



DS-поток, 3 курс, осень 2022

## Статистика

Лекция 8

## 4. Проверка статистических гипотез

### 4.4. p-value

## Гипотезы и критерии (напоминание)

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

$H_0: P \in \mathcal{P}_0$  — основная гипотеза;

$H_1: P \in \mathcal{P}_1$  — альтернативная гипотеза.

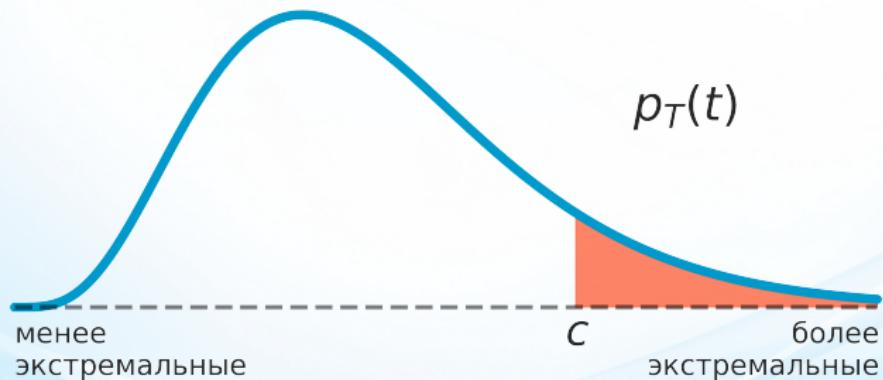
$S \subset \mathcal{X}$  — критерий уровня значимости  $\alpha$  для проверки  $H_0$  vs.  $H_1$ ,  
если  $P(X \in S) \leq \alpha, \forall P \in \mathcal{P}_0$ .

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  $S = \{T(x) \geq c\}$ ,  
где  $T(X)$  — статистика критерия.



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

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

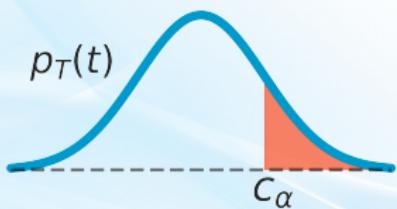
## Гипотезы и критерии (напоминание)

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

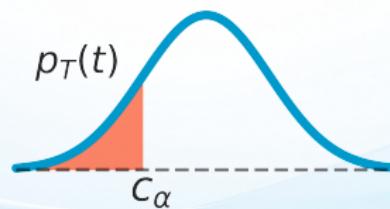
$\alpha$  выбирается **ДО** эксперимента,

$c_\alpha$  вычисляется из условия  $P_0(T(X) > c_\alpha) \leq \alpha$ .

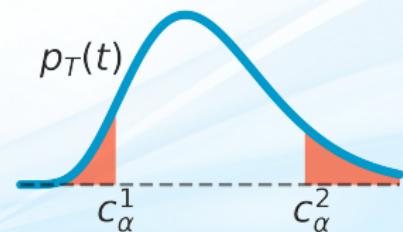
$$S = \{T(x) > c_\alpha\}$$



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



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



Замечание. Выбирать  $\alpha$  после эксперимента неправильно.

Так можно подогнать результат под желаемый.

"Статистика может доказать что угодно, даже истину."

## Пример: АВ-тест

Пользователи делятся случайно на две независимые группы:

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

Что может быть результатом?

- ▶ Клик по рекламе.
- ▶ Регистрация пользователя на сервисе.
- ▶ Покупка какой-либо услуги.
- ▶ и т.д.

Гипотезы:

$H_0: p_1 = p_2$  — отсутствие эффекта

$H_1: p_1 < p_2$  — эффект присутствует

## Пример: АВ-тест

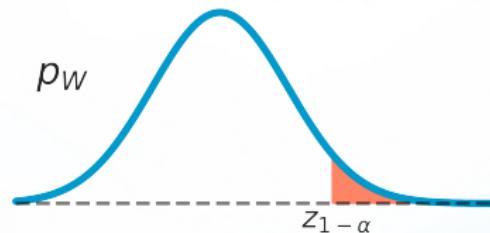
Из ЦПТ можем получить:

$$\hat{p}_1 = \bar{X} \stackrel{d}{\approx} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right), \quad \hat{p}_2 = \bar{Y} \stackrel{d}{\approx} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$$

При справедливости  $H_0$  получаем

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

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



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

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

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

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

$H_0$  отвергается  $\iff 0 \notin C$ .

## Пример: AB-тест

1. 1 группа:  $n = 30$  человек, 21 совершили действие  $\Rightarrow \hat{p}_1 = 0.7$   
2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 2 \Rightarrow H_0$  отвергается, результат стат. значим  
дов. интервал  $(0.036, 1)$
  
2. 1 группа:  $n = 30$  человек, 15 совершили действие  $\Rightarrow \hat{p}_1 = 0.5$   
2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 3.76 \Rightarrow H_0$  отвергается, результат стат. значим  
дов. интервал  $(0.225, 1)$
  
3. 1 группа:  $n = 10$  человек, 7 совершили действие  $\Rightarrow \hat{p}_1 = 0.7$   
2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 1.54 \Rightarrow H_0$  не отвергается, результат стат. незнач.  
дов. интервал  $(-0.017, 1)$

## Пример: AB-тест

1. 1 группа:  $n = 30$  человек, 21 совершили действие  $\Rightarrow \hat{p}_1 = 0.7$   
 2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 2 \Rightarrow H_0$  отвергается, результат стат. значим  
 дов. интервал  $(0.036, 1)$  ← слабая уверенность в результате
  
2. 1 группа:  $n = 30$  человек, 15 совершили действие  $\Rightarrow \hat{p}_1 = 0.5$   
 2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 3.76 \Rightarrow H_0$  отвергается, результат стат. значим  
 дов. интервал  $(0.225, 1)$  ← хорошая уверенность в результате
  
3. 1 группа:  $n = 10$  человек, 7 совершили действие  $\Rightarrow \hat{p}_1 = 0.7$   
 2 группа:  $m = 30$  человек, 27 совершили действие  $\Rightarrow \hat{p}_2 = 0.9$   
 $W(x, y) \approx 1.54 \Rightarrow H_0$  не отвергается, результат стат. незнач.  
 дов. интервал  $(-0.017, 1)$  ← нет результата

## p-value (достигаемый уровень значимости)

$$H_0: P \in \mathcal{P}_0$$

$x_1, \dots, x_n$  — реализация выборки

$T(X)$  — статистика критерия

$t = T(x_1, \dots, x_n)$  — реализация стат.

**p-value** — вероятность получить при справедливости  $H_0$  такое значение статистики  $t = T(x)$  или еще более экстремальное.

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

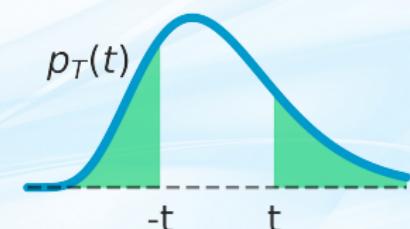
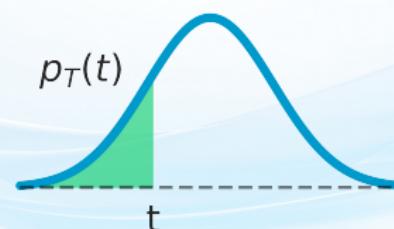
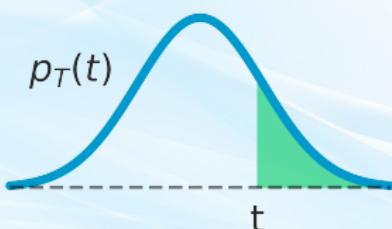
$$p(x) = P_0(T(X) \geq t),$$

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

$$p(x) = P_0(T(X) \leq t),$$

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

$$p(x) = P_0(T(X) \geq |t|) + P_0(T(X) \leq -|t|),$$



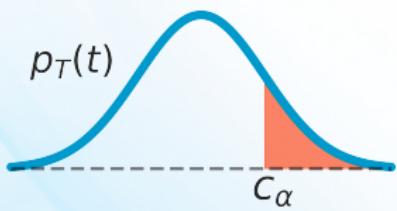
Замечание. Если распр.  $T(X)$  при  $H_0$  не одинаково, то нужно добавить  $\sup_{P \in \mathcal{P}_0}$ .

