

© 2018 г.

И.Ф. ДЕВЯТКО

## НОВЫЕ ДАННЫЕ, НОВАЯ СТАТИСТИКА: ОТ КРИЗИСА ВОСПРОИЗВОДИМОСТИ К НОВЫМ ТРЕБОВАНИЯМ К АНАЛИЗУ И ПРЕДСТАВЛЕНИЮ ДАННЫХ В СОЦИАЛЬНЫХ НАУКАХ

---

*ДЕВЯТКО Инна Феликсовна – доктор социологических наук, профессор, зав. кафедрой Национального исследовательского университета «Высшая школа экономики», главный научный сотрудник Института социологии ФИНСЦ РАН, Россия, Москва (deviatko@gmail.com).*

---

**Аннотация.** В статье анализируются причины и последствия развернувшегося в социальных науках параллельно с «революцией данных» междисциплинарного кризиса воспроизводимости и достоверности результатов научных исследований. Этот кризис выражается не только в растущей озабоченности ученых вопросами достоверности исследовательских результатов и создания практик открытого доступа к эмпирическим данным и к используемому для их анализа статистическому программному обеспечению, но и в спорах, затрагивающих устоявшиеся и отраженные в стандартных руководствах по методологии научных исследований каноны статистического вывода и проверки гипотез. Описаны формирующиеся подходы к обеспечению воспроизводимости научных результатов, достоверности и открытости данных. Рассматриваются основания критики рутинного использования процедуры статистической проверки нулевой гипотезы и некоторые предложения по более широкому внедрению байесовского подхода к статистическому выводу. В контексте описания новых подходов к обеспечению устойчивости статистической модели при анализе эмпирических данных в социологии обсуждаются вопросы соотношения теории и управляемого данными подхода к поиску паттернов и закономерностей.

**Ключевые слова:** данные • кризис воспроизводимости в социальных науках • открытость данных • публикационное смещение • частотный вывод и байесовский вывод • робастность модели • «ведомый данными» подход и социологическая теория

DOI: 10.31857/S013216250003163-2

Социальные науки с начала 2000-х гг. переживают период больших надежд и оптимистических ожиданий. Они связаны с растущим доступом к новым типам и источникам исследовательских данных. Речь здесь о поведенческих данных, не основанных на опросных методах и вербальных самоотчетах респондентов относительно их установок или поведения и представленных как в форме поведенческих «цифровых следов» онлайн-активности (в некоторых случаях нереактивных или малореактивных, то есть лишь в незначительной мере подверженных эффектам социальной желательности и установки на ответ [Девятко, 2016]), так и в виде масштабных онлайн-архивов частных или публичных

документов. Еще один потенциальный источник роста исследовательских возможностей в социальных науках – появление новых методов сбора, комбинирования, визуализации и анализа целенаправленно собираемых (designed) опросных данных или официальной статистики и «органических», то есть не производимых в исходно исследовательских целях поведенческих параданых и административных записей, в фоновом режиме генерируемых в ходе опосредуемых информационными технологиями взаимодействий между индивидуальными и корпоративными акторами, а также цифровых данных социальных медиа и сенсоров электронных устройств.

Однако параллельно «революции данных» развивается масштабный методологический кризис, который с некоторой долей упрощения может быть обозначен как междисциплинарный *кризис воспроизводимости и достоверности результатов эмпирических исследований*. Этот кризис выражен в растущей озабоченности ученых, работающих в столь разных областях, как психология, социология и политология, биостатистика и биомедицинские исследования, вопросами достоверности исследовательских результатов и создания практик открытого доступа к эмпирическим данным и используемому для их анализа статистическому программному обеспечению, в непримиримых, на первый взгляд, спорах, затрагивающих устоявшиеся и отраженные в стандартах методологии научных исследований каноны статистического вывода и проверки гипотез (см. [Kerr, 1998; Gigerenzer, 2004; Ioannidis, 2005; de Groot, 2014; Doucouliago and Stanley, 2013]).

