DOI: https://doi.org/10.17816/humeco97249

Интерпретация величины *p* и альтернативы её использованию в биомедицинских исследованиях

А.М. Гржибовский[1,2,3,4], А.Н. Гвоздецкий[5]

[1] Северный государственный медицинский университет, г. Архангельск, Российская Федерация
[2] Казахский национальный университет имени Аль-Фараби, г. Алматы, Казахстан
[3] Западно-Казахстанский медицинский университет имени Марата Оспанова, г. Актобе, Казахстан
[4] Северо-восточный федеральный университет имени М.К. Аммосова, г. Якутск, Российская Федерация
[5] Северо-Западный государственный медицинский университет имени И.И. Мечникова Минздрава России, г. Санкт-Петербург, Российская Федерация

АННОТАЦИЯ

Существенные проблемы с интерпретацией результатов статистического анализа в биомедицинских исследованиях часто упоминают в литературе в качестве одной из причин кризиса воспроизводимости научных результатов. Закономерно появились предложения по улучшению ситуации, в том числе за счёт полного отказа от представления величины *p* в публикациях.

В настоящей работе рассмотрены причины сложившейся ситуации в контексте исторически различных подходов к проверке статистических гипотез и представлены альтернативы использованию значения *p* — доверительные интервалы и величина эффекта. Приведены аргументы «за» и «против» высказываемого в зарубежных источниках литературы предложения по изменению критического уровня значимости с 0,05 до 0,005. Для профилактики ошибочной интерпретации результатов статистического анализа сформирован список наиболее популярных заблуждений о смысле величины *p*, которые разбираются в ведущих журналах по статистике.

В статье предложены практические рекомендации для молодых учёных, следование которым может существенно сократить случаи некорректной интерпретации результатов статистического анализа в биомедицинских исследованиях.

Ключевые слова: величина *p*; уровень значимости; величина эффекта; доверительный интервал; биомедицинские исследования; статистический анализ.

Как цитировать:
Гржибовский А.М., Гвоздецкий А.Н. Интерпретация величины *p* и альтернативы её использованию в биомедицинских исследованиях // Экология человека. 2022. Т. 29. № 3. С. 209–218. DOI: https://doi.org/10.17816/humeco97249

Рукопись получена: 24.01.2022 **Рукопись одобрена:** 25.01.2022 **Опубликована:** 14.06.2022

ЭКО•ВЕКТОР

DOI: https://doi.org/10.17816/humeco97249

Interpretation of and alternatives to *p*-values in biomedical sciences

Andrej M. Grjibovski[1,2,3,4], Anton N. Gvozdeckii[5]

[1] Northern state medical university, Arkhangelsk, Russian Federation
[2] Al-Farabi Kazakh national university, Almaty, Kazakhstan
[3] West Kazakhstan Marat Ospanov medical university, Aktobe, Kazakhstan
[4] North-Eastern federal university, Yakutsk, Russian Federation
[5] Mechnikov North-Western state medical university, St. Petersburg, Russian Federation

ABSTRACT

Existing difficulties in interpretation of the results of statistical analysis have been repeatedly mentioned as one of the factors behind poor reproducibility of research findings in biomedical sciences followed by a series of publications presenting alternatives to improve the situation including a abandonment of *p*-values and significance testing. In this paper we briefly present the scope of the problem as well as Fischer and Neyman–Pearson approaches to hypothesis testing. Moreover, we present confidence intervals and effect size calculation as alternatives to dichotomization of the results as significant or not significant using a certain cut-off level. In addition, we summarize the pros and cons of suggestion to change the cut-off value from traditional 0.05 to 0.005. We also present a list of the most common misunderstandings of *p*-values discussed in international statistical literature.

We conclude the paper with brief recommendations on careful interpretation of the results of statistical analysis to prevent misinterpretation and misuse of *p*-values in biomedical studies.

Keywords: *p*-value; significance level; effect size; confidence interval; biomedical research; statistical analysis.

To cite this article:
Grjibovski AM, Gvozdeckii AN. Interpretation of and alternatives to p-values in biomedical sciences. *Ekologiya cheloveka (Human Ecology).* 2022;29(3): 209–218. DOI: https://doi.org/10.17816/humeco97249

Received: 24.01.2022 **Accepted:** 25.01.2022 **Published:** 14.06.2022

ВВЕДЕНИЕ

Развитие персональной компьютерной техники обусловило широкое внедрение методов статистического анализа в естественные и гуманитарные науки, такие как психология, медицина, биология, социология и даже философия [1]. Появление технической возможности анализа эмпирических данных привело к ускоренному накоплению знаний, полученных из результатов исследований, в различных областях науки. Формирование клинических рекомендаций, создание сложного диагностического оборудования, разработка психометрических тестов были бы невозможны без доступных средств анализа данных. Вместе с тем закономерно образовалась проблема некорректного и местами даже нецелевого использования статистического инструментария, которая рассматривается в качестве одной из причин кризиса воспроизводимости в науках о здоровье [2–4]. Воспроизводимость результатов вне зависимости от места их получения является весомым аргументом при обосновании той или иной теоретической концепции. Множество проблем связано не со сложностью применяемых методов анализа, а с некорректной, часто необоснованно оптимистичной интерпретацией результатов. Как в зарубежной, так и в отечественной литературе встречаются неверные подходы к интерпретации полученных результатов и, как следствие, ошибочные выводы [5–8]. В отечественной литературе используются слова «значимость», «достоверность» без должной академической строгости [9]. Из-за масштабов проблемы некорректного понимания и избыточного использования величины *p* в зарубежной статистической литературе стали появляться предложения по изменению подходов к планированию исследований, статистическому анализу данных, их интерпретации, а также написанию и рецензированию научных работ [1, 10] вплоть до полного запрета на использование *p*-значения [11]. В русскоязычной литературе, особенно биомедицинской, данная проблема обсуждается крайне редко, поскольку отечественные биомедицинские исследования не входят в число наиболее востребованных в международном научном сообществе.

