Министерство науки и высшего образования Российской Федерации

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

А. П. Ковалевский

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

НОВОСИБИРСК
2024

УДК 519.21(075.8)
К-822

УДК 519.21(075.8)

Оглавление

Глава 1

Статистики и оценки

§1. Выборка и выборочные характеристики

В математической статистике рассматривается ситуация, когда распределение наблюдаемой в случайном эксперименте величины X неизвестно (хотя бы частично), зато исследователь располагает результатами эксперимента (статистическими данными), по которым он должен сделать выводы о неизвестном распределении случайной величины X. К этому стоит добавить, что задачей математической статистики является использовать результаты эксперимента по возможности оптимальным образом.

Основным объектом исследования в математической статистике является **выборка** $\vec{X} = (X_1, X_2, ..., X_n)$, то есть набор значений случайной величины X, полученных в результате n независимых воспроизведений эксперимента. Иначе говоря, выборка представляет собой случайный вектор, координаты которого — **элементы выборки** $X_1, X_2, ..., X_n$ — независимые случайные величины, имеющие общее распределение с функцией распределения $F(t)$.

Конкретный набор числовых значений случайных величин $X_1, X_2, ..., X_n$, полученный в результате эксперимента, будем называть **реализацией** выборки и обозначать $\vec{x} = (x_1, x_2, ..., x_n)$.

Если элементы выборки $X_1, \ldots, X_n$ упорядочить по возрастанию, то получится новый набор случайных величин, называемый **вариационным рядом**:

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n-1)} \le X_{(n)}.$$

В частности, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Эмпирической функцией распределения $F_n^*(t)$ называется относительная частота элементов выборки, не больших заданного t. Эмпирическая функция распределения, соответствующая выборке $\vec{X} = (X_1, X_2, ..., X_n)$, может быть построена по этой выборке с помощью любой из следующих

формул:

$$F_n^*(t) = \frac{\text{число } X_i \le t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(X_i \le t), \tag{1.1}$$

где функция

$$\mathbf{I}(X_i \le t) = \begin{cases} 1, & \text{если } X_i \le t, \\ 0 & \text{иначе,} \end{cases}$$

— индикатор события $\{X_i \le t\}$.

Заметим, что эмпирическая функция распределения, соответствующая случайной выборке $\vec{X}$, сама является случайной, поскольку определяется через элементы выборки $X_1,\ X_2, \ldots, X_n$, являющиеся случайными величинами. В то же время любая реализация $\vec{x} = (x_1, x_2, \ldots, x_n)$ выборки $\vec{X}$ порождает соответствующую реализацию эмпирической функции распределения (по той же формуле (1.1)), которая является обычной (а не случайной) функцией распределения.

Эмпирическая функция распределения $F_n^*(t)$ является выборочным аналогом неизвестной теоретической функции распределения $F(t)$, ее называют также **оценкой** для $F(t)$. Выборочным аналогом для теоретической плотности распределения $f(t)$ является **гистограмма**, или **эмпирическая плотность распределения**, которая строится по выборке $\vec{X} = (X_1, X_2, ..., X_n)$ следующим образом.

Пусть $h > 0$ — произвольное число. Разобьем область значений изучаемой случайной величины (например, всю числовую ось) на промежутки $\Delta_k = [z_{k-1}, z_k)$ длины h и построим ступенчатую функцию $f_n^*(t)$, которая на каждом промежутке Δ_k принимает постоянное значение, вычисляемое по любой из формул:

$$f_n^*(t) = \frac{\text{число } X_i \in \Delta_k}{nh} = \frac{1}{nh}\sum_{i=1}^{n} \mathbf{I}(X_i \in \Delta_k) = \frac{\nu_k}{nh},\ t \in \Delta_k, \tag{1.2}$$

где ν_k — число элементов выборки, попавших в промежуток Δ_k. Так построенная функция называется гистограммой с шагом h и имеет график, изображенный на рис. 1.1.

Заметим, что площадь каждого прямоугольника гистограммы равна $\frac{\nu_k}{n}$, то есть частоте попадания в соответствующий интервал Δ_k.

Иногда шаг гистограммы h выбирают следующим образом. Сначала находят число интервалов K по *формуле Стеджеса*

$$K = [\log_2 n] + 1. \tag{1.3}$$

Здесь $[\cdot]$ — целая часть числа. Потом длина интервала рассчитывается по

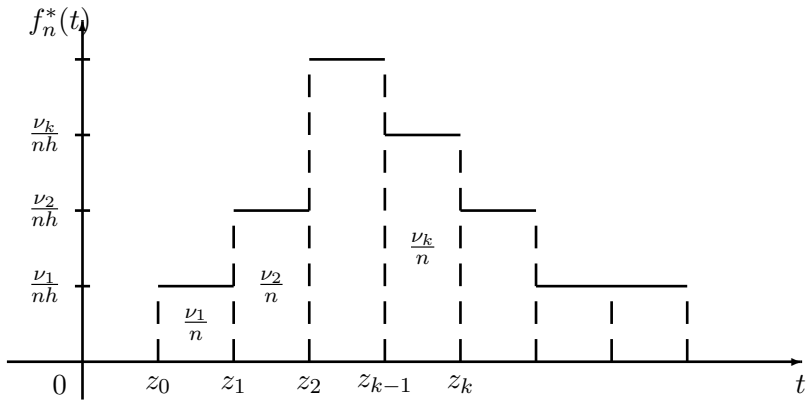

Рис. 1.1. Гистограмма $f_n^*(t)$.

формуле

$$h = \frac{X_{(n)} - X_{(1)}}{K}.$$

При построении гистограммы последний промежуток выбирается замкнутым: $\Delta_K = [z_{K-1};\ z_K]$. Величину $X_{(n)} - X_{(1)} = \max\{X_i\} - \min\{X_i\}$ называют размахом выборки.

Задачи для самостоятельного решения

1.1 По данной реализации выборки $\mathbf{x} = (0, 0, 1, 1, 0, 0, 0, 0, 0, 1)$:

а) построить график эмпирической функции распределения;

б) построить график гистограммы;

в) вычислить выборочное среднее, выборочную дисперсию, несмещенную выборочную дисперсию.

1.2 Пассажир маршрутного такси измерил 8 раз время ожидания такси и получил следующие результаты (в минутах): 8; 4; 5; 4; 2; 15; 1; 6. У него есть две гипотезы относительно графика движения такси: либо график движения соблюдается, и время ожидания имеет равномерное распределение на отрезке $[0,\ \theta]$, либо график движения не соблюдается, и время ожидания имеет показательное распределение с параметром λ.

а) Вычислить реализации оценок параметров θ и λ, использовав оценки $\widetilde{\theta} = (n+1)X_{(n)}/n$ и $\widetilde{\lambda} = \frac{n-1}{n\overline{X}}$.

б) Построить на одном графике реализацию эмпирической функции распределения и теоретические функции распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок.

в) Построить на одном графике реализацию гистограммы и теоретические плотности распределения равномерного и показательного законов, в которые вместо неизвестных параметров подставлены реализации их оценок. На основании проведенного исследования сделать вывод о том, какая из гипотез

выглядит более соответствующей экспериментальным данным.

Индивидуальные домашние задания

Каждой букве своих имени, отчества и фамилии сопоставьте ее номер в алфавите в соответствии с таблицей. По полученной выборке постройте вариационный ряд, графики эмпирической функции распределения и гистограммы. Среди классиков русской литературы найдите того, у которого распределение букв имени, отчества и фамилии лучше всего соответствует Вашему. Постройте для него тоже вариационный ряд, графики эмпирической функции распределения и гистограммы.

а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33

§2. Оценки параметров

Задача оценивания параметров возникает в ситуации, когда распределение F не является полностью неизвестным, а известен его математический вид $F = F(t, \theta)$, содержащий неизвестный параметр θ (или несколько, тогда θ - многомерный параметр). Задача состоит в том, чтобы по выборке $\vec{X}$ вычислить приближенное значение $\theta^*(\vec{X})$ для неизвестного параметра, причем сделать это в том или ином смысле оптимальным образом. Это задача **точечного оценивания**. Другой подход состоит в построении по выборке $\mathbf{X}$ интервала $(\theta_-(\vec{X});\ \theta_+(\vec{X}))$, который накрывает неизвестное значение параметра θ с заданной (высокой) вероятностью. Этот подход называется **интервальным оцениванием**, а $(\theta_-(\vec{X});\ \theta_+(\vec{X}))$ называется **доверительным интервалом**.

Пусть $\vec{X} \Subset F(t, \theta)$, причем параметр θ может принимать значения из множества Θ, которое называется **параметрическим множеством**. Будем называть **статистикой** любую случайную величину вида $T(\vec{X})$, которая является функцией **только от элементов выборки**. **Оценкой параметра** θ называется статистика $\tilde{\theta} = \tilde{\theta}(\vec{X})$, которая принимает значения из параметрического множества Θ.

Оценка $\tilde{\theta}$ называется **несмещенной** оценкой параметра θ, если для любого $\theta \in \Theta$ выполнено

$$\mathbf{E}\tilde{\theta} = \theta. \tag{1.4}$$

Договоримся указывать в обозначении статистики объем выборки, если это необходимо подчеркнуть: $\tilde{\theta} = \tilde{\theta}_n$.

Оценка $\tilde{\theta}_n$ называется **(сильно) состоятельной оценкой параметра** θ, если для любого $\theta \in \Theta$ при $n \to \infty$ имеет место сходимость

$$\tilde{\theta}_n \overset{\text{п. н.}}{\longrightarrow} \theta, \tag{1.5}$$

то есть $\mathbf{P}\{\tilde{\theta}_n \to \theta\} = 1$.

По выборке $\vec{X} = (X_1, X_2, ..., X_n)$ можно построить эмпирические (выборочные) аналоги числовых характеристик распределения. Наиболее употребительными являются выборочное математическое ожидание, или **выборочное среднее**, $\overline{X}$, и **выборочная дисперсия** S^2:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{1.6}$$

Подобно выборочным среднему и дисперсии определяются выборочные моменты порядка k

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

которые являются эмпирическими аналогами моментов $\alpha_k = \mathbf{E}X_i^k$.

Пример 1.1 *Предполагая известными соответствующие теоретические моменты, доказать, что* $\mathbf{E}\overline{X^k} = \alpha_k$.

Приведенное соотношение означает, что математические ожидания эмпирических моментов совпадают с соответствующими теоретическими моментами. Это свойство называется **несмещенностью**: говорят, что эмпирические моменты являются **несмещенными оценками** для соответствующих теоретических.

Решение. Из свойств математического ожидания получаем:

$$\mathbf{E}\overline{X^k} = \mathbf{E}\frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}X_i^k = \frac{1}{n}\sum_{i=1}^{n}\alpha_k = \frac{1}{n}\cdot n\alpha_k = \alpha_k.$$

В то же время центральные эмпирические моменты являются смещенными оценками для своих теоретических аналогов.

Отметим, что выборочная дисперсия вычисляется аналогично дисперсии.

Следствие 1.1 $S^2 = \overline{X^2} - (\overline{X})^2$.

Доказательство.

Раскроем скобки в определении S^2:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 - 2\frac{\overline{X}}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n}(\overline{X})^2 = \overline{X^2} - 2(\overline{X})^2 + (\overline{X})^2 = \overline{X^2} - (\overline{X})^2.$$

Вычислим математическое ожидание статистики S^2:

$$\mathbf{E}S^2 = \mathbf{E}\overline{X^2} - \left(\mathbf{E}\overline{X}\right)^2 = \mathbf{E}\overline{X^2} - (\mathbf{E}\overline{X})^2 - \mathbf{Var}\overline{X} = \frac{n-1}{n}\mathbf{Var}X_1.$$

Итак, эта оценка является асимптотически несмещенной.

Для того, чтобы получить несмещенную оценку дисперсии, делят S^2 на $\frac{n-1}{n}$.

Несмещенная выборочная дисперсия — это статистика

$${S_0}^2 = \frac{n}{n-1}S^2.$$

Для нее выполнено свойство

$$\mathbf{E}{S_0}^2 = \mathbf{Var}X_1.$$

Отметим, что корень из несмещенной выборочной дисперсии S_0 не является несмещенной оценкой для стандартного отклонения σ_X, так как для любой невырожденной случайной величины Y в силу неравенства Йенсена выполнено $\mathbf{E}\sqrt{Y} < \sqrt{\mathbf{E}Y}$.

Задачи для самостоятельного решения

2.1 Для выборки из равномерного распределения $U[0, \theta]$ предлагаются две оценки параметра $\theta > 0$:

$$\theta^* = c\overline{X}, \quad \widehat{\theta} = X_{(n)}.$$

Найдите такое значение константы c, чтобы оценка θ^* была несмещенной. Найдите математическое ожидание оценки $\widehat{\theta}$. Как ее исправить, чтобы она стала несмещенной? По реализации выборки $\mathbf{x} = (2, 2, 1, 1, 4)$ постройте график эмпирической функции распределения и оценок функции распределения; постройте график гистограммы и оценок плотности распределения.

2.2 По реализации выборки $\mathbf{x} = (2, 3, 1, 3, 4, 3, 2, 5)$ из нормального распределения постройте на одном графике эмпирическую функцию распределения и оценку функции распределения. Постройте на одном графике гистограмму с шагом, равным среднеквадратическому (стандартному) отклонению, и оценку плотности распределения.

2.3 Найдите функции numpy, вычисляющие выборочное среднее, выборочную дисперсию, несмещенную выборочную дисперсию и выборочное стандартное отклонение.

2.4 Измерен рост (в см) студентов одной учебной группы. Результаты измерений дали выборку (171; 186; 164; 190; 158; 181; 176; 180; 174; 157; 176; 169; 164; 186).

а) Построить реализацию гистограммы.

б) Вычислить реализации выборочного среднего, выборочной дисперсии и выборочного стандартного отклонения S. На одном графике с гистограммой построить график плотности нормального закона с параметрами $\overline{X}$, S^2.

Индивидуальные домашние задания

Каждой букве своих имени, отчества и фамилии сопоставьте ее номер в алфавите в соответствии с таблицей. Составьте две новые выборки: одну — попарным сложением номеров букв, другую — вычитанием. Для первой выборки используйте модель нормального распределения с неизвестными параметрами, для второй — модель нормального распределения с нулевым математическим ожиданием и неизвестной дисперсией. Подставляя вместо неизвестных параметров их точечные оценки, запишите выражения для оценки функции и плотности распределения. Постройте на одном графике эмпирическую функцию распределения и оценку функции распределения. Постройте на одном графике гистограмму с шагом, равным среднеквадратическому (стандартному) отклонению, и оценку плотности распределения.

§3. Оценки методом моментов

Пусть $\theta \in \Theta$ - одномерный параметр, и $g : R \longrightarrow R$ некоторая числовая функция. Тогда по данной выборке $\vec{X} = (X_1, X_2, ..., X_n)$ можно построить выборку $g(\vec{X}) = (g(X_1), g(X_2), ..., g(X_n))$. Обозначим

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

выборочное среднее этой выборки. С другой стороны, можно найти теоретическое среднее

$$m(\theta) = \mathbf{E}g(X_1).$$

Определение 1.1 *Оценкой метода моментов* (ОММ) называется такое значение θ^*, при котором теоретическое среднее выборки $\mathbf{g(X)}$ совпадает с выборочным средним:

$$m(\theta^*) = \overline{g(X)}, \tag{1.7}$$

то есть ОММ является решением уравнения (1.7) относительно неизвестного θ^*.

В качестве функции g чаще всего выбирают степенные функции: $g(x) = x^k$, где $k = 1, 2, ...$ Оценка по методу моментов в этом случае называется *оценкой по k-тому моменту* и обозначается θ_k^*.

Задачи для самостоятельного решения

3.1 По выборке $(X_1,\ \ldots,\ X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0;\ 1)$ построить оценки параметра p:

а) по первому моменту;

б) по второму моменту;

в) по произвольному k-му моменту.

Можно ли отдать предпочтение какой-либо из построенных оценок?

3.2 При каких значениях параметра $\theta > 0$ распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1, \\ 0, & t < 1 \end{cases}$$

существует оценка параметра по первому моменту?

3.3 По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценки методом моментов:

а) параметра p по первому и по второму моменту при известном $m > 0$;

б) параметров p и m.

Исследовать состоятельность построенных оценок.

3.4 По выборке $(X_1,\ \ldots,\ X_n)$ из распределения Лапласа с плотностью $f_\lambda(t) = \frac{\lambda}{2}e^{-\lambda|t|},\ \ t \in \mathbf{R}$, построить оценку параметра $\lambda > 0$ методом моментов.

3.5 Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод моментов, построить оценки

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

3.6 Используя метод моментов, оценить параметр θ равномерного распределения на отрезке

а) $[-\theta;\ \theta]$, $\theta > 0$; б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

3.7 С помощью метода моментов построить оценку параметра $\theta > 0$, если распределение выборки имеет плотность

а) $\theta t^{\theta-1}$ при $t \in [0;\ 1]$; б) $2t/\theta^2$ при $t \in [0;\ \theta]$.

Исследовать полученные оценки на состоятельность.

3.8 Дана выборка из распределения с плотностью

$$f_\theta(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ 1], \\ 0, & t \notin [0;\ 1]. \end{cases}$$

Найти оценку параметра $\theta > 0$ методом моментов, исследовать ее на несмещенность и состоятельность.

3.9 Методом моментов найти оценку параметра $\alpha > 0$ по выборке из показательного распределения с плотностью $f_\alpha(t) = \alpha e^{-\alpha t},\ t > 0$. Будет ли оценка несмещенной и состоятельной?

3.10 По выборке $(X_1, \ldots, X_n)$ методом моментов найти две различные оценки параметра $p \in (0, 1)$, если известно, что

$P\{X_1 = 1\} = p/2,\ P\{X_1 = 2\} = p/2,\ P\{X_1 = 3\} = 1 - p.$

Будут ли полученные оценки несмещенными и состоятельными?

3.11 В тексте задачи через N обозначен номер студента по списку группы.

1. Для выборки $X_1,\ \ldots,\ X_n$ из равномерного распределения на $[0,\ \theta]$ получить оценки параметра θ методом моментов на основании первого, второго, $(N{+}2)$-го момента.

2. Генерировать реализацию выборки объема $n = 100{+}N$ из равномерного распределения на $[0,\ \theta]$, приняв $\theta = N$.

3. Вычислить реализации всех полученных оценок, а также оценки $\breve{\theta} = \frac{n+1}{n} X_{(n)}$. Подсчитать абсолютные погрешности оценивания и ранжировать оценки по абсолютной погрешности.

