# Гармонизация статистических доказательств и предсказаний

Тишков Артем Валерьевич

Никита Николаевич Хромов-Борисов

Кафедра физики, математики и информатики
ПСПбГМУ им. акад. И.П. Павлова

# Обработка количественных данных

- Эпидемиологи смотрят на мир сквозь решетку таблицы 2×2. При этом надо помнить, что результат обследования является бинарным (дихотомическим): либо положительным, либо отрицательным.
- Для обработки количественных данных, измеряемых или подсчитываемых, используются также определенный набор статистических величин и внушительный арсенал доказательных и предсказательных статистических методов.

# Интерфероны и диагностика ЗВУР - задержки внутриутробного развития

Королева Людмила Илларионовна,
НИИ АГ им.Д.О.Отта

# ЗВУР

- **Термин задержка внутриутробного развития плода (ЗВУР) используется для описания плода, масса которого гораздо меньше ожидаемой для данного гестационного возраста.**
- **Согласно последним отечественным данным частота (распространенность) ЗВУР находится в пределах 3,5 – 8,5%.**
- **Плод с задержкой внутриутробного развития подвержен повышенному риску внутриутробной гибели или неонатальной смерти, асфиксии до или во время родов.**

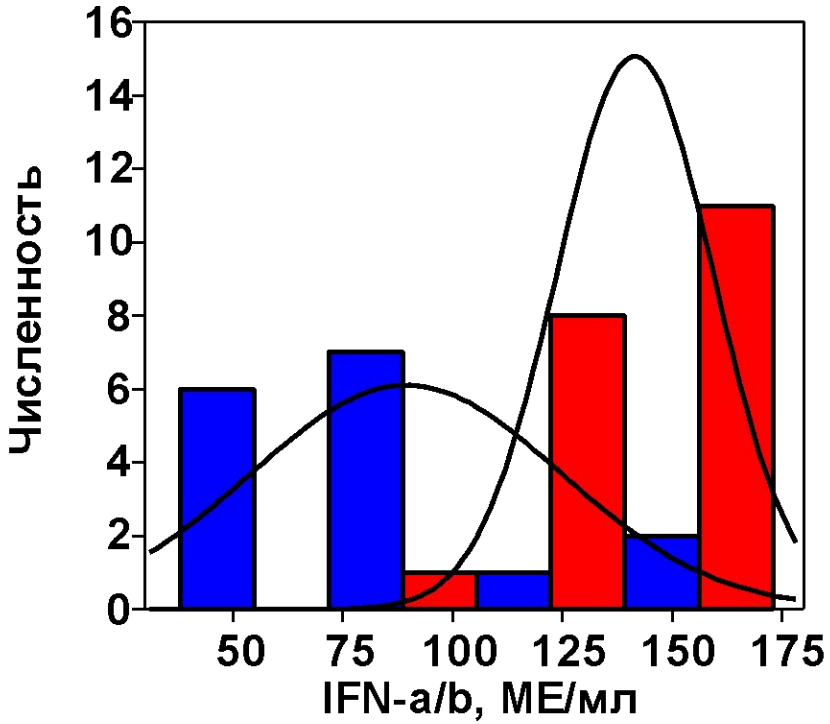
# Содержание INF-α/β у 16 здоровых матерей здоровых детей и у 20 матерей доношенных новорожденных с ЗВУР (Королева Л.И.)

| Здоровые | | | | ЗВУР | | | |
|---|---|---|---|---|---|---|---|
| № | IFN-α/β, МЕ/мл | № | IFN-α/β, МЕ/мл | № | IFN-α/β, МЕ/мл | № | IFN-α/β, МЕ/мл |
| 1 | 38 | 9 | 92 | 1 | 104 | 11 | 144 |
| 2 | 42 | 10 | 93 | 2 | 121 | 12 | 146 |
| 3 | 58 | 11 | 94 | 3 | 123 | 13 | 147 |
| 4 | 59 | 12 | 101 | 4 | 123 | 14 | 149 |
| 5 | 70 | 13 | 103 | 5 | 127 | 15 | 151 |
| 6 | 71 | 14 | 115 | 6 | 130 | 16 | 153 |
| 7 | 81 | 15 | 159 | 7 | 132 | 17 | 162 |
| 8 | 86 | 16 | 170 | 8 | 134 | 18 | 168 |
| | | | | 9 | 134 | 19 | 171 |
| | | | | 10 | 140 | 20 | 173 |

# Гистограмма

- **Гистограмма**
- **(от др.-греч. ἱστός — столб + γράμμα — черта, буква, написание)**
- **— столбиковая диаграмма**
- **— способ графического представления табличных данных.**

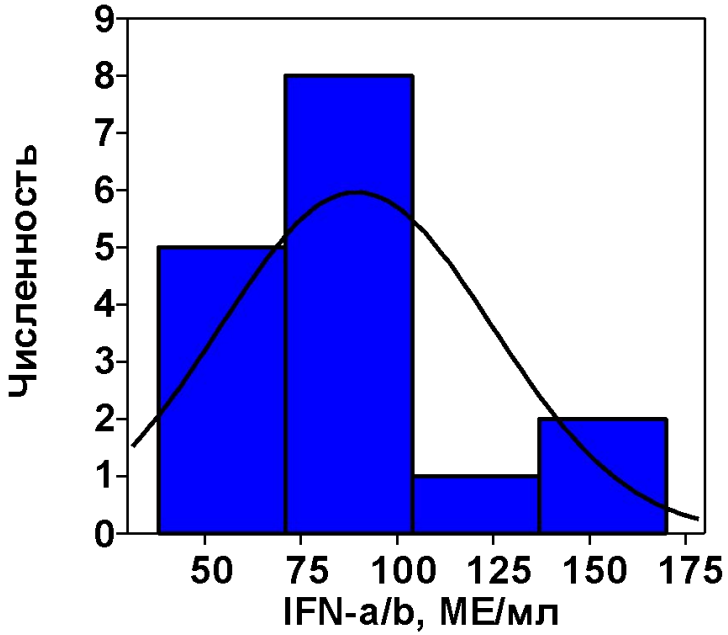
# Сопоставление гистограмм содержания INF-α/β у здоровых матерей здоровых детей и матерей доношенных новорожденных с ЗВУР



