12 МАРТА 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ

ЕВГЕНИЙ БАКИН | МИФ №1:
СТАТИСТИКА – НАУКА ТОЧНАЯ, И В НЕЙ НЕТ МЕСТА МИФАМ

HTTPS://T.ME/CHAT_BIOSTAT_R

Евгений Бакин

к.т.н., руководитель трека
по биостатистике
в Институте биоинформатики

https://bioinf.me/
https://t.me/chat_biostat_R

Предыстория

1,500 scientists lift the lid on reproducibility

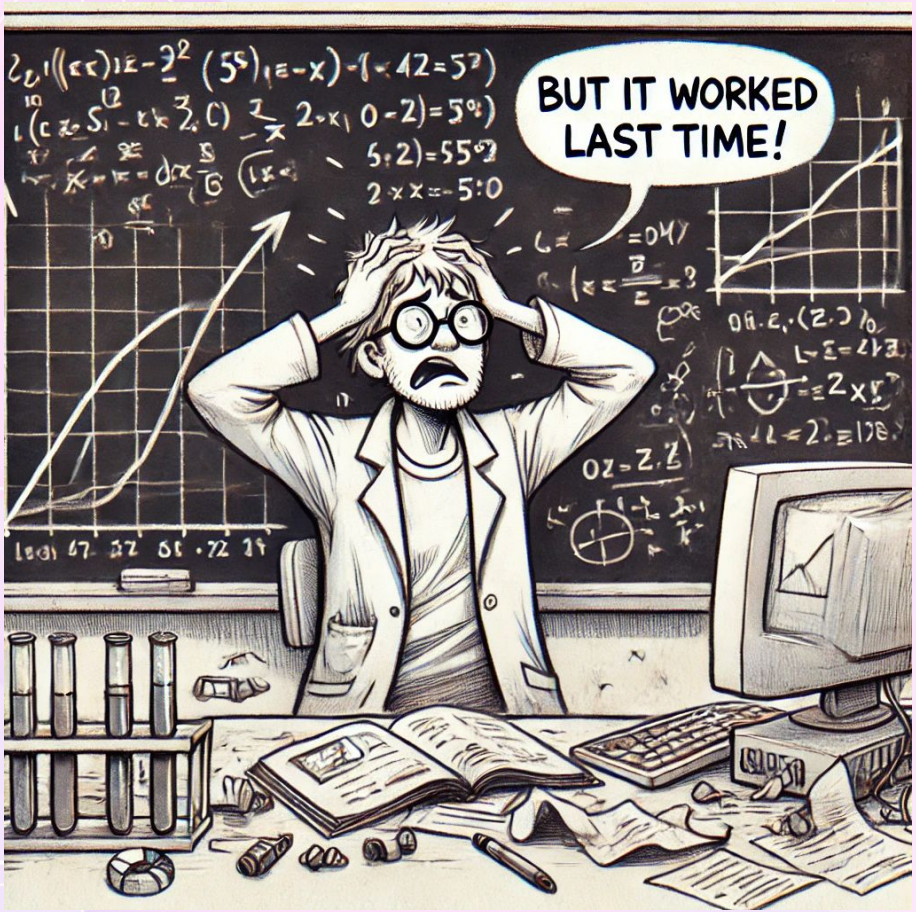
Monya Baker

Nature **533**, 452–454 (2016) | Cite this article

229k Accesses | **2424** Citations | **5159** Altmetric | Metrics

https://www.nature.com/articles/533452a

“**More than 70%** of researchers have tried and failed to reproduce another scientist's experiments, and **more than half** have failed to reproduce their own experiments.”



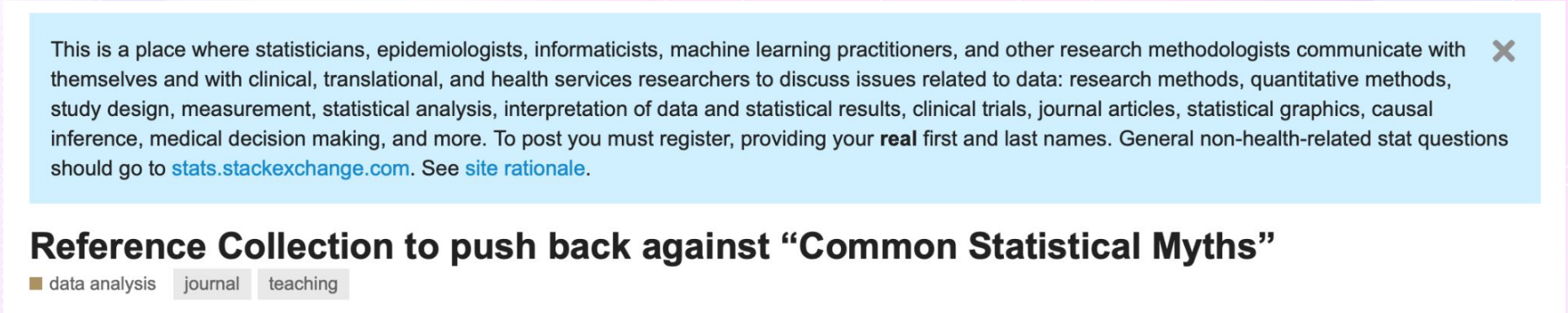
Причины кризиса

1. Прямая фальсификация данных.
2. Умышленное применение сомнительных способов «накрутки» статистической значимости (p-hacking, HARKing).
3. **Непреднамеренные ошибки, вызванные недопониманием сути статистических методов.**

Непреднамеренные ошибки

- **Дизайн исследования:**
  - Как цель исследования влияет на выбор схемы исследования?
  - Что (не)дает расчет выборки?
- **Обработка результатов исследования:**
  - Как выбрать “правильный” статистический метод?
  - Какая мера ассоциации мне подходит?
  - Все ли способы трансформации данных одинаково полезны?
  - В РКИ с выбором стат. методов все просто?
- **Интерпретация результатов:**
  - Что такое значимость, доверительный интервал, точечная оценка?
  - Как понять, почему эксперимент не удался?
- **Публикация результатов исследования:**
  - Какие результаты достойны обнародования?
  - Как надо описывать результаты анализа?
  - Как сравнивать свои результаты с известными?

