Психология. Журнал Высшей школы экономики.
2009. Т. 6, № 4. С. 66–86.

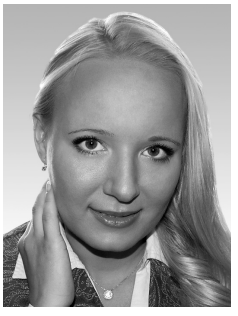
Статистические методы в психологии

О РЕФОРМЕ СТАТИСТИЧЕСКОГО ВЫВОДА В ПСИХОЛОГИИ

Сомнительная значимость статистической значимости

С.В. СИВУХА, А.А КОЗЯК

Сивуха Сергей Викентьевич — кандидат психологических наук, доцент, доцент кафедры социальной коммуникации Белорусского государственного университета. Сфера профессиональных интересов — социальная психология, социология, методология качественных и количественных исследований.
Контакты: e-mail: sergei. sivuha@gmail.com



Козяк Анастасия Александровна — кандидат психологических наук, оргконсультант компании Eurowest (Республика Словения). Сфера профессиональных интересов — социальная психология, методология качественных и количественных исследований.
Контакты: e-mail: nastusek@yandex.ru

Резюме

С середины XX века научная психология ассоциируется с эмпирическими исследованиями и статистическим анализом данных. Критика проверки статистической значимости (расчета значения p и его интерпретации) началась одновременно с популяризацией этой процедуры в психологии. Сведение научного вывода к (часто неверному) толкованию значения p является одной из серьезных помех развитию психологии. В статье обсуждаются некоторые заблуждения исследователей по поводу проверки статистической значимости и затрагиваются некоторые обсуждаемые в литературе альтернативы.

Ключевые слова: *статистический вывод, статистическая значимость, значение p, нуль-гипотеза, проверка значимости нуль-гипотезы*

Введение

В течение нескольких десятилетий статистическая проверка нуль-гипотезы служит основным средством оценки полученных результатов. В сущности, «значение *p*» или даже символ «*» рядом с величиной коэффициента корреляции используется как индикатор важности результата и тем самым как оценка вклада автора исследования в психологическую науку: работы с недостаточно большим количеством «*» могут быть не приняты к печати или не пройти экспертизу перед защитой. Роль значений *p* в развитии академической психологии сопоставима с ролью экспериментального метода. По данным Р. Хаббарда и П. Райан (Hubbard, Ryan, 2000), с 1965 по 1998 г. проверка статистической значимости с помощью значения *p* использовалась более чем в 90% эмпирических статей в 12 журналах Американской психологической ассоциации.

Начиная с 1940-х годов специалисты в области статистической методологии проявляют серьезное беспокойство по поводу необоснованно широкого применения проверки значимости нуль-гипотез (ПЗНГ) и неверной интерпретации результатов этой проверки. Как показал Р. Клайн, количество критических публикаций увеличивается экспоненциально (Kline, 2004). Острая и хорошо аргументированная статья Д. Бейкана (Bakan, 1966) давно разошлась на цитаты. В 1994 г. в журнале «American Psychologist» появилась статья Дж. Коэна под ироничным заголовком «Земля круглая ($p < 0.05$)» (Cohen, 1994), оживившая споры по поводу проверки нуль-гипотезы. Несколько журналов («Psychological Science», «Educational and Psychological Measurement», «Journal of Experimental Education», «Research in the Schools», «Theory & Psychology») посвятили этой проблеме специальные выпуски. Наиболее радикальные авторы (Carver, 1978; Oakes, 1986 и др.; см.: Nix, Barnett, 1998) предлагали запретить ПЗНГ. Одним из результатов дискуссии по поводу проверки статистической значимости стало создание специальной рабочей группы Американской психологической ассоциации по подготовке новых

требований к использованию статистических методов в научных публикациях[1]. Задача группы состояла в том, чтобы «прояснить некоторые спорные вопросы, касающиеся применения статистики, включая проверку значимости и альтернативы ей...» (Wilkinson and TFSI, 1999, p. 594). Проблема статистического вывода обсуждалась в журналах, на съездах и конференциях. Ожидалось, что Американская психологическая ассоциация внесет революционные изменения в «Руководство по подготовке публикаций». О напряженности дискуссий можно судить по началу статьи в академическом журнале: «Неправда, что группа радикальных активистов захватила в заложники десять статистиков и шесть редакторов на ежегодном съезде Американской психологической ассоциации... и скандировала: “Запретите проверку гипотез!”, “Долой нуль-гипотезу!”» (Abelson, 1997).

Отчет рабочей группы (Wilkinson and TFSI, 1999) лишь незначительно повлиял на предписания 5-го издания «Руководства по подготовке публикаций Американской психологической ассоциации» (APA, 2001). Разумеется, ПЗНГ не была запрещена. Однако широкое обсуждение этой проблемы серьезно изменило редакционную политику ведущих журналов ряда американских профессиональных ассоциаций и способствовало распространению новой философии и новых практик статистического анализа. К сожалению, российские психологические журналы в основном обошли молчанием проблему необоснованно широкого применения и неправильной интерпретации результатов ПЗНГ. Данная статья в определенной степени восполняет этот пробел. Мы хотим привлечь внимание психологического сообщества к неверным практикам использования и интерпретации критериев статистической значимости. Это обсуждение не является ни полным, ни беспристрастным, поскольку мы считаем себя умеренными противниками практики ПЗНГ. Мы убеждены в том, что последствия этой практики сегодня являются наиболее острой методологической проблемой психологии и смежных наук. Планируетсяя вторая часть статьи, которая будет посвящена обсуждению оценивания статистических эффектов как одной из альтернатив проверке статистической значимости нуль-гипотез. Эта альтернатива не лишена недостатков и, возможно, менее продуктивна, чем построение доверительных интервалов, эксплораторный анализ, байесова статистика, повторное извлечение выборок и др. Каждый из этих подходов заслуживает отдельного обсуждения. Мы выделили статистические эффекты в качестве главного предмета по причине того, что они являют собой ближайшую, самую простую и

[1] В группу вошли крупнейшие методологи, авторы учебников, преподаватели, статистики, редакторы журналов, специалисты в компьютерных науках. Сопредседатели группы: Robert Rosenthal, Robert Abelson, Jacob Cohen. Члены группы: Leona Aiken, Mark Appelbaum, Gwyneth Boodoo, David A. Kenny, Helena Kraemer, Donald Rubin, Bruce Thompson, Howard Wainer, Leland Wilkinson. Приглашенные советники: Lee Cronbach, Paul Meehl, Frederick Mosteller, John Tukey.

наиболее естественную альтернативу ПЗНГ.

«Теория» статистического вывода

Мы считаем избыточным подробно излагать основания проверки статистических гипотез, однако небольшой экскурс в историю полезен для иллюстрации главного тезиса этого раздела — о смешении двух статистических подходов как одной из основных ошибок при анализе данных (Gigerenzer, 1993; Huberty, 1993; см. также: Стерн, Смит, 2003).

