OTUS

Статистические гипотезы. A/B-тесты

otus.ru

REC Проверить, идет ли запись

Меня хорошо видно && слышно?

Маршрут вебинара

- Узнаем про A/B тест
- Тест на разницу средних
- Тест Манна-Уитни
- Практика

Цели вебинара

После занятия вы сможете

1.	Познакомиться с этапами проведения A/B тестов
2.	Узнать, как понять, верна ли ваша гипотеза
3.	Узнать математические основы проверки гипотез

Важность A/B-тестов

Оффлайн оценки качества моделей не достаточно, так как

- она проводится на исторических данных без взаимодействия с пользователем
- важно оценить бизнес метрики

Пример:

- **Гипотеза**: Новая модель увеличит количество кликов на карточку товара.
- **Реализация**: Проведите A/B-тест, в ходе которого половина пользователей увидит рекомендации от старой модели (группа A), а другая половина - от новой (группа B).
- **Рассчет**: Отследите количество кликов для обеих групп за определенный период.

Проблемы при A/B-тестировании

- Метрики и их чувствительность

Проблема — проводить эксперименты так, чтобы быстрее зафиксировать статистически значимое изменение метрики.

Решение - повышение чувствительности A/B-тестов.

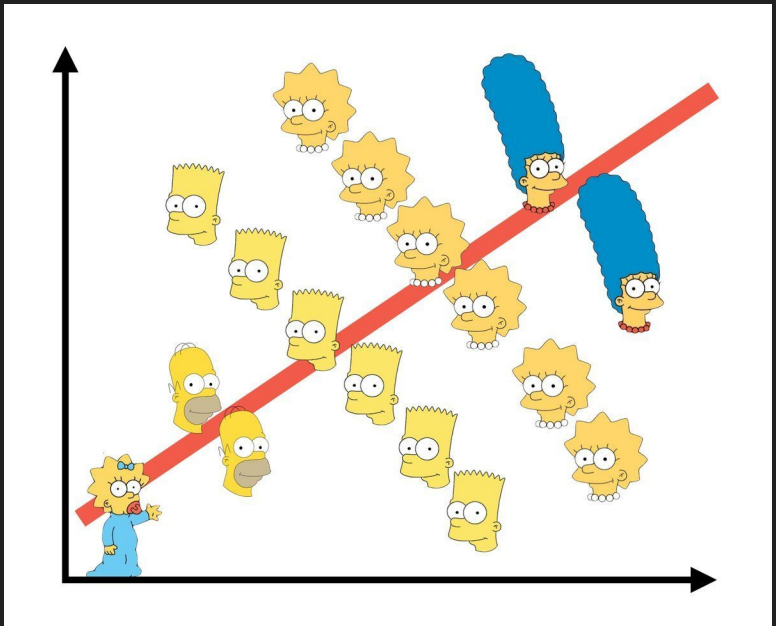
- Волатильность метрик и некорректное разбиение на группы

Метрики, на которые смотрят во время тестирования, не всегда стационарны.

Решение A/A, A/A/B-тесты.

Проблемы при A/B-тестировании

- Парадокс Симпсона



Результаты, полученные при анализе данных по разным группам, могут противоречить результатам, полученным при анализе данных по объединенным группам.
Это может привести к искажению выводов о взаимосвязи между переменными или оценки эффективности какого-либо воздействия.

Парадокс Симпсона

Конверсия	Не видел рекомендации	Видел рекомендации
Да	4000	4800
Нет	400	320
Коэффициент конверсии	9%	6%

	Мобильный		Десктоп	
Конверсия	Не видел рекомендации	Видел рекомендации	Не видел рекомендации	Видел рекомендации
Да	1600	4200	2400	600
Нет	40	180	360	140
Коэффициент конверсии	2%	4%	13%	19%

