ОБ АВТОРЕ

Лидия Дэнуорт (Lydia Denworth) — пишущий редактор *Scientific American*, автор книги «Дружба: эволюция, биология и невероятная сила основополагающей связи жизни» (*Friendship: The Evolution, Biology, and Extraordinary Power of Life's Fundamental Bond*, в печати).

СТАТИСТИКА

Значимая проблема

Стандартные научные методы служат мишенью для нападок. Изменится ли что-нибудь?

Лидия Дэнуорт

В 1925 г. британский генетик и статистик Рональд Фишер опубликовал книгу «Статистические методы для исследователей». По заголовку книгу не отнесешь к бестселлерам, но она имела огромный успех, и Фишера стали называть отцом современной статистики. В своей работе Рональд Фишер взялся за решение следующего вопроса: как исследователи могут применить статистические критерии к числовым данным, чтобы сделать выводы о том, что они обнаружили, и определить, следует ли принимать во внимание эти результаты? Фишер дает ссылку на статистический критерий, обобщающий согласованность данных с предложенной моделью, и представляет *p*-уровень значимости. Фишер предполагает, что исследователи могут рассматривать *p*-уровень, равный 0,05, как удобный руководящий принцип: «Представляется удобным взять это значение в качестве критерия для суждений о существенности или несущественности данного отклонения». Рассматривайте результаты

ОСНОВНЫЕ ПОЛОЖЕНИЯ

- Почти вековое использование *p*-уровня для определения статистической значимости экспериментальных результатов привело к возникновению иллюзии достоверности и кризису воспроизводимости результатов во многих областях науки.
- Все больше растет решимость реформировать статистический анализ, но между исследователями нет согласия в том, должна ли это быть просто корректировка или капитальная перестройка. Некоторые предлагают изменить статистические методы, а другие — ликвидировать порог для определения «значимых» результатов.
- В конечном счете *p*-уровень значимости играет на руку человеческой потребности в уверенности. Поэтому, возможно, и для ученых, и для общества настало время принять дискомфорт неопределенности.

Illustration by Mark Allen Miller

с *p*-уровнем ниже этого порога, советует Фишер, и не тратьте время на результаты с *p*-уровнем выше. Так родилась идея о том, что значение критерия *p* менее 0,05 равноценно тому, что называется статистической значимостью, — математическому определению «значимых» результатов.

Почти столетие спустя во многих областях науки значение *p*-критерия менее 0,05 рассматривается как золотой стандарт для определения достоинств эксперимента. Он открывает двери к важнейшим составляющим научного мира — финансированию и публикациям — и таким образом подкрепляет большинство опубликованных научных выводов. И тем не менее даже Фишер понимал, что концепция статистической значимости и подкрепляющего ее значения *p*-уровня имеет существенные ограничения. Наличие подобных ограничений признавали в течение десятилетий. «Чрезмерная опора на проверку значимости, — писал психолог Пол Мил (Paul Meehl) в 1978 г., — плохой способ заниматься наукой». *P*-уровни значимости постоянно неверно истолковывают, а статистическая и практическая значимость — это не одно и то же. Более того, методологические решения, необходимые в любом исследовании, позволяют экспериментатору сознательно или неосознанно повышать или понижать *p*-уровень. «Как говорится, с помощью статистики можно доказать что угодно», — говорит статистик и эпидемиолог Сандер Гринленд (Sander Greenland), почетный профессор Калифорнийского университета в Лос-Анджелесе, один из ведущих ученых, выступающих за реформу. Исследования, в которых полагаются только на достижение статистической значимости или доказательство ее отсутствия, постоянно приводят к ошибочным утверждениям: в них доказывается истинность ложных вещей и ложность настоящих фактов. После того как Фишер вышел в отставку (на тот момент он работал в Австралии), ему задали вопрос, было ли что-нибудь за его долгую карьеру, о чем он сожалеет. Говорят, он отрезал: «Что вообще упомянул о 0,05».