Замечание. Для асимпт. критерия берется предельное распределение.

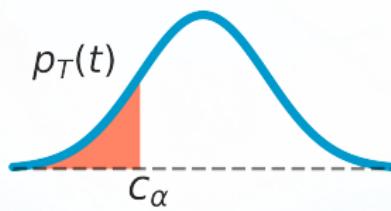
В чем же разница? Графики одинаковые!!!

Еще раз посмотрим на них:

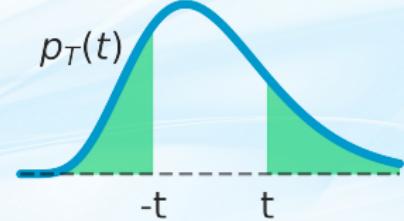
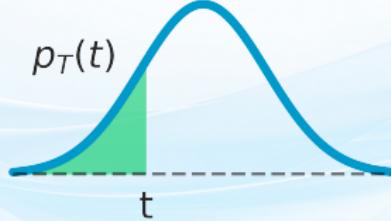
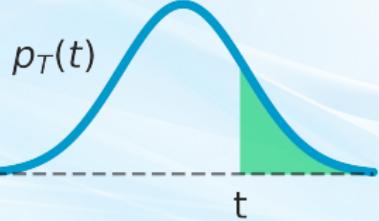
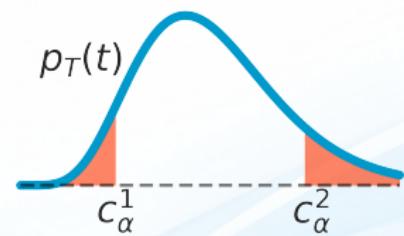
$$S = \{T(x) > c_\alpha\}$$



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



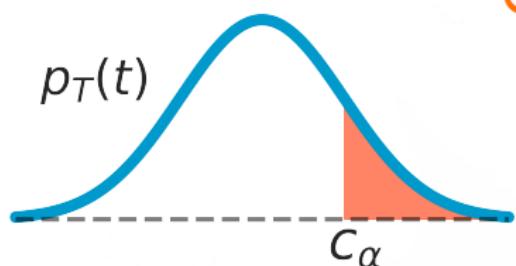
$$S = \{|T(x)| > c_\alpha\}$$



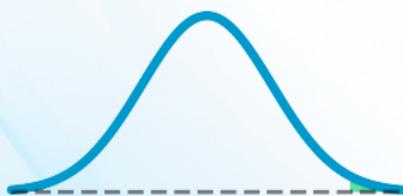
Рассмотрим случай  $S = \{T(x) > c_\alpha\}$

Критическое множество (слева) фиксировано.

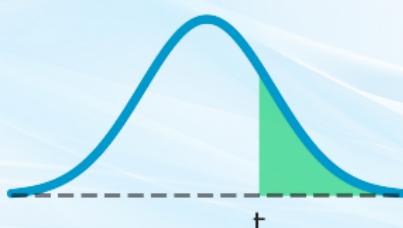
Ниже p-value для различных реализаций.



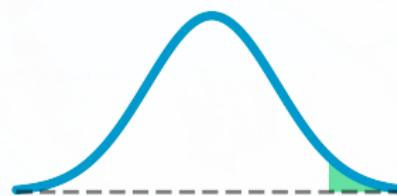
$p\text{-value}(t) = 0.014$



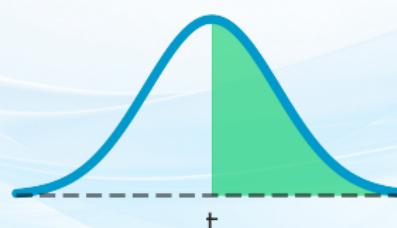
$p\text{-value}(t) = 0.212$



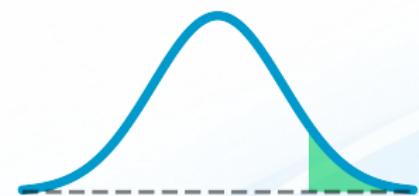
$p\text{-value}(t) = 0.036$



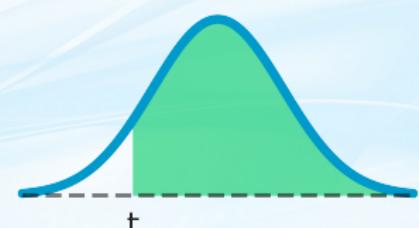
$p\text{-value}(t) = 0.500$



$p\text{-value}(t) = 0.081$



$p\text{-value}(t) = 0.903$



## Вывод:

$H_0$  отвергается  $\iff p\text{-value} \leq \alpha$

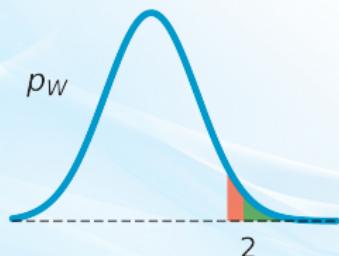
В этом случае  $p\text{-value}$  —  
степень уверенности в отвержении  $H_0$ .  
(чем  $p\text{-value}$  меньше, тем увереннее)

## Пример: АВ-тест

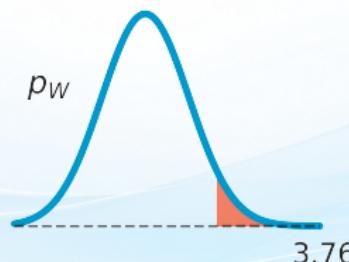
Критерий  $S = \{W(x, y) > z_{1-\alpha}\}$ , где  $W(X, Y) \xrightarrow{d} \mathcal{N}(0, 1)$ .

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

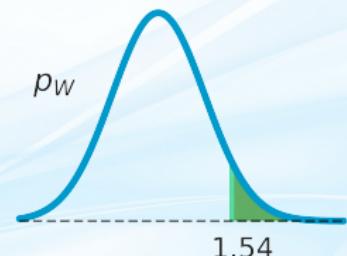
$$w = W(x) = 2$$
$$p(w) = 0.0228$$



$$w = W(x) = 3.76$$
$$p(w) = 0.00008$$



$$w = W(x) = 1.54$$
$$p(w) = 0.0618$$



## Свойство p-value

**Утверждение.**

Если при справедливости  $H_0$

распр. статистики  $T(X)$  одинаково и непрерывно,

то  $p(T(X)) \sim U[0, 1]$  при  $H_0$ .

**Доказательство.**

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

Функция распределения стат.  $T(X)$  при справедливости  $H_0$ :

$$F_T(t) = P(T(X) \leq t) = \sup_{P \in P_0} P(T(X) \leq t) = p(t),$$

где использовано то, что распр. статистики  $T(X)$  одинаково при  $H_0$ .