Индивидуальные домашние задания

Выбрать роман на русском языке (не менее 20 000 слов). Вычислить, сколько союзов «и» содержится в каждом предложении. Рассмотреть следующие вероятностные модели:

1. Пуассоновское распределение

$\mathbf{P}(X_1 = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ k \geq 0,\ \lambda > 0.$

2. Сдвинутое геометрическое распределение

$\mathbf{P}(X_1 = k) = p(1-p)^k,\ k \geq 0,\ 0 < p < 1.$

Найти оценки параметров моделей по первому моменту. Вычислить оценки вероятностей принять значения от 0 до 10 и оценки функции распределения. Построить на одном графике относительные частоты и оценки вероятностей. Построить на другом графике эмпирическую функцию распределения и оценки функции распределения. Сделать вывод, какая из вероятностных моделей лучше соответствует эмпирическим данным.

§4. Оценки максимального правдоподобия

Пусть $\vec{X} \Subset F(t,\theta),\ \theta \in \Theta$. Предположим, что теоретическое распределение либо абсолютно непрерывно с плотностью $f(t,\theta) = f_{X_i}(t)$, либо дискретно, при этом для ряда распределения будем использовать то же обозначение: $f(t,\theta) = \mathbf{P}(X_i = t)$. **Функцией правдоподобия, соответствующей выборке** $\vec{X}$, называется функция

$$\Pi(\theta) = \Pi(\vec{X},\ \theta) = \prod_{i=1}^{n} f(X_i,\ \theta). \tag{1.8}$$

Оценкой максимального правдоподобия (ОМП) называется такое значение параметра $\theta = \hat{\theta}(\vec{X})$, при котором функция правдоподобия принимает наибольшее значение, то есть

$$\Pi(\vec{X},\ \hat{\theta}) = \max_{\theta\in\Theta} \Pi(\vec{X},\ \theta). \tag{1.9}$$

Если функция правдоподобия дифференцируема при всех $\theta \in \Theta$, то значение $\theta = \hat{\theta}$ должно быть решением уравнения

$$\Pi'(\theta) = 0, \tag{1.10}$$

которое называется уравнением правдоподобия, или эквивалентного уравнения

$$\frac{d}{d\theta} \ln \Pi(\theta) = 0 \iff \sum_{i=1}^{n} \frac{d}{d\theta} \ln f(X_i,\theta) = 0. \tag{1.11}$$

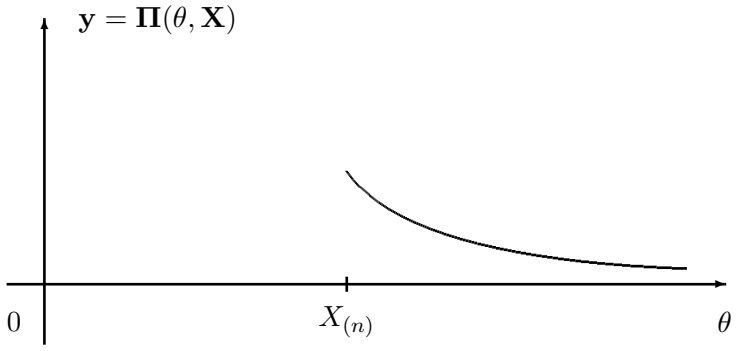

Рис. 1.2. Функция правдоподобия для $\vec{X} \Subset U_{[0,\theta]}$.

Рассмотрим теперь случай многомерного параметра, предположив опять для простоты, что $\theta = (\theta_1,\theta_2)$ - двумерный параметр. Тогда для нахождения ОМП нужно найти точку $\hat{\theta} = (\hat{\theta}_1,\hat{\theta}_2)$ наибольшего значения функции двух переменных $\Pi(\theta_1,\theta_2)$. В частности, если функция правдоподобия дифференцируема, то для решения этой задачи, вместо уравнения правдоподо-

бия (1.10) или (1.11), нужно найти решение следующей системы уравнений:

$$\begin{cases} \frac{\partial \Pi(\theta_1,\theta_2)}{\partial \theta_1} = 0, \\ \\ \frac{\partial \Pi(\theta_1,\theta_2)}{\partial \theta_2} = 0. \end{cases} \tag{1.12}$$

Решение типовых примеров

Пример Для распределения Пуассона найти ОМП неизвестного параметра λ.

Решение. Для распределения Пуассона Π_λ ряд распределения имеет вид:

$$f(t,\lambda) = \mathbf{P}(X_i = t) = e^{-\lambda}\frac{\lambda^t}{t!}.$$

Искомая оценка должна быть решением уравнения правдоподобия (1.10) или (1.11). Для решения этого уравнения вычислим последовательно:

$$f(X_i,\lambda) = e^{-\lambda}\frac{\lambda^{X_i}}{X_i!}, \quad \ln f(X_i,\lambda) = -\lambda + X_i \ln\lambda - \ln(X_i!),$$

$$\frac{d}{d\lambda}\ln f(X_i,\lambda) = -1 + \frac{1}{\lambda}X_i,$$

$$\sum_{i=1}^{n}\frac{d}{d\lambda}\ln f(X_i,\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = -n + \frac{1}{\lambda}n\overline{X}.$$

Тогда уравнение (1.11) и его решение имеют вид:

$$-n + \frac{1}{\lambda}n\overline{X} = 0 \iff \lambda = \overline{X}.$$

Заметим, что вторая производная логарифмической функции правдоподобия

$$\frac{d^2}{d\lambda^2}\ln\Pi(\lambda) = -\frac{1}{\lambda^2}n\overline{X} < 0$$

при всех λ, так как при нашем предположении $\vec{X} \in \Pi_\lambda$ все элементы выборки $X_1, X_2, ..., X_n$, а значит, и выборочное среднее $\overline{X}$, с вероятностью единица неотрицательны. Значит, найденное решение $\lambda = \overline{X}$ уравнения правдоподобия является единственной точкой максимума функций $\Pi(\lambda)$ и $\ln\Pi(\lambda)$, а следовательно, статистика $\hat{\lambda} = \overline{X}$ является ОМП параметра λ.

Пример Пусть $\vec{X} \sim U_{[0,\theta]}$, где $\theta > 0$. Найти ОМП для параметра θ.

Решение. Найдем функцию правдоподобия, соответствующую выборке $\vec{X}$ из равномерного распределения $U_{[0,\theta]}$. Плотность распределения закона $U_{[0,\theta]}$ при $t = X_i$ равна:

$$f(X_i,\theta) = \begin{cases} \frac{1}{\theta}, & \text{если } X_i \in [0,\theta], \\ 0, & \text{если } X_i \notin [0,\theta] \end{cases} \tag{1.13}$$

Тогда функция правдоподобия вычисляется следующим образом:

$$\Pi(\theta) = \Pi(\vec{X},\ \theta) = \prod_{i=1}^{n} f(X_i, \theta) = \begin{cases} \frac{1}{\theta^n}, & \text{если } X_i \in [0, \theta] \text{ для всех } i = 1, 2, ..., n; \\ 0, & \text{иначе.} \end{cases}$$

Это соотношение можно переписать в следующих равносильных формах:

$$\Pi(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{если } \max\limits_{1 \le i \le n} X_i \le \theta; \\ 0, & \text{иначе.} \end{cases} \iff \Pi(\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{если } \theta > X_{(n)}; \\ 0, & \text{иначе.} \end{cases}$$

Последнее задание функции $\Pi(\theta) = \Pi(\vec{X},\ \theta)$ позволяет легко изобразить ее график (см. рис. 1.2). Из графика видно, что своего наибольшего значения функция $\Pi(\theta)$ достигает при $\theta = X_{(n)}$. Следовательно, оценка максимального правдоподобия имеет вид: $\hat{\theta}(\vec{X}) = X_{(n)}$.

Пример Пусть $\vec{X} \sim U_{[a,b]}$, где $a < b$. найти ОМП неизвестного двумерного параметра (a, b).

Решение. Найдем функцию правдоподобия, соответствующую выборке $\vec{X}$ из равномерного распределения $U_{[a,b]}$. Так как плотность распределения закона $U_{[a,b]}$ при $t = X_i$ равна

$$f(X_i, \theta) = \begin{cases} \frac{1}{b-a}, & \text{если } X_i \in [a, b], \\ 0, & \text{если } X_i \notin [a, b], \end{cases}$$

то функция правдоподобия представляется в виде

$$\Pi(a, b) = \prod_{i=1}^{n} p(X_i, \theta) = \begin{cases} \frac{1}{(b-a)^n}, & \text{если все} X_i \in [a, b]; \\ 0, & \text{иначе.} \end{cases}$$

Или по-другому:

$$\Pi(a, b) = \begin{cases} \frac{1}{(b-a)^n}, & \text{если } b \ge X_{(n)},\ a \le X_{(1)}; \\ 0, & \text{иначе.} \end{cases} \tag{1.14}$$

Из последнего равенства видно, что функция правдоподобия отлична от нуля (более того, строго положительна) лишь при значениях (a, b), удовлетворяющих неравенствам:

$$a \le X_{(1)} \le X_{(n)} \le b.$$

Значит, своего наибольшего значения функция $\Pi(a, b)$ достигает лишь при таких (a, b). Однако при таких значениях (a, b) разность $(b - a)$ принимает свое наименьшее значение $(X_{(n)} - X_{(1)})$ при $b = X_{(n)}$, $a = X_{(1)}$. А значит, функция $\Pi(a, b)$ принимает свое наибольшее значение при тех же значениях (a, b), то есть искомая ОМП имеет вид: $\hat{b} = X_{(n)},\ \hat{a} = X_{(1)}$.

Задачи для самостоятельного решения

4.1 По выборке $(X_1,\ \ldots,\ X_n)$ из бернуллиевского распределения B_p с неизвестным параметром $p \in (0;\ 1)$ построить оценку параметра p методом максимального правдоподобия. (Указание: показать, что вероятность попадания в точку t для элементов выборки равна $f(t,\ p) = p^t(1-p)^{1-t}$, где t может принимать только два значения: 0 и 1). Исследовать состоятельность и несмещенность полученной оценки.

4.2 По выборке $(X_1,\ \ldots,\ X_n)$ из биномиального распределения $B_{m,p}$ построить оценку максимального правдоподобия параметра p при известном $m > 0$. Исследовать состоятельность и несмещенность оценки.

4.3 По выборке из показательного распределения E_α построить оценку максимального правдоподобия параметра $\alpha > 0$. Исследовать состоятельность оценки.

4.4 Построить оценку максимального правдоподобия по выборке из распределения Парето с плотностью

$$f_\theta(t) = \begin{cases} \frac{\theta}{t^{\theta+1}}, & t \geq 1, \\ 0, & t < 1 \end{cases}.$$

Доказать состоятельность полученной оценки.

4.5 Пусть дана выборка из нормального распределения с параметрами α и σ^2. Используя метод максимального правдоподобия, построить оценки

а) неизвестного математического ожидания α;

б) неизвестной дисперсии σ^2, если α известно;

в) неизвестной дисперсии σ^2, если α неизвестно.

Исследовать полученные оценки на несмещенность и состоятельность.

4.6 Используя метод максимального правдоподобия, оценить параметр θ равномерного распределения на отрезке

а) $[-\theta;\ \theta]$, $\theta > 0$; б) $[\theta;\ \theta + 1]$.

Исследовать полученные оценки на несмещенность и состоятельность.

4.7 По выборке $(X_1, \ldots, X_n)$ методом максимального правдоподобия найти оценку параметра $p \in (0, 1)$, если известно, что

$P\{X_1 = 1\} = p/2,\ P\{X_1 = 2\} = p/2,\ P\{X_1 = 3\} = 1 - p.$

Будет ли полученная оценка несмещенной и состоятельной?

4.8 По реализации $\vec{x} = (0;\ 2;\ 0;\ 3)$ выборки из распределения Пуассона с параметром $\lambda > 0$ найти реализации оценок параметра λ по первому и второму моментам и оценки максимального правдоподобия.

Индивидуальные домашние задания

Найти оценку максимального правдоподобия неизвестного параметра по

выборке $(X_1, \ldots, X_n)$ из заданного семейства распределений. Проверить ее состоятельность.

Вариант 1. Плотность распределения равна

$$f(x) = \begin{cases} \frac{3x^2}{\theta} e^{-x^3/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 2. Плотность распределения равна

$$f(x) = \begin{cases} \theta x^{-(\theta+1)} & \text{при } x > 1; \\ 0 & \text{при } x \le 1. \end{cases}$$

Вариант 3. Плотность распределения равна

$$f(x) = \begin{cases} \frac{2x}{\theta} e^{-x^2/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 4. Плотность распределения равна

$$f(x) = \begin{cases} \frac{3\sqrt{x}}{2\theta} e^{-x\sqrt{x}/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 5. Плотность распределения равна

$$f(x) = \begin{cases} \frac{1}{2\theta\sqrt{x}} e^{-\sqrt{x}/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 6. Плотность распределения равна

$$f(x) = \begin{cases} \frac{1}{2\theta\sqrt{x}} e^{-\sqrt{x}/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 7. Плотность распределения равна

$$f(x) = \begin{cases} \frac{x}{\theta^2} e^{-x/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 8. Плотность распределения равна

$$f(x) = \begin{cases} \frac{1}{\theta} x^{-(\theta+1)/\theta} & \text{при } x > 1; \\ 0 & \text{при } x \le 1. \end{cases}$$

Вариант 9. Плотность распределения равна

$$f(x) = \begin{cases} (\theta - 1) x^{-\theta} & \text{при } x > 1; \\ 0 & \text{при } x \le 1. \end{cases}$$

Вариант 10. Плотность распределения равна

$$f(x) = \begin{cases} \frac{2}{\sqrt{\pi\theta}} e^{-x^2/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 11. Плотность распределения равна

$$f(x) = \begin{cases} \frac{x^2}{2\theta^3} e^{-x/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 12. Плотность распределения равна

$$f(x) = \begin{cases} \frac{x}{\theta^2} e^{-x/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 13. Плотность распределения равна

$$f(x) = \begin{cases} \frac{1}{\theta-1} x^{-\theta/(\theta-1)} & \text{при } x > 1; \\ 0 & \text{при } x \le 1. \end{cases}$$

Вариант 14. Плотность распределения равна

$$f(x) = \begin{cases} \frac{4x^3}{\theta} e^{-x^4/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 15. Плотность распределения равна

$$f(x) = \begin{cases} \frac{2}{\theta} e^{-2x/\theta} & \text{при } x > 0; \\ 0 & \text{при } x \le 0. \end{cases}$$

Вариант 16. Плотность распределения равна

$$f(t) = \begin{cases} e^{\theta - t}, & t \ge \theta; \\ 0, & t < \theta. \end{cases}$$

Вариант 17. Плотность распределения равна

$$f(t) = \frac{\lambda}{2} e^{-\lambda|t|}, \quad t \in \mathbf{R}.$$

Вариант 18. Плотность распределения равна

$$\theta t^{\theta-1}, \quad t \in [0;\ 1].$$

Вариант 19. Плотность распределения равна

$$2t/\theta^2, \quad t \in [0;\ \theta].$$

Вариант 20. Плотность распределения равна

$$f(t) = \begin{cases} 3t^2\theta^{-3}, & t \in [0;\ 1], \\ 0, & t \notin [0;\ 1]. \end{cases}$$

§5. Сравнение оценок

Пусть $\vec{X} \sim F(t,\theta),\ \ \theta \in \Theta,$ и $\widetilde{\theta} = \widetilde{\theta}(\vec{X})$ — какая-нибудь оценка параметра θ. Так как оценка является случайной величиной, то даже свойство несмещенности не гарантирует близость ее конкретной реализации $\widetilde{\theta}(\vec{x})$ к оцениваемому параметру. Если оценка является состоятельной, то такая близость гарантируется с заданной вероятностью, но только при достаточно больших объемах выборки n. При фиксированном объеме выборки наиболее распространенной «мерой близости» оценки к оцениваемому параметру является **квадратическая характеристика**, или среднее значение квадрата отклонения $\mathbf{E}(\widetilde{\theta} - \theta)^2$.

Из двух оценок $\widetilde{\theta}_1$ считается **лучше**, чем $\widetilde{\theta}_2$, если при всех $\theta \in \Theta$ выполняется неравенство

$$\mathbf{E}(\widetilde{\theta}_1 - \theta)^2 \le \mathbf{E}(\widetilde{\theta}_2 - \theta)^2, \tag{1.15}$$

а хотя бы для одного θ неравенство в (1.15) оказывается строгим.

Заметим, что квадратическая характеристика оценки не меньше ее дисперсии, и равенство достигается для несмещенных оценок:

$$\mathbf{E}(\widetilde{\theta} - \theta)^2 = \Big(\mathbf{E}(\widetilde{\theta} - \theta)\Big)^2 + \mathbf{Var}(\widetilde{\theta} - \theta) = \Big(\mathbf{E}\widetilde{\theta} - \theta\Big)^2 + \mathbf{Var}\,\widetilde{\theta} \ge \mathbf{Var}\,\widetilde{\theta}.$$

Если $\widetilde{\theta}$ — несмещенная оценка параметра θ, то есть $\mathbf{E}\widetilde{\theta} = \theta$, то для нее

$$\mathbf{E}(\widetilde{\theta} - \theta)^2 = \Big(\mathbf{E}\widetilde{\theta} - \theta\Big)^2 + \mathbf{Var}\,\widetilde{\theta} = \mathbf{Var}\,\widetilde{\theta}.$$

Отметим, что при среднеквадратическом подходе к сравнению оценок нельзя найти наилучшую в классе всех оценок (в частности, существуют несравнимые оценки).

Теорема В невырожденной статистической задаче (то есть в ситуации, когда по выборке нельзя однозначно определить неизвестный параметр) не существует наилучшей в классе всех оценок.

Доказательство

Предположим, что существует наилучшая в классе всех оценок параметра θ оценка $\check{\theta}$. Рассмотрим следующую дурацкую оценку $\tilde{\theta}_d = \theta_0$. Эта оценка равняется константе θ_0 для любых выборочных значений, то есть никак не использует информацию, представленную выборкой. Однако если θ_0 — возможное значение параметра θ, то нельзя исключить ситуации, когда $\theta = \theta_0$ (дурацкая оценка может оказаться наиболее верной, если правильно угадывает значение неизвестного параметра). В этом случае квадратичная характеристика этой оценки равна 0:

$$\mathbf{E}(\tilde{\theta}_d - \theta)^2 = \mathbf{E}(\theta_0 - \theta_0)^2 = 0.$$

Но если оценка $\check{\theta}$ не хуже, чем $\tilde{\theta}_d$, то ее квадратичная характеристика всегда не больше, чем квадратичная характеристика оценки $\tilde{\theta}_d$ для любого θ. В частности, при $\theta = \theta_0$ получаем:

$$\mathbf{E}(\check{\theta} - \theta_0)^2 \leq \mathbf{E}(\theta_0 - \theta_0)^2 = 0.$$

Отсюда $\mathbf{E}(\check{\theta} - \theta_0)^2 = 0$, то есть $\check{\theta} = \theta_0$ с вероятностью 1. Это противоречит невырожденности статистической задачи — предполагалось, что по выборке нельзя однозначно определить неизвестный параметр. Итак, не может существовать наилучшей в классе всех оценок. Теорема доказана.

