ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ
МГУ имени М. В. Ломоносова

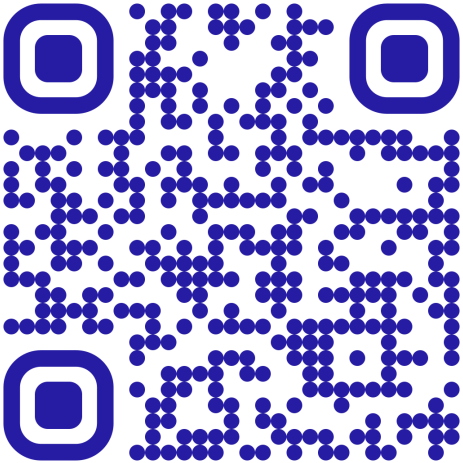
Эксперименты и проблема множественных сравнений

Ставнийчук Анна,
н.с. ЭФ МГУ
annastavnychuk@gmail.com

HSE R Meet Up

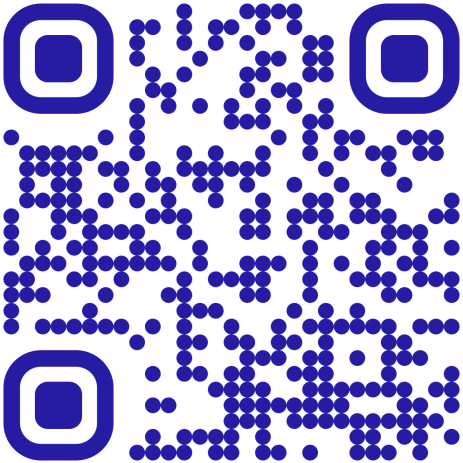
2 ноября 2024 г.

Ссылка на тг-канал



https://t.me/causal_channel

Ссылка на GitHub



https://github.com/annastavniychuk

Содержание

Эксперименты

- **Золотой стандарт оценки** – и по качеству, и про ресурсоемкости проведения
- Хорошая рандомизация позволят зафиксировать прочие равные
- Тот случай, когда важны 90% подготовительных усилий
- **Славятся внутренней валидностью и критикуются за внешнюю**

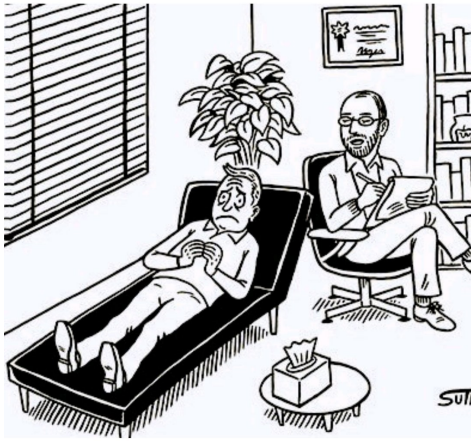
Эксперименты

- В 2019 году Нобелевскую премию по экономике получили Абхиджит Банерджи, Эстер Дюфло и Майкл Кремер за их подход к «снижению глобальной бедности»
- Нобелевскую премию 2021 года получили Дэвид Кард, Джошуа Ангрист и Гвидо Имбенс за вклад в методологию и использование **естественных экспериментов**

Они показали, как можно использовать естественные эксперименты — ситуации, в которых случайные факторы создают условия, аналогичные экспериментальным, — для выявления причинно-следственных связей

Фундаментальная проблема причинного вывода

- Чтобы оценить эффект воздействия для конкретного индивида, мы должны знать потенциальные исходы **сразу для двух его состояний мира**
- Реально мы наблюдаем только одно из них – либо, если индивид подвергся воздействию, либо, если он ему не подвергался
- **Оценка индивидуального эффекта требует доступа к данным, которых у нас физически не может быть!**
- Если с распределением индивидуального эффекта воздействия (treatment effect) работать не получается, будем довольствоваться **средними величинами**



And are the potential outcomes in the room with us now?

Средние эффекты

- **И средние, и индивидуальный эффект воздействия нельзя напрямую рассчитать, но мы будем пробовать их оценить**
- Самая простая идея для оценки ATE (average treatment effect), которая всем придет в голову, взять простую разницу в средних:

$$\mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=0]$$

- Но тут всё не так просто, после небольших преобразований мы получим следующее (доказательство тут):

$$\mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=0] =$$

$$= \underbrace{\mathbb{E}[Y_1] - \mathbb{E}[Y_0]}_{\text{ATE}} + \underbrace{\mathbb{E}[Y_0|T=1] - \mathbb{E}[Y_0|T=0]}_{\text{Selection Bias}} + \underbrace{(1-\pi)(ATT - ATnT)}_{\text{Heterogeneous treatment effect bias}}$$

Средние эффекты

$$\mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=0] =$$

$$= \underbrace{\mathbb{E}[Y_1] - \mathbb{E}[Y_0]}_{\text{ATE}} + \underbrace{\mathbb{E}[Y_0|T=1] - \mathbb{E}[Y_0|T=0]}_{\text{Selection Bias}} + \underbrace{(1-\pi)(ATT - ATnT)}_{\text{Heterogeneous treatment effect bias}}$$

- **ATE** – интересующий нас эффект
- **Selection Bias** – смещение, возникающее из-за того, что контрольная группа и группа воздействия различались, даже если бы на них не было оказано воздействие, то есть имеет место некоторый дисбаланс
- **Heterogeneous treatment effect bias** – различие в интенсивности эффекта для тритмент и контрольной группы, взвешенное на долю выборки $(1-\pi)$, которая попала в контрольную группу

