

# ОСНОВЫ ПРИКЛАДНОЙ СТАТИСТИКИ

## для ДПО ЦНО НИУ ВШЭ

подготовил Илья Аброскин (allen.ilya@gmail.com)  
основано на <http://statistics.zone/>

# Содержание

<b>1</b>	<b>Случайные события</b>	<b>3</b>
1.1	Основные свойства и тождества . . . . .	3
<b>2</b>	<b>Случайные величины</b>	<b>3</b>
2.1	Функции, описывающие случайные величины . . . . .	3
2.2	Совместное распределение . . . . .	3
<b>3</b>	<b>Характеристики случайных величин</b>	<b>4</b>
3.1	Математическое ожидание . . . . .	4
3.2	Дисперсия . . . . .	4
3.3	Ковариация и корреляция . . . . .	4
<b>4</b>	<b>Обзор основных распределений</b>	<b>5</b>
4.1	Дискретные распределения . . . . .	5
4.2	Непрерывные распределения . . . . .	5
4.3	Свойства некоторых распределений . . . . .	5
<b>5</b>	<b>Статистический вывод</b>	<b>6</b>
5.1	Выборочные статистики . . . . .	6
<b>6</b>	<b>Параметрические Д.И. и тестирование гипотез</b>	<b>7</b>
6.1	Введение . . . . .	7
6.1.1	Доверительные интервалы . . . . .	7
6.1.2	Тестирование гипотез . . . . .	7
6.2	Асимптотические Д.И. на основе ЦПТ . . . . .	9
6.2.1	Д.И. для мат. ожидания выборок из любого распределения	9
6.2.2	Д.И. для теоретической доли распределения Бернулли .	9
6.3	Точные Д.И. для нормальных выборок . . . . .	10
6.3.1	Д.И. для мат. ожидания . . . . .	10
6.3.2	Д.И. для дисперсии . . . . .	10
6.3.3	Д.И. для разности мат. ожиданий . . . . .	11
6.3.4	Д.И. для отношения дисперсий . . . . .	11
6.3.5	Д.И. для разности мат. ожиданий в зависимых выборках	12
6.4	Некоторые дополнительные тесты . . . . .	12
6.4.1	Тест о значимости корреляции . . . . .	12
6.4.2	Тест о равенстве пропорций в зависимых выборках . . .	12

# 1 Случайные события

Определения

- Пространство элементарных исходов:  $\Omega$
- Элементарный исход:  $\omega \in \Omega$
- Случайное событие:  $A \subseteq \Omega$

## 1.1 Основные свойства и тождества

Вероятность противоположного события

$$\mathbb{P}[\neg A] = 1 - \mathbb{P}[A]$$

Вероятность объединения событий

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$

Независимость событий  $\perp\!\!\!\perp$

$$A \perp\!\!\!\perp B \iff \mathbb{P}[A \cap B] = \mathbb{P}[A] \mathbb{P}[B]$$

Условная вероятность

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} \quad \mathbb{P}[B] > 0$$

Формула полной вероятности

$$\mathbb{P}[B] = \sum_{i=1}^n \mathbb{P}[B|A_i] \mathbb{P}[A_i] \quad \Omega = \bigsqcup_{i=1}^n A_i$$

Теорема Байеса

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[B | A] \mathbb{P}[A]}{\mathbb{P}[B]}$$

# 2 Случайные величины

Случайная величина (random variable, RV) – функция, значения которой численно выражают исходы  $\omega$  случайного эксперимента.

$$X : \Omega \rightarrow \mathbb{R}$$

Например, для исход  $\omega$  = «родилась девочка» соответствует значению случайной величины  $X = 1$ , а исход  $\omega$  = «родился мальчик» соответствует  $X = 0$ . Случайная величина, которая принимает только 2 возможных значения и описывает только 2 возможных исхода, называется случайной величиной из распределения Бернулли.

