

Интенсив: день 3

по анализу А/В-тестирований

День 3

Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Аналитику Лёне достались в руки новые результаты А/В-теста. Менеджер Петя тестировал влияние промо-пушей приложения доставки пиццы. Он выделил три когорты пользователей: первая – контроль, второй отправил пуш на скидку 30%, третьей отправил пуш на бесплатную пиццу до 35 см к заказу от двух пицц. Вопрос, кто на 3 месяц принесет больше денег?



Контроль



-30%



2+1 FREE

Чтобы ответить, нам поможет ...

Какие тут проблемы?



Контроль



-30%



2+1 FREE

Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Проблема 1:

Уровень False Positive растет соразмерно $1-(1-\alpha)^m$

где m – количество гипотез

Вкратце, проблема заключается в том, что при одновременной проверке большого числа гипотез на том же наборе данных вероятность сделать неверное заключение в отношении хотя бы одной из этих гипотез значительно превышает изначально принятый уровень значимости



Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Если сделать N тестов, то вероятность совершить хотя бы одну ошибку I рода в группе тестов (family-wise error rate, FWER) значительно возрастает согласно формуле

$$1 - (1 - \alpha)^m$$

, где m – количество гипотез. В случае с 3 когортами у нас 3 попарных сравнения, т.е. мы проверяем 3 гипотезы: A/B, A/C, B/C. Это означает, что при уровне значимости 95%, альфа будет

$$1 - (1 - 0.05)^3 = 0.142$$



Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Вероятность того, что тест не ошибется равна

$$0.95^{41} = 0.12$$

Или, если он ошибется хотя бы 1 раз

$$1 - 0.95^{41} = 0.88$$

Можно задать уровень значимости, и при 10 гипотезах можно получить такие результаты (при разных альфа):

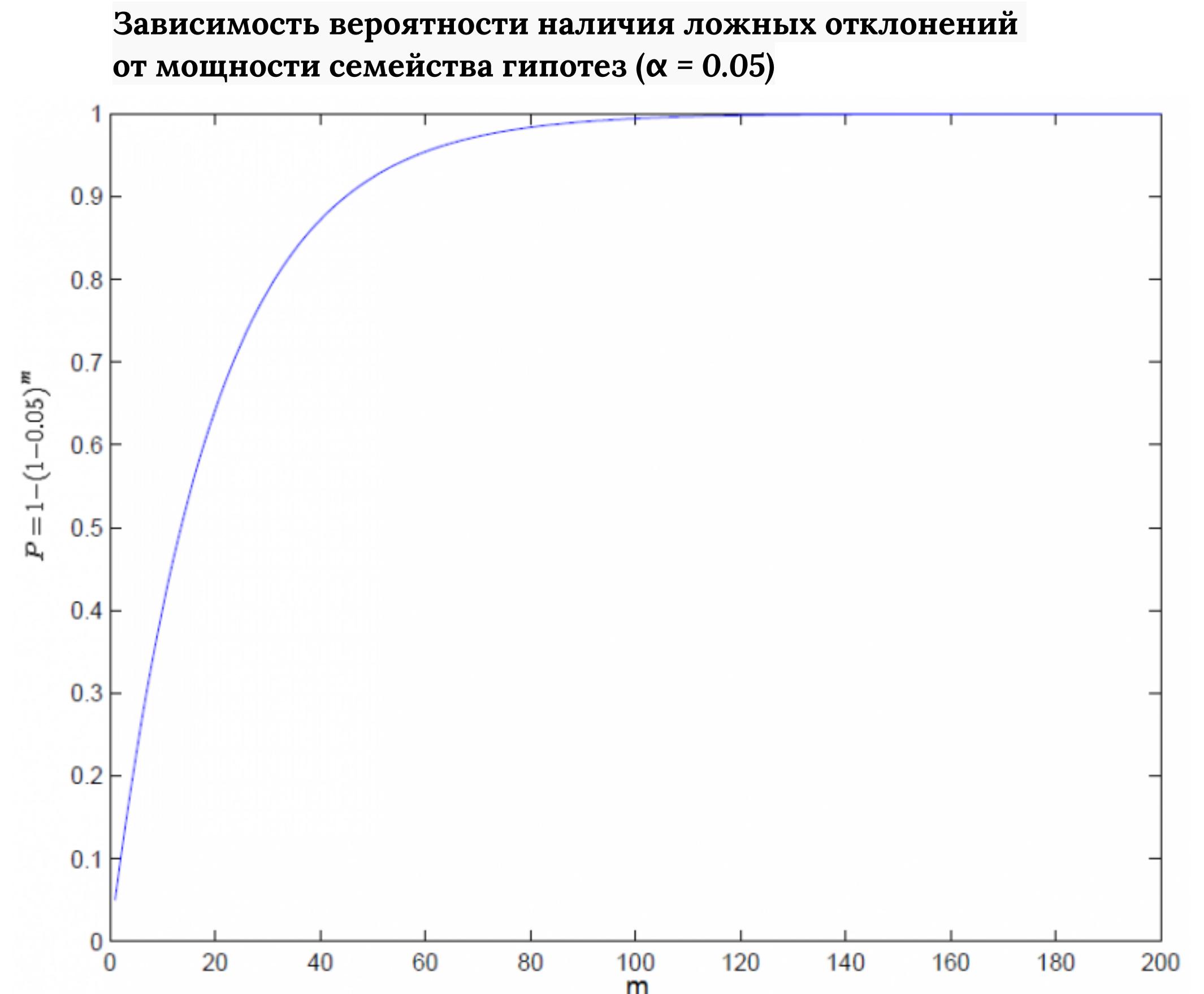
$$95\% = 1 - (1 - 0.05)^{10} = 0.401$$

$$99\% = 1 - (1 - 0.01)^{10} = 0.095$$

$$99.5\% = 1 - (1 - 0.005)^{10} = 0.095$$

$$99.9\% = 1 - (1 - 0.001)^{10} = 0.01$$

Т.е. для $\alpha = 0.05$ мы получим 40% ошибку, а вовсе не 5%, как изначально задается параметром



Проблема 2:

Чем больше гипотез проверяется в один момент,
тем сложнее проинтерпретировать данные

«У вас есть мобильная версия и десктопная, 50 стран, примерно 20 значимых источников реферрального трафика (поиск Google, партнерские ссылки и т.д.). Всего получается $2 \times 50 \times 20 = 2000$ сегментов. Предположим, что каждый сегмент идентичен каждому другому сегменту. Если сегментировать данные, получится $0,05 \times 2000 = 100$ статистически значимых результатов чисто случайно. Так совпало, что пользователи Android из Кентукки, перенаправленные Google, пользователи iPhone, перенаправленные jzyQvh8z, и пользователи, зашедшие через ПК в Нью-Джерси выбрали редизайн. Удивительно!»

Самый простой, но самый жесткий способ коррекции множественных сравнений – Поправка Бонферонни. Зная число тестов, можно вычислить скорректированный уровень значимости и использовать его

$$\alpha^* = \frac{\alpha}{N}$$

Например, чтобы сохранить в группе из 10 тестов вероятность ошибки I рода 0.05, нужно проводить каждый тест при $\alpha = 0.005$.

При этом резко возрастает вероятность не найти различий там, где они есть.

Для случаев, когда нам важнее сохранить истинно-положительные результаты, чем не допустить ложноположительных – используется контроль False Discovery Rate: $FP / (FP + TP)$

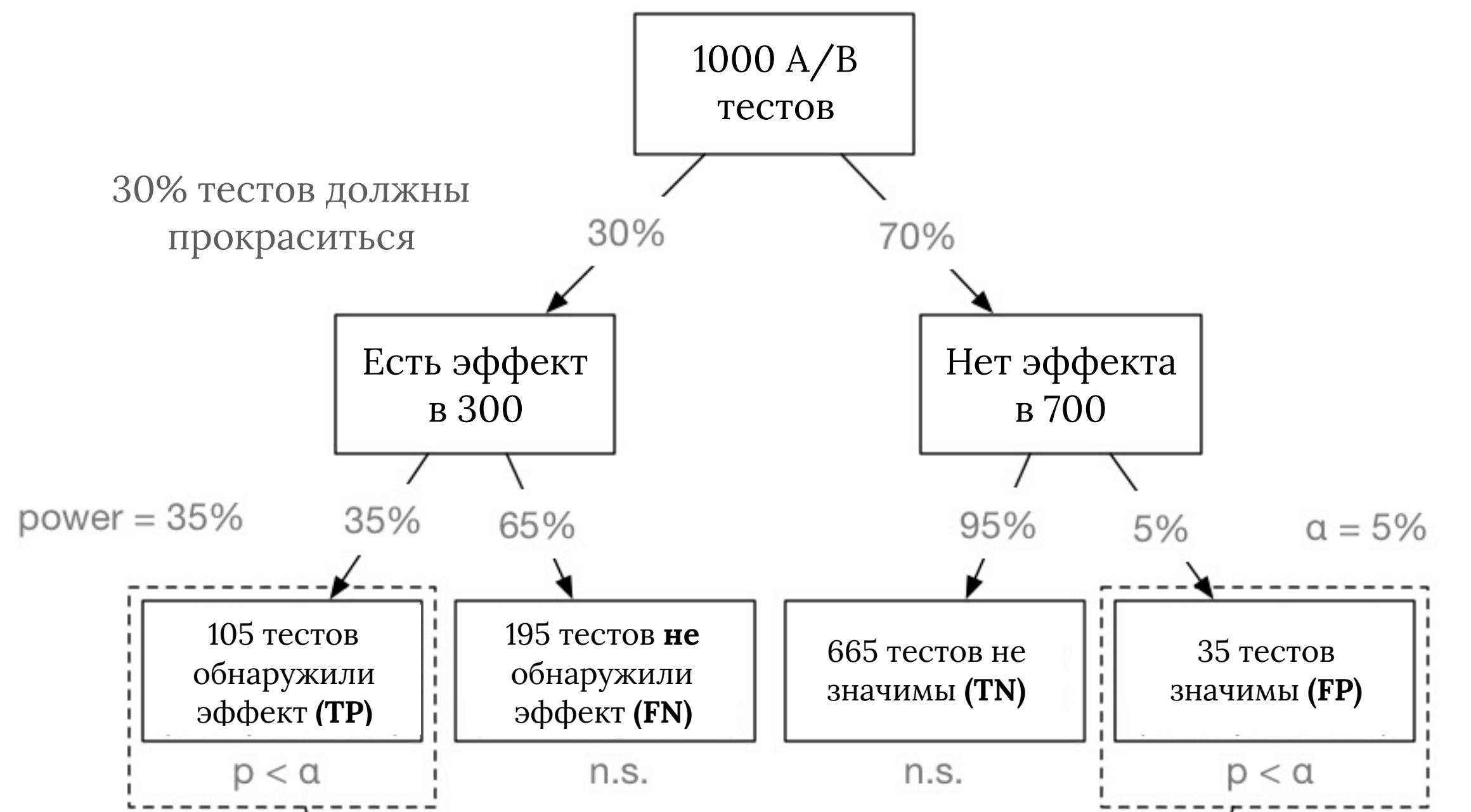
С помощью FDR мы задаем не количество ошибок первого рода в принципе, а количество ложноположительных (fp) результатов в отношении к истинно-положительным (tp) и ложноположительным (fp) (далее это число обозначается как γ)

Для контроля FDR используется поправка Бенджамини-Хохберга

Зачем рассчитывать все эти вероятности?

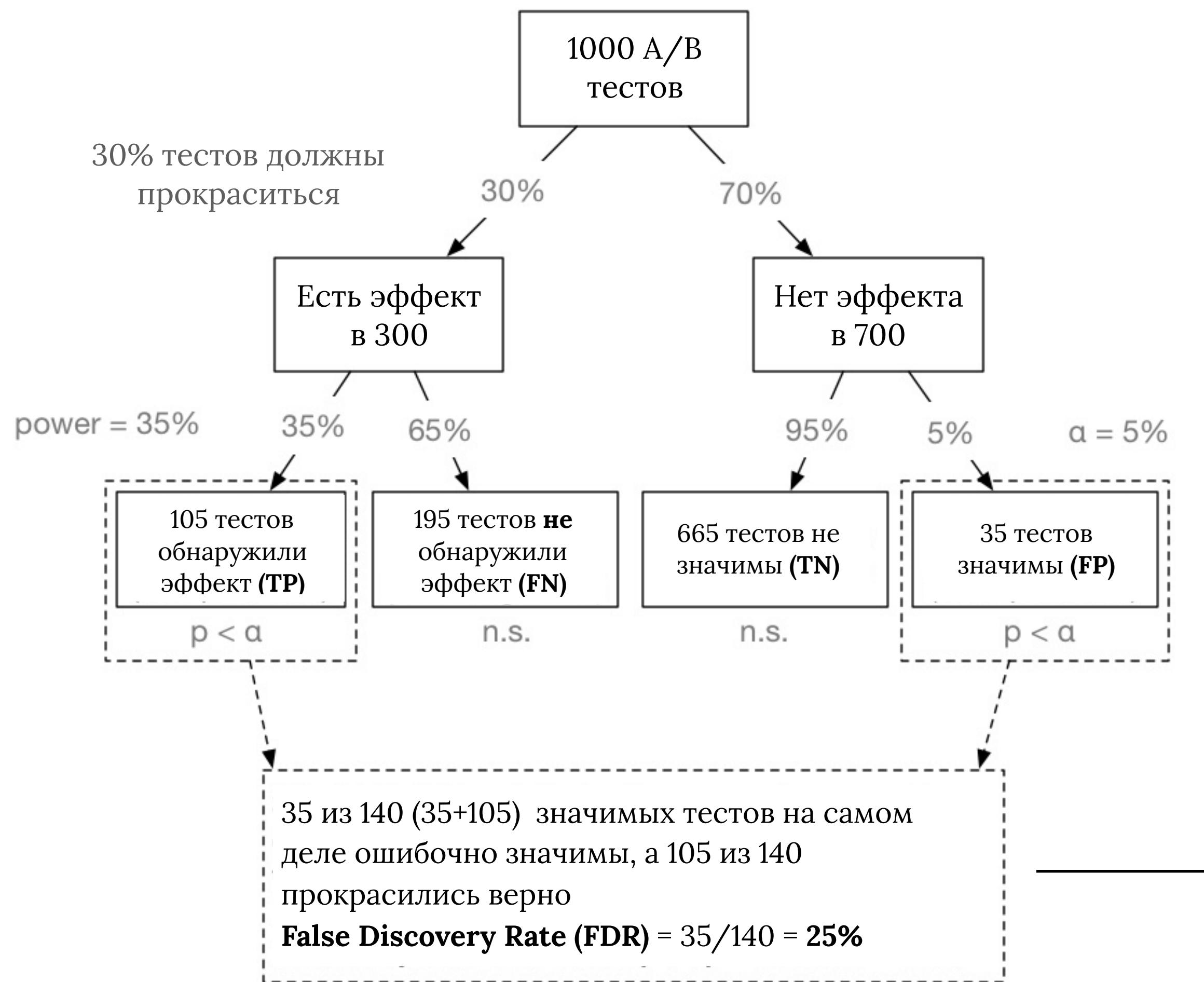
Множественная проверка гипотез и проблема подглядывания

e^xperiment fest



Множественная проверка гипотез и проблема подглядывания

e^xperiment fest



Для контроля FDR используется поправка
Бенджамини-Хохберга

Поправка Бенджамини-Хохберга. Чтобы зафиксировать $FDR < \gamma$ (γ – уровень FDR, обычно берут $0.1 = 10\%$) :

- Сортируем в порядке возрастания N значений p , полученные в тестах, и присваиваем им ранги j от 1 до N

$$p_1 \leq p_2 \leq \dots \leq p_{N-1} \leq p_N$$

- Находим p с наибольшим рангом j^* , такое чтобы

$$p_j \leq \gamma(j/N)$$

- Все тесты с рангами j меньше j^* считаем значимыми.

Для каждого теста можно вычислить q – минимальное значение FDR, при котором результат конкретного теста можно считать значимым

Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Всего $N = 5$ тестов. Частота ложноположительных результатов FDR = 0.1.

Сравниваем q и α .

Ранг j	p	$p^* = \gamma \cdot (j/N)$	$q = \gamma \cdot p^N / j$	Решение
1	0,005	$0.1 \cdot (1/5) = 0.02$	$0.1 \cdot 0.005 / 0.02 = 0.0250$	Отвергаем H_0
2	0,011	$0.1 \cdot (2/5) = 0.04$	$0.1 \cdot 0.011 / 0.04 = 0.0275$	Отвергаем H_0
3	0,02	$0.1 \cdot (3/5) = 0.06$	$0.1 \cdot 0.02 / 0.06 = 0.0333$	Отвергаем H_0
4	0,04	$0.1 \cdot (4/5) = 0.08$	$0.1 \cdot 0.04 / 0.08 = 0.0500$	Отвергаем H_0
5	0,13	$0.1 \cdot (5/5) = 0.10$	$0.1 \cdot 0.13 / 0.1 = 0.1300$	Сохраняем H_0

FWER или FDR?

- FDR (обычно $FDR < 0.1$): Выше мощность и контроль ложных срабатываний
- FWER (обычно $FWER < 0.05$): строгий контроль за вероятностью ошибок первого рода

Основной вывод

FDR для продуктовых реалий является более полезной метрикой для контроля, т.к. нам бы не хотелось выкатывать нерабочие изменения (когда мы говорим, что они рабочие), а наоборот, быть уверенными в реальных эффектах.