В последнее десятилетие спор о статистической значимости вспыхнул с необычной силой. В одной публикации непрочную основу статистического анализа назвали «самым грязным секретом науки». В другой упоминали о «ряде серьезных недостатков» в определении значимости. Экспериментальная экономика, биомедицинские исследования и особенно психология охвачены вызывающим споры кризисом повторяемости результатов: обнаружилось, что существенная часть опубликованных данных о новых открытиях невоспроизводима. Один из самых известных примеров — идея о властных позах. Утверждение о том, что агрессивный язык тела меняет не только ваше отношение, но и уровень гормонов, базировалось на статье, от которой с тех пор отрекся один из ее

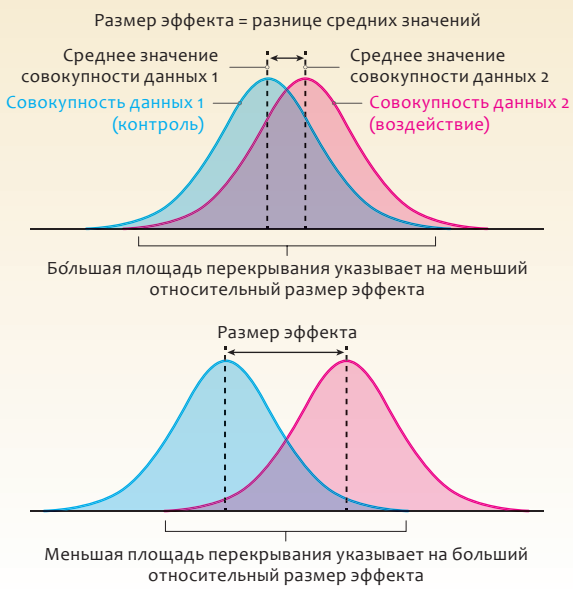
Статистическая значимость

Представьте, что вы выращиваете тыкву в своем саду. Повлияет ли использование удобрения на размер плодов? Исходя из своего опыта без применения удобрений вы знаете, насколько различается вес тыкв и что в среднем масса тыквы составляет 10 фунтов (4,54 кг). Вы решаете вырастить 25 тыкв (это ваша выборка), применяя удобрение. Средняя масса этих 25 тыкв оказалась равна 13,2 фунта (5,99 кг). Как определить, стало ли увеличение средней массы на 3,2 фунта (1,45 кг) по сравнению с прежним весом — гипотетическим «нулевым» уровнем — случайным или при применении удобрения действительно вырастают более крупные тыквы?

Предложенное статистиком Рональдом Фишером решение этой загадки связано с выполнением мысленного эксперимента: представьте, что вы собрались вырастить по 25 тыкв много раз. Каждый раз вы получите разную среднюю массу из-за случайной изменчивости отдельной тыквы. Затем вы строите кривую распределения этих средних и рассматриваете вероятность (*p*-уровень) того, что полученные вами данные были бы возможны, если бы удобрение не оказывало влияния. По договоренности *p*-уровень, равный 0,05, стал порогом для определения значимых результатов: в данном случае таких, на основе которых исследователь сделает вывод о том, что удобрение не имеет эффекта. Здесь мы рассмотрим некоторые концепции, лежащие в основе мысленного эксперимента для определения статистической значимости.

РАЗМЕР ЭФФЕКТА

Размер эффекта — это различия между средним результатом, полученным, когда применяется воздействие, по сравнению со средним значением, когда воздействие не применяется. Концепцию можно использовать для сравнения средних в выборках или «истинных» средних для всех распределений. Размер эффекта может выражаться в тех же единицах (например, в случае с массой тыкв — в фунтах/килограммах), что и результат. Но для многих результатов, таких как ответы на вопросы в опросниках по физиологии, не существует натуральных единиц измерения. В этом случае исследователь может использовать относительные размеры эффекта. Один из способов измерения относительного размера эффекта основан на перекрывании кривых контрольного и экспериментального распределения.

