AB-тестирование

Ph@DS, осень 2023

Лектор — Бруттан Мария

Что такое AB-тестирование?

Примеры в рекламе

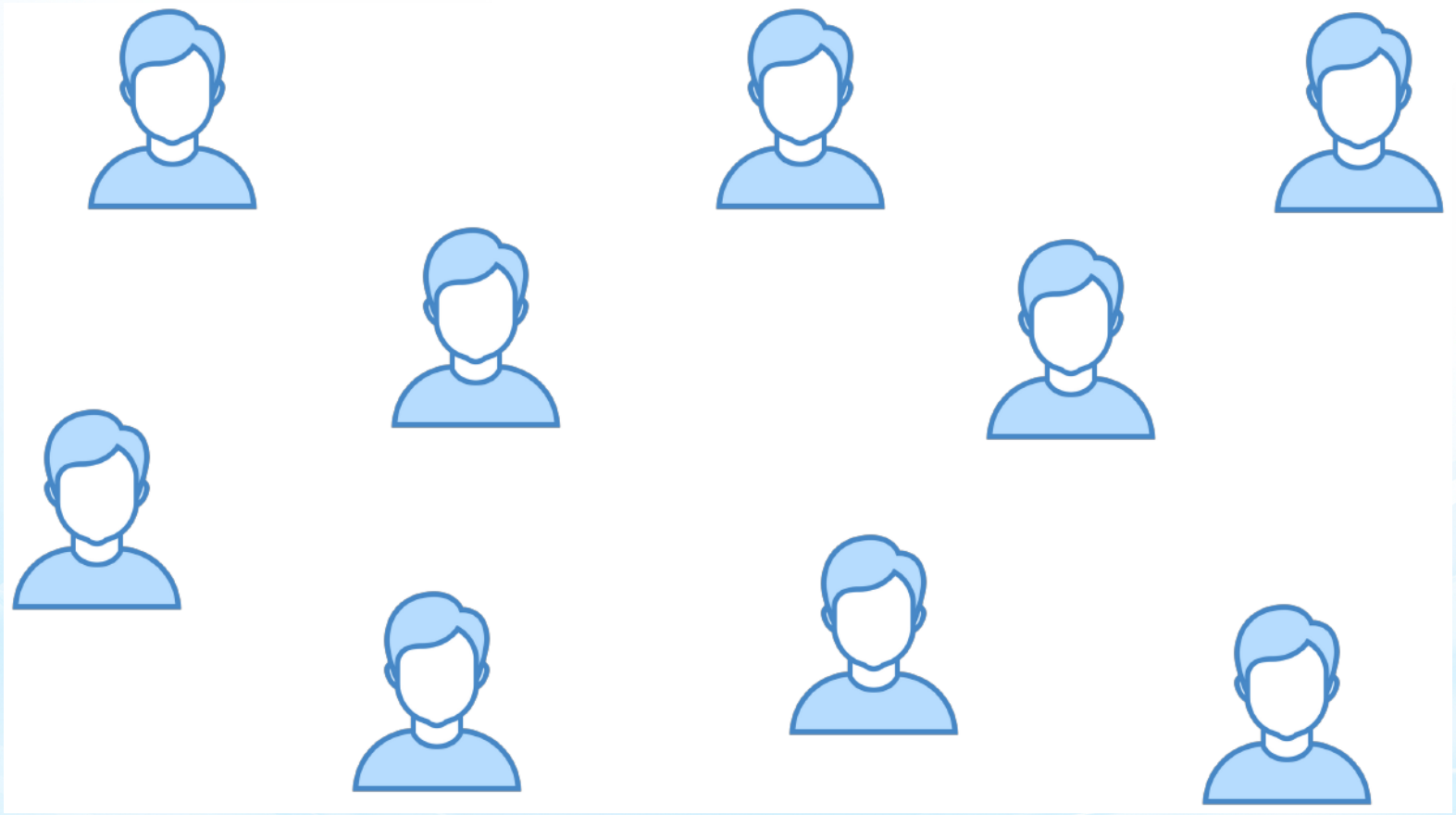
Горные велосипеды. Распродажа 60%
В наличии более 1 000 моделей!
bike.ru

Распродажа! Sale! Rebajas! Saldi!
Не пропустите! До 1 числа продаем горные велосипеды по смешным ценам! :)
bike.ru

Офисные кресла
Скидка от 10 кресел - 15%! Скидка в шоу-рум до 50%! Доставка – бесплатно!!!

Офисные кресла
Скидка от 10 кресел - 15%! Скидка в шоу-рум до 50%! Доставка – бесплатно!!!

