РАЗДЕЛ: Науки об обществе
Направление: Экономические науки

DOI 10.37539/2949-1991.2026.39.4.023

Антипов Сергей Константинович,
к.э.н., старший преподаватель,
ФГАОУ ВО СПбПУ

Кузьмицкая Екатерина Евгеньевна,
Студент, ФГАОУ ВО СПбПУ

Жукова Мария Александровна,
студент, ФГАОУ ВО СПбПУ

ЭВОЛЮЦИЯ СТАТИСТИЧЕСКОГО ВЫВОДА В ЭМПИРИЧЕСКИХ ИССЛЕДОВАНИЯХ: ОТ P-VALUE ФИШЕРА К СОВРЕМЕННЫМ ПОДХОДАМ ОЦЕНКИ ДОСТОВЕРНОСТИ РЕЗУЛЬТАТОВ

Аннотация. В статье представлен историко-методологический анализ эволюции концепции p-value – от её введения Р. Фишером в 1925 г. до современного кризиса воспроизводимости в эмпирических науках. Рассмотрены ключевые этапы развития теории проверки гипотез, включая противостояние фишеровского и неймано-пирсоновского подходов и формирование гибридной парадигмы NHST. Проанализированы типичные ошибки интерпретации, количественные масштабы p-hacking и низкой статистической мощности в психологии и биомедицине. Особое внимание уделено современным альтернативам и дополнениям к p-value, включая S-value, false positive risk и байесовскую калибровку. Предложен практический чек-лист ответственного использования p-value в биомедицинских исследованиях. Сформулированы конкретные сценарии, в которых p-value сохраняет свою полезность, вводит в заблуждение либо требует замены на альтернативные метрики.

Ключевые слова: P-value, проверка статистических гипотез, кризис воспроизводимости, статистическая значимость, p-hacking, статистическая мощность.

Современная статистика имеет богатый инструментарий, позволяющий получать качественные выводы, однако достоверность исследования зависит не только от метода оценки и применяемых инструментов, но и от корректности интерпретации полученных результатов [1]. Одним из самых востребованных методов оценки гипотез в настоящее время, безусловно, является p-value, который, однако, при всех своих достоинствах имеет как ряд ограничений, так и статистических ловушек, приводящих к серьезным заблуждениям и ошибкам [2]. Цель настоящего исследования состояла в описании этих особенностей в контексте эволюции статистического вывода и разборе возможных альтернатив.

Концепция p-value неразрывно связана с именем британского статистика и генетика Рональда Эйлмера Фишера. В 1925 г. в труде «Statistical Methods for Research Workers» он предложил использовать p-value как меру противоречия между наблюдаемыми данными и нулевой гипотезой. Хотя в исторической литературе первые расчёты p-value иногда приписывают Карлу Пирсону (статья 1900 г. в Biometrika), именно Фишер формализовал эту концепцию и распространил её среди экспериментаторов [3]. Работая на Ротамстедской сельскохозяйственной станции, Фишер решал прикладные задачи оценки эффективности удобрений и селекции растений, где требовался объективный критерий для отделения реальных эффектов от случайных колебаний. Согласно фишеровской логике, p-value оценивает силу доказательства против нулевой гипотезы в одном конкретном исследовании [4]. Фишер понимал p-value не как жёсткое правило, а как гибкий инструмент, помогающий

исследователю интерпретировать данные в контексте знаний о предметной области. Порог 0,05 был эмпирическим наблюдением, а не строгим предписанием.

В 1930-х гг. возникла альтернативная парадигма, разработанная Ежи Нейманом и Эгоном Пирсоном [5]. В отличие от фишеровского подхода, ориентированного на оценку силы свидетельств против конкретной нулевой гипотезы, теория Неймана-Пирсона вводила две конкурирующие гипотезы – нулевую (H_0) и альтернативную (H_1) – а также понятия ошибок первого рода (α) и второго рода (β). Исследователь до начала эксперимента фиксирует уровень α (обычно 0,05) и после получения данных принимает бинарное решение: отвергнуть H_0 или не отвергнуть. Как отмечают А.М. Гржибовский и А.Н. Гвоздецкий, Фишер резко критиковал этот подход за механистичность, тогда как Нейман и Пирсон указывали на отсутствие у фишеровского подхода чётких операциональных критериев [6].

