УДК 51: 101.8

В. М. Резников

Институт философии и права СО РАН
ул. Николаева, 8, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: rvm@philosophy.nsc.ru

МЕТОДОЛОГИЯ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ *

Показано, что базовые принципы методов проверки гипотез не являются логически корректными. Так, для некоторых структур данных минимальная ошибка первого рода не гарантирует ни от отвержения абсолютно правильной гипотезы, ни от большой ошибки второго рода. Несмотря на эти и другие недостатки методов проверки гипотез, их широкую критику и существенное уменьшение использования этих методов, они все еще применяются в практике научных исследований. Полная дискредитация методов предполагает наличие альтернативных совершенных подходов, однако все известные подходы имеют собственные существенные недостатки.

Ключевые слова: методология, статистическая проверка гипотез, принцип Курно, ошибка первого рода, ошибка второго рода.

Проверка статистических гипотез – это заключительный этап обработки данных средствами математической статистики. Традиционное название «проверка гипотез» не вполне корректное. Средствами стандартной математической статистики гипотезу можно отвергнуть или, как говорят, фальсифицировать, а если на основе анализа имеющихся данных у исследователя нет оснований для фальсификации гипотезы, тогда она принимается. С операциональной точки зрения понятие проверки гипотез не состоятельно, более уместно название фальсификации гипотез. В этой статье мы будем равноправно использовать термины «проверка гипотез» и «фальсификация гипотез».

Проблемы проверки гипотез представляют интерес для операционализации базовых положений фальсификационизма Карла Поппера. Философия К. Поппера предполагает определение фальсификаторов для исследуемой гипотезы с целью ее опровержения. В случае опровержения гипотезы последняя приобретает статус научной гипотезы. Если гипотеза выдержала серьезные испытания и не была опровергнута, то она тоже заслуживает доверия. В любом случае теория Поппера предполагает осуществление фальсификации. Известный специалист в области философии науки Д. Майо считает, что методы проверки гипотез представляют интерес для всех концепций философии науки, в которых разрабатываются основания для смены парадигм [Mayo, 1996]. Кроме того, методы статистической проверки гипотез представляют широкий интерес для науки в целом, так как эти методы используются во множестве различных дисциплин. Несмотря на широкое использование методов проверки гипотез в практике исследований различных наук и институализацию методов проверки гипотез во многих науках, тем не менее, в ряде научных областей, в том числе в психологии, медицине, экономике и других преимущественно нетехнических науках существует множество работ, в которых критикуется

* Работа выполнена при финансовой поддержке РФФИ (проект № 11-07-00560а).

© В. М. Резников, 2012

неадекватность этих методов проверки гипотез и их оснований [1]. В отечественной методологической литературе нам неизвестны работы, посвященные специально основаниям и принципам методологии проверки гипотез. Наша работа посвящена исследованию обоснованности оснований и принципов проверки гипотез.

В математической статистике различают задачи проверки адекватности данных одной гипотезе и определение наиболее адекватной гипотезы, когда их несколько. В идейном плане эти задачи одинаковы, в техническом отношении проще анализ одной гипотезы. Вначале рассмотрим проблему проверки единственной гипотезы, кстати, в философии Поппера рассматривается поиск фальсификаторов именно для единственной гипотезы. Но прежде чем исследовать основные направления критики проверки гипотез, кратко рассмотрим основные идеи метода проверки статистических гипотез. Методология проверки гипотез основана на сильном принципе А. Курно. Согласно этому принципу, гипотеза опровергается, если при условии правильности исходной гипотезы, в единственном проведенном эксперименте наблюдается событие, имеющее низкую вероятность появления.

Практическая реализация принципа Курно заключается в построении областей опровержения гипотезы. Это области реализации событий, имеющих низкую вероятность появления при условии правильности гипотезы. Если событие, связанное с исследуемой гипотезой, попадает в область опровержения гипотезы, то гипотеза опровергается. Однако попадание случайной величины в область низких вероятностей не означает неверность гипотезы. События с низкими вероятностями вопреки сильному принципу Курно не являются ни физически, ни логически невозможными. Особенностью редких событий является то, что они редко происходят. Оправданием принципа Курно в теории вероятностей является то, что при выполнимости этого принципа эмпирические частоты обеспечивают точную аппроксимацию теоретических вероятностей в теореме закона больших чисел. Оправданием принципа Курно в математической статистике является то, что на основании этого принципа исследователь будет редко опровергать гипотезу, когда она является корректной. Вероятность опровержения гипотезы, когда она является корректной, совпадает с вероятностью попадания случайной величины в построенную исследователем область опровержения гипотезы. Обычно гипотезы для исследователя не являются одинаково значимыми. Наиболее интересную для исследователя гипотезу называют основной, ее обозначение H_0, еще основную гипотезу называют нулевой. Событие, заключающееся в опровержении основной гипотезы, когда она фактически является правильной, называется ошибкой первого рода. В наших прежних работах мы отмечали фактическую неверность принципа Курно, при его использовании для проверки гипотез [Резников, 2009; 2012], однако нами не исследовалась ни логическая некорректность этого принципа, ни неверные многочисленные его интерпретации в статистических исследованиях. Поэтому данная работа посвящена анализу логической обоснованности принципа, а также анализу корректности различных интерпретаций проверки гипотез. В работе рассмотрены пять оснований для критики методологии проверки гипотез.

