Анализ результатов АБ-теста

7 вопросов

1 point

1.

В данном задании вам нужно будет

- проанализировать АБ тест, проведенный на реальных пользователях Яндекса
- подтвердить или опровергнуть наличие изменений в пользовательском поведении между контрольной (control) и тестовой (exp) группами
- определить характер этих изменений и практическую значимость вводимого изменения
- понять, какая из пользовательских групп более всего проигрывает / выигрывает от тестируемого изменения (локализовать изменение)

Описание данных:

- userID: уникальный идентификатор пользователя
- browser: браузер, который использовал userID
- slot: в каком статусе пользователь участвовал в исследовании (exp = видел измененную страницу, control = видел неизменную страницу)
- n_clicks: количество кликов, которые пользоваль совершил за n_queries
- n_queries: количество запросов, который совершил userID, пользуясь браузером browser
- n_nonclk_queries: количество запросов пользователя, в которых им не было совершено ни одного клика

Обращаем ваше внимание, что не все люди используют только один браузер, поэтому в столбце userID есть повторяющиеся идентификаторы. В предлагаемых данных уникальным является сочетание userID и browser.

ab_browser_test.csv

Основная метрика, на которой мы сосредоточимся в этой работе, — это количество пользовательских кликов на web-странице в зависимости от тестируемого изменения этой страницы.

Посчитайте, насколько в группе exp больше пользовательских кликов по сравнению с группой control в **процентах** от числа кликов в контрольной группе.

Полученный процент округлите до третьего знака после точки.

Введите ответ здесь

1 point

2.
Давайте попробуем посмотреть более внимательно на разницу между двумя группами (control и exp) относительно количества пользовательских кликов.

Для этого постройте с помощью бутстрепа 95% доверительный интервал для средних значений и медиан количества кликов в каждой из двух групп. Отметьте все верные утверждения.

- [ ] 95% доверительный интервал для разности медиан не содержит ноль, похоже, медианы отличаются статистически значимо
- [ ] 95% доверительный интервал для разности средних не содержит ноль, похоже, средние отличаются статистически значимо
- [ ] 95% доверительный интервал для разности медиан содержит ноль, похоже, медианы существенно не отличаются
- [ ] Применение bootstrap на выборках такого большого размера неправомерно, потому что bootstrap делает псведовыборки с возвращениями, а с ростом объема исходной выборки псведовыборки с возвращениями становятся более похожими на псведовыборки без возвращения.
- [ ] 95% доверительный интервал для разности средних содержит ноль, похоже, средние существенно не отличаются

1 point

3.
Поскольку данных достаточно много (порядка полумиллиона уникальных пользователей), отличие в несколько процентов может быть не только практически значимым, но и значимым статистически. Последнее утверждение нуждается в дополнительной проверке.

Посмотрите на выданные вам данные и выберите все верные варианты ответа относительно проверки гипотезы о равенстве среднего количества кликов в группах.

- [ ] Для проверки гипотезы о равенстве средних в данной задаче можно использовать только параметрические критерии, потому что непараметрические, как известно, с увеличением размера выборки могут давать непредсказуемые результаты в силу случайности в определении их нулевого распределения.
- [ ] Все ответы неверны
- [ ] Гипотезу о равенстве средних между двумя выборками можно проверить с помощью построения доверительного интервала для среднего объединенной выборки, потому что в силу большого объема выборки этот интервал будет очень точным, и мы сможем надежно оценить необходимый доверительный интервал

- [ ] Используя центральную предельную теорему, мы можем заключить, что с ростом объема выборки любое исследуемое распределение, становится похожим на нормальное, а значит, учитывая большой объем нашей выборки, оптимальным критерием в нашей задаче будет z-критерий.

- [ ] Для проверки гипотезы о равенстве средних категорически нельзя использовать t-критерий Стьюдента ни при каком размере выборки, потому что количество кликов, которые мы исследуем, больше походят на распределение Пуассона, которое *сильно* отличается от нормального.