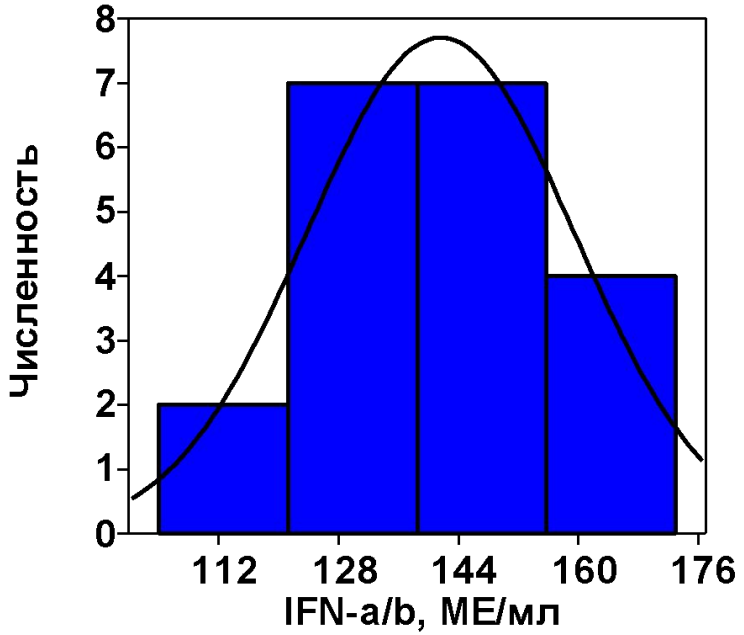
# Гистограммы содержания INF-α/β у здоровых матерей здоровых детей и матерей доношенных новорожденных с ЗВУР.