1. Критический анализ логических оснований проверки гипотез на основе принципа Курно.

В основе принципа Курно лежит следующая логическая схема:

> Если верна гипотеза H, то событие z имеет низкую вероятность реализации. Событие z реализовалось. Тогда по принципу Курно гипотеза H опровергается.

Данная схема является вероятностным вариантом принципа modus tollens. Согласно принципу modus tollens:

$$A \to B, \neg B \Rightarrow \neg A.$$

Принцип modus tollens корректен в математической логике, однако в вероятностных рассуждениях он неверен. Некорректность принципа для вероятностных рассуждений

[1] *Fidler F.* From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology. Submitted in total fulfillment of the requirements of the degree of Doctor of Philosophy November 2005. URL: http://www.latrobe.edu.au/psy/publication_docs/fidler-phd_aug06.pdf

легко понять на следующем известном примере.

Если Джон американец, то маловероятно, что он член конгресса США. Однако Джон – член конгресса США, тогда по принципу Курно гипотеза, что Джон – американец, опровергается. Фактически отвергаем истинную гипотезу, так как член конгресса США обязательно является американцем. В данном примере посылка в первом суждении вызывает заключение с невысокой вероятностью, однако в нем не принимается во внимание, что утверждение, представляющее собой отрицание заключения в первой посылке рассуждения, в свою очередь влечет, причем всегда, суждение, являющееся отрицанием посылки исходного суждения. Поэтому в рассматриваемом примере появление низковероятностного следствия не опровергает гипотезу, а, наоборот, ее доказывает. Другими словами, редкое событие, что Джон является членом конгресса США, или тем более президентом США, не опровергает, гипотезу, что он американец, а полностью доказывает эту гипотезу [Cohen, 1994].

2. Категорический характер проверки гипотез.

В результате применения стандартной статистической теории Фишера, предназначенной для проверки одной гипотезы, эта гипотеза или опровергается, или, в случае, если нет оснований для ее опровержения, она принимается. В некоторых случаях принимается решение о продолжении исследований. Однако это скорее уход от принятия решения. По существу, имеется только два возможных ответа, один ответ это опровержение гипотезы, второй ответ это принятие гипотезы. Однако теория Фишера не говорит, в какой степени данные подтверждают гипотезу, или в какой степени они ее опровергают.

3. Неадекватность методологии проверки гипотез.

Методология проверки гипотез основана на знании условной вероятности некоторого события, связанного с изучаемой гипотезой, в то время, когда последняя верна. Как отмечают многие исследователи, в действительности их интересует, как правило, не вероятность события при условии правильности гипотезы, а обратная вероятность. Исследователь заинтересован в определении обоснованности изучаемой гипотезы, степени ее подтверждения при получении новых данных. Однако знание вероятности этих данных при условии верности гипотезы в рамках теории Фишера не приближает исследователя к определению вероятности гипотезы, сформулированной на основе этих же данных.

Для последующего анализа оснований проверки гипотез нам понадобятся некоторые понятия, относящиеся к проверке нескольких гипотез. Проверка нескольких гипотез изучается в теории Неймана – Пирсона, которая является развитием теории Фишера. В теории Неймана – Пирсона рассматривается основная гипотеза H_0 и одна или несколько альтернативных гипотез H_1, H_2, …, H_k, в простейшем случае $k = 1$. В теории Неймана – Пирсона к введенному Фишером понятию ошибки первого рода добавляется понятие ошибки второго рода. Событие, заключающееся в принятии гипотезы H_0, когда в действительности верна одна из альтернативных гипотез, назовется ошибкой второго рода. Невозможно одновременно достигнуть минимальных ошибок первого и второго рода. Реалистичный подход заключается в минимизации ошибки второго рода, когда ошибка первого рода фиксирована.

4. Минимальная вероятность данных, на которых гипотеза отвергается, не является гарантией корректности гипотезы.