В статье я кратко (по возможности избегая технических подробностей) резюмирую аргументы и предлагаемые методологические решения, обсуждаемые именно в контексте широко определяемого диспута о кризисе воспроизводимости и достоверности данных, об обоснованности статистических выводов в социальных и поведенческих науках, не останавливаясь на ранее обсуждавшихся и все еще актуальных специфических угрозах репрезентативности, надежности и валидности онлайн-данных и «Больших данных» (см., в частности, [Девятко, 2016; Китчин, 2017]). Я также кратко коснусь некоторых последствий споров о том, как должны выглядеть проверка научных гипотез и оценка устойчивости предлагаемых статистических моделей, каковы должны быть минимальные требования к научным публикациям в социологии, психологии и родственных дисциплинах, для сложившихся представлений о соотношении теории и эксплораторного, «ведомого данными» (data-driven) анализа.

Одна из методологических проблем, прямо связанных с кризисом воспроизводимости эмпирических результатов, известна достаточно давно и широко обсуждалась в разных контекстах, включая анализ причин заметного снижения качества публикаций в науках о человеке и обществе при экспоненциальном росте их количества. Это публикационное смещение (publication bias), проистекающее из поддерживаемого современной системой рецензирования статей в научных журналах и ориентацией на библиометрические показатели стремления редакторов публиковать лишь контринтуитивные, яркие и демонстрирующие статистически значимые различия и эффекты исследовательские результаты, что, в свою очередь, порождает у многих ученых мотивацию избирательно искать в данных новые значимые различия и эффекты, нередко являющиеся лишь статистическими артефактами (см. далее), откладывая при этом в дальний ящик стола или на дальнюю полку картотеки любые эмпирические результаты, не позволяющие отвергнуть нулевую гипотезу (file drawer problem – термин, предложенный Р. Розенталем [Rosenthal, 1979]). В результате методологически безупречные и статистически корректные исследования, содержащие отрицательные результаты или не подтверждающие необычных эффектов и впечатляющих, пусть и невозпроизводимых «открытий», имеют незначительные шансы на публикацию. Страницы же ведущих журналов переполнены «новаторскими» статьями, если не прямо основанными на фальсификации данных (что не редкость в социальных науках<sup>1</sup>),

<sup>1</sup> См., напр., недавний пример из области социальной психологии и поведенческой экономики – обнаружение множественных фальсификаций и манипуляции данными в работах голландского

то построенными на некорректной интерпретации результатов проверки статистической значимости, избирательном описании полученных сырых данных без предоставления достаточной дескриптивной статистики, пренебрежении вопросами статистической мощности и, наконец, полученными в результате явного злоупотребления множественными процедурами анализа и математико-статистическими моделями с целью поиска статистически значимых различий, эффектов и корреляций в уже собранных данных *post factum*, к их «прочесыванию» в погоне за статистической значимостью, без предварительной явной формулировки исходных гипотез и планирования предполагаемых методов анализа. Чаще всего последняя стратегия ведет к формулировке гипотез *после* такого многократного и ситуативно модифицируемого анализа результатов, так сказать, «по факту» обнаружения значимых различий или связей, без явного признания исследователями экспериментального (то есть разведочного), а не конфирматорного, то есть нацеленного на статистическую проверку явно сформулированных гипотез, характера исследования – для ее обозначения часто используется английская аббревиатура HARKing для выражения «*hypothesizing after the results are known*», т.е. построение гипотез по уже известным результатам, отсылающая также к слову *harking*, описывающему такую «погоню» за значимостью [Kerr, 1998].

