

КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ИНСТИТУТ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И  
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Кафедра математической статистики

Информационный дайджест:

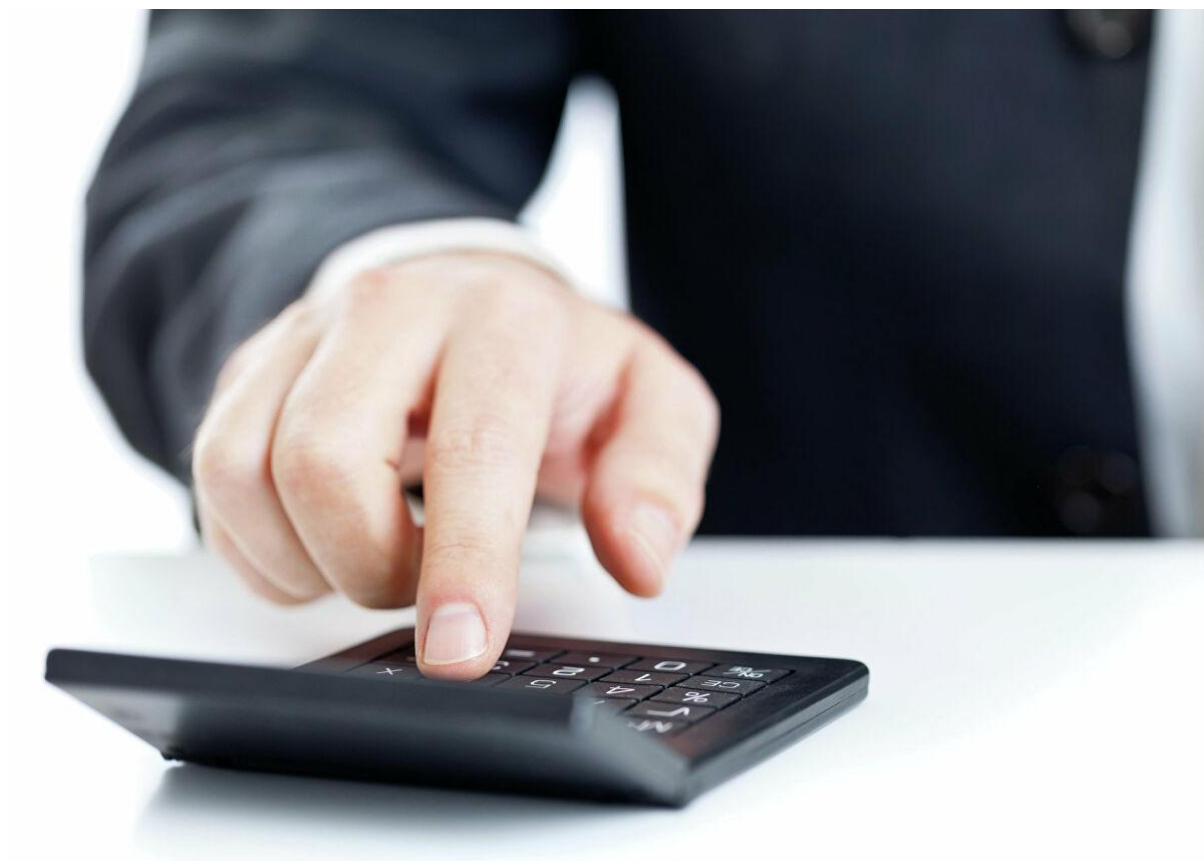
Математическая статистика

17 – 23 июня 2019 г.

13 ноября 2013 г., источник: РИА Наука Для КМС ИВМиИТ подготовил Казанцев А.В.

## **Статистическая значимость и ее роль в неудачах науки**

[14:20 13.11.2013](#) (обновлено: 10:20 14.11.2013)



**Статистическая значимость обычно воспринимается так, как будто она является волшебной палочкой.**

Материал подготовлен редакцией [ИноСМИ](#) специально для раздела РИА Наука>>

Хильда Бастиан (Hilda Bastian)



### **Hilda Bastian**

Hilda Bastian was a health consumer advocate in Australia in the '80s and '90s. Controversies riddled with ideology and vested interests drove her to science. Epidemiology and effectiveness research have kept her hooked ever since.