В послевоенные годы в учебниках и исследовательской практике сформировался гибридный подход – Null Hypothesis Significance Testing (NHST), механически объединивший фишеровский p-value с неймано-пирсоновским порогом $\alpha = 0{,}05$. Формула «если $p < \alpha$, то отвергаем нулевую гипотезу» стала универсальным ритуалом. Гибридизация двух несовместимых подходов породила системную проблему: исследователи начали интерпретировать p-value одновременно и как фишеровскую меру свидетельства, и как неймано-пирсоновский инструмент контроля ошибок первого рода. Как подчёркивают специалисты, «уравнение фишеровского p-value с неймано-пирсоновской ошибкой первого рода представляет собой фундаментальную категориальную ошибку» [4]. Распространению гибрида способствовала его кажущаяся простота и давление публикационных требований, поощряющих «положительные» находки.

Многолетние опросы выявляют устойчивый набор заблуждений относительно смысла p-value. В работе А.М. Гржибовского и А.Н. Гвоздецкого систематизированы наиболее распространённые из них: (1) p-value как вероятность истинности нулевой гипотезы; (2) $p > 0{,}05$ как доказательство отсутствия эффекта; (3) отождествление статистической значимости с практической важностью; (4) прямое сравнение p-value из разных исследований [6]. С.В. Сивуха и А.А. Козяк отмечают, что критика проверки статистической значимости началась одновременно с популяризацией этой процедуры в психологии, а сведение научного вывода к толкованию p-значения является одной из серьёзных помех развитию психологической науки.

Ключевым событием стала публикация в 2015 г. результатов проекта Open Science Collaboration: из 100 повторений психологических экспериментов уровень успешной репликации составил лишь 36-47%, для социальной психологии – 25%. В биомедицинских науках мета-анализ 663 исследований показал, что примерно 50% работ имеют статистическую мощность в диапазоне 0-20%, что значительно ниже стандарта в 80%. Систематический обзор Э. Дюма-Малле и соавторов выявил среднюю мощность менее 20% в ряде областей биомедицины [7]. Как подчёркивается в отечественной литературе, существенные проблемы с интерпретацией результатов статистического анализа в биомедицинских исследованиях часто упоминаются в качестве одной из причин кризиса воспроизводимости научных результатов [8].

Отдельное внимание следует уделить p-hacking – совокупности приёмов для достижения статистической значимости. Систематический обзор выделяет 12 типичных стратегий: досбор данных, исключение «неудобных» наблюдений, множественное тестирование без поправок и другие. Эмпирическая оценка М. Хеда и соавторов с помощью методов интеллектуального анализа текста показала широкое распространение p-hacking в научной литературе, хотя его влияние на итоговые оценки величины эффекта оказалось относительно слабым по сравнению с реальными эффектами. В более поздних исследованиях, например в области визуализационной диагностики, систематических свидетельств p-hacking

обнаружено не было, однако авторы подчёркивают, что использованные методы не способны выявить все формы этой практики [7, 8].

В 2016 г. Американская статистическая ассоциация (ASA) опубликовала заявление с шестью принципами: (1) p-value может указывать на несовместимость данных с моделью; (2) не измеряет вероятность истинности гипотезы; (3) выводы не должны основываться исключительно на пороге; (4) требуется прозрачность; (5) p-value не измеряет величину или важность эффекта; (6) сам по себе не даёт надёжной меры доказательности [10]. Документ предостерегал от редукционистского подхода, но не призывал к полному отказу от p-value. ASA подчеркнула, что «широкое использование «статистической значимости» (обычно интерпретируемой как $p < 0{,}05$) в качестве лицензии на научное утверждение приводит к значительному искажению научного процесса» [11].