---

1 point

4.

t-критерий Стьюдента имеет множество достоинств, и потому его достаточно часто применяют в AB экспериментах. Иногда его применение может быть необоснованно из-за сильной скошенности распределения данных.

Давайте постараемся понять, когда t-критерий можно применять и как это проверить на реальных данных.

Для простоты рассмотрим одновыборочный t-критерий. Его статистика имеет вид $\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$, то есть чтобы действительно предположения t-критерия выполнялись необходимо, чтобы:

- $\bar{X}$ — среднее значение в выборке — было распределено нормально $\mathcal{N}(\mu, \frac{\sigma^2}{n})$
- $\frac{n}{\sigma^2} S^2$ — несмещенная оценка дисперсии с масштабирующим коэффициентом — была распределена по хи-квадрат с $n-1$ степенями свободы $\chi^2(n-1)$

Простое доказательство необходимости и достаточности этого требования можно посмотреть в самом последнем абзаце этого вопроса. Усвоение этого доказательства не обязательно для выполнения задания.

Оба этих предположения можно проверить с помощью бутстрепа. Ограничимся сейчас только контрольной группой, в которой распределение кликов будем называть *данными* в рамках данного вопроса.

Поскольку мы не знаем истинного распределения генеральной совокупности, мы можем применить бутстреп, чтобы понять, как распределены среднее значение и выборочная дисперсия. Для этого

1. Получите из данных n_boot_samples псевдовыборок.
2. По каждой из этих выборок посчитайте среднее и сумму квадратов отклонения от выборочного среднего (control_boot_chi_squared)
3. Для получившегося вектора средних значений из n_boot_samples постройте q-q plot с помощью scipy.stats.probplot для нормального распределения
4. Для получившегося вектора сумм квадратов отклонения от выборочного среднего постройте qq-plot с помощью scipy.stats.probplot для хи-квадрат распределения с помощью команды

```
scipy.stats.probplot(control_boot_chi_squared, dist="chi2",
                     sparams=(n-1), plot=plt)
```

Где sparams=(n-1) означают число степеней свободы = длине выборки - 1.

Чтобы получить такой же ответ, как у нас, зафиксируйте seed и количество псевдовыборок:

```
1  np.random.seed(0)
2  n_boot_samples = 500
```

В качестве ответа отметьте верные утвердения о значениях R^2, которые генерирует scipy.stats.probplot при отображении qq-графиков: одно с графика для среднего и одно с графика для выборочной суммы квадратов отклонения от выборочной суммы.

Почему мы проверяем именно такие условия?

В исходной постановке t-критерий требует нормальности распределения X_i. Именно из-за этого предположения мы имеем, что $\sum_i X_i \sim \mathcal{N}(n\mu, n\sigma^2)$ в силу линейности матожидания, независимости всех X_i между собой и того факта, что сумма нескольких нормальных случайных величин также нормальна.

Поэтому, пользуясь опять формальными свойствами матожидания и дисперсии можем записать, что $\frac{1}{n}\sum_i X_i - \mu \ \sim \ \mathcal{N}(0, \frac{\sigma^2}{n})$

Заметим теперь, что следующие распределения эквивалентны $\mathcal{N}(0, \frac{\sigma^2}{n}) \sim \sqrt{\frac{\sigma^2}{n}}\mathcal{N}(0,1)$

То есть другими словами мы получили, что исходная статистика $\frac{\bar{X}-\mu}{\sqrt{S^2/n}}$ распределена как $\frac{\mathcal{N}(0,1)}{\sqrt{\frac{S^2 n}{n\sigma^2}}} \ \sim \ \frac{\mathcal{N}(0,1)}{\sqrt{\frac{S^2}{\sigma^2}}}$

Вспомним, что распределение Стьюдента с $n-1$ степенями свободы определяется как $\frac{\mathcal{N}(0,1)}{\sqrt{\chi^2/(n-1)}}$. Поэтому и требования, которые накладываются, имеют вид, указанный в начале этого вопроса.

