

А/Б Тестирование

Статистические методы

Contents

1 Т-тест	2
2 Бутстрэп	2
2.1 Перцентильный Д.И.	3
2.2 Обратный перцентильный Д.И.	3
2.3 Бутстрэп t-статистики	4
2.4 Пуассоновский бутстрэп	5
3 Тест Манна-Уитни	6
4 Difference in Difference (DnD)	9
5 Мэтчинг	9
6 Ratio-метрики	11
7 Дельта-метод	12
8 Линеаризация	14
9 Перевзвешивание	16
10 Бакетизация	16
11 Стратификация	17
11.1 Стратификация	19
11.2 Постстратификация	19
12 CUPED	19
13 Множественное тестирование	22
13.1 Поправка Бонферонни	23
13.2 Метод Холма	23
13.3 FDR	23
14 Последовательное тестирование	24

1 Т-тест

Для проверки гипотез о равенстве средних в двух выборках используются статистические критерии. Наиболее популярным является t-критерий Стьюдента, который проверяет гипотезу:

$$\begin{cases} H_0 : \bar{X} = \bar{Y} \\ H_A : \bar{X} \neq \bar{Y} \end{cases}$$

С помощью критерия:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}} \sim t_{n_X+n_Y-2}$$

Однако есть несколько предпосылок для t-критерия Стьюдента:

1. Выборки должны быть независимы между собой
2. У обеих выборок одинаковая дисперсия
3. Отсутствуют выбросы в данных (это означает, что дисперсия конечна)

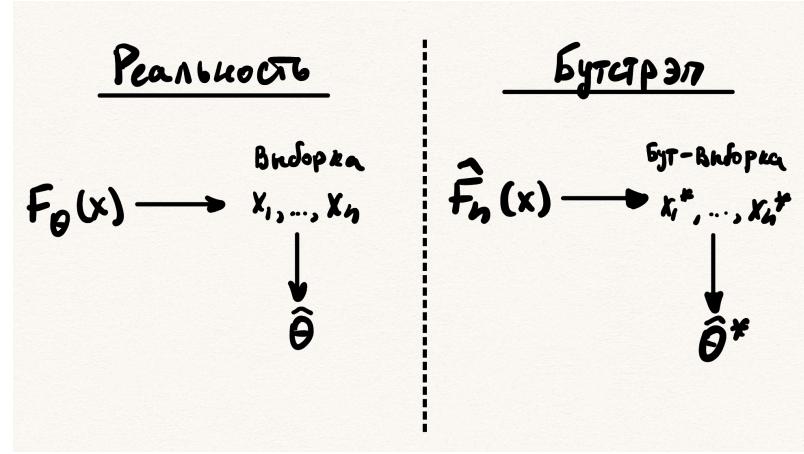
Но случай с одинаковой дисперсией неправдоподобен, поэтому в таком случае лучше использовать тест Уэлча, который учитывает, что дисперсия в обеих выборках может быть разной:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim t(df), \quad df = \frac{\left(\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)}{\frac{1}{n_X-1}\left(\frac{\sigma_X^2}{n_X}\right)^2 + \frac{1}{n_Y-1}\left(\frac{\sigma_Y^2}{n_Y}\right)^2}$$

2 Бутстрэп

Бутстрэп - асимптотический метод, который может быть использован для проверки гипотез, оценки сложных статистик, их доверительных интервалов и т.д. Единственным условием является наличие большой выборки.

Есть некий истинный закон распределения с неизвестными нам параметрами, из которого генерируются данные. Из этого распределения как раз и была получена выборка, по которой мы считаем оценки статистик. Поэтому с помощью бутстрэпа мы стараемся сэмплировать истинный процесс генерации данных с помощью представления выборки в виде генеральной совокупности и получить бут-выборки, по которым будем получать бут-оценки.

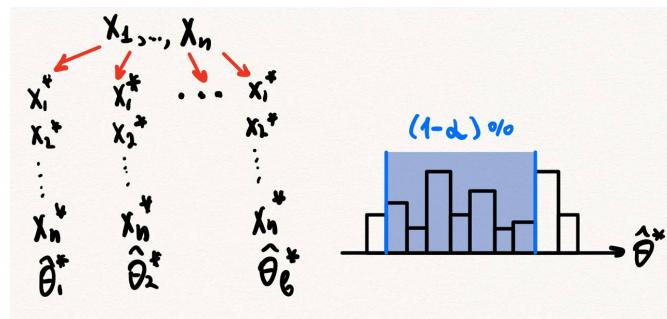


Важно: бут-выборки обязательно должны быть такого же размера, что и исходная выборка. Иначе все оценки, построенные по бут-выборкам, будут смещеными.

Есть несколько способов построить доверительный интервал / проверить гипотезу с помощью бутстрэпа:

2.1 Перцентильный Д.И.

Есть выборка x_1, \dots, x_n , генерируем из нее b бут-выборок и по каждой считаем интересующую нас статистику. После того, как получили вектор из оценок, полученных из бут-выборок, строим гистограмму и выбираем ту область, которая покрывает интересующий нас уровень значимости ($(1 - \alpha\%)$ наблюдений).



2.2 Обратный перцентильный Д.И.

У нас уже есть оценка параметра, полученная по исходной выборке $\hat{\theta}$. Теперь будем генерировать b бут-выборок, по ним считать $\hat{\theta}^*$ и параллельно считать статистику $\hat{q}^* = \hat{\theta}^* - \hat{\theta}$. После того, как мы получили вектор $\hat{q}_1^*, \dots, \hat{q}_b^*$ получаем интервал:

$$CI = [\hat{\theta} - \hat{q}_{1-\frac{\alpha}{2}}^*; \hat{\theta} - \hat{q}_{\frac{\alpha}{2}}^*]$$

2.3 Бутстрэп t-статистики

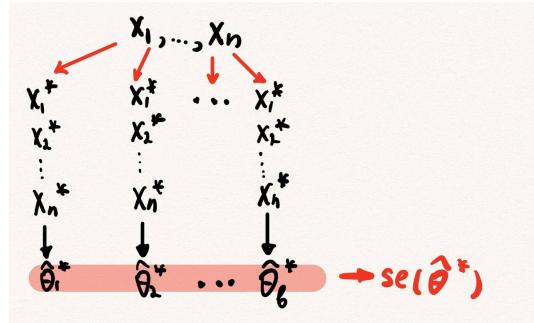
Считаем оценку параметра $\hat{\theta}$ по исходной выборке, затем генерируем b бут-выборок, получаем оценку параметра по бут-выборке $\hat{\theta}^*$ и считаем что-то похожее на t-статистику:

$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{se(\hat{\theta}^*)}$$

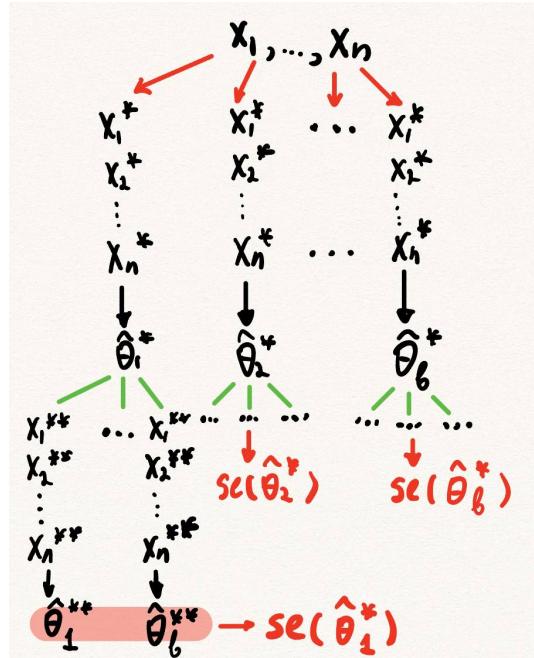
После этого получаем распределение t-статистик t_1^*, \dots, t_b^* и строим доверительный интервал:

$$CI = [\hat{\theta} - se(\hat{\theta}) \times t_{1-\frac{\alpha}{2}}^*, \hat{\theta} - se(\hat{\theta}) \times t_{\frac{\alpha}{2}}^*]$$

Вопрос: как считать $se(\hat{\theta})$? 1-й способ, поскольку мы генерировали b бут-выборок, то и получили вектор из b оценок $\hat{\theta}^*$, по нему и можем оценить $se(\hat{\theta}^*)$.



