

# Школа Data analyst

## Занятие 13

# Статистический анализ

## Тема 3



# Disclaimer

Все формулировки далее нестрогие, за более строгими определениями обращайтесь к специализированной литературе

# А/В тестирование





# План занятия

- АБ тестирование
  - Общие слова
  - Статистические критерии

# Что такое А/В тест?

Бизнес и процессы нуждаются в постоянном улучшении/изменении.

**Откуда идеи ?** – рынок, поведение пользователей, потребности, видение...

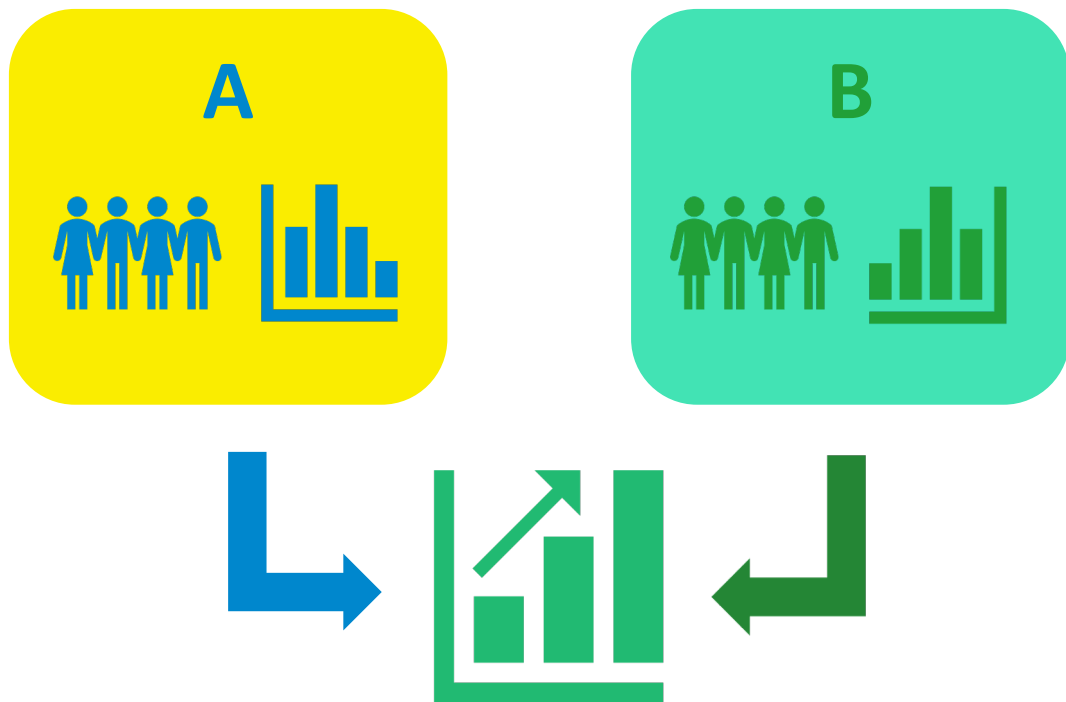
## Как проверить идею?

Как выбрать успешные идеи достоверно и с минимальными затратами?

- Здравый смысл?
- Опросы?
- Фокус-группы?
- Экспертиза?
- Интуиция?
- Оккультные практики?



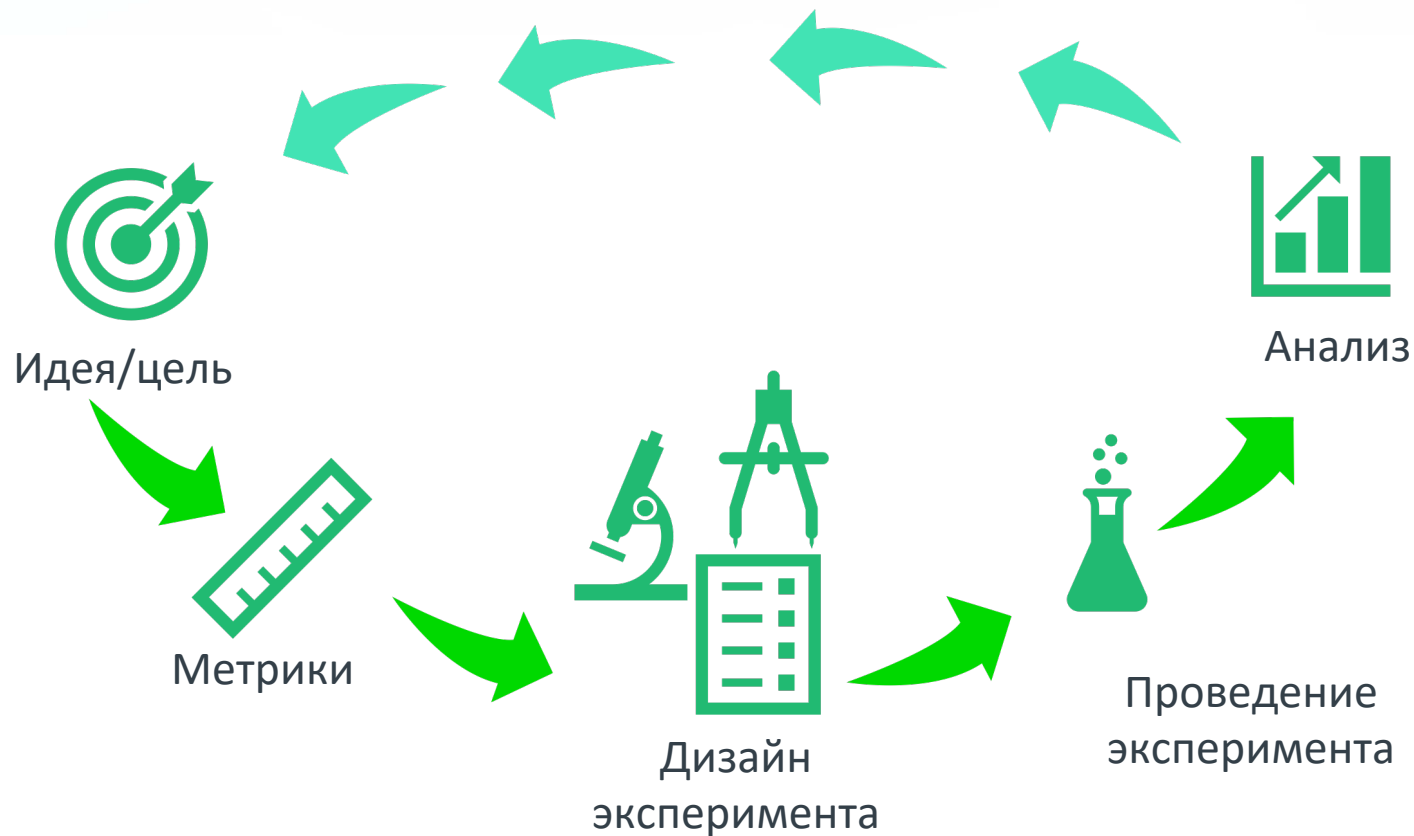
## Что такое A/B тест?



Тестируется:

- Внешний вид (вкус, запах, тембр, пр.)
- Алгоритмы и их влияние
- Новая функциональность
- Пользовательский опыт
- ...
- Ухудшения

# Что такое A/B тест?



## Планирование:

- описание теста
- группировка участников теста
- продолжительность
- пр.

## Проверка гипотез





# Метрики

Прежде чем начать тестирование, необходимо определиться с бизнес показателями по которым мы будем ориентироваться при принятии решений.

Что должно значимо улучшиться?

Например, “количество заказов”, “количество денег”, “количество клиентов” и т.д., и т.п..

Есть ли какие-то проблемы с такими метриками?

Как их корректно рассчитать?



# Метрики

## Proху – метрики:

достаточно чувствительны и хорошо согласуются с бизнес показателями

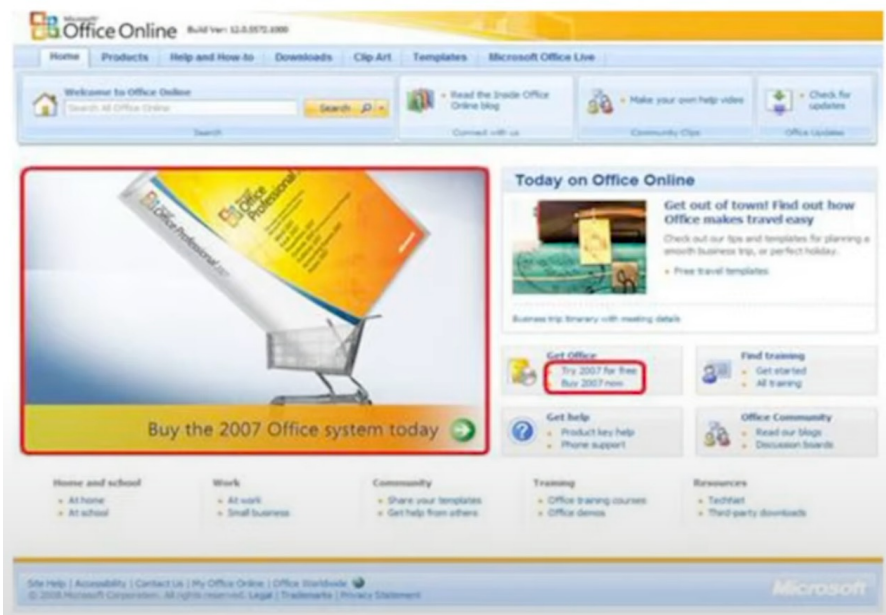
*Пример: число уникальных пользователей или репостов в социальных сетях на сайте яхт-клуба*

