experiment fest

Бутстрап

повторные выборки и децильные методы оценки А/Б тестов

Что узнаем?

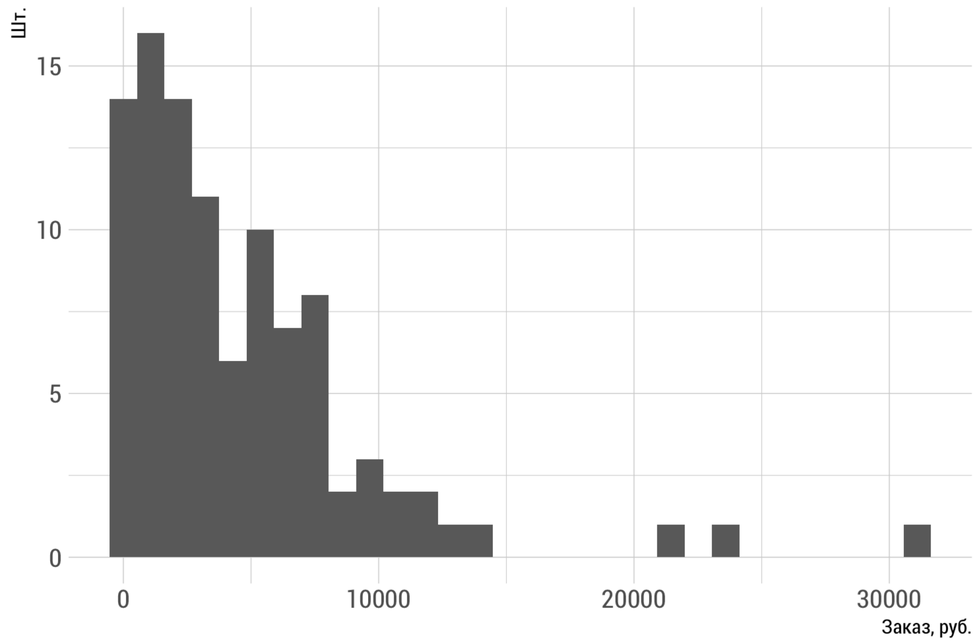
- Что такое бутстрап?
- Оценка среднего и медианы
- Проверка гипотез с помощью бутстрапа

Что такое бутстрап?

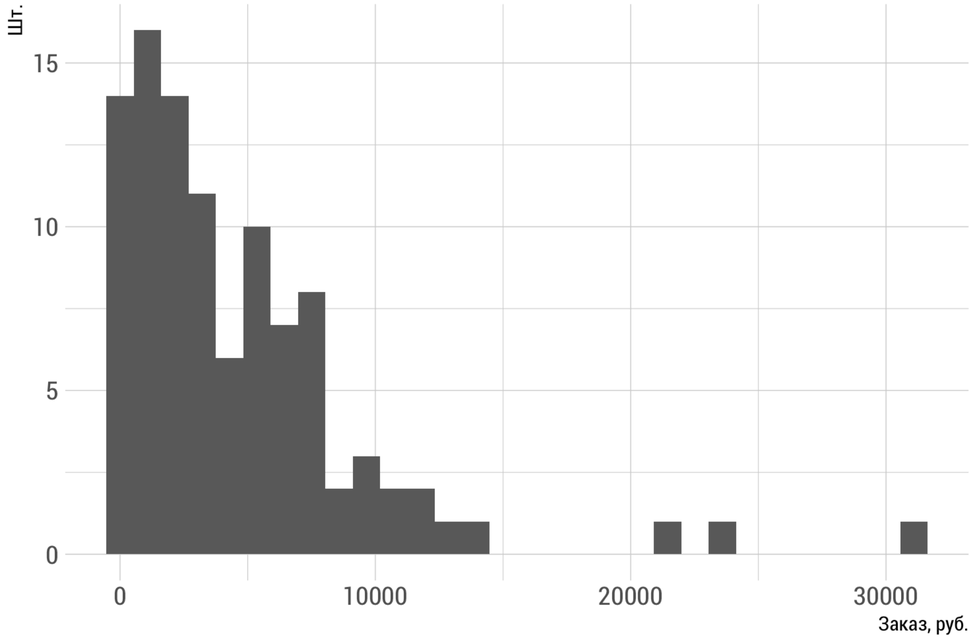
Интернет-магазин продуктов

На графике покупки (заказы) и их сумма

Номер заказа	Сумма, руб
1	2438.86
80	1168.33
85	432.22
37	7971.52
16	31092.87
…	…



Как лучше оценить?