Определение и цель A/B-тестирования

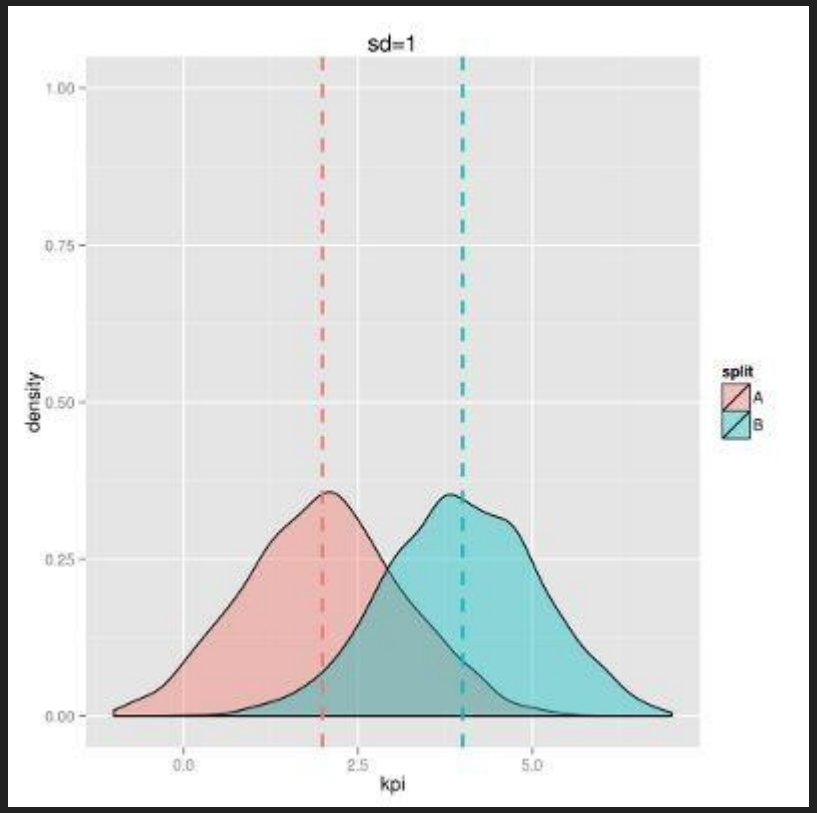
A/B-тестирование представляет собой статистический метод, используемый для сравнения двух версий дизайна/модели/..., обычно путем проверки реакции субъекта на вариант A по сравнению с вариантом B.

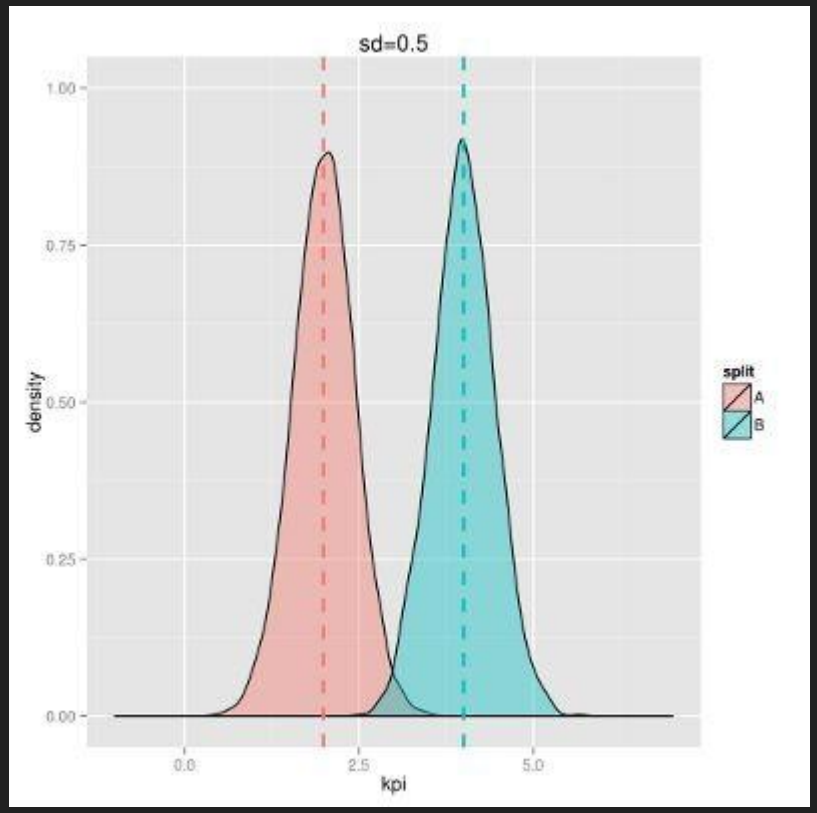
Основная цель A/B-тестирования — принятие решений на основе данных и оптимизация процессов для достижения лучших результатов.

A/B тест

Результаты:

- Группа A (старая модель): 2% кликов
- Группа B (новая модель): 4% кликов





Понимание основ A/B-тестирования

A/B-тестирование включает в себя сравнение двух версий (A и B) моделей или других маркетинговых стратегий, чтобы определить, какая из них работает лучше. Процесс обычно включает в себя:

1. **Рандомизация**: Пользователи случайным образом распределяются по группам A и B, чтобы обеспечить справедливое сравнение.

