

# Математическая статистика

15 июня 2020 г.

## Содержание

<b>1</b>	<b>Постановка задач математической статистики</b>	<b>3</b>
1.1	Задачи теории вероятностей . . . . .	3
1.2	Задачи математической статистики . . . . .	3
<b>2</b>	<b>Частота как оценка вероятности события и её свойства. Построение доверительного интервала для вероятности события на базе асимптотической нормальности частоты.</b>	<b>5</b>
<b>3</b>	<b>Постановка выборочной статистической модели. Точечная оценка параметра и характеристики.</b>	<b>7</b>
<b>4</b>	<b>Функции потерь и функции риска, состоятельность оценки характеристики, достаточное условие для состоятельности оценки.</b>	<b>8</b>
<b>5</b>	<b>Вид квадратичного риска в случае одномерной характеристики.</b>	<b>10</b>
<b>6</b>	<b>Постановка задачи доверительного оценивания, доверительный интервал.</b>	<b>11</b>
<b>7</b>	<b>Определение несмещенности и асимптотической нормальности оценки характеристики. Построение доверительного интервала для характеристики на базе асимптотической нормальности ее оценки.</b>	<b>12</b>
<b>8</b>	<b>Постановка задачи проверки гипотез</b>	<b>13</b>
<b>9</b>	<b>Ошибки первого и второго рода и их вероятности как критерий качества критерия (теста) проверки гипотез. Подход Неймана-Пирсона.</b>	<b>14</b>
<b>10</b>	<b>Эмпирическая функция распределения (ЭФР). Построение, свойства ЭФР при фиксированном значении аргумента (использовать свойства частоты).</b>	<b>15</b>
<b>11</b>	<b>Свойства ЭФР в целом. Расстояние Колмогорова, Смирнова. Теоремы Гливленко-Кантелли, Колмогорова, Мизеса-Смирнова. Построение доверительной полосы для функции распределения.</b>	<b>16</b>

<b>12 Критерии согласия Колмогорова и Мизеса-Смирнова.</b>	<b>18</b>
12.1 Критерий согласия Колмогорова . . . . .	18
12.2 Критерий Мизеса-Смирнова . . . . .	18
12.3 Прикладной алгоритм: . . . . .	19
<b>13 Выборочный метод построения оценок одномерных характеристик. Асимптотическая нормальность оценки. Построение асимптотического доверительного интервала на базе асимптотической нормальности.</b>	<b>20</b>
13.1 Описание выборочного метода . . . . .	20
<b>14 Основные выборочные оценки и их свойства. Выборочное математическое ожидание. Выборочная дисперсия. Выборочные моменты. Выборочные медиана и квантили. Выборочные оценки ковариации и коэффициента корреляции.</b>	<b>21</b>
14.1 Выборочное среднее / М.О. . . . .	21
14.2 Выборочная дисперсия . . . . .	22
14.3 Несмещенная выборочная дисперсия . . . . .	22
14.4 Выборочные моменты . . . . .	23
14.4.1 Выборочные начальные моменты . . . . .	23
14.4.2 Выборочные центральные моменты . . . . .	23
14.5 Выборочная медиана . . . . .	23
14.6 Выборочная ковариация и корреляция . . . . .	24
14.6.1 Выборочная ковариация . . . . .	24
14.7 Выборочная корреляция . . . . .	25

# 1 Постановка задач математической статистики

*Сравним задачи теории вероятностей и математической статистики*

## 1.1 Задачи теории вероятностей

Заданы:

- Вероятностное пространство  $\langle \Omega, \Sigma, P \rangle$ .
- Случайная величина  $X : \Omega \rightarrow \mathbb{R}^n$ .

Требуется получить различного рода характеристики величины  $X$  и величин, получающихся из  $X$ .

## 1.2 Задачи математической статистики

**Определение.** Статистическим экспериментом называется четверка

$$\langle \mathcal{X}, \mathcal{A}, P_\theta, \Theta \rangle.$$

Здесь:

- $\mathcal{X}$  – множество наблюдений.
- $\mathcal{A}$  –  $\sigma$ -алгебра подмножеств  $\mathcal{X}$ .
- $P_\theta$  – известная с точностью до неизвестного параметра  $\theta$  вероятностная мера – закон распределения наблюдаемых данных.
- $\Theta$  – множество допустимых значений неизвестного параметра, то есть  $\theta \in \Theta$ .

Задачей математической статистики является получение той или иной информации о законе распределения наблюдаемых данных  $P = P_\theta$ .

**Определение.** Статистикой называется измеримая функция

$$f : \mathcal{X} \rightarrow A.$$

Для произвольного  $A$ .

**Определение.** Пусть

$$\bar{X} = \langle X_1, \dots, X_n \rangle.$$

Где  $X_i \sim X$  – одинаково распределенные случайные величины. Соответствующая модель называется моделью независимой однородной выборки.

**Определение.** Гипотезой  $H$  называется подмножество  $\Theta$ :

$$H \subseteq \Theta.$$

Перечислим некоторые задачи математической статистики.