## 2.1 Функции, описывающие случайные величины

**Функция вероятности** (probability mass function, **PMF**) для дискретных случайных величин

$$\mathbb{P}[X = x]$$

**Функция плотности** (probability density function, **PDF**) для непрерывных случайных величин

$$f_X(x) \text{ such that: } \mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x) dx$$

**Функция распределения** (cumulative distribution function, **CDF**) для любых случайных величин

$$F_X(x) = \mathbb{P}[X \leq x]$$

Свойства

1. Не убывает:  $x_1 < x_2 \implies F(x_1) \leq F(x_2)$
2. Принимает значения от 0 до 1
3.  $f(x) = F'(x)$
4.  $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$

## 2.2 Совместное распределение

Несколько случайных величин описываются совместным распределением

- PMF:  $\mathbb{P}[X = x, Y = y]$  (иногда используются таблицы)
- PDF:  $f_{X,Y}(x, y)$  (функция от двух аргументов)
- CDF:  $F_{X,Y}(x, y) = \mathbb{P}[X \leq x, Y \leq y]$

Независимость

- для любых случайных величин:  $F_{X,Y}(x, y) = F_X(x)F_Y(y)$
- для непрерывных величин:  $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

## 3 Характеристики случайных величин

### 3.1 Математическое ожидание

Определение

$$\mathbb{E}[X] = \mu_X = \begin{cases} \sum_x x \cdot \mathbb{P}[x] & X \text{ дискретная} \\ \int_{-\infty}^{+\infty} x \cdot f_X(x) dx & X \text{ непрерывная} \end{cases}$$

Свойства

- $\mathbb{E}[aX] = a \mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- $\mathbb{E}[\varphi(Y)] \neq \varphi(\mathbb{E}[X])$  (Неравенство Йенсена)
- $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$  (если  $X \perp Y$  – независимы)
- $\mathbb{E}[XY] = \begin{cases} \sum_{x,y} x \cdot y \cdot \mathbb{P}[X=x, Y=y] & X \text{ дискретная} \\ \int x \cdot y \cdot f_{X,Y}(x, y) dx dy & X \text{ непрерывная} \end{cases}$

Медиана

$$\text{Median}(X) : \mathbb{P}(X < \text{Median}(X)) = \mathbb{P}(X > \text{Median}(X)) = 0.5$$

Мода

$$\text{Mode}(X) = \begin{cases} \underset{x}{\operatorname{argmax}} \mathbb{P}[x] & X \text{ дискретная} \\ \underset{x}{\operatorname{argmax}} f(x) & X \text{ непрерывная} \end{cases}$$

Квантиль уровня  $\gamma$

$$\text{quantile}(X) = q : \mathbb{P}(X \leq q) = \gamma$$

### 3.2 Дисперсия

Определение

$$\mathbb{V}\text{ar}[X] = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Стандартное (среднеквадратичное) отклонение

$$\text{std}[X] = \sigma_X = \sqrt{\mathbb{V}\text{ar}[X]}$$

Свойства

- $\mathbb{V}\text{ar}[X + a] = \mathbb{V}\text{ar}[X]$
- $\mathbb{V}\text{ar}[a \cdot X] = a^2 \cdot \mathbb{V}\text{ar}[X]$
- $\mathbb{V}\text{ar}[X \pm Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y]$  (если  $X \perp Y$  независимы)
- $\mathbb{V}\text{ar}[X \pm Y] = \mathbb{V}\text{ar}[X] + \mathbb{V}\text{ar}[Y] \pm 2 \cdot \text{Cov}[X, Y]$
- $\mathbb{V}\text{ar}[aX + bY] = a^2 \mathbb{V}\text{ar}[X] + b^2 \mathbb{V}\text{ar}[Y] + 2ab \cdot \text{Cov}[X, Y]$

### 3.3 Ковариация и корреляция

Определение

$$\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]$$

Свойства

- $\text{Cov}[X, a] = 0$
- $\text{Cov}[X, X] = \mathbb{V}\text{ar}[X]$
- $\text{Cov}[X, Y] = \text{Cov}[Y, X]$
- $\text{Cov}[aX, bY] = ab \text{Cov}[X, Y]$
- $\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$