## Программа PAST (URL: http://folk.uio.no/ohammer/past/)

**Здоровые**

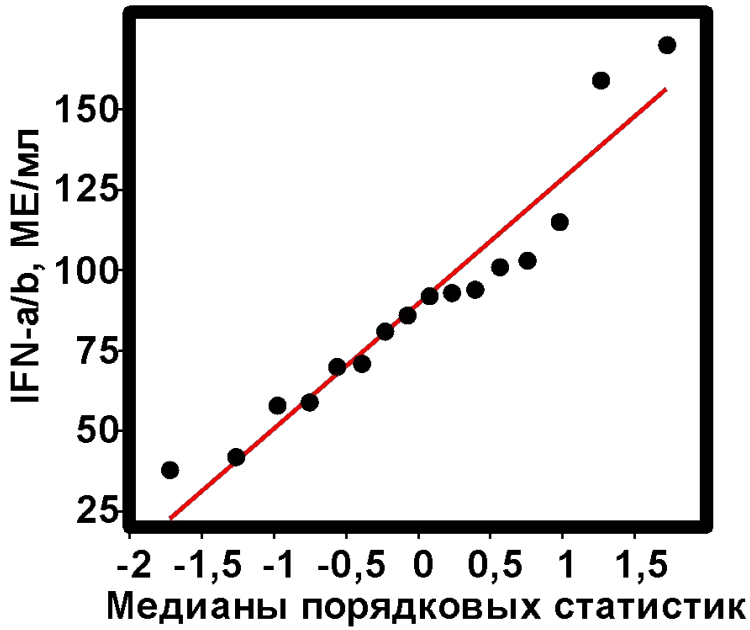



**ЗВУР**

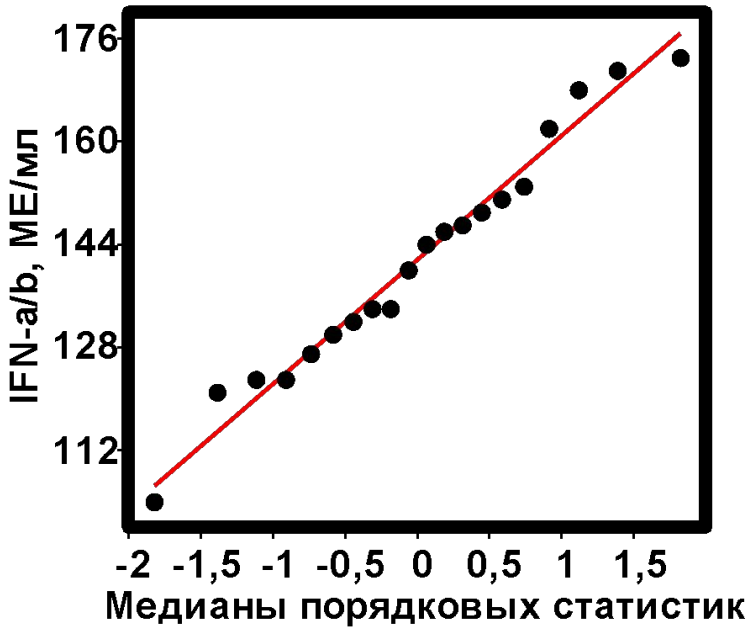


# Нормальные вероятностные графики

**Здоровые**




**ЗВУР**



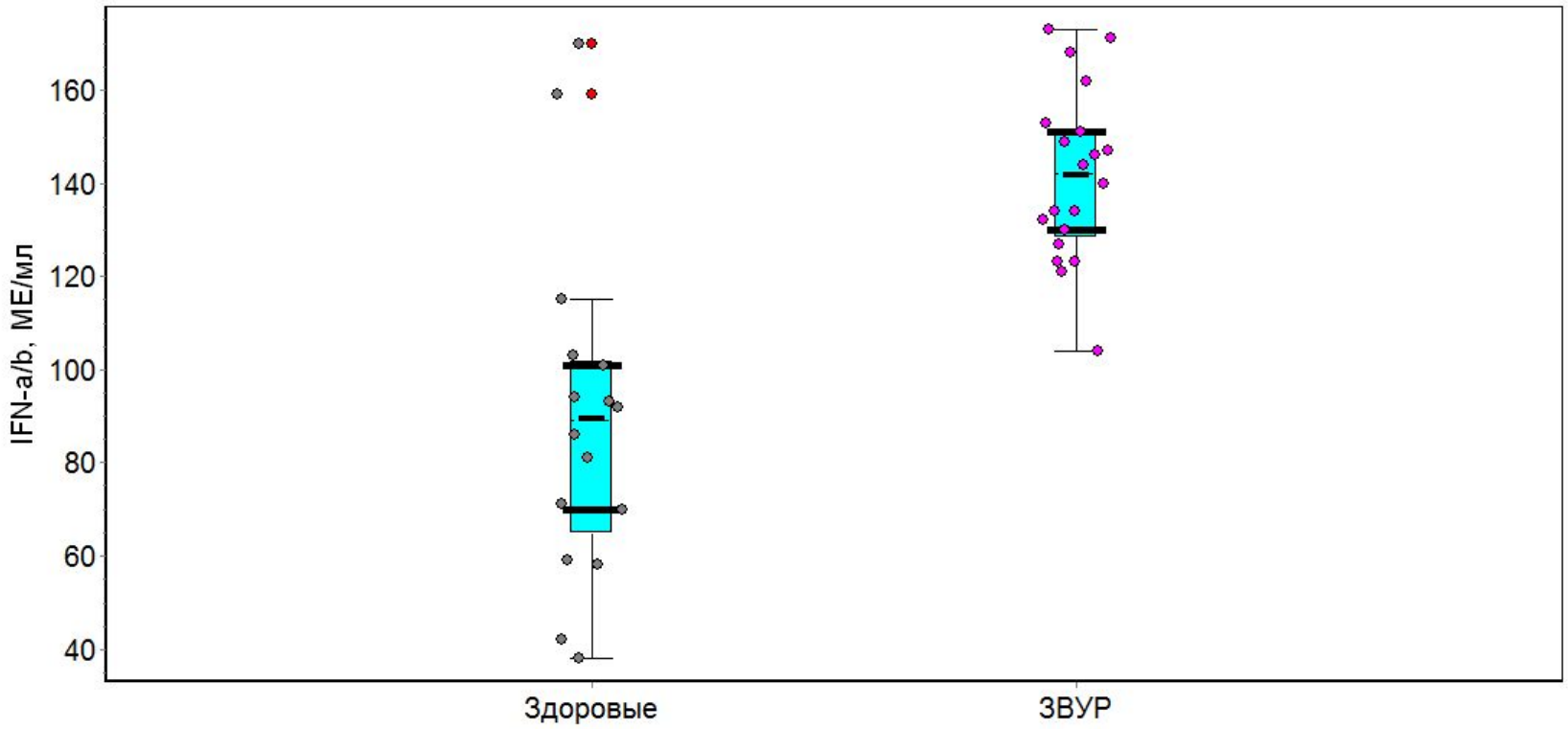
# Проверка нормальности (гауссовости) распределения у матерей здоровых детей и детей с ЗВУР

| Статистический критерий | Наблюдаемые *P*-значения, $P_{val}$ | |
|---|---|---|
| | Здоровые | ЗВУР |
| Андерсона-Дарлинга | 0,25 | 0,15 |
| Шапиро-Уилка | 0,19 | 0,21 |
| Коэффициента асимметрии | 0,059 | 0,46 |
| Коэффициент эксцесса | 0,23 | 0,34 |
| Жарка-Бера | 0,42 | 0,14 |
| Гири | 0,17 | 0,26 |
| Д'Агостино | 0,068 | 0,45 |
| Эппса-Палли | 0,17 | 0,048 |

**Практические все *P*-значения превышают пороговое значение 0,05 или почти равны ему. Следовательно у нас нет оснований сомневаться в гипотезе о нормальности распределения, порождающего наблюдаемые данные.**

# Диаграммы «короб с усами» для данных об уровне индуцированной продукции IFN-α/β у здоровых матерей здоровых детей и у матерей доношенных новорожденных с ЗВУР.

**Программа Instat+ (URL: http://www.reading.ac.uk/ssc/n/n_instat.htm)**



# Исключение резко выделяющихся наблюдений

- С рекомендаций по отбрасыванию выскакивающих (экстремальных) наблюдений («выбросов», «засорений») начинаются многие руководства по прикладной статистике.
- Очень часто авторы и (или) пользователи забывают, что большинство таких процедур предназначено для отбрасывания одного и только одного такого значения.
- Тем не менее, можно найти тексты, в которых, скажем, из 6-и наблюдений отбрасываются три.
- Это совершенно недопустимо.

# Резко выделяющиеся значения – «выбросы»

- **Выскакивающие значения можно и нужно выявлять.**
- **Но отбрасывать их следует на основе внестатистических соображений.**
- **Например, если записано значение для артериального давления 1100, то очевидно, что здесь опечатка: лишняя 1 или лишний 0.**

# Сжатие (свертка, редукция) статистических данных

- **Статистика – любая функция от случайных величин, порождающих получаемые статистические данные.**
- **Простейший пример - выборочное среднее:**

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i$$

# Основная логика статистического оценивания: интервальные оценки

- **Понятно, что если мы многократно повторим эксперимент, то вычисленные средние значения неизбежно будут варьировать.**
- **Поэтому задача математиков – вывести математический закон (вероятностное распределение), которому подчиняется варьирование этих выборочных средних.**
- **Если такой закон найден, то тогда можно построить доверительные интервалы (ДИ) для оцениваемого среднего с заданной доверительной вероятностью (1 – $\alpha$).**

# Статистические гипотезы

- **В обычном языке слово «гипотеза» означает предположение.**
- **В том же смысле оно употребляется и в научном языке для предположений, которые подлежат экспериментальной проверке, в ходе которой гипотеза либо подтверждается, либо опровергается.**
- **В математической статистике, термин «гипотеза» означает предположение о тех или иных свойствах распределений, которые служат моделями для получаемых данных.**
- **Проверка статистической гипотезы состоит в выяснении того, насколько совместима эта гипотеза с имеющимися данными.**

