Корнеев А.А., Рассказова Е.И., Кричевец А.Н., Койфман А.Я. Критика методологии проверки нулевой гипотезы: ограничения и возможные пути выхода. Часть I

-

English version: Korneev A.A., Rasskazova E.I., Krichevets A.N., Koyfman A.Ya. Criticism of Null Hypothesis Significance Testing: Limitations and Possible Ways Out. Part I

Московский государственный университет имени М.В.Ломоносова, Москва, Россия

Сведения об авторах
Литература
Ссылка для цитирования

Статья посвящена критическому обсуждению проблемы статистического вывода и методологии проверки нулевой гипотезы. В ней рассматриваются основные недостатки этого подхода к статистическому оцениванию данных. Выделяются несколько уровней критики проверки нулевой гипотезы: собственно статистический, связанный с процедурами и допущениями, стоящими за этой методологией, уровень социальных последствий, связанных с доминированием данного подхода в статистике, приводящий к ошибкам в интерпретации получаемых результатов, и, наконец, уровень соотнесения статистического и содержательного (психологического) анализа. Далее рассматриваются основные альтернативы, предлагающиеся в настоящее время для преодоления проблем, вызванных использованием методологии проверки нулевой гипотезы, дается их критическая оценка. Формулируется предварительный вывод о недостаточности изменения способов оценивания изолированного исследования.

Ключевые слова: проверка нулевой гипотезы, статистическое оценивание, значимость, мета-анализ

В мировой психологии на протяжении десятилетий множатся публикации, критикующие доминирующий в современной психологической науке подход к статистическому оцениванию результатов исследований и отбору работ для публикации. В последнее время, похоже, критика стала достигать цели, и количество практических следствий увеличивается. Так еще в 1999 году в рамках Американской психологической ассоциации была создана the Task Force on Statistical Inference. Этой рабочей группой обсуждались изменения правил оформления публикаций Американской психологической ассоциации, были сформулированы новые требования к описанию результатов статистического анализа: включение в работы отчетов о величине статистического эффекта, доверительных интервалов и т.п. [Wilkinson, 1999]. В 2014 году в журнале Nature вышла статья, содержащая критику использования традиционного уровня значимости в статистическом выводе [Nuzzo, 2014]. В 2015 году редакция достаточно известного психологического журнала Basic and Applied Social Psychology объявила, что перестает принимать статьи с обсуждением результатов, полученных с помощью традиционного статистического вывода на основании уровня значимости [Trafimow, Marks, 2015].

В русскоязычном сегменте психологических исследований стало обычным пенять коллегам на недостаточное внимание к современным тенденциям в мировой психологии. Мы опираемся на довольно большой массив англоязычных публикаций на нашу тему, однако ставим несколько более

тонкую задачу. Происходящая в настоящее время смена требований к представлению отчетов об исследованиях в ведущие журналы и без нашего скромного вклада изменит способы обработки данных в русскоязычном сообществе, однако мы опасаемся, что в таком случае в массовой науке[1] произойдет замена одной ошибки на другую. Опасность заключается в упрощенном понимании глубокой проблемы и в переписывании системы знаний в набор рецептов, которые всегда оказываются под угрозой, точно описанной в народной сказке «Таскать вам, не перетаскать».

В обсуждаемых статьях предлагаются различные варианты и аспекты реформы требований к статистической обработке данных, из которых мы поддержим три:

1) предварительная регистрация заявок на проведение исследования с описанием предполагаемых условий исследований, объемов выборок и т.п.;

2) более подробное описание данных, условий их сбора и способа обработки;

3) публикация не только статистически значимых результатов, но и результатов, в которых нуль-гипотеза не отвергнута.

Остальные предложения: публикация результатов, описанных в терминах доверительных интервалов и величины статистического эффекта (Effect Size, см. [Kelley, Peacher, 2012]), полный отказ от описания в терминах тестирования нуль-гипотезы (мы будем далее использовать общепринятую аббревиатуру NHST – Null Hypothesis Significance Testing), предпочтение байесовских оценок, использование методов рандомизации – представляются нам более или менее разумными, в зависимости от того, будут ли они сопровождаться углублением понимания проблем в массовой науке.

Мы также разберем (в основном, в части II статьи) вопросы, связанные с растущим распространением мета-анализа и надеждами на него как на радикальное средство решения методологических проблем статистики в психологии и других областях исследований.

В чем проблема?

Распространенная и пока господствующая в области обработки психологических данных парадигма NHST справедливо обвиняется в многообразных грехах. С нашей точки зрения, критика эта неоднородна и относится к трем разным уровням.

Первый уровень составляет собственно *критика статистических процедур и допущений, стоящих за NHST*. Возможен ли статистический вывод в принципе, насколько исследователь может быть в нем уверен и насколько полезную информацию он получает?

