19 И 26 МАРТА, 10 АПРЕЛЯ 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ

ОЛЬГА МИРОНЕНКО И МАКСИМ КУЗНЕЦОВ | МИФ №2: ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ И P-ЗНАЧЕНИЯ – ЭТО ТО, ЧЕМ ОНИ КАЖУТСЯ

HTTPS://T.ME/CHAT_BIOSTAT_R

Основные источники о «мифах»

- **A Dirty Dozen**: Twelve P-Value Misconceptions
  - Goodman S
  - Seminars in Hematology, 2008:45(3)
- The **ASA Statement** on *p*-Values: Context, Process, and Purpose
  - Wasserstein RL, Lazar NA
  - The American Statistician, 2016:70(2)
  - + Supplements
- Statistical Tests, P Values, Confidence Intervals, and Power: **a Guide to Misinterpretations**
  - Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG
  - European Journal of Epidemiology, 2016:31
- The 2017 **ASA Symposium** on Statistical Inference
- Statistical Inference in the 21st Century: **A World Beyond p < 0.05**
  - The American Statistician, 2019:73(sup1)
- Scientists rise up against statistical significance (**Retire statistical significance**)
  - Amrhein V, Greenland S, McShane B
  - Nature, 2019:567(7748)

План лекции

- Статистический вывод и статистическая модель
- Доверительные интервалы (ДИ)
  - Определение и интерпретация
  - Заблуждения относительно ДИ
- *p*-значения
  - Определение и интерпретация
  - Заблуждения относительно p-значений
- Рекомендации по интерпретации и представлению результатов исследований

Ограничения

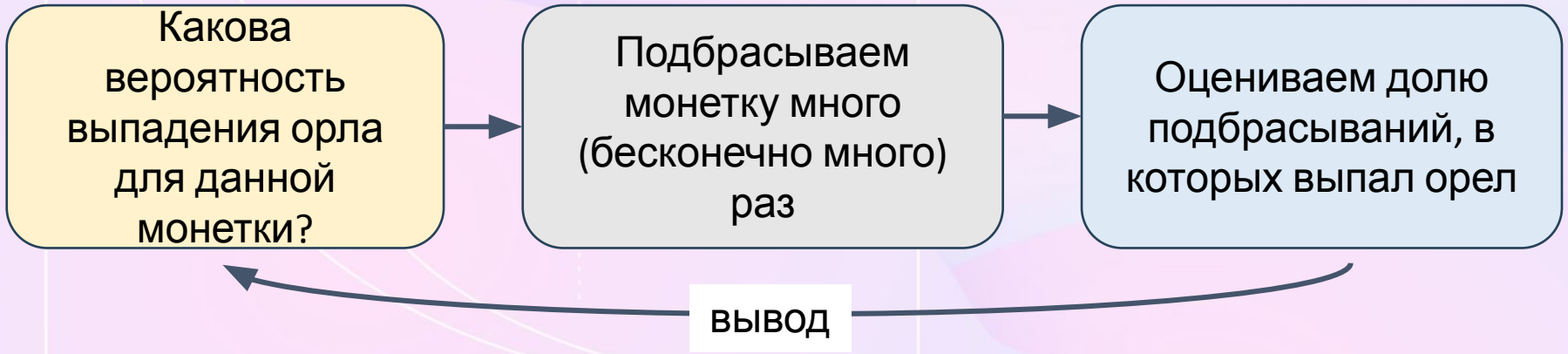
- Целевая аудитория – прежде всего, «нестатистики» (исследователи из разных областей науки)
- Инструменты статистического вывода (statistical inference) в рамках фриквентистской статистики (frequentist statistics)

Статистический вывод (statistical inference)

- Оценка параметра распределения случайной величины по выборочным наблюдениям с целью сделать вывод о значении этого параметра в генеральной совокупности и/ или проверить гипотезу о нем
  - Генеральная совокупность – **ГС** (например, взрослые с сахарным диабетом)
  - Случайная величина – ***Y*** (например, систолическое артериальное давление)
  - Истинное (в ГС) значение параметра распределения – $\boldsymbol{\theta}$ (например, среднее значение)
- Элементы (задачи) статистического вывода:
  - **Оценивание**
    - Точечная оценка – $\widehat{\boldsymbol{\theta}}$ (например, среднее значение САД по данным выборки)
    - Интервальная оценка – доверительный интервал (**ДИ**) (95% ДИ для среднего САД)
  - **Проверка гипотез** в отношении интересующего нас параметра
    - Например, H0: $\theta = 140$

Частотная (frequentist) статистика

- В основе – идея о *многократном (бесконечном) повторении эксперимента*
- Истинное значение параметра θ является фиксированным (не является случайной величиной) – нам просто неизвестно его значение



Операционализация

- Исследовательский вопрос → какой показатель нужно/ можно измерить, какой параметр его распределения оценить и каким образом с его помощью ответить на поставленный вопрос?
- См. примеры из Лекции 1:
  - Оценка частоты септических состояний
  - Выбор темы для грантовой заявки
  - Загруженность коек в больнице
- Операционализация может подразумевать ряд допущений, например:
  - Суррогатная конечная точка сильно коррелирует с жесткой
  - Условия эксперимента не меняются от наблюдения к наблюдению

Статистическая модель – основа для статистического вывода

- **Модель генерации данных** – математическое описание взаимосвязи между значениями изучаемой случайной величины и параметром интереса – **в ГС!**
  - Параметр интереса
  - Дополнительные параметры (например, для конфаундеров)
  - Примеры:
    - Оценка среднего значения – $Y_i = \theta + \epsilon_i$
    - Оценка разницы в средних – $Y_i = \mu_A + \theta \cdot [TRT_i = B] + \epsilon_i$ или $Y_i = \mu_A + \theta \cdot [TRT_i = B] + \beta \cdot Z_i + \epsilon_i$
- **Нулевая гипотеза** о значении параметра в ГС (в случае ее проверки)
- **Эстиматор** для оценки параметра/ **статистический тест** для проверки гипотезы
- **Допущения –** для получения оценки «хорошего» качества, контроля ошибки I рода:
  - О параметрах модели генерации данных
  - Допущения выбранного эстиматора/ теста
  - «Неявные» допущения (относительно отбора наблюдений, назначения воздействия, сбора данных, цензурирования и т.п.)

Эстиматор

Способ оценки параметра интереса по выборочным данным в соответствии со статистической моделью (например, формула для расчета среднего значения или выборочной дисперсии, МНК для оценки коэффициентов линейной регрессии, процедура оценки ДИ и т.п.)

Три распределения

Population

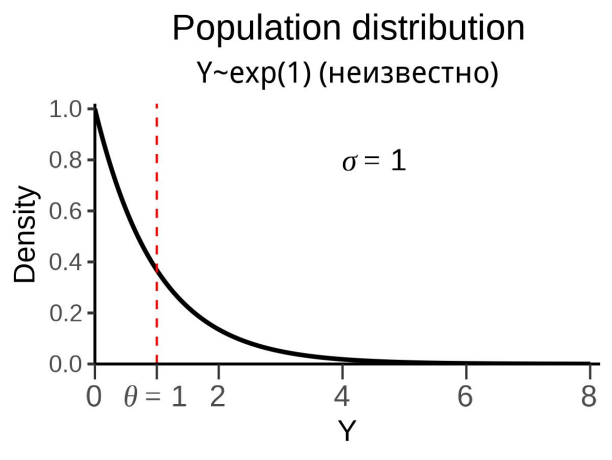
распределение случайной величины в генеральной совокупности → истинное значение параметра (θ)

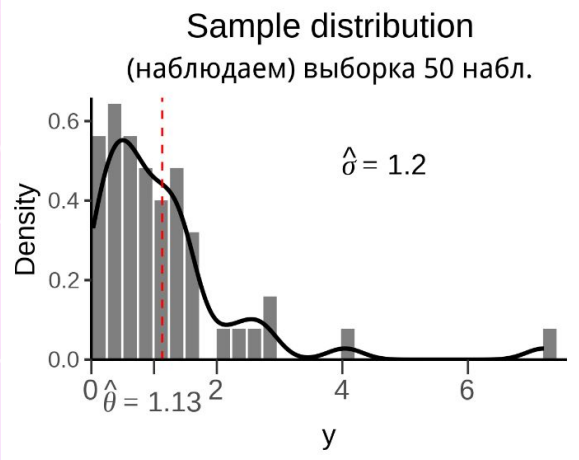
Sample

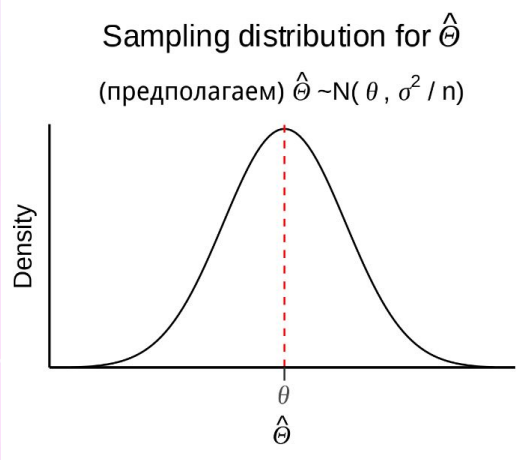
распределение признака в выборке → точечная оценка параметра ($\hat{\theta}$)

Sampl*ing*

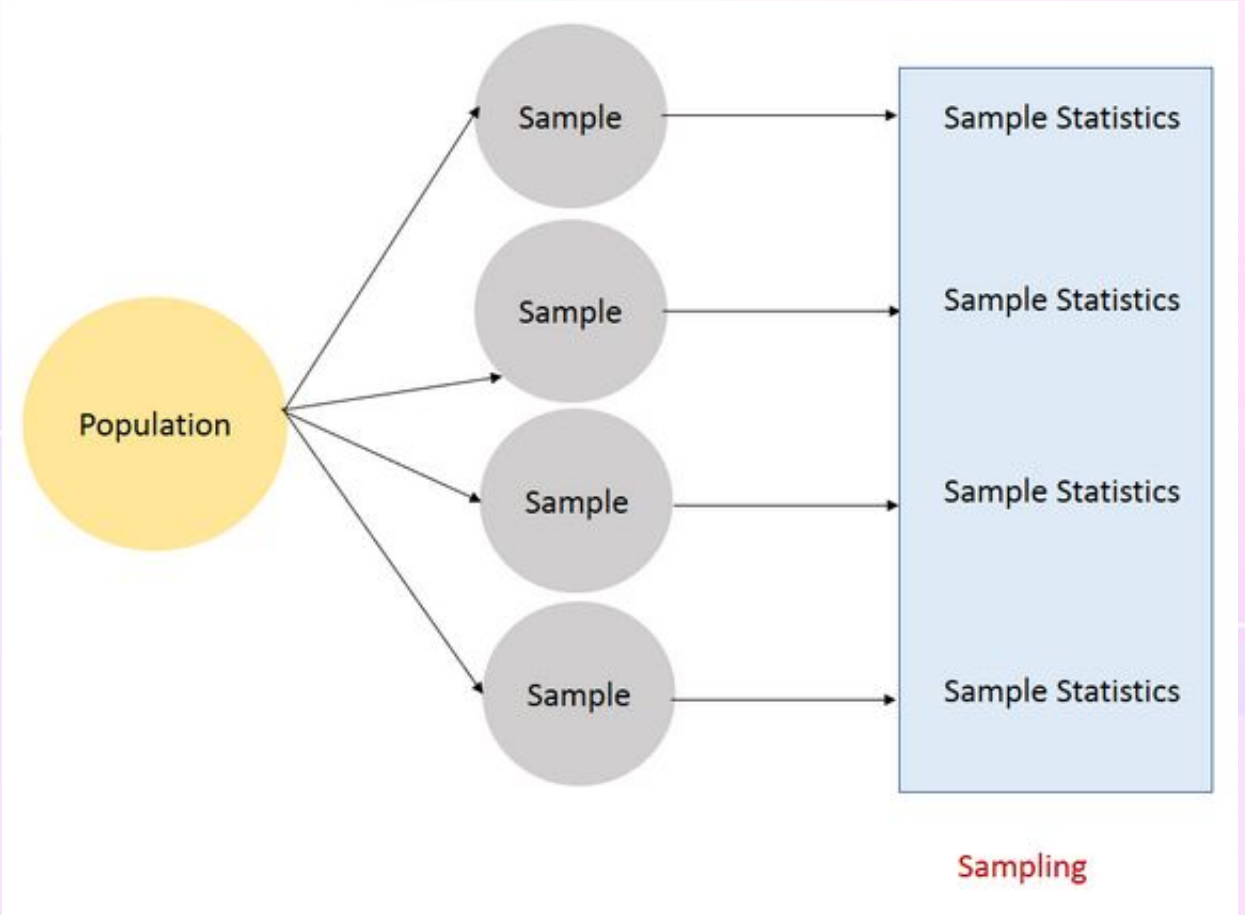
распределение $\hat{\theta}$, полученных *при многократном повторении эксперимента в идентичных условиях* (выборки одинакового размера из одной генеральной совокупности) → интервальная оценка параметра / проверка гипотезы о значении параметра







Три распределения

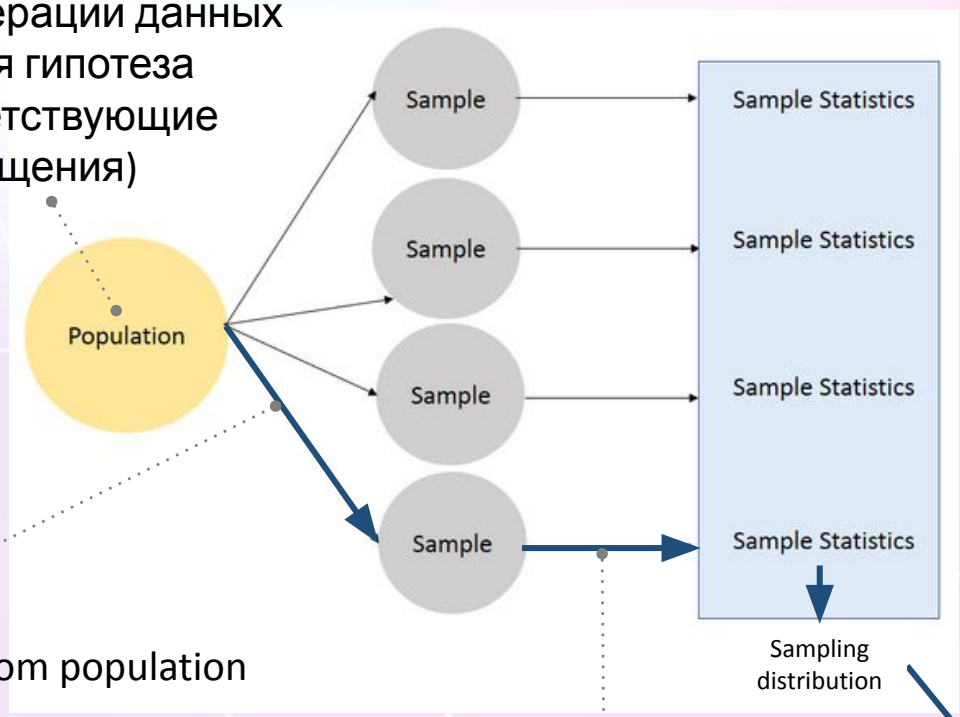


Статистическая модель и ее контекст

Исследовательский вопрос / гипотеза

Операционализация
(и соответствующие допущения)

Модель генерации данных
Нулевая гипотеза
(и соответствующие допущения)



Отбор наблюдений (sampling from population distribution)
Назначение воздействия (assignment)
Сбор данных, измерение показателей
(и соответствующие допущения)

Точечный эстиматор
(и соответствующие допущения)

Интервальный эстиматор
Статистический тест
(и соответствующие допущения)

Статистический вывод (statistical inference)

Отбор результатов для репортирования/ публикации
Принятие решений
Интерпретация
(и соответствующие допущения)

