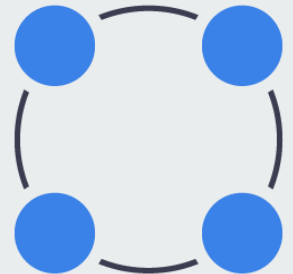




# Статистика и анализ данных

Лекция 5. Тестирование гипотез

(15.10.2022)



---

# Тестирование гипотез

# Примеры гипотез



## Нулевая гипотеза

Средний объем легких у курящих и не курящих людей не различается

## Альтернативная гипотеза

### *Односторонняя*

### *Двусторонняя*

Объем легких у курящих людей  
меньше/больше, чем у некурящих

Объем легких у курящих и не  
курящих людей различается

# Сравнение средних. z-критерий

Формулировка:

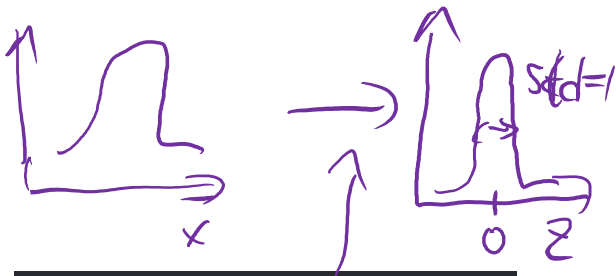
$$H_0: \bar{x} = \mu$$

$$H_1: \bar{x} \neq \mu$$

$$X \sim N(\mu, \sigma)$$

$\sigma$  известны

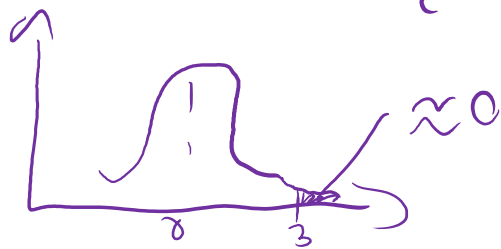
$$Z \sim N(0, 1)$$



$$Z_N = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

# Сравнение средних. z-критерий

Распределение статистики:

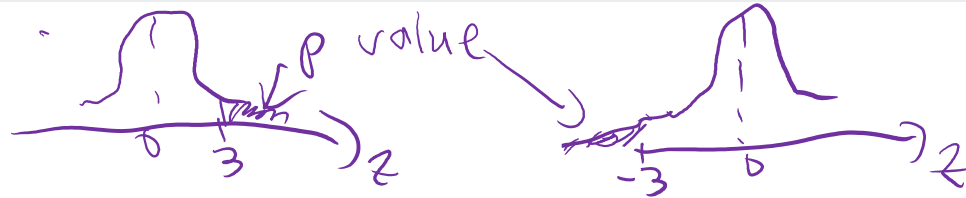


$$Z_N = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

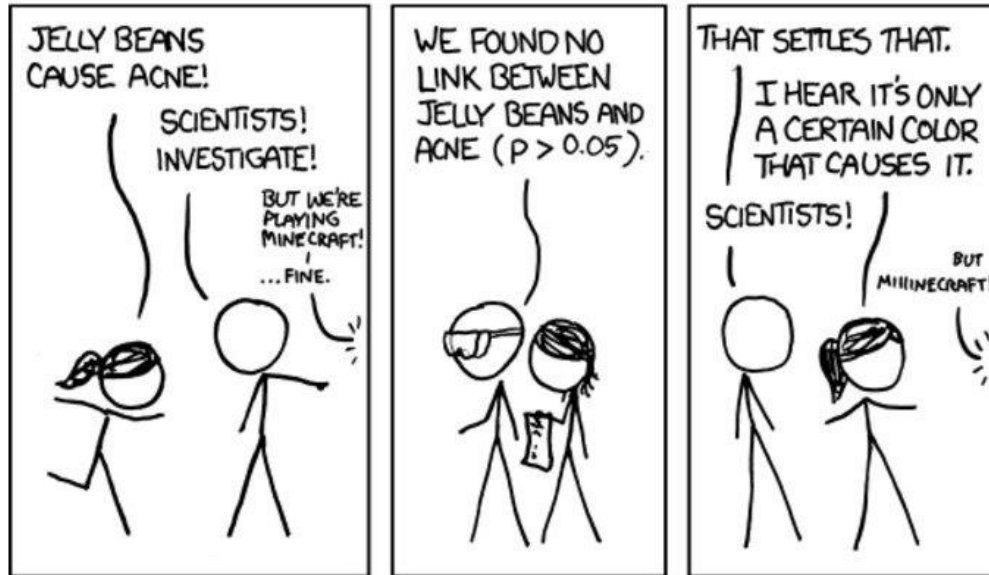
0.1  
3.

