

Сравнение мощности статистических критериев в связи с обсуждением критерия воспроизводимости

В отечественном сегменте методологии науки обсуждается содержание такого критерия научности как воспроизводимость, а также обсуждается специфика его функционирования в постнеклассической науке. Вместе с тем, отечественная критика паранаучных и предположительно паранаучных проектов демонстрирует ситуацию, когда критики избегают в своей аргументации статистических выкладок. В связи с этим возникает ряд вопросов. В частности: «что такое воспроизводимость?» и «какова математическая формулировка критерия воспроизводимости?» Из литературного обзора мы видим, что к настоящему времени зарубежными коллегами предложено пять показателей воспроизводимости, которые апробируются и обсуждаются, но не существует пока общей математической формулировки критерия воспроизводимости (интегрального критерия, охватывающего указанные показатели), и тем более эти показатели еще не стали стандартом. В настоящей работе сравниваются два статистических критерия, относящихся к одному из пяти указанных показателей воспроизводимости.

Цель исследования. Целью настоящей работы является сравнение мощности двух статистических критериев, которые могут быть использованы для выявления эффекта с учетом требования воспроизводимости результатов исследований. При этом воспроизводимость оценивается по показателю «значимость». В соответствии с первым критерием эффект считается выявленным, если значимой оказалась величина эффекта во всех исследованиях (т. е. если значимость величины эффекта воспроизводится во всех исследованиях). В соответствии со вторым критерием эффект считается выявленным, если значимой оказалась взвешенная средняя величина эффекта, полученная в результате метаанализа (при этом значимость величины эффекта может отсутствовать в индивидуальных исследованиях).

Материалы и методы. Для достижения поставленной цели используются методы математической статистики. Мощности двух указанных критериев сравниваются по двум оценкам. Первая оценка является теоретической. Вторая оценка получена в ходе статистического эксперимента. Мощности вычисляются: 1) при различных значениях величины эффекта по Коэну: «малая», «средняя» и «большая», 2) при различной степени гетерогенности: нулевая (модель с фиксированным эффектом), «малая», «средняя» и «большая», 3) при различном числе первичных исследований (от 2 до 8).

Результаты. Мощность первого критерия меньше либо много меньше, чем мощность второго критерия. Мощность первого критерия убывает с ростом числа первичных исследований, а мощность второго возрастает. С учетом конвенционального значения мощности равного 80% первый критерий оказывается непригодным к употреблению при рассмотренных значениях параметров первичных исследований (т. е. если для выяснения значимости величины эффекта в индивидуальных исследованиях будет использован двусторонний t -критерий с уровнем значимости 0,05 и с двумя выборками типичной длины $n = 25$), в то время как мощность второго критерия может быть при необходимости увеличена путем увеличения числа первичных исследований, включаемых в метаанализ.

Вывод. Если критерий воспроизводимости, известный из философии науки, имеет своей целью подтвердить существование эффекта (связи) или, иными словами, выявить эффект, то в условиях, когда в процессе измерения присутствует существенная случайная составляющая, целесообразно применять не первый, а второй критерий.

Ключевые слова: воспроизводимость, показатели воспроизводимости, значимость, метаанализ, мощность критерия

Vadim E. Osipov

Omsk State Technical University, Omsk, Russia

Comparison of the power of statistical tests in connection with the discussion about the reproducibility criterion

The criterion of reproducibility, as well as its functioning in post-non-classical science, are discussed in the Russian methodology of science. At the same time, critics avoid statistical calculations in their arguments. This raises the following questions: "What is reproducibility?" and "What is the mathematical formulation of the reproducibility criterion?" Literature review has identified five indicators of reproducibility, which was proposed by foreign colleagues. These indicators are being tested and discussed. However, there is no General mathematical formulation of the reproducibility criterion (an integral criterion covering these indicators), and these indicators have not yet become a standard. In the present work, we compare two statistical tests, related to one of these five indicators of reproducibility.

Purpose of the study. The aim of this paper is to compare the powers of two tests of statistical significance that can be used to

reveal the effect with the requirement of reproducibility of research results. In this case, the reproducibility is estimated by the indicator "significance". In accordance with the first criterion, the effect is considered to be revealed if the effect size in all studies is significant (i.e. if the significance of the effect size is reproduced in all studies). In accordance with the second criterion, the effect is considered to be revealed if the weighted mean of the effect size obtained as a result of meta-analysis is significant (the significance of the effect size may be absent in individual studies).

Materials and methods. Methods of mathematical statistics are used to achieve this goal. The powers of two tests are compared by two estimates. The first estimate is theoretical. The second one was obtained during a statistical experiment. The powers are calculated: 1) for different values of the Cohen's effect size: "small", "medium" and "large", 2) for different degree of heterogeneity: zero (fixed-effect

model), “small”, “medium” and “large”, 3) for different number of primary studies (from 2 to 8).

Results. The power of the first test is less or much less than the power of the second one. The power of the first test decreases with the growth of the number of primary studies, and the power of the second one increases. Taking into account the conventional power value equal to 80%, the first criterion is unsuitable for use in the considered values of the parameters of primary studies (that is, if a two-tailed *t*-test with the significance level of 0.05 and with two samples of the typical length $n=25$ is used to determine the significance of the effect size in individual studies), while the power of the second test can be

increased if necessary by increasing the number of primary studies included in the meta-analysis.

Conclusion. If the criterion of reproducibility, known from the philosophy of science, is intended to confirm the existence of the effect (connection) or, in other words, to reveal the effect, in conditions where there is a significant random component in the measurement process, it is advisable to apply not the first, but the second test.

Keywords: reproducibility, reproducibility indicators, significance, meta-analysis, the power of a statistical test

Введение

Оценки подобные следующей «результаты парапсихических исследований и экспериментов не воспроизводимы повторно» [1, с. 133] вызывают ряд вопросов. В частности — следующие два вопроса. Что такое воспроизводимость? Каким критерием или критериями следует руководствоваться, чтобы выяснить наличие воспроизводимости?

В литературе мы находим не так уж и много публикаций, проясняющих указанные вопросы. В особенности мало публикаций по второму вопросу.

Насколько можно понять, термин «воспроизводимость» в контексте науки имеет следующие два значения. Первое значение — *инвариантность действий* в научном методе. Второе значение — *инвариантность результатов*, полученных научным методом. Воспроизводимость считают *критерием научности знания* [2, с. 25] (знания об объекте, исследуемом с помощью метода, либо знания о самом методе). Воспроизводимость — «инвариантность действий для любого субъекта в тождественно-подобной ситуации» — является неотъемлемым признаком метода науки [3, с. 35–36]. «...Критерий воспроизводимости, утверждающий обязательность получения в идентичных условиях идентичных результатов» [3, с. 436]. Используются также термины «воспроизводимость экспериментов» [3, с. 518], «повторяемость результатов» [3, с. 623]. Как мы видим, в своих работах

Ильин В.В. касается данного понятия неоднократно.