Дисбаланс в представленности в научных публикациях положительных результатов проверки статистической значимости в сравнении с отрицательными результатами или же с результатами проверки гипотез об инвариантности параметров и отсутствии значимых различий сам по себе является условием воспроизводства кризиса воспроизводимости в науках о человеке: добытые с помощью избирательной чистки и просеивания экспериментальных, квазиэкспериментальных, обсервационных или опросных данных великолепные результаты едва ли могут быть воспроизведены в другой лаборатории или другим исследовательским коллективом, а эффект публикационного смещения практически гарантирует невозможность появления критических статей, содержащих сведения об отрицательных результатах последующих попыток репликации исходного эффекта [Yong, 2012]. Сравнительный переизбыток публикаций, в которых обнаруживаются статистически значимые различия и эффекты в сравнении с публикациями, содержащими негативные результаты, сам по себе свидетельствует о масштабах проблемы. По данным междисциплинарного метаанализа, в последние годы доля позитивных результатов в публикациях из области социальных наук составляет чуть более 85%, что сопоставимо с их долями в химии или иммунологии и, хотя и не достигает рекордных 96–97%, характерных для психологии и психиатрии, заметно превосходит таковую в физике или геологии [Ibid.: 299–300].

Однако проблема воспроизводимости ранее продемонстрированных статистически значимых эффектов, обнаруживающая себя в различных дисциплинарных областях от биомедицинских исследований и клинических испытаний эффективности популярных антипсихотических лекарств до исследований в областях эволюционной биологии и психологии, – отнюдь не сводима к нежелательным эффектам публикационного смещения или недостаточной добросовестности отдельных исследователей, затрагивая основания нашей уверенности в научном методе или, скорее, уверенности в наших текущих представлениях о том, каковы его требования. Интеллектуальная загадка «изнашивающихся» при воспроизведении научных истин несколько лет назад даже привлекла внимание авторов, пишущих для широкой образованной аудитории, выходящей за пределы аудитории научно-популярных журналов. Опубликовавший в 2010 г. статью о снижающихся при попытках воспроизведения размерах обнаруживаемых в научных исследованиях эффектов автор журнала «Нью-Йоркер» Дж. Лерер на примерах продемонстрировал универсальный характер проблемы и собрал некоторые иллюстрации несводимости причин неуклонного снижения

---

исследователя Д. Стапела, в том числе в отозванной статье, исходно опубликованной в «Science», описывающей якобы подтвержденную серией полевых и лабораторных экспериментов положительную связь беспорядка в физическом окружении и мусора на улицах и в публичных пространствах с проявлениями дискриминации и стереотипизирования [Stapel, Lindenberg, 2011].

исходно наблюдавшегося эффекта к публикационному смещению, статистической регрессии оцениваемых показателей к среднему по мере увеличения количества наблюдений и малой статистической мощности изначальных исследований [Lehrer, 2010]. Нередко авторы не могут добиться репликации исходных результатов собственных оригинальных работ в последующих многочисленных экспериментах или наблюдениях. Например, открытый психологом Дж. Скулером впечатляющий эффект «вербального затенения» (худшего распознавания стимулов, которые после первого предъявления испытуемым нужно было описывать словами, в сравнении с точностью распознавания в группе тех, кто не получал задания описывать увиденное) или обнаруженное биологами в начале 1990-х гг. влияние симметрии фенотипических черт самцов на получение преимуществ в половом отборе проверялись тщательно и неоднократно. Результаты метаанализа множества исследований показали, что неуклонное угасание в репликационных исследованиях величины изначально выявленных статистически значимых оценок воздействия изучавшихся факторов на зависимые переменные не может быть полностью сведено к вышеупомянутым причинам. Ситуации, когда желанный значимый результат при более близком рассмотрении и накоплении больших массивов новых данных начинает выглядеть как «слабый сигнал в море шума» [Graham, Özener, 2016], заставляют исследователей переосмысливать канонические подходы к статистическому выводу и представлению эмпирических результатов.