## Виды метрик:

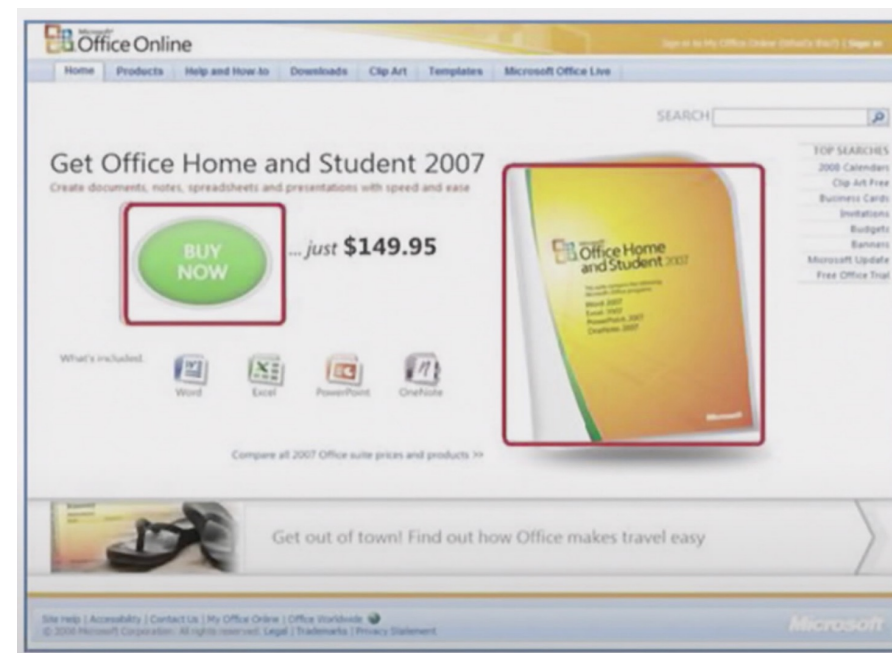
- Предварительные метрики (*до эксперимента*)
- Экспериментальные
- Бизнес-метрики



# Метрики



VS



Объясняет ли фича конверсию?



# Тестирование на исторических данных

Пример: Есть сайт пиццерии, на сайте есть регистрация, пользователи идентифицируются достаточно точно. Имеется также довольно внушительная история заказов пользователей, а также существующий алгоритм выдачи предложений по “сопутствующим товарам”, исходя из того, что заказывает пользователь. Допустим мы думаем над внедрением нового алгоритма. Какие показатели стоит измерить на исторических данных?



# Дизайн эксперимента

- Как правильно выбрать пользователей?
  - Стратификация – снижаем дисперсию
  - Рандомизация – обеспечиваем репрезентативность
- Какие получены артефакты?
- А что если мы хотим провести несколько экспериментов одновременно?
  - Как оценить ошибку теста?



# Дизайн эксперимента

- Как правильно выбрать пользователей?
  - Стратификация
  - Рандомизация
- А что если мы хотим провести несколько экспериментов одновременно?
  - Обычно лучше разбить на непересекающиеся группы
  - Не проводить взаимоисключающие эксперименты
  - В связанных экспериментах продумывать последовательность вариантов
  - Подождать завершения другого эксперимента
  - Хорошо подумать над влиянием фичей друг на друга, оценить связанность метрик и запустить эксперимент на свой страх и риск
  - Скоринг



# Дизайн эксперимента

- Скоринг
  - ROI
  - Видение продукта
  - Запросы обратной связи (исторические данные)
  - Гигиена (фича у всех есть но нам не нужна)
  - Wow-фактор
  - Сложность внедрения и поддержки



# Устойчивость

- Видеть значимые изменения где они есть
- Не видеть значимых изменений где их нет



# Устойчивость

UC Berkeley case

Факультет	Мужчины			Женщины		
	Поступало	Поступило	%	Поступало	Поступило	%
A	825	512	62, 1	108	89	82,4
B	560	353	63	25	17	68
C	325	120	36,9	593	202	34
D	417	138	33,1	375	131	34,9
E	191	53	27,8	393	94	23,9
F	272	16	5,9	341	24	7
Итого	2590	1192	46	1835	557	30,4



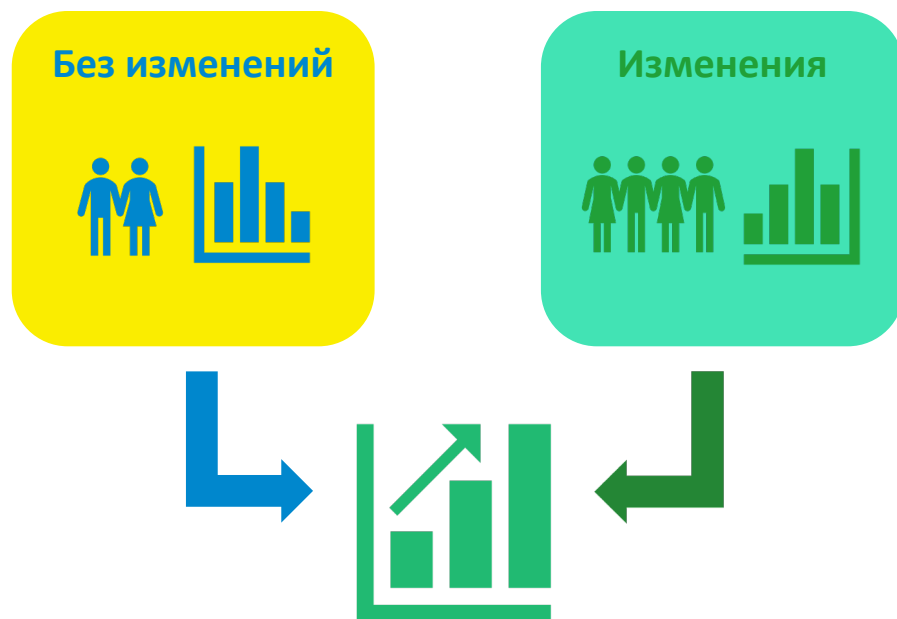
# Устойчивость

		Пятница	Суббота	Всего
А	Пользователи	990 000	500 000	1 490 000
	Конверсии	20 000	5 000	25 000
	%	2.02	1.00	1. 68
В	Пользователи	10 000	500 000	510 000
	Конверсии	230	6 000	6 230
	%	2.30	1.20	1.22