- Оценивание параметра  $\theta$  или какой-либо функции  $g(\theta)$ , то есть построение статистики  $\hat{g}: \mathcal{X} \rightarrow \Theta$ . Оценивание может быть:
  - *точечным*, то есть указание численной оценки  $g(\theta)$
  - *доверительным*, то есть указание множества, с фиксированной вероятностью содержащего  $g(\theta)$
- Проверка гипотез. Пусть имеется разбиение  $\Theta$  на гипотезы:  $\Theta = \bigsqcup_{n \in N} H_n$ . Тогда проверкой гипотезы назовем построение *теста (критерия)*, то есть отображения

$$\varphi: \mathcal{X} \rightarrow N.$$

Которое по наблюдению выдает номер гипотезы, которому это наблюдение “соответствует”.

Естественно, перечисленные задачи можно оценивать с точки зрения качества. В этом смысле всегда требуется с точки зрения какой-либо метрики построить “лучшую” оценку.

## 2 Частота как оценка вероятности события и её свойства. Построение доверительного интервала для вероятности события на базе асимптотической нормальности частоты.

**Теорема 2.1.** (Яков, Бернулли)

Пусть имеется  $\xi_i \sim \xi$  – последовательность одинаково распределенных и попарно независимых случайных величин. Пусть

$$\bar{\xi}_n = \frac{1}{n} \sum_{i=1}^n \xi_i = \frac{k_n}{n}.$$

Тогда

$$\bar{\xi}_n \xrightarrow[n \rightarrow +\infty]{} p.$$

**Теорема 2.2.** (Центральная предельная теорема, простейший вариант)

Пусть случайные величины  $X_i \sim X$  независимы и одинаково распределены, причем  $\exists E(X), D(X)$ . Тогда для случайной величины

$$Y_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sigma(\bar{X}_n)}.$$

Верно:

$$F_{Y_n} \xrightarrow[\mathbb{R}]{} F_{N(0,1)}.$$

**Теорема 2.3.** (Свойства частоты как оценки  $p$ )

Пусть  $\xi \sim B(p)$ . Тогда

$$\hat{p} = \frac{k_n}{n}$$

Является несмещенной асимптотически нормальной оценкой  $p$ , то есть

$$E(\hat{p}) = p,$$

$$\sqrt{n} \cdot (\hat{p} - p) = Y_n \xrightarrow{P_{n,\theta}} Y \sim N(0, \Delta^2(p)), \Delta^2(p) = p(1-p).$$

*Доказательство.*

- Покажем несмещенность:

$$E(\hat{p}) = E\left(\frac{k_n}{n}\right) = \frac{1}{n}np = p.$$

- Асимптотическая нормальность с нормирующим множителем  $\Delta^2(p) = p(1-p)$  следует непосредственно из центральной предельной теоремы.

■

На базе асимптотической нормальности можно построить доверительный интервал. Проделаем это на примере частоты. Выпишем определение асимптотической нормальности:

$$Y_n = \frac{\sqrt{n} \cdot (\hat{p} - p)}{\sqrt{p(1-p)}} \rightarrow N(0, 1).$$

Это буквально означает:

$$P_{n,\theta}(Y_n < t) \rightarrow F_{N(0,1)}(t).$$

Раскроем определение  $Y_n$ , возьмем его по модулю и воспользуемся квантилью:

$$P_{n,\theta}\left(\left|\frac{\sqrt{n} \cdot (\hat{p} - p)}{\sqrt{p(1-p)}}\right| < t_\gamma\right) \rightarrow \gamma \iff P_{n,\theta}\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\gamma + \hat{p} > p > -\frac{\sqrt{p(1-p)}}{\sqrt{n}}t_\gamma + \hat{p}\right) \rightarrow \gamma.$$

Здесь  $\gamma = P(|\xi| < t_\gamma)$ ,  $\xi \sim N(0, 1)$ . Построим д

### 3 Постановка выборочной статистической модели. Точечная оценка параметра и характеристики.

**Определение.** Напомним, что *точечной оценкой* параметра  $\theta$  или какой-либо функции  $g(\theta)$  называют численную оценку этой величины.

Пусть  $\hat{g}_n$  является некоторой точечной оценкой  $g = g(\theta)$ .

**Определение.**  $\hat{g}_n$  называется *несмещенной*, если  $E(\hat{g}_n) = g(\theta)$ .

**Определение.**  $\hat{g}_n$  называется *состоятельной*, если  $\hat{g}_n \xrightarrow{p} g(\theta)$  при  $n \rightarrow \infty$ .

**Определение.**  $\hat{g}_n$  называется *асимптотически нормальной*, если

$$\frac{\sqrt{n}(\hat{g}_n - g(\theta))}{\sigma(g(\theta))} \xrightarrow{n \rightarrow \infty} N(0, 1).$$

**Определение.**  $\hat{g}_n$  называется *эффективной* в классе оценок  $K$ , если для любой другой оценки  $\hat{g}_n^* \in K$  имеет место неравенство:

$$E(\hat{g}_n - g(\theta))^2 \leq E(\hat{g}_n^* - g(\theta))^2.$$

## 4 Функции потерь и функции риска, состоятельность оценки характеристики, достаточное условие для состоятельности оценки.

