Skillbox

AB-тестирование: часть 2

Алла Тамбовцева

Skillbox

Урок 1. Формулировка гипотез

Что такое статистическая гипотеза?

Гипотеза — это некоторое утверждение об интересующем нас показателе, которое мы хотим проверить.

Пример: конверсия сайта в новом дизайне и конверсия сайта в старом дизайне не отличаются друг от друга

Типы гипотез

Нулевая гипотеза (H_0) — утверждение о некотором параметре генеральной совокупности или параметрах генеральной совокупности, которое необходимо проверить.

Пример 1

Конверсия сайта в новом дизайне и конверсия сайта в старом дизайне не отличаются друг от друга

Формальная запись

$$H_0 : p_{old} = p_{new}$$

Пример 2

Средний возраст посетителей сайта женского пола не отличается от среднего возраста посетителей сайта мужского пола

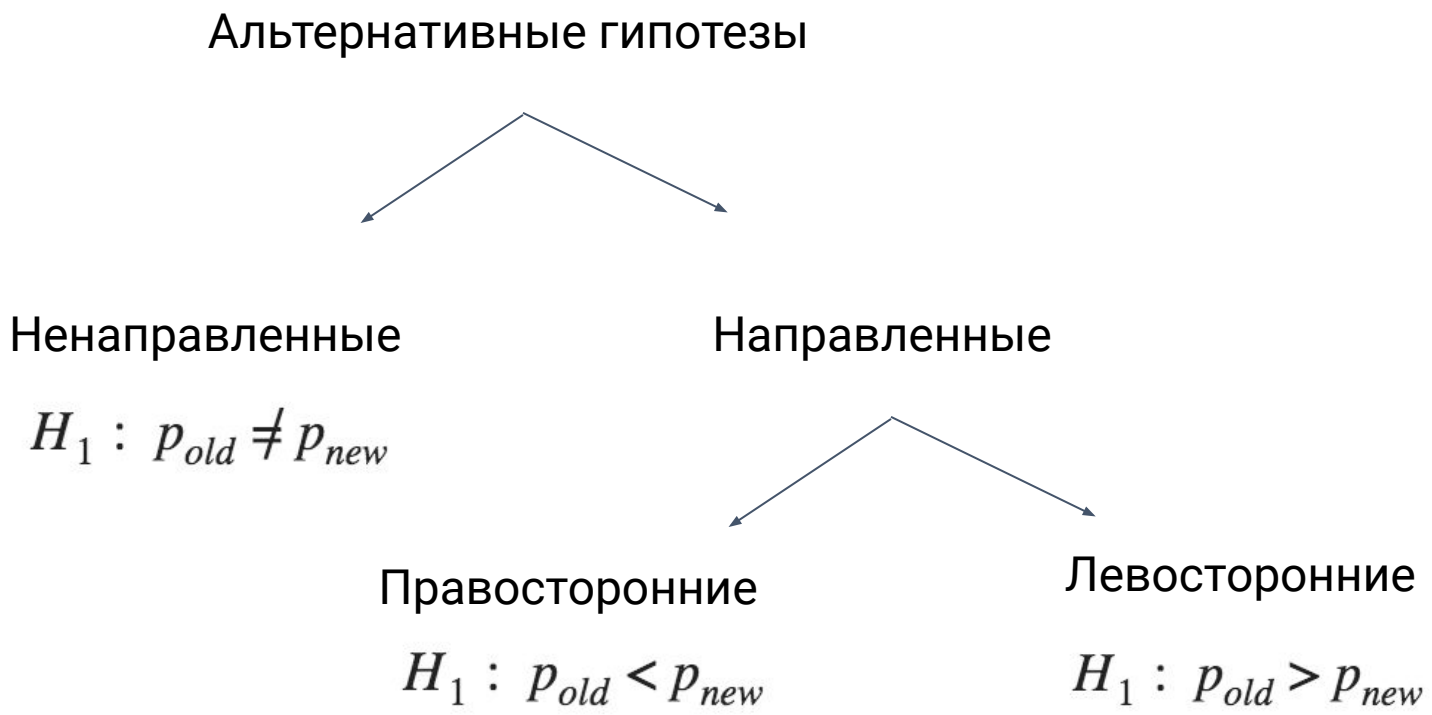
Формальная запись

$$H_{0:} \mu_{female} = \mu_{male}$$

Типы гипотез

Альтернативная гипотеза (***H*₁**) — утверждение, противоположное нулевой гипотезе, которое выдвигается, но не тестируется.

