GeekBrains

A/Б тестирование

Курс “Введение в продуктовую аналитику”

Оглавление

Введение

На этой лекции вы найдете ответы на такие вопросы как:

- Что такое А/Б тесты и зачем они нужны
- Какие метрики проверяют АБ тесты
- Способы формирования выборки
- Калькулятор А/Б тестирования
- Проблемы при проведении тестов

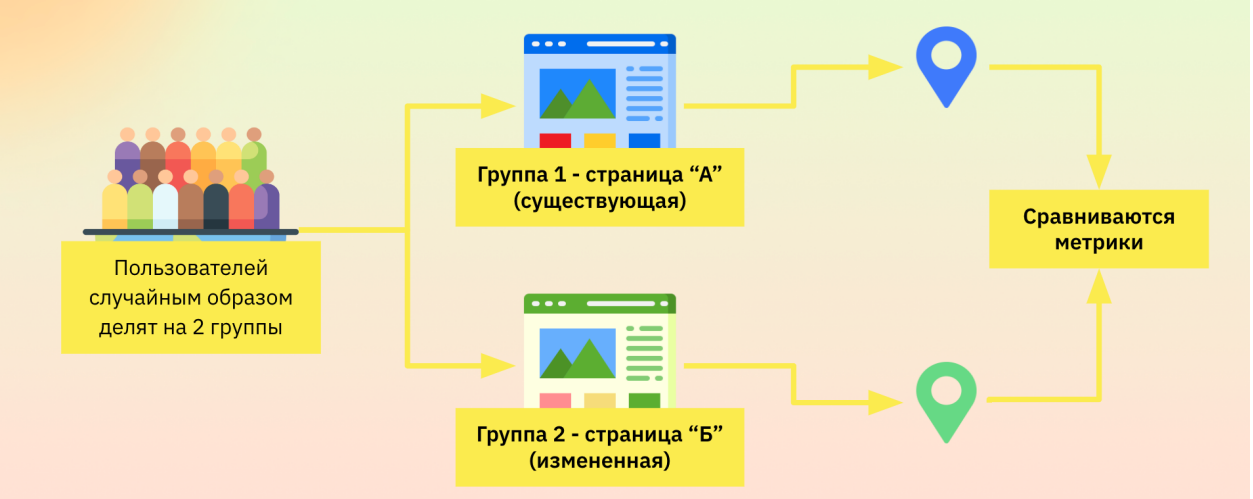
Термины, используемые в лекции

- Генеральная совокупность – это 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 или выборочная совокупность — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).
- Уровень значимости (p-value) статистического теста – это вероятность отклонить нулевую гипотезу , когда на самом деле она верна.
- Уровень доверия означает вероятность того, что доверительный интервал содержит истинное значение оцениваемого параметра распределения.
- Простая случайная выборка (Simple Random Sampling — SRS) – вероятностный метод выборки, согласно которому каждый элемент генеральной совокупности имеет известную и равную вероятность отбора.
- Стратифицированная выборка (stratified sample) – метод семплирования из генеральной совокупности, который позволяет улучшить точность статистических результатов при разбиении всего пространства событий на несколько областей-страт и независимой работе с этими стратами.

- Групповая выборка (cluster sample) – вероятностная выборка, для которой характерна следующая двухступенчатая процедура: 1) генеральная совокупность делится на ряд непересекающихся исчерпывающих ее подмножеств; 2) производится случайный отбор подмножеств
- Доверительный интервал – это способ оценки конверсии, в результате которого мы получаем не одно единственное значение, а интервал значений, внутри которого может содержаться реальное значение конверсии.
- Нулевая гипотеза (H0) – это проверяемое предположение, которое обычно формулируется как отсутствие различий, отсутствие влияние фактора, отсутствие эффекта, равенство нулю значений выборочных характеристик и т.п
- Ошибка первого рода (False Positive) - отклонение нулевой гипотезы, хотя гипотеза верная
- Ошибка второго рода (False Negative) - принятие нулевой гипотезы, хотя верна альтернативная
- ARPU (Average Revenue Per User) - средний доход с одного юзера
- LTV (Lifetime Value) - доход, который получает продукт за всё время работы с клиентом
- Дисперсия – это мера разброса значений случайной величины X относительно её математического ожидания.
- Сезонность характеризуется периодическими колебаниями, которые повторяются каждый сезон по предсказуемой схеме. Они могут быть ежегодными, реже — ежеквартальными или ежемесячными. Чем выше амплитуда, то есть, больше период, тем значительнее сами колебания в спросе и других параметрах.