Для того, чтобы избежать необходимости сравнивать получаемые оценки с вырожденными оценками (рассмотренными в доказательстве теоремы), нужно ограничить класс рассматриваемых оценок. Как правило, сравнивают только несмещенные оценки. Среди несмещенных оценок наилучшая оценка параметра для заданного параметрического семейства может существовать. Ее называют *эффективной* оценкой. Эффективная оценка имеет наименьшую дисперсию из всех несмещенных оценок.

К сожалению, такое определение эффективной оценки непригодно для практического использования, так как для проверки оптимальности одной оценки требуется сравнивать дисперсии всех оценок всех несмещенных оценок. Поэтому желательно иметь критерий, позволяющий проверять оптимальность оценки на основании характеристик распределения только этой оценки. Один из таких критериев основан на неравенстве Рао-Крамера, которое сформулировано ниже.

Для формулировки точного результата введем дополнительное условие. Пусть функция распределения $F(t, \theta)$ рассматриваемой модели имеет плотность или ряд распределения, которые мы, как и прежде, обозначаем одинаково: $f(t, \theta)$). Будем предполагать, что функция $f(t, \theta)$ удовлетворяет некоторым аналитическим условиям, которые будем называть **условиями регулярности** (условия (R)), и суть которых заключается в возможности менять порядок дифференцирования по θ и интегрирования по $\vec{x}$ функции правдоподобия $\Pi(\vec{x},\ \theta)$, соответствующей $f(t, \theta)$. Точная формулировка этих условий довольно сложна, приведем в качестве примера условие из [1], достаточное для (R): функция $\sqrt{f(t,\ \theta)}$ дифференцируема по $\theta \in \Theta$, и функция $\mathbf{i}(\theta)$, называемая информацией по Фишеру и определяемая равенством

$$\mathbf{i}(\theta) = \mathbf{E}\left(\frac{\partial \ln f(X_i, \theta)}{\partial \theta}\right)^2, \tag{1.16}$$

существует, строго положительна и непрерывна по θ для всех $\theta \in \Theta$.

Примером модели, для которой не выполнены условия регулярности, является модель $\vec{X} \Subset U_{[0;\ \theta]}$, $\theta > 0$.

Теорема (неравенство Рао-Крамера.)
Пусть $\vec{X} \Subset F(t,\theta)$, $\theta \in \Theta$, выполнены условия регулярности. Тогда для любой несмещенной оценки $\tilde{\theta}$ параметра θ выполняется неравенство:

$$\mathbf{Var}\,\tilde{\theta} \geq \frac{1}{n\mathbf{i}(\theta)}. \tag{1.17}$$

Если для некоторой несмещенной оценки $\check{\theta}$ окажется, что ее дисперсия совпадает с правой частью неравенства Рао-Крамера (говорят, что для этой оценки в неравенстве Рао-Крамера достигается равенство), то оценка $\check{\theta}$ является эффективной, так как для любой другой несмещенной оценки $\tilde{\theta}$ неравенство (1.17) продолжает выполняться и, следовательно, $\mathbf{Var}\,\tilde{\theta} \geq \mathbf{Var}\,\check{\theta}$ для всех $\theta \in \Theta$.

Пример Пусть $\vec{X} \Subset U_{[0;\,\theta]}$, $\theta > 0$. Сравнить с помощью среднеквадратического подхода оценки параметра θ : $\theta_1^* = 2\overline{X}$ и $\hat{\theta} = X_{(n)}$.

Решение. Проверим сначала свойство несмещенности обеих оценок. Вычисляем математические ожидания:

$$\mathbf{E}\theta_1^* = \mathbf{E}(2\overline{X}) = 2\mathbf{E}\overline{X} = 2\frac{\theta}{2} = \theta, \quad \mathbf{E}\hat{\theta} = \mathbf{E}X_{(n)} = \frac{n}{n+1}\theta.$$

Видим, что из двух оценок $\theta_1^* = 2\overline{X}$ является несмещенной, а $\hat{\theta} = X_{(n)}$ — смещенной. Чтобы выяснить, какая из оценок лучше, вычислим для каждой квадратичную характеристику. Для несмещенной оценки она совпадает с дисперсией:

$$\mathbf{E}(\theta_1^* - \theta)^2 = \mathbf{Var}\,\theta_1^* = \mathbf{Var}\,(2\overline{X}) = 4\mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = 4\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}X_i =$$

$$= 4\frac{1}{n^2}n\mathbf{Var}X_1 = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}. \tag{1.18}$$

При вычислении квадратичной характеристики оценки $\hat{\theta}_n = X_{(n)}$ мы будем использовать плотность ее распределения.

$$\mathbf{E}(X_{(n)} - \theta)^2 = \mathbf{E}X_{(n)}^2 - 2\theta\mathbf{E}X_{(n)} + \theta^2 = \int_0^\theta y^2\frac{ny^{n-1}}{\theta^n}dy - 2\theta\frac{n\theta}{n+1} + \theta^2 =$$

$$= \frac{n}{n+2}\theta^2 - \frac{2n}{n+1}\theta^2 + \theta^2 = \frac{2\theta^2}{(n+1)(n+2)}. \tag{1.19}$$

Сравнивая квадратичные характеристики, вычисленные в (1.18) и (1.19), видим, что

$$\frac{\theta^2}{3n} \geq \frac{2\theta^2}{(n+1)(n+2)}$$

для всех $\theta > 0$ и для всех $n \geq 1$.

Следовательно, ОМП $\hat{\theta} = X_{(n)}$ лучше в среднеквадратичном, чем ОММ $\theta_1^* = 2\overline{X}$.

Пример Исследовать с помощью неравенства Рао-Крамера оптимальность оценки $\overline{X}$ в моделях:

а) $\vec{X} \sim E_{\frac{1}{\theta}}, \quad \theta > 0;$

б) $\vec{X} \sim \Pi_\lambda, \quad \lambda > 0.$

Решение. а) Вычислим дисперсию оценки, применяя свойства дисперсии и учитывая, что $\mathbf{Var}\, X_i = \theta^2$:

$$\mathbf{Var}\, \theta^* = \mathbf{Var}\, \overline{X} = \mathbf{Var} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{Var}\, X_i = \frac{1}{n^2} n \mathbf{Var}\, X_1 = \frac{\theta^2}{n}. \tag{1.20}$$

Найдем правую часть неравенства Рао-Крамера для рассматриваемой модели, для этого вычислим последовательно:

$$f(X_i, \theta) = \frac{1}{\theta} e^{-\frac{X_i}{\theta}}; \quad \ln f(X_i, \theta) = -\ln \theta - \frac{X_i}{\theta}; \quad \frac{\partial \ln f(X_i, \theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{X_i}{\theta^2};$$

$$\mathbf{i}(\theta) = \mathbf{E} \left(\frac{\partial \ln f(X_i, \theta)}{\partial \theta} \right)^2 = \mathbf{E} \left(-\frac{1}{\theta} + \frac{X_i}{\theta^2} \right)^2 = \mathbf{E} \left(\frac{X_i - \theta}{\theta^2} \right)^2 =$$

$$= \frac{1}{\theta^4} \mathbf{E} (X_i - \theta)^2 = \frac{\mathbf{Var}\, X_i}{\theta^4} = \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2};$$

$$\frac{1}{n \mathbf{i}(\theta)} = \frac{\theta^2}{n}. \tag{1.21}$$

Сравнивая (1.20) и (1.21), видим, что для оценки $\theta^* = \overline{X}$ в неравенстве Рао-Крамера достигается равенство, следовательно, она эффективна.

б) Прежде всего вспомним, что для распределения Пуассона Π_λ математическое ожидание и дисперсия равны $\mathbf{E} X_i = \lambda, \quad \mathbf{Var}\, X_i = \lambda$. Тогда дисперсия нашей оценки равна:

$$\mathbf{Var}\, \lambda^* = \mathbf{Var}\, \overline{X} = \mathbf{Var} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{Var}\, X_i = \frac{1}{n^2} n \mathbf{Var}\, X_1 = \frac{\lambda}{n}. \tag{1.22}$$

Аналогично пункту а), вычисляем правую часть неравенства Рао-Крамера:

$$f(X_i, \lambda) = e^{-\lambda} \frac{\lambda^{X_i}}{X_i!}; \quad \ln f(X_i, \lambda) = -\lambda + X_i \ln \lambda - \ln(X_i!); \quad \frac{\partial \ln f(X_i, \lambda)}{\partial \lambda} = -1 + \frac{X_i}{\lambda};$$

$$\mathbf{i}(\lambda) = \mathbf{E} \left(\frac{\partial \ln f(X_i, \lambda)}{\partial \lambda} \right)^2 = \mathbf{E} \left(\frac{X_i}{\lambda} - 1 \right)^2 = \mathbf{E} \left(\frac{X_i - \lambda}{\lambda} \right)^2 =$$

$$= \frac{1}{\lambda^2}\mathbf{E}(X_i - \lambda)^2 = \frac{1}{\lambda^2}\mathbf{Var}\, X_i = \frac{1}{\lambda};$$

$$\frac{1}{n\mathbf{i}(\lambda)} = \frac{\lambda}{n}. \tag{1.23}$$

Сравнивая (1.22) и (1.23), видим, что для оценки $\lambda^* = \overline{X}$ в неравенстве Рао-Крамера достигается равенство, следовательно, она эффективна.

Пример Найти оценки параметра p по первому и второму моментам и методом максимального правдоподобия. Проверить их несмещенность и состоятельность. Для несмещенных оценок проверить их эффективность. Вычислить реализации всех оценок по реализации выборки (1, 1, -1, 0, 0).

k	-1	0	1
$\mathbf{P}(X_1 = k)$	p	$1/2 + p$	$1/2 - 2p$

Решение. Вычислим математическое ожидание X_1:

$$\mathbf{E}X_1 = -1 \cdot p + 0 \cdot (1/2 + p) + 1 \cdot (1/2 - 2p) = 1/2 - 3p.$$

Заменяя первый момент $\mathbf{E}X_1$ на первый выборочный момент $\overline{X}$, а параметр p на оценку по первому моменту p_1^*, получаем уравнение для оценки по первому моменту

$$\overline{X} = 1/2 - 3p_1^*.$$

Отсюда оценка по первому моменту $p_1^* = 1/6 - \overline{X}/3$.

Так как

$$\mathbf{E}p_1^* = 1/6 - \mathbf{E}\overline{X}/3 = 1/6 - \mathbf{E}X_1/3 = 1/6 - (1/2 - 3p)/3 = p,$$

то p_1^* является несмещенной оценкой параметра p.

Так как $\overline{X} \to \mathbf{E}X_1$ с вероятностью 1 согласно закону больших чисел Колмогорова, то

$$p_1^* \to 1/6 - \mathbf{E}X_1/3 = 1/6 - (1/2 - 3p)/3 = p$$

с вероятностью 1, и p_1^* является состоятельной оценкой параметра p.

Вычислим второй момент случайной величины X_1:

$$\mathbf{E}X_1^2 = (-1)^2 \cdot p + 0^2 \cdot (1/2 + p) + 1^2 \cdot (1/2 - 2p) = 1/2 - p.$$

Заменяя второй момент $\mathbf{E}X_1^2$ на второй выборочный момент $\overline{X^2}$, а параметр p на оценку по второму моменту p_2^*, получаем уравнение для оценки по второму моменту

$$\overline{X^2} = 1/2 - p_2^*.$$

Отсюда оценка по второму моменту $p_2^* = 1/2 - \overline{X^2}$.

Так как

$$\mathbf{E}p_2^* = 1/2 - \mathbf{E}\overline{X^2} = 1/2 - \mathbf{E}X_1^2 = 1/2 - (1/2 - p) = p,$$

то p_2^* является несмещенной оценкой параметра p.

Так как $\overline{X^2} \to \mathbf{E}X_1^2$ с вероятностью 1 согласно закону больших чисел Колмогорова, то

$$p_2^* \to 1/2 - \mathbf{E}X_1^2 = 1/2 - (1/2 - p) = p$$

с вероятностью 1, и p_2^* является состоятельной оценкой параметра p.

Для вычисления оценки максимального правдоподобия запишем параметрическое семейство плотностей

$$f(t,p) = \begin{cases} p, & \text{если } t = -1, \\ 1/2 + p, & \text{если } t = 0, \\ 1/2 - 2p, & \text{если } t = 1. \end{cases}$$

Это регулярное параметрическое семейство плотностей. Его удобно переписать в виде

$$f(t,p) = p^{\mathbf{I}\{t=-1\}}(1/2+p)^{\mathbf{I}\{t=0\}}(1/2-2p)^{\mathbf{I}\{t=1\}},$$

где $\mathbf{I}\{t = k\}$ — это индикатор того, что $t = k$.

Согласно свойствам логарифма,

$$\ln f(t,p) = \mathbf{I}\{t=-1\}\ln p + \mathbf{I}\{t=0\}\ln(1/2+p) + \mathbf{I}\{t=1\}\ln(1/2-2p).$$

Вычислим производную по параметру p:

$$\frac{\partial}{\partial p}\ln f(t,p) = \mathbf{I}\{t=-1\}\frac{1}{p} + \mathbf{I}\{t=0\}\frac{1}{1/2+p} - 2\mathbf{I}\{t=1\}\frac{1}{1/2-2p}.$$

Теперь мы должны записать уравнение

$$\sum_{i=1}^{n}\frac{\partial}{\partial p}\ln f(X_i,p) = 0.$$

Для этого введем обозначение $N_k = \sum_{i=1}\mathbf{I}\{X_i = k\}$. Это число элементов выборки, принявших значение k.

Подставляя X_i вместо t и суммируя по всем X_i, получаем уравнение

$$\sum_{i=1}^{n}\frac{\partial}{\partial p}\ln f(X_i,p) = \frac{N_{-1}}{p} + \frac{N_0}{1/2+p} - \frac{2N_1}{1/2-2p} = 0.$$

Приводим к общему знаменателю и приравниваем числитель к нулю:

$$N_{-1}(1/2+p)(1/2-2p) + N_0 p(1/2-2p) - 2N_1 p(1/2+p) = 0.$$

Раскрывая скобки, получаем:

$$N_{-1}/4 - pN_{-1}/2 - 2p^2N_{-1} + pN_0/2 - 2p^2N_0 - pN_1 - 2p^2N_1 = 0.$$

Так как $N_{-1} + N_0 + N_1 = n$, то, умножая на –4 и приводя подобные, получаем уравнение

$$8np^2 + 2(N_{-1} - N_0 + 2N_1)p - N_{-1} = 0.$$

Так как по условию $p \geq 0$, то выбираем решение со знаком плюс перед дискриминантом и получаем оценку максимального правдоподобия

$$\widehat{p} = \frac{-(N_{-1} - N_0 + 2N_1) + \sqrt{(N_{-1} - N_0 + 2N_1)^2 + 8nN_{-1}}}{8n}.$$

Подставим $N_{-1} = n - N_0 - N_1$:

$$\widehat{p} = \frac{-(n - 2N_0 + N_1) + \sqrt{(n - 2N_0 + N_1)^2 + 8n(n - N_0 - N_1)}}{8n}.$$

Так как по закону больших чисел Колмогорова

$$N_0/n \to \mathbf{P}(X_1 = 0) = 1/2 + p, \quad N_1/n \to \mathbf{P}(X_1 = 1) = 1/2 - 2p$$

с вероятностью 1, то (также с вероятностью 1)

$$\widehat{p} \to \frac{-(1 - 2(1/2 + p) + 1/2 - 2p) + \sqrt{(1/2 - 4p)^2 + 8(1 - (1/2 + p) - (1/2 - 2p))}}{8}$$

$$= \frac{-(1/2 - 4p) + \sqrt{(1/2 - 4p)^2 + 8p}}{8} = \frac{-(1/2 - 4p) + \sqrt{(1/2 + 4p)^2}}{8} = p,$$

то есть оценка максимального правдоподобия $\widehat{p}$ является состоятельной оценкой параметра p.

Но для невырожденной положительной случайной величины Y выполнено неравенство $\mathbf{E}\sqrt{Y} < \sqrt{\mathbf{E}Y}$, и поэтому оценка $\widehat{p}$ не является несмещенной оценкой параметра p.

Найдем информацию Фишера для параметрического семейства плотностей $f(t, p)$:

$$\mathbf{i}(p) = \mathbf{E}\left(\frac{\partial}{\partial p}\ln f(X_1, p)\right)^2$$

$$= \mathbf{E}\left(\mathbf{I}\{X_1 = -1\}\frac{1}{p} + \mathbf{I}\{X_1 = 0\}\frac{1}{1/2 + p} - 2\mathbf{I}\{X_1 = 1\}\frac{1}{1/2 - 2p}\right)^2$$

$$= \frac{p}{p^2} + \frac{1/2 + p}{(1/2 + p)^2} + \frac{4(1/2 - 2p)}{(1/2 - 2p)^2} = \frac{1}{p} + \frac{1}{1/2 + p} + \frac{4}{1/2 - 2p}.$$

Оценка по первому моменту $p_1^* = 1/6 - \overline{X}/3$ является несмещенной, ее дисперсия

$$\mathbf{Var}\, p_1^* = \mathbf{Var}(1/6 - \overline{X}/3) = \frac{1}{9}\mathbf{Var}\overline{X} = \frac{1}{9n}\mathbf{Var}X_1$$

$$= \frac{1}{9n}\left(\mathbf{E}X_1^2 - (\mathbf{E}X_1)^2\right) = \frac{1}{9n}\left(1/2 - p - (1/2 - 3p)^2\right).$$

Для этой оценки не выполняется равенство в неравенстве Рао-Крамера:

$$\mathbf{Var}\, p_1^* \neq \frac{1}{n\mathbf{i}(p)}.$$

Действительно, если устремить p к нулю, то информация Фишера устремится к бесконечности, то есть правая часть неравенства устремится к нулю, а предел левой части неравенства при $p \to 0$ равен $\frac{1}{36n} > 0$.

Следовательно, p_1^* не является эффективной оценкой.

Оценка по второму моменту $p_2^* = 1/2 - \overline{X^2}$ является несмещенной, ее дисперсия

$$\mathbf{Var}\, p_2^* = \mathbf{Var}(1/2 - \overline{X^2}) = \mathbf{Var}\overline{X^2} = \frac{1}{n}\mathbf{Var}X_1^2$$

$$= \frac{1}{n}\left(\mathbf{E}X_1^4 - (\mathbf{E}X_1^2)^2\right) = \frac{1}{n}\left((-1)^4 \cdot p + 0^4 \cdot (1/2 + p) + 1^4 \cdot (1/2 - 2p) - (1/2 - p)^2\right)$$

$$= \frac{1}{n}\left(1/2 - p - (1/2 - p)^2\right)$$

Для этой оценки не выполняется равенство в неравенстве Рао-Крамера:

$$\mathbf{Var}\, p_2^* \neq \frac{1}{n\mathbf{i}(p)}.$$

Действительно, если устремить p к нулю, то информация Фишера устремится к бесконечности, то есть правая часть неравенства устремится к нулю, а предел левой части неравенства при $p \to 0$ равен $\frac{1}{4n} > 0$.