Корреляция

$$\rho[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\mathbb{V}\text{ar}[X] \mathbb{V}\text{ar}[Y]}}$$

Независимость

$$X \perp Y \implies \rho[X, Y] = 0 \iff \text{Cov}[X, Y] = 0 \iff \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$$

Исключение: если  $X$  и  $Y$  имеют совместное нормальное распределение

$$\text{Cov}[X, Y] = 0 \implies X \perp Y$$

## 4 Обзор основных распределений

### 4.1 Дискретные распределения

	Notation	$\mathbb{P}[X = x]$	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Бернулли	$\text{Bern}(p)$	$p^x (1-p)^{1-x}$	$p$	$p(1-p)$
Биномиальное	$\text{Bin}(n, p)$	$\frac{n!}{x!(n-x)!}$	$np$	$np(1-p)$
Пуассон	$\text{Pois}(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda$	$\lambda$

### 4.2 Непрерывные распределения

	Notation	$F_X(x)$	$f_X(x)$	$\mathbb{E}[X]$	$\mathbb{V}\text{ar}[X]$
Равномерное	$U(a, b)$	$\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$	$\frac{I(a < x < b)}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Нормальное	$\mathcal{N}(\mu, \sigma^2)$	$\Phi(x) = \int_{-\infty}^x \phi(t) dt$	$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$\mu$	$\sigma^2$
Экспоненциальное	$\text{Exp}(\beta)$	$1 - e^{-x/\beta}$	$\frac{1}{\beta} e^{-x/\beta}$	$\beta$	$\beta^2$

### 4.3 Свойства некоторых распределений

Взаимосвязь между распределениями

- $X_i \sim \text{Bern}(p) \implies \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$
- $\text{Bin}(n, p) \xrightarrow{n \rightarrow \infty} \text{Pois}(np)$  ( $n$  велико,  $p$  мало)
- $\text{Bin}(n, p) \xrightarrow{n \rightarrow \infty} \mathcal{N}(np, np(1-p))$  ( $n$  велико,  $p$  далеко от 0 и 1)
- $\text{Pois}(\lambda) \xrightarrow{\lambda \rightarrow \infty} \mathcal{N}(\lambda, \lambda^2)$  ( $\lambda$  велико)

Свойства нормального распределения

- $X \sim \mathcal{N}(\mu, \sigma^2) \implies Z = \left(\frac{X-\mu}{\sigma}\right) \sim \mathcal{N}(0, 1)$
- $X \sim \mathcal{N}(\mu, \sigma^2) \wedge Z = aX + b \implies Z \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- $\mathbb{P}[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$ , где  $\Phi$  – CDF  $\mathcal{N}(0, 1)$
- $\Phi(-x) = 1 - \Phi(x)$
- Нижний квантиль  $\mathcal{N}(0, 1)$ :  $z_\alpha = \Phi^{-1}(\alpha)$
- Верхний квантиль  $\mathcal{N}(0, 1)$ :  $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$
- Симметрия  $\mathcal{N}(0, 1)$ :  $|z_\alpha| = z_{1-\alpha}$

## 5 Статистический вывод

Рассмотрим выборку (sample) из некоторого распределения  $X_1, \dots, X_n \stackrel{iid}{\sim} F$

- iid - identically independently distributed (независимы и одинаково распределены)

Эмпирическая CDF

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n} \approx \mathbb{P}[X < x]$$

$$I(X_i \leq x) = \begin{cases} 1 & X_i \leq x \\ 0 & X_i > x \end{cases}$$

Эмпирическая PDF (гистограмма)

$$\hat{f}_i(x) = \frac{\sum_{i=1}^n I\{x_{i-1} \leq x \leq x_i\}}{n \cdot (x_i - x_{i-1})} \approx \mathbb{P}(X \in (x_{i-1}; x_i])$$

### 5.1 Выборочные статистики

- Любая функция, посчитанная по выборке,  $g(X_1, \dots, X_n) = g(X_n)$  называется статистикой
- Любая статистика  $g(X_n)$  является случайной величиной до тех пор, пока речь не идёт о реализации выборки (то есть о конкретном датасете)
- Если речь идет о реализации выборки, то иногда используются маленькие латинские буквы вместо заглавных
- Часто используется крышечка, чтобы обозначить, что речь идет о выборочных статистиках

