# Прикладная статистика и анализ данных

## A/B тестирование

Апрель 2020

# Краткое содержание

# Напоминание: проверка гипотез

1. Основная и альтернативная гипотезы
   $H_0, H_1$
2. Статистика критерия
   $T(X)$
3. p-value
   $p(x)$, где $x$ — реализация выборки
4. Уровень значимости
   $P_{H_0}(p(X) \geqslant \alpha) \geqslant \alpha$
5. Мощность
   $P_{H_1}(p(X) \geqslant \alpha) \longrightarrow \max$

# Вопрос-ответ

Вопрос с собеседований и от менеджеров:
как связаны pvalue и вероятность нулевой гипотезы?

# Вопрос-ответ

Вопрос с собеседований и от менеджеров:
как связаны pvalue и вероятность нулевой гипотезы?

**Hint:** представьте, что гипотеза случайна.

# Вопрос-ответ

Вопрос с собеседований и от менеджеров:
как связаны pvalue и вероятность нулевой гипотезы?

**Hint:** представьте, что гипотеза случайна.

Нестрого формулой Байеса:

$$\mathrm{P}(\mathrm{H}_0 \mid T \geqslant t) = \mathrm{P}(T \geqslant t \mid \mathrm{H}_0)\frac{\mathrm{P}(\mathrm{H}_0)}{\mathrm{P}(T \geqslant t)} \sim \mathit{pvalue} \cdot \mathrm{P}(\mathrm{H}_0)$$

pvalue нельзя интерпретировать как вероятность нулевой гипотезы! Надо об этом помнить и напоминать заказчикам при презентации результатов.

# Основные тесты

Расскажите какие критерии вам известны?

# Бернуллиевские выборки

1. **Независимые выборки**
   $X_1, ..., X_n \sim Bern(p)$ и $Y_1, ..., Y_m \sim Bern(q)$.
   $H_0 : p = q$ *vs.* $H_1 : p <\neq> q$
   $Z(X, Y) = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{\sigma}}$, где $\widehat{\sigma} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$.

2. **Связные выборки**
   $X_1, ..., X_n \sim Bern(p)$ и $Y_1, ..., Y_n \sim Bern(q)$.
   $H_0 : p = q$ *vs.* $H_1 : p \{<, \neq, >\} q$

| | $Y_i = 1$ | $Y_i = 0$ |
|---|---|---|
| $X_i = 1$ | $e$ | $f$ |
| $X_i = 0$ | $g$ | $h$ |

$$Z(X, Y) = \frac{\widehat{p} - \widehat{q}}{\sqrt{\frac{f+g}{n^2} + \frac{(f-g)^2}{n^3}}}$$

# Независимые нормальные выборки

$X_1, ..., X_n \sim \mathcal{N}(a_1, \sigma_1^2)$, $Y_1, ..., Y_m \sim \mathcal{N}(a_2, \sigma_2^2)$.

$H_0: a_1 = a_2$ *vs.* $H_1: a_1 \ <\neq> \ a_2$

1. $\sigma_1$ **и** $\sigma_2$ **известны**

$$Z(X,Y) = \frac{\overline{X}-\overline{Y}}{\sqrt{\sigma_1^2/n+\sigma_2^2/m}} \overset{H_0}{\sim} \mathcal{N}(0,1)$$

2. $\sigma_1 = \sigma_2 = \sigma$ **неизвестны**

$$T(X,Y) = \frac{\overline{X}-\overline{Y}}{S_{tot}\sqrt{1/n+1/m}} \overset{H_0}{\sim} T_{n+m-2}$$

где $S_{tot}^2 = \frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$

3. $\sigma_1 \neq \sigma_2$ **и неизвестны**

$$T(X,Y) = \frac{\overline{X}-\overline{Y}}{\sqrt{S_X^2/n+S_Y^2/m}} \overset{H_0}{\underset{\text{прибл.}}{\sim}} T_\nu$$

$$\nu = \left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2 \Bigg/ \left(\frac{S_X^4}{n^2(n-1)}+\frac{S_Y^4}{m^2(m-1)}\right)$$

# Непараметрические критерии

1. **Критерий Смирнова**

   $X_1, ..., X_n$ и $Y_1, ..., Y_m$ — выборки с функ. распр. $F$ и $G$.

   $\mathrm{H}_0 : F = G \quad vs. \quad \mathrm{H}_1 : F \neq G$

   Статистика $D_{nm} = \sup\limits_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - \widehat{G}_m(x) \right|$

2. **Критерий Мана-Уитни**

   $X_1, ..., X_n$ и $Y_1, ..., Y_m$ — выборки с функ. распр. $F$ и $G$.