Мы представляем теоретические рассуждения по данной проблеме, а также предлагаем пути решения с целью улучшения методологического качества отечественных научных публикаций.

ПАРАДИГМА ПРОВЕРКИ НУЛЕВОЙ ГИПОТЕЗЫ

Подход Фишера — проверка значимости нулевой гипотезы

Тестирование гипотезы с использованием подхода Фишера можно разложить на несколько шагов [12]:

- шаг 1 — выбрать необходимый статистический критерий, который соответствует исследовательскому вопросу и имеющимся данным;
- шаг 2 — определиться с нулевой гипотезой;
- шаг 3 — в зависимости от выбранных теоретических допущений рассчитать вероятность получения наблюдаемых результатов относительно нулевой гипотезы.

В ходе расчётов получается *p*-значение — вероятность того, что наблюдаемая статистика или её более экстремальные значения извлечены из такого распределения, которое соответствует нулевой гипотезе [13]. Следует обратить внимание, что это не точечное значение, а кумулятивная сумма вероятностей от наименьшего значения до наблюдаемой границы [12]. Данное значение численно равно площади под кривой распределения (рис. 1). При таком ходе рассуждений можно утверждать, что величина *p*-значения является количественной (вероятностной) мерой доказательства против нулевой гипотезы [12], иными словами, это вероятность обнаружить выявленные или ещё более выраженные различия, если их на самом деле не существует. В итоге исследователю необходимо

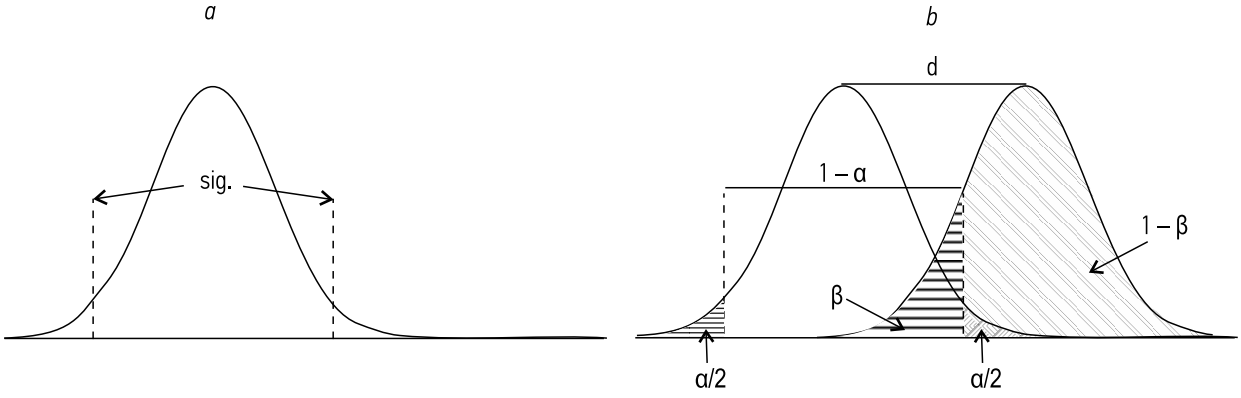

Рис. 1. Графическое сравнение подходов к тестированию статистических гипотез: a — тестирование значимости по Фишеру; b — тестирование принятия по Нейману–Пирсону; sig. — уровень значимости, d — величина эффекта, α — вероятность ошибки первого рода, β — вероятность ошибки второго рода.
Fig. 1. Graphical comparison of approaches to statistical hypothesis testing: a — Fisher significance testing; b — Neymann–Pearson acceptance testing; sig. — significance level, d —effect size, α — probability of Type I error, β — probability of Type II error.

определиться, принять или отклонить нулевую гипотезу. Порог принятия решений остаётся на усмотрение исследователя, но в большинстве биомедицинских статей принимается равным 0,05, что достаточно регулярно подвергается критике в специализированных изданиях.

Подход Неймана–Пирсона — механизм принятия решений

Несмотря на логичность подхода Фишера к тестированию статистической гипотезы, было выполнено множество исследований, в ходе которых предлагались различные улучшения. В научную практику вошёл принцип тестирования нулевой гипотезы Неймана–Пирсона. Основное его отличие от подхода Фишера — чётко проговариваемая альтернативная гипотеза [14], в качестве которой выступает распределение с параметрами, отличными от изучаемого.

Другое существенное отличие от подхода Фишера — принцип контроля частоты ошибок [14]. Напомним, что различают два рода ошибок, относящихся к проверке статистических гипотез:

- ошибка первого рода (type I error, или альфа-ошибка) характеризуется отклонением верной нулевой гипотезы, т.е. когда исследователь делает заключение о том, что различия есть, а на самом деле их нет;
- ошибка второго рода (type II error, бета-ошибка) — это принятие ложной нулевой гипотезы, т.е. исследователь делает вывод о том, что различий нет, а на самом деле они есть.

