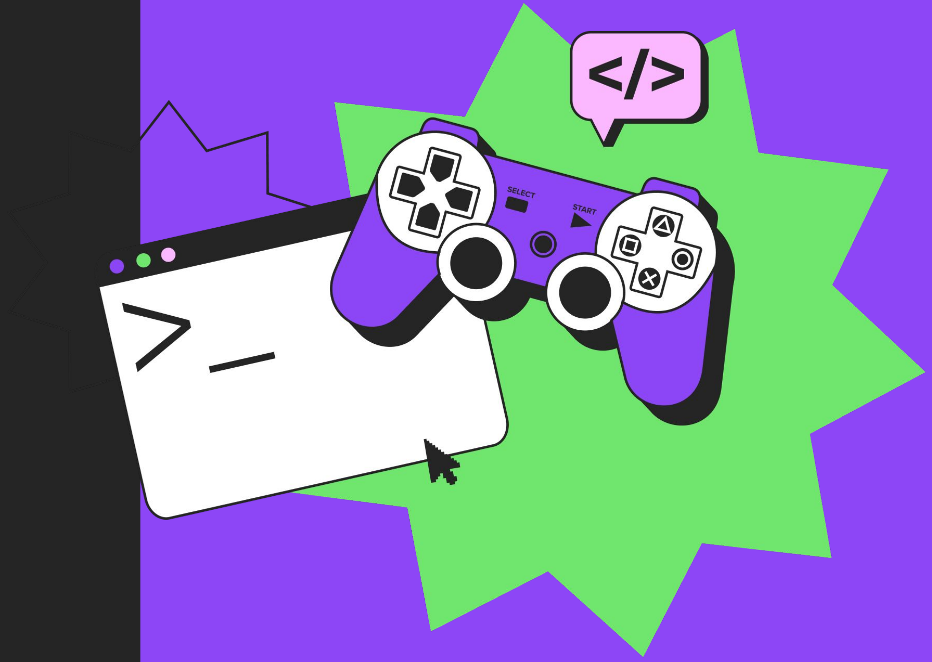
GeekBrains

A/B тестирование

Урок 5

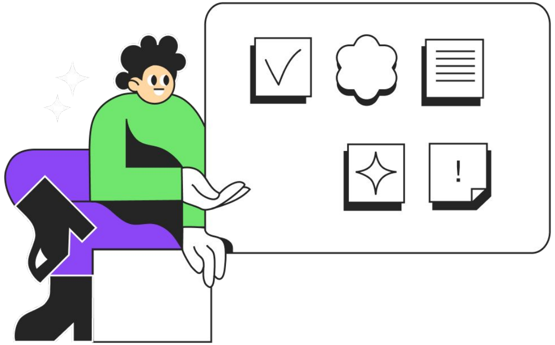
Математическая статистика часть 1



Математическая статистика часть 1

На этом уроке мы разберем

- 💡 Поговорим о важности статистики для A/B тестов
- 💡 Пройдемся по базовым понятиям статистики
- 💡 Разберем как оценивать по выборке в каких границах лежат реальные значения ваших метрик
- 💡 Статистические критерии
- 💡 Алгоритм проверки гипотез
- 💡 Обзор калькуляторов для подсчета результатов

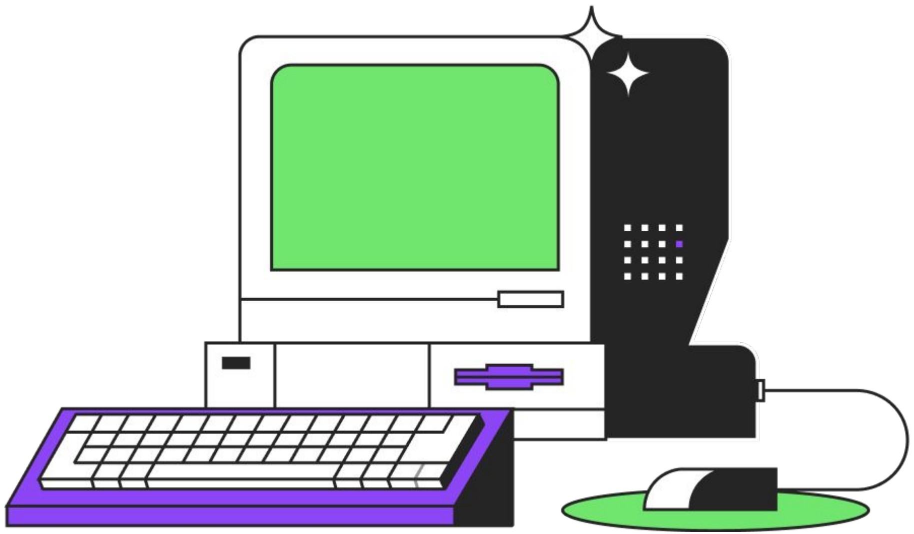


Важность статистики в A/B-тестах

Важность статистики в A/B-тестах

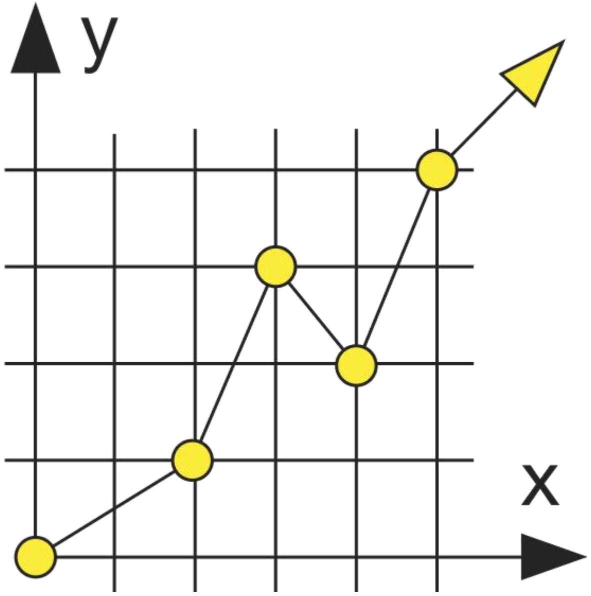
Математическая статистика
это раздел математики, в котором разрабатываются различные методы для описания и анализа наблюдений с цель использования для научных и практических выводов.

Математическая статистика - фундамент A/B тестов, без правильного понимания которого резко возрастает риск принятия неверных решений в продукте. И в этом мы ни раз убедимся в рамках курса.



Что нужно, чтобы проводить A/B-тестирование?

1. Уметь рассчитывать размер выборки для теста
2. Понимать, что означает мощность теста
3. Понимать, насколько страшны ошибки I и II рода
4. Понимать, что означает p-value и доверительный интервал
5. Знать основные статистические критерии
6. Уметь корректно посчитать результаты теста



Базовые понятия

Выборка и ген.совокупность

Генеральная совокупность - совокупность всех объектов или наблюдений, относительно которых исследователь намерен делать выводы при решении конкретной задачи. В ее состав включаются все объекты, которые подлежат изучению.