   $\mathrm{H}_0 : F = G \quad vs. \quad \mathrm{H}_1 : F \geqslant G$

   $S_j$ — ранг $Y_j$ в вар. ряду по выборке $(X_1, .., X_n, Y_1, .., Y_m)$.

   $V = S_1 + ... + S_m$ — статистика критерия.

$$\frac{V - \mathrm{E}V}{\sqrt{\mathrm{D}V}} \xrightarrow{d_0} \mathcal{N}(0, 1),$$

где $\mathrm{E}V = \frac{m(n+m+1)}{2}$, $\mathrm{D}V = \frac{nm(n+m+1)}{12}$ при $\mathrm{H}_0$.

# Пример AB теста

**Горные велосипеды. Распродажа 60%**
В наличии более 1 000 моделей!
bike.ru

**Распродажа! Sale! Rebajas! Saldi!**
Не пропустите! До 1 числа продаем горные велосипеды по смешным ценам! :)
bike.ru

**Офисные кресла**
Скидка от 10 кресел - 15%! Скидка в шоу-рум до 50%! Доставка – бесплатно!!!

**Офисные кресла**
Скидка от 10 кресел - 15%! Скидка в шоу-рум до 50%! Доставка – бесплатно!!!

# Практические вопросы

Давайте разберём несколько практических моментов, с которыми каждый, кто проводит АБ тесты, рано или поздно сталкивается.

# Разбиение на тестовые группы

Нужно разбить пользователей на две группы
Как лучше это сделать?
Как проверить, что полученное разбиение — хорошее?

# Разбиение на тестовые группы

Нужно разбить пользователей на две группы
Как лучше это сделать?
Как проверить, что полученное разбиение — хорошее?

Для разбиения обычно используют хэш от id-шника пользователя конкатенированного со строкой общей для эксперимента (солью).
Как вы думаете, зачем соль?

# Разбиение на тестовые группы

Првостепенная задача - проверить корректность разбиения.
Что можно сделать:

- Сравнить статические фичи (пол, возраст и т.п.) распределены одинаково — Критерии однородности и др
- Сравнить исторические фичи (конверсии за какой-то период, покупки и и.д.). Распределения врядли будут прям совпадать, но стоит проверить разные статистики (среднее, медианы, дисперсии)
- Провести AA-тест. Разбиение всё равно может оказаться плохим, поэтому стоит провести AA тест, в рамках которого убедиться, что в группах нет значимых различий между целевыми метриками.

Кстати, похожий процесс происходит, когда придумывают новую метрику. Ведь нужно убедиться, что она корректно красится (ещё одно сленговое слово) в AA тестах.

# Сроки теста

Никто не будет запускать тест просто так.
Нужно заранее оценить на какой срок нужно запускать тест.
Как это сделать?

# Сроки теста

Никто не будет запускать тест просто так.
Нужно заранее оценить на какой срок нужно запускать тест.
Как это сделать? Часто встречаются следующие две стратегии:

- Из соображений мощности: нужно взять такое $n$, чтобы $\mathrm{P}_{H_1}(p(X) \leqslant \alpha) \geqslant \beta$
- Из соображений эффекта: нужно взять такое $n$, чтобы $p(X) \leqslant \alpha$ при условии, что есть эффект. Например, чтобы в Z-тесте *pvalue* $\leqslant 0.05$ при условии, что $\widehat{p}_2 - \widehat{p}_1 > 0.01$

Как вы думаете, какая оценка даёт меньшие сроки?

# Досрочная остановка теста

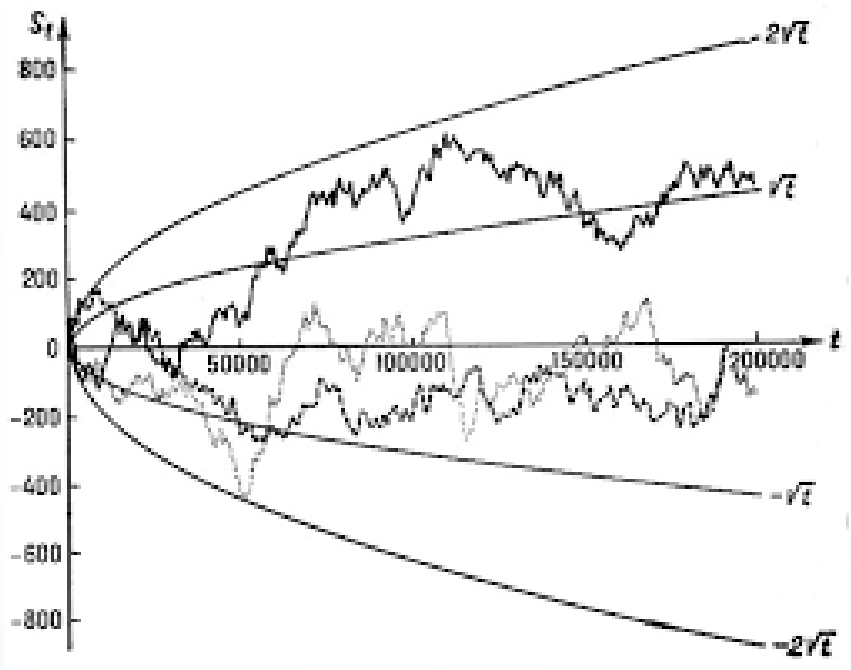
Вы распланировали АБ тест.