# Проверяемая гипотеза

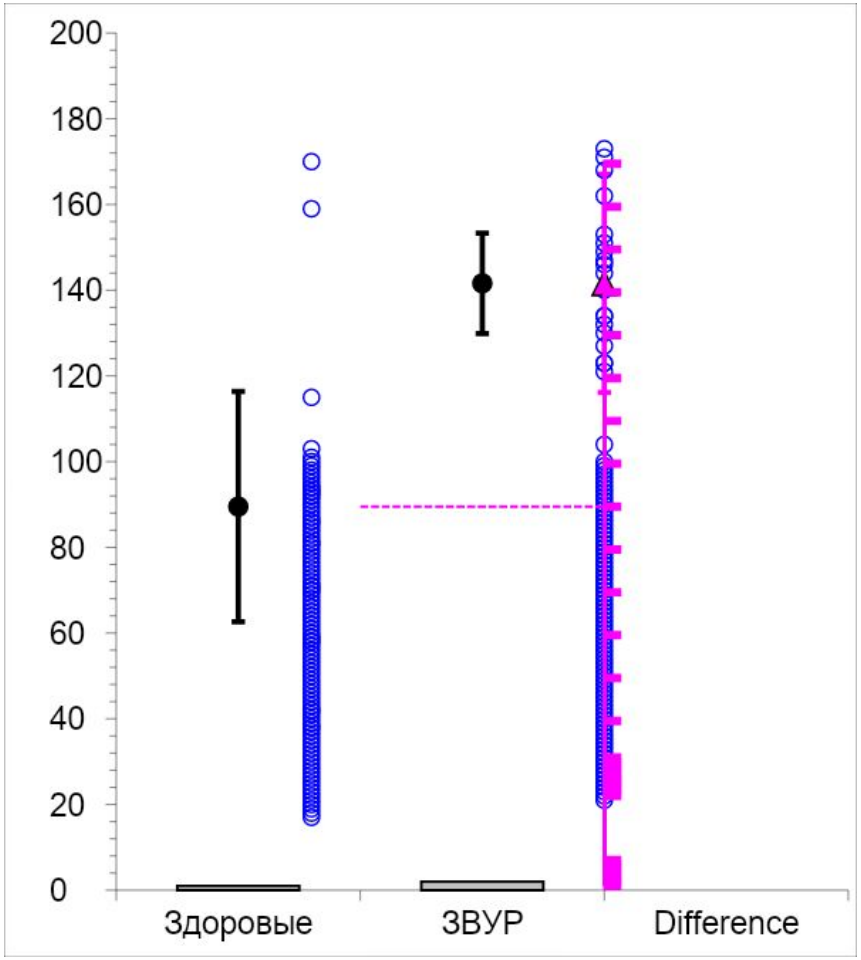
- **В подавляющем большинстве реальных ситуаций проверяемая статистическая гипотеза является гипотезой об отсутствии того или иного эффекта:**
- **об отсутствии различий, например, о равенстве средних, т.е. о равенстве нулю разности средних;**
- **об отсутствии связей, соответствий, зависимостей и т.п.**
- **Поэтому проверяемую гипотезу принято назвать нулевой и обозначать символом $H_0$.**

# Использование доверительных интервалов (ДИ) для проверки нулевых гипотез

- Например, для проверки нулевой гипотезы о равенстве двух средних:
  - $H_0: M_1 - M_2 = 0$
- можно построить ДИ для разности средних.
- Тогда, если вычисленный $100(1 - \alpha)\%$-й ДИ не накрывает постулируемое этой гипотезой значение 0, то отклонение оцениваемой разности от 0 можно признать статистически значимым на заранее выбранном уровне значимости $\alpha$.

# Визуализация результатов проверки статистических гипотез с помощью доверительных интервалов для размера эффекта

# Графическое представление результатов статистического сравнения групп матерей здоровых детей и детей с ЗВУР, 1-α = 0,99. Программа ESCI JSMS.xls http://www.latrobe.edu.au/psy/esci/




- 99%-й ДИ для разности средних не накрывает значение 0.
- Следовательно оцениваемое этим интервалом неизвестное нам значение разности средних статистически значимо отличается от 0 на уровне значимости 0,01.
- Соответственно мы можем взять на себя смелость отклонить нулевую гипотезу о равенстве средних и принять альтернативную.

# Статистики критериев (тестовые статистики)

- **Тестовая статистика – статистика, используемая для проверки конкретной статистической гипотезы.**
- **Пример: статистика *t*-критерия Стьюдента**

$$\tilde{t} = \frac{\tilde{M}_1 - \tilde{M}_2}{\tilde{S}_{(M_1 - M_2)}}, \; df = n_1 + n_2 - 2$$

- **В этом случае проверка гипотезы $H_0$ о равенстве двух средних: $H_0$: $M_1 - M_2 = 0$ сводится к проверке гипотезы о том, что $t = 0$.**
- **Когда эта нулевая гипотеза верна, то распределение этой статистики известно – это *t*-распределение Стьюдента с параметром (числом степеней свободы), равным *df*.**

# Проблема Беренса-Фишера

- **Если дисперсии сравниваемых двух независимых случайных величин не равны, то, то следует использовать модификацию *t*-критерия Стьюдента, которая называется критерием Уэлча:**
- $$t_W = \frac{\Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- **Статистика Уэлча приближенно имеет *t*-распределение Стьюдента, но со степенью свободы $\nu_w$, который задается выражением:**