Выборка - часть генеральной совокупности с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании

Чтобы переносить выводы с выборки на генеральную совокупность, выборка должна быть **репрезентативной,** отражать пропорции и особенности генеральной совокупности.

Оценка параметров на основе выборки

Случайная величина (ξ) – это математическое понятие, служащее для представления случайных явлений, когда для них может быть определена их вероятность, то есть мера возможности наступления.

По сути это переменная со значениями. Для каждого значения есть своя вероятность исхода

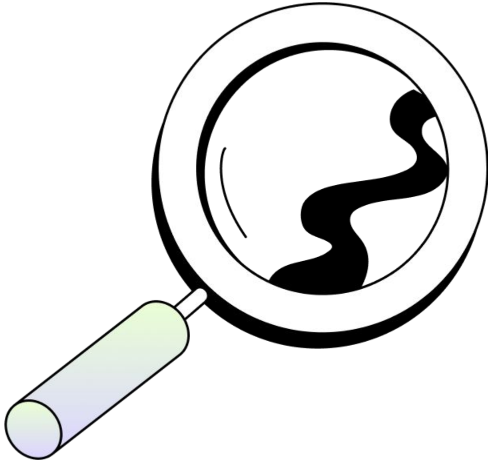
Примеры случайных величин:

-Оценка студента на экзамене

-Цифра выпавшая при броске игральной кости

-Время, которое провел юзер на странице за сеанс

-Цена акции на бирже



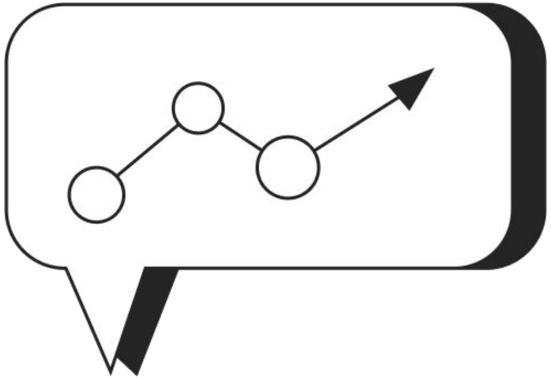
Любая выборка представляет собой значения какой-либо случайной величины.

Оценка параметров на основе выборки

Для точечного оценивания параметров случайной величины используются различные статистики. **Статистика** — это любая измеримая функция от выборки.

Пусть дана выборка Y = (y1, y2, . . . , yi) значений случайной величины

Как ее можно описать с помощью статистик?



Оценка параметров на основе выборки.

Математическое ожидание (*μ*, *M*) (выборочное среднее)— это среднее арифметическое значение случайной величины.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Дисперсия (*S2*) – рассчитанное расстояние, на которое значения случайной величины находятся вокруг его математического ожидания

$$\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Стандартное отклонение (*SD*)— это квадратный корень от дисперсии

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

Медиана это та ... ке, что ровно половина из элементов выборки больше него либо равна, а другая половина меньше него либо равна.

1, 3, 5, 5, 8, 9, **11,** 17, 18, 24, 77, 218, 633

Блюдо	Цена руб
1	500
2	450
3	400
4	470

Мат ожидание	Дисперсия	Стандартное отклонение
455	1 767	42

Рассчитаем дисперсию:

((500-455)^2 + (450-455)^2 + (400-455)^2 + (470-455)^2) / (4-1) = 1767

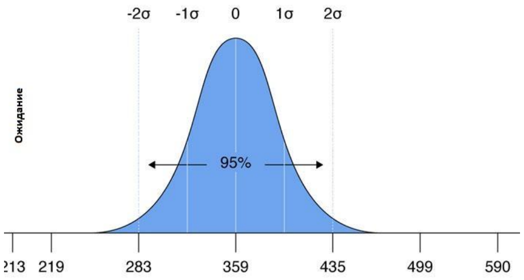
Описательные статистики

Меры центральной тенденции

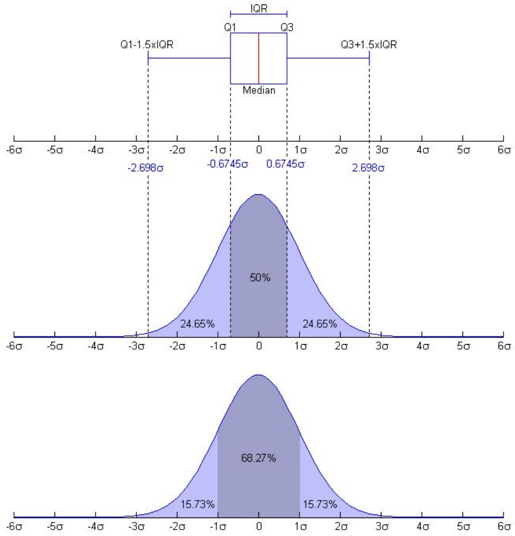
- Среднее значение
- Медиана
- Мода

Меры разброса данных

- Дисперсия
- Стандартное отклонение

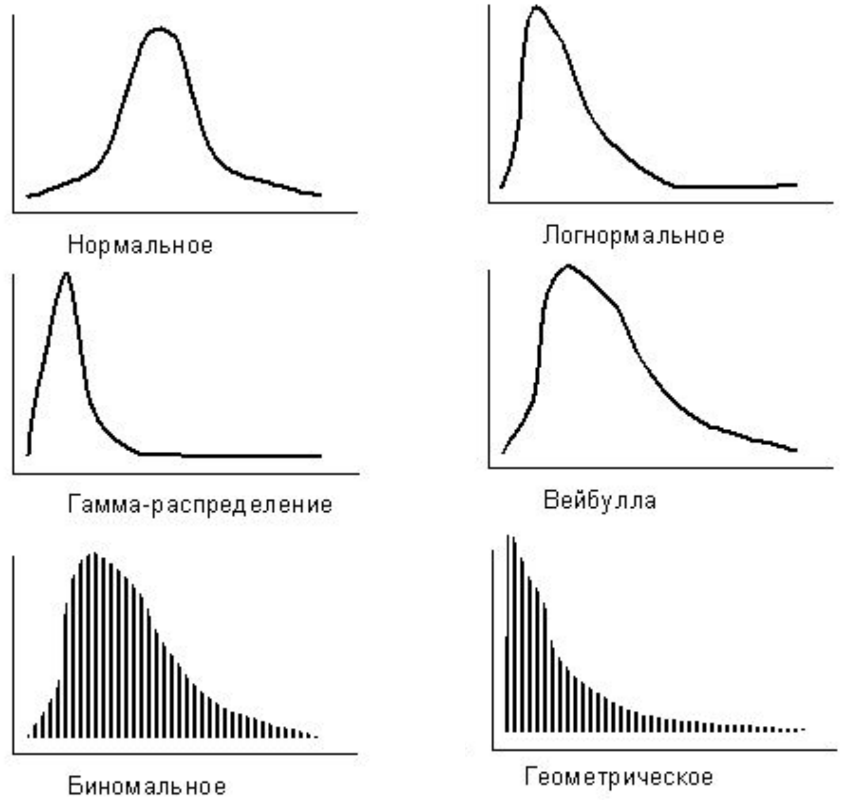


Квантили



Законы распределения

Законом распределения случайной величины называется соотношение, устанавливающие связь между возможными значениями случайной величины и соответствующими им вероятностями.