Считается, что идея статистического вывода впервые появилась в работе Дж. Арбутнота почти 300 лет назад (см., напр.: Huberty, Pike, 1999). Значение **p** впервые использовал К. Пирсон в 1990 г. для критерия χ^2 под названием «p, χ^2 критерий». Популяризация этой идеи и разработка первой теории статистического вывода связаны с именем Рональда Фишера (1925, русское издание Фишер, 1958). В основе этой теории лежит логика доказательства от противного. Утверждение, которое предполагается опровергнуть, называется нуль-гипотезой. На основании ряда допущений, в том числе теоретического выборочного распределения статистики при условии, что нуль-гипотеза верна, исследователь рассчитывает вероятность получения данного результата — значение p. Оно рассматривалось Р. Фишером как показатель силы доказательств в пользу ошибочности нуль-гипотезы[2]. Пороговые значения p не обосновывались: «Если вероятность P <соответствующая значению статистического критерия> содержится в широком промежутке от 0.1 до 0.9, то у нас не будет никаких оснований сомневаться в проверяемой гипотезе; если же вероятность P становится, например, ниже 0.02, то это прямо указывает на несостоятельность данной гипотезы. Риск впасть в ошибку не будет слишком большим, если мы проведем пограничную линию у $P = 0.05$ и будем считать, что значение <статистического критерия>, лежащего выше этой линии, указывает на наличие существенных и реальных отклонений» (Фишер, 1956, с. 70–71). Как можно видеть, Р. Фишер не проводил ясного дихотомического разделения на значимое–незначимое и не обосновывал жесткие пороги вроде 0.05 или 0.01 (Стерн, Смит, 2003). Он полагал, что единственное определенное решение имеет место тогда, когда исследователь отклоняет нуль-гипотезу. Если оснований для ее отклонения нет, исследователь попадает в неопределенное положение. Нуль-гипотеза не может быть «принята» или «подтверждена», просто для ее отвержения пока нет достаточных оснований. Позже Р. Фишер смягчил свою позицию, но не прояснил ее (см.: Gigerenzer, 1993).

Чтобы избежать субъективной интерпретации значения p и избавиться от других недостатков фишеровского подхода (недостаточности нуль-гипотезы для описания исследовательской ситуации и неопределенности в

[2] В следующем разделе с помощью байесовой статистики показано, что эта сила доказательств в фишеровской модели существенно завышена.

случае невозможности отвергнуть нуль-гипотезу), Джерси Нейман и Эгон Пирсон обосновали необходимость альтернативной гипотезы, H_1. Ключевая идея концепции: гипотеза всегда проверяется по отношению к альтернативной. Не указав альтернативу H_0, невозможно обосновать выбор оптимального статистического критерия. Дж. Нейман и Э. Пирсон также ввели понятие ошибки первого рода α (отклонение нулевой гипотезы, в то время как она верна) и ошибки второго рода β (принятие нулевой гипотезы, в то время как она неверна). Критическое значение статистического критерия, используемое для решения в пользу H_0 или H_1, зависит от того, какой тип ошибок исследователь хочет минимизировать, а это определяется целью исследования. Выбор значения α основан на информированных суждениях исследователя. Важнейшим понятием теории Неймана-Пирсона является мощность статистического критерия, равная 1-β, т. е. дополнению вероятности ошибки первого рода до 1. Часто на маленьких выборках нуль-гипотеза не может быть отклонена ввиду недостаточной статистической мощности, и все известные статистические критерии становятся избыточно мощными на больших выборках. Без анализа мощности исследователь не увидит опасности ошибочных решений.

Строго говоря, концепция Неймана-Пирсона не является теорией статистического вывода, это теория принятия статистических решений. Авторы предпочитали называть свою концепцию «проверкой гипотез» (в противоположность «проверке значимости» Р. Фишера). Опираясь на полученные данные и контролируя риск неверных действий, исследователь принимает решение в пользу H_0 или H_1 и ведет себя так, как если бы одна из конкурирующих гипотез была верна. Различия между теориями Фишера и Неймана-Пирсона подробно реконструированы в работах Г. Гигеренцера, К. Хьюберти и Р. Хаббарда (см., напр.: Gigerenzer, Murray, 1987; Gigerezer, 1993; Hubbard, 2004; Hubbard, Lindsay, 2008; Huberty, 1993).

Во-первых, теория Фишера принципиально индуктивна и не допускает иной логики: «индуктивный вывод — единственный известный нам процесс, с помощью которого появляется существенно новое знание» (Fisher, 1966, p. 6). Напротив, Дж. Нейман и Э. Пирсон довольно холодно отзывались о неопределенности и неполноте индуктивного рассуждения; их дедуктивная концепция предписывает действия от общего (правил принятия статистических решений) к частному (принятию решения). Терминологическая путаница может быть связана с тем, что авторы называли свою модель теорией «индуктивного поведения», имея тем самым в виду, что проверка гипотез позволяет «корректировать свои действия полученными частотами событий, чтобы избежать нежелательных последствий» (цит. по: Hubbard, 2004, p. 301).

Во-вторых, согласно Р. Фишеру, формулируется только нуль-гипотеза и рассчитывается показатель ее (не)правдоподобия, хотя, как отмечают критики, Р. Фишер все-таки рассматривал некоторую неясную альтернативу, когда оценивал «сензитивность эксперимента» (см.: Hubbard,

Bayarri, 2003, p. 172)[3]. По Дж. Нейману и Э. Пирсону, формулируются нулевая и альтернативная гипотезы, и после статистической проверки одна из них признается истинной.

В-третьих, модель Фишера сводится к проверке значимости — расчету вероятности получения данного или более экстремального значения статистического критерия в предположении, что нуль-гипотеза верна, а выборка извлечена из бесконечной генеральной совокупности случайно. Этой вероятности — значению p — Р. Фишер давал эпистемологическое толкование (см. критику: Gigerezer, 1993; Hubbard, 2004; Rosnow, Rosenthal, 1989), рассматривая его как меру доказательности против нуль-гипотезы и тем самым как средство кумуляции знаний. В случае малого значения p делается вывод, что либо произошло редкое событие, либо нуль-гипотеза ложна. По Дж. Нейману и Э. Пирсону исследователь сначала устанавливает уровень ошибки I рода и выбирает статистический критерий, минимизирующий (для данной α) вероятность ошибки II рода, а затем проверяет нуль-гипотезу. Если значение статистического критерия попадает в критическую область выборочного распределения (α), H_0 отвергается в пользу H_1. Точная вероятность получения данного или более экстремального значения критерия не имеет смысла и не рассчитывается. По этой причине теория не имеет дела с силой доказательств против нуль-гипотезы. Недостатком термина «ошибка I рода» является то, что он не отражает условную природу вероятности (Falk, Greenbaum, 1995).

В-четвертых, теория Фишера позволяет получать доказательства против нуль-гипотезы в единичном эксперименте. Согласно Дж. Нейману и Э. Пирсону ошибка I рода, задающая критическую область статистического критерия, интерпретируется как доля неверных отвержений H_0 при большом количестве извлечений выборки равного объема с возвратом из одной и той же ясно определенной генеральной совокупности. Поэтому только α получает частотную интерпретацию (см. ниже).

