

Лекция 5

Тестирование гипотез. Непараметрика

Курс: Введение в DS на УБ и МиРА (весна, 2022)

Преподаватель: Владимир Омелянчик

25 апреля 2022 г.

В предыдущих сериях

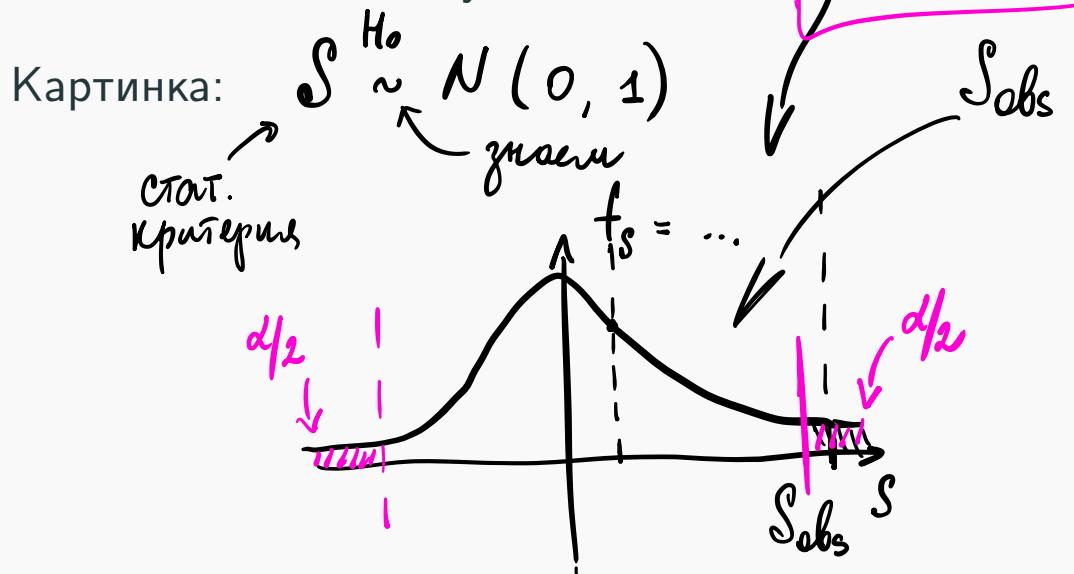
- Ковариация и корреляция.
- Формулировка основной гипотезы.
- Ошибки I и II рода.

Статистический тест и распределения

- Статистический тест (критерий) – математическое правило, в соответствии с которым отвергается или не отвергается проверяемая гипотеза.
- Бывают параметрическими (проверка параметров распределений) и непараметрическими (независимость).
- Обычно рассчитается какая-то статистика, которая при верной H_0 имеет какое-то распределение. Далее смотрят, насколько вероятно при верной H_0 увидеть такое значение статистики.

Статистическая значимость

- Напоминание: ошибка I рода (False Positive) - ситуация, когда отвергнута верная нулевая гипотеза.
- Хотим, чтобы $\mathbb{P}(\text{ош. I рода}) \leq \alpha$, где α выбираем сами.
Обычно используются значения 0.01, 0.05, 0.1.



Z-тест для одной выборки

X_1, \dots, X_N – выборка из N независимых, одинаково распределённых **нормальных** случайных величин с **известной** дисперсией.

$$X_i \sim N(\mu, \sigma^2)$$

известна

известн.

Тогда Z-статистика для гипотезы

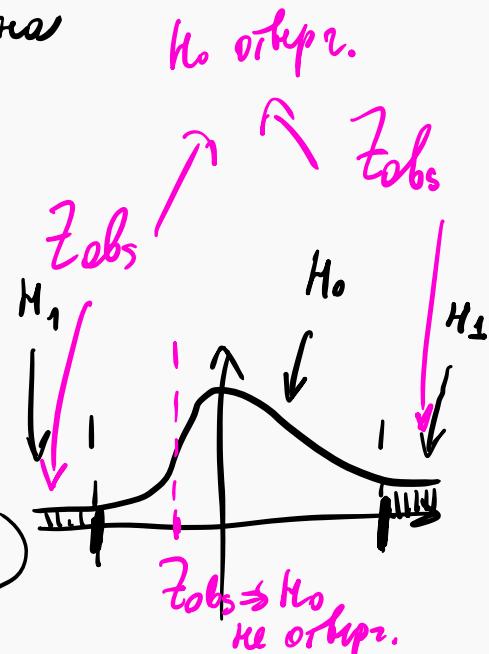
$$\hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{\sqrt{N}}\right) - \text{можно показ.}$$