Что здесь лучше будет отражать центральную тенденцию? – Медиана

Можем ли использовать ЦПТ, чтобы построить ДИ для медианы? – Нет. Тогда как?

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

Выборка	**Повторная бут-выборка**
2.11	3.12
9.4	
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	
9.4	
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
9.4	
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
	9.4
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
9.4	9.4
17.9	
8.6	
4.1	

Выборка	Повторная бут-выборка
	3.12
3.12	9.4
9.4	9.4
17.9	2.11
8.6	
4.1	

Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
9.4	9.4
17.9	2.11
8.6	4.1

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	

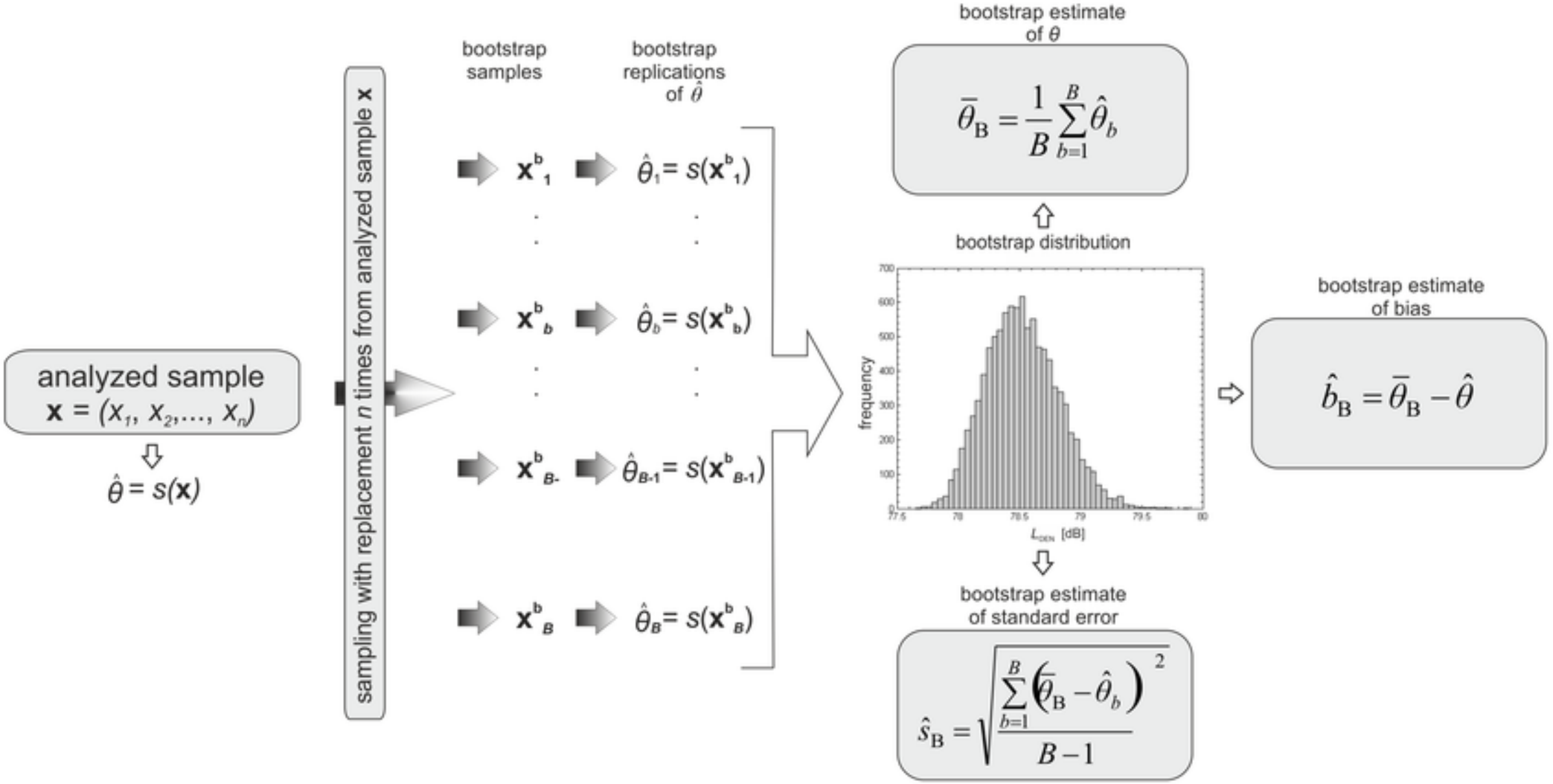
Выборка	Повторная бут-выборка
2.11	3.12
3.12	9.4
9.4	9.4
17.9	2.11
	4.1
4.1	8.6