Стоя на плечах гигантов



https://discourse.datamethods.org/t/reference-collection-to-push-back-against-common-statistical-myths/1787

Дисклеймер

- Мы сами много и увлеченно ошибались, чему, наверняка, в Интернете можно найти свидетельства :)
- В чем-то мы ошибаемся и сейчас (но пока этого не поняли).
- Какие-то ошибки нас ещё ждут.

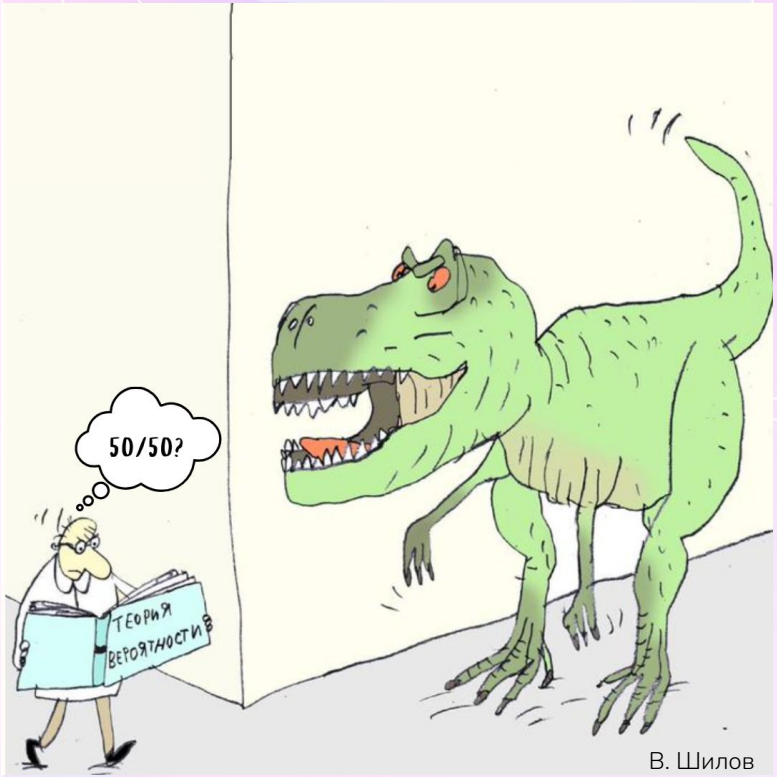
Ошибаться – не страшно. Главное – не упорствовать в заблуждениях и делиться своим опытом с комьюнити.

Надеемся на плодотворное обсуждение!

➢ https://t.me/chat_biostat_R

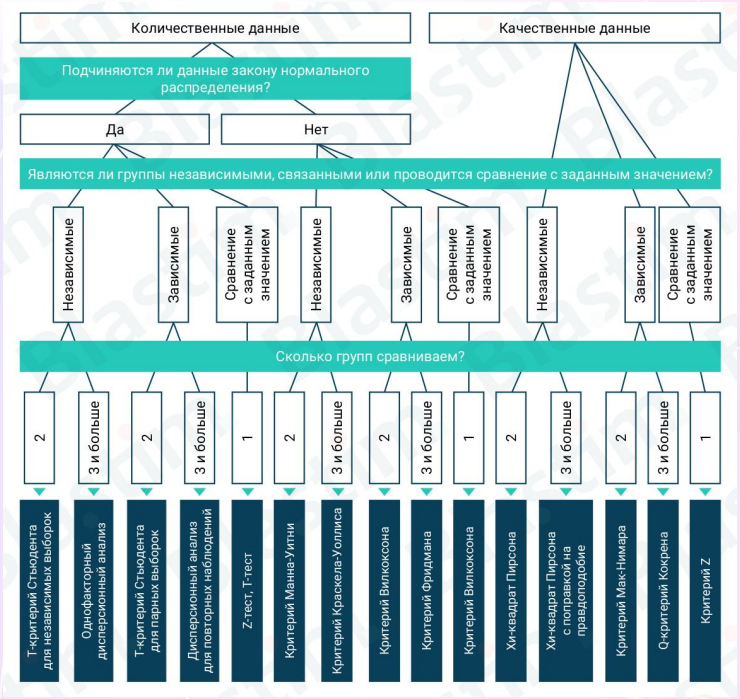
Откуда берутся ошибки?