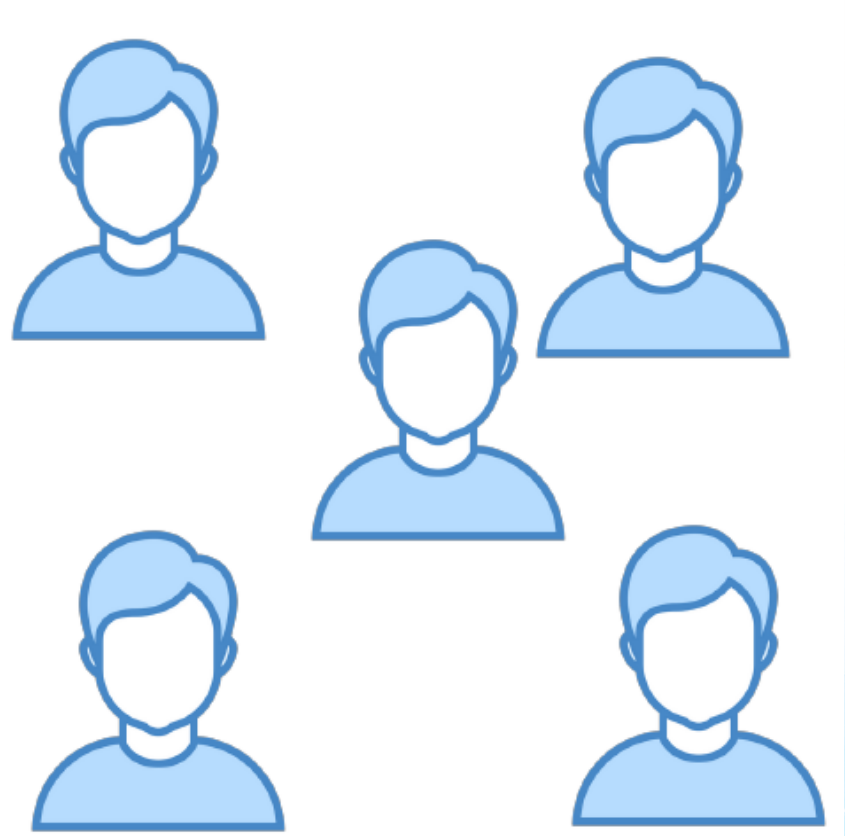
Общая схема



Общая схема

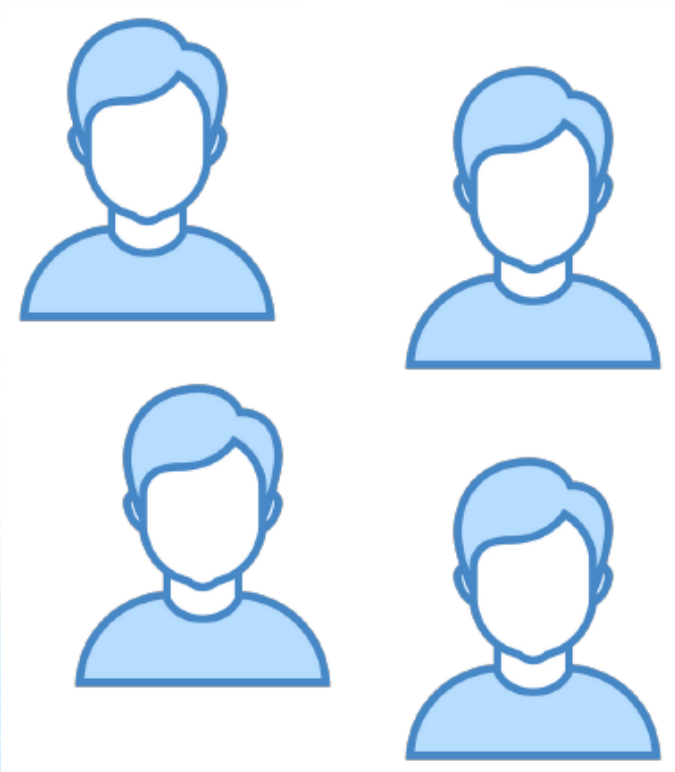
Группа А

Контрольный сегмент



Группа Б

Тестовый сегмент



Пайплайн исследования

I. Дизайн

- Фильтрация,
- Валидация,
- Семплирование.

II. Эксперимент

III. Оценка

- Фильтрация,
- Валидация,
- Оценка.

Критерии AB-тестирования (дисперсионный анализ)

Типы рассматриваемых задач

1. **Независимые выборки**

   Две группы пациентов. Одним дают одно лекарство,

   другим — другое. Верно ли, что первое лекарство эффективнее?

2. **Связные выборки**

   Пациент проходит испытание, принимает средство,

   затем снова проходит испытание. Отличается ли эффект?

- Методы для задач 2 типа можно использовать для задач 1 типа.

  При этом теряется важная информация.
- Методы для задач 1 типа *нельзя* использовать для задач 2 типа.

Независимые выборки

Человек	Препарат	Изменение температуры
Петя	Апотивадом	-0.9
Вася	Апотивадом	-0.6
Катя	Апотивадом	-1.0
Миша	Апотивадом	-0.3
Ира	Волымикер	-2.6
Света	Волымикер	-1.9
Коля	Волымикер	-0.7

Значимо ли отличается эффект от приема препаратов?

Связные выборки

Каждый человек применяет один и тот же препарат.

Человек	**Температура до**	**Температура после**
Петя	38.2	37.6
Вася	37.6	38.0
Катя	38.5	37.1
Миша	38.0	36.9
Ира	37.9	37.1
Света	39.4	37.3

Есть ли эффект от приема препарата?

Другие вопросы на практике

1. Отличаются ли гены по степени экспрессии?
2. Какие факторы влияют на появление дефектов при производстве сенсоров?
3. Увеличивается ли эффективность преобразования энергии в солнечных батареях при использовании модификаций материалов катодной поверхности?
4. многие другие...

Немного повторим

Гипотезы и критерии (*напоминание*)

$X = (X_1, ..., X_n)$ — выборка из неизвестного распределения $\mathsf{P} \in \mathscr{P}$.

$\mathsf{H}_0 \colon \mathsf{P} \in \mathscr{P}_0$ — основная гипотеза;

$\mathsf{H}_1 \colon \mathsf{P} \in \mathscr{P}_1$ — альтернативная гипотеза.

$S \subset \mathscr{X}$ — критерий уровня значимости α для проверки H_0 *vs*. H_1,

если $\mathsf{P}(X \in S) \leqslant \alpha, \quad \forall\, \mathsf{P} \in \mathscr{P}_0$.

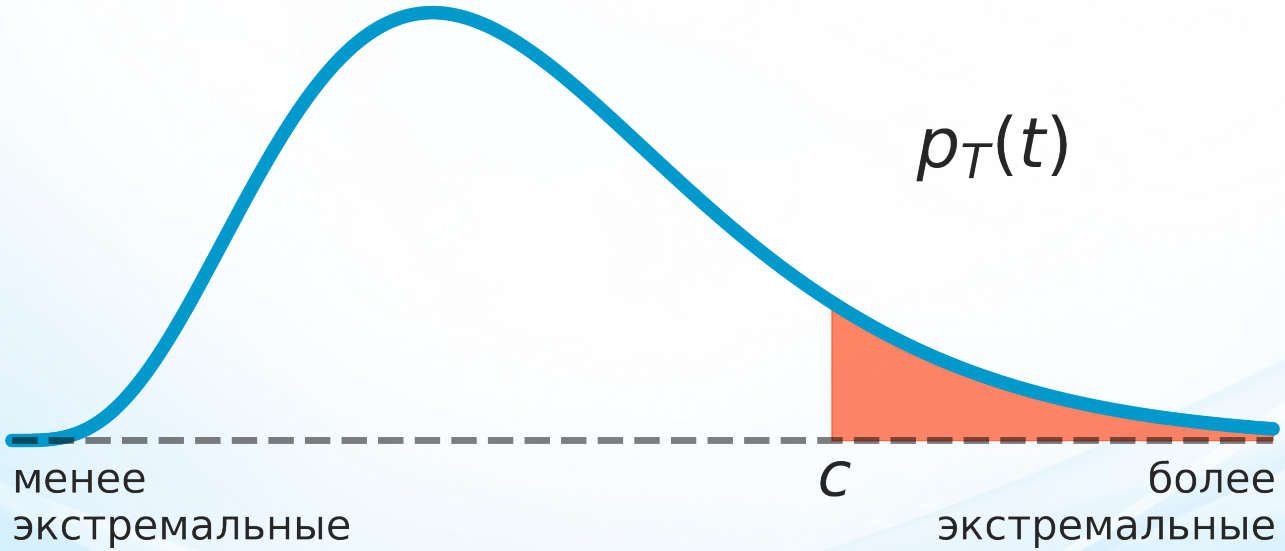
Варианты ответа:

1. $X \in S \implies \mathsf{H}_0$ отвергается $\implies$ результат стат. значим;
2. $X \notin S \implies \mathsf{H}_0$ **не отвергается** $\implies$ результат не стат. значим

Гипотезы и критерии (*напоминание*)

Часто критерий имеет вид $S = \{T(x) \geqslant c\}$,

где $T(X)$ — статистика критерия.



H_0 отвергается $\iff T(X) \geqslant c_\alpha$.

Для S значение t_1 **более экстремально**, чем t_2, если $t_1 > t_2$.

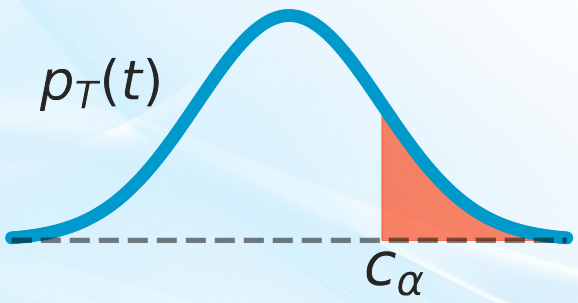
Гипотезы и критерии (*напоминание*)

Часто критерий имеет вид $S = \{T(x) \geqslant c_\alpha\}$,
где $T(X)$ — статистика критерия.

