notin \Theta$, тогда как $\theta \in \Theta$. В этом случае оценку корректируют. Например, в качестве ОММ берут ближайшую к $h^{-1}\left(\overline{g(X)}\right)$ точку из Θ или из замыкания Θ .

Пример 3.

Пусть X_1, \dots, X_n — выборка объема n из нормального распределения $N_{a,1}$ с неотрицательным средним $a \geq 0$. Ищем оценку для a по первому моменту:

$$\mathbb{E}_a X_1 = a, \quad \text{поэтому} \quad a^* = \bar{X}.$$

Однако по условию $a \geq 0$, тогда как \bar{X} может быть и отрицательно. Если $\bar{X} < 0$, то в качестве оценки для a более подойдет 0. Если же $\bar{X} > 0$, в качестве оценки нужно брать \bar{X} . Итого: $a^* = \max\{0, \bar{X}\}$ — «исправленная» оценка метода моментов.

2.5. Методы нахождения оценок: метод максимального правдоподобия

Метод максимального правдоподобия — еще один разумный способ построения оценки неизвестного параметра. Состоит он в том, что в качестве «наиболее правдоподобного» значения параметра берут значение θ , максимизирующее вероятность получить при n опытах данную выборку $\mathbf{X} = (X_1, \dots, X_n)$. Это значение параметра θ зависит от выборки и является искомой оценкой.

Решим сначала, что такое «вероятность получить данную выборку», т.е. что именно нужно максимизировать. Вспомним, что для абсолютно непрерывных распределений \mathcal{F}_θ их плотность $f_\theta(y)$ — «почти» (с точностью до dy) вероятность попадания в точку y . А для дискретных распределений \mathcal{F}_θ вероятность попасть в точку y равна $P_\theta(X_1 = y)$. И то, и другое мы будем называть плотностью распределения \mathcal{F}_θ . Итак,

Определение 5.

Функцию

$$f_\theta(y) = \begin{cases} \text{плотность } f_\theta(y), & \text{если распределение } \mathcal{F}_\theta \text{ абсолютно непрерывно,} \\ P_\theta(X_1 = y), & \text{если распределение } \mathcal{F}_\theta \text{ дискретно.} \end{cases}$$

мы будем называть **плотностью** распределения \mathcal{F}_θ .

Определение 6.

Функция (случайная величина при фиксированном θ)

$$f(\mathbf{X}, \theta) = f_\theta(X_1) \cdot f_\theta(X_2) \cdot \dots \cdot f_\theta(X_n) = \prod_{i=1}^n f_\theta(X_i)$$

называется **функцией правдоподобия**. Функция (тоже случайная)

$$L(\mathbf{X}, \theta) = \ln f(\mathbf{X}, \theta) = \sum_{i=1}^n \ln f_\theta(X_i)$$

называется **логарифмической функцией правдоподобия**.

В дискретном случае функция правдоподобия $f(x_1, \dots, x_n; \theta)$ есть вероятность выборке X_1, \dots, X_n в данной серии экспериментов равняться x_1, \dots, x_n . Эта вероятность меняется в зависимости от θ :

$$f(x, \theta) = \prod_{i=1}^n f_{\theta}(x_i) = P_{\theta}(X_1 = x_1) \cdot \dots \cdot P_{\theta}(X_n = x_n) = P_{\theta}(X_1 = x_1, \dots, X_n = x_n).$$

Определение 7.

Оценкой максимального правдоподобия $\hat{\theta}$ неизвестного параметра θ называют значение θ , при котором функция $f(x, \theta)$ достигает максимума (как функция от θ при фиксированных x_1, \dots, x_n):

$$\hat{\theta} = \arg \max_{\theta} f(x, \theta).$$

Пример 4.

Пусть X_1, \dots, X_n — выборка объема n из нормального распределения N_{μ, σ^2} , где $\mu \in \mathbb{R}$, $\sigma > 0$; и оба параметра μ , σ^2 неизвестны.