На втором уровне – *критика социальных последствий применения NHST* – основное внимание уделяется тому, что и сами процедуры NHST, и их повсеместное принятие в науке провоцируют ошибки в интерпретации результатов, как бы направляя исследователя по ложному пути (misleadings). Ключевой вопрос этого уровня – какие ошибки склонны делать люди, интерпретируя результаты правильно проведенных процедур NHST, и как эти ошибки сказываются на развитии психологии?

Содержательно примыкает к нему третий уровень критики, связанный с *трудностями соотнесения статистического и содержательного (психологического) уровней анализа*. Переход к содержательному психологическому выводу на основе принятого статистически решения выводит исследователя за пределы статистики и требует учета дизайна исследования, практической значимости, проблем репрезентативности и т.п.

1. Центральное основание критики статистических процедур и допущений, стоящих за NHST, – в чрезмерной узости статистического вывода, если вывод этот сделан правильно.

– Общая логика подхода, строго говоря, не соответствует тому вопросу, который интересует исследователя. В известной работе Дж.Коэна [Cohen, 1994] отмечается, что уровень значимости не отвечает на вопрос «Насколько вероятно, что верна гипотеза H_0?».

– Некоторые авторы называют саму методологию проверки нулевой гипотезы «логически бессмысленной», так как эта проверка не является доказательством истинности или ложности статистической гипотезы. В рамках подхода Р.Фишера нулевую гипотезу можно отвергнуть, но нельзя принять [Gigerenzer, 2004]. Это отражается в формулировках, принятых в английском языке. При проверке нулевой гипотезы мы можем либо ее «отвергнуть» (reject), либо «не суметь отвергнуть» (fail to reject), но никогда – «принять» (accept). В рамках подхода Дж.Неймана и Э.Пирсона выбор нулевой или альтернативной гипотезы основан на допущении, что одна из них безусловно истинна. То есть в рамках NHST исследователю приходится принимать однозначное решение по типу «да / нет», хотя граница между этими вариантами условна [Cohen, 1994; Gigerenzer, 2004; Nuzzo, 2014 и др.].

– В рамках NHST анализируются результаты отдельного исследования, а на основании полученных статистических показателей делаются широкие обобщения. При этом не учитываются предшествующие исследования и вообще какие-либо данные за пределами полученного набора данных [Carver, 1978; Robinson, Wainer, 2001; Schneider, 2015]. На уровне теоретического обсуждения это, как правило, не так, но на уровне статистики при проверке гипотез в рамках NHST работа осуществляется в вакууме, без учета уже существующих исследований.

– Классический вариант проверки нулевой гипотезы очень часто требует труднопроверяемых допущений о распределении оцениваемого параметра, которые трудно обосновать, а иногда они вовсе противоречат интуиции и математической модели. Кроме того, есть требование строгой рандомизации выборок, которое соблюдается далеко не всегда [Purssell, While, 2011].

2. Узость статистического вывода имеет социальные последствия: исследователь настолько хочет узнать, верна ли его гипотеза, что пытается интерпретировать полученные значения именно в таком ключе.

Ошибка состоит в том, что исследователь делает однозначный вывод об истинности одной из гипотез и ложности другой; при этом он «подменяет» оценку условной вероятности, которую он может рассчитать (вероятности получить имеющиеся эмпирические результаты при условии, что нулевая гипотеза верна), обратной вероятностью – что нулевая гипотеза верна при имеющихся результатах.

– Еще одна проблема связана с выбором исследователем критического значения a (например, 0,05), который обосновывается апелляцией к традициям и социальным конвенциям. На практике это приводит к интерпретации «близких» уровней значимости как «субзначимых» (marginally significant), что, в свою очередь, нарушает исходные положения проверки нулевой гипотезы. При анализе публикаций в трех ведущих психологических журналах за год было показано, что количество работ, в которых получен уровень значимости немногим ниже 0,05 (0,04–0,05), больше, чем ожидалось бы статистически, исходя из количества других уровней значимости [Masicampo, Lalande, 2012]. То есть работы с уровнем значимости «чуть ниже 0,05» публикуются чаще, чем можно было бы предполагать. Одно из возможных объяснений: чрезмерный акцент на важности получения значимости ниже 0,05 может провоцировать исследователей добиваться достижения желаемой границы, а издателей – к большей готовности издавать эти работы.

3. Если предыдущая группа аргументов относится к ситуации, когда вывод на основе NHST искажается самим исследователем, то следующая относится к критике недостаточной

соотнесенности статистического и содержательного уровней анализа.

– Ключевой аргумент носит теоретический характер: опора на однозначные «да / нет» решения приводит к тому, что психология сохраняет и приумножает только выводы о наличии эффекта, без учета его величины, практической значимости и условий получения. Это значит, что возможности накопления и обобщения знаний крайне малы [Cohen, 1994]. Дополнительный довод связан с тем, что в психологическом исследовании нулевая гипотеза обычно формулируется как гипотеза о равенстве нулю некоторой характеристики случайных величин (математического ожидания, корреляции). В реальности эти характеристики всегда отличаются от нуля хотя бы на ничтожную величину, и при достаточно больших выборках это отличие окажется значимым [Cohen, 1994]. Автор отмечает, что оба эти недостатка могут быть преодолены, например, при формулировке нулевой гипотезы о попадании тестируемого значения в некоторый интервал или оценке и сопоставлении величин статистического эффекта (мета-анализ).