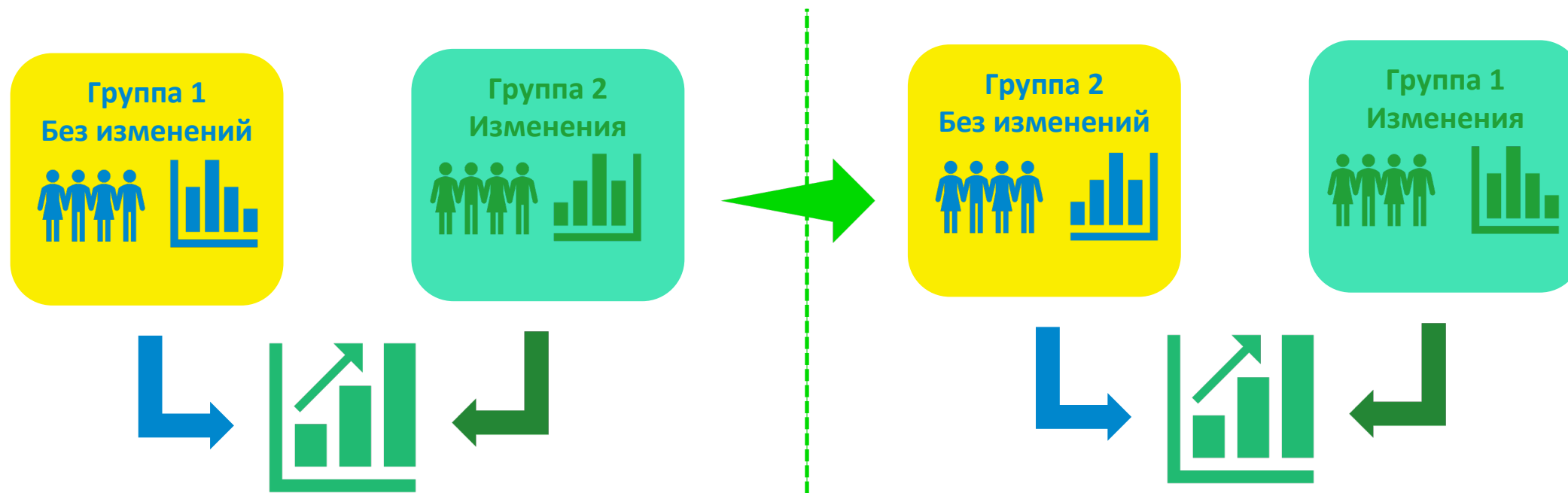
# Устойчивость

Обратный эксперимент: часть пользователей не видит изменений



# Устойчивость

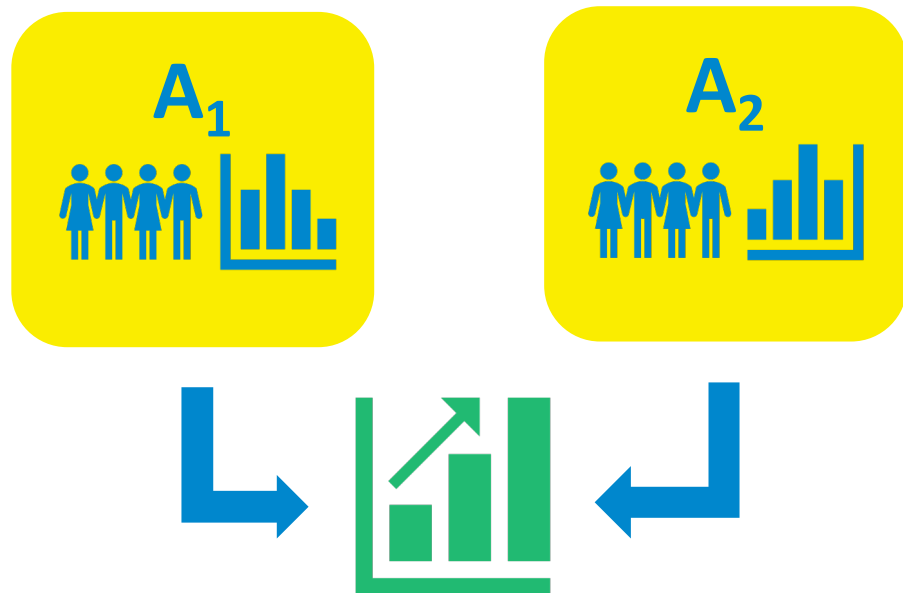
Перекрестный эксперимент: группы пользователей чередуются





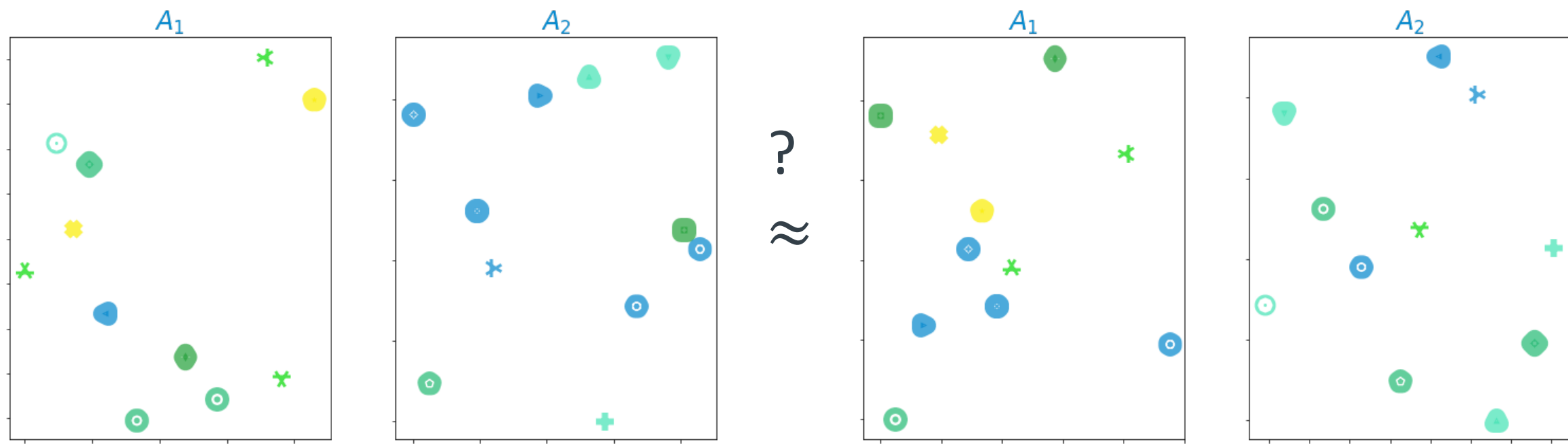
# Устойчивость

АА-тест: АВ-тест, но разным группам демонстрируется одно и то же решение



# Устойчивость

АА-тест: использование ЦПТ (пост-стратификация)



# Устойчивость

ААВ-тест: совмещенный АА и АВ-тесты





# Объём выборки

- Минимальный размер эффекта
- Допустимые вероятности ошибок 1-ого и 2-ого рода
- Статистические инструменты
- Продолжительность теста

Число участников/событий → On-line калькуляторы, например:

<https://raschitat-online.ru/raschet-doveritelnogo-interval/> + формулы

<https://socioline.ru/rv.php>

<https://fdfgroup.ru/poleznaya-informatsiya/stati/vyborka-tipy-vyborok-raschet-oshibki-vyborki/>



# Ошибки 1-ого и 2-ого рода

		Верная гипотеза	
		$H_0$	$H_1$
Ответ теста	$H_0$	$H_0$ принята	$H_0$ неверно принята (ошибка 2-ого рода, $\beta$ ): вероятность получить имеющиеся данные довольно высока при истинности $H_0$
	$H_1$	$H_0$ неверно отвергнута (ошибка 1-ого рода, $\alpha$ ): вероятность получить имеющиеся данные при истинности $H_0$ слишком мала	$H_0$ отвергнута



# Ошибки 1-ого и 2-ого рода

Ошибка 1-ого рода критичнее

Вероятность ошибки первого рода  $P(H_0 \text{ отв.} | H_0 \text{ вер.})$  жестко ограничивается

$$P(H_0 \text{ отвергнута} | H_0 \text{ верна}) = P(p \leq \alpha | H_0) \leq \alpha$$

Вероятность ошибки второго рода  $P(H_0 \text{ принимаем} | H_1 \text{ вер.})$  мягко минимизируется

$$row = P(H_0 \text{ отвергнута} | H_1 \text{ верна}) = 1 - P(H_0 \text{ принимаем} | H_1 \text{ верна})$$



# p-value

Достигаемый уровень значимости — это вероятность при справедливости нулевой гипотезы получить такое же значение статистики, как в эксперименте, или ещё более экстремальное

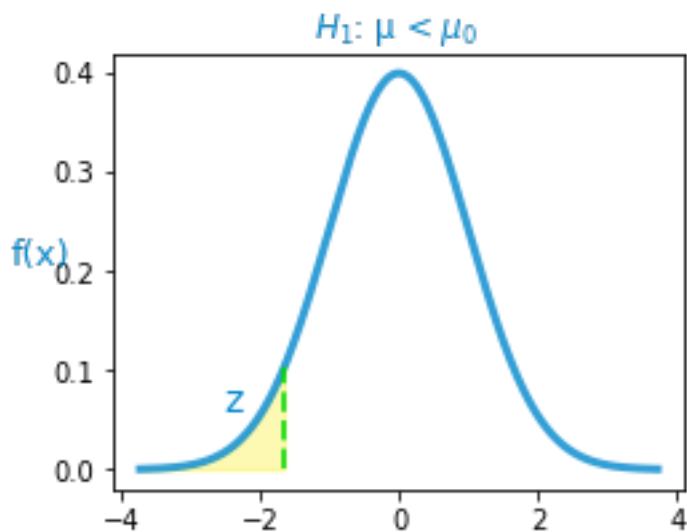
Чем ниже **p** тем сильнее данные свидетельствуют против **H<sub>0</sub>** в пользу **H<sub>1</sub>**

$$T(X) = t$$
$$p = \mathbb{P}(T \geq t | H_0)$$

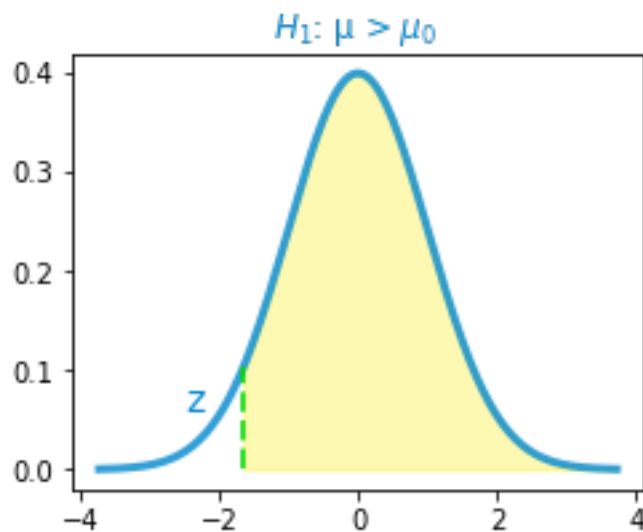


# p-value

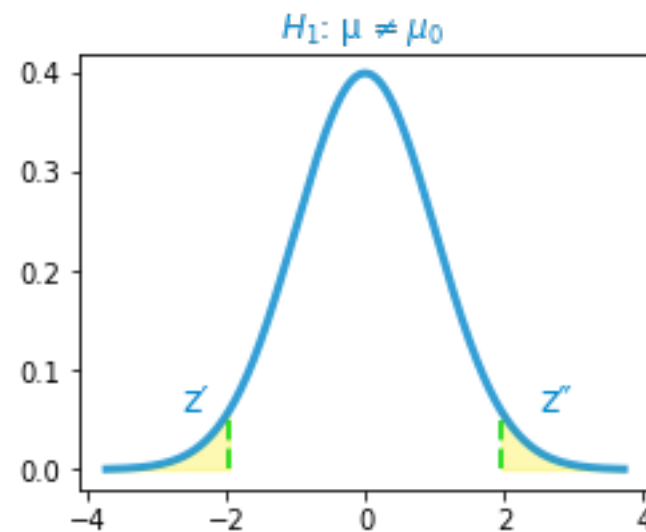
$$p = FN_{(0,1)}(z)$$



$$p = 1 - F_{N(0,1)}(z)$$



$$p = 2(1 - F_{N(0,1)}(|z|))$$





# Размер эффекта

*p-value?*



*Какова вероятность верного предсказания?  
Насколько различаются суммы в чеках?  
и т.д.*

