Ratio метрика

T-test для ratio метрики напрямую не работает.

Как тестировать?

1. Бутстреп
2. Дельта-метод
3. Линеризация

Бутстреп

Как с помощью бутстрапа проверить гипотезу с метрикой отношения?

1. Делаем подвыборку пользователей (семплируем по объектам, а не по наблюдениям!)

2. Считаем в подвыборках метрику отношения в обеих группах

3. Считаем разницу между подвыборками и получаем распределение разницы метрики-отношения контроля и теста

4. Смотрим, попадает ли 0 в доверительный интервал. Если да, то нулевая гипотеза не отвергается на заданном уровне значимости

Дельта-метод

T-test не работает из-за зависимых данных. В чем причина?

Причина в неверной оценке дисперсии

$$t = \frac{\mathcal{R}_B - \mathcal{R}_A}{\sqrt{\mathbb{V}(\mathcal{R}_A) + \mathbb{V}(\mathcal{R}_B)}} \xrightarrow{N \to \text{inf}} \mathcal{N}(0, 1)$$

R_B – ratio метрика в группе B

$V(R_B)$ – дисперсия в группе B

Как правильно оценить дисперсию V(R) ?

$$V(R) = \frac{1}{N\mu_y^2} V(X) - 2\frac{\mu_x}{\mu_y^3}\text{cov}(X, Y) + \frac{\mu_x^2}{N\mu_y^4} V(Y)$$

(вывод можно посмотреть в лекции)

Формула выше для i.i.d. X,Y (в случае зависимых с.в. перед ковариацией 1/N тоже должно стоять)

Если будем использовать верную оценку дисперсии (выше), то T-test будет корректно работать

Линеризация

$$Lx, y, k(U) = X(u) - kY(u)$$

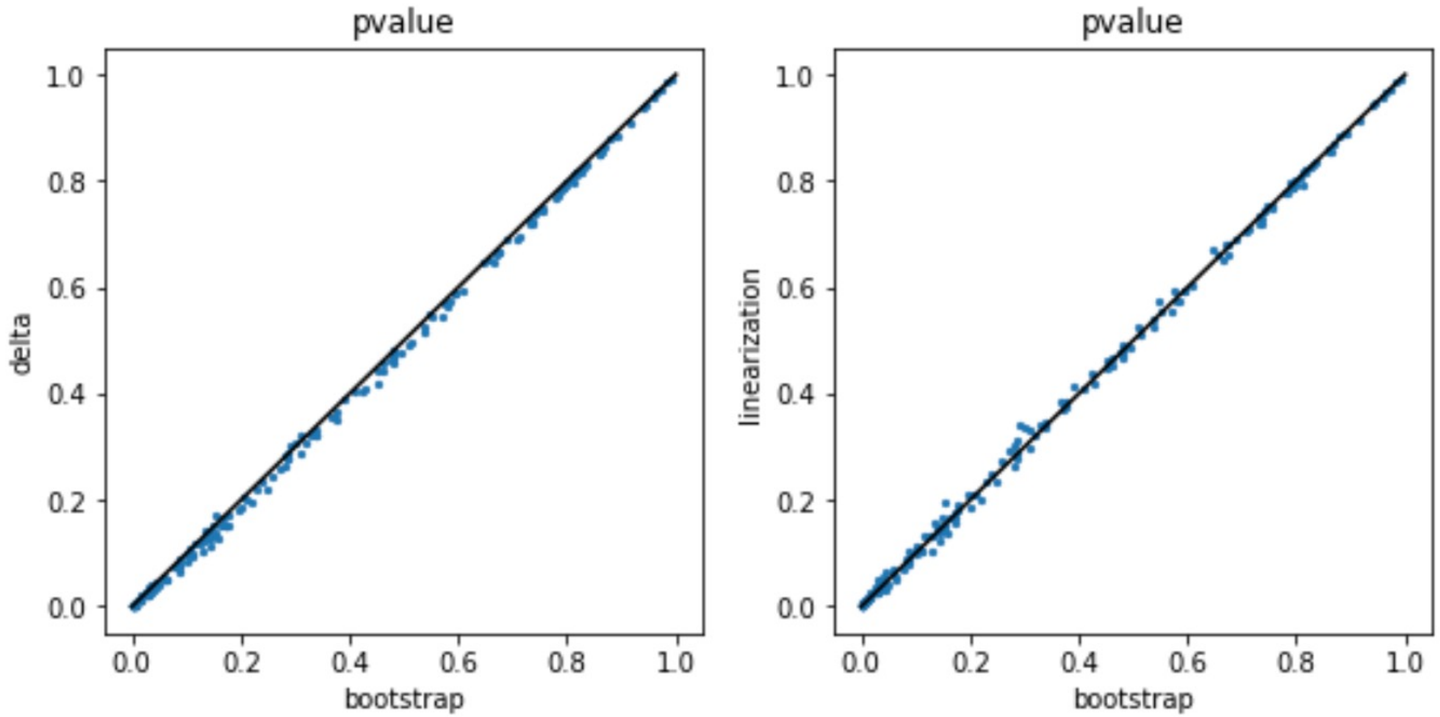
K – ratio по контролю
X – клики пользователя
Y – просмотры пользователя