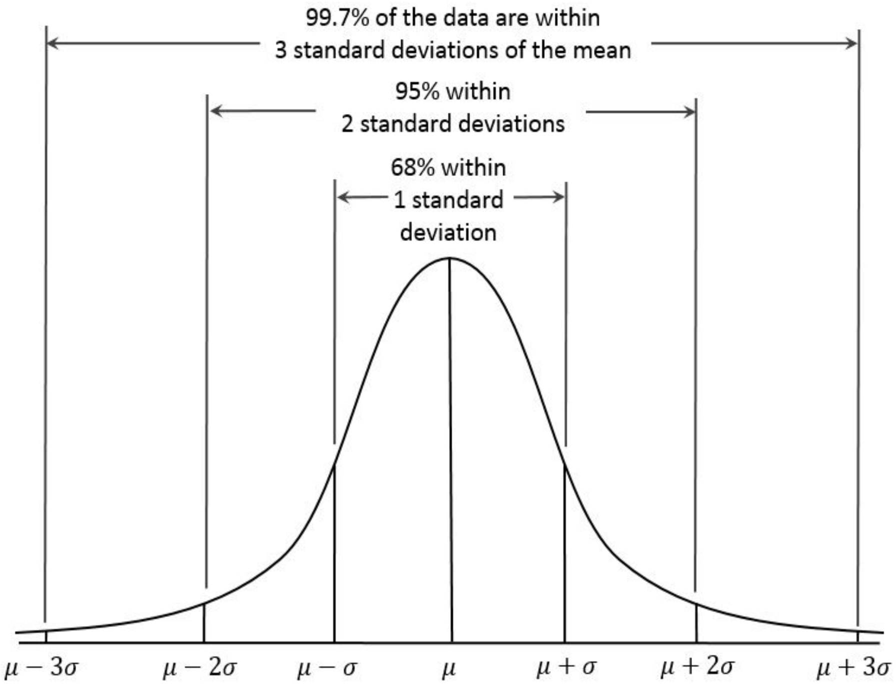


Нормальное распределение

Функция плотности распределения:

Вероятность того, что случайная величина X будет лежать в отрезке (x, y), равна площади под графиком функции плотности f(x) в пределах от x до y.

Общая площадь под графиком функции f (x) равна 1.



Нормальное распределение

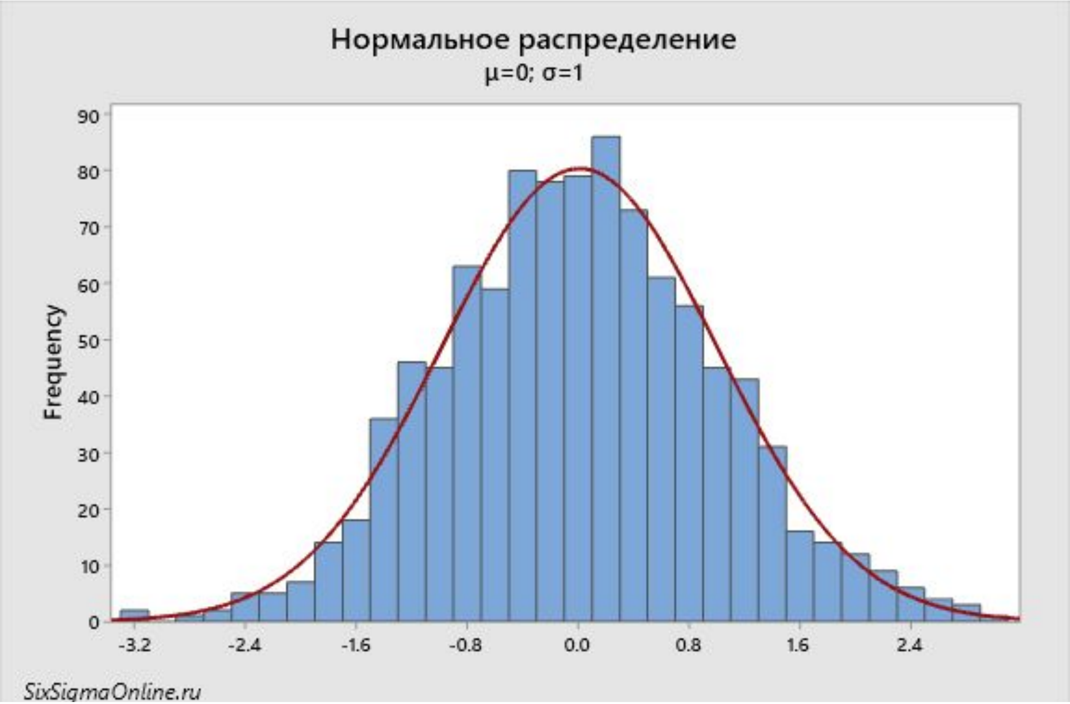
Гистограмма распределения:

По оси x откладываются все значения выборки.

Вся ось x разбивается на заданное число одинаковых отрезков.

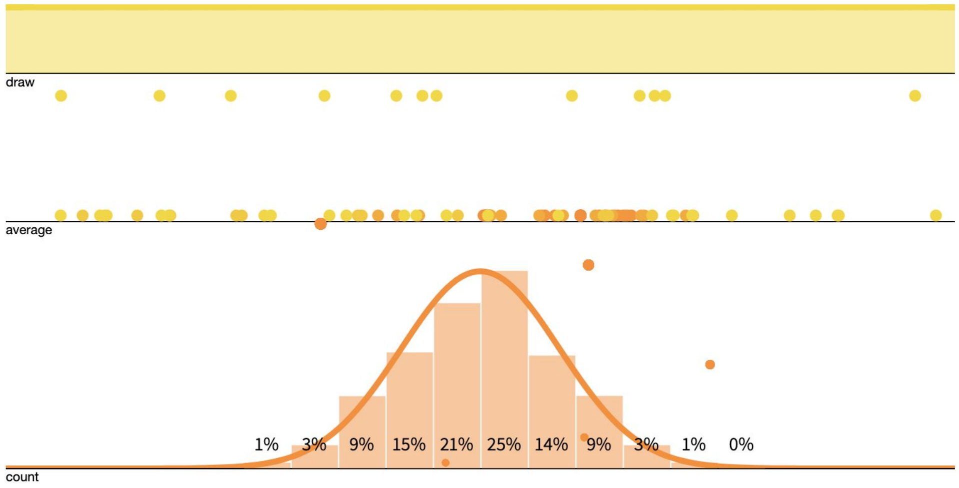
Для каждого отрезка вычисляется кол-во значений выборки, которые лежат в этом отрезке, и это кол-во откладывается по оси y

Гистограмма по форме напоминает график распределения вероятностей случайной величины.