- Резкое увеличение числа биомедицинских исследований
- Недостаток профессиональных биостатистиков
- Необходимость специалистам других специальностей (врачам, биологам) самостоятельно проводить статистический анализ
- Серьезное освоение методов статистики требует существенных трудозатрат, на которые, как правило, нет времени и сил.
- **Приобретают популярность гайдлайны и образовательные курсы, на которых применение методов статистики сводится к использовании простых схем.**
- **Доступность статистических программ.**

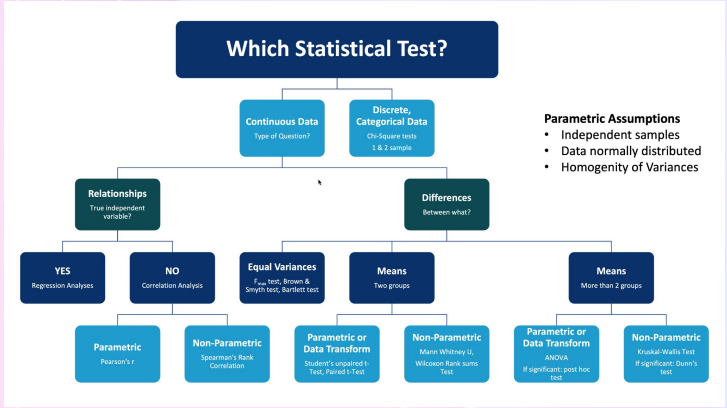


Гайдлайны

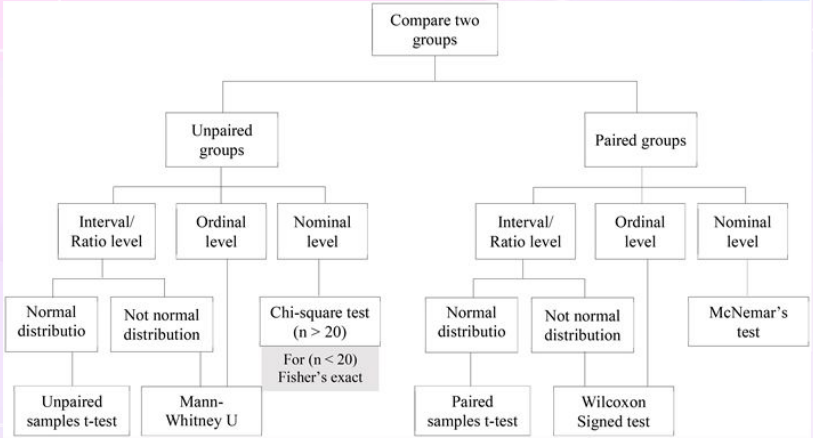
1



2



3



1. Гайд по выбору статистического метода. Blastim.
2. Multidisciplinary Research Methods for Engineers. TU Delft OpenCourseWare. https://ocw.tudelft.nl/course-readings/4-2-5-selecting-a-statistical-method/
3. A Comprehensive Guide for Selecting Appropriate Statistical Tests: Understanding When to Use Parametric and Nonparametric Tests. DOI: 10.4236/ojs.2023.134023

Ложное ощущение, что статистика – это просто.

Появился специфический *правдоподобный* язык, на котором все объяснения звучат легко и «понятно».

Правда или вымысел?

1. Для использования теста X необходимо проверить его допущения (“требования”), несоблюдение которых приведет к неверным результатам.
2. Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.
3. В силу малого объема выборки, рекомендуется использовать непараметрические тесты.
4. Чем меньше p-значение, тем достовернее результаты.
5. Т.к. $p < 0.05$, полученные результаты – достоверны.
6. Т.к. $p = 0.07$, полученные результаты – недостоверны, но имеют тенденцию к значимости.
7. Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.
8. Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.
9. С вероятностью 95% значения показателя попадают в доверительный интервал.
10. Перед началом анализа удалим из выборки имеющиеся выбросы, обнаруженные с помощью боксплота.
11. Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.
12. В регрессионную модель необходимо включать только значимые признаки.
13. Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.

Неверная интерпретация базовых терминов!

1. Для использования теста X необходимо проверить его **допущения** (“требования”), несоблюдение которых приведет к неверным результатам.
2. Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.
3. В силу малого объема выборки, рекомендуется использовать непараметрические тесты.
4. Чем меньше p-значение, тем достовернее результаты.
5. Т.к. $p < 0.05$, полученные результаты – достоверны.
6. Т.к. $p = 0.07$, полученные результаты – недостоверны, но имеют тенденцию к **значимости**.
7. Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.
8. Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.
9. С вероятностью 95% значения показателя попадают в **доверительный интервал**.
10. Перед началом анализа удалим из выборки имеющиеся выбросы, обнаруженные с помощью боксплота.
11. Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.
12. В регрессионную модель необходимо включать только значимые признаки.
13. Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.

Использование терминов, не имеющих принятого определения!

1. Для использования теста X необходимо проверить его допущения (“требования”), несоблюдение которых приведет к неверным результатам.
2. Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.
3. В силу малого объема выборки, рекомендуется использовать непараметрические тесты.
4. Чем меньше p-значение, тем **достовернее** результаты.
5. Т.к. $p < 0.05$, полученные результаты – **достоверны**.
6. Т.к. $p = 0.07$, полученные результаты – **недостоверны**, но имеют **тенденцию** к значимости.
7. Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.
8. Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.
9. С вероятностью 95% значения показателя попадают в доверительный интервал.
10. Перед началом анализа удалим из выборки имеющиеся **выбросы**, обнаруженные с помощью боксплота.
11. Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.
12. В регрессионную модель необходимо включать только значимые признаки.
13. Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.

Наделение методов дополнительными функциями

1. Для использования теста X необходимо проверить его допущения (“требования”), несоблюдение которых приведет к неверным результатам.
2. Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.
3. В силу малого объема выборки, рекомендуется использовать непараметрические тесты.
4. Чем меньше p-значение, тем достовернее результаты.
5. Т.к. $p < 0.05$, полученные результаты – достоверны.
6. Т.к. $p = 0.07$, полученные результаты – недостоверны, но имеют тенденцию к значимости.
7. **Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.***
8. **Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.**
9. С вероятностью 95% значения показателя попадают в доверительный интервал.
10. Перед началом анализа удалим из выборки имеющиеся выбросы, обнаруженные с помощью боксплота.
11. Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.
12. **В регрессионную модель необходимо включать только значимые признаки.**
13. Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.