**Определение.** Оценкой  $g(\theta)$  называется статистика вида

$$\hat{g}: \mathcal{X} \rightarrow g(\Theta).$$

**Определение.** Пусть  $\hat{g}(\theta)$  – оценка  $g(\theta)$ . Тогда функцией потерь называется неотрицательная функция  $l(\hat{g}, g(\theta))$ , характеризующая “близость” оценки к настоящему значению.

**Замечание.** Обычно в качестве функции потерь рассматривают функцию вида

$$l(\hat{g}, g(\theta)) = \omega(\|\hat{g}, g(\theta)\|).$$

Здесь  $\omega$  – неотрицательная монотонно возрастающая функция,  $\omega(0) = 0$ .

**Замечание.**  $l$  является случайной величиной.

**Определение.** Риском называется функция

$$R(\hat{g}, \theta) \stackrel{\text{def}}{=} E_{\theta}(l(\hat{g}, g(\theta))).$$

**Замечание.** Риск – функция параметра  $\theta$  и способа оценивания  $\hat{g}$ .

Опишем самые важные для нас виды функции потерь и риска.

**Определение.** Определим функцию потерь индикатором отклонений:

$$l^{\delta}(\hat{g}, g(\theta)) = \omega^{\delta}(\|\hat{g}, g(\theta)\|).$$

Где

$$\omega(t) = \mathbb{1}_{\delta}(t) = \begin{cases} 0, & t < \delta \\ 1, & t \geq \delta \end{cases}.$$

Соответствующий риск будет вероятностью отклонения:

$$R^{\delta}(\hat{g}, \theta) = E_{\theta}(l^{\delta}(\hat{g}, g(\theta))) = 0 \cdot P_{\theta}(\|\hat{g} - g(\theta)\| < \delta) + 1 \cdot P_{\theta}(\|\hat{g} - g(\theta)\| \geq \delta) = P_{\theta}(\|\hat{g} - g(\theta)\| \geq \delta).$$

**Определение.** При асимптотическом подходе оценка называется *состоятельной*, если

$$\forall \delta > 0 \quad R^{\delta}(\hat{g}_n, \theta) = P_{n,\theta}(\|\hat{g}_n - g(\theta)\| \geq \delta) \xrightarrow{n \rightarrow +\infty} 0.$$

Или, что то же самое:

$$\hat{g}_n \xrightarrow[n \rightarrow +\infty]{P_{n,\theta}} g(\theta).$$



**Определение.** Квадратичной функцией потерь называется функция

$$l_2(\hat{g}, g(\theta)) = \|\hat{g} - g(\theta)\|^2.$$

Соответствующий ей риск называется квадратичным:

$$R_2(\hat{g}, \theta) = E_\theta(\|\hat{g} - g(\theta)\|^2).$$

**Теорема 4.1.** (Достаточное условие для состоятельности оценки)

$R_2(\hat{g}_n, \theta) \xrightarrow[n \rightarrow +\infty]{} 0 \implies$  оценка состоятельна.

*Доказательство.*

$$\begin{aligned} \forall \delta > 0 \quad R^\delta(\hat{g}_n, \theta) &= P(\|\hat{g}_n - g(\theta)\| \geq \delta) = P(\|\hat{g}_n - g(\theta)\|^2 \geq \delta^2) \\ &\leq \frac{E_\theta(\|\hat{g}_n - g(\theta)\|^2)}{\delta^2} = \frac{R_2(\hat{g}_n, \theta)}{\delta^2} \xrightarrow[n \rightarrow +\infty]{} 0. \end{aligned}$$

■

## 5 Вид квадратичного риска в случае одномерной характеристики.

**Определение.** Смещением оценки называется величина

$$b(\hat{g}, \theta) = g(\theta) - E_{\theta}(\hat{g}).$$

**Определение.** Оценка называется несмещенной, если  $b(\hat{g}, \theta) = 0$ .

**Теорема 5.1.**  $R_2(\hat{g}, \theta) = D_{\theta}(\hat{g}) + b^2(\hat{g}, \theta)$ .

*Доказательство.*

$$\begin{aligned} R_2(\hat{g}, \theta) &= E_{\theta}(\|\hat{g} - g(\theta)\|^2) = E_{\theta}(\hat{g} - E_{\theta}(\hat{g}) - (g(\theta) - E_{\theta}(\hat{g})))^2 \\ &= E_{\theta}(\hat{g} - E_{\theta}(\hat{g}))^2 + (g(\theta) - E_{\theta}(\hat{g}))^2 - \underbrace{2(g(\theta) - E_{\theta}(\hat{g}))(E_{\theta}\hat{g} - E_{\theta}\hat{g})}_0 \\ &= D_{\theta}(\hat{g}) + b^2(\hat{g}, \theta). \end{aligned}$$

■

**Следствие 5.2.** Для одномерных несмещенных оценок квадратичный риск в точности равен дисперсии оценки:

$$R_2(\hat{g}, \theta) = D_{\theta}(\hat{g}).$$

## 6 Постановка задачи доверительного оценивания, доверительный интервал.

При оценивании параметров или характеристик распределений мы в качестве результата получаем числовое значение  $\hat{g}(X) \in g(\Theta)$ . Такой способ оценивания мы называем *точечной оценкой*. Заранее не понятно, насколько результат соответствует действительности. Для того, чтобы можно было оценивать качество результата, нужно предъявлять не точку, а подмножество в  $g(\Theta)$ , содержащее в некотором смысле наиболее подходящие значения.