Мы хотим понять отклонение метрики в эксперименте относительно ratio в контроле.

Иначе говоря – смотрим, что изменилось в эксперименте относительно ситуации в контроле

Свойства линеризации

- p-value L-метрики сонаправлен с p-value для bootstrap и дельта метода
- Сохраняется поюзерная направленность
- На L метрику можно использовать методы снижения дисперсии, что делает тест более чувствительным



Множественная проверка гипотез

Мы онлайн магазин, который продает мыло.

Сделали промо-пуши приложения доставки мыла. Выделили три когорты пользователей: первая – контроль, второй отправили пуш на скидку 30%, третьей отправил пуш на бесплатное третье мыло при заказе двух.

Вопрос, кто на 3 месяц принесет больше денег?

Контроль

-30%

2+1 Free

В чем проблема?

Контроль

-30%

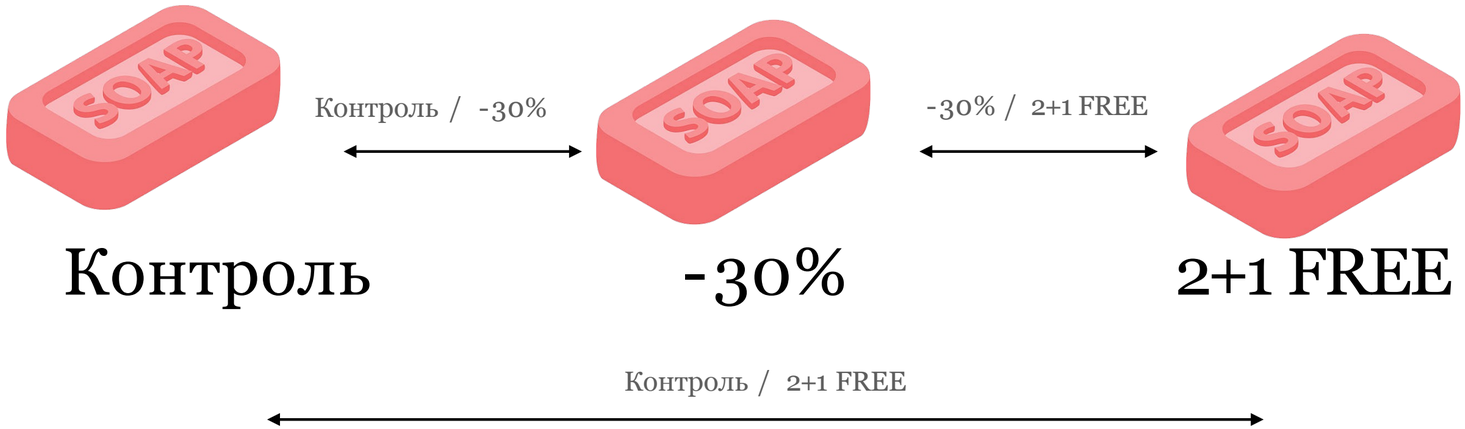
2+1 Free

В чем проблема?

False Positive Rate растет соразмерно $1-(1-\alpha)^m$

где m – количество гипотез

Проблема заключается в том, что при одновременной проверке большого числа гипотез на том же наборе данных вероятность сделать неверное заключение в отношении хотя бы одной из этих гипотез значительно превышает изначально принятый уровень значимости



FWER

Если сделать *N* тестов, то вероятность совершить хотя бы одну ошибку I рода в группе тестов (family-wise error rate, *FWER*) значительно возрастает согласно формуле