Представьте себе возможность существования одной статистической характеристики, которую все могли бы использовать с любым набором данных и которая была бы способна отделять истинное от ложного. О! Какие тогда мы могли бы узнать вещи! Однако было бы нереалистичным ожидать такого рода чародейства, разве не так?

Тем не менее статистическая значимость обычно воспринимается так, как будто она является волшебной палочкой. Возьмите нулевую гипотезу или поищите любую ассоциацию между факторами в наборе данных и абракадаброй. Возьмите «Р-значение» (p value) больше или меньше 0,05, и вы можете быть уверенным на 95% в том, что это либо счастливая случайность, либо нет. Вы в состоянии устранить игру случайности! Вы способны отделить сигнал от шума!

Но проблема в том, что вы не можете этого сделать. И, на самом деле, этим не занимается тестирование статистической значимости. В этом и состоит загвоздка.

Проведение тестирования на статистическую значимость оценивает лишь вероятность получения похожего результата с другим набором данных при сохранении тех же самых условий. Однако оно предоставляет ограниченную картину такой вероятности, поскольку в расчет принимается ограниченное количество информации относительно исследуемых данных. И оно само по себе не способно вам сказать, являются ли основные положения исследования верными и будут ли подтверждены полученные результаты в различных условиях.

Более того, обнаружение самой статистической значимости может оказаться «счастливой случайностью», и подобный вариант становится более вероятным при работе с большим количеством данных, чем при проведении теста на множественные сравнения с теми же данными.

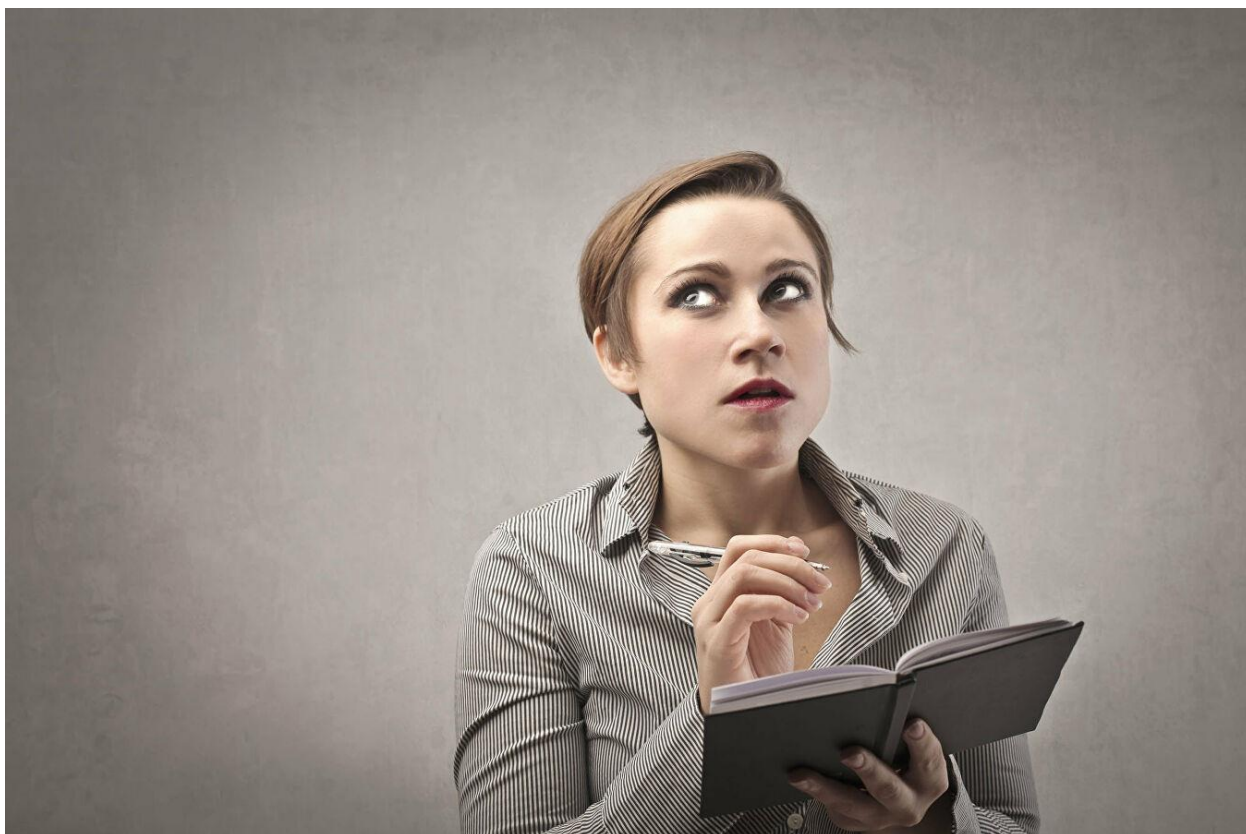
Тестирование статистической значимости может показаться легким занятием, как будто речь идет об отделении зерен от плевел, но одного этого оказывается недостаточно – и оно может развалиться, столкнувшись с большим количеством вызовов. Да и не все тесты на статистическую значимость работают одинаково с разным набором данных. Кроме того,