Центральная Предельная Теорема

ЦПТ говорит, что если мы возьмем достаточно большую выборку из независимых, одинаково распределенных случайных величин(i.i.d из генеральной совокупности , то среднее значение будет нормально распределено с ***μ и SD .***

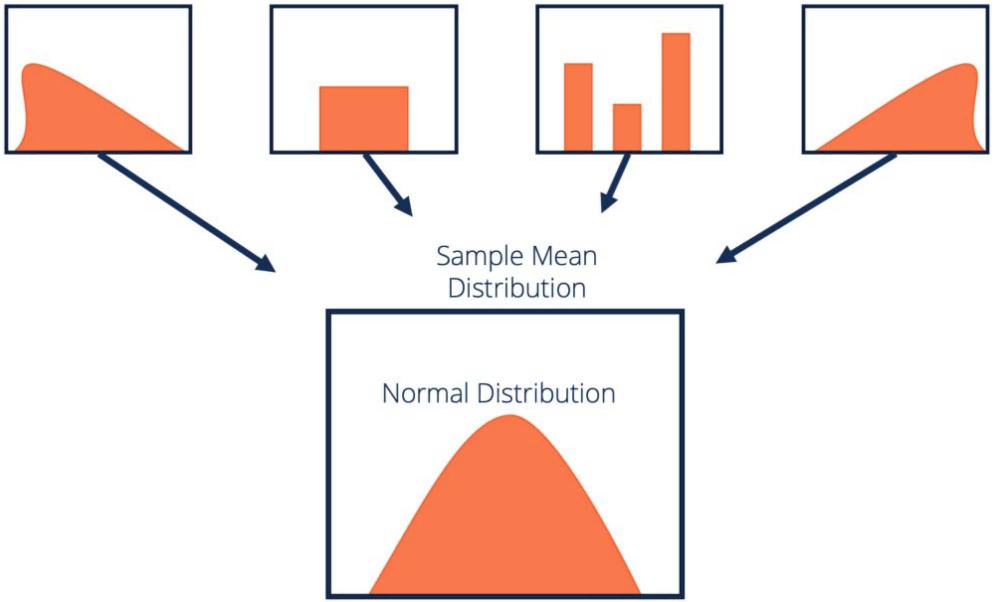


Берем и многократно извлекаем выборки определенного размера и считаем по ним среднее. Распределение средних при соблюдении предпосылок выше будет нормально распределенным и мы сможем оценивать истинное значение в ГС.

ЦПТ

Что нам позволяет делать ЦПТ ?

ЦПТ дает возможность строить доверительные интервалы и проверять статистические гипотезы.



Доверительный интервал

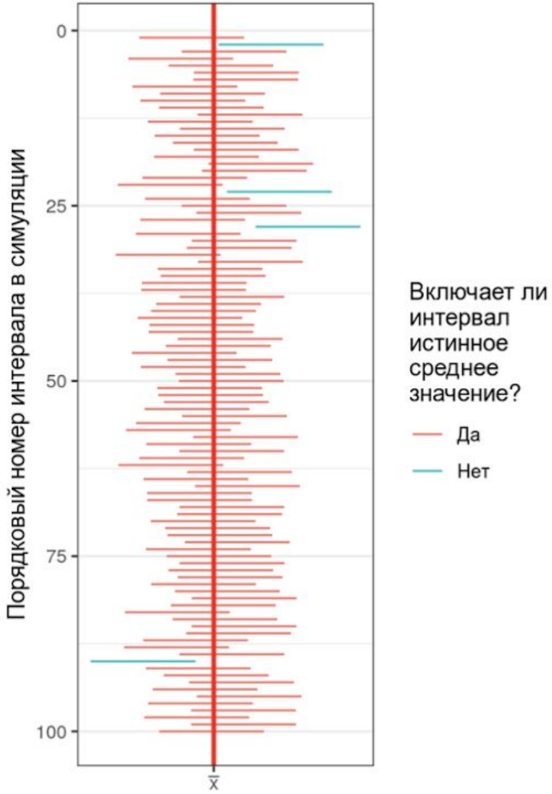
Доверительный интервал

$$M \pm 1,96 \times (SD \div \sqrt{n})$$

Доверительный интервал – Что это такое ? Это способ оценки метрики, используя который, мы получим диапазон значений [x,y] , внутри которого будет лежать истинное значение метрики ГС в 95% случаев.

(Если провести очень большое количество независимых экспериментов с аналогичным построением доверительного интервала, то в 95% экспериментов доверительный интервал будет содержать оцениваемый параметр ген совокупности.

В оставшихся 5% экспериментов доверительный интервал не будет содержать параметр ген совокупности.)



Виды метрик

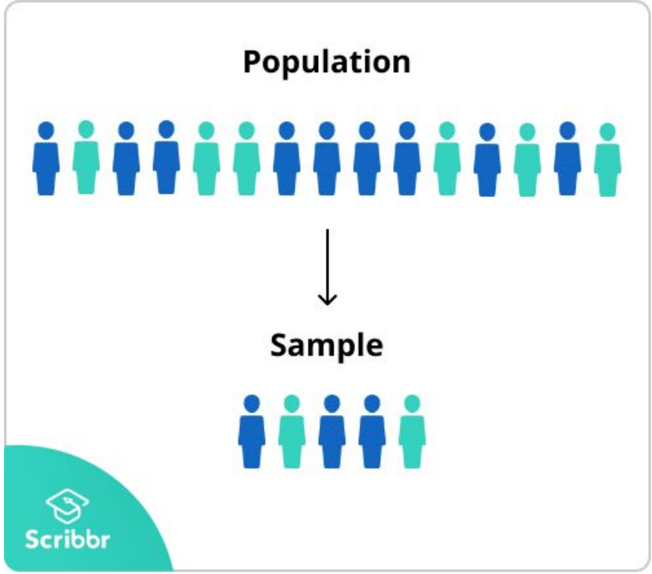
Типы метрик которые бывают в экспериментах:

1)Доли - (ретеншн, конверсии) [0,1,0,0,0,1]

2)Непрерывные - (таймспент в сек / деньги)

3)Отношения - (клики на сессию)

Оценка параметров на основе выборки.



[Доверительный интервал 95% для метрик долей:]

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Ошибки I и II рода

Базовые определения:

Нулевая гипотеза – принимаемое предположение о том, что не существует связи между наблюдениями в двух (или более) событиях (выборках, феноменах, совокупностях). Гипотезу отвергают, если данные показывают разницу между выборками.