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

, где m – количество гипотез. В случае с 3 когортами у нас 3 попарных сравнения, т.е. мы проверяем 3 гипотезы: A/B, A/C, B/C. Это означает, что при уровне значимости 95%, альфа будет

$$1-(1-0.05)^3 = 0.142$$

FWER

Вероятность того, что тест не ошибется равна

$$0.95^{50} = 0.077$$

Или, если он ошибется хотя бы 1 раз

$$1-0.95^{50} = 0.92$$

FWER

Можно задрать уровень значимости, и при 10 гипотезах получим такие результаты (при разных альфа):

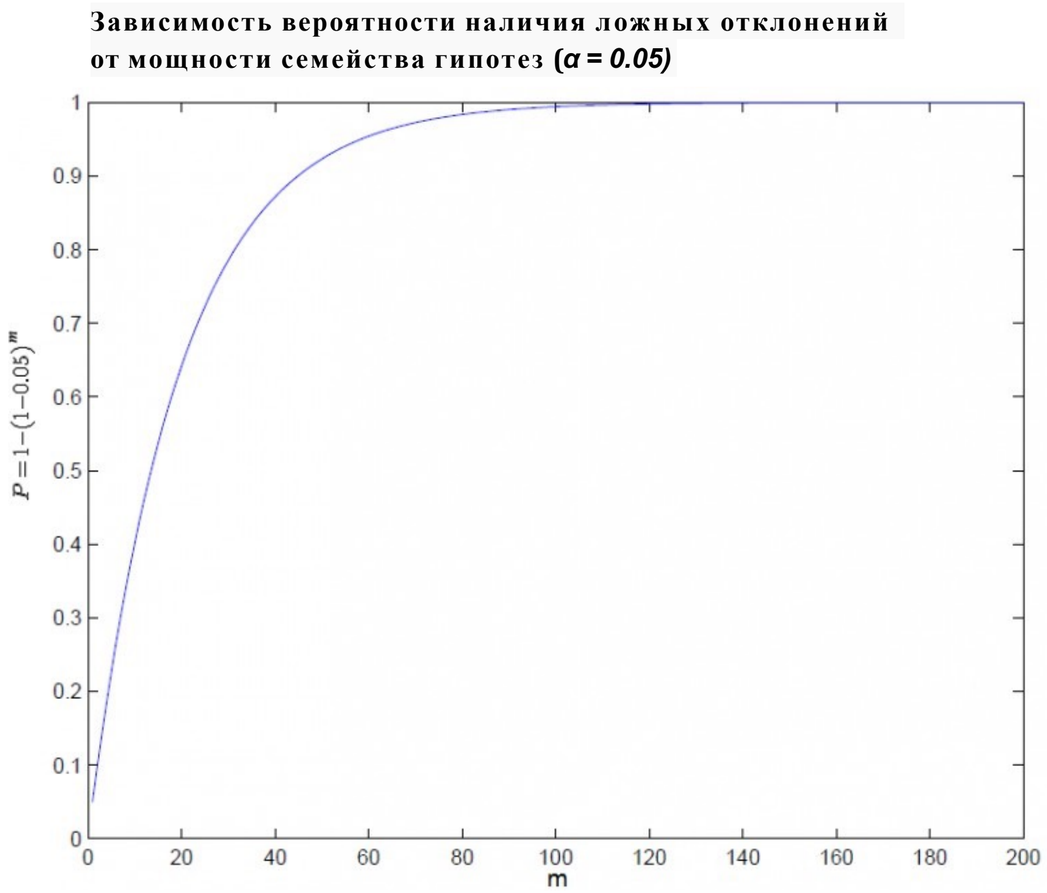
95% = 1-(1-0.05)^{10} = 0.401

99% = 1-(1-0.01)^{10} = 0.095

99.5% = 1-(1-0.005)^{10} = 0.095

99.9% = 1-(1-0.001)^{10} = 0.01

Т.е. для *α = 0.05* мы получим *40%* ошибку, а вовсе не *5%*, как изначально задается параметром



FWER

Пример. По результатам эксперимента вы получили несколько разных pvalue. Какова вероятность ошибиться в хотя бы в одной из метрик?