Доверительные интервалы дают информацию о диапазоне совместимых значений параметра. Величина эффекта (Cohen's d, отношение шансов) оценивает практическую значимость. Байесовский фактор напрямую оценивает относительную поддержку гипотез, показывая, во сколько раз данные более вероятны при одной гипотезе по сравнению с другой. Калибровка p-value по Т. Селлке, М. Баярри и Дж. Бергеру преобразует p-value в нижнюю границу байесовского фактора [12]. S-value ($s = -\log_2(p)$) интерпретирует p-value в терминах количества информации [13]. False positive risk (FPR) Д. Колкухуна оценивает вероятность того, что значимый результат является ложноположительным, используя байесовский подход с учётом априорной вероятности [14]. S-value переформулирует ту же информацию в более наглядной форме, тогда как FPR даёт принципиально иную – байесовскую – оценку риска.

В современной методологической литературе сформировался взвешенный консенсус: p-value остаётся полезным при сообщении вместе с величиной эффекта и доверительным интервалом; порог 0,05 не трактуется как магическая граница; исследователь осознаёт ограничения, связанные с мощностью, множественными сравнениями и предвзятостью публикаций. Журнал Basic and Applied Social Psychology в 2015 г. полностью запретил публикацию p-values, а The American Statistician в 2019 г. выпустил специальный номер, призвав «прекратить использовать термин “статистически значимый”».

Практический фреймворк: чек-лист для биомедицинских исследований. На основе рекомендаций EQUATOR Network предлагается следующий чек-лист:

1. Предварительный расчет мощности. Мощность не менее 80% для обнаружения клинически значимого эффекта.

2. Обязательное сопровождение p-value. Точечная оценка величины эффекта и 95% доверительный интервал.

3. Коррекция на множественные сравнения. Поправка Бонферрони, метод Холма или контроль FDR.

4. Предрегистрация. Протокол с первичными и вторичными конечными точками регистрируется до сбора данных.

5. Анализ чувствительности. Оценка false positive risk с реалистичными априорными вероятностями.

6. Открытые данные и код. Размещение в публичных репозиториях.

Инструменты автоматического анализа создают риски «алгоритмического p-hacking». Без предрегистрации и протокола множественного тестирования возможны массовые ложноположительные находки. Предрегистрация и открытый код создают «аудиторский след», позволяющий отличить эксплораторный анализ от конфирматорного. Исследования показывают, что предрегистрация теоретически предотвращает p-hacking и HARKing, однако её эффективность в борьбе с сомнительными исследовательскими практиками остаётся предметом дискуссий [15, 16].

Анализ трендов за 2015–2025 гг. показывает заметную трансформацию риторики. После заявления ASA в методологических разделах статей всё чаще подчёркиваются ограничения p-value. Журналы JAMA и Nature ввели требования, исключающие изолированное представление p-values. С 2016 по 2021 г. доля статей, использующих доверительные интервалы в дополнение к p-value, в ведущих медицинских журналах выросла примерно на 15%. В российском научном пространстве дискуссия развивалась медленнее, но к 2020-м гг. появились публикации, анализирующие кризис воспроизводимости.

Подводя итоги, можно отразить ситуации, в которых p-value достаточно: (1) в рандомизированных контролируемых испытаниях с мощностью ≥80% и предрегистрированным протоколом; (2) в исследованиях, где p-value дополняется величиной эффекта и доверительным интервалом, а порог 0,05 не трактуется бинарно; (3) в скрининге гипотез для дальнейшей проверки.

Также следует отметить ситуации, в которых p-value вводит в заблуждение и требует замены: (1) в эксплораторных исследованиях с множественным тестированием без поправок – целесообразны байесовские методы или кросс-валидация; (2) в исследованиях с низкой мощностью (<50%) – $p > 0,05$ не несёт информации об отсутствии эффекта, $p < 0,05$ вероятно ложноположителен; рекомендуется оценка величины эффекта с широкими доверительными интервалами; (3) при сравнении невложенных моделей – предпочтительны информационные критерии (AIC, BIC) или байесовский фактор; (4) в мета-анализах без коррекции на публикационную смещённость – рекомендуется использование моделей, корректирующих publication bias.