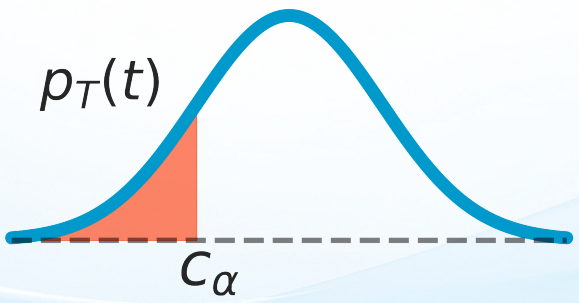
α выбирается **ДО** эксперимента,

c_α вычисляется из условия $\mathrm{P}_0(T(X) > c_\alpha) \leqslant \alpha$.

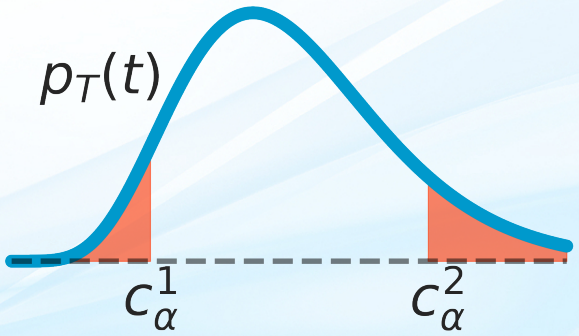
$S = \{T(x) > c_\alpha\}$



$S = \{T(x) < c_\alpha\}$



$S = \{|T(x)| > c_\alpha\}$



Замечание. Выбирать α после эксперимента неправильно.
Так можно подогнать результат под желаемый.
"*Статистика может доказать что угодно, даже истину.*"

Пример: *AB-тест*

Пациенты делятся случайно на две независимые группы:

1. *Контрольная группа A* — принимает **старый препарат**;
   $X = (X_1, ..., X_n), X_i \sim Bern(p_1)$ — результаты.
2. *Исследуемая группа B* — принимает **новый препарат**;
   $Y = (Y_1, ..., Y_m), Y_i \sim Bern(p_2)$ — результаты.

Что может быть результатом?

- Факт выздоровления,
- Факт отсутствия каких-либо симптомов,
- Факт нормализации какого-либо параметра,
- и т.д.

Гипотезы:

$H_0 : p_1 = p_2$ — отсутствие эффекта

$H_1 : p_1 < p_2$ — эффект присутствует

Пример: *AB-тест*

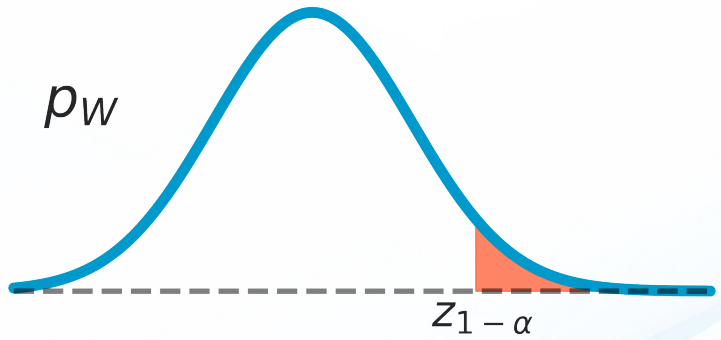
Из ЦПТ можем получить:

$$\widehat{p}_1 = \overline{X} \overset{d}{\approx} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right), \qquad \widehat{p}_2 = \overline{Y} \overset{d}{\approx} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

При справедливости H_0 получаем

$$W(X,Y) = \frac{\widehat{p}_2 - \widehat{p}_1}{\widehat{\sigma}} \overset{d}{\approx} \mathcal{N}(0,1),$$

где $\widehat{\sigma} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$.



Критерий Вальда $S = \{W(x,y) > z_{1-\alpha}\}$.

$\alpha = 0.05 \quad \Longrightarrow \quad z_{1-\alpha} \approx 1.64, \quad S = \{W(x,y) > \mathbf{1.64}\}$.

Дов. интервал для $p_2 - p_1$ равен $C = (\widehat{p}_2 - \widehat{p}_1 - z_{1-\alpha}\widehat{\sigma},\ 1)$.

H_0 отвергается $\Longleftrightarrow 0 \notin C$.

Пример (влияние нового препарата на выздоровление)

Испытуемые делятся случайно на две группы:

1. *Исследуемая группа* — принимает новый препарат;

   $X = (X_1, ..., X_n) \sim Bern(p_1)$ — результаты лечения.

2. *Контрольная группа* — принимает плацебо;

   $Y = (Y_1, ..., Y_m) \sim Bern(p_2)$ — результаты лечения.

$H_0 : p_1 = p_2$ — отсутствие эффекта

$H_1 : p_1 > p_2$ — эффект присутствует

Пример (влияние нового препарата на выздоровление)

1. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 21 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 2$, $pvalue = 0.0228$, дов. интервал $(0.036, 1)$

2. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 15 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.5$

   $W(x, y) \approx 3.76$, $pvalue = 0.00008$, дов. интервал $(0.225, 1)$

3. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 10$ человек, 7 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 1.54$, $pvalue = 0.0618$, дов. интервал $(-0.017, 1)$

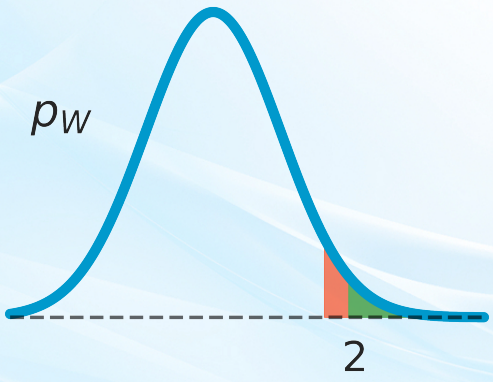
Пример: *AB-тест*

Критерий $S = \{W(x, y) > z_{1-\alpha}\}$, где $W(X, Y) \xrightarrow{d} \mathcal{N}(0, 1)$.

p-value: $p(w) = \mathrm{P}(W(X, Y) \geqslant w) = \text{scipy.stats.norm.sf}(w)$.