$\hat{\mu}_{obs}$

$$\begin{cases} H_0 : \underline{\mu} = \mu_0, \\ H_1 : \underline{\mu} \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$Z_{obs} = \frac{(\hat{\mu}) - (\mu_0)}{\frac{\sigma}{\sqrt{N}}} \stackrel{H_0}{\sim} N(0, 1)$$



t-тест для одной выборки

X_1, \dots, X_N – выборка из N независимых, одинаково распределённых нормальных случайных величин с неизвестной дисперсией.

$X_i \sim N(\mu, \sigma^2)$ *не знаем*
не знаем

$$N \geq 100 \Rightarrow N(0, 1)$$

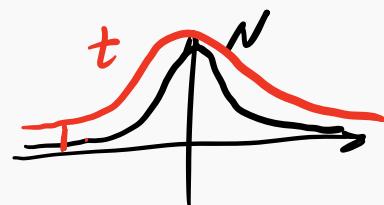
Тогда t-статистика для гипотезы

$$\begin{cases} H_0 : \mu = \mu_0, \\ H_1 : \mu \neq \mu_0 \end{cases}$$

рассчитывается по формуле

$$t_{obs} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}} \sim t_{N-1}$$

$$Z_{obs} = \frac{\hat{\mu} - \mu_0}{\sigma / \sqrt{N}} \sim N(0, 1)$$



NB!: $t_{N-1} \rightarrow Z$ при $N \rightarrow \infty$.

степени свободы

Пример: проверка гипотезы о среднем

$$\begin{aligned} & X_1, \dots, X_{300} \sim \mathcal{N}(\mu, \sigma^2) \\ & \sum_{i=1}^{300} X_i = 60, \quad \sum_{i=1}^{300} X_i^2 = 4000 \end{aligned}$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 0.3, \\ H_1 : \mu \neq 0.3 \end{cases} = \mu_0$$

на уровне значимости $5\% = \alpha$

Для справки: $\tilde{Z}_{0.975} \approx \tilde{t}_{299, 0.975} \approx 1.96$

Пример: проверка гипотезы о среднем

$$t_{obs} = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\sigma^2}{N}}}$$

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n x_i}{N} = \frac{60}{300} = \frac{1}{5} = 0.2$$

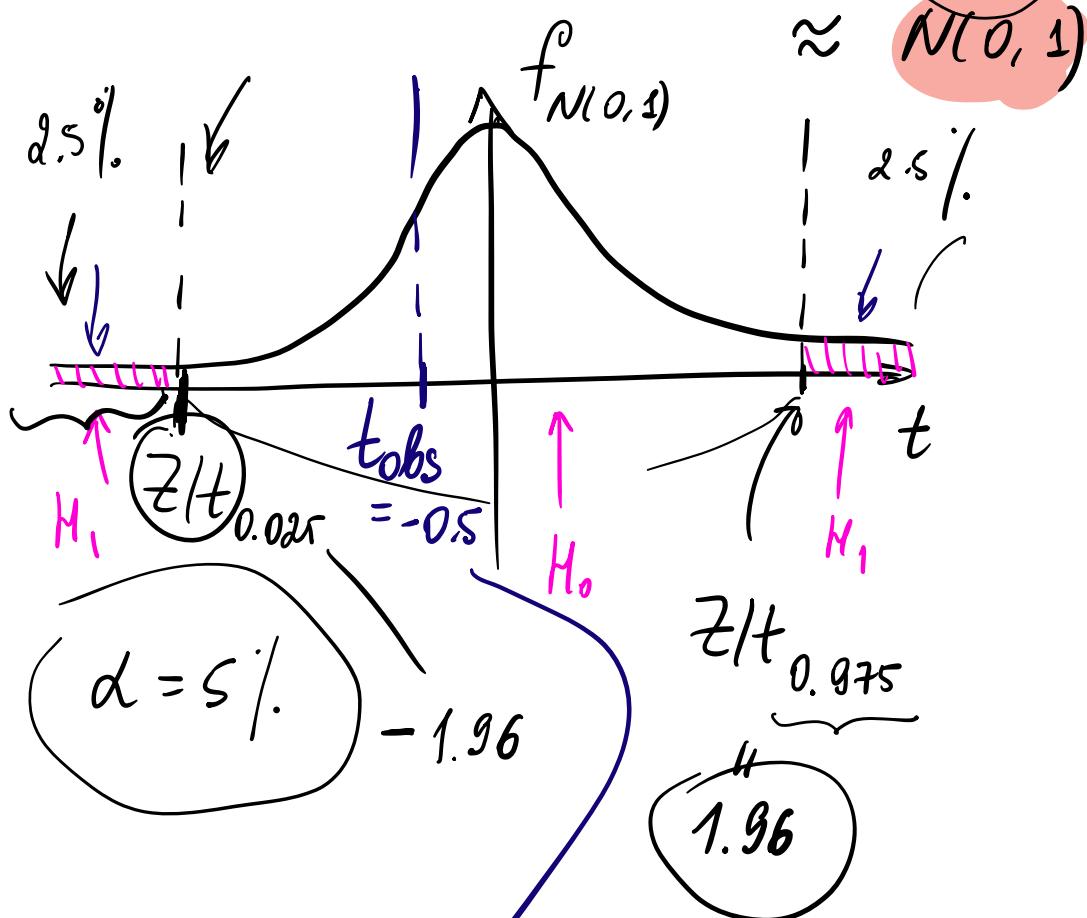
$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2 \Leftrightarrow$$

