

# **Множественное тестирование**

# История о зомби-лососе

- В 2012 году ряд авторов получил Шнобелевскую премию по нейробиологии
- Надо было протестировать аппарат МРТ
- Для этого обычно в него кладут шарик с маслом и сканируют его



- <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- <https://habr.com/ru/company/ods/blog/325416/>

# История о зомби-лососе

- Это скучно, поэтому авторы решили купить на рынке мёртвого лосося и просканировать его мозг
- Лососю показывали фотографии людей и проверяли, есть ли у него в мозгу активность
- Оказалось, что активность есть



- ▶ <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- ▶ <https://habr.com/ru/company/ods/blog/325416/>

# История о зомби-лососе

- Аппарат МРТ возвращает много данных
- Чтобы убедиться, что в мозгу нет реакции, надо проверить много гипотез об отсутствии активности на каждом маленьком участке мозга

**Проблема множественного тестирования:**  
если мы проверяем несколько гипотез подряд,  
уровень значимости выходит из-под контроля

Мы начинаем чаще отвергать верные гипотезы,  
чем нам хотелось бы

- <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>
- <https://habr.com/ru/company/ods/blog/325416/>

# Множественная проверка гипотез

Проверяем две гипотезы:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Каждую на уровне значимости  $\alpha$

Можно ошибиться сразу в двух местах:

$\mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$

$$= 1 - \mathbb{P}(\text{не ошибиться ни в одной}) = 1 - (1 - \alpha)^2$$

$$= 1 - (1 - 2\alpha + \alpha^2) = 2\alpha - \alpha^2 > \alpha$$

$$\alpha_i = 0.05 \Rightarrow \alpha = 0.1 - 0.025 = 0.075 > 0.5$$

❗ Вероятность ошибки первого рода накапливается и выходит из-под контроля

# Множественная проверка гипотез

**Пример:** показ на странице сервиса нескольких новых элементов

- Изменения взаимосвязаны и их можно протестировать только на одном временном промежутке
- В такой ситуации мы сталкиваемся с множественным тестированием
- С ростом числа гипотез, вероятность получить ошибку растёт экспоненциально:  $1 - (1 - \alpha)^n$



Нужно взять уровень значимости под контроль

# Неравенство Бонферрони

- Нужно как-то скорректировать исходный уровень значимости, в этом помогает неравенство Бонферрони:

$$\mathbb{P}(A + B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

- То есть каждую гипотезу из двух надо проверять на уровне значимости  $\frac{\alpha}{2}$

$\alpha = \mathbb{P}(\text{ошибочно отвергнуть хотя бы одну из } H_0)$

$$\leq \mathbb{P}(\text{ош. в 1}) + \mathbb{P}(\text{ош. во 2}) = \frac{\alpha_i}{2} + \frac{\alpha_i}{2} = \alpha_i$$

- Если гипотез  $k$ , берём уровень значимости  $\frac{\alpha}{k}$  для каждой

# Неравенство Бонферрони

- Из-за коррекции уровня значимости возникают проблемы с мощностью тестов
- Чем больше гипотез проверяется, тем ниже шансы отклонить неверные гипотезы
- Более того, из-за презумпции нулевой гипотезы для более низкого уровня значимости нам нужно собрать большее число наблюдений, чтобы зафиксировать значимое отклонение от нулевой гипотезы

⇒ процедуру надо улучшить,  
чтобы мощность стала выше

# Матрица ошибок

Рассмотрим случай, когда мы проверяем  $n$  гипотез

	верных $H_{0i}$	неверных $H_{0i}$
не отвергнутых $H_{0i}$	$U$	$T$
отвергнутых $H_{0i}$	$V$	$S$

- Неверно отклонили  $V$  гипотез, неверно не отклонили  $T$  гипотез
- На практике пытаются контролировать обобщения ошибки первого рода, например: FWER и FDR

# Family-Wise Error Rate (FWER)

Рассмотрим случай, когда мы проверяем  $n$  гипотез

	верных $H_{0i}$	неверных $H_{0i}$
не отвергнутых $H_{0i}$	$U$	$T$
отвергнутых $H_{0i}$	$V$	$S$

