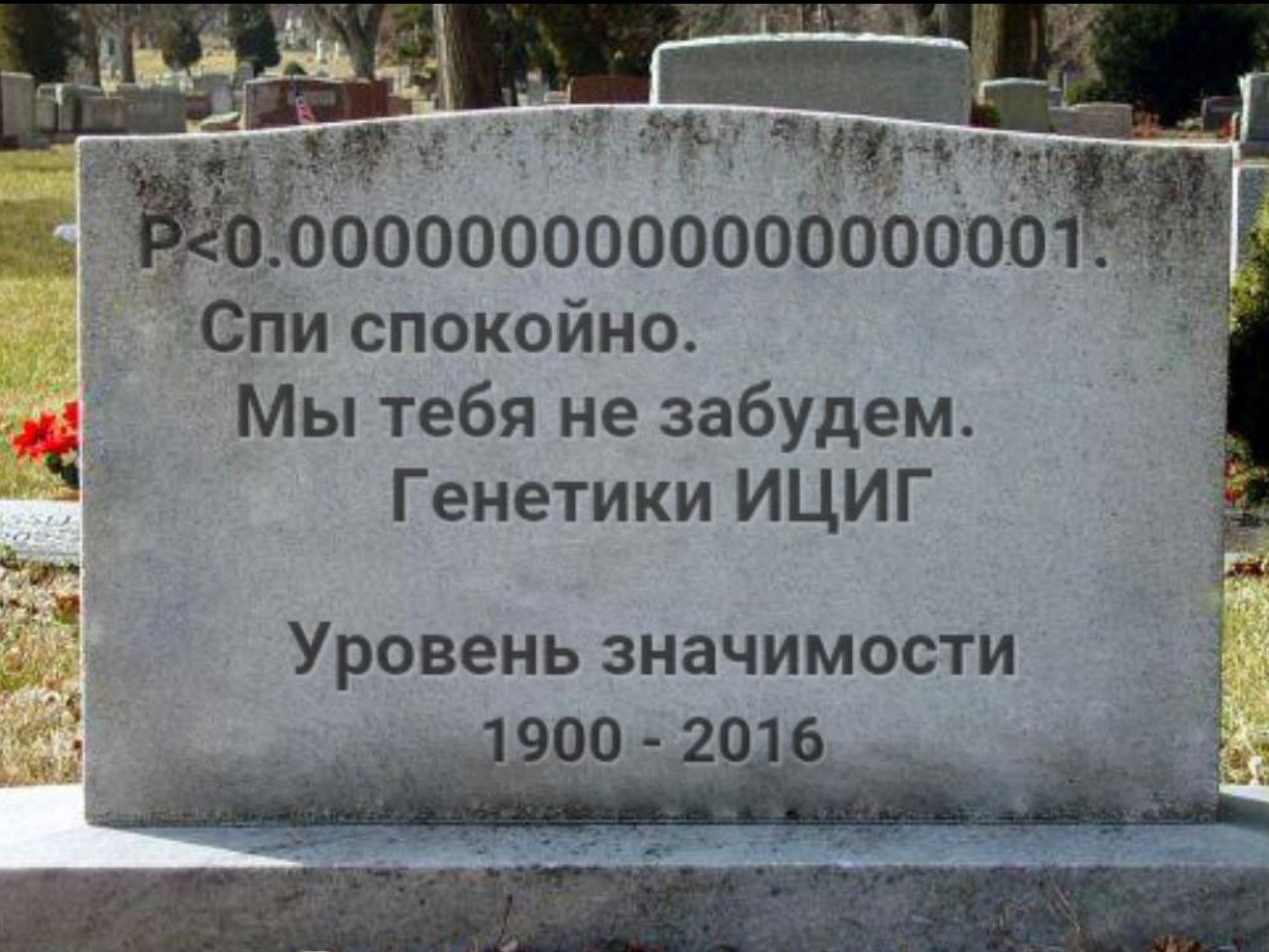

P<0.00000000000000000001.
Спи спокойно.
Мы тебя не забудем.
Генетики ИЦИГ
Уровень значимости
1900 - 2016

NATURE | RESEARCH HIGHLIGHTS: SOCIAL SELECTION

Psychology journal bans *P* values

Test for reliability of results 'too easy to pass', say editors.

Chris Woolston

26 February 2015 | Clarified: 09 March 2015

PDF | Rights & Permissions

A controversial statistical test has finally met its end, at least in one journal. Earlier this month, the editors of *Basic and Applied Social Psychology* (*BASP*) announced that the journal would no longer publish papers containing *P* values because the statistics were too

ACCEPTED MANUSCRIPT

ASA Statement on Statistical Significance and P-values

February 5, 2016

Edited by Ronald L. Wasserstein, Executive Director

On behalf of the American Statistical Association Board of Directors

American Statistical Association (ASA) – крупнейшая в мире ассоциация профессиональных статистиков

Introduction

Increased quantification

recent years have expan

avenues for scientific pr

research data. The valid

more than the statistical

Taylor & Francis

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://amstat.tandfonline.com/loi/utas20

The ASA's statement on p-values: context, process, and purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician, DOI: 10.1080/00031305.2016.1154108

To link to this article: http://dx.doi.org/10.1080/00031305.2016.1154108

«*Неправильное понимание и использование статистического вывода – лишь одна из причин "кризиса воспроизводимости результатов", но, по мнению нашего сообщества, это важная причина*»

«*Хотя значение P может быть полезной статистической мерой, его часто неправильно используют и интерпретируют*»

Из заявления ASA

Научное сообщество:

Наши статистически значимые результаты не воспроизводятся

Профессиональные статистики:

Значения *P* не измеряют вероятность того, что гипотеза истинна,

следовательно,

они не измеряют потенциальную воспроизводимость результатов

> 30 экспертов участвовали в обсуждении

«*Некоторые из них имеют взгляды частично или полностью противоположные тому, что изложено в заявлении*»

Консенсус:

«*ASA должна высказаться по этим вопросам*»

Принципы

1. Значения P способны указывать, насколько данные несовместимы с определенной статистической моделью

2. Значения P не измеряют вероятность того, что исследуемая гипотеза истинна, или вероятность того, что данные являются следствием только случайных причин.