Следовательно, p_2^* не является эффективной оценкой.

Оценка максимального правдоподобия $\widehat{p}$ не является несмещенной и потому не может быть эффективной.

Вычислим реализации оценок по реализации выборки (1, 1, -1, 0, 0):

$$p_1^* = 1/6 - \overline{X}/3 = 1/6 - 1/15 = 0{,}1,$$

$$p_2^* = 1/2 - \overline{X^2} = 1/2 - 3/5 = -0{,}1.$$

Отметим, что реализация оценки по второму моменту выходит за рамки допустимых значений параметра p.

Для вычисления реализации оценки максимального правдоподобия найдем реализации $N_{-1}=1$, $N_0=2$, $N_1=2$, $n=5$ и подставим в формулу:

$$\widehat{p}=\frac{-(n-2N_0+N_1)+\sqrt{(n-2N_0+N_1)^2+8n(n-N_0-N_1)}}{8n}$$

$$=\frac{-(5-4+2)+\sqrt{(5-4+2)^2+40(5-2-2)}}{40}=\frac{\sqrt{49}-3}{40}=0{,}1.$$

Задачи для самостоятельного решения

5.1 Дана выборка $\vec{X}\sim \mathrm{U}_{[0,\theta]}$, $\theta>0$ — неизвестный параметр. Сравнить, какая из оценок для параметра θ лучше в среднеквадратичном: $\theta_1^*=2\overline{X}$, $\theta_2^*=\frac{n+1}{n}X_{(n)}$.

5.2 Пусть $\vec{X}\Subset F(t,\theta)$, где $\theta=\mathbf{E}_\theta X_1$, $\mathbf{Var}X_1<\infty$. Показать, что оценка $\theta_1^*=\overline{X}$ является наилучшей в среднеквадратичном среди всех несмещенных оценок вида

$$\theta^*=C_1X_1+C_2X_2+\cdots+C_nX_n,\quad C_1+C_2+\cdots+C_n=1.$$

5.3 Исследовать с помощью неравенства Рао-Крамера оптимальность ОМП для неизвестного параметра в моделях

а) $\vec{X}\sim \mathrm{B}_p$, $0<p<1$;

б) $\vec{X}\sim \mathrm{B}_{m,p}$, $0<p<1$, m — известно.

в) $\vec{X}\sim \mathrm{N}_{\theta,1}$, $-\infty<\theta<\infty$.

г) $\vec{X}\sim \mathrm{N}_{0,\theta}$, $0<\theta<\infty$.

д) $\vec{X}\sim \mathrm{G}_{1/\theta}$, $\theta>1$.

Индивидуальные домашние задания

Найти оценки параметра θ по первому и второму моментам и методом максимального правдоподобия. Проверить их несмещенность и состоятельность. Для несмещенных оценок проверить их эффективность. Вычислить реализации всех оценок по реализации выборки (2, 1, 3, 3, 4, 2, 4, 3, 4, 4).

Вариант 1

k	1	2	3	4
$\mathbf{P}(X_1=k)$	θ	θ	θ	$1-3\theta$

Вариант 2

k	1	2	3	4
$\mathbf{P}(X_1=k)$	$1/2-\theta$	θ	θ	$1/2-\theta$

Вариант 3

k	1	2	3	4
$\mathbf{P}(X_1=k)$	3θ	$1/3-\theta$	$1/3-\theta$	$1/3-\theta$

Вариант 4

k	1	2	3	4
$\mathbf{P}(X_1=k)$	θ	θ	2θ	$1-4\theta$

Вариант 5

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	θ	$1/3 - \theta$	$2/3 - \theta$

Вариант 6

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	θ	$1 - 3\theta$	θ

Вариант 7

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/2 - \theta$	θ	$1/2 - \theta$	θ

Вариант 8

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/3 - \theta$	3θ	$1/3 - \theta$	$1/3 - \theta$

Вариант 9

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	θ	$1 - 4\theta$	2θ

Вариант 10

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	$1/3 - \theta$	θ	$2/3 - \theta$

Вариант 11

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	$1 - 3\theta$	θ	θ

Вариант 12

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	$1/2 - \theta$	θ	$1/2 - \theta$

Вариант 13

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/3 - \theta$	$1/3 - \theta$	3θ	$1/3 - \theta$

Вариант 14

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1 - 4\theta$	θ	θ	2θ

Вариант 15

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/3 - \theta$	θ	θ	$2/3 - \theta$

Вариант 16

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1 - 3\theta$	θ	θ	θ

Вариант 17

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	θ	$1/2 - \theta$	$1/2 - \theta$	θ

Вариант 18

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/3 - \theta$	$1/3 - \theta$	$1/3 - \theta$	3θ

Вариант 19

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1 - 4\theta$	θ	θ	2θ

Вариант 20

k	1	2	3	4
$\mathbf{P}(X_1 = k)$	$1/3 - \theta$	$2/3 - \theta$	θ	θ

Глава 2

Доверительные интервалы и статистические критерии

§6. Точные доверительные интервалы

Пусть имеется выборка объема n из распределения, известного с точностью до параметра: $\vec{X} \sim F(t, \theta)$, $\theta \in \Theta$. *Точным доверительным интервалом с уровнем доверия* $1 - \varepsilon$ для неизвестного параметра θ называют случайный интервал $(\theta_-;\ \theta_+) \subset \Theta$, построенный по выборке, который накрывает неизвестное значение параметра с вероятностью, равной $1 - \varepsilon$, то есть

$$\mathbf{P}\{\theta \in (\theta_-;\ \theta_+)\} = 1 - \varepsilon. \tag{2.1}$$

θ_-, θ_+ — это оценки параметра θ, называемые *нижней и верхней доверительными границами*. Число $1 - \varepsilon \in (0;\ 1)$ — уровень доверия, или доверительная вероятность, — выбирается заранее и отражает «степень готовности мириться с возможностью ошибки»: чем менее мы готовы мириться с возможной ошибкой, тем большее (более близкое к единице) значение $1 - \varepsilon$ должны устанавливать.

При построении доверительных интервалов для параметров нормального распределения мы будем использовать два специальных распределения, связанных с нормальным: распределение хи-квадрат и распределение Стьюдента. Название «распределение Стьюдента» связано с именем английского статистика К. Госсета, который подписывал свои работы псевдонимом «Стьюдент».

Случайная величина Z_n имеет *распределение хи-квадрат с n степенями свободы*, если

$$Z_n = {X_1}^2 + \ldots + {X_n}^2,$$

где $X_1,\ \ldots,\ X_n$ — независимые случайные величины со стандартным нормальным распределением. Отметим, что «число степеней свободы» — это

просто традиционное название для параметра n распределения хи-квадрат. Параметр n — положительное целое число. В частности, при $n = 1$ получаем квадрат одной случайной величины со стандартным нормальным распределением: $Z_1 = X^2$, где $X \sim N_{0,\,1}$.

Будем использовать следующее обозначение: $Z_n \sim \chi^2_n$.

Отметим следующие свойства распределения хи-квадрат.

Следствие 2.1 *Пусть $Z_n \sim \chi^2_n$. Тогда*
1) $\mathbf{E}Z_n = n$;
2) $Z_n/n \to 1$ с вероятностью 1 при $n \to \infty$.

Доказательство.
Во-первых,

$$\mathbf{E}Z_1 = \mathbf{E}X^2 = \mathbf{Var}\, X + (\mathbf{E}X)^2,$$

где X имеет стандартное нормальное распределение, и потому $\mathbf{E}X = 0$, $\mathbf{Var}\, X = 1$. Следовательно,

$$\mathbf{E}Z_1 = 1 + 0^2 = 1.$$

1) По определению распределения хи-квадрат,

$$\mathbf{E}Z_n = \mathbf{E}({X_1}^2 + \ldots + {X_n}^2) = \mathbf{E}{X_1}^2 + \ldots + \mathbf{E}{X_n}^2 = n\mathbf{E}{X_1}^2 = n \cdot 1 = n.$$

2) Так как Z_n — сумма независимых одинаково распределенных случайных величин, то справедлив закон больших чисел Колмогорова:

$$Z_n/n = \frac{{X_1}^2 + \ldots + {X_n}^2}{n} \to \mathbf{E}{X_1}^2 = 1$$

почти наверное при $n \to \infty$. Доказательство завершено.

Случайная величина Y_n имеет *распределение Стьюдента с n степенями свободы*, если

$$Y_n = \frac{X}{\sqrt{Z_n/n}},$$

где случайные величины X и Z_n независимы, причем X имеет стандартное нормальное распределение, а Z_n имеет распределение хи-квадрат с n степенями свободы. Здесь, как и у распределения хи-квадрат, n — это просто положительный целый параметр.

Будем использовать следующее обозначение: $Y_n \sim T_n$.

Отметим следующие свойства распределения Стьюдента.

Следствие 2.2 *Пусть $Y_n \sim T_n$. Тогда*
1) для любого t выполнено $\mathbf{P}\{Y_n < -t\} = \mathbf{P}\{Y_n > t\}$, то есть распределение Стьюдента симметрично;

2) $Y_n \to X$ почти наверное при $n \to \infty$, где X имеет стандартное нормальное распределение.

Доказательство.

1) Симметрия следует из симметрии стандартного нормального распределения:

$$\mathbf{P}\{Y_n < -t\} = \mathbf{P}\{X < -t\sqrt{Z_n/n}\} = \mathbf{P}\{X > t\sqrt{Z_n/n}\} = \mathbf{P}\{Y_n > t\}.$$

2) Сходимость следует из свойства сходимости почти наверное, непрерывности функции $x/\sqrt{y}$ и из свойства распределения хи-квадрат. Доказательство завершено.

Наиболее распространенной ситуацией, когда возможно построение точных доверительных интервалов, является случай нормального распределения: $\vec{X} \sim \mathrm{N}_{a,\sigma^2}$, — когда хотя бы один из его параметров неизвестен. В этом случае известно совместное распределение наиболее употребительных оценок $\overline{X}$ и S^2 параметров a и σ^2, с помощью которого и строятся соответствующие доверительные интервалы. Основные результаты содержатся в следующей теореме, которую примем без доказательства.

Теорема 2.1 (Теорема Фишера) *Пусть $\vec{X} \sim \mathrm{N}_{a,\sigma^2}$. Тогда верны следующие 4 факта:*

1) $\dfrac{\sqrt{n}(\overline{X} - a)}{\sigma} \sim N_{0,1}$.

2) $\dfrac{\sum_{i=1}^{n}(X_i - a)^2}{\sigma^2} \sim \chi^2_n$.

3) $\dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$, *причем* $\overline{X}$ *и* S^2 *независимы.*

4) $\dfrac{\sqrt{n-1}\,\big(\overline{X} - a\big)}{S} \sim T_{n-1}$.

Отметим, что первое утверждение теоремы следует сразу же из свойств нормального распределения, второе — из определения распределения хи-квадрат с учетом того факта, что $\frac{X_i - a}{\sigma}$ имеет стандартное нормальное распределение. Третье утверждение — нетривиальный факт, прокомментировать который можно следующим образом. Вспомним, что

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

Если выборка состоит из одного элемента X_1, то есть $n = 1$, то

$$\frac{nS^2}{\sigma^2} = \frac{(X_1 - X_1)^2}{\sigma^2} = 0,$$

— случайная величина имеет распределение, вырожденное в точке ноль, которое можно по определению считать распределением хи-квадрат с нулевым числом степеней свободы.

Если выборка состоит из двух элементов $(X_1,\ X_2)$, то есть $n = 2$, то

$$\frac{nS^2}{\sigma^2} = \frac{(X_1 - (X_1 - X_2)/2)^2 + (X_2 - (X_1 - X_2)/2)^2}{\sigma^2} = \frac{(X_1 - X_2)^2}{2\sigma^2}.$$

Отметим, что $\mathbf{E}(X_1 - X_2) = a - a = 0$. Так как X_1 и X_2 независимы, то $\mathbf{Var}(X_1 - X_2) = \mathbf{Var}\, X_1 + \mathbf{Var}\, X_2 = 2\sigma^2$. Согласно свойствам нормального распределения, $X_1 - X_2 \sim \mathrm{N}_{0,2\sigma^2}$, то есть $X_1 - X_2 = \sqrt{2}\sigma X$, где X имеет стандартное нормальное распределение. Поэтому

$$\frac{nS^2}{\sigma^2} = \frac{2\sigma^2 X^2}{2\sigma^2} = X^2 \sim \chi^2_1.$$

Доказательство для $n \geq 2$ оказывается существенно более сложным, и приводить его здесь мы не будем.

Четвертое утверждение следует из третьего и определения распределения Стьюдента.

Пример 2.1 *Пусть* $\vec{X} \Subset \mathrm{N}_{a,\sigma^2}$, $a \in R$. *Построить доверительный интервал* $(a_-;\ a_+)$ *для параметра* a, *считая* σ^2 *известным. Вычислить реализацию доверительного интервала с уровнем доверия* $\gamma = 0,95$, *располагая данными:* $n = 10,\ \overline{X} = 2,7,\ \sigma^2 = 4$.

Решение. Для построения доверительного интервала используем оценку $\overline{X}$, распределение которой известно. Для заданной доверительной вероятности γ найдем такое $A > 0$, что

$$\gamma = \mathbf{P}\left\{\left|\sqrt{n}\frac{\overline{X} - a}{\sigma}\right| < A\right\} = \mathbf{P}\left\{-\frac{\sigma A}{\sqrt{n}} < \overline{X} - a < \frac{\sigma A}{\sqrt{n}}\right\}. \qquad (2.2)$$

Таким образом, нужно искать $\varepsilon_1 = -\frac{\sigma A}{\sqrt{n}},\ \varepsilon_2 = \frac{\sigma A}{\sqrt{n}}$ такие, что выполняется равенство:

$$\mathbf{P}\left\{\varepsilon_1 < \overline{X} - a < \varepsilon_2\right\} = \gamma.$$

Для этого вернемся к (2.2). В силу теоремы 2.1, случайная величина, стоящая под знаком модуля, имеет стандартное нормальное распределение, поэтому вероятность в правой части можно выразить через функцию распределения $\Phi(t)$ закона $N_{0,1}$, и тогда уравнение (2.2) приобретает вид:

$$2\Phi(A) - 1 = \gamma \iff \Phi(A) = \frac{1+\gamma}{2}, \qquad (2.3)$$

где $\Phi(t)$ — функция Лапласа, значения которой представлены в таблице приложения в конце книги. Заметим, что значение $A = A_{\frac{1+\gamma}{2}}$, удовлетворяющее (2.3), представляет квантиль уровня $\frac{1+\gamma}{2}$ распределения $N_{0,1}$. Найдя его по таблице и подставив в (2.2), получим равенство:

$$\gamma = \mathbf{P}\left\{\left|\sqrt{n}\frac{\overline{X}-a}{\sigma}\right| < A_{\frac{1+\gamma}{2}}\right\} = \mathbf{P}\left\{-A_{\frac{1+\gamma}{2}} < \sqrt{n}\frac{a-\overline{X}}{\sigma} < A_{\frac{1+\gamma}{2}}\right\} \iff$$

$$\iff \gamma = \mathbf{P}\left\{\overline{X} - \sigma\frac{A_{\frac{1+\gamma}{2}}}{\sqrt{n}} < a < \overline{X} + \sigma\frac{A_{\frac{1+\gamma}{2}}}{\sqrt{n}}\right\}, \tag{2.4}$$

откуда искомый γ-доверительный интервал:

$$(a_-;\ a_+) = \left(\overline{X} - \sigma\frac{A_{\frac{1+\gamma}{2}}}{\sqrt{n}},\ \overline{X} + \sigma\frac{A_{\frac{1+\gamma}{2}}}{\sqrt{n}}\right).$$

Подставляя сюда конкретные данные из условия, вычисляем реализацию доверительного интервала:

$$(a_-;\ a_+) \approx \left(2,7 - 2\frac{1,96}{\sqrt{10}},\ 2,7 + 2\frac{1,96}{\sqrt{10}}\right) \approx (1,46; 3,94) \iff$$

$$\iff \mathbf{P}(1,46 < \theta < 3,94) = 0,95.$$

Замечание 2.1 *Построенный доверительный интервал оказывается симметричным относительно выборочного среднего* $\overline{X}$ *и имеет длину* $2A\frac{\sigma}{\sqrt{n}}$, *пропорциональную значению* A, *которое было найдено из условия* (2.2).

Пример 2.2 *Пусть* $\vec{X} \in N_{a,\sigma^2}$, *где* $a \in \mathbf{R}$, $\sigma^2 > 0$ — *два неизвестных параметра. Построить доверительный интервал для параметра* σ^2. *Вычислить реализацию доверительного интервала с уровнем доверия* $\gamma = 0,9$, *располагая данными:* $n = 10$, $S^2 = 4$.

Решение. Используя лемму Фишера, проще всего построить так называемый односторонний доверительный интервал. Для этого по заданной доверительной вероятности $\gamma = 0,9$ найдем такое $B > 0$, что

$$\gamma = \mathbf{P}\left\{\frac{nS^2}{\sigma^2} > B\right\} \iff \mathbf{P}\left\{\frac{nS^2}{\sigma^2} \le B\right\} = 1 - \gamma. \tag{2.5}$$

Учитывая, что случайная величина $\frac{nS^2}{\sigma^2}$ имеет распределение χ^2_{n-1}, нетрудно видеть, что искомое B есть не что иное, как квантиль $\chi^2_{1-\gamma,n-1}$ этого

распределения. Подставляя ее в (2.5) и разрешая неравенство под знаком вероятности относительно σ^2, находим доверительный интервал:

$$\gamma = \mathbf{P}\left\{\frac{nS^2}{\sigma^2} > \chi^2_{1-\gamma,n-1}\right\} \iff \gamma = \mathbf{P}\left\{\frac{nS^2}{\chi^2_{1-\gamma}, n-1} > \sigma^2\right\} \iff$$

$$\iff (\sigma^2_-;\ \sigma^2_+) = \left(0, \frac{nS^2}{\chi^2_{1-\gamma,n-1}}\right).$$

Подставляя конкретные данные, находим реализацию доверительного интервала:

$$\chi^2_{0.1,9} = 4,17; \quad (\sigma^2_-;\ \sigma^2_+) = (0;\ 8,63).$$

Чтобы построить двусторонний доверительный интервал, вместо (2.5) используем следующее уравнение:

$$\gamma = \mathbf{P}\left\{x_1 < \frac{nS^2}{\sigma^2} < x_2\right\} \iff \mathbf{P}_\theta\left\{\frac{nS^2}{x_2} < \sigma^2 < \frac{nS^2}{x_1}\right\} = \gamma, \qquad (2.6)$$

где $0 < x_1 < x_2$, удовлетворяющие (2.6), находим по известному распределению χ^2_{n-1} случайной величины $\frac{nS^2}{\sigma^2}$. В общем случае эта задача не имеет единственного решения. Если обратиться к графику плотности распределения χ^2_{n-1}, представленному на рис. 2.1, то $x_1 < x_2$ следует выбирать таким образом, чтобы сумма вероятностей, представленных площадями заштрихованных областей под графиком плотности, равнялась $1-\gamma$. Ясно, что это можно сделать многими способами.