**Групповая вероятность ошибки, FWER (Family-Wise Error Rate)**

– это вероятность совершить хотя бы одну ошибку первого рода

$$FWER = \mathbb{P}(V > 0)$$

# False Discovery Rate (FDR)

Рассмотрим случай, когда мы проверяем  $n$  гипотез

	верных $H_{0i}$	неверных $H_{0i}$
не отвергнутых $H_{0i}$	$U$	$T$
отвергнутых $H_{0i}$	$V$	$S$

**Ожидаемая доля ложны отклонения, FDR (False Discovery Rate)** – это математическое ожидание числа ошибок первого рода к общему числу отклонений нулевой гипотезы

$$FDR = \mathbb{E} \left( \frac{V}{V + S} \right)$$

# Метод Холма

- Поправка Бонферрони пытается контролировать FWER (вероятность хотя бы одной ошибки 1 рода)
- **Бонферрони:** проверяем  $k$  гипотез на уровнях значимости

$$\alpha_1 = \alpha_2 = \cdots = \alpha_k = \frac{\alpha}{k}$$

- **Метод Холма** – улучшение поправки Бонферрони, обладает более высокой мощностью
- Проверяем  $k$  гипотез, но уровни значимости пытаемся выбирать разными

# Метод Холма

- Отсортируем гипотезы по получившимся  $P$ -значениям по возрастанию:  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$
- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{\alpha}{k-1}, \dots, \alpha_{(i)} = \frac{\alpha}{k-i+1}, \dots, \alpha_{(k)} = \alpha$$

- Если  $p_{(1)} \geq \alpha_{(1)}$ , все нулевые гипотезы не отвергаются, иначе отвергаем первую и продолжаем
- Если  $p_{(2)} \geq \alpha_{(2)}$ , все оставшиеся нулевые гипотезы не отвергаются, иначе отвергаем вторую и продолжаем
- Идём, пока не кончатся гипотезы

## Метод Холма

- Метод Холма обеспечивает контроль  $FWER$  на уровне  $\alpha$
- Метод Холма оказывается мощнее корректировки Бонферрони, так как его уровни значимости меньше

# Метод Бенджамини-Хохберга

- Отсортируем гипотезы по получившимся  $P$ -значениям по возрастанию:  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(k)}$
- Возьмём для них

$$\alpha_{(1)} = \frac{\alpha}{k}, \alpha_{(2)} = \frac{2\alpha}{k}, \dots, \alpha_{(i)} = \frac{i\alpha}{k}, \dots, \alpha_{(k)} = \alpha$$

- Если  $p_{(k)} < \alpha_{(k)}$ , отвергнуть все гипотезы, иначе не отвергнуть  $k$  – ую и продолжить
- Если  $p_{(k-1)} < \alpha_{(k-1)}$ , отвергнуть все оставшиеся гипотезы, иначе не отвергнуть  $(k - 1)$  – ую и продолжать
- Идём, пока не кончатся гипотезы

# Метод Бенджамини-Хохберга

- Для любой процедуры множественного тестирования гипотез  $FDR \leq FWER$
- Метод Бенджамини-Хохберга обычно оказывается более мощным, чем методы контролирующие  $FWER$
- Он отвергает не меньше гипотез с теми же  $\alpha_i$
- Это происходит за счёт того, что метод позволяет допустить большее число ошибок первого рода

# Специальные тесты

Альтернатива для процедур множественного тестирования – разработка специальных тестов, которые проверяют гипотезы сразу о нескольких ограничениях

## Примеры:

- Тест отношения правдоподобий (обсудим позже)
- ANOVA – равенство сразу же нескольких математических ожиданий
- Тест Бартлетта – равенство нескольких дисперсий

# Резюме

- Если сделать поправку, мёртвый лосось остаётся мёртвым
- До 2010 около 40% статей по нейробиологии не использовали поправки при множественном тестировании гипотез
- Благодаря работе о лососе и Шнобелевской премии за неё удалось уменьшить число таких статей до 10%
- Корректировка уровня значимости помогает держать под контролем ложно-положительные результаты, это приводит к росту ложно-отрицательных результатов