3. Научные заключения и деловые или политические решения не должны основываться только на том, проходит ли значение P определенный порог.

4. Правильное заключение требует полной отчетности и прозрачности.

5. Значение P, или статистическая значимость, не измеряет размер эффекта или важность результата.

6. Само по себе значение P не обеспечивает хорошую меру доказательства в отношении модели или гипотезы

Что меняется?

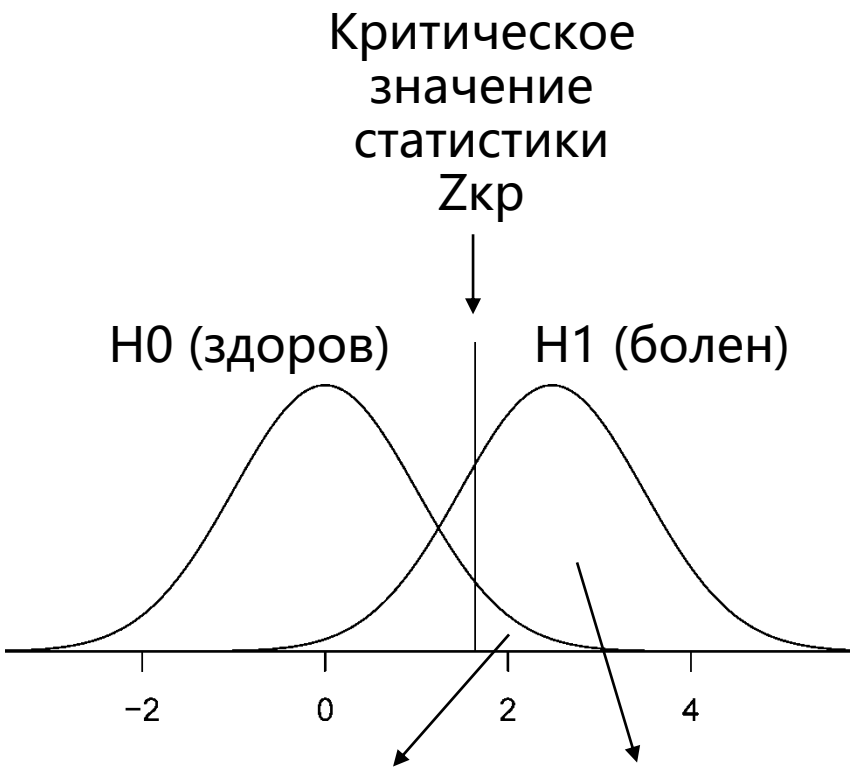
Вопрос о неправильной интерпретации значений P из маргинального становится основным на какое-то время

Значения P больше не будут единственной и высоколиквидной валютой на рынке научных публикаций

Почему значения P часто интерпретируются неправильно?

В чем проблема интерпретации?

Если бы врач тестировал больных так, как тестируются статистические гипотезы?



$\alpha = Pr(Z > Zкр \mid H0)$ (уровень значимости, ошибка 1 рода)

$\beta = Pr(Z > Zкр \mid H1)$ (мощность)

Хороший тест:
$\beta = 0.8$ при $\alpha = 0.05$

Результат анализа:
$P = 0.047 < \alpha = 0.05$
Нулевая гипотеза отвергается

Интерпретация:
«Если бы пациент был здоров, то вероятность получить такие результаты анализа была бы < 5%»

Болен или здоров пациент на самом деле?

Знаем: $\Pr(Z > Z_{кр} \mid \text{здоров}) = \alpha = 0.05$ $= D / G$

Хотим знать: $\Pr(\text{здоров} \mid Z > Z_{кр}) = ?$ $= D / F$

на самом деле ->	H0 (здоров)	H1 (болен)	Всего
формально ↓			
H0 не отвергается (здоров)	A	B	C
H0 отвергается (болен)	D	E	F
Всего	G	H	I

$$\Pr(\text{здоров} \mid Z > Z_{кр}) = \frac{\alpha \Pr(\text{здоров})}{\alpha \Pr(\text{здоров}) + \beta (1 - \Pr(\text{здоров}))} = Q \text{ value}$$

Pr (здоров) – это априорная (до опыта) вероятность гипотезы (G / I)

Для врача Pr (здоров) ~= 1 – (частота заболевания в популяции)

частота	Q при α = 0.05	
0.0001	0.998	наверняка здоров
0.01	0.861	скорее здоров
0.1	0.360	скорее болен
0.5	0.059	вероятно, болен

Что если исследователь не знает Pr(H0)?