Рузавин Г.И. и Кезин А.В. связывают понятие воспроизводимости с понятием intersubjectivity в том смысле, что результаты научных исследований должны допускать воспроизведение любым ученым соответствующей области знания, т.е. не должны зависеть от субъекта, проводящего исследование (см.: [4, с. 34; 5, с. 12–13]).

Черникова И.В. [6, с. 92–93] относит в разряд классической науки такое понимание термина воспроизводимости как возможность повторения опыта в любом месте, в любое время и любым экспериментатором с одинаковым результатом.

В работе [7, с. 25] воспроизводимость рассматривается в трактовке классической науки: результаты экспериментальной проверки не должны зависеть ни от места выполнения, ни от времени, ни от установки, ни от наблюдателя. В работе [8, с. 184–185] критерий воспроизводимости излагается также с точки зрения классической научной рациональности.

Баяндин А.В. считает, что проверка теории (гипотезы) не сводится только к постановке одного или нескольких экспериментов; требуется всесторонний анализ условий возникновения экспериментального результата, а также перекрестные опыты и проверки на воспроизводимость [9]. Однако в данной работе Баяндина мы не видим ответа на практический вопрос: каков конкретно должен быть объем

проведенного всестороннего анализа, т.е. когда можно остановить проверки и сделать заключение о наличии воспроизводимости.

В работах Штанько В.И. и Дж. Холтона [10, с. 52; 11, с. 73] мы встречаем термин «тенденция к воспроизводимости результатов» но не находим пояснений, каким образом следует выявлять указанную тенденцию.

По мнению авторов работы [12] критерий «воспроизводимость эмпирического материала» означает, что факты, которые использовались при создании теории, должны статистически устойчиво повторяться в наблюдении или воспроизводиться в эксперименте. Вместе с тем, в работе [12] мы не находим статистических критериев, выявляющих повторяемость.

Авторы работ [2; 5; 13; 14; 15] указывают на ограниченную повторяемость эксперимента в социально-гуманитарных науках.

В работах [16, с. 156; 17, с. 95] критерий воспроизводимости только упоминается.

В работе [18] рассматриваются общеметодологические аспекты и проблемы с воспроизводимостью («репликацией») экспериментов в психологии, но не обсуждаются статистические критерии.

Таким образом, обзор отечественной литературы по методологии науки, осуществленный в настоящей работе, дал ответ на вопрос о том, что такое воспроизводимость (самое общее понимание), но не дал ответа на вопрос о том,

какова математическая формулировка критерия воспроизводимости, которая позволяет на основании анализа имеющихся эмпирических данных сделать вывод о наличии или отсутствии воспроизводимости.

Среди зарубежных публикаций особое внимание привлекает работа большого авторского коллектива под руководством Брайана Носека. Брайан Носек с соавторами пишут, что «не существует единого стандарта для оценки воспроизводимости» [19, с. аас4716-2]. В их работе мы находим пять использованных ими показателей (indicators) воспроизводимости: 1) значимость, 2) *p*-значения, 3) величины эффекта, 4) субъективные оценки команды экспериментаторов, 5) метаанализ величин эффекта. При этом по каждому из показателей воспроизводимость проверяется одним или несколькими способами (с помощью различных статистических критериев). Вместе с тем, в их работе мы не видим математической формулировки критерия воспроизводимости в общем виде (например по типу критерия пригодности в квалиметрии), которая позволяет дать интегральную оценку воспроизводимости. Таким образом, насколько мы можем понять, не выработан общий критерий воспроизводимости, но предлагается ряд показателей и частных критериев, которые апробируются и обсуждаются. Это обсуждение и является поводом для написания настоящей статьи, и далее в настоящей работе мы сконцентрируем внимание на первом показателе воспроизводимости («значимость»).

Можно назвать два (частных) критерия, проверяющих воспроизводимость по первому показателю («значимость»), который имеет в виду статистическую значимость величины эффекта. *Один* из критериев описан в работе [19]. Авторы указанной работы имели в ка-

честве эмпирических данных две связанные выборки. Одна выборка представляла 100 оригинальных исследований, а вторая – их однократные повторения. Воспроизводимость оценивалась по критерию Макнимара для парных номинальных данных. Однако если представить себе другую ситуацию, когда у нас имеется одно оригинальное исследование и одно или несколько исследований, повторяющих это оригинальное исследование, тогда критерий Макнимара, как представляется, неприменим. В этом случае напрашивается *другой*, интуитивно понятный критерий, в соответствии с которым общий результат считается значимым только тогда, когда значимыми оказались результаты одновременно и оригинального, и всех повторных исследований. В работе [20] рассматривается подобный критерий, который анализирует значимость каждого из двух имеющихся исследований, и выдает общую значимость при условии, если результаты сразу обоих исследований оказываются значимыми (Studies 1 & 2 both achieve .05).

Определение 1. *Критерием 1* будем называть критерий, направленный на выявление истинного эффекта в ряде исследований, при этом результат считается значимым, если в каждом индивидуальном исследовании величина эффекта является значимой.

По сути, Критерий 1 требует жесткой воспроизводимости значимости величины эффекта.

Как представляется, критерий воспроизводимости известный в философии науки направлен на статистическую проверку заявленного эффекта (связи) на устойчивость, необходимость, направлен на защиту от случайных, приводящих обстоятельств. В конечном счете, критерий воспроизводимости направлен на выяснение того, существует ли заявленный эффект (связь), или

не существует. Если поставить более узкую задачу – выявление таких статистических закономерностей как ненулевая истинная величина эффекта (для модели с фиксированным эффектом) или ненулевое математическое ожидание истинной величины эффекта (для модели со случайными эффектами), – то представляется, что Критерий 1 имеет довольно низкую способность к выявлению таких эффектов, и что в значительно большей степени на это способен критерий, оценивающий значимость суммарной величины эффекта, полученной в результате метаанализа.

Определение 2. *Критерием 2* будем называть критерий, направленный на выявление истинного эффекта в ряде исследований, при этом результат считается значимым, если значимым окажется взвешенное среднее величины эффекта, полученное в результате метаанализа.

Предположение. Критерий 2 является более мощным, чем Критерий 1.

Если требование воспроизводимости, известное из философии науки, служит для статистического выявления эффекта (связи между явлениями), то нет смысла использовать Критерий 1, если он является менее мощным, чем Критерий 2.

В работе [20] сравниваются мощности Критерия 1, Критерия 2 и ряда других критериев, а оценки мощности характеризуются следующим образом:

- 1) оценки получены путем статистического эксперимента;
- 2) рассматривается модель с фиксированным эффектом;
- 3) число первичных исследований равно 2;
- 4) объем объединенной выборки в индивидуальных исследованиях устанавливается равным 50 или 80 (в различных комбинациях).

