

Математическая статистика.

Андрей Тищенко @AndrewTGk

2024/2025

Лекция 10 января

Преамбула

Статистика. Мнения о появлении этого слова:

1. Статистиками в Германии назывались люди, собирающие данные о населении и передающие их государству.
2. В определённый день в Венеции народ выстаивался для выплаты налогов (строго фиксированных, в зависимости от рода действий). Государство собирало данные обо всём населении. Это происходило до появления статистиков в Германии, поэтому мы будем считать, что статистика пошла из Венеции.

Задача статистики — по результатам наблюдений построить вероятностную модель наблюдаемой случайной величины.

Основные определения

Определение

Однородной выборкой объёма n называется случайный вектор $X = (X_1, \dots, X_n)$, компоненты которого являются независимыми и одинаково распределёнными. Элементы вектора X называются элементами выборки.

Определение

Если элементы выборки имеют распределение $F_\xi(x)$, то говорят, что выборка соответствует распределению $F_\xi(x)$ или порождена случайной величиной ξ с распределением $F_\xi(x)$.

Определение

Детерминированный вектор $x = (x_1, \dots, x_n)$, компоненты которого x_i являются реализациями соответствующих случайных величин X_i ($i = \overline{1, n}$), называется реализацией выборки.

Уточнение

Если X — однородная выборка объёма n , то его реализацией будет вектор x , каждый элемент x_i которого является значением соответствующей ему случайной величины (элемента выборки) X_i .

Определение

Выборочным пространством называется множество всех возможных реализаций выборки

$$X = (X_1, \dots, X_n)$$

Пример

У вектора $X = (X_1, \dots, X_{10})$ каждый элемент X_i которой порождён случайной величиной $\xi \sim U(0, 1)$, выборочным пространством является \mathbb{R}^{10} (так как X_i может принять любое значение на \mathbb{R})

Определение

Обозначим $x_{(i)}$ — i -ый по возрастанию элемент, тогда будет справедливо:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Обозначим $X_{(k)}$ случайную величину, реализация которой при каждой реализации x выборки X принимает значение $x_{(k)}$. Тогда последовательность $X_{(1)}, \dots, X_{(n)}$ называется вариационным рядом выборки.

Определение

Случайная величина $X_{(k)}$ называется k -ой порядковой статистикой выборки.

Определение

Случайные величины $X_{(1)}, X_{(n)}$ называются эстремальными порядковыми статистиками.

Определение

Порядковая статистика $X_{([n \cdot p])}$ называется выборочной квантилью уровня p , где $p \in [0, 1]$

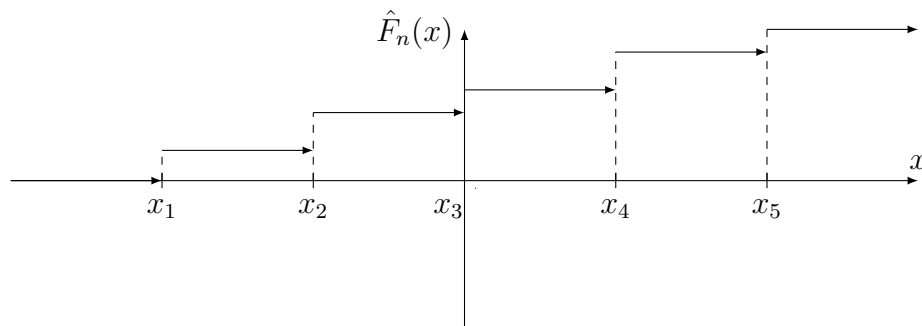
Определение

Пусть каждый элемент выборки X объёма n имеет распределение $F_\xi(x)$. Эмпирической функцией распределения такой выборки называется

$$\hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n I(X_k \leq x)$$

I — индикаторная функция. $I = \begin{cases} 1, & \text{если аргумент верен} \\ 0, & \text{иначе} \end{cases}$

Пусть x_1, \dots, x_n — реализация выборки X_1, \dots, X_n



Свойства $\hat{F}_n(x)$

$$1. \forall x \in \mathbb{R} \quad E\hat{F}_n(x) = E\left(\frac{1}{n} \sum_{k=1}^n I(X_k \leq x)\right) = \frac{1}{n} \sum_{k=1}^n EI(X_k \leq x) = P(X_1 \leq x) = F_\xi(x)$$

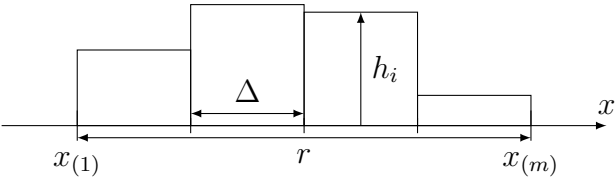
2. По усиленному закону больших чисел (УЗБЧ)

$$\forall x \in \mathbb{R} \quad \hat{F}_n(x) = \frac{1}{n} \sum_{k=1}^n I(X_k \leq x) \xrightarrow[n \rightarrow \infty]{\text{п. н.}} EI(X_k \leq x) = F_\xi(x)$$

Гистограмма

Разбить \mathbb{R} на $(m + 2)$ непересекающихся интервала. Рассматриваются $x_{(1)}, \dots, x_{(m)}$

Название	Обозначение	Формула
Количество интервалов	m	—
Размах выборки	r	$r = x_{(m)} - x_{(1)}$
Ширина интервала	Δ	$\Delta = \frac{r}{m}$
Количество попаданий на i -ый интервал	ν_i	—
Частота попаданий на i -ый интервал	h_i	$h_i = \frac{\nu_i}{\Delta}$



Лекция 17 января

Определение

Пусть $X_1, \dots, X_n \sim F(x, \theta)$. k -ым начальным выборочным моментом называется

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k \in \mathbb{N}$$

Выборочным средним называется:

$$\hat{\mu}_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Определение

k -ым центральным выборочным моментом называется

$$\hat{\nu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 2, 3, \dots$$

$\hat{\nu}_2 = S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ называется выборочной дисперсией

Пусть $(x_1, y_1), \dots, (x_n, y_n)$ соответствует распределению $F(x, y, \theta)$

Определение

Выборочной ковариацией называется

$$\hat{K}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Определение

Выборочным коэффициентом корреляции называется

$$\hat{\rho}_{xy} = \frac{\hat{K}_{xy}}{\sqrt{S_x^2 S_y^2}}$$

Свойства выборочных моментов

1. $E\hat{\mu}_k = E\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n} \sum_{i=1}^n EX_i^k = EX_1^k = \mu_k$
2. $E\bar{X} = m_x$
3. $\mathcal{D}\hat{\mu}_k = \mathcal{D}\left(\frac{1}{n} \sum_{i=1}^n x_i^k\right) = \frac{1}{n^2} \sum_{i=1}^n \mathcal{D}X_i^k = \frac{1}{n} \mathcal{D}X_1^k = \frac{1}{n} (EX_1^{2k} - (EX_1^k)^2) = \frac{1}{n}(\mu_{2k} - \mu_k^2)$
4. $\mathcal{D}\bar{x} = \frac{\sigma_{x_1}^2}{n}$
5. По УЗБЧ

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n x_i^k \xrightarrow[n \rightarrow \infty]{\text{п. н.}} E\hat{\mu}_k = \mu_k$$

$$\hat{\nu}_k \xrightarrow[n \rightarrow \infty]{\text{п. н.}} \nu_k$$

6. По ЦПТ

$$\frac{\hat{\mu}_k - \mu_k}{\sqrt{\frac{\mu_{2k} - \mu_k^2}{n}}} \xrightarrow[n \rightarrow \infty]{d} U, U \sim N(0, 1)$$

$$\frac{\sqrt{n}(\bar{x} - m_{x_1})}{\sigma} \xrightarrow[n \rightarrow \infty]{d} U$$

$$7. ES^2 = E\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right) = \frac{1}{n} E\left(\sum_{i=1}^n (x_i^2 - 2x_i\bar{X} + \bar{X}^2)\right) = E(x^2) - \frac{2}{n} \sum_{i=1}^n E(x_i\bar{X}) + \frac{1}{n} \sum_{i=1}^n E\bar{X}^2 =$$

$$= E(x^2) - \frac{2}{n} \sum_{i=1}^n E x_i \sum_{j=1}^n x_j + \frac{1}{n} \sum_{i=1}^n E\left(\sum_{j=1}^n x_j\right)^2 = E(x^2) - \frac{2}{n} E \sum_{i=1}^n x_i \sum_{j=1}^n x_j + = \frac{n-1}{n} \sigma^2$$

$$8. E\hat{K}_{xy} = \frac{n-1}{n} \text{cov}(x, y)$$

Определение

Оценкой $\hat{\theta}$ параметра θ , называется функция:

$$\hat{\theta} = T(x_1, \dots, x_n), \text{ не зависящая от } \theta$$

Например, отвратительная оценка среднего роста людей в аудитории.

$$\hat{m} = \frac{2x_2 + 5x_5 + 10x_{10}}{3}$$

Определение

Оценка $\hat{\theta}$ называется несмещённой, если $E\hat{\theta} = \theta$ для любых возможных значений этого параметра.