Предпосылки

Чтобы оценка ATE была несмещенной, нам необходимо выполнение предпосылок:

1. **Экзогенность воздействия (Independence assumption)** – распределение объекта в тритмент или контрольную группы осуществляется случайно и независимо от его изначальных характеристик $(Y_1, Y_0, X)_i \perp T_i$
2. Отсутствие «внешних эффектов» воздействия (**SUTVA** – Stable unit treatment value assumption)
   1. воздействие оказывается только на один объект и внешние эффекты у него отсутствуют
   2. воздействие гомогенно – существует только один тип тритмента
   3. SUTVA невозможно проверить формальными статистическими тестами. ~~В неё можно только верить~~. Её выполнение зависит от грамотно продуманного дизайна эксперимента

Хорошая рандомизация, а следовательно, и выполнение предпосылок, позволяет нам очистить эффект воздействия от двух типов смещения:

$$ATE = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y_1|T=1] - \mathbb{E}[Y_0|T=0] \xrightarrow{p} \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{i1} - \frac{1}{N_0}\sum_{i=1}^{N_0} Y_{i0}$$

Содержание

Проверка гипотез

- Гипотеза – утверждение, которое мы хотим проверить на данных
- Проверить гипотезу – оценить, не противоречат ли ей данные
- Данные – это выборки, они случайные
- **Главная мысль: при любом объеме выборки можно совершить ошибку**

Проверка гипотез

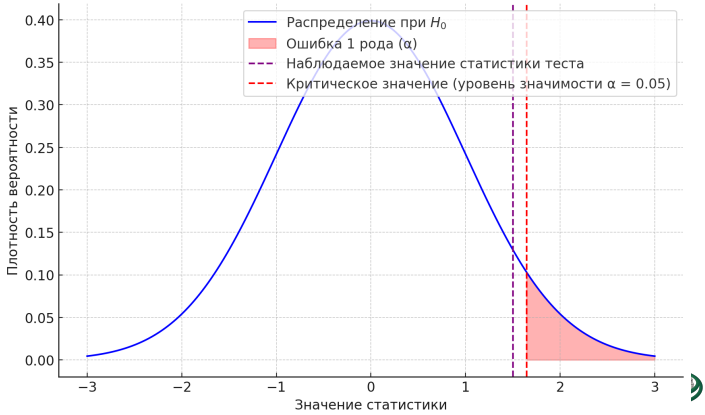
Нулевая гипотеза (H_0) — предположение, которое исследователь выдвигает в начале статистического теста, обычно предполагает отсутствие эффекта или связи между переменными

- Нулевая гипотеза предполагает отсутствие эффекта, влияния или различий. Она выступает как «статус-кво» или «условие по умолчанию»
- В статистике цель теста заключается в том, чтобы либо опровергнуть, либо не опровергнуть нулевую гипотезу, что позволяет судить о наличии значимых эффектов или различий в данных
  - Если данные подтверждают предположение нулевой гипотезы, **мы «не можем её отвергнуть»** (сказать, что мы ее доказали или приняли, будет некорректно, новая выборка может изменить ситуацию)
  - Если данные показывают значимое различие, то нулевая гипотеза **отвергается** в пользу альтернативной гипотезы H_1, которая предполагает наличие эффекта или связи

Проверка гипотез

Уровень значимости α = вероятность ошибочно отвергнуть H_0 = Ошибка 1 рода

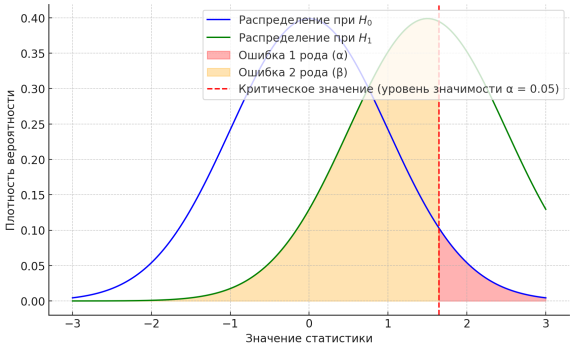
- Если мы отвергаем нулевую гипотезу, когда она верна, мы совершаем ошибку
- Выбирая уровень значимости, мы фиксируем вероятность совершить такую ошибку



Проверка гипотез

- Ошибка **первого рода**: отвергаем нулевую гипотезу H_0, хотя она на самом деле истинна
- Ошибка **второго рода**: отвергаем нулевую гипотезу H_0, хотя на самом деле истинна альтернативная гипотеза H_1.

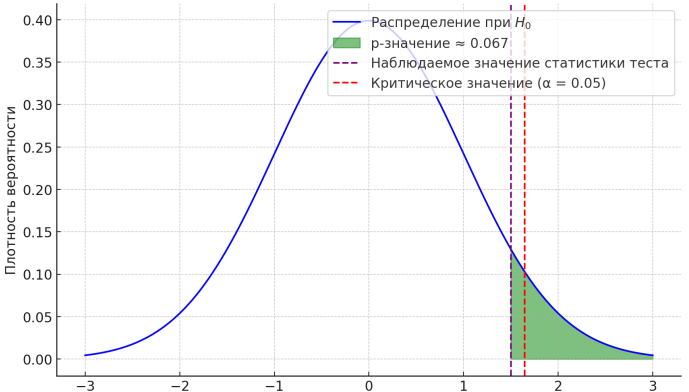
	H_0 не отвергается	H_0 отвергается
H_0 верна	+	Ошибка первого рода (α)
H_0 неверна	Ошибка второго рода (β)	+



Проверка гипотез

P-value – вероятность получить такое же значение наблюдаемой статистики, как в эксперименте (или более экстремального) при условии, что нулевая гипотеза верна

- p-value – уровень значимости, который нужно взять, чтобы гипотеза впервые отверглась
- Это мера, показывающая, насколько данные согласуются с нулевой гипотезой.