Задача доверительного оценивания ставится следующим образом: задана величина  $\gamma \in (0, 1)$ , называемая *уровнем надежности*. По заданному наблюдению  $X$  и значению надежности требуется построить доверительную область надежности.

**Определение.** Доверительной областью надежности называется  $\tilde{G}_\gamma \subseteq G = g(\Theta)$ , обладающая свойством:

$$\forall \theta \in \Theta P_\theta(g(\theta) \in \tilde{G}_\gamma) \geq \gamma.$$

То есть множество, с достаточной вероятностью содержащее оцениваемую величину.

**Определение.** В случае одномерной оценки чаще всего доверительные области надежности выбирают в виде промежутков, которые называются *доверительными интервалами*.

**Определение.** В асимптотическом случае (когда имеется последовательность оценок и статистических экспериментов) последовательность *асимптотических областей надежности*  $\tilde{G}_{n,\gamma}$  задается условием:

$$\forall \theta \in \Theta \lim P_{n,\theta}(g(\theta) \in \tilde{G}_{n,\gamma}) \geq \gamma.$$

**Определение.** Аналогично задается последовательность асимптотических доверительных интервалов в случае одномерной характеристики.

## 7 Определение несмещенности и асимптотической нормальности оценки характеристики. Построение доверительного интервала для характеристики на базе асимптотической нормальности ее оценки.

**Определение.** Напомним, оценка называется *несмещенной*, если

$$b(\hat{g}, \theta) = g(\theta) - E_{\theta}(\hat{g}) = 0.$$

**Определение.** Последовательность оценок  $\hat{g}_n$  называется *асимптотически нормальной*, если

$$\sqrt{n} \cdot (\hat{g}_n - g(\theta)) = Y_n \xrightarrow{P_{n,\theta}} Y \sim N(0, \Delta^2(\theta)).$$

**Определение.** Величина  $\Delta(\theta)$  из определения асимптотически нормальной оценки называется *нормирующим множителем*.

**Замечание.** Определение асимптотически нормальной оценки можно переписать так:

$$\frac{\sqrt{n} \cdot (\hat{g}_n - g(\theta))}{\Delta(\theta)} \xrightarrow{P_{n,\theta}} Y \sim N(0, 1).$$

На базе асимптотической нормальности можно построить доверительный интервал. Выпишем определение асимптотической нормальности:

$$Y_n = \frac{\sqrt{n} \cdot (\hat{g} - g(\theta))}{\Delta(\theta)} \rightarrow N(0, 1).$$

Это буквально означает:

$$P_{n,\theta}(Y_n < t) \rightarrow F_{N(0,1)}(t).$$

Раскроем определение  $Y_n$ , возьмем его по модулю и воспользуемся квантилью:

$$P_{n,\theta}\left(\left|\frac{\sqrt{n} \cdot (\hat{g} - g(\theta))}{\Delta(\theta)}\right| < t_{\gamma}\right) \rightarrow \gamma \iff P_{n,\theta}\left(\frac{\Delta(\theta)}{\sqrt{n}}t_{\gamma} + \hat{g} > g(\theta) > -\frac{\Delta(\theta)}{\sqrt{n}}t_{\gamma} + \hat{g}\right) \rightarrow \gamma.$$

Здесь  $\gamma = P(|\xi| < t_{\gamma})$ ,  $\xi \sim N(0, 1)$ .

## 8 Постановка задачи проверки гипотез

**Определение.** *Гипотезой* называется множество предполагаемых зафиксированных значений некоторого подмножества неизвестных параметров:

$$H : \theta \in \Theta_H \subseteq \Theta.$$

**Определение.** Гипотезу называют *простой*, если  $|H| = 1$ .

**Определение.** Гипотезу называют *сложной*, если  $|H| > 1$ .

**Определение.** Гипотезами *согласия* называют набор из двух гипотез: основной  $H_0$  и альтернативы  $H_1$ , причем  $H_0 = \overline{H_1}$ .

**Определение.** Правило принятия или отклонения основной гипотезы  $H_0$  называют *тестом (критерием)* проверки гипотезы:

$$\varphi(X) : X_n \rightarrow \{0, 1\}.$$

При этом:

- $X_{n,0}$  называют *допустимым множеством*.
- $X_{n,1}$  называют *критическим множеством*.
- $X_{n,0} \sqcup X_{n,1} = X_n$ .