Постановка задачи. Для проверки предположения в

настоящей работе сравниваются: мощность Критерия 1 и мощность Критерия 2; оценки мощности характеризуются следующим образом:

1) рассматривается два рода оценок: а) теоретические оценки, б) оценки, полученные путем статистического эксперимента;

2) рассматриваются две модели: а) модель с фиксированным эффектом, б) модель со случайными эффектами с различной степенью гетерогенности (малая, средняя, большая);

3) число первичных исследований изменяется от 2 до 8;

4) объем объединенной выборки в индивидуальных исследованиях равен 50 (две выборки по 25).

Значимость величины эффекта в индивидуальном исследовании проверяется по t -критерию. Для справки рассматриваются указанные оценки мощности t -критерия.

1. Симуляция эмпирических данных, расчетные соотношения, средства компьютерной математики

1.1. Симуляция эмпирических данных

В настоящей работе используются данные двух типов. Первый тип данных соответствует модели с фиксированным эффектом, а второй тип данных – модели со случайными эффектами. Данные получены с помощью генератора случайных чисел.

Для модели с фиксированным эффектом в ряде исследований задается одинаковая фиксированная истинная величина эффекта. При этом, частично повторяя статистический эксперимент, описанный в работе [20], в настоящей работе взяты две генеральные совокупности с нормальным распределением и одинаковым стандартным отклонением $\sigma = 1$; из каждой совокупности извлекается по одной выборке; объемы выборок одинаковы:

$n_1 = n_2 = 25$ и типичны для психологии (объем объединенной выборки равен 50). Задаются три различных средних значения по первой генеральной совокупности (m_1): на 0,2, на 0,5 и на 0,8 больше, чем среднее по второй совокупности (m_2), чтобы истинное значение величины эффекта по Коэну

$$d = \frac{m_1 - m_2}{\sigma} \quad (1)$$

было соответственно: 0,2 («малая» величина эффекта), 0,5 («средняя» величина эффекта) и 0,8 («большая» величина эффекта) [21, с. 40].

Для модели со случайными эффектами среднее по первой генеральной совокупности неизменно в рамках индивидуального исследования, а от одного индивидуального исследования к другому изменяется по случайному закону: $m_1 \in N(m, \tau^2)$, где τ^2 – дисперсия истинного эффекта между исследованиями; m – переменная, значение которой устанавливается также на 0,2, на 0,5 и на 0,8 больше, чем среднее по второй совокупности (m_2), образуя математическое ожидание истинного эффекта

$$\mu = \frac{m - m_2}{\sigma} \quad (2)$$

равное соответственно: 0,2 («малая» величина общего среднего эффекта), 0,5 («средняя» величина общего среднего эффекта) и 0,8 («большая» величина общего среднего эффекта).

1.2. Значимость результата индивидуального исследования

Значимость величины эффекта в первичных исследованиях оценивается по двустороннему t -критерию с двумя выборками и уровнем значимости $\alpha = 0,05$:

$$\hat{t} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \quad (3)$$

$$v = 2n - 2, \quad (4)$$

$$n = n_1 = n_2 = 25,$$

где \hat{t} – эмпирическое значение t -критерия; \bar{x}_1, \bar{x}_2 – выборочные средние; s_1^2, s_2^2 – выборочные дисперсии; v – число степеней свободы.

Уровень значимости результата вычисляем следующим образом: $p = 2 * \text{tcdf}(t, \text{nu}, \text{'upper'})$, где p – уровень значимости величины эффекта в t -форме в рассматриваемом индивидуальном исследовании; “*” – операция умножения; $\text{tcdf}()$ – функция кумулятивного распределения Стьюдента из пакета прикладных программ MATLAB; t – эмпирическое значение критерия Стьюдента, полученное по формуле (3); nu – число степеней свободы, найденное по формуле (4); ‘upper’ – параметр, указывающий, что следует найти дополнение кумулятивного распределения (т.е. площадь под кривой в сторону верхнего «хвоста»). Если $p \leq \alpha = 0,05$, то исследование маркируем как значимое; а если окажется, что $p > \alpha$, то исследование маркируем как незначимое.

1.3. Значимость результата по Критерию 1

В соответствии с данным критерием итоговый результат ряда исследований, количество которых равно k , считаем значимым, если число значимых исследований равно k .

Назвать данный критерий статистическим можно не вполне, потому что, с одной стороны, критерий, конечно же, опирается на непроблематизируемые критерии значимости, применяемые в индивидуальных исследованиях, но, с другой стороны, научные основы процедуры статистического синтеза, описанной в предыдущем абзаце, пока не просматриваются, о чем свидетельствует отсутствие соответствующих формул. Тем не менее, было принято решение написать настоящую работу в связи с тем, что философский дискурс по вопросам воспроизводимости индуцирует интуи-

тивное представление, сформулированное в Критерии 1, и такое представление возникает не только у автора настоящей статьи, о чем свидетельствует факт рассмотрения данного критерия в работе [20].

1.4. Значимость результата по Критерию 2

Опираясь методику, изложенную в работе [22], в метаанализе мы будем рассматривать значения величины эффекта в d -форме, наблюдаемые в первичных исследованиях:

$$\hat{d} = \frac{\bar{x}_1 - \bar{x}_2}{S_{within}},$$

$$S_{within} = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}},$$

где \hat{d} – значение величины эффекта, наблюдаемое в первичном исследовании; S_{within} – стандартное отклонение в объединенной выборке.

Дисперсия наблюдаемой величины эффекта (аппроксимация):

$$V = \frac{n_1 + n_2}{n_1 n_2} + \frac{\hat{d}^2}{2(n_1 + n_2)}. \quad (5)$$

Проверка на гетерогенность производится для того, чтобы выбрать, по какой модели вести расчет: по модели с фиксированным эффектом либо по модели со случайными эффектами. Для этого используют следующую статистику:

$$Q = \sum_{i=1}^k W_i \hat{d}_i^2 - \frac{\left(\sum_{i=1}^k W_i \hat{d}_i \right)^2}{\sum_{i=1}^k W_i},$$

$$W_i = \frac{1}{V_i}, \quad (6)$$

где W_i – вес i -го исследования; \hat{d}_i – величина эффекта, наблюдаемая в i -ом исследовании; V_i – дисперсия наблюдаемой величины эффекта в i -ом исследовании; k – число первичных исследований. Для нулевой гипотезы статистика Q приблизительно подчиняется распределению χ^2 с $df = k - 1$ степенями свободы [23, с. 181], поэтому полученное значе-

ние Q проверяем по критерию хи-квадрат. Если проверка по критерию дает незначимый результат $p(Q, df) > 0,10$ [22, с. 112], то дальнейший расчет ведем по модели с фиксированным эффектом. Если $p(Q, df) \leq 0,10$, то дальнейший расчет ведем по модели со случайными эффектами.