В практическом смысле эти ошибки эквивалентны понятиям ложноположительного и ложноотрицательного результата соответственно. Контроль заключается в наличии заранее определённых величин ошибок, которые будут лежать в основе принятия решений. Так, число альфа (α) является вероятностью совершить ошибку первого рода, а число бета (β) — вероятностью совершить ошибку второго рода. Значение (1–β) называется статистической мощностью. В сущности мощность — это вероятность корректного отклонения нулевой гипотезы в пользу альтернативной [15]. Простыми словами — это вероятность отклонить нулевую гипотезу (сделать заключение о наличии различий, если они на самом деле есть). В процессе планирования исследований важно держать под контролем вероятность альфа- и бета-ошибок. Для этого рассчитывается необходимый объём выборки. Способ принятия решений по Нейману–Пирсону отражён в табл.

Данный подход прекрасно объясняется «житейскими» примерами из различных областей жизнедеятельности.

В качестве первого примера рассмотрим процесс принятия решения при проведении судебного разбирательства. По умолчанию (нулевая гипотеза) считается, что человек невиновен (презумпция невиновности). Альтернативной гипотезой для невиновности является виновность подсудимого. Обвинению необходимо предоставить доказательства против невиновности подсудимого. В идеале доказательств должно быть достаточно, чтобы изменить первоначальное представление о невиновности, тем самым отклонив исходное допущение, т.е. нулевую гипотезу. Если доказательств окажется достаточно для подтверждения вины подсудимого (хотя в действительности он не совершал преступления), будет совершена ошибка первого рода, или альфа-ошибка. Если же доказательств о виновности подсудимого недостаточно, то он не будет признан виновным, даже совершив преступление, что является ошибкой второго рода, или бета-ошибкой.

Разберём другой пример. Для обеспечения транспортной безопасности в аэропорту необходимо проходить через арочный металлодетектор. По умолчанию рамка не издаёт никаких звуков, но, если в её зону попадает металлический предмет, генерируется звуковой сигнал. Кроме корректного звукового сигнала в случае наличия металлического предмета и корректного отсутствия сигнала в случае отсутствия металлического предмета могут быть ещё два нежелательных, т.е. ошибочных исхода. Рамка может издать звуковой сигнал в случае, когда у пассажира нет никаких металлических предметов (альфа-ошибка). Возможно также отсутствие реакции металлодетектора на проносимый металлический предмет (бета-ошибка).

В качестве третьего примера обсудим процесс диагностики изучаемого медицинского состояния с помощью интересующего нас диагностического маркёра. На основании исходных знаний о патогенезе данного состояния допускаем, что при отсутствии изучаемого состояния у человека нет интересующего нас маркёра. Обнаружение маркёра является свидетельством в пользу наличия заболевания. Как и в предыдущих примерах, возможны следующие ситуации: первый вариант — биомаркёр обнаружен, несмотря на отсутствие изучаемого состояния

Таблица. Взаимосвязь гипотез и ошибок при принятии решения
Table. Relationship between hypotheses and errors in decision-making

Принимаемое решение Decision made	Истинное состояние / True state	
	нулевая гипотеза верна / true null hypothesis	нулевая гипотеза ошибочна / false null hypothesis
Принимаем нулевую гипотезу	Корректное решение	Ошибка второго рода
Отклоняем нулевую гипотезу	Ошибка первого рода	Корректное решение

DOI: https://doi.org/10.17816/humeco77320

(альфа-ошибка); второй вариант — биомаркёр не выявлен, но имеется чёткая клиническая картина, соответствующая изучаемому состоянию (бета-ошибка). Можно привести также и другие примеры.

КОНЦЕПТУАЛЬНАЯ ПРОБЛЕМА *p*-ЗНАЧЕНИЯ И ПУТИ ЕЁ РЕШЕНИЯ

Согласно Рональду Фишеру, в основе тестирования статистической гипотезы лежит доказательство от противного и *p*-значение оценивает силу доказательства против нулевой гипотезы в одном исследовании. Таким образом, p-значение не подразумевает частотной интерпретации и относится только к наблюдаемому набору данных. Однако величину ошибки первого рода целесообразно установить заранее. Этому есть частотное объяснение: при большом количестве проверок гипотез с использованием данных, извлечённых из одной и той же генеральной совокупности, истинная нулевая гипотеза ошибочно будет отклонена в некотором проценте случаев. В таком случае *p*-значение не является силой доказательства против нулевой гипотезы: она или верна, или нет [12]. Желание контролировать ошибку второго рода тоже понятно, так как отсутствие результата из-за нехватки достаточного набора наблюдений — это неэффективное расходование ресурсов, включая временные. При использовании подхода Неймана–Пирсона заранее определяются допустимые уровни альфа- и бета-ошибок, а также те различия, которые мы считаем важными с практической точки зрения. На основании значений ошибок и ожидаемых различий рассчитывается необходимый объём выборки, на которой и выполняется исследование. В этом случае расчёт достигнутого уровня значимости не нужен, так как он будет меньше критического при выявлении ожидаемых или более выраженных различий.

Графическое различие между двумя подходами показано на рис. 1. Подход Фишера можно описать как частный случай подхода Неймана–Пирсона, в котором значительное количество факторов (ошибки, величина эффекта) не контролируются. В настоящее время наблюдается смешение вышеупомянутых подходов. Сложившаяся ситуация может приводить к ошибкам на этапе планирования, а также вести к некорректной интерпретации результатов, что в свою очередь даёт повод усомниться в качестве исследования и целесообразности использования его результатов в практической деятельности. Специалистами предлагались различные пути решения проблемы. Мы кратко остановимся только на некоторых.