Непонимание границ применения методов

1. Для использования теста X необходимо проверить его допущения (“требования”), несоблюдение которых приведет к неверным результатам.
2. **Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.**
3. **В силу малого объема выборки, рекомендуется использовать непараметрические тесты.**
4. Чем меньше p-значение, тем достовернее результаты.
5. Т.к. $p < 0.05$, полученные результаты – достоверны.
6. Т.к. $p = 0.07$, полученные результаты – недостоверны, но имеют тенденцию к значимости.
7. Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.*
8. Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.
9. С вероятностью 95% значения показателя попадают в доверительный интервал.
10. Перед началом анализа удалим из выборки имеющиеся выбросы, обнаруженные с помощью боксплота.
11. **Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.**
12. В регрессионную модель необходимо включать только значимые признаки.
13. Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.

Упование на «черные ящики»

1. Для использования теста X необходимо проверить его допущения (“требования”), несоблюдение которых приведет к неверным результатам.
2. Если данные имеют распределение, отличное от нормального, необходимо использовать непараметрические тесты, созданные именно для такого случая.
3. В силу малого объема выборки, рекомендуется использовать непараметрические тесты.
4. Чем меньше p-значение, тем достовернее результаты.
5. Т.к. $p < 0.05$, полученные результаты – достоверны.
6. Т.к. $p = 0.07$, полученные результаты – недостоверны, но имеют тенденцию к значимости.
7. Т.к. $p > 0.05$, мы принимаем нулевую гипотезу.*
8. Полученное $p < 0.0001$ свидетельствует о высокой эффективности нового препарата.
9. С вероятностью 95% значения показателя попадают в доверительный интервал.
10. Перед началом анализа удалим из выборки имеющиеся выбросы, обнаруженные с помощью боксплота.
11. Для корректного сравнения, группы должны быть сопоставимы и все признаки должны незначимо отличаться.
12. В регрессионную модель необходимо включать только значимые признаки.
13. **Для отбора признаков в регрессионную модель целесообразно использовать эффективные автоматические методы.**

Запомним:

1. Неверная интерпретация базовых терминов.
2. Использование терминов, не имеющих принятого определения.
3. Наделение методов дополнительными функциями.
4. Непонимание границ применения методов.
5. Упование на «черные ящики».

\+ Смешение корректных и некорректных утверждений.

Нулевая гипотеза

Строго сформулированное утверждение **о генеральной совокупности.**

НЕ ВЫБОРКЕ!

Например:

1. В **генеральной совокупности** средний рост мужчин равен среднему росту женщин.
2. В **генеральной совокупности** у пациентов, принимающих препарат А, 5-летняя выживаемость такая же, как и у пациентов, принимающих препарат Б
3. В **генеральной совокупности** вероятность рождения мальчика – 45%.

Статистический тест

Алгоритм, оценивающий степень совместимости полученной выборки с нулевой гипотезой.

Степень совместимости может оцениваться разными способами:

- статистика критерия
- *p*-значение
- отношение правдоподобия
- Байесовский фактор

Допущения статистического теста

Набор условий, соблюдение которых гарантирует то, что исследователь сможет контролировать качество работы теста (например, ошибку 1 рода).

Обратное – НЕ ВЕРНО.

То есть из того, что условия нарушаются, **не следуе**т, что качество работы теста обязательно будет неприемлемым.

Первично

Функционал теста (соответствует ли его нулевая гипотеза клиническому вопросу)

Вторично

Качество выполнения функционала

На чем вы поедете на работу?

Еще несколько комментариев

1. Статистический тест делает именно то, что он делает: проверяет нулевую гипотезу с помощью выбранной метрики (*p*-значение, тестовая статистика и пр.)
2. Эта метрика **не должна** решать другие задачи.
3. Например:

H0: В **генеральной совокупности** средний Hb одинаков в группах T и R.

p-значение не будет отражать то, на сколько сильно это различие клинически.

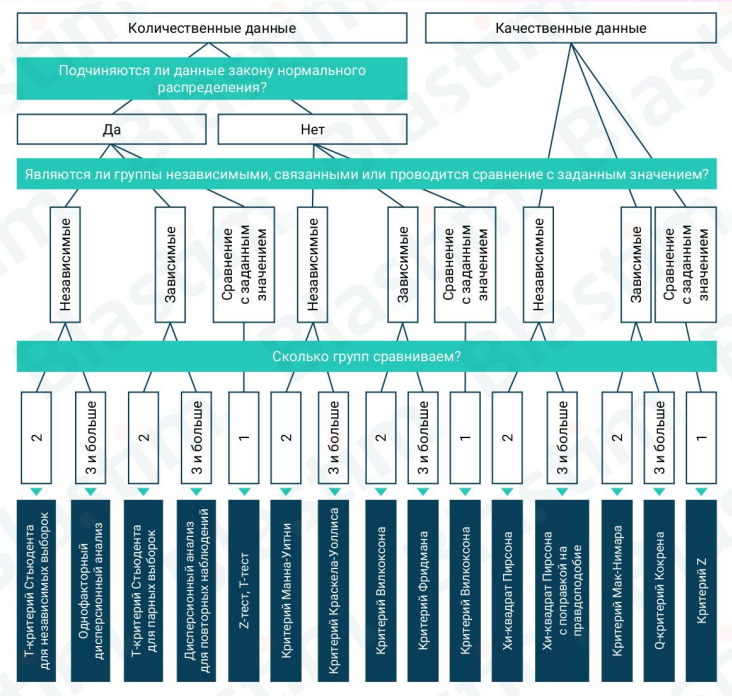
Для этого нужны другие инструменты, например, доверительный интервал.