2. **Эксперимент**: Каждая группа подвергается воздействию различных вариантов (A или B) тестируемой переменной.

3. **Измерение**: Для каждой группы измеряются ключевые показатели эффективности (KPI), чтобы оценить эффективность вариантов.

4. **Статистический анализ**: Статистические методы используются для определения того, являются ли наблюдаемые различия в эффективности статистически значимыми.

Формулирование гипотезы

В A/B-тестировании гипотезы — это утверждения, которые формулируют ожидаемое влияние изменений на конкретный показатель. Хорошо сформулированная гипотеза обычно следует следующей структуре: «Изменение переменной с «А» на «Б» приведет к «ожидаемому результату».

Пример:
- Гипотеза: получение рекомендаций от новой модели увеличит конверсию на 2%.

Связь гипотез с бизнес-целями

Крайне важно согласовать гипотезы с общими бизнес-целями. A/B-тестирование должно способствовать достижению конкретных целей, таких как увеличение доходов, повышение вовлеченности пользователей или оптимизация коэффициентов конверсии.

Пример:
- Бизнес-цель: повысить общую конверсию на веб-сайте.
- Гипотеза: изменение макета главной страницы приведет к увеличению количества регистраций пользователей на 15%.

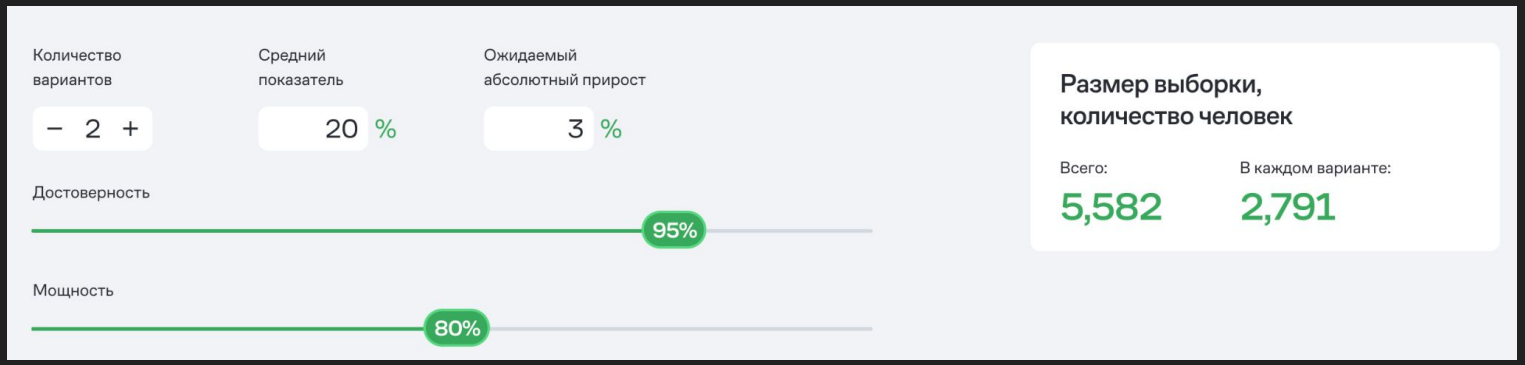
Выбор подходящего размера группы

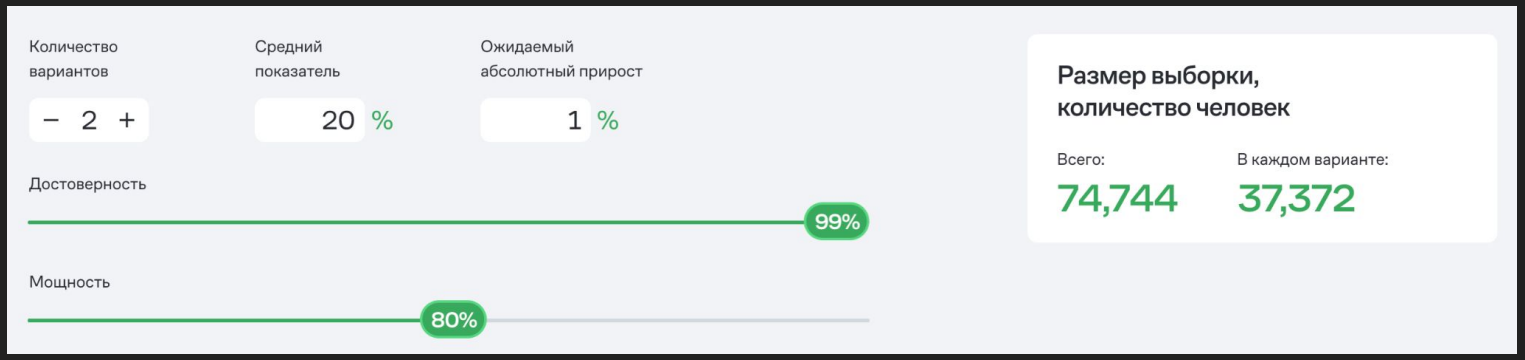
1. Размер эффекта

Размер эффекта представляет собой величину исследуемой разницы. Он используется исследователем для определения ожидаемого эффекта в эксперименте и расчета оптимального размера выборки. Больший размер эффекта часто требует меньшего размера выборки для достижения статистической значимости.

https://mindbox.ru/tools/ab-test-calculator

Выбор подходящего размера группы





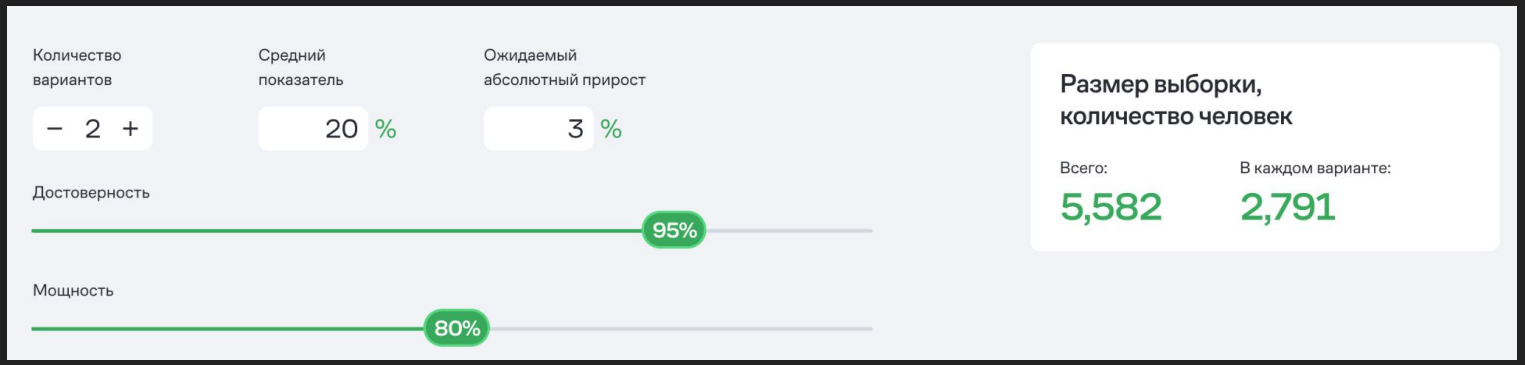
Выбор подходящего размера группы

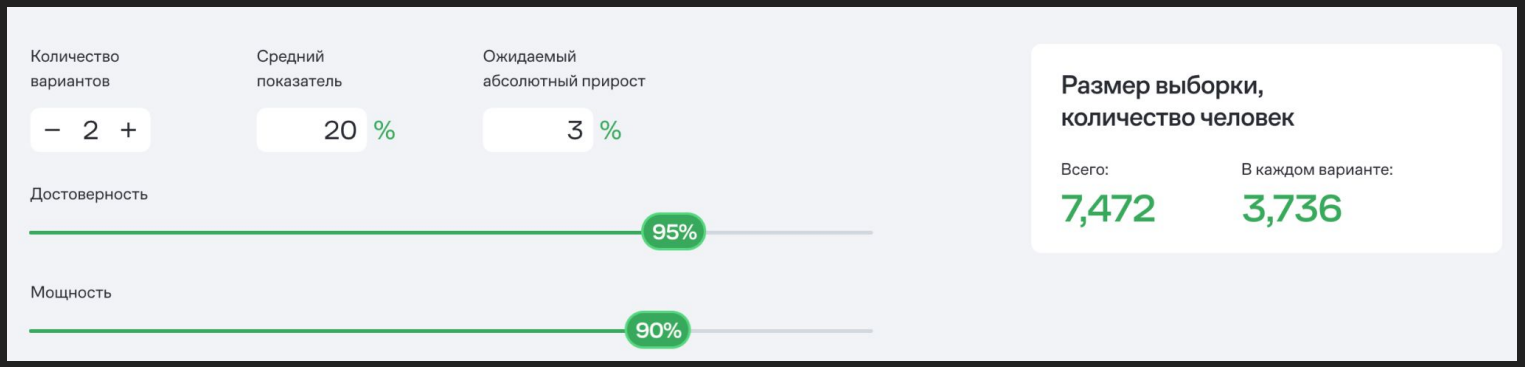
2. Статистическая мощность
Статистическая мощность — это вероятность обнаружения эффекта, если он существует.