**Важно:**

Выборка данных формируется случайно

Оценка размера эффекта по выборке это случайная величина

**p** показывает вероятность случайного получения такой оценки



***p-value* зависит от размера эффекта и размера выборки**

*На малых выборках эффект менее заметен,  $H_0$  не отвергается*



## Вопросы по $\alpha$ (или p-value)

Определяет ли  $\alpha$  вероятность справедливости нулевой гипотезы  $H_0$ ?

p-value – это такое значение статистики при справедливой  $H_0$  с вероятностью  $\alpha$ ?

Отсутствуют ли различия между группами при  $p\text{-value} > 0.05$ ?

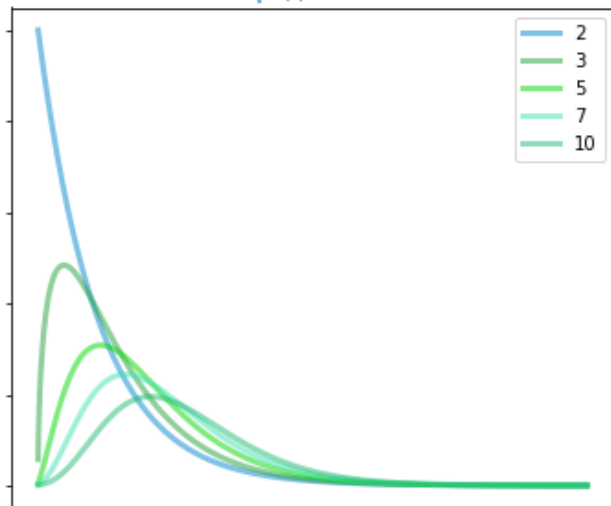
**Имеет ли результат практическое значение?**

# Распределения, производные от нормального

$$X_1, X_2 \dots X_k \sim N(\mu, \sigma^2)$$

$$X = \sum_{i=1}^k X_i^2 \sim \chi^2_k$$

Распределение  $\chi^2$

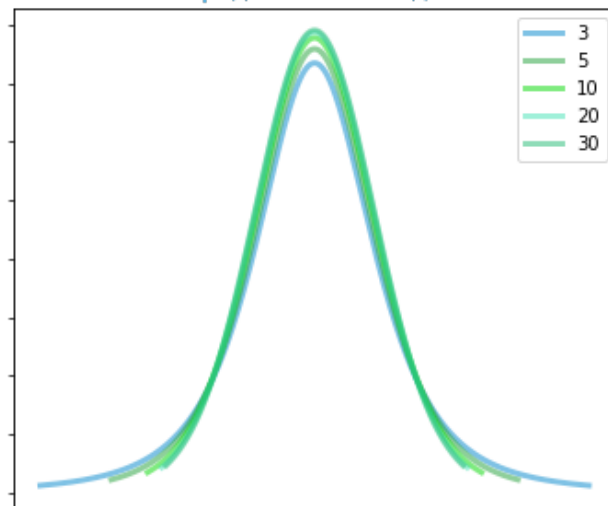


$$X_1 \sim N(0, 1)$$

$$X_2 \sim \chi^2_k$$

$$X = \frac{X_1}{\sqrt{X_2/k}} \sim St(k)$$

Распределение Стьюдента

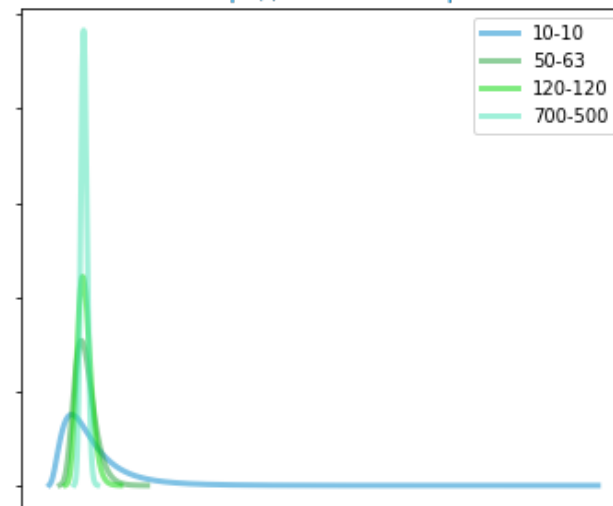


$$X_1 \sim \chi^2_{k_1}$$

$$X_2 \sim \chi^2_{k_2}$$

$$X = \frac{X_1/k_1}{X_2/k_2} \sim F(k_1, k_2)$$

Распределение Фишера



# Распределения, производные от нормального

$$X_1, X_2 \dots X_n \sim N(\mu, \sigma^2) \rightarrow X^n = (X_1, X_2 \dots X_n)$$

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

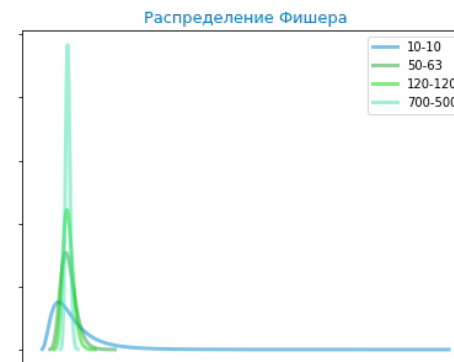
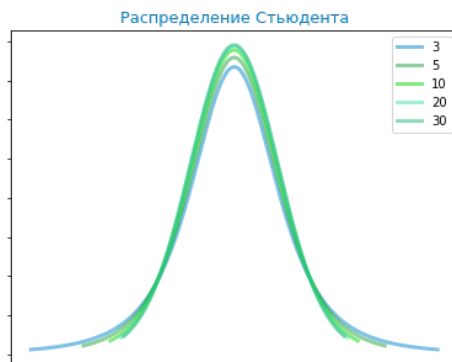
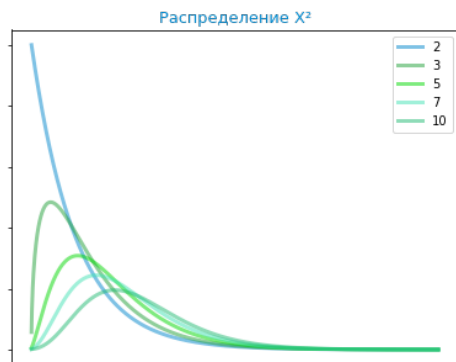
$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \Rightarrow \chi_{n-1}^2$$

$$T = \frac{\bar{X}_n - \mu}{S_n / \sqrt{n}} \sim St(n-1)$$

$$X_1 \sim N(\mu_1, \sigma_1^2) \rightarrow \chi_{n_1}^2$$

$$X_2 \sim N(\mu_2, \sigma_2^2) \rightarrow \chi_{n_2}^2$$

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$







# Статистические критерии

Для среднего:

- Данные распределены нормально? Известна дисперсия  $\rightarrow$  z-критерий
- Дисперсия неизвестна  $\rightarrow$  t-критерий (при большом количестве данных можно использовать квантили нормального распределения)

Для частоты:

- Критерий согласия Пирсона (Хи-квадрат)

Для дисперсии:

- Среднее известно  $\rightarrow$  критерий Хи-квадрат
- Неизвестно  $\rightarrow$  критерий [Фишера](#) (z-критерий)