Дальнейшее обсуждение – на встрече 19 марта

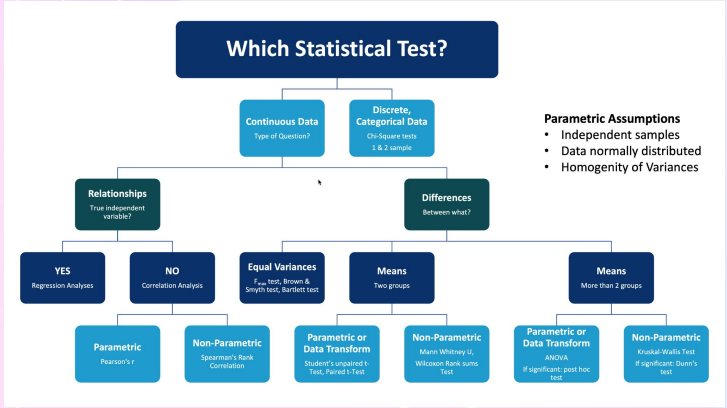
Миф №2: Доверительные интервалы и p-значения – это то, чем они кажутся

Гайдлайны

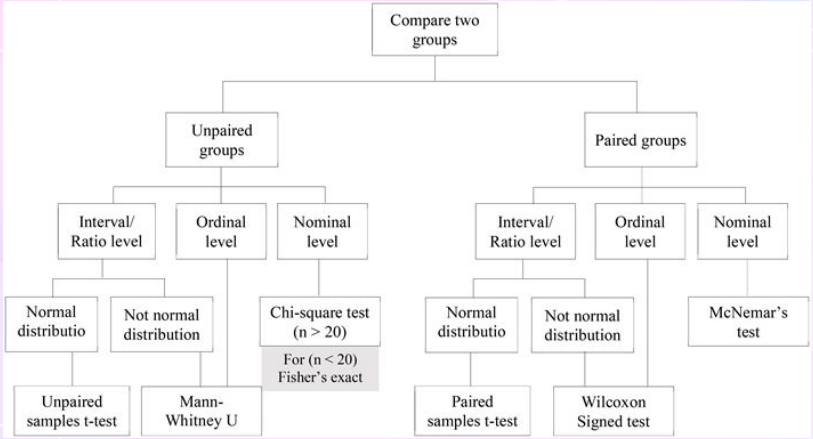
1



2

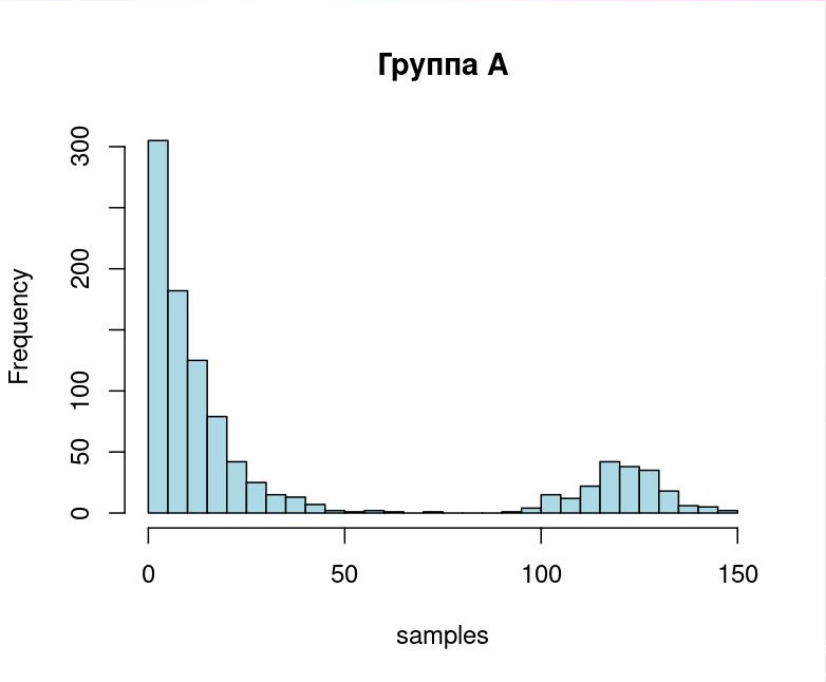


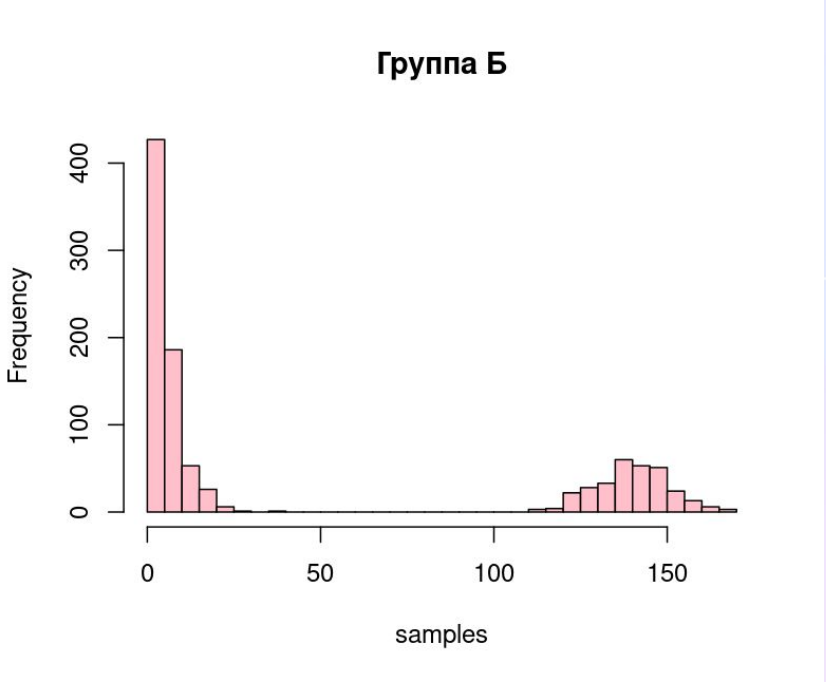
3



1. Гайд по выбору статистического метода. Blastim.
2. Multidisciplinary Research Methods for Engineers. TU Delft OpenCourseWare. https://ocw.tudelft.nl/course-readings/4-2-5-selecting-a-statistical-method/
3. A Comprehensive Guide for Selecting Appropriate Statistical Tests: Understanding When to Use Parametric and Nonparametric Tests. DOI: 10.4236/ojs.2023.134023

Пришел коллега и спросил, какой тест применить к таким данным (сравнение двух групп)





Кейс № 1

Коллега: реаниматолог

Выборка: биомаркер сепсиса

Что на самом деле волнует: как часто возникают септические состояния?

Разумный вариант решения:

Разделить на группы сепсис/не сепсис по рефентному порогу

и сравнить частоты.

Кейс № 2

Коллега: аспирант медицинского ВУЗа

Выборка: баллы за грантовые заявки по тематикам А и Б

Что на самом деле волнует: какая из двух тем имеет больше шанс на выигрыш?

$$H0: Pr\{A > Б\} = Pr\{A < Б\} ?$$

Разумный вариант решения:

Манна-Уитни* (Бруннера-Мюнцеля)

Кейс № 3

Коллега: главврач