Для модели с фиксированным эффектом находим следующие величины.

$$M = \frac{\sum_{i=1}^k W_i \hat{d}_i}{\sum_{i=1}^k W_i},$$

где M – взвешенное среднее величины эффекта (или суммарная величина эффекта, оценка истинного значения величины эффекта d , установленного по формуле (1)).

$$V_M = \frac{1}{\sum_{i=1}^k W_i} \quad (7)$$

где V_M – оценка дисперсии суммарного эффекта.

$$SE_M = \sqrt{V_M},$$

где SE_M – оценка стандартной ошибки суммарного эффекта.

$$Z = \frac{M}{SE_M},$$

где Z – z -оценка суммарного эффекта. Уровень значимости среднего взвешенного значения величины эффекта находим по следующей формуле:

$$p = 2 \cdot [1 - \Phi(|Z|)],$$

где $\Phi(Z)$ – кумулятивное стандартное нормальное распределение. Если $p \leq \alpha = 0,05$, то результат метаанализа маркируем как значимый; а если окажется, что $p > \alpha$, то результат метаанализа маркируем как незначимый.

Для модели со случайными эффектами находим следующие величины.

$$T^2 = \max \left\{ 0, \frac{Q - (k - 1)}{\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}} \right\},$$

где T^2 – оценка величины τ^2 ;

$$M^* = \frac{\sum_{i=1}^k W_i^* \hat{d}_i}{\sum_{i=1}^k W_i^*},$$

$$W_i^* = \frac{1}{V_i + T^2},$$

где M^* – взвешенное среднее величины эффекта, оценка величины μ , установленной по формуле (2); W_i^* – вес i -го исследования в модели со случайными эффектами. Также вычисляем оценку дисперсии суммарного эффекта:

$$V_{M^*} = \frac{1}{\sum_{i=1}^k W_i^*};$$

оценку стандартной ошибки суммарного эффекта:

$$SE_{M^*} = \sqrt{V_{M^*}};$$

z -оценку суммарной величины эффекта:

$$Z^* = \frac{M^*}{SE_{M^*}};$$

уровень значимости результата для двустороннего критерия:

$$p^* = 2 \cdot [1 - \Phi(|Z^*|)].$$

Если $p^* \leq \alpha = 0,05$, то результат метаанализа маркируем как значимый; а если окажется, что $p^* > \alpha$, то результат метаанализа маркируем как незначимый.

1.5. Мощность t -критерия

Мощность t -критерия находим с помощью функции нецентрального распределения Стьюдента в MATLAB.

Для модели с фиксированным эффектом мощность находим по следующей формуле:

$$\text{PowerT} = \text{nctcdf}(\text{Tcr}, \text{nu}, \text{lambda1}, \text{'upper'}) + \text{nctcdf}(-\text{Tcr}, \text{nu}, \text{lambda1}), \quad (8)$$

где PowerT – мощность t -критерия для модели с фиксированным эффектом; nctcdf() – функция кумулятивного нецентрального распределения Стьюдента; Tcr – критическое значение, квантиль уровня $(1 - 0,05 / 2)$ кумулятивного центрального распределения

Стьюдента; λ_1 – параметр нецентральности, представляющий собой значение истинной величины эффекта в t -форме, рассчитываем по формуле

$$\lambda_1 = \frac{|m_1 - m_2|}{\sqrt{2\sigma^2/n}}. \quad (9)$$

В модели с фиксированным эффектом все величины в формуле (9) фиксированы, а в модели со случайными эффектами величина m_1 , как было сказано выше, имеет нормальное распределение. Поэтому для модели со случайными эффектами находим приближенное значение мощности методом Монте-Карло следующим образом. Формируем массив из 20 000 элементов $m_1 \in N(m, \tau^2)$, подставляем его в (9) и получаем массив элементов λ_1 , подставляя которые в (8), получаем массив значений мощности. Вычисляя среднее арифметическое массива мощностей, получаем приближенное значение мощности t -критерия для модели со случайными эффектами:

$$\begin{aligned} \lambda_{1_Monte} &= \text{abs}(\text{normrnd}(m, \tau, 1, N) - m_2) / \\ &\quad \text{sqrt}(2 * \sigma^2 / n); \\ T_{cr} &= \text{tinv}(1 - 0.05 / 2, nu); \\ \text{PowerT_Monte} &= \text{mean}(\text{nctcdf}(T_{cr}, nu, \lambda_{1_Monte}, \\ &\quad \text{'upper'}) + \text{nctcdf}(-T_{cr}, nu, \lambda_{1_Monte})). \end{aligned}$$

1.6. Мощность Критерия 1

Мощность Критерия 1 будем искать, исходя из следующих соображений. Мощность t -критерия в нашем случае есть вероятность такого простого события, когда случайная величина – величина эффекта, наблюдаемая в индивидуальном исследовании, – попадает в критическую область. И если считать, что значения величины эффекта, наблюдаемые в различных исследованиях, есть независимые случайные величины, то вероятность сложного события, состоящего в том, что случайные величины ока-

жутся в критической области одновременно во всех исследованиях, есть произведение вероятностей соответствующих простых событий. То есть мощность Критерия 1 есть мощность t -критерия, возведенная в степень равную числу исследований. Для модели с фиксированным эффектом имеем:

$$\text{Power1} = \text{PowerT}^k, \quad (10)$$

где Power1 – мощность Критерия 1 для модели с фиксированным эффектом; k – число первичных исследований; PowerT найдено по формуле (8). Для модели со случайными эффектами имеем:

$$\begin{aligned} \text{Power1_Monte} &= \\ &= \text{PowerT_Monte}^k \end{aligned} \quad (11)$$

где Power1_Monte – мощность Критерия 1 для модели со случайными эффектами; PowerT_Monte – мощность t -критерия для модели со случайными эффектами.

1.7. Мощность Критерия 2

Модель с фиксированным эффектом рассчитываем с учетом следующих формул, приведенных в работе [22, с. 268]:

$$\begin{aligned} \text{Power2} &= 1 - \Phi(c_\alpha - \lambda) + \\ &\quad + \Phi(-c_\alpha - \lambda), \end{aligned} \quad (12)$$

$$\lambda = \frac{\delta}{\sqrt{V_\delta}}, \quad (13)$$

$$c_\alpha = 1,96,$$

где Power2 – мощность двустороннего метааналитического критерия (т.е. мощность Критерия 2) для модели с фиксированным эффектом; c_α – критическое значение нормированного отклонения для уровня значимости α ; λ – истинное значение нормированного отклонения, соответствующее альтернативной гипотезе; δ – истинное значение величины эффекта; V_δ – генеральная дисперсия суммарной величины эффекта.