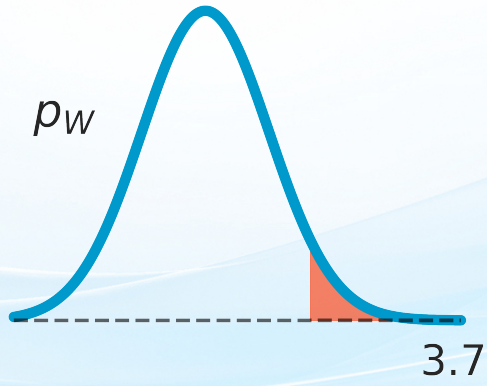
$w = W(x) = 2$

$p(w) = 0.0228$



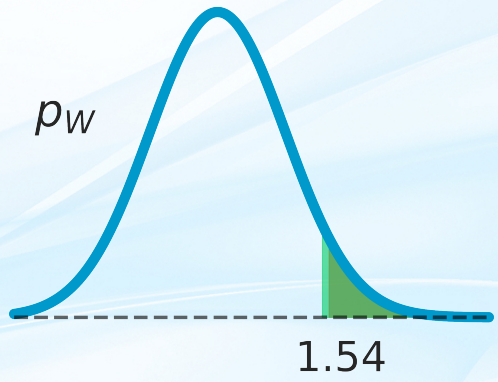
$w = W(x) = 3.76$

$p(w) = 0.00008$



$w = W(x) = 1.54$

$p(w) = 0.0618$



Класс критериев **t-test**

Связные выборки: частный случай

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, ..., Y_n \sim \mathcal{N}(a_2, \sigma_2^2)$.

$H_0 : a_1 = a_2$ *vs.* $H_1 : a_1 \{<, \neq, >\} a_2$

Сведение к задаче с одной выборкой:

Рассмотрим выборку $\delta_1, ..., \delta_n$, где $\delta_i = X_i - Y_i$.

Тогда $H_0 : E\delta_i = 0$ *vs.* $H_1 : E\delta_i \{<, \neq, >\} 0$

Применяем критерий Вальда:

$$T(X, Y) = \sqrt{n}\,\overline{\delta}/S_\delta \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Почему не точный?

Если $X_i \sim \mathcal{N}(a_1, \sigma_1^2)$ и $Y_i \sim \mathcal{N}(a_2, \sigma_2^2)$ зависимы,

то разность не обязана быть нормальной.

Связные выборки: общий случай

$X_1, ..., X_n$ и $Y_1, ..., Y_n$ — произвольные выборки.

$H_0 : EX_1 = EY_1$ *vs.* $H_1 : EX_1 \ \{<, \neq, >\} \ EY_1$

Сведение к задаче с одной выборкой:

Рассмотрим выборку $\delta_1, ..., \delta_n$, где $\delta_i = X_i - Y_i$.

Требование: $\delta_1, ..., \delta_n$ — выборка с конечной дисперсией.

Тогда $H_0 : E\delta_i = 0$ *vs.* $H_1 : E\delta_i \ \{<, \neq, >\} \ 0$
Применяем критерий Вальда:

$$T(X, Y) = \sqrt{n}\ \overline{\delta}/S_\delta \xrightarrow{d_0} \mathcal{N}(0, 1),$$

$$S = \{|T(X, Y)| > z_{1-\alpha/2}\},$$

$$\left(\overline{X} - \overline{Y} \pm z_{1-\alpha/2} S_\delta/\sqrt{n}\right).$$

Независимые выборки: общий случай

$X_1, ..., X_n$ и $Y_1, ..., Y_m$ — произвольные выборки.

$\mathrm{H}_0 \colon \mathrm{E}X_1 = \mathrm{E}Y_1$ *vs.* $\mathrm{H}_1 \colon \mathrm{E}X_1 \ \{<, \neq, >\} \ \mathrm{E}Y_1$

Тогда справедлива сходимость

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2 / n + S_Y^2 / m}} \xrightarrow{d_0} \mathcal{N}(0, 1).$$

$$S = \{|T(X, Y)| > z_{1-\alpha/2}\},$$

Доверительный интервал для $\mathrm{E}X_1 - \mathrm{E}Y_1$ ур. дов. $1 - \alpha$

$$\left(\overline{X} - \overline{Y} \pm z_{1-\alpha/2}\sqrt{S_X^2 / n + S_Y^2 / m}\right).$$

Независимые выборки: частные случаи

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

$H_0: a_1 = a_2$

$H_1: a_1 \{<, \neq, >\} a_2$

Рассуждения:

$\overline{X} \sim \mathcal{N}\left(a_1, \sigma_1^2/n\right)$

$\overline{Y} \sim \mathcal{N}\left(a_2, \sigma_2^2/m\right)$

$\overline{X} - \overline{Y} \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma_1^2/n + \sigma_2^2/m\right)$

Случай 1. σ_1 и σ_2 известны

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

Случай 2. $\sigma_1 = \sigma_2 = \sigma$ **не**известны

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{S_{tot}\sqrt{1/n + 1/m}} \overset{H_0}{\sim} T_{n+m-2},$$

где $S_{tot}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$ — несмещенная оценка σ, как взвешенное усреднение дисперсий: S_X^2, S_Y^2 — *несмещ.* оценки дисп.

Критерий

$$S = \{|T(X, Y)| > T_{n+m-2,1-\alpha/2}\}$$

Дов. интервал для $a_1 - a_2$ ур. дов. $1 - \alpha$

$$\left(\overline{X} - \overline{Y} \pm T_{n+m-2,1-\alpha/2} S_{tot}\sqrt{1/n + 1/m}\right)$$

Независимые выборки: частные случаи

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$

$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

$H_0: a_1 = a_2$

$H_1: a_1 \{<, \neq, >\} a_2$

Рассуждения:

$\overline{X} \sim \mathcal{N}\left(a_1, \sigma_1^2/n\right)$

$\overline{Y} \sim \mathcal{N}\left(a_2, \sigma_2^2/m\right)$

$\overline{X} - \overline{Y} \overset{H_0}{\sim} \mathcal{N}\left(0, \sigma_1^2/n + \sigma_2^2/m\right)$

Случай 1. σ_1 и σ_2 известны

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

Случай 3. $\sigma_1 \neq \sigma_2$ и **не**известны

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2/n + S_Y^2/m}} \overset{H_0 \text{ прибл.}}{\sim} T_\nu$$

$$\nu = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \Big/ \left(\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}\right),$$

где S_X^2, S_Y^2 — *несмещ.* оценки дисп.

Критерий $S = \{|T(X, Y)| > T_{\nu, 1-\alpha/2}\}$

Дов. интервал для $a_1 - a_2$ ур. дов. $1 - \alpha$

$$\left(\overline{X} - \overline{Y} \pm T_{\nu, 1-\alpha/2} \sqrt{S_X^2/n + S_Y^2/m}\right).$$

Посмотрим на то, что мы получили

1. Норм. независ. выборки

$$T(X,Y) = \frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \overset{H_0}{\sim} \mathcal{N}(0,1)$$

$$T(X,Y) = \frac{\overline{X}-\overline{Y}}{S_{tot}\sqrt{1/n+1/m}} \overset{H_0}{\sim} T_{n+m-2}$$

$$T(X,Y) = \frac{\overline{X}-\overline{Y}}{\sqrt{S_X^2/n+S_Y^2/m}} \overset{H_0}{\underset{\text{прибл.}}{\sim}} T_\nu$$

2. Норм. связные выборки

$$T(X,Y) = \sqrt{n}\,\overline{\delta}/S_\delta \xrightarrow{d_0} \mathcal{N}(0,1),$$

где $\delta_i = X_i - Y_i$.