Выборочное среднее

$$\hat{\mu} = \bar{X}$$

(Смещенная) выборочная дисперсия

$$\hat{s}^2 = \widehat{\text{Var}}_{\text{biased}}[X_n] = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \bar{X}^2 - (\bar{X})^2$$

Несмещенная выборочная дисперсия

$$\hat{\sigma}^2 = \widehat{\text{Var}}[X_n] = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Выборочная ковариация

$$\widehat{\text{Cov}}[X_n, Y_n] = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}_n) = \frac{n-1}{n} \cdot (\overline{XY} - \bar{X} \cdot \bar{Y})$$

Выборочная корреляция Пирсона

$$\hat{\rho}_P = \frac{\widehat{\text{Cov}}[X_n, Y_n]}{\sqrt{\widehat{\text{Var}}[X_n]} \sqrt{\widehat{\text{Var}}[Y_n]}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}} \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\hat{\sigma}_X \cdot \hat{\sigma}_Y}$$

Выборочная корреляция Спирмена (корреляция Пирсона между рангами)

$$\hat{\rho}_S(X, Y) = \hat{\rho}_P(R_X, R_Y)$$

## 6 Параметрические Д.И. и тестирование гипотез

Этот раздел посвящен **параметрическим** доверительным интервалам (Д.И.) и тестированию **параметрических** гипотез. Это означает, что мы будем строить Д.И. и тестировать гипотезы для **параметров** распределения элементов выборки, а также тестировать гипотезы о равенстве этих **параметров** некоторым числам.

### 6.1 Введение

#### 6.1.1 Доверительные интервалы

Рассмотрим выборку  $X_1 \dots, X_n \stackrel{iid}{\sim} F_X$  из некоторого распределения (не обязательно нормального) с мат. ожиданием  $\mu$  и дисперсией  $\sigma^2$ , тогда, пользуясь ЦПТ, получим следующий критерий (тестовую статистику)  $Z_n$  на основе статистики  $\bar{X}$ , которая будет иметь стандартное нормальное распределение

$$Z_n = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \stackrel{d}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

Для этой тестовой статистики мы можем записать

$$\mathbb{P} \left[ -z_{1-\frac{\alpha}{2}} < Z_n < z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

И получить асимптотический доверительный интервал для  $\mu$

$$\mathbb{P} \left[ \bar{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\sigma}^2}{n}} \right] = 1 - \alpha$$

Или более компактно

$$\mu \in \left\{ \bar{X} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right\}$$

Примечание: мы можем заменить  $\sigma^2$  на  $\hat{\sigma}^2$ , т.к.  $\hat{\sigma}^2 \xrightarrow{d} \sigma^2$

Исходя из подобной процедуры строятся все доверительные интервалы

1. Рассматривается некоторая статистика, связанная с параметром распределения, например, с мат. ожиданием, дисперсией, разностью мат. ожиданий, отношением дисперсий и т. д.
2. Подбирается переход к некоторому известному распределению
3. Строится доверительный интервал на основе квантилей этого распределения

#### 6.1.2 Тестирование гипотез

На подобной логике основана процедура тестирования гипотез. Допустим нас интересует гипотеза о том, равно ли мат. ожидание некоторому числу  $\mu_0$ .

$H_0 : \mu = \mu_0$  основная гипотеза

$H_1 : \mu \neq \mu_0$  альтернативная гипотеза

Для тестирования гипотезы рассмотрим статистику  $\bar{X}$ , поскольку  $\mathbb{E}[\bar{X}] = \mu$ . Попробуем найти некоторый критерий (тестовую статистику) на основе статистики  $\bar{X}$ . Таким критерием снова окажется  $Z_n$ , однако теперь мы предположим, что  $\mu = \mu_0$

$$Z_n = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \stackrel{H_0}{=} \frac{\bar{X} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \stackrel{H_0}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

Если наша гипотеза противоречит данным, то окажется, что тестовая статистика  $Z_n$  принимает какие-то экстремальные, нетипичные ей значения. В качестве таких значений выберем значения, которые будут лежать ниже квантиля  $Z_{\frac{\alpha}{2}} = -Z_{1-\frac{\alpha}{2}}$  и выше квантиля  $Z_{1-\frac{\alpha}{2}}$ . Будем считать эту область критической (rejection region), вероятность попасть в неё будет равна  $\alpha$ .