Числитель формулы (13) есть истинная величина эффекта: $\delta = d$, найденная по

формуле (1), а дисперсию V_δ рассчитываем следующим образом. Подставляя в (5) истинную величину эффекта d , получаем генеральную дисперсию величины эффекта, наблюдаемой в индивидуальном исследовании:

$$V_{pop} = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}. \quad (14)$$

Далее, подставляя (14) в (6), а (6) в (7), получаем формулу генеральной дисперсии суммарного (наблюдаемого) эффекта для модели с фиксированным эффектом [22, с. 268]:

$$V_\delta = V_{pop} / k. \quad (15)$$

Для модели со случайными эффектами имеем следующее нормированное отклонение, соответствующее истинной величине эффекта (альтернативной гипотезе):

$$\lambda^* = \frac{\delta^*}{\sqrt{V_\delta^*}}, \quad (16)$$

где δ^* , V_δ^* – истинное среднее величины эффекта и истинная дисперсия суммарного эффекта (are the true mean effect size and its variance for the summary effect). Числитель формулы (16) находим по следующей формуле:

$$\delta^* = \mu = \frac{m - m_2}{\sigma}.$$

Говоря о дисперсии в знаменателе формулы (16), Боренстин с соавторами пишут: «Дисперсия инкорпорирует дисперсию внутри исследований и дисперсию между исследованиями (variance within studies and variance between studies). Рассмотрим простую ситуацию, в которой все индивидуальные исследования имеют одинаковые дисперсии наблюдаемой величины эффекта (each study has the same within-study variance), скажем, V_Y . Тогда дисперсия может быть записана как

$$V_\delta^* = \frac{V_Y + \tau^2}{k}. \quad (17)$$

Правдоподобные значения дисперсии внутри иссле-

дования (of the within-study variance), V_Y , могут быть получены с использованием тех же процедур, что были использованы для модели с фиксированным эффектом» [22, с. 270]. И здесь мы сталкиваемся с противоречием. Согласно формуле (5), приведенной в [22, с. 27], при фиксированных размерах выборок n_1, n_2 дисперсия величины эффекта V может быть «той же» только при фиксированной величине эффекта d , формула (14), однако мы теперь рассматриваем модель со случайными эффектами. Поэтому, не находя ничего лучшего, по указанию Боренстина будем оценивать генеральное значение внутриэкспериментальной дисперсии, используя ту же процедуру, что и в модели с фиксированным эффектом. То есть в формулу (17) будем подставлять дисперсию, найденную следующим образом:

$$V_Y = \frac{n_1 + n_2}{n_1 n_2} + \frac{\mu^2}{2(n_1 + n_2)}.$$

Наконец, мощность Критерия 2 для модели со случайными эффектами вычисляем по следующей формуле:

$$Power2 = 1 - \Phi(c_\alpha - \lambda^*) + \Phi(-c_\alpha - \lambda^*). \quad (18)$$

В заключение теоретической части следует отметить, что соглашением (см.: [22, с. 270]) были установлены три степени гетерогенности: малая, средняя и большая, подразумевающие, что значение τ^2 составляет соответственно: 0,33; 0,67 и 1,0 от дисперсии величины эффекта в индивидуальном исследовании V_Y . В настоящей работе для модели со случайными эффектами варьировались три параметра: общее среднее величины эффекта μ («малый» эффект, «средний» эффект и «большой» эффект), степень гетерогенности: малая ($\tau^2 = \frac{V_Y}{3}$), средняя ($\tau^2 = \frac{2 \cdot V_Y}{3}$), большая ($\tau^2 = V_Y$) и число исследований (от 2 до 8).

2. Результаты моделирования

2.1. Модель с фиксированным эффектом

Значение, полученное по формуле (8), совпадает со значением, приведенным Коэном, и с результатами статистических экспериментов (табл. 1).

Теоретические и экспериментальные оценки мощности

Критериев 1 и 2, полученные в настоящей работе (н.р.) и в работе [20] для «средней» величины эффекта ($d = 0,5$), приведены в табл. 2. По каждой строке табл. 2 было выполнено 40 000 повторений с подсчетом относительных частот тех событий, при которых удовлетворялось требование критерия. (В работе [20] вы-

Таблица 1

Значения мощности t -критерия для фиксированного эффекта «среднего» размера ($d = 0,5$)

Параметр	Значение
Теоретическая оценка мощности в работе [21, с. 36] для $d = 0,5$ при уровне значимости 0,05 и длине выборок $n = 25$; нулевая гипотеза: $m_1 = m_2$	0,41
$PowerT$ – теоретическая оценка мощности, полученная в настоящей работе по (8).	0,41
Экспериментальная оценка мощности в работе [20, с. 336] (относительные частоты значимых первичных исследований Study 1, Study 2)	0,42 0,41
Экспериментальная оценка мощности в настоящей работе (относительная частота значимых первичных исследований)	0,41

Таблица 2

Теоретические и экспериментальные оценки мощности критериев для «среднего» фиксированного эффекта ($d = 0,5$), выраженные в процентах

Число исследований, k	Критерий 1			Критерий 2		
	Опыт [20]	Опыт (н.р.)	Теория (н.р.)	Опыт [20]	Опыт (н.р.)	Теория (н.р.)
1	2	3	4	5	6	7
2	17	16,70	16,82	69	69,7	69,2
3	–	6,93	6,90	–	85,8	85,4
4	–	2,77	2,83	–	93,9	93,6
5	–	1,14	1,16	–	97,5	97,3
6	–	0,53	0,48	–	99,1	98,9
7	–	0,17	0,20	–	99,6	99,6
8	–	0,08	0,08	–	99,9	99,8

Таблица 3

Теоретические и экспериментальные оценки мощности критериев для «малого» и «большого» фиксированных эффектов, выраженные в процентах

Число исследований, k	Для «малого» эффекта ($d = 0,2$)				Для «большого» эффекта ($d = 0,8$)			
	Критерий 1		Критерий 2		Критерий 1		Критерий 2	
	Опыт	Теория	Опыт	Теория	Опыт	Теория	Опыт	Теория
2	1,05	1,14	16,2	16,9	62,50	62,64	97,6	97,1
3	0,14	0,12	22,7	23,1	49,60	49,58	99,8	99,7
4	0,02	0,01	28,8	29,2	39,21	39,24	100,0	100,0
5	0,00	0,00	34,3	35,1	30,72	31,05	100,0	100,0
6	0,00	0,00	40,6	40,8	24,63	24,58	100,0	100,0
7	0,00	0,00	45,3	46,3	19,45	19,45	100,0	100,0
8	0,00	0,00	51,2	51,4	15,16	15,40	100,0	100,0

полнено 10 000 повторений.) Относительные частоты, умноженные на 100 %, вошли в столбцы, обозначаемые словом «опыт». Значения в столбце 4 рассчитаны по формуле (10), а в столбце 7 – по формуле (12), и все выражены в процентах.