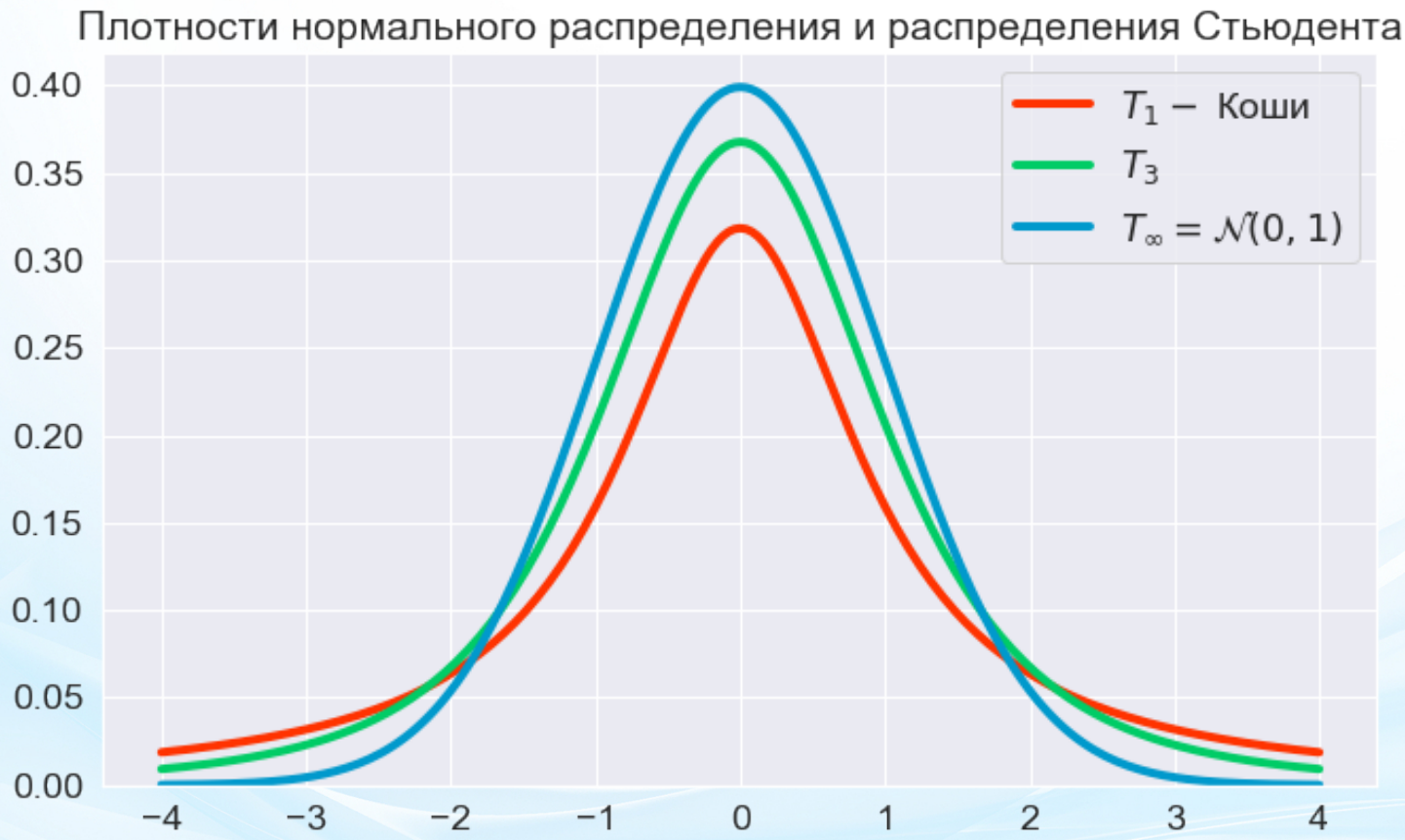
3. Берн. независ. выборки

$$T(X,Y) = \frac{\widehat{p}_1-\widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n}+\frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}} \xrightarrow{d_0} \mathcal{N}(0,1)$$

Общий вид:

$$\frac{\overline{X}-\overline{Y}}{\widehat{\sigma}} \xrightarrow{d_0} \mathcal{N}(0,1)$$

Сравнение распределений



Абсолютный t-test

Общий вид:

$$\frac{\overline{X} - \overline{Y}}{\widehat{\sigma}} \xrightarrow{d_0} \mathcal{N}(0,1),$$

например, $\widehat{\sigma} = \sqrt{S_X^2/n + S_Y^2/m}$.

1. Подобное выражение верно для многих других распределений.
   Главное требование: конечная дисперсия распределений.

2. T-распределение имеет более тяжелые хвосты
   $\Rightarrow$ его квантили больше по модулю.
   Для более надежного контроля за уровнем значимости
   используют T-квантили вместо Z-квантилей.
   Отсюда название: t-test.

3. Идеален с точки зрения интерпретации,
   позволяет сравнивать именно средние.

4. Неустойчив к выбросам.
   Обычно это недостаток, но иногда можно интерпретировать как преимущество.

Доверительный интервал

Общий вид:

$$\frac{\overline{X} - \overline{Y}}{\widehat{\sigma}} \xrightarrow{d_0} \mathcal{N}(0,1),$$

На практике рекомендуется строить доверительный интервал

$$\left(\overline{X} - \overline{Y} \pm z_{1-\alpha/2}\widehat{\sigma}\right)$$

Пример

- +10 мкг/л, p-value=0.01, **результат стат. значим**
- Более информативно: $+(10 \pm 5)$ мкг/л

А много это или мало?

- Если начальное значение равно 100 мкг/л, тогда $+(10 \pm 5)\%$
- Если начальное значение равно 1000 мкг/л, тогда $+(1 \pm 0.5)\%$

Относительный t-test для независимых выборок

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$ — тестовая группа

$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$ — контрольная группа

$\mathrm{H}_0 : a_1 = a_2$ *vs.* $\mathrm{H}_1 : a_1 \{<, \neq, >\} a_2$

Рассмотрим статистику

$$R = \frac{\overline{X} - \overline{Y}}{\overline{Y}}$$

Асимптотически можно получить приближения

$$a_R = \mathrm{E}R \approx \frac{a_1 - a_2}{a_2}, \qquad \sigma_R^2 = \mathrm{D}R \approx \frac{\sigma_1^2}{a_2^2} + \frac{a_1^2}{a_2^4}\sigma_2^2$$

Используя соответствующие оценки, получаем

$$\sqrt{n}\frac{R}{\widehat{\sigma}_R} \xrightarrow{d_0} \mathcal{N}(0, 1)$$

На практике рекомендуется строить доверительный интервал

$$\left(R \pm z_{1-\alpha/2}\widehat{\sigma}_R\right)$$

Вспомним, как это было в Python

Непараметрические критерии

Непараметрический == свободный от семейства распределений

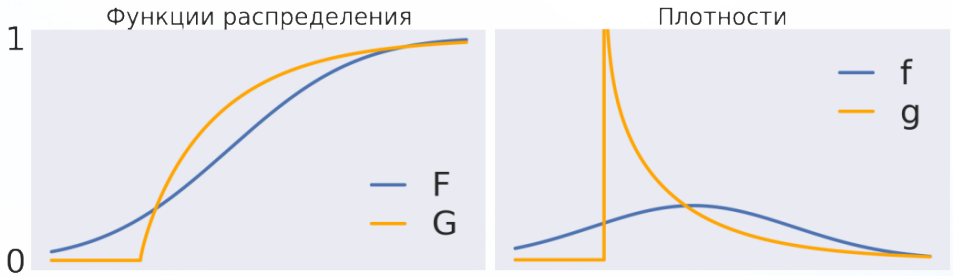
Альтернативы

$X_1, ..., X_n$ и $Y_1, ..., Y_m$ — две выборки из неизвестных **непрерывных** распределений с функциями распределений F и G.