– Категоричность выводов на основе NHST провоцирует ошибки смешения статистической и содержательной значимости. Содержательная значимость результата связана, скорее, с надежностью и воспроизводимостью результата, а эти вопросы проверка нулевой гипотезы никак не проясняет [Fraley, Marks, 2007]. При этом возможности современных статистических методов обработки провоцируют исследователя интерпретировать данные без учета того, как они получены, – репрезентативности выборок, дизайна исследования, практической значимости. Один из путей решения проблемы – терминологический, использование более «осторожных» терминов на статистическом уровне анализа, в меньшей степени провоцирующих безосновательные выводы о причинно-следственной связи и механизмах (например, «эффект» вместо «предсказание», «непрямой эффект» вместо «медиации» и т.п., см. [Little, 2013]).

Критика и альтернативы первого уровня

С нашей точки зрения, разведение уровней критики важно, поскольку позволяет понять разницу во взглядах на возможности преодоления недостатков NHST. На первом, статистическом, уровне критики никаких альтернатив, полностью снимающих ограничения NHST, нет [Cohen, 1994] – отчасти потому, что вывод всегда остается вероятностным, отчасти потому, что, как мы постараемся показать ниже, другие альтернативы также обладают рядом серьезных ограничений. Перевод проблемы в социальную плоскость открывает значительно больше возможностей и позволяет даже говорить о необходимости глобальных изменений в психологическом сообществе. Сюда относятся как способы привлечь внимание к вероятностной природе статистического вывода за счет лучшего обучения или требования дополнять публикации указанием доверительных интервалов и величины статистического эффекта, так и способы, направленные на уточнение языка статистического вывода таким образом, чтобы он соотносился с выводом эмпирическим. Именно здесь мы видим путь к улучшению ситуации, поскольку, по нашему мнению:

1) все недостатки господствующей практики вытекают из одного – неправильной интерпретации NHST подхода, которая захватывает не только слой массовой науки, но проникает в учебники, написанные специалистами самого высокого уровня в нашем сообществе, см. обсуждения [Gliner et al., 2002; Sotos et al., 2007];

2) правильная интерпретация подхода обескураживающе бедна, но это не следствие каких-то недоработок или ошибок, а связано с самой сущностью проблемы индуктивного вывода и не может быть легко преодолено на тех путях, которые предлагают ее оппоненты – альтернативы столь же обескураживающе бедны при их корректном применении.

Критика, не предлагающая альтернатив

В англоязычном журнальном пространстве количество цитирований работ, радикально

критикующих процедуру NHST, измеряется сотнями. На русском языке поднимаемые нами проблемы обсуждаются в единичных работах [Сивуха, Козяк, 2009; Сивуха 2014; Алексеев, 2012], которые к тому же не порождают дискуссий. В первой из этих работ дана содержательная критика самой процедуры NHST. Вслед за классическими обзорами по проблеме методологии NHST [Gigerenzer, 2004] авторы утверждают, что в процедуре NHST смешаны два подхода (оценка значимости Фишера и проверка гипотез Неймана–Пирсона), что полученная «смесь» противоречива и что в результате этого выводы, сделанные с помощью этой процедуры, слишком часто оказываются ложными. Соглашаясь с последним тезисом, заметим, что вся цепочка рассуждений кажется нам слишком радикальной и не совсем точной. Основные трудности NHST проистекают не из смешения подходов и не могут быть решены их строгим разведением.

Авторы цитируют работу Фишера, где говорится о принятии или отвержении гипотезы (которая в современных терминах обозначается H0) без противопоставления ей определенной альтернативы. В этом контексте определяется значимость, измеряемая вероятностью получения данного или более «экстремального» для гипотезы H_0 значения выбранной статистики. Значимость говорит о *степени* основательности решения об отвержении нулевой гипотезы.

В то же время подход Неймана–Пирсона предлагает обоснование выбора из двух определенных альтернатив, и решение принимается при заранее фиксированном критическом значении, разделяющем гипотезы. Ошибка смешения двух подходов «состоит в том, что исследователь выбирает уровень значимости (допустимую величину *p*) и в случае получения значимого результата, то есть 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 *p* установлена до проверки гипотезы, решение дихотомично: H0 верна или неверна» [Сивуха, Козяк, 2009, с. 72].

Отметим, что цитата из Фишера касается специального случая критерия хи-квадрат, оценивающего согласованность полученных результатов с гипотезой о данном теоретическом распределении породившей их случайной величины. В этом случае нет необходимости определять альтернативу: *всякое* другое распределение будет в среднем порождать более высокие значения статистики хи-квадрат [Фишер, 1958, с. 71].

