KARPOV.COURSES >>>
КОНСПЕКТ

> Конспект > 9 урок > СТАТИСТИКА

> Оглавление

Скачать Презентацию

> Что такое бутстрап: введение

Бутстрап – целое семейство методов, позволяющее проверять гипотезы с помощью повторных выборок.

Пример: кейс из фуд-ритейла

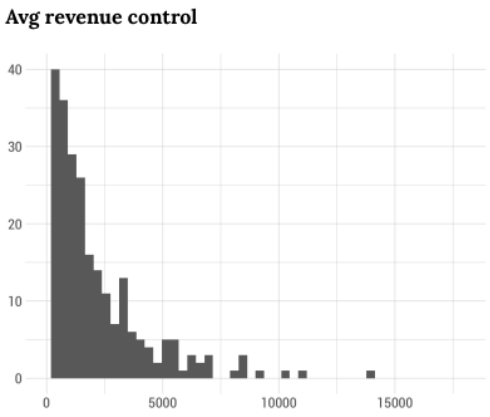
Добавили новую витрину “Ваши прошлые покупки” на чекаут. Интересно, как изменился средний чек. Как оценить влияние эксперимента на прибыль? Можем

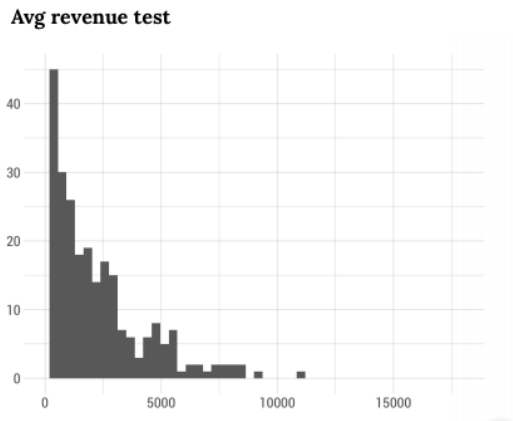
посмотреть на распределение, описательные статистики, подобрать тест и проверить значимость различий.

На глаз кажется, что в тестовой группе мы стали зарабатывать больше. Но: бизнесу захочется понять, чем объясняется эта изменчивость (разница, которую мы наблюдаем), за счет какой аудитории достигли эффекта?

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

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





Что можем сделать дальше?

Что лучше будет отражать центральную тенденцию? Медиана.

Можем ли использовать ЦПТ, чтобы построить ДИ для медианы? На данный момент не сможем, поэтому нужен какой-то метод для этой задачи.

> Что такое 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.

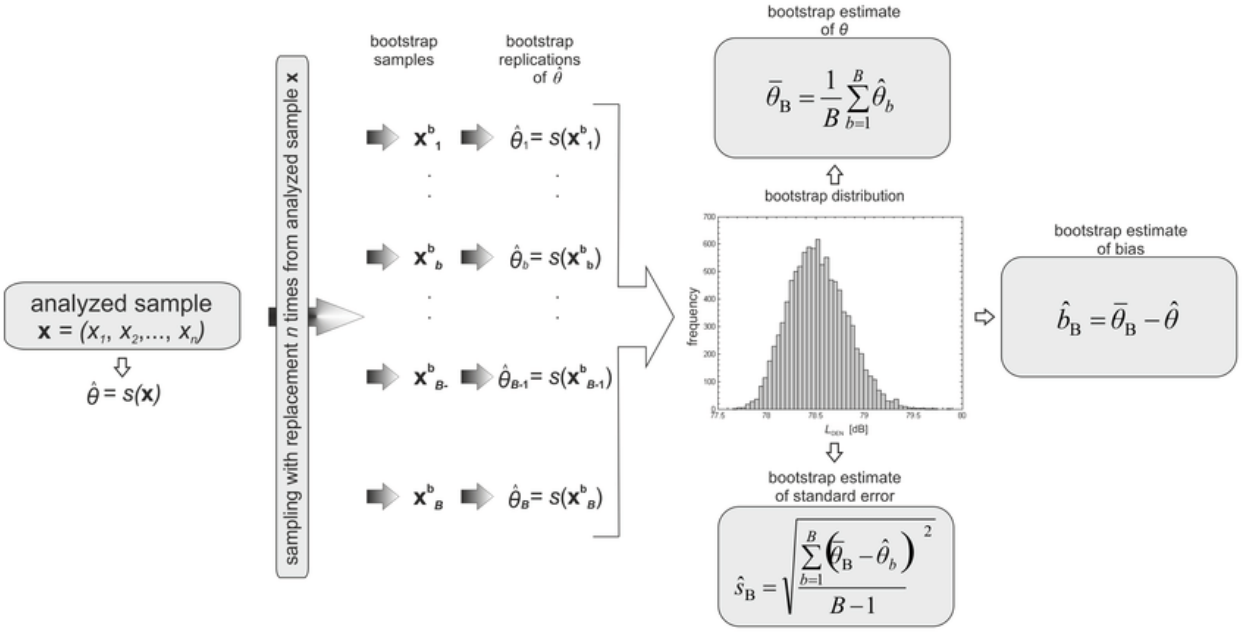
1. Берем значение из выборки, кладем в "корзину" (потом вернем в исходную выборку)
2. Возвращаем, повторяем операцию – снова берем какое-либо значение из выборки

