Дисперсионный анализ

Ph@DS, весна 2023

Дисперсионный анализ (критерии АВ-тестирования)

Типы задач

1. **Независимые выборки**

   Провели эксперимент несколько раз разными методами.

   Действительно ли получились одинаковые результаты?

2. **Связные выборки**

   Человеку дали препарат для снижения температуры.

   Отличается ли температура до и после?

- 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

№	Метод	Результат
1	Колебания	10.1
2	Колебания	9.7
3	Колебания	9.9
4	Колебания	9.5
1	Полет	10
2	Полет	10.5
3	Полет	9.8

Значимо ли отличаются результаты разных методов?

Связные выборки

Рассмотрим испытуемых, которые приняли лекартста.

Человек	**T до**	**T после**
Петя	37.8	37.2
Вася	38.3	36.6
Катя	36.9	36.2
Миша	37.1	36.8
Ира	36.7	36.8
Света	37.5	37.1

Есть ли эффект от препарата?

Другие вопросы на практике

1. Изменился ли сигнал от звезды?
2. Есть ли эффект от введения вакцины?
3. Отличаются ли гены по степени экспрессии?
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 $\mathsf{P}(X \in S) \leqslant \alpha, \quad \forall\, \mathsf{P} \in \mathscr{P}_0$.

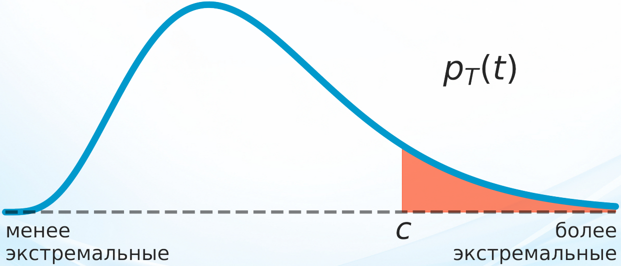
Варианты ответа:

1. $X \in S \implies$ H_0 отвергается $\implies$ результат стат. значим;
2. $X \notin S \implies$ 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 $\iff$ $T(X) \geqslant c_\alpha$.

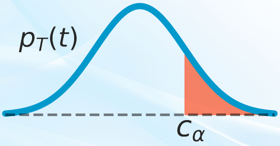
Гипотезы и критерии (*напоминание*)

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

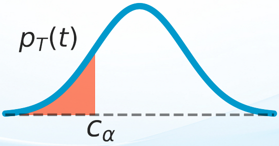
α выбирается **ДО** эксперимента,

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

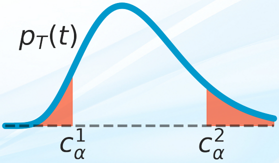
$S = \{T(x) > c_\alpha\}$



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



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



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

Пример: *AB-тест*

Пользователи 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 проявления каких-нибудь симптомов.
- и т.д.

Гипотезы:

$\mathrm{H}_0: p_1 = p_2$ — отсутствие эффекта
$\mathrm{H}_1: p_1 < p_2$ — эффект присутствует

Пример: *AB-тест*

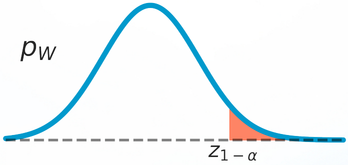
Из ЦПТ можем получить:

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

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



Сходимость $W(X,Y) \xrightarrow{d} \mathcal{N}(0,1)$ при $n, m \to +\infty$ можно доказать строго.

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

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

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: *AB-тест*

1. 1 группа: $n = 30$ человек, 21 выздоровели $\Longrightarrow$ $\widehat{p}_1 = 0.7$

   2 группа: $m = 30$ человек, 27 выздоровели $\Longrightarrow$ $\widehat{p}_2 = 0.9$

   $W(x, y) \approx 2$ $\Longrightarrow$ H_0 отвергается, результат стат. значим

   дов. интервал $(0.036, 1)$ $\leftarrow$ **слабая уверенность в результате**

2. 1 группа: $n = 30$ человек, 15 выздоровели $\Longrightarrow$ $\widehat{p}_1 = 0.5$

   2 группа: $m = 30$ человек, 27 выздоровели $\Longrightarrow$ $\widehat{p}_2 = 0.9$

   $W(x, y) \approx 3.76$ $\Longrightarrow$ H_0 отвергается, результат стат. значим

   дов. интервал $(0.225, 1)$ $\leftarrow$ **хорошая уверенность в результате**

3. 1 группа: $n = 10$ человек, 7 выздоровели $\Longrightarrow$ $\widehat{p}_1 = 0.7$

   2 группа: $m = 30$ человек, 27 выздоровели $\Longrightarrow$ $\widehat{p}_2 = 0.9$

   $W(x, y) \approx 1.54$ $\Longrightarrow$ H_0 не отвергается, результат стат. незнач.