Что касается смешения подходов, то, по нашему мнению, ситуацию можно назвать продуктивной прививкой элементов фишеровского подхода на «ствол» Неймана–Пирсона. При этом введение обычая приводить «точное значение *p* как меру доказательности ошибочности нуль-гипотезы», по всей видимости, никак не связано с Фишером, а вызвано переходом от сравнения результатов с бумажными таблицами к обработке компьютерными пакетами, для которых вычисление точного значения *p* (вероятности получить данное или большее значение статистики при условии истинности нуль-гипотезы) не представляет принципиальной трудности.

В терминах Неймана–Пирсона (хорошее изложение основ подхода можно найти в старой книге [Артемьева, Мартынов, 1975]) можно интерпретировать приводимое значение *p* в современной «эклектической» стандартной процедуре так. Для примера рассмотрим односторонний критерий и *p*, равное *0,02*: во-первых, подразумевается, что используемая статистика корректна в смысле Неймана–Пирсона – выполнены все условия, касающиеся предполагаемого распределения результатов испытаний. Во-вторых, при установлении критерия для любого уровня значимости большего, чем 0,02, полученное значение статистики попадет в критическую область, и нулевая гипотеза будет отвергнута в пользу альтернативы. В-третьих, поскольку среди этих больших значений имеется социально маркированное значение 0,05, полученный результат можно интерпретировать как социально приемлемое подтверждение альтернативной гипотезы, и отчет может быть опубликован.

На наш взгляд, первые два пункта безупречны, под вопросом только третий пункт. Именно социальная организация, определяющая дальнейшую судьбу подобных результатов, задает те

негативные последствия, которые вызывают справедливую критику, часто, однако, ошибающуюся адресом. То, что величина уровня значимости нередко ошибочно интерпретируется (особенно студентами) как величина эффекта, не есть недостаток NHST как таковой, но относится к психолого-педагогическому аспекту его понимания и применения.

Возможные альтернативы

Подобно тому, как отвержение нулевой гипотезы имеет смысл ввиду какой-то подразумеваемой альтернативы, так и отвержение самой процедуры NHST подразумевает альтернативные процедуры, которые также имеют свои недостатки. Кратко остановимся на основных альтернативах, предлагаемых критиками NHST.

Критики утверждают, что отчеты, содержащие информацию о величине эффекта и доверительных интервалах, акцентирующие оценку параметров вместо проверки гипотез, истолковывающие оценки в байесовском духе, блокируют вредную практику рассматривать статистическое оценивание как доказательство истинности или ложности гипотез (дихотомическое мышление, как называют его критики, см. [Hoekstra et al., 2006]). Мы согласны с этим тезисом, однако укажем теперь на аналогичные опасности и предлагаемых альтернатив.

Байесовский подход

Изначально он направлен на преодоление ошибки интерпретации значимости[2] *p* как вероятности истинности нулевой гипотезы. Критики (в частности, и авторы упоминавшейся работы [Сивуха, Козяк, 2009] справедливо указывают, что в этом случае смешивают условные вероятности $p(D|H_0)$ и $p(H_0|D)$ (где D – обозначение выборки). Исследователя интересует вопрос о степени подтверждения / опровержения гипотезы H_0, то есть вторая из этих вероятностей, в то время как уровень значимости *p* представляет первую[3].

В качестве иллюстрации того, что эти вероятности могут существенно различаться, авторы пользуются примером, взятым из классической работы [Cohen, 1994][4]. Речь идет о шизофреническом расстройстве и тесте, его диагностирующем: «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, то есть $P(D|H_0) = 0{,}03$. 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 $p = 0{,}03$!» [Там же. С. 74].

Пример ясно показывает, что нет прямой связи между значимостью *p* и вероятностью истинности гипотезы H_0. Однако он показывает также, что байесовский подход корректно работает лишь в очень специфических ситуациях, когда вероятности, входящие в формулу, вычисляются «классическим» образом, например, исходя из генеральных совокупностей. Использование формулы Байеса в иных случаях опирается на так называемые субъективные вероятности и приводят в итоге к ним же. Простая модификация приведенного выше примера показывает, что

