

КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И  
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Кафедра математической статистики

Информационный дайджест:

Математическая статистика

24 – 30 июня 2019 г.

21 марта 2019 г., источник: N+1

Для КМС ИВМиИТ подготовил Казанцев А.В.

## 850 ученых и один журнал высказались против статистической значимости



Ученые должны перестать подчиняться «диктату» статистической значимости в ее традиционной интерпретации (когда показатель статистической значимости  $p$  должен быть меньше 0,05) и перестать использовать эти категории в исследованиях, считают редакторы специального выпуска журнала *The American Statistician*. [Комментарий](#)

**нескольких авторов из этого выпуска для *Nature*, по их словам, поддержали более 850 человек (ни одного подписанта из России в списке нет).**

Как правило, в научном исследовании проверяется истинность гипотезы, которую обычно называют [нулевой](#), о том, что какой-либо эффект, связь, влияние факторов, разница между группами или другое явление в генеральной совокупности на самом деле отсутствуют — именно эта гипотеза считается верной, пока не доказано обратное. Анализируя имеющиеся данные, ученые с помощью статистических тестов оценивают, насколько вероятно, что эта гипотеза верна. Уровнем значимости теста называется вероятность так называемой ошибки первого рода: отклонения нулевой гипотезы, которая на самом деле верна.

Обычно пороговым значением, позволяющим ученому говорить о статистически значимом результате, считается [значение  \$p < 0,05\$](#) , но для некоторых тестов, например, в физике элементарных частиц, применяются более жесткие критерии. Научные журналы [предпочитают публиковать](#) именно статистически значимые результаты, и этот перекокс иногда приводит к тому, что авторы работ используют статистические манипуляции, чтобы подогнать результаты под этот критерий. Недавний пример разоблачения таких манипуляций — история Брайана Вонсинка, которую мы рассказывали в [блоге «Вы просто не умеете готовить»](#). Кроме того, критерий статистической значимости часто используется как показатель успешности всего исследования: мы [писали о неудаче](#) крупнейшего на данный момент эксперимента по нераспространению ВИЧ, вывод о провале которого был сделан именно потому, что разница между экспериментальной и контрольной группой оказалась не значимой. Наконец, статистически значимые результаты важны при перепроверке исследования: именно этот критерий считается критерием [воспроизводимости](#) работы.

В специальном выпуске журнала *The American Statistician* специалисты по статистике из разных стран в 43 статьях рассказывают, почему стандартный подход к статистической значимости изжил себя и требует переосмысления и реформы. Ранее Американская статистическая ассоциация уже выступала с [похожим призывом](#) не злоупотреблять статистической значимостью в 2016 году.

«Основываясь на нашем анализе статей в этом специальном выпуске и литературы по теме в целом, мы заключаем, что пришло время перестать использовать термин „статистически значимый“... Ни одно значение  $p$  не гарантирует убедительность, наличие, истинность или важность какой-либо связи или эффекта. Таким образом, ярлык статистической значимости не означает и не предполагает, что эта связь или эффект вероятны, реальны, истинны или важны. В свою очередь, ярлык статистической незначимости не означает, что они невероятны, отсутствуют, ложны или не важны. Для честности процесса научных публикаций и распространения информации о научных исследованиях, таким образом, не нужно оценивать, превышает ли значение  $p$  некий произвольно выставленный порог, чтобы решить, какие работы публиковать и продвигать», — говорится в редакционной статье.

[Одну из статей](#) в номере написал известный ученый-медик, статистик и научный активист Джон Иоаннидис, который называет использование оценки статистической значимости «наиболее часто (зло)употребляемым способом получения статистических выводов». Иоаннидис, в частности, подчеркивает, что снижение порога статистической значимости, к примеру, с 0,05 до 0,005 — это временная мера, которая лишь позволит частично сдержать вал некачественных работ и отсрочить то, что он называет «смертью от статистической значимости», когда эти на самом деле некачественные, но «статистически значимые» работы затопят собой науку.

В комментарии для *Nature* Блейк Макшейн (Blake McShane) из Северо-западного университета в США, который ранее [уже высказывался](#) по этому же поводу, и его коллеги напоминают, что статистически не значимый результат на самом деле не доказывает истинность нулевой гипотезы, а статистически значимый — не доказывает истинность некоей альтернативной гипотезы. Эти распространенные заблуждения, по их словам, искажают истинное положение вещей и даже приводят к научным конфликтам там, где их быть не должно.

Они приводят пример двух статистических анализов последствий приема противовоспалительных препаратов для сердечно-сосудистой системы. Они обнаружили одинаковую [точечную оценку](#) эффекта (20-процентное повышение риска мерцательной аритмии для пациентов, принимавших препараты), но из-за особенностей дизайна исследования [доверительный интервал](#) в одном из них составлял от минус 3 процентов до 48 процентов ( $p$ , по расчетам авторов комментария, равное 0,091, то есть результат не значим), а во втором — от 9 до 33 процентов ( $p=0,0003$ , то есть результат значим). Таким образом, считают авторы, хотя оба исследования указывают на то, что эффект, скорее всего, есть, одно из них можно использовать для обоснованного опровержения этого эффекта.

Они также напоминают, что нельзя недооценивать фактор чистой случайности в результатах: по этой самой случайности вы всегда можете совершить ошибку второго рода, то есть принять нулевую гипотезу в качестве верной, хотя на самом деле это не так. Оценить вероятность такой ошибки помогает параметр [статистической мощности](#) (чем она выше, тем вероятность меньше), а мощность, помимо заданного уровня значимости, зависит от величины самого искомого эффекта и величины выборки. Увеличение мощности — одна из задач метаанализов; подробнее об этой теме можно почитать в нашем [блоге «Я такая мета-мета»](#).

Ученые не призывают полностью отказаться от проверки уровня статистической значимости результата, но хотят, чтобы, во-первых, эта информация помещалась в более широкий контекст интерпретации выводов, а во-вторых, не расценивалась как стопроцентное подтверждение истинности того или иного утверждения, не оставляя пространства для естественной неопределенности в научном процессе.

Не все ученые согласны с призывом отказаться от понятия статистической значимости. Радиостанция NPR [цитирует](#) Дебору Майо, философа науки из Политехнического университета штата Виргиния, которая говорит, что отказ от этого критерия, в частности, позволит ученым занижать важность отрицательных результатов и также откроет дорогу к манипуляциям выводами. Кроме того, сам

журнал *Nature* в сопровождающей комментарий ученых редакционной заметке подчеркнул, что не собирается изменять свой подход к проверке статистического анализа в научных работах.

Другие интересные новости и большие материалы о качестве научных исследований можно найти в нашей постоянной [рубрике «Научные закрытия»](#).

*Ольга Добровидова*

<https://nplus1.ru/news/2019/03/21/p-value-protest>