$$\begin{aligned} \Leftrightarrow \frac{1}{N-1} \sum_i \left[x_i^2 - 2x_i \bar{x} + (\bar{x})^2 \right] &= [(\bar{x})^2 \left(\sum_{i=1}^N 1 \right)] \\ &= \frac{1}{N-1} \left(\sum_i x_i^2 - \sum_i 2\bar{x} x_i + \sum_i (\bar{x})^2 \right) = \\ &= \frac{1}{N-1} \left(\sum_i x_i^2 - 2\bar{x} \sum_i x_i + N \cdot (\bar{x})^2 \right) = \end{aligned}$$

$$= \frac{1}{299} (4000 - 2 \cdot 0.2 \cdot 60 + 300 \cdot 0.04) \approx 13$$

$$t_{\text{obs}} = \frac{0.2 - 0.3}{\sqrt{\frac{13}{300}}} = \frac{-0.1}{0.2} = -0.5$$

Наші висновки H_0 : $t \sim t_{N-1} = t_{299}$



$\Rightarrow H_0$ не отвергнена з п. знат.

Проверка гипотезы о доле

X_1, \dots, X_N – выборка из N независимых, одинаково распределённых случайных величин Бернулли

$$X_i \sim \text{Bern}(\cancel{p}). \quad \underset{\curvearrowright}{\qquad} \quad X_i = \begin{cases} 0, & 1-p \\ 1, & \cancel{p} \end{cases}$$

Тогда Z -статистика для гипотезы

$$\begin{cases} H_0 : p = \cancel{p_0} \\ H_1 : p \neq p_0 \end{cases}$$

$$\mathbb{E}(X_i) = 1 \cdot \cancel{p} + 0 \cdot (1 - \cancel{p})$$

рассчитывается по формуле

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \rightarrow \mathcal{N}(0, 1)$$

NB!: Можно проверять только при большом N . t -статистику использовать нельзя.