Вероятность  $\alpha$  называют **уровнем значимости** (significance level), она отражает вероятность отвергнуть  $H_0$ , когда она верна. Также её называют вероятностью ошибки 1-го рода (Type I Error)

Если **наблюдаемая (observed) статистика**, посчитанная по данным, окажется больше некоторого **критического (critical) уровня**  $Z_{\text{crit}} = Z_{1-\frac{\alpha}{2}}$  (то есть попадёт в критическую область), то мы сделаем вывод, что гипотеза  $H_0 : \mu = \mu_0$  противоречит данным и мы отвергаем основную гипотезу **на уровне значимости**  $\alpha$ . В противном случае у нас нет оснований отвергать основную гипотезу.

$$|Z_{\text{obs}}| = \left| \frac{\bar{x} - \mu_0}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \right| > Z_{\text{crit}} = Z_{1-\frac{\alpha}{2}},$$

Рассмотренная гипотеза  $H_0 : \mu = \mu_0$  называется **двусторонней**, и для неё критическая область:  $\mathcal{X} = \{|Z_{\text{obs}}| > Z_{\text{crit}}\}$ .

Существуют также **односторонние** гипотезы:

##### Левосторонняя

- $H_0 : \mu \leq \mu_0$  VS  $H_1 : \mu > \mu_0$
- $\mathcal{X} = \{Z_{\text{obs}} \leq Z_{\text{crit}}\}$ ,  $Z_{\text{crit}} = Z_{\alpha}$

##### Правосторонняя

- $H_0 : \mu \geq \mu_0$  VS  $H_1 : \mu < \mu_0$
- $\mathcal{X} = \{Z_{\text{obs}} \geq Z_{\text{crit}}\}$ ,  $Z_{\text{crit}} = Z_{1-\alpha}$

## Принятие решений

	Приняли $H_0$	Отвергли $H_0$
$H_0$ верна	✓	Ошибка 1-го рода ( $\alpha$ )
$H_0$ не верна	Ошибка 1-го рода ( $\beta$ )	✓

Вероятность ошибки 1-го рода фиксируется для любого теста, а вероятность ошибки 2-го рода минимизируется по остаточному принципу. Иногда тесты сравнивают исходя из **мощности теста** – величины равной  $1 - \beta$ .

**P-value** отражает максимальную вероятность допустить ошибку, когда отвергается нулевая гипотеза

$$\text{p-value} = \mathbb{P}[|Z_n| \geq Z_{\text{obs}} \mid H_0 \text{ верна}]$$

P-value удобно использовать, чтобы абстрагироваться от  $Z_{\text{obs}}$  и  $Z_{\text{crit}}$ . Можно пользоваться следующим правилом при принятии решения:

p-value	evidence
$< 0.01$	very strong evidence against $H_0$
$0.01 - 0.05$	strong evidence against $H_0$
$0.05 - 0.1$	weak evidence against $H_0$
$> 0.1$	little or no evidence against $H_0$

Тестирование гипотез естественным образом связано с доверительными интервалами. Если тестируемое значение  $\mu$  не попало в доверительный интервал для  $\mu$ , то автоматически можно сделать вывод о том, что гипотеза  $H_0 : \mu = \mu_0$  отвергается. Поэтому далее наряду с формулами доверительных интервалов также будут приводиться тестовые статистики для тестирования гипотез. Для расчета тестовой статистики вместо истинного параметра  $\mu$  нужно подставить тестируемый параметр  $\mu_0$ .