Содержание

Проблема множественного тестирования гипотез

- Исследовательский вопрос может быть таким, что вам интересно оценить воздействия разных типов тритмента, то есть у вас есть несколько экспериментальных групп и одна контрольная
- При такой постановке мы хотим проверить не одну, а сразу много статистических гипотез о различиях в группах.
- При проверке любой гипотезы существует вероятность совершить ошибку первого рода (отклонить нулевую гипотезу, если она верна = обнаружить эффект, которого нет)
- Особенность множественного тестирования гипотез состоит в том, что чем больше гипотез мы проверяем на одних и тех же данных, тем больше будет вероятность допустить как минимум одну ошибку первого рода – эффект множественных сравнений (multiple comparisons/testing)

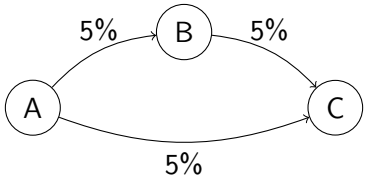
Проблема множественного тестирования гипотез

Источниками множественного тестирования могут быть:

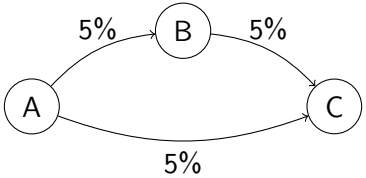
- Несколько типов воздействия (Multiple treatment arms)
- Гетерогенное воздействие (Heterogeneous treatment effects)
- Несколько способов оценки (Multiple estimators)
- Несколько зависимых переменных (Multiple outcomes), эффект на которые мы хотим оценить

Проблема множественного тестирования гипотез

- Предположим, что у нас есть 3 группы (A, B и C), в которых мы хотим сравнить среднее значение переменной интереса
- Будем использовать t-тест Стьюдента
  - Если мы получили достаточно большое значение t-статистики такое, что p-value < 0.05, то мы отклоняем нулевую гипотезу и заключаем, что группы статистически различаются по переменной интереса
  - Отсечка p-value < 0.05 значит, что вероятность ошибочного вывода о различии между групповыми средними не превышает 0.05
  - Это будет работать именно так, когда у нас всего две группы, но в случае множественного тестирования вероятность будет больше 5%



Проблема множественного тестирования гипотез



- Выполняя тест Стьюдента, исследователь проверяет нулевую гипотезу об отсутствии разницы между двумя группами.
- Сравнивая группы A и B, он может ошибиться с вероятностью 5%, B и C – 5%, A и C – тоже 5%.
- Соответственно, вероятность ошибиться хотя бы в одном из этих трех сравнений составит:

$$P = 1 - (1 - \alpha)^n = 1 - 0.95^3 \approx 0.14 > 0.05$$

- такая ошибка называется family-wise error rate.

Проблема множественного тестирования гипотез

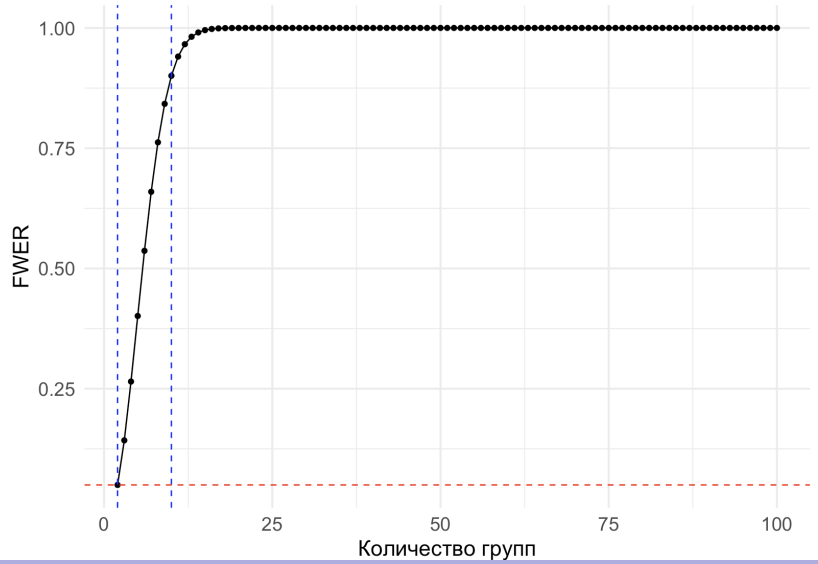
Если бы групп было бы 5:

$$FWER = 1 - (1 - \alpha)^n = 1 - 0.95^{10} \approx 0.4 > 0.05$$

Для нахождения числа групп (сочетаний) из k элементов в выборке из n элементов без повторений:

$$C(n, k) = \frac{n!}{k! \cdot (n - k)!}$$

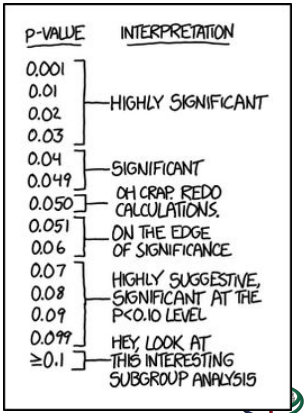
Проблема множественного тестирования гипотез



Проблема множественного тестирования гипотез

p-hacking — это практика манипуляции данными или проведением анализа с целью достижения статистически значимого результата