Pvals, если проверяем одну метрику

$$CR_1 = 0.021$$

Отклоняем нулевую гипотезу при $\alpha = 0.05$

Pvals, если проверяем одну из нескольких метрик

$$CR_1 = 0.021$$

$$CR_2 = 0.014$$

$$CR_3 = 0.045$$

Отклоняются все нулевые гипотезы при $\alpha = 0.05$ (если проверять любую метрику, не смотря на другие)

Pvals, если проверяем несколько метрик единовременно

$$CR_1 = 1 - (1 - 0.021)^3 = 0.061$$

$$CR_2 = 1 - (1 - 0.014)^3 = 0.041$$

$$CR_3 = 1 - (1 - 0.045)^3 = 0.129$$

Отклоняются нулевая гипотеза только у CR_2 при $\alpha = 0.05$ (т.к. проверяем ошибется ли тест, при одновременной проверке на других метриках)

FWER

Самый простой, но самый жесткий способ коррекции множественных сравнений – *Поправка Бонферонни*. Зная число тестов, можно вычислить скорректированный уровень значимости и использовать его

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

Например, чтобы сохранить в группе из 10 тестов вероятность ошибки I рода 0.05, нужно проводить каждый тест при $\alpha = 0.005$.

При этом резко возрастает вероятность не найти различий там, где они есть.

Также есть и другие поправки, например, метод Холма. О них можно почитать самостоятельно

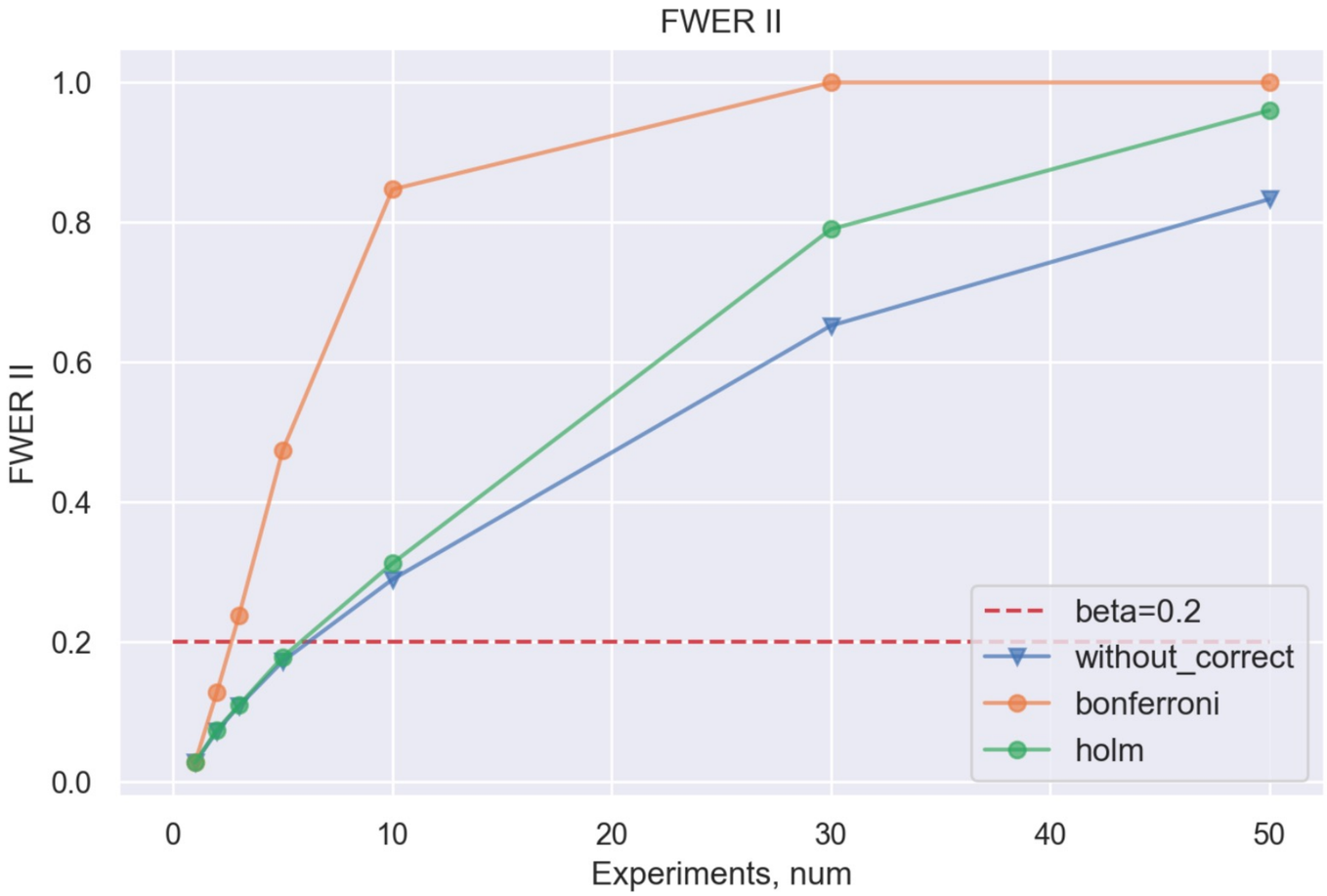
В чем проблема с методами, контролирующими FWER?

