Дисперсионный анализ

Ph@DS, весна 2025

Дисперсионный анализ (критерии АВ-тестирования)

Типы рассматриваемых задач

1. **Независимые выборки**

   Провели эксперимент несколько раз разными методами.

   Действительно ли получились одинаковые результаты?

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

№	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

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. Изменился ли сигнал от звезды?
2. Отличаются ли гены по степени экспрессии?
3. Есть ли эффект от введения вакцины?
4. Какие факторы влияют на появление дефектов при производстве сенсоров?
5. Увеличивается ли эффективность преобразования энергии в солнечных батареях при использовании модификаций материалов катодной поверхности?
6. И 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 $\mathsf{P}(X \in S) \leqslant \alpha, \quad \forall\, \mathsf{P} \in \mathscr{P}_0$.

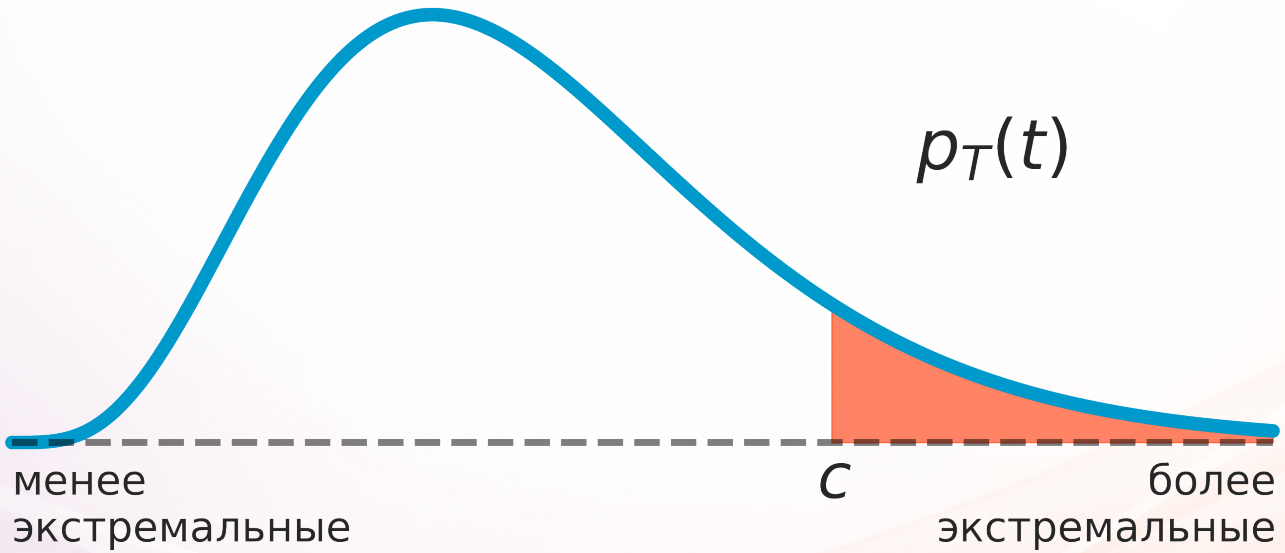
Варианты ответа:

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 $t_1 > t_2$.

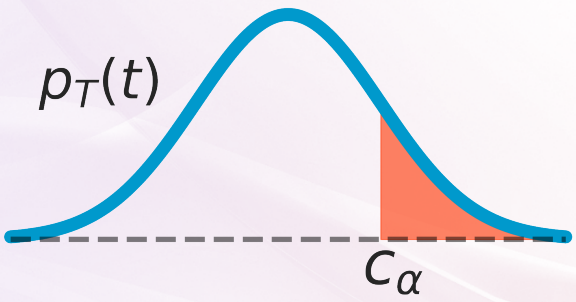
Гипотезы и критерии (*напоминание*)