178  
180

# p-value



Вероятность получить более экстремальное значение статистики, когда нулевая гипотеза верна



# Заблуждения о p-value



1. p-value вероятность того, что верна сама  $H_0$  — нет, расчет производится при условии, что  $H_0$  верна
2. p-value — это вероятность получить такое значение статистики при справедливой  $H_0$  — такое или более экстремальное
3.  $p > 0.05$ , то различий между группами на самом деле нет (у нас просто нет оснований отклонить нулевую гипотезу)

# Ошибки при тестировании гипотез

	H0 верна	H0 не верна
Отклонить H0	Ошибка I рода	✓
Принять H0	✓	Ошибка II рода

Какая ошибка хуже?...Зависит от ситуации.

В науке считается, что ошибка I рода опаснее, так как нарушает бритву Оккама “не плодить сущности сверх необходимого”.

Но в медицине обе ошибки могут стоить очень дорого.



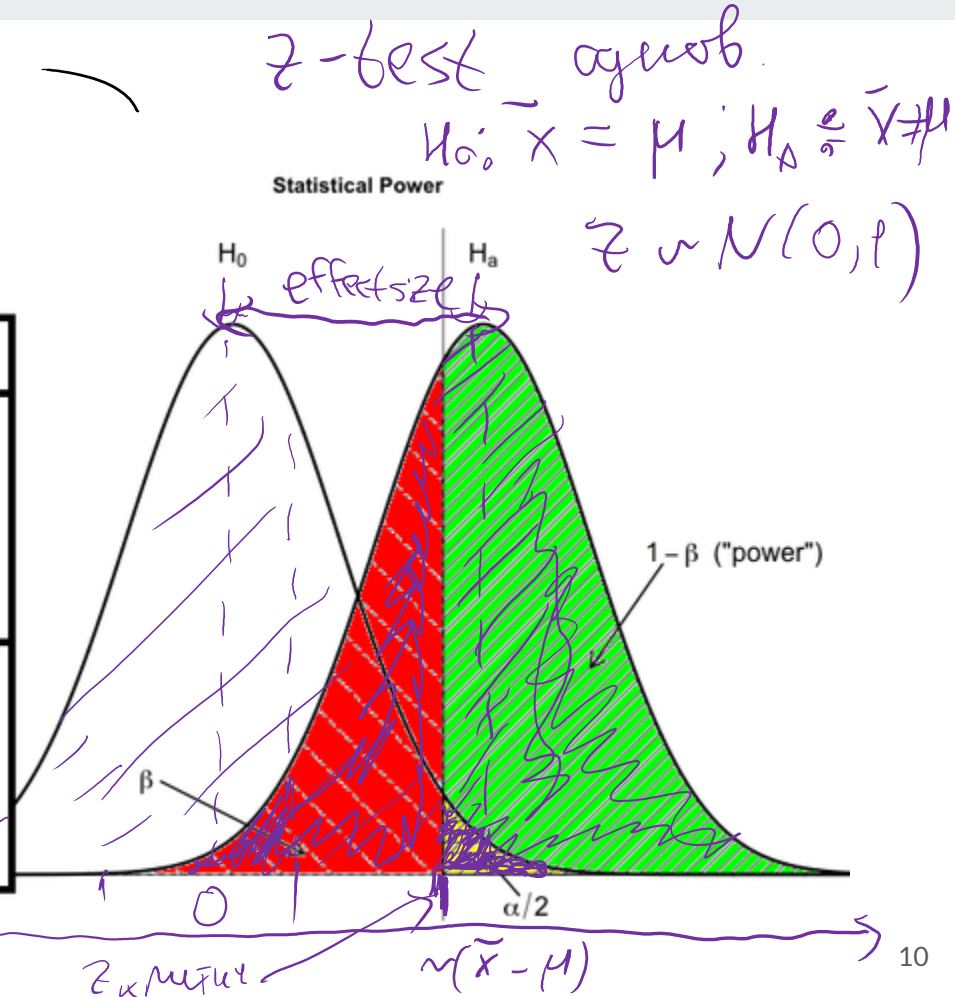
---

# Мощность теста

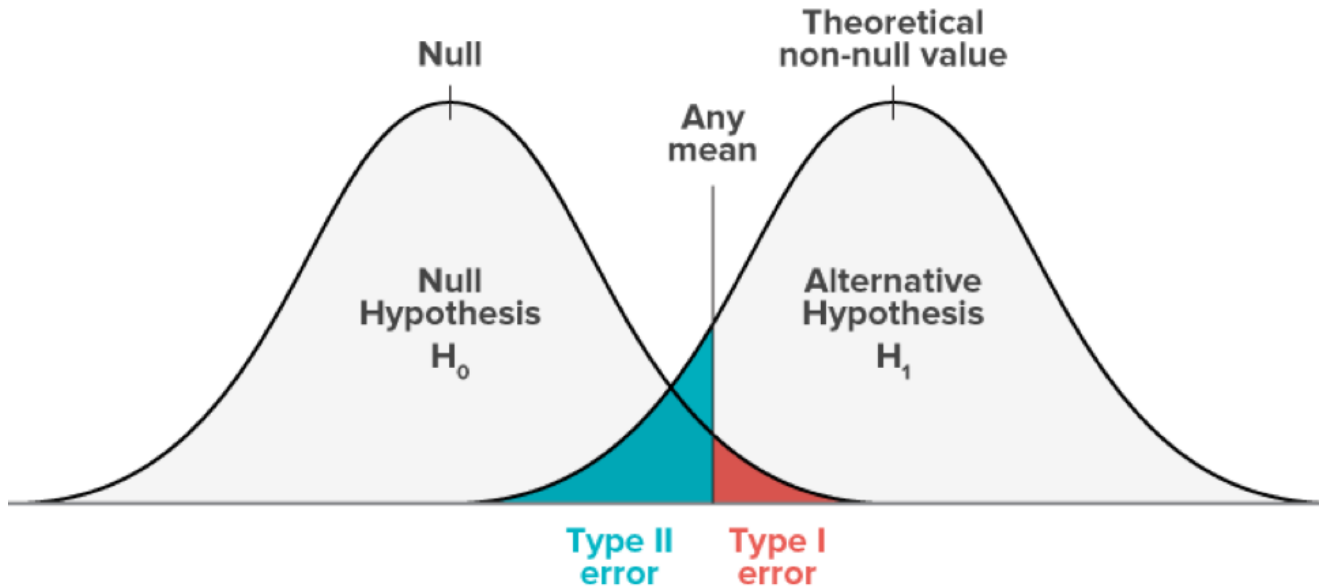


# Мощность теста

		Decision	
		Accept $H_0$	Reject $H_0$
Null Hypothesis ( $H_0$ )	True	<b>Correct</b> "Confidence Level" Probability = $1 - \alpha$	<b>Type I Error</b> "False Positive" Probability = $\alpha$
	False	<b>Type II Error</b> "False Negative" Probability = $\beta$	<b>Correct</b> "Statistical Power" Probability = $1 - \beta$