Медиана	*Медиана*
6.35	6.35
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

Медиана	*Медиана*	*Медиана*	*Медиана*	*Медиана*	
6.35	6.35	5.86	6.35	3.61	
Выборка	**Повторная бут-выборка**	**Повторная бут-выборка**	**Повторная бут-выборка**	**Повторная бут-выборка**	...
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	

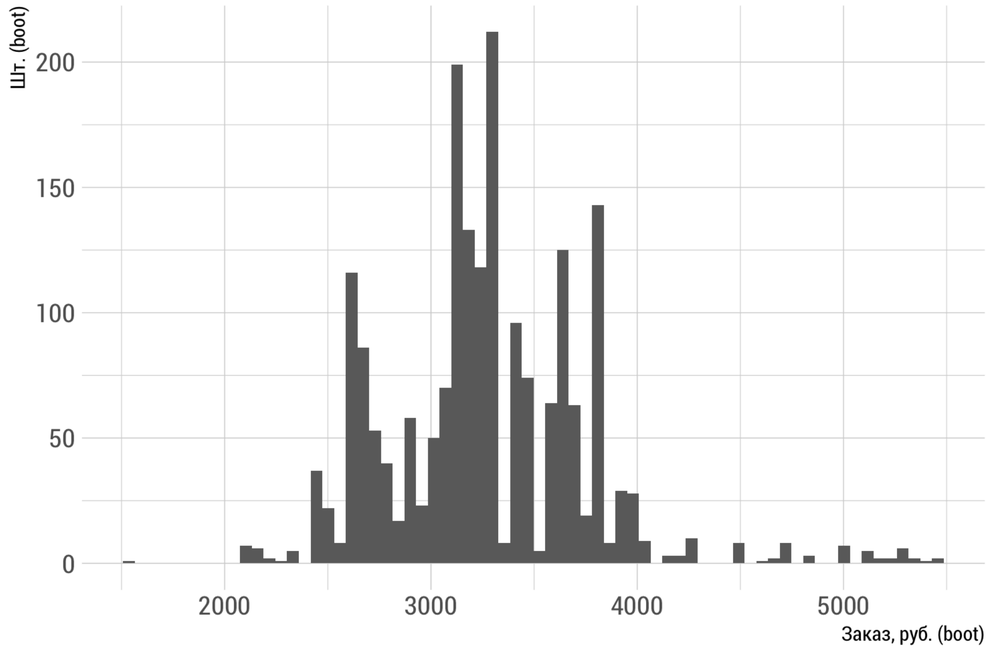
Бутстрап распределение медиан

6.35 5.86 6.35 3.61 → считаем доверительный интервал → получаем эмпирическую оценку параметра распределения (медианы)


analyzed sample
x = (x1, x2,..., xn)
θ̂ = s(x)
sampling with replacement n times from analyzed sample x
bootstrap samples
bootstrap replications of θ̂
x^b_1
θ̂1 = s(x^b_1)
x^b_b
θ̂b = s(x^b_b)
x^b_B-
θ̂B-1 = s(x^b_B-1)
x^b_B
θ̂B = s(x^b_B)
bootstrap estimate of θ
θ̄B = 1/B Σ_{b=1}^{B} θ̂b
bootstrap distribution
frequency
L_DEN [dB]
bootstrap estimate of bias
b̂B = θ̄B − θ̂
bootstrap estimate of standard error
ŝB = √(Σ_{b=1}^{B} (θ̄B − θ̂b)^2 / (B − 1))