True Positive = говорим истина, когда по факту истина (факт)**True Negative** = говорим не истина, когда по факту тоже не истина (факт)

False Positive (ошибка I рода) = говорим истина, когда по факту не истина. Отклонение верной нулевой гипотезы. Риск совершить такую ошибку равен выбранному уровню статистической значимости (например, $\alpha=0.05$) (ошиблись)

False Negative (ошибка II рода) = говорим не истина, когда по факту истина (ошиблись). Принятие неверной нулевой гипотезы. Вероятность отклонить реально работающее изменение

Ошибки I и II рода

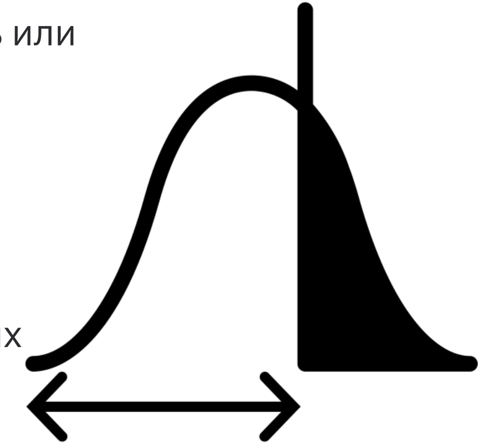
Если бы влияние на конверсию было значительным, но мы это не обнаружили

GeekBrains

Проверка статистических гипотез

Основные понятия

Статистическая гипотеза — выдвигаемое предположение о свойствах случайной величины/виде ее распределения, которое можно подтвердить или опровергнуть на основании имеющихся данных.



Примеры:

-Между конверсиями в покупку в двух группах нет статистически значимых различий

-Между Retention 7 дня в двух группах нет статистически значимых различий

-Случайная величина имеет нормальное распределение

Основные шаги при проверке гипотез

1) Формулируются нулевая и альтернативная гипотезы.

Нулевая гипотеза – принимаемое предположение о том, что не существует связи между наблюдениями в двух (или более) выборках. Гипотезу отвергают, если данные показывают разницу между выборками. (чаще всего в A/B тестах используют двухсторонние гипотезы)

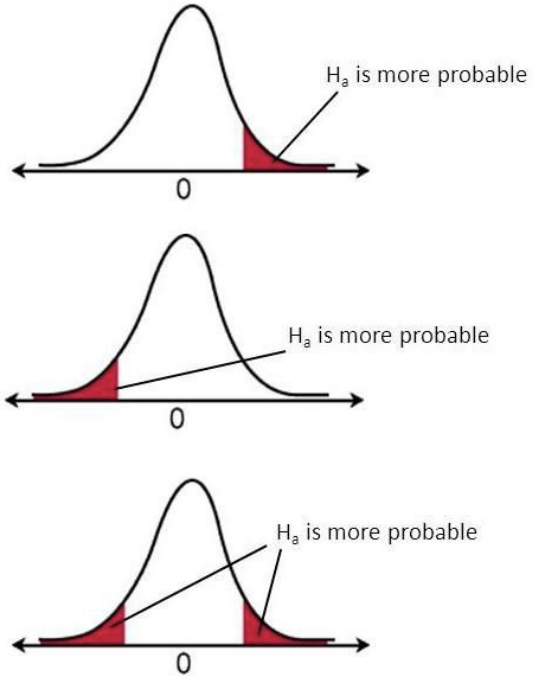
В зависимости от задачи альтернативные гипотезы бывают левосторонние, правосторонние или двухсторонние.

1) Задаётся статистика (функция от выборки) F(Y), которая в условиях справедливости нулевой гипотезы H0 имеет известное распределение

1) Фиксируется уровень значимости α (false positive) — допустимая для данной задачи вероятность ошибки первого рода (чаще всего 0.01, 0.05 или 0.1).

1) Проводится статистический тест: для выборки(выборок) Y считается значение F(Y), и если оно принадлежит критической области, то заключаем, что данные противоречат гипотезе H0, и принимается гипотеза H1.

Основные шаги при проверке гипотез



Right-tail test

H_a: $\mu >$ value

Left-tail test

H_a: $\mu <$ value

Two-tail test

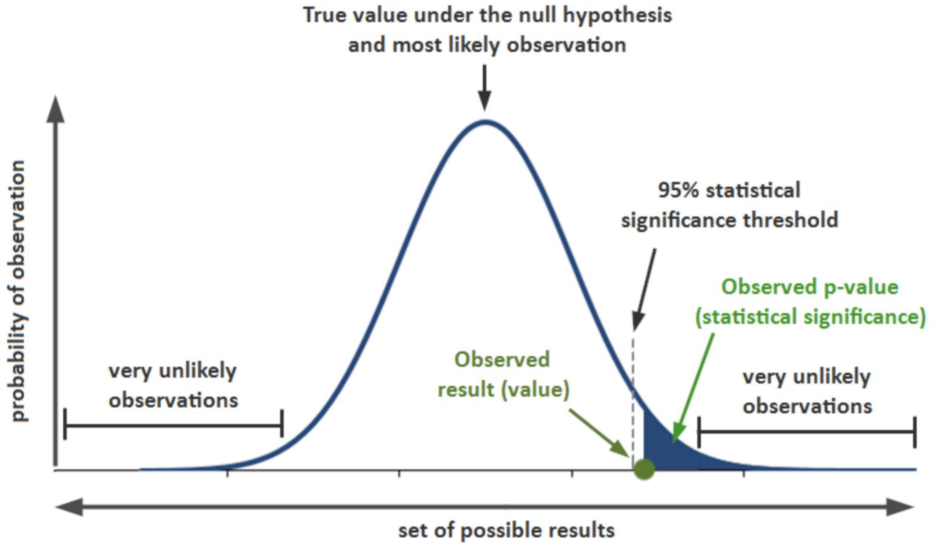
H_a: $\mu \neq$ value

В зависимости от задачи альтернативные гипотезы бывают левосторонние, правосторонние или двусторонние.