2-й способ: для каждого $\hat{\theta}_i^*$ генерировать еще несколько бут-выборок, по которым будем считать $\hat{\theta}^{**}$ и по вектору $\hat{\theta}_1^{**}, \dots, \hat{\theta}_b^{**}$ оценивать $se(\hat{\theta}^*)$.



Только теперь когда генерируем выборки $x_1^{**}, \dots, x_n^{**}$, то мы семплируем из выборки x_1^*, \dots, x_n^* , а не из изначальной. Далее считаем статистики $t = \frac{\hat{\theta} - \hat{\theta}_i^*}{se(\hat{\theta}_i^*)}$. Однако в доверительном

интервале мы также считаем $se(\hat{\theta})$ по бут-оценкам $\hat{\theta}_1^*, \dots, \hat{\theta}_b^*$.

2.4 Пуассоновский бутстрэп

Метод, придуманный для ускорения и распараллеливания бутстрэпа. Идея заключается в следующем: мы будем генерировать какое число раз попался i -й элемент из выборки из биномиального распределения. Также у нас есть условие, что сумма частот каждого элемента должна равняться исходному размеру выборки. Итого постановка следующая:

$$\begin{cases} Y_i \sim Bin(n, \frac{1}{n}) \\ Y_1 + \dots + Y_n = n \end{cases}$$

Тогда можно сказать, что распределение частот элементов имеет мультиномиальное распределение и для каждой бут-выборки генерируется следующий случайный вектор:

$$Y_1, \dots, Y_n \sim MultiNom(n, \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$$

Но минус в том, что необходимо следить сколько наблюдений осталось набрать, чтобы в сумме получилось n . Но можно воспользоваться фактом, что $plim_{n \rightarrow \infty} Bin(n, \frac{1}{n}) = Poiss(1)$. Поэтому частоту каждого элемента можно генерировать из распределения Пуассона $Y_i \sim Poiss(1)$ и можно не следить за тем, чтобы сумма частот равнялась n . Понятно, что выборки скорее всего не будут того же размера, что и n , но доказано, что от этого бутстрэп не ломается.

Ускорение заключается в том, что мы заранее генерируем сколько раз мы берем тот или иной элемент для каждой выборки и создаем бут-выборку на основе получившихся частот.

Также с помощью бутстрэпа можно проверять гипотезы. Пусть у нас есть гипотеза:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_A : \theta > \theta_0 \end{cases}$$

Мы по изначальной выборке считаем t-статистику $t = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$ и получаем t_{Obs} . Далее с помощью бутстрэпа мы бутстрапируем t-статистику $t^* = \frac{\hat{\theta}^* - \theta}{se(\hat{\theta}^*)}$ и по получившемуся вектору t_1^*, \dots, t_b^* считаем квантиль $t_{1-\alpha}^*$ - что и будет являться нашим t_{Crit} . Далее сравниваем t_{Obs} и t_{Crit} и делаем вывод.

В рамках А/Б теста мы хотим проверять гипотезу для двух выборок. И это можно сделать с помощью бутстрэпа. Пусть есть выборка x_1, \dots, x_n и посчитанные по ней статистики

\bar{X} и $\hat{\sigma}_X^2$. То же самое и для другой выборки y_1, \dots, y_m - \bar{Y} и $\hat{\sigma}_Y^2$. Тестируется следующая гипотеза:

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_A : \mu_X \neq \mu_Y \end{cases}$$

С помощью t-статистики $t_{Obs} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{n} + \frac{\hat{\sigma}_Y^2}{m}}}$. Чтобы внедрить сюда бутстрэп используется следующий алгоритм: 1. Для каждого наблюдения из обеих выборок делаем преобразование:

$$x'_i = x_i - \bar{X} + \bar{Z}$$

$$y'_i = y_i - \bar{Y} + \bar{Z}$$

Где \bar{Z} - среднее, посчитанное по объединенной выборки x_1, \dots, x_n и y_1, \dots, y_m .

2. Генерируем из x'_1, \dots, x'_n бут-выборку $(x')_1^*, \dots, (x')_n^*$ и из y'_1, \dots, y'_m - $(y')_1^*, \dots, (y')_m^*$.
3. Считаем $t^* = \frac{\bar{(X')}^* - \bar{(Y')}^*}{\dots}$, где дисперсию можно считать одним из предыдущих способов, описанных ранее.
4. Сравниваем t_{Obs} и $t^*_{1-\frac{\alpha}{2}}$ и делаем вывод.

3 Тест Манна-Уитни

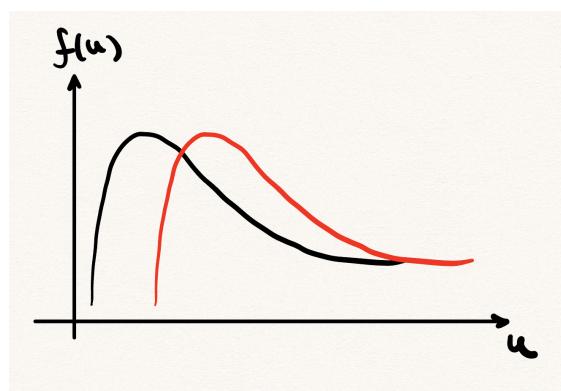
Работает при небольшом количестве наблюдений и не распределенных нормально. Если были бы распределены нормально, то тогда бы можно было использовать тест Уэлча.

Пусть есть две выборки Y_1, \dots, Y_{n_y} и X_1, \dots, X_{n_x} , у каждой выборки есть свое распределение.

Тогда формуируем нулевую гипотезу, что эти распределения совпадают (в терминах плотностей):

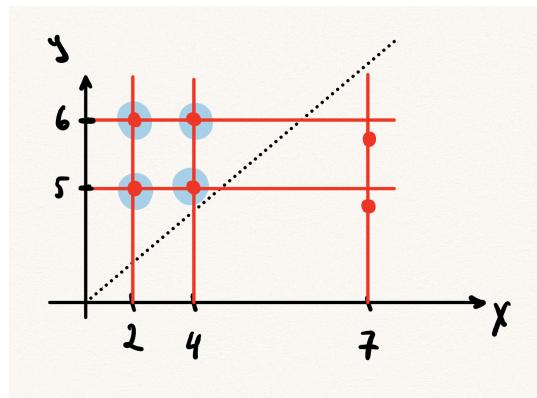
$$H_0 : f_Y(t) = f_X(t)$$

$$H_A : f_Y(t) \neq f_X(t + \Delta)$$



Здесь если бы мы хотели проверить гипотезу с помощью проверки равенства средних, то ДИ был бы неправильный из-за того, что распределение статистики не имеет нормальное распределение и значение 1.96 в $[\bar{Y} - \bar{X} - 1.96 \times se(\bar{Y} - \bar{X}); ...]$ не будет корректным.

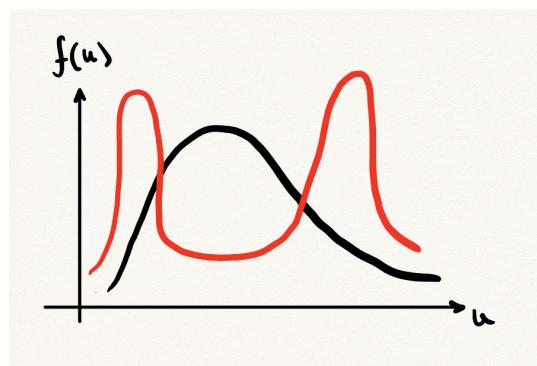
Рассмотрим следующую ситуацию. Есть две выборки: $X = (2, 4, 7)$ и $Y = (5, 6)$. Нарисуем их на графике в виде сетки:



Статистика Манна-Уитни: $U = N(Y_j > X_i)$ - сколько раз Y оказались больше X во всех возможных парах. Далее отмечаем те точки, где Y больше X (точки, которые оказались выше биссектрисы, отмечены синим). В данном примере количество всевозможных пар равно 6, статистика U равна 4.

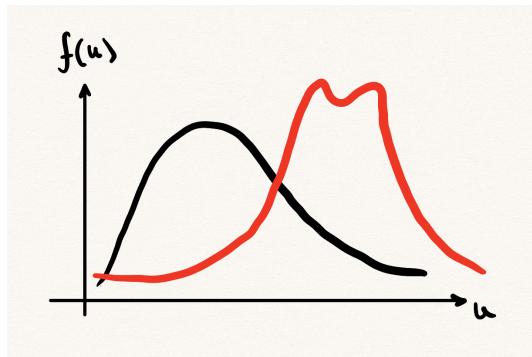
Процедура принятия решений стандартная: если $U > U_{crit}$, то отвергаем H_0 . Для маленьких значений выборок есть таблица с соответствием размеров выборок с U_{crit} .