Предлагаемые способы решения проблемы «исчезающих» или откровенно невоспроизводимых результатов включают изменения в стандартных требованиях к научным публикациям. А именно, обязательное указание доверительных интервалов и статистической мощности, приведение оценок величины обнаруженных эффектов, детальное описание всех процедур не только сбора, но и анализа данных, включая «чистку» от выбросов и неполных наблюдений и другие манипуляции с данными, приведение в статьях дескриптивной статистики для всех используемых показателей, а также поощрение политики открытого доступа к данным и использованному для анализа ПО [Munafò et al., 2017]. Среди предложений по необходимому радикальному улучшению подготовки исследователей в области методологии и статистики, к некоторым аспектам которого мы обратимся далее, очевидны те, которые нацелены на более глубокое понимание природы статистического вывода и основных правил планирования эмпирического исследования. В частности, рекомендуются: 1) использование различных форм «слепого» контроля для уменьшения смещений, вызванных осознаваемым или неосознаваемым воздействием исследователя на результаты, 2) применение рандомизации и контролбалансирования в экспериментальных и квазиэкспериментальных планах, 3) по возможности более широкое и корректное<sup>2</sup> использование внутрисубъектных экспериментальных планов, позволяющих резко повысить статистическую мощность и, соответственно, снизить вероятность ошибки второго рода (т.е. ложно-негативного вывода об отсутствии различий или эффекта, когда они в действительности присутствуют)<sup>3</sup> [ibid.: 2].

Формулируются и более радикальные предложения организационного характера, например постпубликационное рецензирование статей, более активное вовлечение в исследовательские проекты специалистов по методологии и статистике или же создание инфраструктуры для независимой поддержки и консультирования в области методов исследования, привлечение студентов к большим исследованиям-репликациям в рамках межуниверситетских учебно-научных коллабораций [ibid.], или даже квотирование числа публикуемых каждым ученым научных статей, основанных на эмпирических данных, в целях

<sup>2</sup>То есть с использованием поправок при проведении множественных сравнений и соответствующих множественных проверок статистической значимости, что увеличивает вероятность ошибки первого рода и, следовательно, шансы ошибочного отклонения нулевой гипотезы.

<sup>3</sup>Для обсервационных и опросных исследований, в отличие от экспериментальных, уместнее при оценке меры поддержки нулевой или любой альтернативной гипотезы конкретным набором данных использование Байесовых факторов (коэффициентов) или отношения правдоподобия [Wagenmakers et al., 2015].

снижения оппортунистической мотивации со стороны исследователей и академических менеджеров, поощряющих гонку за количеством публикаций, и повышения ценности и качества отдельных работ. Например, критика предложения по облегчению процесса публикации исследований, содержащих отрицательные или нулевые результаты, что позволило бы уменьшить масштаб смещения, возникающего из-за отказов от попыток публикации подобных данных, которое искажает действительную картину в пользу позитивных результатов, Л. Нелсон и соавт. в качестве своего рода мысленного эксперимента предлагают ученым ограничиваться одной статьей в год, поскольку в ином случае в результате продвигаемой реформы у некоторых возникнет соблазн публикации чего угодно, включая данные, явно не пригодные с точки зрения вышеупомянутых сугубо методологических требований [Nelson, Simmons, Simonsohn, 2012].

Некоторые инициативы, направленные на повышение прозрачности способов получения и анализа данных, достоверности публикуемых результатов исследований, уже получили воплощение в виде масштабных проектов в целях проверки воспроизводимости в ряде дисциплинарных областей – например, можно упомянуть основанный на сотрудничестве большого количества ученых проект по проверке воспроизводимости данных психологических исследований, в ходе которого была продемонстрирована возможность репликации не более половины опубликованных в 2008 г. результатов [Open Science Collaboration, 2015]. Другим примером может служить принятие новых статистических рекомендаций для научных журналов или профессиональных сообществ [Appelbaum et al., 2018].

Ставшие весьма острыми в последнее десятилетие споры о достоверности, открытости и воспроизводимости научных результатов создали почву для обсуждения тех подходов к статистическому анализу данных и представлению основанных на нем выводов, которые примерно с 1950-х гг. окончательно закрепились сначала в психологии, а затем в социологии и других социальных, поведенческих и биомедицинских дисциплинах в качестве стандартизированного и выполняемого без особых размышлений статистического ритуала проверки значимости для нулевой гипотезы, ПЗНГ [NHST, null hypothesis significance testing], воспроизводимого в соответствии с упрощенными рецептами из учебников и статистических руководств или передаваемого с разной степенью точности новым поколениям исследователей в процессе университетского обучения [Gigerenzer, 2004: 589–590]. Нерефлективное следование ритуалу закономерным образом порождает как концептуальную путаницу, так и множественные ошибки и рассогласования в сообщаемых уровнях значимости, значениях статистических тестов и степенях свободы. Это было, например, убедительно продемонстрировано в недавнем исследовании массива публикаций восьми ведущих психологических журналов с 1985 по 2013 г. [Nuijten et al., 2016].