- **Множественное тестирование**: исследователь тестирует большое количество гипотез и выбирает только те, которые дали «значимый» результат
- **Изменение критериев**: исследователь может начать с анализа всей выборки, а затем, не получив значимого результата, выделить подгруппы или исключить выбросы
- **Игра с методами**: исследователь может экспериментировать с различными методами анализа, чтобы выбрать тот, который дает желаемый результат.
- **Применение различных ковариат**: добавление или удаление ковариат для изменения результата теста на значимость

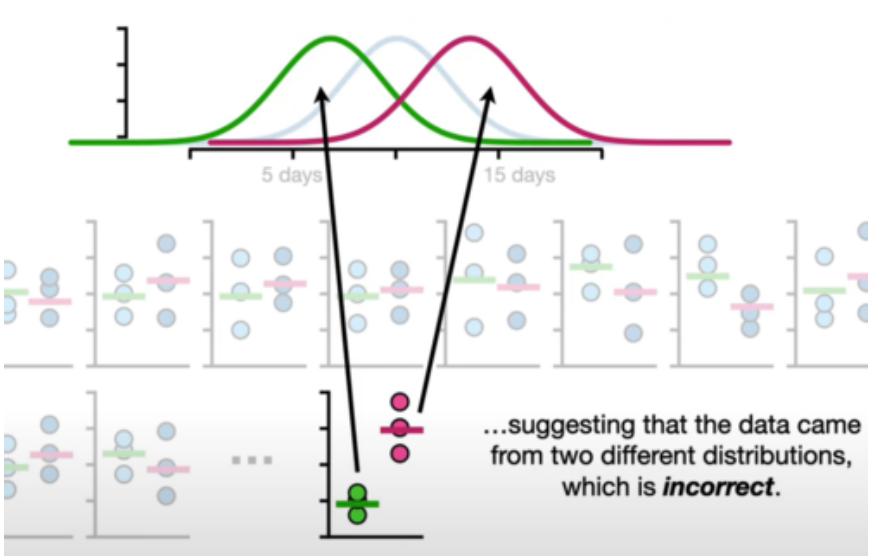


Проблема множественного тестирования гипотез

«If You Torture the Data Long Enough, It Will Confess» (Ronald H. Coase)

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Проблема множественного тестирования гипотез



Проблема множественного тестирования гипотез

Проблема множественного тестирования гипотез

К счастью, существует несколько методов, позволяющих преодолеть эту сложность:

- Корректировка p-value (p-value adjustments)
- Планирование эксперимента и фиксирование его условий (pre-analysis plans)
- Повтороное проведение эксперимента (replication)

Содержание

Контроль ошибок первого и второго рода

	Число принятых нулевых гипотез (p-value $> \alpha$) $\Rightarrow \hat{\tau} = 0$	**Число отвергнутых нулевых гипотез** (p-value $< \alpha$) $\Rightarrow \hat{\tau} \neq 0$	**Всего гипотез**
Число верных нулевых гипотез $\hat{\tau} = 0$	Число безошибочно принятых нулевых гипотез (TN, true negatives)	Число ошибочно отвергнутых нулевых гипотез (FP, false positives) – **ошибка первого рода**	m_0 – Число верных нулевых гипотез (true null hypotheses)
Число неверных нулевых гипотез $\hat{\tau} \neq 0$	Число ошибочно принятых нулевых гипотез (FN, false negatives) – ошибка второго рода	Число безошибочно отвергнутых нулевых гипотез (TP, true positives)	$m - m_0$ – Число истинных альтернативных гипотез (true alternative hypotheses)
Всего гипотез	$m - R$ – Общее число принятых гипотез	R – Общее число отвергнутых гипотез	m – всего гипотез

Групповая вероятность ошибки первого рода (family-wise error rate)

- При одновременной проверке семейства статистических гипотез мы хотим, чтобы количество наших ошибок (FP и FN) было минимальным
  - Традиционно исследователи пытаются **минимизировать величину ошибочно отвергнутых гипотез** FP
  - Это вполне логично, поскольку ложно отвергнутая нулевая гипотеза грозит нам ложноположительным найденным эффектом, которого реально может не быть
- **Если $FP \geq 1$, мы совершаем как минимум одну ошибку первого рода**
- Вероятность допущения такой ошибки при множественной проверке гипотез называют **групповой вероятностью ошибки** (familywise error rate, FWER или experiment-wise error rate). По определению,

$$FWER = P(FP \geq 1)$$

  – вероятность ошибочно отклонить хотя бы одну нулевую гипотезу во всех тестах
- Когда мы говорим, что хотим **контролировать групповую вероятность ошибки** на определенном уровне значимости α, мы подразумеваем, что должно выполняться неравенство $FWER \leq \alpha$.

Коррекция Бонферрони

Вернемся к нашему примеру, когда мы сравнили 3 группы A, B и C с помощью t-теста. Предположим, что мы получили следующие p-значения: 0.001, 0.01 и 0.04.

Как было сказано выше, мы хотим, чтобы групповая вероятность ошибки была не больше уровня значимости $FWER \leq \alpha$. Согласно методу Бонферрони, мы должны сравнить каждое из полученных p-значений не с α, а с $\frac{\alpha}{n}$, где n — число проверяемых гипотез.