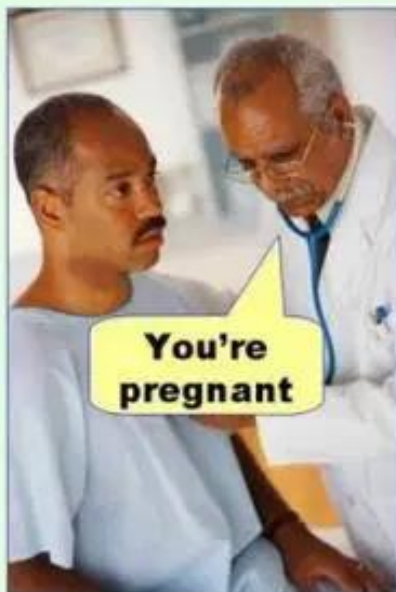
# Мощность теста



При увеличении  
выборки — возрастает

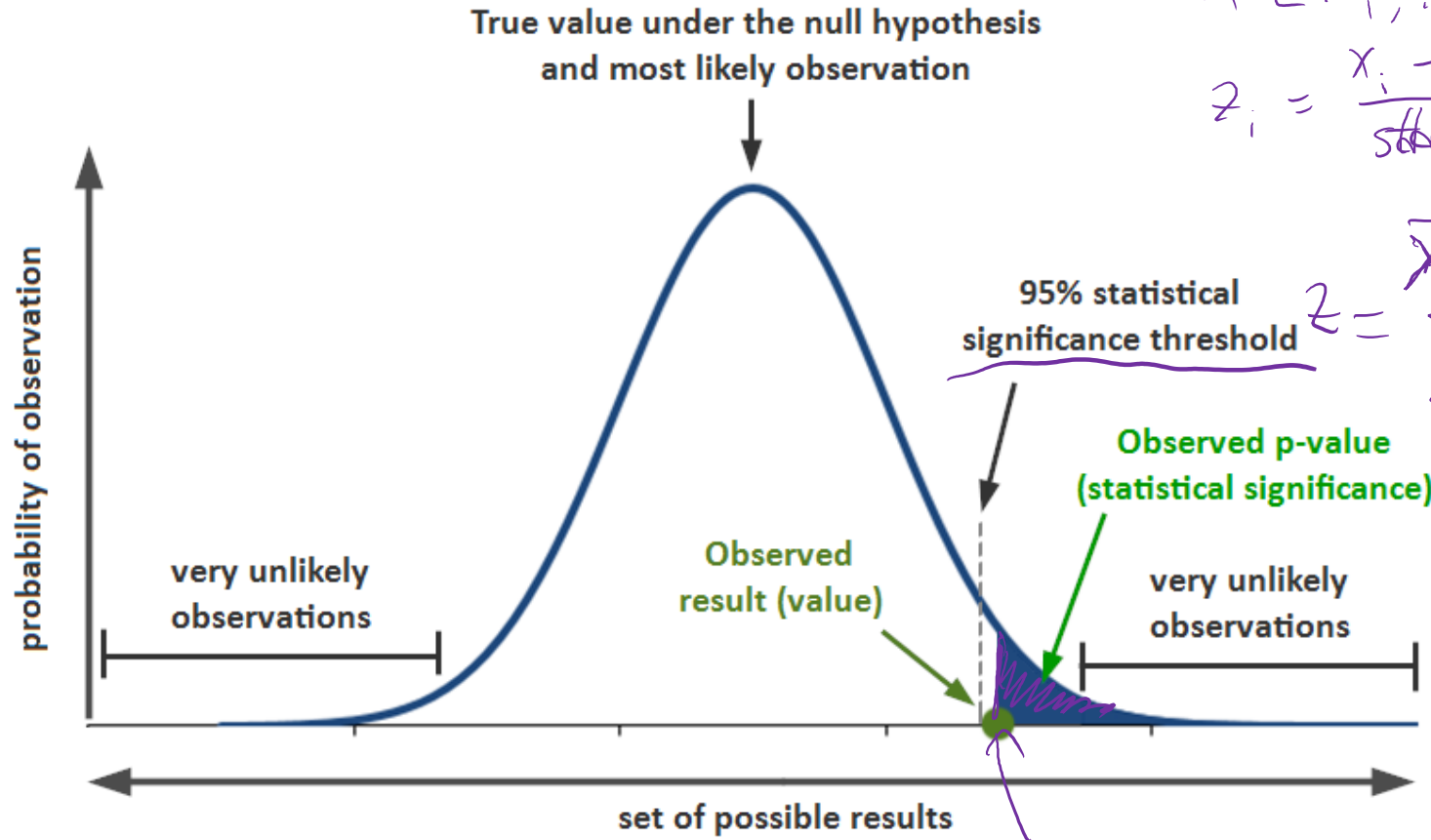
# Ошибки при тестировании гипотез

**Type I error**  
(false positive)



**Type II error**  
(false negative)





$$X [x_1, \dots, x_n]$$

$$z_i = \frac{x_i - \bar{x}}{\text{std}(X)}$$

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

# Анализ мощности

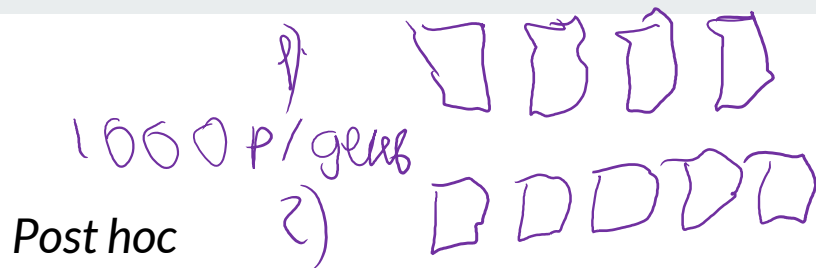
## *A priori*

- какой нужен объем выборки, чтобы найти различия с разумной долей уверенности?
- различия какой величины мы можем найти, если известен объем выборки?

1000 p/genf

Post hoc


2)



- смогли бы мы найти различия при помощи нашего эксперимента ( $\alpha, n$ ), если бы величина эффекта была  $X$ ?

# Анализ *a priori*



тест	t-критерий
уровень значимости	$\alpha=0.05$
желаемая мощность теста	0.8
ожидаемая величина	??? 

# Величина эффекта

Коэффициент Коэна

$$d = \frac{\bar{\mu}_1 - \bar{\mu}_2}{\sigma}$$