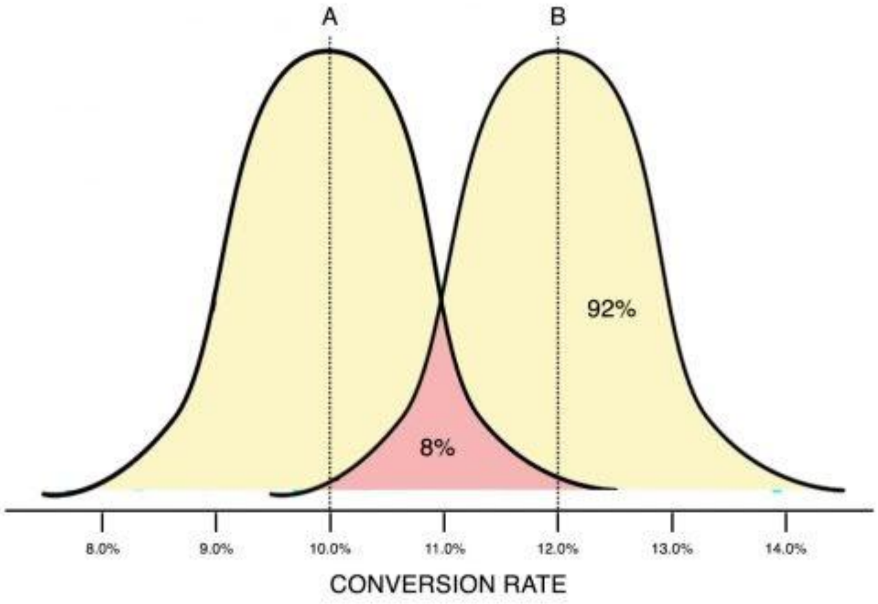
Основные понятия

p-value – вероятность получить наблюдаемое или еще большее отклонение оценки от гипотезы, если она (гипотеза) верна. Геометрически это площадь под кривой, которая начинается от статистического критерия в сторону от гипотезы (от центра).

Если p-value < alpha - нулевая гипотеза отвергается



Сравнение средних



У каждой из метрик в двух выборок есть сигнал (разница средних) и шум (дисперсия). Мы хотим понять несмотря на наличие шума а есть ли действительная разница между средними ?

В этом нам помогают статистические критерии.

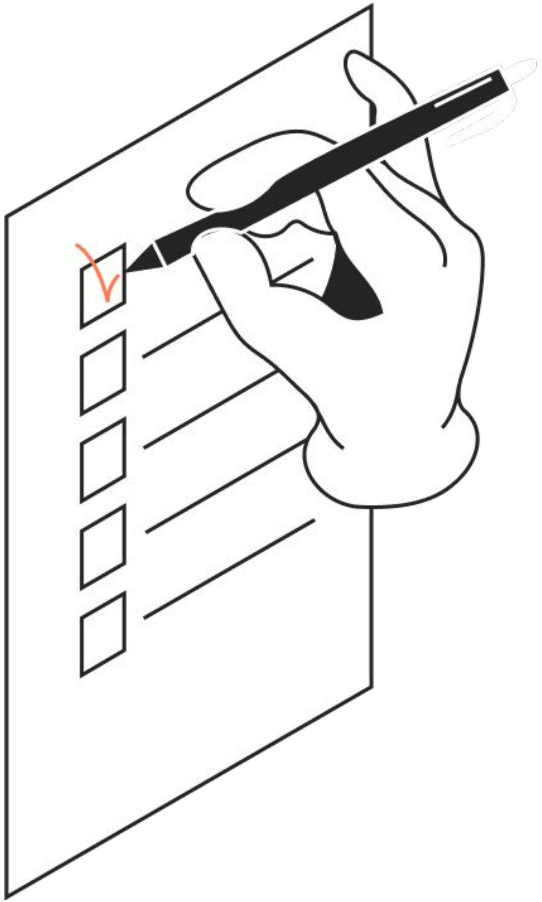
Статистические критерии

Статистический критерий — математическое правило, в соответствии с которым отвергается либо не отвергается та или иная статистическая гипотеза с заданным уровнем значимости.

Для того, чтобы проверить гипотезу о равенстве показателей, применяется два типа критериев оценки: параметрические и непараметрические.

Параметрическими называются критерии, в которых мы можем сделать предположение о распределении , относящееся к какой-то выборке. В большинстве случаев в качестве распределения используется нормальное.

Непараметрические не используют предположения о распределении, а оперируют *рангами* и *частотами.*



Основные понятия

Статистическая мощность (True Positive) — это вероятность, что тест правильно засечёт эффект там, где он и правда есть. (т.е. $1-\beta$)

Чтобы найти хороший критерий для проверки гипотезы H0 vs H1 нужно из всех корректных критериев выбрать критерий с максимальной мощностью. У непараметрических критериев мощность меньше по сравнению с параметрическими -при возможности лучше использовать параметрические критерии.

Даже при ненормальности изначального распределения - ЦПТ работает для распределения средних на больших выборках

Типы данных:

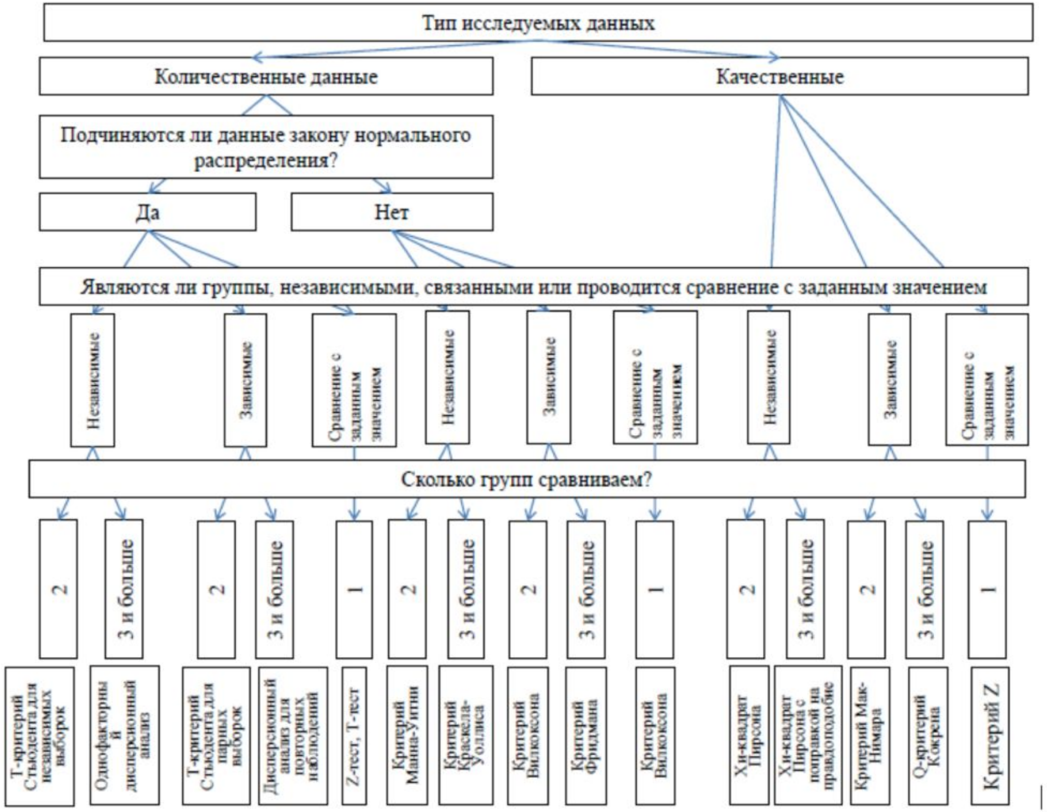
1. **Количественные**

- Непрерывные (средний чек , таймспент итд)
- Дискретные (число детей, число мотоциклов итд)