Доверительный интервал

Несмотря на то, что доверительный интервал и *p*-значение имеют тесную связь, они несут различную информацию [16]. Доверительный интервал по величине противоположен вероятности ошибки первого рода (1–β). Из этого следует, что в нём не содержатся значения, которые мы бы хотели отклонить при заданном уровне альфа-ошибки. Это его роднит с точечными оценками, которые основаны на вычислении *p*-значения. Вместе с тем доверительный интервал позволяет оценить ожидаемый размер эффекта, что явно более информативно. В сущности доверительный интервал подразумевает, что если выборки взяты из одной и той же генеральной совокупности (популяции) с использованием одинакового метода извлечения (сбора) данных, то заданный процент их доверительных интервалов будет включать истинное значение интересующего параметра.

В качестве типового примера приведём оценку среднего значения для нормального распределения с использованием не менее типового 95% доверительного интервала (1–0,05=0,95). При проведении единичного эксперимента вычисляются среднее арифметическое значение и доверительный интервал. При повторе эксперимента бесконечно большое количество раз 95% вновь вычисленных доверительных интервалов будут включать искомое значение. Демонстрация данного принципа приведена на рис. 2. При помощи генератора случайных чисел 100 раз создавалась выборка размером в 100 наблюдений со средним значением 0, стандартным отклонением 1. В данном простом симуляционном эксперименте оказалось, что 6% значений не содержат ожидаемого среднего.

Тем не менее доверительный интервал не является панацеей, так как его интерпретация не всегда корректна. В литературе часто объясняют доверительный интервал как вероятность того, что интересующий параметр будет принимать значение в заданных границах 95% времени [14]. Для примера среднего значения некорректная, но часто встречающаяся трактовка звучит так: 95% доверительного интервала среднего значения, вычисленного по выборке, включает среднее значение совокупности (популяции) с вероятностью 95%. Ошибка связана с расчётом доверительного интервала на реальных данных (10, 100, 1000 наблюдений и т.д.), которые не равны общей совокупности [16]. Изменение ширины доверительного интервала оцениваемого параметра в зависимости от увеличения количества наблюдений приведено на рис. 3. Увеличение размера выборки повышает точность наших оценок, что положительно сказывается на мощности статистических критериев.

Величина эффекта

Проверка статистической значимости не несёт никакой информации о том, насколько выявленные различия сильны. Мысль о том, что связь, для которой $p < 0{,}001$ сильнее или имеет более высокую клиническую значимость, чем связь, для которой $p = 0{,}043$, является глубоко ошибочной. Подобная интерпретация силы связи по величине абсолютного значения достигнутого уровня значимости довольно часто встречается в отечественной медицинской научной литературе.

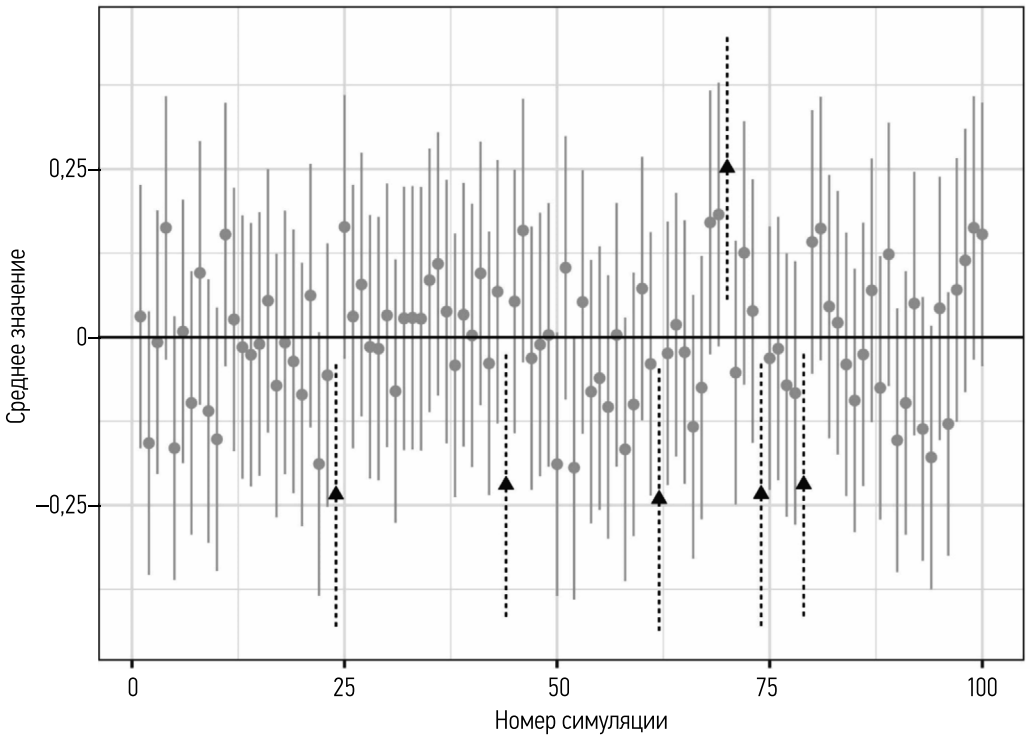

Рис. 2. Демонстрация концепции доверительного интервала.
Fig. 2. Demonstration of the confidence interval concept.

Доверительный интервал ситуацию радикально не исправляет, так как содержит информацию только о возможном диапазоне значений. Куда более важно наличие «линейки» (единиц измерения), при помощи которой можно объективизировать «расстояние» между интересующими данными точно так же, как это делается при измерении массы, температуры, силы тока и т.п. Величина (сила) эффекта — это количественная характеристика ошибочности нулевой гипотезы [17]. Если нулевая гипотеза верна (оцениваемые параметры одинаковы, нет никакой ассоциации или связи между признаками), то сила эффекта равна нулю. Наглядное отображение концепции силы эффекта приведено на рис. 4.