байесовская статистика приводит к проблемам не менее трудным. Пусть наш тест выявляет способность ясновидения и состоит в угадывании результатов десяти испытаний какой-то бинарной случайной величины, например подбрасывания симметричной монеты. Пусть гипотеза H_0 в этом случае состоит в том, что вероятность каждого из десяти угадываний равна 0,5, а гипотеза H_1 – что она равна 0,9 (хотя обычно мы не в состоянии столь точно сформулировать H_1). Предположим, что испытуемый угадал все 10 результатов. Мы можем вычислить $P(10|H_0) = 0{,}001$ и $P(10|H_1) = 0{,}348$. Это значит, что если мы присвоим априори $P(H_0) = P(H_1) = 0{,}5$, выражая этим полную индифферентность к гипотезам, то после байесовского пересчета мы получим $P(H_1|10) = 1 – p(H_0|10) = 348/349$. То есть данный испытуемый является ясновидящим с вероятностью 0,997. При этом ни априорная вероятность гипотезы H_1, ни апостериорная ее вероятность не могут быть истолкованы ни в терминах классической вероятности, ни в частотных терминах. Речь идет только о субъективной вере. В психологических исследованиях вопрос о данном конкретном испытуемом чаще всего не ставится, обсуждаются более общие утверждения – в данном случае теоретическая гипотеза, стоящая за экспериментом, могла бы быть гипотезой о возможности ясновидения как такового. Один из нас категорически не верит в ясновидение и предлагает принять такое априорное распределение вероятностей гипотез: $P(H_0) = 1$, $P(H_1) = 0$. Тогда байесовские апостериорные вероятности равны априорным, то есть эксперимент не поколеблет его уверенности.

Правильнее всего было бы в нашем случае говорить о том, что каждый результат эксперимента задает байесовский оператор, преобразующий любой априорный набор субъективных вер в набор апостериорных субъективных вер – при полном отсутствии оснований для предпочтения тех или иных априорных вер. Исследователи, обсуждающие байесовский подход, говорят о возможности опоры на предшествующие исследования как основания для введения тех или иных априорных вероятностей, но вполне ясной процедуры не предлагается, и в целом они действительно рассматриваются как результат «убеждений» исследователя («personal belief of the researcher», [Eidswick, 2012, p. 8]). Вряд ли такая интерпретация может быть принята как норма при подготовке статей. Если вместо ясновидения взять более привычные для психолога задачи, например оценку эффективности какого-то воздействия, то логическая структура аргументов окажется не затронутой, лишь изменится эмоциональный фон поставленной задачи.

К примеру, приведенному Сивухой и Козяк, можно было бы добавить и соображения, связанные с *ценой* ошибок первого и второго рода[5], которые могут быть приняты во внимание в некоторых случаях. Такие соображения вписываются в байесовский подход даже, пожалуй, хуже, чем в NHST, поскольку цены ошибок надо перевести здесь в априорные вероятности. В случае ясновидения соответствующие соображения могли бы выглядеть так: «признание ясновидения противоречит преобладающим представлениям, цена принятия H_1 слишком высока (например, придется переписывать учебники), поэтому будем считать априорную вероятность H_1 близкой к нулю – 0,000001». Для NHST изменение цены ошибки будет отражаться только в изменении уровня значимости критерия отвержения H_0.

Как мы видим, логические основания для принятия общих решений, опираясь на какие-то эмпирические результаты, весьма бедны (это верно и для NHST, и для байесовского подхода), и оставляют слишком много неопределенных аспектов, которые могут доопределяться различным образом, исходя из соображений прагматических, мировоззренческих и иных. Мы настаиваем, что хотя эти соображения не могут быть приведены в стройную систему, порождающую в конкретных случаях конкретные рекомендации, они не должны игнорироваться и вытесняться, как произошло в случае NHST. Попытки превратить использование статистики в процедуры, описываемые алгоритмического типа инструкциями, привели к последствиям, вызывающим столько нареканий. Выбор универсального уровня значимости 0,05 – только один из этих грубых просчетов. Лишь постольку, поскольку этот показатель рассматривается как универсальный, возникает соблазн упрощать суждения до уровня «значение *p* равно 0,02, следовательно, гипотеза H_0 отвергается, а

H_1, тем самым, истинна».

Доверительные интервалы

Одно из предложений реформаторов – перейти от NHST к оцениванию параметров с помощью доверительных интервалов. В ранних работах доверительные интервалы предлагаются как форма отчета о величине эффекта, привлекающая внимание к тому, что даже высоко статистически значимые результаты и оценки могут быть случайными [Cohen, 1994]. В статье С.В.Сивухи и А.А.Козяк [Сивуха, Козяк, 2009] эта часть работы Дж.Коэна рассматривается только в свете связи уровня значимости и объема выборки, однако, сам Дж.Коэн указывает, что проблема шире: непредсказуемое действие фактора «корреляционного шума»[6] приводит к тому, что многие «случайные» с точки зрения содержательной интерпретации и проведения исследования корреляции принимают довольно большие по модулю значения [Cohen, 1994, p. 1000]. С его точки зрения, доверительные интервалы не являются способом преодоления этой проблемы, а лишь удачным напоминанием исследователю о ее существовании.

Действительно, математически эквивалентно сообщить значимость отличия, средние значения, дисперсии и объемы выборок или дополнить эту информацию доверительным интервалом, убрав какое-нибудь из чисел первого набора. Различие, однако, между этими представлениями имеется и относится к психолого-педагогическому аспекту. Дело в том, что ошибочное смешение, совершенно аналогичное описанному выше смешению $P(H_0|D)$ и $P(D|H_0)$, приводит к довольно интересным и не столь страшным последствиям.