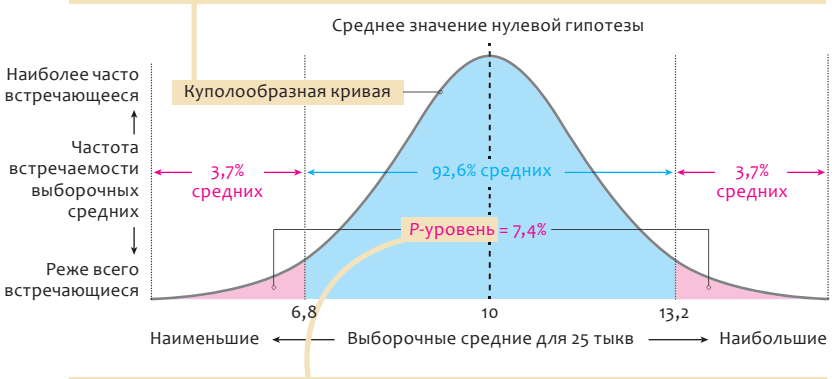


Graphic by Amanda Montañez (graphs) and Heather Krause

P-УРОВЕНЬ ЗНАЧИМОСТИ

Для вычисления *p*-уровня нам нужно сравнить среднее значение (13,2 фунта), действительно наблюдаемое нами в выборке из 25 тыкв, с распределением случайных средних значений, если бы мы взяли множество новых выборок из 25 тыкв.

Куполообразная кривая показывает распределение случайных средних значений массы для выборок из 25 тыкв при нулевой гипотезе о том, что удобрение не обладает эффектом.

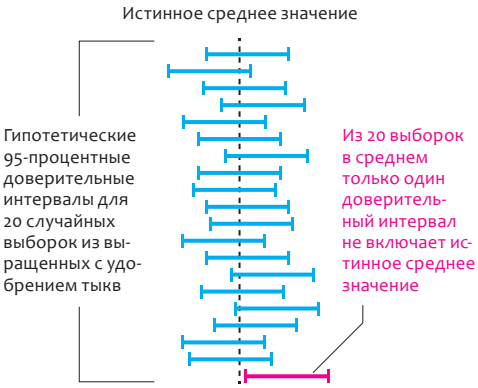


P-уровень значимости — это вероятность получить случайное среднее значение массы, отличающееся от 10 так же, как отличается действительно наблюдаемая средняя, 13,2. Поскольку 13,2 – 10 = 3,2, мы говорим о вероятности получить среднее значение ≥ 13,2 или ≤ 6,8 (6,8 = 10 – 3,2). В этом случае такая вероятность равна 0,074 — действительно наблюдаемый *p*-уровень для вашей выборки. Поскольку значение больше 0,05, ваш результат не будет считаться значимым доказательством того, что разница в весе тыкв связана с применением удобрения.

В примере показана двусторонняя проверка, «с двумя хвостами», где при определении значения *p*-уровня принимается во внимание вероятность получить среднюю массу больше 13,2 фунтов и меньше 6,8 (10 – 3,2 = 6,8). При определенных условиях исследователь может выбрать выполнение односторонней проверки, «с одним хвостом». В этом случае *p*-уровень был бы лишь 0,037, и поскольку он меньше 0,05, то считался бы значимым. Данный пример иллюстрирует один из способов, используя который, исследователи могут изменить свои заявленные намерения в отношении исследования, чтобы получить другие *p*-уровни при тех же данных.