Споры о преимуществах и недостатках ПЗНГ в кругу специалистов по методологии и прикладной статистике велись давно. В социологии, например, такого рода диспут с конца 1950-х гг. в течение примерно десятилетия развивался вокруг полемической статьи Х. Селвина [Selvin, 1957], обратившего, в частности, внимание на возможность некорректных результатов при применении данной процедуры к неэкспериментальным данным. С начала 2000-х гг. стали накапливаться доказательства того, что бездумное следование сложившемуся статистическому ритуалу не просто ведет в некоторых случаях к невалидным статистическим выводам, но и основано на глубоко укоренившихся когнитивных иллюзиях, распространенных не только среди большинства студентов и исследователей, но и среди преподавателей прикладной статистики. Например, в исследованиях ([Oakes, 1986; Haller, Kraus, 2002]; цит. по: [Gigerenzer, 2004: 594–597]) показано, что даже последние принимали в качестве верных в среднем два из шести откровенно неверных или некорректно сформулированных утверждений, описывающих результаты использования  $t$ -теста в простом случае сравнения средних для экспериментальной и контрольной групп в гипотетическом эксперименте, в результате которого было обнаружено значение  $p = 0,01$ . Не вдаваясь в неуместные в данной статье статистические детали, можно привести следующий пример: большинство опрошенных соглашались с утверждениями, что

полученный результат проверки значимости показывает «вероятность того, что нулевая гипотеза верна» или «позволяет сделать вывод о вероятности экспериментальной гипотезы». Поскольку, с некоторой долей упрощения, можно определить значение  $p$  как вероятность полученных в наблюдении данных при предположении, что нулевая гипотеза верна  $p(D|H_0)$ , описанные утверждения неверны в силу того, что они утверждают что-то о другой величине, вероятности гипотезы при условии полученных данных ( $p(H_0|D)$ ).

В действительности, при использовании иного подхода к статистическому выводу постановка вопроса об оценке сравнительного правдоподобия двух гипотез, в частности нулевой и альтернативной, во многих случаях может иметь смысл и быть выполнимой. Однако это требует от исследователей информированного и осознанного выбора, а не слепого следования ритуалу<sup>4</sup>. Например, вариант байесовского подхода к сравнению средних при помощи того же  $t$ -теста с использованием Байесова фактора как отношения правдоподобия двух гипотез [Rouder et al., 2009] позволяет не только прямо сравнить нулевую и альтернативную гипотезы с точки зрения их правдоподобия, но и открывает принципиальную возможность сделать по результатам такого сравнения вывод об инвариантности или отсутствии различий (то есть сформулировать прямое утверждение относительно правдоподобия самой нулевой гипотезы для данного набора данных, что невозможно при стандартном подходе, предполагающем асимметрию статистического вывода). Как отмечают Дж. Раудер и соавторы, стандартная процедура ПЗНГ порождает риски ложного отвержения нулевой гипотезы (напр., при множественных сравнениях или многократном эксплораторном поиске связей в одной и той же базе данных), не позволяя оценить содержательные предположения об инвариантности некоторых параметров, об отсутствии различий или эффектов, например осмысленное в ряде случаев предположение о том, что пол не оказывает влияния на некоторые когнитивные или социальные характеристики индивидов: «Возможно, самая большая преграда в оценке инвариантностей – методологическая. Инвариантности соответствуют нулевой гипотезе о равенстве, но общепринятый тест значимости не позволяет аналитику представить доказательства в пользу нулевой гипотезы. Если инвариантность, даже приблизительная, имеет место, тогда наилучшим исходом проверки значимости будет невозможность отвергнуть нулевую гипотезу, что интерпретируется как состояние незнания» [ibid.: 226].