Мораторий на множественные сравнения

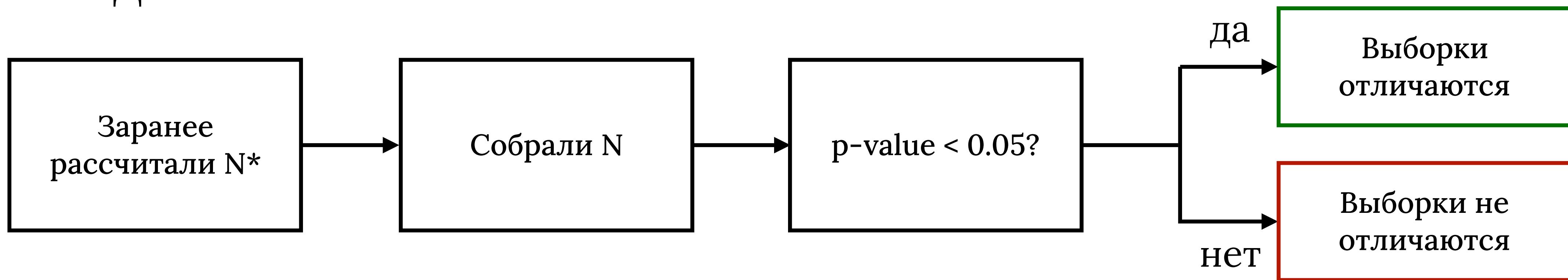
В некоторых продуктовых компаниях принят мораторий на множественные эксперименты из-за уже озвученных проблем:

- 1) сложная интерпретация
- 2) растущий false positive

Если можно не проводить множественный тест, то лучше воспользоваться такой возможностью

Fixed horizon

Как должен быть



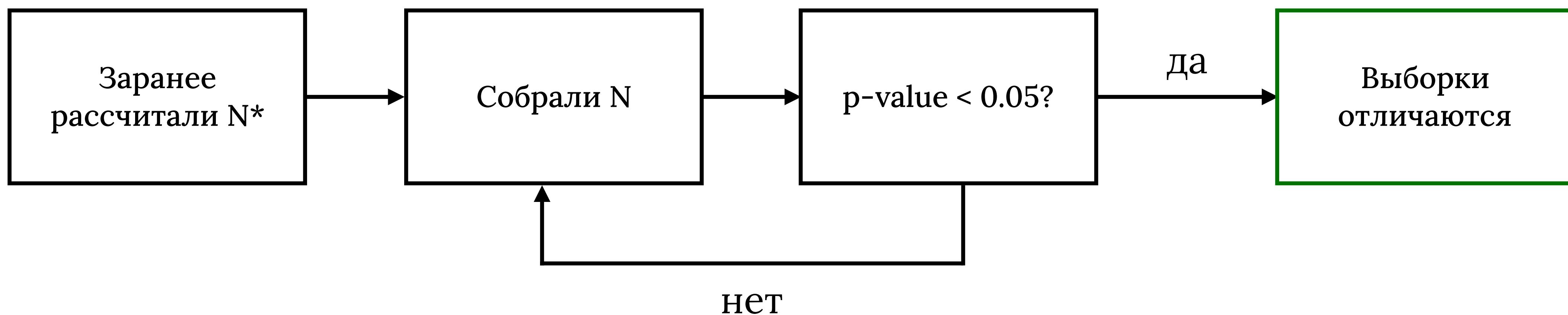
*согласно параметрам ТР (мощность), TN (уровень значимости), MDE (minimum detectable effect)

Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Fixed horizon

Как есть



*согласно параметрам ТР (мощность), TN (уровень значимости), MDE (minimum detectable effect)

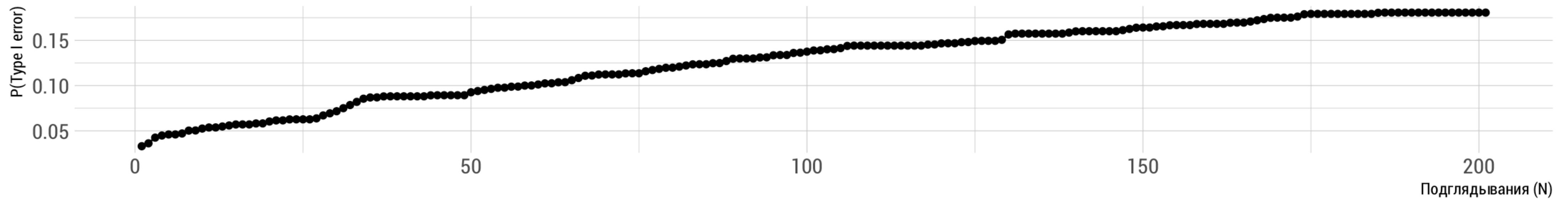
Множественная проверка гипотез и проблема подглядывания

e^xperiment fest

Проблема подглядываний в эксперименте

Так накапливается ошибка каждый раз, если бы мы подглядывали в эксперимент (после +1 пользователя).

При параметрах теста, где нет разницы между А и Б, критерий должен давать ложный результат только в 5% случаев. В нашем случае это не так.



Итого

- Подход Fixed horizon близок к медицинским исследованиям, где бюджет определяет допустимую выборку на эксперимент
- Основные положения по расчету выборки описаны в работах J. Cohen'a в 1988*
- Однажды применив статистический критерий, и приняв ту или иную гипотезу, мы не можем продолжить эксперимент, чтобы набрать побольше пользователей и проверить ещё раз, т.к. увеличиваем FDR**

*гугл Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale,NJ: Lawrence Erlbaum.

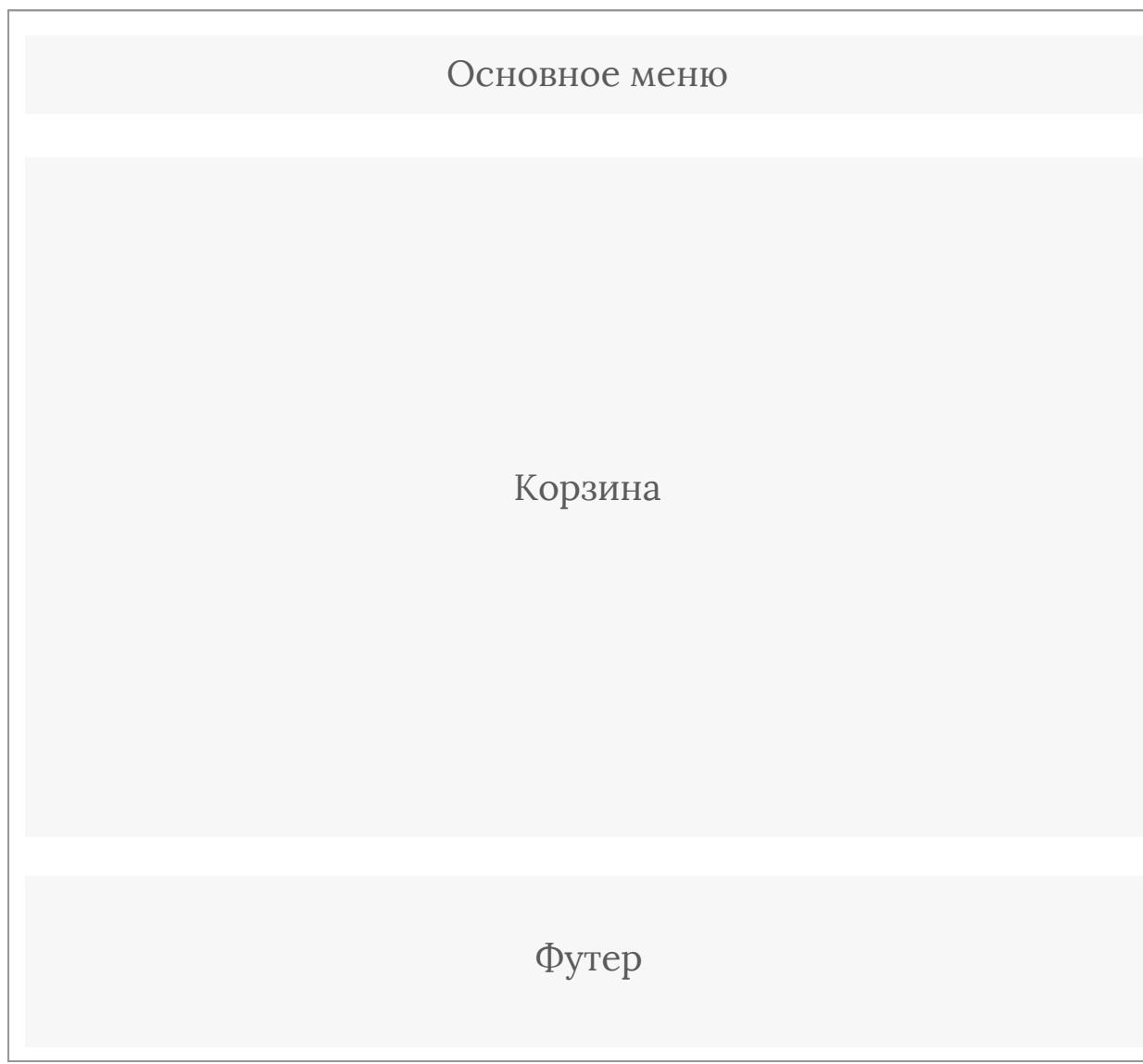
** False Discovery Rate – ложноположительные оценки из всех положительных ($fp/tp+fp$)

День 3

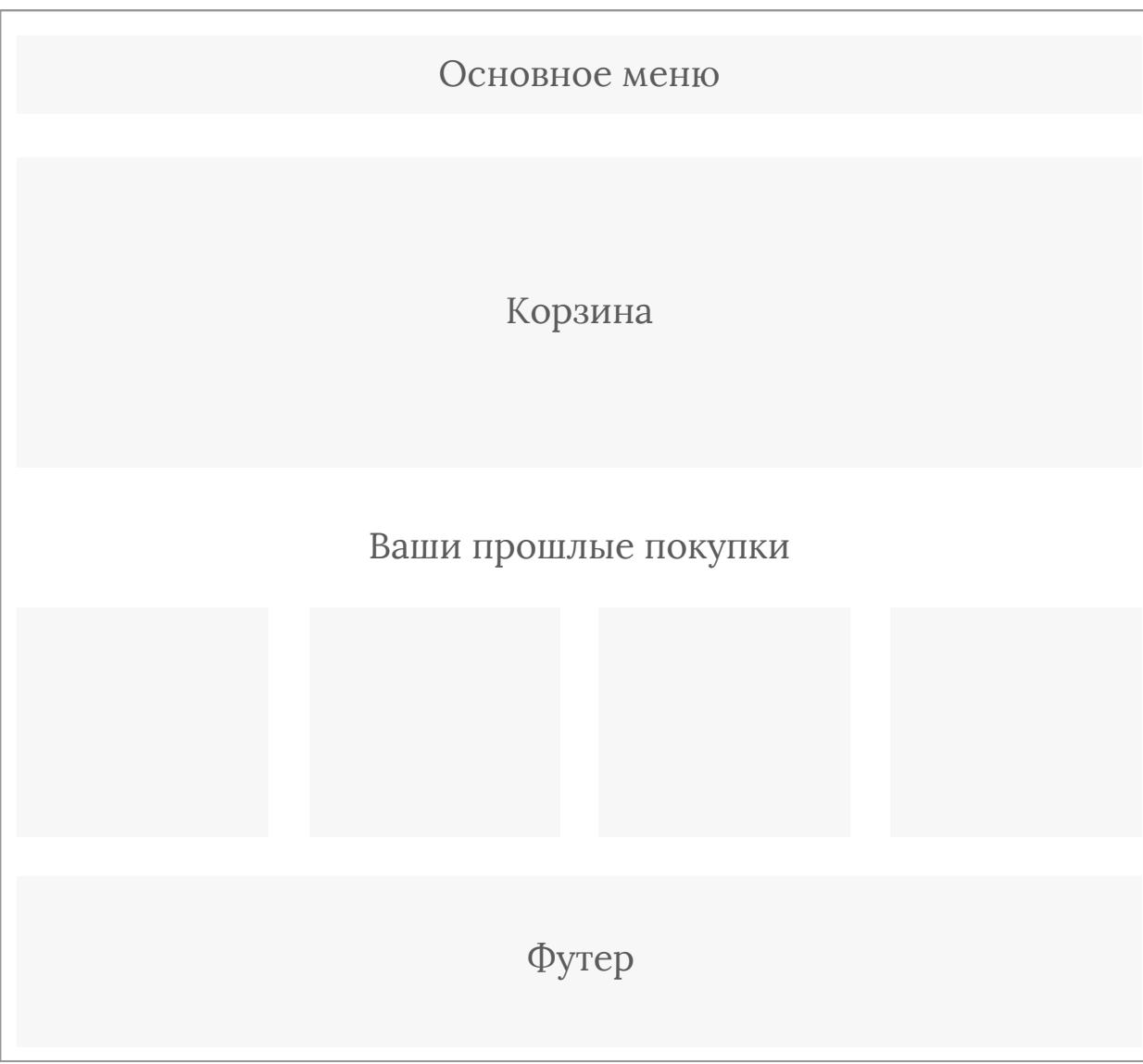
Бутстрап: повторные выборки и децильные методы оценки А/Б тестов

e^xperiment fest

Контроль



Тест



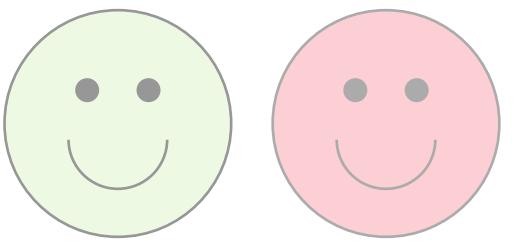
Кейс из фуд-ритейла:

Добавили новую витрину «Ваши прошлые покупки» на чекаут.

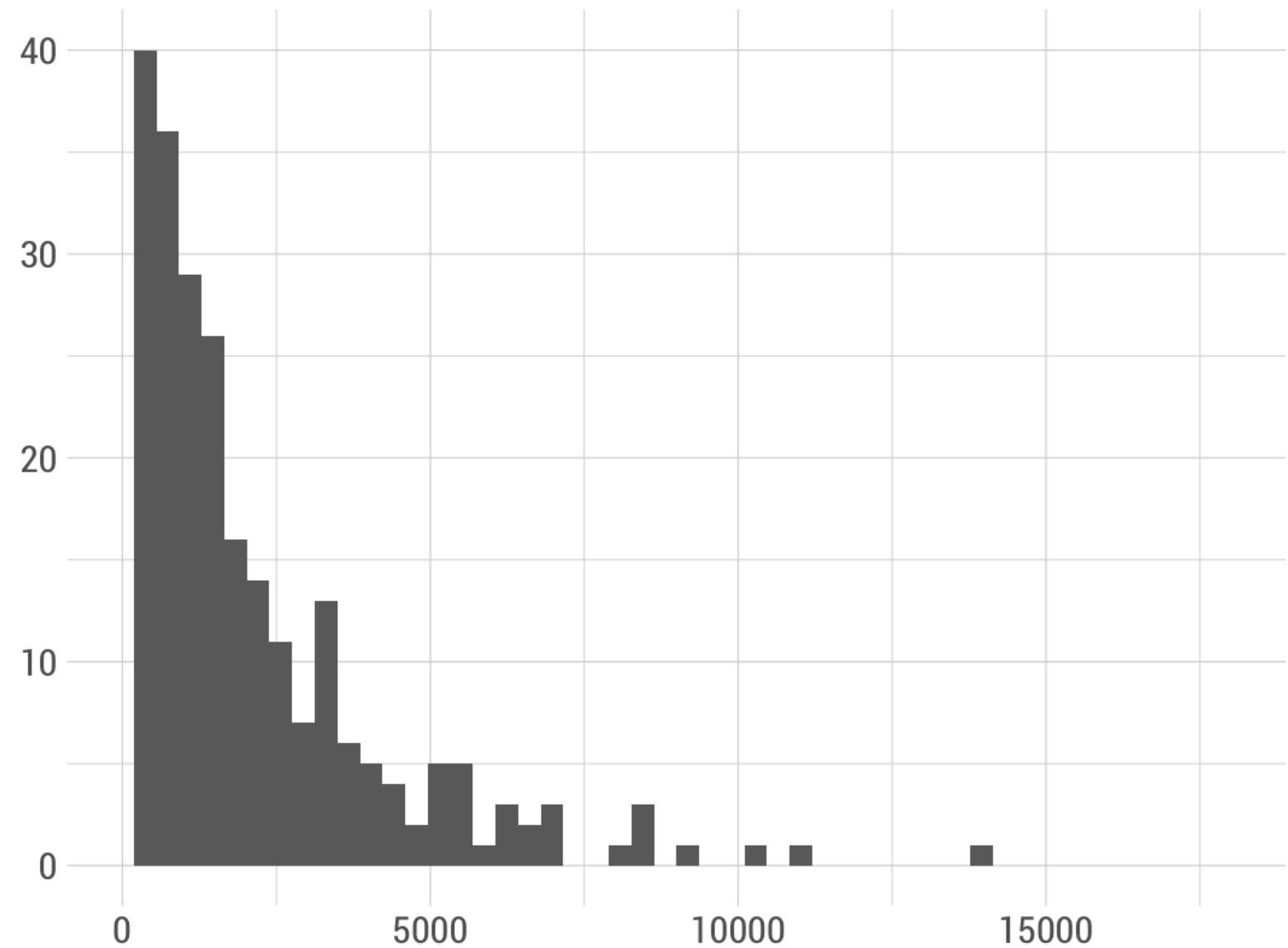
По классике, бизнесу интересно узнать как изменился средний чек

Как оценить влияние эксперимента на прибыль?