На вопрос «Чем заменять p-value без потери интерпретируемости?» можно предложить в качестве ответа следующие решения: (1) доверительные интервалы; (2) байесовский фактор (при обоснованных априорных распределениях); (3) S-value (интуитивно понятная мера); (4) false positive risk (при явном указании априорной вероятности).

Дальнейшая эволюция статистической практики будет определяться формированием культуры методологической прозрачности.

Список литературы:

1. Антипов С.К. Методика ведения статистического учета для оценки устойчивого развития регионов Арктики // Сборник статей International scientific conference, Гуманитарный национальный исследовательский институт НАЦРАЗВИТИЕ, Санкт-Петербург – 2025. С. 46-51
2. Антипов С.К. Статистика Том. 1. Основы статистического анализа // учебное пособие, издательство ФГАОУ ВО СПбПУ Политех-пресс, 2022
3. Fisher R.A. Statistical Methods for Research Workers. – Edinburgh: Oliver and Boyd, 1925.
4. Сивуха С.В., Козяк А.А. О реформе статистического вывода в психологии // Психология. Журнал Высшей школы экономики. – 2009. – Т. 6, № 4. – С. 66-86.
5. Neyman J., Pearson E.S. On the Problem of the Most Efficient Tests of Statistical Hypotheses // Philosophical Transactions of the Royal Society of London. Series A. – 1933. – Vol. 231. – P. 289-337.
6. Гржибовский А.М., Гвоздецкий А.Н. Интерпретация величины p и альтернативы её использованию в биомедицинских исследованиях // Экология человека. – 2022. – № 6. – С. 361–372. DOI: 10.33396/1728-0869-2022-6-361-372.
7. Dumas-Mallet E., Button K.S., Boraud T., Gonon F., Munafò M.R. Low statistical power in biomedical science: a review of three human research domains // Royal Society Open Science. – 2017. – Vol. 4, № 2. – 160254. DOI: 10.1098/rsos.160254.

8. Head M.L., Holman L., Lanfear R., Kahn A.T., Jennions M.D. The Extent and Consequences of P-Hacking in Science // PLoS Biology. – 2015. – Vol. 13, № 3. – e1002106. DOI: 10.1371/journal.pbio.1002106.

9. Hubbard R. Alphabet Soup: Blurring the Distinctions Between p's and α's in Psychological Research // Theory & Psychology. – 2004. – Vol. 14, № 3. – P. 295–327.

10. Gigerenzer G. Mindless statistics // The Journal of Socio-Economics. – 2004. – Vol. 33, № 5. – P. 587–606.

11. Open Science Collaboration. Estimating the reproducibility of psychological science // Science. – 2015. – Vol. 349, № 6251. – aac4716. DOI: 10.1126/science.aac4716.

12. Sellke T., Bayarri M.J., Berger J.O. Calibration of p Values for Testing Precise Null Hypotheses // The American Statistician. – 2001. – Vol. 55, № 1. – P. 62–71. DOI: 10.1198/000313001300339950.

13. Wasserstein R.L., Schirm A.L., Lazar N.A. Moving to a World Beyond "p < 0.05" // The American Statistician. – 2019. – Vol. 73, sup1. – P. 1–19. DOI: 10.1080/00031305.2019.1583913.

14. Colquhoun D. The reproducibility of research and the misinterpretation of p-values // Royal Society Open Science. – 2017. – Vol. 4, № 12. – 171085. DOI: 10.1098/rsos.171085.

15. Wasserstein R.L., Lazar N.A. The ASA Statement on p-Values: Context, Process, and Purpose // The American Statistician. – 2016. – Vol. 70, № 2. – P. 129–133. DOI: 10.1080/00031305.2016.1154108.

16. Raftery A.E. Bayesian Model Selection in Social Research // Sociological Methodology. – 1995. – Vol. 25. – P. 111–163. DOI: 10.2307/271063.