Типы гипотез



Skillbox

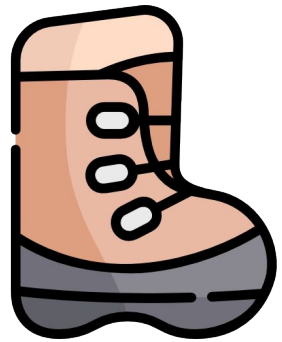
Урок 2. Проверка статистических гипотез

Проверка гипотез

Два подхода к проверке статистических гипотез:

- Использование статистического критерия (теста)
- Бутстрэп (bootstrap)

Бутстрэп



- Название происходит от *“to pull oneself over a fence by one’s bootstraps”* = «перебраться через ограду, потянув за ремешки на ботинках».

- Симуляция многократного повторения эксперимента с целью имитировать генеральную совокупность, которую исследовать непосредственно не получится в силу ограниченности ресурсов.

Бутстрэп

В симуляциях мы создаем много новых выборок путем *случайного выбора с возвращением*.

Пример 1. Есть выборка – значения возраста посетителей сайта:

23, 46, 32, 33, 29

Создадим на ее основе выборку из 10 элементов – случайным образом «надергаем» из нее элементы и запишем их.

46, 23, 32, 23, 46, 29, 33, 33, 23, 32

Выбор с возвращением: элемент, который мы уже отобрали для новой выборки, мы не «выбрасываем», а «возвращаем», чтобы была была возможность выбрать его еще раз.

Бутстрэп

Пример 2. Есть две выборки из 0 и 1, которые содержат индикаторы того, кликнул ли пользователь на кнопку *Оформить заказ* или нет в старом и новом дизайне сайта:

старый: 1, 0, 0, 0, 1, 1, 0, 1

новый: 0, 1, 1, 1, 0, 0, 1, 1

Получим на их основе новые выборки того же размера:

для старого: 1, 0, 0, 1, 0, 0, 0, 1

для нового: 1, 1, 1, 0, 0, 1, 1, 1

Алгоритм проверки гипотезы с помощью бутстрэпа

Итак, у нас сформулированы нулевая и альтернативная гипотезы:

$$H_0 : p_{old} = p_{new}$$

$$H_1 : p_{old} < p_{new}$$

Для проверки нулевой гипотезы мы хотим реализовать бутстрэп.

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг 1

На входе имеем тестовую и контрольные выборки – два набора из 0 и 1, где 1 соответствует пользователю, который кликнул на кнопку *Оформить заказ*, а 0 – тому, кто не кликнул.

Пример на R:

```
old <- c(1, 0, 0, 0, 1, 1, 0, 1)
new <- c(0, 1, 1, 1, 0, 0, 1, 1)
p_old <- sum(old) / length(old) # конверсия 1
p_new <- sum(new) / length(new) # конверсия 2
p_new - p_old  # разница в долях-конверсиях
0.125
```

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг **2**

Фиксируем число новых выборок, которые мы хотим получать на основе старых с помощью случайного выбора с возвращением (число итераций алгоритма). Обычно рекомендуется генерировать не менее 1000 выборок.

Пример на R:

```
N <- 1000
```

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг **3**

Для каждой пары новых выборок считаем разницу в долях единиц – разницу в конверсии нового и старого сайта.

Пример на R:

```
differences <- rep(NA, N) # 1000 пустых элементов
for(i in 1:N){
 s1 <- sample(old,replace=TRUE) # выборка для old с возвращением
 s2 <- sample(new,replace=TRUE) # выборка для new с возвращением
 p1 <- sum(s1)/length(s1) # доля 1 в old
 p2 <- sum(s2)/length(s2) # доля 1 в new
 p_diff <- p2 - p1 # разница в долях
 differences[i] <- p_diff # записываем ее в differences
 }
```