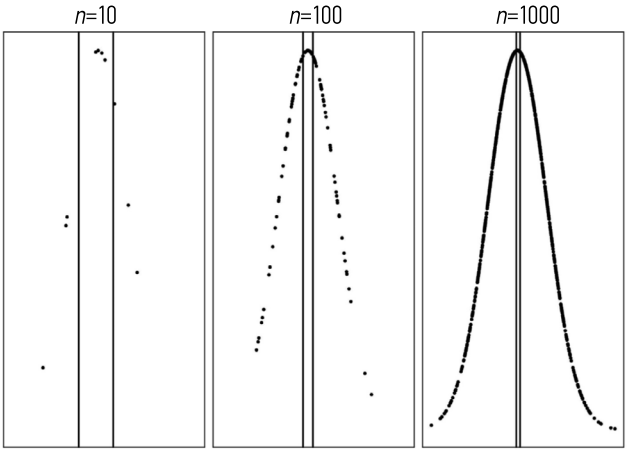

Рис. 3. Изменение ширины доверительного интервала в зависимости от объёма наблюдений.
Fig. 3. Change in the width of the confidence interval depending on the number of observations.

Существует большое количество мер силы эффекта [15, 18]: d Коэна, отношение шансов (odds ratio), относительный риск (relative risk), r (коэффициент корреляции Пирсона) и т.д. Вне зависимости от меры силы эффекта при планировании следует опираться на минимально значимый эффект или минимально клинически важный эффект. Это та граница, для преодоления которой необходима адекватная статистическая мощность критерия (для мощности рассчитывается минимально приемлемый объём выборки). При возрастании объёма выборки мощность увеличивается [19], за счёт чего можно выявить сколь угодно малый статистически значимый эффект [18], который может иметь ничтожное значение с точки зрения практической деятельности. Наоборот, если сила эффекта достаточно высока, нет нужды собирать большой массив данных и тратить лишние ресурсы. Яркой демонстрацией разницы между силой эффекта и *p*-значением является наблюдаемая в практике пропасть между клинической эффективностью и статистически значимыми результатами [15]. Напомним, что возможны четыре варианта:

- клинически незначимо и статистически незначимо;
- статистически значимо, клинически незначимо;
- статистически значимо, клинически значимо;
- статистически незначимо, клинически значимо.

Перечень вариантов напоминает табл., так как это точно та же проблема принятия решений. В данном направлении

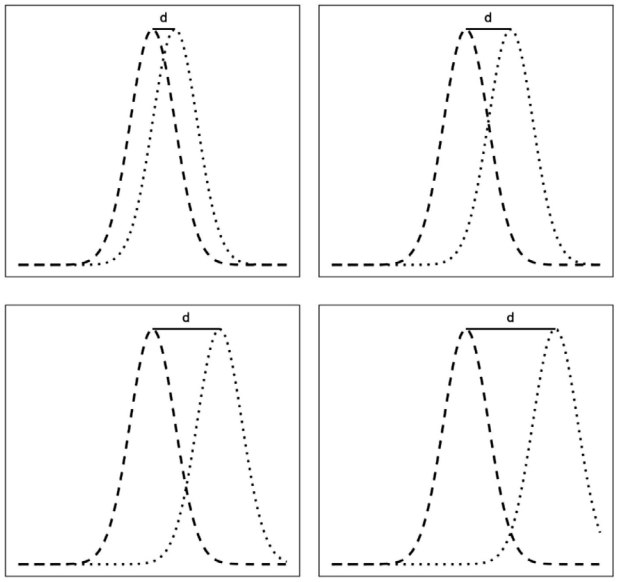

Рис. 4. Демонстрация концепции силы эффекта.
Fig. 4. Demonstration of the effect size concept.

есть предложения по более детальной классификации с учётом накопленного опыта [20], но принципиальная суть остаётся той же.

Сила эффекта — это именно то, что интересует исследователя. Самым наглядным примером важности силы эффекта является соотношение «сигнал–шум» из теории обнаружения сигналов, которое касается не только работы металлодетектора, но и мистических и паранормальных явлений [21]. Однако широкому внедрению силы эффекта в повседневную практику, на наш взгляд, мешает ряд проблем. Во-первых, отсутствие должного освещения данного вопроса в образовательных программах уменьшает вероятность использования силы эффекта в реальной практике. Отсутствие примеров использования силы эффекта в публикациях приводит к невостребованности изучения данной темы. Невостребованность показателя в практике ограничивает внедрение расчётов силы эффекта в программное обеспечение. Во-вторых, несмотря на достаточное количество литературы, в основном англоязычной, поиск информации о расчётах различных мер силы эффекта может быть сложной задачей. Этот пробел в отечественной литературе нуждается в восполнении, однако отправной точкой поиска может стать список литературы под основным текстом статьи. В-третьих, достаточно сложные и комплексные исследовательские вопросы требуют такого же трудоёмкого анализа, который не ограничивается традиционным набором статистических критериев. Это приводит к ещё большему повышению требований к специалистам в вопросах владения навыками статистического анализа, а также умения составить концепцию исследования, грамотно определить измеримый научный вопрос, разработать и реализовать математическую модель и грамотно интерпретировать полученные результаты.