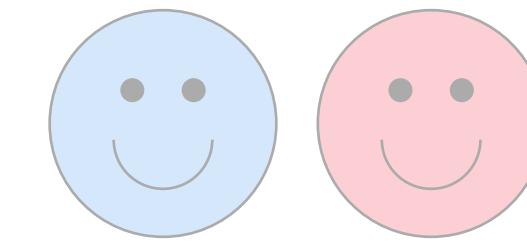
Average revenue_{control} = 5253



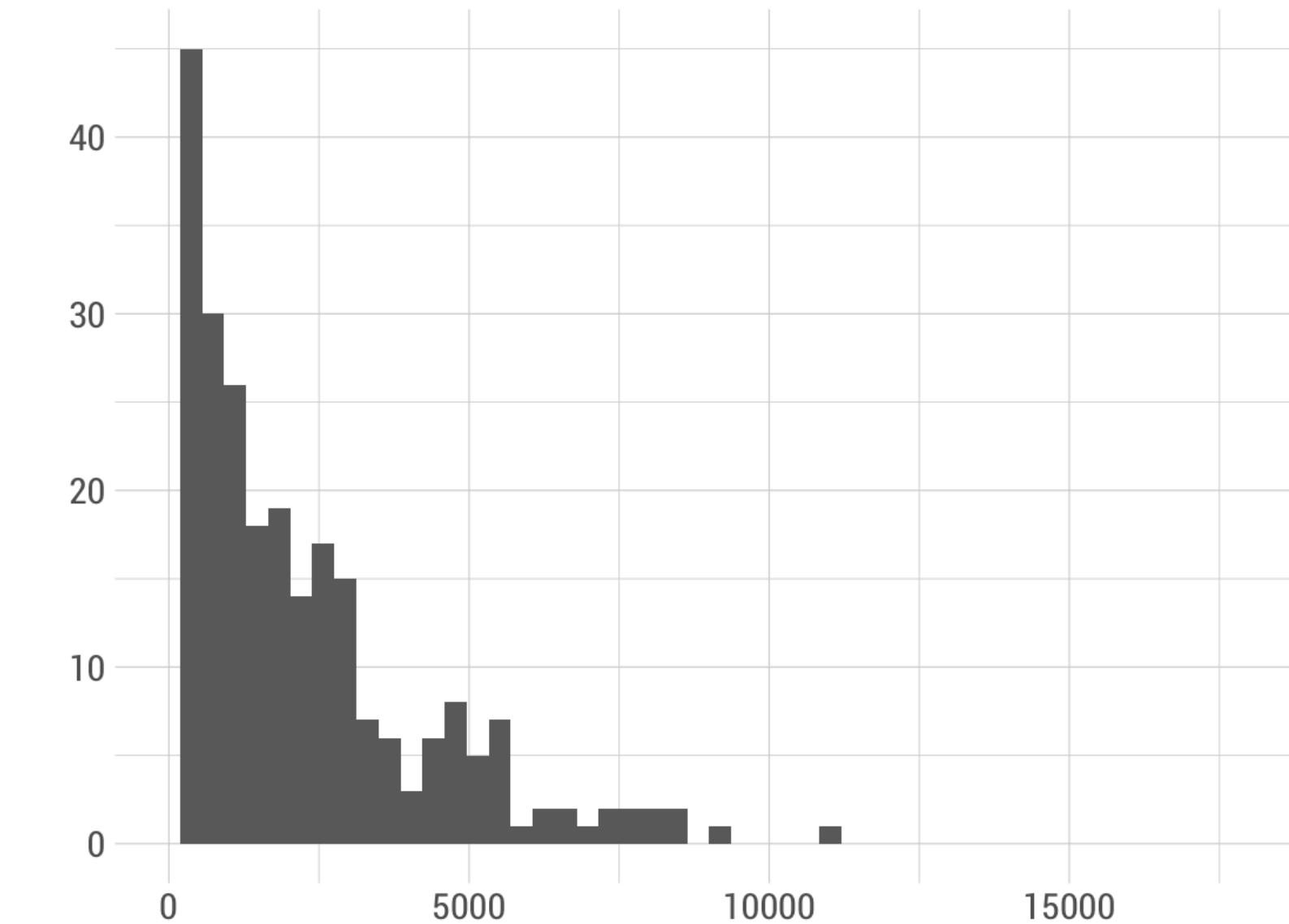
Avg revenue control



Average revenue_{test} = 5486



Avg revenue test



Bootstrap

e^xperiment fest

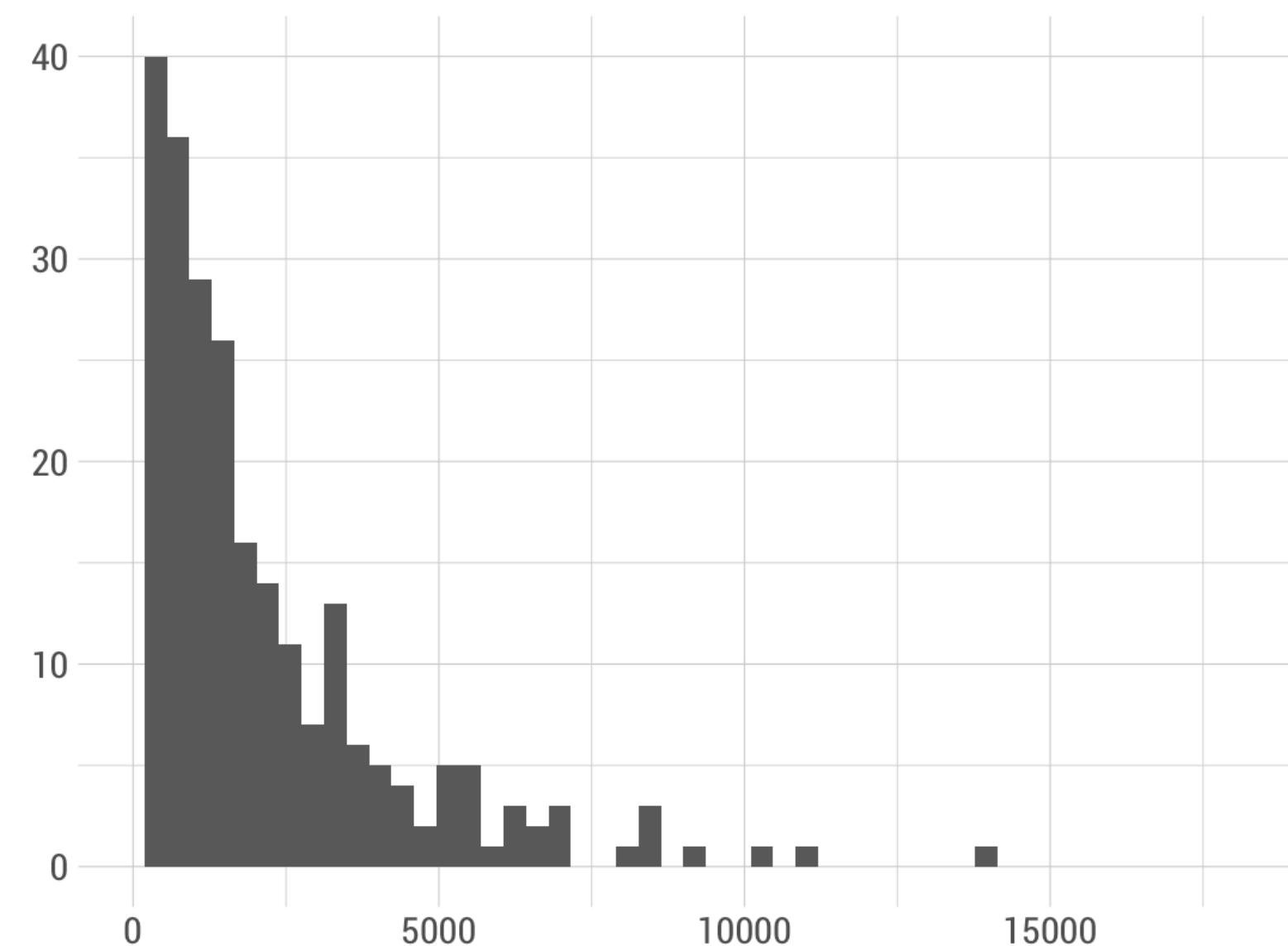
Что можем сделать?

**Что лучше будет отражать
центральную тенденцию? –**

Медиана

**Можем ли использовать ЦПТ,
чтобы построить ДИ для
медианы? – Нужен какой-то метод
для этой задачи**

Avg revenue control



Bootstrap

e^xperiment fest

Бутстрап

Суть

Обладая только данными по имеющейся выборке, существует возможность оценить любой ее параметр, построив эмпирическое *распределение параметра*.

В контексте нашей задачи с медианой – получить распределение медиан и далее по ним вычислить доверительный интервал.

Давайте разберемся как это делается

Выборка

2.11

3.12

9.4

17.9

8.6

4.1

Bootstrap

e^xperiment fest

Выборка

2.11

9.4

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

e^xperiment fest

Выборка

2.11

3.12

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

9.4

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

9.4

e^xperiment fest

Выборка

2.11

3.12

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

9.4

9.4

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

8.6

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

9.4

9.4

e^xperiment fest

Выборка

**Повторная
бут-выборка**

3.12	3.12
9.4	9.4
17.9	2.11
8.6	
4.1	

Bootstrap

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

8.6

**Повторная
бут-выборка**

3.12

9.4

9.4

2.11

4.1

Bootstrap

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

8.6

4.1

**Повторная
бут-выборка**

3.12

9.4

9.4

2.11

4.1

Bootstrap

e^xperiment fest

Выборка

2.11

3.12

9.4

17.9

4.1

Bootstrap

**Повторная
бут-выборка**

3.12

9.4

9.4

2.11

4.1

8.6

e^xperiment fest

<i>Среднее</i>	<i>Среднее</i>
7.53	6.12
Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
9.4	9.4
17.9	2.11
8.6	4.1
4.1	8.6

Bootstrap

e^xperiment fest

<i>Среднее</i>	<i>Среднее</i>	<i>Среднее</i>	<i>Среднее</i>	<i>Среднее</i>	<i>Среднее</i>
Выборка	Повторная бут-выборка	Повторная бут-выборка	Повторная бут-выборка	Повторная бут-выборка	Повторная бут-выборка
7.53	6.12	6.12	5.98	4.9	
2.11	3.12	3.12	8.60	4.10	
3.12	9.4	9.40	9.40	9.40	
9.4	9.4	9.40	8.60	3.12	
17.9	2.11	3.12	2.11	2.11	
8.6	4.1	3.12	4.10	8.60	
4.1	8.6	8.60	3.12	2.11	

Бутстррап распределение средних

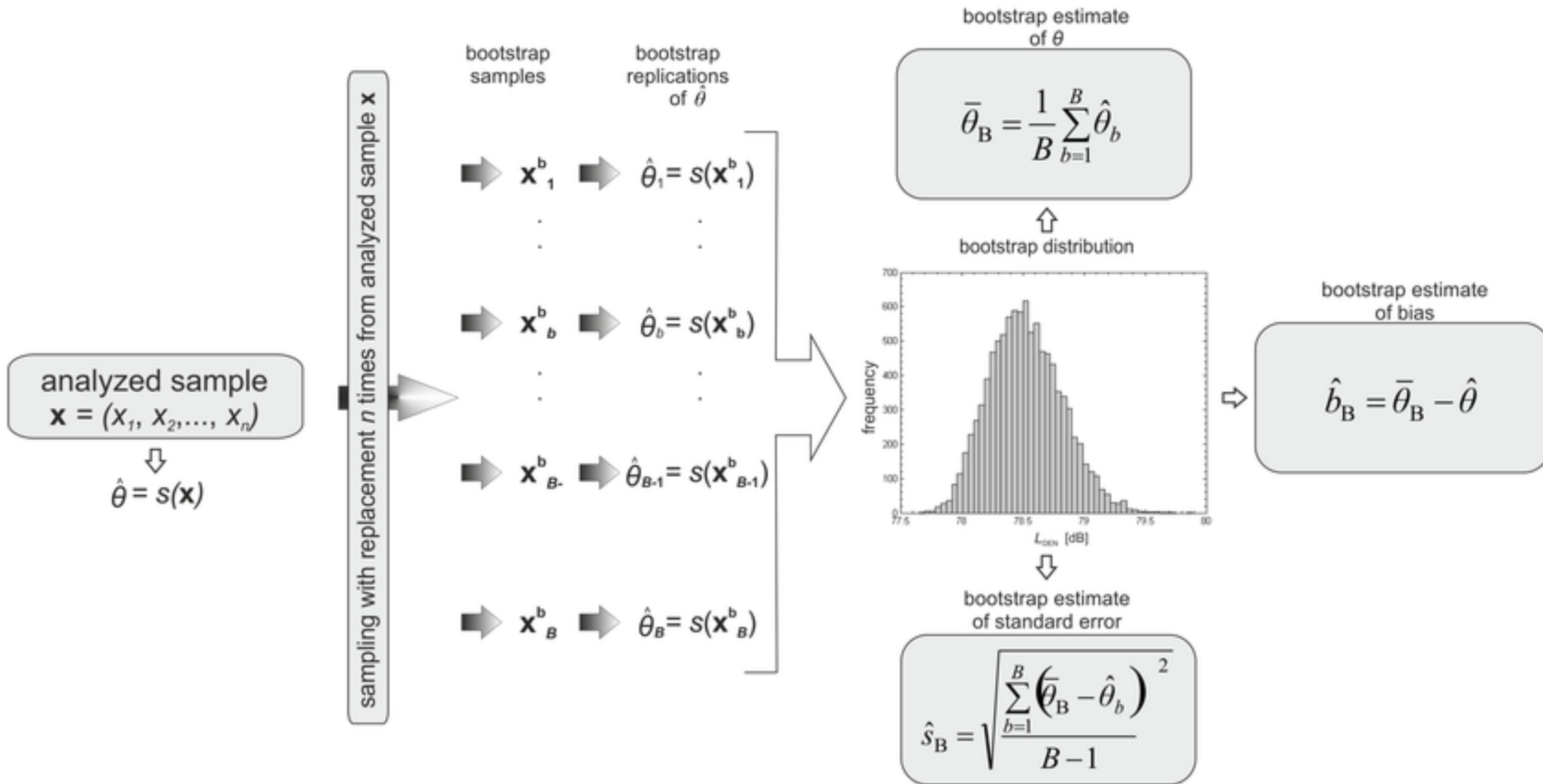
6.12 6.12 5.98 4.9

→ считаем доверительный интервал →

получаем эмпирическую оценку
параметра распределения

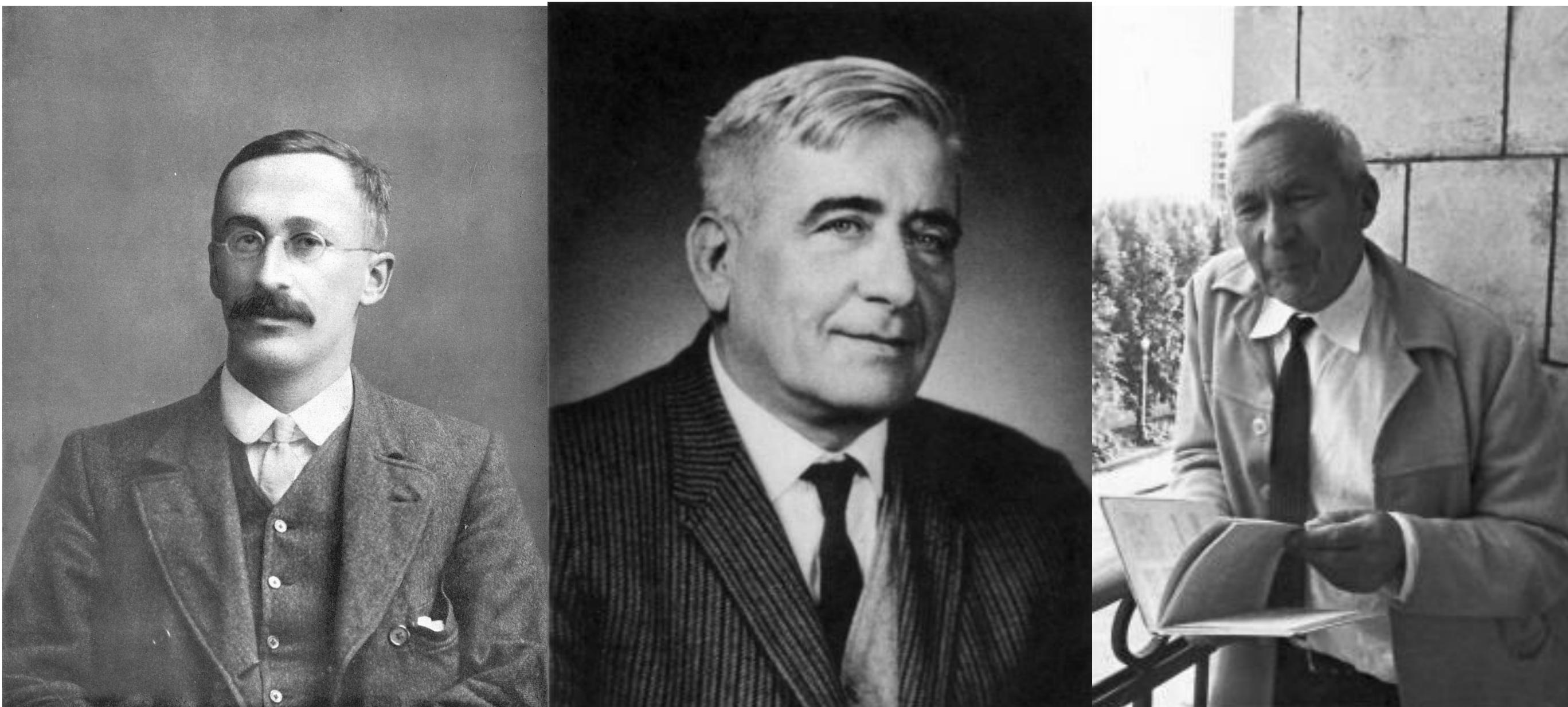
Bootstrap

e^xperiment fest



Демонстрация

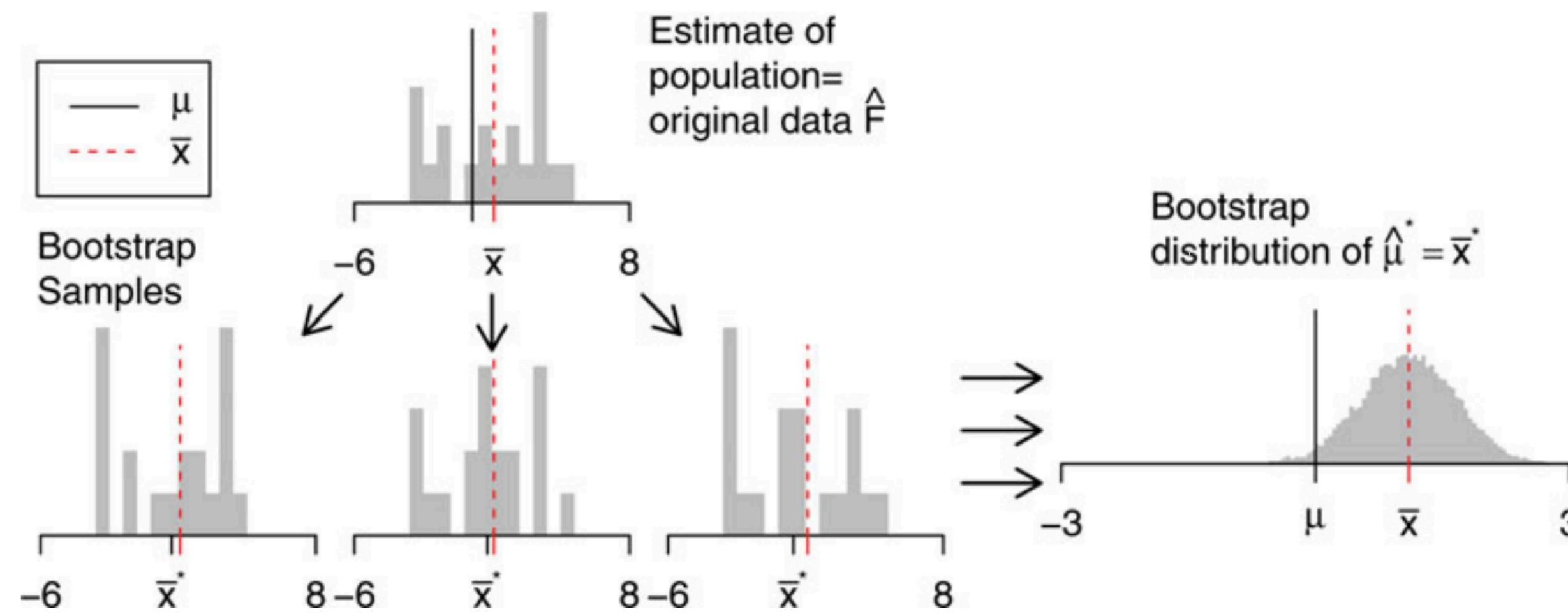
Что нам мешает просто использовать классический критерий?