Выпишем плотность, функцию правдоподобия и логарифмическую функцию правдоподобия.
Плотность:

$$f_{(\mu, \sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right),$$

функция правдоподобия:

$$l(x, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}\right),$$

логарифмическая функция правдоподобия:

$$L(\mathbf{X}, \mu, \sigma^2) = \ln f(\mathbf{X}, \mu, \sigma^2) = -\ln(2\pi)^{n/2} - \frac{n}{2} \ln \sigma^2 - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}.$$

В точке экстремума (по (μ, σ^2)) гладкой функции L обращаются в нуль обе частные производные:

$$\frac{\partial}{\partial \mu} L(\mathbf{X}, \mu, \sigma^2) = \frac{2 \sum_{i=1}^n (X_i - \mu)}{2\sigma^2} = \frac{n\bar{X} - n\mu}{\sigma^2}; \quad \frac{\partial}{\partial \sigma^2} L(\mathbf{X}, \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^4}.$$

Оценка максимального правдоподобия $(\hat{\mu}, \hat{\sigma}^2)$ для (μ, σ^2) — решение системы уравнений

$$\frac{n\bar{X} - n\mu}{\sigma^2} = 0; \quad -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{2(\sigma^2)^2} = 0.$$

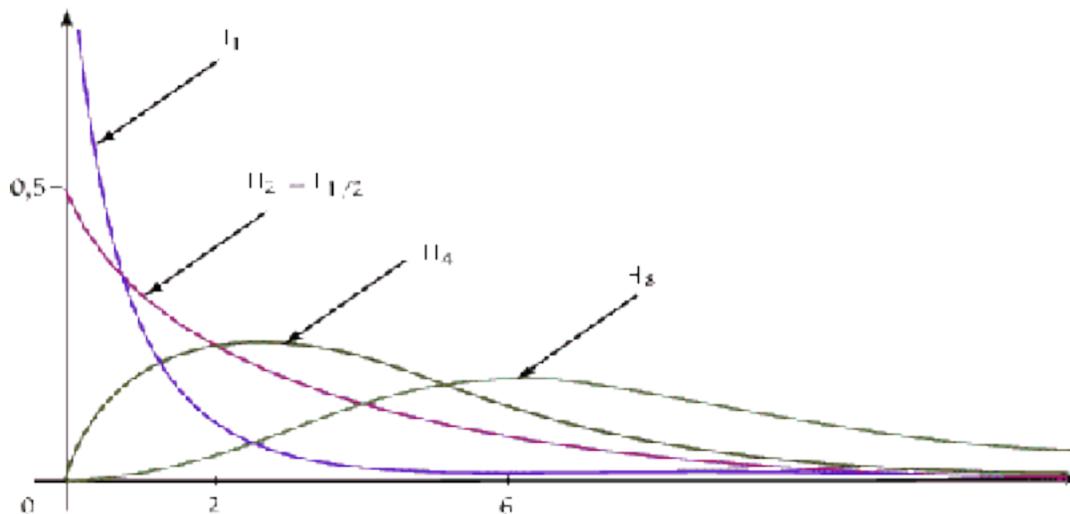
Решая, получим хорошо знакомые оценки:

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = S^2.$$

3.1. Распределение «хи-квадрат» и его свойства

Определение 8.

Распределение суммы k квадратов независимых стандартных нормальных случайных величин называют **распределением «хи-квадрат»** с k степенями свободы и обозначают χ_k^2 .



Вид плотности χ^2 -распределения в зависимости от числа степеней свободы

Мы часто будем обозначать через χ_k^2 случайную величину с распределением H_k .

Рассмотрим свойства χ^2 -распределения:

1. Устойчивость по суммированию.

Пусть случайная величина χ_k^2 имеет распределение H_k , случайная величина χ_m^2 имеет распределение H_m , причем эти случайные величины независимы. Тогда их сумма $\chi_k^2 + \chi_m^2$ имеет распределение H_{k+m} .

2. Моменты распределения χ^2 .

Если χ_k^2 имеет распределение H_k , то $E\chi_k^2 = k$ и $D\chi_k^2 = 2k$.

Следствие 1.

Если ξ_1, \dots, ξ_k независимы и имеют нормальное распределение N_{a, σ^2} , то

$$\chi_k^2 = \sum_{i=1}^k \left(\frac{\xi_i - a}{\sigma} \right)^2$$

имеет χ^2 -распределение H_k с k степенями свободы.