2. **Качественные**

- номинативные (пол , названия групп, именна итд)
- ранговые (оценка в психологическом исследовании , оценка асессора итд)

Независимые выборки – это те выборки, в которых вероятность отбора любого респондента одной выборки не зависит от отбора любого из респондентов другой выборки. Пример: рандомно взятые новые пользователи



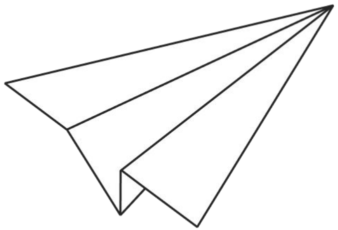
Выбор критерия light version

- **Конверсии** (Хи-квадрат на однородность распределения в двух ген совокупностях или Z критерий долей)

- **Средние**
  если нормальное распределение то t критерий
  если не нормальное распределение то критерий Манна-Уитни

Пример:

Допустим мы запустили эксперимент направленный на улучшение конверсии на одном из кусков сайта:



Группа	Посетители	Кликнувшие посетите...	Конверсия %
Тестовая группа	2400	178	7,42%
Контрольная группа	2400	199	8,29%

Нулевая гипотеза – Между конверсией в двух группах нет статистически значимых различий

Альтернативная гипотеза – Между конверсией в двух группах есть статистически значимые различия

$$\hat{p}_2 - \hat{p}_1 \pm 1.96 \times \sqrt{\hat{p}_1 \times (1 - \hat{p}_1) \div n_1 + \hat{p}_2 \times (1 - \hat{p}_2) \div n_2}$$

0.0742-0.0829 ± 1.96 * (0.0742 × (1 − 0.0742) ÷ 2400 + 0.0829 × (1 − 0.0829) ÷ 2400) ^1/2 = [-2.392% ; 0.652%]

0 входит в интервал - значит на основе собранных данных статистически значимой разницы между группами не обнаружено.

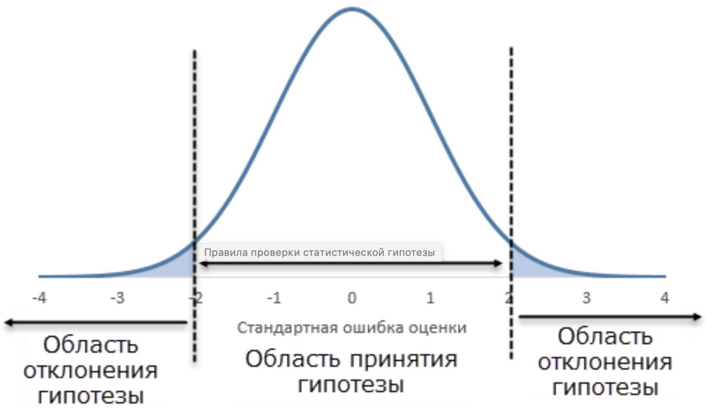
Алгоритм проверки гипотез

Проверка гипотез

1. Выбираем метрику и формулируем нулевую и альтернативную гипотезы
2. Выбираем параметр alpha (например 5%) равный вероятности допустить ошибку первого рода
3. Выбираем критерий, подходящий под наши условия
4. Считаем p - value и(или) доверительный интервал и делаем вывод:

Если p-value < alpha - разница между группами стат.значима . Либо если доверительный интервал для разницы не включает 0.

5. Даем (рекомендации лицам принимающим решения /принимаем решение) выкатывать или не выкатывать новое изменение



Обзор калькуляторов

Выбор критерия light version

- **Конверсии** (Хи-квадрат на однородность распределения в двух ген совокупностях или Z критерий долей)