Недавно вышедшая статья К. Янга [Young, 2018] идет дальше в направлении предлагаемой трансформации требований к анализу данных в социальных науках. Автор призывает бороться с еще одним «слепым пятном» в анализе данных – *неопределенностью статистической модели*, позволяющей исследователям пробовать множество подходов к анализу одних и тех же данных до достижения желаемого (и повышающего шансы на публикацию) эффекта. Надо признать: в «индустрии вторичного анализа опросных данных», процветающей в социальных и экономических науках, такое использование всех возможных спецификаций модели и методов анализа приобрело чрезвычайную популярность. В опубликованной в русском переводе статье Ф. Шродта [Шродт, 2016] среди перечисляемых «семи смертных грехов» современного количественного анализа упоминается не основанное на каких-то их теоретических предсказаниях «прочесывание» баз неэкспериментальных данных (чаще всего речь о вторичном анализе данных опросных исследований, но это вполне может быть отнесено и к «майнингу» цифровых следов) с добавлением все новых ковариат в модели до получения желаемых уровней значимости на всё тех же данных, порождающее множественные статистические артефакты, и некорректная интерпретация частотной статистики как прямой оценки относительного правдоподобия гипотез.

<sup>4</sup>Иногда «классический» подход к ПЗНГ приписывают Р. Фишеру. В действительности он является неким сложным гибридом рекомендаций, расходившихся во взглядах на цели статистического вывода Р. Фишера, с одной стороны, и Э. Пирсона и Е. Неймана – с другой. Фишер отрицательно относился к установке единого фиксированного уровня значимости, выступающего критерием для принятия однозначного решения об отвержении гипотезы [Gigerenzer, 2004].

Однако Янг идет несколько дальше, включая в список пунктов, которые должны быть проконтролированы для обеспечения робастности итогового решения и снижения неопределенности, все решения и процедуры, ведущие к финальной модели, включая чистку и кодирование данных, подверженный осознанно или неосознанно вносимым исследователем смещением выбор контрольных переменных и методов оценивания модели и т.п. Как оценить влияние выбора конкретной спецификации модели или способа анализа на полученный исследователями результат? Для оценки робастности, устойчивости результатов выбора модели данный автор предлагает использовать некую усовершенствованную комбинацию двух стратегий: 1) ведение прилагаемого к публикации своего рода «супер-журнала регистрации» (*uber log file*), автоматически фиксирующего все, что исследователь когда-либо делал/а с «данными» (своего рода стратегия «Большого статистического брата»), и 2) создание кооперативно-конкурентных «рабочих групп», делающих с данными все, что только возможно, под взаимным надзором, и публикующих, наряду с тем, что признается авторами коллаборации основным и устойчивым результатом, результаты, полученные для всех других использовавшихся спецификаций модели. Для реализации такой непростой и весьма емкой в вычислительном отношении комбинированной стратегии мультимодельного анализа предлагается авторское ПО для Stata.

В случае реализации предлагаемый подход сильно ограничит авторский произвол. Однако возникает опасение, что при действительно масштабном применении в рамках описанной процедуры разных спецификаций и разных методов анализа даже компактно изложенные результаты отдельного исследования со всеми возможными «разветвлениями» базовой модели могут превзойти способность читательской аудитории их адекватно воспринять. Так что с избирательностью реализованного в людях как биологических организмах интеллекта мы, возможно, столкнемся теперь уже на стороне реципиентов научных результатов, решающих, что считать убедительным доказательством робастности итоговой модели.