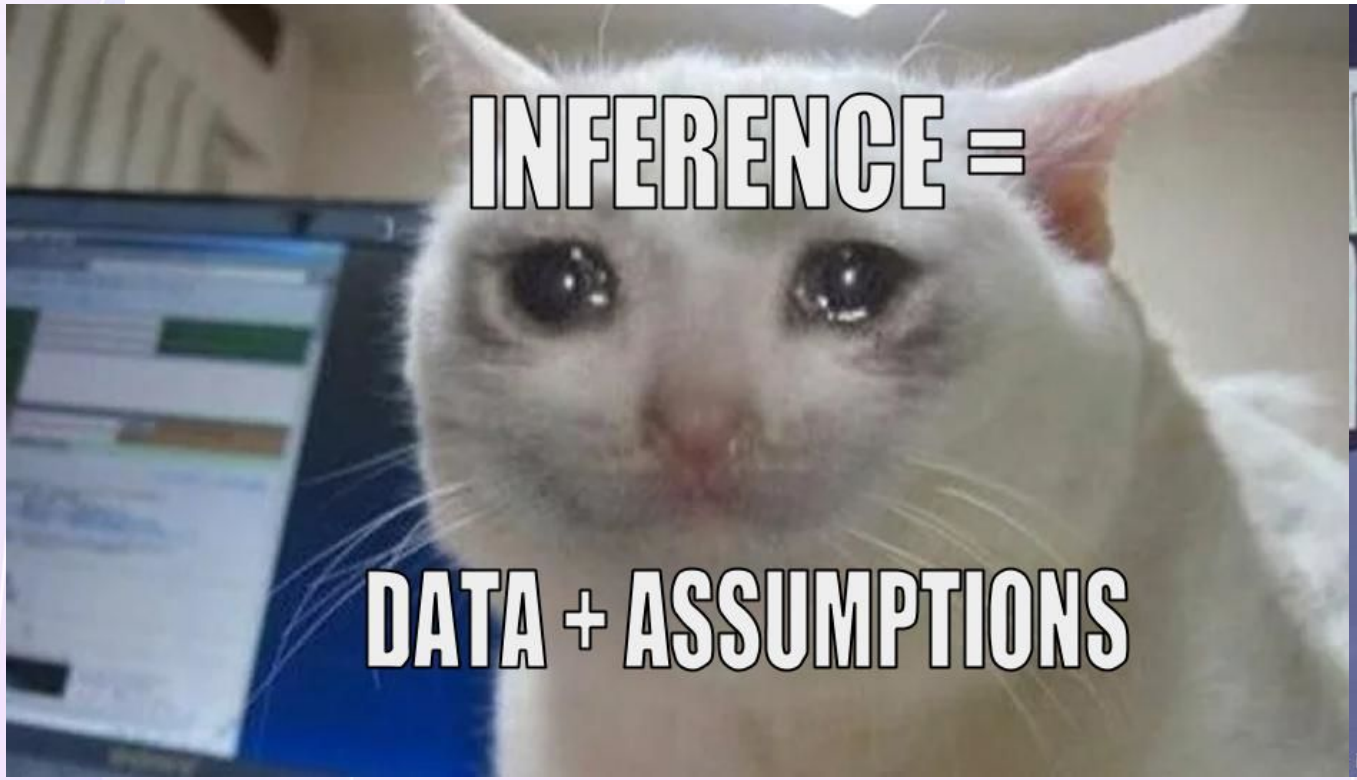
Исследовательский вывод (scientific inference)

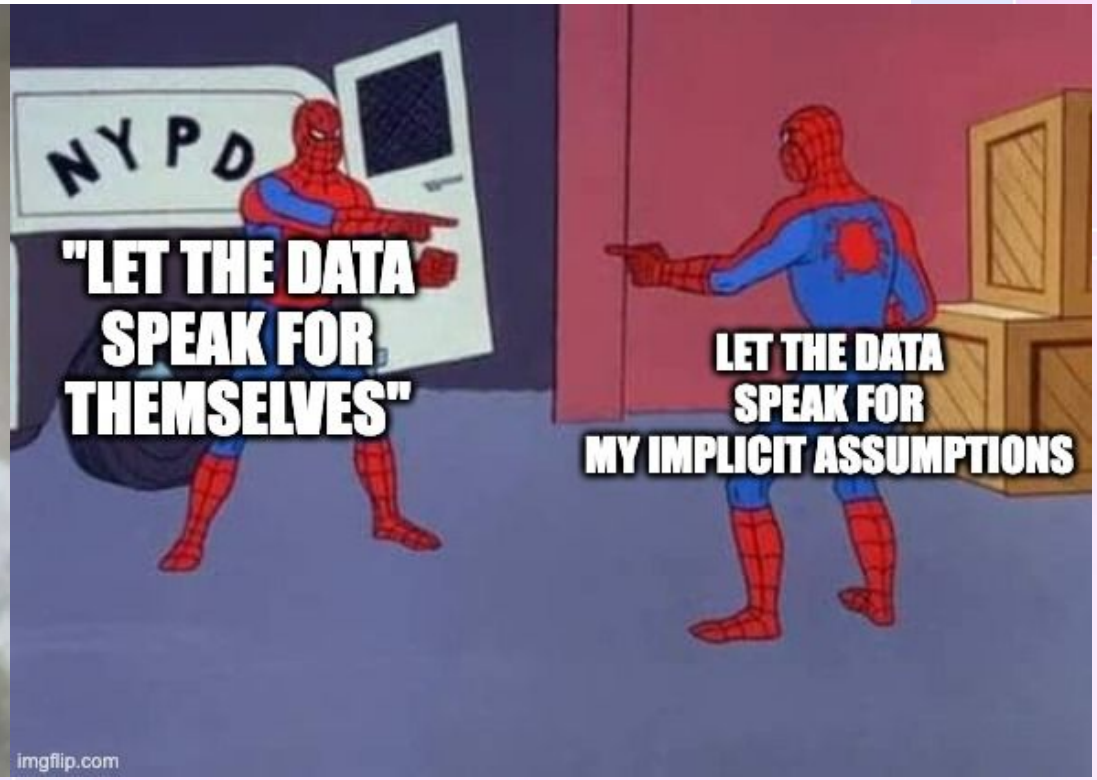
Статистическая модель и ее контекст

Исследовательский вопрос / гипотеза

?

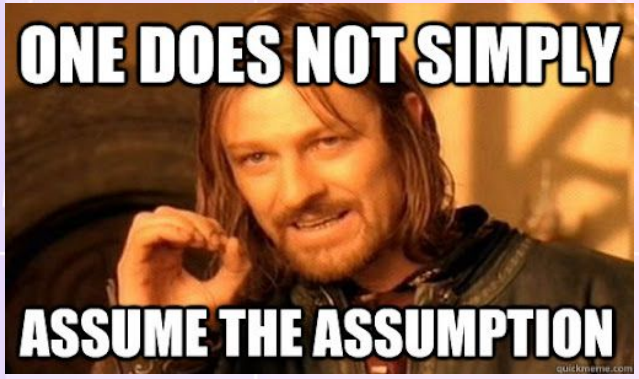
Статистический вывод (statistical inference)

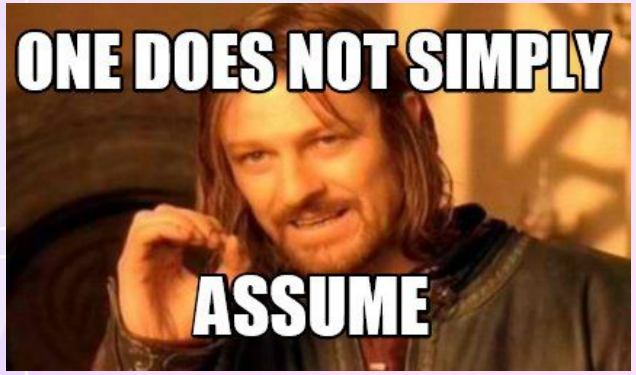


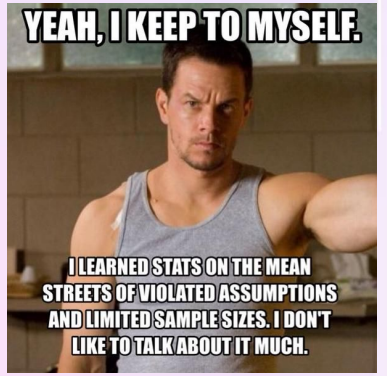


Допущения: великие и ужасные

- См. Лекцию 1
- Harrell FE. (2024) What Does a Statistical Method Assume?
  - Ситуации, когда мы считаем, что в статистической процедуре/ эстиматоре S используется допущение A
  - Как оценивать «качество» S
  - Неявные допущения
  - Примеры допущений для различных тестов (в том числе непараметрических) / моделей и возможных последствий их невыполнения
- Часто допущения, которые можно проверить статистическими тестами, воспринимаются как самые критичные, хотя неявные обычно важнее







Точечная оценка $\hat{\theta}$

- Best guess о значении θ в генеральной совокупности по имеющимся данным
- Возможны разные эстиматоры с разными свойствами и допущениями
- **Состоятельность** – при *увеличении объема выборки* вероятность больших отклонений $\hat{\theta}$ от θ стремится к 0
  - Несостоятельность ← систематические ошибки (не решаются увеличением выборки)
- **Несмещенность** – если бы могли *многократно повторить эксперимент в идентичных условиях*, то *среднее значение из* $\hat{\theta}$, полученных в этих экспериментах, было бы $\approx \theta$
  - Необходимо для допущений о sampl***ing*** distribution $\Rightarrow$ оценке ДИ/ проверки гипотез
- Состоятельность и несмещенность – это не свойства значения $\hat{\theta}$, полученного в данном исследовании, а свойства эстиматора (которые могут не выполняться)
  - Проверить состоятельность и несмещенность оценки невозможно – можно только ее обеспечить (не только выбором корректного эстиматора)

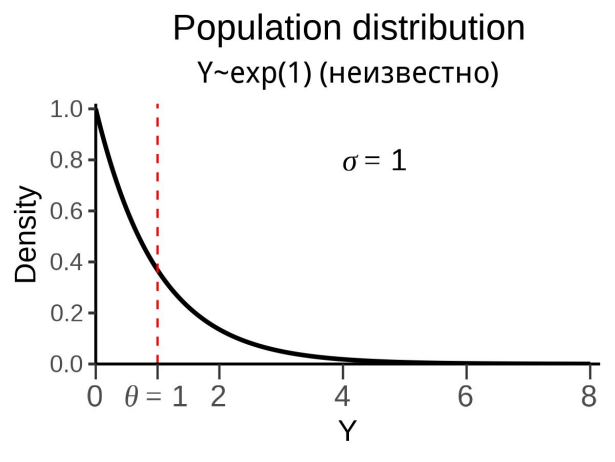
Интервальная оценка

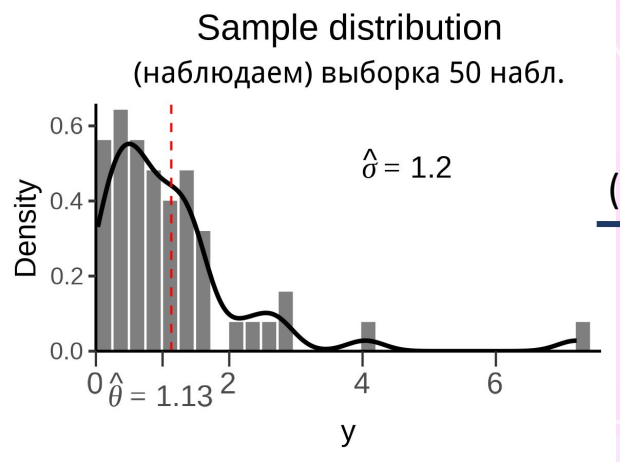
- Почему недостаточно точечной оценки?
  - Выборочные данные вместо ГС $\Rightarrow$ Необходимо учесть sampl***ing*** variability (степень нашей неуверенности относительно точечной оценки, связанную с использованием выборочных данных для вывода о ГС)
- Интервальная оценка – ДИ:
  - Sampl***ing*** distribution (распределение $\hat{\theta}$) $\rightarrow$ различные эстиматоры для стандартной ошибки (sampl***ing*** standard deviation) и квантилей распределения

Чему доверяем в доверительном интервале?

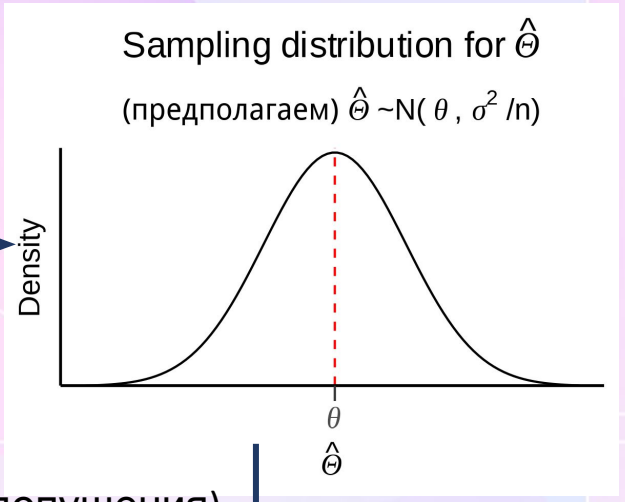
- При многократном повторении эксперимента в идентичных условиях, с использованием выбранной статистической модели и выбранного эстиматора ожидаем, что в $X\%$ случаев ДИ накроет истинное значение параметра (❄)
- $\Rightarrow$ Наше доверие – не к тому конкретному ДИ, который мы получили по нашим данным, а к методу оценки ДИ – к тому, что он сможет соблюсти условие ❄
- Проверить соблюдение условия ❄ по данным невозможно – можно только обеспечить его выполнение (допущения)

Интервальная оценка – пример

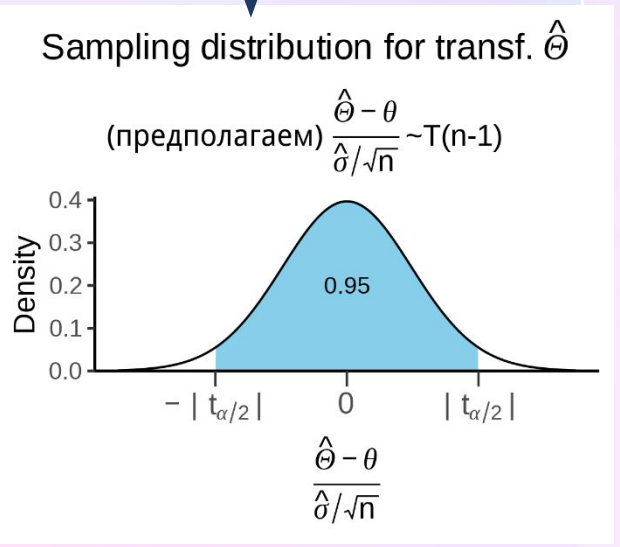




(допущения)



(допущения)



Мы бы хотели найти такой интервал значений, что в $X\%$ экспериментах с идентичными условиями θ попадала бы в этот интервал (в $\alpha \cdot 100\% = 100\% - X\%$ случаев ДИ «промахнется»)

ДИ для θ с уровнем доверия $(1-\alpha)\cdot 100\%$ по 1 выборке:

$$\hat{\theta} \pm t_{1-\frac{\alpha}{2},n-1} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

(в примере на этом слайде 95% ДИ для среднего: 0.79-1.47)

«Хорошая» интервальная оценка

- **Точность** – чем у̇же ДИ, тем точнее оценка параметра, ширина ДИ зависит от:
  - Уровня доверия $X\%$
  - Объема выборки
  - Вариации данных в выборке
  - ! Выбранного эстиматора и выполнения его допущений
- **Обеспечение уровня покрытия ДИ заявленному**
  - Проверить соблюдение этого свойства по данным невозможно – только обеспечить (допущения)

Интерпретация ДИ – пример

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ
«Центр экспертизы и контроля качества медицинской помощи»
Министерства здравоохранения Российской Федерации
(ФГБУ «ЦЭККМП» Минздрава России»)

Утверждено приказом
ФГБУ «ЦЭККМП» Минздрава России

от «29» декабря 2017 г. № 181-од

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ПРОВЕДЕНИЮ

НЕПРЯМЫХ СРАВНЕНИЙ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ

Доверительный интервал (ДИ) – интервал значений признака, рассчитанный для какого-либо параметра распределения (например, среднего) по выборке и с определенной вероятностью (например, 95% – для 95% ДИ) включающий истинное значение этого параметра во всей популяции [5].