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

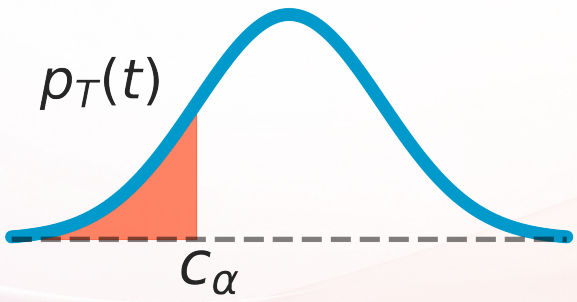
α выбирается **ДО** эксперимента,

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

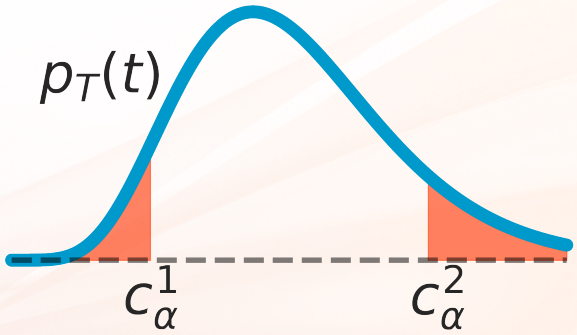
$S = \{T(x) > c_\alpha\}$



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



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



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*

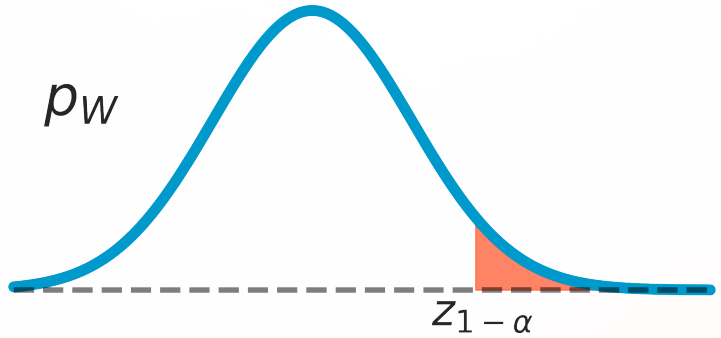
Из ЦПТ можем получить:

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

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



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

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

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

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: *AB-тест*

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

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

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

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

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

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

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

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

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

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

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

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

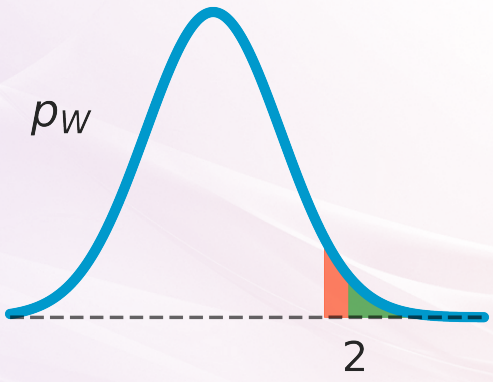
Пример: *AB-тест*

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

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

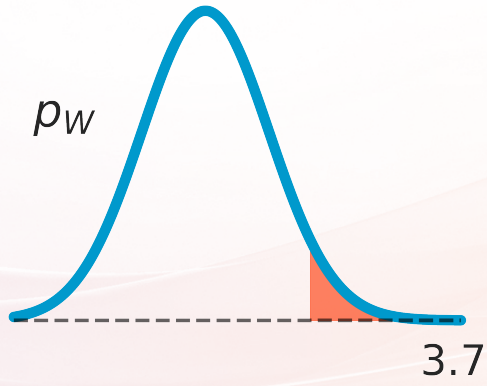
$w = W(x) = 2$

$p(w) = 0.0228$



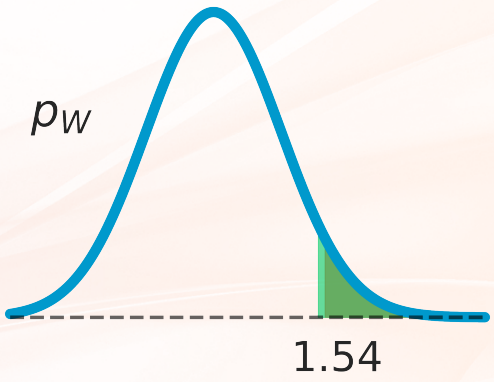
$w = W(x) = 3.76$

$p(w) = 0.00008$



$w = W(x) = 1.54$

$p(w) = 0.0618$



Класс критериев **t-test**

Связные выборки: частный случай

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

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

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

Сведение к задаче с одной выборкой:

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

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

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.