Проблемы FWER

Методы, контролирующие FWER, сильно понижают мощность

То есть сильно перестраховываясь в отношении ошибки 1 рода, мы пропускаем истинные эффекты, то есть уменьшаем мощность эксперимента с ростом числа гипотез

Если посмотреть на графике на групповую вероятность ошибки II рода, то видно, что у методов FWER она выше



Контроль групповой вероятности ошибки II рода методами Бонферрони / Холма

FDR

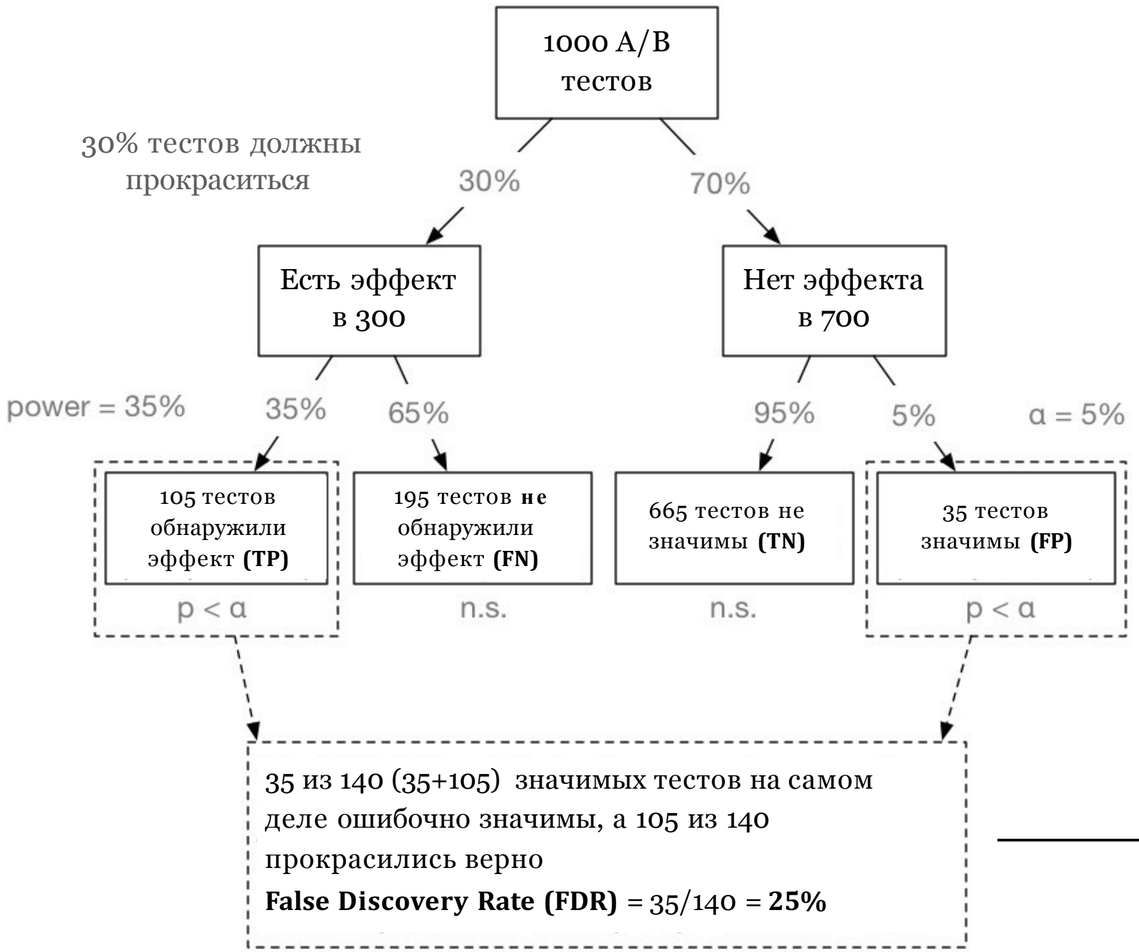
Для случаев, когда нам важнее сохранить истинно-положительные результаты, чем не допустить ложноположительных – используется контроль

False Discovery Rate = FP / (FP + TP)

С помощью FDR мы задаем не количество ошибок первого рода в принципе, а количество ложноположительных (fp) результатов в отношении к истинно- положительным (tp) и ложноположительным (fp) (далее это число обозначается как γ)