Пример: проверка гипотезы о доле

$$0/1 \quad 0/1 \\ " \quad "$$

$X_1, \dots, X_{300} \sim \text{Bern}(p)$

$$\underbrace{\sum_i X_i = 60,}_{\text{...}}$$

$$\hat{p} = \frac{60}{300} = \frac{1}{5}$$

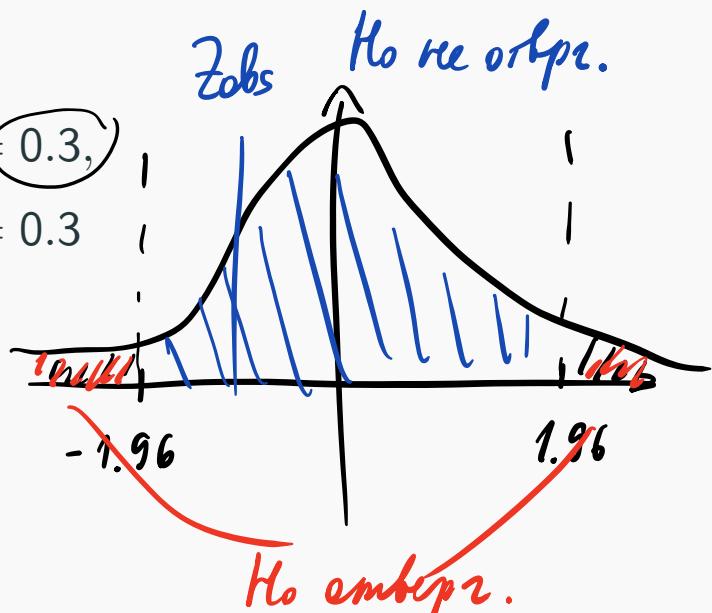
$$Z_{\text{obs}} = \dots$$

Проверьте гипотезу

$$\begin{cases} H_0 : p = 0.3, \\ H_1 : p \neq 0.3 \end{cases}$$

на уровне значимости 5%.

Для справки: $Z_{0.975} \approx 1.96$



p-value

p-value

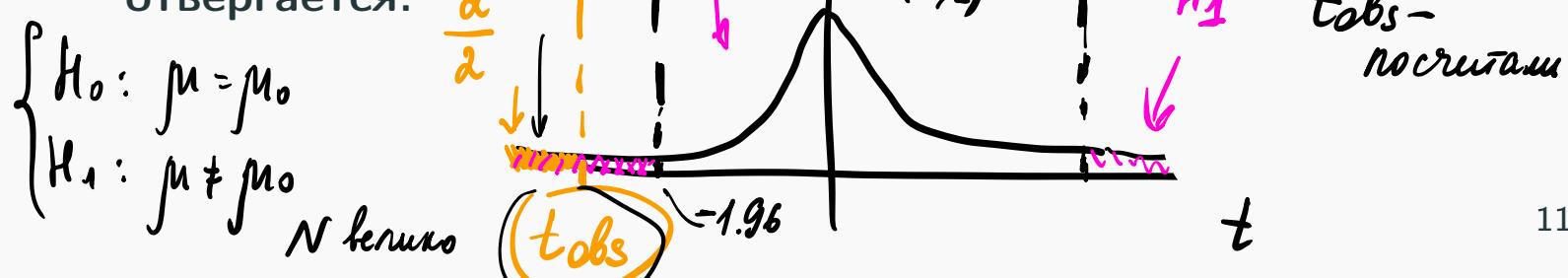
Минимальный уровень значимости, при котором нулевая гипотеза не отвергается.

Пусть проверяем гипотезу о равенстве средних или долей против гипотезы о неравенстве. Посчитали Z_{obs} или t_{obs} . Тогда

$$\tilde{\alpha} = p\text{-value} = 2 \mathbb{P}(t \leq t_{obs})$$

$$2 \int_{-\infty}^{Z_{obs}} f_z(z) dz$$
$$\frac{p\text{-val}}{2} < \frac{\alpha}{2}$$

Основной результат: $p\text{-value} < \alpha \Rightarrow$ нулевая гипотеза отвергается.



Пример: проверка гипотезы о доле

Замечание: общая формула теста

$$\left\{ \begin{array}{l} H_0: \hat{\Theta} = \Theta_0 \\ H_1: \hat{\Theta} \neq \Theta_0 \end{array} \right.$$

↑
проверь.
пар.р распр.

$\hat{\Theta}$ - момент посчит.

$$\hat{Var}(\hat{\Theta}) \checkmark$$

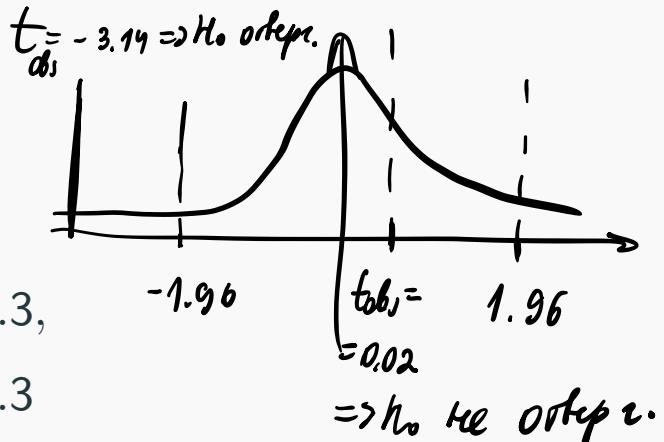
$$\frac{\hat{\Theta} - \Theta_0}{\sqrt{\hat{Var}(\hat{\Theta})}} \sim t_{\infty} \sim N(0, 1)$$

Проверка гипотез при помощи p-value

Пример 1

Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 0.3, \\ H_1 : \mu \neq 0.3 \end{cases}$$



на уровне значимости 5% (используйте ± 1.96 как критическое значение), если

1. Оказалось, что $t_{obs} = -3.14$.
2. Оказалось, что $t_{obs} = 0.02$.