Г. Гигеренцер и Д. Мюррей видят причины проблем современной статистики в том, что после Второй мировой войны подходы Фишера и Неймана-Пирсона были неправильно поняты и механически объединены статистиками (Gigerenzer, Murray, 1987). Обычно исследовали формально следуют процедуре проверки гипотез, но дают своим результатам толкование в духе проверки значимости[4]. Вторая особенность принятой модели состояла в дихотомическом принятии решения (значимо–незначимо), хотя идеи Р. Фишера допускали и другую интерпретацию. Модель

[3] Сензитивность эксперимента определяется его логическим дизайном, конструктной валидностью измерений и особенно объемом выборки, от которого зависит статистическая мощность (Meehl, 1978).

[4] Р. Фишер и Дж. Нейман до последних дней боролись против объединения двух теорий и переноса понятий из одной модели в другую (см.: Gigerenzer, Murray; 1987, Sedlmeier, Gigerenzer, 1989). Э. Пирсон в перепалке не участвовал.

ПЗНГ стала доминировать в англоязычной психологии в 1940–1955 гг., вытеснив другие подходы к описанию результатов исследования и тщательный анализ данных. Всплеск публикаций с использованием ПЗНГ получил название «революции вывода» (Gigerenzer, 1993; Hubbard, Ryan, 2000). Причудливый гибрид двух теорий в большинстве учебников по прикладной статистике излагается как догма.

Хотя в исторической перспективе гибридная форма статистического вывода имела определенные преимущества (см.: Hubbard, Parsa, Luthy, 1997), смешение двух несовместимых теорий имело прямые негативные следствия (Finch, Cumming, 2001), что проявилось прежде всего в неверной интерпретации p и α. Напомним, что, по Р. Фишеру, p оценивает силу доказательств против нуль-гипотезы в единичном исследовании и не допускает частотной интерпретации. Это случайная переменная, основанная на данных. Неправильно называть значение p ошибкой I рода. Напротив, ошибка I рода, α – фиксированное значение, установленное исследователем. Эта ошибка имеет простую частотную интерпретацию: при заданном уровне α на большом числе выборок, извлеченных из одной и той же генеральной совокупности, истинная H_0 будет ошибочно отвергнута не более чем в $\alpha \cdot 100\%$ случаев. Ошибка I рода не зависит от данных и является свойством критерия. Многие авторы пытаются дать значению p частотную интерпретацию по аналогии с α. Другая распространенная ошибка состоит в том, что исследователь выбирает уровень значимости (допустимую величину α) и в случае получения значимого результата, т. е. когда значение статистического критерия попадает в критическую область, приводит точное значение p как меру доказательности ошибочности нуль-гипотезы. С позиций Дж. Неймана и Э. Пирсона это неправильно. Поскольку α установлена до проверки гипотезы, решение дихотомично: H_0 верна или неверна. Точное положение значения критерия в критической области совершенно неважно. Парадокс теории Неймана–Пирсона в том, что новые данные, полученные после неудачной попытки отклонения нуль-гипотезы, не позволяют ее отвергнуть в следующий раз. Более точно, если α установлена равной 0.05, при второй проверке ошибка I рода она будет обусловлена первой неудачей и составит не 0.05, а $0.0975 = 0.05 + 0.05 \cdot 0.095$ (см.: Hubbard, 2004, p.310).

Одной из самых распространенных ошибок является так называемая скользящая α, когда в одной публикации автор приводит несколько уровней значимости, отмечая количеством звездочек $p < 0.05$, $p < 0.01$ и т. д. В действительности уровень значимости, или ошибка I рода устанавливается единожды, до начала исследования, и не корректируется в зависимости от величины коэффициентов или разностей средних (Tryon, 2001)[5]. Следует либо приводить

[5] Одной из причин смешения p и α, как сообщает в биографии У. Госсета (Стьюдента) Эгон Пирсон, явилось нежелание Р. Фишера перепечатывать в своих книгах точные таблицы

рассчитанное в процессе анализа точное значение *p*, не интерпретируя его в частотных терминах, либо сообщать заранее установленный уровень ошибки I рода, и в этом случае ограничиваться дихотомическим решением и не рассчитывать точное значение *p*.

По данным Р. Хаббарда, смешение *p* и α допускают практически все известные авторы, в том числе большинство авторов, цитированных в данной статье. Оно также присутствует во многих англоязычных учебниках и в подавляющем большинстве эмпирических статей (см.: Hubbard, 2004; Hubbard, Armstrong, 2006). Это касается и публикаций на русском языке, в том числе популярных учебников (Наследов, 2006; Сидоренко, 2003 и др.).

Заблуждения, связанные с проверкой статистической значимости

Большая часть документированных в литературе ошибок связана с доминирующей фишеровской традицией, основанной на оценке вероятности получения результата при условии, что нуль-гипотеза верна. По этой причине в названии раздела мы использовали фишеровское словосочетание, а не неймановскую «проверку статистической гипотезы». Даже определяя значение *p* правильно, многие авторы его неправильно интерпретируют. Среди распространенных ошибок в процедуре ПЗНГ — неверная интерпретация значения *p* как 1) показателя истинности нуль-гипотезы, 2) вероятностной меры случайного получения результатов, 3) объективного средства проверки нуль-гипотезы, 4) меры доказательства истинности альтернативной гипотезы, 5) меры воспроизводимости результатов и 6) показателя практической важности полученного результата. Мы остановимся на этих ошибках очень кратко, лишь для обоснования поиска альтернативных практик статистического вывода, и отсылаем заинтересованных читателей к цитированным здесь публикациям (Cohen, 1994; Sohn, 1998; Thompson, 2006 и др.)[6].

Заблуждение 1: p как показатель истинности нуль-гипотезы

Основной недостаток общепринятой процедуры ПЗНГ с использованием значения *p* («статистической значимости») состоит в том, что *p* «не говорит нам того, что мы хотим знать», т. е. вероятности того, что нуль-гипотеза верна (Cohen, 1994, p. 977). Эту мысль тридцатью годами ранее высказал Д. Бейкан: «Критерий значимости не дает информации о психологических феноменах, которая ему обычно приписывается»

критерия хи-квадрат, подготовленные его недругом Карлом Пирсоном и сотрудниками. Р. Фишер рассчитал собственные таблицы, составленные так, чтобы значения критерия сообщались для некоторых «круглых» значений *p* (см.: Hubbard, 2004). Со временем эти «пороговые значения» закрепились на практике.