Поскольку распределение стат.  $T(X)$  непрерывно, то

$$p(T(X)) = F_T(T(X)) \sim U[0, 1].$$

# Свойство p-value

**Утверждение.**

Если при справедливости  $H_0$

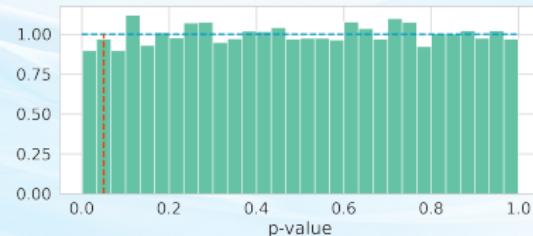
распр. статистики  $T(X)$  одинаково и непрерывно,

то  $p(T(X)) \sim U[0, 1]$  при  $H_0$ .

**Замечание.**

Часто на практике это верно, т.к. многие критерии так и строятся.

Перед применением критерия на практике стоит проверить его корректность с помощью многократного семплирования в предположении справедливости  $H_0$ . В частности, p-value обычно ведет себя так:





## Свойство p-value

**Утверждение.**

Если при справедливости  $H_0$

распр. статистики  $T(X)$  одинаково и непрерывно,

то  $p(T(X)) \sim U[0, 1]$  при  $H_0$ .

*Замечание.*

Часто на практике это верно, т.к. многие критерии так и строятся.

*Следствие.*

Значения 0.2 и 0.9 одинаковы с точки зрения справедливости  $H_0$ ,  
т.е. p-value не есть степень уверенности в справедливости  $H_0$ .

Что возможно, если p-value большой?

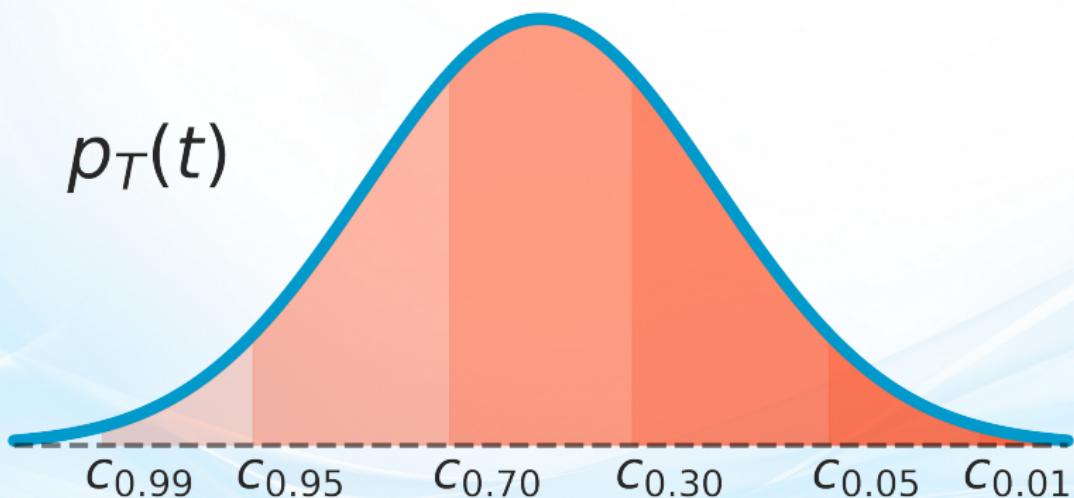
1.  $H_0$  верна;
2. Критерий недостаточно мощный.

## Общий случай p-value

$\{S_\alpha \mid \alpha \in [0, 1]\}$  — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$  — критерий

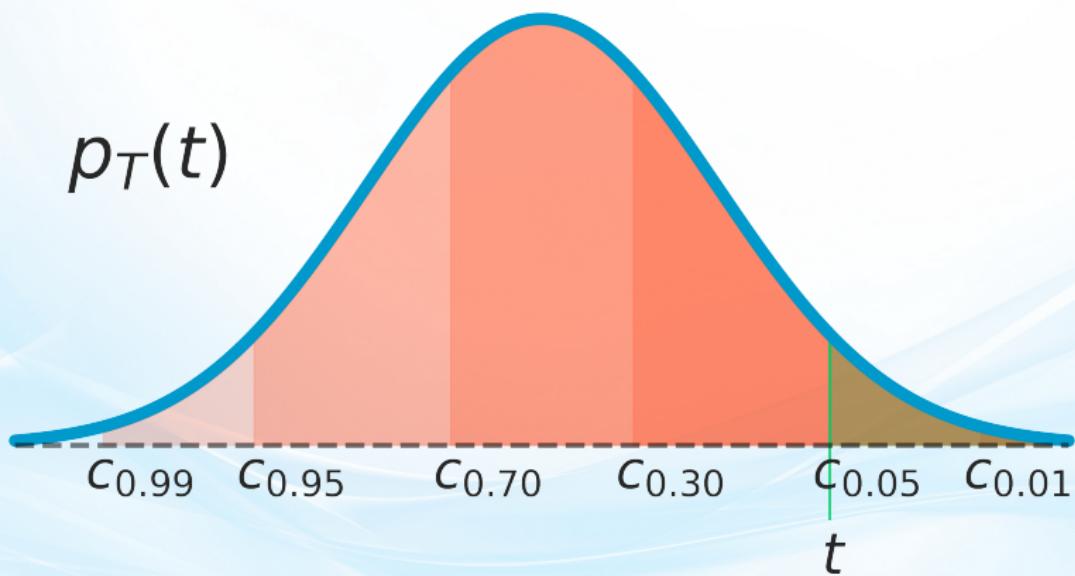
$$p_T(t)$$



## Общий случай p-value

$\{S_\alpha \mid \alpha \in [0, 1]\}$  — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$  — критерий

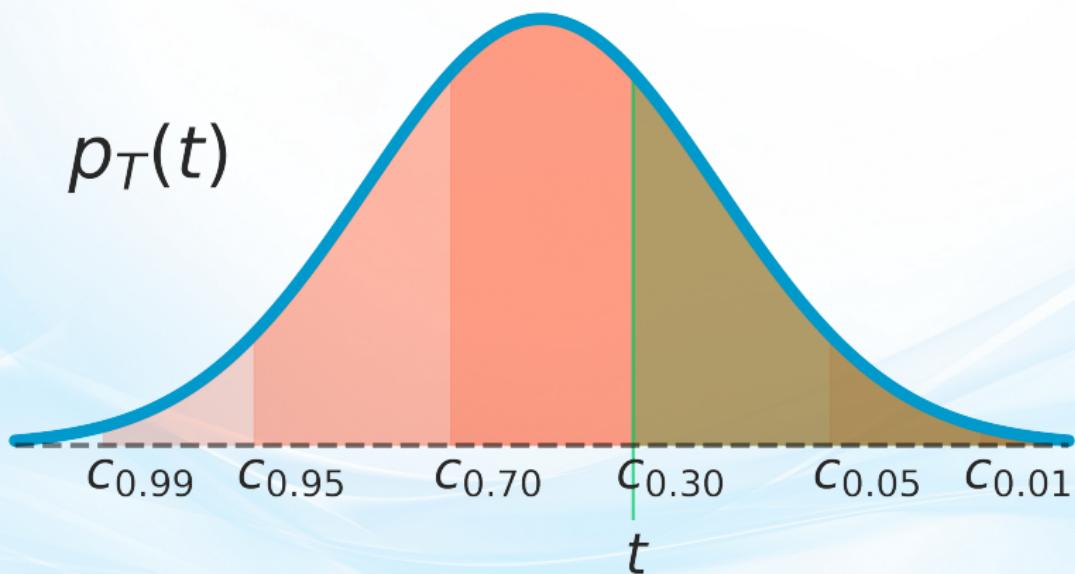


$$p\text{-value}(t) = 0.05$$

## Общий случай p-value

$\{S_\alpha \mid \alpha \in [0, 1]\}$  — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$  — критерий