Определение и применение A/B тестирования

A/B тестирование - это метод исследования, где сравнивается 2 разных варианта, не обязательно для онлайна.



A/B-тестирование или иначе сплит-тестирование — метод маркетингового исследования, в котором:

1. Пользователей случайным образом делят на 2 группы
2. Группа 1 направляется на страницу “А” (существующую)
3. Группа 2 направляется на страницу “Б” (измененную)
4. Сравниваются метрики по обеим группам

Примеры А/Б тестов в офлайне:

1. Медицина: тестирование нового лекарства
2. Ритейл: тестирование выкладки товаров на полки
3. Производство FMCG:

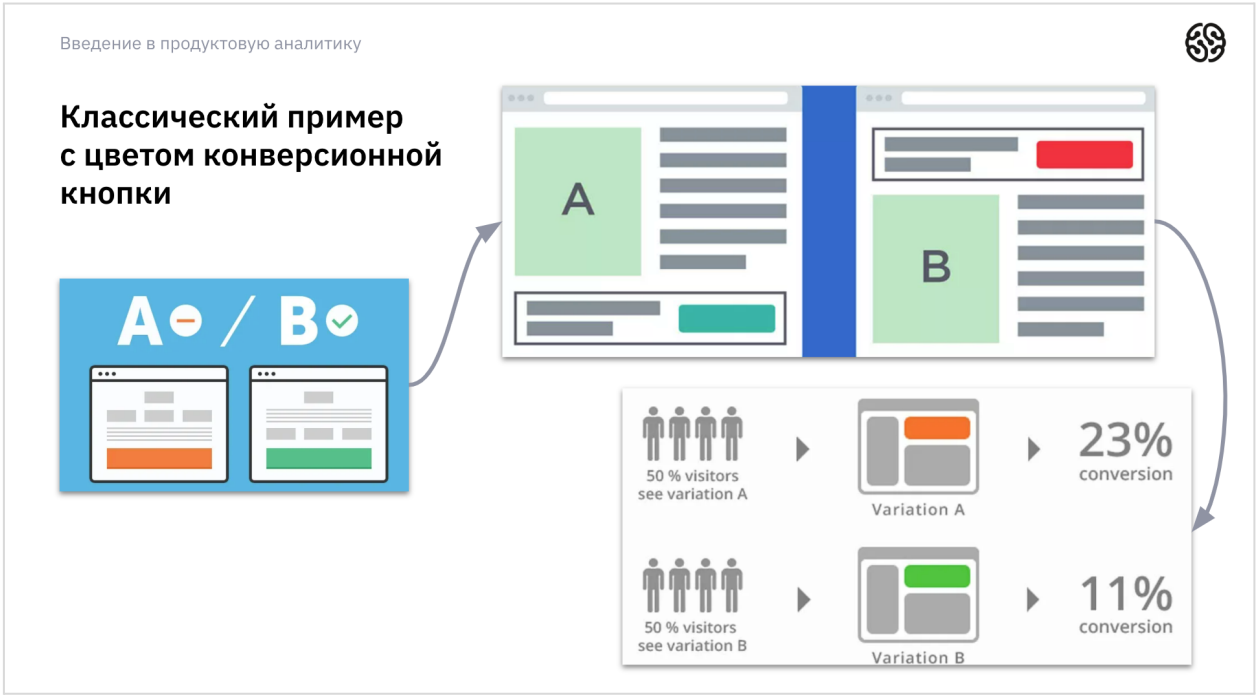
- тестирования разных вариантов упаковки по цвету, форме и т.д.
- тестирования состава продукта (количество сахара, например)

Особенности данного типа тестирования:

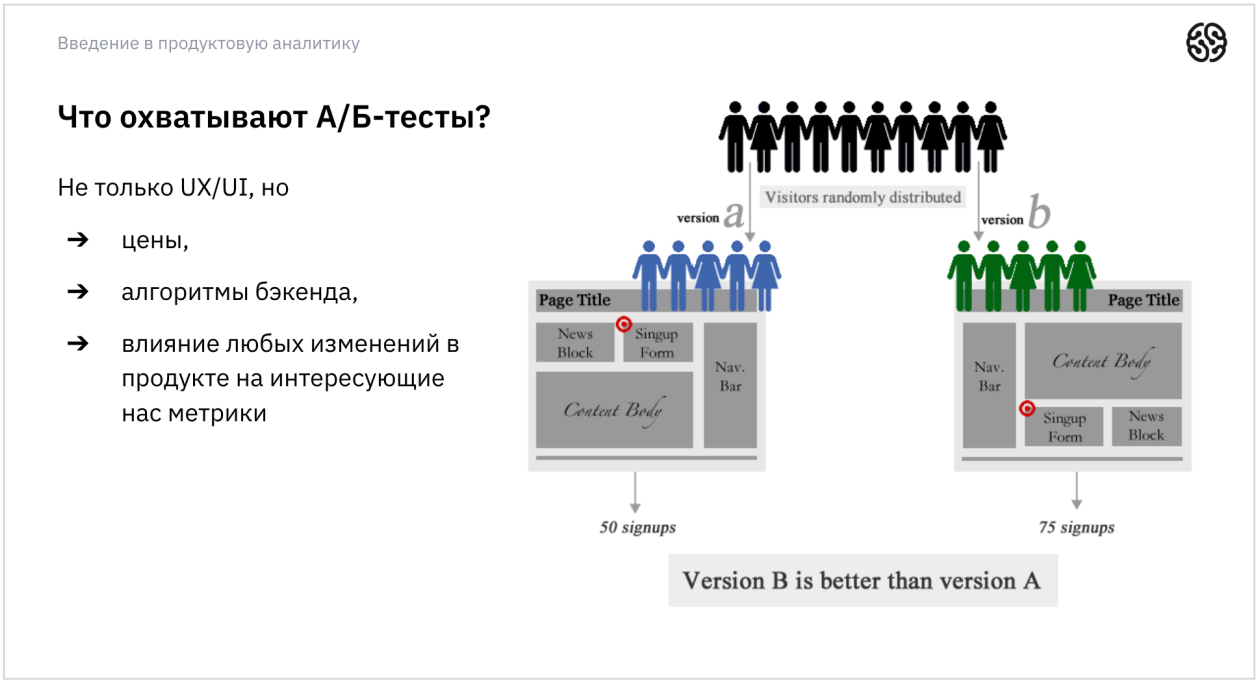
- Варианты могут различаться по разным параметрам аудитории
- Изменения, должны иметь понятный фокус
- Такого рода тестирование помогает “докручивать” результат

- В идеале тестирование нужно проводить на регулярной основе

Классический пример с цветом конверсионной кнопки.



Что охватывают A/B тесты?



A/B тесты применяться в анализе не только таких показателей, как UX/UI, но и

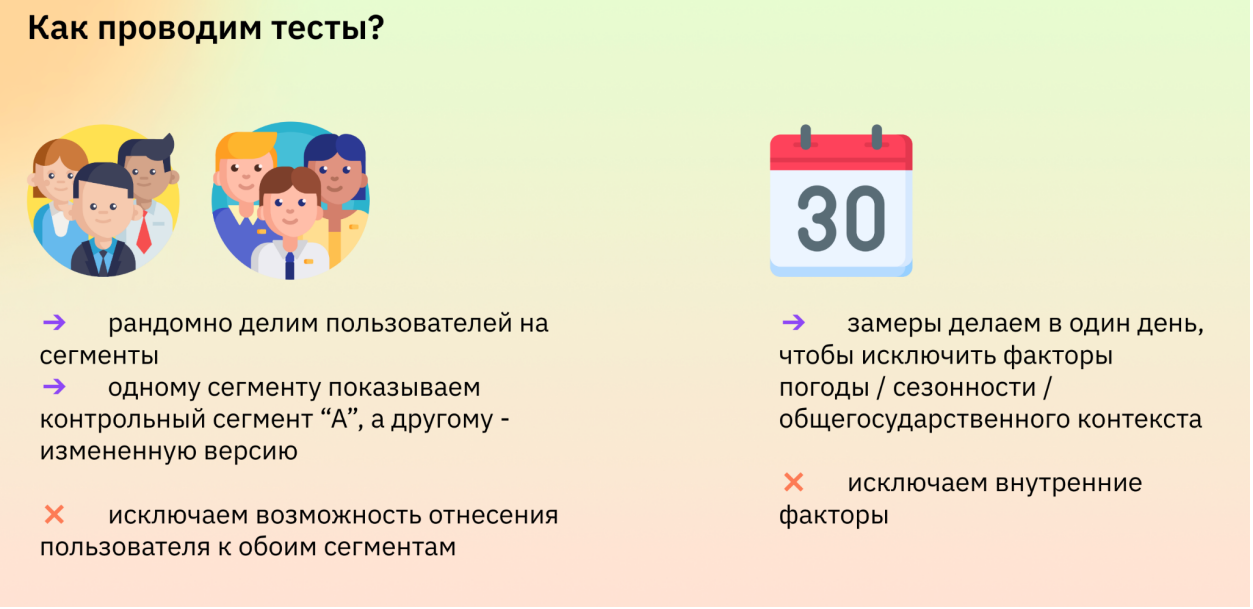
- цены,

- алгоритмы бэкенда,
- влияние любых изменений в продукте на интересующие нас метрики