**Бутстрап**

# Бутстррап

- Не для всех описательных статистик можно найти распределение в аналитическом виде (медиана, эксцесс, куртосис)
- Бутстррап помогает решить эту проблему

# Бутстррап

- **Идея метода:** имеющаяся выборка – это единственная информация об истинном распределении данных
- Давайте приблизим истинное распределение эмпирическим, то есть “сами себя вытащим”
- Предполагается, что бутстррап-распределение окажется похожим на реальное распределение

# Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики  $\bar{x}$

Выборки с повторениями:

$$1,4 \Rightarrow \bar{x} = 2.5$$

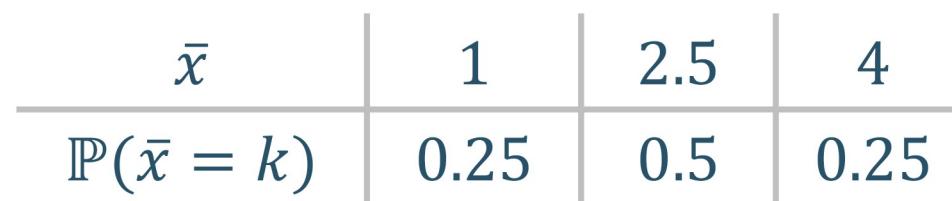
$$4,1 \Rightarrow \bar{x} = 2.5$$

$$1,1 \Rightarrow \bar{x} = 1$$

$$4,4 \Rightarrow \bar{x} = 4$$

Всего вариантов выборок:

$$n^n = 2^2 = 4$$



# Пример

- У Саши есть выборка из двух наблюдений: 1 и 4
- Нужно построить по этой выборке бутстррап-распределение для статистики  $\bar{x}$

Всего вариантов выборок:

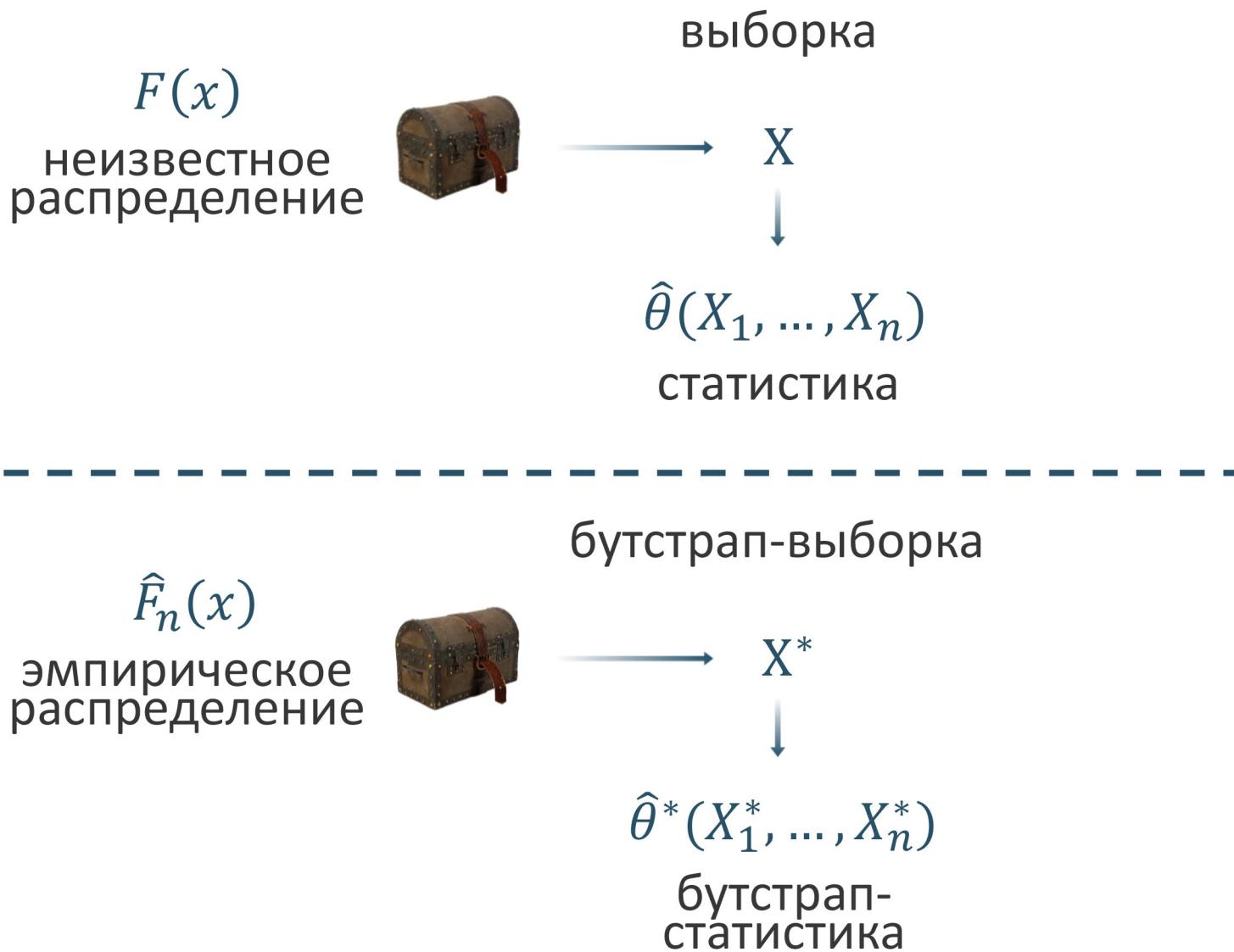
$$n^n = 2^2 = 4$$

- Строить такие распределения для больших выборок дорого
- Задача слишком сложная, а точность излишняя
- **Выход:** симуляции