Деление исходного уровня значимости α на n — это и есть поправка Бонферрони. В рассматриваемом примере каждое из полученных p-значений необходимо было бы сравнить с $\frac{0.05}{3} \approx 0.017$

- $p\text{-}value_1 = 0.001 < \alpha_{\text{adjusted}} = 0.017$ — гипотеза отклонена
- $p\text{-}value_2 = 0.01 < \alpha_{\text{adjusted}} = 0.017$ — гипотеза отклонена
- $p\text{-}value_3 = 0.04 > \alpha_{\text{adjusted}} = 0.017$ — гипотеза принята

Коррекция Бонферрони

Вместо деления уровня значимости на число гипотез, мы могли бы умножить каждое p-значение на это число и получить точно такие же выводы (эта эквивалентная процедура реализована в R):

- $p\text{-}value_{1,\text{adjusted}} = 0.001 \cdot 3 = 0.003 < \alpha = 0.05$ — гипотеза отклонена
- $p\text{-}value_{2,\text{adjusted}} = 0.01 \cdot 3 = 0.03 < \alpha = 0.05$ — гипотеза отклонена
- $p\text{-}value_{3,\text{adjusted}} = 0.04 \cdot 3 = 0.12 > \alpha = 0.05$ — гипотеза принята

Иногда при домножении p-значений результат может получиться больше единицы. Из теории вероятностей мы знаем, что вероятность не может быть больше одного, поэтому в таких случаях p-значение принимают равным за единицу

Коррекция Бонферрони

Пусть у нас есть m независимых гипотез, которые мы тестируем с заданным уровнем значимости α

1. **Цель:** Мы хотим контролировать вероятность хотя бы одной ошибки первого рода среди m тестов на уровне α.
2. **Решение:** В методе Бонферрони корректируем уровень значимости для каждого отдельного теста до $\alpha' = \frac{\alpha}{m}$. Это означает, что каждый отдельный тест теперь имеет уровень значимости α', а не α.

Коррекция Бонферрони

Доказательство:

- Вероятность того, что хотя бы один из m тестов даст ложноположительный результат, если нулевая гипотеза верна для всех тестов, ограничена сверху суммой вероятностей ложноположительных результатов для каждого теста (неравенство Бонферрони):

$$P(\text{хотя бы один ложноположительный результат}) \leq$$

$$\sum_{i=1}^{m} P(\text{ложноположительный результат для теста}_i) = m \cdot \alpha' = m \cdot \frac{\alpha}{m} = \alpha$$

- Поскольку каждый тест теперь имеет уровень значимости α', вероятность ложноположительного результата для одного теста составляет α'.
- Таким образом, метод Бонферрони гарантирует, что вероятность ошибки первого рода среди всех m тестов не превышает α

Реализация в R

- **Base R:** В базовом R есть функция p.adjust, которая поддерживает большинство методов коррекции
- **stats:** Пакет stats, встроенный в R, также содержит функцию p.adjust (поскольку она является частью base R) и ряд других тестов
- **rstatix:** более новая и навороченная версия, содержит кроме прочего ряд других интересных тестов

Коррекция Бонферрони

```
p.adjust(c(0.001, 0.01, 0.04), method = "bonferroni")
```

```
[1] 0.003 0.030 0.120
```

Можно на выходе сразу получить выводы относительно гипотез при $\alpha = 5$:

```
alpha <- 0.05
p.adjust(c(0.001, 0.01, 0.04), method = "bonferroni") < alpha # отклоняем H_0 (есть
```

```
[1]  TRUE  TRUE FALSE
```

Коррекция Бонферрони

- Важно помнить об **уязвимости коррекции Бонферрони** – с ростом числа гипотез **мощность метода уменьшается**
- Чем больше гипотез мы хотим проверить, тем сложнее нам будет их отвергать (даже если они реально должны быть отвергнуты)
- Например, для 5 групп (10 гипотез), применение поправки Бонферрони привело бы к снижению исходного уровня значимости до $0.01/10 = 0.001$
- Соответственно, для отклонения гипотез, соответствующие p-значения должны быть меньше 0.001, а это довольно жесткая отсечка
- Из этого делаем вывод, что при большом числе гипотез коррекцию Бонферрони лучше не использовать

Нисходящая процедура Хольма

Метод Хольма позволяет побороть недостатки метода Бонферрони

- Сначала p-значения сортируются по возрастанию $p\text{-}value_1 \leq p\text{-}value_2 \leq \cdots \leq p\text{-}value_n$.
- Затем проверяется условие для первого из p-значений: $p\text{-}value_1 \geq \frac{\alpha}{n-i+1} = \frac{\alpha}{n}$,
  - если условие выполнено, то все нулевые гипотезы принимаются, и процедура останавливается, иначе первая из гипотез отвергается, и начинается следующий шаг
- На следующем шаге проверяется условие $p\text{-}value_2 \geq \frac{\alpha}{n-i+1} = \frac{\alpha}{n-1}$,
  - если условие выполнено, то все гипотезы, начиная со второй, принимаются, иначе первые две гипотезы отклоняются и начинается следующий шаг
- На последнем шаге проверяется условие вида $p\text{-}value_n \geq \frac{\alpha}{n-n+1}$,
  - если оно выполнено, то последняя гипотеза принимается, если нет — отклоняется, на этом процедура заканчивается

Нисходящая процедура Хольма

Метод Хольма называют нисходящей (step-down) процедурой. Он начинается с наименьшего p-значения в упорядоченном ряду и последовательно "спускается" вниз к более высоким значениям. На каждом шаге соответствующее значение $p\text{-}value_i$ сравнивается со скорректированным уровнем значимости

$$\alpha_{\text{adjusted}} = \frac{\alpha}{n + i - 1}.$$

Аналогично коррекции Бонферрони можно вместо корректировки уровня значимости корректировать p-значения (эта эквивалентная процедура реализована в R)

$$p\text{-}value_{i,\text{adjusted}} = p\text{-}value_i \cdot (n - i + 1)$$

Возвращаясь к нашему примеру:

- $p\text{-}value_{1,\text{adjusted}} = 0.001 \cdot (3 - 1 + 1) = 0.003 < \alpha = 0.01$ — гипотеза отклонена
- $p\text{-}value_{2,\text{adjusted}} = 0.01 \cdot (3 - 2 + 1) = 0.02 > \alpha = 0.01$ — гипотеза принята
- $p\text{-}value_{3,\text{adjusted}} = 0.04 \cdot (3 - 3 + 1) = 0.04 > \alpha = 0.01$ — гипотеза принята

Нисходящая процедура Хольма

```
p.adjust(c(0.001, 0.01, 0.04), method = "holm")
```

```
[1] 0.003 0.020 0.040
```

И результаты проверки гипотез при $\alpha = 5$:

```
alpha <- 0.05
p.adjust(c(0.001, 0.01, 0.04), method = "holm") < alpha # отклоняем H_0 (есть эффе
```

```
[1] TRUE TRUE TRUE
```

Средняя доля ложных отклонений (false discovery rate)

- FWER методы обеспечивают контроль над групповой вероятностью ошибки первого рода
- Эти методы чересчур жестко работают, когда нужно одновременно проверить слишком много гипотез (падает статистическая мощность)
  - Под «недостаточной мощностью» понимается сохранение многих нулевых гипотез, которые потенциально могут представлять исследовательский интерес и которые, соответственно, следовало бы отклонить – съедаем эффект, который потенциально есть
- Недостаточная мощность традиционных процедур множественной проверки гипотез привела к разработке новых методов, например, метода Бенджамини-Хохберга

Средняя доля ложных отклонений (false discovery rate)

- Для преодоления недостаточной мощности FWER методов был предложен новый подход к проблеме множественных проверок статистических гипотез
- Суть подхода заключается в том, что вместо контроля над групповой вероятностью ошибки первого рода выполняется **контроль над ожидаемой долей ложных отклонений (false discovery rate, FDR)** среди всех отклоненных гипотез
- В терминах таблицы выше эта ожидаемая доля может быть записана следующим образом:

$$FDR = \left(\frac{FP}{R}\right) \quad \text{(считают, что если } R = 0, \text{ то } FDR = 0)$$

- Часто можно встретить запись через мат. ожидание

$$FDR = \mathbb{E}\left(\frac{FP}{R}\right).$$

FDR – ожидаемая доля ложных отклонений среди всех отклоненных гипотез

Средняя доля ложных отклонений (false discovery rate)

- В отличие от уровня значимости α, каких-либо общепринятых значений FDR не существует
- Многие исследователи по аналогии контролируют FDR на уровне 5%
- Интерпретация порогового значения FDR очень проста: например, если в ходе анализа данных отклонено 1000 гипотез, то при $q = 0.10$ ожидаемая доля ложно отклоненных гипотез не превысит 100

Восходящая процедура Бенджамини — Хохберга

В статье (Benjamini, Hochberg, 1995) описание процедуры контроля над FDR выглядит так:

- Сначала p-значения сортируются по возрастанию $p\text{-}value_1 \leq p\text{-}value_2 \leq \cdots \leq p\text{-}value_n$.
- Находят максимальное значение k среди всех индексов $i = 1, \ldots, n$, для которого $p\text{-}value_i \leq \frac{i}{n} q$ выполняется неравенство.
- Отклоняют все гипотезы H_i с индексами $i = 1, \ldots, k$.

Восходящая процедура Бенджамини — Хохберга

В качестве примера рассмотрим следующий ряд из 15 упорядоченных по возрастанию p-значений (из оригинальной статьи Benjamini and Hochberg 1995):

```
p.adjust(c(0.0001, 0.0004, 0.0019, 0.0095,  0.0201, 0.0278, 0.0298, 0.0344, 0.0459
```

```
 [1] 0.00150000 0.00300000 0.00950000 0.03562500 0.06030000 0.06385714
 [7] 0.06385714 0.06450000 0.07650000 0.48600000 0.58118182 0.71487500
[13] 0.75323077 0.81321429 1.00000000
```

И результаты проверки гипотез при $\alpha = 5$:

```
alpha <- 0.05
p.adjust(c(0.0001, 0.0004, 0.0019, 0.0095,  0.0201, 0.0278, 0.0298, 0.0344, 0.0459
```

```
 [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE
```

Эквивалентная процедура, реализованная в R, отличается тем, что вместо нахождения максимального индекса k, исходные p-значения корректируются следующим образом:

$$q_i = \frac{p_i n}{i}.$$

Восходящая процедура Бенджамини — Хохберга

Интерпретация этих p-значений с поправкой (в большинстве литературных источников их называют q-значениями) такова:

- Допустим, что мы хотим контролировать долю ложно отклоненных гипотез на уровне FDR = 0.05
- Все гипотезы, q-значения которых $q\text{-}value \leq 0.05$, отклоняются
- Среди всех этих отклоненных гипотез доля отклоненных по ошибке не превышает 5%

Коррекция p-значений по методу Бенджамини-Хохберга работает особенно хорошо в ситуациях, когда необходимо принять общее решение по какому-либо вопросу при наличии информации (=проверенных гипотез) по многим параметрам

Следует помнить, что описанный здесь метод контроля над ожидаемой долей ложных отклонений предполагает, что все **тесты**, при помощи которых получают p-значения, **независимы**. На практике в большинстве случаев это условие выполняться не будет, но есть хорошая новость...