$$\frac{1}{\nu_W} = \frac{C^2}{n_1} + \frac{(1-C)^2}{n_2}$$

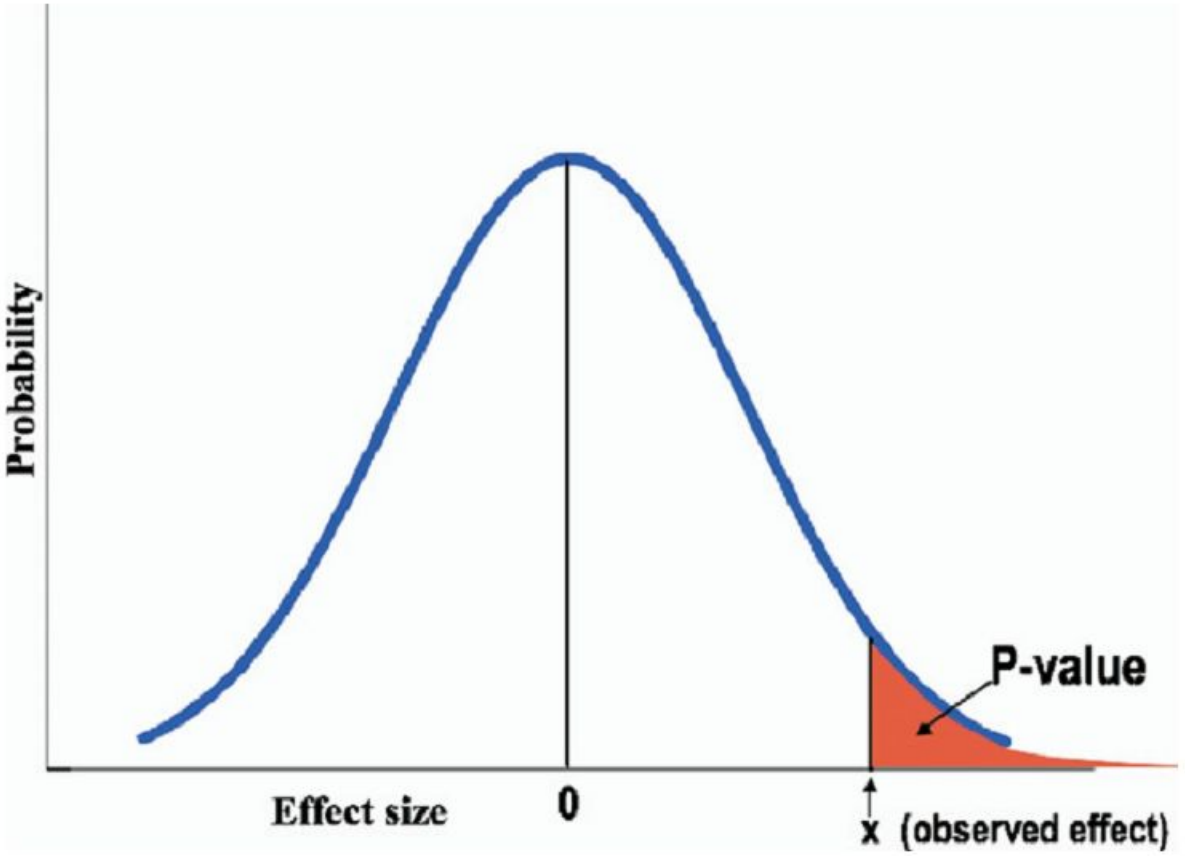
- **где**

$$C = \frac{s_1^2}{n_1} : \left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)$$

# *P*-значение

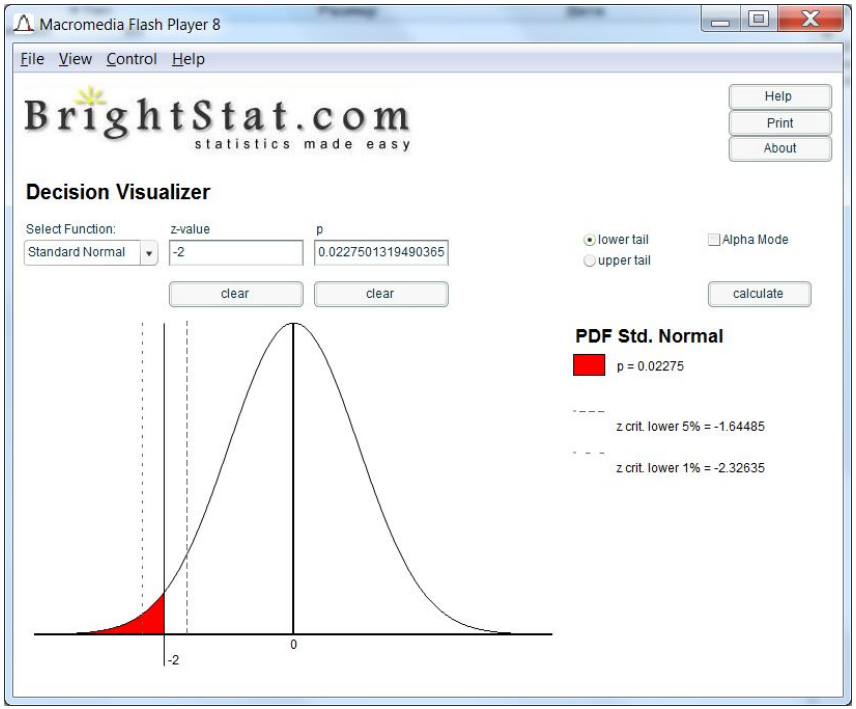
- **Для проверки нулевых гипотез с помощью статистических критериев основным приемом является вычисление значения вероятности, которое называется *P*-значением.**
- ***P*-значение есть условная вероятность, а именно:**
- **Вероятность получить наблюдаемое значение $t_{набл.}$ статистики некоего критерия *T* и все остальные еще менее вероятные значения этой статистики (или значения, еще более отклоняющиеся от ожидаемых) ПРИ УСЛОВИИ, что верна нулевая гипотеза $H_0$:**
  - $P_{val} = \Pr\{|T| \geq |t_{набл.}| \mid H_0\}$.
- **Тут следует обратить внимание на то, что «еще менее вероятные данные» не являются «данными», мы их не наблюдаем.**
- **Мы их додумываем из всех возможных значений статистики критерия *T* в рамках выбранной нами (нулевой) модели.**

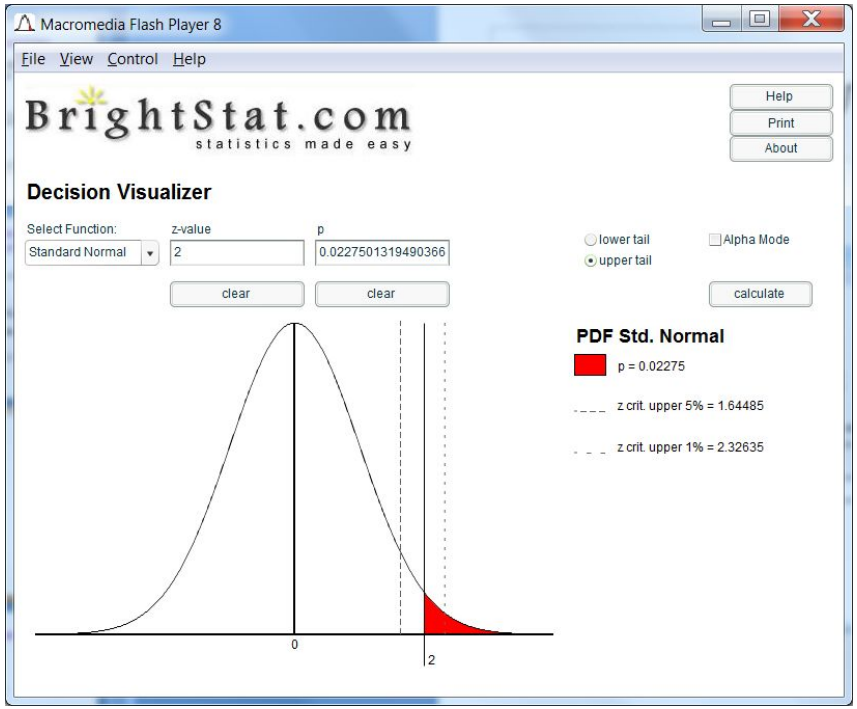
**_P_-значение есть вероятность наблюдать исход (_x_), плюс все «еще более экстремальные исходы». Они представлены затушеванной областью хвоста распределения, соответствующего нулевой модели**