Ошибка второго рода (бета) — ситуация, когда принята неверная нулевая гипотеза.

Мощность критерия = 1 - бета. Чем выше мощность критерия, тем меньше вероятность совершить ошибку второго рода.

Выбор подходящего размера группы





Выбор подходящего размера группы

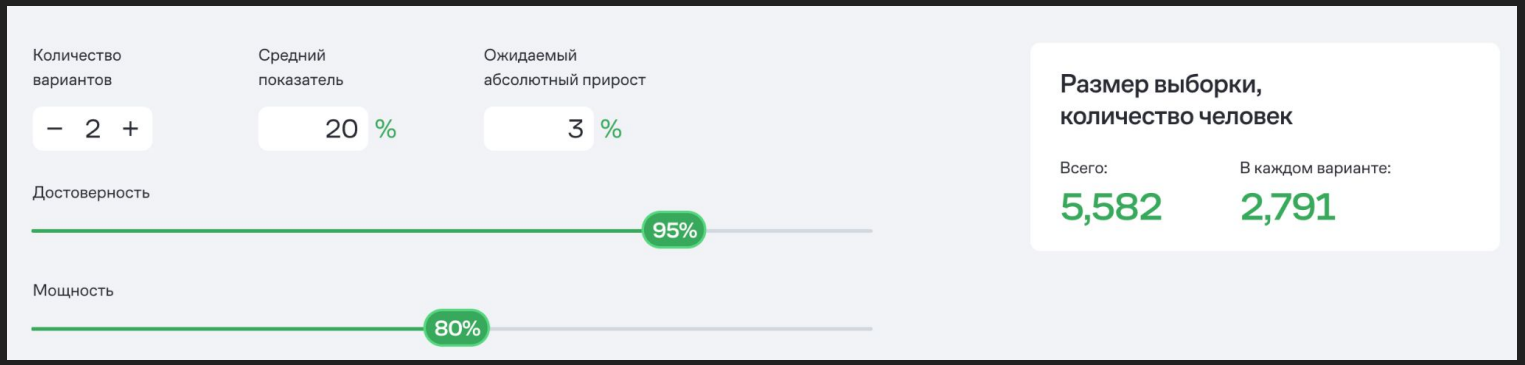
3. Уровень значимости

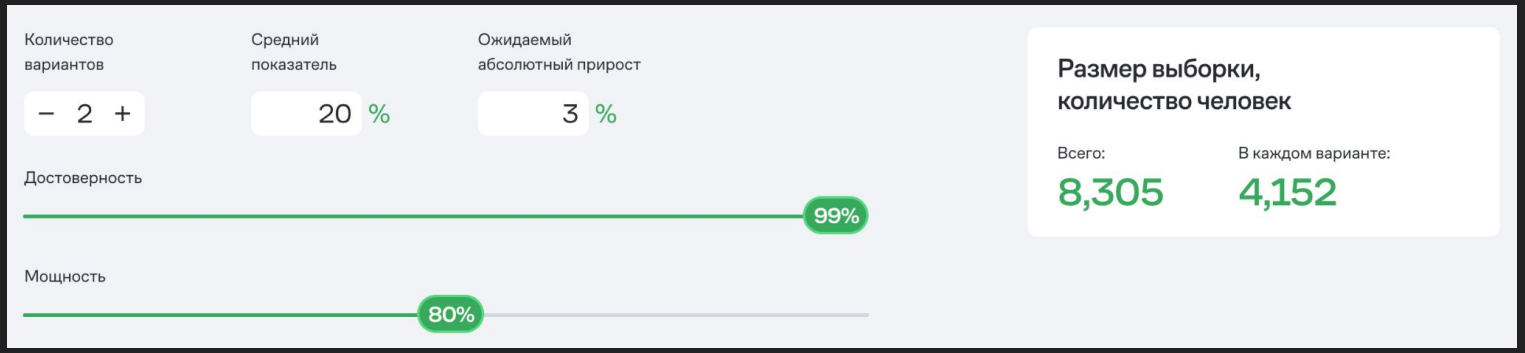
Уровень значимости (альфа) — это вероятность отклонения нулевой гипотезы, которая на самом деле верна.