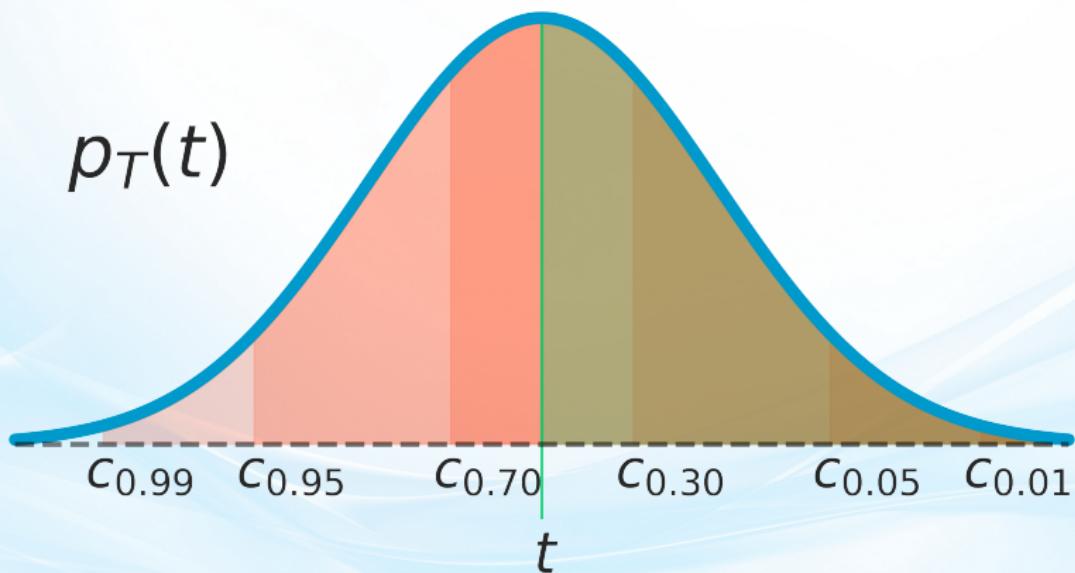


$$p\text{-value}(t) = 0.30$$

## Общий случай p-value

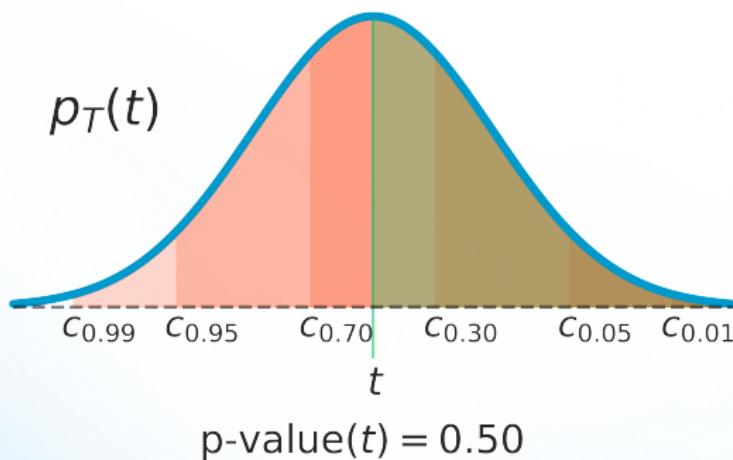
$\{S_\alpha \mid \alpha \in [0, 1]\}$  — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$  — критерий



$$p\text{-value}(t) = 0.50$$

## Общий случай p-value



$t = c_{0.5} \implies$  при  $\alpha \geq 0.5$  гипотеза  $H_0$  отвергается.

при  $\alpha < 0.5$  гипотеза  $H_0$  не отвергается.

**Ключевое наблюдение:** Если отвергнуть  $H_0$  можно только совершив большую ошибку, то скорее ее не стоит отклонять.

## Общий случай p-value

### Вывод:

p-value — наименьший уровень значимости, при котором  $H_0$  можно отвергнуть для данной реализации выборки  $x$ .

$$p(x) = \inf\{\alpha \in [0, 1] \mid x \in S_\alpha\}$$

# Что не есть p-value

Величина p-value не является

- ▶ уровнем значимости, реальным уровнем значимости, вероятностью ошибки первого рода;  
**не зависят от выборки**
- ▶ вероятностью  $H_0$ , вероятностью  $H_0$  при условии выборки;  
**она либо верна, либо нет**
- ▶ многое еще.

# Что не есть p-value (Пример)

На ЧМ по футболу в 2010 г. осьминог Пауль предсказывает результаты матчей с участием сборной Германии, выбирая кормушку с флагом страны-победителя.



$X_1, \dots, X_n \sim Bern(\theta)$  — результаты предсказания (правильно/нет).

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

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

p-value:  $p(t) = \frac{1}{2^n} \sum_{j=t}^n C_n^j$ .

13 матчей: Пауль верно угадывает исход матча в 11 случаях.

$p(11) = 2^{-13} (C_{13}^{11} + C_{13}^{12} + C_{13}^{13}) \approx 0.0112 < 0.05 \implies H_0$  отвергается;

0.0112 не является вероятностью того, что Пауль выбирает кормушку наугад

## Выводы:

$H_0$  отвергается  $\iff p\text{-value} \leq \alpha$

Если  $H_0$  отвергается, то  $p\text{-value}$  –  
степень уверенности в отвержении  $H_0$ .  
(чем  $p\text{-value}$  меньше, тем увереннее)

Если  $H_0$  не отверг., то ничего сказать нельзя,  
 $p\text{-value}$  не есть степень уверенности в справ.  $H_0$ .

Если  $H_0$  верна, то обычно  $p(T(X)) \sim U[0, 1]$ .

## 4. Проверка статистических гипотез

### 4.5. Практическая значимость результата

# Большие выборки

$X_1, \dots, X_n \sim Bern(\theta)$  — результаты испытания Пауля.

$H_0: \theta = 1/2$  vs.  $H_1: \theta > 1/2$

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

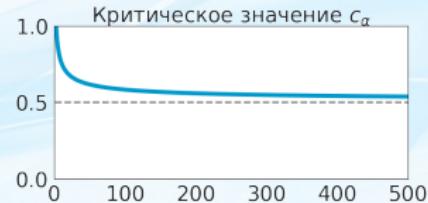
Как критическое значение  $c_\alpha$  зависит от  $n$ ?