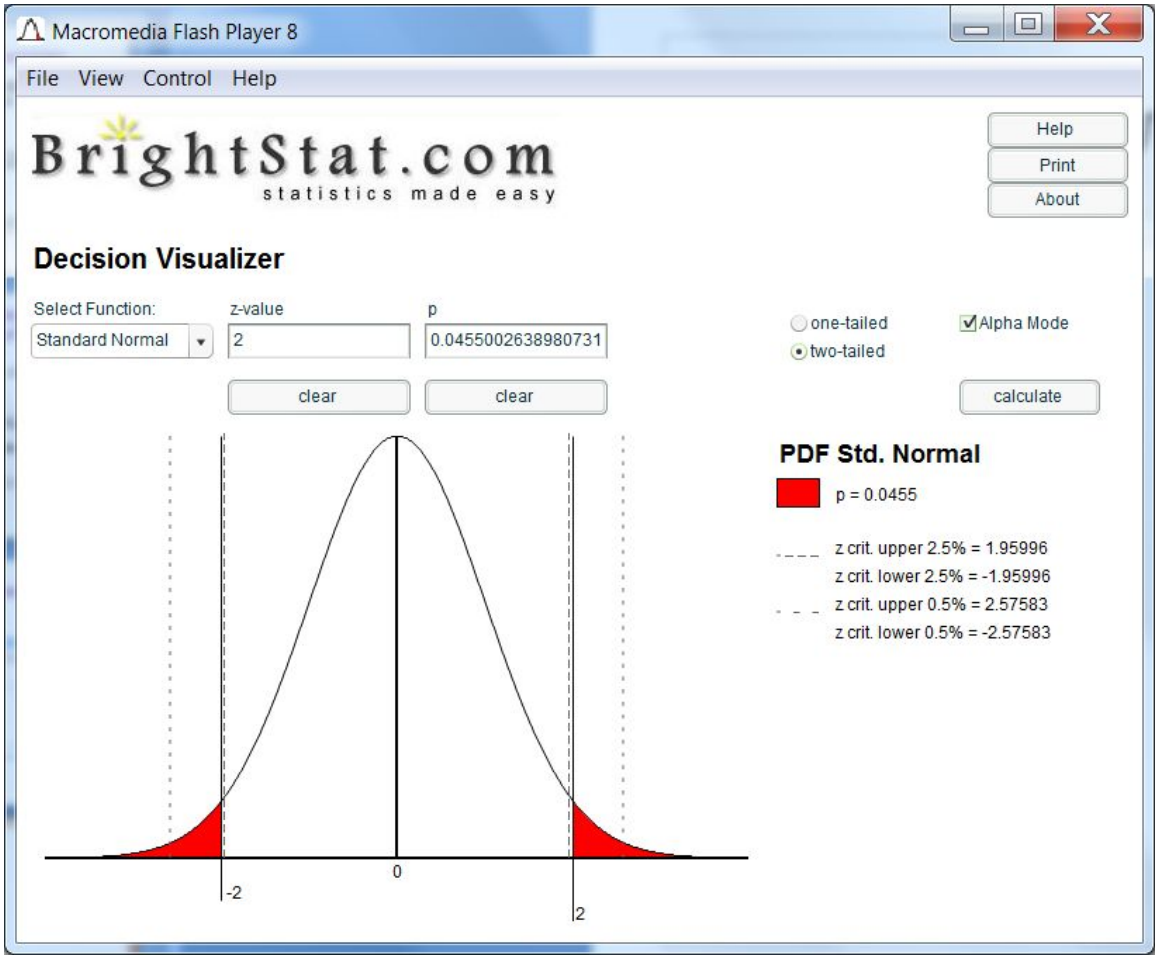

Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Semin. Hematol., 2008. – Vol. 45. – P. 135-140.

# Односторонние *P*-значения



# Двухстороннее *P*-значение



- **Основная логика использования наблюдаемого значения величины *P* состоит в том, что если оно малó, то считается, что малоправдоподобно получить имеющиеся данные при условии, что справедлива нулевая гипотеза.**
- **Как следствие делается вывод, что в таком случае малоправдоподобна и сама нулевая гипотеза.**
- **Это считается достаточным аргументом для того, чтобы отклонить $H_0$ и принять альтернативную гипотезу $H_1$.**

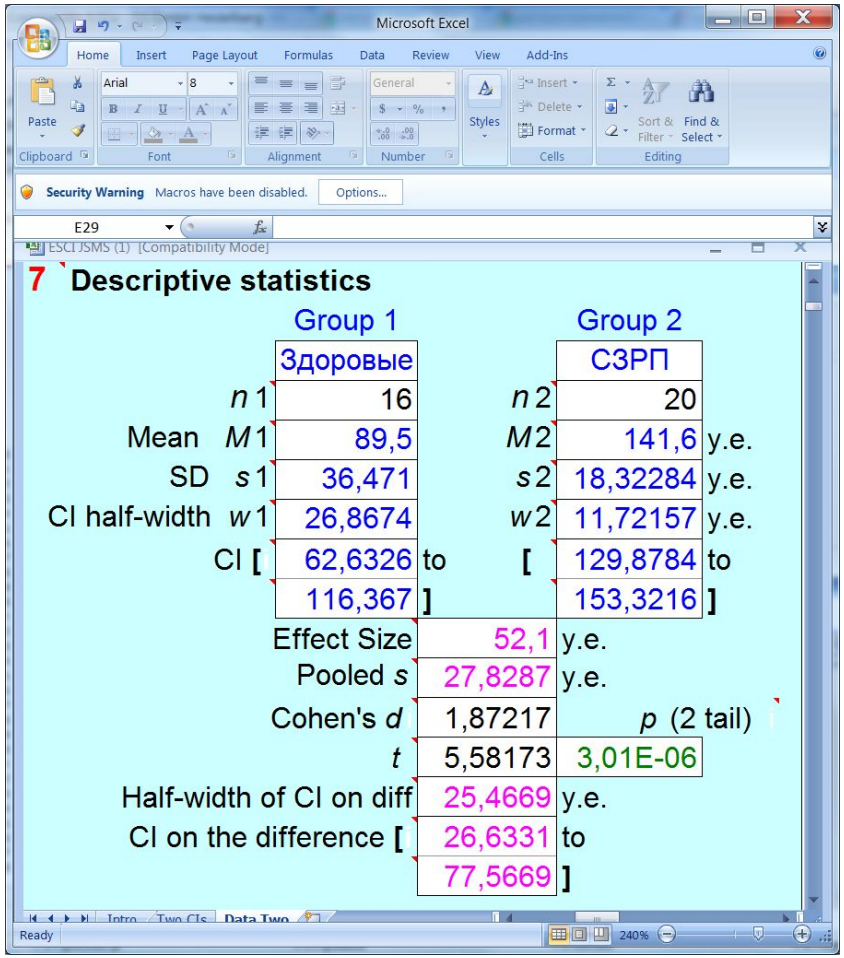
# Выбор порога для значения *P*, и можно ли его обосновать?