Пример 2

p-value $< \alpha$ $\Rightarrow H_0 \text{ отфр.}$
 $1\% < 5\%$

(a) p-value (t_{obs}) = 0.01

(б) p-value (t_{obs}) = 0.1

$10\% > \alpha = 5\% \Rightarrow$
 $H_0 \text{ не отфр.}$

Доверительный интервал

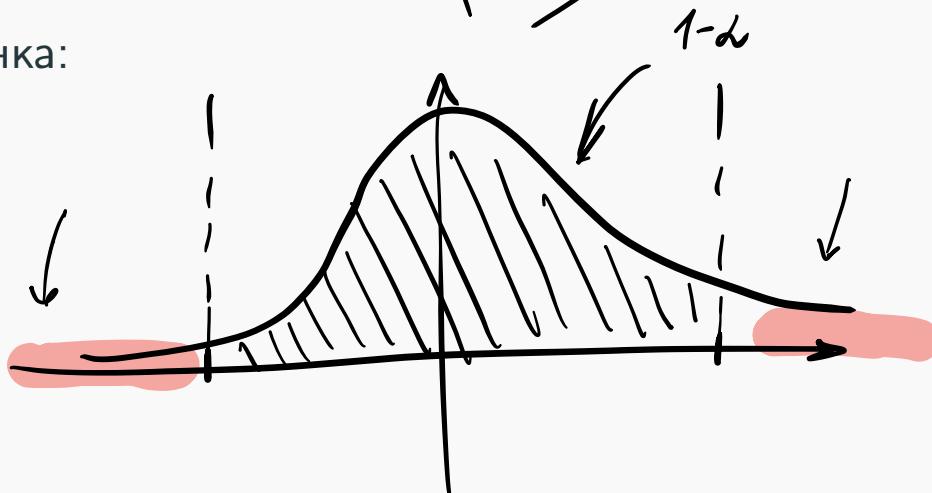
Доверительный интервал

Интервал со случайным границами, такой что

$$\mathbb{P}(T_l(X) < \theta < T_r(X)) \geq 1 - \alpha$$

α → 1%
 α → 10%
 α → 5%

Картина:



Проверка гипотез при помощи доверительного интервала

Заметим, что H_0 не отвергается на уровне значимости 5% при использовании Z-теста, когда

$$\begin{aligned} & \text{Zobs} \\ & -1.96 \leq \frac{\hat{\mu} - \mu_0}{\sigma / \sqrt{N}} \leq 1.96 \\ & -1.96 \frac{\sigma}{\sqrt{N}} \leq \hat{\mu} - \mu_0 \leq 1.96 \frac{\sigma}{\sqrt{N}}, \\ & \hat{\mu} - 1.96 \frac{\sigma}{\sqrt{N}} \leq \mu_0 \leq \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{N}} \end{aligned}$$

↑
↓

Qи:
95%.

~~наблюдаемое~~

Доверит
интервал
для μ

Основной результат

Если наблюдаемое значение статистики попадает в

$(1 - \alpha)$ -процентный доверительный интервал, то нулевая гипотеза не отвергается на уровне значимости α .

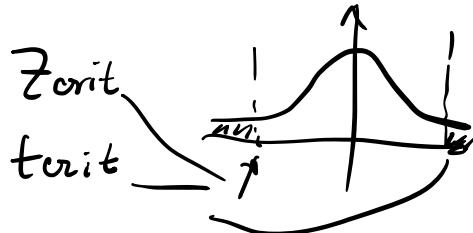
Θ - прошл
параметр

$$\text{Qи: } \hat{\Theta} \in [\hat{\Theta} - Z_{\text{crit}} \cdot \sqrt{\text{Var}(\hat{\Theta})}; \hat{\Theta} + Z_{\text{crit}} \cdot \sqrt{\text{Var}(\hat{\Theta})}]$$