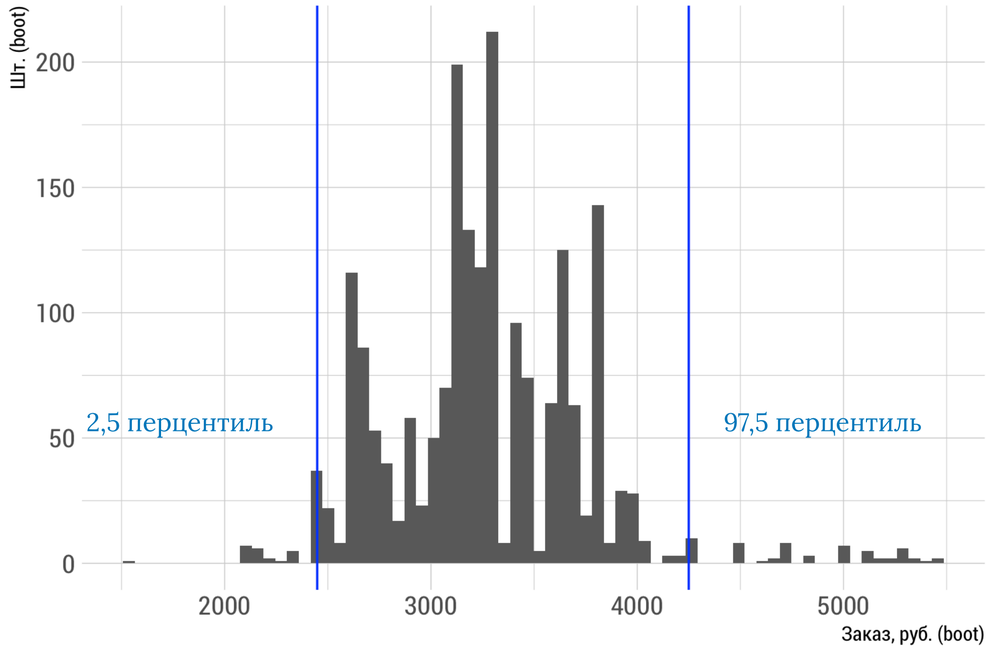
Демонстрация

Распределение медиан (медиана = 3268.88)



Распределение медиан (медиана = 3268.88)

95% доверительный интервал [2448.51, 4249.82]



Перцентильный метод ДИ:

Берем 95% площади распределения (для 95% уровня значимости):

1. Ищем перцентиль 2,5 в бут-распределении $\frac{\alpha}{2}$
2. Ищем перцентиль 97,5 в бут-распределении $\frac{1-\alpha}{2}$

Это и будет ДИ

Демонстрация python

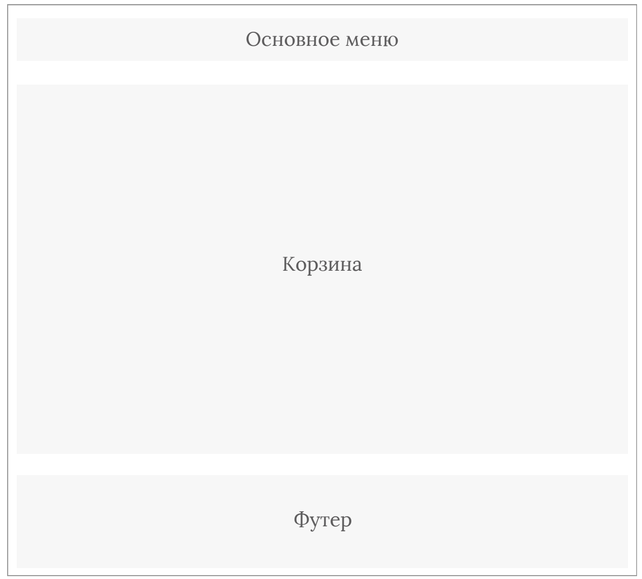
Другие способы расчета ДИ:

https://arch.readthedocs.io/en/latest/bootstrap/confidence-intervals.html

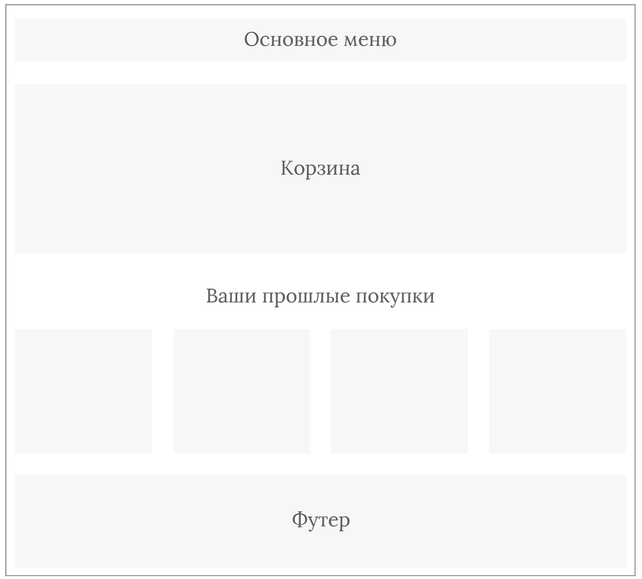
- Percentile CI
- Normal CI
- Basic CI
- BCA

Проверка гипотез с помощью бутстрапа

Контроль



Тест



Интернет магазин продуктов

Добавили новую витрину «Ваши прошлые покупки» на чекаут.

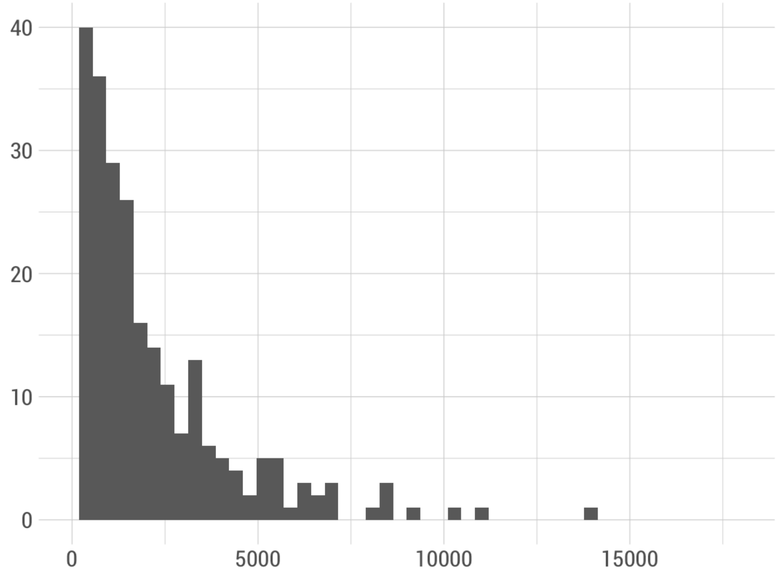
По классике, бизнесу интересно узнать как изменился средний чек

Как оценить влияние эксперимента на прибыль?