[6] Список из 402 работ, посвященных ПЗНГ, оставлен Б. Томпсоном в 2001 г. Режим доступа: http://welcome.warnercnr.colostate.edu/~anderson/thompson1.html

(Bakan, 1966, p. 423). Значение *p* повсеместно и *неверно* интерпретируется как вероятность того, что нуль-гипотеза верна в свете полученных данных. Обозначим эту (желаемую) условную вероятность как $P(H_0|D)$, где *D* означает «данные». Если вероятность истинности нуль-гипотезы мала (менее 0.05), мы отклоняем ее в пользу альтернативной гипотезы. Именно эта логика ПЗНГ представлена в большинстве учебников для психологов. В действительности ПЗНГ дает нам обратную условную вероятность: $P(D|H_0)$. *Значение p — это рассчитанная вероятность того, что данный или более экстремальный результат для выборки данного объема может быть получен при условии, что нуль-гипотеза верна в генеральной совокупности*[7] (Finch, Cumming, Thomason, 2001, p. 183; Thompson, 2006), и, как отмечают Л. Кронбах и Р. Сноу, только байесова статистика позволяет делать утверждения о $P(H_0|D)$ (см.: Carver, 1978, p. 384). Начинающие исследователи полагают, что ПЗНГ позволяет делать выводы о неизвестных параметрах генеральной совокупности. На самом деле эта процедура предполагает определенные характеристики генеральной совокупности истинными и проверяет выборочные статистики, а не основывается на них.

Смешение двух условных вероятностей свойственно авторам многих учебников по статистике (Brewer, 1985), не говоря уже об авторах диссертаций и эмпирических статей, хотя эти вероятности могут заметно различаться. Дж. Коэн показывает это на примере диагностики шизофренического расстройства (Cohen, 1994, p. 998–999). Пусть распространенность этого заболевания в популяции взрослых составляет 2% ($P(H_1) = 0.02$ и $P(H_0) = 0.98$). Обозначим позитивный результат теста символом *D*. Чувствительность теста (способность правильно определять наличие расстройства, или $P(D|H_1)$) составляет 0.95; специфичность теста (способность правильно определять отсутствие расстройства, или $P(\text{не-}D|H_0)$) равна 0.97, т. е. $P(D|H_0) = 0.03$. По теореме Байеса вероятность того, что нуль-гипотеза верна, при условии получения позитивного результата теста равна:

$$P(H_0|D) = \frac{P(H_0)\cdot P(D|H_0)}{P(H_0)\cdot P(D|H_0) + P(H_1)\cdot P(D|H_1)} = \frac{0.98\cdot 0.03}{0.98\cdot 0.03 + 0.02\cdot 0.95} = 0.607$$

Вероятность того, что индивид с диагностированным шизофреническим расстройством здоров, превышает 0.6, хотя значение $p = 0.03$! Критики незамедлительно указали на неубедительность примера с шизофренией. Д. Зон, противник байесова подхода в статистике, считает полученный результат следствием того, что шизофренические расстройства редки, и подобные задачи якобы «не типичны для психологии» (Sohn, 1998, p. 295). Однако работы Дж. Бергера и Т. Селке (см. обзор и ссылки на сайт с программой Дж. Бергера:

[7] Ряд статистиков выступает против того, чтобы при принятии решения об отклонении нуль-гипотезы учитывать вероятность прямо не наблюдаемых («более экстремальных») результатов.

Hubbard, Lindsay, 2008) показывают, что $P(H_0|D)$ существенно превышает p как раз в типичных для психологии задачах. Так, при уровне значимости p = 0.05 будет ошибочно отвергнуто как минимум 22% истинных нуль-гипотез. По совокупности исследований, проведенных на реальных и симулированных данных, можно уверенно утверждать, что значения p существенно преувеличивают доказательность свидетельств против нуль-гипотез. Это обстоятельство заставляет критически воспринимать публикации, в которых делается вывод о наличии «тренда» или «тенденции» на том основании, что полученный результат «близок к статистически значимому» (см.: Tryon, 2001).

Неверная интерпретация p как вероятности истинности H_0 отчасти объясняется ошибкой переноса логики доказательства от противного на вероятностные утверждения. Условный деструктивный силлогизм, так называемый modus tollens, лежащий в основании метода фальсификации, имеет форму «Если А, то В; не-В; следовательно, не-А» (Ивин, 2003). Отрицание следствия ведет к отрицанию основания. Однако этот силлогизм, сконструированный в вероятностных терминах, приводит к неверному выводу: «Если X — человек, маловероятно, что X — психолог; X — психолог; следовательно, маловероятно, что X является человеком» (примеры подобных «силлогизмов» в применении к проверке нуль-гипотез см.: Cohen, 1994; Gigerenzer, 1993). Логическое и вероятностное доказательства «являются разными системами, подчиняющимися разным правилам» (Falk, Greenbaum, 1995, p. 81). Р. Фишер хорошо понимал проблему обратного вывода — необходимость делать вывод о нуль-гипотезе, т. е. $P(H_0)$, имея в наличии значение $P(D|H_1)$, но не смог ее решить без обращения к байесовой статистике.

Заблуждение 2: p как показатель вероятностной меры случайного получения результатов

Нечувствительность к условно-вероятностной природе значения p побуждает многих авторов учебников неправильно определять его как вероятность случайного получения результата. Подобное неверное определение p — как «вероятности того, что ... <результаты> носят случайный характер» — приведено в одном из лучших современных российских учебников (Наследов, 2006, с. 99). По этому поводу ясно высказался Р. Карвер: «Значение p не может быть вероятностью того, что разность между средними значениями... обусловлена случайностью, поскольку (a) значение p рассчитано в предположении, что вероятность случайного получения различия между средними равна единице, и (b) значение p используется для решения о принятии или отвержении идеи о том, что вероятность того, что случайность обусловила разность между средними, равна единице» (Carver, 1978, p. 383).

Заблуждение 3: p как объективное средство проверки нуль-гипотезы

Популярность ПЗНГ часто связывают с тем, что эта процедура будто бы обеспечивает объективность выводов (Hubbard, Parsa, Luthy, 1997; Kirk, 1996; Kline, 2004) и позволяет

уменьшить различия в ценностях (Thompson, 1999). Объективность значения *p* как меры доказательства ошибочности нуль-гипотезы подчеркивал Р. Фишер: «Чувство, вызванное проверкой значимости, имеет объективную основу в том, что вероятностное утверждение, на котором оно базируется, есть факт, сообщаемый и проверяемый другими рассудительными умами» (цит. по: Hubbard, Lindsay, 2008, p. 71). Между тем значения *p* зависят не только от полученных данных, но и от факторов, находящихся под контролем исследователя: объема выборки, выбранного статистического критерия, дизайна исследования, формулировки нуль-гипотезы. Так, на основании субъективных представлений исследователь может делить значение *p* пополам, чтобы использовать односторонний статистический критерий или корректировать его в процедуре множественных сравнений. Интерпретация *p* зависит от формулировки нулевой (точечная или интервальная) и альтернативной (одно- или двусторонняя) гипотезы.

Необъективность показателя *p* проявляется и в его чувствительности к объему выборки. Так, статистически значимый коэффициент корреляции Пирсона (двусторонний критерий *t*) для N=10 должен превышать 0.632, а для N=1000 может иметь пренебрежимо малое значение 0.062. В этом смысле можно утверждать, что *p* служит критерием величины выборки (Daniel, 1998, p. 26). На больших выборках, типичных для социологических опросов и диссертационных исследований в психологии, даже бессмысленные коэффициенты корреляции статистически значимы, Д. Ликкен назвал это корреляционным шумом (см.: Cohen, 1994). Ситуация осложняется тем, что не только обычные люди, но и академические психологи нечувствительны к роли объема выборки в задачах статистического характера (Тверски, Канеман, 2005).