**Определение.** Случайная величина  $L(\bar{X}) : X_n \rightarrow \mathbb{R}$  называется *тестовой статистикой*, если она служит порогом для правила принятия или отклонения основной гипотезы:

$$\varphi(\bar{X}) = \begin{cases} 0, & L(\bar{X}) < T(H_0) \\ 1, & L(\bar{X}) \geq T(H_0) \end{cases}.$$

Где  $T$  называют *порогом принятия решения*.

## 9 Ошибки первого и второго рода и их вероятности как критерий качества критерия (теста) проверки гипотез. Подход Неймана-Пирсона.

**Определение.** *Ошибкой I рода* называют отклонение основной гипотезы, в то время как она была верна.

**Определение.** *Ошибкой II рода* называют принятие основной гипотезы, в то время как она не была верна.

**Определение.**  $\alpha$  называют *вероятностью ошибки I рода*:

$$\alpha(\varphi, \theta) \stackrel{\text{def}}{=} P_{\theta}(X_{n,1}), \theta \in \Theta_{H_0}.$$

**Определение.** *Уровнем значимости теста* называют верхнюю границу вероятности ошибки I рода по всем возможным наблюдаемым значениям неизвестных параметров, отвечающих основной гипотезе:

$$\alpha(\varphi) \stackrel{\text{def}}{=} \sup_{\theta \in \Theta_{H_0}} \alpha(\varphi, \theta).$$

**Определение.**  $\beta$  называют *вероятностью ошибки II рода*:

$$\beta(\varphi, \theta) \stackrel{\text{def}}{=} P_{\theta}(X_{n,0}), \theta \in \Theta_{H_1}.$$

**Определение.** *Мощностью теста* называют следующую величину:

$$\gamma(\varphi, \theta) \stackrel{\text{def}}{=} 1 - \beta(\varphi, \theta).$$

### Подход Неймана-Пирсона.

Зафиксируем  $\alpha \in (0, 1)$  (обычно выбирают малое значение). Будем считать это значение минимальной допустимой величиной ошибки I рода (*допустимый уровень значимости*).

Рассмотрим множество всех тестов таких, что:

$$\bar{\Phi}_{\alpha} = \{\varphi = \varphi(x) \mid \alpha(\varphi) \leq \alpha\}.$$

Среди этих тестов выбирается тест с минимальным значением  $\beta$ .

В асимптотических задачах ограничения накладываются на предельные значения.

## 10 Эмпирическая функция распределения (ЭФР). Построение, свойства ЭФР при фиксированном значении аргумента (использовать свойства частоты).

**Определение.** Эмпирической функцией распределения (ЭФР) называют следующую оценку функции распределения генеральной совокупности:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}.$$

Иными словами, значение ЭФР в точке  $t$  равно отношению числа наблюдений, меньших  $t$ , к их общему числу  $n$ .

**Свойства ЭФР:**

1. ЭФР кусочно-постоянна.
2. Скачки ЭФР имеют вид  $\frac{k}{n}$  для некоторого  $k \in (1; n)$ .
3. Область принимаемых значений:  $[0; 1]$ .
4. Частота может служить как оценка функции распределения генеральной совокупности. При фиксированном  $t = t_0$ :

$$F_x(t_0) \approx F_n(t_0) = \xi_1 + \dots + \xi_n = \frac{k_n}{n} - \text{частота}.$$

5.  $F_n(t)$  является состоятельной оценкой:

$$F_n(t_0) = \bar{\xi}_n : F_n(t_0) \xrightarrow{p=1} F_x.$$

6.  $F_n(t)$  является асимптотически нормальной оценкой.

## 11 Свойства ЭФР в целом. Расстояние Колмогорова, Смирнова. Теоремы Гливленко-Кантелли, Колмогорова, Мизеса-Смирнова. Построение доверительной полосы для функции распределения.

Со свойствами ЭФР можно ознакомиться в предыдущем разделе.

**Определение.** Расстояние Колмогорова:

$$\rho_{\infty}(F_n, F_x) = \sup_t |F_n(t) - F_x(t)|.$$

**Определение.** Расстояние Смирнова:

$$\rho_2^2(F_n, F_x) = \int_{\mathbb{R}} (F_n(t) - F_x(t))^2 dF_x(t).$$

**Теорема 11.1.** (Гливленко-Кантелли)

Пусть  $\mathcal{F}$  – множество функций распределения. Тогда  $\forall F_x(t) \in \mathcal{F}$  верно:

$$\rho_{\infty}(F_n, F_x) \xrightarrow[n \rightarrow \infty]{p=1} 0 \Rightarrow \rho_{\infty}(F_n, F_x) \xrightarrow[p]{} 0.$$

**Замечание.**  $F_n(t)$  – состоятельная оценка  $F_x(t)$  в расстояниях Колмогорова и Смирнова.

Пусть  $\mathcal{F}_c$  – множество всех непрерывных функций распределения.