Определение

Оценка $\hat{\theta}(x_1, \dots, x_n)$ называется асимптотически несмещённой оценкой θ , если

$$\lim_{n \rightarrow \infty} E\hat{\theta}(x_1, \dots, x_n) = \theta$$

$$\lim_{n \rightarrow \infty} ES^2 = \lim_{n \rightarrow \infty} \frac{n-1}{n} \sigma^2 = \sigma^2$$

Определение

Несмещённой выборочной (или исправленной) выборочной дисперсией называется

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

Оценки

$$\hat{m}_1 = \frac{x_1 + x_2 + x_3}{3}$$

$$\hat{m}_2 = \frac{\sum_{i=1}^{10} x_i}{10}$$

$$\hat{m}_3 = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Являются несмещёнными.

Определение

Оценка $\hat{\theta}(x_1, \dots, x_n)$ называется:

Состоятельной оценкой θ , если

$$\hat{\theta}(x_1, \dots, x_n) \xrightarrow[n \rightarrow \infty]{p} \theta$$

Сильно состоятельной оценкой, если

$$\hat{\theta}(x_1, \dots, x_n) \xrightarrow[n \rightarrow \infty]{\text{п. н.}} \theta$$

Определение

Пусть $\hat{\theta}$ — несмещённая оценка параметра θ . Если $\mathcal{D}\hat{\theta} \leq \mathcal{D}\theta^*$, где θ^* — любая несмещённая оценка параметра θ . Тогда $\hat{\theta}$ называется эффективной оценкой параметра θ .

R-эффективные оценки

Рассматриваем выборку $X_1, \dots, X_n \sim f(x, \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^1$. Назовём модель $(S, f(x, \theta))$ регулярной, если она удовлетворяет следующим условиям:

1. $\forall x \in S$ функция $f(x, \theta) = f(x_1, \dots, x_n, \theta) > 0$ и дифференцируема по θ .

$$2. \begin{cases} \frac{\delta}{\delta\theta} \int_S f(x, \theta) dx = \int_S \frac{\delta}{\delta\theta} f(x, \theta) dx \\ \frac{\delta}{\delta\theta} \int_S T(x) f(x, \theta) dx = \int_S \frac{\delta}{\delta\theta} T(x) f(x, \theta) dx \end{cases}$$

Пусть $\hat{\theta} = T(x) = T(x_1, \dots, x_n)$ — несмещённая оценка параметра θ :

$$\int_S \frac{\delta}{\delta\theta} f(x, \theta) dx = \frac{\delta}{\delta\theta} \int_S f(x, \theta) dx = \frac{\delta}{\delta\theta} 1 = 0$$

$$\int_S \frac{\delta}{\delta\theta} T(x) f(x, \theta) dx = \frac{\delta}{\delta\theta} \int_S T(x) f(x, \theta) dx = \frac{\delta}{\delta\theta} ET(x) = \frac{\delta}{\delta\theta} \theta = 1$$

Определение

Информацией Фишера о параметре θ , содержащейся в выборке X_1, \dots, X_n называется величина

$$I_n(\theta) = E \left(\frac{\delta \ln(f(x, \theta))}{\delta\theta} \right)^2 = \int_S \left(\frac{\delta \ln(f(x, \theta))}{\delta\theta} \right)^2 f(x, \theta) dx$$

Неравенство Рао-Крамера

Если S , $f(x, \theta)$ — регулярная модель и $\hat{\theta}$ — несмещённая оценка θ , то

$$\mathcal{D}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

Доказательство

Выпишем некоторые равенства (пригодятся в доказательстве):

$$\int_S \frac{\delta}{\delta \theta} f(x, \theta) dx = \int_S \frac{\delta f(x, \theta)}{\delta \theta} \frac{f(x, \theta)}{f(x, \theta)} dx \stackrel{*}{=} \int_S \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = 0$$

Пояснение $\stackrel{*}{=}$. Логарифм — сложная функция. По правилу дифференцирования сложной функции:

$$\frac{\delta \ln f(x, \theta)}{\delta \theta} = \frac{1}{f(x, \theta)} \cdot \frac{\delta f(x, \theta)}{\delta \theta}$$
$$\int_S \frac{\delta}{\delta \theta} T(x) f(x, \theta) dx = \int_S T(x) \frac{\delta}{\delta \theta} f(x, \theta) \frac{f(x, \theta)}{f(x, \theta)} dx = \int_S T(x) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = 1$$

Чуть преобразуем последнее полученное равенство:

$$\int_S T(x) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = \int_S T(x) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx - \underbrace{\theta \int_S \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx}_{=0} =$$
$$= \int_S (T(x) - \theta) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = 1 \Rightarrow 1 = 1^2 = \left(\int_S (T(x) - \theta) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx \right)^2$$

Далее нам понадобится неравенство Коши-Буняковского, которое выглядит так:

$$\left(\int \varphi_1(x) \varphi_2(x) dx \right)^2 \leq \int \varphi_1^2(x) dx \int \varphi_2^2(x) dx$$

Подгоним полученное равенство ($f(x, \theta) > 0 \Rightarrow f(x, \theta) = \sqrt{f(x, \theta)^2}$):

$$\left(\int_S (T(x) - \theta) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx \right)^2 = \left(\int_S \underbrace{(T(x) - \theta) \sqrt{f(x, \theta)}}_{\varphi_1(x)} \cdot \underbrace{\frac{\delta \ln f(x, \theta)}{\delta \theta} \sqrt{f(x, \theta)}}_{\varphi_2(x)} dx \right)^2 = 1$$

И применим неравенство Коши-Буняковского:

$$1 = \left(\int_S \underbrace{(T(x) - \theta) \sqrt{f(x, \theta)}}_{\varphi_1(x)} \cdot \underbrace{\frac{\delta \ln f(x, \theta)}{\delta \theta} \sqrt{f(x, \theta)}}_{\varphi_2(x)} dx \right)^2 \leq$$
$$\leq \int_S \left((T(x) - \theta) \sqrt{f(x, \theta)} \right)^2 dx \cdot \int_S \left(\frac{\delta \ln f(x, \theta)}{\delta \theta} \sqrt{f(x, \theta)} \right)^2 dx =$$
$$= \underbrace{\int_S (T(x) - \theta)^2 f(x, \theta) dx}_{=\mathcal{D}\hat{\theta}} \cdot \underbrace{\int_S \left(\frac{\delta \ln f(x, \theta)}{\delta \theta} \right)^2 f(x, \theta) dx}_{=I_n(\theta)}$$

Получаем:

$$1 \leq \mathcal{D}(\theta) \cdot I_n(\theta) \Rightarrow \mathcal{D}(\theta) \geq \frac{1}{I_n(\theta)}$$

Определение

Оценка $\hat{\theta}$ называется R-эффективной, если $E\hat{\theta} = \theta$ и $\mathcal{D}\hat{\theta} = \frac{1}{I_n(\theta)}$

Лекция 24 января

Замечание 1

$$I_n(\theta) = \mathcal{D} \left(\frac{\delta \ln f(x, \theta)}{\delta \theta} \right)$$

Замечание 2

$$I_n(\theta) = nI_1(\theta)$$

$$f(x, \theta) = f(x_1, \dots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

$$\begin{aligned} E \left(\frac{\delta \ln f(x, \theta)}{\delta \theta} \right)^2 &= E \left(\sum_{i=1}^n \frac{\delta \ln f(x_i, \theta)}{\delta \theta} \right)^2 = \sum_{i \neq j} E \left(\frac{\delta \ln f(x_i, \theta)}{\delta \theta} \cdot \frac{\delta \ln f(x_j, \theta)}{\delta \theta} \right) + n E \left(\frac{\delta \ln f(x_1, \theta)}{\delta \theta} \right)^2 = \\ &= \sum_{i \neq j} \left(\underbrace{E \left(\frac{\delta \ln f(x_i, \theta)}{\delta \theta} \right)}_{=0} \cdot \underbrace{E \left(\frac{\delta \ln f(x_j, \theta)}{\delta \theta} \right)}_{=0} \right) + n E \left(\frac{\delta \ln f(x_1, \theta)}{\delta \theta} \right)^2 = n E \left(\frac{\delta \ln f(x_1, \theta)}{\delta \theta} \right)^2 = n I_1(\theta) \end{aligned}$$