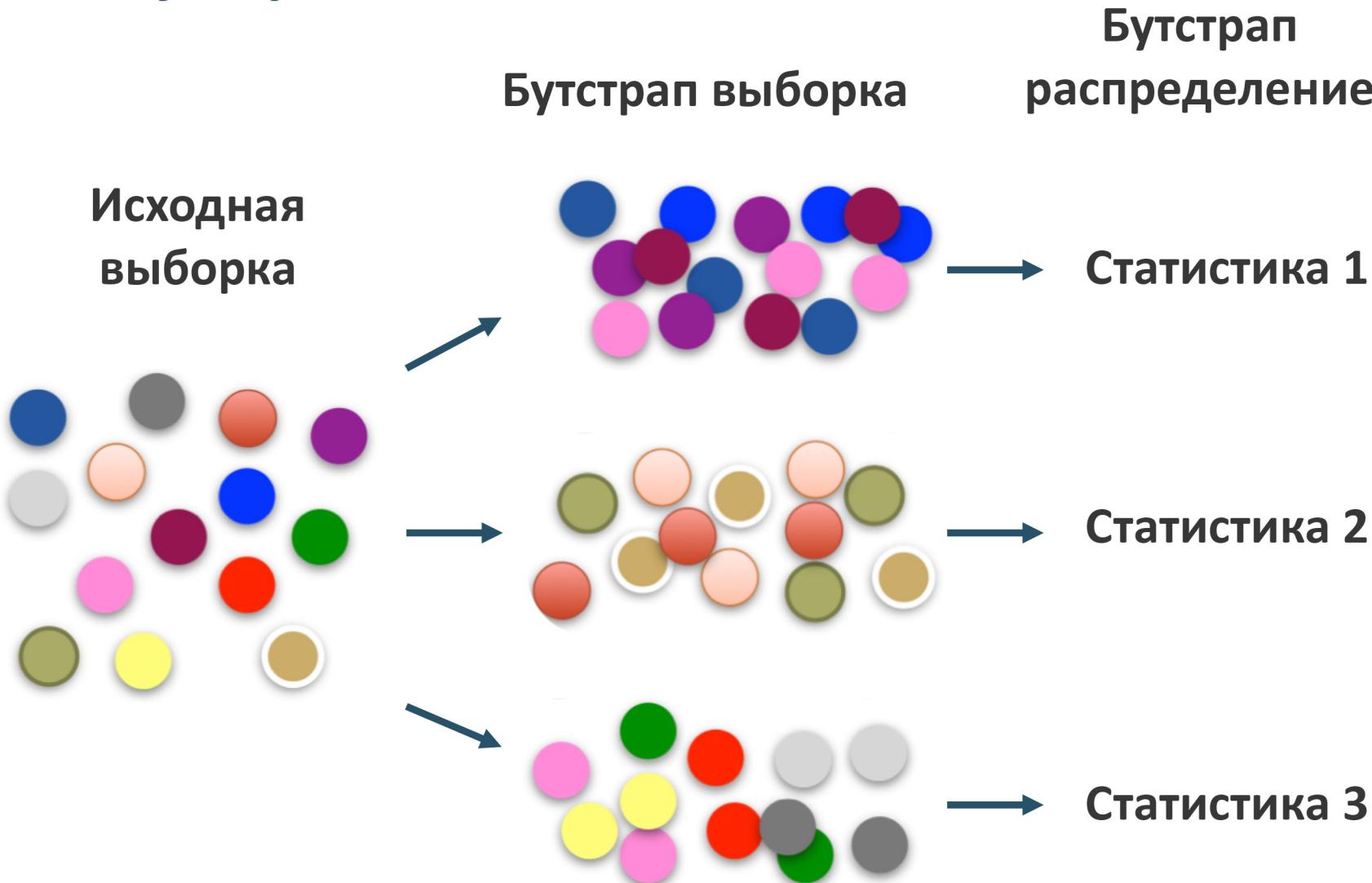
## Схема бутстрата

- Извлечение выборок из генеральной совокупности – сэмплирование из неизвестного распределения  $F(x)$
- У нас есть оценка для  $F(x)$  –  $\hat{F}_n(x)$
- Сэмплировать из такого распределения – то же самое, что брать выборки с повторениями объёма  $n$

# Схема бутстрата



# Схема бутстрата



# Схема бутстрата

- Генерируем  $B$  “псевдовыборок” с повторениями объёма  $n$  и оцениваем настоящее распределение “псевдоэмпирическим”
- По каждой выборке вычисляем интересующую нас статистику, получаем для неё бутстрап-распределение
- **Эфронов доверительный интервал:** находим выборочные квантили бутстрап-распределения
- В таком доверительном интервале возникает смещение  $\Rightarrow$  более сложные техники бутстрата

# Доверительный интервал Эфрана

Бутстранируем: оценку неизвестного параметра

Сэмплируем:  $x_1^*, \dots, x_n^*$

Считаем:  $\hat{\theta}^*$

Повторяем:  $B$  раз

Строим распределение:  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Интервал:  $[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*]$

# Доверительный интервал Эфрана

Бутстранируем: оценку неизвестного параметра

Сэмплируем:  $x_1^*, \dots, x_n^*$

Считаем:  $\hat{\theta}^*$

Повторяем:  $B$  раз

Строим распределение:  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Интервал:  $[\hat{\theta}_{\frac{\alpha}{2}}^*; \hat{\theta}_{1-\frac{\alpha}{2}}^*]$

- ! Если распределение выборки несимметрично, такой доверительный интервал усиливает смещение, присущее изначальной выборке

# Центрирование

- Чтобы не создавать искусственное смещение между  $\hat{\theta}$  и  $\theta$ , нужно бутстрартировать центрированную статистику
- Хотим бутстрартировать разность  $\hat{\theta} - \theta$
- Бутстрарпируя  $\hat{\theta}$ , мы подсчитываем её каждый раз заново, но на бутстрарповских выборках,  $\hat{\theta}^*$
- Бутстрарповским аналогом для  $\theta$  будет  $\hat{\theta}$ , так как мы сэмплируем данные их эмпирического распределения
- Бутстрарповский аналог для  $\hat{\theta} - \theta$  – это  $\hat{\theta}^* - \hat{\theta}$