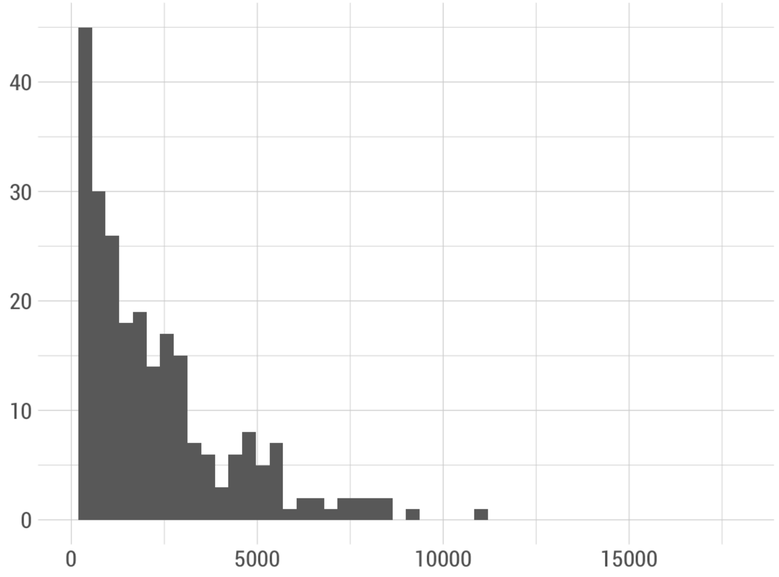
$$Average\ revenue_{control} = 5253$$

$$Average\ revenue_{test} = 5486$$

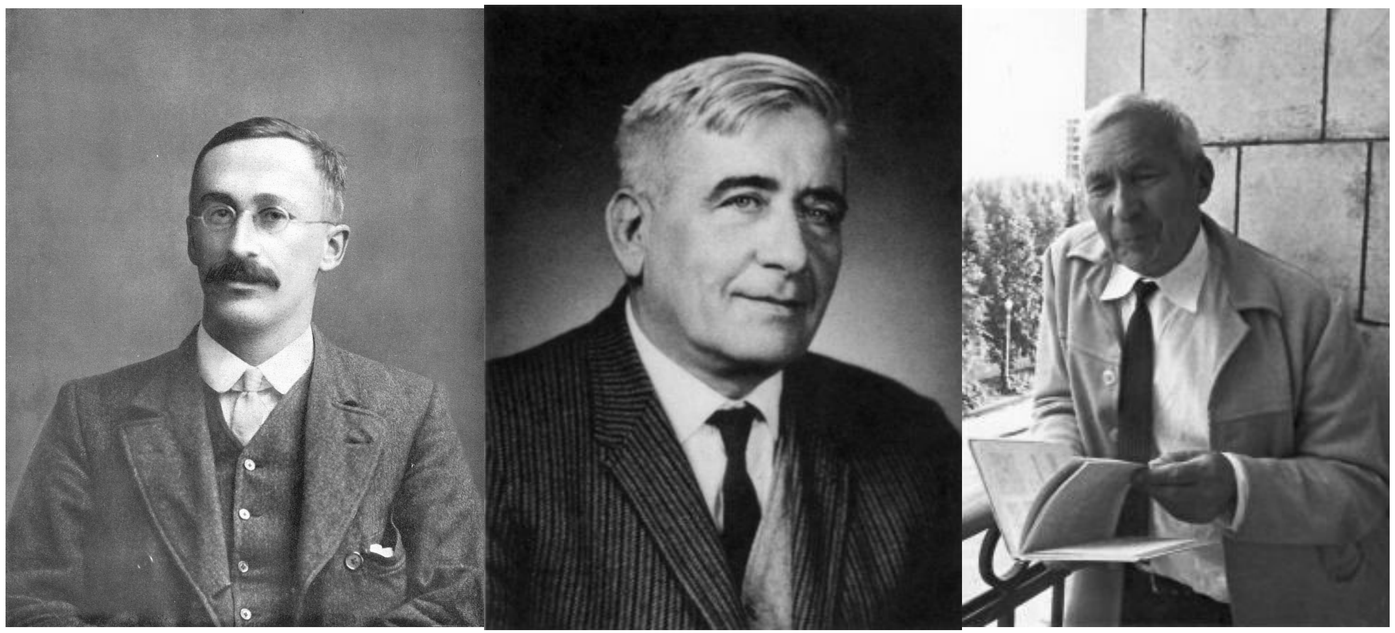
Avg revenue control



Avg revenue test



Что нам мешает просто использовать классический критерий?



Что нам мешает просто использовать классический критерий?