«значимость» еще не означает, что это важно. Разброс воздействия может опуститься ниже порога в 5%. Вскоре мы вернемся к вопросу о том, что это означает на практике.

Общий подход к тестированию статистической значимости было так просто понять и так легко было его провести еще даже до появления компьютеров, что он стремительно завладел научным миром. Как показывает Стивен Стиглер (Stephen Stigler) в своей работе о Фишере (Fischer) и 5-процентном уровне, «он открыл миру экспериментаторов и исследователей тайную область статистических вычислений».

Однако это также привело к своего рода лавине злоупотреблений. Излишне упрощенный подход к статистической значимости несет ответственность за многое. Как отмечает по этому поводу Джон Иоаннидис (John Ioannidis), существует серьезный игрок, имеющий отношение к неспособности науки воспроизводить результаты.

Прежде чем пойти дальше, я должна кое в чем признаться. Я не являюсь статистиком, но я уже в течение долгого времени занимаюсь объяснением статистических концепций. Я также очень долго пользовалась простым путем в отношении этого предмета. Но теперь я считаю, что увековечивание чрезмерно простых способов объяснения в большом количестве тренингов является главной составной частью проблемы.



© [Fotolia / olly](#)

Необходимость добиваться лучших результатов в отношении того, что означает статистическая значимость, и что она не означает, привела к проведению «часа вопросов» в ходе нашей панельной дискуссии, посвященной числам, на недавней ежегодной встрече во Флориде Национальной ассоциации авторов, пишущих о науке (National Association of Science Writers).

Такой же энтузиаст в области статистики и блоггер на портале SciAm Кэтлин Рейвен (Kathleen Raven) организовала и провела панельную дискуссию с участием меня, математического блоггера SciAm Эвелины Лэм (Evelyn Lamb), профессора статистики Реджины Наззо (Regina Nuzzo) и математика Джона Полоса (John Allen Paulos). Рейвен в настоящее время занимается организацией постоянного блога под названием «Шум и числа» (Noise and Numbers) при участии именно этой команды веселых и пишущих о науке чудаков.

Два затронутых мною там вопроса имеют отношение и к обсуждаемой теме. Во-первых, нам следует избегать чрезмерной точности и принимать в расчет интервалы доверия и стандартные отклонения. Если у вас имеются данные для интервалов доверия, то вы обладаете лучшей картиной, чем вам может дать статистическая значимость Р-показателя (p value). Кроме того, это намного более интересно и намного более интуитивно.

Во-вторых, важно не принимать в расчет информацию из одного исследования изолированно, и этой темой я более обстоятельно занималась в другой своей публикации. Одно исследование само по себе недостаточно для того, чтобы получить искомый ответ.