## 6.2 Асимптотические Д.И. на основе ЦПТ

### Предпосылки

- Выборка достаточно велика  $n \rightarrow \infty$
- Наблюдения независимы
- Нет выбросов

### Основа

- Центральная предельная теорема (ЦПТ)
- Свойства нормального распределения

### 6.2.1 Д.И. для мат. ожидания выборок из любого распределения

Напомним, что для одной выборки мы получили тестовую статистику

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \stackrel{d}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

и доверительный интервал

$$\mu \in \left\{ \bar{X} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right\}$$

По аналогии рассмотрим две **выборки**  $X_1, \dots, X_{n_x} \stackrel{iid}{\sim} F_X$  и  $Y_1, \dots, Y_{n_y} \stackrel{iid}{\sim} F_Y$  из некоторых **независимых** друг от друга распределений (необязательно нормальных) с мат. ожиданиями  $\mu_x, \mu_y$  и дисперсиями  $\sigma_x^2, \sigma_y^2$ . Тогда, пользуясь ЦПТ, получим следующую тестовую статистику  $Z$  на основе  $\bar{X}$  и  $\bar{Y}$  для разности мат. ожиданий  $\mu_x - \mu_y$

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \stackrel{d}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

и доверительный интервал

$$\mu_x - \mu_y \in \left\{ \bar{X} - \bar{Y} \pm Z_{crit} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}} \right\}$$

### Примечания

1. При тестировании гипотезы  $H_0 : \mu_x = \mu_y$  в тестовой статистике  $Z$  разность  $\mu_x - \mu_y$  должна быть заменена на 0
2. Подобные гипотезы, где сравниваются показатели двух выборок, также называют гипотезами об **однородности** (homogeneity)

### 6.2.2 Д.И. для теоретической доли распределения Бернулли

Рассмотрим выборку  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p)$  из распределения Бернулли с мат. ожиданием  $p$  и дисперсией  $p(1-p)$ , тогда, пользуясь ЦПТ, получим следующий критерий (тестовую статистику)  $Z_n$  на основе статистики  $\hat{p} = \bar{X}$ , которая будет иметь стандартное нормальное распределение

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \stackrel{d}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

и доверительный интервал

$$p \in \left\{ \hat{p} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right\}$$

Рассмотрим **две выборки**  $X_1, \dots, X_{n_x} \stackrel{iid}{\sim} \text{Bern}(p_x)$  и  $Y_1, \dots, Y_{n_y} \stackrel{iid}{\sim} \text{Bern}(p_y)$  из **независимых** друг от друга распределений Бернулли с мат. ожиданиями  $p_x, p_y$  и дисперсиями  $p_x(1-p_x), p_y(1-p_y)$ . Тогда, пользуясь ЦПТ, получим следующую тестовую статистику  $Z$  на основе  $\bar{X}$  и  $\bar{Y}$  для разности долей  $p_x - p_y$

$$Z = \frac{\hat{p}_x - \hat{p}_y - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_x)}{n_x} + \frac{p_y(1-p_y)}{n_y}}} \stackrel{d}{\underset{n \rightarrow \infty}{\sim}} N(0, 1)$$

и доверительный интервал

$$p_x - p_y \in \left\{ \hat{p}_x - \hat{p}_y \pm Z_{crit} \cdot \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}} \right\}$$

**Примечание:** при тестировании гипотезы  $H_0 : p_x = p_y$  в тестовой статистике  $p_x, p_y$  должны быть заменены на некоторую общую долю, например, на  $p = (m_x + m_y)/(n_x + n_y)$ , где в знаменатели сумма единиц по обоим выборкам.