Если разведочный, управляемый данными поиск паттернов во всегда конечной совокупности наблюдений или даже слабые формулировки бинарных предположений («наличествует–отсутствует», «позитивный–негативный») о возможных паттернах различий и связей останутся в ближайшие годы основной исследовательской стратегией в социальных науках, описанная перспектива распределенного контроля над анализом данных и проверкой моделей может оказаться безальтернативной, как и описанные выше риски понимания и интерпретации результатов. В этой ситуации отнюдь не умозрительный характер приобретет вопрос: можно ли и дальше доверять такое серьезное дело, как проверка неопределенного множества гипотез, потенциально соответствующих имеющемуся очень большому, но конечному набору наблюдений, живым мыслящим существам, руководимым – осознанно или нет – своими интересами и предпочтениями? Социальные науки, вероятно, вынуждены будут пройти до конца путь к полной автоматизации научного поиска и оценки правдоподобия множества генерируемых из базовой модели с латентными переменными гипотез, который был намечен к началу 1990-х гг. [Glymour et al., 1987] и анализировался, например, в [Девятко, 1991]. Другой, видимо, желательной перспективой остается разработка более обязывающих теорий и выводимых из них точечных гипотез и предсказаний.

## СПИСОК ЛИТЕРАТУРЫ [REFERENCES]

Девятко И.Ф. От «виртуальной лаборатории» до «социального телескопа»: метафоры тематических и методологических инноваций в онлайн-исследованиях // *Онлайн-исследования в России: тенденции и перспективы* / Под общ. ред. А.В. Шашкина, И.Ф. Девятко, С.Г. Давыдова. М.: ООО «Онлайн маркет интеллидженс», 2016. Гл. 1. С. 19–33. [Deviatko I.F. (2016) From «Virtual Lab» to «Social Telescope»: Metaphors of Theoretical and Methodological Innovations in Online Research. In: *Online-research in Russia: Trends and Prospects*. Ed. by A.V. Shashkin, I.F. Deviatko, S.G. Davydov. Moscow: Online Market Intellidgenсe: 19–33. (In Russ.)]

- Девятко И.Ф. TETRAD-методология: завершение процедурной эпистемы? // Вестник Академии наук СССР. 1991. № 2. С. 79–94. [Deviatko I.F. (1991) TETRAD-methodology: A Completion of Procedural Episteme? *Vestnik Akademii nauk SSSR* [Herald of the Russian Academy of Sciences]. No. 2: 79–94. (In Russ.)]
- Китчин Р. Большие данные, новые эпистемологии и смена парадигм / Пер. с англ. // Социология: методология, методы, математическое моделирование (4М). 2017. № 44. С. 111–152. [Kitchin R. (2017) Big Data, New Epistemologies and Paradigm Shifts. *Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie* (4M) [Sociology: Methodology, Methods, Mathematical Modeling (4M)]. No. 44: 111–152 (In Russ.)]
- Шродт Ф.А. Семь смертных грехов современного количественного анализа в политической науке / Пер. с англ. // Социология: методология, методы, математическое моделирование (4М). 2016. № 43. С. 154–210. [Schrodt Ph.A. (2016) Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie* (4M) [Sociology: Methodology, Methods, Mathematical Modeling (4M)]. No. 43: 154–210 (In Russ.)]
- Appelbaum M., Kline R., Nezu A., Cooper H., Mayo-Wilson E., Rao S.M. (2018) Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*. Vol. 73 (1): 3–25. DOI: 10.1037/amp0000191.
- de Groot A.D. (1956/2014) The Meaning of «Significance» for Different Types of Research. [Transl. and annot. by E.-J. Wagenmakers et al.] *Acta Psychologica*. Vol. 148 (May): 188–194.
- Doucoulagos, C., Stanley T.D. (2013) Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity. *Journal of Economic Surveys*. No. 27(2): 316–39.
- Gigerenzer G. (2004) Mindless Statistics. *The Journal of Socio-Economics*. Vol. 33(5): 587–606.
- Glymour C., Scheines R., Spirtes P. and Kelly K. (1987) *Discovering Causal Structure*. San Diego, CA: Academic Press.
- Graham J.H., Özener B. (2016) Fluctuating Asymmetry of Human Populations: A Review. *Symmetry*. Vol. 8(154): 1–36.
- Haller H., Kraus S. (2002). Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research*. No. 7(1): 1–20.
- Ioannidis J.P.A. (2005) Why Most Published Research Findings Are False. *PLoS Med*. No. 2(8): e124. DOI: 10.1371/journal.pmed.0020124.
- Kerr N.L. (1998) HARKing: Hypothesizing after the Results Are Known. *Personality and Social Psychology Review*. Vol. 2(3): 196–217.
- Lehrer J. (2010) The Truth Wears off: Is There Something Wrong with the Scientific Method? *The New Yorker*. December, 13.
- Munafò M.R., Nosek B.A., Bishop D.V.M., Button K.S., Chambers C.D., Percie Du Sert N., Simonsohn U., Wagenmakers E.J., Ware J.J., Ioannidis J.P.A. (2017) A Manifesto for Reproducible Science. *Nature Human Behaviour*. Vol. 1(0021). DOI:10.1038/s41562-016-0021.
- Nelson L., Simmons J.P., Simonsohn U. (2012) Let's Publish Fewer Papers. *Psychological Inquiry*. Vol. 23(3): 291–293.
- Nuijten M.B., Hartgerink C.H., van Assen M.A., Epskamp S., Wicherts J.M. The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods*. Vol. 48: 1205–1226.
- Oakes M. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley, 1986.
- Open Science Collaboration (Nosek B.A. et al.) (2015) Estimating the Reproducibility of Psychological Science. *Science*. 2015. No. 349(6251): aac4716. DOI: 10.1126/science.aac4716.
- Rosenthal R. (1979) The «File Drawer Problem» and Tolerance for Null Results. *Psychological Bulletin*. Vol. 86(3): 838–641.
- Rouder J.N., Speckman P.L., Sun D., Morey R.D., Iverson G. (2009) Bayesian t Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review*. No. 16(2): 225–237. DOI:10.3758/PBR.16.2.225.
- Selvin H.C. (1957) A Critique of Tests of Significance in Survey Research. *American Sociological Review*. Vol. 22(5): 519–527.
- Stapel D.A., Lindenberg S. (2011) Coping with Chaos: How Disordered Contexts Promote Stereotyping and Discrimination. *Science*. No. 332(6026): 251–253.
- Wagenmakers E.-J., Verhagen J., Ly A., Bakker M., Lee M.D., Matzke D., Rouder J.N., Morey R.D. (2015). A Power Fallacy. *Behavior Research Methods*. No. 47(4): 913–917. DOI: 10.3758/s13428-014-0517-4.
- Yong E. (2012) Replication Studies: Bad Copy. *Nature*. No. 485(7398): 298–300.
- Young C. (2018) Model Uncertainty and the Crisis in Science. *Socius: Sociological Research for a Dynamic World*. Vol. 4: 1–7. DOI: 10.1177/2378023117737206.