**Теорема 11.2.** (Колмогоров)

$$P_{n,F}(\sqrt{n}\rho_{\infty}(F_n, F_x) < u) \xrightarrow[n \rightarrow \infty]{} \mathcal{K}(u) = \begin{cases} 0, & u = 0 \\ \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2(ju)^2}, & u > 0 \end{cases}.$$

**Теорема 11.3.** (Мизес, Смирнов)

$$P_{n,F}(\sqrt{n}\rho_2^2(F_n, F_x) < u) \xrightarrow[n \rightarrow \infty]{} \mathcal{S}(u),$$

где  $\mathcal{S}(u)$  есть функция распределения следующей случайной величины:

$$\mathcal{U} = \sum_{j=1}^{\infty} \frac{\xi_j^2}{j^2 \pi^2}, \quad \xi_j \sim N(0, 1), \text{ независимые.}$$

**Замечание.** Используя теорему Колмогорова, можно построить доверительную полосу для функции распределения.

**Определение.** Доверительной полосой называют часть плоскости, в которую с надежностью  $\gamma$  попадает функция распределения генеральной совокупности:

$$\begin{cases} F_n^-(t) = \max\left(0, F_n(t) - \frac{u_{\gamma}}{\sqrt{n}}\right) \\ F_n^+(t) = \min\left(1, F_n(t) + \frac{u_{\gamma}}{\sqrt{n}}\right) \end{cases}, \text{ где } \mathcal{K}(u_{\gamma}) = \gamma.$$



**Утверждение 11.4.**

$$P_x(F_n^-(t) \leq F_x(t) \leq F_n^+(t)) \xrightarrow{n \rightarrow \infty} \gamma.$$

*Доказательство.*  $0 \leq F_x(t) \leq 1$  всегда, тогда:

$$\begin{aligned} P_x(F_n^-(t) \leq F_x(t) \leq F_n^+(t)) &= P_x(F_n(t) - \frac{u_\gamma}{\sqrt{n}} \leq F_x(t) \leq F_n(t) + \frac{u_\gamma}{\sqrt{n}}) \stackrel{\forall t}{=} \\ &\stackrel{\forall t}{=} P_x(\sqrt{n}|F_x(t) - F_n(t)| \leq u_\gamma) \stackrel{\forall t}{=} \\ &\stackrel{\forall t}{=} P_x(\sup_t |F_x(t) - F_n(t)| \leq u_\gamma) \xrightarrow{\text{th. Колмогорова}} \mathcal{K}(u_\gamma) = \gamma. \end{aligned}$$

■

## 12 Критерии согласия Колмогорова и Мизеса-Смирнова.

Пусть  $F_0(t)$  – заданная непрерывная функция распределения.

Поставим задачу проверки согласия:

$$H_0 \equiv (F_x(t) \equiv F_0(t)).$$

### 12.1 Критерий согласия Колмогорова

Определим тестовую статистику:

$$L(\bar{X}) = \sqrt{n} \rho_{\infty}(F_0, F_n).$$

По th. Колмогорова:

$$P(L(\bar{X}) < z) \xrightarrow{n \rightarrow \infty} \mathcal{K}(z),$$

где  $\mathcal{K}$  – распределение Колмогорова. Тогда порогом принятия решения при уровне значимости  $\alpha$  является квантиль распределения Колмогорова порядка  $1 - \alpha$  (далее  $u_{1-\alpha}$ ).

Таким образом, определим тест:

$$\varphi(\bar{X}) = \begin{cases} 0, & \sqrt{n} \rho_{\infty}(F_0, F_n) < u_{1-\alpha} \\ 1, & \sqrt{n} \rho_{\infty}(F_0, F_n) \geq u_{1-\alpha} \end{cases}.$$

### 12.2 Критерий Мизеса-Смирнова

Определим тестовую статистику:

$$L(\bar{X}) = \sqrt{n} \rho_2^2(F_0, F_n).$$

По th. Мизеса-Смирнова:

$$P(L(\bar{X}) < z) \xrightarrow{n \rightarrow \infty} \mathcal{S}(z),$$

где  $\mathcal{S}$  – распределение Мизеса-Смирнова. Тогда порогом принятия решения при уровне значимости  $\alpha$  является квантиль распределения Мизеса-Смирнова порядка  $1 - \alpha$  (далее  $s_{1-\alpha}$ ).

Таким образом, определим тест:

$$\varphi(\bar{X}) = \begin{cases} 0, & \sqrt{n} \rho_2^2(F_0, F_n) < w_{1-\alpha} \\ 1, & \sqrt{n} \rho_2^2(F_0, F_n) \geq w_{1-\alpha} \end{cases}.$$

### 12.3 Прикладной алгоритм:

1. Строится ЭФР
2. Считается статистика критерия. Поскольку ЭФР является кусочно-постоянной, расстояние Колмогорова / Мизеса-Смирнова можно считать как верхнюю границу по соответствующим значениям расстояний в точках скачка.
3. Для заданного уровня значимости  $\alpha$  находится квантиль распределения Колмогорова / Мизеса-Смирнова порядка  $1 - \alpha$ .
4. Если значение тестовой статистики меньше полученного квантиля, следует принять нулевую гипотезу, иначе – отклонить.

## 13 Выборочный метод построения оценок одномерных характеристик. Асимптотическая нормальность оценки. Построение асимптотического доверительного интервала на базе асимптотической нормальности.