Что можно подвергать A/B тестированию

- Сайт, приложение, e-mail
- Элементы: цвет, размер, расположение
- Формат текста: шрифт, размер шрифта, цвет шрифта
- Новый функционал
- Разные блоки информации
- Контент: разные текстовые формулировки, призывы к действию, заголовки и прочее

Алгоритм проведения теста



1. Пользователи рандомно делятся на сегменты
2. Одному сегменту показывают контрольный сегмент “А”, а другому - измененную версию

3. Должна быть исключена возможность отнесения пользователя к обоим сегментам
4. Замеры производятся в один день, для исключения факторов погоды / сезонности / общегосударственного контекста
5. Исключения влияния внутренних факторов

Кому и для чего нужны A/B тесты?

A/B тесты нужны всем — продакт-менеджерам, маркетологам, продуктовым дизайнерам,

всем, кто хочет улучшать свой продукт, делать его удобнее и приятнее для целевой аудитории.

Когда нужны A/B тесты

A/B тестирование - действительно оптимальный вариант получения нужных сведений для принятия какого-либо решения, когда:

- необходимо получить объективное мнение о качестве изменений
- достаточное количество пользователей и данных
- достаточно времени и ресурсов для дизайна и проведения теста

Какие бывают тесты?

А/Б тесты

Классический вариант

А/А тесты

Используются для исключения ложноположительного результата.

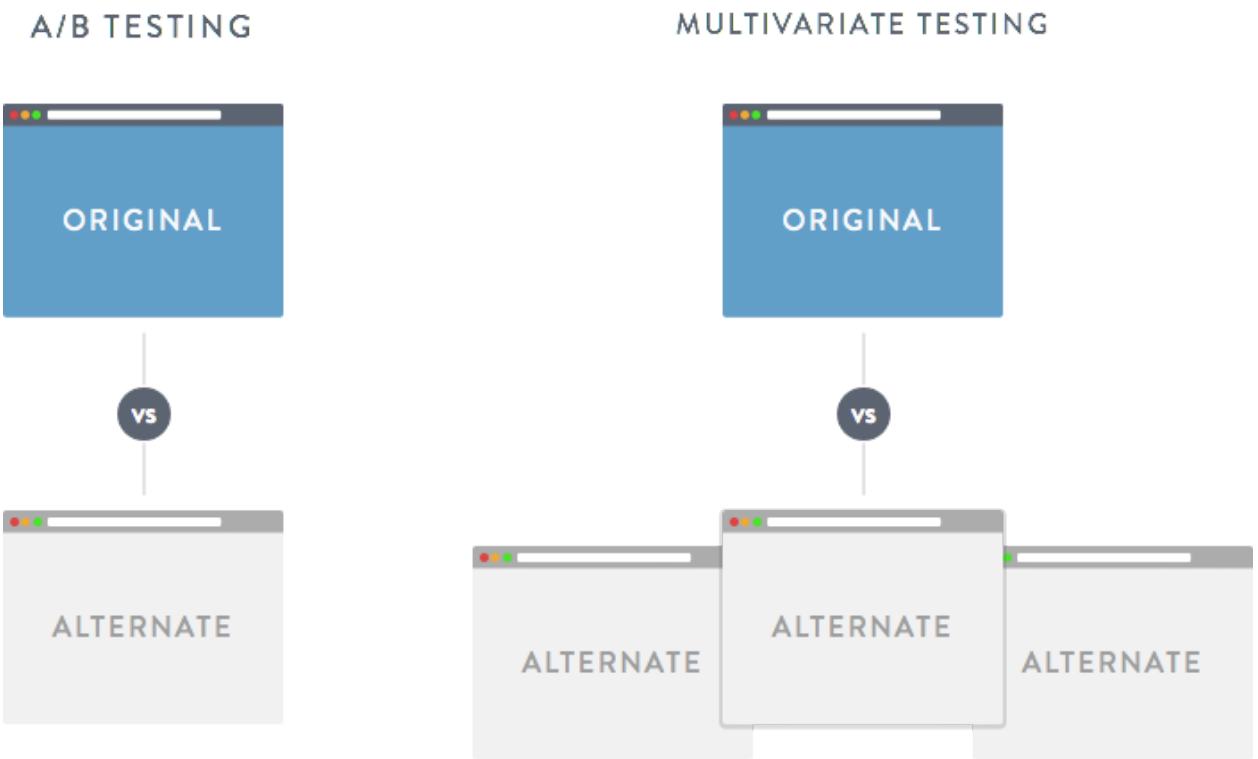
Используются для проверки сплитования, чтобы проверить репрезентативность выборки и избежать ошибки 1 рода (т.е. когда принимаем альтернативную гипотезу, а верна нулевая)