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

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 $\mathrm{H}_0 : \mathrm{E}\delta_i = 0$ *vs.* $\mathrm{H}_1 : \mathrm{E}\delta_i \ \{<, \neq, >\} \ 0$
Применяем критерий Вальда:

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

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

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

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. $1 - \alpha$

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

Посмотрим на то, что мы получили

1. Норм. независ. выборки

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

2. Норм. связные выборки

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

где $\delta_i = X_i - Y_i$.

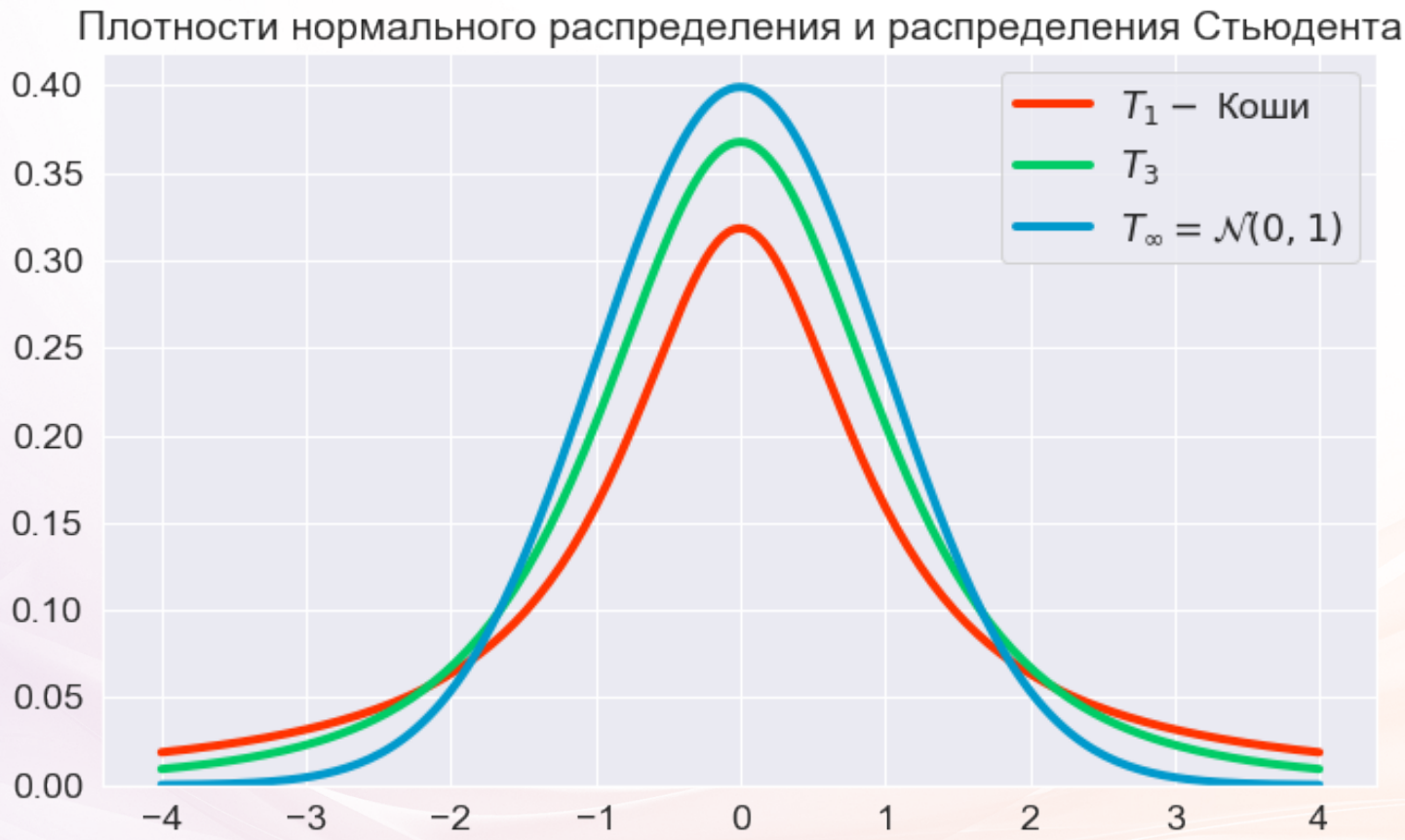
3. Берн. независ. выборки

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

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, **результат стат. значим**
- Более информативно: (10 ± 5) дней

А много это или мало?