   дов. интервал $(-0.017, 1)$ $\leftarrow$ **нет результата**

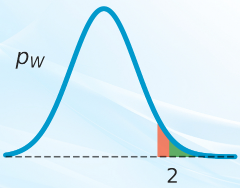
Пример: *AB-тест*

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

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

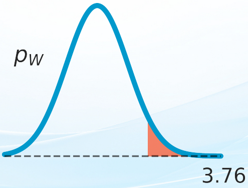
$w = W(x) = 2$

$p(w) = 0.0228$



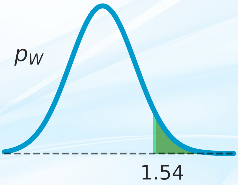
$w = W(x) = 3.76$

$p(w) = 0.00008$



$w = W(x) = 1.54$

$p(w) = 0.0618$



Класс критериев **t-test**

Связные выборки: частный случай

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

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

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

Сведение к задаче с одной выборкой:

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

Тогда $\mathrm{H}_0\colon \mathrm{E}\delta_i = 0$ $vs.$ $\mathrm{H}_1\colon \mathrm{E}\delta_i \ \{<, \neq, >\}\ 0$

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.

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

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 $\mathrm{H}_0\colon \mathsf{E}\delta_i = 0$ *vs.* $\mathrm{H}_1\colon \mathsf{E}\delta_i \ \{<, \neq, >\}\ 0$
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.

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

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 $\mathsf{E}X_1 - \mathsf{E}Y_1$ 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

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

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

$\mathsf{H}_0: a_1 = a_2$

$\mathsf{H}_1: a_1 \{<, \neq, >\} a_2$

Рассуждения:

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

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

$\overline{X} - \overline{Y} \overset{\mathsf{H}_0}{\sim} \mathcal{N}\left(0, \sigma_1^2/n + \sigma_2^2/m\right)$

Случай 1. σ_1 и σ_2 известны

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{\mathsf{H}_0}{\sim} \mathcal{N}(0, 1)$$

Случай 2. $\sigma_1 = \sigma_2 = \sigma$ **неизвестны**

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{S_{tot}\sqrt{1/n + 1/m}} \overset{\mathsf{H}_0}{\sim} T_{n+m-2},$$

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

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$
$Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

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

Рассуждения:

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

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

$\overline{X} - \overline{Y} \overset{\mathrm{H}_0}{\sim} \mathcal{N}\left(0, \sigma_1^2/n + \sigma_2^2/m\right)$

Случай 1. σ_1 и σ_2 известны

Статистика критерия

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \overset{\mathrm{H}_0}{\sim} \mathcal{N}(0, 1)$$

Случай 3. $\sigma_1 \neq \sigma_2$ и **неизвестны**

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2/n + S_Y^2/m}} \overset{\text{H}_0 \text{ прибл.}}{\sim} T_v$$

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

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

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

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

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

Посмотрим на то, что мы получили

1. Норм. независ. выборки

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

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

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

2. Норм. связные выборки

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

где $\delta_i = X_i - Y_i$.

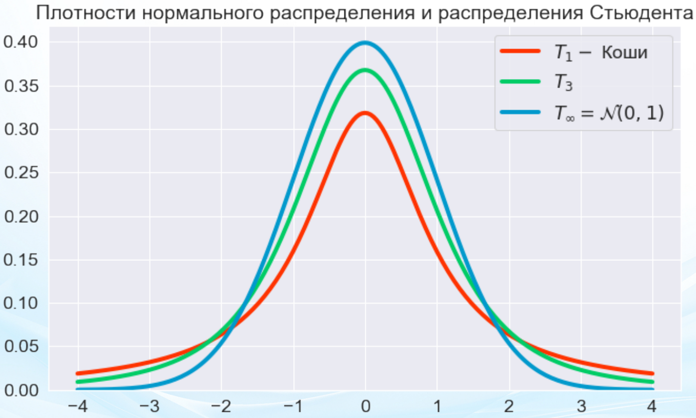
3. Берн. независ. выборки

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

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

- Лечение быстрее на 10 дней., p-value=0.01, рез-т стат. значим
- Более информативно: 10 ± 5 дней.

А много это или мало?

- Если до этого лечили 100 дней, то $+10 \pm 5\%$
- Если до этого лечили 20 дней, то $+50 \pm 25\%$

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

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

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

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

Бутстрепные тесты

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

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

Рассматриваем статистику $T(X, Y)$.

Возможные проблемы

- Распределение статистики недостаточно похоже на нормальное распред., например, мало данных или слишком тяжелые хвосты.
- В выборке есть зависимости, в следствие чего дисперсия среднего оценивается неправильно.

Можно применить бутстреп.

1. Получить бутстрепную выборку статистик $T(X, Y)$.
2. Построить бутстрепный доверительный интервал и сравнить с 0.

Валидация критериев

АА-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 без учета специфики.

   Но это не гарантирует корректность на реальных данных.

2. **Исторические данные.**

   Если есть данные

   - из других работ
   - прошлых экспериментов
   - полученные с помощью моделирования

   Является более адекватной проверкой критерия.

Рекомендация: для грубой проверки можно проверять критерий на искусственных данных. Перед непосредственным применением критерия лучше выполнить проверку на реальных данных.


ВСЁ!