# Статистические критерии

Одновыборочный критерий Стьюдента

Сравнение непрерывных величин

Среднее/Медиана/Мода,  $\sigma^2$  — известна

**Выборка:**

$$X^n = (X_1, \dots, X_n), \\ X \sim \mathcal{N}(\mu, \sigma^2)$$

**Нулевая гипотеза:**

$$H_0: \mu = \mu_0$$

**Альтернатива:**

$$H_1: \mu < \neq > \mu_0$$

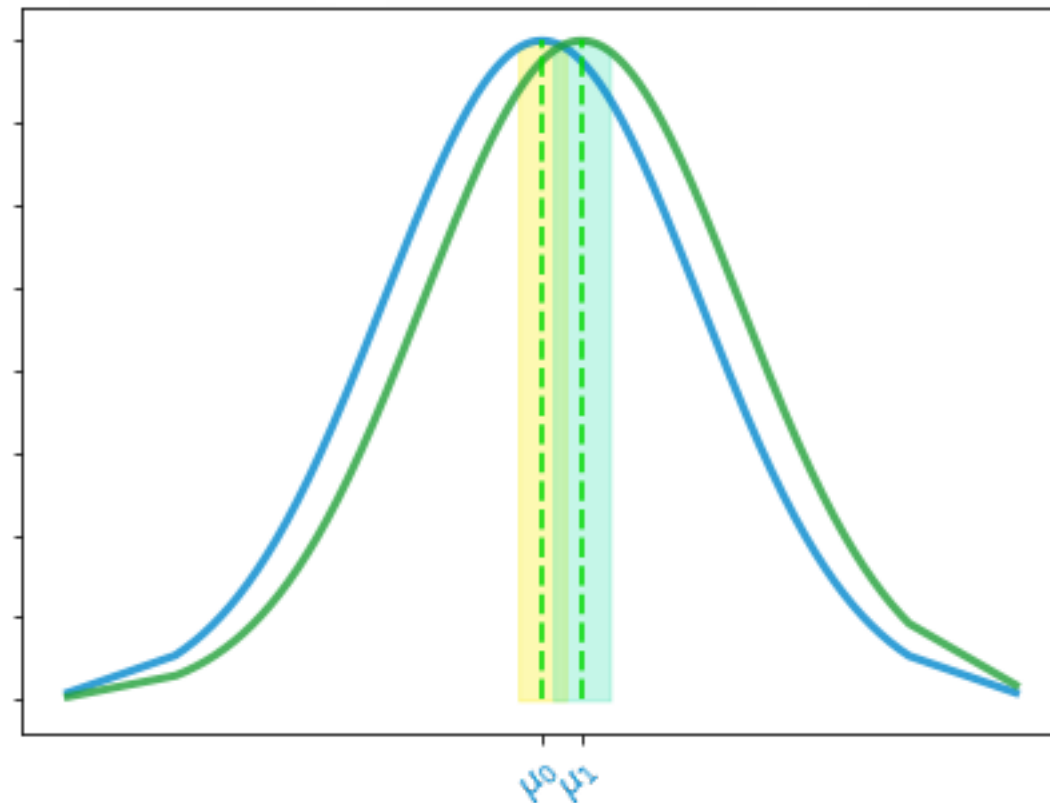
**Статистика:**

$$Z(X^n) = \frac{X_i - \mu_0}{\sigma/\sqrt{n}}$$

**Нулевое**

**распределение:**

$$Z(X^n) = N(0,1)$$



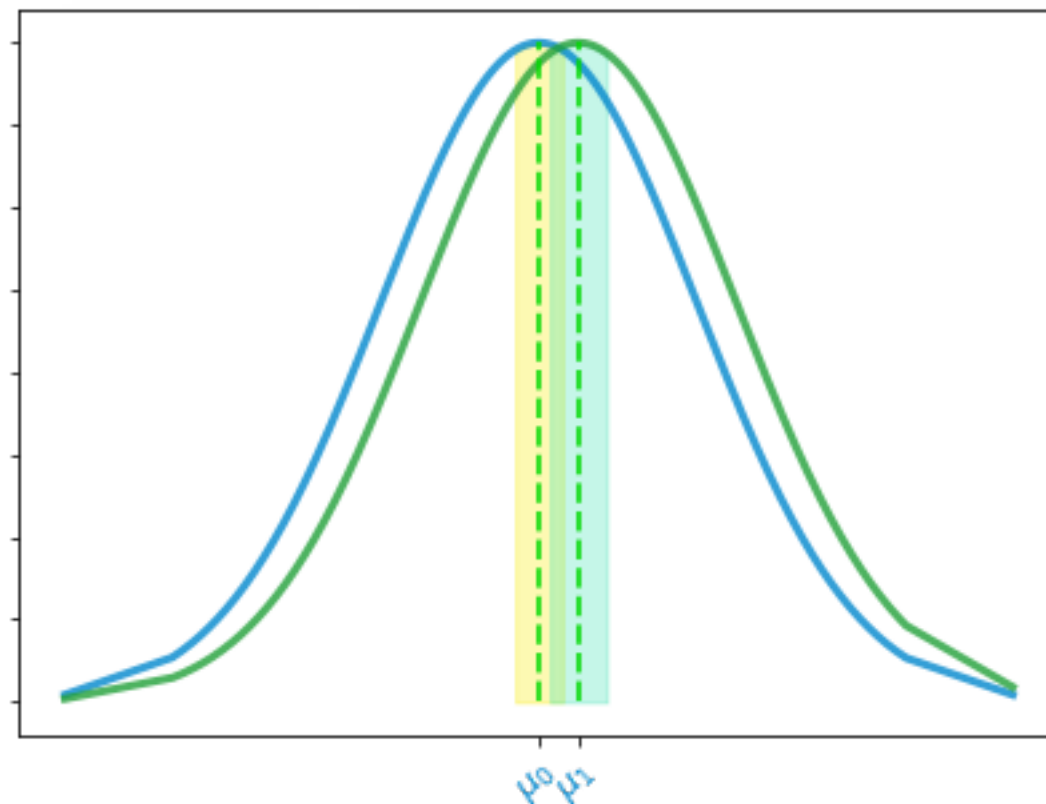
# Статистические критерии

Одновыборочный критерий Стьюдента

Сравнение непрерывных величин

Среднее/Медиана/Мода,  $\sigma^2$  — неизвестна

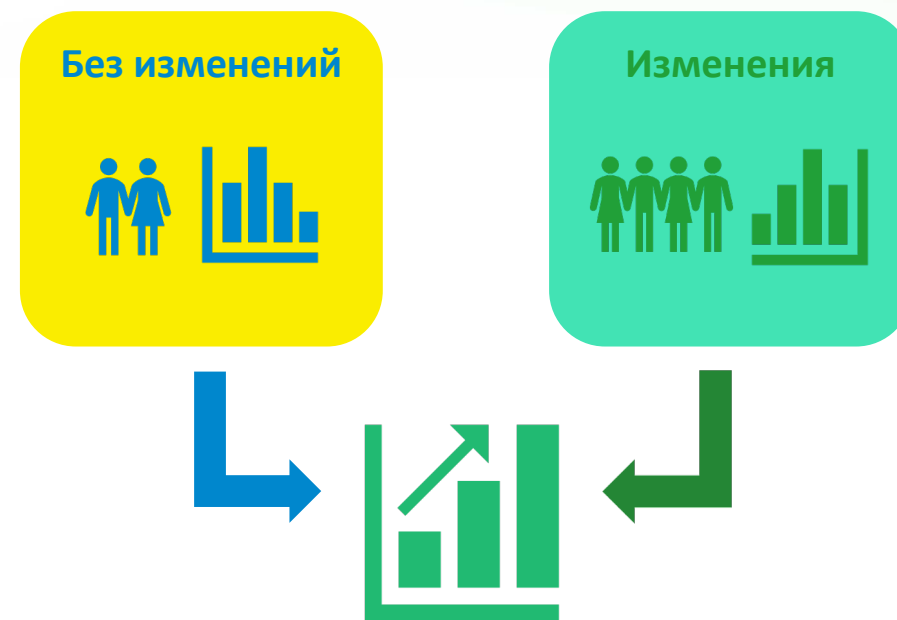
Выборка:	$X^n = (X_1, \dots, X_n)$ , $X \sim \mathcal{N}(\mu, \sigma^2)$
Нулевая гипотеза:	$H_0: \mu = \mu_0$
Альтернатива:	$H_1: \mu < \neq > \mu_0$
Статистика:	$T(X^n) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$
Нулевое распределение:	$T(X^n) \sim St(n - 1)$



# Статистические критерии

Двухвыборочный критерий Стьюдента  
Независимые выборки, Z - критерий

Выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$ $X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$ $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
Нулевая гипотеза:	$H_0: \mu_1 = \mu_2$
Альтернатива:	$H_1: \mu_1 < \neq > \mu_2$
Статистика:	$Z(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Нулевое распределение:	$Z(X_1^{n_1}, X_2^{n_2}) = N(0,1)$