Проиллюстрируем это положение на известном примере [Cohen, 1994] исследования диагностики шизофрении для подростков. Вводятся две гипотезы, они описываются переменными: H_0 и H_1.

H_0 – подросток является нормальным;

H_1 – подросток является шизофреником;

D описывает, что исследуемые данные свидетельствуют в пользу шизофрении;

P – символ вероятности.

Известно, что

$$P(H_1) = 0{,}02, \quad P(D/H_1) > 0{,}95.$$

X описывает, что данные свидетельствуют о нормальности подростка,

P(диагностики нормальности/H_0) $\approx 0{,}97$.

С учетом введенных обозначений получаем

$$P(X/H_0) \approx 0{,}97.$$

Тогда $P(\neg X/H_0) \approx 0{,}03 < 0{,}05$. Отсюда появление события $\neg X$ при условии корректности гипотезы X имеет вероятность меньше 5 %. Тогда

$$P(D/H_0) \approx 0{,}03 < 0{,}05.$$

На первый взгляд получили хороший результат, что только в трех процентах диагностика свидетельствует о наличии шизофрении, когда в действительности подросток здоров. Является ли гарантией нахождение данных, обеспечивающих при условии корректности гипотезы низкую вероятность, что правильная гипотеза будет редко отвергаться? Для ответа на этот вопрос необходимо определить вероятность $P(H_0/D)$. Это вероятность того, что подросток будет определен как нормальный, в то время как на самом деле этот подросток по тестам определен как шизофреник.

$$P(H_0/D) = P(H_0 \wedge D)/P(D),$$
$$P(H_0 \wedge D) = P(H_0) \times P(D/H_0),$$
$$P(D) = P(H_0) \times P(D/H_0) + P(H_1) \times P(D/H_1).$$

Вероятность $P(D)$ – это вероятность того, что подросток является шизофреником. Она равна сумме двух вероятностей. Первая равна произведению вероятности быть нормальным на условную вероятность по тестам быть определенным как шизофреник, когда в действительности испытуемый подросток является нормальным. Вторая вероятность равна произведению вероятности быть шизофреником на условную вероятность по тестам быть определенным как шизофреник, при условии, что тестируемый в действительности является шизофреником.

$$P(H_0/D) = 0{,}98 \times 0{,}03 = 0{,}0294,$$
$$P(D) = 0{,}0294 + 0{,}02 \times 0{,}95 =$$
$$= 0{,}0294 + 0{,}0190 = 0{,}0484,$$
$$P(H_0/D) = 0{,}607.$$

Получили, что низкая вероятность реализации данных, свидетельствующих о шизофрении при условии, что подросток здоров, не является гарантией для корректности гипотезы о том, что подросток здоров. В рассматриваемом примере вероятность, что подросток здоров при условии, что данные свидетельствуют в пользу шизофрении, очень велика. Таким образом, несмотря на низкую вероятность $P(D_0/H)$ вероятность $P(H_0/D)$ высока, что наглядно свидетельствует о принципиальном различии событий D_0/H и H_0/D и нередуцируемости первого из этих событий ко второму [Cohen,1994; McCloskey, Ziliak, 2009].

5. При большом количестве данных опровергается практически любая гипотеза даже при небольшом отклонении данных от модели. Это было обнаружено в конце 1930-х гг. Такая особенность статистического аппарата воспринималась как некоторый курьез. Вплоть до 1980-х гг. считалось, что много данных не бывает, чем больше данных, тем лучше, так как статистические методы имеют асимптотический характер. Развитие техники, в том числе вычислительной, привело к тому, что современные носители информации обеспечивают хранение огромных массивов данных, их сохранность, выборку нужных данных по самым различным основаниям. В настоящее время актуальной проблемой является разработки алгоритмов, предназначенных для анализа огромных массивов данных. Актуальность связана с тем, что в некоторых научных дисциплинах накоплены огромные массивы данных, например в области маркетинга, в то же время там отсутствуют законы в количественной форме. Характерной особенностью этой области исследований является то, что принимаемые решения на основе статистического анализа являются ответственными. Хранение, обработка и статистический анализ огромных массивов данных обеспечиваются исследованиями в рамках новой прикладной статистической науки под названием Data Mining [Резников, 2006].