- Манна-Уитни лучшего всего подойдет для задачи. Он дает ответ на вопрос, значимо ли различаются распределения или нет. Хотелось бы понимать **где именно** эта разница

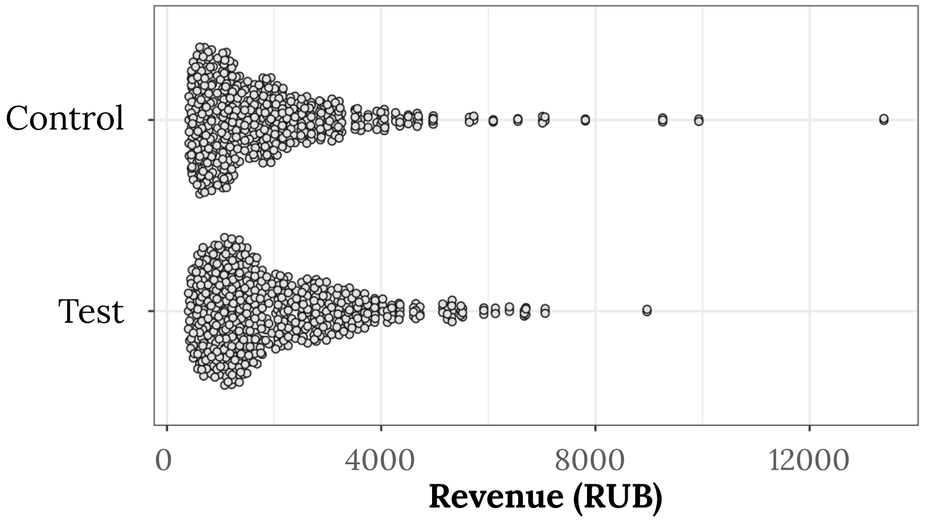
- К тому же, у каждого критерия свое аналитическое решение, которое требует придерживаться ряда допущений (например, одинаковая дисперсия/ одинаковый размер выборки/одинаковая форма распределений и т.п). Такая возможность не всегда имеется

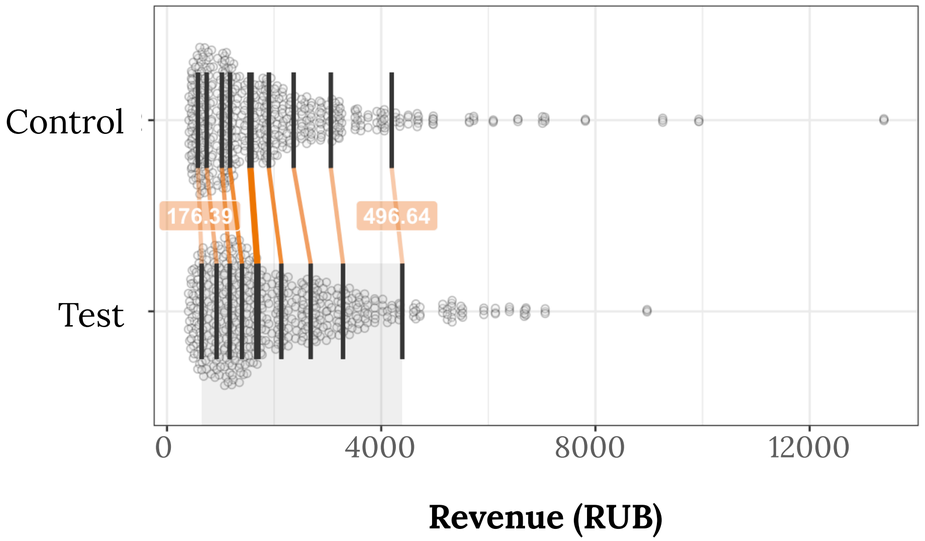
Мы можем пойти дальше и проверить каждую дециль распределения