стационарный вектор $\Rightarrow X_1 \dots X_n \sim F(\text{нап-ор})$

$$\begin{cases} H_0: \Theta = \Theta_0 \\ H_1: \Theta \neq \Theta_0 \end{cases}$$

1) Z_{obs} и сравнив
 t_{obs}



2) p-value для t_{obs} и сравнив с α

3) Несколько для $g(\Theta)$

Две выборки

1. Выборки независимы, равенство средних (долей):
 - 1.1 Дисперсии известны, либо обе выборки большие.
 - 1.2 Дисперсии неизвестны, но предполагаем, что равны.
 - 1.3 Дисперсии неизвестны и не предполагаем, что равны (сложно).
 2. Выборки зависимы: связанные пары.
-

$$\left\{ \begin{array}{l} N_1 \geq 100 \\ N_2 \geq 100 \end{array} \right.$$

$$\left\{ \begin{array}{l} H_0: \mu_x = \mu_y \\ H_1: \mu_x \neq \mu_y \end{array} \right.$$

X_1, \dots, X_{N1} и Y_1, \dots, Y_{N2} – две независимые выборки.

Предположим, что дисперсии этих выборок известны или $N1$ и $N2$ велики. Тогда для проверки гипотезы о равенстве средних (долей) можно использовать статистику

$$Z_{obs} = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{N1} + \frac{\sigma_y^2}{N2}}} \stackrel{>0}{\sim} \mathcal{N}(0, 1).$$

$$\hat{b} = \mu_x - \mu_y$$

$$\hat{b} = \bar{X} - \bar{Y}$$

$$\left\{ \begin{array}{l} H_0: b = 0 \\ H_1: b \neq 0 \end{array} \right.$$

X_1, \dots, X_{N1} и Y_1, \dots, Y_{N2} – две независимые выборки.

Предположим, что дисперсии этих выборок неизвестны и равны. Тогда для проверки гипотезы о равенстве средних можно использовать статистику

$$t = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\hat{\sigma} \sqrt{\frac{1}{N1} + \frac{1}{N2}}} \sim t_{N1+N2-2},$$

$$\hat{\sigma} = \sqrt{\frac{(N1 - 1)\hat{\sigma}_X^2 + (N2 - 1)\hat{\sigma}_Y^2}{N1 + N2 - 2}}$$

Тесты для парных выборок

Пусть есть N наблюдений над одними объектами ДО и ПОСЛЕ проведения эксперимента: X_1, \dots, X_N и Y_1, \dots, Y_N . Составим разности $d_i = y_i - x_i$ и проверим гипотезу о равенстве средних до и после.

x	y	$y - x$
x_1	y_1	$y_1 - x_1$
\vdots	\vdots	\vdots
x_N	y_N	$y_N - x_N$

$$\left\{ \begin{array}{l} H_0 : \mu_d = 0, \\ H_1 : \mu_d \neq 0 \end{array} \right\} \left\{ \begin{array}{l} H_0 : \mu_x = \mu_y \\ (\text{то} \rightarrow \text{после}) \\ H_1 : \mu_x \neq \mu_y \end{array} \right\}$$

Тогда гипотезу о разности средних ДО и ПОСЛЕ можно проверить при помощи статистики

$$\bar{d} = \frac{\sum_{i=1}^N (Y_i - X_i)}{N}$$

где

$$t_{\text{obs}} = \frac{\bar{d} - \mu_d}{\frac{\sigma_d}{\sqrt{N}}} \sim t_{N-1},$$

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

$$\hat{\sigma}_d = \sqrt{\frac{1}{N-1} \left(\sum_i d_i^2 - \frac{(\sum_i d_i)^2}{N} \right)}.$$

Непараметрика: задачи

Непараметрическая статистика

Охватывает методы, которые не полагаются на данные, относящиеся к какому-либо конкретному распределению. При этом параметры (средние, медианы и проч.) присутствуют, но не фиксированы заранее.

Примеры задач:

- Проверка независимости двух выборок (здесь же аналоги корреляций для категориальных переменных).
- Проверка того, пришли ли выборки из одного семейства распределений.
- Гистограмма и ядерная оценка плотности.