Для контроля FDR используется *поправка Бенджамини-Хохберга*

FDR



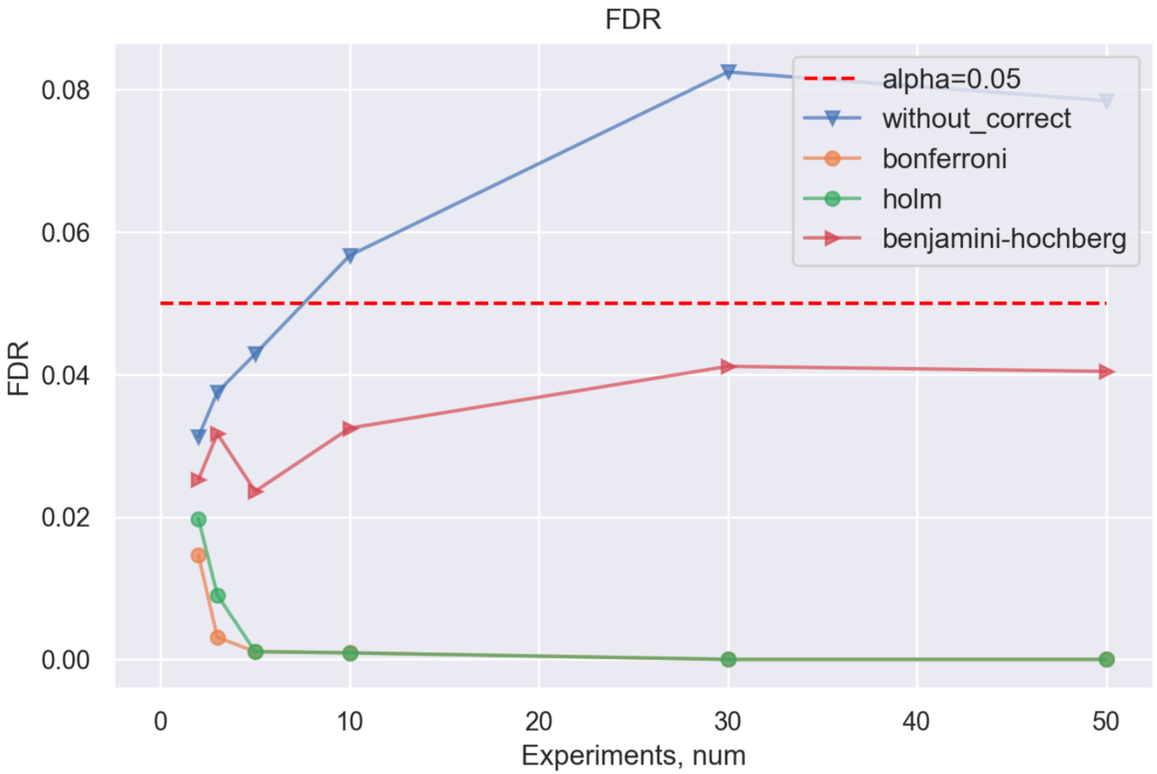
То есть мы «мягко» контролируем ошибку I рода (FP) и при этом учитываем ошибку II рода (TP). За счет этого методы FDR более мощные, но не так жестко контролируют ошибку I рода

Для контроля FDR используется поправка Бенджамини-Хохберга

FDR

Метод Холма и Бонферонни более жестко контролируют FDR за счет своей консервативной природы, но метод Бенджамини-Хохберга также контролирует FDR на заданном уровне.

Также он более мощный (на семинаре рассмотрим пример)



Контроль FDR в зависимости от числа гипотез в случае, когда половина гипотез с эффектом

FWER или FDR

- FDR (обычно FDR < 0.1): Выше мощность и контроль ложных срабатываний
- FWER (обычно FWER < 0.05): строгий контроль за вероятностью ошибок первого рода

Основной вывод

FDR для продуктовых реалий является более полезной метрикой для контроля, т.к. нам бы не хотелось выкатывать нерабочие изменения (когда мы говорим, что они рабочие), а наоборот, быть уверенными в реальных эффектах.

НО!

Если тестируется очень важный функционал, который влияет на ключевые метрики и хочется «перестраховаться», то надежней использовать FWER в силу его большей консервативности в отношении ошибки I рода

Самый главный совет по множественному тестированию

Если можно не проводить множественный тест, то лучше его не проводить, так как:

1) Сложная интерпретация и реализация

2) Растущий false positive rate