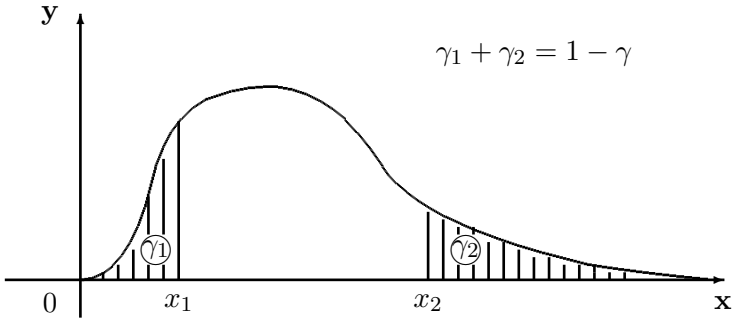

Рис. 2.1. Плотность распределения χ^2_{n-1}

Чтобы сделать решение однозначным, выберем $x_1 < x_2$ так, чтобы каждая из заштрихованных площадей равнялась $\frac{1-\gamma}{2}$, тогда нетрудно видеть, что x_1, x_2 выражаются через квантили распределения χ^2_{n-1}:

$$x_1 = \chi^2_{\frac{1-\gamma}{2},n-1}, \quad x_2 = \chi^2_{\frac{1+\gamma}{2},n-1}.$$

Подставляя это в (2.6), находим искомый двусторонний доверительный интервал:

$$\mathbf{P}_\theta\left\{\frac{nS^2}{\chi^2_{\frac{1+\gamma}{2},n-1}}<\sigma^2<\frac{nS^2}{\chi^2_{\frac{1-\gamma}{2},n-1}}\right\}=\gamma \iff (\sigma_-^2;\ \sigma_+^2)=\left(\frac{nS^2}{\chi^2_{\frac{1+\gamma}{2},n-1}},\ \frac{nS^2}{\chi^2_{\frac{1-\gamma}{2},n-1}}\right).$$

Находя по таблице распределения хи-квадрат $\chi^2_{\frac{1-\gamma}{2},n-1}$, $\chi^2_{\frac{1+\gamma}{2},n-1}$ — квантили распределения χ^2_{n-1} для конкретных значений $\gamma=0,9$, $n=10$, и данных в условии численных значений, находим реализацию доверительного интервала:

$$\chi^2_{\frac{1+0,9}{2},9}=16,9;\ \chi^2_{\frac{1-0,9}{2},9}\approx 3,325;$$

$$(\sigma_-^2,\ \sigma_+^2)=(2,37;\ 12,03).$$

Задачи для самостоятельного решения

6.1 Пусть $\vec{X}\sim N(\theta_1,\theta_2)$, $\theta_1\in R$, $\theta_2>0$. Построить центральный доверительный интервал для параметра θ_1. Вычислить реализацию доверительного интервала с уровнем $\gamma=0,95$, располагая данными: n=10, $\overline{X}=2,7;\ S^2=4$. Сравнить с результатом примера 2.1.

6.2 Пусть $\vec{X}\sim N(a,\theta)$, $\theta>0$, a- известно. Построить точные доверительные интервалы (односторонний и центральный двухсторонний) на основе статистики $S_1^2=\frac{1}{n}\sum_{i=1}^n(X_i-a)^2$. Вычислить реализации построенных интервалов с уровнем $\gamma=0,9$, располагая данными: n=10, $S_1^2=4$. Сравнить с результатами примера 2.2.

6.3 Пусть $\vec{X}$ имеет плотность распределения $\frac{1}{\pi(1+(t-\theta)^2)}$, $\theta\in R$. Построить оптимальный точный доверительный интервал для параметра θ по одному наблюдению (n=1).

Индивидуальные домашние задания

Взять 10 толстых книг разных авторов. Записать авторов и названия книг. Измерить линейкой толщину каждой книги (без обложки) в миллиметрах. Найти число листов (число страниц, деленное на 2). Вычислить толщину листа, поделив толщину книги на число листов. Предполагая, что толщина листа имеет нормальное распределение, найти точечные оценки математического ожидания и стандартного отклонения толщины листа. Найти доверительные интервалы для них уровня доверия 0,95 и уровня доверия 0,9.

§7. Асимптотические доверительные интервалы

Пусть имеется выборка объема n из распределения, известного с точностью до параметра: $\vec{X} \sim F(t,\theta),\ \theta \in \Theta$. *Асимптотическим доверительным интервалом с уровнем доверия* $1-\varepsilon$ для неизвестного параметра θ называют случайный интервал $(\theta_-;\ \theta_+) \subset \Theta,$ построенный по выборке, который накрывает неизвестное значение параметра с вероятностью, стремящейся к γ с ростом объема выборки, то есть

$$\mathbf{P}\{\theta \in (\theta_-;\ \theta_+)\} \to 1-\varepsilon \tag{2.7}$$

при $n \to \infty$.

Если распределение не является нормальным, точный доверительный интервал, как правило, не удается построить. Поэтому строят асимптотический доверительный интервал, применяя центральную предельную теорему, которая утверждает, что для всех $t_1, t_2 \in \mathbf{R}\ (t_1 < t_2)$ выполнено

$$\lim_{n\to\infty} \mathbf{P}\left(t_1 \le \frac{n\overline{X} - na}{\sigma\sqrt{n}} < t_2\right) = \Phi(t_2) - \Phi(t_1), \quad n \to \infty$$

то есть центрированные и нормированные суммы случайных величин $n\overline{X} = X_1 + \ldots + X_n$ сходятся по распределению к случайной величине, имеющей стандартное нормальное распределение.

Здесь $a = \mathbf{E}X_1,\ \sigma^2 = \mathbf{Var}\, X_1$ — математическое ожидание и дисперсия элементов выборки.

Если выбрать $t_2 = -t_1 = A$ и принять доверительный уровень равным γ, то

$$\lim_{n\to\infty} \mathbf{P}\left(-A \le \frac{n\overline{X} - na}{\sigma\sqrt{n}} < A\right) = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = \gamma, \quad n \to \infty$$

откуда получаем

$$\Phi(A) = (\gamma+1)/2. \tag{2.8}$$

По заданному γ можно найти A с помощью таблиц нормального распределения или программных приложений. Отметим без доказательства следующее свойство сходимости по распределению: если Y_n сходится по распределению к Y, а Z_n сходится почти наверное к 1, то их произведение Y_nZ_n сходится по распределению к Y. Выберем

$$Y_n = \frac{n\overline{X} - na}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{X} - a)}{\sigma}, \qquad Z_n = \frac{\sigma}{S}.$$

Вспомним, что $S = \sqrt{\overline{X^2} - (\overline{X})^2} \to \sigma$ почти наверное, и по свойству сходимости почти наверное $Z_n \to 1$ п. н. Следовательно,

$$Y_n Z_n = \frac{\sqrt{n}(\overline{X} - a)}{\sigma} \cdot \frac{\sigma}{S} = \frac{\sqrt{n}(\overline{X} - a)}{S}$$

сходится по распределению к стандартной нормальной случайной величине, то есть

$$\lim_{n\to\infty} \mathbf{P}\left(-A \le \frac{\sqrt{n}(\overline{X} - a)}{S} < A\right) = \Phi(A) - \Phi(-A) = 2\Phi(A) - 1 = \gamma, \;\; n \to \infty, \tag{2.9}$$

где константа A выбирается по формуле 2.8.

Чтобы для неизвестного параметра θ найти доверительный интервал асимптотического уровня γ, нужно для исследуемого однопараметрического семейства распределений найти зависимость $\mathbf{E}X_1 = a = a(\theta)$ и подставить полученное выражение в 2.9, то есть решить относительно параметра θ двойное неравенство

$$-A \le \frac{\sqrt{n}(\overline{X} - a(\theta))}{S} < A \tag{2.10}$$

(для этого нужно, чтобы функция $a(\theta)$ была непрерывной и строго монотонной). Получившиеся границы дверительного интервала будем обозначать через θ_- и θ_+.

Пример 2.3 *Пусть* $\vec{X} \Subset E_\alpha$, $\alpha > 0$. *Построить асимптотический доверительный интервал для параметра* α.

Решение. Так как для показательного распределения с параметром α математическое ожидание равняется $\mathbf{E}X_1 = 1/\alpha$, то согласно формуле 2.10

$$-A \le \frac{\sqrt{n}(\overline{X} - 1/\alpha)}{S} < A.$$

Последовательно выразим

$$-\frac{AS}{\sqrt{n}} \le \overline{X} - \frac{1}{\alpha} < \frac{AS}{\sqrt{n}};$$

$$\overline{X} - \frac{AS}{\sqrt{n}} < \frac{1}{\alpha} \le \overline{X} + \frac{AS}{\sqrt{n}};$$

$$\frac{1}{\overline{X} - \dfrac{AS}{\sqrt{n}}} > \alpha \ge \frac{1}{\overline{X} + \dfrac{AS}{\sqrt{n}}}.$$

Итак, мы получили доверительный интервал $(\alpha_-;\ \alpha_+)$, где

$$\alpha_- = \frac{1}{\overline{X} + \dfrac{AS}{\sqrt{n}}}, \qquad \alpha_+ = \frac{1}{\overline{X} - \dfrac{AS}{\sqrt{n}}},$$

а константа A выбирается в соответствии с равенством (2.8).

Задачи для самостоятельного решения

7.1 $\vec{X} \sim B_p,\ 0 < p < 1$. Построить приближенные доверительные интервалы для параметра p на основе оценки $p^* = \overline{X}$.

7.2 $\vec{X} \sim {}_\lambda,\ \lambda > 0$. Построить асимптотический доверительный интервал для параметра λ с помощью оценки $\lambda^* = \overline{X}$.

7.3 Пусть $\vec{X} \sim U_{[0,\theta]}$, где $\theta > 0$. С помощью статистик $\overline{X}$ $\overline{X^2}$ построить асимптотические доверительные интервалы (соответственно (θ_1^-, θ_1^+) и (θ_2^-, θ_2^+)) уровня $1-\varepsilon$ и показать, что случайный интервал (θ_2^-, θ_2^+) асимптотически короче соответствующего (θ_1^-, θ_1^+).

Индивидуальные домашние задания

Возьмите два романа одного автора. Для каждого романа найдите доверительный интервал для вероятности появления слова «он» асимптотического уровня доверия 0,95. Пересекаются ли эти доверительные интервалы, построенные для разных романов? Сделайте то же для асимптотического уровня доверия 0,99.

§8. Статистические критерии

Пусть $\vec{X} = (X_1, X_2, ..., X_n)$ — выборка, $\vec{X} \sim \mathbf{F}$, где $\mathbf{F}$ - полностью или частично неизвестное распределение отдельного наблюдения X_i.

Определение 2.1 *Статистической гипотезой* будем называть всякое утверждение о виде или свойствах неизвестного распределения $\mathbf{F}$.

Пример 2.4 *Пусть* $\mathbf{F}$ *— полностью неизвестное распределение. Примерами гипотез являются*

H : $\mathbf{F} = \mathbf{F}_0$, *где* $\mathbf{F}_0$ *— полностью определенное распределение;*

H : $\mathbf{F} \in \widehat{\mathbf{F}}_0$, *где* $\widehat{\mathbf{F}}_0$ *— множество распределений (например,* $\widehat{\mathbf{F}}_0 = N_{a,\sigma^2}$ *или* $\widehat{\mathbf{F}}_0 = B_p$*).*

В этих примерах наблюдения имеют распределения из некоторого одно- или двухпараметрического семейства. Но могут быть непараметрические множества, например, $\mathbf{F}_0 \in \{\mathbf{F} : \ \mathbf{E}X_i > 0\}$ *— класс распределений с положительными математическими ожиданиями.*

Пример 2.5 *Пусть* $\mathbf{F}$ *— частично известное распределение. Например,* $\mathbf{F} = U_{[a,\,b]}$ *(наблюдения имеют равномерное распределение). В этом случае примеры гипотез:*

H : $a = 0,\ b = 1$ *(распределение равномерное на* $[0,\,1]$*);*

H : $a = 0$ *(распределение равномерное на* $[0,\,b]$*);*

H : $a < b - 1$ *(распределение равномерное на отрезке длины более 1).*

Гипотеза называется *простой*, если она однозначно определяет распределение $\mathbf{F}$, в противном случае гипотеза называется *сложной*. В приведенных выше примерах простыми являются гипотезы:

H : $\mathbf{F} = \mathbf{F}_0$ и H : $a = 0,\ b = 1$ (последняя в случае, когда известно, что распределение равномерное на $[a,\,b]$).

Остальные гипотезы являются сложными.

Мы будем рассматривать ситуацию, когда гипотез всего две. Одну из них называют *основной*, а другую — *альтернативной*, обозначая соответственно H_1 и H_2.

Определение 2.2 *Статистическим критерием* называют всякое правило, позволяющее на основании наблюдаемого выборочного вектора $\vec{X}$ принять одну из гипотез: основную или альтернативную.

При применении статистического критерия могут возникнуть ошибки двух родов. Ошибка первого рода состоит в том, что отвергается верная первая гипотеза. Ошибка второго рода — отвергается верная вторая гипотеза. Вообще ошибка i-го рода состоит в том, что статистический критерий отвергает верную i-ю гипотезу.

принимаемая гипотеза	верна гипотеза H_1	верна гипотеза H_2
H_1	нет ошибки	ошибка 2-го рода
H_2	ошибка 1-го рода	нет ошибки

Критерий характеризуется вероятностями ошибок:

$$\alpha_1 = \mathbf{P}_{H_1}(H_1 \text{ отвергается}), \qquad \alpha_2 = \mathbf{P}_{H_2}(H_2 \text{ отвергается}).$$

Здесь нижний индекс у символа вероятности указывает, при выполнении какой гипотезы подсчитывается вероятность. Из всевозможных критериев надо выбирать такие, у которых вероятности ошибок по возможности малы. К сожалению, в невырожденной статистической задаче не существует критерия, для которого обе вероятности ошибок равны нулю. Как правило, чем меньше вероятность ошибки нулевого рода, тем больше вероятность ошибки первого рода.

Рассмотрим введенные понятия на следующем примере.

Пример 2.6 *Студенты группы A считают, что они играют в шахматы вдвое лучше, чем студенты группы B. В свою очередь, студенты группы B считают, что они играют в шахматы втрое лучше, чем студенты группы A. Для решения спора назначается шахматный матч между группами A и B. С каждой стороны участвуют 3 студента, выбираемые по жребию. Решено считать справедливым мнение группы, выигрывшей матч, то есть набравшей не менее 2 очков в 3 партиях. Предполагается, что ничьих нет. Найти, в чем состоят ошибки нулевого и первого рода. Вычислить вероятности этих ошибок.*

Решение. Предполагаем, что первая гипотеза (соответствующая мнению студентов группы A) состоит в том, что вероятность выигрыша каждого студента группы A у студента группы B вдвое больше вероятности проигрыша, то есть вероятность выигрыша равна $2/3$. Согласно второй гипотезе (мнению студентов группы B), вероятность выигрыша каждого студента группы A втрое меньше вероятности проигрыша, то есть равняется $1/4$.

Итак, проводятся три испытания схемы Бернулли с вероятностью успеха p, гипотеза $H_1: \ p = 2/3$; гипотеза $H_2: \ p = 1/4$.

Критерий (исход матча) предписывает принять гипотезу H_1, если число успехов в схеме Бернулли равняется двум или трем, а в противном случае принять гипотезу H_2.

Ошибка первого рода состоит в том, что критерий предписывает считать вероятность выигрыша студента первой группы равной $1/4$ в то время, как она равняется $2/3$. Ошибка второго рода описывает противоположную ситуацию: вероятность выигрыша студента первой группы равняется $1/4$, а критерий предписывает считать ее равной $2/3$.

Вычислим вероятности ошибок.

Вероятность ошибки первого рода α_1 — это вероятность отвергнуть верную нулевую гипотезу, то есть получить ноль или один успех в схеме Бернулли, которая предполагает 3 испытания с $p = 2/3$ в каждом. Вычислим

эту вероятность на основании формулы Бернулли:

$$\alpha_1 = \mathbf{P}_{H_1}(H_1 \text{ отвергается}) =$$

$$= C_3^0(2/3)^0(1/3)^3 + C_3^1(2/3)^1(1/3)^2 = 1/27 + 2/9 \approx 0,25.$$

Вероятность ошибки второго рода α_2 — это вероятность получить два или три успеха в схеме Бернулли, которая предполагает 3 испытания с $p = 1/4$ в каждом.

$$\alpha_2 = \mathbf{P}_{H_2}(H_2 \text{ отвергается}) =$$

$$= C_3^2(1/4)^2(3/4)^1 + C_3^3(1/4)^3(3/4)^0 = 9/64 + 1/64 \approx 0,15.$$

Задачи для самостоятельного решения

8.1 Крупная партия товаров может содержать долю дефектных изделий. Поставщик полагает, что эта доля составляет 3%, а покупатель — 10%. Условия поставки: если при проверке 20 случайным образом отобранных товаров обнаружено не более одного дефектного, то партия принимается на условиях поставщика, в противном случае — на условиях покупателя. Требуется определить:

1) каковы статистические гипотезы, статистика критерия, область ее значений;

2) какое распределение имеет статистика критерия, в чем состоят ошибки первого и второго рода и каковы их вероятности.

8.2 Имеется выборка объема 1 из нормального распределения $\mathrm{N}_{a,1}$. Проверяются простые гипотезы $H_1: \ a = 0, \ H_2: \ a = 1$. Используется следующий критерий (при заданной постоянной c):

$$H_1 \Leftrightarrow X_1 \leq c.$$

Вычислить, в зависимости от c, вероятности ошибок первого и второго рода.

8.3 Построить критерий, обладающий нулевыми вероятностями ошибок, для проверки гипотез $\mathrm{H}_1 : \vec{X} \sim \mathrm{N}_{0,1}$ против $\mathrm{H}_2 : \vec{X} \sim \Pi_\lambda$.

8.4 Пусть $\vec{X} \sim \mathrm{N}_{a,1}$. Для проверки гипотез $\mathrm{H}_1 : a = 0$ против $\mathrm{H}_2 : a = 1$ используется следующий критерий: H_1 принимается, если $X_{(n)} < 3$, и отвергается в противном случае. Найти вероятности ошибок.