Таблица сопряжённости

Сразу пример: 2 участка: узл 1 и узл 2

3 в. дрр: эбл, групп, слово

Пример из головок

	кн2	убл	Чп	Сн	
Ког Нер 1					- Свободная таблица
→ узл 1		20	50	10	- pivot-table
узл 2	30	60	80		

"Участок": $\begin{bmatrix} "узл 1" & "узл 2" \end{bmatrix}$

"Тип аргумента": $\begin{bmatrix} "убл" & "Чп" & "Сн" \end{bmatrix}$

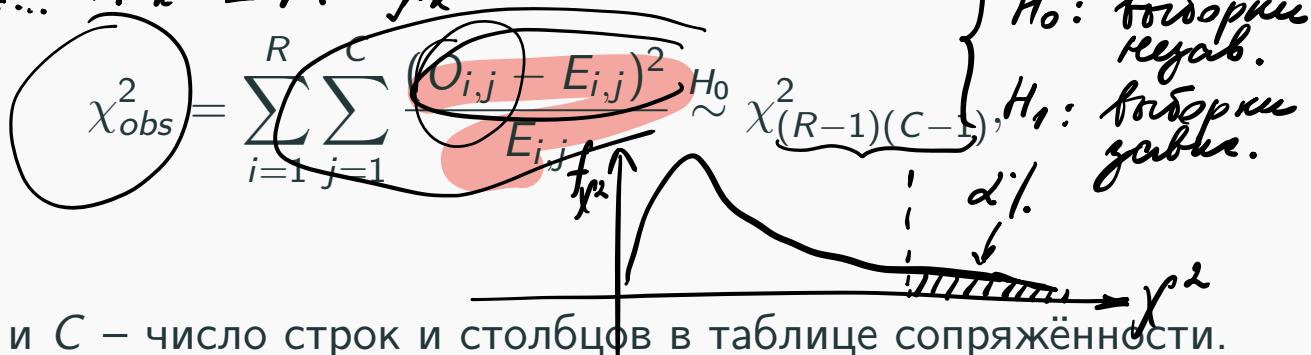
Критерий согласия для проверки независимости

$$\chi_1^2 + \chi_2^2 + \dots + \chi_k^2 = M \sim \chi_k^2$$

$$\chi_i \sim N(0,1)$$

χ_i незав.

где



- R и C – число строк и столбцов в таблице сопряжённости.
- $O_{i,j}$ – количество наблюдений в клетке (i,j) таблицы сопряжённости.
- N – число наблюдений в выборке.

$$E_{i,j} = N p_{i \cdot} p_{\cdot j}$$

$$p_{i \cdot} = \sum_{j=1}^C \frac{O_{i,j}}{N}$$

$$p_{\cdot j} = \sum_{i=1}^R \frac{O_{i,j}}{N}$$

	"1"	"2"	"3"
"a"	O_{11}	O_{12}	O_{13}
"б"	O_{21}	O_{22}	O_{23}

$$N = \sum_i \sum_j O_{i,j}$$

$$p_{1 \cdot} = \frac{O_{11} + O_{12} + O_{13}}{N}$$

прн. теоретич. частота встречаем.

$$p_{\cdot 1} = \frac{O_{11} + O_{21}}{N}$$

Пример: садоводство

На участке 1 собрали 25 яблок низкого качества, 50 яблок среднего качества и 25 яблок высокого качества, а на участке 2 собрали 52 яблока низкого качества, 41 яблоко среднего качества и 7 яблок высокого качества. Существует ли зависимость между типом участка и качеством урожая?

Используйте уровень значимости 5%. Для справки:

$$\chi^2_{2,0.95} = 5.99.$$

		низк	ср	выс	сумма	
ур 1	низк	25	50	25	100	\downarrow
	ср	52	41	7	100	
ур 2	сумма	77	91	32	$N = 200$	
→	сумма	77	91	32	$N = 200$	

$P_{низк} = \frac{100}{200}$

$P_{ср} = \frac{100}{200}$

$P_{выс} = \frac{77}{200}$

$P_{ср} = \frac{91}{200}$

$P_{выс} = \frac{32}{200}$

Пример: садоводство

E_{ij}	нч	ср	внс	$E_{ij} = N p_i \cdot p_j$
y_1	$\frac{7700}{200}$	$\frac{9100}{200}$	$\frac{3200}{200}$	
	$\frac{7700}{200}$	$\frac{9100}{200}$	$\frac{3200}{200}$	

Рук. Рнч. $\cdot N$

$$y_{obs}^2 = \left(d5 - \frac{7700}{200} \right)^2 + \left(50 - \frac{9100}{200} \right)^2 + \dots + \left(7 - \frac{3200}{200} \right)^2 / \frac{3200}{200} = 20.48$$