Рассмотрим асимптотически эквивалентный критерий Вальда

$$W(X) = \sqrt{n} \frac{\bar{X} - 1/2}{\sqrt{1/4}} \xrightarrow{d_{1/2}} \mathcal{N}(0, 1).$$

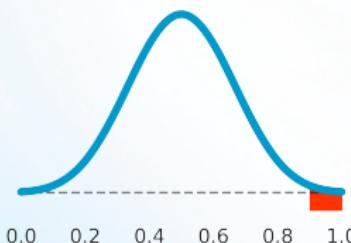
Тогда критерий

$$S_W = \{W(x) > z_{1-\alpha}\} = \left\{ \bar{x} > \frac{1}{2} + \frac{z_{1-\alpha}}{2\sqrt{n}} \right\}.$$

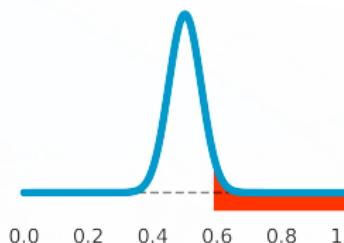


# Большие выборки

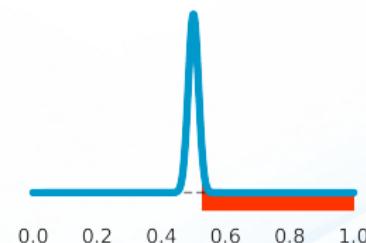
$$n = 10, c_\alpha = 0.9$$



$$n = 100, c_\alpha = 0.59$$



$$n = 1000, c_\alpha = 0.527$$



**Вывод:** при  $n \rightarrow +\infty$  мощность критерия сходится к 1.

**Теория:** это замечательно!

**Практика:** не совсем...

**Теория:** даже при небольшом отличии истины от  $H_0$  мы ее отклоним!

**Практика:** Хахаха, какой смысл в осьминоге, который угадывает с вероятностью 0.51? На гриль его!

# Вывод с точки зрения практики

Как правило, на практике:

## 1. При малом размере выборки:

Почти ничего не отклоняется, т.к. мощность небольшая.

*Недообучение*

## 2. При большом размере выборки:

Отклоняется почти все, т.к. небольшие отличия от  $H_0$  есть почти всегда.

*Переобучение*

# Практическая значимость

**Размер эффекта** — величина, оценивающая по данным, насколько основная гипотеза отличается от истины.

## Пример 1.

В течении трех лет женщины выполняли физические упражнения:

1. группа А — не менее часа в день;
2. группа В — не более 20 минут в день.

$H_0$ : изменение веса в обоих группах одинаково.

$H_1$ : в группе А уменьшение веса больше, чем в группе В.

$p\text{-value} < 0.001 \implies$  результат *статистически* значим.

Разница в весе **150 грамм**  $\implies$  результат *практически* не значим.

# Практическая значимость

**Размер эффекта** — величина, оценивающая по данным, насколько основная гипотеза отличается от истины.

## Пример 2.

В 2002 году проводились клинические испытания гормонального препарата Премарин, которые были досрочно прерваны.

На **0.08%** увеличивается риску развития рака груди;

На **0.08%** увеличивается риск инсульта;

На **0.07%** увеличивается риск инфаркта.

С учетом численности населения есть практическая значимость.

# Практическая значимость

	Есть практ. значимость	Нет практ. значимости
Есть стат. значимость	$H_0$ отвергается: Эффект присутствует и доказан статистически	Скорее всего в полученном результате смысла нет
Нет стат. значимости	Эффект присутствует, но не доказан статистически; нужно продолжать эксперимент	$H_0$ не отвергается: результата нет

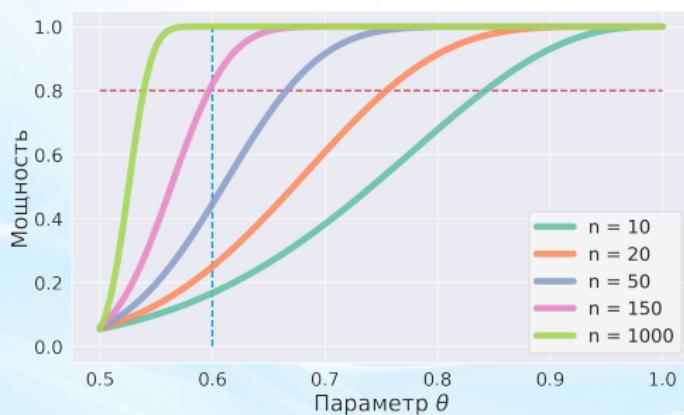
# План эксперимента

Как определить размер выборки **до** эксперимента?

$$X_1, \dots, X_n \sim Bern(\theta)$$

$$H_0: \theta = 1/2 \text{ vs. } H_1: \theta > 1/2$$

Графики мощности для критерия  $S = \{\sum X_i > c_\alpha\}$ :



$\alpha = 0.05$  — ур. знач.

Желаемые значения:

$\beta = 0.8$  — мощность;

$\theta \geq 0.6$  — значимый эффект.

Выбираем  $n$ , для которого  
кривая мощности проходит  
через точку  $(0.6, 0.8)$ .

## 4. Проверка статистических гипотез

### 4.6. Множественная проверка гипотез

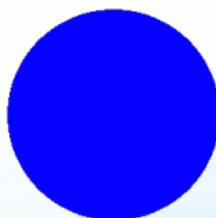
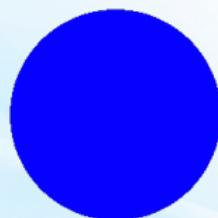
# Поиск экстрасенсов

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



# Поиск экстрасенсов

**Этап 1:** ответы.



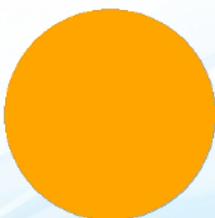
## Поиск экстрасенсов

**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.

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

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

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

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

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

## 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...

$$\begin{aligned} P_{1/2} (\text{хотя бы один из } 1000 \text{ угадает } 9 \text{ или } 10 \text{ карт верно}) &= \\ &= 1 - \left(1 - C_{10}^9 / 2^{10} - C_{10}^{10} / 2^{10}\right)^{1000} = 1 - \left(1 - 11/2^{10}\right)^{1000} \approx 0.99997 \end{aligned}$$

# Гипотезы и критерии

## Проверка гипотез:

$X = (X_1, \dots, X_n)$  — выборка из неизвестного распр.  $P_j$ .

$H_0: P \in \mathcal{P}_0$  — проверяемая гипотеза.

$S$  — критерий для проверки  $H_0$ , если  $H_0$  отвергается  $\Leftrightarrow X \in S$ .

$P(I_S) \leq \alpha$  — уровень значимости.

## Размножим по $j \in \{1, \dots, m\}$ :

$X_j = (X_{j,1}, \dots, X_{j,n_j})$  —  $j$ -ая выборка из неизвестного распр.  $P_j$ .

$H_j: P_j \in \mathcal{P}_j$  — проверяемая гипотеза.

$S_j$  — критерий для проверки  $H_j$ , если  $H_j$  отвергается  $\Leftrightarrow X_j \in S_j$ .

$P(I_{S_j}) \leq \alpha$  — уровень значимости.

# Обобщение ошибки

**Групповая ошибка первого рода** (*familywise error rate*)

Вероятность отвергнуть хотя бы одну верную гипотезу.

$$FWER = P(V_{P,S} > 0),$$

$V_{P,S}$  — количество верных гипотез, которые были отвергнуты критерием  $S$  для распределения  $P$  (задает верные гипотезы).

**Что мы знаем?**

Пусть  $H_1, \dots, H_{m_0}$  — верные гипотезы. ( $m_0$  не знаем)

$FWER = P(\text{произошла хотя бы одна ошибка I рода}) =$

$$= P\left(\bigcup_{j=1}^{m_0} \{X_j \in S_j\}\right) \leq \sum_{j=1}^{m_0} P(X_j \in S_j) \leq \alpha m_0$$

А нам нужно  $FWER \leq \alpha$ .

# Методы контроля FWER

## Метод Бонферрони

Каждый критерий имеет уровень значимости  $\alpha/m$ .

## Метод Холма

Пересортируем гипотезы и критерии в порядке возрастания p-value:

$p_1 \leq \dots \leq p_m$  — p-value.

$H_1, \dots, H_m$  — соответствующие гипотезы.

$\alpha_j = \frac{\alpha}{m-j+1}$  — уровень значимости критерия  $S_j$ .