Замечание 3

Пример: $X_1, \dots, X_n \sim N(\theta, \sigma^2)$

Рассмотрим оценку $\hat{\theta} = \bar{X}$, её дисперсия $\mathcal{D}\bar{X} = \frac{\sigma^2}{n}$. Посчитаем информацию Фишера:

$$\begin{aligned} I_1(\theta) &= E \left(\frac{\delta \ln f(x, \theta)}{\delta \theta} \right)^2 = E \left(\frac{\delta}{\delta \theta} \ln \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \right) \right)^2 = E \left(\frac{\delta}{\delta \theta} \ln \left(\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\theta)^2}{2\sigma^2} \right) \right)^2 = E \left(\frac{x-\theta}{\sigma^2} \right)^2 = \\ &= \frac{1}{\sigma^4} E(x-\theta)^2 = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2} \Rightarrow I_n(\theta) = \frac{n}{\sigma^2} \end{aligned}$$

Знаем, что $\mathcal{D}\hat{\theta} \geq \frac{1}{nI_1(\theta)} = \frac{\sigma^2}{n} = \mathcal{D}(\bar{X}) \Rightarrow$ оценка $\hat{\theta} = \bar{X}$ является R-эффективной.

Критерий эффективности $X_1, \dots, X_n \sim F_\xi(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^1$ выполнены условия регулярности, то есть

$$\int T(x) \frac{\delta f(x, \theta)}{\delta \theta} dx = \frac{\delta}{\delta \theta} \int T(x) f(x, \theta) dx = E\hat{\theta}$$

Определение

Функцией вклада выборки X_1, \dots, X_n называется

$$U(x, \theta) = \sum_{i=1}^n \frac{\delta \ln f(x_i, \theta)}{\delta \theta}$$

Пусть $0 < U(x, \theta) < \infty$.

$\hat{\theta} = T(x_1, \dots, x_n)$ — R-эффективная оценка $\theta \Leftrightarrow \hat{\theta} - \theta = a(\theta)U(x, \theta)$, где $a(\theta) = \mathcal{D}\hat{\theta}$

Доказательство \Rightarrow :

Пусть $\hat{\theta} - \theta = a(\theta)U(x, \theta) \Rightarrow \hat{\theta}$ — R-эффективная оценка θ .

Посчитаем математическое ожидание частей равенства:

$$E(\hat{\theta} - \theta) = a(\theta)EU(x, \theta) = a(\theta) \int \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = 0$$

Посчитаем дисперсию частей:

$$\mathcal{D}(\hat{\theta} - \theta) = a^2(\theta)\mathcal{D}U(x, \theta) = \underbrace{a^2(\theta)}_{=(\mathcal{D}(\hat{\theta}))^2} I_n(\theta) \Rightarrow \mathcal{D}(\hat{\theta}) = (\mathcal{D}(\hat{\theta}))^2 I_n(\theta) \Rightarrow 1 = \mathcal{D}(\theta) I_n(\theta)$$

Значит оценка является R-эффективной.

Доказательство \Leftarrow :

Пусть $\hat{\theta}$ — R-эффективная оценка $\Rightarrow \hat{\theta} - \theta = a(\theta)U(x, \theta)$. Хотим доказать, что $\rho(\hat{\theta}, U(x, \theta)) = 1$.

Для подсчёта корреляции нужно посчитать ковариацию:

$$\text{cov}(\hat{\theta}, U(x, \theta)) = E(\hat{\theta} - \theta)U(x, \theta) = E\hat{\theta}U(x, \theta) - \underbrace{\theta EU(x, \theta)}_{=0} =$$

$$= \int_S T(x)U(x, \theta)f(x, \theta) dx = \int_S T(x) \frac{\delta \ln f(x, \theta)}{\delta \theta} f(x, \theta) dx = 1$$

Так как $\hat{\theta}$ — R-эффективная оценка, то $\mathcal{D}\hat{\theta} = \frac{1}{I_n(\theta)}$. Знаем, что $\mathcal{D}U(x, \theta) = I_n(\theta)$, тогда:

$$\rho(\hat{\theta}, U(x, \theta)) = \frac{\text{cov}(\hat{\theta}, U(x, \theta))}{\sqrt{\mathcal{D}\hat{\theta}\mathcal{D}U(x, \theta)}} = \frac{1}{\sqrt{\frac{I_n(\theta)}{I_n(\theta)}}} = 1 \Rightarrow$$

$$\Rightarrow \hat{\theta} = c_1 + c_2 U(x, \theta)$$

$E\hat{\theta} = c_1 + Ec_2 U(x, \theta) = c_1 + 0 = \theta$, так как оценка эффективная

$$\mathcal{D}\hat{\theta} = c_2^2 I_n(\theta) = \frac{1}{I_n(\theta)} \Rightarrow c_2^2 = \frac{1}{I_n^2} \Rightarrow c_2 = \frac{1}{I_n} = \mathcal{D}\hat{\theta} = a(\theta).$$

Итак, $\hat{\theta} = \theta + a(\theta)U(x, \theta) \Rightarrow \hat{\theta} - \theta = U(x, \theta)$.

Метод моментов

$X_1, \dots, X_n \sim F_\xi(x, \theta)$, $\theta \in \Theta \subset R^k$

$$\exists \mu_j < \infty, j = \overline{1, k} \quad \underbrace{\mu_j}_{=\mu_j(\theta)} = E\xi^j = \int_{-\infty}^{+\infty} x^j f(x, \theta) dx = 1$$

Тогда можно получить систему уравнений:

$$\begin{cases} \hat{\mu}_1 = \mu_1(\theta) \\ \vdots \\ \hat{\mu}_k = \mu_k(\theta) \end{cases} \quad (1)$$

Если система уравнений (1) однозначно разрешима относительно $\theta_1, \dots, \theta_k$, то решения $\hat{\theta}_1, \dots, \hat{\theta}_k$ называется равной $\theta_1, \dots, \theta_k$ по методу моментов.

Пример

$X_1, \dots, X_n \sim N(\theta_1, \theta_2^2)$, $\theta = (\theta_1, \theta_2^2)$, тогда:

$$\begin{cases} \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n x_i = \theta_1 \Rightarrow \hat{\theta}_1 = \bar{X} \\ \hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n x_i^2 = \theta_2^2 + \theta_1^2, (E\xi^2 = \mathcal{D}\xi + (E\xi)^2) \Rightarrow \hat{\theta}_2^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{X}^2 \end{cases}$$

Метод максимального правдоподобия (ММП)

Определение

Функцией правдоподобия для X_1, \dots, X_n , порождённых случайной величиной ξ , называется функция

$$L(x_1, \dots, x_n, \theta) = \begin{cases} \prod_{i=1}^n f(x_i, \theta), & \text{если } \xi \text{ — непрерывная случайная величина} \\ \prod_{i=1}^n P(\xi = x_i, \theta), & \text{если } \xi \text{ — дискретная случайная величина} \end{cases}$$

Определение

Реализацией оценки максимального правдоподобия (ОМП) называется значение $\hat{\theta} \in \Theta$, такое что:

$$\hat{\theta} = \operatorname{argmax} L(x_1, \dots, x_n, \theta), \text{ где } \theta \in \Theta$$

Для нахождения точки максимума нужно взять частные производные по всем составляющим θ от функции правдоподобия. Однако считать производную произведения нам впадлу, поэтому мы введём следующую вещь:

Определение

Функция $\ln L(x_1, \dots, x_n, \theta)$ называется логарифмической функцией правдоподобия.

Итак, получаем систему уравнений:

$$\begin{cases} \frac{\delta \ln L(x_1, \dots, x_n, \theta)}{\delta \theta_1} = 0 \\ \vdots \\ \frac{\delta \ln L(x_1, \dots, x_n, \theta)}{\delta \theta_k} = 0 \end{cases}$$

Логарифм монотонный, поэтому его argmax совпадёт с argmax функции $L(x_1, \dots, x_n, \theta)$ (НАУКА!).