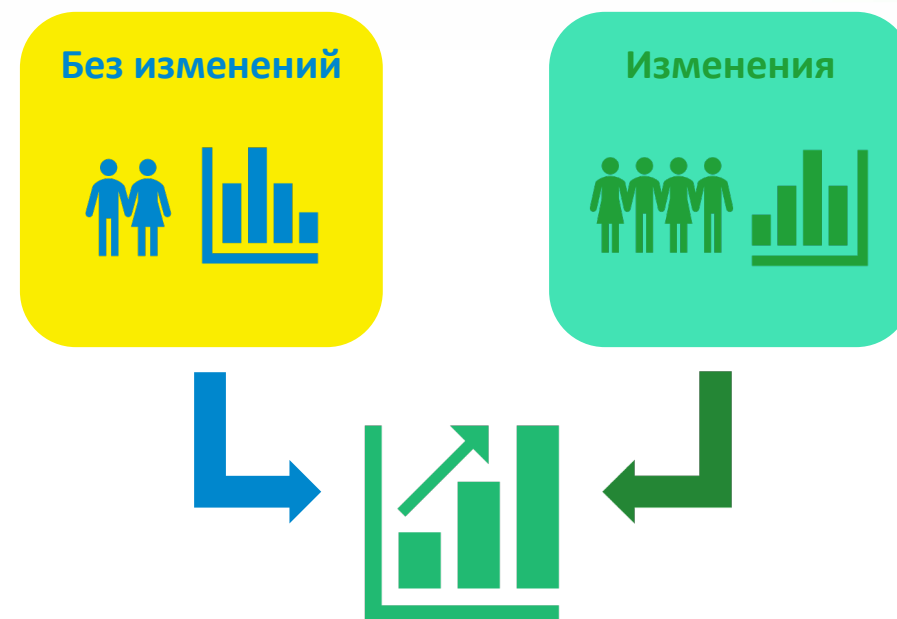


$$\sigma_1 < \sigma_2$$
$$n_1 < n_2$$

# Статистические критерии

Двухвыборочный критерий Стьюдента  
Независимые выборки, Т - критерий

Выборки:	$X_1^{n_1} = (X_{11}, \dots, X_{1n_1}),$ $X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$ $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
Нулевая гипотеза:	$H_0: \mu_1 = \mu_2$
Альтернатива:	$H_1: \mu_1 < \neq > \mu_2$
Статистика:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Нулевое распределение:	$T(X_1^{n_1}, X_2^{n_2}) \approx \sim St(v)$



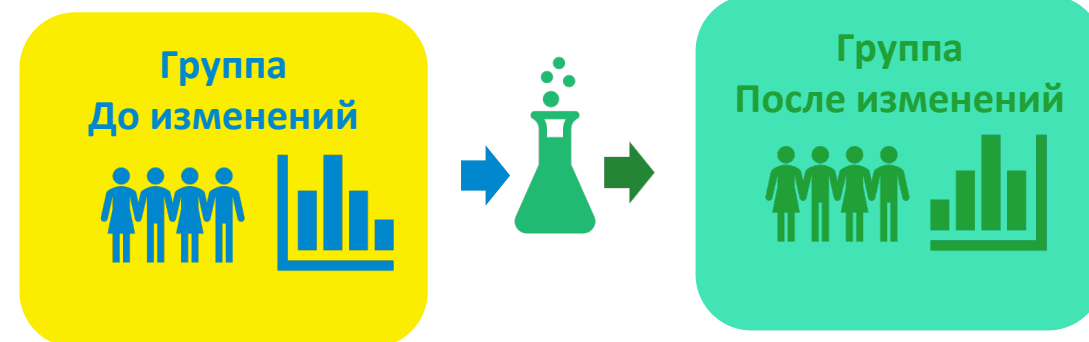
$$\sigma_1 < \sigma_2$$

$$n_1 < n_2$$

# Статистические критерии

Двухвыборочный критерий Стьюдента  
Связанные выборки, Т - критерий

Выборки:	$X_1^n = (X_{11}, \dots, X_{1n}),$ $X_2^n = (X_{21}, \dots, X_{2n})$ $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
Нулевая гипотеза:	$H_0: \mu_1 = \mu_2$
Альтернатива:	$H_1: \mu_1 < \neq > \mu_2$
Статистика:	$T(X_1^n, X_2^n) = \frac{\bar{X}_1 - \bar{X}_2}{S/\sqrt{n}}$
Нулевое распределение:	$T(X_1^n, X_2^n) \sim St(n-1)$



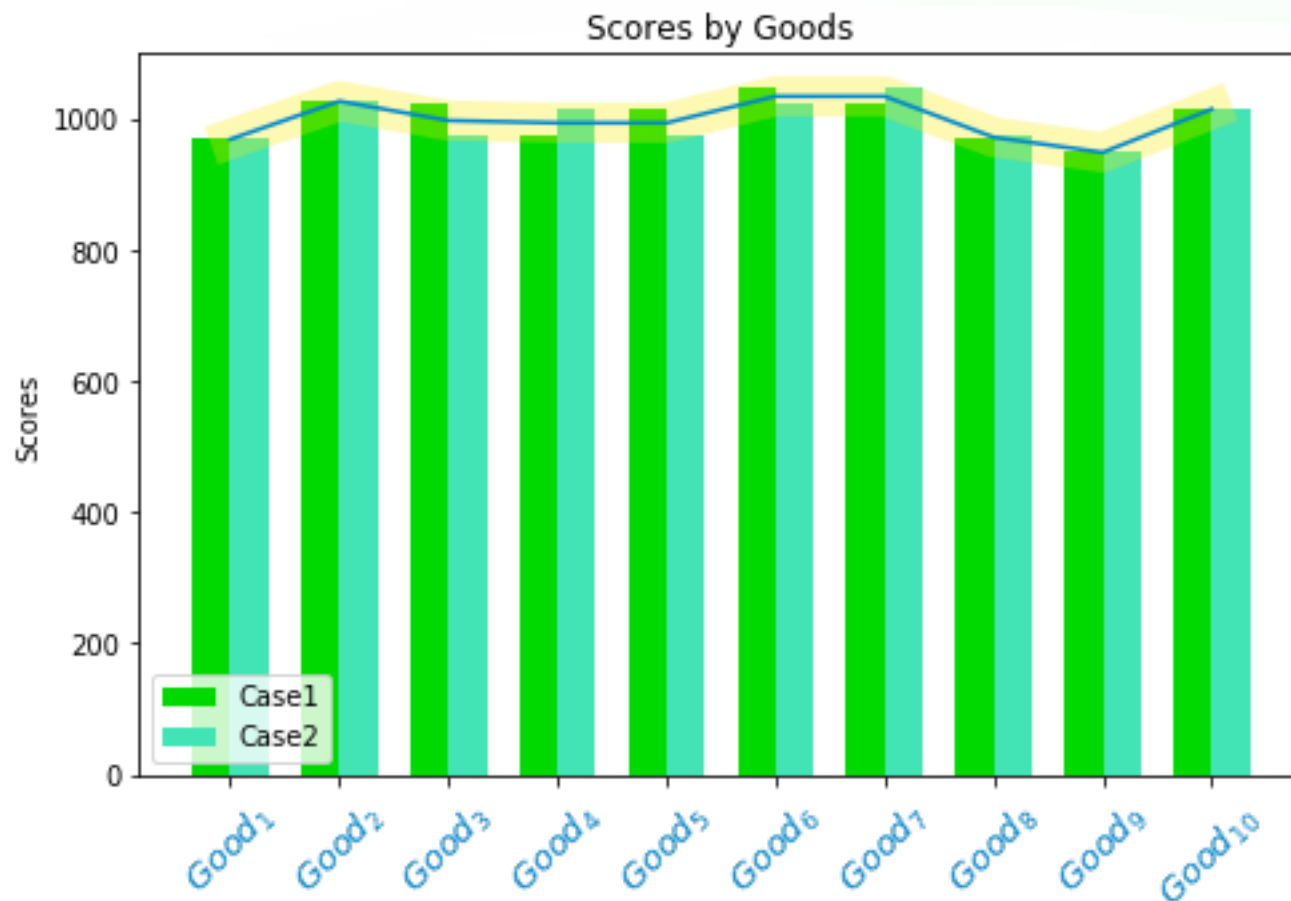
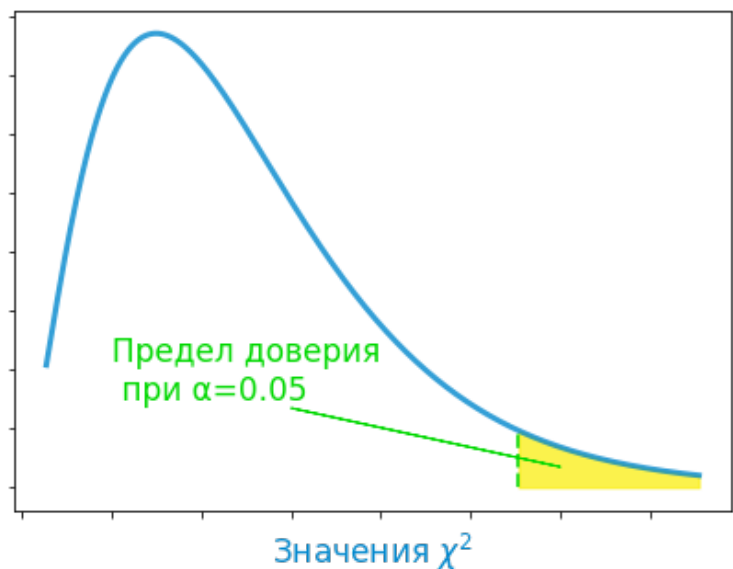
$$^* S = \sqrt{\frac{\sum_{i=0}^n (D_i - \bar{D})^2}{n-1}}, D_i = X_{1i} - X_{2i}$$