Пусть  $j = \min\{j \mid p_j > \alpha_j\}$  — номер первой неотвергнутой гипотезы.

Отвергаем гипотезы  $H_1, \dots, H_{j-1}$ .

Всегда мощнее метода Бонферрони.

# Обобщение ошибки

**Ожидаемая доля ложных отклонений** (*false discovery rate*)

$$FDR = E_P \frac{V_{P,S}}{\max(R_S, 1)},$$

$V_{P,S}$  — количество верных гипотез, которые были отвергнуты,

$R_S$  — количество отвергнутых гипотез

**Утверждение:**  $FDR \leq FWER$

## Методы Бенджамини-Хохберга и Бенджамини-Иекутиели

Пересортируем гипотезы и критерии в порядке возрастания p-value.

Пусть  $j = \max\{j \mid p_j \leq \alpha_j\}$  — номер последней отвергнутой гипотезы

Отвергаем гипотезы  $H_1, \dots, H_j$ .

$\alpha_j = \frac{\alpha j}{m}$  — метод Б.-Х. (если статистики критериев независимы)

$\alpha_j = \frac{\alpha j}{m} / \sum_{j=1}^m \frac{1}{j}$  — метод Б.-И. (работает всегда)

## Замечания

Зависимости:

1.  $V_{P,S}$  зависит от совместн. распр.  $P$ , общего критерия  $S$  и выборок  $X$ ;
2.  $R_S$  зависит от общего критерия  $S$  и выборок  $X$ ;
3.  $FWER$  и  $FDR$  зависят от совместного распр.  $P$  и общего критерия  $S$ .

Контроль  $FWER$  на уровне  $\alpha$  означает:

for  $\forall P_1$  — распределение выборки  $X_1$  (в т.ч. не из  $H_1$ ):

...

for  $\forall P_m$  — распределение выборки  $X_m$  (в т.ч. не из  $H_m$ ):

проверить  $FWER \leq \alpha$ .

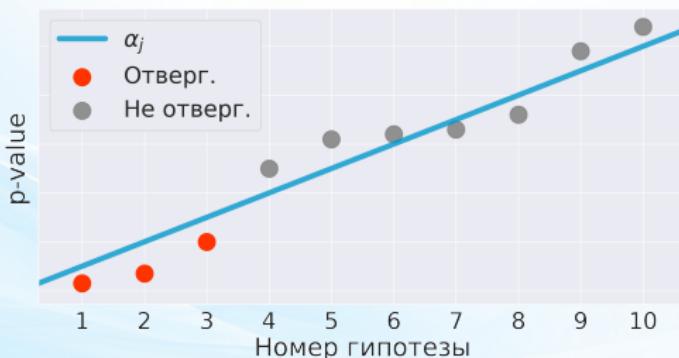
Максимум FWER не обязательно достигается  
при справедливости всех  $H_1, \dots, H_m$ .

# Итерационные процедуры

## Нисходящая процедура

В методе Холма можно выполнять следующие итерации.

1. Если  $p_1 \leq \alpha_1$ , то отвергнуть  $H_1$  и продолжить, иначе не отвергнуть  $H_1, \dots, H_m$  и остановиться;
2. Если  $p_2 \leq \alpha_2$ , то отвергнуть  $H_2$  и продолжить, иначе не отвергнуть  $H_2, \dots, H_m$  и остановиться;
3. и т.д.

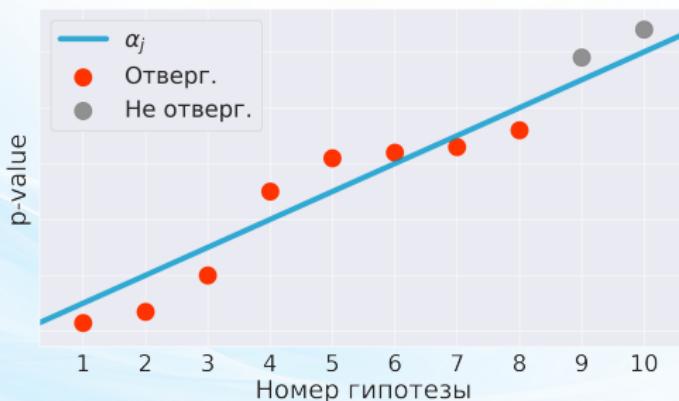


# Итерационные процедуры

## Восходящая процедура

В методах Б.-Х. и Б.-И. можно выполнять следующие итерации.

1. Если  $p_m > \alpha_m$ , то не отвергать  $H_m$  и продолжить, иначе отвергнуть  $H_m, \dots, H_1$  и остановиться;
2. Если  $p_{m-1} > \alpha_{m-1}$ , то не отвергать  $H_{m-1}$  и продолжить, иначе отвергнуть  $H_{m-1}, \dots, H_1$  и остановиться;
3. и т.д.



## Скорректированные p-value

$p_j$  — p-value, сравниваем с  $\alpha_j$  [+процедура]

$\tilde{p}_j$  — скорректированные p-value, хотим сравнивать с  $\alpha$

### Метод Бонферрони

Запишем пограничные случаи:

$$\begin{cases} p_j = \alpha_j \quad (= \alpha/m) \\ \tilde{p}_j = \alpha \end{cases} \implies \tilde{p}_j = mp_j$$

Чтобы  $\tilde{p}_j \in [0, 1]$  поправим их:  $\tilde{p}_j = \min(1, mp_j)$

### Метод Холма

Аналогично получаем  $\tilde{p}_j = (m - j + 1)p_j$

Если  $H_{j-1}$  не отверглась, то дальше не отвергаем

$$\Rightarrow \tilde{p}_j = \max(\tilde{p}_{j-1}, (m - j + 1)p_j)$$

Чтобы  $\tilde{p}_j \in [0, 1]$  поправим их:  $\tilde{p}_j = \min(1, \max[\tilde{p}_{j-1}, (m - j + 1)p_j])$

# 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	$\alpha_j$	p-value adj	Отвергаем?
$H_{(1)}$	0.001	0.0071	0.007	True
$H_{(2)}$	0.005	0.0071	0.035	True
$H_{(3)}$	0.009	0.0071	0.063	False
$H_{(4)}$	0.013	0.0071	0.091	False
$H_{(5)}$	0.014	0.0071	0.098	False
$H_{(6)}$	0.015	0.0071	0.105	False
$H_{(7)}$	0.8	0.0071	1.0	False

Ошибок I рода: 1

Верных отвержений: 1

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 Холма:

Гипотеза	p-value	$\alpha_j$	p-value adj	Отвергаем?
$H_{(1)}$	0.001	0.0071	0.007	True
$H_{(2)}$	0.005	0.0083	0.03	True
$H_{(3)}$	0.009	0.0100	0.045	True
$H_{(4)}$	0.013	0.0125	0.052	False
$H_{(5)}$	0.014	0.0167	0.052	False
$H_{(6)}$	0.015	0.0250	0.052	False
$H_{(7)}$	0.8	0.0500	0.8	False