Пример

Для Гауссовской величины $N(\theta_1, \theta_2^2)$:

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x_i - \theta_1)^2}{2\theta_2^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\theta_2}\right)^n e^{-\frac{(x - \theta_1)^2}{2\theta_2^2}}$$

Логарифмируем:

$$\ln L(x_1, \dots, x_n, \theta) = \ln \left(\frac{1}{\sqrt{2\pi}}\right)^n - n \ln \theta_2 - \frac{\sum_{i=1}^n (x_i - \theta_1)^2}{2\theta_2^2}$$

Возьмём частные производные:

$$\begin{cases} \frac{\delta \ln L(x_1, \dots, x_n, \theta)}{\delta \theta_1} = \frac{\sum_{i=1}^n (x_i - \hat{\theta}_1)}{\hat{\theta}_2^2} \\ \frac{\delta \ln L(x_1, \dots, x_n, \theta)}{\delta \theta_2} = -\frac{n}{\hat{\theta}_2} + \frac{\sum_{i=1}^n (x_i - \hat{\theta}_1)^2}{\hat{\theta}_2^3} \end{cases}$$

Посчитаем θ_1, θ_2 :

$$\begin{cases} \sum_{i=1}^n (x_i - \hat{\theta}_1) = 0 \Rightarrow \hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{X} \\ -n\hat{\theta}_2^2 + \sum_{i=1}^n (x_i - \bar{x})^2 = 0 \Rightarrow \hat{\theta}_2^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \end{cases}$$

Лекция 31 января.

Робастные оценки

От слова robust.

Определение

Пусть оценка $\hat{\theta}_n$ построена по выборке X_1, \dots, X_n . Затем добавлено наблюдение x и построена оценка $\hat{\theta}_{n+1}$, тогда кривой чувствительности, изучающей влияние наблюдения x на оценку $\hat{\theta}$ называется функция:

$$SC_n(x) = \frac{\hat{\theta}_{n+1} - \hat{\theta}_n}{\frac{1}{n+1}} = (n+1) (\hat{\theta}_{n+1} - \hat{\theta}_n)$$

Определение

Оценка $\hat{\theta}$ называется B -робастной, если $SC_n(x)$ ограничена.

Пример

Пусть $\hat{\theta} = \bar{X}$

$$SC_n(x) = (n+1) \left(\frac{1}{n+1} \left(\sum_{i=1}^n (x_i) + x \right) - \frac{1}{n} \sum_{i=1}^n x_i \right) = \sum_{i=1}^n x_i + x - \left(\sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n x_i \right) = x - \bar{X}$$

Это линейная функция от x , то есть кривая чувствительности неограничена.

Пусть $\hat{\theta} = \hat{\mu}$ (выборочная медиана)

$$\hat{\mu} = \begin{cases} X_{(k+1)}, & n = 2k + 1 \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k \end{cases}$$

Определение

Пороговой точкой (BP) ε_n^* оценки $\hat{\theta}$, построенной на выборке X_1, \dots, X_n называется:

$$\varepsilon_n^* = \frac{1}{n} \max \left\{ m : \max_{i_1, \dots, i_m} \sup_{y_1, \dots, y_m} |\hat{\theta}(z_1, \dots, z_m)| < \infty \right\}$$

Где выборка z_1, \dots, z_m получена заменой значений X_{i_1}, \dots, X_{i_m} на произвольные значения y_1, \dots, y_m

Доверительные интервалы

Определение

Пусть для $X_1, \dots, X_n \sim F(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^1$ построены статистики $T_1(x_1, \dots, x_n)$ и $T_2(x_1, \dots, x_n)$, такие что

$$\begin{cases} T_1(x) < T_2(x) \\ P(T_1(x) < \theta < T_2(x)) = 1 - \alpha, \quad 0 < \alpha < 1 \end{cases}$$

Тогда интервал $(T_1(x), T_2(x))$ называется доверительным интервалом уровня надёжности (доверия) $1 - \alpha$ параметра θ .

Определение

Случайная функция $G(x_1, \dots, x_n, \theta) = G(x, \theta)$ называется центральной (опорной) статистикой, если

1. $G(x, \theta)$ непрерывна и монотонна по θ
2. $F_G(x)$ не зависит от θ

Односторонние доверительные интервалы:

$$P(G(x, \theta) < Z_{1-\alpha}) = 1 - \alpha$$

$$P(Z_\alpha < G(x, \theta)) = 1 - \alpha$$

Квантили не зависят от θ , с их помощью можно выразить односторонние доверительные интервалы.

Центральным доверительным интервалом будет:

$$P(Z_{\frac{\alpha}{2}} < G(x, \theta) < Z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

Определение

Пусть случайные величины $\xi_1, \dots, \xi_m \sim N(0, 1)$ и независимы.

Тогда случайная величина $\eta = \sum_{i=1}^m \xi_i^2 \sim \chi^2(m)$ (удовлетворяет распределению хи-квадрат (χ^2) с m степенями свободы).

Определение

Пусть $\xi_0, \xi_1, \dots, \xi_m \sim N(0, 1)$ и независимы.

Тогда случайная величина $\zeta = \frac{\xi_0}{\sqrt{\frac{1}{m} \sum_{i=1}^m \xi_i^2}} \sim t(m)$ (распределение Стьюдента с m степенями свободы)

Определение

Пусть случайная величина $\xi_1 \sim \chi^2(m)$, $\xi_2 \sim \chi^2(n)$ и ξ_1 и ξ_2 — независимы. Тогда случайная величина $F = \frac{\frac{1}{m}\xi_1}{\frac{1}{n}\xi_2} \sim F(m, n)$ (распределение Фишера со степенями свободы n, m)

Теорема Фишера

Пусть X_1, \dots, X_n порождены случайной величиной $X \sim N(m, \sigma^2)$, тогда:

1. $\frac{nS^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 \sim \chi^2(n-1)$ (так как мы знаем \bar{X} , и все наблюдения, а по $n-1$ наблюдению и \bar{X} можно восстановить последнее наблюдение)
2. \bar{X} и S^2 — независимые случайные величины.

Пример 1

$X_1, \dots, X_n \sim N(\theta, \sigma^2)$, σ^2 — известно. Построить доверительный интервал для θ

$$\hat{\theta} = \bar{X} \sim N(\theta, \frac{\sigma^2}{n})$$

$$\frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} \sim N(0, 1)$$

$$P\left(Z_{\frac{\alpha}{2}} < \frac{\sqrt{n}(\bar{X} - \theta)}{\sigma} < Z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

Поскольку по середине стоит стандартное гауссовское распределение: $Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$

$$P\left(-\frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} - \bar{X} < -\theta < \frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} - \bar{X}\right) = 1 - \alpha$$

$$P\left(\bar{X} - \frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}} < \theta < \bar{X} + \frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Итак, доверительный интервал: $\left(\bar{X} - \frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{X} + \frac{Z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$

Пример 2

$X_1, \dots, X_n \sim N(m, \theta_2^2)$. Построить доверительный интервал для θ_2^2

$$\sum_{i=1}^n \left(\frac{x_i - m}{\theta_2} \right)^2 \sim \chi^2(n)$$

$$P\left(\chi_{n, \frac{\alpha}{2}}^2 < \frac{\sum_{i=1}^n (x_i - m)^2}{\theta_2^2} < \chi_{n, 1-\frac{\alpha}{2}}^2\right) = 1 - \alpha$$

$$P\left(\frac{\sum_{i=1}^n (x_i - m)^2}{\chi_{n, 1-\frac{\alpha}{2}}^2} < \theta_2^2 < \frac{\sum_{i=1}^n (x_i - m)^2}{\chi_{n, \frac{\alpha}{2}}^2}\right) = 1 - \alpha$$

Здесь $\chi_{n, \alpha}^2$ — квантиль уровня α распределения $\chi^2(n)$

Пример 3

Если нам неизвестны оба параметра $N(\theta_1, \theta_2^2)$. Заменяем m на \bar{X} : Доверительный интервал для θ_2 :

$$P\left(\frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\chi_{n, 1-\frac{\alpha}{2}}^2} < \theta_2^2 < \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\chi_{n, \frac{\alpha}{2}}^2}\right) = 1 - \alpha$$

Доверительный интервал для θ_1 :

$$\frac{\sqrt{n} \left(\frac{\bar{X} - \theta}{\sigma} \right)}{\sqrt{\frac{1}{n-1} \sum \left(\frac{x_i - \bar{X}}{\sigma} \right)^2}} = \frac{\sqrt{n}(\bar{X} - \theta_1)}{\tilde{S}} \sim t(n-1)$$