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$$

*std<sub>1</sub>* *std<sub>2</sub>*

<i>Effect size</i>	<i>d</i>
Very small	0.01
small	0.20
Medium	0.50
Large	0.80
Very large	1.20
Huge	2.0



# Как оценить ожидаемую величину эффекта



- Пилотные исследования
- Литература
- Общебиологические знания
- Технические требования

---

t-тест



# t-критерий Стьюдента (одновыборочный)



Смоделируем ситуацию:

В статье показано, что в среднем программист пишет 100 строчек кода в день. Мы провели собственное исследование и мы получили среднее = 110 строк,  $s = 5.1$  ( $N = 21$ )

Продуктивнее ли наши программисты?

$H_0$  — Наши программисты работают также как и все программисты в мире

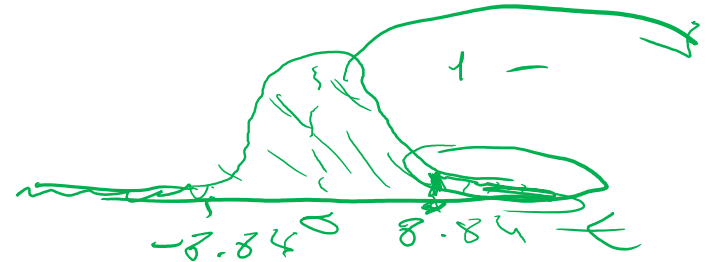
$H_A$  — Наши программисты продуктивнее обычных программистов

# t-критерий Стьюдента (одновыборочный)

Подставляем значения в формулу и получаем t значение = 8.84

Много это или мало?