Интерпретация ДИ – пример

Доверительный интервал (ДИ) — интервал значений признака, рассчитанный для какого-либо параметра (например, среднего значения признака) по выборке и с определенной вероятностью (например, 95%) включающий истинное значение этого параметра во всей генеральной совокупности.

ДИ всегда связан с каким-либо уровнем доверия, уверенности. Напомним, что все оценки параметров признаков генеральной совокупности, полученные на основе анализа данных выборки, не являются абсолютно истинными. Они истинны лишь с некоторой долей вероятности. Так, если мы выбираем доверительный коэффициент (ДК; степень уверенности, выраженная в процентах; вероятность того, что данный интервал содержит истинное значение параметра) равным 95%, то это означает, что в 95 выборках из 100, сделанных таким же способом из генеральной совокупности объектов исследования, оценка параметра признака будет находиться в рассчитанном нами ДИ.

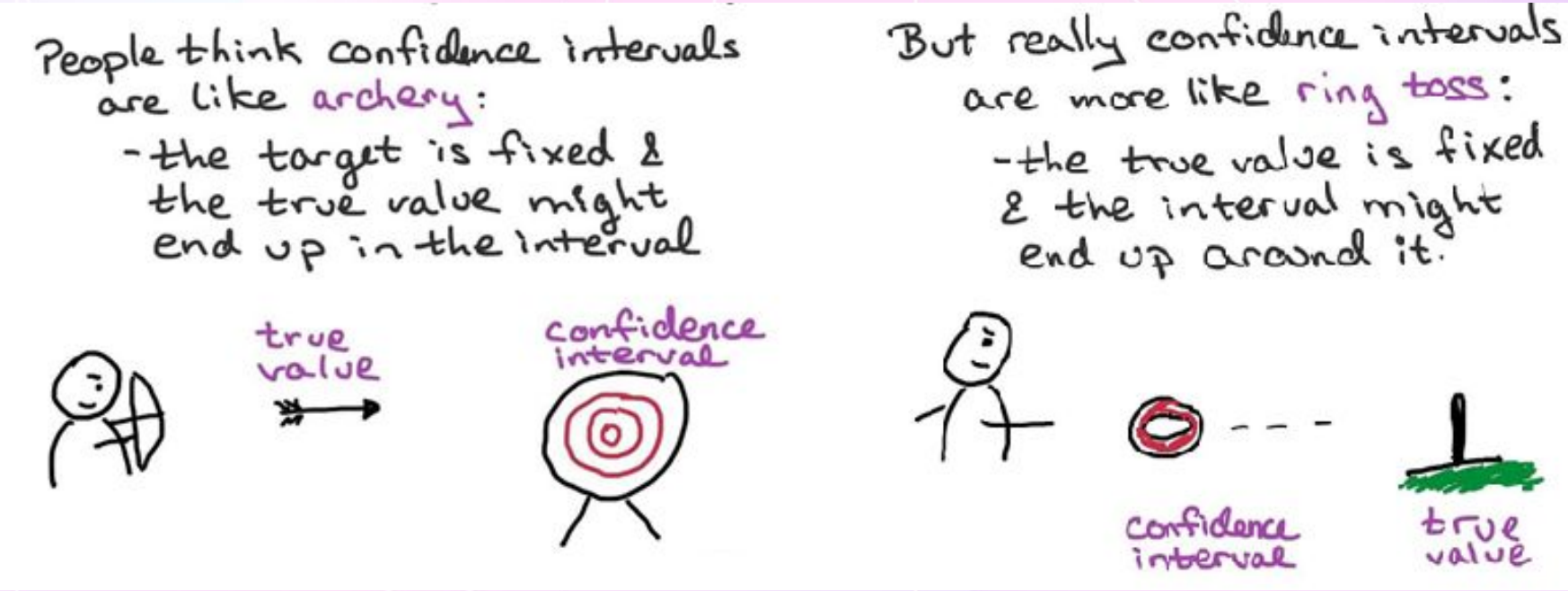
Миф про ДИ № 1

- Истинное значение параметра с 95% вероятностью попало в полученный 95% ДИ
- Мы на 95% уверены, что полученный нами 95% ДИ содержит истинное значение параметра

Что не так?

- Истинное значение параметра фиксировано (не является случайной величиной), оно нам неизвестно ⇒ полученный нами ДИ либо накрыл его, либо нет
- 95% – это наша уверенность в том, что при многократном повторении исследования по выборкам аналогичного размера из той же генеральной совокупности, 95% полученных нами 95% ДИ будут покрывать истинное значение параметра, если все допущения, сделанные в процессе оценки этих интервалов, соблюдаются

Анти-Миф про ДИ № 1 – иллюстрация 1

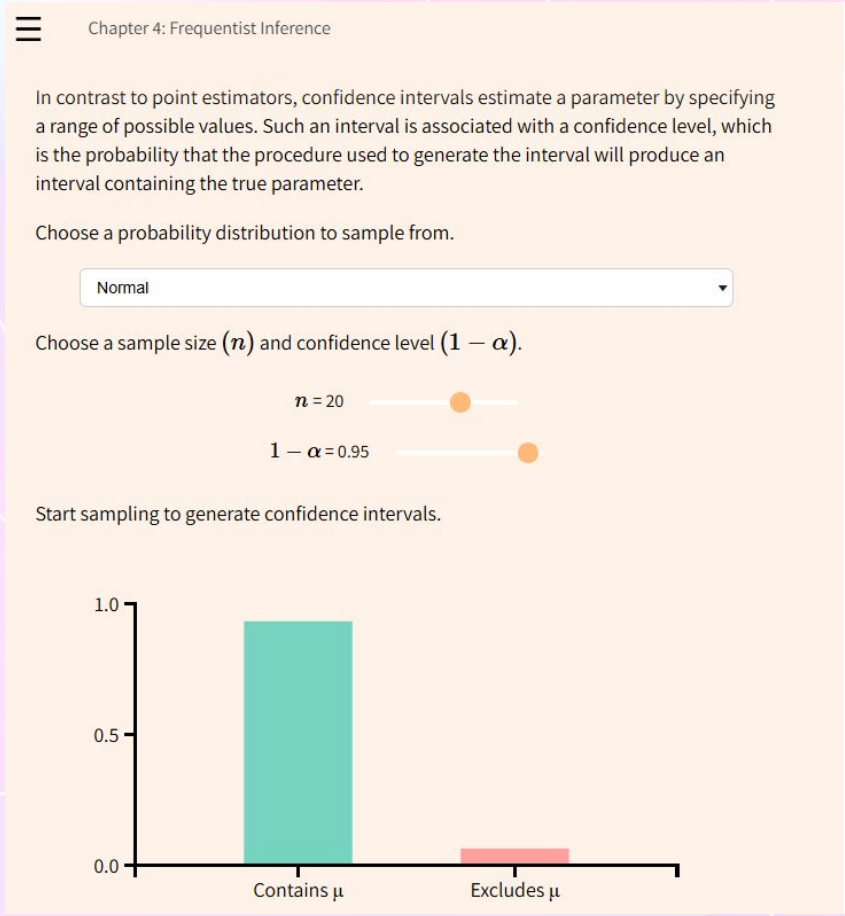


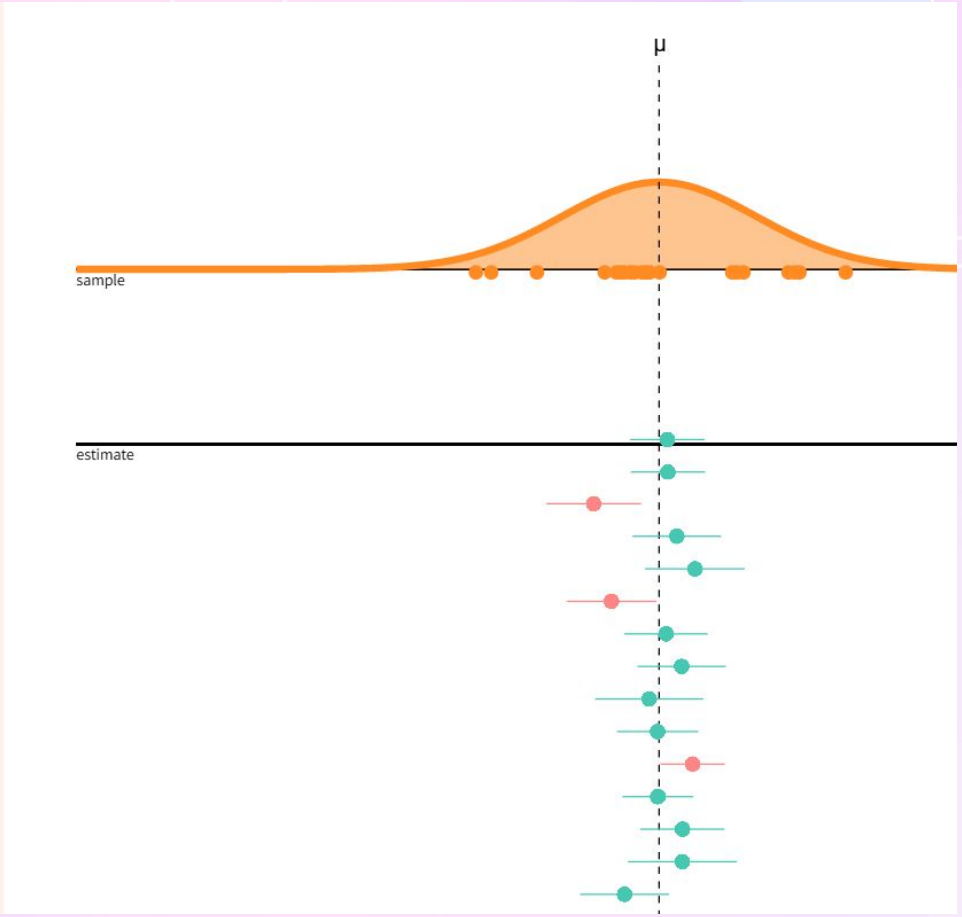
Murray E. (2019) Having confidence in confidence intervals

Анти-Миф про ДИ № 1 – иллюстрация 2

«Почувствовать» определение ДИ:

- Seeing theory
- R <- psychologist





Миф про ДИ № 2

- При многократном повторении исследования в 95% повторов оценка параметра попадет в 95% ДИ, который мы получили в данном исследовании

Что не так?

- 95% – это частота покрытия истинного значения *доверительными интервалами* при повторах, если все допущения, сделанные в процессе оценки этих интервалов, соблюдаются (к точечной оценке при повторах не имеет отношения)

Анти-Миф про ДИ № 2

- Обычно вероятность попадания точечной оценки при повторе в 95% ДИ предыдущего исследования < 95%
- Пример: несмещенный эстиматор, нормальное sampl***ing*** распределение, в двух независимых исследованиях одинаковая стандартная ошибка → вероятность того, что среднее значение в исследовании 2 попадет в 95% ДИ из исследования 1 составляет 83%

A Guide to Misinterpretations

Анти-Миф про ДИ № 2 в деталях

Условия: несмещенный эстиматор, нормальное sampl***ing*** распределение, в двух независимых исследованиях одинаковая стандартная ошибка

1. Пусть $\hat{\theta}_1$ и $\hat{\theta}_2$ – точечные оценки параметра из исследования 1 и 2, соответственно
2. По условию: $\hat{\theta}_1 \sim N(\theta, se(\hat{\theta}_1)), \hat{\theta}_2 \sim N(\theta, se(\hat{\theta}_2)), se(\hat{\theta}_1) = se(\hat{\theta}_2) = \hat{\sigma}$
3. Если $\hat{\theta}_2$ попало в 95% ДИ из исследования 1, то $|\hat{\theta}_1 - \hat{\theta}_2| \leq 1.96 \cdot \hat{\sigma}$
4. $2 \Rightarrow \hat{\theta}_1 - \hat{\theta}_2 \sim N(0, 2 \cdot \hat{\sigma}^2) \Rightarrow \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{2} \cdot \hat{\sigma}} \sim (0,1)$
5. $3 \Rightarrow \frac{|\hat{\theta}_1 - \hat{\theta}_2|}{\sqrt{2} \cdot \hat{\sigma}} \leq \frac{1.96}{\sqrt{2}}$
6. $4,5 \Rightarrow \mathrm{P}\left(\frac{|\hat{\theta}_1 - \hat{\theta}_2|}{\sqrt{2} \cdot \hat{\sigma}} \leq \frac{1.96}{\sqrt{2}}\right) = 1 - 2 \cdot \Phi\left(-\frac{1.96}{\sqrt{2}}\right) \approx 0.83$

Мифы про ДИ № 1 и 2 – выводы

⇒ ДИ, полученный в данном исследовании, ничего не говорит об истинном значении параметра

⇒ ДИ, полученный в данном исследовании, не позволяет точно предсказать результаты будущих исследований

- Как все-таки корректно интерпретировать ДИ, обсудим чуть дальше

Проверка гипотезы

- **Нулевая статистическая гипотеза H0**
  - Точечная или составная
  - Необязательно о равенстве 0 (null ≢ nil)
- **Альтернативная статистическая гипотеза H1**
  - Точечная или составная
  - Односторонняя или двусторонняя
  - Не исследуется при проверке гипотезы – нужна для понимания того, что именно считать несогласованностью данных с H0 / «экстремальными» значениями оценки параметра
- H0 либо верна, либо нет – нам просто это неизвестно, но мы исходим из предположения о том, что она верна
- Проверка H0 путем оценки того, насколько полученные данные согласуются с ней
  - **! При выполнении допущений статистической модели**
  - Если сильно не согласуются, то сомневаемся в том, что H0 верна и отвергаем ее
  - Понимание того, как устроена альтернатива (отклонение от нулевой гипотезы) в ГС важно при выборе эстиматора/ теста для контроля ошибки II рода

Проверка H0 – порядок действий

- Формулируем H0 и H1
- Задаем заранее вероятность ошибки I рода (α) – вероятность отклонить нулевую гипотезу, когда она верна
- Выбираем статистический тест → тестовая статистика (функция от выборочных данных – «расстояние»** от данных до того, что нам предсказывает статистическая модель при верной H0*)
- Предположения о распределении тестовой статистики при верной H0 **и соблюдении всех остальных допущений статистической модели**
  - Если бы многократно повторяли эксперимент в идентичных условиях при верной H0* и в каждом из них оценивали бы тестовую статистику, то как выглядело бы распределение этих значений
- Оцениваем тестовую статистику по выборочным данным

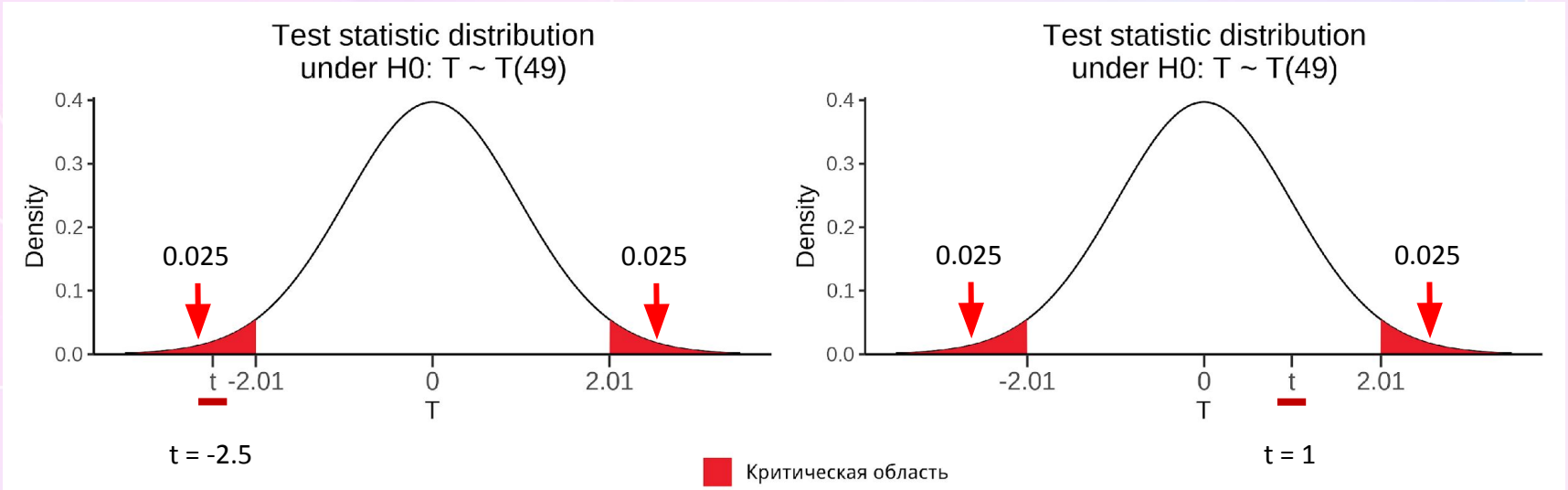
* - при соблюдении допущений статистической модели
** - A Guide to Misinterpretations