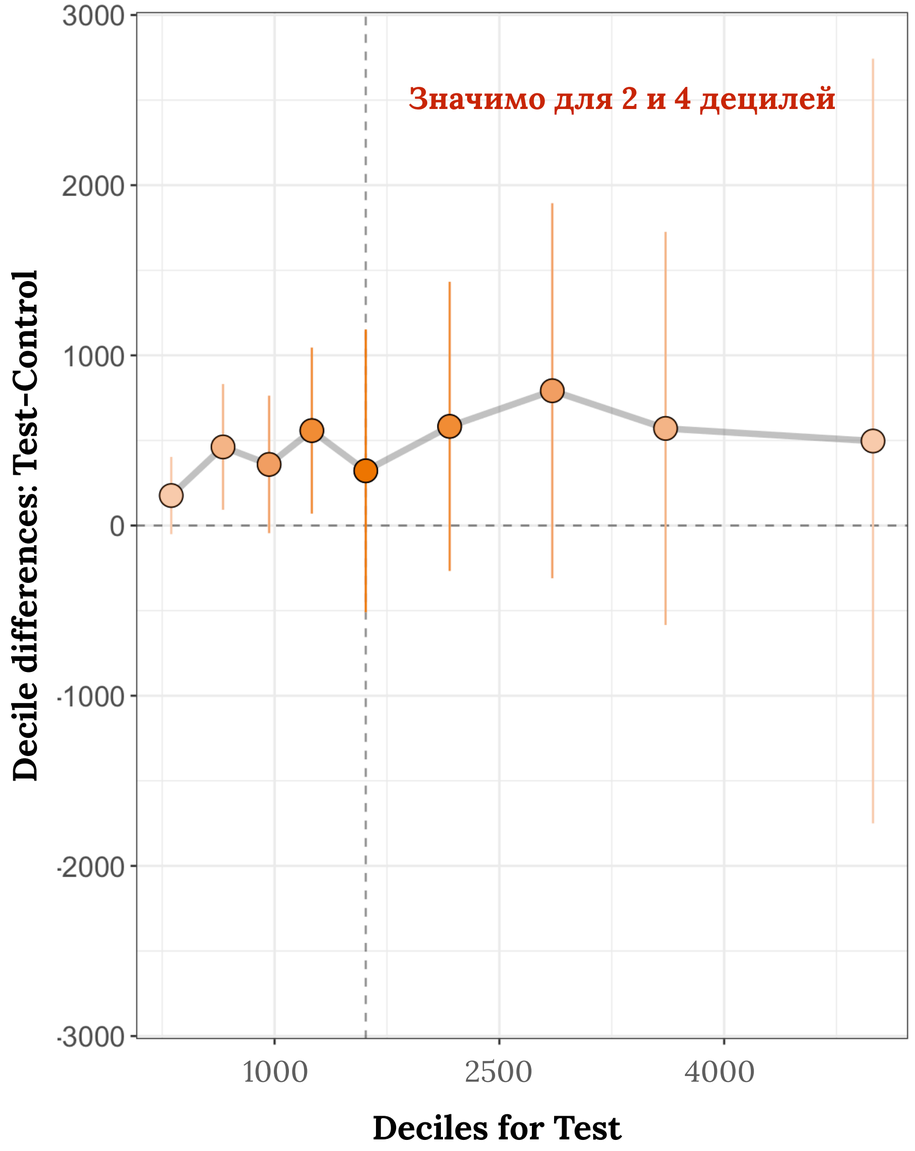
Как проверяются гипотезы с помощью бутстрапа?

1. Строите бутстрап-распределения параметра в А и Б
2. Вычисляете их разницу
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 python

Рекомендации, ограничения и выводы

Выводы

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

Ограничения

- Бутстрап не требует соблюдать предположения, но ему все еще требуется достаточно репрезентативная выборка
- Чем выше количество бутстрап итераций, тем дольше придется ждать результат расчета
- Чем исходная выборка, тем дольше придется ждать результат расчета

Также бутстрап

- Не способ нормализации выборки
- Не способ получения большего количества наблюдений
- Не дает больше получить больше информации об исходной выборке

Можно ли нормализовать распределение метрики с помощью бутстрапа, 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 $X_1, X_2, \ldots, X_n$, 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 бутстрапа так построен, что одно наблюдение может встретиться много-много раз в одном и том же распределении

e^xperiment $f$est

Мирмахмадов Искандер