Уменьшение порогового значения

На сегодняшний день общепринятое критическое значение уровня значимости составляет 0,05, по крайней мере в биомедицинских исследованиях, с редкими исключениями. Допускается, что в каждом двадцатом случае исследователь может получить ложноположительный результат. Есть предложение изменить сложившуюся практику и уменьшить значение критического уровня значимости в 10 раз, т.е. сделать его не 0,05, а 0,005 для профилактики таких ложных открытий [22]. Данное предложение вытекает из байесовского подхода пересчёта вероятностей, который, впрочем, не лишён проблем и ограничений [23]. Расчёты демонстрируют, что при p=0,005 истинная нулевая гипотеза может быть отклонена только в 6,7% случаев [22]. Аналогичные расчёты показывают, что использование 0,05 в качестве критического значения может привести к ложным открытиям или ошибочному принятию справедливой нулевой гипотезы в 28,9% случаев [19]! Последствиями ложных открытий может стать назначение не более эффективного лекарства взамен уже использующегося и часто более дешёвого препарата; написание бесполезных профилактических программ, связанных с устранением влияния несуществующего фактора риска, и т.д.

Сдвиг точки разделения для уровня значимости не решает принципиальную проблему восприятия контекста и непрерывности p-значения. Мощность статистического критерия зависит от размера выборки, поэтому легко смоделировать ситуацию, когда p=0,005 на малой выборке будет отсекать интересующий нас эффект, в то время как с ростом выборки даже такого значения не хватит, чтобы отсечь неинформативные случайные находки. Понимание данной проблемы привело к предложению делать критическое значение индивидуальным для каждого конкретного случая [24]. Это является следствием идеи, что искомый эффект — клинический, экономический или любой другой — может варьировать от слабого до сильного. Никакого толка от того, будет ли преодолён порог 0,05, 0,005 и т.д., нет, пока отсутствует количественная оценка интересующего исследователей эффекта [24].

Следует также упомянуть и другие аргументы против данной позиции. При 80% мощности сдвиг порогового значения может приводить к увеличению необходимой выборки на 70% [25], что весьма затратно. Это усугубит проблему невозможности доказать эффективность клинически важных эффектов из-за их сложности и часто слабой выраженности [26], что приведёт к ещё большему расхождению между теорией и практикой. В работе A. Vexler [27] утверждается справедливость «народной мудрости» при выборе значения p=0,05 на основе ещё более сложных вычислительных экспериментов.

ЗАБЛУЖДЕНИЯ, КАСАЮЩИЕСЯ *p*-ЗНАЧЕНИЯ

Ниже перечислены наиболее часто встречающиеся заблуждения, однако перечень не является исчерпывающим [10]:

- *p* говорит о вероятности того, что отклонение нулевой гипотезы объясняется случайностью;
- статистическая значимость устанавливает наличие важного эффекта;
- *p* <0,05 доказывает, что у нас есть поддержка проверяемой гипотезы;
- *p* <0,05 — это «значимый» результат, *p* <0,01 — «очень значимый», а *p* <0,001 — «высоко значимый» (чаще даже встречается не «значимый», а «достоверный»);
- *p* является подходящей метрикой для тех, кто заинтересован в развитии теории, а размер эффекта имеет значение только тогда, когда речь идёт о практическом применении;
- уровень *p* указывает на вероятность того, что результат не повторится, если исследование будет повторено;
- уровень *p* предсказывает количество статистических результатов, которые были бы значимы случайно;
- нулевая гипотеза — это научная гипотеза;
- отклонение нулевой гипотезы означает, что альтернатива верна;
- *p* — это то же самое, что и альфа-ошибка;
- проверка значимости нулевой гипотезы всё чаще рассматривает надёжность как замену валидности.

ЗАКЛЮЧЕНИЕ

Несмотря на декларирование приверженности концепции принятия решений Неймана–Пирсона, на практике большинство исследователей в науках о здоровье следуют в русле подхода Фишера. Ошибочная интерпретация *p*-значения во многом является следствием недостаточного качества преподавания (это предположение справедливо для большинства российских вузов, по крайней мере медицинских). Помимо объяснения сущности величины *p* и её значения следует представлять информацию и о других аспектах проверки статистических гипотез.

В нашей работе не рассмотрено много других «подводных камней», связанных с *p*-значением (например, проблемы множественных сравнений). Из всего набора полезных рекомендаций по повышению качества исследований мы хотели бы отметить критически важную: из четырёх рассматриваемых параметров (вероятность ошибки первого рода, вероятность ошибки второго рода, сила эффекта, количество наблюдений) три необходимо выбрать до проведения исследования и соответственно вычислить четвёртый.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ / ADDITIONAL INFORMATION

Вклад авторов. Оба автора подтверждают соответствие своего авторства международным критериям ICMJE. А.М. Гржибовский и А.Н. Гвоздецкий оба участвовали в разработке концепции, проведении исследования и подготовке первого варианта рукописи, внесли изменения во все последующие варианты рукописи, прочли и одобрили финальную версию перед публикацией.

Author contribution. Both authors participated in the development of the concept of the article, the preparation of the first version of the manuscript, amendments to all subsequent versions of the manuscript, and approved the final version of the text.

Финансирование. Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

Funding source. This study was not supported by any external sources of funding.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Competing interests. The authors declare that they have no competing.