Статистика Манна-Уитни никак не учитывает законы распределений выборок, только берет во внимание их взаимное расположение. Может быть случай, что выборки X и Y будут иметь разные распределения. Тогда есть 2 варианта работы теста. Первый случай, если плотность распределения после эффекта (то есть второй выборки) осталась примерно на той же позиции, но поменяла вид, тогда в этом случае тест Манна-Уитни будет плохо работать.



Тут красная линия - плотность выборки пользователей после эффекта воздействия. На примере видно, что нельзя однозначно определить - есть ли эффект. На концах распределений заметно, что эффект оказал влияние в лучшую сторону, но в центре - нет. Поэтому для данного примера тест Манна-Уитни не поймает эффект.

Если плотность распределения выборки после эффекта и поменяло вид, но заметно, что оно сдвинулось, то этот эффект тест Манна-Уитни поймает.



Если в данном случае попарно сравнивать выборки, то видно, что точки из красного распределения будут больше превосходить точки из чёрного распределения.

Одной из предпосылок теста Манна-Уитни является отсутствие большого количества повторяющихся значений.

Одно из преимуществ статистики Манна-Уитни - быстрая сходимость к нормальному. Значения $n_X = 7$ и $n_Y = 7$ уже достаточно для того, чтобы можно было пользоваться нормальным распределением:

$$\frac{U - \mathbb{E}(U)}{\sqrt{\text{Var}(U)}} \xrightarrow{d} \mathbb{N}(0, 1)$$

Есть быстрый способ подсчета статистики U вместо того, чтобы перебирать все возможные комбинации пар. Если U - число преимуществ Y над X , то также можно посчитать $U_{Y>X} = R_Y - \frac{n_Y(n_Y+1)}{2}$, где R_Y - ранги Y .

Как считать ранги? Объединяем две выборки X и Y в одну и сортируем в порядке возрастания (сверху таблицы минимальный элемент), указывая какой элемент пришел из выборки X , а какой из Y . Нумеруем ранги, т.е. на какой позиции находится тот или иной элемент. Теперь R_X - сумма рангов X , R_Y - сумма рангов Y . Расчет для выборок из примера выше:

X: 2, 4, 7		Y: 5, 6
X\Y		Ранг
2	X	1
4	X	2
5	Y	3
6	Y	4
7	X	5

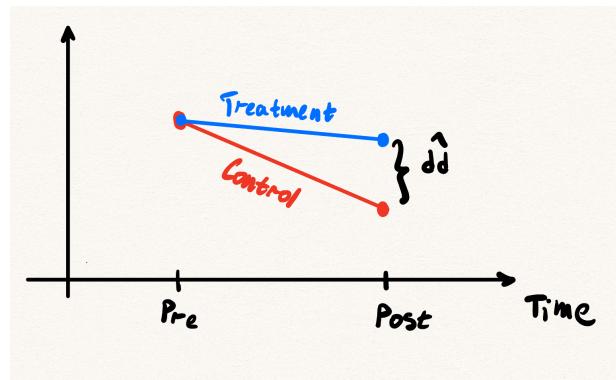
4 Difference in Difference (DnD)

Есть две группы наблюдений: группа контроля и группа воздействия; и 2 периода: до проведения эксперимента и после проведения эксперимента. Для каждой комбинации группы и периода считается значение метрики:

	Pre	Post
Control	$\bar{Y}_{Pre}^{(C)}$	$\bar{Y}_{Post}^{(C)}$
Treatment	$\bar{Y}_{Pre}^{(T)}$	$\bar{Y}_{Post}^{(T)}$

Тогда оценкой эффекта будет являться $\hat{d} = (\bar{Y}_{Post}^{(T)} - \bar{Y}_{Pre}^{(T)}) - (\bar{Y}_{Post}^{(C)} - \bar{Y}_{Pre}^{(C)})$

Данный метод работает, поскольку позволяет учесть сторонние эффекты, не связанные с экспериментом, например, экономические шоки и тд. Тогда они повлияют сразу на обе группы и мы все равно сможем оценить эффект от нашего воздействия на пользователей.



5 Мэтчинг

До проведения эксперимента для каждого индивида существуют 2 потенциальных исхода: $Y_i(0)$ - значение Y_i , если бы i -й индивид не получил воздействие, и $Y_i(1)$ - значение Y_i , если

бы i -й индивид получил воздействие. Проблема - для каждого индивида мы наблюдаем только одно из состояний в зависимости от воздействовали ли мы на него или нет.

В идеале мы бы хотели узнать $ATE = \mathbb{E}[Y_i(1) - Y_i(0)]$. Мы бы могли оценить ATE с помощью $\widehat{ATE} = \frac{\sum Y_i(1) - Y_i(0)}{n}$, но не можем получить эту из-за того, что не наблюдаем оба потенциальных исхода для одного индивида.

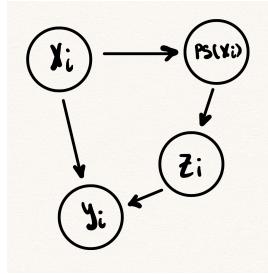
Также нас может интересовать $CATE = \mathbb{E}[Y_i(1) - Y_i(0)|X]$, где X - набор параметров. То есть хотим оценить зависимость эффекта воздействия от каких-то параметров. Для оценки $CATE$ можно использовать обычную линейную регрессию: $\widehat{Y_i(1) - Y_i(0)} = \hat{\beta}_1 + \hat{\beta}_2 x_i + \dots$

В рандомизированных испытаниях мы предполагаем, что переменная воздействия z_i не зависит от x_i , $Y_i(1)$ и $Y_i(0)$. При этих предпосылках для оценки ATE можно пользоваться обычной регрессией на переменную воздействия, регрессией на переменную воздействия + контрольные переменные, CUPED.

Идея мэтчинга: считаем разницу $Y_i(1) - Y_i(0)$ между близкими по значению контрольных переменных наблюдениями. Важно, чтобы у двух выбранных наблюдений отличалась переменная воздействия. Чаще всего в мэтчинге используется KNN, но есть теорема, что чем больше количество предикторов в KNN, тем медленнее сходимость \widehat{ATE} к ATE .

Y_i	z_i	X_i	$\widehat{ATE} = \frac{1}{n} ((5-1) + (2-6) + (3-7))$
5	1	21	
6	0	70	
3	1	40	
2	1	45	
1	0	21	
7	0	73	

Также можно учитывать вероятность, что индивид при определенном наборе характеристик сам совершил целевое действие (т.е. сам решил попасть в группу воздействия). Propensity score (вероятность совершить целевое действие) считается как $ps(x_i) = \mathbb{P}(z_i = 1|x_i)$. В рандомизированном эксперименте $ps(x_i) = \frac{1}{2}$. В реальности нам приходится оценивать ps с помощью логистической регрессии, ML моделей и тд.



В данной схеме видно, что x_i влияет на переменную воздействия z_i только через $ps(x_i)$. Поэтому когда фиксируем $ps(x_i)$, то мы избавляемся от влияния x_i на z_i . Отсюда и рождается идея **Propensity score + Matching**. Оцениваем вероятности $ps(x_i)$ и используем их в качестве предикторов в KNN для мэтчинга.

6 Ratio-метрики

Может произойти такой случай при тестировании гипотезы, что наблюдения зависимы между собой и тогда предпосылки t-теста о независимости ломаются и проверять гипотезу с помощью него будет иррационально.

Например, средний чек: при расчете среднего чека в данных могут наблюдаться покупки одного и того же пользователя \Rightarrow такие наблюдения будут зависимы. В таком случае сломается оценка дисперсии, она станет смещенной из-за того, что выскочат ковариации между наблюдениями и увеличится ошибка I-го рода. Также кол-во заказов коррелирует с суммой заказов: чем больше чеков, тем больше и сумма.

В таких случаях вводят ratio метрики:

$$R_C = \frac{X_1^C + \dots + X_N^C}{Y_1^C + \dots + Y_N^C} = \frac{\sum_{u \in C} X(u)}{\sum_{u \in C} Y(u)}$$

$$R_T = \frac{X_1^T + \dots + X_N^T}{Y_1^T + \dots + Y_N^T} = \frac{\sum_{u \in T} X(u)}{\sum_{u \in T} Y(u)}$$

Где X_i - сумма покупок i -го пользователя, Y_i - общее количество покупок i -го пользователя.