- Когда наблюдаемое значение ***P*** мало, то появляется соблазн отвергнуть $H_0$.
- Однако нет никаких ***статистических*** соображений, какое значение ***P*** следует считать настолько малым, чтобы смело отклонить $H_0$.
- Это решение является ***внестатистическим***.
- На практике ***решение отклонить или принять*** $H_0$ ***должно зависеть от обстоятельств.***
- Исследователь в каждой конкретной ситуации должен сам сделать этот выбор.

# Традиционная интерпретация значений *P* (шкала Michelin)

| Рначение *P* | Статистическая значимость | Шкала Мишлена |
|---|---|---|
| > 0,05 | Незначимо | |
| 0,05 – 0,01 | Умеренно значимо | * |
| 0,01 – 0,001 | Значимо | ** |
| < 0,001 | Высоко значимо | *** |

# Результаты статистического сравнение групп матерей здоровых детей и детей с ЗВУР, 1-α = 0,99. Программа ESCI JSMS.xls http://www.latrobe.edu.au/psy/esci/




- **В данном случае**
- $P_{val}$ = 3,0E-06 ≡ $3 \cdot 10^{-6}$.

- **Вывод:**
- **различие в содержании IFN-α/β у матерей здоровых детей и детей с ЗВУР статистически высоко значимо;**
- **во второй группе оно выше, чем в первой.**

# Акт интеллектуальной смелости

- Когда значение *P* очень мало, мы берем на себя смелость отклонить нулевую гипотезу (и принять альтернативную).
- Всякий раз, принимая решение отклонить или принять нулевую гипотезу, мы совершаем акт интеллектуальной смелости.
- И этот акт является внестатистическим.

# Распространенный соблазн

- **Квинтэссенцию традиционных (частотнических) заключений при проверке статистических гипотез принято интерпретировать так:**
- ***чем меньше значение P, тем весомее доводы против нулевой гипотезы $H_0$, которые предоставляют нам имеющиеся данные; тем больше у нас оснований сомневаться в $H_0$.***
- **Отсюда невольно (и вроде бы естественно) возникает соблазн интерпретировать значение *P* как вероятность нулевой гипотезы.**

# Распространенное заблуждение

- **Значение *P* не есть вероятность нулевой гипотезы !**
- **Поскольку *P-значение* вычисляется при условии,**
- **что справедлива нулевая гипотеза $H_0$:**
  - $P_{val} = \Pr\{|T| \geq |t_{набл.}| \,|\, H_0\}$,
- **то оно никак не может быть вероятностью нулевой гипотезы:**
  - $P\{t \,|\, H_0\} \neq P\{H_0 \,|\, t\}$

- ***P*-значение потому столь привлекательно для ученых, что с ним очень легко получить «значимый» («достоверный») результат, даже когда на самом деле эффекта нет.**

# «Цена» значения *P*

| значение *P* | Нижняя граница для вероятности нулевой гипотезы $P(H_0)$ | Верхняя граница для вероятности воспроизведения $P_{repr}$ |
|---|---|---|
| 0,05 | > 30% | < 50% |
| 0,01 | > 10% | < 73% |
| 0,001 | > 2% | < 90% |

**Для наглядности значения в таблице округлены до первой значащей цифры. Более точно значения для $P(H_0)$ (сверху вниз) равны 29%, 11% и 1,8%.**

**Posavac E.J. Using *p* values to estimate the probability of statistically significant replication // Understanding Statistics, 2002. – Vol. 1. – No. 2. – P. 101-112.**

# Бейзовская интерпретация значения *P*

- **Обычно принято интерпретировать значения *P* как меру доказательства, предоставляемого имеющимися данными, против нулевой гипотезы.**
- **Однако с точки зрения бейзовской статистики значение *P* есть всего лишь вероятность того, что при повторении эксперимента будет получена разность средних с противоположным знаком.**
- **При такой интерпретации понятно, что значение *P* ничего не говорит ни о вероятности нулевой гипотезы $P\{H_0|t\}$, ни о размере эффекта, в данном случае о разности средних.**

# Привычка свыше нам дана

- **Это прекрасно понимал Р.А. Фишер:**
- ***«Критерий значимости не позволяет нам делать какие-либо выводы о проверяемой гипотезе в терминах математической вероятности»*** **(Fisher R.A. The design of experiments. Edinburgh: Oliver & Boyd, 1935).**
- **Тем не менее многие исследователи (авторы) имеют дурную привычку обращать внимание исключительно на значение *P*,**
- **игнорируя практическую (клиническую) важность полученных ими результатов, игнорируя размер эффекта.**

# Статистическая значимость и размер эффекта

- **Эффект (различие, связь, риск, польза, ассоциация и т. п.) может быть статистически значимым, но его практическая (например, клиническая) ценность может оказаться ничтожной.**
- **«Статистически значимый» не означает «значительный», «практически важный», «ценный».**
- **Эффекты могут быть реальными, неслучайными, но практически пренебрежимо малыми.**

# Размер эффекта

- **Вопрос о клинической (практической) ценности (важности) наблюдаемого размера эффекта**
- **является ключевым при интерпретации результатов биомедицинских исследований, таких как диагностические исследования, клинические испытания и т. п.**
- **Размер эффекта можно выражать в реальных единицах, а можно сделать его безразмерным – Стандартизированным.**