Многие критики отмечают бессмысленность типичной формулировки нуль-гипотезы, касающейся нулевого статистического эффекта, т. е. равенства средних значений или равенства коэффициента корреляции нулю. По Р. Фишеру, нуль-гипотеза (null hypothesis) — это все лишь утверждение, которое предполагается отвергнуть (нуллифицировать, nullify). Утверждение о равенстве параметра или эффекта нулю Дж. Коэн предлагает называть более точно — nil hypothesis, что можно перевести как «совершенно нулевая гипотеза». По поводу этого утверждения известно, что оно «всегда ложно» — на достаточно большой выборке и при достаточно точном уровне измерения (см.: Cohen, 1994; Daniel, 1998; Meehl, 1978; Tukey, 1991). Другими словами, для «совершенно нулевой» гипотезы вероятность ошибки первого рода почти всегда равна нулю, а большие выборки и вовсе не дают ей шанса уцелеть. Трудно оценить количество таких артефактных открытий в психологии.

Заблуждение 4: p как показатель меры воспроизводимости результатов

Авторы часто полагают, что маленькое значение *p* (меньше 0.05) свидетельствует в пользу альтернативой гипотезы (см.: Brewer, 1985;

Carver, 1978; Oakes, 1986). Однако «никто не смог предложить убедительного доказательства того, что малая вероятность $P(D|H_0)$ является достаточным основанием для (а) отклонения H_0 и (б) подтверждения H_1» (Sohn, 1998, p. 298). Еще более серьезная ошибка связана с выводом об истинности проверяемой теории вследствие отвержения нуль-гипотезы. Дело здесь не только в подмене силлогизмов (см.: Cohen, 1994, p. 999), но и в нарушении базовой логики построения научных теорий (Meehl, 1990). Проверяемая содержательная теория основана на вспомогательных теориях и предположениях, и именно следствия из этих предположений формулируются в виде альтернативной гипотезы, противостоящей нуль-гипотезе (Meehl, 1978). Отвержение последней может быть индикатором ложности любого из вспомогательных предположений, но не содержательной теории. Будучи статистическими гипотезами, H_0 и H_1 сами по себе не имеют содержательного смысла и совместимы с различными содержательными гипотезами (Kline, 2004). Статистическая значимость не дает информации о том, какой из факторов ответствен за полученный результат. Как выразил эту мысль С. Чоу, «<Содержательная> теория истинна лишь в той степени, в которой она выдерживает согласованные попытки фальсифицировать ее с помощью тщательно спроектированной и выполненной серии конвергентных операций» (Chow, 1998, p. 325), и вопросы об истинности имеют не статистический, а концептуальный характер. Выборочная ошибка, возможное влияние которой на результат проверяется с помощью критериев статистической значимости, является лишь одним из 15 угроз валидности исследования (Carver, 1978). Отклонив эту угрозу в случае статистически значимого результата, исследователь должен позаботиться еще о 14 угрозах.

Заблуждение 5: p как показатель меры воспроизводимости результатов

Еще одна ошибка, совершаемая явно или лишь предполагаемая в рассуждениях, состоит в трактовке значения $(1 - p)$ как меры воспроизводимости полученных результатов (Oakes, 1986). Это «когнитивная западня» связана с тем, что «поскольку H_0 есть гипотеза о случайности, а значимый результат означает отклонение гипотезы, почти неизбежно возникает соблазн сделать вывод о том, что этот эффект не является случайным и может появиться снова» (Falk, Greenbaum, 1985, p. 90). На самом деле ПЗНГ не оценивает воспроизводимость результатов, а предполагает ее (Thompson, 1999, 2006). Репликация результатов зависит не от условной вероятности $P(D|H_0)$, а от дизайна исследования, выборки и величины полученного эффекта. Д. Зон напоминает, что ПЗНГ основана на идее бесконечного (пусть и воображаемого) извлечения выборок равного объема, необходимого для построения выборочного распределения статистики, поэтому маленькое значение p может верифицируемо предсказать результат лишь при бесконечных воспроизведениях исследования. Поскольку последнее условие невыполнимо, статистическая значимость не может указывать

на воспроизводимость (Sohn, 1998, p. 300). Д. Зон также считает процедуру ПЗНГ самодостаточной: «Парадокс практики проверки значимости — в обстоятельствах *эмпирической* проверки... делающей последующие *эмпирические* свидетельства избыточными... Проверка значимости по духу фундаментально антиэмпирична. <Она> ... позволяет ученому не проводить исследований, направленных на воспроизведение» (Sohn, 1998, p. 301). Дж. Миллер в недавней статье показывает, что вероятность воспроизведения статистически значимого эффекта принципиально не может быть рассчитана (Miller, 2009). Если соглашаться с тем, что цель науки состоит в кумуляции знания, в создании обобщенных теорий, проверка статистической значимости явно бесполезна. О воспроизводимости результатов свидетельствуют либо реальные (в терминологии Б. Томпсона «внешние») воспроизведения исследований, либо статистические процедуры вроде кросс-валидизации, расшнурованной (bootstrap) выборки или выборки «складного ножа» (jackknife) — так называемые «внутренние воспроизведения» (Daniel, 1998; Thompson, 1993, 2006).

Заблуждение 6: p как показатель практической важности полученного результата

Недопустимость интерпретации статистически значимых результатов как важных обсуждается методологами в течение восьми десятилетий, т. е. на протяжении всей истории проверки статистической значимости нуль-гипотез (Daniel, 1998). Распространенной практикой является подмена терминов, когда статистически значимые результаты, например, в корреляционной матрице отмечаются звездочками, а в конце статьи обсуждаются как просто значимые, большие и важные (Cohen, 1994, p. 1001). Хорошо известно критическое отношение практиков к результатам академических исследований, и эта критичность отчасти связана с различными критериями важности результатов (см.: Kline, 2004, ch. 1). Социальная, практическая и клиническая значимость имеют мало общего со значимостью статистической (Thompson, 2002). Так, при оценке клинической значимости обращают внимание, например, на долю пациентов, состояние которых улучшилось, и на индивидуально-ориентированные показатели, но не на значение p (Henson, 2006). Как отмечает Б. Томпсон, суждение о важности результата непременно носит ценностный характер, и ссылка на якобы объективные статистические критерии не снимает ответственности с исследователя (Thompson, 1999). Оценка важности полученных результатов, полагает Р. Кирк, есть задача ученого, а не побочный продукт статистических процедур: «Исследователь, собравший и проанализировавший данные, находится в лучшем положении для принятия решения о том, являются ли результаты тривиальными» (Kirk, 1996, p. 755). Смешение статистической и практической значимости во многом связано с неудачным оборотом, особенно когда авторы опускают слово «статистически» и пишут, например, «значимые различия» или «значимая корреляция». П. Мил характеризует последнее прилагательное

довольно сильным эпитетом – «злокачественное» (Meehl, 1978). В литературе неоднократно высказывались предложения заменить «значимость» на более точное слово, например, «надежность» (см.: Kline, 2004), хотя Р. Карвер считает эту замену крайне неудачной (Carver, 1978).