Проверка H0 – способ 1

- Насколько значение тестовой статистики по выборке «укладывается» в распределение тестовой статистики при верной H0*?
  - Область критических значений тестовой статистики (критическая область) – при верной H0* получение какого-либо значения из этой области является маловероятным (вероятность ≤ α)
  - Если тестовая статистика по выборке попадает в критическую область, отвергаем H0 на уровне значимости α
  - В противном случае – не отвергаем H0

$$H0: \theta = \theta_0$$
$$H1: \theta \neq \theta_0$$
$$T = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}$$
$$\alpha = 0.05$$
$$n = 50$$

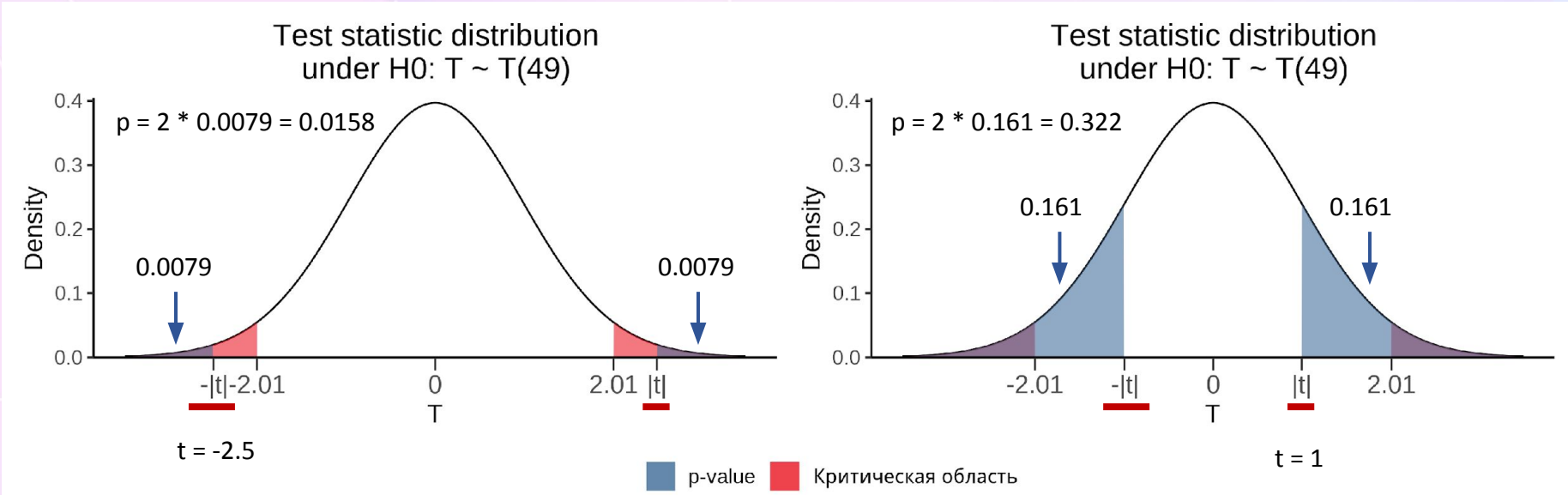
* - при соблюдении допущений статистической модели



Проверка H0 – способ 2

- Какова вероятность получить значение тестовой статистики такое же, как в выборке, или более экстремальное при верной H0* – *p*-значение?
  - Что значит «более экстремальное», определяем заранее (на основе H1)
  - Если $p < \alpha$, отвергаем H0 на уровне значимости α
  - В противном случае – не отвергаем H0
- Вывод аналогичен способу 1

$$H0: \theta = \theta_0$$
$$H1: \theta \neq \theta_0$$
$$T = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}$$
$$\alpha = 0.05$$
$$n = 50$$



* - при соблюдении допущений статистической модели

p-значение – определение

- Вероятность получить значение тестовой статистики такое же, как в выборке (t), или более экстремальное при верной H0* (*при многократном повторении эксперимента в идентичных условиях*); эта вероятность оценивается по распределению тестовой статистики T

$$p = \Pr(T \geq |t|; H0^*) + \Pr(T \leq -|t|; H0^*)$$

- Поскольку тестовая статистика обычно рассчитывается по $\hat{\theta}$, то *p*-значение – это вероятность получить значение оценки параметра такое же, как в выборке, или более экстремальное при верной H0* (*при многократном повторении эксперимента в идентичных условиях*); эта вероятность оценивается по samp***ling*** distribution для $\widehat{\Theta}$

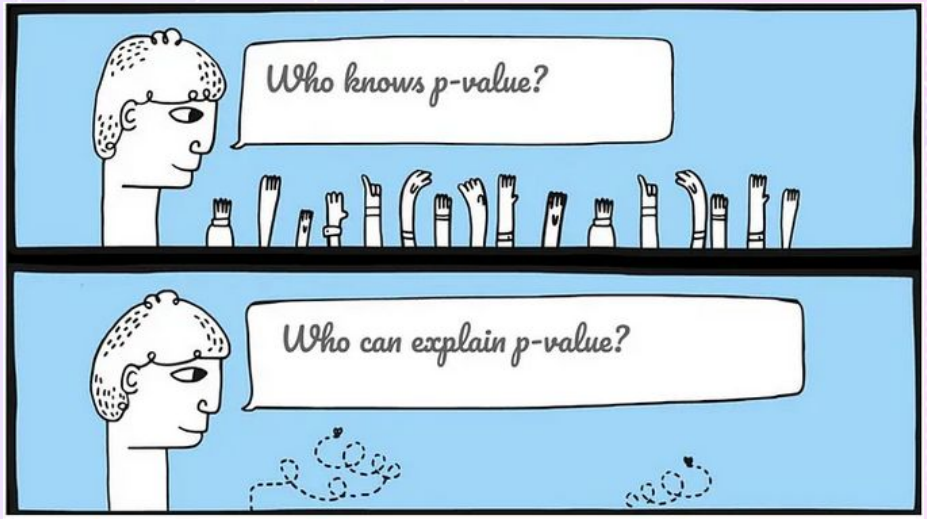
$$p = \Pr\left(\widehat{\Theta} - \theta_0 \geq \left|\hat{\theta} - \theta_0\right|; H0^*\right) + \Pr(\widehat{\Theta} - \theta_0 \leq -|\hat{\theta} - \theta_0|; H0^*)$$

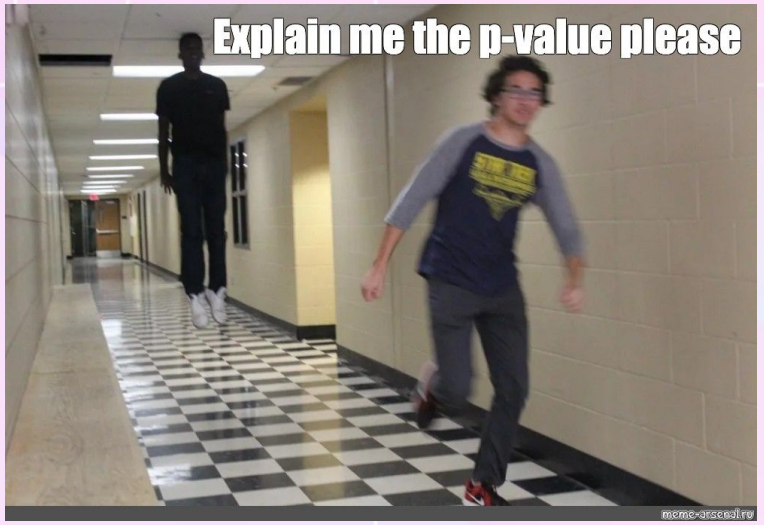
* - при соблюдении допущений статистической модели и такой же выборочной дисперсии

формулы приводятся для двусторонней альтернативной гипотезы

Нельзя просто так взять и дать определение *p*-значению

- Aschwanden C. (2015). Not even scientists can easily explain p-values, video





p-значение – интерпретация

ASA Statement:

- **Степень несогласованности** (incompatibility) между имеющимися данными и выбранной статистической моделью (включая H0 и все допущения)
- Чем *меньше* *p*-значение, тем *меньше* статистическая согласованность между данными и статистической нулевой гипотезой, ***если*** выполняются все допущения, принятые при оценке p-значения
- Несогласованность ⇒ сомнения в H0 ***или*** выполнимости допущений

Berry DA. (2016) P-Values Are Not What They're Cracked Up to Be (supplement to ASA statement):

- Важность «данных» в определении степени несогласованности (манипуляции с данными до оценки *p*-значения)

p-значение – не только для проверки H0

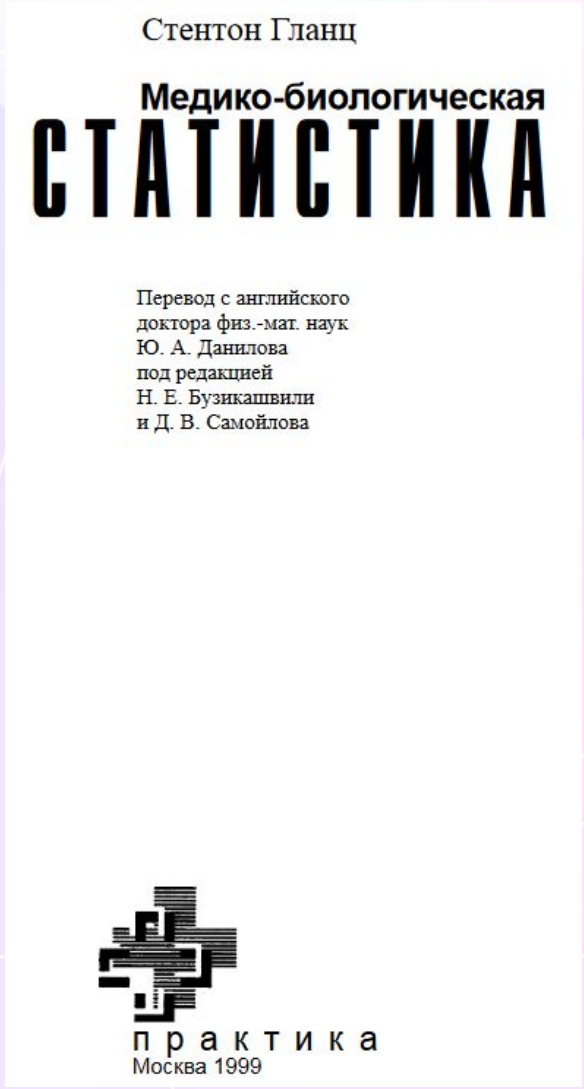
- *p* – мера согласованности между данными и выбранной статистической моделью (включая H0 и все допущения)

⇒ Указание полученного *p*-значения (например, p = 0.023, p = 0.789) вместо записей вида p < 0.05, p > 0.05 более информативно и предпочтительно

Интерпретация *p*-значения – пример (1)

Значение p — это рассчитанная в ходе статистического теста вероятность ошибочного отклонения нулевой гипотезы об отсутствии различий. Другое определение: значение p — это вероятность получить данные анализируемых выборок в случае справедливости нулевой гипотезы (в частности, в случае отсутствия различий групп). Третье определение: значение p — это вероятность справедливости нулевой гипотезы.

Интерпретация *p*-значения – пример (2)



Теперь вернемся к препарату и вычислим значение критерия. Если оно превышает критическое значение, то мы можем утверждать следующее, *если бы нулевая гипотеза была справедлива, то вероятность получить наблюдаемые различия была бы меньше 5%*. В принятой системе обозначений это записывается как $P < 0{,}05$. Отсюда мы заключаем, что гипотеза об отсутствии влияния препарата на температуру вряд ли справедлива, то есть различия статистически значимы (при 5% уровне значимости). Разумеется, этот вывод по сути своей носит вероятностный характер. Не исключено, что мы ошибочно признаем неэффективный препарат эффективным, то есть найдем различия там, где их нет. Однако мы можем утверждать, что вероятность подобной ошибки не превышает 5%.

Дадим определение *P*.

P есть вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

Определение можно сформулировать и по-другому.

P есть вероятность ошибочно отвергнуть нулевую гипотезу об отсутствии различий.

Упрощая, можно сказать, что *P — это вероятность справедливости нулевой гипотезы*. Часто говорят также, что *P* — это вероятность ошибки. В общем, и это верно, однако несколько неточно. Дело в том, что существует два рода ошибок. Ошибка I рода — это ошибочное заключение о существовании различий, которых в действительности нет. Вероятность именно этой ошибки оценивает *P*. Возможна и противоположная ошибка — принять неверную нулевую гипотезу то есть не найти действительно существующее различие. Это гак называемая ошибка II рода. О вероятности этой ошибки *P* ничего не говорит, мы обсудим ее в гл. 6.

Интерпретация *p*-значения – пример (3)

P value

Usually used to test a *null hypothesis*, the *P* value gives the probability of any observed differences having happened by chance. See page 24.

Миф про *p*-value № 1

- *p* – это вероятность того, что H0 верна
- *p* – это вероятность того, что мы совершим ошибку, если отвергнем H0

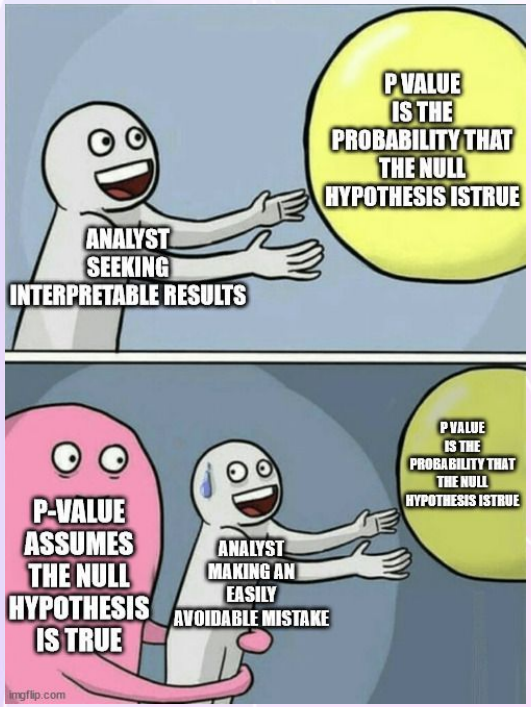
Что не так?

- *p* оценивается при допущении о том, что H0 верна* $\Rightarrow$ оно, по определению, не может быть вероятностью того, что H0 верна
- *p* – про согласованность данных и H0*
- H0 либо верна, либо нет – нам это неизвестно $\Rightarrow$ если H0 верна, а мы ее отвергли, то мы точно ошиблись, если не верна – точно не ошиблись, но в любом случае не знаем об этом
- *p* – не вероятность ошибки, в оценке *p* не используется α, а сравнение с ней не делает *p* вероятностью ошибки

A Dirty Dozen; ASA Statement

* - при соблюдении допущений статистической модели

Источник многих мифов о *p* и ДИ – inverse probability



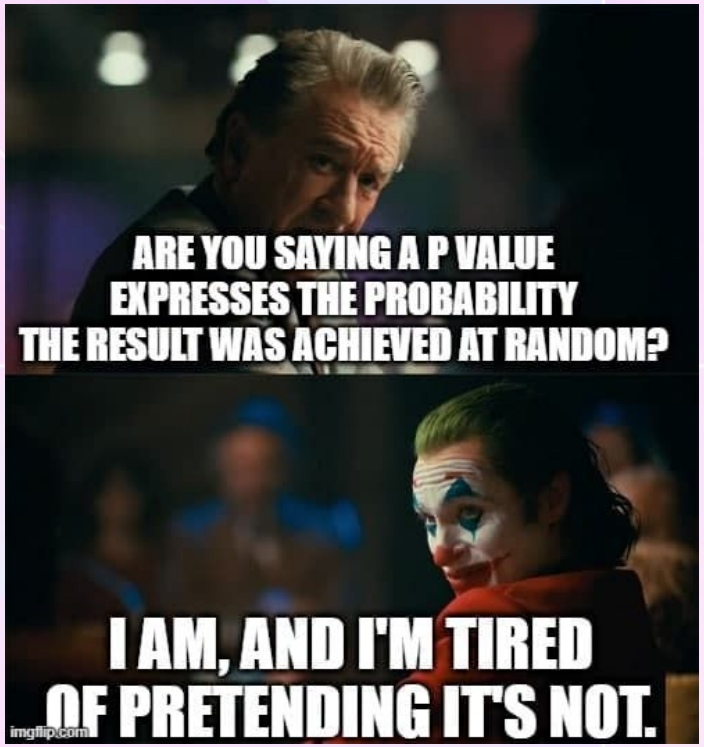
- В идеале хотели бы по *p*-значению (т.е. по данным) сделать вывод о гипотезе, но можем сделать вывод только о данных, их соответствии H0
- *P*-value отвечает на вопрос, который никто не задавал (Senn S.)