СПИСОК ЛИТЕРАТУРЫ

1. Polonioli A., Vega-Mendoza M., Blankinship B., Carmel D. Reporting in experimental philosophy: current standards and recommendations for future practice // Rev Philos Psychol. 2021. Vol. 12, N 1. P. 49–73. doi: 10.1007/s13164-018-0414-3
2. Amrhein V., Trafimow D., Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication // The American statistician. 2019. Vol. 73. Suppl. 1. P. 262–270. doi: 10.1080/00031305.2018.1543137
3. Amrhein V., Korner-Nievergelt F., Roth T. The earth is flat (p >0.05): significance thresholds and the crisis of unreplicable research // PeerJ. 2017. Vol. 5. P. e3544. doi: 10.7717/peerj.3544
4. Szucs D., Ioannidis J. When null hypothesis significance testing is unsuitable for research: a reassessment // Front Hum Neurosci. 2017. Vol. 11. P. 390. doi: 10.3389/fnhum.2017.00390
5. Аканов А.А., Турдалиева Б.С., Изекенова А.К., и др. Оценка использования статистических методов в научных статьях медицинских журналов Казахстана. // Экология человека. 2013. Т. 20, № 5. С. 61–64.
6. Dorey F. The p value: what is it and what does it tell you? // Clin Orthop Relat Res. 2010. Vol. 468, N 8. P. 2297–2298. doi: 10.1007/s11999-010-1402-9
7. Haller H., Krauss S. Misinterpretations of significance: a problem students share with their teachers? // Methods of psychological research. 2002. Vol. 7, N 1. P. 1–20.
8. Palesch Y.Y. Some common misperceptions about p-values // Stroke. 2014. Vol. 45, N 12. P. e244–e246. doi: 10.1161/STROKEAHA.114.006138

9. Зорин Н.А. «Достоверность» или «статистическая значимость» — 12 лет спустя // Педиатрическая фармакология. 2011. Т. 8, № 5. С. 13–19.
10. Kmetz J.L. Correcting corrupt research: recommendations for the profession to stop misuse of p-values // The American statistician. 2019. Vol. 73. Suppl. 1. P. 36–45. doi: 10.1080/00031305.2018.1518271
11. McShane B.B., Gal D., Gelman A., Robert C., Tackett J.L. Abandon statistical significance //The American statistician. 2019. Vol. 73. Suppl. 1. P. 235–245. doi: 10.1080/00031305.2018.1527253
12. Perezgonzalez J.D. Fisher, Neyman–Pearson or NHST? A tutorial for teaching data testing // Front Psychol. 2015. Vol. 6. P. 223. doi: 10.3389/fpsyg.2015.00223
13. Lew M.J. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know p: statistical inference using p-values // Br J Pharmacol. 2012. Vol. 166, N 5. P. 1559–1567. doi: 10.1111/j.1476-5381.2012.01931.x
14. Pernet C. Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice // F1000Research. 2017. Vol. 4. P. 621. doi: 10.12688/f1000research.6963.5
15. Serdar C.C., Cihan M., Yücel D., Serdar M.A. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies // Biochem Med (Zagreb). 2021. Vol. 31. N 1. P. 010502. doi: 10.11613/BM.2021.010502
16. Lee D.K. Alternatives to p value: confidence interval and effect size // Korean J Anesthesiol. 2016. Vol. 69, N 6. P. 555–562. doi: 10.4097/kjae.2016.69.6.555
17. Grissom R.J., Kim J.J. Effect sizes for research. 2nd ed. New York : Routledge; 2012. doi: 10.4324/9780203803233
18. Sullivan G.M., Feinn R. Using effect size — or why the p value is not enough // J Grad Med Educ. 2012. Vol. 4, N 3. P. 279–282. doi: 10.4300/JGME-D-12-00156.1
19. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values // R Soc Open Sci. 2014. Vol. 1, N 3. P. 140216. doi: 10.1098/rsos.140216
20. Stahel W.A. New relevance and significance measures to replace p-values // PLoS One. 2021. Vol. 16, N 6. P. e0252991. doi: 10.1371/journal.pone.0252991
21. Anderson N.D. Teaching signal detection theory with pseudoscience // Front Psychol. 2015. Vol. 6. P. 762. doi: 10.3389/fpsyg.2015.00762
22. Benjamin D.J., Berger J.O., Johannesson M., et al. Redefine statistical significance // Nat Hum Behav. 2018. Vol. 2, N 1. P. 6–10. doi: 10.1038/s41562-017-0189-z
23. Рубанович А.В. Пересмотр критического уровня значимости (0.005 вместо 0.05): байесовский след // Радиационная биология. Радиоэкология. 2018. Т. 58, № 5. С. 453–462. doi: 10.1134/S0869803118050156
24. Betensky R.A. The p-value requires context, not a threshold // The American statistician. 2019. Vol. 73. Supl. 1. P. 115–117. doi: 10.1080/00031305.2018.1529624
25. Lakens D., Adolfi F.G., Albers C.J., et al. Justify your alpha // Nature human behaviour. 2018. Vol. 2, N 3. P. 168–171. doi: 10.1038/s41562-018-0311-x
26. Di Leo G., Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics — reasons for a conservative approach // Eur Radiol Exp. 2020. Vol. 4, N 1. P. 1–8. doi: 10.1186/s41747-020-0145-y
27. Vexler A. Valid p-values and expectations of p-values revisited // Ann Inst Stat Math. 2021. Vol. 73. P. 227–248. doi: 10.1007/s10463-021-00800-8