ПЗНГ как препятствие развитию науки

Ошибки в использовании и интерпретации ПЗНГ столь распространены, что напрашивается вывод о несовершенстве лежащей в ее основе статистической модели. Так, ошибочные идеи по поводу статистического вывода обнаружены в двух десятках англоязычных учебников, опубликованных в 1965–1994 гг., в том числе в книгах, написанных ведущими методологами (см.: Tryon, 2001). Из российских учебников отметим довольно популярное пособие Е.В. Сидоренко (Сидоренко, 2003), ссылки на которое встречаются в текстах диссертационных работ. Написанная остроумно и с явной заботой о студентах, эта книга является коллекцией архаизмов и искренних заблуждений.

Среди причин ошибок в использовании и интерпретации ПЗНГ в литературе называются сила инерции, закрепленная в образовательных практиках, консерватизм редакторов научных журналов, незнание истории становления теории статистического вывода, трудности с усвоением понятий теории вероятности, лингвистические факторы (использование эпитета «значимое» <различие>, сокрытие условной природы вероятностей в «ошибке I рода» и др.), а также отсутствие ясных и безукоризненных альтернатив. Эмпирические исследования показывают, что психологи и социальные исследователи не понимают логики проверки нуль-гипотез (Gigerenzer, Murray, 1987; Mittag, Thompson, 2000).

Многие авторы критикуют доминирующую статистическую модель за дихотомическое решение в пользу нулевой или альтернативной гипотез, что препятствует накоплению научных знаний (Cohen, 1994). Отчасти это обстоятельство имел в виду Р. Кирк, заявляя, что «акцент на значении p и отклонении H_0 на деле отвлекает нас от действительных целей: решения о том, подкрепляют ли данные нашу научную гипотезу и являются ли они практически значимыми или полезными» (Kirk, 1996). Р. Розноу и Р. Розенталь также заявляют, что «дихотомическая проверка значимости не имеет онтологической основы», и иронизируют: «Несомненно, Бог любит [значение] 0.06 не меньше, чем 0.05. Могут ли быть сомнения в том, что Бог рассматривает силу доказательств за нуль-гипотезу или против нее как непрерывную функцию величины p?» (Rosnow, Rosenthal, 1989, p. 1277).

Обзор литературы в данной статье выполнен в жанре критики ПЗНГ. Как бывает в научном споре, оппоненты часто риторически преувеличивают силу своих аргументов. Некоторые авторы чересчур категоричны в своих оценках: «... Почти универсальная опора на простое отвержение нуль-гипотезы как стандартный метод подтверждения содержательных теорий в неточных (soft) науках — ... одна из самых плохих вещей, случившихся в истории

психологии» (Meehl, 1978, p. 817), это «самая порочная форма научного метода» (Carver, 1993, p.288), «Наша наука заплатила высокую цену за свою ритуальную приверженность проверке значимости нуль-гипотезы» (Kirk, 1996, p.756). Однако цель, которую сформулировал Дж. Нелдер (1986) — «самая важная задача ... в развитии статистической науки — уничтожить культуру значений p ...» (цит. по: Hubbard, Lindsay, 2008, p. 9), — нам кажется слишком экстремистской.

Критика ПЗНГ не осталась без ответа. Мы не будем развернуто обсуждать здесь эти контраргументы по трем причинам. Во-первых, выступления защитников проверки значимости были заметно более малочисленными по сравнению с критическими публикациями. Во-вторых, по нашему мнению, они были менее убедительны и каждое серьезное возражение было встречено новой солидной аргументацией и критикой (напр.: Carver, 1993; Schmidt, Hunter, 1997). В-третьих, — и это самое важное — контраргументация неоднородна, и ее содержательный анализ требует отдельной большой публикации. Ограничимся очень кратким изложением этих идей.

А. Гринвалд с соавт. (Greenwald et al., 1996) показывают, что в определенных ситуациях значения p могут свидетельствовать о воспроизводимости результатов. Р. Абельсон (Abelson, 1997) считает, что ПЗНГ идеальным образом подходит для оценки пригодности модели данным, и значение p, по сути, подобно показателю пригодности GOF (goodness-of-feet), используемому, например, в структурных уравнениях или логлинейном анализе. Логика проверки значимости столь проста и элегантна, считает Р. Абельсон, что если бы этого подхода не было, его следовало бы изобрести. Согласно Р. Фрику (Frick, 1996), статистическая проверка нуль-гипотезы как раз бесполезна для проверки моделей. Из-за нелинейных отношений между переменными и ненадежности измерений в психологии практически невозможны количественные предсказания, основанные на величинах статистических эффектов. Никто не говорит, что рост фрустрации на одну единицу приводит к повышению агрессии на 0.5 единиц. Говорят о том, что фрустрация усиливает агрессию или что курение положительно связано с раком. Такие утверждения Р. Фрик называет порядковыми. Основная мысль автора в том, что проверка значимости наилучшим образом подходит для проверки такого рода законов. Эта идея получила развитие (Jones, Tukey, 2000). Вместо нуль-гипотезы о равенстве двух средних, которая, как показано в предыдущем разделе данной статьи, всегда неверна, Л. Джонс и Дж. Тьюки предложили оценивать вероятность справедливости трех утверждений: 1) $\mu_1 - \mu_2 > 0$, 2) $\mu_1 - \mu_2 < 0$ и 3) знак $\mu_1 - \mu_2$ (пока) не определен. Модель Джонса и Тьюки следует рассматривать не как модификацию традиционной ПЗНГ, а как альтернативу ей. С. Чоу (1988) доказывает, что бинарное решение (отклонить/не отклонять нуль-гипотезу) согласуется с идеей поступательного накопления научных данных, если правильно сформулировать и содержательно обосновать нуль-гипотезу.

Следует отметить, что даже самые воинственные противники ПЗНГ не исключали полезности этой процедуры в двух случаях. Первый — так называемая сильная форма проверки (Meehl, 1978; см. также: Cohen, 1994), когда содержательная теория сформулирована ясно и точно, так что при опровержении (фальсификации) центрального предсказания теории она признается ложной. Подобные концепции являются идеалом научности для К. Поппера, и они типичны для точных наук вроде физики. Второй случай касается «сельскохозяйственной модели» науки (Meehl, 1978; Chow, 1988, 1999). Ее суть в том, что экспериментальная манипуляция (внесение удобрения) является содержательной, сущностной, т. е. нет необходимости в гипотетических конструктах и индикаторах. В этих условиях отвержение нуль-гипотезы также идентично отвержению теории и оценка вероятности истинности нуль-гипотезы становится избыточной. Нет необходимости доказывать, что эти два случая не имеют отношения к психологии — «мягкой» науке (Meehl, 1988), — понятия которой гипотетичны, противоречивы и недостаточно ясно определены.