3. Опять возвращаем, тем самым повторяя это boot-количество раз
4. По повторной бут-выборке считаем среднее (или другой интересующий параметр)

Среднее	*Среднее*	*Среднее*	*Среднее*	*Среднее*	
7.53	6.12	6.12	5.98	4.9	
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	

Создаем несколько бут-выборок, для каждой считаем среднее. Обладая средними мы можем взять распределение средних.

Бутстрап распределение средних (6.12, 6.12, 5.98, 4.9) → затем считаем доверительный интервал → получаем эмпирическую оценку параметра распределения.



> Что нам мешает просто использовать классический критерий?

Зачем считать средние, что-то сравнивать, если можно просто применить какой-либо критерий?

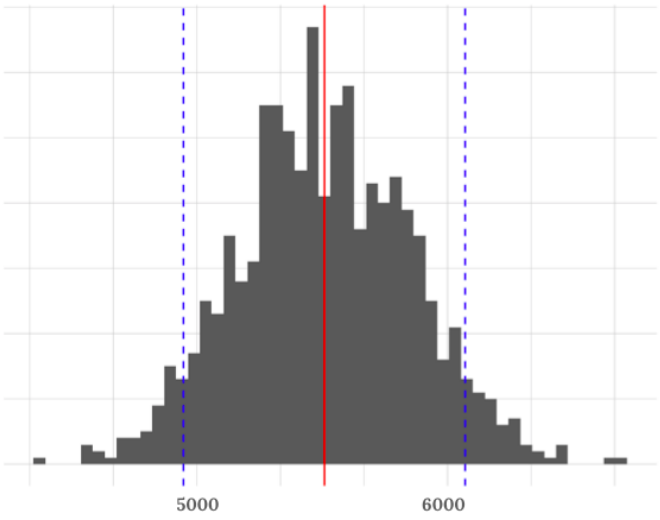
- В данном случае критерий Манна-Уитни лучшего всего подойдет для задачи. Он дает ответ на вопрос, значимо ли различаются распределения или нет. **Но:** хотелось бы понимать **где именно** эта разница
- К тому же, у каждого критерия свое аналитическое решение, которое требует придерживаться ряда допущений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

Если бы мы бутстрапировали среднее из примера, то получили бы следующую картину:

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

Помним, что

$Average\ revenue_{test} = 5486$



- bias = 5
- Взяли 10000 бутстрап-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

Итого:

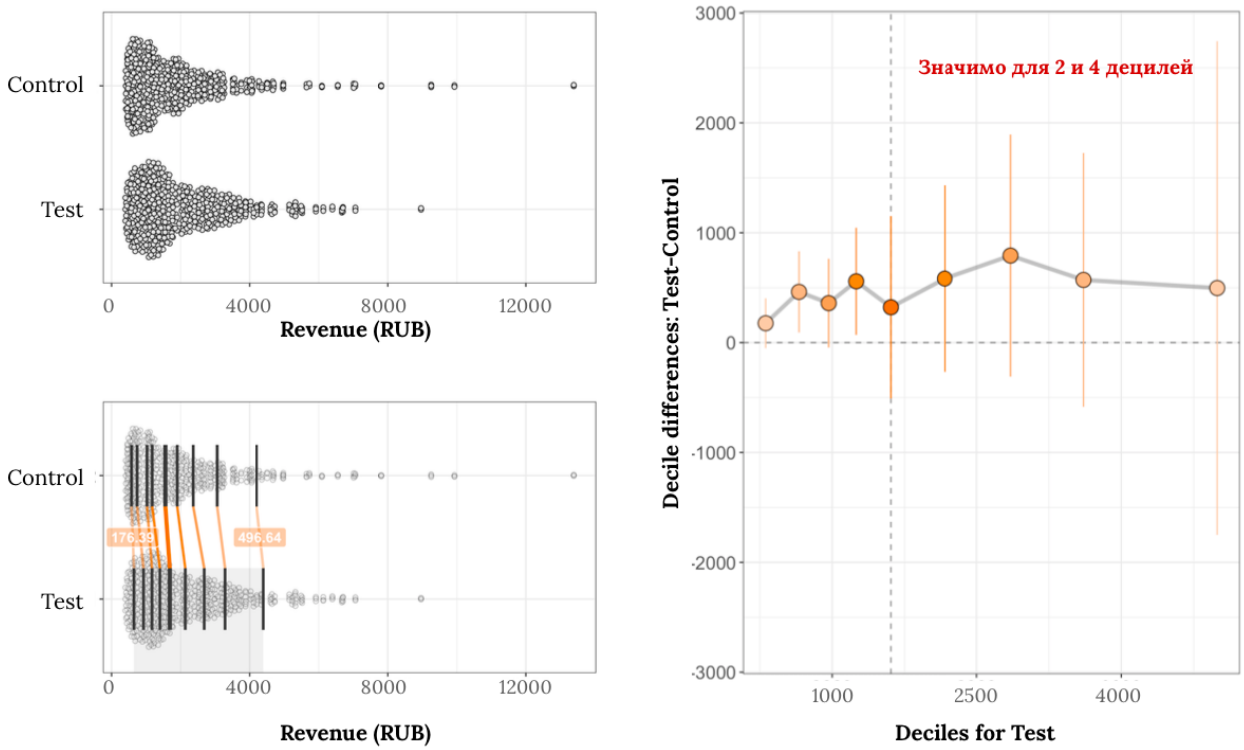
- Бутстрап позволяет строить доверительный интервал для любого параметра распределения, не применяя для этого аналитическую формулу