## 6.3 Точные Д.И. для нормальных выборок

### Предпосылки

- Наблюдения независимы
- Требования на размер выборки нет
- Выборка из нормального распределения

### Основа

- свойства нормального распределения
- распределения хи-квадрат, Стьюдента, Фишера
- теорема Фишера

#### 6.3.1 Д.И. для мат. ожидания

Рассмотрим выборку  $X_1 \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$  из нормального распределения с мат. ожиданием  $\mu$  и дисперсией  $\sigma^2$ , тогда по свойствам нормального распределения получим следующий критерий (тестовую статистику)  $Z_n$  на основе статистики  $\bar{X}$ , которая будет иметь стандартное нормальное распределение, если **дисперсия  $\sigma^2$  известна**

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

и доверительный интервал

$$\mu \in \left\{ \bar{X} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma^2}{n}} \right\}$$

Если **дисперсия  $\sigma^2$  не известна**, то необходимо использовать её оценку, и тогда мы получим распределение **Стьюдента**

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim t(n-1)$$

и доверительный интервал

$$\mu \in \left\{ \bar{X} \pm t_{1-\frac{\alpha}{2}}(n-1) \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} \right\}$$

#### 6.3.2 Д.И. для дисперсии

Построим доверительный интервал для  $\sigma^2$  на основе  $\hat{s}^2$  и  $\hat{\sigma}^2$ , предположим, что **мат. ожидание известно**, тогда следующая тестовая статистика имеет распределение **хи-квадрат по определению**

$$\chi = n \frac{\hat{s}^2}{\sigma^2} \sim \chi^2(n)$$

Доверительный интервал имеет вид

$$\frac{n \cdot \hat{s}^2}{\chi_{1-\frac{\alpha}{2}}^2(n)} \leq \sigma^2 \leq \frac{n \cdot \hat{s}^2}{\chi_{\frac{\alpha}{2}}^2(n)}$$

Если предположить, что **мат. ожидание не известно**, тогда по **теореме Фишера** получим снова распределение хи-квадрат

$$\chi = (n-1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n)$$

Доверительный интервал имеет вид

$$\frac{(n-1) \cdot \hat{\sigma}^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)} \leq \sigma^2 \leq \frac{(n-1) \cdot \hat{\sigma}^2}{\chi_{\frac{\alpha}{2}}^2(n-1)}$$

**Примечание:** квантили не равны по модулю, т.к. распределение хи-квадрат положительно и не симметрично

### 6.3.3 Д.И. для разности мат. ожиданий

Рассмотрим две **выборки**  $X_1, \dots, X_{n_x} \stackrel{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x)$  и  $Y_1, \dots, Y_{n_y} \stackrel{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y)$  из **независимых** друг от друга нормальных распределений с мат. ожиданиями  $\mu_x, \mu_y$  и дисперсиями  $\sigma_x^2, \sigma_y^2$ . Тогда, пользуясь свойствами нормального распределения, получим следующую тестовую статистику  $Z$  на основе  $\bar{X}$  и  $\bar{Y}$  для разности мат. ожиданий  $\mu_x - \mu_y$ , если **дисперсии известны**

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0, 1)$$

и доверительный интервал

$$\mu_x - \mu_y \in \left\{ \bar{X} - \bar{Y} \pm Z_{crit} \cdot \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right\}$$

Предположим, что **дисперсии не известны, но равны**, тогда

$$t = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_o^2}{n_x} + \frac{\hat{\sigma}_o^2}{n_y}}} \sim t(n_x + n_y - 2)$$

где  $\hat{\sigma}_o^2$  - объединенная (pooled) дисперсия:

$$\hat{\sigma}_o^2 = \frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}$$

и доверительный интервал имеет вид

$$\mu_x - \mu_y \in \left\{ \bar{X} - \bar{Y} \pm t_{crit} \cdot \sqrt{\frac{\hat{\sigma}_o^2}{n_x} + \frac{\hat{\sigma}_o^2}{n_y}} \right\}$$

Предположим, что **дисперсии не известны и не равны**, тогда получим примерное распределение Стьюдента (распределение Уэлча),

$$t = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \sim t(d)$$

где  $d$  – примерное число степеней свободы:

$$d = \frac{\left( \frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y} \right)^2}{\frac{\hat{\sigma}_x^4}{n_x^2(n_x - 1)} + \frac{\hat{\sigma}_y^4}{n_y^2(n_y - 1)}}$$

Доверительный интервал имеет вид

$$\mu_x - \mu_y \in \left\{ \bar{X} - \bar{Y} \pm t_{crit} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}} \right\}$$