Миф про *p*-value № 2

- *p* – это вероятность того, что полученные результаты/ выявленные различия/ отклонение от H0 случайны (маленькое *p* – отклонения от H0 неслучайны)

Что не так?

- Если H0 верна, то вероятность того, что какое-либо отклонение от нее будет случайным, равна 1
- Утверждение о том, что обнаруженные отклонения случайны, эквивалентно утверждению о том, что все допущения, принятые при оценке *p*, корректны, включая H0 ⇒ *p* – это вероятность, оцененная из допущения о том, что все отклонения случайны, а не наоборот (еще один пример inverse probability)



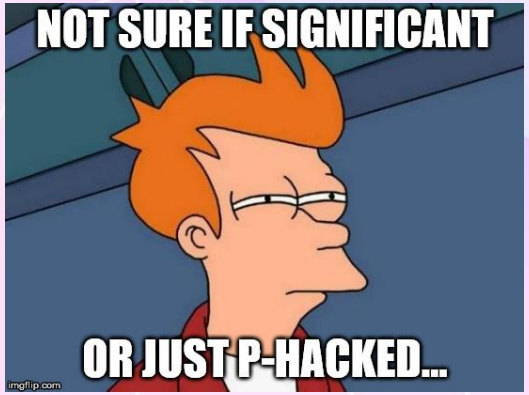
Миф про *p*-value № 3

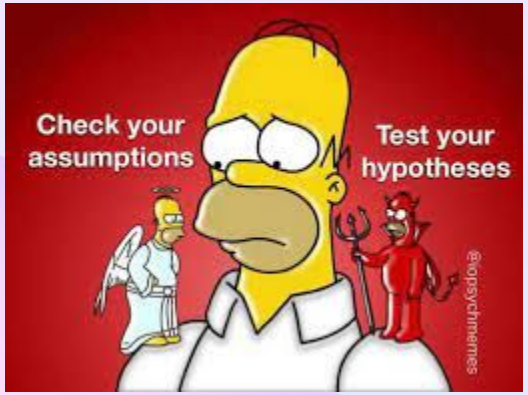
- *p* – это вероятность получить данные, аналогичные нашей выборке / оценку параметра, как по нашей выборке, если H0 верна

Что не так?

- *p* – вероятность получить не только такую же оценку параметра, как по данной выборке, но и «более экстремальные» оценки, при верной H0 и соблюдении допущений статистической модели
  - С точностью до определения того, что считается «более экстремальным»
  - «Более экстремальные» оценки – ненаблюдаемые (наше представление о том, какими они могли бы быть при повторении исследования, при верной H0)

Миф про *p*-value № 4





- $p < \alpha$ означает, что H0 не верна
- Чем меньше *p*, тем больше у нас доказательств в пользу того, что H0 не верна

Что не так?

- H0 либо верна, либо нет – нам это неизвестно
- Маленькое значение *p* – знак того, что данные, с точки зрения модели (при выполнении ее допущений, в том числе H0), являются необычными
- Это может быть результатом того, что H0 не верна, или того, что были нарушены какие-то допущения или быть результатом случайной ошибки
- Меньше *p* – больше *сомнений* в том, что H0 верна (при выполнении допущений статистической модели)

Миф про *p*-value № 5

- $p \geq \alpha$ означает, что H0 верна/ мы можем принять H0

Что не так?

- H0 либо верна, либо нет – нам это неизвестно
- Большое значение *p* – знак того, что данные, с точки зрения модели (при выполнении ее допущений, в том числе H0), не являются необычными
- Это может быть результатом как того, что H0 верна, так и того, что было сделано какое-то жесткое допущение или мощность была недостаточной, так и быть результатом случайной ошибки

Миф про *p*-value № 6

- Чем больше *p*, тем больше у нас доказательств в пользу того, что H0 верна

Что не так?

- *p* – мера согласованности между данными и этой H0*
- Наиболее согласующаяся с данными H0* – та, для которой *p* = 1
- Сравнивая p между разными вариантами H0*, можем делать вывод о том, какая из них лучше согласуется с данными
- Но гипотез (моделей), которые лучше согласуются с данными, чем та H0*, которую мы проверяли, может быть много, при этом идеальная согласованность с какой-либо одной гипотезой (моделью) не означает, что все остальные гипотезы (модели) опровергаются

A Guide to Misinterpretations

* - при соблюдении допущений статистической модели

ДИ – интерпретация

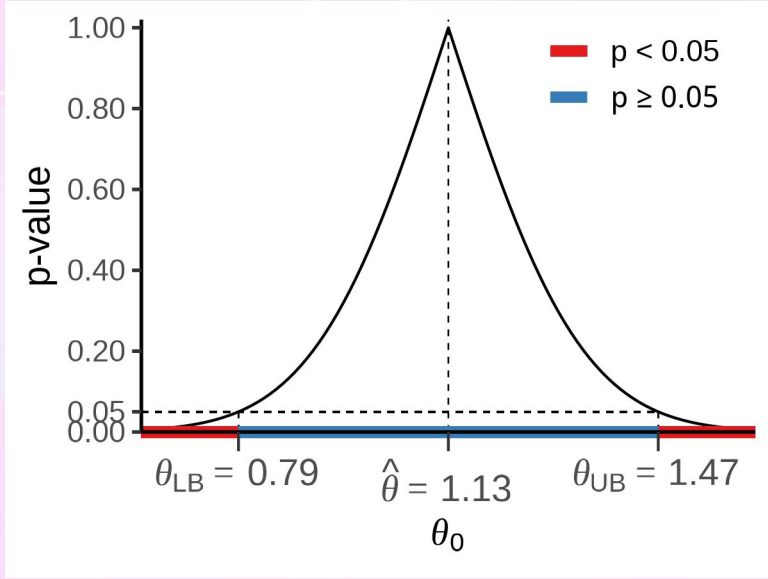
- Оценка ДИ не привязана к проверке H0
- Что, если проверить H0: $\theta = \theta_0$ против H1: $\theta \neq \theta_0$ для всех значений θ_0 внутри полученного $(1-\alpha)\cdot 100\%$ ДИ и для каждой из них рассчитать p?

$\Rightarrow (1-\alpha)\cdot 100\%$ ДИ – это интервал таких значений параметра, проверка H0* для которых даст $p \geq \alpha$

$\Rightarrow (1-\alpha)\cdot 100\%$ ДИ – это способ обобщить результаты проверки гипотез для множества значений параметра

- p – мера согласованности между данными и H0*

$\Rightarrow$ Значения параметра внутри ДИ лучше согласуются с данными, чем значения вне его*



A Guide to Misinterpretations

* - при соблюдении допущений статистической модели

Миф про ДИ № 3

- Значение параметра, равное точечной оценке эффекта, – наиболее вероятно, значения ближе к границам ДИ – менее вероятны
- Значения за пределами ДИ опровергнуты полученными данными

Что не так?

- Истинное значение эффекта фиксировано (не является случайной величиной), оно или попало в полученный ДИ, или нет, а если попало, то неизвестно, в какой части ДИ оно находится
- Значения параметра внутри ДИ (и за его пределами) отличаются друг от друга по совместимости (согласованности) с полученными данными* (чем ближе к полученной точечной оценке, тем согласованность больше, чем дальше от нее в обе стороны – тем меньше)

A Guide to Misinterpretations

* - при соблюдении допущений статистической модели

Связь между ДИ и проверкой H0

1. ДИ для среднего значения θ оцениваем, исходя их sampl***ing*** distribution для преобразования $\widehat{\Theta}$*: $\frac{\widehat{\theta}-\boldsymbol{\theta}}{\widehat{\sigma}/\sqrt{n}} \sim T_{n-1}$
2. H0: $\theta = \theta_0$, H1: $\theta \neq \theta_0$, t-тест $\rightarrow$ распределение тестовой статистики T*: $\frac{\widehat{\theta}-\boldsymbol{\theta_0}}{\widehat{\sigma}/\sqrt{n}} \sim T_{n-1}$

- Если в 1 и 2 используется одинаковое преобразование $\widehat{\theta}$ и предполагается одинаковое распределение для него, тогда способы 1 и 2 дадут такой же вывод при проверке гипотезы, как и способ 3 (**дуальность между тестом и ДИ**)

* - при соблюдении допущений статистической модели

Проверка H0 – способ 3

- Покрывает ли рассчитанный по выборке $(1-\alpha)\cdot 100\%$ ДИ для θ значение параметра из нулевой гипотезы (θ_0)?
  - Если нет, отвергаем H0 на уровне значимости α
  - Если да, не отвергаем H0 на уровне значимости α

Расхождение между способами 1/2 и 3 при проверке H0

- Возможно (test-CI inconsistency), особенно для дискретных случайных величин
- Примеры и объяснения:
  - https://stats.stackexchange.com/a/169149
  - Fay M. (2010). Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data
  - Harrell FE. (2024) The log-rank Test Assumes More Than the Cox Model
  - Lin DY et al. (2016) On confidence intervals for the hazard ratio in randomized clinical trials
- Тест и ДИ решают разные задачи/ отвечают на разные вопросы, поэтому в некоторых случаях расхождение между ними в проверке H0 – это ОК
- Некоторые статистические тесты не имеют соответствующей процедуры получения *интерпретируемой* оценки эффекта – для некоторых (редких) задач это ОК (пример с выбором тематики грантовой заявки в Лекции 1)

Inference – пример (1)

April 18, 2017

Association Between Serotonergic Antidepressant Use During Pregnancy and Autism Spectrum Disorder in Children

Hilary K. Brown, PhD[1,2,3]; Joel G. Ray, MD, MSc, FRCPC[3,4,5]; Andrew S. Wilton, MSc[3]; et al

» Author Affiliations | Article Information

JAMA. 2017;317(15):1544-1552. doi:10.1001/jama.2017.3415

Results There were 35 906 singleton births at a mean gestational age of 38.7 weeks (50.4% were male, mean maternal age was 26.7 years, and mean duration of follow-up was 4.95 years). In the 2837 pregnancies (7.9%) exposed to antidepressants, 2.0% (95% CI, 1.6%-2.6%) of children were diagnosed with autism spectrum disorder. The incidence of autism spectrum disorder was 4.51 per 1000 person-years among children exposed to antidepressants vs 2.03 per 1000 person-years among unexposed children (between-group difference, 2.48 [95% CI, 2.33-2.62] per 1000 person-years; hazard ratio [HR], 2.16 [95% CI, 1.64-2.86]; adjusted HR, 1.59 [95% CI, 1.17-2.17]). After inverse probability of treatment weighting based on the high-dimensional propensity score, the association was not significant (HR, 1.61 [95% CI, 0.997-2.59]). The association was also not significant when exposed children were compared with unexposed siblings (incidence of autism spectrum disorder was 3.40 per 1000 person-years vs 2.05 per 1000 person-years, respectively; adjusted HR, 1.60 [95% CI, 0.69-3.74]).

Conclusions and Relevance In children born to mothers receiving public drug coverage in Ontario, Canada, in utero serotonergic antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child. Although a causal relationship cannot be ruled out, the previously observed association may be explained by other factors.

This matched case–control study revealed a null association between statin use and risk of glioma. While there was a small suggestion that long-term use of statins (≥90 prescriptions, or approximately 10–20 years of use) was associated with protection from risk of glioma, the effect estimates did not reach statistical significance.

Inference – пример (2)

› Eur J Epidemiol. 2016 Sep;31(9):947-52. doi: 10.1007/s10654-016-0145-7. Epub 2016 Apr 4.

Statin use and risk of glioma: population-based case-control analysis

Corinna Seliger [1], Christoph Rudolf Meier [2 3 4], Claudia Becker [2], Susan Sara Jick [3], Ulrich Bogdahn [5], Peter Hau [5], Michael Fred Leitzmann [6]

Affiliations + expand

PMID: 27041698 DOI: 10.1007/s10654-016-0145-7

Abstract

Statins have been reported to decrease the incidence of cancer, but the risk of glioma among statin users has been investigated in only two prior observational studies, both of them suggesting a modest protective effect of statins. We conducted a matched case-control study using data from the UK-based Clinical Practice Research Datalink to analyse use of statins among 2469 cases with glioma and 24,690 controls. We performed conditional logistic regression analysis to calculate relative risks, estimated as odds ratios (ORs) with 95 % confidence intervals (CIs) adjusting for multiple confounding factors. As compared with non-use of statins, use of statins was not associated with risk of glioma (OR for ≥90 prescriptions=0.75; 95 % CI 0.48-1.17). Our findings do not support previous sparse evidence of a possible inverse association between statin use and glioma risk.

Two previous case–control studies [6, 7] investigated the association between statin use and risk of glioma and both studies reported an inverse relation between the two (OR 0.72; 95 % CI 0.52–1.00 [6], OR 0.76; 95 % CI 0.59–0.98 [7]) that was of similar magnitude to the one found in our analysis.

We investigated total statin use in relation to glioma risk (Table 2). As compared with no prior use of statins, the adjusted OR for 2–9 prescriptions was 1.01 (95 % CI 0.80–1.28), for 10–29 prescriptions it was 1.00 (95 % CI 0.81–1.22), for 30–59 prescriptions it was 1.11 (95 % CI 0.89–1.38), for 60–89 prescriptions it was 1.10 (95 % CI 0.80–1.52), and for ≥90 prescriptions it was 0.75 (95 % CI 0.48–1.17). When we removed the terms for diabetes and CHF from the multivariate model, the OR became slightly more pronounced (OR for ≥90 vs. no prescriptions of statins = 0.67; 95 % CI 0.43–1.04).

When we examined the associations between individual statins and risk of glioma (Table 2), the results were null for any type of statin. When we investigated timing of statin use, there was no relation between recent or past use and risk of glioma, and there were also no meaningful changes to the results when we stratified by sex or age (<60/≥60 years), or when we restricted cases to glioblastoma patients (data not shown).

Inference – пример (3)

Annals of Internal Medicine®

Letters | 1 January 1993

Calculation Errors in Meta-Analysis

Authors: Andrea Messori, PharmD, Giovanna Scroccaro, PharmD, and Nello Martini, PharmD | AUTHOR, ARTICLE, & DISCLOSURE INFORMATION

Publication: Annals of Internal Medicine • Volume 118, Number 1 • https://doi.org/10.7326/0003-4819-118-1-199301010-00022

The recent paper by Hommes and colleagues (5) reports a meta-analysis of six randomized trials comparing subcutaneous heparin with continuous intravenous heparin for the initial treatment of deep vein thrombosis. To recalculate the odds ratio (with 95% CI), we used a computer program (Messori A. Unpublished observations) developed at our institution that is based on equations 7.18, 7.20, 7.21 (including the correction for zero values), 7.22, and 7.23 (4). The result of our calculation was an odds ratio of 0.61 (95% CI, 0.298 to 1.251; $P > 0.05$); this figure differs greatly from the value reported by Hommes and associates (odds ratio, 0.62; 95% CI, 0.39 to 0.98; $P < 0.05$). When our recalculations involved the odds ratios (with 95% CI) of the six individual studies, we obtained the same identical results as those reported in Table 2 of Hommes' article.