- 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

> Как проверяются гипотезы с помощью бутстрапа?

1. Строите бутстрап-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



Первый график: распределение сумм заказов.

Второй график (ниже): тот же график, только с отсечками по каждой децили.

Третий график (справа): доверительные интервалы для разницы в конкретной децили. 2 и 4 уходят за ноль, поэтому для них можем отвергнуть нулевую гипотезу о том, что эти децили равны между A и B вариантами.

Итоговая интерпретация: маленькие чеки стали чуть больше, а большие не изменились.

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

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

Механизм бутстрапа так построен, что одно наблюдение может встретиться много-много раз в одном и том же распределении.

> Как считается pvalue в бутстрапе?

В видео мы забыли объяснить метод расчета pvalue, поэтому ниже привели подробное объяснение:

В рассмотренном bootstrap’е считается ресэмплированная разница среднего двух выборок. Чтобы понять, как считать pvalue в bootstrap'е, вспомним как это делается с помощью t-критерия:

- Вспомним нулевую гипотезу t-критерия, она имеет следующий вид:

$$H_0 : \mu_x = \mu_y$$

- Далее, чтобы это проверить, в помощь вступает t-значение:

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

  Оно проверяет отклонение разницы от нуля по t-распределению. Чем больше значение отклонено от нуля (в большую или меньшую сторону), тем меньше вероятность видеть такие результаты в эксперименте, где между средними нет различий.

- На каждой итерации бутстрапа мы считаем разницу, где 0 – это ее мат. ожидание
- Далее нам бы хотелось выяснить, с какой вероятностью мы бы видели такие различия в эксперименте при справедливой нулевой гипотезе. Поэтому мы можем посчитать вероятность такого случая, посчитав сумму случаев с отклонениями от нуля и поделив на количество всех оценок. Это и будет pvalue
- **Первый способ.** Представить в питоне это можно так `pvalue = min( sum(boot <= 0), sum(boot >= 0) ) * 2 / len(boot)` . Выбор минимума из двух оценок в большую и меньшую сторону, а также умножение на 2 тут нужно для двусторонней проверки.

- **Второй способ (используется в коде).** В нашем коде преследуется та же самая суть, но с использованием нормального распределения: у нас есть выборка бутстрапированной разницы, в случае если у нас в первой группе все так же как во второй группе, то эта выборка должна быть нормального распределения по ЦПТ . В `norm.cdf` мы берем среднее `x=0` (опять же из соображений одинаковости средних), а стд. отклонение = 1, потому что применяется правило 3 сигм. Таким образом, мы считаем вероятность отклониться от нуля в центре по нормальному распределению с стд. откл. = 1

Интересно, что и первый и второй способ дают почти одинаковые значения pvalue (при достаточно большом количестве итераций на бутстрапе). Но скорее всего, одинаковых результатов не получится добиться, т.к. есть зависимость от бутстрап итераций (в первом способе это знаменательно). И поэтому не стоит ждать точных расчетов при малом количестве итераций (< 100)

> Дополнительные материалы

- Материал от MIT
- Хотя в этом курсе мы делаем кастомную функцию для бутстрапа, есть и готовые решения. Например, вот это.
- Вдобавок к этому уроку вы можете посмотреть открытый вебинар Анатолия Карпова на схожую тему :)