Мультивариативные

Если самим считать, то возникает family wise error rate и нужно вводить поправки уменьшения ошибки.

Если пользуемся готовым инструментом - то нужно иметь только большое кол-во трафика.

Мультивариантное тестирование (MVT)



Пошаговый план проведения А/Б тестирования

1. Формулировка гипотезы, определение метрик, необходимых для проведения тестирования
2. Создание двух альтернативных вариантов для тестирования (или более в случае мультивариатного тестирования)
3. Подбор части аудитории для эксперимента
4. Разделение выбранной аудитории на равные сегменты
5. Показ тестируемых вариантов выбранной аудитории
6. Анализ метрик, определение варианта-победителя
7. Внедрение варианта-победителя

💡 А/Б тесты не дадут вам заблудиться в развитии продукта!

Какие метрики проверяются А/Б тестами

Виды метрик

Типы метрик которые применяются в экспериментах:

1. Доли - (ретеншн, конверсии)
2. Непрерывные - (таймспент в секундах / деньги)
3. Отношения - (клики на сессию)

Конверсия: все этапы воронки

Экономические метрики

- Средний чек
- Объём выручки
- Прибыль
- LTV и т.д.

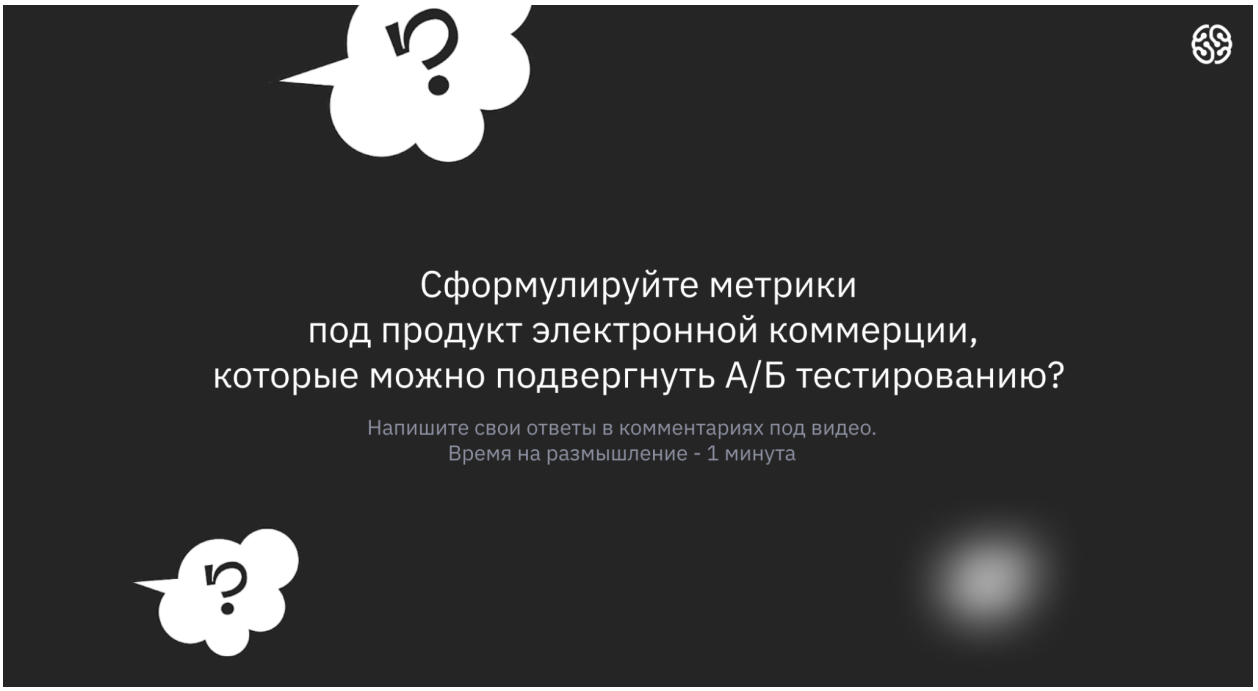
Поведенческие метрики

- Retention
- Отказы

- Глубина просмотра
- Время на сайте и т.д.