Несмотря на многие недостатки методов проверки гипотез и существенное уменьшение объема публикаций на основе этих методов в психологии, медицине, экономике, экологии, образовании и других науках, нельзя говорить, что в практике этих наук полностью отказались от применения статистических критериев проверки гипотез. Почему не произошла смена парадигм в области статистической обработки, несмотря на многолетнюю и широкую критику методов проверки гипотез? Существует целый ряд причин для этого положения дел. Во-первых, в учебниках по математической статистике по-прежнему представлен раздел проверки гипотез, статистические пакеты

программ содержат реализованные алгоритмы различных критериев проверки гипотез, прикладные исследователи знают методы проверки гипотез. Во-вторых, чтобы понять, почему не происходит окончательный отказ от методологии проверки гипотез, необходимо рассмотреть, какие имеются альтернативы проверке гипотез. Для стандартной математической статистики альтернативными подходами являются следующие направления.

1. Байесовский анализ. Это очень популярный вид статического анализа. Он используется в экспертных системах, во многих научных областях, в том числе и в философии. В философии он популярен, так как на его основе разработана нехолистская методология опровержения и подтверждения гипотез [Williamson, 2005]. Байесовский анализ охватывает различные направления от практически субъективистского, где к вероятностям предъявляется единственное требование – они должны быть адекватны колмогоровской аксиоматике, до практически объективистского байесовизма. Однако байесовизм обладает своими недостатками. В рамках байесовизма считается, что исследователь может количественно оценить правдоподобие изучаемых гипотез в конкретной научной дисциплине. Это достаточно спорное положение дел. Признавая, что оценить количественно правдоподобие сложно, байесовисты считают, что правильное вычисление вероятностей гипотез обеспечивается на основе следующей методологии. В случае отсутствия предпочтений о большей реализации одних гипотез по отношению к другим, на основе принципа индифферентности все гипотезы считаются равновероятными. Принцип индифферентности в достаточной степени субъективен и приводит к многочисленным парадоксам. Однако для многих ситуаций удалось справиться с парадоксами. Согласно байесовизму, считается, что исследователь знает не только вероятности гипотез, но и вероятности реализации связанного с гипотезами некоторого события. Если это событие происходит, то пересчитываются условные вероятности гипотез, где в качестве условия берется это событие [Резников, 2007; 2009].

2. Метод доверительных интервалов. Во многих работах отмечается метод доверительных интервалов как замещение или дополнение к методу проверки гипотез. По нашему мнению, правильнее говорить именно о дополнении к методу проверки гипотез. Для того чтобы прояснить свою позицию, отметим, что в статистическом анализе гипотезы подразделяются на полные и неполные. Гипотеза называется неполной, если она задана в виде распределения и исследователь должен оценить неизвестные параметры этого распределения, или, если неизвестно распределение изучаемой случайной величины, то задача исследователя – определить искомое распределение и его параметры. Гипотеза называется полной, если известно и распределение изучаемой случайной величины, и его параметры. В случае если гипотеза задана неполностью и одной из задач является интервальное оценивание параметров, то определение доверительного интервала для оцениваемого параметра – составная часть проблемы проверки гипотез. Каковы преимущества метода доверительных интервалов, если гипотеза задана полностью и нет необходимости решать задачу оценивания параметров распределения? Фидлер отмечает три достоинства метода доверительных интервалов. Во-первых, величина доверительного интервала свидетельствует о точности найденной оценки параметра распределения. Чем шире интервал, тем хуже точность полученной оценки. Во-вторых, различные доверительные интервалы, построенные для одного и того же параметра для разных объемов данных, показывают, как изменяется точность оценивания при увеличении объема используемых данных. В-третьих, использование доверительных интервалов не препятствует определению, фальсифицирована или нет проверяемая гипотеза методами проверки гипотез. Кроме того, метод доверительных интервалов способствует проведению метаанализа [2].

3. Метаанализ. Введен в практику научных исследований Г. Глассом в 1976 г. Метаанализ означает принятие во внимание исследователем всего объема данных по исследуемой проблеме для того, чтобы ограничить или даже избежать предпочтений

[2] *Fidler F.* From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology. Submitted in total fulfillment of the requirements of the degree of Doctor of Philosophy November 2005. URL: http://www.latrobe.edu.au/psy/publication_docs/fidlerphd_aug06.pdf

исследователя в пользу наиболее близких ему гипотез. При этом чтобы избежать избыточного влияния исследований, по которым собраны наибольшие объемы данных, используется показатель под названием «эффект размера». Этот показатель относится к семейству статистических характеристик, которые измеряют влияние условий проведения эксперимента. Причем в отличие от показателей методов проверки гипотез эти характеристики не зависят от объема данных. Формально эффект размера определяется путем вычисления разности между экспериментальной средней величиной и средней величиной в контрольной группе, деленной на стандартное отклонение контрольной группы. Как отмечает Чемберс, существует ряд ограничений для метаанализа. Во-первых, нет смысла подключать данные, полученные в результате применения техники исследований, которая существенно отличается от техники, с помощью которой получены предыдущие данные. Во-вторых, нет смысла подключать данные, полученные в результате менее качественно проведенных исследований, по сравнению с качеством ранее проведенных исследований [3].