Что нам мешает просто использовать классический критерий?

- Манна-Уитни лучше всего подойдет для задачи. Он дает ответ на вопрос, значимо ли различаются распределения или нет. Хотелось бы понимать **где именно** эта разница
- К тому же, у каждого критерия свое аналитическое решение, которое требует придерживаться ряда допущений (например, одинаковая дисперсия / одинаковый размер выборки / одинаковая форма распределений и т.п.). Такая возможность не всегда имеется

Мы не будем загонять результаты в критерий,
а решим задачу с помощью повторных выборок (а
именно Бутстррап), сравнивая **декили распределений**



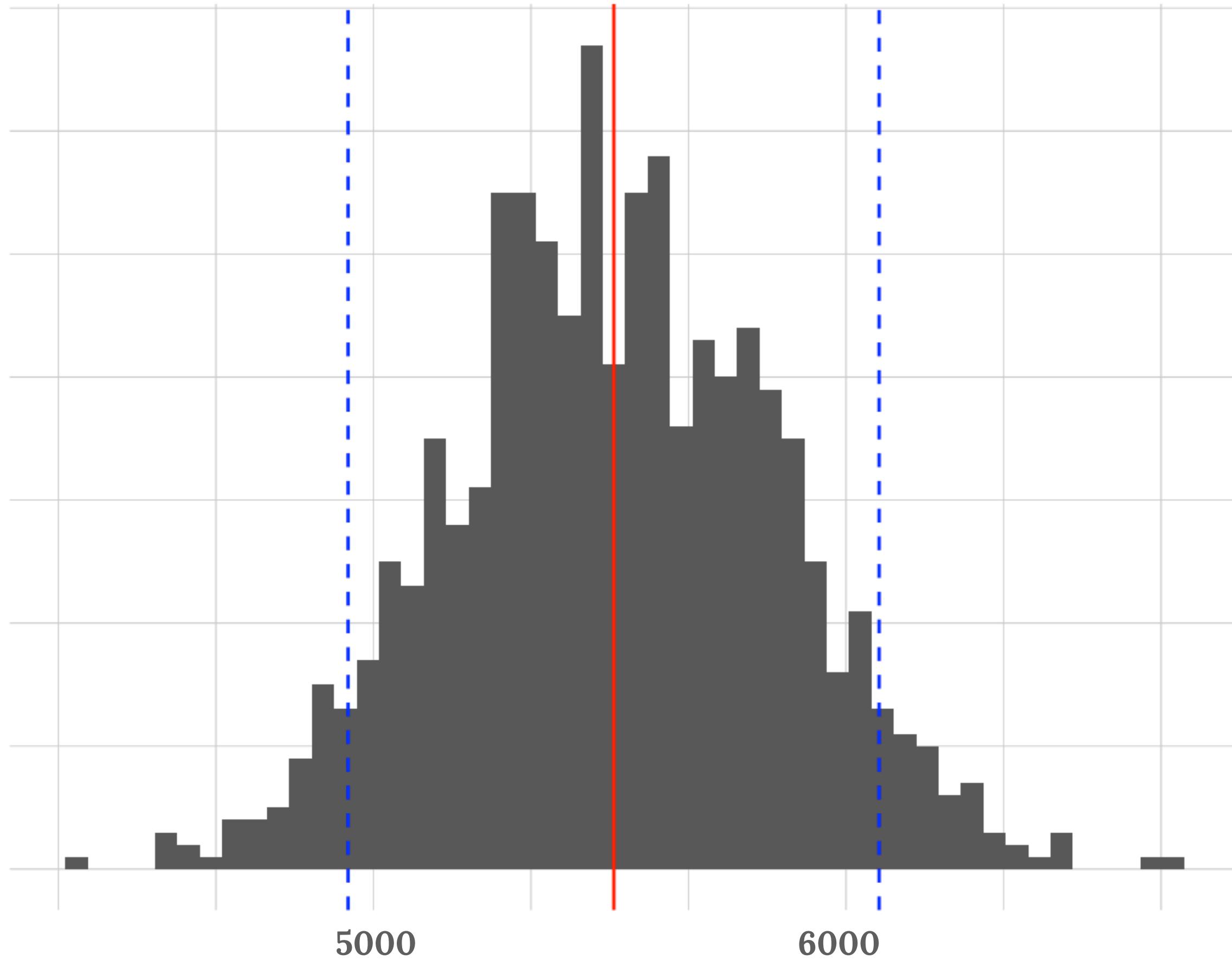
Bootstrap

e^xperiment fest

$$\text{avg}(\hat{\theta}^*)_{\text{test}} = 5481$$

Помним, что
Average revenue_{test} = 5486

51



Bootstrap

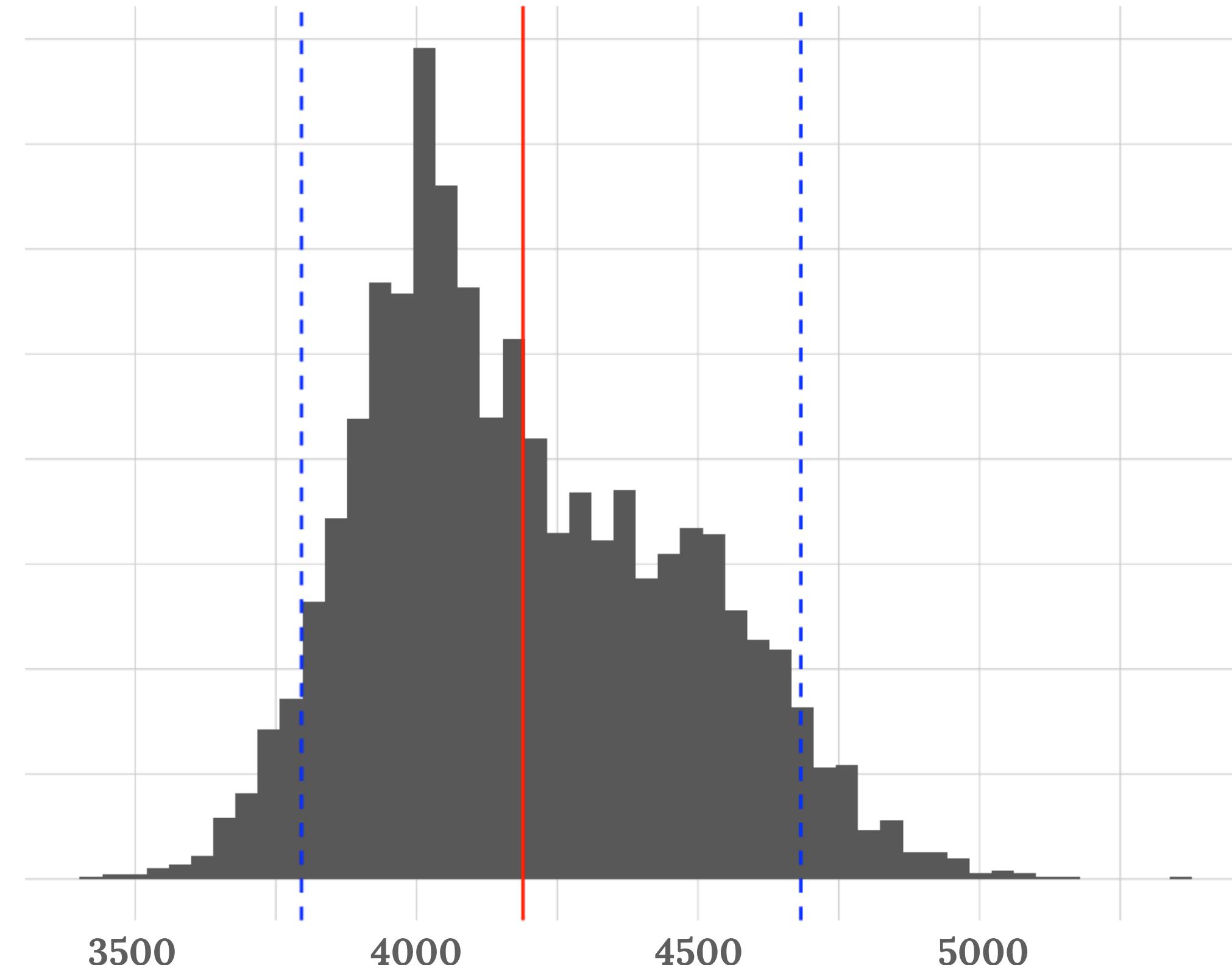
- bias = 5
- Взяли 10000 бутстр-выборок, нашли в них среднее. Это и есть распределение оценочного параметра
- Синим пунктиром ДИ 95%
- bias очень низкий, это хорошо
- Можем найти эффект, посчитав разницу для параметра между двумя оценщиками

e^xperiment fest

$$avg(\hat{\theta}^*)_{test} =$$

Посмотрим на другой параметр
Median revenue test = 4046

52



- bias = 154
- Проблема сложных оценочных параметров (в данном случае медиана) – высокий bias
- Чем меньше объем бутстрэп-выборки, тем выше bias. Но и увеличение объема не решает проблему всегда
- Нужно подогнать оценщик к исходной выборке

Bootstrap

e^xperiment fest

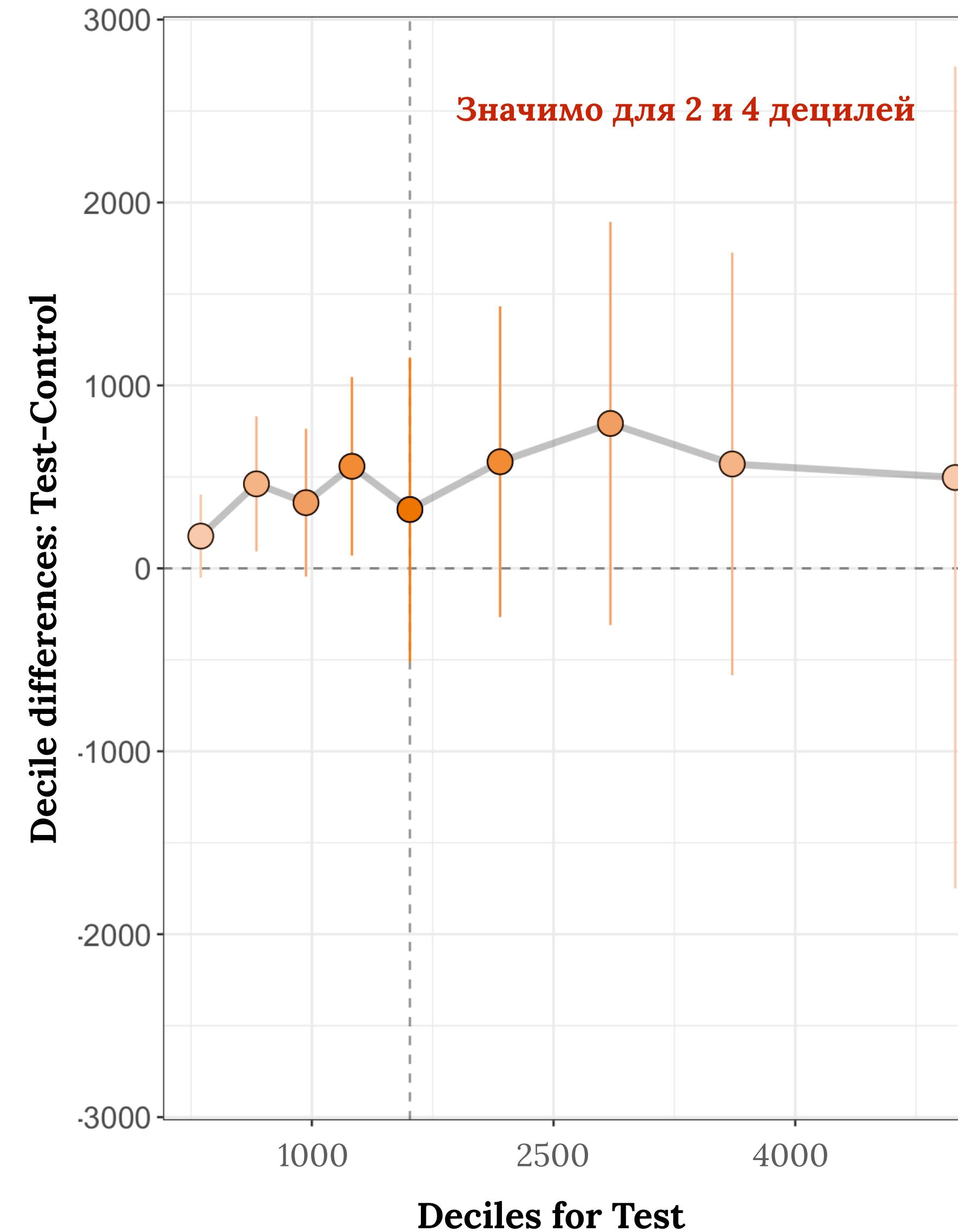
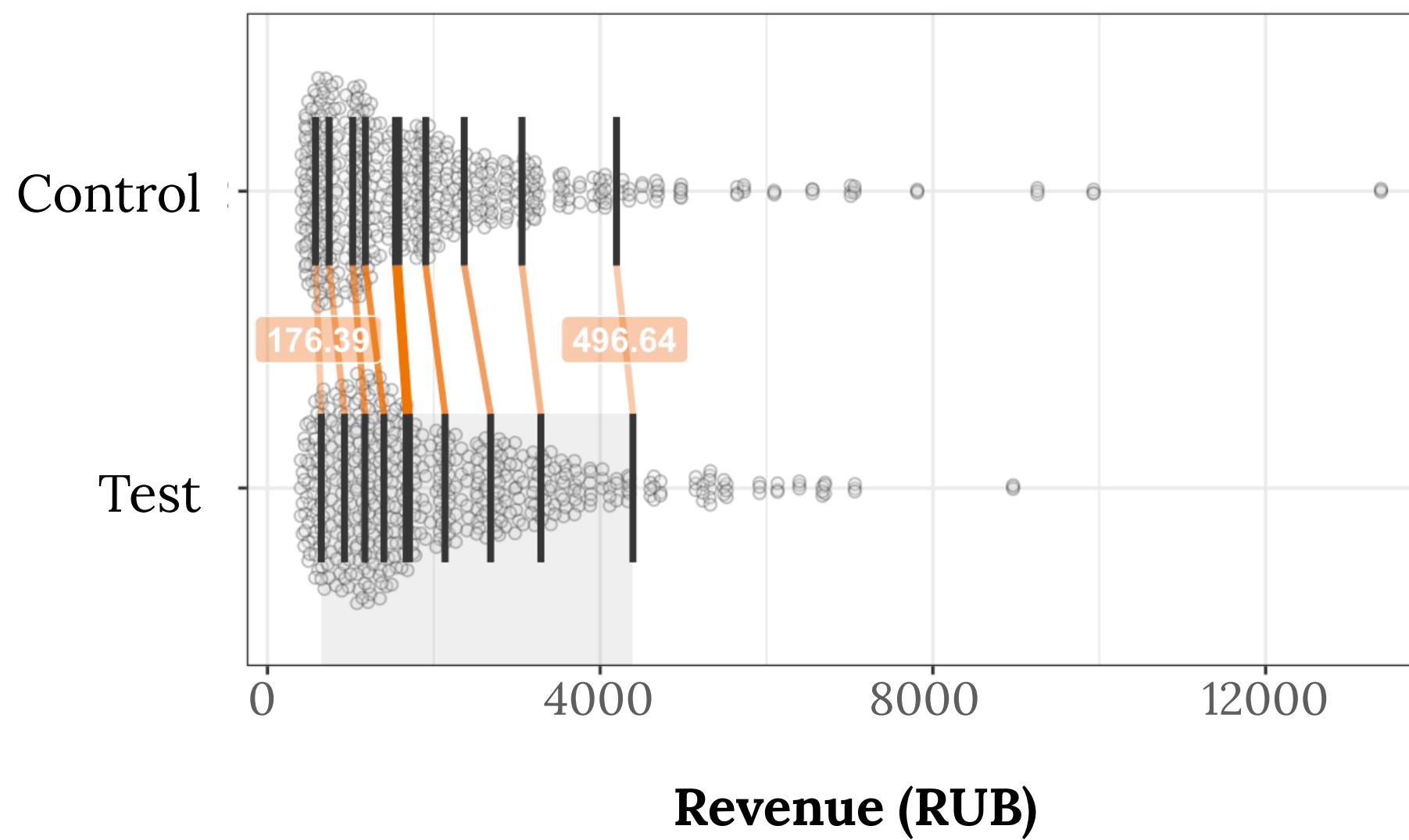
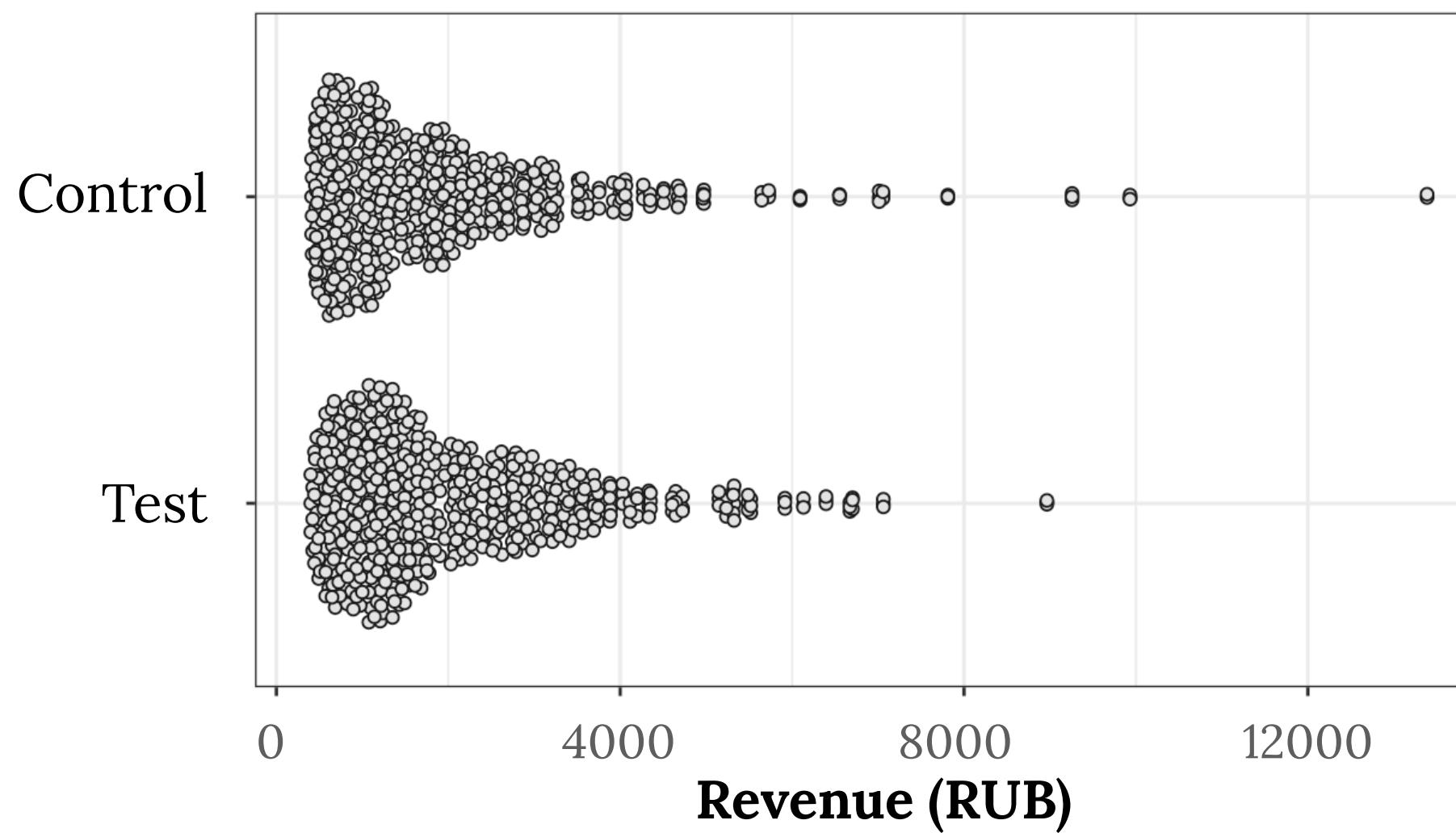
- Бутстррап позволяет строить доверительный интервал для любого параметра распределения, не применяя для этого аналитическую формулу
- Основное преимущество Бутстррап – проверять гипотезы для любых параметров распределения или моделей: Перцентили/Квантили/Децили и т.п.
- Бутстррап проверяет статистические гипотезы без опоры на определенное теоретическое распределение данных (в отличие от классических стат. критериев)
- Бутстррап позволяет сделать оценку любого «сложного» параметра путем нахождения доверительных интервалов для него. А для проверки гипотез – путем вычисления их разницы

**Мы можем пойти дальше и
проверить каждую дециль
распределения**

Как проверяются гипотезы с помощью бутстрата?