Ошибок I рода: 1

Верных отвержений: 2

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

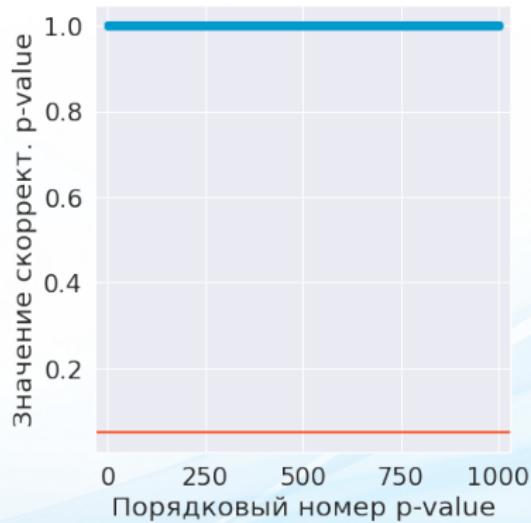
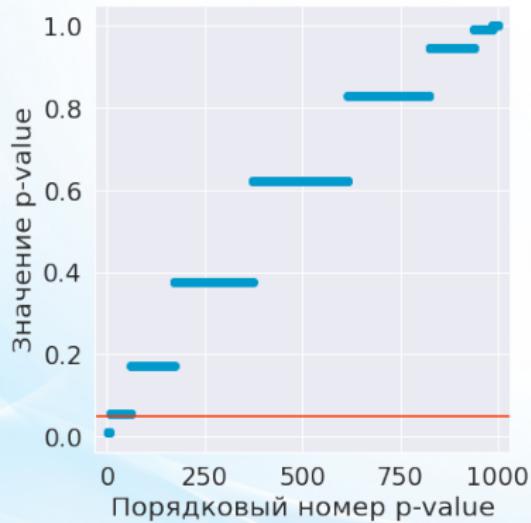
Метод Бенджамина-Иекутиелли:

Гипотеза	p-value	$\alpha_j$	p-value adj	Отвергаем?
$H_{(7)}$	0.8	0.0193	1.0	False
$H_{(6)}$	0.015	0.0165	0.045375	True
$H_{(5)}$	0.014	0.0138	0.045375	True
$H_{(4)}$	0.013	0.0110	0.045375	True
$H_{(3)}$	0.009	0.0083	0.045375	True
$H_{(2)}$	0.005	0.0055	0.045375	True
$H_{(1)}$	0.001	0.0028	0.01815	True

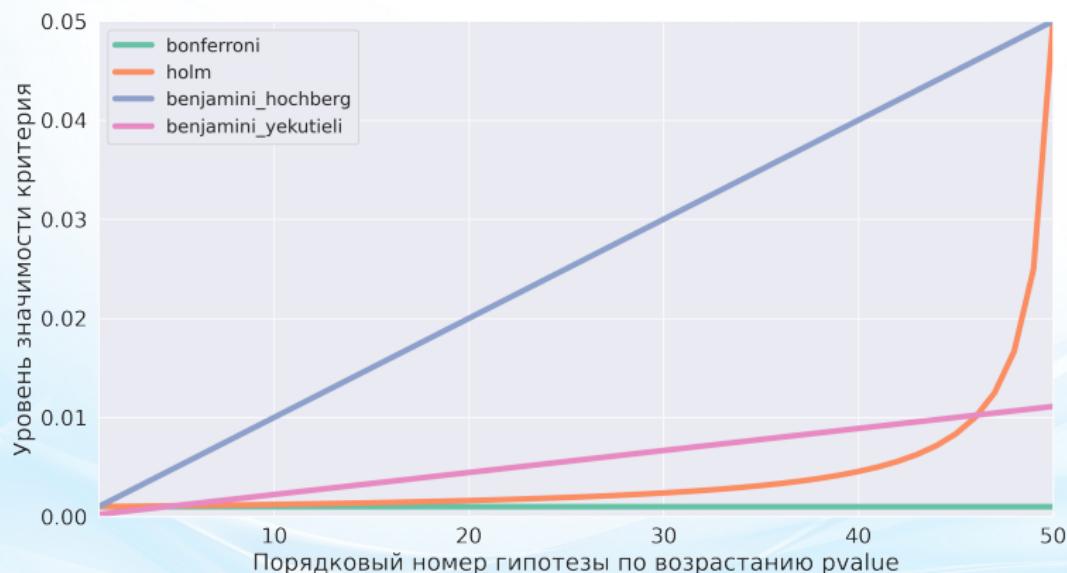
Ошибок I рода: 3

Верных отвержений: 3

# Поиск экстрасенсов



# Сравнение методов МПГ



# Реализация МПГ

```
statsmodels.stats.multitest.multipletests  
(pvals, alpha=0.05, method='hs',  
is_sorted=False, returnsorted=False)
```

- ▶ pvals — значения p-value по всем критериям
- ▶ alpha — желаемый уровень значимости
- ▶ method:
  - ▶ bonferroni
  - ▶ sidak
  - ▶ fdr\_bh
  - ▶ holm
  - ▶ holm-sidak
  - ▶ fdr\_by

Возвращает:

- ▶ reject — для отвергаемых гипотез True
- ▶ pvals\_corrected — скорректированные p-value

# Простой пример

Знакомая задача:

$$X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$$

$$H_0: \theta \geq 0 \text{ vs } H_1: \theta < 0$$

$$\text{РНМК: } S = \{x \in \mathbb{R} \mid \bar{x} \leq c_\alpha\}$$

Пусть теперь две одинаковые задачи с независимыми выборками:

$$X_1, \dots, X_n \sim \mathcal{N}(\theta_1, 1)$$

$$Y_1, \dots, Y_n \sim \mathcal{N}(\theta_2, 1)$$

$$H_1: \theta_1 \geq 0 \text{ vs } H'_1: \theta_1 < 0$$

$$H_2: \theta_2 \geq 0 \text{ vs } H'_2: \theta_2 < 0$$

$$\text{Критерии: } S_1 = \{(x, y) \in \mathbb{R}^2 \mid \bar{x} \leq c_\alpha\}$$