Алгоритм проверки гипотезы с помощью бутстрэпа

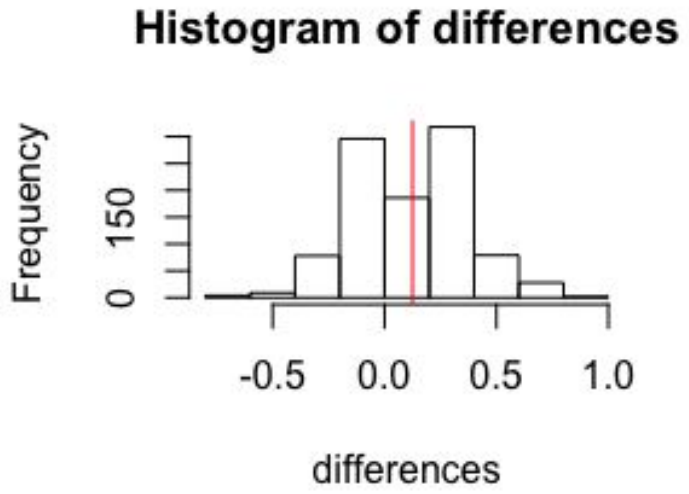
Шаг 4

Считаем долю случаев, когда разность долей оказалась больше 0.125 – значения, полученного на наших настоящих данных в самом начале (больше – так как согласно альтернативной гипотезе ожидаем, что второе значение конверсии больше первого).

Пример на R:

```
hist(differences)
abline(v = 0.125, col = “red”)

sum(differences > 0.125)
0.427
```



Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг **4**

Полученная на этом шаге доля – **p-value**.

Интерпретация p-value=0.427:

Если нулевая гипотеза верна, то вероятность того, что при случайном извлечении контрольной и тестовой выборок по 8 элементов мы получим разницу в значениях конверсии, равную 0.125 или выше, равна 0.427.

P-value – вероятность того, что при многократном повторении эксперимента мы действительно получим тот результат, который уже получили на наших данных или еще более необычный, при условии, что нулевая гипотеза верна.

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг **5**

Определяемся со степенью доверия к нашим данным и делаем вывод.

Фиксируем **уровень значимости** – это вероятность отвергнуть нулевую гипотезу при условии, что она верна, то есть вероятность ошибочно заключить, что различия в конверсии есть, когда их нет. На практике чаще всего выбирают 5% уровень значимости.

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг **5**

Если сильно упрощать, p-value=0.427 – это вероятность того, что различий в конверсии нет. А уровень значимости 0.05 – вероятность того, что мы ошибочно заключим, что различия все-таки есть.

На 5% уровне значимости мы можем заключить, что значение конверсии нового сайта такое же, как и старого.

Алгоритм проверки гипотезы с помощью бутстрэпа

Шаг 5

Общее правило

- p-value < уровень значимости α $\Rightarrow$ H_0 отвергаем на уровне значимости α, на имеющихся данных
- p-value > уровень значимости α $\Rightarrow$ H_0 не отвергаем на уровне значимости α, на имеющихся данных.

Skillbox

Урок 3. Примеры статистических тестов

Статистический тест (критерий)

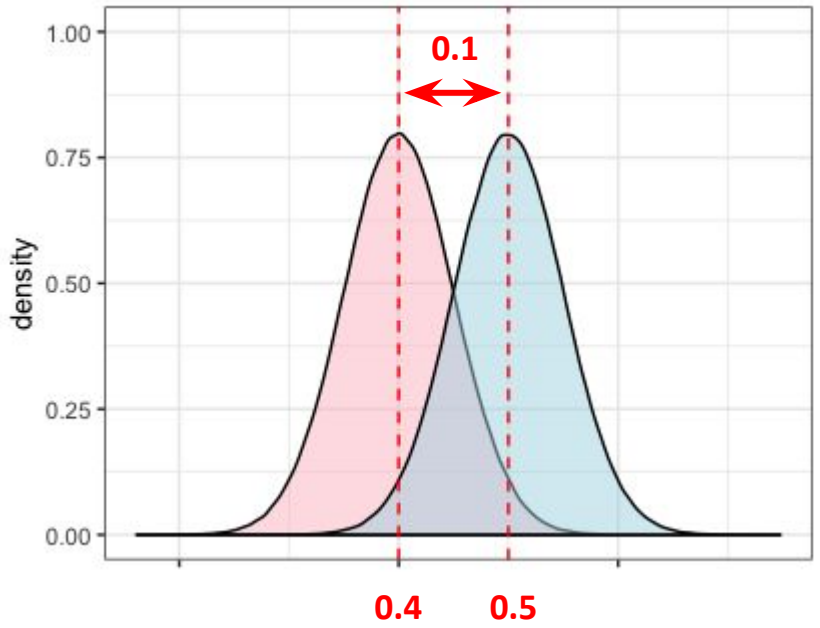
Статистический критерий – правило, которое позволяет делать вывод о том, стоит ли на основе имеющихся данных отвергать нулевую гипотезу или нет, то есть понять, отличается ли конверсия или нет.