$H_0 : F = G$ — гипотеза однородности

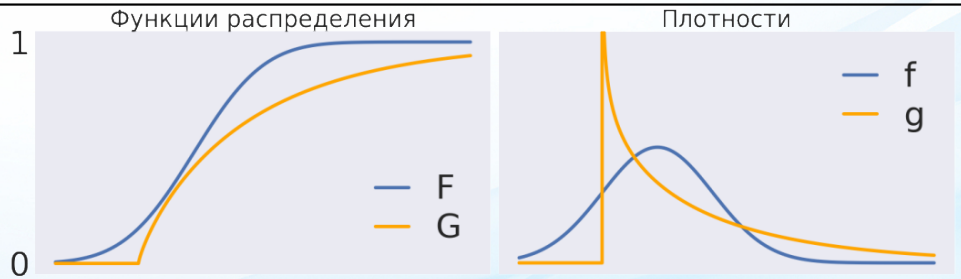
Гипотеза неоднородности:

$H_1 : F \neq G$



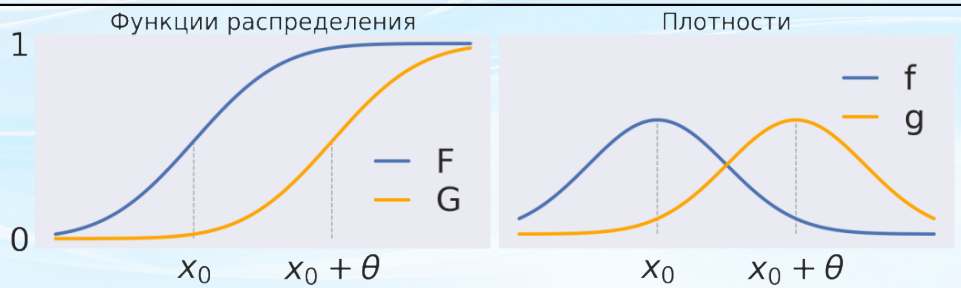
Гипотеза доминирования:

$H_2 : F \geqslant G$



Гипотеза сдвига:

$H_3 : F(x - \theta) = G(x)$



Непараметрический случай

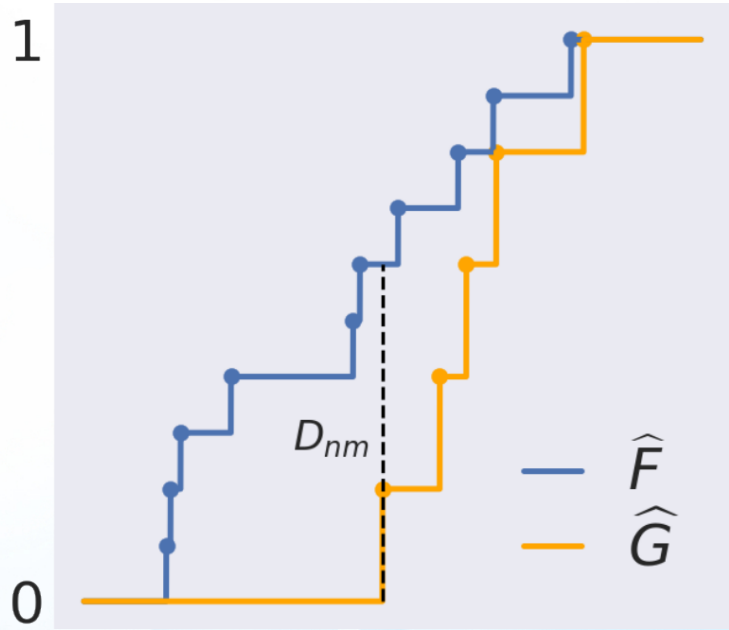
Независимые выборки

I. Критерии на основе ЭФР: Критерий Смирнова

$X_1, ..., X_n$ и $Y_1, ..., Y_m$ — две выборки из неизвестных непрерывных распределений с функциями распределений F и G.

$\mathrm{H}_0 : F = G \quad vs. \quad \mathrm{H}_1 : F \neq G$

Статистика $D_{nm} = \sup\limits_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - \widehat{G}_m(x) \right|$



Вычисление:

$$D_{nm} = \max\{D^+_{nm}, D^-_{nm}\}$$

$$D^+_{nm} = \max_{i=1..n} \left\{ i/n - \widehat{G}_m\left(X_{(i)}\right) \right\}$$

$$D^-_{nm} = \max_{j=1..m} \left\{ j/m - \widehat{F}_n\left(Y_{(j)}\right) \right\}$$

$$\sqrt{\frac{nm}{n+m}} D_{nm} \xrightarrow{d_0} Kolmogorov, \text{ при } n, m \to +\infty \text{ если } n/(n+m) \to \gamma \in (0,1)$$

Приближение точное при $n, m \geqslant 20$

II. Критерий Уилкоксона-Манна-Уитни

Рассматриваем альтернативу $H_2 : F \geqslant G$.

S_j — ранг Y_j в вариационном ряду по выборке $(X_1, .., X_n, Y_1, .., Y_m)$.

$V = S_1 + ... + S_m$ — статистика критерия.

$$\frac{V - EV}{\sqrt{DV}} \xrightarrow{d_0} \mathcal{N}(0, 1),$$

где $EV = \frac{m(n+m+1)}{2}$, $DV = \frac{nm(n+m+1)}{12}$ при H_0.

Идея: если H_0 верна, то значения $Y_{(j)}$ равномерно разбросаны по вар. ряду. Большие значения V указывают на преобладание Y_j над X_i.
Критерий имеет вид $S = \{V > c\}$.

- приближение при $n, m \geqslant 50$;
- если $n, m \geqslant 25$, используется поправка Иман-Давенпорта;
- при малых n и m используются таблицы.

II. Критерий Уилкоксона-Манна-Уитни

Совпадения

- Рассматриваются средние ранги
- Дисперсия

$$\mathrm{D}V = \frac{nm}{12}\left(n+m+1-\frac{1}{(n+m)(n+m-1)}\sum_{k=1}^{g} l_k(l_k-1)\right),$$

g — число групп совпадений

l_k — количество элементов в k-ой группе.

II. Критерий Уилкоксона-Манна-Уитни

Оценка параметра сдвига

В случае альтернативы $H_3\colon F(x-\theta) = G(x)$ оценка

$$\widehat{\theta} = med\{W_{ij} = Y_j - X_i, i = 1..n, j = 1..m\}$$

Свойство: $\sqrt{\frac{nm}{n+m}}\left(\widehat{\theta}-\theta\right) \xrightarrow{d_0} \mathcal{N}\left(0,\sigma^2\right), \quad \left[n,m\to+\infty, \frac{n}{n+m}\to\gamma\in(0,1)\right]$

где $\sigma^{-1} = \sqrt{12}\int\limits_{\mathbb{R}} p^2(x)dx,$

$p(x)$ — плотность ф.р. F.