Задание на закрепление материала:

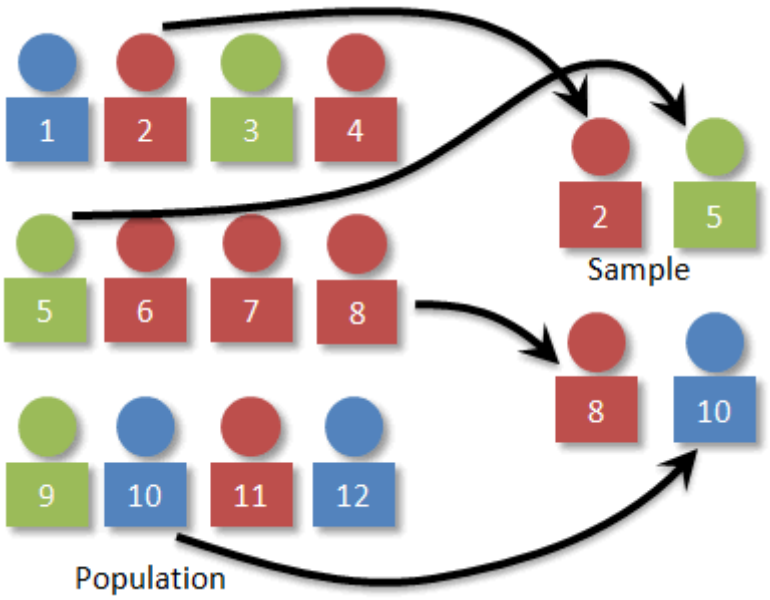
Сформулируйте метрики под продукт электронной коммерции, которые можно подвергнуть А/Б тестированию?



Ответ: LTV, конверсия в корзину, средний чек, повторные продажи

Способы формирования выборки

1) Простая случайная выборка – simple random sample



2) Стратифицированная выборка – stratified sample

3) Групповая выборка – cluster sample

Задание на закрепление материала

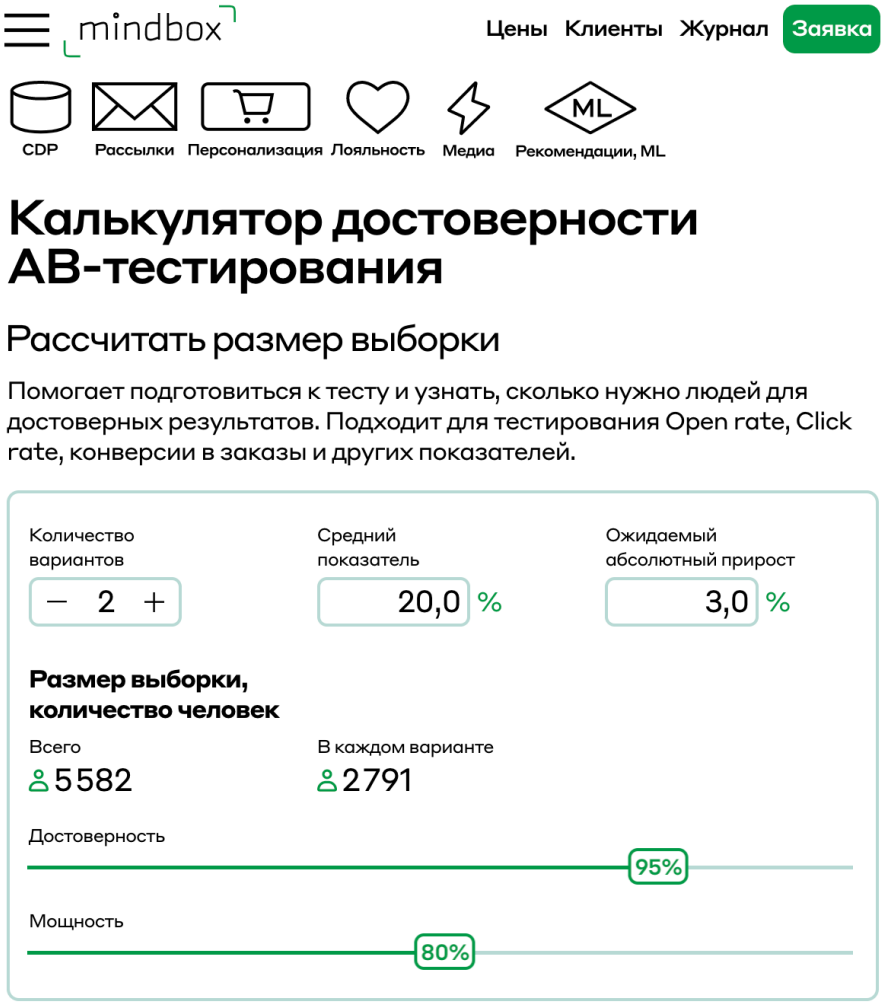
Соотнесите следующие примеры с тем, какой способ формирования выборки он описывает:

1. При тестировании вакцины, учёные разделили больных по тяжести заболевания, получили 4 группы и из каждой выбрали случайным образом по 10 человек. (Ответ: стратифицированная)
2. Для того, чтобы узнать, насколько понятен материал моим студентам, я написал 5-ти из них в Телеграм и узнал их мнение. (Ответ: простая)
3. Чтобы узнать средний уровень знания информатики среди школьников Москвы, было выбрано 10 школ, а из них случайным образом выбрано по 20 учеников. (Ответ: групповая)

Калькулятор и прочие инструменты

Инструменты

- Google optimize для веб
- Firebase для app
- Калькулятор достоверности A/B-тестирования — Mindbox (mindbox.ru)



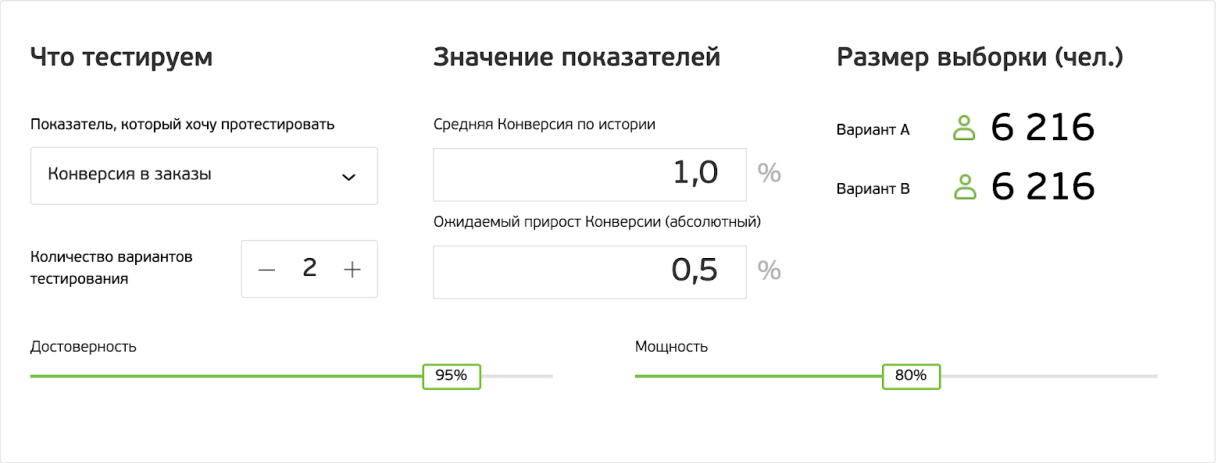
Описание терминов онлайн-калькулятора

- **Среднее значение** тестируемого показателя определяется, например, по историческим данным.
- **Ожидаемый прирост** — минимальный процент, на который планируете увеличить тестируемый показатель. Если он будет слишком маленьким — для теста понадобится много людей. Если будет слишком большим, а реальный прирост окажется меньше — значит не удалось добиться нужного роста и результаты теста не значимы.
- **Достоверность** — процент уверенности в том, что результаты теста верны, если он показал разницу.
- **Мощность** — процент уверенности в том, что результаты теста верны, если он не показал разницу. Если не знаете, какой процент указать, оставьте значения по умолчанию.