$$t = \frac{\bar{X} - m}{s_X / \sqrt{n}}$$

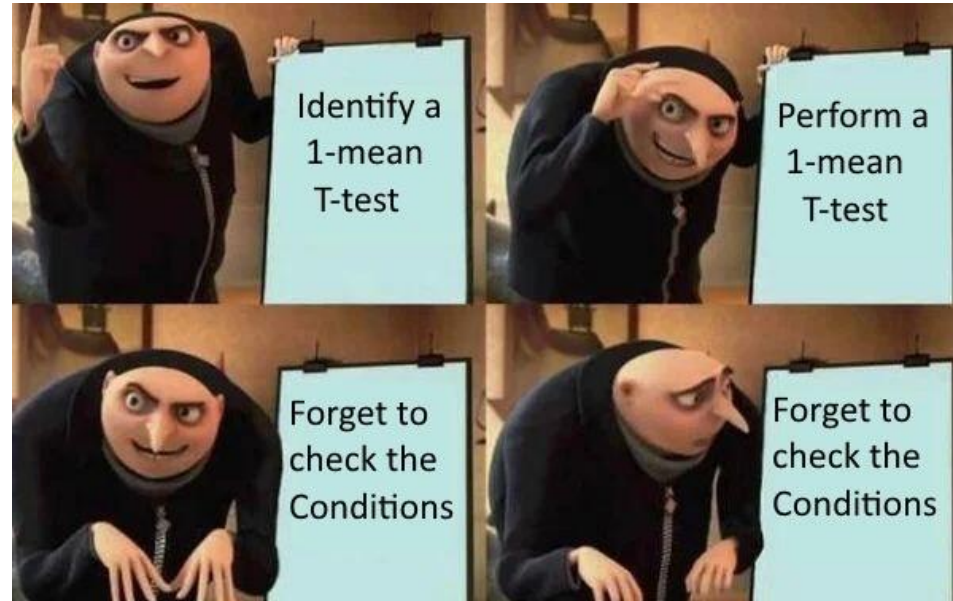


~~2.5%~~  $(1 - \text{pt}(8.84, \text{df} = 21 - 1)) = 0.00000002410724$  - значение p-value

$$\text{df} = n - 1$$

# Критерии данных для t-test

- Наблюдения в выборке должны быть независимы друг от друга.
- Объем выборки достаточно велик **или** величины нормально распределены.



# t-критерий Стьюдента (двухвыборочный)

$H_0: \mu_1 - \mu_2 = 0$  — средние значения не различаются в двух группах

$H_A: \mu_1 - \mu_2 \neq 0$  — средние значения различаются

Нас интересует **разность выборочных средних**, которая будет равна 0 при верной нулевой гипотезе.

$x$  — 1 выборка  $y$  — 2 выборка

~~$\bar{x}$~~

~~$\bar{y}$~~

$$t = \frac{\text{Наблюдаемая величина} - \text{Ожидаемое значение}}{\text{Стандартная ошибка}}$$

$$\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$df = (n_x + n_y - 2)$$

# Критерии данных для двухвыборочного t-test



- Наблюдения независимы друг от друга
- Выборки независимы друг от друга
- Объем выборки достаточно велик или величины нормально распределены

# Разновидности t-test

Двухвыборочный t-тест используется для проверки значимости различий между средними

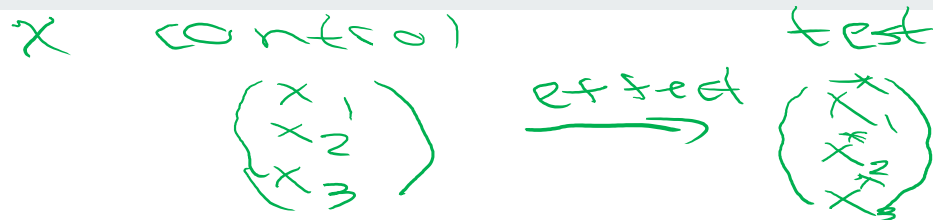
$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$$

стандартная ошибка разности двух средних, может рассчитываться по-разному

- t-тест Стьюдента — если считать, что дисперсии в группах равны
- t-тест Уэлча — если считать, что дисперсии могут быть разными



# Разновидности t-test



Однако, если выборки у вас связанные, то необходимо использовать парный t-test. Это необходимо использовать в случае, если наблюдения в выборках взаимосвязаны:

- прием сначала одного, а затем второго препарата
- сравнение групп до и после воздействия