3.1. Распределение Стьюдента и его свойства

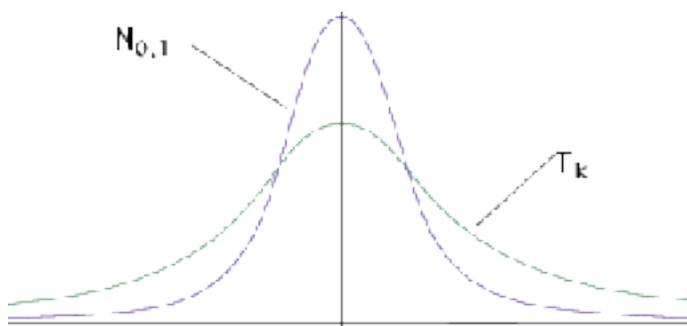
Определение 9.

Пусть $\xi_0, \xi_1, \dots, \xi_k$ независимы и имеют стандартное нормальное распределение.

Распределение случайной величины

$$t_k = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \dots + \xi_k^2)}} = \frac{\xi_0}{\sqrt{\frac{\chi_k^2}{k}}}$$

называют **распределением Стьюдента с k степенями свободы** и обозначают T_k .



Плотность распределения Стьюдента по сравнению с плотностью стандартного нормального распределения.

Плотность распределения Стьюдента с k степенями свободы равна

$$f_k(y) = \frac{\Gamma(k-1)/2}{\sqrt{\pi k} \Gamma(k/2)} \left(1 + \frac{y^2}{k}\right)^{-(k+1)/2}. \quad (2)$$

Свойства распределения Стьюдента:

1. Симметричность.

Если случайная величина t_k имеет распределение Стьюдента T_k с k степенями свободы, то и $-t_k$ имеет такое же распределение.

2. Асимптотическая нормальность.

Распределение Стьюдента T_k слабо сходится к стандартному нормальному распределению при $k \rightarrow \infty$.

У распределения Стьюдента существуют только моменты порядка $m < k$, и не существуют моменты порядка $m \geq k$. При этом все существующие моменты нечетного порядка равны нулю.

4. Проверка гипотез

Если возможно выдвинуть несколько взаимоисключающих «гипотез» о распределении элементов выборки, то возникает задача выбора одной из этих гипотез на основании выборочных данных. Как правило, по выборке конечного объема безошибочных выводов о распределении сделано быть не может, поэтому приходится считаться с возможностью выбрать неверную гипотезу.

Пусть дана выборка $X = (X_1, \dots, X_n)$ из распределения \mathcal{F} . Если не оговорено противное, считается, что все наблюдения имеют одно и то же распределение. В ряде случаев это предположение также нуждается в проверке (см., например, ниже: гипотеза об однородности или гипотеза о случайности) — в таких случаях одинаковая распределенность наблюдений не предполагается. То же касается и независимости наблюдений.

Определение 10.

Гипотезой (H) называется любое предположение о распределении наблюдений:

$$H = \{\mathcal{F} = \mathcal{F}_1\} \quad \text{или} \quad H = \{\mathcal{F} \in \{\mathcal{F}\}\}.$$

Гипотеза H называется простой, если она однозначно определяет распределение, т.е. $H = \{\mathcal{F} = \mathcal{F}_1\}$. Иначе H называется сложной гипотезой. Сложная гипотеза предполагает, что распределение \mathcal{F} — одно из некоторого множества распределений $\{\mathcal{F}\}$.

Если гипотез всего две, то одну из них принято называть основной, а другую — альтернативой или отклонением от основной гипотезы.

Определение 11.

Для заданного критерия $\delta: \mathbb{R}^n \rightarrow \{H_1, \dots, H_k\}$ будем говорить, что произошла **ошибка i -го рода**, если гипотеза H_i отвергнута критерием, в то время как она верна. Вероятностью ошибки i -го рода критерия δ называется

$$\alpha_i(\delta) = P_{H_i}(\delta(X) \neq H_i).$$

Замечание 4.