Значения мощности для величины эффекта «малого» и «большого» размеров (по Коэну) представлены в табл. 3. Из таблиц видно, что расчетные значения мощности совпадают с оценками мощности, полученными в статистическом эксперименте, с точностью до одного процентного пункта.

Теоретическая оценка показывает, что для выявления фиксированного эффекта «малой» величины с вероятностью 99,9% требуется объединить метаанализом не менее 25 исследований, при этом мощность Критерия 1 составит $1,4 \cdot 10^{-18}\%$.

2.2. Модель со случайными эффектами

В модели со случайными эффектами произведены вычисления для девяти сочетаний величин эффекта и гетерогенности (табл. 4). Для каждого сочетания трех условий – величина эффекта, степень гетерогенности, число исследований – было выполнено по 20 000 повторений. В графы «Опыт» по Критерию 1 и Критерию 2 помещены относительные частоты событий, удовлетворяющих соответствующим критериям, умноженные на 100 %. В графы «Теория» по Критерию 1 помещены значения, рассчитанные по формуле (11) и умноженные на 100 %. В графы «Теория» по Критерию 2 помещены значения, рассчитанные по формуле (18) и умноженные на 100 %.

3. Заключение

3.1. Обсуждение

Как видно из табл. 4, наибольшие расхождения опытных и теоретических оценок

Теоретические и экспериментальные оценки мощности критериев для модели со случайными эффектами, выраженные в процентах

μ	$\frac{\tau^2}{V_Y}$	k	Критерий №1		Критерий №2	
			Опыт	Теория	Опыт	Теория
0,2	0,333	2	1,93	2,12	16,4	13,9
		3	0,25	0,30	21,3	18,5
		4	0,02	0,04	25,4	23,1
		5	0,01	0,01	30,3	27,7
		6	0,00	0,00	34,3	32,2
		7	0,00	0,00	39,1	36,6
		8	0,00	0,00	43,0	40,8
		8	0,00	0,00	43,0	40,8
	0,667	2	3,13	3,29	18,0	12,1
		3	0,54	0,61	21,4	15,7
		4	0,07	0,11	25,2	19,4
		5	0,03	0,02	28,3	23,1
		6	0,00	0,00	31,8	26,8
		7	0,00	0,00	34,8	30,4
		8	0,00	0,00	38,5	33,9
		8	0,00	0,00	38,5	33,9
	1,000	2	4,43	4,46	18,3	10,9
		3	0,91	0,96	21,6	13,9
		4	0,23	0,21	24,0	16,9
		5	0,02	0,04	26,4	20,0
		6	0,01	0,01	28,7	23,1
		7	0,00	0,00	31,4	26,1
		8	0,00	0,00	34,7	29,2
		8	0,00	0,00	34,7	29,2
0,5	0,333	2	17,85	17,67	57,7	56,8
		3	7,51	7,47	72,8	74,3
		4	3,16	3,22	83,8	85,4
		5	1,44	1,36	89,9	92,1
		6	0,56	0,57	94,5	95,8
		7	0,31	0,25	96,9	97,9
		8	0,09	0,10	98,5	98,9
		8	0,09	0,10	98,5	98,9
	0,667	2	18,64	18,30	53,8	47,9
		3	8,24	8,17	67,2	64,6
		4	3,27	3,47	77,0	76,9
		5	1,47	1,44	84,3	85,4
		6	0,65	0,65	89,5	91,0
		7	0,29	0,29	93,5	94,6
		8	0,13	0,12	95,7	96,8
		8	0,13	0,12	95,7	96,8
	1,000	2	19,02	19,37	50,1	41,3
		3	8,14	8,63	61,5	56,8
		4	3,73	3,74	70,8	69,2
		5	1,80	1,67	78,4	78,6
		6	0,76	0,73	84,5	85,4
		7	0,34	0,31	89,0	90,3
		8	0,14	0,14	92,6	93,6
		8	0,14	0,14	92,6	93,6
0,8	0,333	2	56,91	57,56	87,0	91,5
		3	43,16	43,51	95,9	98,3
		4	32,80	32,96	99,0	99,7
		5	25,48	25,12	99,8	100,0
		6	19,30	18,69	100,0	100,0
		7	14,01	14,23	100,0	100,0
		8	11,01	11,12	100,0	100,0
		8	11,01	11,12	100,0	100,0
	0,667	2	53,80	53,55	82,1	84,6
		3	39,54	39,30	92,2	95,5
		4	28,81	28,85	97,3	98,8
		5	21,57	21,45	99,2	99,7
		6	15,71	15,16	99,8	99,9
		7	11,93	11,29	99,9	100,0
		8	53,80	53,55	82,1	84,6
		8	53,80	53,55	82,1	84,6
	1,000	2	51,97	51,34	78,2	77,7
		3	36,83	36,01	88,3	91,5
		4	26,20	26,39	94,9	97,1
		5	18,55	18,69	98,0	99,0
		6	13,58	13,19	99,3	99,7
		7	9,27	9,63	99,8	99,9
		8	6,96	6,93	100,0	100,0
		8	6,96	6,93	100,0	100,0

(около 8–9 процентных пунктов) получаются для Критерия 2 при малой и средней величинах эффекта, большой гетерогенности и числе исследований $k = 2$. По-видимому, такое большое расхождение связано с грубостью аппроксимации, принятой в настоящей работе. Однако это не затемняет общей картины: мощность Критерия 1 меньше или гораздо меньше мощности Критерия 2, и при этом с ростом числа исследований мощность Критерия 1 уменьшается, а мощность Критерия 2 возрастает. И если учесть, что по общему соглашению мощность критерия должна быть 80% [22, с. 265], то Критерий 1, как видно из табл. 4, оказывается непригодным к употреблению (при указанных значениях параметров первичных исследований), тогда как мощность Критерия 2 при необходимости может быть увеличена путем увеличения числа исследований, включаемых в метаанализ.

При неблагоприятном сочетании независимых параметров ($k = 2$; $\frac{\tau^2}{V_Y} = 1,000$) мы видим, что мощность t -критерия (как корень степени k из мощности Критерия 1) приблизительно равна мощности Критерия 2:

для $\mu = 0,8$ имеем $\sqrt{0,5134} = 0,717 < 0,777$ (теория);

для $\mu = 0,5$ имеем $\sqrt{0,1937} = 0,440 > 0,413$ (теория);

для $\mu = 0,2$ имеем $\sqrt{0,0443} = 0,210 > 0,183$ (опыт).