Тогда для t-теста гипотеза будет выглядеть как:

$$\begin{cases} H_0 : \theta = R_C - R_T = 0 \\ H_A : \theta = R_C - R_T \neq 0 \end{cases}$$

Тестовая статистика записывается как:

$$T = \frac{R_C - R_T}{\sqrt{\frac{\sigma^2(R_C)}{N_C} + \frac{\sigma^2(R_T)}{N_T}}}$$

Для проверки гипотез с Ratio-метриками можно использовать бутстрэп. Алгоритм будет выглядеть следующим образом:

1. Делаем подвыборку пользователей (единицей рандомизации является пользователь, а не транзакция и тд).
2. Считаем Ratio-метрику в тестовой и контрольной группах.
3. Считаем разницу между подвыборками, строим распределение разницы Ratio-метрик в обеих группах.
4. Если 0 входит в доверительный интервал, то H_0 не отвергается.

Однако как было сказано ранее, бутстрэп очень долгий, поэтому есть другие методы, с помощью которых можно проверить гипотезу с Ratio-метрикой.

7 Дельта-метод

Можно заметить, что дисперсию метрики R сложно посчитать, поскольку имеет вид $f(X, Y) = \frac{X}{Y}$. В таком случае для подсчета ее дисперсии необходимо использовать дельта-метод, который заключается в разложении функции в ряд Тейлора до первого члена в точке истинного значения параметра, в случае ratio метрики истинное значение каждой из компонент функции - математическое ожидание: $\theta = (\mathbb{E}[X], \mathbb{E}[Y]) = (\mu_X, \mu_Y)$. Тогда разложение ratio метрики $f(X, Y)$ это будет выглядеть как:

$$f(X, Y) = f(\theta) + f'_X(\theta)(X - \mu_X) + f'_Y(\theta)(Y - \mu_Y) + \bar{o}(\dots)$$

Посчитаем математическое ожидание, которое далее поможет при расчете дисперсии. Также зафиксируем, что $\mathbb{E}[\bar{o}(\dots)] = 0$:

$$\mathbb{E}[f(X, Y)] = \mathbb{E}[f(\theta)] + f'_X \mathbb{E}[(X - \mu_X)] + f'_Y \mathbb{E}[(Y - \mu_Y)] = \mathbb{E}[f(\theta)] = \frac{\mu_X}{\mu_Y}$$

Теперь считаем дисперсию:

$$\begin{aligned} Var(f(X, Y)) &= \mathbb{E}[(f(X, Y) - \mathbb{E}[f(X, Y)])^2] = \mathbb{E}[(f(X, Y) - f(\theta))^2] = \\ &= \mathbb{E}[(f'_X(\theta)(X - \mu_X) + f'_Y(\theta)(Y - \mu_Y))^2] = \\ &= \mathbb{E}[(f'_X(\theta))^2(X - \mu_X)^2] + \mathbb{E}[(f'_Y(\theta))^2(Y - \mu_Y)^2] + 2\mathbb{E}[f'_X(\theta)f'_Y(\theta)(X - \mu_X)(Y - \mu_Y)] \end{aligned}$$

Все функции производные функции f выносятся из матожидания как детерминированные

и остается, что $\mathbb{E}[(X - \mu_X)^2] = Var(X)$, $\mathbb{E}[(Y - \mu_Y)^2] = Var(Y)$, $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = Cov(X, Y)$:

$$Var(f(X, Y)) = (f'_X(\theta))^2 Var(X) + 2f'_X f'_Y Cov(X, Y) + (f'_Y(\theta))^2 Var(Y) = \\ = \frac{1}{\mu_Y^3} Var(X) + 2\frac{\mu_X}{\mu_Y^3} Cov(X, Y) + \frac{\mu_X^2}{\mu_Y^4} Var(Y)$$

Теперь если проецировать дельта-метод на ratio метрику:

$$R = \frac{X_1 + \dots + X_N}{Y_1 + \dots + Y_n} = \frac{\bar{X}}{\bar{Y}}$$

То ее дисперсия будет выглядеть следующим образом:

$$Var(R) = Var\left(\frac{X_1 + \dots + X_N}{Y_1 + \dots + Y_n}\right) = Var\left(\frac{\bar{X}}{\bar{Y}}\right) \\ Var\left(\frac{\bar{X}}{\bar{Y}}\right) = \frac{1}{\mu_Y^3} Var(\bar{X}) + 2\frac{\mu_X}{\mu_Y^3} Cov(\bar{X}, \bar{Y}) + \frac{\mu_X^2}{\mu_Y^4} Var(\bar{Y}) = \\ = \frac{1}{N\mu_Y^3} Var(X) + 2\frac{\mu_X}{N\mu_Y^3} Cov(X, Y) + \frac{\mu_X^2}{N\mu_Y^4} Var(Y)$$

N во множителе $\frac{\mu_X}{N\mu_Y^3}$ перед $Cov(X, Y)$ возникает из-за того, что X и Y зависимы. Если бы они были независимы, то при подсчете $Cov(\bar{X}, \bar{Y})$ вынеслось бы $\frac{1}{N^2}$ и при подсчете всех возможных пар комбинаций X и Y получилось бы $N^2Cov(X, Y)$, N бы сократилось.

Какое распределение тогда будет иметь статистика $T = \frac{R_C - R_T}{\sqrt{\frac{\sigma^2(R_C)}{N_C} + \frac{\sigma^2(R_T)}{N_T}}}$? Дельта-метод заключается в том, что если статистика имеет асимптотически нормальное распределение, то и функция от этой статистики тоже будет асимптотически нормальна. Так как наша ratio метрика является функцией от двух выборочных средних, которые асимптотически нормальны, то и функция будет асимптотически нормальна, поэтому:

$$R_C - R_T \sim \mathcal{N}(0, \sigma^2(R_C) + \sigma^2(R_T))$$

А значит статистика T тоже будет иметь нормальное распределение:

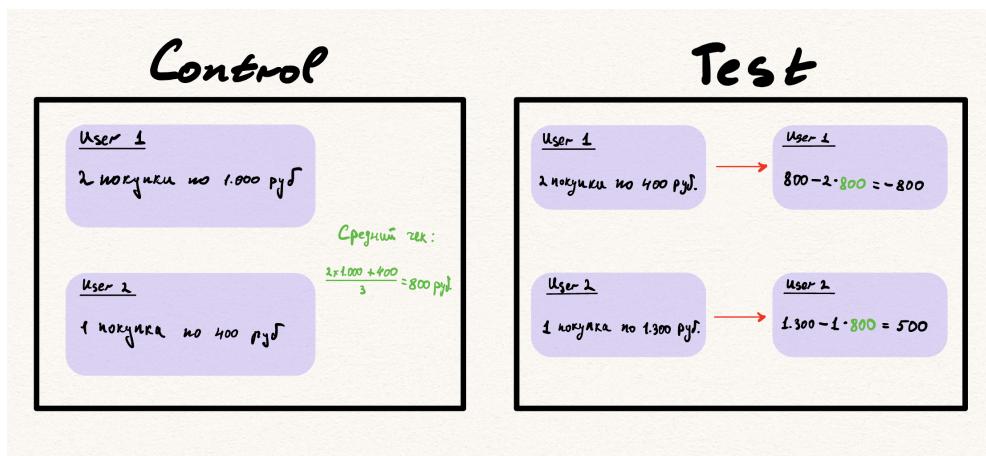
$$T = \frac{R_C - R_T}{\sqrt{\frac{\sigma^2(R_C)}{N_C} + \frac{\sigma^2(R_T)}{N_T}}} \sim \mathcal{N}(0, 1)$$

И гипотезу о равенстве ratio метрик можно проверять с помощью Z-теста.

8 Линеаризация

Также для того, чтобы избавиться от зависимых наблюдений можно перейти к новой поуверенной метрике, которая будет показывать насколько каждый пользователь из тестовой группы отклоняется от "усредненного" поведения в контрольной группе. В этом и суть линеаризации.

Например, есть группы А (контрольная) и Б (тестовая). Считаем по контрольной группе среднее значение метрики (например, для среднего чека $\frac{\sum_{u \in U_{\text{users}}} Price_u}{\# \text{Заказов}}$) и далее для каждого пользователя вычитаем из потраченной им суммы среднее значение метрики по контрольной группе, умноженной на количество заказов, сделанных этим пользователем:



Можно сделать то же самое и для тестовой группы (для каждого юзера посчитать отклонение), но можно посчитать, что сумма всех отклонений в тестовой группе будет равняться 0.