Для критерия определяется соответствующая ему **статистика –** мера, которая показывает, насколько велика разница в конверсии с учетом разброса значений в нашей выборке.

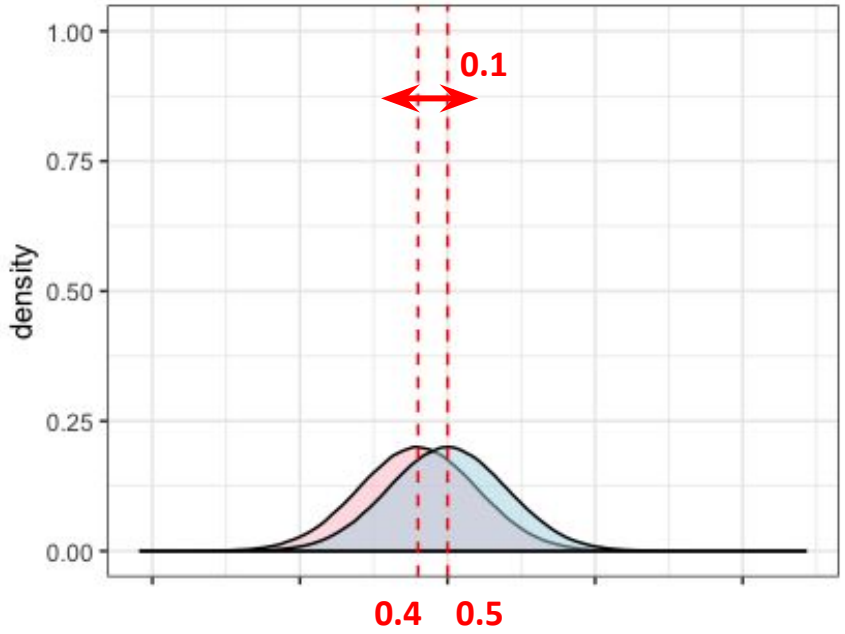
Логика вычисления статистики сводится к одному: оценить разницу в показателях с учетом разброса значений.

Статистический тест (критерий)

Сравним два случая:



Маленький разброс значений → разница в 0.1 значительна



Большой разброс значений → разница в 0.1 незначительна

Статистический тест (критерий)

Чтобы понять, является ли разница между значением параметра в гипотезе и значением оценки, полученной по выборке, в статистике используется уже знакомый нам показатель – **p-value**.

Важно: в выводе относительно отвержения / не-отвержения нулевой гипотезы необходимо указывать уровень значимости, так как от этого зависит результат. Так, например, в случае, если p-value равно 0.02, у нас есть основания отвергнуть нулевую гипотезу на уровне значимости 5% (0.02 < 0.05), и нет оснований отвергнуть ее на уровне значимости 1% (0.02 > 0.01).

Алгоритм проверки гипотез с помощью статистического теста

Шаг 1. Сформулировать нулевую гипотезу.

$$H_0 : p_{old} = p_{new}$$

Шаг 2. Сформулировать альтернативную гипотезу.

$$H_1 : p_{old} < p_{new}$$

Важно понимать, какого типа гипотеза (направленная или ненаправленная), так как от типа альтернативной гипотезы зависит расчет значения p-value.

Алгоритм проверки гипотез с помощью статистического теста

Шаг 3. Запустить критерий, необходимый для проверки нулевой гипотезы.

Если речь идет о сравнении двух долей, обычно используют z-test.

Шаг 4. Посчитать p-value. Сравнить p-value с фиксированным уровнем значимости.

Запускаем в R или Python тест на наших данных, в выдаче результатов теста видим, что p-value = 0.034. Значение 0.034 меньше уровня значимости 0.05, поэтому делаем вывод о том, что нулевую гипотезу об отсутствии различий, необходимо отвергнуть. Различия есть!

Алгоритм проверки гипотез с помощью статистического теста

Шаг 5. Сделать статистический и содержательный вывод.

Пример статистического вывода: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) есть основания отвергнуть нулевую гипотезу в пользу альтернативы.

Пример содержательного вывода: уровень конверсии сайта в новом дизайне выше, чем уровень конверсии в старом дизайне сайта.

Перейдем в R и Python и реализуем z-test!