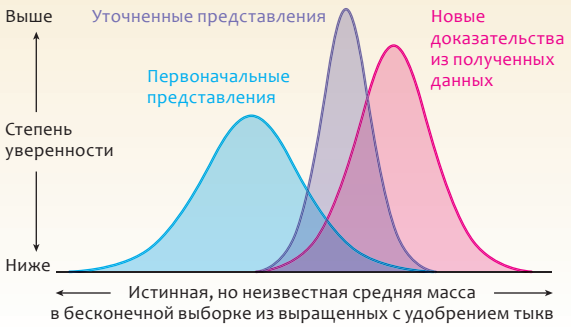
ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ

Мы можем вычислить 95-процентный доверительный интервал для нашей выборки из 25 тыкв. Это догадка о том, какова средняя масса тыкв, выращенных с применением удобрения. Расчет 95%-процентного доверительного интервала включает вычисление, обратное вычислению *p*-уровня, для того чтобы найти все гипотетические значения, при которых *p*-уровень ≥ 0,05. Для нашей выборки из 25 тыкв 95-процентный доверительный интервал — от 9,69 до 16,71. «Истинная» средняя масса выращенных с удобрением тыкв может попасть, а может и не попасть в этот интервал. Мы не можем быть уверены, что же тогда означает «95%». Представьте, что произошло бы, если бы мы многократно выращивали партии из 25 тыкв и брали из них образцы. В каждой выборке был бы случайный, разный доверительный интервал. Мы знаем, что за длительный период времени 95% этих интервалов будут включать истинное значение средней, а 5% — нет. Но что же с нашим конкретным интервалом для первой выборки? Неизвестно, попадает ли он в 95%, которые работают, или в 5%. Это процесс, который верен на протяжении 95% времени.



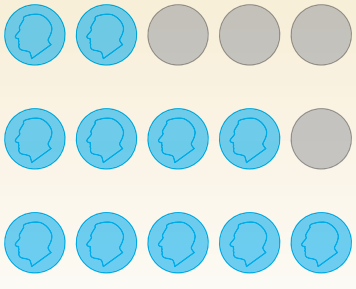
БАЙЕСОВСКИЕ МЕТОДЫ

При байесовском подходе состояние неуверенности человека, касающееся неизвестной величины, представлено распределением вероятности. Теорема Байеса используется для объединения первоначальных убеждений, существующих у индивида, — их распределением до обращения к данным — с информацией, получаемой из данных, которая дает предполагаемое математическое распределение для уточненных представлений. Уточненные представления, полученные в результате одного исследования, становятся новыми начальными представлениями для следующего и т.д. Большинство обсуждений и споров касаются попыток найти «объективные» критерии для первоначальных представлений. Необходимо отыскать способы задания априорных распределений, которые большинство исследователей могли бы считать приемлемыми.



СОБСТВЕННАЯ ИНФОРМАЦИЯ

P-уровень выражает, насколько удивительны наши данные о тыквах, если предполагается, что в действительности применение удобрения не влияет на рост тыкв. Некоторые исследователи полагают, что *p*-уровни не отражают «удивительность» способом, интуитивно понятным для большинства людей. Ученые предлагают вместо этого использовать математическую величину, называемую собственной информацией (мерой «удивительности»), которая известна также как *s*-уровень, или преобразование Шеннона. Она выражает *p*-уровни в битах (как в компьютерных битах). Критерий «собственная информация» можно объяснить с помощью примера с подбрасыванием монеток.



Два орла подряд = 2 бита информации = *p*-уровню 1/22 = 0,25

Четыре орла подряд = 4 бита информации = *p*-уровню 1/24 = 0,0625

Пять орлов подряд = 5 битов информации = *p*-уровню 1/25 = 0,03215

Выборка из 25 тыкв со средней массой 13,2 фунта и *p*-уровнем значимости 0,074 дает количество информации, равное 3,76 бита, поскольку 3,76 = $-\log_2(0{,}074)$.

авторов. Статья об экономике изменений климата (написанная скептиком) «заканчивалась почти таким же числом поправок, сколько в ней было исходных данных, — я не шучу! Но этих поправок для него было недостаточно, чтобы изменить выводы». Так написал статистик из Колумбийского университета Эндрю Гелман (Andrew Gelman) в своем блоге, где он постоянно критикует исследователей за некачественную работу и нежелание признавать наличие проблем в их трудах. Гелман считает: «Нет ничего страшного в том, чтобы заниматься чисто теоретической работой, но тогда не надо сбивать нас с толку данными».