REFERENCES

1. Polonioli A, Vega-Mendoza M, Blankinship B, Carmel D. Reporting in experimental philosophy: current standards and recommendations for future practice. *Rev Philos Psychol*. 2021;12(1):49–73. doi: 10.1007/s13164-018-0414-3
2. Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: there is no replication crisis if we don't expect replication. *The American statistician*. 2019;73(supl. 1):262–270. doi: 10.1080/00031305.2018.1543137
3. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat (p >0.05): significance thresholds and the crisis of unreplicable research. *PeerJ*. 2017;5:e3544. doi: 10.7717/peerj.3544
4. Szucs D, Ioannidis J.P.A. When null hypothesis significance testing is unsuitable for research: a reassessment. *Front Hum Neurosci*. 2017;11:390. doi: 10.3389/fnhum.2017.00390
5. Akanov A, Turdaliyeva BS, Izekenova AK, et al. Assessment of use of statistical methods in scientific articles of the Kazakhstan's medical journals. *Ekologiya cheloveka (Human Ecology)*. 2013;20(5):61–64. (In Russ).
6. Dorey F. The p value: what is it and what does it tell you? *Clin Orthop Relat Res*. 2010;468(8):2297–2298. doi: 10.1007/s11999-010-1402-9
7. Haller H. Misinterpretations of significance: a problem students share with their teachers? *Methods of psychological research*. 2002;7(1):1–20.
8. Palesch YY. Some common misperceptions about p-values. *Stroke*. 2014;45(12):e244–e246. doi: 10.1161/STROKEAHA.114.006138
9. Zorin NA. «Validity» or «significance» — 12 years later. *Pediatric Pharmacology*. 2011;8(5):13–19. (In Russ).
10. Kmetz JL. Correcting corrupt research: recommendations for the profession to stop misuse of p-values. *The American statistician*. 2019;73(supl. 1):36–45. doi: 10.1080/00031305.2018.1518271
11. McShane BB. Abandon statistical significance. *The American statistician*. 2019;73(supl 1):235–245. doi: 10.1080/00031305.2018.1527253
12. Perezgonzalez JD. Fisher, Neyman–Pearson or NHST? A tutorial for teaching data testing. *Front Psychol*. 2015;6:223. doi: 10.3389/fpsyg.2015.00223
13. Lew MJ. Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know p: statistical inference using p-values. *Br J Pharmacol*. 2012;166(5):1559–1567. doi: 10.1111/j.1476-5381.2012.01931.x
14. Pernet C. Null hypothesis significance testing: a guide to commonly misunderstood concepts and recommendations for good practice. *F1000Research*. 2017;4:621. doi: 10.12688/f1000research.6963.5
15. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in

pre-clinical, clinical and laboratory studies. *Biochem Med (Zagreb)*. 2021;31(1)010502. doi: 10.11613/BM.2021.010502
16. Lee DK. Alternatives to p value: confidence interval and effect size. *Korean J Anesthesiol*. 2016;69(6):555–562. doi: 10.4097/kjae.2016.69.6.555
17. Grissom RJ, Kim JJ. *Effect sizes for research*. 2nd ed. New York: Routledge; 2012. doi: 10.4324/9780203803233
18. Sullivan GM, Feinn R. using effect size — or why the p value is not enough. *J Grad Med Educ*. 2012;4(3):279–282. doi: 10.4300/JGME-D-12-00156.1
19. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*. 2014;1(3):140216. doi: 10.1098/rsos.140216
20. Stahel WA. New relevance and significance measures to replace *p*-values. *PLoS One*. 2021;16(6):e0252991. doi: 10.1371/journal.pone.0252991
21. Anderson N.D. Teaching signal detection theory with pseudoscience. *Front Psychol*. 2015;6:762. doi: 10.3389/fpsyg.2015.00762
22. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav*. 2018;2(1):6–10. doi: 10.1038/s41562-017-0189-z
23. Rubanovich AV. Redefining the critical value of significance level (0.005 instead of 0.05): the bayes trace. *Radiation biology. Radioecology*. 2018;58(5):453–462. (In Russ). doi: 10.1134/S0869803118050156
24. Betensky RA. The p-value requires context, not a threshold. *The American statistician*. 2019;73(supl. 1):115–117. doi: 10.1080/00031305.2018.1529624
25. Lakens D, Adolfi, FG, Albers CJ, et al. Justify your alpha. *Nature human behaviour*. 2018;2(3):168–171. doi: 10.1038/s41562-018-0311-x
26. Di Leo G, Sardanelli F. Statistical significance: p value, 0.05 threshold, and applications to radiomics — reasons for a conservative approach. *Eur Radiol Exp.* 2020;4(1):1–8. doi: 10.1186/s41747-020-0145-y
27. Vexler A. Valid p-values and expectations of p-values revisited // *Ann Inst Stat Math.* 2021;73:227–248. doi: 10.1007/s10463-021-00800-8

ОБ АВТОРАХ

***Гржибовский Андрей Мечиславович,** PhD;
адрес: Россия, 163061, Архангельск, Троицкий проспект, д. 51;
ORCID: https://orcid.org/0000-0002-5464-0498;
eLibrary SPIN: 5118-0081;
e-mail: andrej.grjibovski@gmail.com

Гвоздецкий Антон Николаевич, к.м.н.;
адрес: Россия, 191015, Санкт-Петербург, ул. Кирочная, 41;
ORCID: https://orcid.org/0000-0001-8045-1220;
eLibrary SPIN: 4430-6841;
e-mail: gvozdetskiy_an@outlook.com

*Автор, ответственный за публикацию / Corresponding author

AUTHORS INFO

***Andrej M. Grjibovski,** MD, MPhil, PhD;
address: 51 Troitsky avenue, 163061, Arkhangelsk, Russia;
ORCID: https://orcid.org/0000-0002-5464-0498;
eLibrary SPIN: 5118-0081;
e-mail: andrej.grjibovski@gmail.com

Anton N. Gvozdeckii, MD, Cand. Sci. (Med.);
address: 41 Kirochnaya st., 191015, St. Petersburg, Russia;
ORCID: https://orcid.org/0000-0001-8045-1220;
eLibrary SPIN: 4430-6841;
e-mail: gvozdetskiy_an@outlook.com