- Если начальное значение равно 100 дней, тогда $(10 \pm 5)\%$
- Если начальное значение равно 20 дней, тогда $(50 \pm 25)\%$

Относительный t-test для независимых выборок

$X_1, ..., X_n$ и $Y_1, ..., Y_m$ — контрольная группа

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

Рассмотрим статистику

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

Асимптотически можно получить приближения

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

Используя соответствующие оценки, получаем

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

На практике рекомендуется строить доверительный интервал

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

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.**

   Если есть данные:

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

   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.

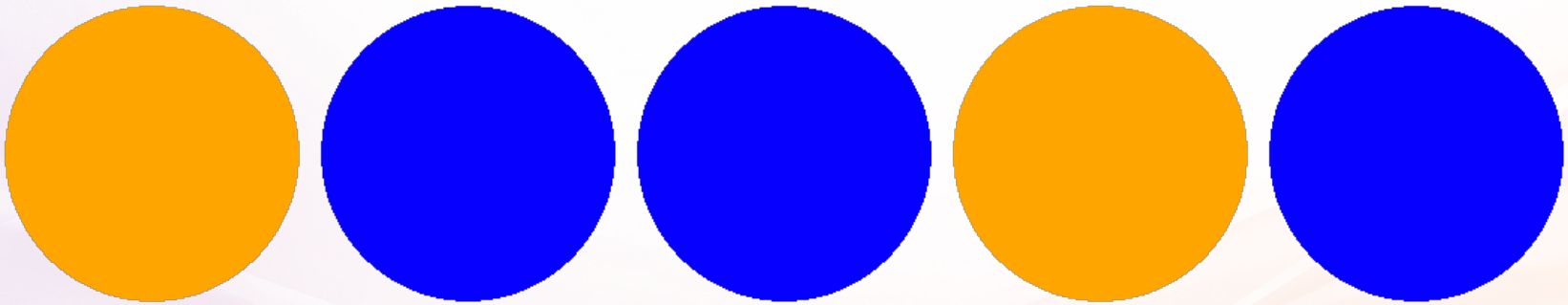
Множественная проверка гипотез

Поиск экстрасенсов

Этап 1: Угадайте цвета (**синий** и **оранжевый**) с учетом порядка.

Поиск экстрасенсов

Этап 1: ответы.

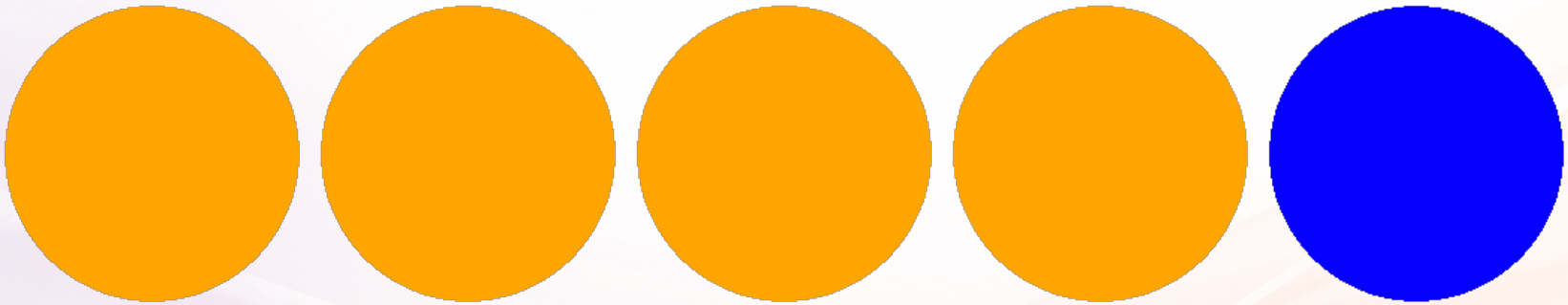


Поиск экстрасенсов

Этап 2: Угадайте цвета (**синий** и **оранжевый**) с учетом порядка.