Концепция статистической значимости, хотя она и не единственный фактор, представляет собой очевидную составляющую проблемы. В последние три года сотни исследователей настойчиво требуют реформы, выступая авторами или одобряя статьи в престижных журналах о пересмотре статистической значимости или отказе от нее вообще. Американская статистическая ассоциация (*ASA*), сделавшая в 2016 г. серьезное и необычное заявление по этой проблеме, выступает за «переход к миру за пределами $p < 0{,}05$». Рональд Вассерштейн (Ronald Wasserstein), исполнительный директор *ASA*, говорит об этом так: «Предполагается, что статистическая значимость указывает всего лишь на определенный уровень интереса. Но, к сожалению, это не то, чем на самом деле стала статистическая значимость. Люди говорят: "Я добился 0,05. Я молодец". Наука останавливается».

Вопрос в том, изменится ли что-нибудь. «Все это не ново. Нас должно отрезвить то, что и в этот раз, возможно, все будет так же, как раньше», — говорит Дэниел Бенджамин (Daniel Benjamin), специалист по поведенческой экономике из Университета Южной Калифорнии, тоже выступающий за реформу. И все же, хотя существуют разногласия по поводу средств, поразительно, как много исследователей сходятся во мнении, что, как написал экономист Стивен Зилиак (Stephen Ziliak), «существующая культура проверки статистической значимости, интерпретации и составления отчетов должна исчезнуть».

Мир, каков он есть

Цель науки — описать, что в природе истинно. Ученые применяют статистические модели, чтобы сделать выводы об истинном: например, определить, более ли эффективно одно лечение по сравнению с другим или отличается ли одна группа от другой. Каждая статистическая модель основывается на ряде допущений о том, как собираются и анализируются данные и как исследователи представляют свои результаты.

Эти результаты почти всегда базируются на статистическом подходе, называемом проверкой значимости нулевой гипотезы, которая дает

p-уровень. Такая проверка не подразумевает непосредственного определения истины. Это лишь косвенный мимолетный взгляд, поскольку проверка значимости предназначена лишь для того, чтобы указать, стоит ли и дальше двигаться в этом направлении в исследовании. «Что мы желаем знать, когда ставим эксперимент, — так это насколько вероятно то, что наша гипотеза верна, — говорит Бенджамин. — Но [проверка значимости] отвечает на сложный альтернативный вопрос: если бы моя гипотеза была ложной, насколько невероятны были бы мои данные?»

Иногда это работает. Поиск бозона Хиггса, частицы, существование которой впервые было теоретически предсказано физиками в 1960-х гг., — исключительный, но удобный пример. Нулевая гипотеза состояла в том, что бозона Хиггса не существует; альтернативная гипотеза — что он должен существовать. Команда физиков из *CERN* провела множество экспериментов на Большом адронном коллайдере и получила эквивалент *p*-уровня настолько исчезающе малый, что это означало: вероятность получить имевшиеся результаты, если бы бозона Хиггса не существовало, была равна 1 к 3,5 млн. Таким образом, нулевая гипотеза стала несостоятельной. Затем исследователи дважды проверили результат, чтобы убедиться, что он не вызван ошибкой. «Единственный способ быть уверенным в научной важности результата — и Нобелевской премии — сообщить, что [они] прошли сквозь огненные кольца, чтобы убедиться, что [ни одна] из потенциальных проблем не могла дать такое малое значение, — говорит Гринленд. — Такой ничтожный уровень говорит, точнее, кричит, что Стандартная модель без бозона Хиггса [не может быть верна]».