Выборка: длительности госпитализаций

Что на самом деле волнует: сколько в среднем коек будет занято в больнице

Формула Литтла:

E[Койки] = E[Длительность] * Интенсивность

Разумный вариант решения:

Тест, оценивающий среднее. Простейший вариант - таки t-тест (но, возможно, есть и более эффективные параметрические тесты).

Жестокий мир статистики

1. Да, *t*-тесту будет плохо (ошибка 1 рода будет далека от желаемой).
2. Но он отвечает на вопрос исследователя.
3. Можно попробовать что-то улучшить ресемплингом (пермутации, бутстреп).
4. Можно попробовать более сложные параметрические тесты.
5. … но тоже без гарантий.

Иногда нужно признать, что при имеющихся данных **нет** адекватного теста, отвечающего на заданный вопрос.

Иногда можно попробовать переформулировать вопрос.
Но сначала - вопрос!

Дальнейшее обсуждение – на встрече 26 марта
Миф №3: Ненормальное распределение требует ненормальных решений

Ещё один пример из анализа категориальных данных

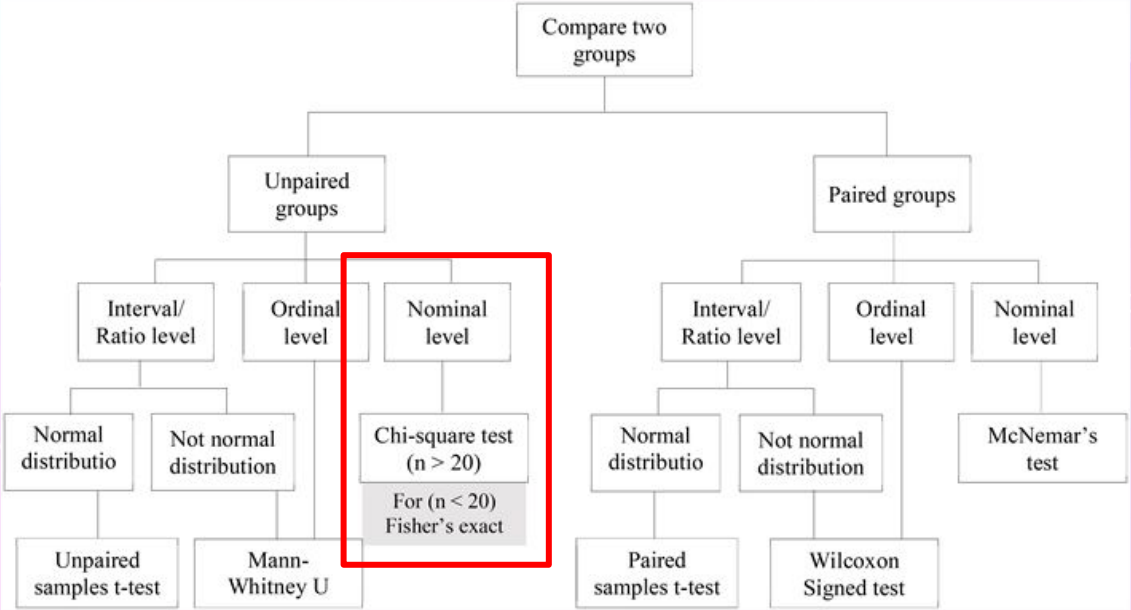
Таблицы сопряженности aka contingency tables (четырех- и многопольные таблицы)

Сводка по двум категориальным признакам:

ID	Группа	Исход
1	Treatment	+
2	Treatment	-
3	Placebo	-
4	Treatment	+
5	Placebo	-
6	Treatment	+
7	Placebo	-

	-	+
Treatment	1	3
Placebo	3	-

Типовой совет по выбору теста:

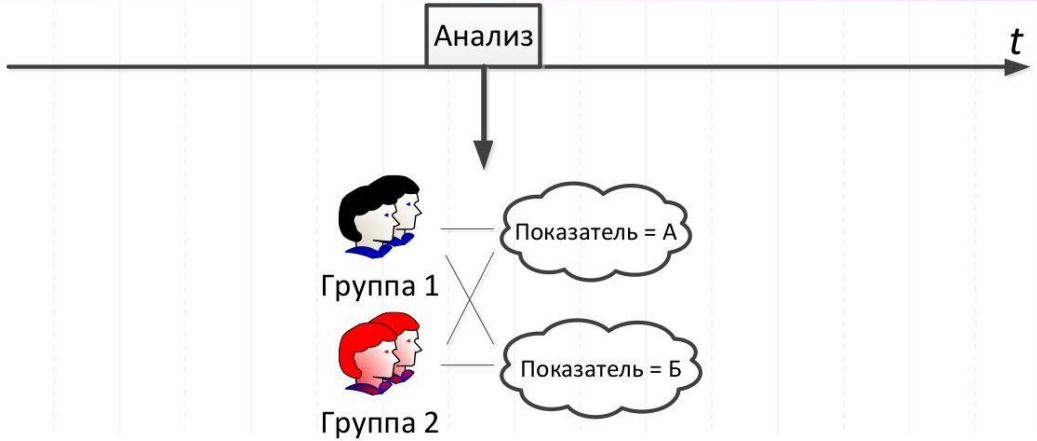


1. A Comprehensive Guide for Selecting Appropriate Statistical Tests: Understanding When to Use Parametric and Nonparametric Tests. DOI: 10.4236/ojs.2023.134023

Вариации: если в каждой ячейке >5 наблюдений, то хи-квадрат, иначе – Фишер.

А как появилась таблица сопряженности?

Поперечное исследование (cross-sectional study)



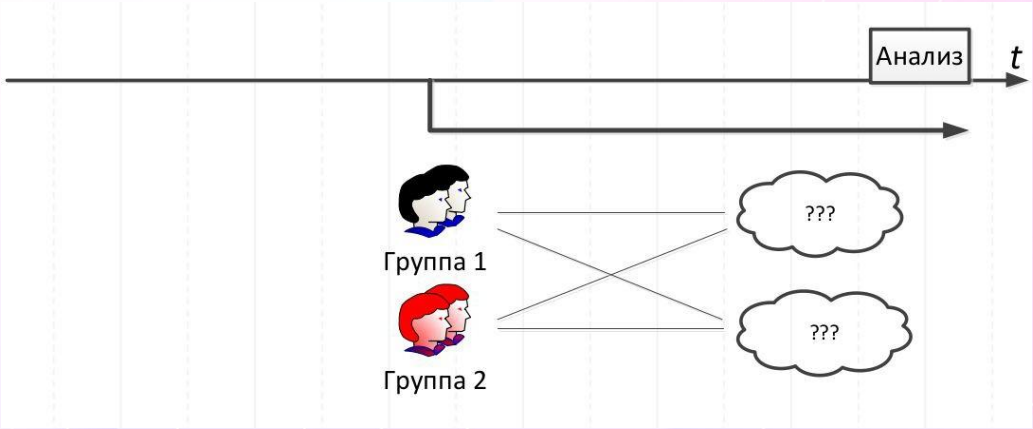
1. Набирается одна (!) группа добровольцев, подходящих под критерии влючения/невключения.
2. Проводится анкетирование (разовое обследование/забор анализов).
3. Находятся взаимосвязи.

Случайно

Случайно

	Показатель +	Показатель -
Группа 1 (принимают витамины)	a	b
Группа 2 (не принимают витамины)	c	d

Когортное исследование (cohort study)



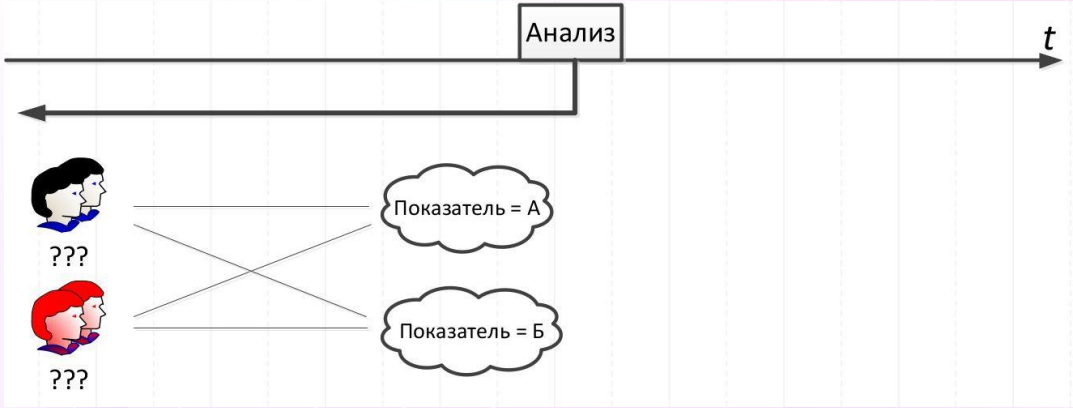
1. Набирается несколько групп интереса (проводится интервенция).
2. По ходу исследования производится замер показателей
3. Анализ проводится по прошествии времени наблюдения.

Случайно

Неслучайно

	Показатель +	Показатель -
Группа 1 (принимают витамины)	a	b
Группа 2 (не принимают витамины)	c	d

Исследование случай-контроль (case-control study)



1. Набирается несколько групп в соответствии с исходом.
2. Обращаясь во времени назад, описываются исходные группы, в которых событие произошло и не произошло.
3. В момент анализа производится сравнение групп между показателями.

Неслучайно

Случайно

	Показатель +	Показатель -
Группа 1 (принимают витамины)	a	b
Группа 2 (не принимают витамины)	c	d

Могут ли распределения и в столбцах и столбика быть неслучайными одновременно?