Доверительный интервал параметра сдвига

$(W_{(k_\alpha+1)}, W_{(nm-k_\alpha)})$,

где $k_\alpha = \left\lfloor nm/2 - 1/2 - z_{1-\alpha}\sqrt{nm(n+m+1)/12}\right\rfloor$

Связь оценки и критерия

Статистика Манна-Уитни:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i \leqslant Y_j\}$$

При отсутствии совпадений $U = V - \frac{m(m+1)}{2}$.

Пусть θ — неизвестный сдвиг.

Тогда $(X_1, ..., X_n)$ и $(Y_1 - \theta, ..., Y_m - \theta)$ однородны.

$\Longrightarrow$ для них распределение U симметрично относительно $\frac{nm}{2}$.

Получаем уравнение

$$\sum_{i=1}^{n} \sum_{j=1}^{m} I\{X_i \leqslant Y_j - \theta\} = \sum_{i=1}^{n} \sum_{j=1}^{m} I\{Y_i - X_i \geqslant \theta\} = \frac{nm}{2}$$

Откуда $\widehat{\theta} = med\{W_{ij} = Y_j - X_i, i = 1..n, j = 1..m\}$

Непараметрический случай

Связные выборки

Связные выборки: модель

$X_1, ..., X_n$ и $Y_1, ..., Y_n$ — связные выборки

Перейдем к **выборке разностей**:

$Z_i = Y_i - X_i = \theta + \varepsilon_i$,

- $\theta > 0$ — интересующий систематический эффект воздействия;
- $\varepsilon_1, ..., \varepsilon_n$ — случайные ошибки.

Предположения об ошибках:

- независимы;
- имеют непрерывные распределения (м.б. разные);
- медиана $= 0$.

Гипотезы: $H_0' : \theta = 0$ *vs.* $H_3' : \theta > 0$

Критерий знаков

Рассмотрим знаки $U_i = I\{Z_i > 0\} \sim Bern(p)$

$H_0' : p = 1/2$ *vs.* $H_3' : p > 1/2$

Статистика критерия $S = U_1 + ... + U_n \overset{H_0'}{\sim} Bin(n, 1/2)$

Критерий $\{S > c\}$.

Аппроксимация при $n > 15$: $\frac{S - n/2 - 1/2}{\sqrt{n/4}} \xrightarrow{d_0} \mathcal{N}(0, 1)$,

Совпадения: выбрасываем соответствующие наблюдения.

Оценка параметра: $\widehat{\theta} = med\{Z_i, i = 1..n\}$

Доверительный интервал для параметра

$(Z_{(k_\alpha+1)}, Z_{(n-k_\alpha)})$ — д.и. уровня доверия $1 - 2\alpha$,

где $k_\alpha = \left\lfloor n/2 - 1/2 - z_{1-\alpha}\sqrt{n/4} \right\rfloor$

Связь оценки и критерия

Знаки $U_i = I\{Z_i > 0\} \sim Bern(p)$

Статистика критерия $S = U_1 + ... + U_n$

Пусть θ — неизвестный сдвиг.

Тогда для $(Z_1 - \theta, ..., Z_n - \theta)$ медиана равна нулю.

Получаем уравнение

$$\sum_{i=1}^{n} I\{Z_i - \theta > 0\} = \sum_{i=1}^{n} I\{Z_i > \theta\} = \frac{n}{2}$$

Откуда $\widehat{\theta} = med\{Z_i, i = 1..n\}$.

Пример: времена реакции (Лагутин)

X_i — время реакции i-го испытуемого на световой сигнал

Y_i — время реакции i-го испытуемого на звуковой сигнал

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	176	163	152	155	156	178	160	164	169	155	122	144
y_i	168	215	172	200	191	197	183	174	176	155	115	163
z_i	-8	+52	+20	+45	+35	+19	+23	+10	+7	0	-7	+19

$S(x) = 9$ — значение статистики критерия

$pvalue = (1 + 11 + 55)/2048 \approx 0.033$

$\widehat{\theta}(x) = 19$

$(7, 35)$ — 90% доверительный интервал

Критерий ранговых сумм Уилкоксона

Оценка параметра сдвига — медиана средних Уолша

$$\widehat{\theta} = med\{V_{ij} = (Z_i + Z_j)/2, 1 \leqslant i \leqslant j \leqslant n\}$$

Свойство: $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d_0} \mathcal{N}\left(0, \sigma^2\right)$,

где $\sigma^{-1} = \sqrt{12}\int\limits_{\mathbb{R}} p^2(x)dx$,

$p(x)$ — плотность ф.р. F.

Доверительный интервал параметра сдвига

$(V_{(k_\alpha+1)}, V_{(n(n+1)/2-k_\alpha)})$ — д.и. уровня доверия $1-2\alpha$,

где $k_\alpha = \left\lfloor n(n+1)/4 - 1/2 - z_{1-\alpha}\sqrt{n(n+1)(2n+1)/24}\right\rfloor$

Связь оценки и критерия

$T = R_1 U_1 + ... + R_n U_n$ — статистика критерия.

Отсутствии нулей и совпадений среди $|Z_i|$ выполнено

$$T = \sum_{i \leqslant j} I\left\{\frac{Z_i + Z_j}{2} > 0\right\}$$

Пусть θ — неизвестный сдвиг.

Тогда для $(Z_1 - \theta, ..., Z_n - \theta)$ распределение статистики T симметрично относительно среднего $\frac{n(n+1)}{4}$

Получаем уравнение

$$\sum_{i \leqslant j} I\left\{\frac{Z_i - \theta + Z_j - \theta}{2} > 0\right\} = \sum_{i \leqslant j} I\left\{\frac{Z_i + Z_j}{2} > \theta\right\} = \frac{n(n+1)}{2}$$

Откуда $\widehat{\theta} = med\{V_{ij} = (Z_i + Z_j)/2, 1 \leqslant i \leqslant j \leqslant n\}$.

Пример: времена реакции (Лагутин)

X_i — время реакции i-го испытуемого на световой сигнал

Y_i — время реакции i-го испытуемого на звуковой сигнал

z_i	-7	7	-8	10	19	19	20	23	35	45	52
$\overline{\lvert z_i \rvert}$	7	7	8	10	19	19	20	23	35	45	52
$\overline{R_i}$	1.5	1.5	3	4	5.5	5.5	7	8	9	10	11
$\overline{U_i}$	0	1	0	1	1	1	1	1	1	1	1

$T(x) = 61.5$ — значение статистики критерия

$T^*(x) = 2.54$ — нормированное значение статистики

$pvalue = 0.006$

$\widehat{\theta}(x) = 19.25$

$(7.5, 31)$ — 90% доверительный интервал

Как еще можно считать доверительные интервалы?

Метод бутстрепа

Генерация бутстрепной выборки $X_1^*, ..., X_n^*$:

упоряд. выбор с возвращением n элементов из мн-ва $\{X_1, ..., X_n\}$.

Другой вид записи:

1. $i_1, ..., i_n \sim U\{1, ..., n\}$.
2. $X^* = (X_1^*, ..., X_n^*) = (X_{i_1}, ..., X_{i_n})$ — бутстрепная выборка.

Этап 1. Процедуру генерации выборок повторить B раз:

$X_b^* = (X_{b1}^*, ..., X_{bn}^*)$, где $1 \leqslant b \leqslant B$.