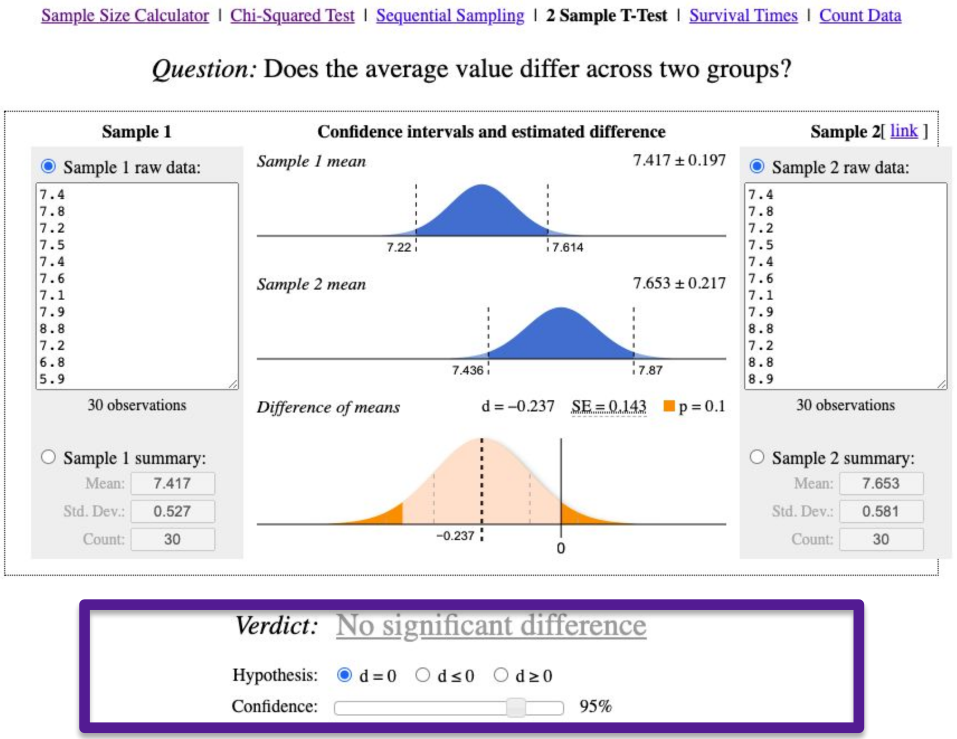
- **Средние**

  если нормальное распределение то t критерий

  если не нормальное распределение то критерий Манна-Уитни

t критерий Стьюдента

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

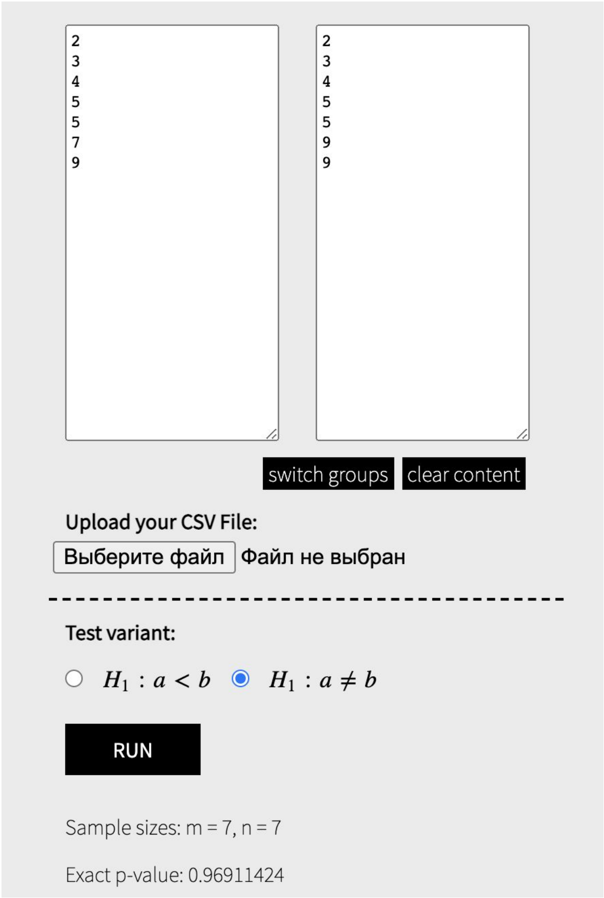


GeekBrains

U критерий Манна-Уитни

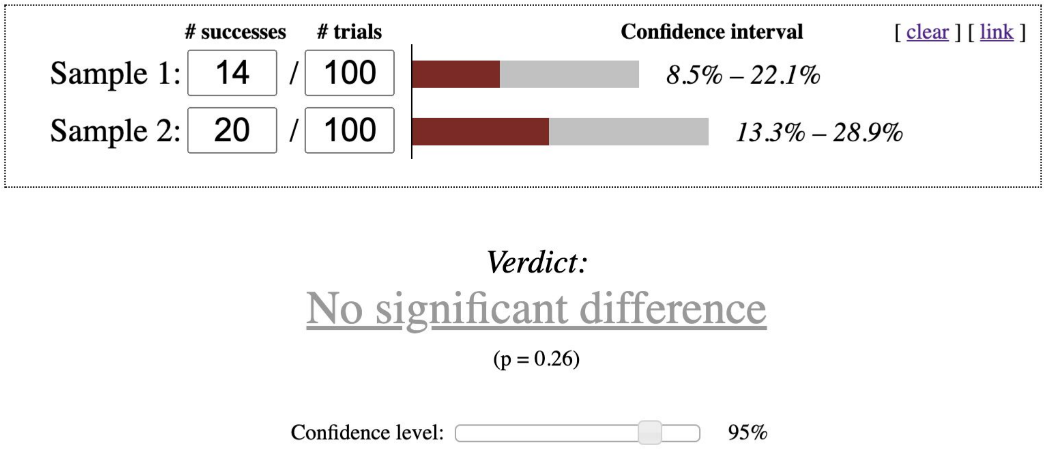
$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U = \min(U_1, U_2)$$



Критерий Хи-квадрат на однородность распределений

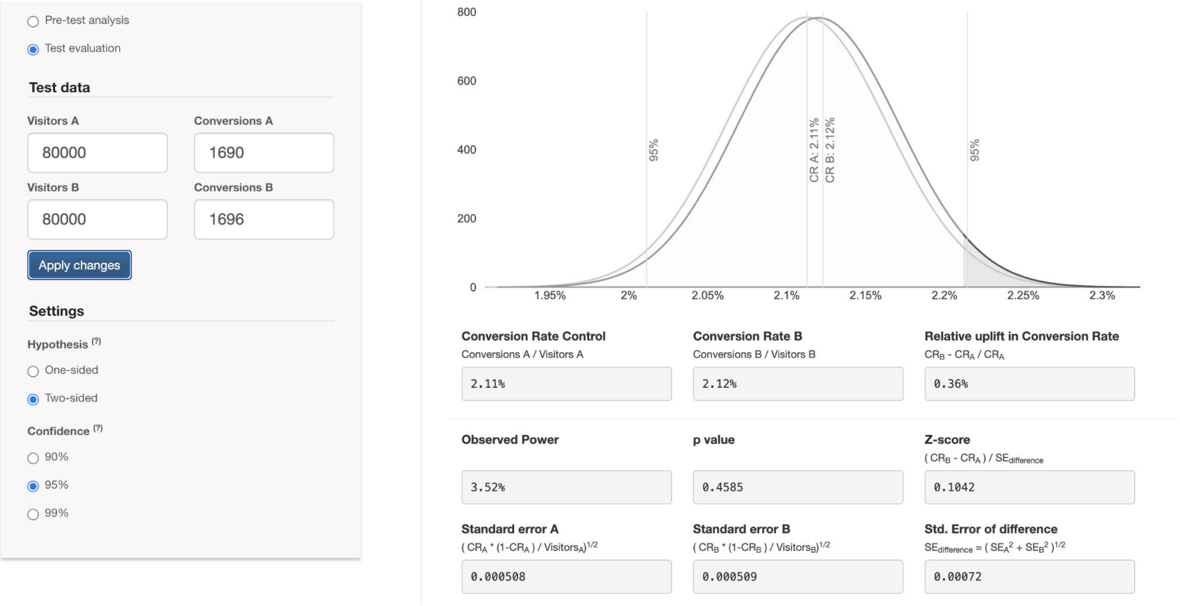
$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$



Z test для долей

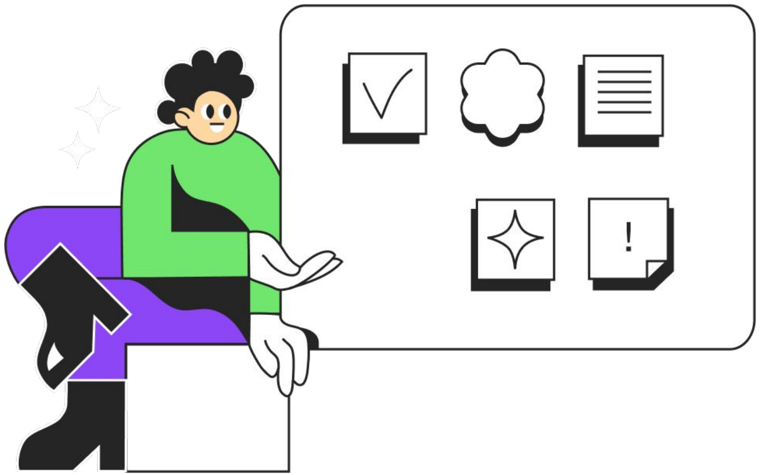
Z-score

$(CR_B - CR_A) / SE_{difference}$



На этом уроке мы разобрали

- 💡 Освежили базовые понятия из статистики
- 💡 Принцип проверки статистических гипотез
- 💡 Популярные статистические критерии



На следующем уроке мы рассмотрим:

- Для чего рассчитывать длительность теста?
- Как понять сколько дней держать а/б тест ?
- Какой объем выборки нужен?
- Когда останавливать тест?