Ошибка состоит в неправильной интерпретации доверительного интервала. Ее можно найти, например, в популярном в России руководстве по обработке данных в SPSS для психологов: «95% Confidence Interval for Mean (Доверительный интервал для среднего значения в 95%) – при большом числе выборок из генеральной совокупности 95% средних значений этих выборок попадут в интервал, определяемый указанными в таблице границами». Иногда ошибочная интерпретация выражается в родственной форме: вероятность того, что истинное математическое ожидание попадает в указанный интервал, равна 0,95. На самом деле в случае доверительного интервала для среднего 0,95 – это вероятность, что при получении выборки испытаниями случайной величины с неким математическим ожиданием это математическое ожидание будет накрыто вычисленным (по известной формуле) в соответствии с выборочными значениями 95%-ным интервалом. Если мы точно знаем, что математическое ожидание случайной величины равно нулю, то 95%-ный доверительный интервал, который мы посчитаем по извлеченной выборке, только в 5% случаев не будет накрывать ноль. Если по имеющимся данным нам выпало несчастье получить интервал из этих 5%, говорить, что вероятность нахождения истинного математического ожидания в полученном интервале равна 0,95, совершенно неверно. Неверно ожидать, и что следующие выборочные средние попадут в этот интервал в 95% случаев. Другие распространенные ошибки (не закрепленные в учебных текстах) таковы: уверенность, что размер доверительного интервала не зависит от размера выборки, уверенность, что перекрывающиеся интервалы для средних двух независимых выборок означают отсутствие значимых различий между ними [Sotos et al., 2007].

Точная, не наводящая на ложные коннотации интерпретация доверительного интервала такова: это совокупность возможных значений $\{a\}$ математического ожидания, каждое из которых не будет отвергнуто при применении к нему Т-теста с гипотезой H_0: математическое ожидание равно a. Например, если $p > 0{,}05$ для гипотезы $a = 0$, то она не будет отвергнута и попадет в доверительный интервал.

Преимущество представления результатов в виде доверительных интервалов, а не значимостей, как мы показали, не может обосновываться теоретическими соображениями – в этом смысле они эквивалентны. Различия проявляются в том, насколько способы представления способствуют или противодействуют дихотомическому / интегративному мышлению. Обоснование преимущества

доверительных интервалов и других альтернативных форм представлений результатов может, следовательно, проводиться в эмпирических исследованиях. Примером такового может служить работа [Sotos et al., 2007], где показано, что студенты, получающие материал по экологической проблематике с иллюстрациями в виде error bars, лучше осознают, что не-отвержение нулевой гипотезы (которая говорит о безопасности некоторого подхода к решению проблем) не означает ее истинности[7]. При этом даже ошибочная интерпретация, сопровождающаяся описанием доверительных интервалов, может быть полезнее ошибки, обычно сопровождающей NHST, так как интервал, показывающий вероятное положение оцениваемого параметра, демонстрирует широкий диапазон согласующихся с выборкой возможностей. Это, по крайней мере, выводит из плоскости «дихотомического мышления».

В заключение раздела отметим, что ошибки «дихотомического мышления» могут не вызывать вопросов у подавляющей части сообщества даже во вполне привычных дискурсах. Так, правила принятия / отвержения гипотезы о нормальности распределения перед проведением дисперсионного анализа демонстрирует сакральный характер магического числа 0,05. Поскольку эмпирические распределения чего угодно не могут быть в точности нормальными, то на достаточно большой выборке на уровне 0,05 будут отвергнуты гипотезы о нормальности для распределений, как угодно мало отличающихся от нормального. В то же время роль этих отклонений в нарушениях точности оценок значимости Т-критерия и дисперсионного анализа с увеличением выборки становится несущественна даже в случае существенной асимметрии распределений [Корнеев, Кричевец, 2011].

Величина статистического эффекта

Следует отметить, что мы не считаем величину статистического эффекта полноценной альтернативой NHST. Как и доверительные интервалы, метод предложен изначально для дополнения результатов NHST показателем, не зависящим от объема выборки, по сути, для напоминания исследователю, что уровень значимости зависит от объема выборки [Cohen, 1994]. В этом смысле он является попыткой совершенствования NHST. Формально же такое представление данных математически эквивалентно «традиционным» для NHST показателям: величина эффекта может быть вычислена на основе средних, стандартных отклонений, степеней свободы, величины статистики[8].

Роль величины статистического эффекта изменилась в связи с развитием мета-анализа, как правило, основанного на применении d Коэна. Требование публикации эффектов стало отвечать содержательной задаче – возможности последующего статистического обобщения различных эмпирических исследований. В этом контексте величина статистического эффекта действительно относится к альтернативе NHST – но включается в мета-анализ в качестве технической процедуры.