- **Размер выборки** показывает, сколько людей должны увидеть каждый вариант, чтобы можно было доверять результату теста. Помогает рассчитать время теста и не выключить его слишком рано или слишком поздно.

Определение численности выборки для доли

- Предположим, что вы предложили запустить A/B-тест, в котором в версии (B) будет другая форма подтверждения заказа. Нынешняя конверсия сайта = 1%.
- Мы предполагаем, что в новой версии конверсия вырастет до 1,5%.
- Сколько юзеров нужно отправить на каждую из версий?
- Воспользуемся калькулятором:



Для лучшего представления аудитории о работе калькулятора предполагается показать скринкаст - смотри лекцию к уроку!

Ошибки первого и второго рода:

Мощность критерия

При планировании эксперимента нужно помнить, что мощность должна быть разумно высокой, чтобы обнаружить корректные отклонения от нулевой гипотезы.

В противном случае, эксперимент проводить не следует. Обычно мы берём уровень мощности в пределах 80%-90%.

Доверительный интервал

Доверительный интервал — это способ оценки конверсии, в результате которого мы получаем не одно единственное значение, а интервал значений, внутри которого может содержаться реальное значение конверсии.

Проблемы при проведении А/Б тестов

Препятствия для проведения А/Б тестирования

- Недостаточный трафик
- Дорогая разработка
- Нет возможности обеспечить правильное распределение трафика
- Нет компетенций и опыта у команды по проведению тестов
- Слабые гипотезы
- Частые ошибки со стороны менеджеров и продактов: принимают решения интуитивно, не хотят тратить время.

Частые ошибки при проведении А/Б тестирования

- множественное сравнение - слишком много метрик
- слишком рано заканчиваем тест
- низкий уровень статистической значимости
- игнорирование внешних или внутренних факторов, которые могут влиять на тест
- неправильная настройка тестирования
- тестирование в период сезонного повышения или повышения спроса, что может искажать результаты эксперимента
- тестирование нескольких элементов одновременно, которые могут оказывать искажающее влияние друг на друга
- тестирование непроверенных гипотез вместо тех, которые были получены получили после анализа данных или проведения качественных исследований
- тестирование гипотез, с неочевидной или минимальной пользой для ключевых метрик.
- Не влюбляйтесь в свои идеи.
- Старайтесь объективно оценивать любые гипотезы и тесты.
- Проблема подглядывания (peeking problem) – слишком ранний сбор и оценка промежуточных результатов тестирования увеличивает шанс принять ошибку первого рода (false positive) пропорционально количеству таких «подкглядываний»
- Недостаточность данных для анализа. Если у вас недостаточно данных, чтобы рассматривать статистическую значимость «корневой метрики», то спускайтесь вниз по воронке, разбивайте её на отдельные шаги. Берите в расчёт только тех юзеров, которые увидели какое-либо изменение на вашем продукте (т.е. дошли до вашего теста).

Для отработки правильности настройки А/Б тестирования можно использовать метод A/A тестирования. A/A-тест — это тест, в котором сравнивается одна и та же версия.

Здесь могут учитывать технические проблемы с организацией теста

Пример проблем, которые помогает решить A/A тестирование:

- Не все сессии отправляются в систему аналитики
- Неправильно подсчитываются события или конверсии
- В тест попадают юзеры, которые не должны в него попасть

Пример неправильного выбора метрики:

Представит проведение теста новой системы рекомендаций для онлайн кинотеатра. В качестве проверочной метрики решено использовать конверсию в подписку. Однако эта метрика не позволит увидеть реальную картину изменения спроса. Необходимо заменить ее на метрики, связанные с вовлеченностью, залипанием и удержанием уже сконвертировавшихся клиентов.

Приоритизация гипотез

Чаще всего в продукте гипотез больше, чем возможно протестировать и образуется беклог.

Приоритизацию можно проводить по классическим методам определения приоритетов:

- Метод ICE
- Метод RICE
- Метод Value vs Cost
- Модель Кано

Дополнительные материалы

1. Что позволит генерировать идеи

2. Супер-презентация по метрикам на основе lean analytics
3. Таблица для сравнения 2-х версий продукта.
4. Видео более продвинутого уровня. Доверительный интервал за 15 мин. Биостатистика.
5. Статья “Интуитивное объяснение проверки гипотез и p-значение”
6. Статья “Почему ученые против статистической значимости”
7. Калькулятор достоверности A/B-тестирования (вкладка "Итоги тестирования")
8. Калькулятор доверительного интервала