# Статистические критерии

Сравнение дискретных величин

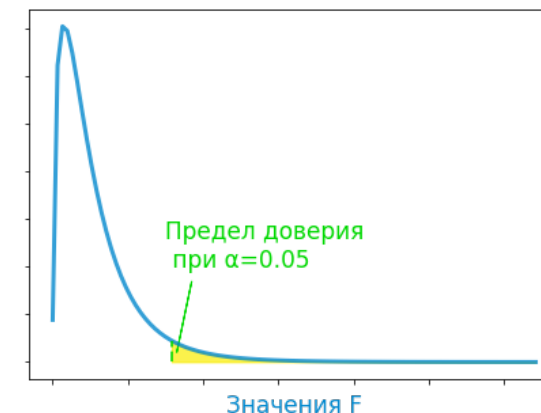
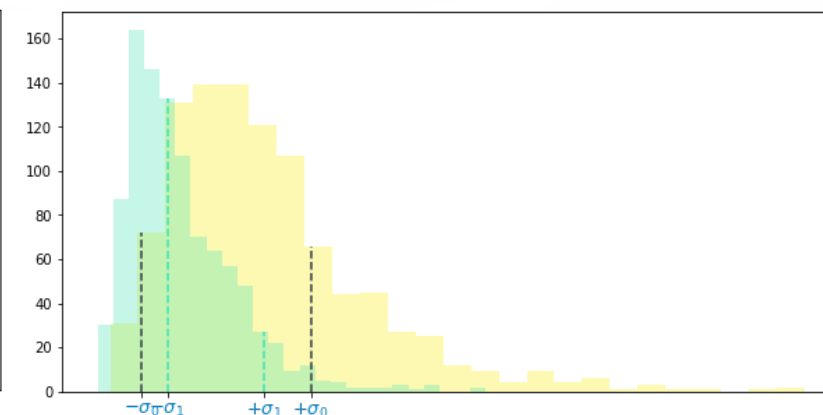
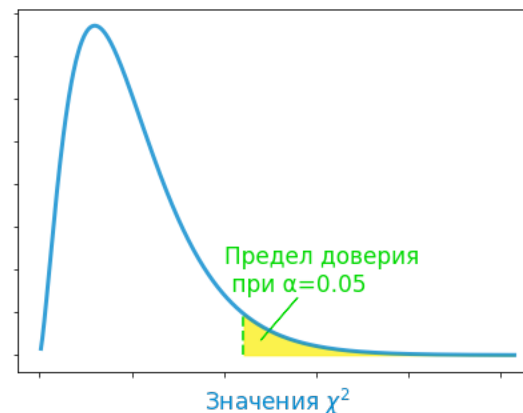
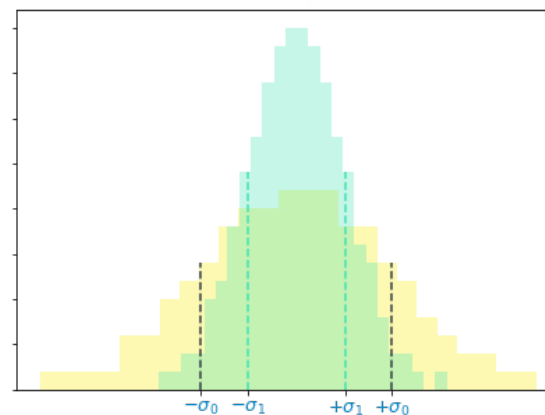
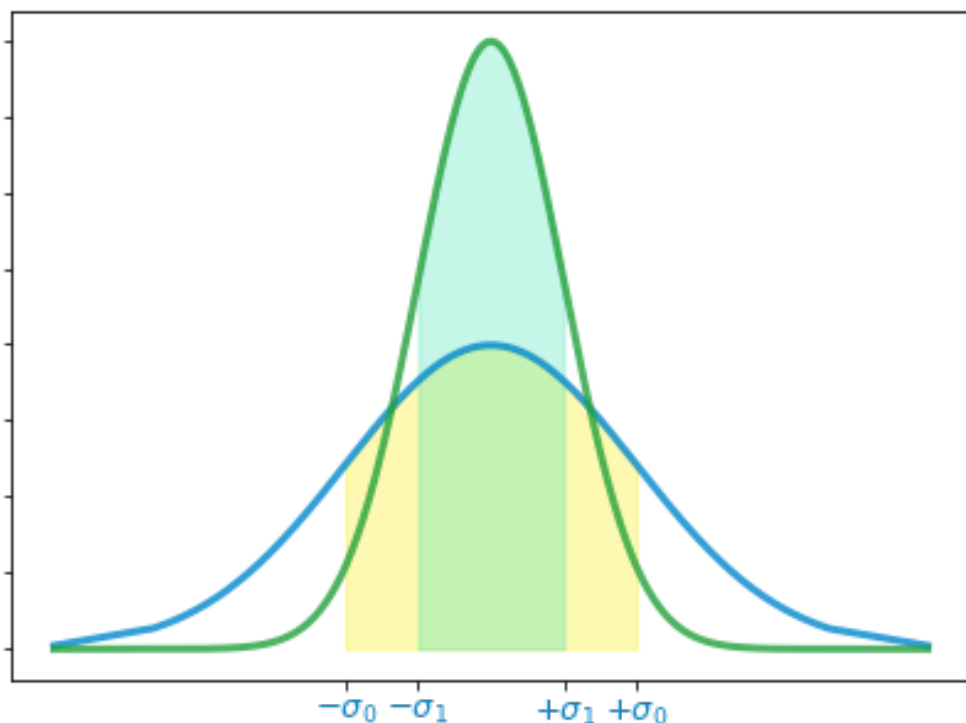
Критерий согласия Пирсона  $\chi^2$

Работает при большом количестве значений в каждой категории



# Статистические критерии

Сравнение непрерывных величин  
Дисперсия/Разброс







# Как не обмануть себя

- Не верить
  - Ждать
  - Проводить обратный эксперимент
  - Не делать АБ тест без предварительного АА теста
  - Использовать статистику!
    - Определиться с размером ожидаемого эффекта
    - Допустимые вероятности ошибок 1-ого и 2-ого рода
    - Грамотно выбирать статистические критерии
- Не забывать правило Паретто

# Механизм проверки гипотез



# Гипотеза - данные - вывод

Прежде чем что-то сравнивать выдвигают конструктивную гипотезу, которую хочется проверить.

Например,

- препараты X и Y по-разному влияют на кровяное давление больных
- продолжительность лекции влияет на успеваемость студента
- постановка вопроса влияет на ответ респондента

На полученных данных мы пытаемся делать выводы об истинности или ложности выдвигаемой гипотезы



## Гипотеза - данные - вывод

Нужно помнить, что проверка статистической гипотезы имеет вероятностный характер.

Точно также, мы не можем быть уверены на все 100, что параметр, оцениваемый по конечной выборке совпадает с реальным значением параметра в генеральной совокупности (вспоминаем доверительные интервалы).



# Гипотеза

Статистическая гипотеза — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки.



# Гипотеза

Пусть в эксперименте наблюдается случайная величина  $X$ , распределение которой  $P$  полностью или частично неизвестно. Любое утверждение относительно  $P$  называют статистической гипотезой. Гипотезы бывают простые и сложные.

- Если гипотеза однозначно определяет  $P$ , т.е  $H: \{P = P_0\}$ , где  $P_0$  это какой-то конкретный закон (например, нормальное распределение с параметрами 0 и 1), то гипотеза **простая**.
- Если же гипотеза утверждает, что  $P$  относится к семейству распределений, то гипотеза **сложная** (например, гипотеза о том, что данные распределены нормально, без фиксации параметров).



# Общий фреймворк проверки гипотез: шаг 1

Формулировка основной гипотезы  $H_0$  и конкурирующей гипотезы  $H_1$ .



## Общий фреймворк проверки гипотез: шаг 2

Задание уровня значимости  $\alpha$  (или p-value), на котором в дальнейшем и будет сделан вывод о справедливости гипотезы. Он равен вероятности допустить *ошибку первого рода  $H_0$* .



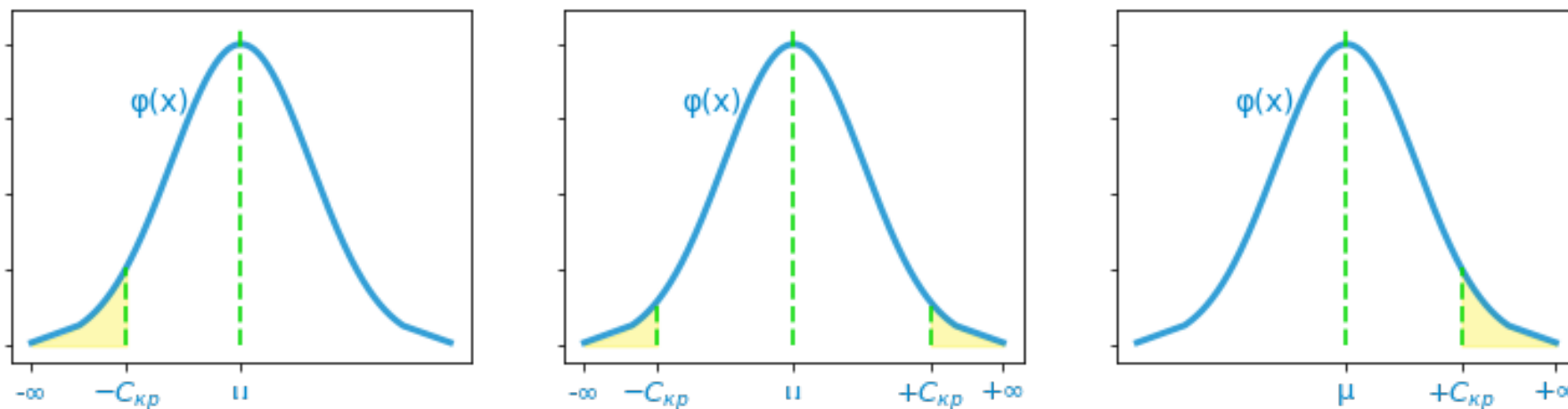
# Общий фреймворк проверки гипотез: шаг 3