Индивидуальные домашние задания

Изучаются гипотезы о вероятности p появления гласной буквы среди букв, составляющих Ваши имя, отчество и фамилию. Первая гипотеза предполагает, что эта вероятность равна $1/4$, а согласно второй гипотезе $p = 1/2$.

Предлагается следующий статистический критерий: если относительная частота гласных букв не превосходит некоторой константы C, то принимается первая гипотеза, а если превосходит, то вторая. Найти такое значение константы C, чтобы сумма вероятностей ошибок первого и второго рода была минимальна. Выяснить, какая из гипотез принимается при этом выборе константы C.

§9. Критерии согласия

Удобно представлять статистический критерий как функцию $\delta(\vec{X})$ от выборочного вектора, принимающую два значения: H_1 и H_2. Наиболее общий подход для построения статистических критериев состоит в следующем.

Пусть $T = T(\vec{X})$ - некоторая статистика, характеризующая отклонение эмпирических данных, представленных выборкой, от теоретических, соответствующих проверяемой гипотезе H_1. Если распределение статистики $T(\vec{X})$ известно (точно или хотя бы приближенно), то для любого $\alpha > 0$ можно найти такое множество T_α значений T, для которого будет выполнено неравенство:

$$\mathbf{P}(T \in \mathrm{T}_\alpha / \mathrm{H}_1) \leq \alpha. \tag{2.11}$$

Пусть $\alpha > 0$ настолько мало, что событие, имеющее вероятность, не превосходящую α, может считаться практически невозможным. Тогда статистический критерий можно задать следующим образом:

$$\delta(\vec{X}) = \begin{cases} \mathrm{H}_2, & \text{если } T(\vec{X}) \in \mathrm{T}_\alpha, \\ \mathrm{H}_1, & \text{если } T(\vec{X}) \notin \mathrm{T}_\alpha. \end{cases} \tag{2.12}$$

Это правило основано на здравом смысле: оно предписывает отвергнуть гипотезу H_1 (то есть принять H_2), если происходит событие $\{T(\vec{X}) \in \mathrm{T}_\alpha\}$, которое не должно произойти, будь гипотеза H_1 справедлива. Число $\alpha > 0$, которое фигурирует в (2.11) - (2.12), называется *уровнем критерия*, или *уровнем значимости*, статистика $T(\vec{X})$ называется *статистикой критерия*, а множество T_α - *критическим множеством*.

От статистики $T = T(\vec{X})$ требуют следующих свойств:

1) при выполнении гипотезы H_1 статистика T имеет известное распределение или, по крайней мере, сходится по распределению к некоторой случайной величине J с известным распределением;

2) при выполнении гипотезы H_2 статистика T сходится почти наверное к бесконечности с ростом объема выборки.

Для того, чтобы получить критерий уровня α, задают критическое множество в виде

$$\mathrm{T}_\alpha = \{T \geq C\},$$

где C — константа, определяемая условием

$$\mathbf{P}\{J \geq C\} = \alpha,$$

то есть $F_J(C) = 1-\alpha$. Ясно, что при таком выборе константы C вероятность ошибки первого рода α_1 либо равна уровню критерия α (в случае, когда статистика T при верной первой гипотезе распределена в точности как J), либо, по крайней мере, сходится к α с ростом объема выборки.

Сходимость статистики T почти наверное к бесконечности при выполненной первой гипотезе гарантирует *состоятельность* критерия, то есть сходимость вероятности ошибки второго рода α_1 к нулю с ростом объема выборки.

Для каждой конкретной выборки $\vec{X}$ можно найти предельное значение уровня $\alpha^* = \alpha^*(\vec{X})$, при котором гипотеза H_1 еще может быть принята. Такое значение называется *(реально) достигаемым уровнем значимости* (пи-значением). Реально достигаемый уровень значимости — это вероятность получить худшее согласие с проверяемой гипотезой, чем реально полученное, если гипотеза H_1 верна. Поэтому чем меньше α^*, тем более это говорит против гипотезы H_1.

Достигаемый уровень значимости вычисляется с помощью распределения статистики J:

$$\alpha^* = \mathbf{P}\{J \geq T(\vec{X})\} = 1 - F_J(T(\vec{X})).$$

В терминах достигаемого уровня значимости критическая область имеет вид

$$\mathrm{T}_\alpha = \{\alpha^* \leq \alpha\},$$

то есть основная гипотеза отвергается на уровне α в случае, когда $\alpha^* \leq \alpha$.

Каждый критерий согласия использует свою статистику, предназначенную для различения основной гипотезы и альтернативы и обладающую нужными свойствами: сходимостью к фиксированному распределению при выполнении основной гипотезы и сходимостью почти наверное к бесконечности при ее невыполнении.

В качестве важных примеров критериев согласия рассмотрим критерий Колмогорова.

Рассмотрим выборку $\vec{X} \Subset \mathrm{F}$ объема n с неизвестной функцией распределения F и простую гипотезу $\mathrm{H}_1 : \mathrm{F} = \mathrm{F}_0$. Альтернативной для H_1 является сложная гипотеза $\mathrm{H}_2 : \mathrm{F} \neq \mathrm{F}_0$.

Критерий Колмогорова применяется в случае, когда функция распределения $\mathrm{F}_1(t)$ непрерывна. Рассматривается следующее расстояние между

эмпирической и теоретической функциями распределения:

$$D_n = D(F_n^*, F_1) = \sup_{-\infty<t<\infty} |F_n^*(t) - F_1(t)| = \max_{-\infty<t<\infty} |F_n^*(t) - F_1(t)|.$$

В качестве статистики критерия Колмогорова выбирается это расстояние, умноженное на $\sqrt{n}$, где n — объем выборки:

$$T_n = \sqrt{n} D_n = \sqrt{n} \max_{-\infty<t<\infty} |F_n^*(t) - F_1(t)|.$$

А. Н. Колмогоров доказал следующие свойства статистики T_n:

1) если гипотеза H_1 верна, то T_n с ростом n сходится к случайной величине J с функцией распределения, называемой функцией распределения Колмогорова:

$$F_J(t) = 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2t^2};$$

2) если гипотеза H_1 неверна, то T_n сходится почти наверное к $+\infty$ при $n \to \infty$.

Таким образом, достигаемый уровень значимости критерия Колмогорова равен

$$\alpha^* = 1 - F_J(T_n) = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2T_n^2} = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2nD_n^2}. \qquad (2.13)$$

Отметим, что для расчетов по этой формуле нужно брать не всю бесконечную сумму, а только несколько слагаемых, при этом ошибка вычислений не превосходит последнего отброшенного слагаемого. Критерий Колмогорова отвергает гипотезу H_1 на уровне α, если $\alpha^* \le \alpha$.

Для практического вычисления статистики $D_n = D_n(\vec{X})$ можно использовать следующую формулу:

$$D_n(\vec{X}) = \max_{1\le i\le n} \max \left(\left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right). \qquad (2.14)$$

Здесь $X_{(i)}$ — это элементы *вариационного ряда*, то есть для этих вычислений выборку следует предварительно *упорядочить по возрастанию.*

Пример 2.7 *Вариационный ряд выборки имеет вид* (1; 2; 3; 4; 5; 6; 7; 8; 9; 10). *Проверить гипотезу о равномерности распределения элементов выборки на отрезке от 0 до 10 с помощью критерия Колмогорова: найти реализацию достигаемого уровня значимости и сделать вывод о принятии гипотезы на уровнях 0,1 и 0,01.*

Решение. Построим на одном графике эмпирическую $F_n^*(t)$ и теоретическую $F_1(t)$ функции распределения.

Эмпирическая функция распределения — это ступенчатая функция, высота ступеньки равна 1/10 в точках 1; ...; 10.

Теоретическая функция распределения равномерного закона на отрезке от 0 до 10 равна

$$\mathrm{F}_1(t) = \begin{cases} 0, & \text{если } t \le 0, \\ t/10, & \text{если } 0 < t \le 10, \\ 1, & \text{если } t > 10. \end{cases} \tag{2.15}$$

Так как функция распределения $\mathrm{F}_1(t)$ непрерывна, то можно применять критерий Колмогорова. Найдем по графику значение D_n — наибольшую по модулю разность между эмпирической и теоретической функциями распределения. Эта разность достигается в точках разрыва эмпирической функции распределения и равна 1/10. Вычислим реализацию достигаемого уровня значимости, вспоминая, что $n = 10$: согласно (2.13),

$$\alpha^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2} \approx$$

$$\approx 2e^{-0,2} - 2e^{-4\cdot 0,2} + 2e^{-9\cdot 0,2} - 2e^{-16\cdot 0,2} + 2e^{-25\cdot 0,2} - 2e^{-36\cdot 0,2} + 2e^{-49\cdot 0,2} \approx 0,99997.$$

Достигаемый уровень значимости оказался близким к 1; это означает, что нет оснований отвергать гипотезу о равномерности выборочных значений. Эту гипотезу следовало бы отвергнуть только в случае, когда достигаемый уровень значимости оказался бы близким к нулю.

В частности, в нашем случае выполнено неравенство $\alpha^* > 0,1$. Следовательно, гипотеза о равномерности принимается на уровне 0,1. Тем более она будет приниматься на уровне 0,01.

Пример 2.8 *Решить пример 2.7 для реализации выборки* (10; 0; 0; 10; 10; 10; 0; 0; 0; 10).

Решение. Упорядочив реализацию выборки по неубыванию, получим реализацию вариационого ряда: (0; 0; 0; 0; 0; 10; 10; 10; 10; 10). Как и в предыдущем примере, построим на одном графике эмпирическую $F_n^*(t)$ и теоретическую $F_1(t)$ функции распределения. В отличие от предыдущего примера, эмпирическая функция распределения здесь имеет всего две ступеньки в точках 0 и 10, высотой по $5/10 = 0,5$. Теоретическая функция распределения остается той же самой и определяется формулой (2.15). Значение D_n достигается в точках разрыва эмпирической функции распределения и

равняется $0,5$. Вычислим реализацию достигаемого уровня значимости:

$$\alpha^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2nD_n^2} \approx 2e^{-2\cdot 10\cdot 0,5^2} = 2e^{-5} \approx 0,0135.$$

Здесь мы взяли только одно слагаемое суммы, так как остальные слагаемые гораздо меньше.

В этом примере достигаемый уровень значимости оказался близким к 0, что говорит против гипотезы H_0. В частности, $\alpha^* < 0,1$, то есть гипотеза однородности отвергается на уровне 0,1. Однако она принимается на более низком уровне 0,01, так как $\alpha^* > 0,01$.

9.1 Используя конструкции доверительного интервала, построить критерий уровня ε для проверки гипотезы $\mathrm{H} : \theta = 1$, если:

а) $\vec{X} \sim \mathrm{N}_{\theta,1}$;

б) $\vec{X} \sim \mathrm{N}_{1,\theta}$;

в) $\vec{X} \sim \mathrm{E}_{\theta}$;

г) $\vec{X} \sim \mathrm{B}_{\theta/2}$;

д) $\vec{X} \sim \Pi_{\theta}$.

Индивидуальные домашние задания

Возьмите 20 книг, запишите автора, название и число страниц. Для каждой книги найдите остаток от деления числа страниц на 100. Например, если в книге 256 страниц, то остаток от деления на 100 равен 0,56. Для этих остатков от деления проверьте гипотезу о равномерности на отрезке от 0 до 1: найдите пи-значение (реально достигнутый уровень значимости) и проверьте, принимается ли гипотеза о равномерности на уровнях 0,1; 0,01; 0,001.

§10. Критерий хи-квадрат Пирсона

Если гипотетическая функция распределения $\mathrm{F}(x)$ не является непрерывной, то критерий Колмогорова неприменим. В этом случае можно воспользоваться χ^2-*критерием Пирсона*. Статистика критерия Пирсона строится после предварительного «группирования» выборочных данных. Для этого все множество S возможных значений случайных величин X_i разбивается на конечное число непересекающихся частей:

$$S = S_1 \cup S_2 \cup \cdots \cup S_r, \quad S_i \cap S_j = \emptyset, i \neq j.$$

Обозначим ν_j — число элементов выборки $\vec{X}$, попавших в множество S_j, а p_j — вероятность попадания случайной величины X_i в множество S_j, вы-

численная с помощью гипотетической функции распределения $\mathrm{F} = \mathrm{F}_0$. Тогда в качестве статистики критерия χ^2 рассматривают следующую предложенную Пирсоном меру отклонения эмпирического распределения от предполагаемого теоретического:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j}. \tag{2.16}$$

Справедлива следующая теорема, позволяющая находить распределение статистики χ^2 при больших значениях n, а стало быть, и строить статистический критерий.

Теорема 2.2 *Если гипотеза* H *однозначно фиксирует вероятности* $p_1, p_2, \ldots, p_r$, *где* $p_j = \mathbf{P}(X_i \in S_j)$, *то при выполнении этой гипотезы статистика* $\chi^2(\vec{X})$ *слабо сходится к распределению* χ^2_{r-1}:

$$\chi^2 \Longrightarrow \chi^2_{r-1}, \quad n \to \infty.$$

При невыполнении основной гипотезы статистика $\chi^2(\vec{X})$ *сходится почти наверное к* $+\infty$.

Для построения критерия, основанного на статистике χ^2, используем распределение χ^2_{r-1}, и по найденному значению $\chi^2(\vec{X})$ отыскиваем достигаемый уровень значимости

$$\alpha^* = 1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X}))$$

по таблице 5 распределения хи-квадрат или с помощью математических пакетов. В пакете Microsoft Excel достигаемый уровень значимости вычисляется формулой

```
=ХИ2РАСП(ячейка;r-1)
```  (2.17)

(в качестве ячейки надо подставить адрес ячейки, в которой вычислена статистика хи-квадрат, а $r-1$ — число степеней свободы).

Тогда критерий Пирсона имеет следующий вид:

$$\mathrm{H}_0 \Leftrightarrow \ \alpha^* > \alpha. \tag{2.18}$$

Заметим, что для практического применения рекомендуется разбиение производить таким образом, чтобы выполнялось условие $np_j \geq 10$. При нарушении этого условия нужно объединить соседние множества S_j. Вероятности p_j надо выбирать по возможности равными.

Критерий хи-квадрат часто используют для проверки сложных гипотез о принадлежности распределения к некоторому параметрическому семейству (например, к нормальному). При этом вместо известных вероятностей

p_j подставляют их оценки p_j^*, полученные путем оценивания неизвестных парметров распределения. Важно понимать, что в этом случае предельное распределение статистики $\chi^2(\vec{X})$ уже не будет распределением χ^2_{r-1}, а будет близко к распределению χ^2_{r-1-s}, где s — число оцениваемых параметров ($s = 2$ для нормального распределения). Более точно, предельная функция распределения заключена между функциями распределения χ^2_{r-1-s} и χ^2_{r-1}.

Достигаемый уровень значимости α^* заключен между $1 - F_{\chi^2_{r-1}}(\chi^2(\vec{X}))$ и $1 - F_{\chi^2_{r-1-s}}(\chi^2(\vec{X}))$, где s — число оцениваемых параметров.

Для того, чтобы получить в точности распределение хи-квадрат с $r-1-s$ степенями свободы, следует оценивать неизвестные параметры методом максимального правдоподобия по *группированной* выборке, но это приводит, как правило, к сложным вычислительным процедурам.

Пример 2.9 *Проверить гипотезу о равномерности на отрезке от 0 до 10 для выборок из двух предыдущих примеров с помощью критерия хи-квадрат Пирсона: найти реализации достигаемых уровней значимости и сделать выводы о принятии гипотезы на уровнях 0,1 и 0,01. Число промежутков группирования выбрать по формуле Стеджеса.*

Решение.

Согласно формуле Стеджеса (1.3), вычисляем целую часть логарифма по основанию 2 от объема выборки и прибавляем единицу:

$$r = [\log_2 n] + 1 = [\log_2 10] + 1 = 3 + 1 = 4,$$

так как $2^3 = 8 < 10 < 2^4 = 16$.

Итак, множество допустимых выборочных значений — отрезок $[0;\ 10]$ — следует разбить на 4 промежутка равной длины:

$$S_1 = [0;\ 2,5), \quad S_2 = [2,5;\ 5), \quad S_3 = [5;\ 7,5), \quad S_4 = [7,5;\ 10].$$

Согласно нулевой гипотезе, распределение равномерное на отрезке от 0 до 10. Следовательно, равны вероятности попадания элемента выборки в отрезки равной длины:

$$p_1 = p_2 = p_3 = p_4 = 1/4 = 0,25.$$

Значения статистики хи-квадрат Пирсона различны для примеров 2.7 и 2.8:

1) В примере 2.7 количества элементов, попавших в каждый из промежутков, равны соответственно

$$\nu_1 = 2, \quad \nu_2 = 2, \quad \nu_3 = 3, \quad \nu_4 = 3.$$

Вычислим статистику хи-квадрат согласно формуле (2.16):

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} =$$

$$= \frac{(2 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(2 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(3 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(3 - 10 \cdot 0,25)^2}{10 \cdot 0,25} = 0,4.$$

Найдем достигнутый уровень значимости по формуле (2.17), используя функцию ХИ2РАСП и подставляя значение 0,4 и число степеней свободы, равное $r - 1 = 4 - 1 = 3$:

$$\texttt{ХИ2РАСП(0,4;3)} \approx 0,94.$$

Итак, здесь достигнут уровень значимости 0,94, что не дает оснований отвергать гипотезу о равномерности ни на уровне $0,1 < 0,94$, ни тем более на уровне 0,01.

2) В примере 2.8 количества элементов, попавших в каждый из промежутков, принимают значения

$$\nu_1 = 5, \quad \nu_2 = 0, \quad \nu_3 = 0, \quad \nu_4 = 5.$$

Как и в пункте (1), вычислим статистику хи-квадрат и найдем достигнутый уровень значимости:

$$\chi^2(\vec{X}) = \sum_{j=1}^{r} \frac{(\nu_j - np_j)^2}{np_j} =$$

$$= \frac{(5 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(0 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(5 - 10 \cdot 0,25)^2}{10 \cdot 0,25} + \frac{(0 - 10 \cdot 0,25)^2}{10 \cdot 0,25} = 10;$$

$$\texttt{ХИ2РАСП(10;3)} \approx 0,0186.$$

В этом примере достигнут низкий уровень значимости 0,0186, что дает основания отвергать гипотезу о равномерности на уровне $0,1 > 0,0186$, но не на уровне 0,01.

Отметим, что для рассмотренных примеров критерии Колмогорова и хи-квадрат Пирсона дают похожие результаты — достигнутые уровни значимости для обоих критериев оказались довольно близкими. В случае, когда основная гипотеза предполагает дискретное распределение, критерий Колмогорова неприменим, и мы будем пользоваться только критерием хи-квадрат Пирсона.