Ошибка первого рода — ситуация, когда отвергнута верная нулевая гипотеза (об отсутствии искомого эффекта).

Вероятность ошибки первого рода при проверке статистических гипотез называют уровнем значимости.

Выбор подходящего размера группы





Статистические критерии

Статистический критерий — математическое правило, в соответствии с которым принимается или отвергается та или иная статистическая гипотеза.

Построение критерия заключается в выборе подходящей функции от результатов наблюдений, которая служит для выявления меры расхождения между результатами в двух группах.

Статистический тест на разницу средних

Для близкого к нормальному распределению. Является параметрическим, так как сравнивает параметры распределений - средние значения.

1. Вычислим выборочное среднее в группе A и группе B

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

2. Вычислим выборочную дисперсию в группе A и группе B

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

3. Вычисление статистики

$$t_{st} = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$

4. По значению статистики определяем p-value

p-value

P-value - это вероятность получить результат, как минимум, столь же экстремальный, как наблюдаемый результат, при условии, что нулевая гипотеза верна.

В контексте A/B тестирования, p-value используется для оценки статистической значимости различий между двумя группами.

p-value

Чем меньше значение p-value, тем более значимым считается различие между группами. Обычно устанавливается критическое пороговое значение alpha (например, 0.05), и если p-value меньше alpha, то различие считается статистически значимым.

Интерпретация p-value:

Если p-value < alpha, то мы отвергаем нулевую гипотезу и заключаем, что наблюдаемые различия статистически значимы.

Если p-value >= alpha, то мы не можем отвергнуть нулевую гипотезу, и различия между группами не считаются статистически значимыми.

Критерий Манна-Уитни

Для ненормального распределения или малого размера выборки.
Является непараметрическим.

Критерий Манна-Уитни применяется для сравнения двух независимых выборок.
Используется, когда данные не удовлетворяют условиям параметрических тестов (например, нормальности распределения).

Для каждого наблюдения из первой выборки подсчитывается количество наблюдений из второй выборки, которые оно превосходит.

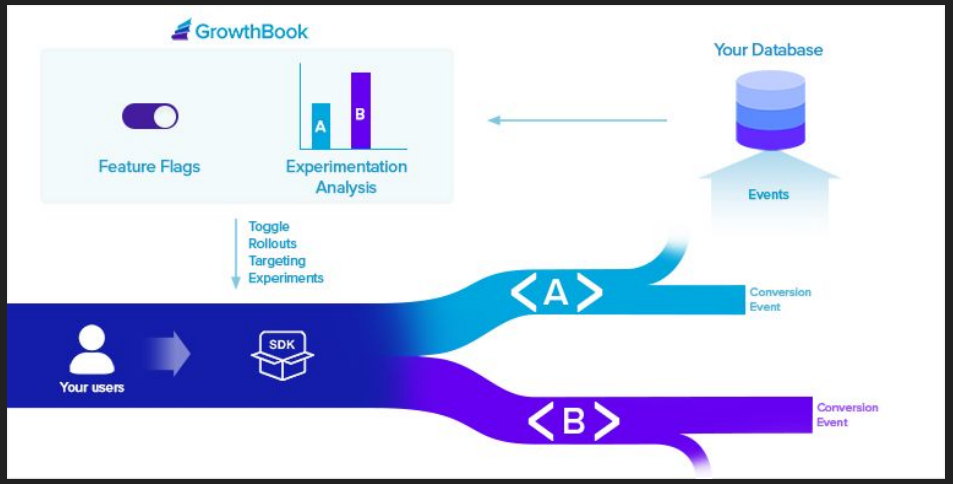
Сумма этих количеств и будет значением статистики U.

Продвинутые методы

Bootstrap
https://habr.com/ru/companies/X5Tech/articles/679842/

Cupped
https://habr.com/ru/companies/X5Tech/articles/780270/

Инструменты для проведения A/B теста



Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Цели вебинара

После занятия вы сможете

1.	Познакомиться с этапами проведения A/B тестов
2.	Узнать как понять верна ли ваша гипотеза
3.	Узнать математические основы проверки гипотез

Заполните, пожалуйста, опрос о занятии
по ссылке в чате и спасибо за внимание