Расчёт статистики  $\varphi$  критерия такой, что:

- её величина зависит от исходной выборки  $X^n = (X_1, \dots, X_n)$   $\varphi = \varphi(X_1, \dots, X_n)$
- по её значению можно делать выводы об истинности гипотезы  $H_0$
- статистика  $\varphi$ , как функция случайной величины  $X$ , также является случайной величиной и подчиняется какому-то закону распределения

## Общий фреймворк проверки гипотез: шаг 4

Построение критической области. Из области значений  $\varphi$  выделяется подмножество  $C$  таких значений, по которым можно судить о существенных расхождениях с предположением. Его размер выбирается таким образом, чтобы выполнялось равенство  $P(\varphi \in C) = \alpha$ . Это множество  $C$  и называется критической областью.





## Общий фреймворк проверки гипотез: шаг 5

Вывод об истинности гипотезы. Наблюдаемые значения выборки подставляются в статистику  $\varphi$  и по попаданию (или не попаданию) в критическую область  $C$  выносится решение об отвержении (или принятии) выдвинутой гипотезы  $H_0$ .

## Итого

Выборка:	$X^n = (X_1, \dots, X_n), \quad X \sim P$
Нулевая гипотеза:	$H_0: P \in \omega$
Альтернатива:	$H_1: P \notin \omega$
Статистика:	$T(X_n), \quad \begin{array}{l} T(X_n) \sim F(x) \text{ при } H_0 \\ T(X_n) \not\sim F(x) \text{ при } H_1 \end{array}$



# Формализация конструктивной гипотезы

Пример с препаратами.

$H_0$ : Реальная разность между средними значениями давлений в двух группах равна 0 ( $\mu_0 - \mu_1 = 0$ )

$H_1$ : Реальная разность между средними значениями давлений в двух группах не равна 0 ( $\mu_0 - \mu_1 \neq 0$ )



## Рассмотрим следующую задачу

В десятизначной записи числа  $\pi$  среди 10000 первых десятичных знаков после запятой цифры 0, 1, ..., 9 встречаются соответственно  $\mathbf{h}=(968, 1026, 1021, 972, 1014, 1046, 1021, 970, 948, 1014)$  раз. Можно ли при уровне значимости 0.05 (величина ошибки 1-го рода) считать эти цифры случайными?



# Colab? Colab!

# p-value

[Про p-value habr](#): Если я живу в мире, где время доставки пиццы составляет 30 минут или меньше (нулевая гипотеза верна), насколько неожиданными являются мои доказательства в реальной жизни?

*p-value* отвечает на этот вопрос числом — вероятностью.

$$T(X) = t$$
$$p = \mathbb{P}(T \geq t | H_0)$$







# Colab? Colab!



# Размер эффекта

Размер эффекта — степень отклонения данных от нулевой гипотезы.

Примеры:

- Вероятность верного предсказания
- Вероятность выздоровления пациента
- Увеличение среднего чека

# Ошибки 1/2-ого рода и размер эффекта

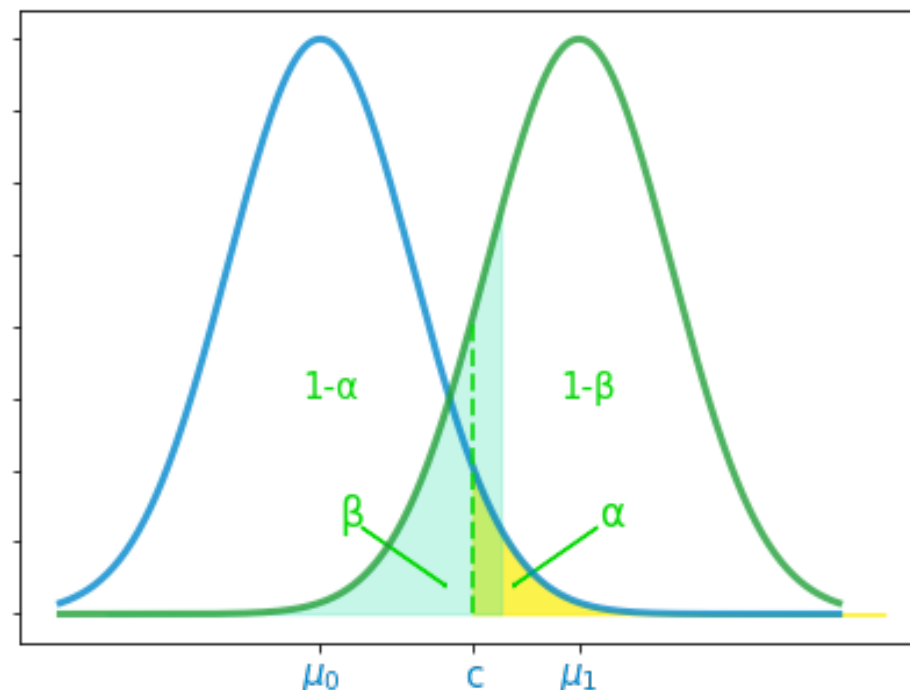
$\mu_1 - \mu_0$  размер эффекта

$\alpha$  ошибка первого рода (или уровень значимости)

$\beta$  ошибка второго рода

$1 - \beta$  мощность критерия

$c$  порог принятия решения



# Ошибки 1/2-ого рода и размер эффекта

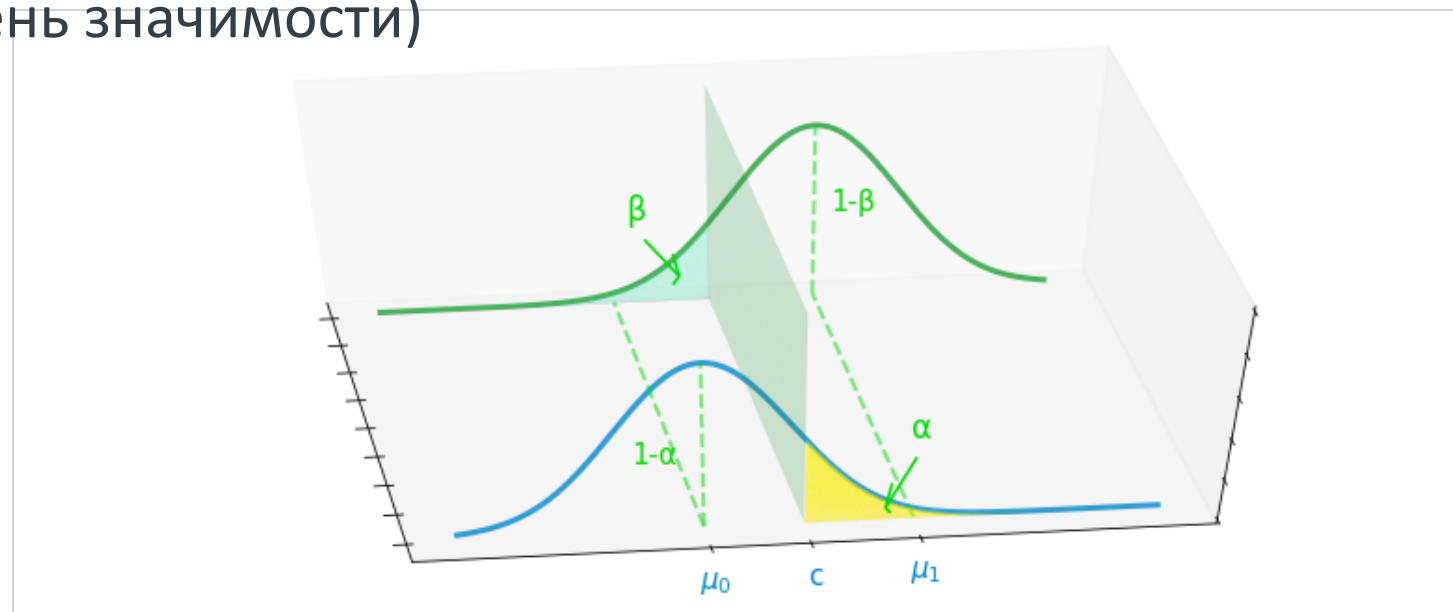
$\mu_1 - \mu_0$  размер эффекта

$\alpha$  ошибка первого рода (или уровень значимости)

$\beta$  ошибка второго рода

$1 - \beta$  мощность критерия

$c$  порог принятия решения





# Colab? Colab!



# Резюме

Обсудили что такое АБ тестирование, и общие принципы при его проведении

Обсудили работу статистических критериев — механизма при принятии решения



Обратная связь

?



Спасибо за внимание!