Предварительные выводы

1. Те обычаи употребления, которые возникли вокруг процедуры NHST к настоящему времени, заслуживают категорического отвержения. Однако сама процедура NHST в этом не виновата: как и всякий другой инструмент, она должна употребляться с необходимой осторожностью.

2. Если неосторожное применение NHST будет заменено на неосторожное применение других подходов (доверительные интервалы, байесовский подход), скорее всего, в результате возникнут другие массовые ошибки – основания для таких опасений были предъявлены в нашей статье.

3. По нашему мнению, изменений требует не метод оценивания, а организация оценивания результатов, о чем пойдет речь во второй части статьи.

Литература

Алексеев А.А. Земля круглая (p < 0,05): Байес, Фишер и другие. В кн.: А.Н. Алехин (Ред.), Методологические проблемы медицинской психологии: материалы семинара кафедры клинической психологии Российского государственного педагогического университета им. А.И.Герцена. Часть II. СПб.: Знание, 2012, С. 69–88.

Артемьева Е.Ю., Мартынов Е.М. Вероятностные методы в психологии. М.: Моск. гос. университет, 1975.

Корнеев А.А., Кричевец А.Н. Условия применимости критериев Стьюдента и Манна–Уитни. Психологический журнал, 2011, 32(1), 97–110.

Кричевец А.Н., Корнеев А.А., Рассказова Е.И. Математическая статистика для психологов. М.: Академия, 2012.

Кричевец А.Н., Шварц А.Ю., Чумаченко Д.В. Перцептивные действия у учащихся и экспертов при использовании визуальной математической модели. Психология. Журнал Высшей школы экономики, 2014, 11(3), 55–78.

Сивуха С.В. Статистические эффекты в публикациях с изложением результатов научных исследований. В кн.: М.М. Горбунов (Ред.), Актуальные проблемы гуманитарных и социально-экономических наук: Сборник материалов Восьмой международной заочной научно-практической конференции. М.: Перо, 2014. Ч. 1, с. 123–25.

Сивуха С.В., Козяк А.А. О реформе статистического вывода в психологии. Психология. Журнал Высшей школы экономики, 2009, 6(4), 66–86.

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

Borenstein M., Hedges L.V., Higgins J.P.T., Rothstein H.R. Introduction to Meta-Analysis. Chichester, UK: John Wiley and Sons, 2009.

Carver R. The case against statistical significance testing. Harvard Educational Review, 1978, 48(3), 378–399.

Cohen J. The earth is round (p < .05): Rejoinder. American Psychologist, 1994, 50(12), 997–1003.

Eidswick J. A Bayesian Alternative to Null Hypothesis Significance Testing. Shinken Research Bulletin, 2012, 16(1), 2–15.

Fraley R.C., Marks M.J. The null hypothesis significance testing debate and its implications for personality research. In: Handbook of research methods in personality psychology. New York, NY: Guilford, 2007. pp. 149–169.

Gigerenzer G. Mindless statistics. The Journal of Socio-Economics, 2004, 33(5), 587–606.

Gliner J.A., Leech N.L., Morgan G.A. Problems with null hypothesis significance testing (NHST): what do the textbooks say? The Journal of Experimental Education, 2002, 71(1), 83–92.

Hoekstra R., Finch S., Kiers H.A., Johnson A. Probability as certainty: Dichotomous thinking and the misuse of p-values. Psychonomic Bulletin and Review, 2006, 13(6), 1033–1037.

Kelley K., Preacher K.J. On effect size. Psychological methods, 2012, 17(2), 137–152.

Little T.D. Longitudinal structural equation modeling. New York, NY: Guilford Press, 2013.

Masicampo E.J., Lalande D.R. A peculiar prevalence of p values just below. 05. The Quarterly Journal of Experimental Psychology, 2012, 65(11), 2271–2279.

Nuzzo R. Statistical errors. Nature, 2014, 506(13), 150–152.

Purssell E., While A. P = nothing, or why we should not teach healthcare students about statistics. Nurse education today, 2011, 31(8), 837–840.

Robinson D.H., Wainer H. On the past and future of null hypothesis significance testing. Research report, 2001, No. 2, 1–20.

Schneider J.W. Null hypothesis significance tests. A mix-up of two different theories: the basis for widespread confusion and numerous misinterpretations. Scientometrics, 2015, 102(1), 411–432.

Sotos A.E.C., Vanhoof S., Van den Noortgate W., Onghena P. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. Educational Research Review, 2007, 2(2), 98–113.

Trafimow D., Marks M. Editorial. Basic and Applied Social Psychology, 2015, 37(1), 1–2.

Wilkinson L. Statistical methods in psychology journals: Guidelines and explanations. American psychologist, 1999, 54(8), 594–604.