Далее по каждой выборке посчитаем значение статистики T,

получив выборку значений $T_1^* = T(X_1^*), ..., T_B^* = T(X_B^*)$.

Этап 2. Бутстрепная оценка:

$$\widehat{v}_{boot} = \frac{1}{B}\sum_{b=1}^{B} T_b^{*2} - \left(\frac{1}{B}\sum_{b=1}^{B} T_b^*\right)^2.$$

Особенности

- Число B стоит брать как можно больше.

- Размер бутстрепной выборки **всегда тот же**, что и у исходной.

  При генерации выборок иного размера распределение статистики T, вообще говоря, может быть другим.

  Например, дисперсия выборочного среднего зависит от размера выборки.

- Генерация бутстр. выборки проводится независимо с повторами.

  Иначе полученный набор даже не является выборкой.

Бутстрепные доверительные интервалы

1. **Нормальный интервал**

   Пусть $\widehat{\theta}$ — а.н.о. θ с ас. дисп. $\sigma^2(\theta)$.

   $\widehat{v}_{boot}$ — бутстрепная оценка дисперсии.

   Бутстрепный дов. интервал для параметра θ имеет вид

$$\left(\widehat{\theta} - z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}, \quad \widehat{\theta} + z_{(1+\alpha)/2}\sqrt{\widehat{v}_{boot}}\right)$$

2. **Центральный интервал**

   $\theta = G(\mathrm{P})$ и $\widehat{\theta} = G(\widehat{\mathrm{P}}_n)$ — оценка методом подстановки.

   $\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

   Бутстрепный доверительный интервал имеет вид

$$C^* = \left(2\widehat{\theta} - \theta^*_{(\lceil B(1+\alpha)/2 \rceil)}, \quad 2\widehat{\theta} - \theta^*_{(\lfloor B(1-\alpha)/2 \rfloor)}\right).$$

Бутстрепные доверительные интервалы

3. **Квантильный интервал**

$\widehat{\theta}$ — некоторая оценка θ.

$\theta_1^*, ..., \theta_B^*$ — оценки по бутстрепным выборкам.

Бутстрепный доверительный интервал имеет вид

$$C^* = \left(\theta^*_{(\lfloor B(1-\alpha)/2 \rfloor)}, \quad \theta^*_{(\lceil B(1+\alpha)/2 \rceil)}\right).$$

Утв. Если существует монотонное преобразование φ, для которого $\varphi(\widehat{\theta}) \sim \mathcal{N}(\varphi(\theta), \sigma^2)$, то $\mathrm{P}\left(\theta \in C^*\right) = \alpha$.

На практике такое преобразование существует редко, но при этом часто может существовать приближенное преобразование.

Пример: построение дов. интервалов для θ

$x = (5, 1, 3, 6, 4)$ — реализация выборки

$\theta = \mathsf{E}X_1$ — параметр, $\quad \widehat{\theta} = \overline{X}$ — оценка, $\quad \widehat{\theta} = 3.8$ — реализация оценки

Реализации оценки параметра по бутстрепным выборкам ($B = 100$):

4.2, 4.2, 2.6, 3.2, 4.2, 3.8, 3.2, 3.6, 3.6, 3.4, 3.8, 4.4, 3.6, 3.2, 4.6, 4.2, 3.0, 3.2, 4.0, 3.0,
3.2, 3.0, 2.6, 3.0, 3.6, 3.4, 5.0, 4.8, 3.4, 2.6, 2.6, 3.6, 3.2, 4.2, 3.2, 3.4, 4.4, 4.2, 4.4, 3.4,
4.0, 2.4, 3.4, 3.8, 2.0, 3.0, 4.6, 3.2, 3.6, 3.6, 4.0, 3.8, 4.0, 3.4, 3.8, 3.8, 4.2, 3.2, 2.8, 4.0,
3.2, 3.4, 3.0, 4.0, 3.6, 3.4, 3.8, 3.2, 3.8, 2.6, 3.4, 5.0, 3.6, 3.0, 4.8, 4.2, 3.4, 5.2, 5.0, 3.4,
3.2, 3.6, 4.2, 3.4, 3.2, 3.8, 3.6, 3.8, 3.0, 2.8, 3.0, 4.0, 3.2, 3.6, 2.6, 3.2, 2.4, 3.6, 4.0, 4.2

1. **Нормальный интервал**

   $\widehat{\theta} = 3.8, v_{boot} = 0.394, z_{0.975} = 1.96$

   $(3.8 \pm 1.96 \cdot \sqrt{0.394}) = (2.57,\ 5.03)$

2. **Центральный интервал**

   $B(1+\alpha)/2 = 100 \cdot 0.975 = 97.5, B(1-\alpha)/2 = 100 \cdot 0.025 = 2.5$

   $\theta^*_{(\lceil 97.5 \rceil)} = 5, \quad \theta^*_{(\lfloor 2.5 \rfloor)} = 2.4$

   $(2 \cdot 3.8 - 5,\ 2 \cdot 3.8 - 2.4) = (2.6,\ 5.2)$

3. **Квантильный интервал**

   $(2.4,\ 5)$

Вспомним, как это было в Python

Валидация критериев

AA-тесты

Пусть S — некоторый критерий уровня значимости α.

Оценка реального уровня значимости (вер-ти ошибки 1 рода)

1. Создаем датасеты с отсутствием эффекта между группами.
2. Для каждого датасета применяем критерий.
3. Вычисляем долю случаев, в которых критерий отклонил основную гипотезу, и строим доверительный интервал $(\widehat{\alpha}_\ell, \widehat{\alpha}_r)$.

Результаты:

- Если $\widehat{\alpha}_\ell \leqslant \alpha \leqslant \widehat{\alpha}_r$, то все хорошо.
- Если $\alpha < \widehat{\alpha}_\ell$, то такой критерий использовать нельзя.
- Если $\alpha > \widehat{\alpha}_r$, то неплохо,
  но скорее всего он недостаточно мощный.

Искусственные AB-тесты

Оценка мощности

1. Создаем датасеты с отсутствием эффекта между группами.
2. Добавить эффект к одной из групп. Он может быть
   - одинаковым для всех точек,
   - случайным с фиксированным мат. ожиданием.
3. Для каждого датасета применяем критерий.
4. Вычисляем долю случаев, в которых критерий отклонил основную гипотезу, и строим доверительный интервал $\left(\widehat{\beta}_\ell, \widehat{\beta}_r\right)$.

Особенности:

- Обычно оценивают мощность для нескольких значений эффекта и определяют минимально детектируемый эффект.
- Из критериев, допустимых по величине вер-ти ошибки 1 рода, выбирают критерий с наибольшей мощностью.

Откуда взять датасеты?

1. **Искусственные данные.**

   Можно быстро сгенерировать сколько угодно датасетов.

   Но это не гарантирует корректность на реальных данных.

2. **Исторические данные.**

   Делим данные по разным факторам:

   - возрастная категория
   - регион
   - дни/месяцы

   Сложно собрать много датасетов, используя сдвиги во времени.

   Способ гарантирует адекватную проверку критерия.

Рекомендация: на начальных этапах исследования лучше проверять критерий на искусственных данных.
Перед непосредственным применением критерия необходимо выполнить проверку на реальных исторических данных.


ВСЁ!