Обозначим $t_{n, \alpha}$ квантиль уровня α распределения $t(n)$, заметим, что $t_{n, 1-\alpha} = t_{n, 1-\frac{\alpha}{2}}$

$$P(t_{n, 1-\frac{\alpha}{2}} < \frac{\sqrt{n}(\bar{X} - \theta_1)}{\tilde{S}} < t_{n, \frac{\alpha}{2}}) = 1 - \alpha$$

$$P\left(\bar{X} - \frac{\tilde{S} \cdot t_{n, 1-\frac{\alpha}{2}}}{\sqrt{n}} < \theta_1 < \bar{X} + \frac{\tilde{S} \cdot t_{n, 1-\frac{\alpha}{2}}}{\sqrt{n}}\right) = 1 - \alpha$$

Лекция 7 февраля

Задача

$X_1, \dots, X_{n_1} \sim N(m_1, \sigma_1^2)$ и $Y_1, \dots, Y_{n_2} \sim N(m_2, \sigma_2^2)$. σ известны, m — неизвестны.
 X_1, \dots, X_n и Y_1, \dots, Y_n независимы. Доверительный интервал для $\theta = m_1 - m_2$

$$T(x, y) = \frac{\bar{X} - \bar{Y} - \theta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Задача

Пусть $X_1, \dots, X_{n_1} \sim N(m_1, \sigma^2)$, $Y_1, \dots, Y_{n_2} \sim N(m_2, \sigma^2)$. σ неизвестна. Выборки независимы.

Утверждение

$$\frac{\sum_{i=1}^{n_1} (x_i - \bar{X})^2}{\sum_{i=1}^{n_2} (y_i - \bar{Y})^2} \sim F(n_1 - 1, n_2 - 1)$$
$$\frac{\bar{X} - \bar{Y} - (m_1 - m_2)}{\sqrt{\hat{\mathcal{D}}(\bar{X} - \bar{Y})}}$$

Посчитаем дисперсию в знаменателе:

$$\mathcal{D}(\bar{X} - \bar{Y}) = \mathcal{D}\bar{X} + \mathcal{D}\bar{Y} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$
$$S^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{X})^2 + \sum_{i=1}^{n_2} (y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Тогда

$$\frac{\bar{X} - \bar{Y} - (m_1 - m_2)}{\sqrt{\hat{\mathcal{D}}(\bar{X} - \bar{Y})}} = \frac{\bar{X} - \bar{Y} - (m_1 - m_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Теперь можно построить доверительный интервал:

$$P \left(-t_{1-\alpha/2, n_1+n_2-2} < \frac{\bar{X} - \bar{Y} - \theta}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\alpha/2, n_1+n_2-2} \right) = 1 - \alpha$$

$$P \left(-t_{1-\alpha/2, n_1+n_2-2} \cdot S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} - (\bar{X} - \bar{Y}) < -\theta < t_{1-\alpha/2, n_1+n_2-2} \cdot S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} - (\bar{X} - \bar{Y}) \right) = 1 - \alpha$$

$$P \left((\bar{X} - \bar{Y}) - t_{1-\alpha/2, n_1+n_2-2} \cdot S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \theta < t_{1-\alpha/2, n_1+n_2-2} \cdot S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} + (\bar{X} - \bar{Y}) \right) = 1 - \alpha$$

Асимптотические доверительные интервалы

Пусть $X_1, \dots, X_n \sim F(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^1$

$\hat{\theta}$ — состоятельная оценка θ .

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} U, U \sim N(0, \sigma^2(\theta))$$

И $\sigma^2(\theta)$ непрерывна по θ .

$$P \left(Z_{\alpha/2} < \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\hat{\theta}_n)} < Z_{1-\alpha/2} \right) \rightarrow 1 - \alpha$$

$$P \left(\hat{\theta}_n - \frac{\sigma(\hat{\theta}_n)Z_{1-\alpha/2}}{\sqrt{n}} < \theta < \frac{\sigma(\hat{\theta}_n)Z_{1-\alpha/2}}{\sqrt{n}} + \hat{\theta}_n \right)$$

Если \exists R-эффективная оценка $\hat{\theta}_n$, то выбирая её $\mathcal{D}\hat{\theta}_n = \frac{1}{I_n(\theta)}$, тогда $\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}} = \sqrt{\mathcal{D}\hat{\theta}_n} = \frac{1}{\sqrt{nI_1(\hat{\theta}_n)}}$

$$P \left(\hat{\theta}_n - \frac{Z_{1-\alpha/2}}{\sqrt{nI_1(\hat{\theta}_n)}} < \theta < \hat{\theta}_n + \frac{Z_{1-\alpha/2}}{\sqrt{nI_1(\hat{\theta}_n)}} \right) \rightarrow 1 - \alpha$$

Пример

$X_1, \dots, X_n \sim Bi(1, \theta)$

АДИ для θ :

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} \text{ — несмещённая, состоятельная, R-эффективная}$$

$\mathcal{D}x_i = \theta(1 - \theta)$.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n \rightarrow \infty]{d} U, U \sim N(0, \theta(1 - \theta))$$

$$P \left(\hat{\theta} - Z_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}} < \theta < \hat{\theta} + Z_{1-\alpha/2} \frac{\sqrt{\hat{\theta}(1 - \hat{\theta})}}{\sqrt{n}} \right) \rightarrow 1 - \alpha$$

Определение

Основная (или нулевая) гипотеза H_0 , с ней конкурируют H_1, H_2, \dots, H_A (альтернативные гипотезы).

Определение

Сложной гипотезой называют гипотезу, которая не определяет параметры распределения или само распределение однозначно.

Например

$$H_1 : \xi \sim N(m, \sigma^2)$$

$$H_2 : \xi \sim N(5, \sigma^2)$$

Простая гипотеза определяет распределение однозначно, например:

$$H_3 : \xi \sim N(5, 36)$$

Односторонние гипотезы выглядят так:

$$H_4 : \xi m < 5$$

$$H_5 : \xi m > 5$$

Двусторонние:

$$H_6 : n \neq 5$$

$$H_7 : m \in [1, 3]$$

А гипотеза $H_8 : \{ \text{“Сегодня хорошая погода”} \}$ не является статистической, ведь не относится к распределению и параметрам.

Определение

Статистическим критерием называют правило, руководствуясь которым, на основании реализации x_1, \dots, x_n выборки X_1, \dots, X_n принимается решение о справедливости/несправедливости гипотезы H_0 .

Делим множество реализаций выборки S на два множества S_0, S_1 , такие что

$$S_0 \cdot S_1 = \emptyset$$

$$S_0 + S_1 = S$$

Назовём S_0 доверительной областью, а S_1 — критической областью. Если реализация попала в S_0 , то мы принимаем H_0 , иначе принимает альтернативную гипотезу.

Тогда ошибкой первого рода (уровнем значимости критерия) называется

$$P(X \in S_1 \mid \text{верна } H_0) = \alpha$$

Ошибкой второго рода называется

$$P(X \in S_0 \wedge \text{верна } H_1) = 1 - \beta$$

Определение

Пусть критерий предназначен для проверки $H_0 : \theta = \theta_0$ против альтернативы $H_1 : \theta \neq \theta_0$, тогда функцией мощности критерия называется

$$\beta(\theta) = P(X \in S_1, \theta)$$

Критерий называется состоятельным, если при отдалении от θ_0 его функция мощности стремится к 1.

Лекция 13 февраля

Проверка статистических гипотез

Если β — функция мощности критерия проверки гипотезы $H_0 : \theta = \theta_0$, тогда $\beta(\theta) = P(X \in S_1, \theta)$ и $\beta(\theta_0) = \alpha$, где α — вероятность ошибки первого рода.

Задача

$H_0 : \theta = \theta_0$ и $H_1 : \theta \in \Theta_1, \theta_0 \notin \Theta_1$. Пусть зафиксировано $\alpha > 0$, тогда критерий называется несмещённым, если:

$$\beta(\theta) \leq \alpha, \text{ если } \theta = \theta_0$$

$$\beta(\theta) > \alpha, \text{ если } \theta \in \Theta_1$$

Определение

Критерий, предназначенный для проверки $H_0 : \theta = \theta_0$ против $H_1 : \theta \in \Theta_1$ называется состоятельным, если

$$\forall \theta \in \Theta_1 \quad \beta(\theta) \xrightarrow{n \rightarrow \infty} 1, \text{ где } n — \text{ количество испытаний}$$

Определение

Критерий β_0 называется равномерно наиболее мощным, если среди всех критериев β :

$$\forall \theta \in \Theta \quad \beta_0(\theta) \geq \beta(\theta)$$

Локально наиболее мощным, если

$$\forall \theta \in \Theta_1 \subseteq \Theta \quad \beta_0(\theta) \geq \beta(\theta)$$

Алгоритм проверки параметрических гипотез

1. Сформулировать проверяемую гипотезу H_0 и альтернативную к ней H_1 .
2. Выбрать уровень значимости α
3. Выбрать статистику T для проверки гипотезы H_0
4. Найти распределение $F(z | H_0)$ статистики T , при условии $\{“H_0 \text{ верна}”\}$
5. Построить, в зависимости от формулировки гипотезы H_1 и уровня значимости α , критическую область \bar{G}
6. Получить реализацию выборки наблюдений x_1, \dots, x_n и вычислить реализацию $t = \varphi(x_1, \dots, x_n)$ статистики T критерия
7. Принять статистическое решение на уровне доверия $1 - \alpha$: если $t \in \bar{G}$, то отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений, а если $t \in G$, то принять гипотезу H_0 как не противоречащую результатам наблюдений.