Однако такой уровень точности, которого позволяет достичь физика, недостижим везде. При тестировании людей, например в психологии, невозможно получить вероятность 1 к 3 млн. При *p*-уровне значимости, равном 0,05, вероятность многократного отклонения правильной гипотезы в результате множества тестов составляет 1 к 20. (*p*-уровень = 0,05 не указывает, как часто считают, что вероятность ошибки каждого отдельного теста составляет 5%). Вот почему статистики давно используют доверительные интервалы как средство для обеспечения представления о величине ошибки или неопределенности в оценках, сделанных учеными. Доверительные интервалы математически связаны с *p*-уровнем значимости. Значение *p*-уровня находится в пределах от 0 до 1. Вычитая 0,05 из 1, получим 0,95, или 95%, общепринятый доверительный интервал. Однако доверительный интервал — это просто удобный способ обобщения результатов проверки гипотезы для разных размеров эффектов. «В них нет ничего, что должно было бы внушать доверие», — отмечает Сандер Гринленд. Тем не менее со временем такие критерии, как *p*-уровни значимости и доверительные интервалы, закрепились, создавая иллюзию достоверности.

P-уровни значимости сами по себе не обязательно представляют проблему. Это удобный инструмент, когда рассматривается в контексте. Как заявляют редакторы журналов, спонсоры науки и регуляторы, так они и поступают. Беспокойство вызывает тот факт, что важность такого показателя, как статистическая значимость, может преувеличиваться или чрезмерно подчеркиваться, что особенно легко сделать при малых выборках. Именно это и привело к существующему кризису повторяемости результатов. В 2015 г. Брайан Носек (Brian Nosek), сооснователь «Центра открытой науки» (*Center for Open Science*), возглавил проект, в котором была предпринята попытка воспроизвести эксперименты, описанные в 100 известных статьях по социальной психологии; выяснилось, что лишь 36,1% результатов можно точно повторить. В 2018 г. сообщалось о результатах точного повторения 21 экспериментального исследования по социологии, статьи о которых были опубликованы в журналах *Nature* и *Science* в период с 2010 г. по 2015 г., в рамках проекта «Социология: повторяемость» (*Social Sciences Replication Project*). Ученые обнаружили значимый эффект того же направления, что и в оригинальном исследовании, для 13 (62%) экспериментов, а размер эффекта в повторных исследованиях в среднем составлял половину от оригинального размера эффекта.

В середине 2000-х гг. в генетике тоже наблюдался кризис повторяемости. После многих споров порог статистической значимости в этой области значительно изменился. «Когда обнаруживается новая генетическая вариация, связанная с какой-либо болезнью или другим фенотипом, стандарт для статистической значимости составляет 5×10^{-8}, то есть 0,05, поделенные на миллион, — говорит Дэниел Бенджамин, который также работал в сфере генетики. — Нынешнее поколение исследований в области генетики человека считается очень надежным».

Однако нельзя сказать то же самое в отношении биомедицинских исследований, где наблюдается тенденция к ложноотрицательным результатам при оценке риска: исследователи сообщают об отсутствии статистической значимости, когда эффекты существуют. Отсутствие доказательств не есть доказательство отсутствия, точно так же как отсутствие обручального кольца на пальце не означает, что человек не женат / не замужем, а лишь свидетельствует о том, что человек не носит кольцо. Иногда такие случаи заканчиваются в суде, когда на карту поставлены корпоративная ответственность и безопасность потребителя.

Размывание четких границ

Так насколько серьезны проблемы в науке? Ученые, работающие в разных областях, в основном согласны в том, что неверная интерпретация и переоценка критерия *p* и статистической значимости представляют реальную проблему, хотя некоторые более сдержанны в оценке ее серьезности, чем другие. «Я провел анализ за длительное время, — рассказывает социальный психолог Блэр Джонсон (Blair T. Johnson) из Университета Коннектикута. — В науке это происходит постоянно. Маятник будет колебаться между экстремумами, и вам придется с этим жить». Преимущество этого этапа, говорит Джонсон, в том, что он служит напоминанием: надо быть сдержанным в выводах. «Если мы, ученые, не будем скромны, то не сможем двигаться дальше».