Based on our recalculation of the overall odds ratio, we concluded that subcutaneous heparin is *not* more effective than intravenous heparin, exactly opposite to that of Hommes and colleagues. This could have resulted from a calculation error or because our technique differed from theirs. However, it would

Inference – пример (4)

Original Investigation | Caring for the Critically Ill Patient FREE

February 17, 2019

Effect of a Resuscitation Strategy Targeting Peripheral Perfusion Status vs Serum Lactate Levels on 28-Day Mortality Among Patients With Septic Shock

The ANDROMEDA-SHOCK Randomized Clinical Trial

Glenn Hernández, MD, PhD[1]; Gustavo A. Ospina-Tascón, MD, PhD[2]; Lucas Petri Damiani, MSc[3]; et al

» Author Affiliations | Article Information

JAMA. 2019;321(7):654-664. doi:10.1001/jama.2019.0071

Design, Setting, and Participants Multicenter, randomized trial conducted at 28 intensive care units in 5 countries. Four-hundred twenty-four patients with septic shock were included between March 2017 and March 2018. The last date of follow-up was June 12, 2018.

Interventions Patients were randomized to a step-by-step resuscitation protocol aimed at either normalizing capillary refill time (n = 212) or normalizing or decreasing lactate levels at rates greater than 20% per 2 hours (n = 212), during an 8-hour intervention period.

Main Outcomes and Measures The primary outcome was all-cause mortality at 28 days. Secondary outcomes were organ dysfunction at 72 hours after randomization, as assessed by Sequential Organ Failure Assessment (SOFA) score (range, 0 [best] to 24 [worst]); death within 90 days; mechanical ventilation–, renal replacement therapy–, and vasopressor-free days within 28 days; intensive care unit and hospital length of stay.

Results Among 424 patients randomized (mean age, 63 years; 226 [53%] women), 416 (98%) completed the trial. By day 28, 74 patients (34.9%) in the peripheral perfusion group and 92 patients (43.4%) in the lactate group had died (hazard ratio, 0.75 [95% CI, 0.55 to 1.02]; *P* = .06; risk difference, −8.5% [95% CI, −18.2% to 1.2%]). Peripheral perfusion–targeted resuscitation was associated with less organ dysfunction at 72 hours (mean SOFA score, 5.6 [SD, 4.3] vs 6.6 [SD, 4.7]; mean difference, −1.00 [95% CI, −1.97 to −0.02]; *P* = .045). There were no significant differences in the other 6 secondary outcomes. No protocol-related serious adverse reactions were confirmed.

Conclusions and Relevance Among patients with septic shock, a resuscitation strategy targeting normalization of capillary refill time, compared with a strategy targeting serum lactate levels, did not reduce all-cause 28-day mortality.

Inference – пример (5)

Objective To investigate the effectiveness of therapies for reducing pain and improving quality of life (QOL) in people with fibromyalgia.

Data Sources Searches were performed in the MEDLINE, Cochrane, Embase, AMED, PsycInfo, and PEDro databases without language or date restrictions on December 11, 2018, and updated on July 15, 2020.

Study Selection All published randomized or quasi-randomized clinical trials that investigated therapies for individuals with fibromyalgia were screened for inclusion.

Data Extraction and Synthesis Two reviewers independently extracted data and assessed risk of bias using the 0 to 10 PEDro scale. Effect sizes for specific therapies were pooled using random-effects models. The quality of evidence was assessed using the Grading of Recommendations Assessment (GRADE) approach.

Main Outcomes and Measures Pain intensity measured by the visual analog scale, numerical rating scales, and other valid instruments and QOL measured by the Fibromyalgia Impact Questionnaire.

Results A total of 224 trials including 29 962 participants were included. High-quality evidence was found in favor of cognitive behavioral therapy (weighted mean difference [WMD], −0.9; 95% CI, −1.4 to −0.3) for pain in the short term and was found in favor of central nervous system depressants (WMD, −1.2 [95% CI, −1.6 to −0.8]) and antidepressants (WMD, −0.5 [95% CI, −0.7 to −0.4]) for pain in the medium term. There was also high-quality evidence in favor of antidepressants (WMD, −6.8 [95% CI, −8.5 to −5.2]) for QOL in the short term and in favor of central nervous system depressants (WMD, −8.7 [95% CI, −11.3 to −6.0]) and antidepressants (WMD, −3.5 [95% CI, −4.5 to −2.5]) in the medium term. However, these associations were small and did not exceed the minimum clinically important change (2 points on an 11-point scale for pain and 14 points on a 101-point scale for QOL). Evidence for long-term outcomes of interventions was lacking.

Conclusions and Relevance This systematic review and meta-analysis suggests that most of the currently available therapies for the management of fibromyalgia are not supported by high-quality evidence. Some therapies may reduce pain and improve QOL in the short to medium term, although the effect size of the associations might not be clinically important to patients.

October 26, 2020

Association of Therapies With Reduced Pain and Improved Quality of Life in Patients With Fibromyalgia

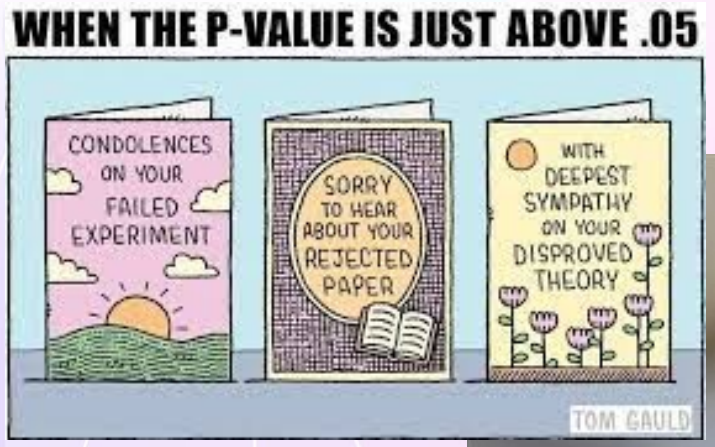
A Systematic Review and Meta-analysis

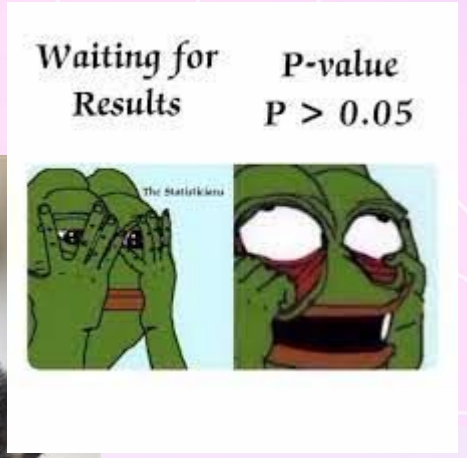
Rodrigo Oliveira Mascarenhas, MSc[1]; Mateus Bastos Souza, BAppSc[2]; Murilo Xavier Oliveira, PhD[2]; et al

» Author Affiliations | Article Information

JAMA Intern Med. 2021;181(1):104-112. doi:10.1001/jamainternmed.2020.5651

$p \geq 0.05$

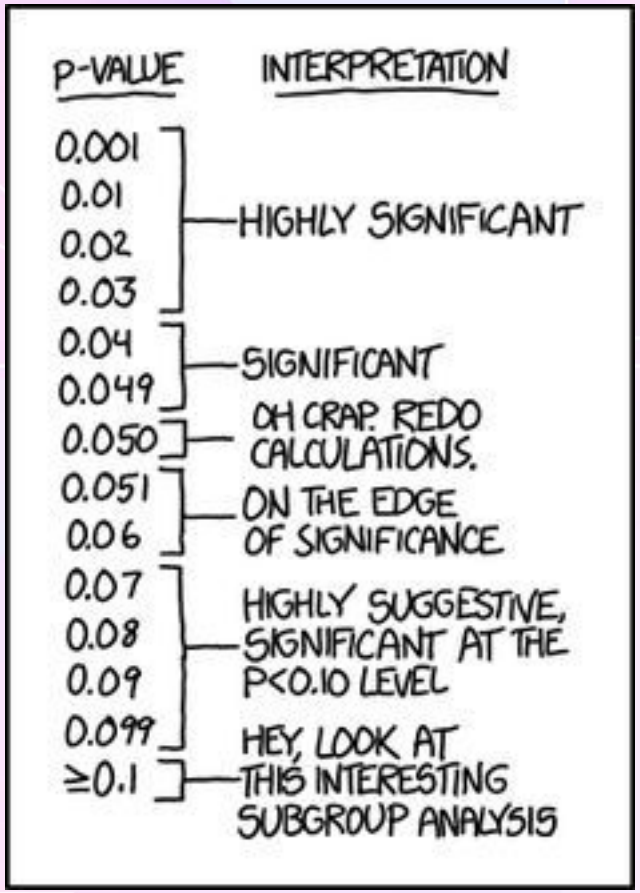

WHEN THE P-VALUE IS JUST ABOVE .05
CONDOLENCES ON YOUR FAILED EXPERIMENT
SORRY TO HEAR ABOUT YOUR REJECTED PAPER
WITH DEEPEST SYMPATHY ON YOUR DISPROVED THEORY
TOM GAULD


Waiting for Results
P-value P > 0.05

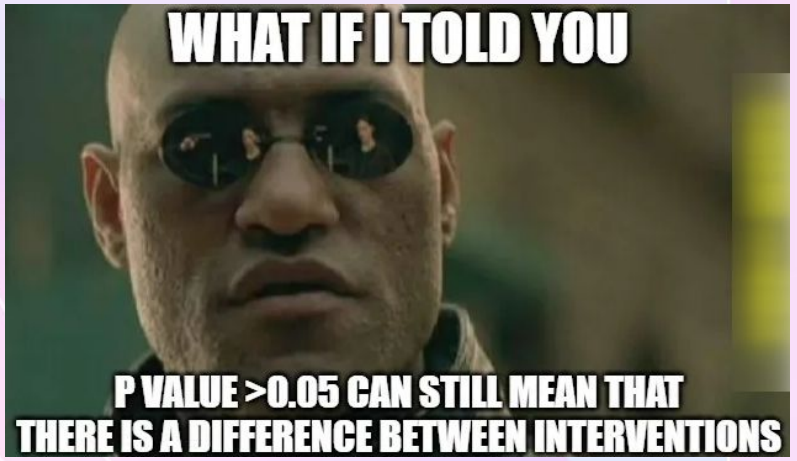

P > 0.05
GAME OVER, TRY AGAIN
imgflip.com


WHAT'S YOUR P-VALUE?
0.051


WHEN YOUR P-VALUE IS > 0.05
SO YOU DESTROY YOU RESEARCH AND START OVER.
makeameme.org


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

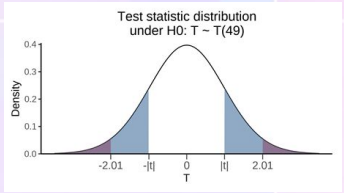
Миф про *p*-value № 7



- $p \geq 0.05$ означает, что эффект отсутствует/ группы сопоставимы (nullism)
- Большое *p* свидетельствует о маленьком размере эффекте

Что не так?

- Само по себе *p*-значение ничего не говорит о размере эффекта – ни того, который мы наблюдаем в выборке ($\hat{\theta} - \theta_0$ или $\hat{\theta}/\theta_0$), ни того, который есть в генеральной совокупности ($\theta - \theta_0$ или θ/θ_0)
- При небольшом объеме выборки и большой вариабельности данных в ней даже большой размер эффекта может оказаться статистически незначимым («сигнал утонет в шуме»)



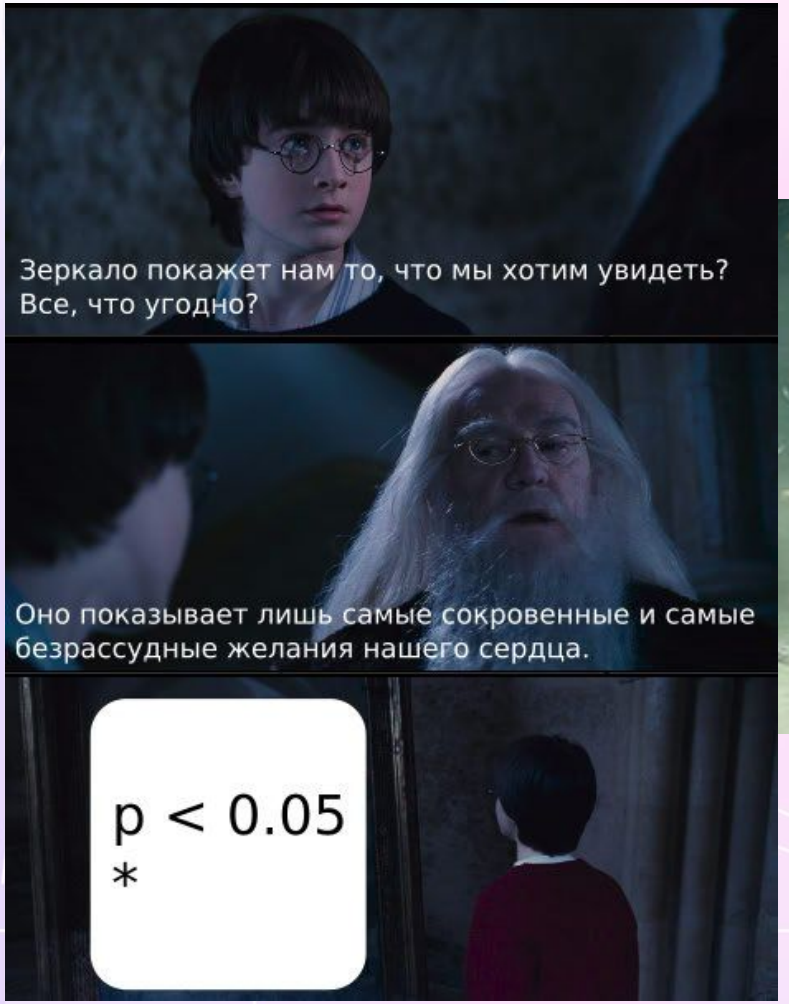
$$T = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}}$$

ICH E9:

Concluding equivalence or non-inferiority based on observing a non-significant test result of the null hypothesis that there is no difference between the investigational product and the active comparator is inappropriate.

A Guide to Misinterpretations

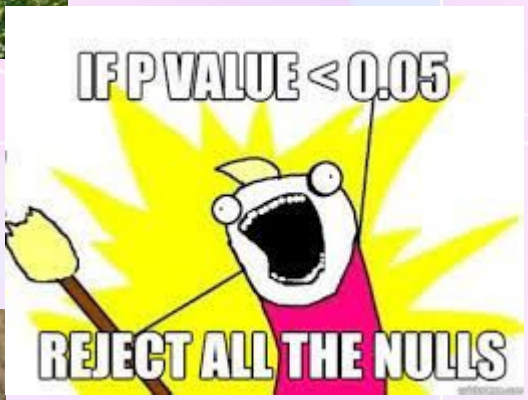
# $p < 0.05$


Зеркало покажет нам то, что мы хотим увидеть?
Все, что угодно?
Оно показывает лишь самые сокровенные и самые безрассудные желания нашего сердца.
p < 0.05
*


THIS IS WHAT
P<.05
FEELS LIKE


WHEN YOUR P-VALUE
IS LESS THAN 0.05


WHEN YOUR P-VALUE IS ACTUALLY LESS THAN .05
SO YOU DON'T EVEN HAVE TO P-HACK


IF P VALUE < 0.05
REJECT ALL THE NULLS

Миф про *p*-value № 8

- $p < 0.05$ свидетельствует об обнаружении важного эффекта
- Маленькое *p* свидетельствует о большом размере эффекте

Что не так?

- Само по себе *p*-значение ничего не говорит о размере эффекта
- При большом объеме выборки даже маленький размер эффекта может оказаться статистически значимым (данные могут быть необычны, с точки зрения статистической модели, но их необычность может не представлять практического (клинического) интереса)
- При небольшом объеме выборки статистически значимый эффект может в действительности быть сильно завышенным (по абсолютной величине) по сравнению с истинным («выброс») - winner’s curse

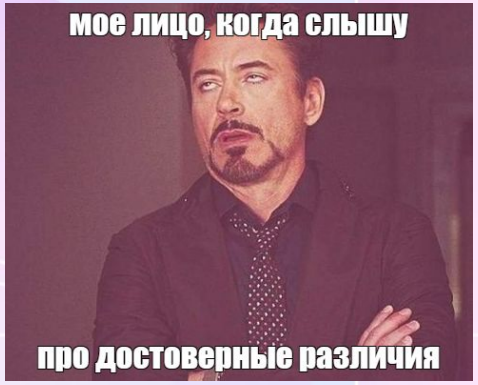
A Guide to Misinterpretations; Benjamin et al. (2018) Redefine Statistical Significance

Миф про *p*-value № 8 – русскоязычное дополнение

- $p < 0.05$ свидетельствует об обнаружении достоверных различий
- $p \geq 0.05$ свидетельствует об отсутствии достоверных различий

Что не так?