1. Строите бутстррап-распределения параметра в А и Б
2. Вычисляете их разницу (вычитание матриц)
3. В получившемся распределении разницы считаете доверительный интервал
4. Смотрите, попадает ли доверительный интервал в 0. И если да, то нулевая гипотеза на заданном уровне значимости принимается

Результаты для эксперимента с добавлением блока с прошлыми покупками



Демонстрация в R

Bootstrap

e^xperiment fest

Можно ли нормализовать
распределение метрики с помощью
бутстрата, а потом использовать
критерий?

Assumptions [\[edit \]](#)

Most test statistics have the form $t = \frac{Z}{s}$, where Z and s are functions of the data.

Z may be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a **scaling parameter** that allows the distribution of t to be determined.

As an example, in the one-sample t -test

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

where \bar{X} is the **sample mean** from a sample X_1, X_2, \dots, X_n , of size n , s is the **standard error of the mean**, $\hat{\sigma}$ is the estimate of the **standard deviation** of the population, and μ is the **population mean**.

The assumptions underlying a t -test in its simplest form are that

- \bar{X} follows a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$
- s^2 follows a χ^2 distribution with $n - 1$ **degrees of freedom**. This assumption is met when the observations used for estimating s^2 come from a normal distribution (and i.i.d for each group).
- Z and s are **independent**.

Для применения параметрических критериев (напр., t -критерий Стьюдент), требуется соблюдать предположение о независимости выборок.

Механизм бутстрата так построен, что одно наблюдение может встретиться **МНОГО-МНОГО** раз в одном и том же распределении

Кейс Школа. Подготовка школьников к ЕГЭ. Пакеты разбиты на 10,15,30,60 занятий подготовки к ЕГЭ (любые предметы). Пользователь приобретает пакет и далее занимается с репетитором. Добавили новый пакет на 5 уроков, что случится с экономикой?

Контроль

60 занятий
30 занятий
15 занятий
10 занятий

Тест

60 занятий
30 занятий
15 занятий
10 занятий
5 занятий

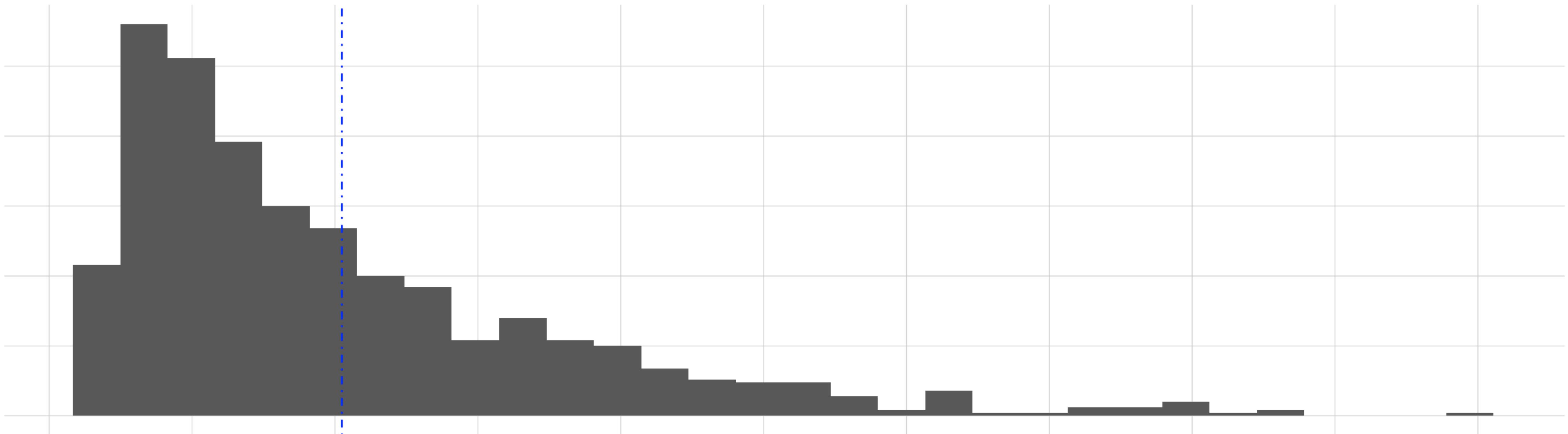
- Итоговая конверсия в С1 (первая покупка) выросла. Но тут не обошлось без каннибализации. Средних пакетов стало чуть меньше, а пакеты на 5 занятий расходились «как пирожки»
- С деньгами история чуть сложнее. В «моменте» С1 деньги (ARPPU) не упали и разницы почти не было. Тест шел недолго, ждать закрытия С2, С3 окон не было времени. Как итог – 5 занятий выкатили на всех

Спустя почти 1 год, выяснилось, что средняя отхоженность занятий упала.
Нам показалось, что это связано с релизом нового пакета на 5 занятий.

В конечном счете, выяснилось, что так оно и есть. Как мы это поняли?

Благодаря децильному сравнению.

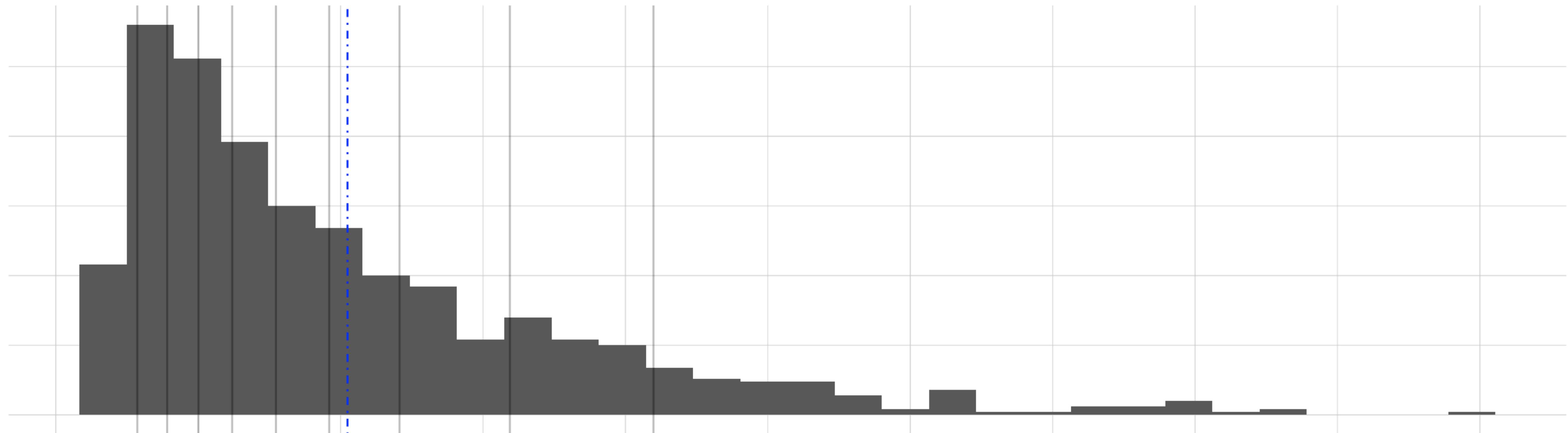
Суть заключается в том, что распределение выборки разделяется на 10 равных децилей. Это каждая 10-ая квантиль (10,20,30,40,50...)



Bootstrap

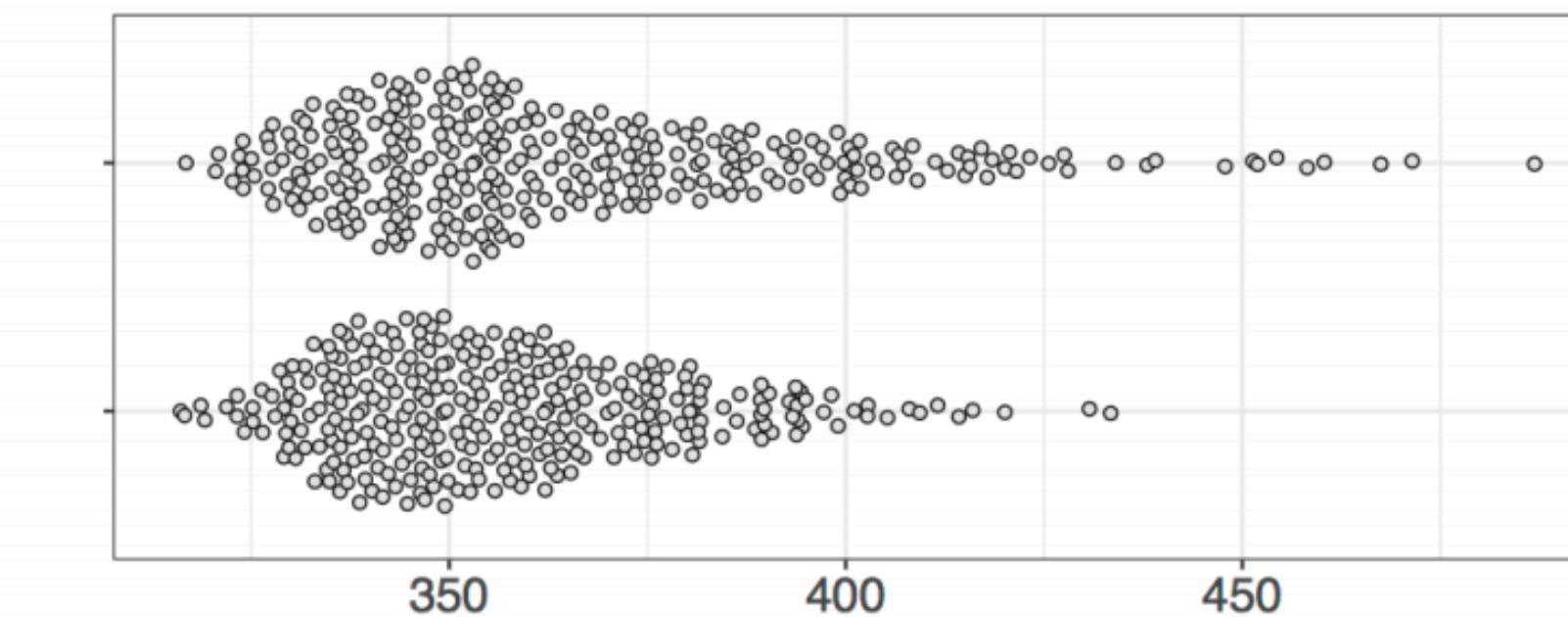
e^xperiment fest

Далее для каждого дециля рассчитывается доверительный интервал с помощью бутстрэпа, путем извлечения рандомных выборок

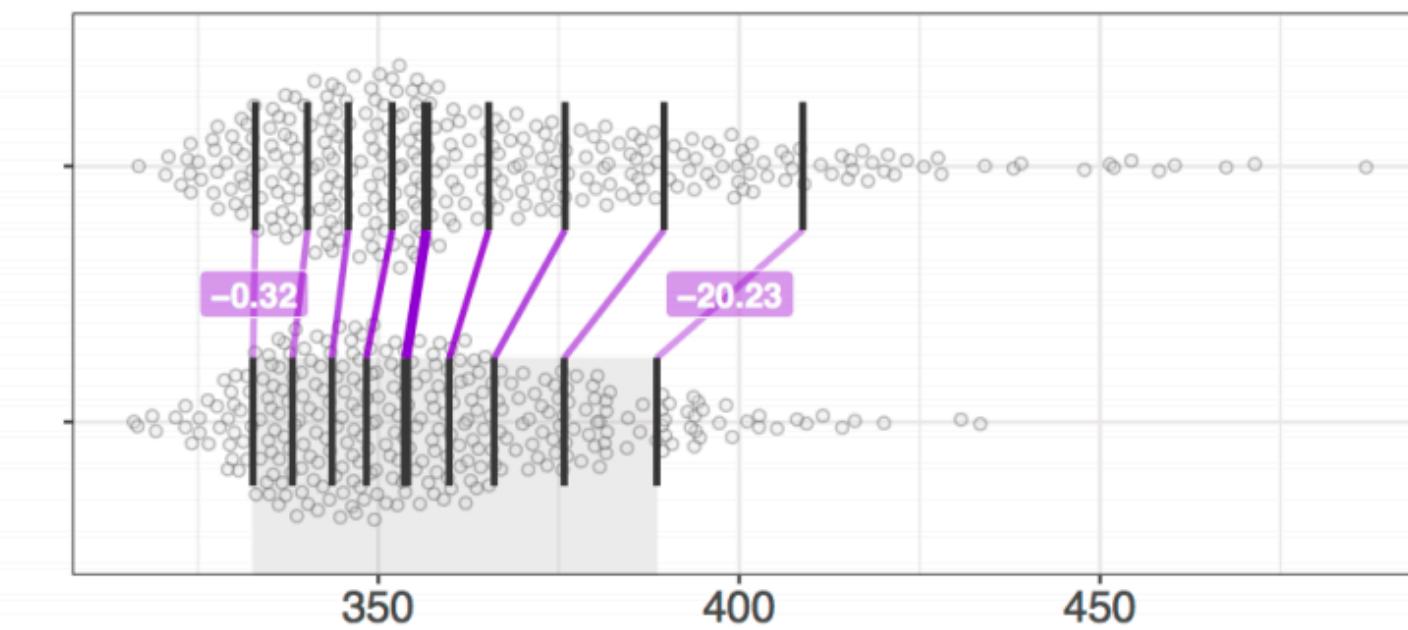


Для сравнения децилей распределения используется Harrell-Davis quantile estimator (в котором используются методы Monte-Carlo)