Говоря « H_i верна» и вычисляя $P_{H_i}(\cdot)$, мы имеем в виду, что распределение выборки именно такое, как предполагает гипотеза H_i , и вычисляем вероятность в соответствии с этим распределением. Если гипотеза H_i простая, т.е. указывает ровно на одно возможное распределение выборки, то $\alpha_i(\delta)$ — число. Если же H_i — сложная гипотеза, то $\alpha_i(\delta)$ будет зависеть от того, при каком именно из распределений \mathcal{F}_i , отвечающих H_i , вычисляется вероятность:

$$\alpha_i(\delta) = \alpha_i(\delta, \mathcal{F}_i] = P_{\mathcal{F}_i}(\delta(X) \neq H_i)$$

5. Критерии согласия

Критериями согласия называют критерии, предназначенные для проверки простой гипотезы $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при сложной альтернативе $H_2 = \{H_1 \text{ неверна}\}$. Мы рассмотрим более широкий класс основных гипотез, включающий и сложные гипотезы, а критериями согласия будем называть любые критерии, устроенные по одному и тому же принципу. А именно, пусть задана некоторая **функция отклонения** эмпирического распределения от теоретического, распределение которой существенно разнится в зависимости от того, верна или нет основная гипотеза. Критерии согласия принимают или отвергают основную гипотезу исходя из величины этой функции отклонения.

Итак, имеется выборка $\mathbf{X} = (X_1, \dots, X_n)$ из распределения \mathcal{F} . Мы сформулируем ряд понятий для случая простой основной гипотезы, а в дальнейшем будем их корректировать по мере изменения задачи. Проверяется простая основная гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ при сложной альтернативе $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

К1.

Пусть возможно задать функцию $\rho(\mathbf{X})$, обладающую свойствами:

а) если гипотеза H_1 верна, то $\rho(\mathbf{X}) \Rightarrow G$, где G — непрерывное распределение;

б) если гипотеза H_1 неверна, то $|\rho(\mathbf{X})| \xrightarrow{P} \infty$ при $n \rightarrow \infty$.

К2.

Пусть такая функция $\rho(\mathbf{X})$ задана. Для случайной величины Π из распределения G определим постоянную C из равенства $\varepsilon = P(|\Pi| \geq C)$.

Построим критерий:

$$\delta(\mathbf{X}) = \begin{cases} H_1, & \text{если } |\rho(\mathbf{X})| < C, \\ H_2, & \text{если } |\rho(\mathbf{X})| \geq C. \end{cases} \quad (3)$$

Мы построили критерий согласия. Он «работает» по принципу: если для данной выборки функция отклонения велика (по абсолютному значению), то это свидетельствует в пользу альтернативы, и наоборот. Убедимся в том, что этот критерий имеет (асимптотический) размер ε и является **состоятельным**.

Определение 12.

Говорят, что критерий δ для проверки простой гипотезы H_1 является критерием **асимптотического** размера ε , если его размер приближается к ε с ростом n :

$$\alpha_1(\delta) = P_{H_1}(\delta(\mathbf{X}) \neq H_1) \rightarrow \varepsilon \text{ при } n \rightarrow \infty.$$

Поскольку альтернатива H_2 всегда является сложной, то, как мы уже отмечали в замечании 4, вероятность ошибки второго рода любого критерия δ есть функция $\alpha_2(\delta, \mathcal{F}_2)$ от конкретного распределения \mathcal{F}_2 из списка возможных альтернатив $\{\mathcal{F}_2 : \mathcal{F}_2 \neq \mathcal{F}_1\}$. Или, при ином виде основной гипотезы, из числа распределений, отвечающих альтернативе H_2 .

Определение 13.

Критерий δ для проверки гипотезы H_1 против сложной альтернативы H_2 называется **состоятельным**, если для любого распределения \mathcal{F}_2 , отвечающего альтернативе H_2 , вероятность ошибки второго рода стремится к нулю с ростом объема выборки:

$$\alpha_2(\delta, \mathcal{F}_2) = P_{\mathcal{F}_2}(\delta(\mathbf{X}) = H_1) \rightarrow 0 \text{ при } n \rightarrow \infty.$$

Свойство 1.