- Достоверное событие – это событие с вероятностью 1
- $p < 0.05$, статистическая значимость, свидетельствует о том, что данные слабо согласуются с H0*, но ничего не говорит о том, что эффект / разница точно есть в ГС
- $p \geq 0.05$ свидетельствует о том, что согласуются с H0*, но ничего не говорит о том, что эффект / разница точно отсутствует



Зорин Н.А. (2011) «Достоверность» или «Статистическая значимость» 12 лет спустя

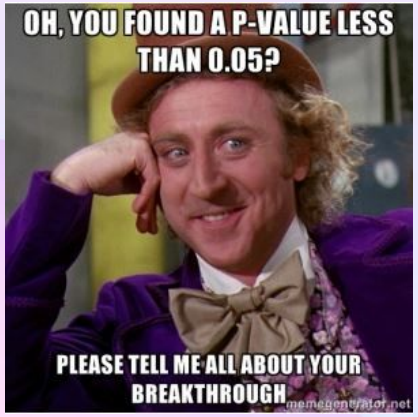
* - при соблюдении допущений статистической модели

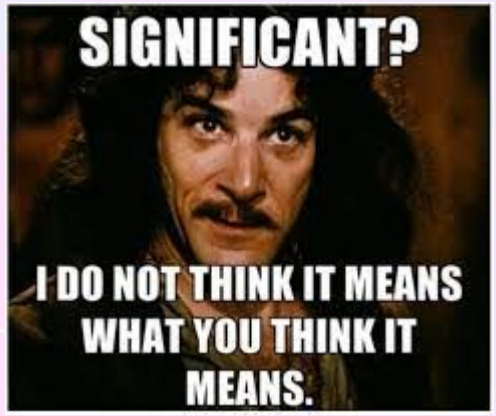
Миф про p-value № 9

- $p < 0.05$ – «обнаружены свидетельства/ доказательства статистической значимости эффекта»
- $p \geq 0.05$ – «статистически значимый эффект не обнаружен»

Что не так?

- Статистическая значимость – это не свойство изучаемого эффекта или ГС, а характеристика результата статистического теста, поэтому «обнаруживать» или «доказывать» статистическую значимость эффекта бессмысленно





Cochrane Handbook for Systematic Reviews of Interventions:

- Review authors should not describe results as 'statistically significant', 'not statistically significant' or 'non-significant' or unduly rely on thresholds for P values, but report the confidence interval together with the exact P value.

A Guide to Misinterpretations

Что есть воспроизводимость результатов?

- Часто под этим понимают воспроизводимость «статистической значимости»
- Dance of the Cis (см. анти-миф про ДИ № 1)
  - На самом деле танцуют все! – *p*-value тоже, и это может быть результатом случайной вариации эффектов между исследованиями (вполне возможно, большей, чем наше представление о sampl***ing*** distribution)
  - If the alternative is correct and the actual power of two studies is 80%, the chance that the studies will both show $P \leq 0.05$ will at best be only $0.80 \cdot 0.80 = 64\%$; furthermore, the chance that one study shows $P \leq 0.05$ and the other does not (and thus will be misinterpreted as showing conflicting results) is $2 \cdot 0.80 \cdot 0.20 = 32\%$*
- «Невоспроизводимость» усугубляется публикационным смещением и смещением репортирования

Supplements to ASA Statement; Supplements to A World beyond $P < 0.05$; Retire Statistical Significance
* - A Guide to Misinterpretations

Миф про ДИ № 4

- Если 95% ДИ пересекаются, различия между группами / исследованиями статистически незначимы на уровне значимости 0.05

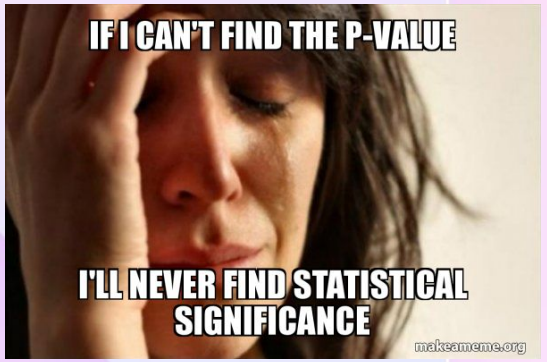
Что не так?

- Для сравнения групп необходимо оценивать ДИ / проверять H0 для параметра, характеризующего различия между группами
  - Если 95% ДИ не пересекаются, получим $p < 0.05$ для H0 о равенстве разницы 0*
  - Если 95% ДИ покрывает точечную оценку параметра для группы сравнения, получим $p > 0.05$ для H0 о равенстве разницы 0*

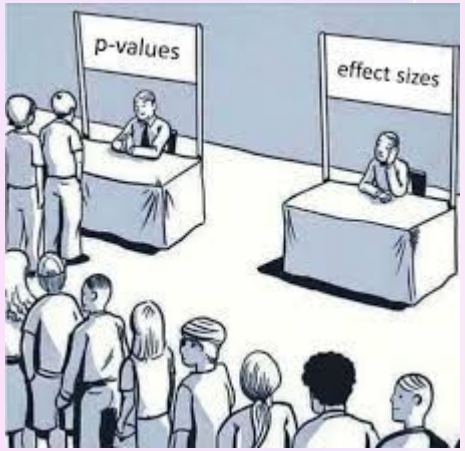
A Guide to Misinterpretations

* - если p оценивается из тех же допущений, что и оба 95% ДИ

Inference – выводы (1)



- **Научные выводы** (scientific inference) – вот что важно для ответа на исследовательский вопрос
- **ДИ – более информативны для inference, чем *p***:
  - По ДИ можно судить о том, какие значения размера эффекта наилучшим образом согласуются с данными* (**compatibility intervals**)
  - ДИ содержат в себе информацию о неопределенности
  - С ДИ невозможно допустить ошибку «отсутствие доказательств эффекта = доказательство отсутствия эффекта»
  - Из ДИ можно получить оценку p, а по p рассчитать ДИ – нет
- Для выводов по результатам исследования необходимы **допущения о MCID – минимальном клинически значимом размере эффекта** (не только для проверки гипотезы)

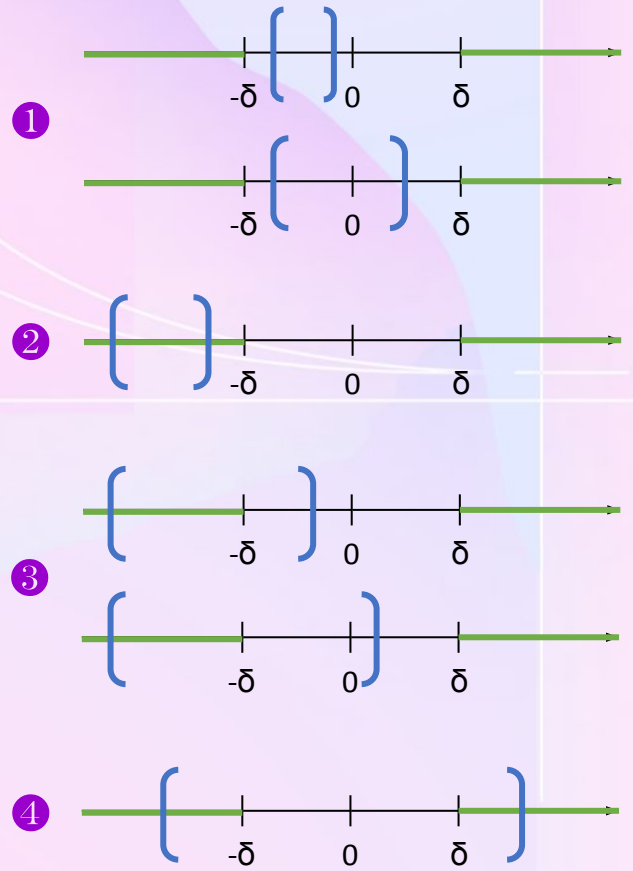


* - при соблюдении допущений статистической модели

A Dirty Dozen; A Guide to Misinterpretations; Retire Statistical Significance; Harrell FE Biostatistics for Biomedical Research

Inference – рекомендации (1)

- Содержит ли ДИ *практически (клинически)* значимые значения эффекта – независимо от того, содержит ли значение, соответствующее отсутствию эффекта (nil)
  - Если таковых в нем нет, можно сделать вывод, что полученные результаты согласуются с отсутствием практически (клинически) значимого эффекта*
- Какие практические выводы можно сделать при различных значениях параметра, попавших в ДИ (особенно при точечной оценке параметра и границах ДИ)
  - Все значения параметра внутри ДИ достаточно хорошо согласуются с данными – выделять среди них какое-либо одно (например, nil, значение θ_0 из H0 или точечную оценку эффекта $\hat{\theta}$) не имеет смысла



δ > 0 : MCID
( ) : ДИ
— : клинически значимые значения

Retire Statistical Significance

* - при соблюдении допущений статистической модели

Inference – рекомендации (2)

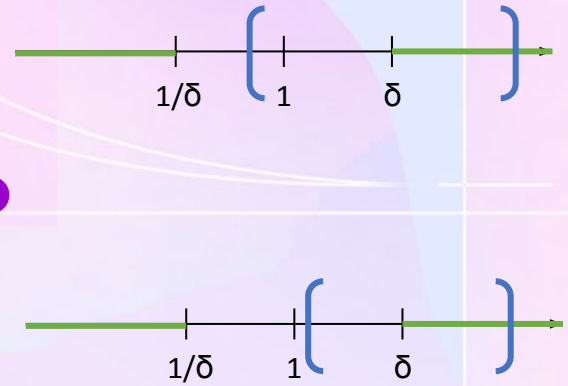
- Значения вне ДИ тоже согласуются с данными – просто в меньшей степени, чем значения внутри него*. Кроме того, значения вне ДИ, близкие к его границам, в практическом смысле мало отличаются от значений внутри ДИ, поэтому **неверно говорить о том, что ДИ включает все возможные значения эффекта**

Retire Statistical Significance

* - при соблюдении допущений статистической модели

Inference – пример (1)

- RR = 1.2, 95% ДИ 0.97-1.48, p = 0.091
  - «Результаты нашего исследования указывают на 20%-ное увеличение риска. Тем не менее, разница рисков, варьирующаяся от 3%-ного снижения (небольшая отрицательная взаимосвязь) до 48%-ного увеличения (существенная положительная взаимосвязь), также вполне совместима с нашими данными, с учетом сделанных нами предположений»*
  - Если бы 95% ДИ был бы 1.01-1.52, вывод строился бы аналогичным образом
  - Дальше можно обсудить, почему разница может быть незначительной, а также то, насколько полученные результаты и выводы зависят от ограничений исследования и какие возможны издержки в связи с такими выводами для пациентов**



δ > 1: MCID
( ): ДИ
—: клинически значимые значения

* - Retire Statistical Significance

** - Amrhei et al. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication

Еще примеры интерпретаций (неправильных и правильных) – см. Language for communicating frequentist results about treatment effects (+ комментарии)

Inference – пример (2)

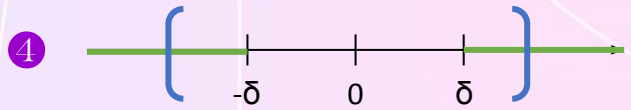
- 95% CI for the ratio of resolution rates for ivermectin vs. placebo extends from 0.87 to 1.32, which means that every rate ratio inside the interval has $p > 0.05$. Thus, using the conventional 5% significance cutoff adopted in the paper, we could say **that the results are most compatible with or provide little evidence against anything from a 13% lower to a 32% higher symptom-resolution rate** from the ivermectin treatment protocol used in the trial.
- To translate the statistics into qualitative clinical terms, we'd have to agree on what percent improvement would be considered "**clinically significant**".
  - If that were 25% and we accepted the traditional 0.05 criterion for deciding how to report this trial, then the results were indecisive: The *P*-value for a 25% improvement is 0.14, very compatible with the data, while even 30% higher and a 10% lower recovery rate for ivermectin also have $p > 0.05$ and thus are reasonably compatible with the data by the 0.05 convention.

❸

Greenland et al. (2022). To curb research misreporting, replace significance and confidence by compatibility: A Preventive Medicine Golden Jubilee article

Inference – пример (3)

- Широкий ДИ, включающий значения, совместимые как с отрицательным, так и с положительным клинически значимым эффектом
  - «Данных нашего исследования оказалось недостаточно, чтобы сделать какие-либо уверенные выводы о частоте нежелательных явлений со стороны почек при сравнении ибупрофена с парацетамолом; для ответа на вопрос о безопасности ибупрофена необходимы дополнительные данные»*



δ > 0 : MCID
( ) : ДИ
— : клинически значимые значения

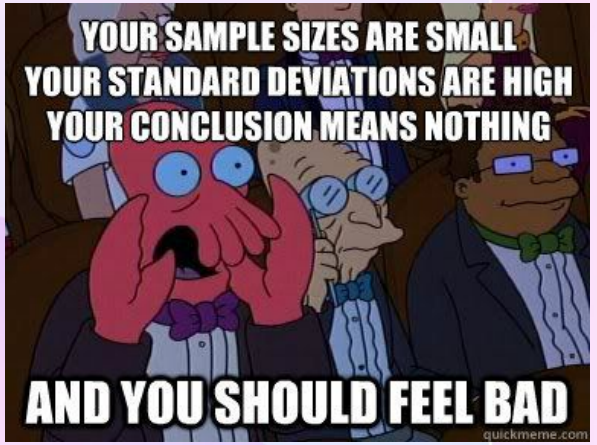
* - Greenland S. (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values

Inference – рекомендации (3)

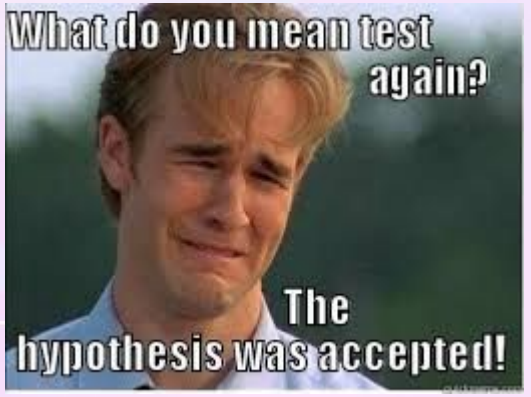
- Можно задавать (заранее) **другой уровень доверия**, помимо «общепринятого» 95%, как и выбирать **другое значение** α, помимо 0.05, при проверке гипотезы – в зависимости от доменной области и исследуемого вопроса («цена ошибки»)
- Оценка согласованности зависит от корректности допущений статистической модели, используемой для оценки ДИ и *p*. На практике эти предположения, в лучшем случае, сопровождаются значительной неопределенностью. **Опишите эти допущения** как можно более понятным образом и **проверьте** те, которые можно (графики, оценка альтернативных моделей, анализ чувствительности), а затем **сообщите все результаты**
  - ДИ и *p* невозможно интерпретировать корректным образом без информации о том, какие гипотезы проверялись и в каком количестве, как данные были получены и предобработаны и каким образом отбирались результаты для итоговых выводов
  - ~~Cherry picking, data dredging, significance chasing, significance questing, p-hacking, HARKing, selective inference/ reporting, multiple testing~~

ASA Statement
Retire Statistical Significance

Inference – рекомендации (4)



- С точки зрения дальнейших исследований:
  - ❶ Узкий ДИ, не содержащий клинически значимых размеров эффекта ⇒ полученные результаты согласуются с отсутствием клинически значимого эффекта* ⇒ возможно, повтор нерационален
  - ❹ Широкий ДИ, совместимый с клинически значимыми отрицательными и положительными эффектами ⇒ не можем сделать уверенных выводов* ⇒ “get more data”
  - ❷❸ ДИ, совместимый с клинически значимым позитивным эффектом* ⇒ “repeat the experiment”
    - Особенно при небольшом размере выборки