# Доверительный интервал Холла

**Бутстрапируем:** Отклонение оценки от истинного значения

**Сэмплируем:**  $x_1^*, \dots, x_n^*$

**Считаем:**  $\hat{q}_i^* = \hat{\theta}^* - \hat{\theta}$

**Повторяем:**  $B$  раз

**Строим распределение:**  $\hat{q}_1^*, \dots, \hat{q}_B^*$

**Интервал:**  $[\hat{\theta} - \hat{q}_{1-\frac{\alpha}{2}}^*; \hat{\theta} - \hat{q}_{\frac{\alpha}{2}}^*]$

# t-процентильный доверительный интервал

Бутстранируем: t - статистику

Сэмплируем:  $x_1^*, \dots, x_n^*$

Считаем:  $t_i^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$

Повторяем:  $B$  раз

Строим распределение:  $t_1^*, \dots, t_B^*$

Интервал:  $[\hat{\theta} - t_{1-\frac{\alpha}{2}}^* \cdot se(\hat{\theta}); \hat{\theta} + t_{\frac{\alpha}{2}}^* \cdot se(\hat{\theta})]$

❗ Если бутстрарировать  $|\hat{\theta}^* - \hat{\theta}|$ , можно  
получить симметричный интервал

# Проверка гипотез при помощи бутстрата

$$H_0: \theta = \theta_0$$

$$H_a: \theta > \theta_0$$

t-статистика:

$$t = \frac{(\hat{\theta} - \theta_0)}{se(\hat{\theta})}$$

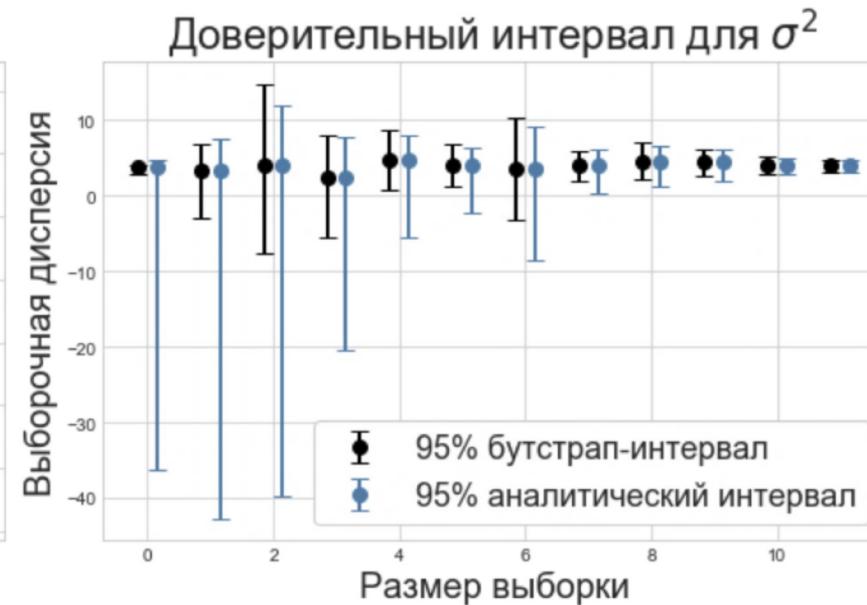
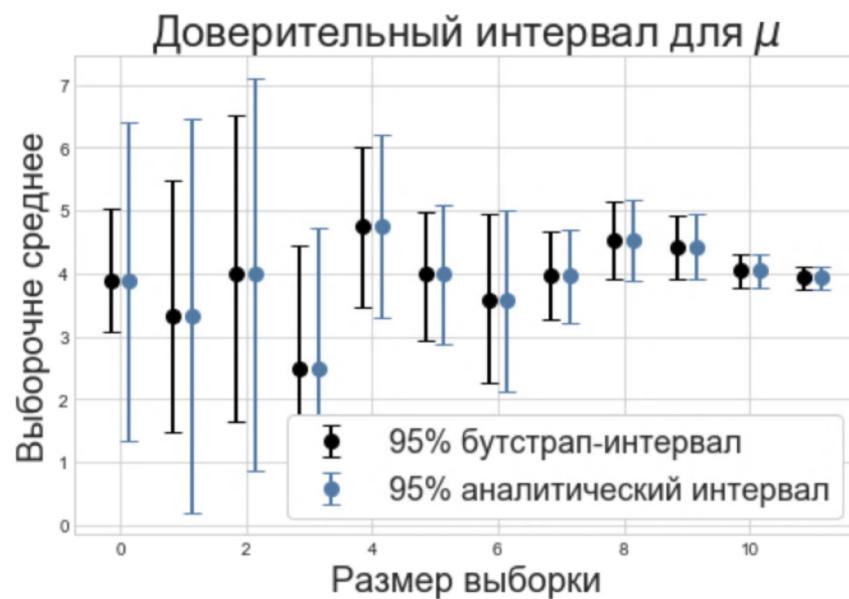
Бутстррап-аналог:

$$t^* = \frac{(\hat{\theta}^* - \hat{\theta})}{se(\hat{\theta}^*)}$$

- Гипотеза отвергается, если  $t_{obs} > t_{1-\alpha}^*$
  - По аналогии можно проверять гипотезы против других альтернатив
  - Для более сложных гипотез есть специальные бутстрраповские алгоритмы проверки
- <http://quantile.ru/03/03-SA.pdf>

# Проблемы бутстрата

- Чтобы бутстррап сработал, выборка должна быть репрезентативной



- Если исходная выборка маленькая, бутстрраповский доверительный интервал будет уже аналитического, так как в выборке недостаточно “неопределенности”

# Проблемы бутстрата

- Если в данных есть структура (регрессия, временные ряды), бутстррап нужно устроить так, чтобы учитывать её ⇒ разные виды бутстрата
- Бутстррап ненадёжно работает в хвостах распределения из-за маленького числа наблюдений: мы можем хорошо оценить медиану, но не 99% квантиль
- Если у распределения тяжёлые хвосты, бутстррап может работать некорректно и в средиземье

# Сколько процентов выборки берем

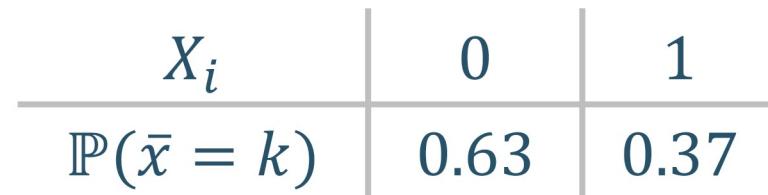
У Винни-Пуха есть 100 песенок (кричалок, вопелок, пыхтелок и сопелок). Каждый день он поёт одну равновероятно наугад. Одну и ту же песенку он может петь несколько раз. Сколько в среднем песенок не будут спеты ни разу за 100 дней?

- Вероятность конкретной песенки  $\frac{1}{n}$
- В конкретный день не споёт эту песенку с вероятностью  $1 - \frac{1}{n}$
- Не споёт песенку ни разу с вероятностью

$$\left(1 - \frac{1}{n}\right)^n \rightarrow e^{-1} = \frac{1}{e} \approx 0.37 \text{ при } n \rightarrow \infty$$

# Пример

- Случайная величина  $X_i = 1$ , если песенка не была ни разу спета за  $n$  дней



- Случайная величина  $Y = X_1 + \dots + X_n$  – число песенок, которое Винни-Пух ни разу не споёт за  $n$  дней

$$\mathbb{E}(Y) = \mathbb{E}(X_1 + \dots + X_n) = n \cdot \mathbb{E}(X_1) = 0.37 \cdot n$$

! В одной бутстрраповской выборке будет в среднем оказываться 63% наблюдений

# Резюме

- Бутстррап – это метод получения критических значений статистики
- Процедура может требовать много времени для оценки
- При некоторых ограничениях бутстррап даёт состоятельные оценки, но не в общем случае
- Плохо работает для статистик, значение которых зависит от небольшого числа элементов выборки