И это приводит нас, наконец, к Томасу Байесу (Thomas Bayes), математику и священнику 18-го века, чьи идеи являются исключительна важными для дебатов относительно вычисления и интерпретации вероятности. Байес полагал, что нам нужно учитывать наше предыдущее знание, когда мы рассматриваем вероятность, а не только принимать в расчет частотность специфического набора данных, находящихся перед нами, в сравнении с фиксированной, неизменной величиной вне зависимости от самого вопроса.

Более подробно познакомиться со статистическими взглядами Байеса можно в Википедии. Там приводится следующий пример: предположим, кто-то сказал вам, что они говорят с каким-то человеком. Вероятность того, что этим человеком может быть женщина, обычно составляет 50%. Но если эти люди скажут вам, что они говорят с человеком с длинными волосами, то тогда полученное знание может увеличить вероятность того, что этот человек является женщиной. И вы получаете возможность вычислить новую вероятность, основанную на имеющемся знании.

Статистиков часто подразделяют на сторонников Байеса и приверженцев частотного подхода (frequentists).

Если строго придерживаться значения  $p < 0.05$  (или 0.001), независимо от чего бы то ни было, то это будет классический частотный подход. Важная причина для озабоченности по поводу данного подхода состоит в недостаточности большей части наших предварительных знаний, а также в осознании того, что люди могут быть очень сильно предубеждены и способны безответственно обращаться с данными, если они прочно не зафиксированы на своем месте, как стойки ворот.

Байесианство испытывало несколько раз взлеты и падения, однако возрастающая статистическая изоциренность и мощь компьютеров позволяют ему выдвигаться на передовые позиции в 21-м веке. Но далеко не все находятся в одном или в другом лагере – существует также много вариантов «слияния».

Недавно Вэлен Джонсон (Valen Johnson) в Сборнике Национальной академии наук США (Proceedings of the National Academy of Sciences in the USA) подчеркнул, что методы Байеса для вычисления статистической значимости дошли в своем развитии до такой точки, что они уже готовы воздействовать на практику. Смысл состоит в том, что порог

статистической значимости, по мнению Джонсона, следует установить значительно ниже – ближе к 0,005, чем 0,05. Какой ужас. И последствия этого для размеров образцов, необходимых для проведения исследования, будут весьма значительными.

Однако не все зависит от порога Р-значения. Статистически значимые результаты могут оказаться важными или не важными по целому ряду причин. Одно из эмпирических правил состоит в следующем: если результаты не достигают подобных числовых значений, исследуемые данные все равно демонстрируют определенные показатели, но они всегда должны быть помещены внутрь более широкого контекста. Такие факторы, например, как величина явного результата, а также наличие или отсутствие интервала доверия свидетельствуют о том, является ли полученная оценка слишком общим планом или нет.

Однако споры по поводу уровня статистической значимости не означают, что отсутствие статистической значимости не является важным фактором. Данные, не достигающие до уровня статистической значимости, слишком слабы для того, чтобы войти в наши заключения. Но и наличие статистической значимости не означает, что нечто обязательно является «истинным», и отсутствие достаточного количества доказательств не обязательно означает, что нечто является «ложным».

Споры, в которых сторонники Байеса выступают против приверженцев частотного подхода и тестирования гипотез, являются живым напоминанием о том, что статистическая область является динамичной – как и другие области науки. Не каждый статистик будет одинаковым образом оценивать имеющиеся данные. Между собой будут соревноваться различные теории и практики, а знание будет при этом развиваться. Есть много способов изучения данных и интерпретации их значений, и нет особого смысла в том, чтобы смотреть на данные сквозь призму только одного показателя. Р-величина не является тем значением, которое управляет всеми теориями и практиками.

Оригинал публикации **[Statistical significance and its part in science downfalls](https://ria.ru/20131113/976570804.html)**

[https://ria.ru/20131113/976570804.html?](https://ria.ru/20131113/976570804.html)