Задача

Дамы оценивают чай. Могут ли из двух чашек выбрать чашку с хорошим чаем?

Проводятся наблюдения $X_1, \dots, X_n \sim Bi(1, p)$

1. $H_0 : p = p_0 = 0.5$, $H_1 : p > 0.5$. То есть H_0 — дамы не могут выбрать (просто пытаются угадать).
2. $\alpha = 0.05$. Так как специально указано не было, берём стандартное значение.
3. $T(x) = \sum_{i=1}^n x_i$
4. $T(x | H_0) \sim Bi(n, \frac{1}{2})$. Если n велико:

$$\frac{T(x) - np_0}{\sqrt{np_0(1-p_0)}} = \tilde{T}(x) \sim N(0, 1)$$

5. Доверительная область: $[0, Z_{0,95}] = [0, 1.65]$. Критическая область: $(1.65, +\infty)$
6. Пусть у нас есть данные $n = 30$, $\sum_{i=1}^{30} x_i = 20 = T(x)$

$$\tilde{T}(x) = \frac{20 - 30 \cdot \frac{1}{2}}{\sqrt{30 \cdot 0.5 \cdot 0.5}} \approx 1.82574$$

7. Попали в критическую область, значит принимаем H_1 на уровне доверия $1 - \alpha = 0.95$

Задача

А если у нас есть две серии различных испытаний Бернулли?

Пусть $\xi_1 \sim Bi(n_1, p_1)$ и $\xi_2 \sim Bi(n_2, p_2)$. Хотим проверить $H_0 : p_1 = p_2$ против альтернатив $H_1 : p_1 < p_2$, $H_2 : p_1 > p_2$, $H_3 : p_1 \neq p_2$.

Введём обозначение $\hat{p}_1 = \frac{\sum_{i=1}^{n_1} x_{i1}}{n_1}$, $\hat{p}_2 = \frac{\sum_{i=1}^{n_2} x_{i2}}{n_2}$, тогда:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\mathcal{D}(\hat{p}_1 - \hat{p}_2)}} \sim N(0, 1)$$

$$\text{Посчитаем } \mathcal{D}(\hat{p}_1 - \hat{p}_2) = \mathcal{D}(\hat{p}_1) + \mathcal{D}(\hat{p}_2) - \underbrace{2\text{cov}(\hat{p}_1, \hat{p}_2)}_{=0} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} = pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right).$$

Оценим p :

$$\hat{p} = \frac{\sum_{i=1}^{n_1} x_{i1} + \sum_{i=1}^{n_2} x_{i2}}{n_1 + n_2}$$

Тогда $\tilde{T}(x) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$. По этой статистике уже можем принимать решения.

Лекция 21 февраля

Лемма Неймана-Пирсона

Пусть $X_1, \dots, X_n \sim f(x, \theta)$, параметр θ неизвестен. Проверяется простая гипотеза $H_0 : \theta = \theta_0$ против простой альтернативной гипотезы $H_1 : \theta = \theta_1$ (БОО $\theta_1 > \theta_0$).

Существует наиболее мощный критерий для проверки H_0 против H_1 с критической областью $S_{1\alpha}^* = \{(x_1, \dots, x_n) \mid T(x_1, \dots, x_n) \geq c_\alpha\}$, где $T(x_1, \dots, x_n) = \frac{L(x_1, \dots, x_n, \theta_1)}{L(x_1, \dots, x_n, \theta_0)} = \frac{\prod_{i=1}^n f(x_i, \theta_1)}{\prod_{i=1}^n f(x_i, \theta_0)}$, а c_α такое что $P_{\theta_0}(T(x) \geq c_\alpha) = \alpha$

Доказательство

Пусть есть критерий с критической областью $S_{1\alpha}$ лучше (более мощный) предложенного нашей леммой. Тогда (под x далее понимается вектор (x_1, \dots, x_n)):

$$\beta(\theta_1, S_{1\alpha}) = \int \prod_{i=1}^n f(x_i, \theta_1) dx_1 \dots dx_n = \int_{S_{1\alpha}} L(x, \theta_1) dx = \int_{S_{1\alpha} S_{1\alpha}^*} L(x, \theta_1) dx + \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} L(x, \theta_1) dx =$$

По определению функции $T(x)$:

$$T(x) = \frac{L(x, \theta_1)}{L(x, \theta_0)} \Rightarrow T(x)L(x, \theta_0) = L(x, \theta_1)$$

Подставим это в сумму:

$$= \int_{S_{1\alpha} S_{1\alpha}^*} T(x)L(x, \theta_0) dx + \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} T(x)L(x, \theta_0) dx$$

По определению $\beta(\theta, S_1) = P(X \in S_1, \theta)$ то есть правдоподобие попадания случайной величины в критическую область при заданном параметре.

$$\begin{aligned} \beta(\theta_1, S_{1\alpha}^*) &= \int_{S_{1\alpha}^*} L(x, \theta_1) dx = \int_{S_{1\alpha} S_{1\alpha}^*} L(x, \theta_1) dx + \int_{\bar{S}_{1\alpha} S_{1\alpha}^*} L(x, \theta_1) dx = \\ &= \int_{S_{1\alpha} S_{1\alpha}^*} T(x)L(x, \theta_0) dx + \int_{\bar{S}_{1\alpha} S_{1\alpha}^*} T(x)L(x, \theta_0) dx \end{aligned}$$

Чуток пошаманим с выведенными формулами:

$$\begin{aligned} \beta(\theta_1, S_{1\alpha}) - \beta(\theta_1, S_{1\alpha}^*) &= \left(\int_{S_{1\alpha} S_{1\alpha}^*} T(x) L(x, \theta_0) dx + \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} T(x) L(x, \theta_0) dx \right) - \\ &\quad - \left(\int_{S_{1\alpha} S_{1\alpha}^*} T(x) L(x, \theta_0) dx + \int_{\bar{S}_{1\alpha} S_{1\alpha}^*} T(x) L(x, \theta_0) dx \right) \Rightarrow \\ \Rightarrow \beta(\theta_1, S_{1\alpha}) - \beta(\theta_1, S_{1\alpha}^*) &= \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} \underbrace{T(x)}_{< c_\alpha} L(x, \theta_0) dx - \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} \underbrace{T(x)}_{\geq c_\alpha} L(x, \theta_0) dx \end{aligned}$$

Теперь можно составить равенство:

$$\beta(\theta_1, S_{1\alpha}) = \beta(\theta_1, S_{1\alpha}^*) + \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} \underbrace{T(x)}_{< c_\alpha} L(x, \theta_0) dx - \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} \underbrace{T(x)}_{\geq c_\alpha} L(x, \theta_0) dx$$

Правый интеграл содержит область $S_{1\alpha}^*$, по заданию это множество таких точек, в которых $T(x) \geq c_\alpha$. Левый интеграл, наоборот, содержит $\bar{S}_{1\alpha}^*$, то есть все точки, в которых $T(x) < c_\alpha$.