Запрет ПЗНГ, за который выступили Р. Карвер, П. Мил, Ф. Шмидт, Дж. Шейвер и др., невозможен хотя бы потому, что принятая практика проверки значимости глубоко укорена, «остается удивительно стабильной в течение полувека» (Finch, Cumming, 2001, p. 197) и в сознании нескольких поколений исследователей связывается со стандартами научности и объективности. Как отмечает Д. Зон, уже во время учебы в университете психологи научаются «рационализировать свои скромные результаты» с помощью статистики (Sohn, 1998, p. 307). У. Трайон делает пессимистичный вывод: «Значительные и повторяющиеся усилия, предпринятые за прошедшие 75 лет для устранения неправильного использования ПЗНГ <...>, не были продуктивными. Нет оснований ожидать, что дальнейшие усилия в этом направлении будут полезными» (Tryon, 2001, p. 372). В редакционной статье в «Journal of Applied Psychology» Дж. Кэмпбелл сетует: «Возможно, значения *p* подобны комарам. Они занимают эволюционную нишу, и никакие почесывания, хлопки или опрыскивания не изгоняют их» (Campbell, 1982, цит. по: Finch, Cumming, 2001, p. 186).

В поисках альтернатив

В этом разделе мы рассмотрим некоторые предложения рабочей группы Американской психологической ассоциации, созданной для разрешения споров по поводу использования статистических методов. Отчет рабочей группы (Wilkinson and TFSI, 1999) составлен как руководство по использованию статистических методов в научной публикации: при планировании, проведении и описании психологических исследований. Структура руководства соответствует типичному формату эмпирических статей в научных журналах: метод (дизайн, определение генеральной совокупности, выборка, распределение испытуемых по условиям, измерение), результаты (поиск аномалий в данных, анализ) и обсуждение результатов (интерпретация и

выводы). Вопросы, рассматриваемые в данной статье, обсуждаются в разделе «Анализ». Априорная оценка статистической мощности обсуждается в разделе «Метод». Пересказ и тем более оценка всех предложений рабочей группы выходит далеко за рамки данной статьи, к тому же в тексте прослеживаются непоследовательность и компромиссы между авторами, стоящими на разных позициях. Было бы полезно опубликовать перевод полного текста отчета в одном из российских психологических журналов. Некоторые рекомендации — особенно по поводу формирования выборки, определения ключевых переменных, более широкого использования методов графического анализа (в том числе для проверки предположений статистических моделей), проверки множественных гипотез, выбора простейшего из возможных методов анализа, каузального вывода — чрезвычайно актуальны для исследовательской практики. Мы извлекли из текста четыре ключевых тезиса по поводу ПЗНГ и альтернативных ей подходов.

1. Руководство предписывает сообщать не только объем выборки, но и «информацию о... процессе принятия решения об объеме выборки», в том числе размеры статистических эффектов. И далее: «Расчет мощности имеет наибольший смысл, если он сделан до сбора данных. Поэтому важно показать, каким образом показатели размера эффекта были получены из предыдущих исследований и теории, с тем чтобы рассеять подозрения в том, что они могли быть рассчитаны на основании данных, использованных в исследовании, или, того хуже, сфальсифицированы с целью оправдать размер выборки» (Wilkinson and TFSI, 1999, p. 596). Априорный анализ мощности требует поиска в литературе наилучших оценок величин статистических эффектов и потому «стимулирует более серьезное отношение авторов к предшествующим исследованиям и теории в их области и лишь немногим оставляет... возможность утверждать, что не существует серьезных работ, кроме данного исследования» (там же)[8]. Отметим, что внимание рабочей группы к статистической мощности свидетельствует о нежелании или неготовности отказаться от ПЗНГ, поскольку в принятой сегодня теории статистического вывода оба понятия тесно связаны: «Если проверка значимости больше не используется, понятие статистической мощности избыточно и не имеет смысла» (Schmidt, 1996, p. 124).

2. Рекомендацию по поводу проверки гипотез приведем целиком: «Сложно представить ситуацию, когда дихотомическое решение «принять–отвергнуть» предпочтительнее сообщения точного значения *p* и особенно доверительного интервала. Не следует использовать неудачное выражение “принимается нулевая гипотеза”. Значения *p* должны всегда сопровождаться информацией о величине эффекта. Статья Коэна

[8]В отечественных публикациях и диссертационных работах бездоказательные утверждения о недостаточной изученности проблемы и об отсутствии убедительных исследований стали практически ритуальными.

(Cohen, 1994) на эту тему будет особенно полезна для психологов» (Wilkinson and TFSI, 1999, p. 599). Этот фрагмент особенно противоречив. ПЗНГ не отвергается, хотя оборот «сложно представить ситуацию» ставит ее под сомнение. Необходимость расчета доверительных интервалов, продвигаемая противниками ПЗНГ, не исключает использования фишеровского значения *p*. Вторая фраза в рекомендации рабочей группы отдает преимущество фишеровской теории над концепцией Неймана-Пирсона. Реформаторский посыл содержится лишь в требовании всегда сообщать величину статистического эффекта наряду с традиционным *p*.

3. Позиция рабочей группы по поводу величин статистических эффектов более определенна: «Следует всегда приводить величины эффектов для основных результатов» (Wilkinson and TFSI, 1999, p. 599). Посетовав на то, что мягкий совет авторам статей сообщать величину эффекта, высказанный в предыдущем (4-м) издании «Руководства Американской психологической ассоциации по подготовке публикаций», не оказал на них никакого влияния, члены рабочей группы повторяют и разъясняют свою позицию: «Мы снова подчеркиваем, что сообщение и интерпретация величин эффектов в контексте ранее полученных результатов является неотъемлемой частью хорошего исследования. Это дает возможность читателю оценить стабильность результатов относительно выборок, дизайнов и способов анализа данных. Сообщение величин эффектов дает информацию для анализа мощности и метаанализа в последующих исследованиях» (Wilkinson and TFSI, 1999, p. 599).

4. Руководство придает особую ценность построению доверительных интервалов «для всех величин эффектов основных результатов», в том числе «для корреляций и других коэффициентов ассоциации и дисперсии» (Wilkinson and TFSI, 1999, p. 599). Многие профессиональные статистики и эксперты в других областях знания предпочитают доверительные интервалы точечным оценкам. Среди прочего это связано с тем, что «сравнение доверительных интервалов, полученных в данном исследовании, с интервалами, приведенными в других исследованиях, способствует привлечению внимания к вопросу стабильности результатов. Накопление данных об интервалах, полученных в различных исследованиях, позволяет установить вероятные области нахождения параметров генеральной совокупности» (там же). К сожалению, построение доверительных интервалов для статистических эффектов выходит за рамки элементарной статистики, так как основано на нецентральных теоретических распределениях.

Обсуждаемые в литературе альтернативы ПЗНГ также не лишены недостатков. По этому поводу Дж. Коэн высказался недвусмысленно: «Не ищите магическую альтернативу статистической проверке нуль-гипотезы... Ее не существует» (Cohen, 1994, p. 1001).

Литература

Ивин А.А. Логика. М.: Фаир-Пресс, 2003.