Сообщение о важном открытии должно не быть в виде одного предложения, а занимать целый параграф, и оно не должно базироваться на единственном исследовании. В конце концов, удачная теория — та, что десятилетиями выдерживала проверку в повторных исследованиях

Тем не менее, для того чтобы по-настоящему двигаться дальше, ученые должны договориться о решениях. Это почти настолько же трудно, насколько сложны и сами методы статистики. Рональд Вассерштейн считает: «Боятся, что с исключением этой давно устоявшейся практики, позволяющей утверждать, что вещи статистически значимы или нет, в процессе наступит своего рода анархия». И все же предложений масса. Они касаются изменения статистических методов, терминов, использующихся для описания этих методов, и способов применения статистического анализа. Наиболее существенные идеи были выдвинуты в серии статей, последовавших после заявления *ASA* в 2016 г., в которых более двух десятков статистиков пришли к общему мнению относительно некоторых принципов реформы. Затем последовал специальный выпуск одного из журналов ассоциации, в котором 45 статей были посвящены способам выхода за рамки статистической значимости.

В 2018 г. группа ученых опубликовала в журнале *Nature Human Behaviour* комментарий под названием «Новое определение статистической значимости», поддерживая изменение порога статистической значимости с 0,05 до 0,005 для объявления о новом открытии. (Результаты при уровне значимости от 0,05 до 0,005 будут называться «предположительными»). Дэниел Бенджамин, ведущий автор этой статьи, рассматривает эту меру как несовершенное и временное решение, однако его можно внедрить немедленно: «Я беспокоюсь, что если не сделать что-нибудь прямо сейчас, то мы в конечном итоге потратим все время на споры об идеальном решении и упустим момент для более серьезных изменений, которые по-настоящему улучшат положение вещей. А между тем вреда будет намного больше». Иными словами, нельзя допустить, чтобы лучшее было врагом хорошего.

Другие ученые полагают, что переопределение порога статистической значимости вообще не даст никаких улучшений, поскольку настоящая проблема заключается в самом существовании такой границы. В марте Сандер Гринленд из Калифорнийского университета в Лос-Анджелесе, Валентин Амрхайн (Valentin Amrhein), зоолог из Базельского университета, и Блэйкли Макшейн (Blakeley McShane), статистик и эксперт по маркетингу из Северо-Западного университета, опубликовали в журнале *Nature* заметку, в которой выступают за отказ от концепции статистической значимости. Исследователи предлагают использовать *p*-уровень как непрерывную переменную среди других элементов доказательств и переименовать доверительные интервалы в «интервалы совместимости» для отражения того, на что они действительно указывают: на совместимость с данными, а не на достоверность результатов. Гринленд с соавторами попросили поддержать их идеи в *Twitter*, и за высказались 800 ученых, в том числе и Бенджамин.

Несомненно, существуют лучшие — или по крайней мере более эффективные — статистические методы. Эндрю Гелман, часто критикующий статистические подходы других ученых, в своей работе вообще не использует проверку значимости нулевой гипотезы. Он предпочитает байесовскую методологию, прямой статистический метод, при котором у ученого уже имеется исходное мнение, затем добавляются новые свидетельства и первоначальные представления пересматриваются. Сандер Гринленд продвигает использование такого критерия, как собственная информация (или метрика *surprisal*, мера «удивительности»), — математической величины, выражающей *p*-уровни в битах (как компьютерные биты) информации. *P*-уровень значимости 0,05 — это всего 4,3 бита

информации по сравнению с нулем. «Это равносильно тому, что при подбрасывании монетки орел выпадет четыре раза подряд, — рассказывает Гринленд. — Достаточное ли это свидетельство против идеи о том, что подбрасывание монетки было честным? Нет. Вы увидите, что так происходит постоянно. Вот почему *p*-уровень значимости 0,05 — слабый стандарт». Если бы исследователи определяли количество информации для каждого *p*-уровня, считает Гринленд, им бы пришлось придерживаться более высокого стандарта. Помог бы также акцент на размерах эффектов, указывающих на величину обнаруженных различий.