Пример 2.10 *При* 4040 *бросаниях монеты Бюффон получил* $\nu_1 = 2048$ *выпадений герба и* $\nu_2 = n - \nu_1 = 1992$ *выпадений решетки. Согласуется ли это с гипотезой о том, что монета правильная, при уровне*

значимости $\alpha = 0,1$*? С каким предельным уровнем значимости может быть принята эта гипотеза?*

Решение. Можно считать, что мы имеем дело со статистической моделью $\vec{X} \in \mathrm{B}_p$, где неизвестен параметр p - вероятность выпадения герба. Проверяемая гипотеза $\mathrm{H}_0 : \ p = 0,5$. Поскольку выборочные данные уже сгруппированы ($\nu_1 = 2048$ — число значений $X_i = 1$, ν_2 — число значений $X_i = 0$), то можем вычислить наблюдаемое значение статистики χ^2:

$$p_1 = \mathbf{P}_{H_0}(X_i = 1) = 0,5; \quad p_2 = \mathbf{P}_{H_0}(X_i = 0) = 0,5;$$

$$\frac{(\nu_1 - np_1)^2}{np_1} = \frac{(2048 - 2020)^2}{2020} = 0,285;$$

$$\frac{(\nu_2 - np_2)^2}{np_2} = \frac{(1992 - 2020)^2}{2020} = 0,388; \qquad \chi^2 = 0,285 + 0,388 = 0,673.$$

Число множеств разбиения $r = 2$, поэтому достигнутый уровень значимости

$$\texttt{ХИ2РАСП(0,673;1)} \approx 0,412.$$

Достигнутый уровень значимости довольно высок. В частности, $0,412 > 0,1$, то есть гипотеза о симметричности монеты принимается на уровне 0,1.

10.1 При n=4000 независимых испытаний события $A1$, $A2$, $A3$, составляющие полную группу, осуществились соответственно 1905, 1015 и 1080 раз. Проверить, согласуются ли эти данные при уровне значимости 0,05 с гипотезой H_0: $p_1 = 1/2$, $p_2 = p_3 = 1/4$, где $p_j = \mathbf{P}(A_j)$. Найти достигнутый уровень значимости.

10.2 В экспериментах с селекцией гороха Мендель наблюдал частоты различных видов семян, полученных при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей по теории наследственности приведены в следующей таблице:

Семена	Частота	Вероятность
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16
Σ	n=556	1

Следует проверить гипотезу H_0 о согласовании частотных данных с теоретическими вероятностями (на уровне значимости 0,1) и найти достигнутый уровень значимости.

10.3 В таблице приведены числа m_i участков равной площади 0,25 км^2 южной части Лондона, на каждый из которых приходилось по i попаданий самолетов-снарядов во время второй мировой войны. Проверить согласие опытных данных с законом распределения Пуассона, приняв за уровень значимости $\alpha = 0,05$:

i	0	1	2	3	4	5 и более	Итого
m_i	229	211	93	35	7	1	$\Sigma m_i = 576$

Индивидуальные домашние задания

Проверьте следующие две гипотезы с помощью подходящих критериев хи-квадрат.

Простая гипотеза: буквы Ваших имени, отчества и фамилии с одинаковыми вероятностями имеют номера с 1 по 11, с 12 по 22, с 23 по 33.

Сложная гипотеза: вероятность появления гласной буквы среди букв Вашего имени такая же, как среди букв Вашего отчества, и такая же, как среди букв Вашей фамилии.

Для каждой гипотезы найдите реально достигаемый уровень значимости (пи-значение, p-value) и сделайте вывод о том, принимается ли гипотеза на уровнях 0,01; 0,02; 0,05; 0,1; 0,2.

Глава 3

Регрессионный и дисперсионный анализ

§11. Однопараметрическая и парная регрессия

В модели однопараметрической регрессии предполагается, что

$$y_i = \theta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

Здесь y_i — отклики, θ — неизвестный параметр регрессии, x_i — регрессоры, ε_i — независимые и одинаково распределенные регрессионные ошибки, $\mathbf{E}\,\varepsilon_i = 0$, $\mathbf{Var}\,\varepsilon_i = \sigma^2 > 0$.

В модели нормальной регрессии дополнительно предполагается нормальное распределение регрессионных ошибок: $\varepsilon_i \sim N_{0,\sigma^2}$.

Оценка параметра θ может быть найдена по методу наименьших квадратов:

$$\sum_{i=1}^{n} (y_i - \theta x_i)^2 \to \min_{\theta}.$$

Продифференцировав по θ и приравняв производную нулю, получаем уравнение, из которого выражается оценка $\widehat{\theta}$ параметра θ:

$$\widehat{\theta} = \overline{xy}/\overline{x^2}.$$

На основании этой оценки вычисляются прогнозные значения $\widehat{y}_i = \widehat{\theta} x_i$ и регрессионные остатки

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i.$$

В модели нормальной регрессии оценка $\widehat{\theta}$ является оценкой максимального правдоподобия, и оценка $\widehat{\sigma^2}$ дисперсии σ^2 по методу максимального правдоподобия

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{\varepsilon}_i)^2.$$

Качество регрессионной модели нередко характеризуют коэффициентом детерминации

$$R^2 = 1 - \widehat{\sigma^2}/S_y^2,$$

где $S_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2 = \overline{y^2} - (\overline{y})^2$ — выборочная дисперсия откликов.

Отметим, что в модели однопараметрической регрессии значение коэффициента детерминации R^2 может быть и отрицательным, так как основано на сравнении регрессионной модели с моделью $y_i = const + \varepsilon_i$.

Упражнение По двум точкам (1, 1) и (2, 0) найдите оценки параметров однопараметрической регрессии, прогнозные значения и регрессионные остатки, коэффициент детерминации. Изобразите на чертеже линию регресии и точки данных. Найдите на чертеже прогнозные значения и регрессионные остатки.

Модель простой линейной регрессии также предполагает, что даны точки $(x_1,\ y_1), \ldots, (x_n,\ y_n)$, но регрессоры и отклики связаны соотношением

$$y_i = kx_i + b + \varepsilon_i, \quad i = 1, \ldots, n.$$

Оценивание параметров по методу наименьших квадратов состоит в отыскании прямой $y = \widehat{k}x + \widehat{b}$, наилучшим образом приближающей эти точки в следующем смысле: значение суммы квадратов отклонений значений y_i от соответствующих значений $\widehat{k}x_i + \widehat{b}$ достигает минимума:

$$\sum_{i=1}^n (y_i - \widehat{k}x_i - \widehat{b})^2 \to \min.$$

Решение задачи дается следующими формулами:

$$\widehat{k} = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - (\overline{x})^2};$$

$$\widehat{b} = \overline{y} - k\overline{x}.$$

Вероятностная постановка задачи, приводящая к тому же ответу, состоит в следующем: случайные величины y_i независимы и имеют нормальное распределение с одной и той же дисперсией и с математическим ожиданием $kx_i + b$. Тогда $\widehat{k}$ и $\widehat{b}$ — это оценки параметров k, b.

На основании этой оценки вычисляются прогнозные значения $\widehat{y}_i = \widehat{k}x_i + \widehat{b}$ и регрессионные остатки

$$\widehat{\varepsilon}_i = y_i - \widehat{y}_i.$$

Как и в однопараметрическом случае, оценка $\widehat{\sigma^2}$ дисперсии σ^2 по методу максимального правдоподобия

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (\widehat{\varepsilon}_i)^2.$$

Качество регрессионной модели нередко характеризуют коэффициентом детерминации

$$R^2 = 1 - \widehat{\sigma^2}/S_y^2,$$

где $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - (\overline{y})^2$ — выборочная дисперсия откликов.

Для этой модели всегда $R^2 \geq 0$.

Задачи для самостоятельного решения

11.1 Методом наименьших квадратов оценить неизвестные параметры регрессии.

x_i	1	2	3
y_i	5	-1	2

11.2 Методом наименьших квадратов оценить неизвестные параметры регрессии.

x_i	1	2	3	4
y_i	2	4	3	5

Индивидуальные домашние задания

Постройте модель линейной регрессии числа букв своих имени, отчества, фамилии на число гласных букв (а, е, е, и, о, у, ы, э, ю, я).

Оцените параметры регрессии и постройте на одном графике точки данных и прямую регрессии.

x_i — число гласных букв в имени (x_1), отчестве (x_2), фамилии (x_3).

y_i — число букв в имени (y_1), отчестве (y_2), фамилии (y_3).

Примечание. Если число гласных букв в имени, отчестве, фамилии одинаково ($x_1 = x_2 = x_3$), то используйте однопараметрическую модель с $b = 0$. При этом прямая регрессии $y = \widehat{k}x$, где $\widehat{k} = \frac{\overline{xy}}{\overline{x^2}}$.

§12. Общая регрессионная модель

Общая модель линейной регрессии имеет вид

$$\vec{y} = X\vec{\theta} + \vec{\varepsilon}.$$

Здесь $\vec{y} = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$ — вектор отклика, $X = \begin{pmatrix} x_{11} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nm} \end{pmatrix}$ — матрица регрессора, $\vec{\theta} = \begin{pmatrix} \theta_1 \\ \dots \\ \theta_m \end{pmatrix}$ — вектор неизвестных параметров, $\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$ —

вектор регрессионных ошибок (шума).

Перемножая матрицу X на вектор θ, запишем это равенство по строкам:

$$y_i = x_{i1}\theta_1 + \ldots + x_{im}\theta_m + \varepsilon_i, \quad i = 1, \ldots, n.$$

В модели нормальной регрессии предполагается, что $\varepsilon_1, \ldots, \varepsilon_n$ независимы и имеют нормальное распределение с нулевым математическим ожиданием и неизвестной дисперсией σ^2.

Будем также предполагать, что матрица $X^T X$ невырождена, то есть $\det X^T X \neq 0$. Здесь X^T означает транспонирование матрицы X.

Оценка максимального правдоподобия $\widehat{\theta}$ векторного параметра θ (совпадающая с оценкой метода наименьших квадратов)

$$\widehat{\theta} = (X^T X)^{-1} X^T \vec{y}.$$

Вектор прогнозных значений $\widehat{y} = X\widehat{\theta}$, вектор остатков регрессии $\widehat{\varepsilon} = \vec{y} - \widehat{y}$.

Оценка параметра σ^2 методом максимального правдоподобия

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (\widehat{\varepsilon}_i)^2.$$

Как и для частных случаев из предыдущего параграфа, аккуратность приближения данных можно характеризовать коэффициентом детерминации

$$R^2 = 1 - \widehat{\sigma^2}/S_y^2,$$

где $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - (\overline{y})^2$ — выборочная дисперсия откликов.

Сам термин «регрессия» придуман Гальтоном при изучении зависимости роста детей от роста родителей. Гальтон обнаружил, что средний рост детей (отдельно для сыновей и дочерей) вычисляется как линейная комбинация роста отца и матери. Эта модель прогнозирует, что если отец очень высокий, то рост его детей будет в среднем меньше его роста, и Гальтон назвал это явление «регрессией к среднему». Поясним результат Гальтона на следующем условном примере. Целые числа взяты для упрощения вычислений.

Пример Предположим, что у отца ростом 1 м и матери ростом 1 м рост взрослого сына составил 1,4 м; у отца ростом 2 м и матери ростом 1 м рост взрослого сына составил 1,6 м; у отца ростом 2 м и матери ростом 2 м рост взрослого сына составил 1,8 м.

Модель Гальтона предполагает, что

$$y_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \varepsilon_i,$$

где i — номер наблюдения, y_i — рост сына, x_{i1} — рост отца, x_{i2} — рост матери, ε_i — шум, то есть случайное отклонение от линейной зависимости, вызванное внутренними и внешними факторами. Неизвестные коэффициенты θ_1 и θ_2 характеризуют вклад роста отца и роста матери в рост сына. Найдем оценки этих коэффициентов, для этого запишем вектор откликов и матрицу регрессора:

$$\vec{y} = \begin{pmatrix} 1{,}4 \\ 1{,}6 \\ 1{,}8 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 2 \end{pmatrix}.$$

Тогда

$$X^T = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 2 \end{pmatrix} = \begin{pmatrix} 9 & 7 \\ 7 & 6 \end{pmatrix},$$

$$\det(X^T X) = \det \begin{pmatrix} 9 & 7 \\ 7 & 6 \end{pmatrix} = 54 - 49 = 5 \neq 0,$$

$$(X^T X)^{-1} = \frac{1}{\det(X^T X)} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^T = \frac{1}{5} \begin{pmatrix} 6 & -7 \\ -7 & 9 \end{pmatrix}^T = \frac{1}{5} \begin{pmatrix} 6 & -7 \\ -7 & 9 \end{pmatrix},$$

здесь A_{ij} — алгебраические дополнения (миноры с соответствующими знаками). Далее,

$$X^T \vec{y} = \begin{pmatrix} 1 & 2 & 2 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1{,}4 \\ 1{,}6 \\ 1{,}8 \end{pmatrix} = \begin{pmatrix} 1{,}4 + 3{,}2 + 3{,}6 \\ 1{,}4 + 1{,}6 + 3{,}6 \end{pmatrix} = \begin{pmatrix} 8{,}2 \\ 6{,}6 \end{pmatrix},$$

$$\widehat{\theta} = (X^T X)^{-1} X^T \vec{y} = \frac{1}{5} \begin{pmatrix} 6 & -7 \\ -7 & 9 \end{pmatrix} \begin{pmatrix} 8{,}2 \\ 6{,}6 \end{pmatrix} = \begin{pmatrix} 0{,}6 \\ 0{,}4 \end{pmatrix}.$$

Итак, $\widehat{\theta}_1 = 0{,}6$, $\widehat{\theta}_2 = 0{,}4$, то есть оценка вклада роста отца в рост сына 60%, а вклада роста матери 40%. Вычисляем прогнозные значения

$$\widehat{y} = X\widehat{\theta} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} 0{,}6 \\ 0{,}4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1{,}6 \\ 2 \end{pmatrix}$$

и остатки регрессии

$$\vec{y} - \widehat{y} = \begin{pmatrix} 1{,}4 \\ 1{,}6 \\ 1{,}8 \end{pmatrix} - \begin{pmatrix} 1 \\ 1{,}6 \\ 2 \end{pmatrix} = \begin{pmatrix} 0{,}4 \\ 0 \\ -0{,}2 \end{pmatrix}.$$

Оценка дисперсии регрессионных ошибок

$$\widehat{\sigma^2} = \frac{1}{3}(0{,}4^2 + 0^2 + (-0{,}2)^2) = 1/15,$$

коэффициент детерминации отрицательный:

$$R^2 = 1 - \frac{1}{15/3((-0{,}2)^2 + 0^2 + 0{,}2^2)} = 1 - 2{,}5 = -1{,}5 < 0.$$

Задачи для самостоятельного решения

12.1 Методом наименьших квадратов оценить неизвестные параметры регрессии.

x_{i1}	1	2	3
x_{i2}	2	2	3
y_i	5	4	2

12.2 Методом наименьших квадратов оценить неизвестные параметры регрессии.

x_{i1}	1	2	3	4
x_{i2}	1	1	3	3
y_i	2	4	3	5

Индивидуальные домашние задания

Возьмите 5 разных книг разной высоты, для каждой книги запишите название, автора, высоту книги в мм (x_{i1}), число страниц (x_{i2}), толщину книги в мм вместе с обложкой (y_i). Найдите оценки коэффициентов регрессии и проинтерпретируйте их. Найдите коэффициент детерминации.

§13. Корреляционный анализ

При использовании модели простой нормальной регрессии

$$Y_i = a + bX_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

где $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}_{0,\sigma^2}$ и независимы, возникает вопрос о том, не равен ли коэффициент b нулю.

Если $b = 0$, то $Y_1, \ldots, Y_n$ одинаково распределены, и нет никакой вероятностной связи между значениями иксов и игреков.

Для проверки гипотезы о том, что $b = 0$, строятся различные статистические критерии согласия. Основная гипотеза утверждает, что $b = 0$, а альтернативная состоит в том, что $b \neq 0$.

Изучим критерии согласия, основанные на выборочных коэффициентах корреляции Пирсона и Спирмена.

Выборочный коэффициент корреляции Пирсона вычисляется по формуле

$$r_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X S_Y},$$

где S_X, S_Y — выборочные среднеквадратические (стандартные) отклонения иксов и игреков:

$$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sqrt{\overline{X^2} - (\overline{X})^2},$$

$$S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \sqrt{\overline{Y^2} - (\overline{Y})^2}.$$

Если $X_1, \ldots, X_n$ независимы и одинаково распределены в соответствии с нормальным распределением, то при выполнении основной гипотезы $b = 0$ статистика

$$t_{n-2} = r_n\sqrt{\frac{n-2}{1-r_n^2}}$$

имеет распределение Стьюдента с $(n-2)$ степенями свободы. Поэтому реально достигаемый уровень значимости (пи-значение) для конкретного наблюдаемого значения r_n

$$\alpha^* = 2\left(1 - F_{T_{n-2}}\left(|t_{n-2}|\right)\right).$$

Пример

X_i	1	2	2	3
Y_i	3	3	1	1

Здесь $n = 4$, $\overline{X} = \overline{Y} = 2$, $\overline{X^2} = 4{,}5$, $\overline{Y^2} = 5$, $S_X = \sqrt{0{,}5}$, $S_Y = 1$, $\overline{XY} = 3{,}5$, $r_n = -\sqrt{2}/2$, $t_{n-2} = -\sqrt{2}$, α^*=2(1-t.dist(КОРЕНЬ(2);2;1))≈ 0,2929.

Полученное пи-значение (реально достигнутый уровень значимости) не является маленьким, поэтому нет оснований отвергать основную гипотезу.

Выборочный коэффициент корреляции Спирмена для двумерной выборки $(\vec{X}, \vec{Y})$ вычисляется по формуле

$$\rho_n = \frac{\overline{UV} - \overline{U}\,\overline{V}}{S_U S_V},$$

где $\vec{U}$, $\vec{V}$ — векторы рангов, то есть номеров значений векторов $\vec{X}$ и $\vec{Y}$, упорядоченных по неубыванию. Если несколько значений совпадают, то для

них вычисляется средний ранг. При выполнении гипотезы о независимости выборок иксов и игреков статистика

$$\tau_{n-2} = \rho_n \sqrt{\frac{n-2}{1-\rho_n^2}}$$

имеет лишь приближенно распределение Стьюдента с $(n-2)$ степенями свободы, и реально достигаемый уровень значимости (пи-значение) для конкретного наблюдаемого значения ρ_n

$$\alpha^* \approx 2\left(1 - F_{T_{n-2}}\left(|\tau_{n-2}|\right)\right).$$

Пример

X_i	1	2	2	3
Y_i	3	3	1	1

Для этого упорядочим значения по неубыванию.

$X_{(i)}$ $1 \le 2 \le 2 \le 3$

i 1 2 3 4

Значению 1 соответствует ранг 1, значению 2 средний ранг 2,5, значению 3 ранг 4.

X_i	1	2	2	3
U_i	1	2,5	2,5	4

Проделаем то же для игреков.