Поиск экстрасенсов

Этап 2: ответы.



Поиск экстрасенсов

В 1950 г. проводились испытания

возможности экстрасенсорного восприятия.

Этап 1: поиск экстрасенсов — испытуемому нужно угадать цвет 10 карт.

$X_1, ..., X_{10} \sim Bern(\theta)$ — результаты (правильно / нет).

$H_0 : \theta = 1/2$ *vs.* $H_1 : \theta > 1/2$ — наугад *vs.* не наугад

Критерий $S = \{T(x) \geqslant c_\alpha\}$, где $T(X) = \sum_{i=1}^{n} X_i \sim Bin(n, \theta)$.

c	7	8	9	10
$P_{1/2}(T(X) \geqslant c)$	0.172	0.055	0.010	0.001

Берем $c_\alpha = 9$, т.е. H_0 отклоняется если $\sum X_i \geqslant 9$.

Поиск экстрасенсов

Вывод: если человек верно отгадывает хотя бы 9 карт из 10,

то он становится предполагаемым экстрасенсом.

В эксперименте приняли участие 1000 человек, при этом

- 9 карт верно отгадали 9 человек;
- 10 карт верно отгадали 2 человека.

В дальнейшем ни один из них не подтвердил свои способности...

$$\mathrm{P}_{1/2}\,(\text{хотя бы один из 1000 угадает 9 или 10 карт верно}) =$$

$$= 1 - \left(1 - C_{10}^{9}/2^{10} - C_{10}^{10}/2^{10}\right)^{1000} = 1 - \left(1 - 11/2^{10}\right)^{1000} \approx 0.99997$$

Материал на доске

Численный пример

Гипотезы, верность и полученные результаты:

Гипотеза	H_1	H_2	H_3	H_4	H_5	H_6	H_7
Верность	Нет	Да	Да	Нет	Да	Нет	Да
p-value	0.015	0.005	0.014	0.009	0.013	0.001	0.8

Верность известна тем, кто сгенерировал выборку, а не аналитикам :)

Перегруппируем:

Гипотеза	$H_{(1)}$	$H_{(2)}$	$H_{(3)}$	$H_{(4)}$	$H_{(5)}$	$H_{(6)}$	$H_{(7)}$
Верность	Нет	Да	Нет	Да	Да	Нет	Да
p-value	0.001	0.005	0.009	0.013	0.014	0.015	0.8

Гипотеза	$H_{(1)}$	$H_{(2)}$	$H_{(3)}$	$H_{(4)}$	$H_{(5)}$	$H_{(6)}$	$H_{(7)}$
Верность	Нет	Да	Нет	Да	Да	Нет	Да
p-value	0.001	0.005	0.009	0.013	0.014	0.015	0.8

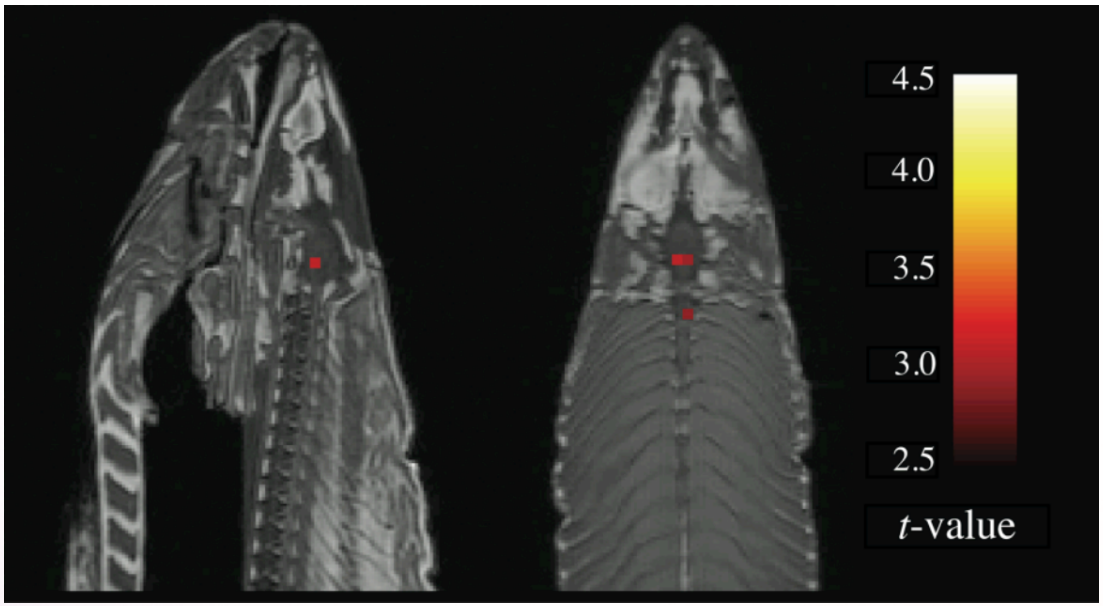
Метод Бонферрони:

Гипотеза	p-value	α_j	Отвергаем?
$H_{(1)}$	0.001	0.0071	True
$H_{(2)}$	0.005	0.0071	True
$H_{(3)}$	0.009	0.0071	False
$H_{(4)}$	0.013	0.0071	False
$H_{(5)}$	0.014	0.0071	False
$H_{(6)}$	0.015	0.0071	False
$H_{(7)}$	0.8	0.0071	False

Ошибок I рода: 1
Верных отвержений: 1

Удивительные открытия

2009 год. МРТ мозга мертвого самца лосося:



МРТ дает 3D-изображение на 130 000 вокселей.

Эксперимент: Лососю показывали фото и просили его пояснить, какие эмоции испытывают люди с картинки.

Обработка: Для каждого вокселя тестируется гипотеза о наличии активации этого участка мозга.

Удивительные открытия

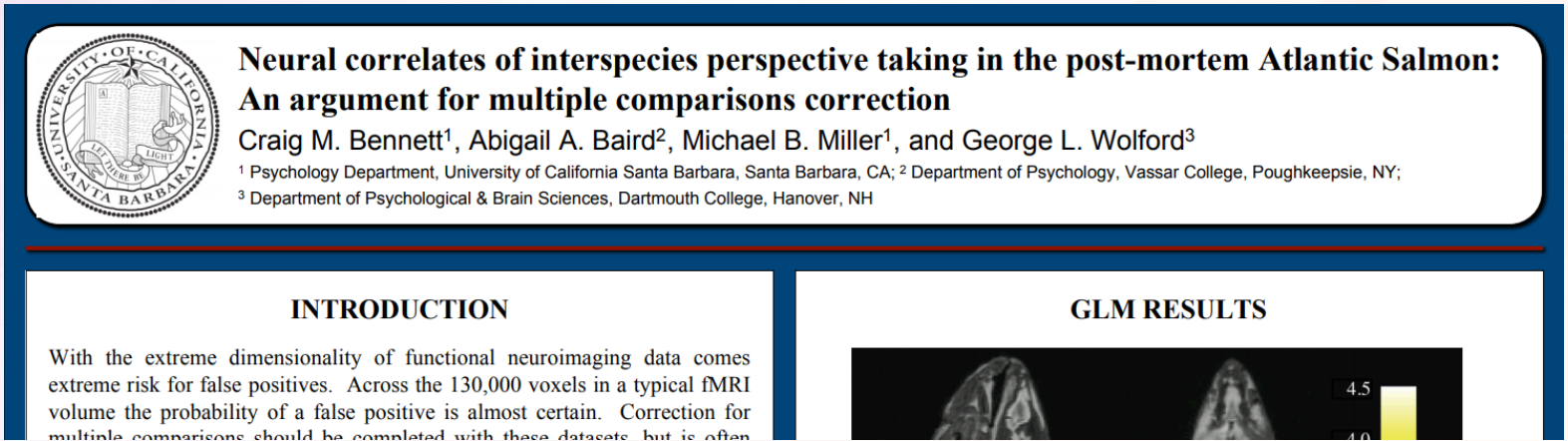
Результат: Для каждой картинки для нескольких вокселей мозга p-value оказывалось меньше 0.001.

Вывод: мертвый лосось реагирует на фотки!!!

Авторы удостоились Шнобелевской премии (2012 год)
за открытие в области неврологии.

При применении МПГ лосось переставал на что-либо реагировать...

http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for

GLM RESULTS


ВСЁ!