Есть три метода построения оценок:

1. Выборочный метод
2. Метод моментов
3. Метод максимального правдоподобия

### 13.1 Описание выборочного метода

Этот метод основывается на знании того, что ЭФР  $F_n(t)$  является “хорошей” оценкой функции распределения  $F_x(t)$ .

ЭФР  $F_n(t)$  является функцией распределения дискретной случайной величины  $Y$ , имеющей следующий ряд распределения:

$$\begin{array}{c|c|c|c|c} Y_i & x_{(1)} & x_{(2)} & \dots & x_{(n)} \\ \hline p_i & 1/n & 1/n & \dots & 1/n \end{array}$$

где  $x_{(1)}, \dots, x_{(n)}$  упорядоченная выборка:

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

## 14 Основные выборочные оценки и их свойства. Выборочное математическое ожидание. Выборочная дисперсия. Выборочные моменты. Выборочные медиана и квантили. Выборочные оценки ковариации и коэффициента корреляции.

Здесь  $X$  – произвольная рассматриваемая случайная величина.

### 14.1 Выборочное среднее / М.О.

**Определение.** Случайную величину  $\bar{X}_n = EY = \sum_{i=1}^n X_{(i)} \cdot \frac{1}{n} = \frac{1}{n} \sum_{j=1}^n X_j$  называют *выборочным средним* некоторой выборки  $X_{[n]}$  из генеральной совокупности  $X$ .

Выборочное среднее является выборочной точечной оценкой  $EX$ .

**Свойства.**

- Выборка является набором одинаково распределенных независимых случайных величин, из чего по закону больших чисел:

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{P} EX, \text{ если } \exists EX.$$

Поэтому выборочное среднее является *состоятельной* оценкой  $EX$ .

- Выборочное среднее является *несмещенной* оценкой  $EX$ :

$$E_x \bar{X}_n = \frac{1}{n} \sum_{i=1}^n E(X_i) = EX, \text{ оо св-ву М.О.}$$

- Если  $\exists EX, DX$ , то по центральной предельной теореме:

$$Y_n = \frac{\bar{X}_n - E_x(\bar{X}_n)}{\sigma_x(\bar{X}_n)} = \frac{\bar{X}_n - EX}{\sigma(X)} \sqrt{n} \xrightarrow[n \rightarrow \infty]{F} Y \sim N(0, 1),$$

или

$$F_{Y_n}(t) \xrightarrow[n \rightarrow \infty]{\mathbb{R}} F_Y(t).$$

Из этого следует, что центрированное нормированное выборочное среднее сходится по распределению к стандартному нормальному распределению. Следовательно, выборочное среднее является *асимптотически нормальной* оценкой  $EX$ .

## 14.2 Выборочная дисперсия

**Определение.** Случайную величину  $S_n^2 = D(X_{[n]}) = \sum_{i=1}^n (X_{(i)} - EY)^2 \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  называют *выборочной дисперсией* некоторой выборки  $X_{[n]}$  из генеральной совокупности  $X$ .

Выборочная дисперсия является выборочной точечной оценкой  $DX$ .

**Свойства.**

- Выборочная дисперсия является *состоятельной* оценкой  $DX$ :

$$S_n^2 \xrightarrow[n \rightarrow \infty]{P} E(X^2) - (EX)^2 = DX.$$

- Выборочная дисперсия является *смещенной* оценкой  $DX$ .  
(Ниже, не теряя общности, будем считать  $EX = 0$ , инвариантность  $DX$  относительно сдвига):

$$E(X^2) = DX, \quad E_x(\bar{X}_n) = EX = 0$$

$$\Rightarrow E_x(\bar{X}_n^2) = D_x(\bar{X}_n) + (E_x(\bar{X}_n))^2 = \frac{DX}{n}$$

$$\Rightarrow E_x S_n^2 = E_x \left( \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) = \frac{1}{n} n E(X^2) - \frac{DX}{n} = DX - \frac{DX}{n} = \frac{n-1}{n} DX.$$

- Выборочная дисперсия является *асимптотически нормальной* оценкой  $DX$  – без доказательства.

## 14.3 Несмещенная выборочная дисперсия

**Определение.** Чаше вместо  $S_n^2$  используют *несмещенную (исправленную) оценку дисперсии*:

$$\sigma_n^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

Несмещенная выборочная дисперсия является выборочной точечной оценкой  $DX$ .

**Свойства.**

- *Состоятельность* следует из состоятельности  $S_n^2$ :

$$\sigma_n^2 \xrightarrow[n \rightarrow \infty]{p=1} DX.$$

- *Несмещенность* очевидна из доказательства смещенности выборочного среднего.
- Несмещенная оценка дисперсии является *асимптотически нормальной* оценкой  $DX$  – без доказательства.

## 14.4 Выборочные моменты

### 14.4.1 Выборочные начальные моменты

**Определение.** Выборочным начальным моментом порядка  $k$  называется статистика:

$$m_{n,k} = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots$$

Эти выборочные характеристики можно считать выборочным средним для случайной величины  $Z = X^k$ :

$$m_{n,k} = \bar{Z}_k.$$

Следовательно, если  $\exists E(X^k)$ , то  $m_{n,k}$  является состоятельной и несмещенной оценкой  $E(X^k)$ .