Теперь внутри каждой группы усредняем все отклонения. Так как сумма всех отклонений в контрольной группе равна 0, то и среднее отклонений тоже будет равно 0, но в тестовой группе среднее уже не будет нулевым и можно проверять гипотезу.

Формально, значение линеаризованной метрики можно записать как:

$$L_C = \frac{\sum_u l(u)}{|\{u : u \in C\}|}$$

$$L_T = \frac{\sum_u l(u)}{|\{u : u \in T\}|}$$

$$l(u) = X(u) - kY(u), \quad k = R_c$$

Также можно показать, что линеаризованная метрика будет сонаправлена с исходной метрикой и проверять гипотезы на линеаризованных метриках корректно. То есть нужно показать, что $\Delta(L) = Y_T \Delta(R)$, $\Delta(\bigodot) = \bigodot_C - \bigodot_T$:

$$\Delta(L) = L_C - L_T = \frac{1}{U_C} \sum_{u \in C} (X(u) - R_C Y(u)) - \frac{1}{U_T} \sum_{u \in T} (X(u) - R_C Y(u))$$

$$U_C = |\{u : u \in C\}|, U_T = |\{u : u \in T\}|$$

Введем некоторые обозначения:

$$X_C = \frac{1}{U_C} \sum_u X(u)$$

$$X_T = \frac{1}{U_T} \sum_u X(u)$$

$$Y_C = \frac{1}{U_C} \sum_u Y(u)$$

$$Y_T = \frac{1}{U_T} \sum_u Y(u)$$

И перепишем разницу линеаризованной метрики как:

$$\Delta(L) = X_C - R_C Y_C - X_T + R_C Y_T = (X_C - X_T) - R_C (Y_C - Y_T)$$

$$R_C = \frac{X_C}{Y_C}$$

$$\Delta(L) = (X_C - X_T) - \frac{X_C}{Y_C} (Y_C - Y_T) = X_C - X_T - X_C + \frac{X_C Y_T}{Y_C} =$$

$$= Y_T \left(\frac{X_T}{Y_T} - \frac{X_C}{Y_C} \right) = Y_T \Delta(R)$$

Поскольку мы избавились от зависимости между наблюдениями и перешли к поозерным метрикам (поведение юзеров не зависит друг от друга), то можно использовать Z-тест:

$$T = \frac{L_C - L_T}{\sqrt{\frac{\sigma^2(L_C)}{N_C} + \frac{\sigma^2(L_T)}{N_T}}} \sim \mathcal{N}(0, 1)$$

9 Перевзвешивание

Один из недостатков линеаризации - все пользователи вносят одинаковый вклад в метрику, однако мы знаем, что у каких-то пользователей больше заказов, а у каких-то меньше. Идея перевзвешивания заключается в том, чтобы модернизировать линеаризацию с помощью добавления информации о количестве заказов каждого пользователя. То есть при подсчете метрики мы хотим учесть "активность" каждого пользователя.

Теперь рассмотрим пример с CTR:

$$CTR = \frac{\sum_u Clicks_u}{\sum_u Shows_u}$$

Мы хотим добавить информацию о том какая доля показов конвертировалась в клик для каждого пользователя. Поэтому в числителе для каждого пользователя появится множитель $\frac{Clicks_u}{Shows_u}$ - CTR каждого пользователя. Но нам необходимо как-то учесть и количество показов. Нельзя просто при суммировании домножать на $Shows_u$, потому что получится обычный CTR . Поэтому приняли решение домножать на что-то пропорциональное $Shows_u$, а именно использовать $\sqrt{Shows_u}$. Тогда новая формула для CTR будет считаться как:

$$RCTR = \frac{\sum_u \sqrt{Shows_u} \frac{Clicks_u}{Shows_u}}{\sum_u \sqrt{Shows_u}}$$

Также можно применить и линеаризацию к полученной метрике:

$$LRCTR = \frac{\sum_u \sqrt{Shows_u} (Clicks_u - k \times Shows_u)}{\sum_u \sqrt{Shows_u}}$$

Но тут может возникнуть вопрос как считать коэффициент k : сначала сделать перевзвешивание и посчитать k по новой метрике или вначале посчитать k по исходной метрике и сделать перевзвешивание? Ответ: сначала сделать перевзвешивание, а затем посчитать k .

10 Бакетизация

Смысл бакетизации - разделение пользователей по бакетам. То есть создаются m бакетов и все пользователи рандомно распределяются по этим бакетам. Затем внутри каждого бакета считается целевая метрика, получаем m целевых метрик, которые мы будем считать независимыми друг от друга, поэтому в этом случае применим t-тест.

Число бакетов - гиперпараметр, который стоит подбирать перед проведением теста.

Алгоритм бакетизации:

-
1. Выбирается идентификатор (например, user_id).
 2. Алгоритм использует хеш-функцию для распределения пользователей в бакеты (например, по остатку от деления хеша на количество бакетов).
 3. Каждый бакет привязывается к группе эксперимента (например, если бакет номер чётный — это контрольная группа, если нечётный — тестовая группа).

11 Стратификация

Делим выборку на страты, при этом знаем размер каждой страты. Далее генерируем случайную выборку внутри каждой страты. Либо же можно просто включить индикатор попадания индивида в какой либо кластер при оценки регрессии.

Пример, Uber. В рамках одного города они допускали зависимость между наблюдениями, но корреляция между разными городами отсутствовала. В этом примере кластеры - города. Далее при оценке уравнения линейной регрессии использовали робастные кластерные ошибки.

Введем L - кол-во страт, i_l - конкретное наблюдение из L -ой страты ($i_l = \overline{1, N_L}$), $\sum_{l=1}^L N_l = N$. Это было в терминах генеральной совокупности, для выборок будет аналогично: i_l - конкретное наблюдение из L -ой страты ($i_l = \overline{1, n_L}$), $\sum_{l=1}^L n_l = n$, где n - размер выборки, N - размер генеральной совокупности. Введем также вес каждой страты: $w_L = \frac{N_L}{N}$.

Среднее будет считаться как

$$\mu = \frac{1}{N} \sum_{l=1}^L \sum_{i=1}^{N_l} x_{il}$$

Видоизменим это выражение $\mu = \sum_{l=1}^L \frac{N_l}{N} \frac{1}{N_l} \sum_{i=1}^{N_l} x_{il}$, где $\frac{N_l}{N}$ - вес каждой страты (w_l), $\frac{1}{N_l} \sum_{i=1}^{N_l} x_{il}$ - математическое ожидание внутри каждой страты (μ_l). Итого:

$$\mu = \sum_{l=1}^L w_l \mu_l$$

Также можно и посчитать дисперсию:

$$\sigma^2 = \frac{1}{N} \sum_{l=1}^L \sum_{i=1}^{N_l} (x_{il} - \mu)^2 = \sum_{l=1}^L \frac{N_l}{N} \frac{1}{N_l} \sum_{i=1}^{N_l} (x_{il} - \mu_l + \mu_l - \mu)^2 = \sum_{l=1}^L w_l \frac{1}{N_l} \sum_{i=1}^{N_l} (x_{il} - \mu_l)^2 + \sum_{l=1}^L w_l \frac{1}{N_l} \sum_{i=1}^{N_l} (\mu_l - \mu)^2$$

$\frac{1}{N_l} \sum_{i=1}^{N_l} (x_{il} - \mu_l)^2$ - это дисперсия внутри страты σ_l^2 , а во втором слагаемом мы суммируем N_l раз одно слагаемое, поэтому сумма $\frac{1}{N_l} \sum_{i=1}^{N_l} (\mu_l - \mu)^2$ превратится в $(\mu_l - \mu)^2$. Итоговая

дисперсия будет выглядеть как

$$\sigma^2 = \sum_{l=1}^L w_l \sigma_l^2 + \sum_{l=1}^L w_l (\mu_l - \mu)^2$$

, где первое слагаемое - внутригрупповая дисперсия, а второе слагаемое - межгрупповая дисперсия. Мы не знаем вес каждой страты w_l и хотим найти их таким образом, чтобы понять, в каких пропорциях выбирать людей из каждой страты так, чтобы общая дисперсия уменьшилась.