По вашим оценкам (вы использовали биномиальный тест) за 81 день отклонение изменяемой величины на 1 процент является значимым.

Вы ежедневно мониторили результаты теста и через 9 дней обнаружили отклонение в 5 процентов, что является значимым для теста длиной 9 дней.

Можно в этом случае досрочно завершить АБ тест?

# Последовательный анализ




**Как с этим жить**: Нужно применять Статистический последовательный анализ (Sequential analysis). Он даёт, во-первых, корректные, а во-вторых, более мощные критерии (точнее более быстрые прокраски), пользуясь дополнительным знанием о потоковости данных.

# Поиск разных эффектов

Допустим, что вы проверяете средний чек, среднее число товаров в чеке, среднее число аксессуаров в чеке.

Для каждой из этих величин вы составили свой критерий для проверки гипотезы о наличие эффекта.

Каков уровень значимости для такой одновременной проверки гипотез?

# Множественная проверка гипотез

Поскольку величина чека, число товаров в нём и число аксессуаров в нём — зависимые величины, то нельзя в точности найти уровень значимости, но можно его оценить:

$$\alpha \leqslant \mathsf{P}_{\mathsf{H}_0}(p_1 \leqslant \alpha \text{ or } p_2 \leqslant \alpha \text{ or } p_3 \leqslant \alpha) \leqslant \sum_j \mathsf{P}_{\mathsf{H}_0}(p_j \leqslant \alpha) = 3\alpha$$

# Множественная проверка гипотез

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

**Как с этим жить**: Нужно применять методы Множественной проверки гипотез. Самый простой способ — уменьшить $\alpha$ в число гипотез раз, но есть и более сложные подходы. Есть хорошая реализация в Python — statsmodels.sandbox.stats.multicomp.multipletests.

Какие подходы вы знаете?

# Пример реального теста

- Вы сравниваете конверсию пользовательской сессии в приложении в клик по рекламному блоку.
- Тест длится около месяца.
- Вы для всех сессий имеет величину 0 или 1 — сконвертировалась ли она в клик.
- Значимость оценивается с помощью Z-теста на выборках из вышеописанных величин

Оцените насколько хороша такая схема теста.

# Пример реального теста

- Вы сравниваете конверсию пользовательской сессии в приложении в клик по рекламному блоку.
- Тест длится около месяца.
- Вы для всех сессий имеет величину 0 или 1 — сконвертировалась ли она в клик.
- Значимость оценивается с помощью Z-теста на выборках из вышеописанных величин

Оцените насколько хороша такая схема теста.

Объекты выборки зависимы. Из-за этого вы обычно будете занижать pvalue!

На практике эта ситуация сплошь и рядом, поэтому давайте разберёмся как решить эту проблему.

# Постановка задачи

## Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^{n}$, где $k_i$ – ключ $i$-го события, $x_i$ – произвольная статистика события (может несколько чисел). События **не независимы**, независимость есть только между событиями **разных ключей**.

# Постановка задачи

## Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^{n}$, где $k_i$ – ключ $i$-го события, $x_i$ – произвольная статистика события (может несколько чисел). События **не независимы**, независимость есть только между событиями **разных ключей**.

## Примеры

1. Конверсия сессии в поездку, ключ – id пользователя
2. Стоимость заказа, ключ – id пользователя
3. Конверсия предложения заказа в его принятие, ключ – id водителя
4. Количество заказов в день, ключ – id водителя

# Постановка задачи

## Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^{n}$, где $k_i$ – ключ $i$-го события, $x_i$ – произвольная статистика события (может несколько чисел). События **не независимы**, независимость есть только между событиями **разных ключей**.

## Метод обсчёта

Функция $f$, которая для произвольного набора $\{(k_j, x_j)\}_{j=1}^{m}$ находит значение статистики

# Постановка задачи

## Данные

Две "выборки" вида $\{(k_i, x_i)\}_{i=1}^{n}$, где $k_i$ – ключ $i$-го события, $x_i$ – произвольная статистика события (может несколько чисел). События **не независимы**, независимость есть только между событиями **разных ключей**.