4. Методы теории информации. Это направление связано с методами сжатия информации. Если на основе гипотезы H_1 удается сжать данные в большей степени, чем на основе гипотезы H_2, то принимается гипотеза H_1. Это направление наиболее интересное для философии науки, так как в рамках этого направления исследователь не только принимает или отвергает гипотезу, но и имеет возможность оценить прогресс в достижении экономичного представления информации. Методы сжатия информации пока не являются универсальными статистическими методами, предназначенными для проверки гипотез в различных областях знания. Эти методы преимущественно используются в теории информации для целей безопасного представления информации, кодирования, раскодирования информации. Широкому использованию методов сжатия информации в практике статистических исследований препятствует отсутствие методов вычисления колмогоровской сложности для выборок конечного объема данных. Однако методы сжатия информации позволяют модифицировать и улучшить некоторые традиционные статистические критерии проверки гипотез, в частности критерий Пирсона хи-квадрат [Ryabko, Monarev, 2005].

Итак, нами показано, что основания проверки статистических гипотез не являются логически обоснованными. В частности, если даже ошибка первого рода является минимальной, то для определенных структур данных это не гарантирует ни от отвержения абсолютно корректной гипотезы, ни от принятия ошибочной гипотезы, т. е. от ошибки второго рода. Описанные нами сложности и другие проблемы проверки гипотез привели к критике методологии проверки гипотез специалистами в области математической статистики, а также к критике этого раздела статистики в работах психологов, медиков, экономистов, активно применяющих статистические подходы. Несмотря на критику и существенное уменьшение применения методов проверки гипотез во многих областях знания, нельзя говорить, что произошла смена парадигм и методы проверки гипотез являются полностью дискредитированными. Смена парадигм возможна, когда появляются новые, явно более совершенные подходы к решению старых проблем. Однако все известные альтернативные подходы к методам проверки гипотез обладают собственными несовершенствами.

[3] *Chambers E.* An Introduction to Meta-Analysis with Articles from the Journal of Educational Research (1992–2002). URL: http://www-psychology.concordia.ca/fac/kline/601/chambers.pdf

Список литературы

Резников В. М. Методологические проблемы применения статистических критериев // Вестн. Новосиб. гос. ун-та. Серия: Философия. 2009. Т. 7, вып. 3. С. 18–23.

Резников В. М. Критический анализ условий применения теории вероятностей по Колмогорову // Вестн. Новосиб. гос. ун-та. Серия: Философия. 2012. Т. 10, вып. 1. С. 19–24.

Резников В. М. Некоторые подходы к проблеме ошибки модели в системах Data Mining // Философия науки. 2006. № 2 (29). С. 65–74.

Резников В. М. Проблема точных вычислений в байесовской концепции // Вестн. Новосиб. гос. ун-та. Серия: Философия. 2007. Т. 2, вып. 1. С. 22–28.

Cohen J. The Earth Is Round ($p < 0{,}05$) // American Psychologist. 1994. Vol. 49. No. 12. P. 997–1003.

Mayo D. Error and the Growth of Experimental Knowledge. Chicago; L.: The Univ. of Chicago Press, 1996.

McCloskey D. N., Ziliak S. The Unreasonable Ineffectiveness of Fisherian «Tests» in Biology, and Especially in Medicine // Biological Theory. Konrad Lorenz Institute for Evolution and Cognition Research. Altenberg, 2009. Vol. 4. No. 1. P. 44–53.

Ryabko B. Ya., Monarev V. A. Using Information Theory Approach to Randomness Testing // Journal of Statistical Planning and Inference. 2005. Vol. 133. P. 95–110.

Williamson J. Bayesian Nets and Causality. Philosophical and Computational Foundations. Oxford: Oxford Univ. Press, 2005.

Материал поступил в редколлегию 02.04.2012

V. M. Reznikov

METHODOOGICAL ANALYSIS OF STATISTICAL HYPOTHESES TESTING

The article shows that the basic principles of hypothesis testing are not logically correct. For example, for some data structures the minimal error of the first kind prevents neither rejection of an absolutely correct hypothesis nor the large error of the second kind. In spite of these and other shortcomings of the methods of hypotheses testing, their serious criticism and a considerable decrease in their application they are still used in scientific research. The full discredit of these methods would require the existence of alternative ideal methods, but all known statistical approaches have their own shortcomings.

Keywords: methodology, statistical hypotheses testing, Cournot's principle, error of the first kind, error of second kind.