Восходящая процедура Бенджамини-Йекутили

Для преодоления ограничения независимости тестов при проверке гипотез в работе (Benjamini and Yekutieli 2001) был предложен усовершенствованный метод, учитывающий наличие корреляции между проверяемыми гипотезами.
Процедура Бенджамини-Йекутиили очень похожа на процедуру Бенджамини-Хохберга. Основное отличие заключается во введении поправочной константы

$$c_n = \sum_{i=1}^{n} \frac{1}{i},$$

далее аналогично:

- Сначала p-значения сортируются по возрастанию $p\text{-}value_1 \leq p\text{-}value_2 \leq \cdots \leq p\text{-}value_n$.
- Находят максимальное значение k среди всех индексов $i = 1, \ldots, n$, для которого $p\text{-}value_i \leq \frac{i}{n}\frac{q}{c_n}$.
- Отклоняют все гипотезы H_i с индексами $i = 1, \ldots, k$.

Восходящая процедура Бенджамини-Йекутили

```
p.adjust(c(0.0001, 0.0004, 0.0019, 0.0095,  0.0201, 0.0278, 0.0298, 0.0344, 0.0459
```

```
 [1] 0.004977343 0.009954687 0.031523175 0.118211908 0.200089208 0.211892623
 [7] 0.211892623 0.214025770 0.253844518 1.000000000 1.000000000 1.000000000
[13] 1.000000000 1.000000000 1.000000000
```

И результаты проверки гипотез при $\alpha = 5$:

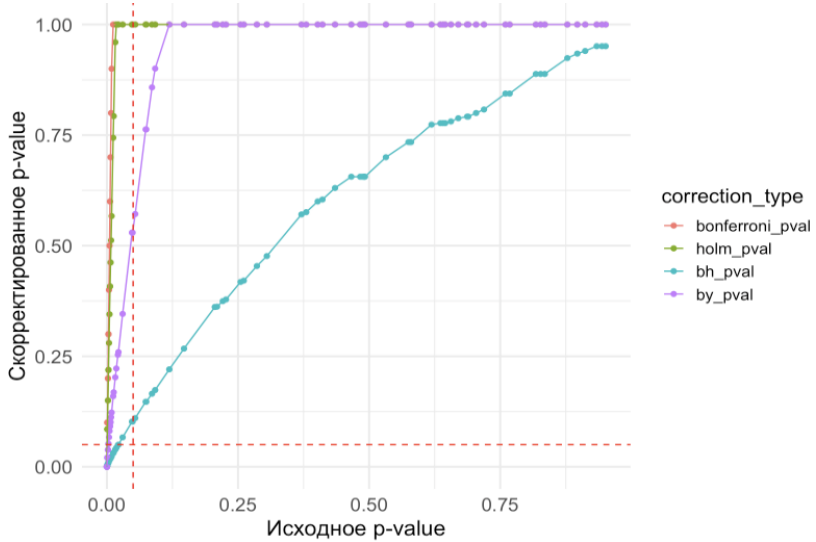
```
alpha <- 0.05
p.adjust(c(0.0001, 0.0004, 0.0019, 0.0095,  0.0201, 0.0278, 0.0298, 0.0344, 0.0459
```

```
 [1]  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[13] FALSE FALSE FALSE
```

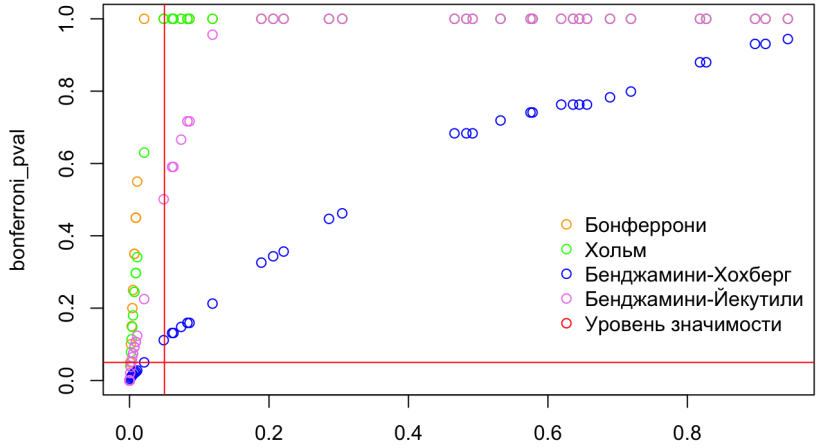
Эквивалентная процедура, реализованная в R, отличается тем, что вместо нахождения максимального индекса k, исходные p-значения корректируются следующим образом:

$$q_i = \frac{p_i \cdot n \cdot c_n}{i}.$$

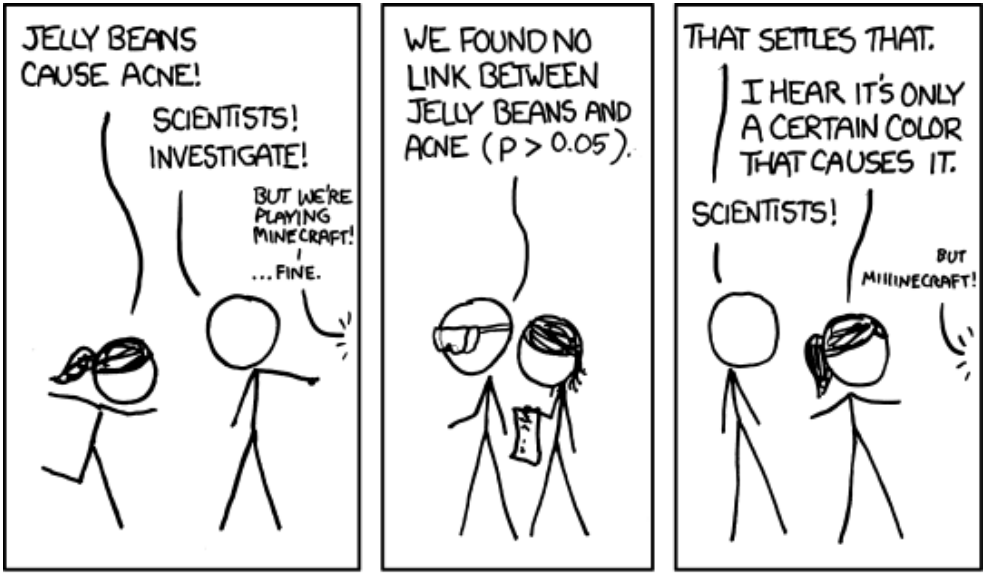
Сравнение коррекций

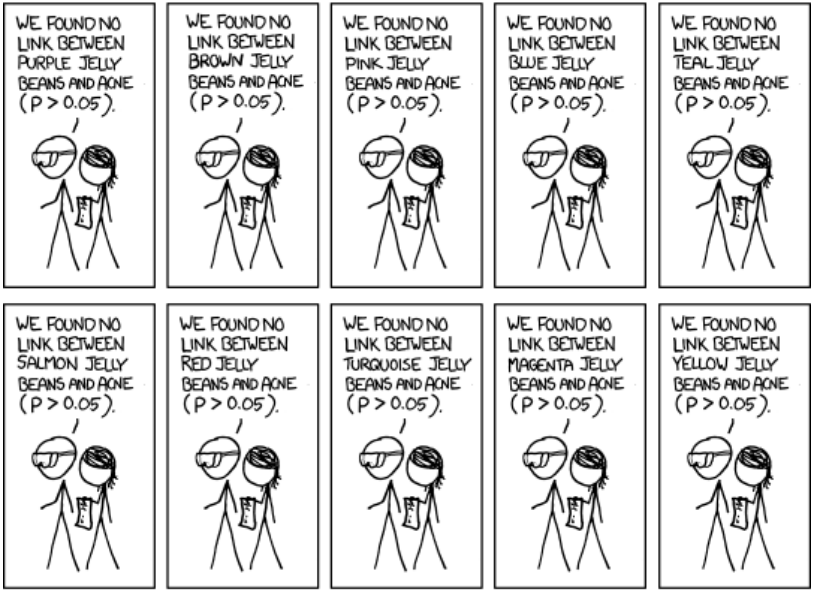


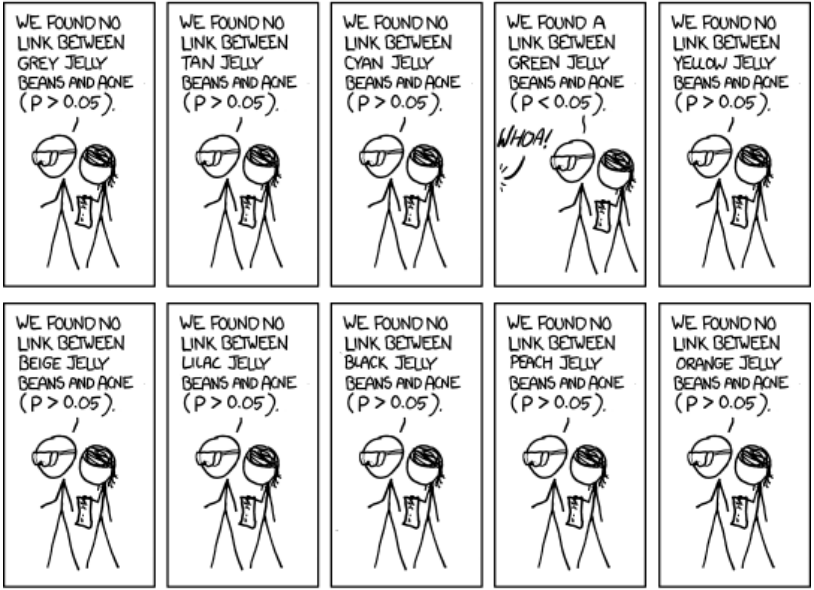
Сравнение коррекций

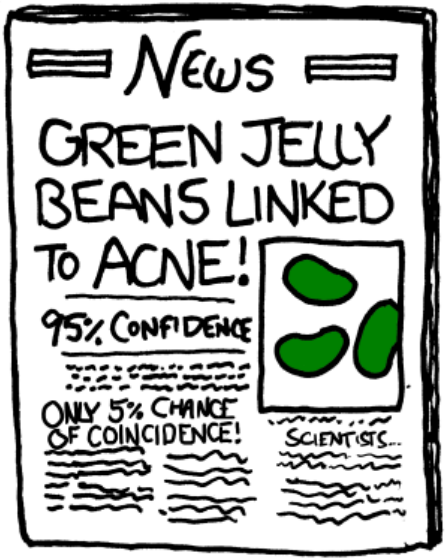


Содержание


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Пример

Предположим, что мы хотим изучить влияние употребления желейных бобов на акне. Также предположим, что мы располагаем информацией о регулярности употребления желейных бобов испытуемыми и качестве их кожи (наличии акне).
Пусть

- 90% испытуемых регулярно едят желейные бобы, то есть тритмент распределен как $eating \sim Bern(0, 9)$;
- состояние акне участника рапсределено равномерно $acne\_condition \sim U(0, 1)$

1. Оцените эффект от употребления желейных бобов на качество кожи
2. Проверьте гипотезу для разных цветов
3. Сформулируйте выводы из анализа