И это понятно. Для модели с фиксированным эффектом такого быть не может, поскольку и в индивидуальных исследованиях, и в метаанализе мы измеряем одну и ту же истинную величину эффекта d , и дисперсия взвешенного среднего в соответствии с формулой (15) кратно меньше дисперсии результата индивидуального исследования за счет кратно большего объема объединенной выборки при метаанализе. Для модели со случайными эффектами мы путем метаанализа измеряем общее среднее μ , вследствие чего в дополнение к генеральной дисперсии индивидуального исследования появляется дисперсия истинной величины эффекта τ^2 . Тогда при большой степени гетерогенности ($\tau^2 = V_Y$) и двух исследованиях ($k = 2$) из формулы (17) получаем генеральную дисперсию взвешенного среднего равную генеральной дисперсии величины эффекта в индивидуальном исследовании:

$$V_{\delta}^* = \frac{V_Y + \tau^2}{k} = \frac{V_Y + V_Y}{2} = V_Y.$$

Поэтому мощности Критерия 2 и t -критерия приблизительно равны при указанном сочетании параметров.

3.2. Выводы

Известны два частных критерия, проверяющих воспроизводимость по показателю «значимость». Один критерий, рассмотренный в работе Брайана Носека, работает с выборкой пар исследований, где одно исследование в паре является оригинальным, а другое – повторным. Другой критерий, подробно рассмотренный в настоящей статье и обозначенный здесь как Критерий 1, работает с выборкой из двух и более исследований, где одно исследование можно условно считать оригинальным, а остальные – повторными; и этот критерий требует, чтобы во всех исследованиях величина эффекта была значимой. Кроме того, в настоящей

работе рассмотрен критерий, требующий значимости суммарной величины эффекта, полученной в результате метаанализа (Критерий 2).

Критерий 1 по сравнению с Критерием 2 малоприменим для цели статистического выявления эффекта по следующим причинам. Во-первых, мощность Критерия 1 меньше или много меньше мощности Критерия 2, и с ростом числа первичных исследований, охватываемых анализом, мощность Критерия 1 убывает, а мощность Критерия 2 возрастает. Во-вторых, с учетом конвенционального значения мощности равного 80% Критерий 1 оказывается непригодным к употреблению при рассмотренных значениях параметров первичных исследований (т.е. если в индивидуальных исследованиях для выяснения значимости величины эффекта будет использован двусторонний t -критерий с уровнем значимости 0,05 и с двумя выборками типичной длины $n = 25$), в то время как мощность Критерия 2 может быть при необходимости увеличена путем увеличения числа первичных исследований, включаемых в метаанализ.

Мощность Критерия 2, как правило, больше мощности t -критерия, применяемого к индивидуальному исследованию, благодаря сравнительно большому объему объединенной выборки в метааналитическом исследовании. Исключением из этого правила являются случаи с большой гетерогенностью и малым числом исследований. Мощность Критерия 1 меньше мощности t -критерия, применяемого к индивидуальному исследованию, вследствие перемножения мощностей t -критериев.

Литература

1. Лешкевич Т. Г. *Философия: курс лекций*. М.: ИНФРА-М, 2000. 240 с.
2. Ильин В. В. *Критерии научности знания: монография*. М.: Высш. шк., 1989. 128 с.
3. Ильин В. В. *Философия: учебник*. В 2 т. Т. 1. Ростов н/Д: Феникс, 2006. 832 с.
4. Кезин А. В. *Научность: эталоны, идеалы, критерии. Критический анализ методологического редукционизма и плюрализма*. М.: Издательство московского университета, 1985. 128 с.
5. Рузавин Г.И. *Методология научного познания: учеб. пособие для вузов*. М.: ЮНИТИ-ДАНА, 2012. 287 с.
6. Черникова И. В. *Природа науки и критерии научности // Гуманитарный вектор*. 2012. № 3 (31). С. 89–96.
7. Карпин В.А., Кузьмина Н.В., Добрынина И.Ю. и др. *Методология научного исследования // Вестник СурГУ. Медицина*. 2015. № 2 (24). С. 24–27.
8. Полякова Я. С. *Трансформация идеала и критериев научности в науке // Физическое воспитание и тренировка*. 2011. № 1. С. 183–186.
9. Баяндин А. В. *К проблеме критериев научности // Философия науки*. 2000. № 1 (7). С. 13.
10. Холтон Дж. *Что такое «антинаука»? // Вопросы философии*. № 2. 1992. С. 26–58.
11. Штанько В.И. *Философия и методология науки. Учебное пособие для аспирантов и магистрантов естественнонаучных и технических вузов*. Харьков: ХНУРЭ, 2002. 292 с.
12. Губанов Н. И., Губанов Н. Н., Третьяков Н. Г. *Критерии некоторых аспектов научного знания // Вестник северо-восточного федерального университета им. М.К. Аммосова. Серия: Педагогика. Психология. Философия*. 2016. № 1 (01). С. 25–35.
13. Губанов Н.И., Губанов Н.Н., Волков А.Э. *Особенности критериев научности в социально-гуманитарных науках // Гуманитарное образование в креативно-антропологическом измерении: сборник научных статей 12-й Всероссийской научно-практической конференции, 19–20 нояб. 2015 г., Екатеринбург / науч. ред. С. З. Гончаров, Рос. гос. проф.-пед. ун-т. Екатеринбург: Издательский Дом «Ажур», 2015. С. 15–23.*

References

1. Leshkevich T. G. *Filosofiya: kurs lekt-siy = Philosophy: a course of lectures*. Moscow: INFRA-M; 2000. 240 p. (In Russ.)
2. Il'in V. V. *Kriterii nauchnosti znaniya: monografiya = Century Criteria of scientific knowledge: monograph*. Moscow: Higher School; 1989. 128 p. (In Russ.)
3. Il'in V. V. *Filosofiya: uchebnik = Philosophy: a textbook*. In 2 vol. Vol. 1. Rostov on Don: Phoenix; 2006. 832 p. (In Russ.)