**Примечание:** Приближение работает хорошо, если  $n_x = n_y$  или  $n_x < n_y$  и  $\sigma_x < \sigma_y$

### 6.3.4 Д.И. для отношения дисперсий

Построим доверительный интервал для отношения дисперсий, которое имеет распределение Фишера

$$\frac{\hat{\sigma}_x^2/\sigma_x^2}{\hat{\sigma}_y^2/\sigma_y^2} \sim F(n_x - 1, n_y - 1)$$

Доверительный интервал имеет вид

$$\frac{1}{F_{\frac{\alpha}{2}}(n_x - 1, n_y - 1)} \cdot \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \cdot F_{\frac{\alpha}{2}}(n_y - 1, n_x - 1)$$

#### Примечания

1. квантили не равны по модулю, т.к. распределение Фишера положительно и не симметрично
2. При тестировании гипотезы  $H_0 : \sigma_x^2 = \sigma_y^2$  статистика упрощается до отношения выборочных дисперсий

### 6.3.5 Д.И. для разности мат. ожиданий в зависимых выборках

Рассмотрим две **выборки**  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x)$  и  $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y)$  из **зависимых** друг от друга нормальных распределений с мат. ожиданиями  $\mu_x, \mu_y$  и дисперсиями  $\sigma_x^2, \sigma_y^2$ . Например, мы делаем измерения на **одних и тех же объектах** в 2 момента времени. Рассмотрим прирост на каждом объекте

$$d_i = X_i - Y_i$$

с мат. ожиданием

$$\mathbb{E}[d] = \Delta$$

Дисперсию оценим по выборке

$$\hat{\sigma}_\Delta^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

Тогда снова получим распределение **Стьюдента**

$$t = \frac{\bar{d} - \Delta}{\sqrt{\frac{\hat{\sigma}_\Delta^2}{n}}} \sim t(n-1)$$

и доверительный интервал

$$\Delta \in \left\{ \bar{d} \pm t_{\text{crit}} \cdot \sqrt{\frac{\hat{\sigma}_\Delta^2}{n}} \right\}$$

## 6.4 Некоторые дополнительные тесты

### 6.4.1 Тест о значимости корреляции

Для выборочного коэффициента корреляции Пирсона можно протестировать гипотезу о равенстве нулю этого коэффициента (то есть об отсутствии линейной взаимосвязи двух переменных)

$$H_0 : \hat{\rho}_{x,y} = 0 \text{ нет линейной взаимосвязи}$$

$$H_1 : \hat{\rho}_{x,y} \neq 0 \text{ есть}$$

Для этого снова понадобится  $t$ -статистика

$$t = \hat{\rho}_{x,y} \sqrt{\frac{n-2}{1-\hat{\rho}_{x,y}^2}} \sim t(n-2)$$

### 6.4.2 Тест о равенстве пропорций в зависимых выборках

Делаем измерение двух бинарных признаков на одних и тех объектах и тестируем гипотезу о наличии взаимосвязи между двумя признаками. Составим таблицу сопряженности с частотами.

	$X_1 = 1$	$X_1 = 0$
$X_2 = 1$	$n_{11}$	$n_{21}$
$X_2 = 0$	$n_{12}$	$n_{22}$

Гипотеза об однородности может быть сформулирована как

$$H_0 : n_{1+} = n_{+1}, \quad \text{где } n_{1+} = n_{11} + n_{12}, \quad n_{+1} = n_{11} + n_{21} \quad \text{или } H_0 : n_{12} = n_{21}$$

Тестовая статистика имеет биномиальное распределение

$$n_{12} \sim \text{Bin}(n^*, 0.5), \quad \text{где } n^* = n_{12} + n_{21}$$

которое сходится к стандартному нормальному или хи-квадрат распределению

$$Z = \frac{n_{12} - 0.5n^*}{\sqrt{0.5(1-0.5)n^*}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \sim N(0, 1) \quad \text{или} \quad Z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi_1^2$$

Для стандартного нормального распределения также существует коррекция

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}}$$