Теперь на примере с выборкой. Пусть для каждой страты есть какие-то наблюдения:

$$\forall l : n_l, x_{il}, \dots, x_{nl}, n_1 + \dots + n_L = n$$

Выпишем статифицированное среднее:

$$\bar{X}_{strat} = \sum_{l=1}^L w_l \bar{X}_l, \quad \bar{X}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{il}, \quad \sum_{l=1}^L w_l = 1$$

и распишем его математическое ожидание и дисперсию:

$$\mathbb{E}[\bar{X}_{strat}] = \sum_{l=1}^L w_l \frac{1}{n_l} \sum_{i=1}^{n_l} \mathbb{E}[x_{il}] = \sum_{l=1}^L w_l \frac{1}{n_l} \sum_{i=1}^{n_l} \mu_l = \sum_{l=1}^L w_l \mu_l = \mu$$

$$Var(\bar{X}_{strat}) = \sum_{l=1}^L w_l^2 \frac{1}{n_l^2} \sum_{i=1}^{n_l} Var(x_{il}) = \sum_{l=1}^L w_l^2 \frac{1}{n_l^2} \sum_{i=1}^{n_l} \sigma_l^2 = \sum_{l=1}^L w_l^2 \frac{\sigma_l^2}{n_l^2}$$

Если же рассмотреть дисперсию обычного среднего, то получится:

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \sum_{l=1}^L \frac{w_l}{n} \sigma_l^2 + \sum_{l=1}^L \frac{w_l}{n} (\mu_l - \mu)^2$$

И интуитивно кажется, что $Var(\bar{X})$ больше, чем $Var(\bar{X}_{strat})$. Тогда итоговая задача стратификации для нахождения оптимальных весов записывается как:

$$\begin{cases} Var(\bar{X}_{strat}) = \sum_{l=1}^L \frac{w_l^2 \sigma_l^2}{n_l} \rightarrow \min_{w_1, \dots, w_L} \\ n_1 + \dots + n_L = n \end{cases}$$

Тогда выписывая лагранжиан и решая такую задачу оптимизации получаются следующие размеры каждой страты:

$$\mathcal{L} = Var(\bar{X}_{strat}) - \lambda(n_1 + \dots + n_L - n) \Rightarrow n_l^* = \frac{w_l \sigma_l}{\sum_{k=1}^L w_k \sigma_k} n, \quad w_l^* = \frac{n_l}{n}$$

То есть у нас есть генеральная совокупность пользователей, на ней мы посчитали истинные доли каждой страты $\frac{N_l}{N}$. Когда захоти построить выборку размером $n < N$, то количество наблюдений в выборке из каждой страты будет равняться $\frac{N_l}{N}n$.

Есть два способа использования стратификации:

11.1 Стратификация

1. Разбиваем генеральную совокупность на нескоррелированные страты (пол, возраст и тд).
2. Считаем веса w_l каждой страты по генеральной совокупности.
3. Генерируем выборку так, чтобы сохранялись веса каждой страты w_l в выборке.
4. Считаем стратифицированные статистики.

11.2 Постстратификация

1. Разбиваем генеральную совокупность на нескоррелированные страты (пол, возраст и тд).
2. Считаем веса w_l каждой страты по генеральной совокупности.
3. Генерируем выборку случайно без привязки к стратам.
4. Считаем стратифицированные статистики (используем w_l , посчитанные в пункте 2).

12 CUPED

По дефолту все А/Б тесты основываются на проверке равенства средних:

$$t = \frac{\bar{Y}^{(T)} - \bar{Y}^{(C)}}{se(\bar{Y}^{(T)} - \bar{Y}^{(C)})}$$

$n^{(T)}$ и $n^{(C)}$ достаточно велики, поэтому можно пользоваться нормальным распределением и ДИ будет выглядеть как $[\bar{Y}^{(T)} - \bar{Y}^{(C)} - 1.96 \times se(...); ...]$. Хоть наблюдений и много, но мы все равно хотим, чтобы интервал был уже. Этого добиться можно за счет снижения se .

Идея: найти предиктор, который коррелирует с метрикой, но при этом не коррелирует с дамми-переменной воздействия (другими словами предиктор не участвовал при разбиении пользователей на группы контроля и воздействия). Также предполагаем, что воздействие назначается рандомно.

Для каждого индивида считаем $\hat{Y}_{CUPED_i}^{(T)} = Y_i^{(T)} - \hat{\theta}X_i^{(T)}$, $\hat{\theta} = \frac{cov(X^{(T)}, Y^{(T)})}{Var(X^{(T)})}$. Идея в том, что математические ожидания $\mathbb{E}[\bar{Y}^{(T)}]$ и $\mathbb{E}[\hat{Y}_{CUPED}^{(T)}]$ будут совпадать, но заметно снизится стандартная ошибка. В таком случае ДИ будет выглядеть $[\hat{Y}_{CUPED}^{(T)} - \hat{Y}_{CUPED}^{(C)} - 1.96 \times se(...); ...]$.

Тривиальный подход. Рассмотрим базовый случай без использования CUPED. Пусть i - индивид, z_i - переменная воздействия, $z_i \in \{0, 1\}$. Тогда эффект воздействия можно будет оценить с помощью линейной регрессии: $y_i = \beta_1 + \beta_2 \times z_i + u_i$, причем можно оценить при разных предположениях u_i (гомо/гетероскедастичность). С помощью МНК получаем $\hat{\beta}_2$ и проверяем на значимость этот коэффициент. Недостаток такого подхода - он не пользуется всей доступной информацией, поэтому и интервал слишком широкий.

Элегантный способ. Теперь учтем контрольную переменную x_i в уравнении регрессии в виде какой-то (необязательно линейной) функции f : $y_i = \beta_1 + \beta_2 \times z_i + f(x_i) + u_i$. Перед нами не стоит задача верно определить зависимость y_i от $f(x_i)$, нам также интересно проверить на значимость β_2 . Оцениваем регрессию $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \times z_i + \hat{\beta}_3 x_i$, при этом коэффициент $\hat{\beta}_3$ нас не интересует, правильную ли зависимость мы учли или нет, и аналогично проверяем на значимость β_2 . Причем поскольку мы указали, что z_i и x_i независимы, то оценки $\hat{\beta}_2$ в тривиальном и элегантном подходах будут совпадать по теореме Фриша-Бау-Ловелла.

CUPED. Из двух предыдущих способов мы не до конца выжали информацию. Рассмотрим оценку ковариационной матрицы коэффициентов в линейной регрессии:

$$Var(\hat{\beta}|X) = (X'X)^{-1}\sigma^2$$

где $X = \begin{bmatrix} 1 & z_1 & x_1 \\ 1 & z_2 & x_2 \\ 1 & z_3 & x_3 \\ \dots & \dots & \dots \end{bmatrix}$. Для начала умножим и поделим исходное выражение с

$$Var(\hat{\beta}|X) = \left(\frac{X'X}{n}\right)^{-1}\sigma^2 \text{ и распишем } \frac{X'X}{n}:$$

$$\frac{X'X}{n} = \begin{bmatrix} \frac{\sum 1^2}{n} & \frac{\sum 1 \times z_i}{n} & \frac{\sum 1 \times x_i}{n} \\ \frac{\sum z_i \times 1}{n} & \frac{\sum z_i^2}{n} & \frac{\sum z_i \times x_i}{n} \\ \frac{\sum x_i \times 1}{n} & \frac{\sum x_i \times z_i}{n} & \frac{\sum x_i^2}{n} \end{bmatrix}$$

Если отбросить из рассмотрения свободный коэффициент и рассматривать ковариационную матрицу только для коэффициентов $\hat{\beta}_2$ и $\hat{\beta}_3$ (выкидываем первый столбец и первую строку из $\frac{X'X}{n}$), то обращенная матрица будет выглядеть следующим образом:

$$\left(\frac{X'X}{n}\right)^{-1} = \frac{n^2}{(\sum z_i^2)(\sum x_i^2) - (\sum z_i \times x_i)^2} \begin{bmatrix} \frac{\sum x_i^2}{n} & -\frac{\sum z_i \times x_i}{n} \\ -\frac{\sum z_i \times x_i}{n} & \frac{\sum z_i^2}{n} \end{bmatrix}$$