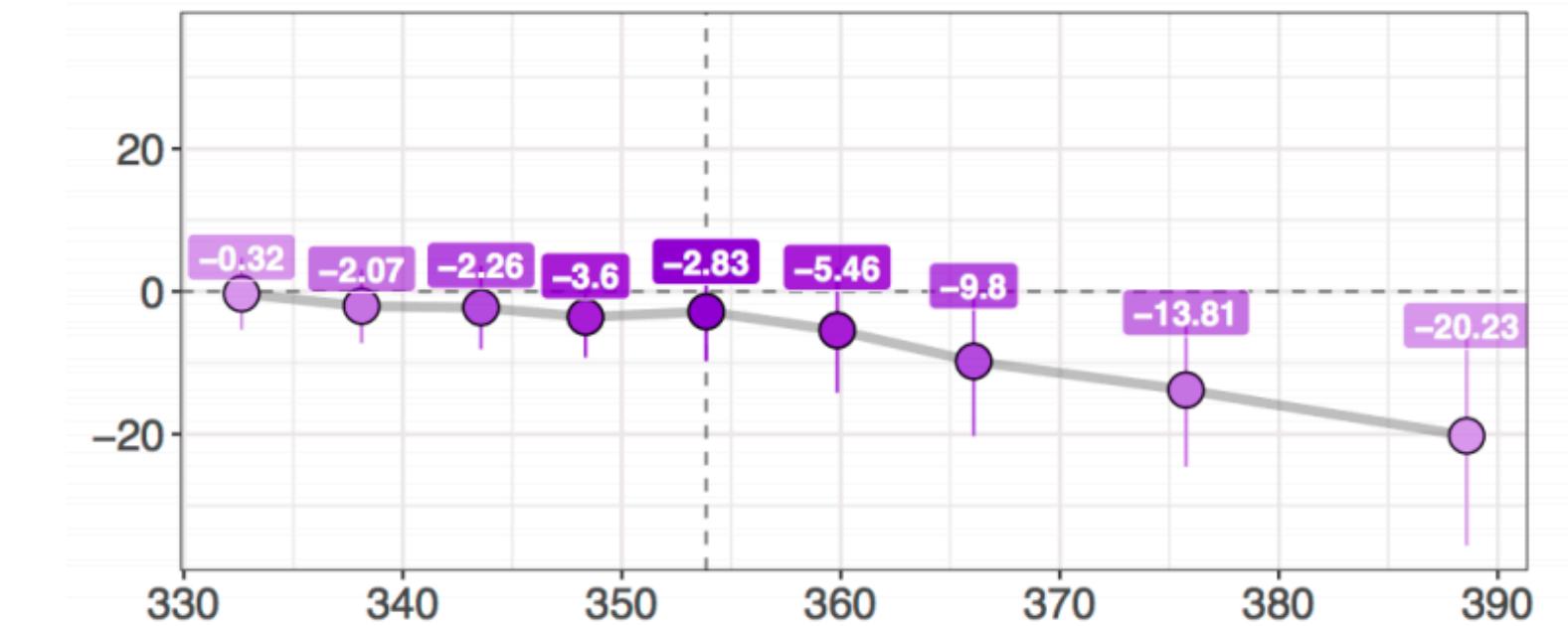
Распределения выборок



Сравнение децилей



Доверительный интервал для децилей



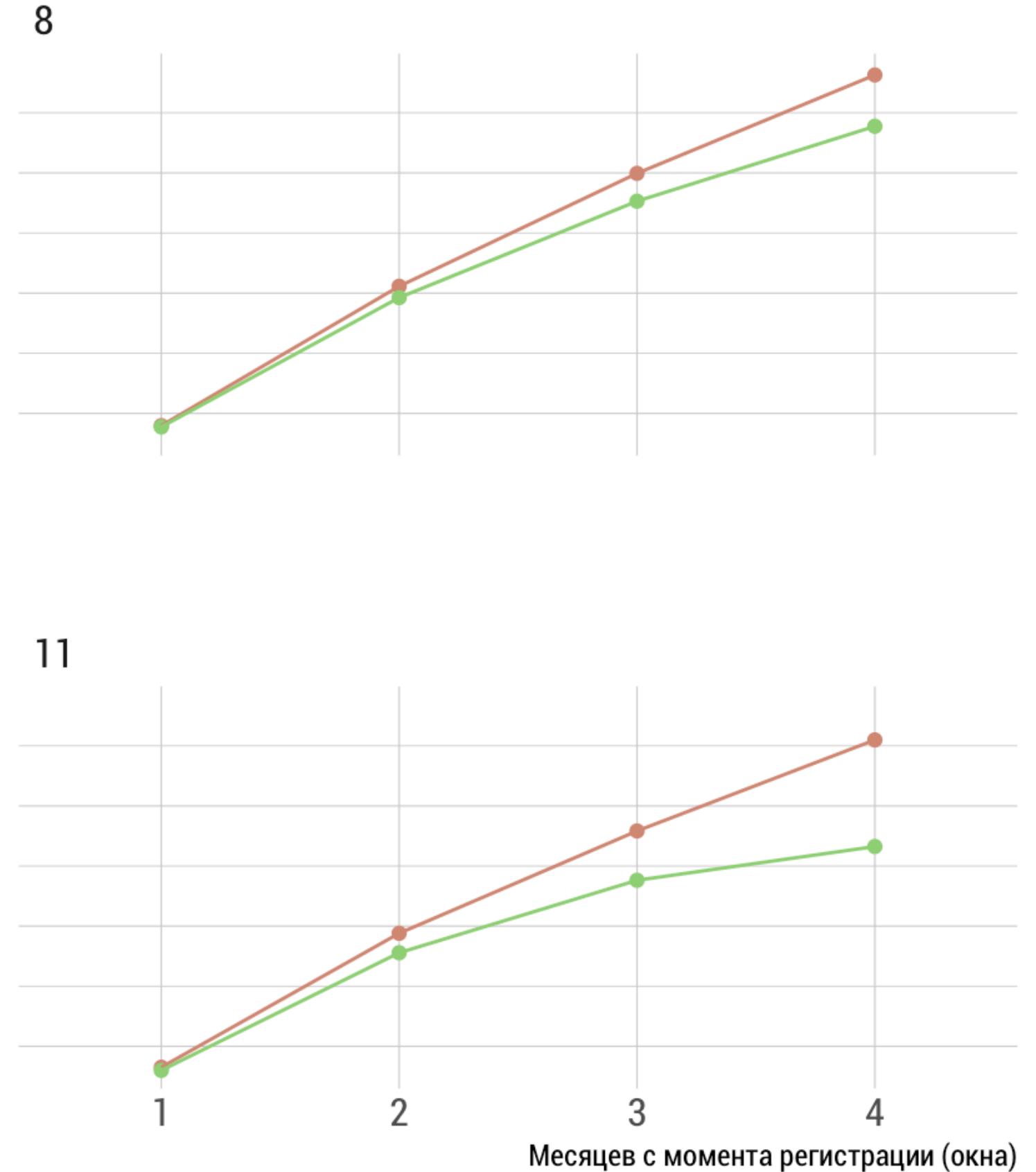
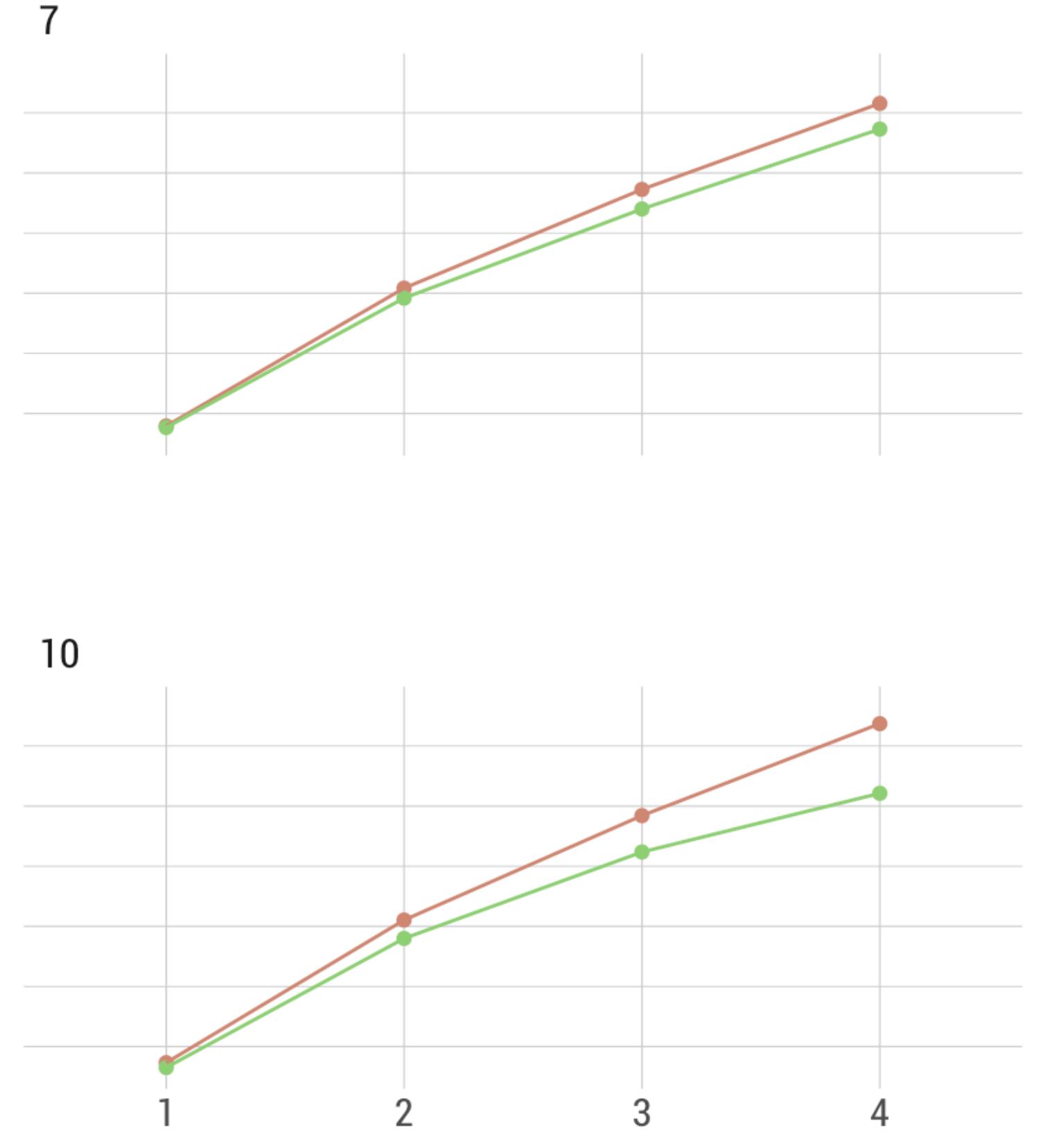
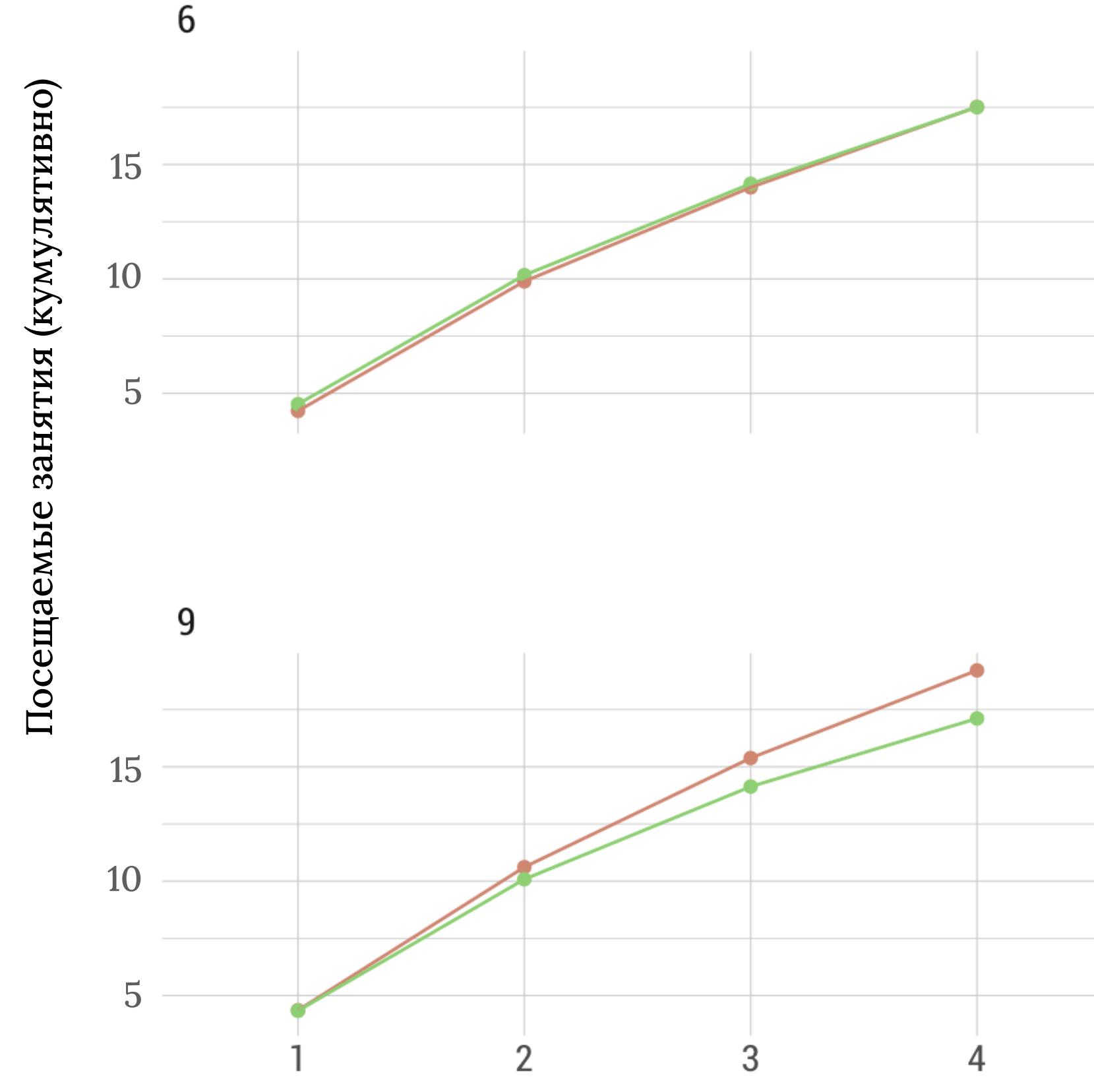
Bootstrap

*Ссылка: Harrell-Davis quantile estimator

e^xperiment fest

YoY закрытые занятия. В июне 2018 был добавлен новый пакет на 5 занятий

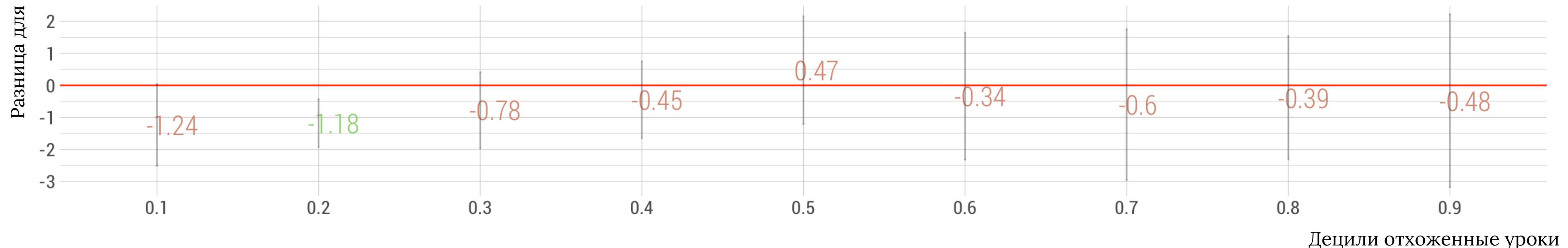
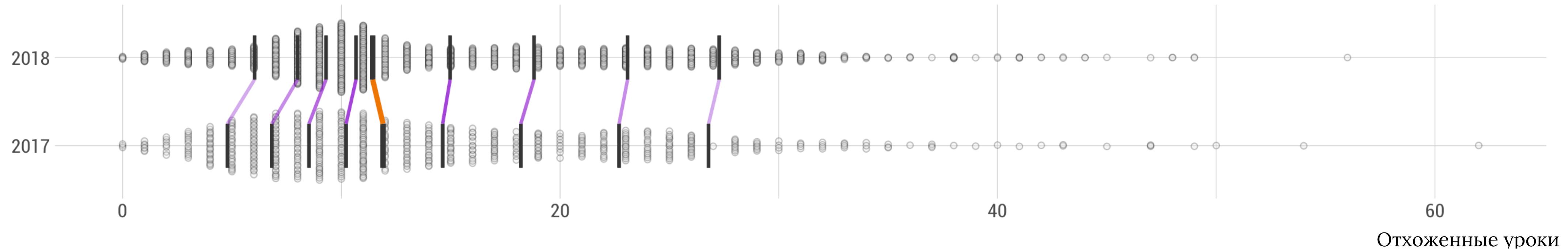
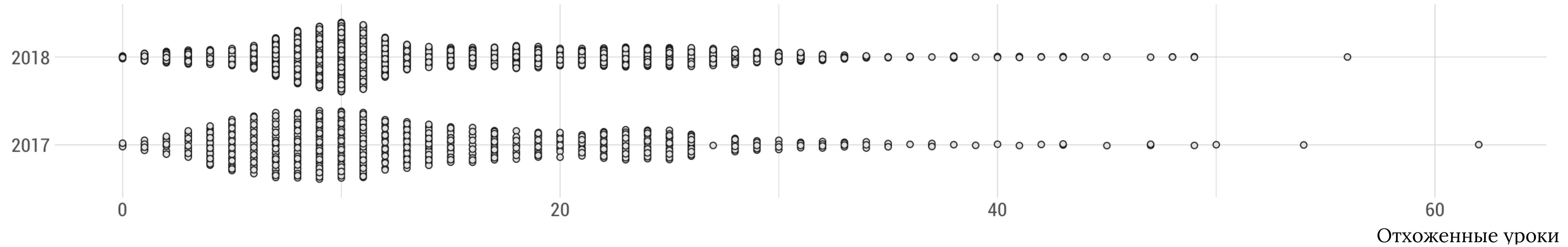
cohort_year 2017 2018