Значит будет справедливо неравенство:

$$\beta(\theta_1, S_{1\alpha}) < \beta(\theta_1, S_{1\alpha}^*) + c_\alpha \left(\int_{S_{1\alpha} \bar{S}_{1\alpha}^*} L(x, \theta_0) dx - \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} L(x, \theta_0) dx \right)$$

Вероятность попадания в критическую область должна быть равна α , тогда верно:

$$\alpha = \int_{S_{1\alpha}} L(x, \theta_0) dx = \int_{S_{1\alpha}^*} L(x, \theta_0) dx$$

При этом

$$\begin{aligned} \int_{S_{1\alpha}} L(x, \theta_0) dx &= \int_{S_{1\alpha} S_{1\alpha}^*} L(x, \theta_0) dx + \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} L(x, \theta_0) dx \\ \int_{S_{1\alpha}^*} L(x, \theta_0) dx &= \int_{S_{1\alpha}^* S_{1\alpha}} L(x, \theta_0) dx + \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} L(x, \theta_0) dx \\ \int_{S_{1\alpha}} L(x, \theta_0) dx - \int_{S_{1\alpha}^*} L(x, \theta_0) dx &= \int_{S_{1\alpha} \bar{S}_{1\alpha}^*} L(x, \theta_0) dx - \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} L(x, \theta_0) dx = \alpha - \alpha = 0 \end{aligned}$$

Тогда в ранее записанном неравенстве:

$$\begin{aligned} c_\alpha \left(\int_{S_{1\alpha} \bar{S}_{1\alpha}^*} L(x, \theta_0) dx - \int_{S_{1\alpha}^* \bar{S}_{1\alpha}} L(x, \theta_0) dx \right) &= 0 \Rightarrow \\ \Rightarrow \beta(\theta_1, S_{1\alpha}) &< \beta(\theta_1, S_{1\alpha}^*) + 0 \Rightarrow \beta(\theta_1, S_{1\alpha}) < \beta(\theta_1, S_{1\alpha}^*) \end{aligned}$$

То есть всякая критическая область, отличная от $S_{1\alpha}^*$, будет менее мощной.

Задача

$X_1, \dots, X_n \sim N(m, \sigma^2)$, дисперсия известна. Построить наиболее мощный критерий для проверки $H_0 : m = m_0$ против $H_1 : m = m_1 > m_0$

Решение (моё)

По лемме Неймана-Пирсона критическая область необходимого нам критерия должна выглядеть так:

$$S_{1-\alpha}^* = \{(x_1, \dots, x_n) \mid T(x) \geq c_\alpha\}, \quad T(x) = \frac{L(x, m_1)}{L(x, m_0)} \geq c_\alpha, \quad P_{m_0}(T(x) \geq c_\alpha) = \alpha$$

$$L(x, m_1) = \prod_{i=1}^n f(x_i, m_1) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i - m_1)^2}{2\sigma^2}}$$

$$L(x, m_0) = \prod_{i=1}^n f(x_i, m_0) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^n \frac{(x_i - m_0)^2}{2\sigma^2}}$$

$$\frac{L(x, m_1)}{L(x, m_0)} = e^{\sum_{i=1}^n \frac{(x_i - m_0)^2 - (x_i - m_1)^2}{2\sigma^2}} = e^{\sum_{i=1}^n \frac{(m_0 - m_1)(2x_i - m_1 - m_0)}{2\sigma^2}}$$

Хотим найти такое c_α , что $P(T(x) \geq c_\alpha) = \alpha$, то есть хотим найти:

$$F_{T(x)}(c_\alpha) = \alpha \Rightarrow \int_{-\infty}^{c_\alpha} e^{\sum_{i=1}^n \frac{(m_0 - m_1)(2x_i - m_1 - m_0)}{2\sigma^2}} dx = \alpha$$

Ответ с лекции

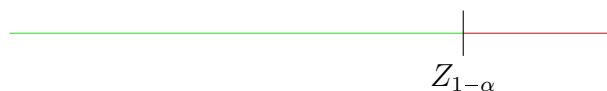
$$S_{1-\alpha}^* \{(x_1, \dots, x_n) \mid \bar{X} \geq m_0 + \frac{Z_{1-\alpha}\sqrt{n}}{\sigma}\} = \{(x_1, \dots, x_n) \mid \frac{(\bar{X} - m_0)\sqrt{n}}{\sigma} \geq Z_{1-\alpha}\}$$

Задача

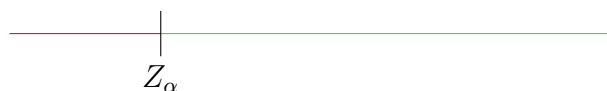
Для проверки гипотезы $H_0 : m = m_0$

$$T(x) = \frac{(\bar{X} - m_0)\sqrt{n}}{\sigma} \Rightarrow T(x)|_{H_0: m=m_0} \sim N(0, 1)$$

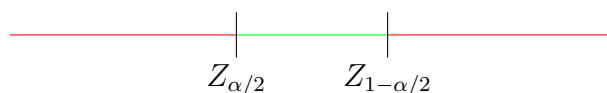
Против гипотезы $H_1 : m > m_0$



Против гипотезы $H_2 : m < m_0$



Против гипотезы $H_3 : m \neq m_0$



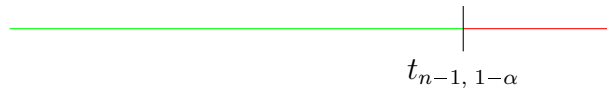
Пояснение: на рисунках зелёным обозначена доверительная область, красным обозначена критическая область.

Задача

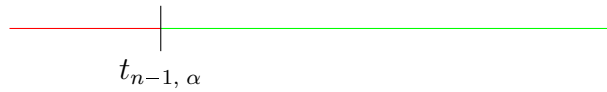
Снова гауссовская выборка, но дисперсия неизвестна. Хотим проверить гипотезу $H_0 : m = m_0$. Тогда нужно поменять статистику на:

$$T(x) = \frac{(\bar{X} - m_0)\sqrt{n}}{\tilde{S}} = \frac{(\bar{X} - m_0)\sqrt{n-1}}{S}$$
$$T(x)|_{H_0} \sim t(n-1)$$

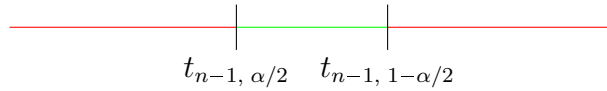
Против гипотезы $H_1 : m > m_0$



Против гипотезы $H_2 : m < m_0$



Против гипотезы $H_3 : m \neq m_0$



Та же самая идея, только разделение идёт по квантилям распределения Стьюдента.

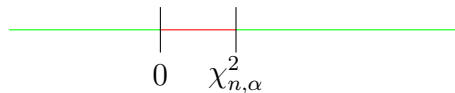
Задача

Теперь строим критерий для оценки дисперсии при известном математическом ожидании. Проверяем гипотезу $H_0 : \sigma = \sigma_0$:

$$T(x) = \frac{\sum_{i=1}^n (x_i - m)^2}{\sigma_0^2}$$

$$T(x)|_{H_0: \sigma = \sigma_0} \sim \chi^2(n)$$

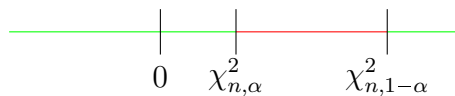
Против гипотезы $H_1 : \sigma < \sigma_0$



Против гипотезы $H_2 : \sigma > \sigma_0$



Против гипотезы $H_3 : \sigma \neq \sigma_0$



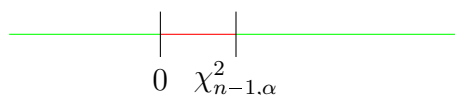
Задача

Если математическое ожидание неизвестно:

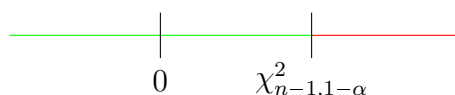
$$T(x) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{\sigma_0^2}$$

$$T(x)|_{H_0: \sigma = \sigma_0} \sim \chi^2(n-1)$$

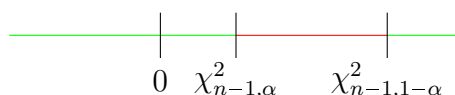
Против гипотезы $H_1 : \sigma < \sigma_0$



Против гипотезы $H_2 : \sigma > \sigma_0$



Против гипотезы $H_3 : \sigma \neq \sigma_0$



Проверка гипотез о распределении случайных величин

Критерий Колмогорова (КАКОЙ ЖЕ ОН КРУТОЙ)

$X_1, \dots, X_n \sim F_\xi(x, \theta_0) = F_0(x)$, θ_0 известна. Проверяем гипотезу $H_0 : \xi \sim F_0(x)$

Колмогоров предложил считать $D_n = \max_{1 \leq i \leq n} |\hat{F}_n(x_i) - F_0(x_i)|$.

Если $n \rightarrow \infty$ (начиная с 20 уже хорошая аппроксимация) и при условии верности H_0 получаем

$$\sqrt{n}D_n \sim K(t)$$

Функция распределения Колмогорова

$$K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j \exp\{-j^2 t^2\}$$

Критерий хи-квадрат

$X_1, \dots, X_n \sim F_\xi(x, \theta_0) = F_0(x)$, θ_0 знаем. Проверяем гипотезу $H_0 : \xi \sim F_0(x)$.