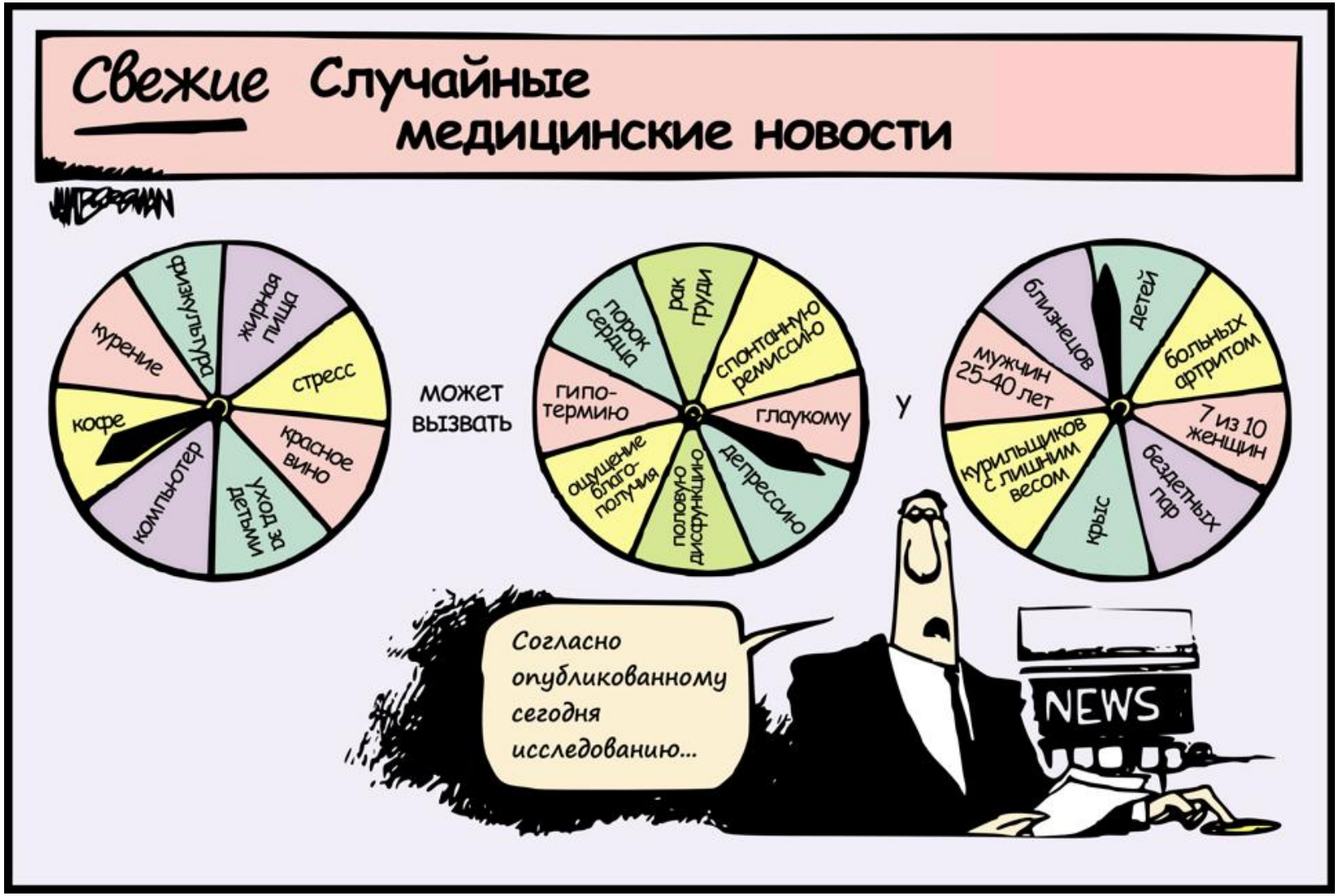
$$Pr\,(H0 \mid Z > Zкр\,) = \frac{\alpha\, Pr\,(H0)}{\alpha\, Pr\,(H0) + \beta\,(1 - Pr\,(H0)\,)}$$

$$Pr\,(H0 \mid Z > Zкр\,) = \frac{0.05 * \text{не знаю}}{0.05 * \text{не знаю} + 0.8 * \text{не знаю}} = \text{не знаю}$$

Но если действительно ничего не известно, то мы можем предположить, что H0 и H1 априори равновероятны:

$$Pr\,(H0 \mid Z > Zкр\,) = \frac{0.05 * 0.5}{0.05 * 0.5 + 0.8 * 0.5} = Q?$$

α	Q при Pr (H0) = 0.5 и β = 0.8
0.05	0.059
0.01	0.012
0.001	0.0012
0.0001	0.00012



Согласно статистике британских ученых, 99% британских ученых одобряют статистику британских ученых (P < 0.000...)

Какова вероятность встретить на улице динозавра?