Повышение уровня образования в области статистики (как для ученых, так и для общества) могло бы сделать язык статанализа более доступным. В то время, когда Фишер использовал концепцию «значимости», это слово имело меньшую смысловую нагрузку. «Термин означал "имеющий значение", а не "важный"», — говорит Гринленд. Поэтому неудивительно, что термин «доверительные интервалы», по-видимому, вызывает чрезмерное доверие.

Восприятие неопределенности

Статистическая значимость подпитывала человеческую потребность в уверенности. «Первородный грех, так сказать, заключается в том, что люди хотят определенности, когда это неуместно», — говорит Эндрю Гелман. Возможно, для нас настало время принять дискомфорт неуверенности. Если мы сможем сделать это, то научная литература будет выглядеть по-другому. «Сообщение о важном открытии должно не быть в виде одного предложения, а занимать целый параграф», — говорит Вассерштейн. И оно не должно базироваться на единственном исследовании. В конце концов, удачная теория — та, что десятилетиями выдерживала проверку в повторных исследованиях.

Небольшие изменения происходят среди властвующих в науке сил. «Мы согласны, что иногда *p*-уровнями злоупотребляют или их неправильно интерпретируют, — говорит Дженнифер Цейс (Jennifer Zeis), представитель *New England Journal of Medicine*. — Выводы о том, что лечение эффективно, если $p < 0{,}05$, и неэффективно, если $p > 0{,}05$, — это упрощенный взгляд на медицину, не всегда отражающий реальность». Дженнифер рассказывает, что в размещаемых в их журнале отчетах об исследованиях теперь реже содержатся ссылки на *p*-уровни и все больше результатов приводится с указанием доверительных интервалов, без *p*-уровня. Журнал также придерживается принципов открытой науки: публикует более детальные протоколы исследований и требует от авторов, чтобы они следовали предварительно определенному плану анализа и сообщали, если отступают от этого плана.

По словам Джона Скотта (John Scott), руководителя отдела биостатистики Управления по санитарному надзору за качеством пищевых продуктов и медикаментов США (*FDA*), предъявляемые *FDA* требования к клиническим испытаниям не изменились. «Маловероятно, что в ближайшее время определение *p*-уровней значимости исчезнет из исследований, связанных с разработкой лекарств, но полагаю, что альтернативные подходы будут применяться все чаще, — говорит Джон Скотт. — Например, среди соискателей растет интерес к использованию байесовских выводов. Современные споры отражают общий рост обеспокоенности в связи с ограниченностью традиционно использующихся статистических выводов».

Блэр Джонсон, новый редактор журнала *Psychological Bulletin*, разделяет взгляды нынешнего редактора, но говорит: «Я собираюсь заставить соответствовать довольно строгим стандартам представления отчетов. Таким образом я буду уверен, что все знают, что происходит и почему, и исследователям будет проще решить, надежны ли их методы или в них есть недостатки». Джонсон также подчеркивает важность хорошо выполненного метаанализа и систематических обзоров как способов снизить зависимость от результатов отдельных исследований.

«Важнее всего, чтобы *p*-уровень значимости не играл роль привратника, — говорит Блэйкли Макшейн. — Давайте придерживаться более целостной точки зрения и в оценках будем внимательны к деталям». Подобное мнение разделяли даже современники Рональда Фишера. В 1928 г. два других гиганта статистики, Юрий (Ежи) Нейман и Эгон Пирсон, писали о статистическом анализе: «Сами по себе проверки не служат окончательным вердиктом, но они помогают исследователю, использующему их в качестве инструментов, принять окончательное решение». ■

Перевод: С.М. Левензон

ДОПОЛНИТЕЛЬНЫЕ ИСТОЧНИКИ

■ Палус Ш. Обеспечить воспроизводимость // ВМН, № 12, 2018.

■ Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. Colin F. Camerer et al. in Nature Human Behaviour, Vol. 2, pages 637–644; September 2018.

■ Moving to a World beyond “p < 0.05.” Ronald L. Wasserstein, Allen L. Schirm and Nicole A. Lazar in American Statistician, Vol. 73, Supplement 1, pages 1–19; 2019.