$$S_2 = \{(x, y) \in \mathbb{R}^2 \mid \bar{y} \leq c_\alpha\}$$

**Частая ошибка:** Выборки независимы  $\implies$  МПГ не нужна.

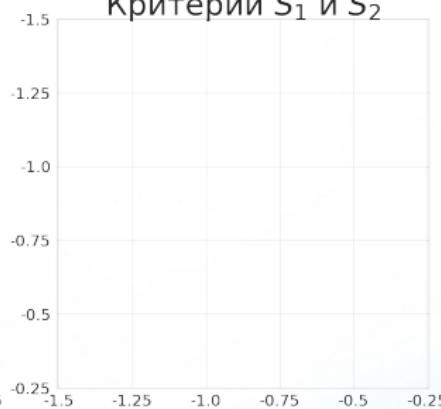
Критерий  $S_1$



Критерий  $S_2$



Критерии  $S_1$  и  $S_2$



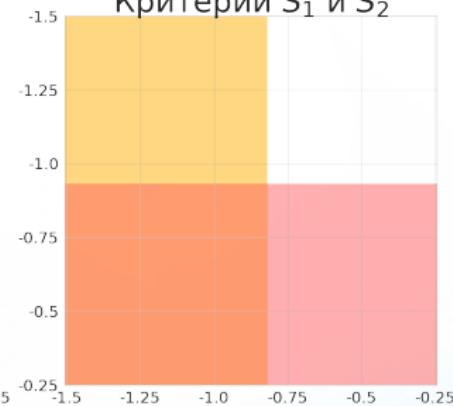
Критерий  $S_1$



Критерий  $S_2$



Критерии  $S_1$  и  $S_2$

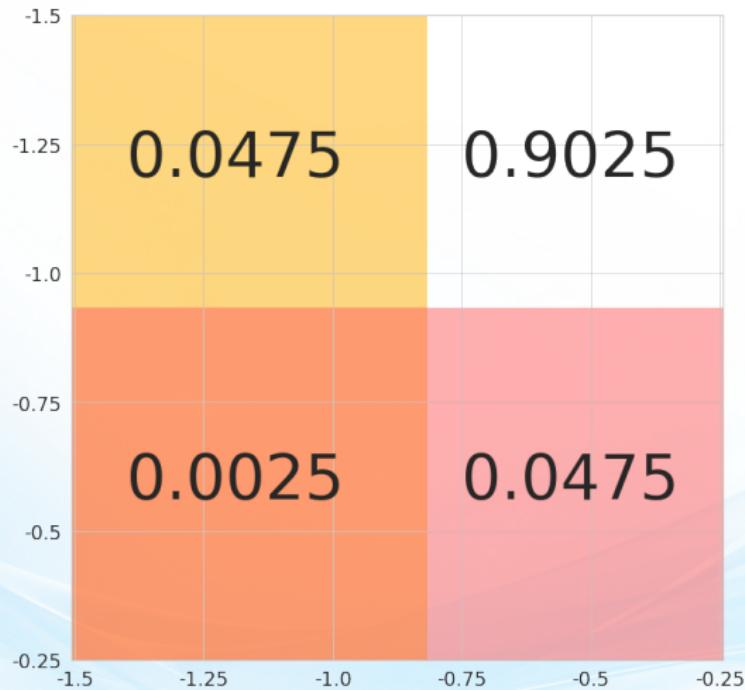


Критерий  $S_1$ Критерий  $S_2$ Критерии  $S_1$  и  $S_2$ 

**Вывод:**

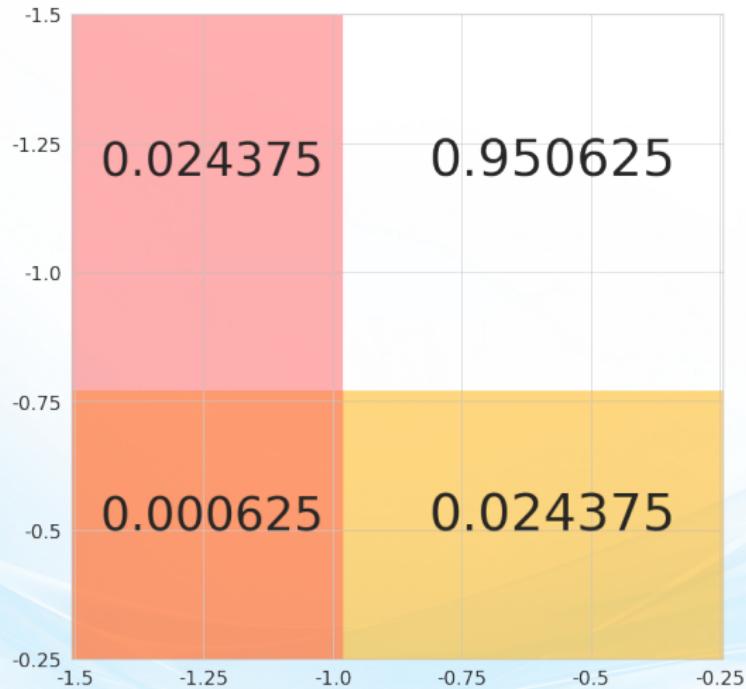
вероятность допустить хотя бы одну ошибку равна 0.0975,  
если обе основные гипотезы верны.

## Сравнение: без корректировки



Вероятности указаны при справедливости  $H_1$  и  $H_2$ .

## Сравнение: метод Бонферрони



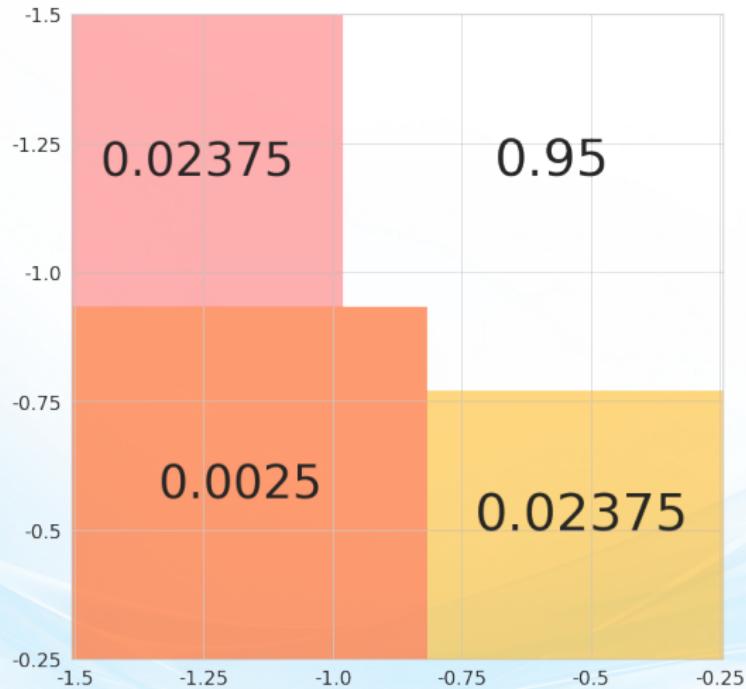
Вероятности указаны при справедливости  $H_1$  и  $H_2$ .

## Сравнение: метод Холма



Вероятности указаны при справедливости  $H_1$  и  $H_2$ .

## Сравнение: метод Бенджамина-Хохберга (не контр. FWER)



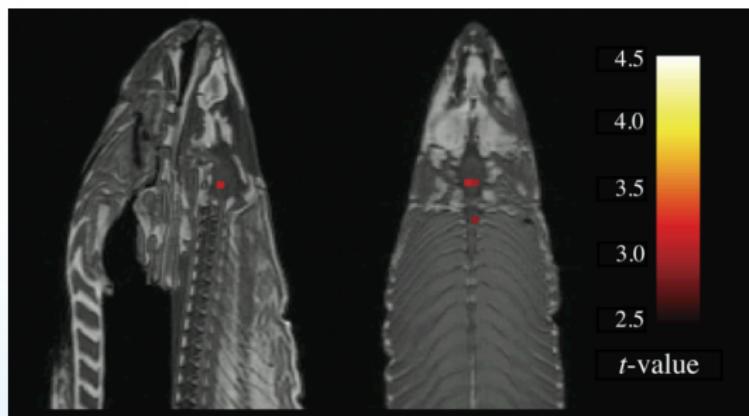
Вероятности указаны при справедливости  $H_1$  и  $H_2$ .

## Какой подход использовать?

- ▶ При первичном анализе данных, при котором только происходит формулировка интересных гипотез, можно вообще не делать поправки на МПГ. При этом **всегда** нужно приводить информацию об общем количестве гипотез и количестве отвергнутых.
- ▶ При проведении исследований и отбора признаков для дальнейшего анализа, который обычно является более сложным и дорогим, следует применять методы, контролирующие **FDR**. Обычно берут  $FDR \leq 0.1$ .
- ▶ На этапе подтверждения выводов следует проводить строгий контроль за вероятностью ошибок первого рода, контролируя **FWER**. Обычно берут  $FWER \leq 0.05$ .

# Удивительные открытия

2009 год. МРТ мозга мертвого самца лосося:



МРТ дает 3D-изображение на 130 000 вокселей.

**Эксперимент:** Лосося 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 все!!!

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<sup>1</sup>, Abigail A. Baird<sup>2</sup>, Michael B. Miller<sup>1</sup>, and George L. Wolford<sup>3</sup>

<sup>1</sup> Psychology Department, University of California Santa Barbara, Santa Barbara, CA; <sup>2</sup> Department of Psychology, Vassar College, Poughkeepsie, NY;

<sup>3</sup> 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 multiple comparisons should be completed with these datasets but is often

### GLM RESULTS



Theta



BCE !