Однако можно заметить, что $\text{plim}_{n \rightarrow \infty} \frac{\sum z_i \times x_i}{n} = \mathbb{E}[z \times x]$ (ЗБЧ). Центрируем все переменные и перепишем матрицу $\frac{X'X}{n}$ (также отбросим свободный коэффициент):

$$\frac{X'X}{n} = \begin{bmatrix} \frac{\sum(z_i - \bar{z})^2}{n} & \frac{\sum(z_i - \bar{z}) \times (x_i - \bar{x})}{n} \\ \frac{\sum(x_i - \bar{x}) \times (z_i - \bar{z})}{n} & \frac{\sum(x_i - \bar{x})^2}{n} \end{bmatrix}$$

И теперь элементы, находящиеся на побочной диагонали по вероятности будет сходиться к $\text{cov}(x, z)$, которая, как нам известно, равна 0 из-за того, что мы подбирали x таким образом, чтобы он не был скоррелирован с переменной воздействия. Тогда матрица будет выглядеть следующим образом:

$$\frac{X'X}{n} = \begin{bmatrix} \frac{\sum(z_i - \bar{z})^2}{n} & 0 \\ 0 & \frac{\sum(x_i - \bar{x})^2}{n} \end{bmatrix}$$

И при подсчете обратной матрицы ее определитель увеличится \Rightarrow дисперсия коэффициента уменьшится:

$$\left(\frac{X'X}{n}\right)^{-1} = \frac{n^2}{(\sum z_i^2)(\sum x_i^2) - 0} \begin{bmatrix} \frac{\sum x_i^2}{n} & 0 \\ 0 & \frac{\sum z_i^2}{n} \end{bmatrix}$$

Алгоритм CUPED выглядит следующим образом:

1. Оцениваем регрессию $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 z_i + \hat{\beta}_3 x_i$.
2. Считаем остатки $r_i = y_i - \hat{\beta}_3 x_i$.
3. Строим новую регрессию $\hat{r}_i = \hat{\gamma}_1 + \hat{\gamma}_2 z_i$.

По той же теореме Фриша-Вау-Ловелла оценки $\hat{\gamma}_1$, $\hat{\gamma}_2$ и $\hat{\beta}_1$, $\hat{\beta}_2$ будут равны соответственно. Глобальный смысл в том, что при оценивании МНК в компьютере он не будет знать структуру ковариационной матрицы для коэффициентов (то, что на побочной диагонали будут нули). Также же можно сказать, что мы "очищаем" метрику от влияния других факторов.

Есть расширения CUPED, можно использовать не один предиктор, а несколько. Такой подход называется **CUMPED**. Идея заключается в том, что оцениваем уравнение регрессии с несколькими контрольными переменными, которые также должны удовлетворять предпосылкам независимости переменной воздействия от этих предикторов: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 z_i + \hat{\beta}_3 x_i + \hat{\beta}_4 a_i + \dots$

Причем в части $\hat{\beta}_3 x_i + \hat{\beta}_4 a_i + \dots$, необязательно использовать линейные модели, можно использовать ML.

Один из актуальных вопросов, а как гарантировать, что переменная воздействия не зависит от выбранных контрольных переменных? Для этого случая в качестве контрольной переменной используют значение метрики до проведения эксперимента, так можно гарантировать независимость с переменной воздействия. В для остальных предикторов на свой страх и риск.

Может возникнуть случай, что нам неизвестно значение метрики до проведения эксперимента для некоторых наблюдений. В таком случае заполняем значение метрики константой для новых пользователей и вводим дамми-переменную d_i , которая будет принимать значение 1, если для индивида известно значение метрики до эксперимента, 0 - если нет. Тогда все предыдущие методы будут работать.

13 Множественное тестирование

Часто возникает потребность в проведении А/Б/С тестов, но тут приходится сравнивать сразу 3 группы. Например, тестируется эффективность пуш-уведомление о скидке. В этом примере будут А группа - нет изменений, Б группа - рассыпается простое уведомление пользователю о приложении (не несет в себе никакой скидки), С группа - пуш-уведомление со скидкой. Мы не можем в группе Б сразу внедрить пуш-уведомление + скидку, потому что не знаем, что повлияло на рост метрики - уведомление или скидка. Тогда мы сравниваем А и Б группы, есть ли эффект от уведомления. Если да, то дальше сравниваем группы Б и С чтобы понять, насколько эффективно наше предложение.

Но в примере выше результат одного теста зависит от результатов другого. Это и называется проблемой множественного тестирования. В таком случае растет ошибка I-го рода: $FPR = 1 - (1 - \alpha)^m$, где m - количество тестируемых гипотез. Это называется *FWER* (Family-Wise Error Rate) - вероятность совершить ошибку I-го рода хотя бы в одном из проводимых тестов:

$$FWER = \mathbb{P}(FP > 0) = 1 - (1 - \alpha)^m$$

Есть несколько способов, с помощью которых мы можем провести множественное тестирование и при этом сохранить вероятность ошибки I-го рода на изначально заданному уровне значимости.

13.1 Поправка Бонферонни

Основной идеей заключается деление изначального уровня значимости на количество проводимых тестов. Таким образом ставим более жесткие рамки для каждой гипотезы и уменьшаем вероятность ошибки I-го рода на каждом из тестов. Итого каждая из m гипотез тестируется на уровне значимости:

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

Но как раз из-за более жестких рамок мы все реже будем отвергать H_0 и реже будем ловить эффект там, где он действительно есть. То есть мощность проводимого теста значительно упадет.

13.2 Метод Холма

Метод Холма равномерно более мощный критерий, чем поправка Бонферонни, поэтому и мощность проводимого теста будет больше. Процедура следующая:

1. Сортируем p-value в порядке возрастания:

$$p-value_{(1)} \leq \dots \leq p-value_{(m)}$$

И ставим в соответствие нулевые гипотезы, которые тестируются $H_{(1)}, \dots, H_{(m)}$.

2. Далее итеративно от наименьшего p-value к большему тестируем гипотезы по следующему правилу:

- Если $p-value_{(1)} \geq \frac{\alpha}{m-1+1}$, принимаем гипотезы $H_{(1)}, \dots, H_{(m)}$. Иначе отвергаем $H_{(1)}$ и проверяем оставшиеся гипотезы на уровне значимости $\frac{\alpha}{m-2+1}$.
- Если $p-value_{(2)} \geq \frac{\alpha}{m-2+1}$, принимаем гипотезы $H_{(2)}, \dots, H_{(m)}$. Иначе отвергаем $H_{(2)}$ и проверяем оставшиеся гипотезы на уровне значимости $\frac{\alpha}{m-3+1}$.
- ...
- Если $p-value_{(i)} \geq \frac{\alpha}{m-i+1}$, принимаем гипотезы $H_{(i)}, \dots, H_{(m)}$. Иначе отвергаем $H_{(i)}$ и проверяем оставшиеся гипотезы на уровне значимости $\frac{\alpha}{m-(i+1)+1}$.

13.3 FDR

Иная метрика, которая учитывает ошибку II-го рода, но не так жестко контролирует ошибку I-го рода. Интерпретируется как ожидаемое количество ложных отклонений H_0 (FP) от всех отклоненных H_0 (FP + TP):

$$FDR = \frac{FP}{FP + TP}$$

Напоминание про проверки гипотез в терминах Confusion Matrix:

	H_0 верна	H_A верна
Отвергли H_0	FP / α	TP / $1 - \beta$
Не отвергли H_0	TN	FN / β

Мы стараемся минимизировать FDR : в числителе стоит ошибка I-го рода (FP), которую мы хотим сделать наименьшей, а в знаменателе стоит мощность (TP), которую мы хотим увеличить.