Для критерия δ , заданного в (3), при $n \rightarrow \infty$:

1.

$$\alpha_1(\delta) = P_{H_1}(|\rho(X)| \geq C) \rightarrow P(|\eta| \geq C) = \epsilon;$$

2.

$$\alpha_2(\delta, \mathcal{F}_2) = P_{\mathcal{F}_2}(|\rho(X)| < C) \rightarrow 0 \text{ для любого распределения } \mathcal{F}_2, \text{ отвечающего } H_2.$$

Иначе говоря, построенный критерий имеет асимптотический размер ϵ и состоятелен.

5.1. Критерии согласия: критерий χ^2 Пирсона

Критерий χ^2 (К. Pearson, 1903) основывается на группированных данных. Область значений предполагаемого распределения \mathcal{F}_1 делят на некоторое число интервалов. После чего строят функцию отклонения ρ по разностям теоретических вероятностей попадания в интервалы группировки и эмпирических частот.

Имеется выборка $X = (X_1, \dots, X_n)$ из распределения \mathcal{F} . Проверяется простая гипотеза $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ против сложной альтернативы $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$.

Пусть A_1, \dots, A_k — интервалы группировки в области значений случайной величины с распределением \mathcal{F}_1 . Обозначим для $j = 1, \dots, k$ через v_j число элементов выборки, попавших в интервал A_j

$$v_j = \{\text{число } X_i \in A_j\} = \sum_{i=1}^n I(X_i \in A_j),$$

и через $p_j > 0$ — теоретическую вероятность $P_{H_1}(X_1 \in A_j)$ попадания в интервал A_j случайной величины с распределением \mathcal{F}_1 . С необходимостью, $p_1 + \dots + p_k = 1$. Как правило, длины интервалов выбирают так, чтобы $p_1 = \dots = p_k = 1/k$.

Пусть

$$\rho(X) = \sum_{j=1}^k \frac{(v_j - np_j)^2}{np_j} \quad (4)$$

Замечание 5.

Свойство K1(б) выполнено далеко не для всех альтернатив. Если распределение выборки $\mathcal{F}_2 \neq \mathcal{F}_1$ имеет такие же, как у \mathcal{F}_1 , вероятности P_i попадания в каждый из интервалов A_i , то по данной функции ρ эти распределения различить невозможно.

Поэтому на самом деле критерий, который мы построим по функции ρ из (4), решает совсем иную задачу. А именно, пусть задан набор вероятностей p_1, \dots, p_k такой, что $p_1 + \dots + p_k = 1$. Критерий χ^2 предназначен для проверки сложной гипотезы

$H'_1 = \{ \text{распределение } X_1 \text{ обладает свойством: } P(X_1 \in A_j) = p_j \text{ для всех } j = 1, \dots, k \}$

против сложной альтернативы $H'_2 = [H'_1 \text{ неверна}]$, т.е.

$H'_2 = \{ \text{хотя бы для одного из интервалов вероятность } P(X_1 \in A_j) \text{ отличается от } p_j \}$.

Покажем, что $\rho(X)$ удовлетворяет условию K1(a).

Теорема 3 (Пирсона).

Если верна гипотеза H'_1 , то при фиксированном k и при $n \rightarrow \infty$

$$\rho(X) = \sum_{j=1}^k \frac{(v_j - np_j)^2}{np_j} \Rightarrow H_{k-1},$$

где, напомним, H_{k-1} есть χ^2 -распределение с $k-1$ степенью свободы.

Замечание 6.

Сходимость по распределению $\rho(X) \Rightarrow H_{k-1}$ обеспечивается ЦПТ, поэтому разница допредельной и предельной вероятностей имеет тот же порядок, что и погрешность нормального приближения

$$|P(\rho(X) \geq C) - P(\chi_{k-1}^2 \geq C)| \leq \text{примерно!} \max \left\{ \frac{b}{\sqrt{\pi p_j(1-p_j)}} \right\}$$

(см. неравенство Берри — Эссеена для погрешности в ЦПТ), где b — некоторая постоянная. Маленькие значения πp_j в знаменателе приведут к тому, что распределение $\rho(X)$ будет существенно отличаться от H_{k-1} . Тогда и реальная вероятность $P(\rho \geq C)$ — точный размер полученного критерия — будет сильно отличаться от ε . Поэтому для выборки объема n число интервалов разбиения выбирают так, чтобы обеспечить нужную точность при замене распределения $\rho(X)$ на H_{k-1} .