Наследов А.Д. Математические методы психологического исследования. Анализ и интерпретация данных. 2-е изд. СПб.: Речь, 2006.

Сидоренко Е.В. Методы статистической обработки в психологии. СПб.: Речь, 2003.

Стерн Дж., Смит Дж. Отсеивание фактов: почему нас не удовлетворяют статистические критерии значимости? // Обзор современной психиатрии. 2003. Вып. 18.

Тверски А., Канеман Д. Вера в закон малых чисел // Под ред. Д. Канемана, П. Словика, А. Тверски. Принятие решений в неопределенности: правила и предубеждения. Харьков: Гуманитарный центр, 2005. С. 39–48.

Фишер Р.А. Статистические методы для исследователей. М.: Госстатиздат, 1958.

Abelson R.P. A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented) // L.L. Harlow, S.A. Mulaik, J.H. Steiger (eds.). What if there were no significance tests? Mahwah, NJ: Lawrence Erlbaum, 1997. P. 117–141.

American Psychological Association. Publication manual of the American Psychological Association. 5th ed. Washington, DC: APA, 2001.

Bakan D. The test of significance in psychological research // Psychological Bulletin. 1966. 66. 6. 423–437.

Brewer J.K. Behavioral statistics textbooks: Source of myths and misconceptions? // Journal of Experimental Education. 1985. 10. 3. 252–268.

Carver R.P. The case against statistical significance testing // Harvard Educational Review. 1978. 48. 3. 378–399.

Carver R.P. The case against statistical significance testing, revisited // Journal of Experimental Education. 1993. 61. 2. 287–292.

Chow S.L. Significance test or effect size? //Psychological Bulletin. 1988. 103. 1. 105–110.

Chow S.L. What statistical significance means Theory & Psychology. 1998. 8. 3. 329–3330.

Cohen J. The Earth is round ($p < .05$) // American Psychologist. 1994. 49. 12. 997–1003.

Daniel L.G. Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals // Research in the Schools. 1998. 5. 2. 23–32.

Falk R., Greenbaum C.W. Significance tests die hard: The amazing persistence of a probabilistic misconception // Theory & Psychology. 1995. 5. 1. 75–98.

Finch S., Cumming G., Thomason N. Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform // Educational and Psychological Measurement. 2001. 61. 2. 181–210.

Fisher R.A. The design of experiments. 8th ed. Edinburgh: Oliver and Boyd, 1966.

Frick R.W. The appropriate use of null hypothesis testing // Psychological Methods. 1996. 1. 4. 379–390.

Gigerenzer G. The superego, the ego, and the id in statistical reasoning // G. Keren, C. Lewis (eds.). A handbook for data analysis in the behavioral sciences: Methodological issues / Hillsdale, N.J.: Erlbaum, 1993. 311–339.

Gigerenzer G., Murray D.J. Cognition as intuitive statistics. Hillsdale, NJ: Erlbaum, 1987.

Greenwald A.G., Gonzalez R., Harris R.J., Guthrie D. Effect size and p-values: What

should be reported and what should be replicated? // Psychophysiology. 1996. 33. 2. 175–183.

Hubbard R. Alphabet soup: Blurring the distinctions between p's and α's in psychological research // Theory & Psychology. 2004. 14. 3. 295–327.

Hubbard R., Armstrong J.S. Why we don't really know what statistcal significance means: Implication for educators // Journal of Marketing Education. 2006. 28. 2. 114–120.

Hubbard R., Bayarri M.J. Confusing over measures of evidence (p's) versus errors (α's) in classical statistical testing // The American Statistician. 2003. 57. 3. 171–178.

Hubbard R., Lindsay R.M. Why P values are not a useful measure of evidence in statistical significance testing // Theory & Psychology. 2008. 18. 1. 69–88.

Hubbard R., Parsa R.A., Luthy M.R. The spread of statistical significance testing in psychology: the case of the Journal of Applied Psychology, 1917–1994 // Theory & Psychology. 1997. 7. 4. 545–554.

Hubbard R., Ryan P.A. The historical growth of statistical significance testing in psychology — and its future prospects // Educational and Psychological Measurement. 2000. 60. 5. 661–681.

Huberty C.J Historical origins of statistical testing practices: The treatment of Fisher versus Neyman–Pearson views in textbooks // Journal of Experimental Education. 1993. 61. 4. 317–333.

Huberty C.J, Pike C.J. On some history regarding statistical testing // B. Thompson (eds.). Advances in social science methodology. Stamford, CT: JAI Press, 1999. 5. 1–22.

Jones L.V., Tukey J.W. A sensible formulation of the significance test // Psychological Methods. 2000. 5. 4. P. 411–414.

Kline R.B. Beyond significance testing: Performing data analysis methods in behavioral research. APA, 2004.

Meehl P.E. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology // Journal of consulting and clinical psychology. 1978. 46. 4. 806–834.

Meehl P. Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it // Psychological Inquiry. 1990. 1. 2. 108–141.

Miller J. What is the probability of replicating a statistically significant effect? // Psychonomic Bulletin & Review. 2009. 16. 4. 617–640.

Mittag K.C., Thompson B. A national survey of AERA members' perceptions of statistical significance tests and other statistical issues // Educational Researcher. 2000. 29. 1. 14–20.

Nix T.W., Barnette J.J. The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing // Research in the Schools. 1998. 5. 1. 3–14.

Oakes M. Statistical inference: A commentary for the social and behavioural sciences. Chichester: Wiley, 1986.

Rosnow R.L., Rosenthal R. Statistical procedures and the justification of knowledge in psychological science // American Psychologist. 1989. 44. 10. 1276–1284.

Schmidt F. Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers // Psychological Methods. 1996. 1. 2. 115–129.

Schmidt F.L., Hunter J.E. Eight common but false objections to the discontinuation of significance testing in the analysis of research data // L.L. Harlow, S.A. Mulaik, J.H. Steiger (eds.). What if there were no significance tests? Mahwah, NJ: Erlbaum, 1997. 37–64

Sedlmeier P., Gigerenzer G. Do studies of statistical power have an effect on the power of studies // Psychological bulletin. 1989. 105. 2.. 309–316.

Sohn D. Statistical significance and replicability: Why the former does not presage the latter? // Theory & Psychology. 1998. 8. 3. 291–311.

Thompson B. The use of statistical significance tests in research: Bootstrap and other alternatives // Journal of Experimental Education. 1993. 61. 4. 361–377.

Thompson B. If statistical significance tests are broken/misused, what practices should supplement or replace them? // Theory and psychology. 1999. 9. 2. 165–181.

Thompson B. «Statistical», «practical» and «clinical»: How many kinds of significance do counselors need to consider? // Journal of Counseling and Development. 2002. 80. 1. 64–71.

Thompson B. Foundations of behavioral statistics: An insight-based approach. N.Y: Guilford, 2006.

Tryon W.W. Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests // Psychological methods. 2001. 6. 4. 371–386.

Tukey J.W. The philosophy of multiple comparisons // Statistical science. 1991. 6. 1. 100–116.

Wilkinson L. and Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations // American Psychologist. 1999. 54. 8. 594–604.