---

**NEW DATA, NEW STATISTICS: FROM REPRODUCIBILITY CRISIS TOWARD NEW REQUIREMENTS TO DATA ANALYSIS AND PRESENTATION IN SOCIAL SCIENCES****DEVIATKO I.F.***National Research University Higher School of Economics, Russia*

---

*Inna F. DEVIATKO, Dr. Sci. (Soc.), Full Professor, National Research University Higher School of Economics; Chief Researcher, Institute of Sociology FCTAS RAS, Moscow, Russia (deviatko@gmail.com).*

---

**Acknowledgment.** The paper is supported by the Russian Science Foundation, grant 17-78-20172.

**Abstract.** The article analyzes main causes and consequences of the interdisciplinary crisis of the reproducibility and reliability of the results of scientific research that has unfolded in the social sciences in parallel with the «data revolution». This crisis is expressed not only in the growing concern of scientists about the reliability of research results and the possibilities to establish the practices securing the transparency of empirical data and the statistical software used for their analysis, but also in disputes on limitations of the routine approach to significance testing and feasibility of alternatives based on Bayesian approach. Some aspects of the relationship between theory and data-driven methods of searching for patterns in empirical data are briefly discussed in the context of describing a new approach to multimodel analysis aiming at evaluation of model robustness and model uncertainty.

**Keywords:** reproducibility crisis in social sciences, transparency of data and open data, publication bias, null hypothesis significance testing and Bayesian approach, model robustness, data-driven approach and sociological theory.

*Received: 31.07.18. Accepted: 20.08.18.*