# Метрики для АБ-тестов

# Метрики

- Показатель для улучшения – метрика
- Метрики бывают разными, они конструируются в зависимости от бизнес-задачи
- Иногда метрики привязаны к деньгам
- Чаще всего денежные метрики грубые (слабо реагируют на изменения либо, надо очень много времени, чтобы их измерить)
- Из-за этого чистым денежным метрикам предпочтительнее промежуточные метрики

**Пример:** Сайт с арендой квартир: число посетителей за день, число уникальных посетителей и тп.

# Желательные свойства метрик

- **Согласованность** – метрика должна быть согласована с целями сервиса и его ключевыми метриками
- **Направленность** – если значение метрики изменилось, должна быть чёткая интерпретация этого изменения (хорошо это или плохо)

# Желательные свойства метрик

- **Чувствительность** (sensitivity) – способность метрики отражать статистически значимую разницу между контрольной и тестовой группами, когда она есть
- Чем выше чувствительность, тем меньше данных нужно, чтобы обнаружить статистически-значимые изменения

**Пример:** метрики, основанные на деньгах слабо реагируют на изменения

# Желательные свойства метрик

- **Стабильность** – метрика должна быть чувствительной и согласованной с тем, что нельзя ломать
- Если у метрики высокая дисперсия, то для того, чтобы уловить значимый эффект, надо собирать много данных

**Пример:** розничный торговый оборот магазина может колебаться в очень широких диапазонах. Чтобы уменьшить его дисперсию, обычно смотрят торговый оборот отдельных отделов.

# Желательные свойства метрик

## Лояльность пользователя:

- Число пользовательских сессий
- Время, которое юзер проводит в сервисе

Имеют чёткую направленность

Хорошие предикторы для долгосрочного успеха продукта

Обладают слабой чувствительностью

► <https://research.yandex.com/tutorials/online-evaluation/sigir-2019>

# Желательные свойства метрик

## Активность пользователя:

- Число кликов за сессию
- Длина пользовательской сессии

Обладают сильной чувствительностью

Обладают неоднозначной направленностью

► <https://research.yandex.com/tutorials/online-evaluation/sigir-2019>

# Желательные свойства метрик

**Пример:** клики пользователей в рекомендательной системе отражают как позитивные, так и негативные сигналы

- С одной стороны, они говорят, что пользователю нравится пользоваться продуктом
  - С другой, они говорят, что у нас много кликбейтного контента
  - Метрики с чёткой интерпретацией часто обладают низкой чувствительностью
- <https://research.yandex.com/tutorials/online-evaluation/sigir-2019>

# Как изучать свойства метрик

- Надо понимать особенности тех метрик, которые используются
- Замерять их характеристики
- Находить модификации, которые улучшают эти характеристики
- Нужен большой пул полезных исторических экспериментов

# Примеры свойств

В метриках могут быть тренд и сезонность, их необходимо от них очищать:

- Линейная регрессия
- Взятие 12-ой разности

Среднее не единственная метрика, которую можно считать:

- Средние чувствительны к выбросам
- Медианы устойчивы к выбросам
- Квантили помогают следить за определённым сегментом

# Математические трюки

Есть различные математические трюки, призванные улучшить свойства метрик:

- Сложные составные метрики с различными весами для составных частей
- CUPED (техника для увеличения чувствительности)
- Стратификация, разбиение пользователей на когорты
- Различные трансформации данных



! Разработка подобных метрик осуществляется под конкретный сервис

# Математические трюки

- После математических трюков, на АА-тестах метрика не должна показывать значимые изменения чаще, чем в  $\alpha$  процентах случаев, иначе мы сделали что-то странное
- Если преобразование оказалось успешным, оно должно быть проинтерпретировано

# Резюме

- Метрики используются, чтобы понять, что изменилось в нашем сервисе
- Метрики обладают различными свойствами
- Нужно аккуратно подбирать их для проведения эксперимента