Полезно, однако, заметить, что можно подставить в числитель и знаменатель любые случайные величины, распределенные по нормальному закону и по Хи-квадрат соответственно, для этого необязательно, чтобы именно X_i были распределены нормально. Поэтому, если мы убедимся в том, что, действительно, числитель и знаменатель распределены образом, указанным выше, то можно смело использовать t-критерий Стьюдента.

- ☐ R^2 для выборочной суммы квадратов отклонения от выборочной суммы получился больше, чем 0.99
- ☐ R^2 для выборочного среднего получился меньше, чем 0.99
- ☐ R^2 для выборочной суммы квадратов отклонения от выборочной суммы получился меньше, чем 0.99
- ☐ R^2 для выборочного среднего получился больше, чем 0.99

1 point

5.

Одним из возможных аналогов t-критерия, которым можно воспрользоваться, является тест Манна-Уитни. На достаточно обширном классе распределений он является асимптотически более эффективным, чем t-критерий, и при этом не требует параметрических предположений о характере распределения.

Разделите выборку на две части, соответствующие control и exp группам. Преобразуйте данные к виду, чтобы каждому пользователю соответствовало суммарное значение его кликов. С помощью критерия Манна-Уитни проверьте гипотезу о равенстве средних. Что можно сказать о получившемся значении достигаемого уровня значимости ? Выберите все правильные ответы

- ☐ Критерий Манна-Уитни в данной задаче применять нельзя, поэтому вопрос о достигаемом уровне значимости некорректен
- ☐ Получившееся значение достигаемого уровня значимости свидетельствует о статистической значимости отличий между двумя выборками
- ☐ $p < 0.01$, поэтому можно сказать, что отличия незначительны на уровне доверия 0.05
- ☐ Согласно полученному значению p-value, мы вынуждены принять нулевую гипотезу

1 point

6.

Проверьте, для какого из браузеров наиболее сильно выражено отличие между количеством кликов в контрольной и экспериментальной группах.

Для этого примените для каждого из срезов (по каждому из уникальных значений столбца browser) критерий Манна-Уитни между control и exp группами и сделайте поправку Холма-Бонферрони на множественную проверку с $\alpha = 0.05$.

Какое заключение можно сделать исходя из полученных результатов ?

В качестве ответа введите количество незначимых изменений с точки зрения результатов, полученных после введения коррекции.

Введите ответ здесь

1 point

7.

Для каждого браузера в каждой из двух групп (control и exp) посчитайте долю запросов, в которых пользователь не кликнул ни разу. Это можно сделать, поделив сумму значений n_nonclk_queries на сумму значений n_queries. Умножив это значение на 100, получим процент некликнутых запросов, который можно легче проинтерпретировать.

Сходятся ли результаты проведенного Вами анализа с показателем процента некликнутых запросов ? Отметьте все верные утверждения.

- [ ] По всем браузерам мы видим незначительное уменьшение доли некликнутых запросов, поэтому делаем вывод о том, что тестируемое изменение приносит больше вреда, чем пользы.

- [ ] По одному из браузеров мы видим значительное уменьшение доли некликнутых запросов, поэтому уже только на этом основании тестируемое изменение можно рекомендовать к применению для всех пользователей.

- [ ] С помощью анализа, проведенного в предыдущем вопросе, мы показали, что тестируемое изменение приводит к статистически значимому отличию только для одного браузера. Для этого браузера на основе данных о доли некликнутых запросов, заключаем, что тестируемое изменение влияет на пользователей позитивно.

- [ ] Тестируемое изменение можно предложить к внедрению только на тот сегмент пользователей, где локализуется изменение, то есть для того браузера, для которого доля некликнутых запросов уменьшилась больше всего. Для прочих браузеров мы не обладаем никакой информацией относительно влияния тестируемого изменения на поведение пользователей.

- [ ] I, **Владислав Лисеев**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.
Узнайте больше о Кодексе чести Coursera

2 вопросов без ответа

Сдать тест