4. Kezin A. V. *Nauchnost': etalony, idealy, krit-eriy. Kriticheskiy analiz metodologicheskogo re-duktsionizma i plyuralizma = Scientific: standards, ideals, criteria. Critical analysis of methodological reductionism and pluralism*. Moscow: Moscow Uni-versity Press; 1985. 128 p. (In Russ.)
5. Ruzavin G.I. *Metodologiya nauchnogo poznaniya: ucheb. posobiye dlya vuzov = Method-ology of scientific knowledge: manual for universi-ties*. Moscow: YUNITI-DANA; 2012. 287 p. (In Russ.)
14. Лекторский В. А. *Научное и вненаучное мышление: скользящая граница // Научные и вненаучные формы мышления. Симпозиум (Москва, 4–9 апреля 1995 г.)*. Москва-Киль, 1996. Режим доступа: <http://philosophy.ru/iphras/library/ruspaper/LEKTORS1.htm> (дата обращения: 29.05.2014).
15. Вачков И. В., Вачкова С. Н. *Воспроиз-водимость психологических экспериментов как проблема постнеклассической науки // Куль-турно-историческая психология*. 2016. Т. 12. № 1. С. 97–101. DOI: 10.17759/chp.2016120110.
16. Федорова А. К. *Критерии научности зна-ния // Вестник научных конференций*. 2015. № 2–4 (2). С. 156–157.
17. Хуббеев Р.И., Семенова Э.Р. *Основные критерии научности знания // Вестник научных конференций*. 2015. № 4–5 (4). С. 95–96.
18. Стебаков Д. А. *Психологический экс-перимент в контексте проблемы сложности // Философия науки и техники*. 2017. Т. 22. № 1. С. 46–60. DOI: 10.21146/2413-9084-2017-22-1-46-60.
19. Nosek B. and oth. *Estimating the reproducibility of psychological science [Элек-трон. ресурс] // Science*. 28 Aug. 2015. Vol. 349. Iss. 6251. DOI: 10.1126/science.aac4716. Режим доступа: <http://datacolada.org/wp-content/uploads/2016/03/5341-Nosek-et-al-Science-2015-Estimating-the-reproducibility-of-psychological-science.pdf> (дата обращения 05.01.2018).
20. Braver S. L., Thoemmes F. J., Rosenthal R. *Continuouslycumulatingmeta-analysisandreplicability // Perspectives on Psychological Science*. 2014. Vol. 9(3). P. 333–342. Режим доступа: <http://journals.sagepub.com/doi/pdf/10.1177/1745691614529796> (дата обращения 20.07.2018)
21. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Erlbaum, 1988. 567 p.
22. Borenstein M., Hedges L., Higgins J., Rothstein H. *Introduction to Meta-Analysis*. New Jersey: Wiley, 2009. 421 p.
23. DerSimonian R., Laird N. *A meta-analysis in clinical trials // Controlled Clinical Trials*. 1986. 7. P. 177–188.

6. Chernikova I. V. The Nature of Science and the Criteria of Science. *Gumanitarnyy vector = The Humanitarian Vector*. 2012; 3 (31): 89–96. (In Russ.)
7. Karpin V.A., Kuz'mina N.V., Dobrynina I.Y. et al. Methodology of scientific research. *Bulletin of SurGU. The medicine*. 2015; 2 (24): 24–27. (In Russ.)
8. Polyakova YA: Transformation of the ideal and criteria of science in science. *Fizicheskoye vospitaniye i trenirovka = Physical education and training*. 2011; 1: 183–186. (In Russ.)
9. Bayandin A. V. To the problem of scientific criteria. *Filosofiya nauki = Philosophy of Science*. 2000; 1 (7): 13. (In Russ.)
10. Kholton D. What is “anti-science”? *Voprosy filosofii = Questions of philosophy*; 2. 1992: 26–58. (In Russ.)
11. Shtan'ko V.I. *Filosofiya i metodologiya nauki. Uchebnoye posobiye dlya aspirantov i magistrantov estestvennonauchnykh i tekhnicheskikh vuzov = Philosophy and methodology of science. A manual for graduate students and undergraduates of natural science and technical universities*. Kharkov: KNURE; 2002. 292 p. (In Russ.)
12. Gubanov N. I., Gubanov N. N., Tret'yakov N. G. Criteria of some aspects of scientific knowledge. *Vestnik severo-vostochnogo federal'nogo universiteta im. M.K. Ammosova. Seriya: Pedagogika. Psikhologiya. Filosofiya. = Bulletin of the Ammosov Northeast Federal University. Series: Pedagogy. Psychology. Philosophy*. 2016; 1 (01): 25–35. (In Russ.)
13. Gubanov N.I., Gubanov N.N., Volkov A.E. Features of the criteria of science in the social sciences and humanities. *Gumanitarnoye obrazovaniye v kreativno-antropologicheskom izmerenii: sbornik nauchnykh statey 12-y Vserossiyskoy nauchno-prakticheskoy konferentsii = Humanitarian education in the creative and anthropological dimension: a collection of scientific articles of the 12th Russian Scientific and Practical Conference*. November 19–20. 2015, Ekaterinburg. Ed. S.Z. Goncharov, Russian state professional pedagogical university Ekaterinburg: Publishing House “Openwork”; 2015: 15–23. (In Russ.)
14. Lektorskiy V. A. Scientific and extra-scientific thinking: a sliding border. *Nauchnyye i vnenauchnyye formy myshleniya. Simpozium = Scientific and non-scientific forms of thinking. Symposium* (Moscow, April 4-9; 1995). Moskva-Kil'; 1996. Available from: <http://philosophy.ru/iphras/library/ruspaper/LEKTORS1.htm> (cited: 29.05.2014). (In Russ.)
15. Vachkov I. V., Vachkova P. N. Reproducibility of psychological experiments as a problem of post-non-classical science. *Kul'turno-istoricheskaya psikhologiya = Cultural-historical psychology*. 2016; 12 (1): 97–101. DOI: 10.17759/chp.2016120110. (In Russ.)
16. Fedorova A. K. Criteria for the science of knowledge. *Vestnik nauchnykh konferentsiy = Bulletin of scientific conferences*. 2015; 2-4 (2): 156–157. (In Russ.)
17. Khubbeyev R.I., Semenova E.R. The main criteria of scientific knowledge. *Vestnik nauchnykh konferentsiy = Bulletin of scientific conferences*. 2015; 4–5 (4): 95–96. (In Russ.)
18. Stebakov D. A. Psychological experiment in the context of the problem of complexity. *Filosofiya nauki i tekhniki = Philosophy of science and technology*. 2017. 22; 1: 46–60. DOI: 10.21146/2413-9084-2017-22-1-46-60. (In Russ.)
19. Nosek B. et al. Estimating the reproducibility of psychological science [Internet]. *Science*. 28 Aug. 2015; 349 (6251). DOI: 10.1126/science.aac4716. Available from: <http://datacolada.org/wp-content/uploads/2016/03/5341-Nosek-et-al-Science-2015-Estimating-the-reproducibility-of-psychological-science.pdf> (cited: 05.01.2018).
20. Braver P. L., Thoemmes F. J., Rosenthal R. Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*. 2014; 9(3): 333–342. Available from: <http://journals.sagepub.com/doi/pdf/10.1177/1745691614529796> (cited: 20.07.2018)
21. Cohen J. *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Erlbaum; 1988. 567 p.
22. Borenstein M., Hedges L., Higgins J., Rothstein H. *Introduction to Meta-Analysis*. New Jersey: Wiley; 2009. 421 p.
23. DerSimonian R., Laird N. A meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986. 7: 177–188.

Сведения об авторе

Вадим Евгеньевич Осипов
 Старший преподаватель
 Омский государственный технический университет, Омск, Россия
 Эл. почта: osvad@list.ru

Information about the author

Vadim E. Osipov
 Senior Lecturer
 Omsk State Technical University,
 Omsk, Russia
 E-mail: osvad@list.ru