$Y_{(i)}$ $1 \le 1 \le 3 \le 3$

i 1 2 3 4

Значению 1 соответствует средний ранг 1,5, значению 3 средний ранг 3,5.

Y_i	3	3	1	1
V_i	3,5	3,5	1,5	1,5

Вместо исходной таблицы значений получаем таблицу рангов:

U_i	1	2,5	2,5	4
V_i	3,5	3,5	1,5	1,5

Здесь $n = 4$, $\overline{U} = \overline{V} = 2{,}5$, $\overline{U^2} = 59/8$, $\overline{V^2} = 29/4$, $S_U = \sqrt{9/8}$, $S_V = 1$, $\overline{UV} = 5{,}5$, $\rho_n = -v\sqrt{2}/2$, $\tau_{n-2} = -\sqrt{2}$, $\alpha^* \approx 2$(1-t.dist(КОРЕНЬ(2);2;1))$\approx$ 0,2929.

Задачи для самостоятельного решения

13.1 Построить точки на чертеже и найти, на каком уровне значимости принимается гипотеза о независимости координат точек по критерию Пирсона.

X_i	1	2	3	4	5
Y_i	1	1	2	3	3

13.2 Построить точки на чертеже и найти, на каком уровне значимости принимается гипотеза о независимости координат точек по критерию Пирсона и по критерию Спирмена.

X_i	1	2	3	3
Y_i	1	2	2	5

Индивидуальные домашние задания

Возьмите 10 разных книг, для каждой книги запишите автора, название, число страниц (X_i), толщину книги в мм без обложки (Y_i). Постройте зависимость на графике. Найдите оценки коэффициентов кореляции Пирсона и Спирмена и реально достигнутые уровни значимости для гипотезы о независимости толщины книги от числа страниц.

§14. Однородность двух выборок

Приведем формулировки для двухвыборочного статистического теста.

Пусть $\vec{x}_{n_1} = (x_1, \ldots, x_{n_1})$ и $\vec{y}_{n_2} = (y_1, \ldots, y_{n_2})$ — конечные последовательности чисел.

Статистическая гипотеза предполагает, что эти конечные последовательности являются конечными реализациями бесконечных последовательностей случайных величин $(X_1, X_2, \ldots)$ и $(Y_1, Y_2, \ldots)$ с заданным совместным распределением.

Более подробно, случайные последовательности заданы на некотором вероятностном пространстве, и для некоторого $\omega \in \Omega$ выполнено:

$x_i = X_i(\omega)$ для всех $i = 1, \ldots, n_1$;

$y_i = Y_i(\omega)$ для всех $i = 1, \ldots, n_2$.

Простыми гипотезами h называются статистические гипотезы, для которых распределения однозначно определены.

Основная гипотеза H и альтернативная гипотеза $\overline{H}$ — это два непересекающихся множества простых гипотез.

Статистикой называется функция

$$J_{n_1,n_2}(\vec{x}_{n_1}, \vec{y}_{n_2}).$$

Для построения статистического теста будем требовать от статистики следующих свойств:

1. Для любой $h \in H$ имеет место слабая сходимость

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) \Rightarrow J$$

при $n_1 \to \infty$, $n_2 \to \infty$, где случайная величина J имеет известную непрерывную функцию распределения $F_J(x)$.

2. Для любой $h \in \overline{H}$ имеет место сильная сходимость

$$J_{n_1,n_2}(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) \overset{a.s.}{\to} +\infty$$

при $n_1 \to \infty$, $n_2 \to \infty$.

Определим пи-значение (p-value, реально достигаемый уровень значимости)

$$\text{p-value} = \varepsilon^* = 1 - F_J(J_{n_1,n_2}(\vec{x}_{n_1}, \vec{y}_{n_2})).$$

Статистический тест (нерандомизированный статистический критерий согласия) состоит в следующем:

$$\text{гипотеза } H \text{ принимается на уровне } \varepsilon \quad \Longleftrightarrow \quad \varepsilon^* < \varepsilon.$$

Уровень ε выбирается в зависимости от задачи или оптимизируется по результатам исследования.

Согласно основной гипотезе H, случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ независимы и одинаково распределены с конечной ненулевой дисперсией.

Согласно альтернативной гипотезе $\overline{H}$, случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ также независимы, но $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ имеют разные распределения с разными математическими ожиданиями и с конечными ненулевыми дисперсиями.

Построим тест на разности средних $\overline{X} - \overline{Y}$, где $\overline{X} = \sum_{i=1}^{n_1} X_i/n_1$, $\overline{Y} = \sum_{i=1}^{n_2} Y_i/n_2$.

Согласно основной гипотезе,

$$\mathbf{Var}(\overline{X} - \overline{Y}) = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Теорема (о t-статистике Уэлча)

1) Пусть случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ независимы и одинаково распределены с конечной ненулевой дисперсией σ^2,

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

$$\overline{X^2} = \sum_{i=1}^{n_1} X_i^2/n_1, \quad \overline{Y^2} = \sum_{i=1}^{n_2} Y_i^2/n_2,$$

$$S_X^2 = \overline{X^2} - (\overline{X})^2, \quad S_Y^2 = \overline{Y^2} - (\overline{Y})^2,$$

$$t_{n_1,n_2} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1-1} + \frac{S_Y^2}{n_2-1}}}.$$

Тогда распределение статистики t_{n_1,n_2} слабо сходится к стандартному нормальному закону при $n_1 \to \infty$, $n_2 \to \infty$.

2) Пусть случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ независимы, но $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ имеют разные распределения с разными математическими ожиданиями и с конечными ненулевыми дисперсиями. Тогда $|t_{n_1,n_2}| \to +\infty$ с вероятностью 1 при $n_1 \to \infty$, $n_2 \to \infty$.

Итак, на этой статистике мы сможем построить статистический тест: реально достигаемый уровень значимости

$$\alpha^* = 1 - F_{|\mathcal{N}_{0,1}|}(|t_{n_1,n_2}|)) = 2\left(1 - \Phi\left(|t_{n_1,n_2}|\right)\right).$$

Здесь $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$ — функция распределения стандартного нормального закона.

Пример Пусть $\vec{x} = (0,1,0,1,0,1,0,1,0,1)$, $\vec{y} = (1,2,1,2,1,2,1,2)$.
Здесь $n_1 = 10$, $n_2 = 8$,

$$\overline{X} = 1/2, \quad \overline{Y} = 3/2, \quad \overline{X^2} = 1/2, \quad \overline{Y^2} = 5/2,$$

$$S_X^2 = 1/4, \quad S_Y^2 = 1/4, \quad t_{n_1,n_2} = \frac{-1}{\sqrt{\frac{1}{36} + \frac{1}{28}}} = -\sqrt{63}/2 \approx -4,$$

$\alpha^* \approx 2\left(1 - \Phi\left(4\right)\right)$=2*НОРМСТРАСП(-4)≈0,00006. Достигнутый уровень значимости является очень маленьким, что говорит против гипотезы об однородности.

Другим статистическим тестом однородности является тест Колмогорова—Смирнова. Основная гипотеза предполагает, что распределение элементов выборки непрерывное, но зато альтернативная гипотеза не требует различия математических ожиданий.

Теорема (о статистике Колмогорова—Смирнова)
1) Пусть случайные величины $X_1, X_2, \ldots$ и $Y_1, Y_2, \ldots$ независимы и одинаково распределены с непрерывной функцией распределения,

$$F_{n_1}^*(t) = \sum_{i=1}^{n_1} \mathbf{1}(X_i \le t)/n_1, \quad G_{n_2}^*(t) = \sum_{i=1}^{n_2} \mathbf{1}(Y_i \le t)/n_2,$$

$$D_{n_1,n_2} = \sup_{t \in \mathbf{R}} |F_{n_1}^*(t) - G_{n_2}^*(t)|, \quad d_{n_1,n_2} = \frac{D_{n_1,n_2}}{\sqrt{1/n_1 + 1/n_2}}.$$

Тогда распределение случайной величины d_{n_1,n_2} сходится к распределению Колмогорова с функцией распределения

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}, \quad t > 0.$$

2) Пусть случайные величины $X_1, X_2, \dots$ и $Y_1, Y_2, \dots$ независимы, но $X_1, X_2, \dots$ и $Y_1, Y_2, \dots$ имеют разные распределения. Тогда $d_{n_1,n_2} \to +\infty$ с вероятностью 1 при $n_1 \to \infty$, $n_2 \to \infty$.

На этой теореме основан статистический тест однородности Колмогорова—Смирнова: реально достигаемый уровень значимости

$$\alpha^* = 1 - K(d_{n_1,n_2}) = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2 d_{n_1,n_2}^2}.$$

Пример Пусть $\vec{x} = (0,1,0,1,0,1,0,1,0,2)$, $\vec{y} = (2,3,2,3,2,3,2,3)$.

Здесь $n_1 = 10$, $n_2 = 8$, максимальная разность эмпирических функций распределения достигается в точке 1 и равняется $9/10$:

$$D_{n_1,n_2} = |F_{n_1}^*(1) - G_{n_2}^*(1)| = \left|\frac{9}{10} - \frac{0}{8}\right| = 9/10.$$

Поэтому

$$d_{n_1,n_2} = \frac{9/10}{\sqrt{1/10 + 1/8}} = 3\sqrt{10}/5,$$

$$\alpha^* = 1 - K(3\sqrt{10}/5) = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-36k^2/5} \approx 2(e^{-36/5} - e^{-144/5}) \approx 0{,}00149.$$

Реально достигнутый уровень значимости очень маленький, основная гипотеза об однородности принимается на уровне меньше 0,002.

Задачи для самостоятельного решения

14.1 Проверить однородность выборок $\vec{x} = (1,2,3,4,1,2,3,4)$ и $\vec{y} = (2,3,4,5,2,3,4,5)$ по критериям Уэлча и Колмогорова—Смирнова.

14.2 Проверить однородность выборок $\vec{x} = (0,1,2,3,4,0,1,2,3,0,1,2,0,1,0)$ и $\vec{y} = (4,5,6,7,5,6,7,6,7,7)$ по критериям Уэлча и Колмогорова—Смирнова.

Индивидуальные домашние задания

Каждой букве своих имени, отчества и фамилии сопоставьте ее номер в алфавите в соответствии с таблицей. Среди классиков русской литературы найдите того, у которого распределение букв имени, отчества и фамилии лучше всего соответствует Вашему. Проверьте гипотезу об однородности (Вашей с классиком) по критериям Уэлча и Колмогорова—Смирнова.

а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33

§15. Критерий наличия разладки

Критерий наличия разладки служит для проверки основной гипотезы о том, что распределение не поменялось за весь период наблюдений против альтернативной гипотезы, утверждающей, что одно распределение сменилось другим.

Введем следующие обозначения. Пусть $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин с распределениями $\mathcal{F}_1$ и $\mathcal{F}_2$ соответственно.

Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^{n}$ задается следующим образом:
$X_i^{(n)} = \xi_i^{(1)}$ при $1 \le i \le [nT]$;
$X_i^{(n)} = \xi_i^{(2)}$ при $[nT]+1 \le i \le n$.

Здесь T — неизвестный параметр, $0 < T < 1$. Для краткости мы будем писать X_i вместо $X_i^{(n)}$.

Основная гипотеза H состоит в том, что разладки не произошло, то есть $\mathcal{F}_1 = \mathcal{F}_2$. Ее альтернатива — в том, что не только распределения, но и их математические ожидания различны. Как при нулевой гипотезе, так и при альтернативе мы будем предполагать, что дисперсии распределений $\mathcal{F}_1$ и $\mathcal{F}_2$ конечны и положительны.

Для решения задачи обнаружения разладки строятся статистики, являющиеся функционалами от *эмпирического моста* $Z_n = \{Z_n(t), 0 \le t \le 1\}$ — случайной ломаной, построенной по точкам

$$\left(\frac{k}{n};\ \frac{n\Delta_k - k\Delta_n}{Sn\sqrt{n}}\right), k = 0, \ldots, n,$$

где $\Delta_k = \sum_{i=1}^{k} X_i$, $\overline{X} = \Delta_n/n$, $S^2 = \overline{X^2} - (\overline{X})^2$.

По определению,

$$Z_n(t) = \frac{n\Delta_k - k\Delta_n}{Sn\sqrt{n}} + \frac{nX_{k+1} - \Delta_n}{S\sqrt{n}}\left(t - \frac{k}{n}\right),$$

$$\frac{k}{n} \le t < \frac{k+1}{n}, \quad k = 0, \ldots, n-1.$$

Обозначим

$$J_n = \max_{t \in [0;\ 1]} |Z_n(t)|$$

— максимум модуля эмпирического моста.

Теорема

1) Если $\mathcal{F}_1 = \mathcal{F}_2$, и дисперсия распределения конечная и ненулевая, то распределение J_n сходится при $n \to \infty$ к распределению Колмогорова с функцией распределения

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2t^2}, \quad t > 0.$$

2) Если распределения $\mathcal{F}_1$ и $\mathcal{F}_2$ различны и, более того, различны их математические ожидания, а дисперсии конечны и положительны, то $J_n \to \infty$ при $n \to \infty$ с вероятностью 1.

На этой теореме основан статистический тест отсутствия разладки: реально достигаемый уровень значимости

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

Пример Проверить последовательность $\vec{x} = (3,4,3,4,3,4,3,4,1,2,1,2,1,2,1,2)$ на наличие разладки.

Здесь $n = 16$, $\Delta_0 = 0$, Δ_k — это последовательные накопленные суммы:

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_k	-	3	4	3	4	3	4	3	4	1	2	1	2	1	2	1	2
Δ_k	0	3	7	10	14	17	21	24	28	29	31	32	34	35	37	38	40

Итак, $\Delta_n = 40$. Вычислим $n\Delta_k - k\Delta_n$ для всех k от 0 до 16:

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X_k	-	3	4	3	4	3	4	3	4	1	2	1	2	1	2	1	2
Δ_k	0	3	7	10	14	17	21	24	28	29	31	32	34	35	37	38	40
$n\Delta_k - k\Delta_n$	0	8	32	40	64	72	96	104	128	104	96	72	64	40	32	8	0

Так как $\overline{X} = 5/2$, $\overline{X^2} = 15/2$, то $S^2 = 15/2 - 25/4 = 5/4$, $Sn\sqrt{n} = 32\sqrt{5}$, $J_n = 4\sqrt{5}/5$,

$$\alpha^* = 1 - K(4\sqrt{5}/5) \approx 2(e^{-32/5} - e^{-128/5}) \approx 0{,}00332.$$

Этот достигнутый уровень значимости является довольно маленьким. В частности, гипотеза об отсутствии разладки отвергается на уровне 0,004.

Задачи для самостоятельного решения

15.1 Проверить на отсутствие разладки последовательность $\vec{x} = (1,2,3,4,1,2,3,4)$.

15.2 Проверить на отсутствие разладки последовательность $\vec{x} = (1, 2, 3, 4, 5, 6, 7, 8, 6, 4, 2, 0)$.

Индивидуальные домашние задания

Каждой букве своих имени, отчества и фамилии сопоставьте ее номер в алфавите в соответствии с таблицей. Проверьте гипотезу об отсутствии разладки полученной последовательности чисел.

а	б	в	г	д	е	ё	ж	з	и	й	к	л	м	н	о	п	р
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33

Список литературы

[1] *Аркашов Н.С., Бородихин В.М., Ковалевский А.П.* Высшая математика. — Новосибирск: НГТУ, 2008. — Т. 4.2: Теория вероятностей и математическая статистика. — 228 с.

[2] *Боровков А.А.* Теория вероятностей. — М.: Эдиториал УРСС, 1999. — 470 с.

[3] *Боровков А.А.* Математическая статистика. — Новосибирск: Наука, 1997. — 772 с.

[4] *Бородин А.Н.* Элементарный курс теории вероятностей и математической статистики. — СПб., 1999. — 223 с.

[5] *Бородихин В.М.* Теория вероятностей и математическая статистика: Практикум. — Новосибирск, 2000. — Ч. 1. — 159 с.

[6] *Бородихин В.М.* Теория вероятностей и математическая статистика: Практикум. — Новосибирск, 2001. — Ч. 2. — 105 с.

[7] *Бородихин В.М., Ковалевский А.П.* Высшая математика. — Новосибирск: НГТУ, 2005. — Т. 4.2: Теория вероятностей и математическая статистика. — 256 с.

[8] *Гнеденко Б.В.* Теория вероятностей. — М.: Наука, 1969. — 400 с.

[9] *Емельянов Г.В., Скитович В.П.* Задачник по теории вероятностей и математической статистике. — Л.: Изд-во Ленинградского ун-та, 1967. — 332 с.

[10] *Ермаков С.М., Михайлов Г.А.* Статистическое моделирование. — М.:

Наука, 1982.

[11] *Зубков А.М., Севастьянов Б.А., Чистяков В.П.* Сборник задач по теории вероятностей. — Л.: Наука, 1989. — 320 с.

[12] *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. — М.: Высшая школа, 1984. — 248 с.

[13] *Ивченко Г.И., Медведев Ю.И., Чистяков А.В.* Сборник задач по математической статистике. — М.: Высшая школа, 1989. — 255 с.

[14] *Коршунов Д.А., Фосс С.Г., Эйсымонт И.М.* Сборник задач и упражнений по теории вероятностей. — СПб., 2004. — 192 с.

[15] *Коршунов Д.А., Чернова Н.И.* Сборник задач и упражнений по математической статистике. — Новосибирск, 2001. — 120 с.

[16] *Лотов В.И.* Теория вероятностей и математическая статистика. — Новосибирск: НГУ, 2006. — 128 с.

[17] *Свешников А.А. и др.* Сборник задач по теории вероятностей, математической статистике и теории случайных функций. — М., 1970. — 656 с.

[18] *Чистяков В.П.* Курс теории вероятностей. — М.: Наука, 1987. — 240 с.

[19] *Чернова Н.И.* Теория вероятностей. — Новосибирск: НГУ, 2007. — 160 с.

[20] *Чернова Н.И.* Математическая статистика. — Новосибирск: НГУ, 2007. — 148 с.

Интернет-источники

1. Лотов В.И. Лекции по теории вероятностей и математической статистике.
http://www.nsu.ru/mmf/tvims/lotov/tv&ms_ff.pdf

2. Коршунов Д.А., Фосс С.Г. Сборник задач и упражнений по теории вероятностей.
http://www.math.nsc.ru/LBRT/v1/dima/ExerciseProbability2.pdf

3. Коршунов Д.А., Чернова Н.И. Сборник задач и упражнений по математической статистике.
http://www.math.nsc.ru/LBRT/v1/dima/ExerciseStatistics2.pdf

4. Чернова Н.И. Лекции по теории вероятностей.
http://www.nsu.ru/mmf/tvims/chernova/tv/index.html

5. Чернова Н.И. Лекции по математической статистике.
http://www.nsu.ru/mmf/tvims/chernova/ms/lec/ms.html