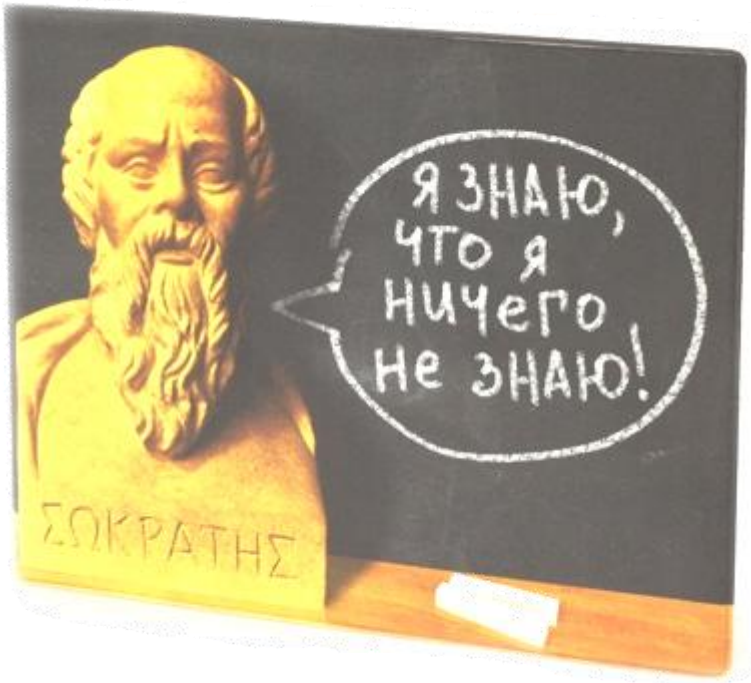
50%
либо встречу,
либо не встречу

Оценка
априорной
вероятности
события

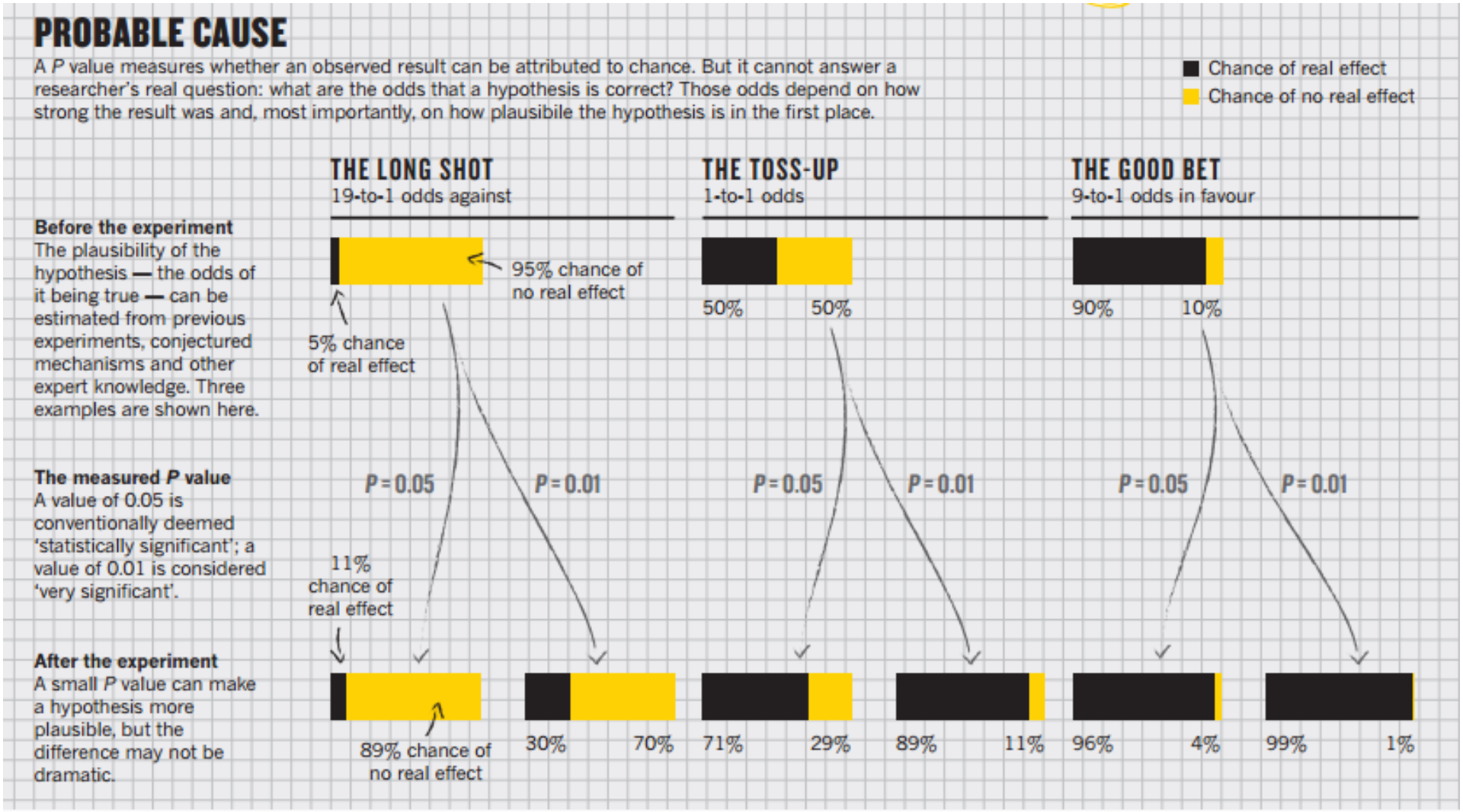
Не знаю

Британские
ученые
доказали

Оценка надежности
вывода



Nature, “Scientific method: statistical errors” (Nuzzo, 2014)



BF min = exp(-Z^2/2)

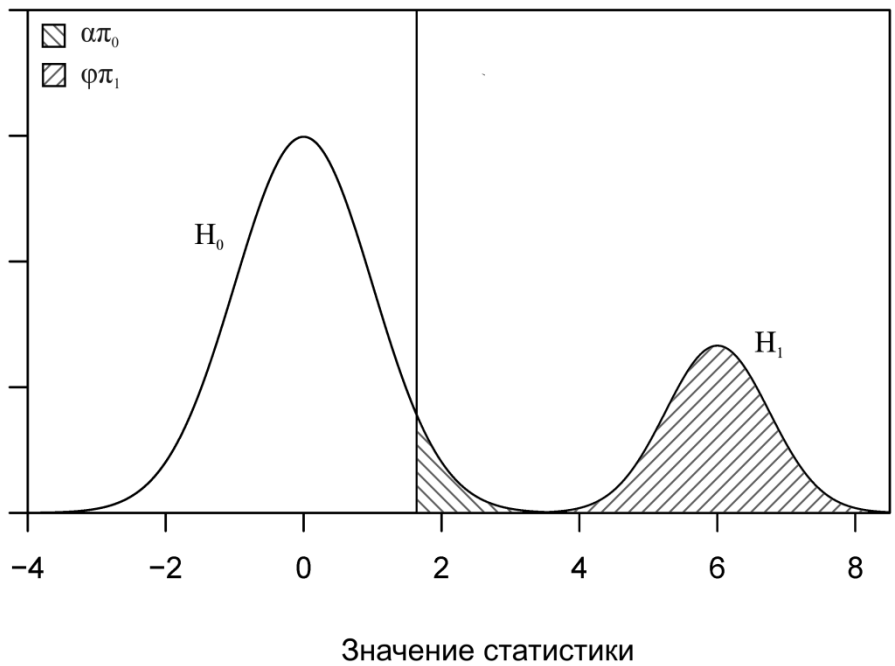
$$\Pr(H0 \mid Z) = \frac{BF\ min * \Pr(H0)}{BF\ min * \Pr(H0) + (1 - \Pr(H0))}$$

Частотный вывод:

$$\Pr(H0 \mid Z > Z_{кр}) = \frac{\alpha \Pr(H0)}{\alpha \Pr(H0) + \beta (1 - \Pr(H0))}$$

Простейший байесовский вывод:

$$\Pr(H0 \mid Z) = \frac{BF\ min * \Pr(H0)}{BF\ min * \Pr(H0) + (1 - \Pr(H0))}$$



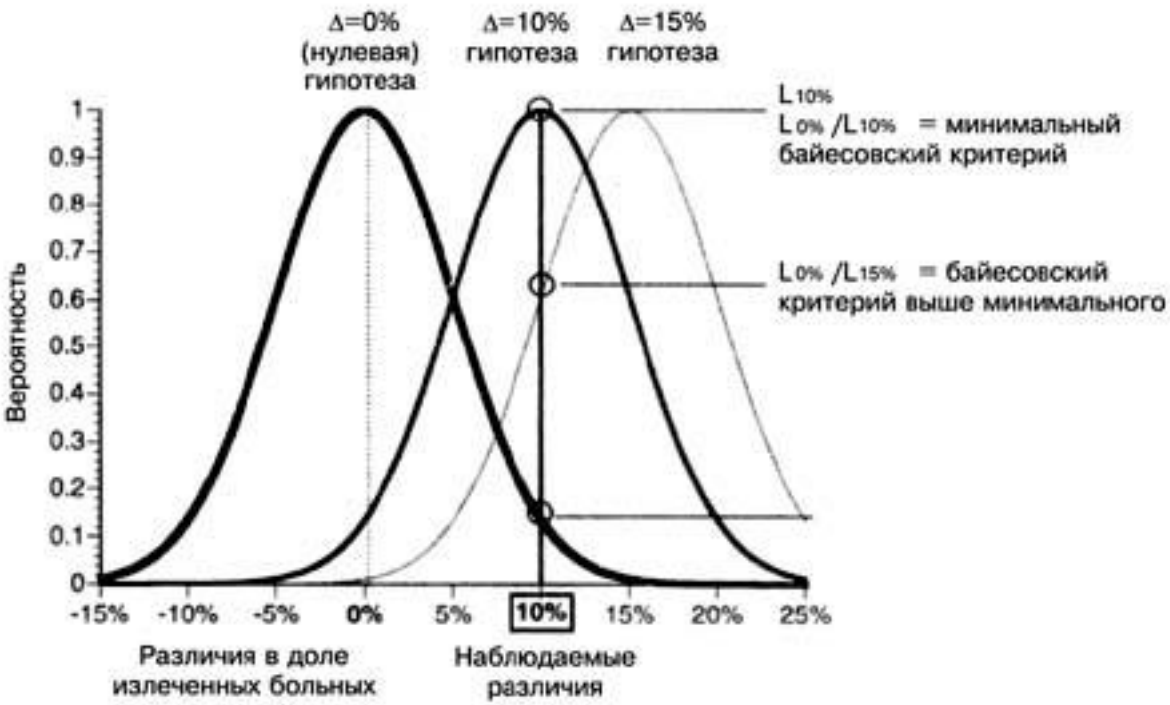
$\Pr(H_1) = 0.25$

$\Pr(H_1 \mid p \leq 0.05) =$
$1 - Q\ (\alpha = 0.05) = 0.87$

$\Pr(H_1 \mid p = 0.047) = 1 \times 10^{-7}$

Байесовская оценка в точке может быть более адекватна, чем частотная

Почему «минимальный» BF?



http://pubhealth.spb.ru/SASDIST/ebmstat2.htm

«К научно-доказательной медицинской статистике, часть 2: байесовский фактор»
Steven N. Goodman, MD, PhD
Ann Intern Med. 1999;130:1005-1013

Есть еще часть 1:
http://pubhealth.spb.ru/SASDIST/ebmstat1.htm
«Заблуждение P-оценки»

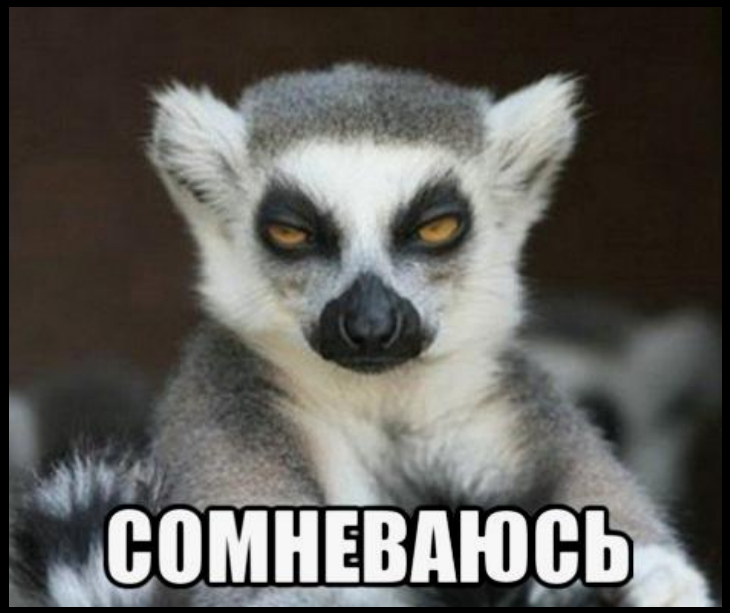
В частотной парадигме
H1: Δ ≠ 0
означает по сути:
H1: Δ = Δ-оцененное
т.е. только тестирование простых гипотез.
Отсюда:
- завышение эффекта статистически значимых результатов
- сложности с метаанализом
- шум от априори невероятных эффектов

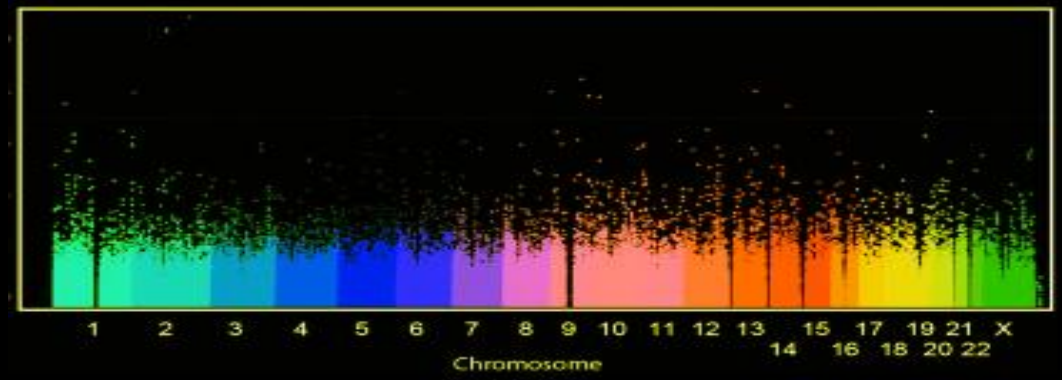
В байесовской парадигме возможно настоящее тестирование сложных гипотез H1: Δ ≠ 0

(Дальнейший шаг: оценивание апостериорной плотности вероятности на пространстве эффектов **вообще без тестирования гипотез**)