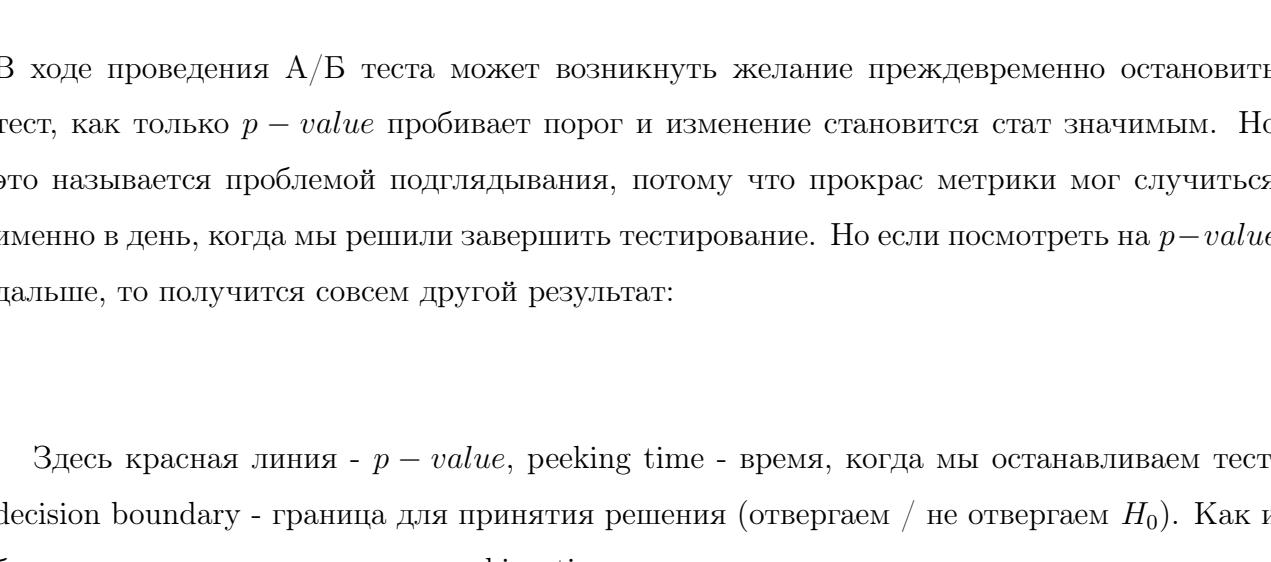
В рамках этой метрики мы сравниваем все поправки и выбираем ту, у которой значение метрики FDR меньше.

Встает вопрос: когда лучше использовать $FWER$, а когда FDR ?

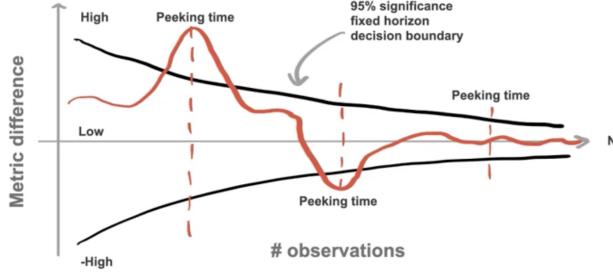
- В терминах А/Б тестирования лучше использовать FDR , потому что мы будем пропускать меньше реальных различий между выборками, учитывая ошибку II рода. При этом строгость в отношении ошибки I рода для нас не будет очень важна.
- Если же нам наоборот важнее ошибка I-го рода (например, в медицине), то стоит выбрать $FWER$ для перестраховки.

14 Последовательное тестирование

В ходе проведения А/Б теста может возникнуть желание преждевременно остановить тест, как только $p-value$ пробивает порог и изменение становится стат значимым. Но это называется проблемой подглядывания, потому что прокрас метрики мог случиться именно в день, когда мы решили завершить тестирование. Но если посмотреть на $p-value$ дальше, то получится совсем другой результат:



Здесь красная линия - $p-value$, *peeking time* - время, когда мы останавливаем тест, *decision boundary* - граница для принятия решения (отвергаем / не отвергаем H_0). Как и было сказано ранее, в момент *peeking time* случился прокрас, но если смотреть дальше, то эффект от нововведения отсутствует.



Если мы будем завершать тестирование раньше расчитанного времени, то значительно повысится ошибка I-го рода. Чтобы этого избежать используется метод последовательного тестирования (SPRT).

Гипотеза для последовательного тестирования используется следующая:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_A : \theta = \theta_1 \end{cases}$$

Которую можно проверить с помощью критерия, основанного на отношении правдоподобия:

$$\Lambda_k = \prod_{i=1}^k \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)}, \quad k = 1, 2, ..$$

Задачей является подобрать такие пороги γ_0 и γ_1 , чтобы при получении нового наблюдения k действовало следующее правило:

- Если $\Lambda_k \geq \gamma_1$, то отвергаем H_0 и останавливаем тест.
- Если $\Lambda_k \leq \gamma_0$, то не отвергаем H_0 и останавливаем тест.

Также пороги должны быть установлены таким образом, чтобы удалось сохранить ошибку I-го рода и мощность.

Рассмотрим выборку $x = (x_1, \dots, x_k)$ и введем область, где отвергается H_0 ($R_1 = \{x : \Lambda_k \geq \gamma_1\}$), и область, где не отвергаем H_0 ($R_0 = \{x : \Lambda_k \leq \gamma_0\}$). Тогда мощность можно записать как:

$$1 - \beta = \int_{R_1} p_{\theta_1}(x) dx = \int_{R_1} \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} p_{\theta_0}(x) dx = \int_{R_1} \Lambda_k p_{\theta_0}(x) dx \geq \gamma_1 \int_{R_1} p_{\theta_0}(x) dx = \gamma_1 \alpha$$

Таким же образом можно расписать и вероятность ошибки I-го рода:

$$\begin{aligned} 1 - \alpha &= 1 - \int_{R_1} p_{\theta_0}(x) dx = \int_{R_0} p_{\theta_0}(x) dx = \int_{R_0} \frac{p_{\theta_0}(x)}{p_{\theta_1}(x)} p_{\theta_1}(x) dx = \\ &\int_{R_0} \Lambda_k^{-1} p_{\theta_1}(x) dx \geq \gamma_0^{-1} \int_{R_1} p_{\theta_0}(x) dx = \gamma_0^{-1} \beta \end{aligned}$$

В итоге получилась следующая система неравенств:

$$\begin{cases} \gamma_1 \leq \frac{1-\beta}{\alpha} \\ \gamma_0 \geq \frac{\beta}{1-\alpha} \end{cases}$$

Можно заменить все неравенства на равенства, ошибки I-го и II-го рода мы заранее зафиксировали, поэтому можно получить границы. Тогда с каждым новым наблюдением мы пересчитываем статистику Λ_k и принимаем решение тогда, когда Λ_k пробивает либо γ_1 , либо γ_0 .

Но проблема в том, что в А/Б тестах альтернативная гипотеза редко точечно задается. Поэтому используется метод mSPRT.

Мы не знаем точный размер эффекта, поэтому введем априорную вероятность размера эффекта $h(\theta)$. Тогда будем брать смесь отношений правдоподобий с учетом этой априорной вероятности. Тогда теперь тест строится следующим образом: мы все также тестируем нулевую гипотезу о равенстве истинного значения параметра θ значению θ_0 против смеси альтернативных гипотез, каждая из которых имеет некоторую вероятность $h(\theta)$. Теперь статистика считается следующим образом:

$$\Lambda_n^H = \int_{\Theta} \left(\frac{f_{\theta}(\theta_n)}{f_{\theta_0}(\theta_n)} \right)^n h(\theta) d\theta$$

Для подбора f_{θ} и $h(\theta)$ используются исторические данные. Проверять гипотезу мы будем с помощью $p-value$, которое считается на каждом шаге как:

$$p-value_n = \inf\{\alpha : T(\alpha) \leq n, \delta(\alpha) = 1\}$$

Где $T(\alpha)$ - размер выборки при остановке теста, $\delta(\alpha)$ - индикатор того, что H_0 отвергнута. Это можно представить в виде:

$$T^H(\alpha) = \inf\{n : \Lambda_n^H \geq \alpha^{-1}\}$$

$$\delta^H(\alpha) = \mathbb{1}\{T^H(\alpha) < \infty\}$$

Изначально мы предполагаем, что размер выборки бесконечный. Поэтому на каждом этапе считаем Λ_n^H и смотрим, превышает ли это значение α^{-1} . Если да, то мы получили конечную выборку, индикатор $\mathbb{1}\{T^H(\alpha) < \infty\}$ принимает значение 1 и мы можем посчитать $p-value$. Также этот p-value можно интерпретировать как "уровень значимости", при котором выполнились оба условия", потому что подбирая α мы по сути подбираем уровень значимости.

Итоговый алгоритм mSPRT выглядит следующим образом:

1. На исторических данных подбираем f_θ и $h(\theta)$.
2. С каждым новым наблюдением пересчитываем Λ_n^H .

3. Пересчитываем $p - value_n$ по формуле выше. Можно сделать итеративный метод:

$$p - value_0 = 1, p - value_n = \min(p - value_{n-1}, \frac{1}{\Lambda_n^H}).$$

4. Как только $p - value_n$ становится меньше заданного уровня значимости, то останавливаем тест и отвергаем H_0 .
5. Если $p - value_n$ не пробивает заданный уровень значимости долгое время, то принудительно останавливаем тест и не отвергаем H_0 .