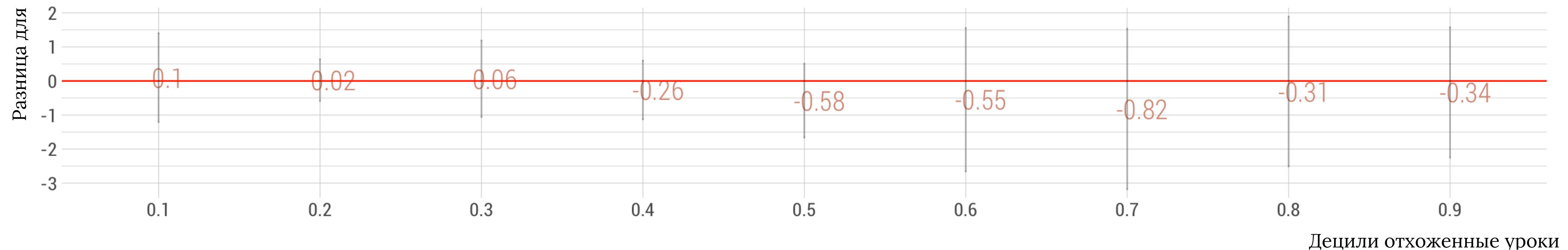
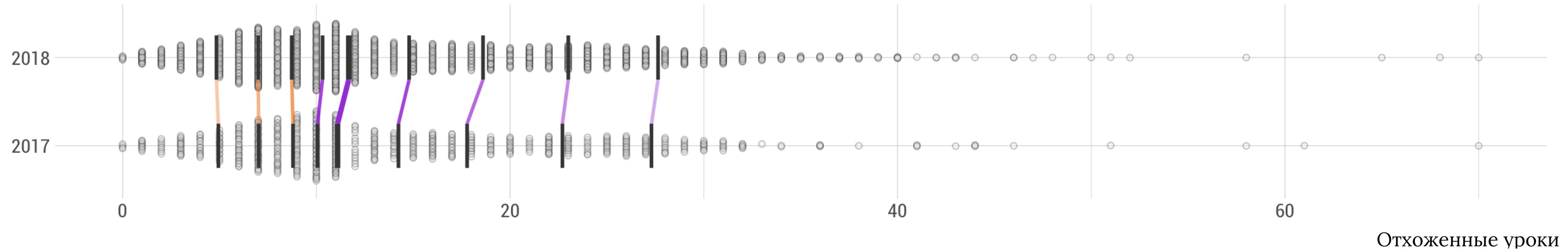
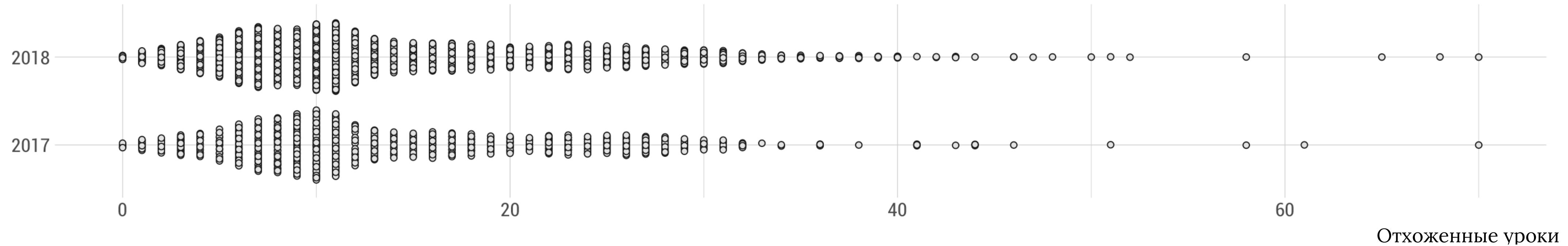
Bootstrap

e^xperiment fest

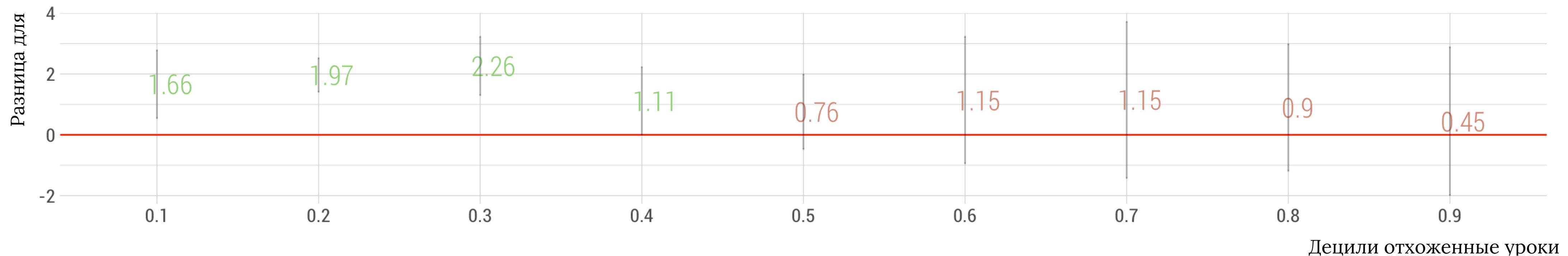
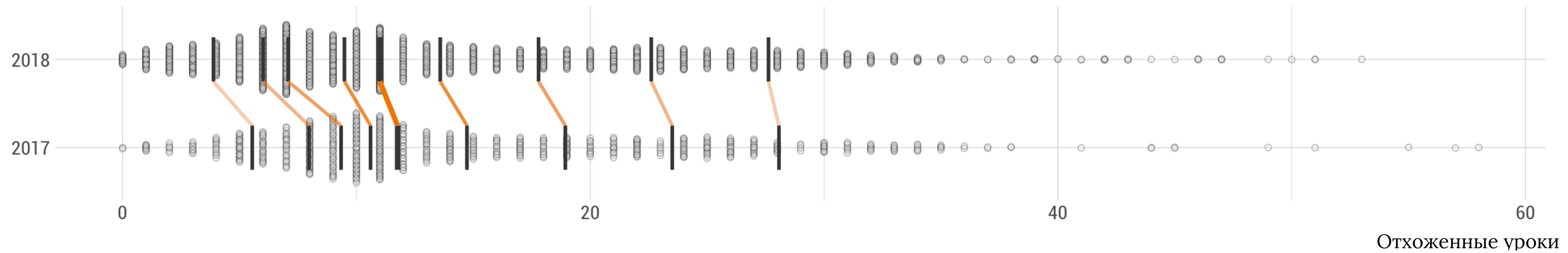
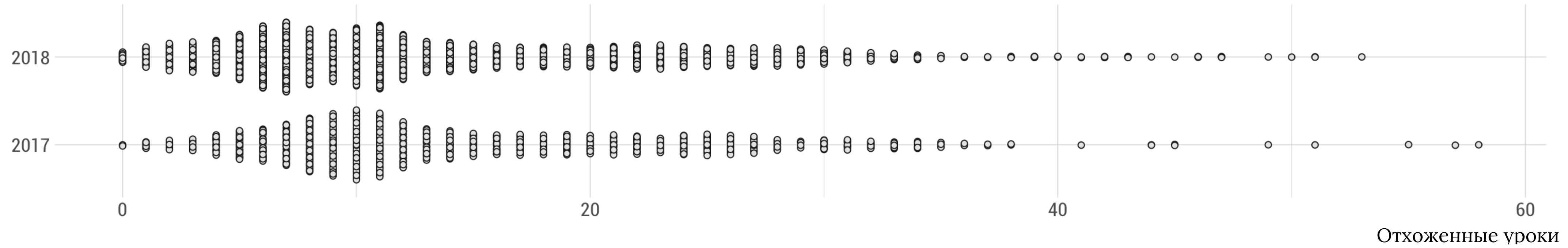
Сравнение отхоженных занятий по децилям (Июнь)



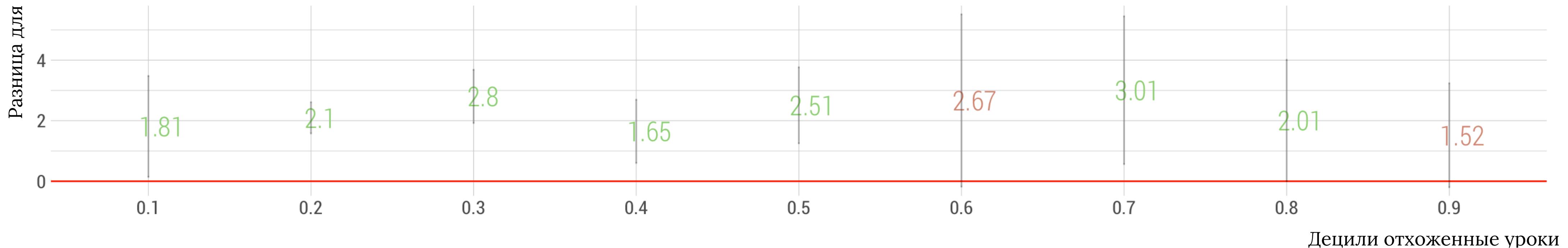
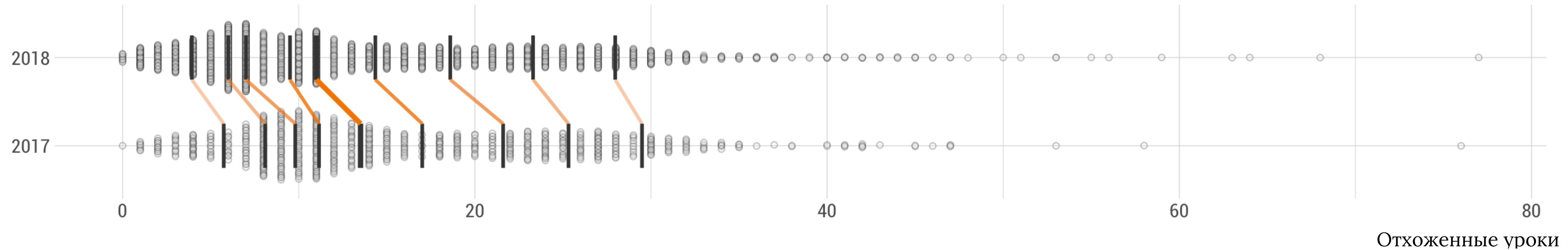
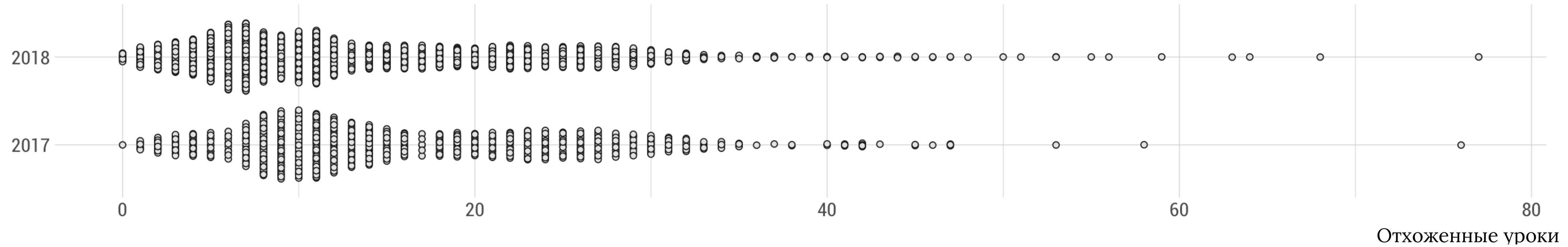
Сравнение отхоженных занятий по децилям (Июль)



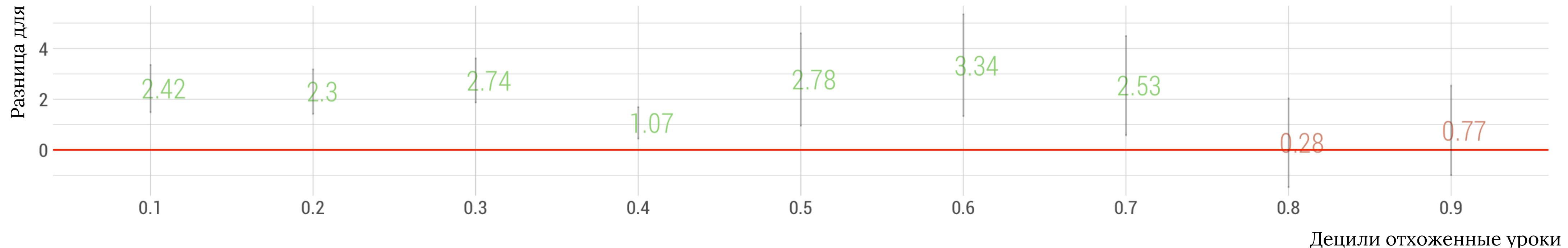
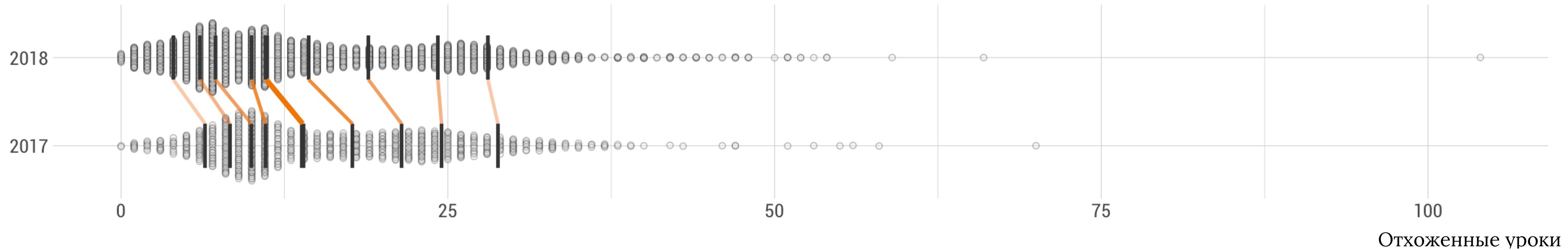
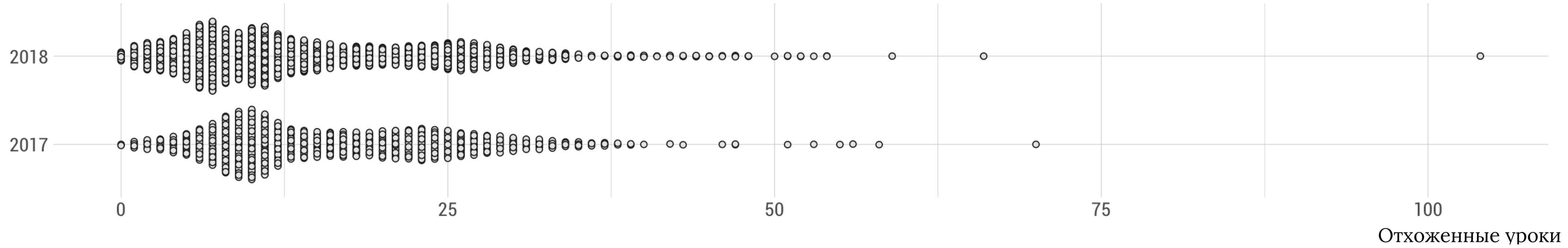
Сравнение отложенных занятий по децилям (Август)



Сравнение отхоженных занятий по децилям (Сентябрь)



Сравнение отхоженных занятий по децилям (Октябрь)



- Пакеты на 5 занятий существенно сдвинули распределение отхоженных занятий **влево** (т.е. уменьшилось)
- В первую очередь это коснулось той части распределения, что ниже **медианы**

День 3

Бакетирование

e^xperiment fest

Можно ли увеличить скорость
бутстрапа?

Проблема

Для бутстреп-выборок нужно задавать размер такой же, как и у изначальной выборки: изменение статистики будет зависеть от размера выборки.

Если мы хотим аппроксимировать эту изменчивость, нам нужно использовать повторные выборки одинакового размера.

Начиная с ~1 млн наблюдений начинаются проблемы, связанные со скоростью вычисления (долго ждать, пока посчитается)

Решение

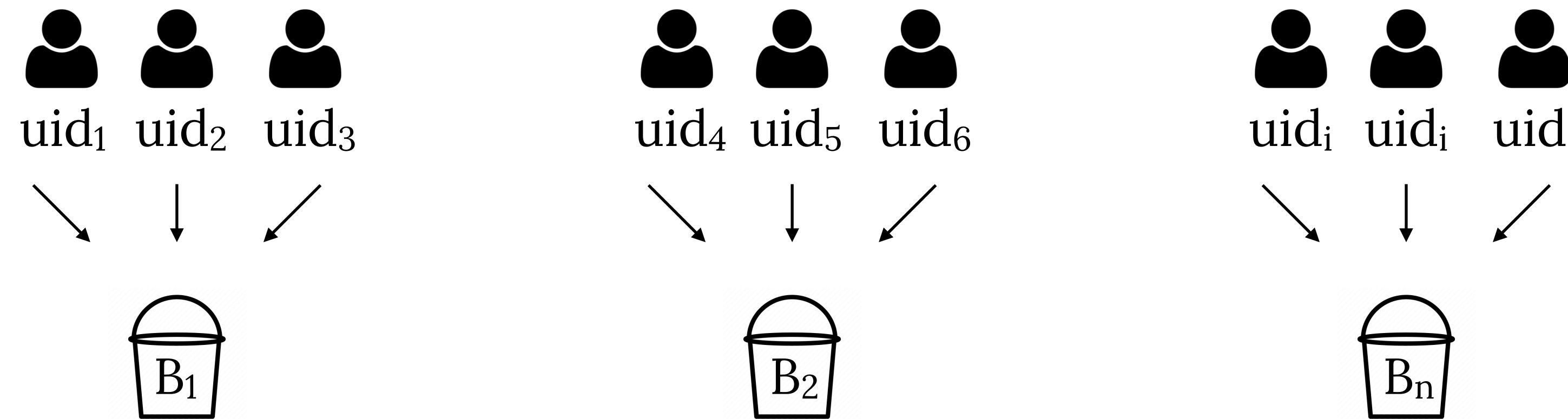
Использовать *бакеты* на корм бутстрапу

Бакетирование

e^xperiment fest

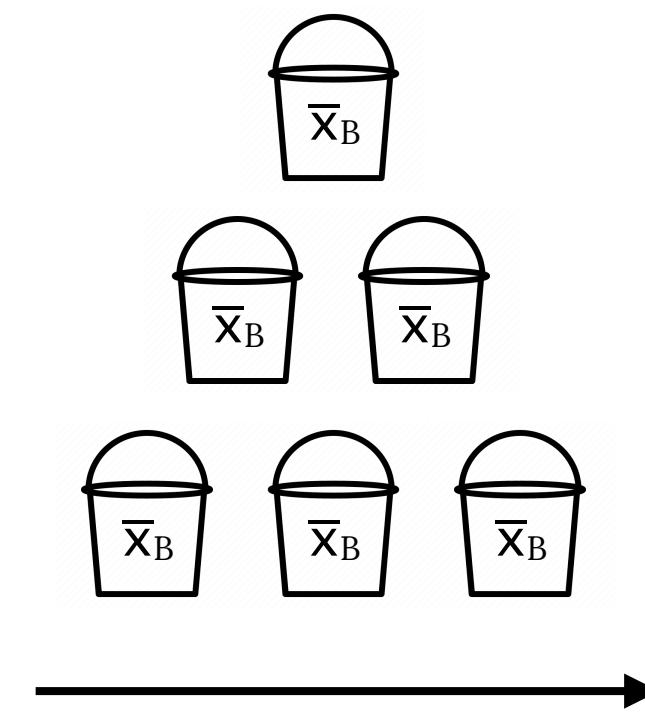
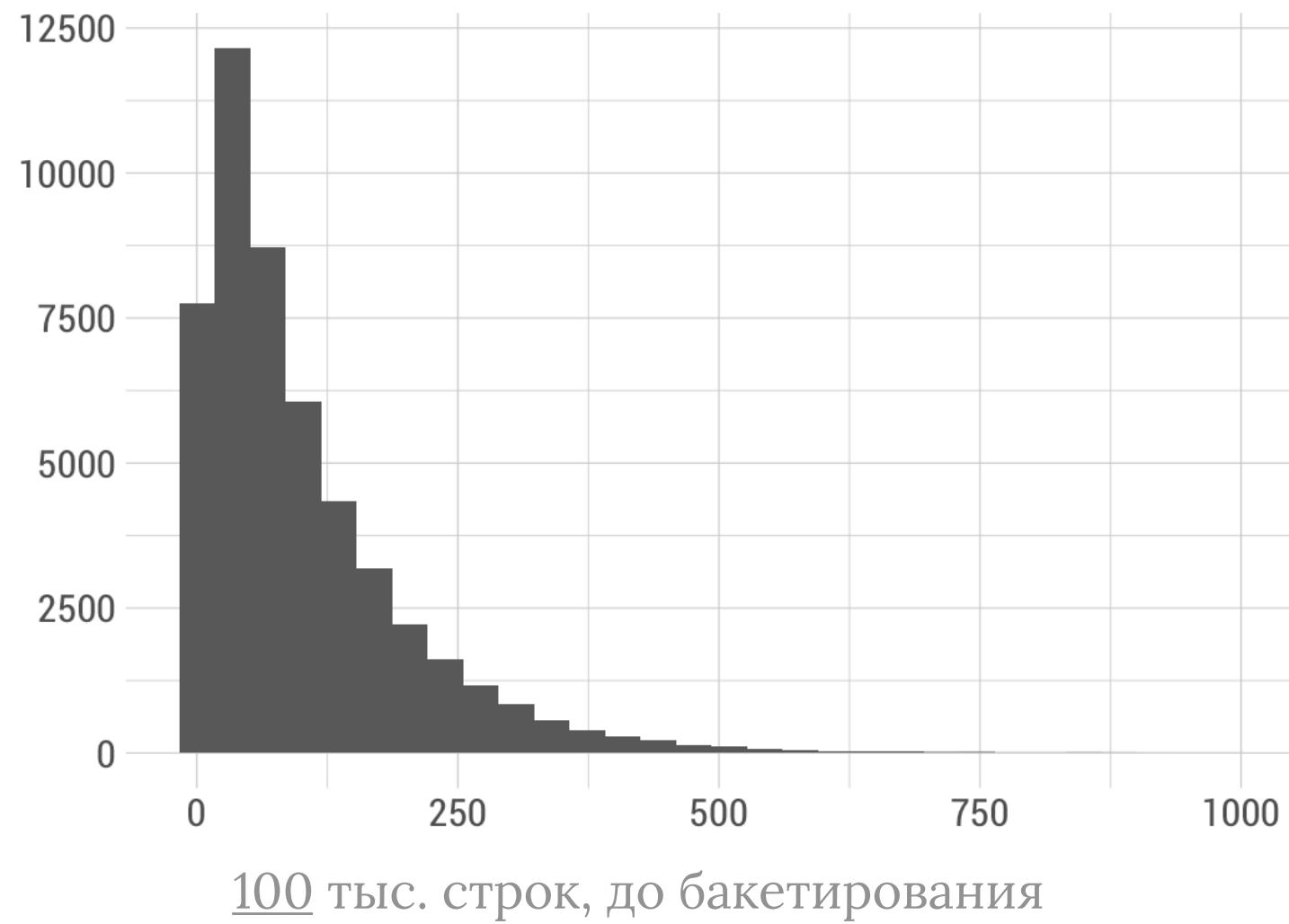
Распределение, отличающиеся от нормального, можно привести к нормальному с помощью техники бакетов:

1. Рандомно присваиваем номер группы от B_1 до B_n (где $B_n =$ оптимальное число групп, которые мы усредним, напр., 500)
2. Усредняем значения в каждой группе
3. Из усредненных значений получаем распределение близкое к нормальному



Распределение, отличающиеся от нормального, можно привести к нормальному с помощью техники бакетов:

1. Рандомно присваиваем номер группы от B_1 до B_n (где B_n = оптимальное число групп, которые мы усредним, напр., 500)
2. Усредняем значения в каждой группе
3. Из усредненных значений получаем распределение близкое к нормальному



Бакетирование

e^xperiment fest

- При таком подходе сохраняется условие о независимости наблюдений благодаря равномерному распределению по бакетам (в отличие от бутстрата)
- Сохраняется информация о метрике в исходящей выборке при ее меньшем размере (полезно для автоматизированных вычислений): дисперсия и сама метрика

$$\frac{s^2}{N} \approx \frac{s_b^2}{B}$$

- Бакетирование полезно когда наблюдений достаточно много (напр., > 1 млн)

Демонстрация в R

Бакетирование

e^xperiment fest

В 148 раз быстрее

- Симуляция при 1 млн строк
- Без бакетов = 40 сек; с бакетами = 0,27 сек
- Тот же самый результат (p-val, quants)

День 3

Проверка качества систем сплитования и А/А-тестирования

e^xperiment fest

Преимущественно, задача A/A тестов заключается
в том, чтобы понять, работает ли система
сплитования корректно или нет

Для чего A/A?

Убедиться в корректности системы сплитования
можно путем двухэтапной проверки:

- **Честное деление пользователей между группами.** Сохраняется репрезентативность по долям и дисперсии: сплитовалка не должна отдавать приоритет какой-либо из групп по какому-либо признаку, в силу чего может произойти дисбаланс -> изменение дисперсии и средних
- **Проверка FPR с помощью бизнес-метрик.** Частота ложных прокрасов метрики (например, конверсия и средний чек) не должна быть выше заданного уровня α

Этапы проверки А/А

1. **Проводим А/А тест.** Время на А/А определяется таким образом, чтобы охватить как можно больше факторов влияния на метрику (например, недельная сезонность)
2. **Симулируем новые А/А.** Тест пересчитывается ≥ 10 тыс. раз при помощи симуляции новых А/А
3. **Считаем стат. значимость.** В каждом тесте считается p-value при помощи статистического оценщика (бутстреп, т-тест и т.п.)
4. **Считаем метрику качества FPR (False Positive Rate)**
5. **Делаем выводы.** Проверяется условие $FPR < \alpha$, и если условие соблюдается, то сплитовалка работает корректно

Показатель FPR

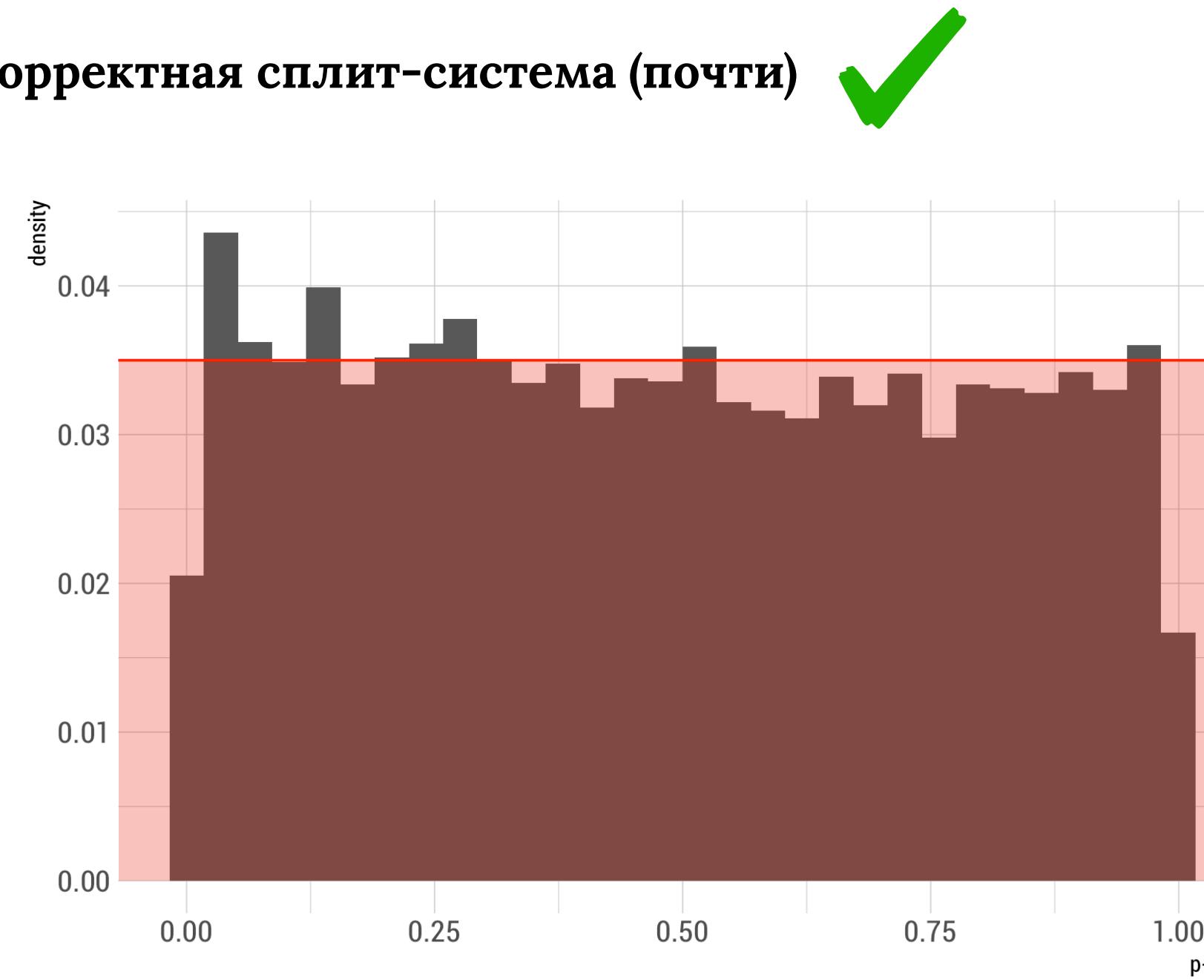
Для проверки качества сплитовалки считаем долю

$$\text{ложных прокрасов (FPR): } \frac{FP}{N} = \frac{FP}{FP + TN}$$

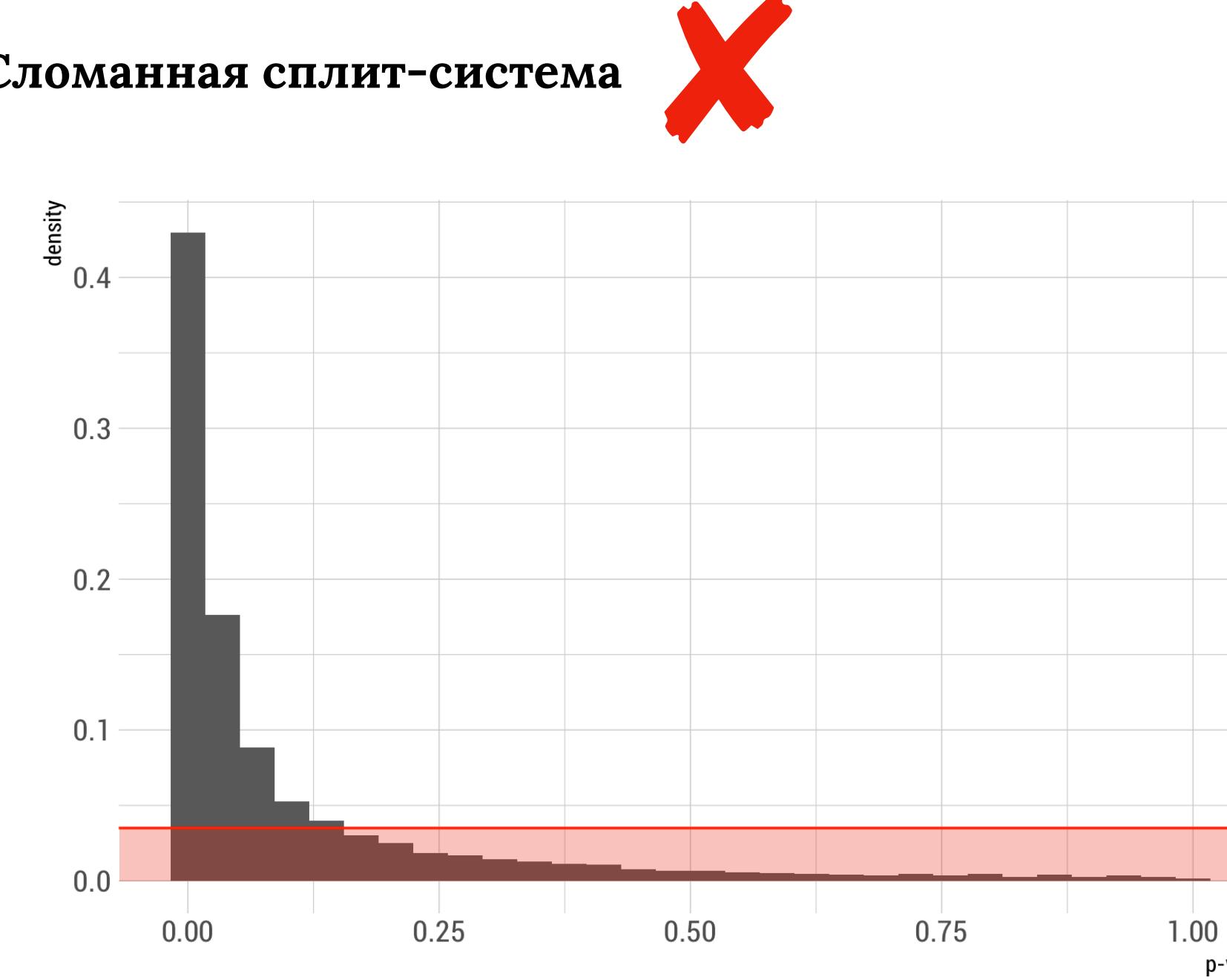
По сути, необходимо проверять FPR на каждом уровне значимости: частота ложных прокрасов не должна быть выше заданного уровня значимости. FPR не должен превышать 0.05 для $\alpha = 0.05$. Соответственно и для 0.01, 0.005 и т.п.

Показатель FPR

Корректная сплит-система (почти)



Сломанная сплит-система



Красная закрашенная область – uniform теоретическое распределение α .

Если бины выше или ниже красной линии, то что-то не так и нужно искать причины.

Проверка качества систем сплитования и A/A-тестирования

e^xperiment fest

Завышенный FPR

Техническая реализация

Основные причины кроются в сломанном сплит-алгоритме.

Причины необходимо искать на стороне где реализован скрипт и его запуск. Частые кейсы:

- Долгое ожидание ответа сервера по присвоению id эксперимента и сплита
- Приоритет той или иной группе
- Не на всех страницах / кейсах реализован сплит-алгоритм
- Банально «сломан» рандом (остаток от деления по сумме хеша?)

Поиск возможной причины

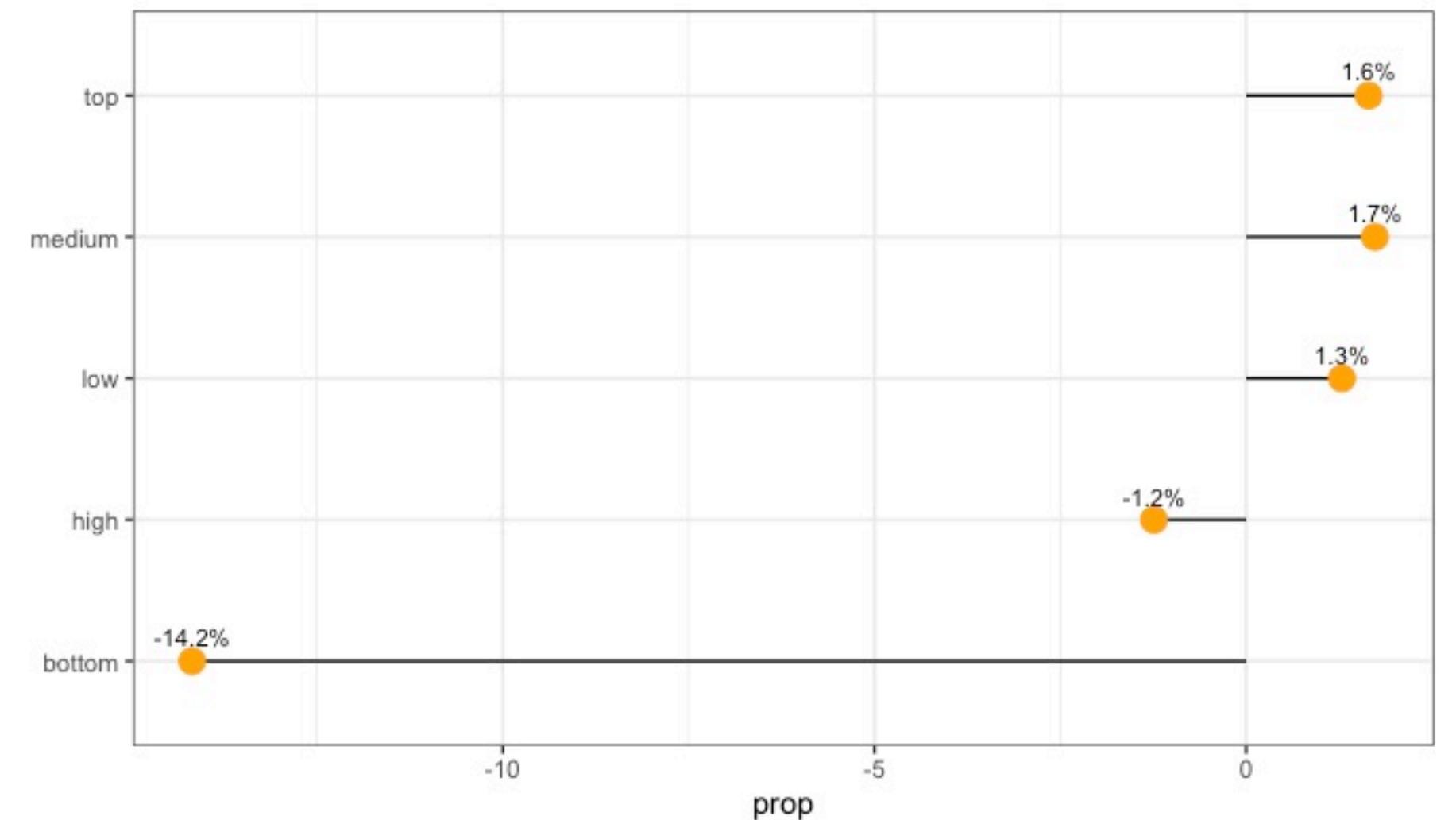
Дисбаланс в группах по описательным признакам.

Первая возможная причина нарушения условия $FPR < \alpha$.

Для поиска дисбаланса необходимо сравнить распределившиеся доли между группами по признакам. Вполне подойдут:

- регионы
- источники трафика
- браузер и т.п.

Сравнение долей RFM сегментов по 2 сплитам (должно быть 0% или незначительное отклонение)



Проверка качества систем сплитования и A/A-тестирования

e^xperiment fest

Поиск возможной причины

Критерий Кохрана-Мантеля-Ханзеля для проверки дисбаланса

Для проверки фактических долей с их теоретическим равномерным распределением используются специализированные критерии согласия.

В ситуации с А/А подойдет критерий СМН (Cochran-Mantel-Haenszel) для проверки таблиц сопряженности $2 \times 2 \times K$,

где К – количество градаций по анализируемому признаку (например браузер 1, браузер 2 и т.п.)

Ограничения и другие моменты

- A/A желательно проводить как можно дольше, чтобы достичь достаточной репрезентативности (охватить недельную сезонность и разные группы пользователей)
- В случае, если нет возможности ждать, то не рекомендуется использовать долгоиграющие метрики для проверки сплита (например, C2)
- Пост-симуляции нужно делать без возвращения наблюдений в сплитах
- Для пост-симуляций лучшим образом подойдет бутстррап, благодаря своей точности

Конец третьего дня

e^xperiment fest

Мирмахмадов Искандер

Черемисинов Виталий

07/2020

experiment-fest.ru