Делим \mathbb{R}^1 на $l+2$ интервала, где $S_0 = -\infty$, $S_{l+1} = +\infty$ тогда $\hat{p}_k = \frac{n_k}{n}$, $p_k^{(0)} = F_0(S_k) - F_0(S_{k-1})$, где $k = \overline{1, l+1}$.

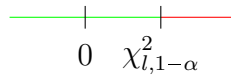
Здесь возникает $\hat{\chi}^2 = \sum_{k=1}^{l+1} \frac{n}{p_k^{(0)}} \left(\hat{p}_k - p_k^{(0)} \right)^2$

Утверждение

Если $0 < p_k^{(0)} < 1$ для $\forall k = \overline{1, l+1}$, $n \rightarrow \infty$ и справедлива H_0 , то

$$\hat{\chi}^2 \sim \chi^2(l)$$

Тогда график будет выглядеть так



Задача

Проверяем теории из биологии

	$p_k^{(0)}$	n_k	$\hat{p}_k = \frac{n_k}{n}$
AB	$\frac{9}{16}$	315	0.556
Ab	$\frac{3}{16}$	108	0.194
aB	$\frac{3}{16}$	101	0.182
ab	$\frac{1}{16}$	32	0.058

Теперь проверим $H_0 : \vec{p}^{(0)} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right)$

Применяя критерий хи-квадрат:

$$\hat{\chi}^2 = 0.49, \text{ посчитали за кадром}$$

$$\hat{\chi}^2|_{H_0} \sim \chi^2(3)$$

Тогда при параметрах $\alpha = 0.05$, $\chi_{3, 0.95}^2 = 7.81 \Rightarrow$ наш результат лежит в доверительной области.

Лекция 28 февраля

Критерий хи-квадрат Пирсона

Имеется выборка $X_1, \dots, X_n \sim F_\xi(x, \theta)$, $\theta \in \Theta \subset \mathbb{R}^n$.

Проверяем гипотезу $H_0 : \xi \sim F_\xi^0(x, \theta)$ (здесь использован верхний индекс для указания на какое-то конкретное распределение).

1. Оценим вектор параметров $\theta = (\theta_1, \dots, \theta_m)$ по методу максимального правдоподобия.
2. Разбиваем \mathbb{R}^1 на $(l + 1)$ непересекающийся интервал.



3. Введём следующие обозначения:

$$\forall k \in [1, l-1] \cap \mathbb{Z} \quad \hat{p}_k = \frac{n_k}{n}$$

$$\forall k \in [0, l] \cap \mathbb{Z} \quad p_k^{(0)}(\hat{\theta}) = P_{H_0}(\xi \in \Delta_k), \text{ (вероятность } \xi \text{ попасть в } k\text{-ый интервал при условии } H_0)$$

$$p_k^{(0)}(\hat{\theta}) = F(s_{k+1}, \hat{\theta}) - F(s_k, \hat{\theta})$$

Тогда справедливо

$$\hat{\chi}^2 = \sum_{k=0}^l \frac{n}{p_k^{(0)}(\hat{\theta})} \left(\hat{p}_k - p_k^{(0)}(\hat{\theta}) \right)^2 = np_0^{(0)}(\hat{\theta}) + \sum_{k=1}^{l-1} \frac{n}{p_k^{(0)}(\hat{\theta})} \left(\hat{p}_k - p_k^{(0)}(\hat{\theta}) \right)^2 + np_l^{(0)}(\hat{\theta})$$

Утверждение

При $n \rightarrow \infty, p_k^{(0)} > 0, \sum_{k=0}^l p_k^{(0)} = 1$ и соблюдении некоторых условий регулярности (про дифференцируемость и существование вторых производных) выполняется

$$\hat{\chi}^2|_{H_0} \sim \chi^2(l+1-1-m)$$

Здесь $l+1$ — количество интервалов, а m — количество оцененных параметров. Доверительным интервалом будет $(0, \chi_{1-\alpha, l-m}^2)$

Определение

Выборки $X_1, \dots, X_m \sim F_x(t)$ и $Y_1, \dots, Y_m \sim F_y(t)$ называются однородными, если

$$\forall t \in \mathbb{R}^1 \quad F_x(t) \sim F_y(t)$$

Для доказательства однородности выборок следует проверять гипотезу $H_0 : \forall t \in \mathbb{R}^1 \quad F_x(t) = F_y(t)$

Пример

Имеется две выборки $X_1, \dots, X_m \sim F(t)$ и $Y_1, \dots, Y_n \sim F(t - \theta)$. Пусть $|EX| < \infty$, тогда

$$\begin{aligned} EY_1 &= \int_{-\infty}^{+\infty} t f_y(t) dt = \int_{-\infty}^{+\infty} t f_x(t - \theta) dt = \left\langle \begin{matrix} t - \theta = z \\ t = z + \theta \end{matrix} \right\rangle = \int_{-\infty}^{+\infty} (z + \theta) f_x(z) dz = \\ &= \underbrace{\int_{-\infty}^{+\infty} z f_x(z) dz}_{EX} + \theta \underbrace{\int_{-\infty}^{+\infty} f_x(z) dz}_1 = EX + \theta \end{aligned}$$

Тогда для проверки однородности могут быть использованы гипотезы:

$$H_0 : \theta = m_y - m_x = 0, \text{ против } \begin{cases} H_1 : \theta < 0 (m_y < m_x) \\ H_2 : \theta > 0 (m_y > m_x) \\ H_3 : \theta \neq 0 (m_y \neq m_x) \end{cases}$$

Критерий Стьюдента

Есть две выборки $X_1, \dots, X_m \sim N(m_x, \sigma^2)$ и $Y_1, \dots, Y_n \sim N(m_y, \sigma^2)$. Выборки независимы и имеют одинаковые (но неизвестные нам) дисперсии.

Тогда для проверки гипотезы $H_0 : m_y - m_x = 0$ подойдёт статистика:

$$T(x, y) = \frac{\bar{Y} - \bar{X}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Здесь $S^2 = \frac{\sum_{i=1}^m (x_i - \bar{X})^2 + \sum_{i=1}^n (y_i - \bar{Y})^2}{m+n-2}$. При верности гипотезы H_0 получаем

$$T(x, y)|_{H_0} \sim t(n + m - 2)$$

Против гипотезы $H_1 : \theta < 0$



Против гипотезы $H_2 : \theta > 0$



Против гипотезы $H_3 : \theta \neq 0$



Ранговые критерии

Определение

Рангом элемента выборки называется его номер в вариационном ряду:

$$R(x_{(k)}) = k$$

Процедура определения рангов элементов выборки называется ранжированием.

Определение

Связкой размера n называют n совпадающих элементов выборки.

Если связке размера m предшествует k элементов, то все элементы связки получают один ранг, равный

$$\frac{1}{m} \sum_{i=k+1}^{m+k} i$$

Ранговый критерий Вилкоксона (1945)

Предполагается $X_1, \dots, X_m \sim F(t)$ и $Y_1, \dots, Y_n \sim F(t - \theta)$. Выборки независимы, $F(t)$ — непрерывное распределение. Проверяем гипотезу $H_0 : \theta = 0$.

Интуитивно понятно, что в случае $\theta \ll 0$ (математическое ожидание Y сильно меньше, чем у X) элементы в вариационном ряду располагаются так:

$$y_{(1)}, \dots, y_{(n)} x_{(1)}, \dots, x_{(m)}$$

И в случае $\theta \gg 0$:

$$x_{(1)}, \dots, x_{(m)} y_{(1)}, \dots, y_{(n)}$$

Для проверки критерия введём следующую статистику:

$$W_{m, n} = \sum_{i=1}^n R_i, \text{ где } R_i - \text{ранг } Y_i \text{ в объединённой выборке}$$

Тогда для случая $\theta \ll 0$

$$\min W_{m, n} = \sum_{i=1}^n R_i = (n+1) \frac{n}{2}$$

Для случая $\theta \gg 0$

$$\max W_{m, n} = \sum_{i=1}^n R_i = (n+2m+1) \frac{n}{2}$$

Если $\theta = 0$, то выборка должна быть перемешана, тогда для статистики справедливо.

$$EW_{m, n}|_{H_0} = (n+m+1) \frac{n}{2}, \quad \mathcal{D}W_{m, n} = \frac{mn}{12}(m+n+1)$$