## Метод обсчёта

Функция $f$, которая для произвольного набора $\{(k_j, x_j)\}_{j=1}^{m}$ находит значение статистики

## Задача

Ответить на вопрос стат. значимо ли $f(\{(k_i^{test}, x_i^{test})\}_{i=1}^{n_{test}})$ отличается от $f(\{(k_i^{control}, x_i^{control})\}_{i=1}^{n_{control}})$

# Возможные решения

## Задача

Ответить на вопрос стат. значимо ли $f(\{(k_i^{test}, x_i^{test})\}_{i=1}^{n_{test}})$ отличается от $f(\{(k_i^{control}, x_i^{control})\}_{i=1}^{n_{control}})$

# Возможные решения

## Задача

Ответить на вопрос стат. значимо ли $f\left(\left\{\left(k_i^{test}, x_i^{test}\right)\right\}_{i=1}^{n_{test}}\right)$ отличается от $f\left(\left\{\left(k_i^{control}, x_i^{control}\right)\right\}_{i=1}^{n_{control}}\right)$

## Простой случай

Все ключи **различные**, $f$ – среднее, медиана, гистограмма.

Тогда можно использовать обычные тесты для соответсвующей нулевой гипотезы

# Бакетное сэмплирование

## Описание

Случайно разобьём события на бакеты (от 1 до $B$) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию $f$, получив $B$ чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

# Бакетное сэмплирование

## Описание

Случайно разобьём события на бакеты (от 1 до $B$) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию $f$, получив $B$ чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

## Основные вопросы

1. Почему в итоге получается выборка?
2. Почему применение теста корректно?
3. Почему мы не потеряем в мощности теста?

# Бакетное сэмплирование

## Описание

Случайно разобьём события на бакеты (от 1 до $B$) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию $f$, получив $B$ чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

## Выборочность. Нестрого

Все зависимые события (события одного ключа) попадают в один бакет. То есть все события разных бакетов независимы.

# Бакетное сэмплирование

## Описание

Случайно разобьём события на бакеты (от 1 до $B$) при помощи некоторой хэш-функции от ключа. Для каждого бакета применим функцию $f$, получив $B$ чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

## Корректность. Нестрого

Если эффекта нет и данных дотаточно много, то $f(bucket)$ будет распределено одинаково в тесте и контроле. Кстати, поэтому требуется, чтобы размер теста и контроля совпадали (ttest-у без разницы, а вот mw страдает).

# Бакетное сэмплирование

## Описание

Случайно разобьём события на бакеты (от 1 до $B$) при помощи какой-то функции от ключа. Для каждого бакета применим функцию $f$, получив $B$ чисел. Теперь у нас две выборки (уже без кавычек). К ним можно применить произвольный статистический тест.

## Мощность. Нестрого

Если все ключи уникальны, то в результе бакетирования количество объектов выборки уменьшается в $\frac{n}{B}$ раз. Но дисперсия значений выборки тоже уменьшется в $\frac{n}{B}$ раз. Эти два эффекта компенсируют друг друга и дают надежду на мощность теста.

# Бакетное сэмплирование на примере

Пусть $k_i = i$, $x_i \sim Bern(p)$, то есть $x_i$ — независимые конверсии. Метод обсчёта — это среднее, а поверх бакетов применяется ttest.

В ttest-е статистика это $\frac{\overline{x}_{test} - \overline{x}_{control}}{\sqrt{\frac{\overline{s^2}_{test}}{n_{test}} + \frac{\overline{s^2}_{control}}{n_{control}}}}$.

$\overline{x}$ из-за бакетов не сильно изменится, так как размеры бакетов примерно одинаковые и среднее средних будет почти совпадать с исходным средним.

$\overline{s^2}$ из-за усреднения бакетов уменьшится примерно в $\frac{n}{B}$ раз. Но и размер выборки будет не $n$, а $B$, то есть знаменатель тоже уменьшится в $\frac{n}{B}$ раз. Получается, что статистика изменится не сильно.

# В сухом остатке

1. Бакетное сэмплирование позволяет быстро сравнивать почти произвольные метрики на выборках больших размеров.
2. Но нужно быть аккуратным с размером групп и с метриками (не везде так хорошо как со средними). Требуется проводить AA тесты.
3. Если у вас небольшая выборка, то лучше использовать классические тесты
4. Практически все метрики, которые используются, можно представить в виде $\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$ (средние, конверсии, пороги) и поэтому бакетный тест можно эффективно реализовать на MapReduce