GWAS

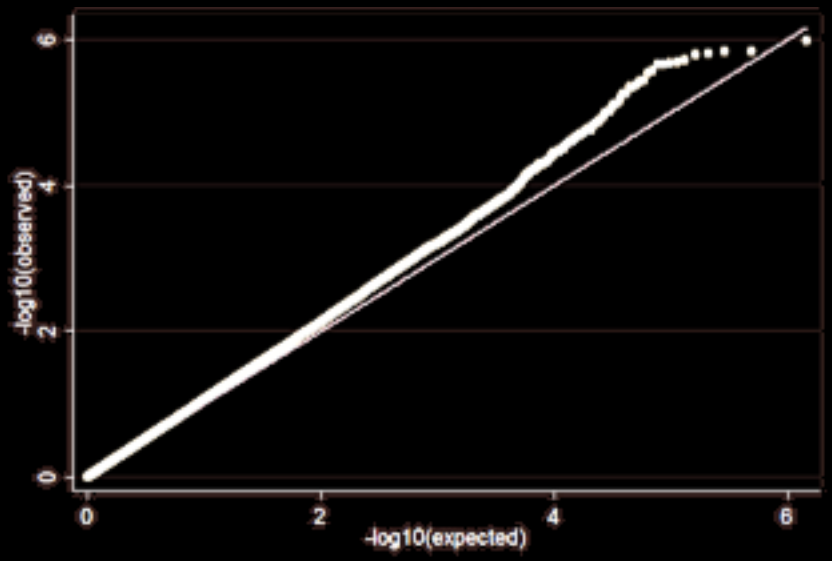
Случайно выбранный SNP в геноме ассоциирован с признаком с априорной вероятностью 50%? При тех размерах эффектов, которые детектируются с высокой мощностью?





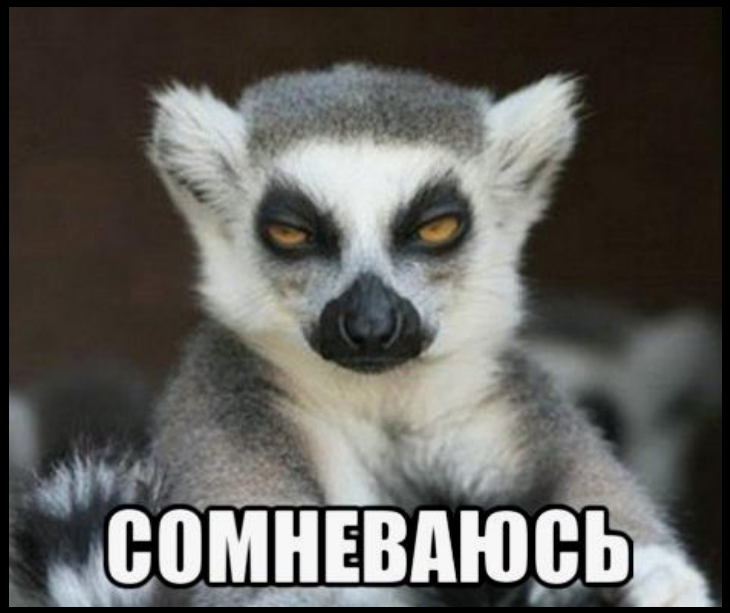
Распределение статистик GWAS ≈
≈ распределению при H0

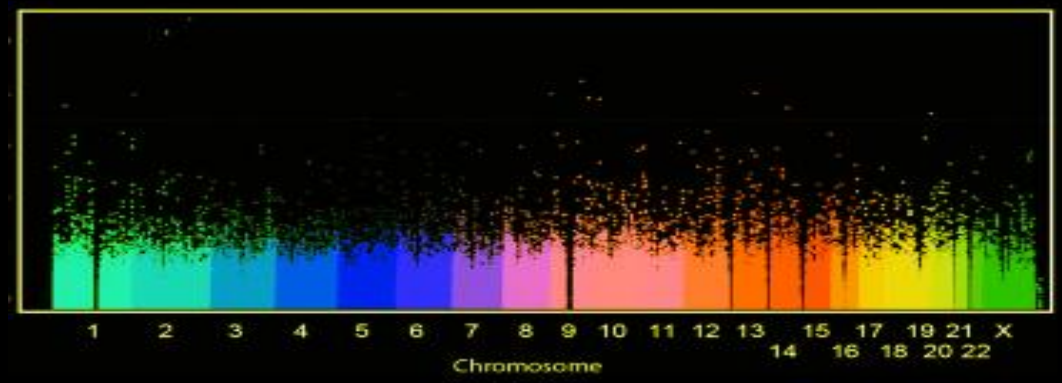
Иначе считается, что что-то не в порядке (геномный контроль)



GWAS

Случайно выбранный SNP в геноме ассоциирован с признаком с априорной вероятностью 50%? При тех размерах эффектов, которые детектируются с высокой мощностью?





При полногеномном уровне значимости ($\alpha = 5 \times 10^{-8}$):

Pr (H1)	Q (β = 0.2)	Pr (H0 | Z)
0.0001	0.0025	0.0068
10^{-5}	0.024	0.065
10^{-6}	0.20	0.41

Распределение статистик GWAS ≈
≈ распределению при H0

Иначе считается, что что-то не в порядке (геномный контроль)

Множественное тестирование

$$m\,\alpha = m\,\alpha$$

$$\frac{m\,\alpha\,Pr\,(H0)}{m\,\alpha\,Pr\,(H0) + m\,\beta\,(1 - Pr\,(H0))} = \frac{\alpha\,Pr\,(H0)}{\alpha\,Pr\,(H0) + \beta\,(1 - Pr\,(H0))}$$

Надежность и воспроизводимость результата
не зависит
от числа тестов

Болен или здоров пациент на самом деле, если теста два?

P1 = 0.047
H0 (здоров) отвергается при α = 0.05

P2 = 0.031 < α = 0.05

но

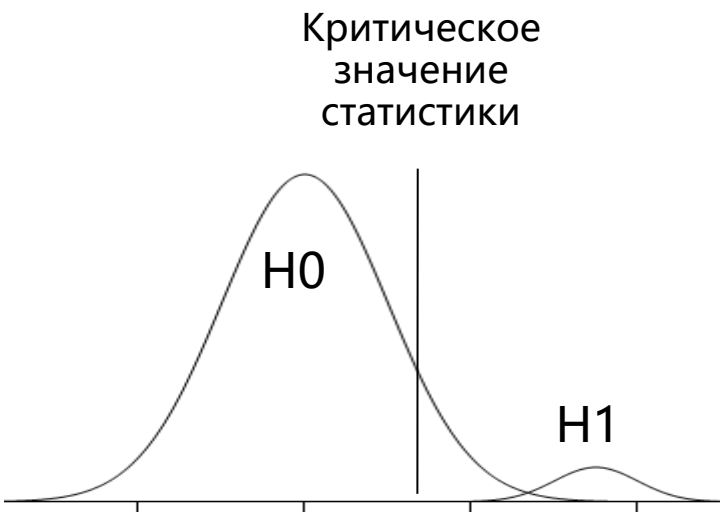
Т.к. теста два, то новый уровень значимости α2 = 0.05 / 2 = 0.025
H0 (здоров) не отвергается