Примечания

[1] Мы предлагаем считать словосочетание «массовая наука» термином со статусом, аналогичным понятию массовой культуры в социальной критике. Без введения в анализ этого аспекта проблемы нам не обойтись.

[2] После наших разъяснений мы для простоты пишем «значимости», поскольку так называют сейчас показатель *p* в статистических пакетах и в текстах, хотя исходно словами «уровень значимости» обозначали параметр критерия, выбранный до исследования.

[3] Точнее говоря, не $p(D|H_0)$, а вероятность того, что «будет получен данный или более экстремальный результат» [Там же]. Авторы не акцентируют внимание на этом различии, и используют одно и то же обозначение $p(D|H_0)$ и в контексте NHST, и в байесовском контексте, что не точно по существу.

[4]Сам Дж.Коэн приводит его не как аргумент в пользу байесовскиго подхода, а чтобы показать ошибочность инверсии вероятностей при выводе на основе NHST. Более того, он специально подбирает редкий случай, когда исследователь знает априорную вероятность (как в случае распространенности шизофрении), показывая, что даже в этом случае подмена грозит ошибками. Пример он предваряет словами: «Но в нормальном случае человек не знает априорную вероятность H_0».

[5] Более подробный разбор этого вопроса в связи с установлением уровня значимости для критерия можно найти в нашем учебнике [Кричевец и др., 2012].

[6] Мы используем предложенный С.В.Сивухой и А.А.Козяк перевод "crud factor".

[7] Надо иметь в виду, что корень проблемы не в том, что тот или иной способ графического представления данных сам по себе обеспечивает должный позитивный эффект. Восприятие иллюстраций должно быть сформировано в обучении [Кричевец и др., 2014]. Таким образом, следует менять не только требования к представлению данных, но системы обучения статистике, в котором подтверждение / опровержение гипотез будет включаться в адекватный контекст.

[8] Формулы существенно меняются в зависимости от критериев; при этом для ряда методов обработки данных существует несколько общепринятых оценок. Например, для Т-критерия Стьюдента для независимых выборок могут быть рассчитаны r Пирсона и d Коэна. Обсуждение видов величин статистических эффектов и способов их расчета не входит в задачи данной статьи (см., например, [Borenstein et al., 2009]).

Поступила в редакцию 12 августа 2015 г. Дата публикации: 26 февраля 2016 г.

Сведения об авторах

Корнеев Алексей Андреевич. Старший научный сотрудник, лаборатория нейропсихологии, факультет психологии, Московский государственный университет имени М.В.Ломоносова, ул. Моховая, д. 11, стр. 9, 125009 Москва, Россия; старший научный сотрудник, лаборатория нейрофизиологии когнитивной деятельности, Институт возрастной физиологии, Российская академия образования, ул. Погодинская, д. 8, корп. 2, 119121 Москва, Россия.
E-mail: korneeff@gmail.com

Рассказова Елена Игоревна. Кандидат психологических наук, доцент, кафедра нейро- и патопсихологии, факультет психологии, Московский государственный университет имени М.В.Ломоносова, ул. Моховая, д. 11, стр. 9, 125009, Москва, Россия; старший научный сотрудник, лаборатория медицинской психологии, Научный центр психического здоровья, Каширское шоссе, д. 34, 115522 Москва, Россия.
E-mail: e.i.rasskazova@gmail.com

Кричевец Анатолий Николаевич. Доктор философских наук, кандидат физико-математических наук, профессор, кафедра методологии психологии, факультет психологии, Московский государственный университет имени М.В.Ломоносова, ул. Моховая, д. 11, стр. 9, 125009 Москва, Россия.
E-mail: ankrich@mail.ru

Койфман Александра Яковлевна. Психолог, кафедра психологии образования и педагогики, факультет психологии, Московский государственный университет имени М.В.Ломоносова, ул. Моховая, д. 11, стр. 9, 125009 Москва, Россия.
E-mail: skoyfman@gmail.com

Ссылка для цитирования

Стиль psystudy.ru
Корнеев А.А., Рассказова Е.И., Кричевец А.Н., Койфман А.Я. Критика методологии проверки нулевой гипотезы: ограничения и возможные пути выхода. Часть I. Психологические исследования, 2016, 9(45), 1. http://psystudy.ru

Стиль ГОСТ
Корнеев А.А., Рассказова Е.И., Кричевец А.Н., Койфман А.Я. Критика методологии проверки нулевой гипотезы: ограничения и возможные пути выхода. Часть I // Психологические исследования. 2016. Т. 9, № 45. С. 1. URL: http://psystudy.ru (дата обращения: чч.мм.гггг).
[Описание соответствует ГОСТ Р 7.0.5-2008 "Библиографическая ссылка". Дата обращения

в формате "число-месяц-год = чч.мм.гггг" – дата, когда читатель обращался к документу и он был доступен.]

Адрес статьи: http://psystudy.ru/index.php/num/2016v9n45/1231-korneev45.html

К началу страницы >>