Могут ли распределения и в столбцах и столбика быть неслучайными одновременно?

МОГУТ!

Достаточно экзотический пример

1. Набрали участников исследования, принимавших и не принимавших витамины (10 + 10).
2. Независимому эксперту предстоит угадать, кто из участников принимал, а кто нет (распределить участников на две группы по 10 человек).

Неслучайно

Неслучайно

	Оценка: принимают витамины	Оценка: не принимают витамины	Итого:
Группа 1 (принимают витамины)	4	6	10
Группа 2 (не принимают витамины)	6	4	10
Итого:	10	10	

Три конфигурации таблиц сопряженности:

Хи-квадрат

1. Оба распределения (по столбцам и строкам) – случайны.
2. Одно распределение – случайно, другое - нет.
3. Оба распределения – неслучайны.

Фишер

Итоги:

1. Оба теста были разработаны для оценки зависимости между величинами, но **для разных дизайнов исследования**.
2. Тест хи-квадрат является асимптотическим (нулевое распределение сходится к хи-квадрат на больших выборках).
3. На малых выборках тест хи-квадрат может не обеспечивать заданную вероятности ошибки I рода из-за погрешности аппроксимации нулевого распределения.
4. Тест Фишера был разработан для специфических условий (двойная фиксированность распределений) и является “точным” именно там. В других условиях может быть излишне консервативным.

Серая зона: малые выборки + Случай I/II.
Хи-квадрат не достиг асимптотики,
а Фишер – тестирует не ту гипотезу.

Since 1976...

G. R. Garside and C. Mack, Actual Type 1 Error Probabilities for Various Tests in the Homogeneity Case of the 2×2 Contingency Table, The American Statistician, Vol. 30, No. 1 (Feb., 1976), pp. 18-21

Случай №2 (кейс-контроль или когортное исследование)

Table of Observed Frequencies

	A	*B*	Row Totals
C	n_1	n_2	*K*
not-*C*	n_3	n_4	*K'*
Column Totals	*M*	*M'*	*N*

Exact Type I error probabilities for cases in which M = M'

Key: F ≡ Fisher's test, Y ≡ Yates' corrected chi-squared test,
U ≡ Uncorrected chi-squared test, B ≡ Boschloo's test,
G ≡ λ_G corrected chi-squared test, γ ≡ Boschloo's 'raised level'

		TEST				
Parameters		F	Y	U	B	G
M = M' = 40 α = 0.05					γ = 0.071	λ_G = 0.07
p	0.1	0.0194	0.0193	0.0544	0.0405	0.0434
	0.2	0.0296	0.0258	0.0509	0.0444	0.0444
	0.3	0.0306	0.0297	0.0529	0.0486	0.0486
	0.4	0.0278	0.0278	0.0474	0.0451	0.0451
	0.5	0.0284	0.0284	0.0466	0.0465	0.0465

Quiz

Пример 1

Идеальное нормальное распределение в том виде, в котором его описал немецкий математик Иоганн Карл Фридрих **Гаусс** – распределение Гаусса – в природе не встречается. Отсюда обязательная задача исследователя: понять, **значимо или не значимо отличается наша выборка от нормальной?**

...

Мы воспользуемся **критерием Шапиро-Уилка**.

1 **Нулевая гипотеза** у этого критерия: «проверяемое распределение подчиняется нормальному закону», как и все в природе.

2 **Альтернативная**: «Проверяемое распределение отлично от нормального».

...

Использовав критерий, мы получили p-значение (p-value) > 0.05, что говорит о незначимом отклонении нашей выборки от нормальной, о том, что она подчиняется нормальному закону распределения.

...

2 Во-вторых, тип распределения данных определяет и методы для их анализа:

- **параметрические методы** работают корректно только в случае нормального распределенной выборки,
- **непараметрические** – устойчивы к нарушению нормальности.

Неверная интерпретация базовых терминов:

- Тест служит для проверки гипотезы касающейся генеральной совокупности, а не выборки.

Использование терминов, не имеющих принятого определения:

- Что такое “в природе”? В генеральной совокупности? Если да, то зачем проверять, ведь в природе оно не встречается.

Непонимание границ применения методов:

- Параметрические методы вполне могут работать с ненормальными распределениями.
- Тем более, что идеальное нормальное распределение “в природе не встречается” :)

Остальные вы найдете после следующих двух встреч!

Пример № 2

Наделение методов дополнительными функциями:

- Складывается впечатление, что тест Манна-Уитни также сравнивает средние, только более “корректно”.

Использование терминов, не имеющих принятого определения:

- Что такое “выброс” в данном случае?
- Откуда они взялись в выборке?
- Почему нужно изменять нулевую гипотезу вместо удаления выбросов?

Хороший пост на эту тему
от Матвея Славенко:
https://t.me/choking_data/36

Разбор ошибок от Елены Убогоевой (https://t.me/stats_for_science):
https://ubogoeva.github.io/R4Analytics/posts/review_of_statistics_course.htm

Ждём вас на следующей лекции
19 марта в 19:00 МСК!

Ольга Мироненко и Максим Кузнецов
Миф №2: Доверительные интервалы и p-значения –
это то, чем они кажутся

Институт биоинформатики в социальных сетях

Чат по биостатистике и R: https://t.me/chat_biostat_R

По всем вопросам: biostat@bioinf.me

Сайт Института: bioinf.me

Институт в VK: vk.com/bioinf

Телеграм-канал Института: t.me/bioinforussia

Чат про образование и карьеру: t.me/bioinf_career

YouTube-канал: www.youtube.com/bioinforussia