При этом значение
Q = Pr (здоров | P < α) не изменилось:

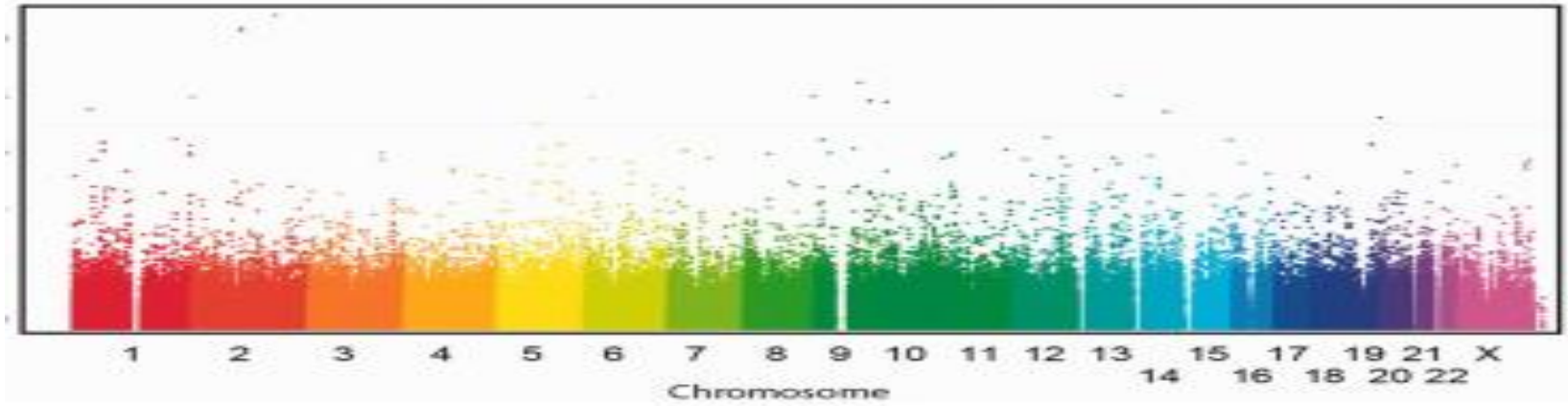
частота	Q1 = Q2, α = 0.05	
0.0001	0.998	наверняка здоров
0.01	0.861	скорее здоров
0.1	0.360	скорее болен
0.5	0.059	вероятно, болен

Поправка Бонферрони «на множественное тестирование» – способ неявного учета априорной вероятности Pr(H0)

Снижение α может косвенно приводить к снижению Q:

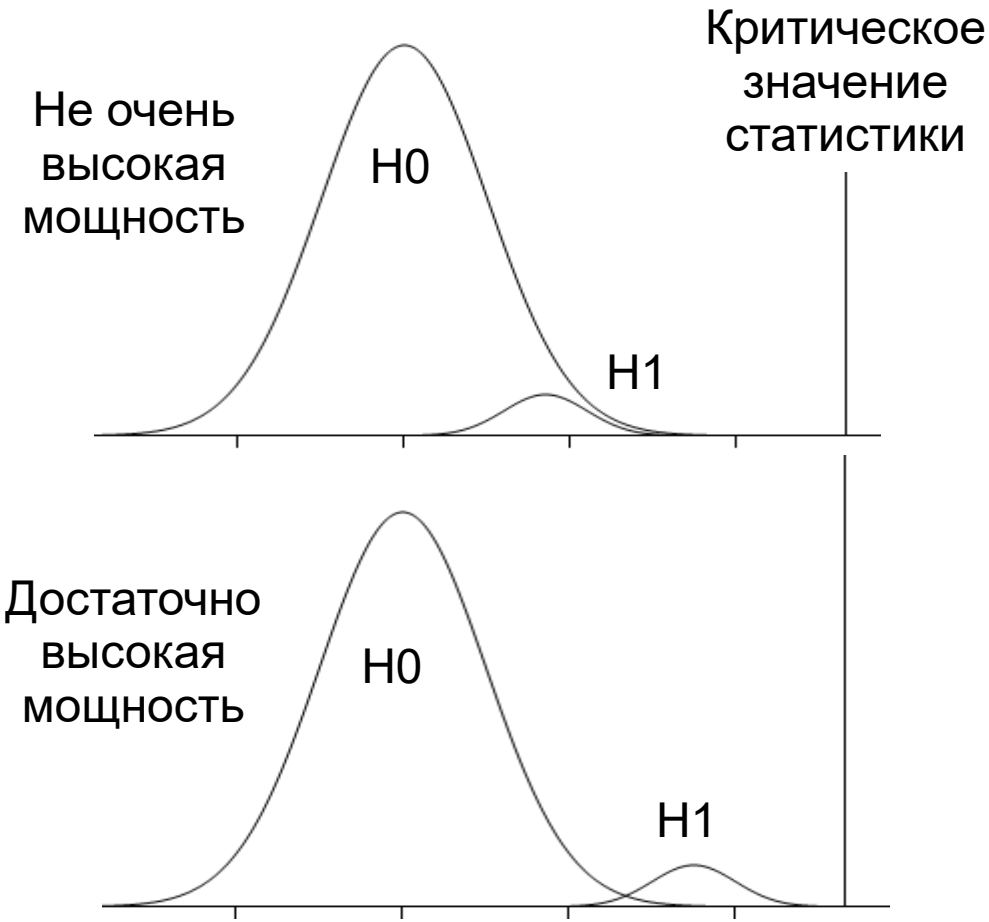


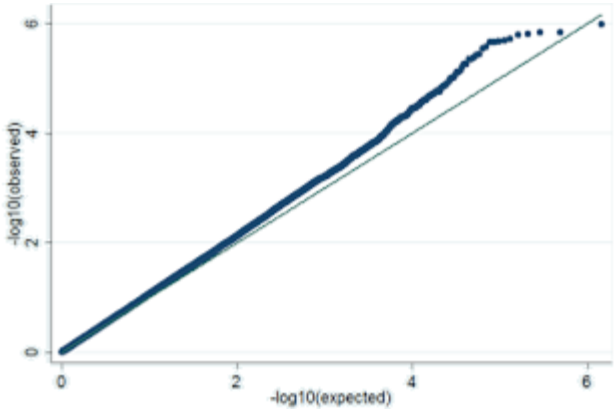
Но никто формально не мешает взять один понравившийся SNP в геноме, тестировать его при высоком уровне значимости и получить, скорее всего, ложноположительный результат



Поправка Бонферрони «на множественное тестирование» – способ неявного учета априорной вероятности Pr(H0)

Снижение α может косвенно приводить к снижению Q.
Но не всегда!