Обычно требуют, чтобы $\pi p_1 = \dots = \pi p_k$ были не менее 5-6.

5.2. Проверка гипотезы независимости: критерий хи-квадрат Пирсона

Есть выборка $(X, Y) = ((X_1, Y_1), \dots, (X_n, Y_n))$ значений двух наблюдаемых совместно случайных величин X и Y в n независимых экспериментах. Проверяется гипотеза $H_1 = \{X \text{ и } Y \text{ независимы}\}$.

Введем k интервалов группировки $\Delta_1, \dots, \Delta_k$ для значений X и m интервалов группировки $\nabla_1, \dots, \nabla_m$ для значений Y .

Посчитаем эмпирические частоты:

$$n_{i,j} = \{\text{число пар } (X_i, Y_i), \text{ попавших в } \Delta_i \times \nabla_j\},$$

$$n_{i.} = \{\text{число } Y_i, \text{ попавших в } \nabla_j\}, n_{.i} = \{\text{число } X_i, \text{ попавших в } \Delta_i\}.$$

Y	∇_1	∇_2	...	∇_m	$\sum_{j=1}^m$
X					
Δ_1	$n_{1,1}$	$n_{1,2}$...	$n_{1,m}$	$n_{1.}$

⋮			...		⋮
Δ_k	$v_{k,1}$	$v_{k,2}$...	$v_{k,m}$	$v_{k,}$
$\sum_{i=1}^k$	$v_{.,1}$	$v_{.,2}$...	$v_{.,m}$	n

Если гипотеза H_1 верна, то теоретические вероятности попадания пары (X, Y) в любую из областей $\Delta_i \times \nabla_j$ равны произведению вероятностей: для всех i и j

$$p_{i,j} = P((X, Y) \in \Delta_i \times \nabla_j) = P(X \in \Delta_i) \cdot P(Y \in \nabla_j) = p_i^x \cdot p_j^y$$

Именно эту гипотезу (назовем ее H'_1) мы в действительности и проверяем.

По ЗБЧ

$$\frac{v_{i,}}{n} \approx p_i^x, \quad \frac{v_{.,j}}{n} \approx p_j^y, \quad \frac{v_{i,j}}{n} \approx p_{i,j}$$

Поэтому значительная разница между $\frac{v_{i,j}}{n}$ и $\frac{v_{i,} \cdot v_{.,j}}{n}$ (или между $\frac{v_{i,j}}{n}$ и $\frac{v_{i,} \cdot v_{.,j}}{n}$) может служить основанием для отклонения гипотезы независимости.

Пусть

$$\rho(X, Y) = n \sum_{i=1}^k \sum_{j=1}^m \frac{(v_{i,j} - (v_{i,} \cdot v_{.,j})/n)^2}{v_{i,} \cdot v_{.,j}} \quad (5)$$

Теорема 4.

Если гипотеза H_1 верна, то $\rho(X, Y) \Rightarrow H_{[k-1][m-1]}$ при $n \rightarrow \infty$.

Критерий согласия асимптотического уровня ϵ строится обычным образом.

6. Исследование статистической зависимости

Часто требуется определить, как зависит наблюдаемая случайная величина от одной или нескольких других величин. Самый общий случай такой зависимости — зависимость статистическая: например, $X = \xi - \eta$ и $Z = \xi - \phi$ зависимы, но эта зависимость не функциональная.

Для зависимых случайных величин имеет смысл рассмотреть математическое ожидание одной из них при фиксированном значении другой (других). Такое условное математическое ожидание показывает, как влияет на среднее значение первой величины изменение значений второй. Скажем, стоимость квартиры зависит от площади, этажа, района и других параметров, но не является функцией от них. Зато в широких предположениях можно считать ее математическое ожидание функцией от этих величин. Разумеется, наблюдать это среднее значение мы не можем — в нашей власти лишь наблюдать значения первой случайной величины при разных значениях остальных. Эту зависимость можно вообразить как вход и выход некоторой машины. Входные данные, или «факторы», как правило, известны. На выходе мы наблюдаем результат преобразования входных данных в ящике по каким-либо правилам.