A Dirty Dozen
Retire Statistical Significance

* - при соблюдении допущений статистической модели

Inference – выводы (2)

Проверка H0 (NHST) и вывод о *статической* значимости («позитивное» / «негативное» исследование) – не для научных выводов и принятия решений о публикации / продолжения исследований

- “It seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an “uncertainty laundering” that begins with data and concludes with success as measured by statistical significance.”*
- “A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.”, “As “statistical significance” is used less, statistical thinking will be used more.”**
- “Misleading use of P-values is so easy and automated that, especially when rewarded with publication and funding, it can become addictive.”***

Retire Statistical Significance; Supplements to ASA Statement; Supplements to A World beyond P < 0.05;
* - The Problems With P-Values are not Just With P-Values (Gelman A); ** - Moving to a World Beyond “p < 0.05.” (Wasserstein et al);
*** - Fit-for-Purpose Inferential Methods: Abandoning/Changing P-values Versus Abandoning/Changing Research (Ioannidis JPA)

Inference – выводы (2)+

Замечание 2. *Часто статистически незначимые результаты несправедливо рассматриваются исследователями как неудача работы, что подчеркивается такими распространенными выражениями как "отрицательный результат", "не удалось достичь статистической значимости". Такие исследования часто не публикуются, что приводит к возникновению систематической ошибки, обусловленной преимущественным опубликованием положительных результатов исследования, когда опубликованными оказываются лишь те исследования, в которых результаты оказались статистически значимыми, т.е. лишь часть от всех выполненных исследований по данной проблеме. Вместе с тем статистически незначимые результаты являются не менее важными в контексте общенаучного процесса.*

Замечание 1. *К сожалению, до настоящего времени очень часто вместо термина "статистически значимый" в отечественных публикациях ошибочно используется термин "достоверный", имеющий в статистике другой смысл.*

Inference – выводы (3)

- Проверка H0 (NHST) и вывод о *статической* значимости («позитивное» / «негативное» исследование) – прежде всего, для принятия решений (например, о регистрации препарата) в условиях (жестко) регулируемой среды
  - H0 должна быть научно обоснована
  - H0 необязательно должна быть о равенстве эффекта нулю. Гипотезы эквивалентности, неменьшей эффективности, превосходства с учетом минимального клинически значимого эффекта более полезны
  - Также полезно оценить p (совместимость данных с) H0 с размером эффекта, использованном в альтернативной гипотезе планирования

- α необязательно принимать на «общепринятом» уровне 0,05 – лучше это делать (заранее) в зависимости от доменной области и исследуемого вопроса
  - α фиксируется заранее (при $\alpha = 0.05$ что-либо не может быть статистически значимым на 10%-ном уровне или «почти значимым»)
- $p < 0.05$ – не единственный критерий для принятия решений

ASA Statement; A Guide to Misinterpretations; Retire Statistical Significance

Retire statistical significance – «общество еще не готово»

nature

Explore content | About the journal | Publish with us | Subscribe

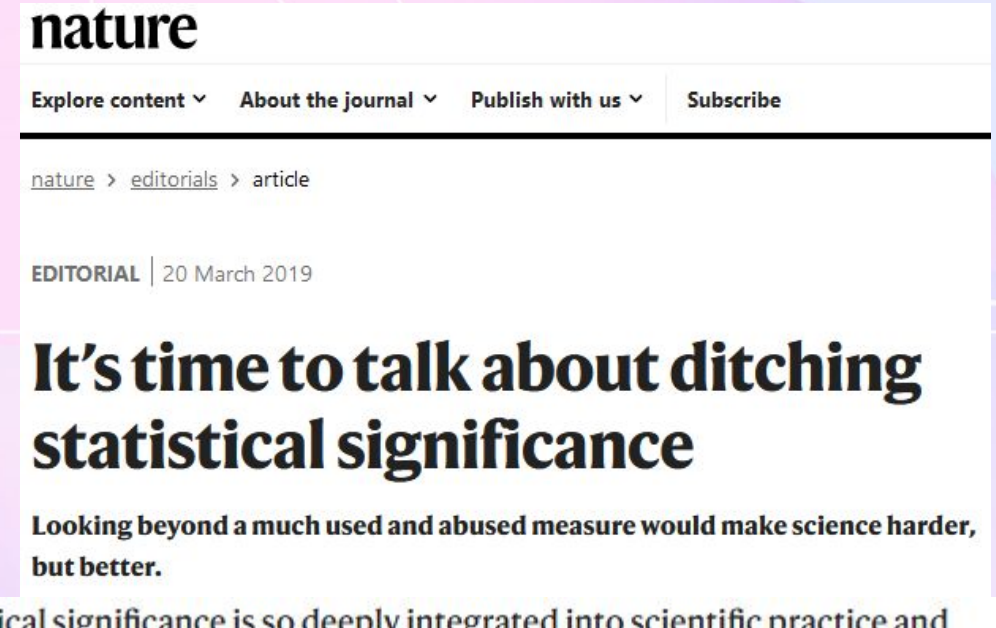
nature › comment › article

COMMENT | 20 March 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

By Valentin Amrhein, Sander Greenland & Blake McShane



nature

Explore content | About the journal | Publish with us | Subscribe

nature › editorials › article

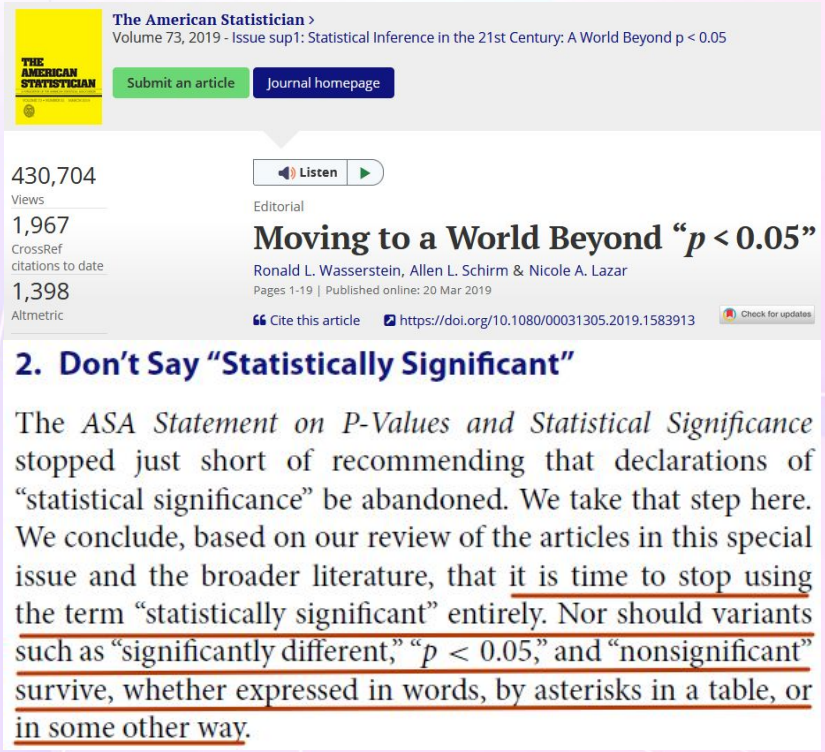
EDITORIAL | 20 March 2019

It's time to talk about ditching statistical significance

Looking beyond a much used and abused measure would make science harder, but better.

Statistical significance is so deeply integrated into scientific practice and evaluation that extricating it would be painful. Critics will counter that arbitrary gatekeepers are better than unclear ones, and that the more useful argument is over which results should count for (or against) evidence of effect. There are reasonable viewpoints on all sides; *Nature* is not seeking to change how it considers statistical analysis in evaluation of papers at this time, but we encourage readers to share their views (see go.nature.com/correspondence).

Abandon statistical significance – «общество еще не готово»



The American Statistician ›
Volume 73, 2019 - Issue sup1: Statistical Inference in the 21st Century: A World Beyond p < 0.05

Submit an article | Journal homepage

430,704 Views
1,967 CrossRef citations to date
1,398 Altmetric

Listen

Editorial

Moving to a World Beyond "$p < 0.05$"

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Pages 1-19 | Published online: 20 Mar 2019

Cite this article | https://doi.org/10.1080/00031305.2019.1583913

2. Don't Say "Statistically Significant"

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

September 2021

The ASA president's task force statement on statistical significance and replicability

Yoav Benjamini, Richard D. De Veaux, Bradley Efron, Scott Evans, Mark Glickman, Barry I. Graubard, Xuming He, Xiao-Li Meng, Nancy Reid, Stephen M. Stigler, Stephen B. Vardeman, Christopher K. Wikle, Tommy Wright, Linda J. Young, Karen Kafadar

Author Affiliations +

Ann. Appl. Stat. 15(3): 1084-1085 (September 2021). DOI: 10.1214/21-AOAS1501 IF: 1.3 Q2

Over the past decade, the sciences have experienced elevated concerns about replicability of study results. An important aspect of replicability is the use of statistical methods for framing conclusions. In 2019 the President of the American Statistical Association (ASA) established a task force to address concerns that a 2019 editorial in *The American Statistician* (an ASA journal) might be mistakenly interpreted as official ASA policy. (The 2019 editorial recommended eliminating the use of "$p < 0.05$" and "statistically significant" in statistical analysis.) This document is the statement of the task force, and the ASA invited us to publicize it. Its purpose is two-fold: to clarify that the use of P-values and significance testing, properly applied and interpreted, are important tools that should not be abandoned, and to briefly set out some principles of sound statistical inference that may be useful to the scientific community.

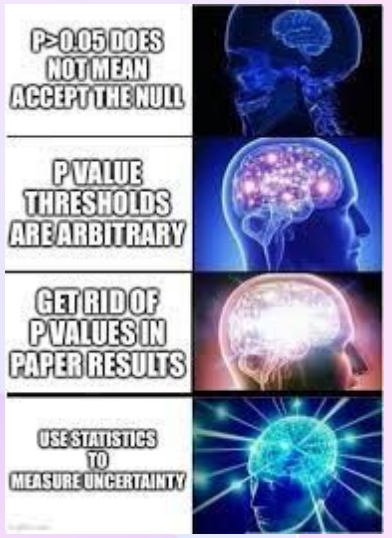
P-values are valid statistical measures that provide convenient conventions for communicating the uncertainty inherent in quantitative results. Indeed, P-values and significance tests are among the most studied and best understood statistical procedures in the statistics literature. They are important tools that have advanced science through their proper application.

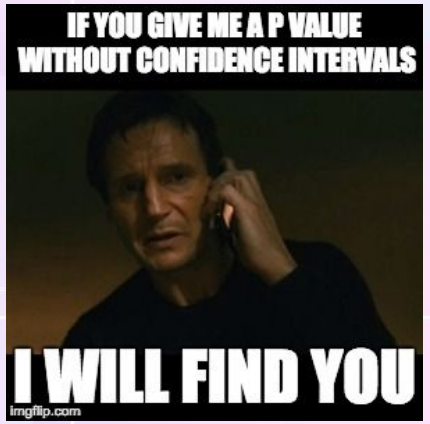
Much of the controversy surrounding statistical significance can be dispelled through a better appreciation of uncertainty, variability, multiplicity, and replicability. The following general principles underlie the appropriate use of P-values and the reporting of statistical significance and apply more broadly to good statistical practice.

Mayo D.: "…it is just implausible to suggest that refraining from talk about statistical significance will appreciably help overcome mechanical decision-making in statistical practice, and lead to a greater engagement with statistical thinking. Such an outcome will require, among other things, the implementation of science education reforms that centre on the conceptual foundations of statistical inference."

Общие рекомендации

- В приоритете – научный вывод, ответ на исследовательский вопрос, статистический вывод – только инструмент
- Качество вывода зависит от ответственного подхода к дизайну исследования, выбору статистической модели, формулировке допущений
- Не отказываемся от *p*-значений – отказываемся от некорректных интерпретаций
- «Читаем» и интерпретируем ДИ (клинически значимые значения параметра?)
- “Repeat” + “get more data”
- Признаем, что иногда статистика бессильна, в деле превращения garbage в конфетку – возможно, всегда





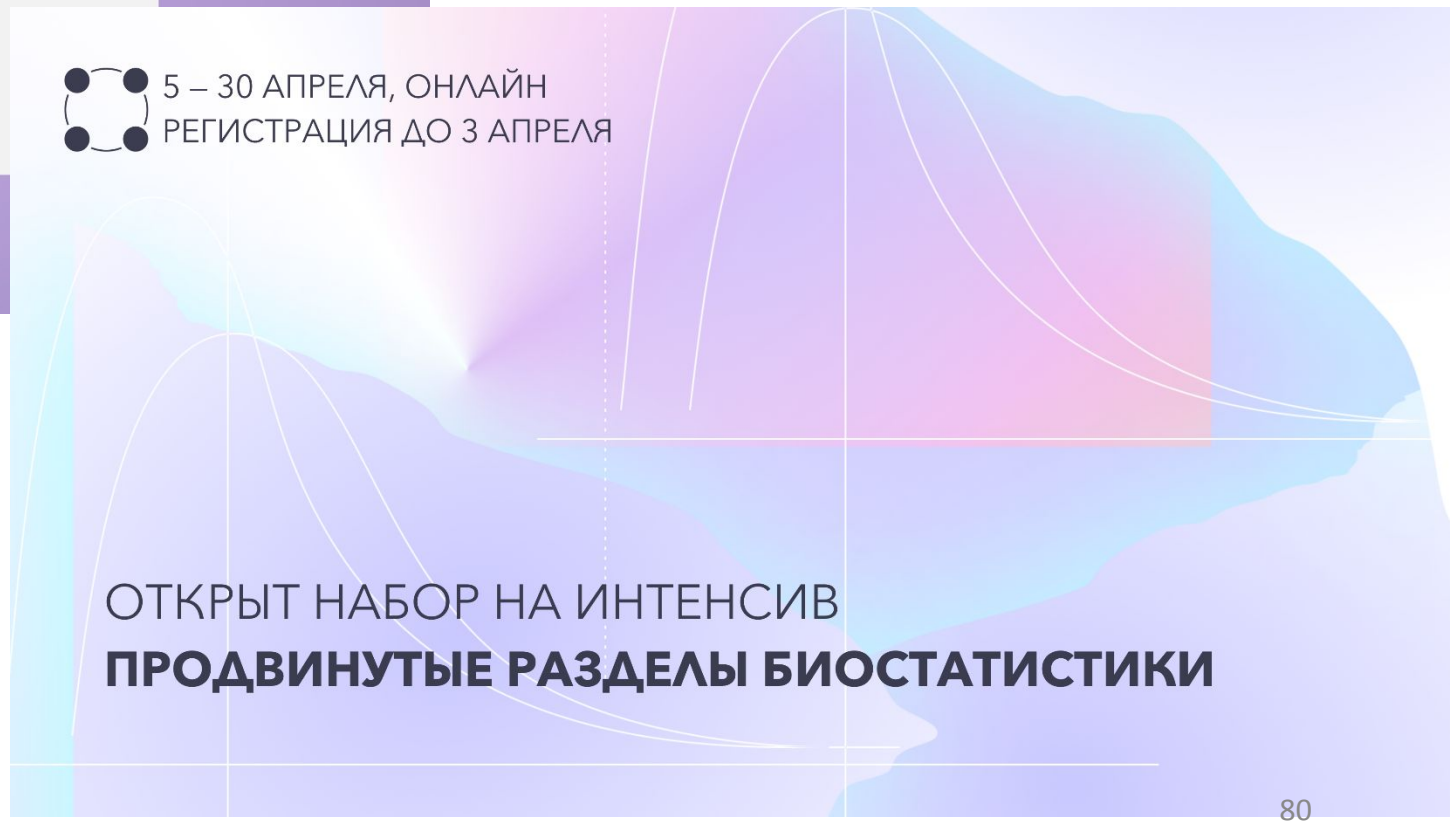
10 АПРЕЛЯ 19:00 МСК
ОНЛАЙН

РАЗРУШИТЕЛИ СТАТИСТИЧЕСКИХ МИФОВ

МАТВЕЙ СЛАВЕНКО | МИФ №3:
НЕНОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ ТРЕБУЕТ НЕНОРМАЛЬНЫХ РЕШЕНИЙ

bioinf.institute/hack2025

bioinf.me/education/workshops/hardstat



Институт биоинформатики в социальных сетях

Разрушители статистических мифов: bioinf.me/stat_myths

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