Если тестов очень много,
то мы никогда не преодолеем порог множественного тестирования, даже с мощным статистическим тестом

Исследователь всё поймет, если посмотрит на тестовое распределение статистики глазами.

Но посмотрит ли?

Нажми на кнопку – получишь результат

Нажимая на эту кнопку, вы автоматически соглашаетесь с условиями и допущениями, лежащими в основе процедуры тестирования статистических гипотез методом сравнения значения P с фиксированным уровнем значимости:

1) Оценка вероятности ошибки первого рода (тестовое значение P) не характеризует надежность или воспроизводимость результата и даже приблизительно не отражает их, если тестируемые гипотезы априори не равновероятны или если тест не имеет достаточной статистической мощности.
2) Контроль вероятности ошибки первого рода является основным и приоритетным критерием качества исследования при данном подходе.
3) В ситуации множественного тестирования контроль числа ошибок первого рода на фиксированном уровне повлечет повышение вероятности ошибки второго рода. В этом случае получить статистически значимый результат будет тем сложнее, чем больше статистических тестов вы собираетесь произвести.
4) Величина значения P не отражает величину эффекта и его реальную значимость.
5) Значение P характеризует ошибку первого рода в точке наиболее вероятной оценки эффекта, полученной на основе анализируемых данных. Оно не характеризует ошибку первого рода в точках, соответствующих другим вероятным значениям эффекта. Оно не характеризует ошибку первого рода для всего пространства ненулевых эффектов.
6) .....
7) .....

... .

" Статистические техники тестирования гипотез... содержат больше изъянов, чем политика конфиденциальности Facebook "

Siegfried, 2014, ScienceNews

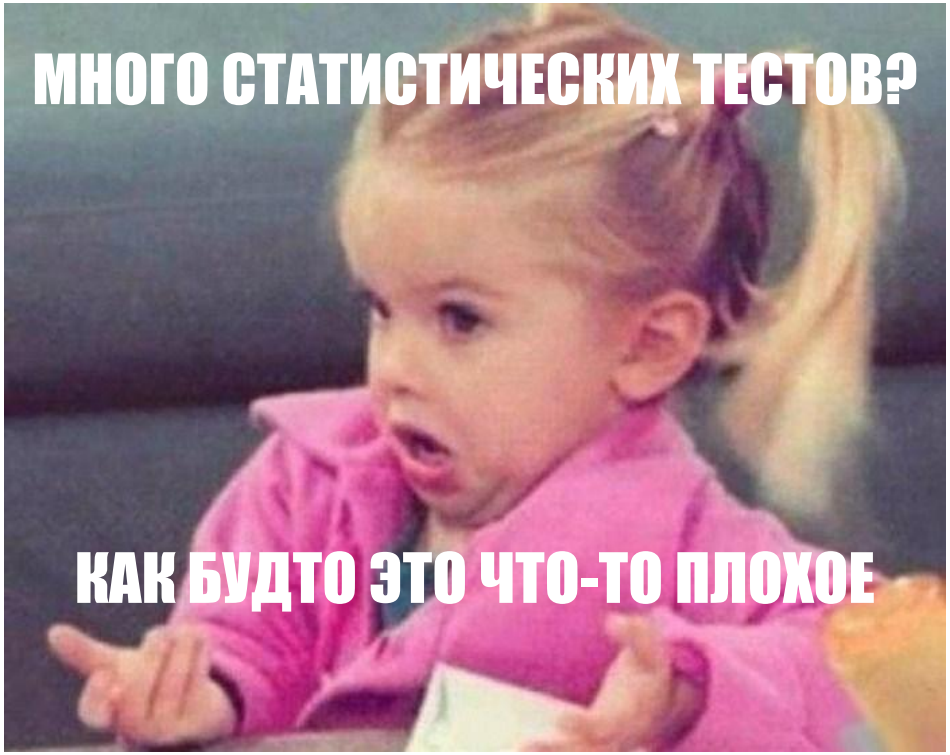
Множественное тестирование

Контроль mα за счет mβ

чем больше тестов, тем сложнее что-то доказать

Контроль вероятности ложноположительного результата (Q value, false discovery rate (FDR))

чем больше тестов, тем точнее оценка Pr(H0), тем надежнее результат



Правдоподобие данных при гипотезе

Априорное знание

Величина P

Если бы нулевая гипотеза была справедлива, то вероятность получить такое или большее отличие была бы не более, чем P

Неосознанно

Осознанно

Байесовский вывод: апостериорная вероятность гипотезы

Британские ученые доказали

Частотный подход: величина Q или FDR

Суждение о вероятности того, что тестируемая гипотеза истинна, результат надежен и воспроизводим

Оценка надежности:

$$P = 0.0000...0000001$$

$\longrightarrow$ 0.00001

$\longrightarrow$ 0.001

$\longrightarrow$ 0.01

«Обменный курс» апостериорной вероятности

Частотный вариант:

$$Pr(H0 \mid Z > Zкр) = \frac{\alpha\, Pr(H0)}{\alpha\, Pr(H0) + \beta\,(1 - Pr(H0))}$$

- может быть выполнен на основе значения P

Байесовский вариант:

$$Pr(H0 \mid Z) = \frac{BF * Pr(H0)}{BF * Pr(H0) + (1 - Pr(H0))}$$

BF min = exp(-Z^2/2)

- короче (без вычисления P)
- надежнее (оценка в точке, а не квантиль)
- в перспективе - больше возможностей