6.1. Математическая модель регрессии

Пусть наблюдаемая случайная величина X зависит от случайной величины или случайного вектора Z . Значения Z мы либо задаем, либо наблюдаем. Обозначим через $f(t)$ функцию, отражающую зависимость среднего значения X от значений Z :

$$E(X | Z = t) = f(t). \quad (6)$$

Функция $f(t)$ называется **линией регрессии X на Z** , а уравнение $x = f(t)$ -- регрессионным уравнением. После n экспериментов, в которых Z последовательно принимает значения $Z = t_1, \dots, Z = t_n$, получим значения наблюдаемой величины X , равные X_1, \dots, X_n .

Обозначим через ϵ_i разницу

$$X_i - E(X | Z = t_i) = X_i - f(t_i)$$

между наблюдаемой в i -м эксперименте случайной величиной и ее математическим ожиданием.

Итак, $X_i = f(t_i) + \varepsilon_i$, $i = 1, \dots, n$, где ε_i — ошибки наблюдения, равные в точности разнице между реальным и усредненным значением случайной величины X при значении $Z = t_i$. Про совместное распределение $\varepsilon_1, \dots, \varepsilon_n$ обычно что-либо известно или предполагается: например, что вектор ошибок ε состоит из независимых и одинаково нормально распределенных случайных величин с нулевым средним.

Нулевое среднее тут необходимо:

$$E \varepsilon_i = E X_i - f(t_i) = E(X | Z = t_i) - f(t_i) = 0.$$

Требуется по значениям t_1, \dots, t_n и X_1, \dots, X_n оценить как можно точнее функцию $f(t)$. Величины t_i не являются случайными, так что вся случайность сосредоточена в неизвестных ошибках ε_i и в наблюдаемых X_i .

Но пытаться в классе всех возможных функций восстанавливать $f(t)$ по «наилучшим оценкам» для $f(t_i)$ довольно глупо — наиболее точными приближениями к $f(t_i)$ оказываются X_i , и функция $f(t)$ будет просто ломаной, построенной по точкам (t_i, X_i) . Поэтому сначала заранее определяют вид функции $f(t)$. Часто предполагают, что $f(t)$ есть полином (редко больше третьей или четвертой степени) с неизвестными коэффициентами. Будем пока предполагать, что функция $f(t)$ полностью определяется неизвестными параметрами $\theta_1, \dots, \theta_k$.

6.2. Метод максимального правдоподобия

Оценки неизвестных параметров находят с помощью метода максимального правдоподобия. Он предписывает выбирать неизвестные параметры так, чтобы максимизировать функцию правдоподобия случайного вектора X_1, \dots, X_n .

Будем, для простоты, предполагать, что вектор ошибок ε состоит из независимых и одинаково распределенных случайных величин с плотностью распределения $h(x)$ из некоторого семейства распределений с нулевым средним и, вообще говоря, неизвестной дисперсией. Очень часто полагают, что ε_i имеют симметричное распределение — нормальное N_{0, σ^2} , Стьюдента, Лапласа, логистическое и т.п. Поскольку X_i от ε_i зависят

линейно, то распределение X_i окажется таким же, как у ε_i , но с центром уже не в нуле, а в точке $f(t_i)$. Поэтому X_i имеет плотность $h(x - f(t_i))$, и функция правдоподобия вектора X_1, \dots, X_n равна, в силу независимости координат,

$$f(X_1, \dots, X_n; \theta_1, \dots, \theta_k) = h(X_1 - f(t_1)) \cdot \dots \cdot h(X_n - f(t_n)) = h(\varepsilon_1) \cdot \dots \cdot h(\varepsilon_n). \quad (7)$$

Если величины ε_i имеют разные распределения, то h следует заменить на соответствующие h_i . В отсутствие независимости произведение плотностей в (7) заменится плотностью совместного распределения координат вектора ε .

Метод максимального правдоподобия предписывает находить оценки неизвестных параметров θ_i функции $f(t)$ и оценки неизвестной дисперсии (или дисперсий) σ^2 , максимизируя по этим параметрам функцию правдоподобия (7). Рассмотрим, во что превращается метод максимального правдоподобия в наиболее частых на практике предположениях.