Если существует  $E(X^{2k})$ , то  $m_{n,k}$  является асимптотически нормальной оценкой  $E(X^k)$  с асимптотической дисперсией  $\Delta^2 = E_x(Z^2) - (E_x(Z))^2$ .

### 14.4.2 Выборочные центральные моменты

**Определение.** Выборочным центральным моментом порядка  $k$  называется статистика:

$$\mu_{n,k} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^k.$$

Данные статистики являются состоятельными, смещенными оценками соответствующих центральных моментов генеральной совокупности.

## 14.5 Выборочная медиана

**Определение.** Медианой  $t_0$  случайной величины  $X$  ( $med(x)$ ) называют такое значение аргумента функции распределения  $F_x(t)$ , что для него выполняются неравенства:

$$\begin{cases} P(X \geq t_0) \geq \frac{1}{2} \\ P(X \leq t_0) \geq \frac{1}{2} \end{cases}.$$

Если  $F_x(t) \in C(\mathbb{R})$ , то  $F_x(t_0) = \frac{1}{2}$ .

**Определение.** Пусть  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  — упорядоченная выборка (вариационный ряд), тогда выборочной медианой  $med_n$  называется следующая случайная величина:

$$med_n = \begin{cases} X_{(k)} = X_{\frac{n-1}{2}}, & \text{при } n = 2k - 1 \\ \frac{X_{(k)} + X_{(k+1)}}{2}, & \text{при } n = 2k \end{cases}.$$

**Свойства.** Пусть генеральная совокупность является непрерывной случайной величиной и  $T = \{t : 0 < F_X(t) < 1\}$ . Если  $f_X(t)$  непрерывна и положительна при  $t \in T$ , то плотность распределения случайной величины  $Y = f_Y(t)$ , где

$$Y = 2\sqrt{n}f_X(t_0)(\text{med}_n - t_0), \quad t_0 = \text{med}(X)$$

при  $n \rightarrow \infty$  стремится к  $f_{N(0,1)}(t) = \frac{1}{\sqrt{2\pi}} \exp -\frac{t^2}{2}$ , а

$$P(a < Y < b) \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx.$$

- Следовательно, выборочная медиана является *состоятельной* оценкой  $\text{med}(X)$ .
- Также видно, что выборочная медиана является *асимптотически нормальной* оценкой  $\text{med}(x)$  с асимптотической дисперсией  $\Delta^2 = \frac{1}{rf_X^2(t_0)}$ .

- Выборочная медиана является  $\sqrt{n}$ -несмещенной оценкой  $\text{med}(X)$ . То есть:

$$\sqrt{n}b_{n,\theta}(\text{med}_n) = \sqrt{n}(E_x(\text{med}_n) - \text{med}(X)) \xrightarrow{n \rightarrow \infty} 0.$$

## 14.6 Выборочная ковариация и корреляция

Выборочная ковариация и корреляция используются при решении вопроса о наличии зависимости между случайными величинами  $X$  и  $Y$ .

В этом случае рассматривается выборка из случайного вектора  $(X, Y)$ . Здесь пары  $\{X_i, Y_i\}_i$  независимы и одинаково распределены. Если случайные величины  $X$  и  $Y$  не являются линейно зависимыми ( $r(X, Y) \neq 1$ ), то для последовательности  $\{X_i, Y_i\}_i$  справедливо утверждение аналогичное центральной предельной теореме.

### 14.6.1 Выборочная ковариация

**Определение.** Выборочной ковариацией называется статистика:

$$K_n = K_n(X, Y) = \frac{1}{n} \sum_{i=1}^n ((X_i - \bar{X}_n)(Y_i - \bar{Y}_n)) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \bar{X}_n \bar{Y}_n.$$

**Свойства.**

- *Состоятельная* оценка.  $X$  и  $Y$ .
- *Смещенная* оценка. Аналогично выборочной дисперсии, можно показать:

$$E_x(K_n) = \frac{n-1}{n} K(X, Y).$$

- *Асимптотически нормальная* оценка.

**Определение.** В приложениях обычно рассматривают *несмещенную* оценку ковариации:

$$\tilde{K}_n = \frac{n}{n-1} K_n = \frac{1}{n-1} \sum_{i=1}^n ((X_i - \bar{X}_n)(Y_i - \bar{Y}_n)).$$



## 14.7 Выборочная корреляция

**Определение.** Выборочной корреляцией  $X$  и  $Y$  называется статистика:

$$r_n = r_n(X, Y) = \frac{K_n}{S_{n,X}S_{n,Y}} = \frac{\tilde{K}_n}{\sigma_{n,X}\sigma_{n,Y}}.$$

**Замечание.** В определении выше предполагается существование всех необходимых моментов:  $EX, EY, K(X, Y), \dots$

**Свойства.**

- *Состоятельная оценка.*
- *Несмещенная оценка.*
- *Асимптотически нормальная оценка.*