6.3. Метод наименьших квадратов

Предположим, что вектор ошибок ε состоит из независимых случайных величин с нормальным распределением $N(0, \sigma^2)$. Функция правдоподобия (7) имеет вид

$$f(\mathbf{X}; \theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_1 - f(t_1))^2}{2\sigma^2}\right\} \cdot \dots \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(X_n - f(t_n))^2}{2\sigma^2}\right\} = \\ = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - f(t_i))^2\right\}.$$

Очевидно, что при любом фиксированном σ^2 максимум функции правдоподобия достигается при наименьшем значении суммы квадратов ошибок $\sum (X_i - f(t_i))^2 = \sum \varepsilon_i^2$.

Определение 14.

Оценкой метода наименьших квадратов (ОМНК) для неизвестных параметров $\theta_1, \dots, \theta_k$ уравнения регрессии называется набор значений параметров, доставляющий минимум сумме квадратов отклонений

$$\sum_{i=1}^n (X_i - f(t_i))^2 = \sum_{i=1}^n \varepsilon_i^2.$$

Найдя оценки для t_i , найдем тем самым оценку $\hat{f}(t_i)$ для $f(t_i)$. Обозначим через $\hat{f}(t_i)$ значения этой функции, и через $\hat{\varepsilon}_i = X_i - \hat{f}(t_i)$ соответствующие оценки ошибок. Оценка максимального правдоподобия для σ^2 , она же точка максимума по σ^2 функции правдоподобия, равна вычислить!

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{f}(t_i))^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2. \quad (8)$$

Мудрый читатель понял, что основная цель рассмотренного выше примера — показать, что метод наименьших квадратов не падает с неба, а есть в точности метод максимального правдоподобия в случае, когда вектор ошибок, а вместе с ним и вектор наблюдаемых откликов регрессии, состоит из независимых и одинаково распределенных случайных величин с [нормальным распределением](#).

6.4. Общая модель линейной регрессии

Введем два вектора: $\mathbf{Z} = (Z_1, \dots, Z_k)$ — факторы регрессии и $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)$ — неизвестные параметры регрессии. Каждый вектор есть вектор-столбец, а изображен по горизонтали для удобства. Обозначать вектора мы, как и ранее, будем жирным шрифтом.

Рассматривается модель регрессии, которая в курсе «Эконометрика» называется [простой \(линейной\) регрессией](#):

$$F(X | Z = \mathbf{t}) = f(\mathbf{t}) = \beta_1 t_1 + \dots + \beta_k t_k, \text{ или } F(X | Z) = f(Z) = \beta_1 Z_1 + \dots + \beta_k Z_k.$$

Пусть в i -м эксперименте факторы регрессии принимают заранее заданные значения $\mathbf{Z}^{(i)} = (Z_1^{(i)}, \dots, Z_k^{(i)})$, где $i = 1, \dots, n$.

После $n \geq k$ экспериментов получен набор откликов $\mathbf{X} = (X_1, \dots, X_n)$, где

$$\begin{cases} X_1 = \beta_1 Z_1^{(1)} + \dots + \beta_k Z_k^{(1)} + \varepsilon_1 \\ X_2 = \beta_1 Z_1^{(2)} + \dots + \beta_k Z_k^{(2)} + \varepsilon_2 \\ \dots \\ X_n = \beta_1 Z_1^{(n)} + \dots + \beta_k Z_k^{(n)} + \varepsilon_n \end{cases}$$

или, в матричной форме, $\mathbf{X} = \mathbf{Z}^T \boldsymbol{\beta} - \boldsymbol{\varepsilon}$, где матрица $\mathbf{Z}(k \times n)$ (матрица плана) равна

$$\mathbf{Z} = \begin{pmatrix} z_1^{(1)} & \dots & z_1^{(n)} \\ \dots & \vdots & \dots \\ z_k^{(1)} & \dots & z_k^{(n)} \end{pmatrix} = (\mathbf{z}^{(1)} \dots \mathbf{z}^{(n)}).$$

Вектор $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ состоит из случайных ошибок в данных экспериментах.

Требуется по данным матрице плана \mathbf{Z} и вектору результатов \mathbf{X} найти оценки для параметров регрессии $\boldsymbol{\beta}$ и параметров распределения вектора ошибок $\boldsymbol{\varepsilon}$.