# Стандартизированный размер эффекта по Коуэну (Cohen) $d_C$

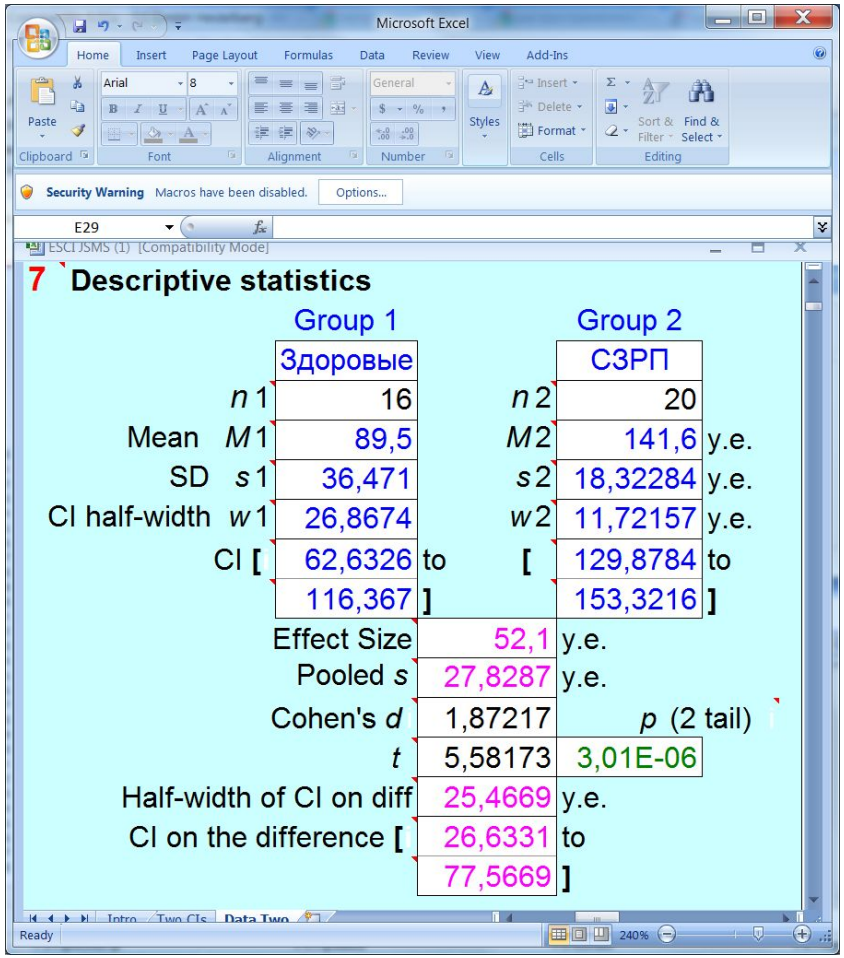
$$d_C = \frac{M_1 - M_2}{S_{pooled}}$$

# Интерпретация стандартизированного размера эффекта $d_c$

http://www.sportsci.org/resource/stats/

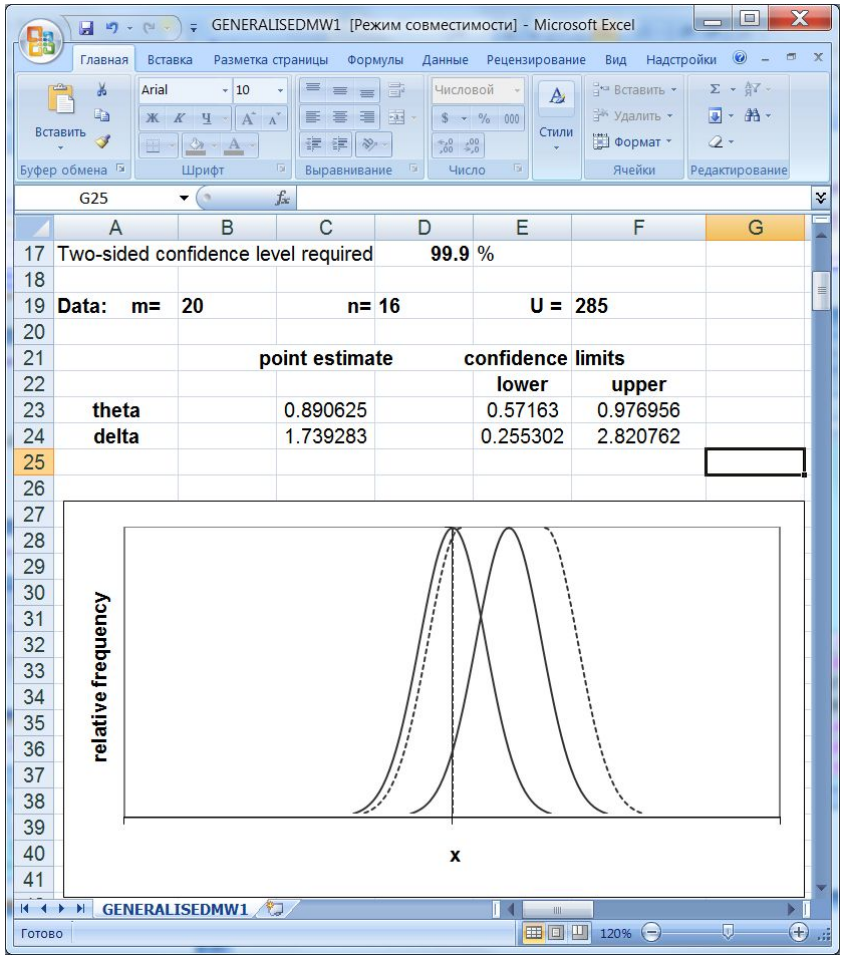
| Размер эффекта, $d_c$ | Градация эффекта |
|---|---|
| 0 – 0,2 | Ничтожный |
| 0,2 – 0,5 | Малый |
| 0,5 – 1,0 | Умеренный |
| 1,0 – 2,0 | Большой |
| 2,0 – 4,0 | Очень большой |
| 4,0 - ∞ | Исключительно большой |

# Результаты статистического сравнения групп матерей здоровых детей и детей с ЗВУР, (1 - α) = 0,99. Программа ESCI JSMS.xls http://www.latrobe.edu.au/psy/esci/




- **В данном примере абсолютный размер эффекта *ES* есть попросту разность средних:**
  - $ES = M_2 - M_1 = {}_{26,6}52,1_{77,6}$ **у.е.**
- **Стандартизированный размер эффекта по Коуэну:**
  - $d_C = 1,87$
- **Его можно интерпретировать как сильный (большой).**

# Непараметрическая оценка $d_c$




- **95%-й ДИ:**
  - $_{0,8}^{1,7}{}_{2,5}$
- **99%-й ДИ:**
  - $_{0,6}^{1,7}{}_{2,6}$
- **99,9%-й ДИ:**
  - $_{0,3}^{1,7}{}_{2,8}$

# Бейзов фактор, *BF*

- **Бейзов фактор *BF* принципиально отличается от значения *P*.**
- **Бейзов фактор не является вероятностью сам по себе, а является отношением вероятностей, и он может варьироваться от нуля до бесконечности.**
- **Он требует знания двух гипотез, тем самым четко указывая, что если есть свидетельства против нулевой гипотезы, то должны существовать свидетельства и в пользу альтернативной гипотезы.**

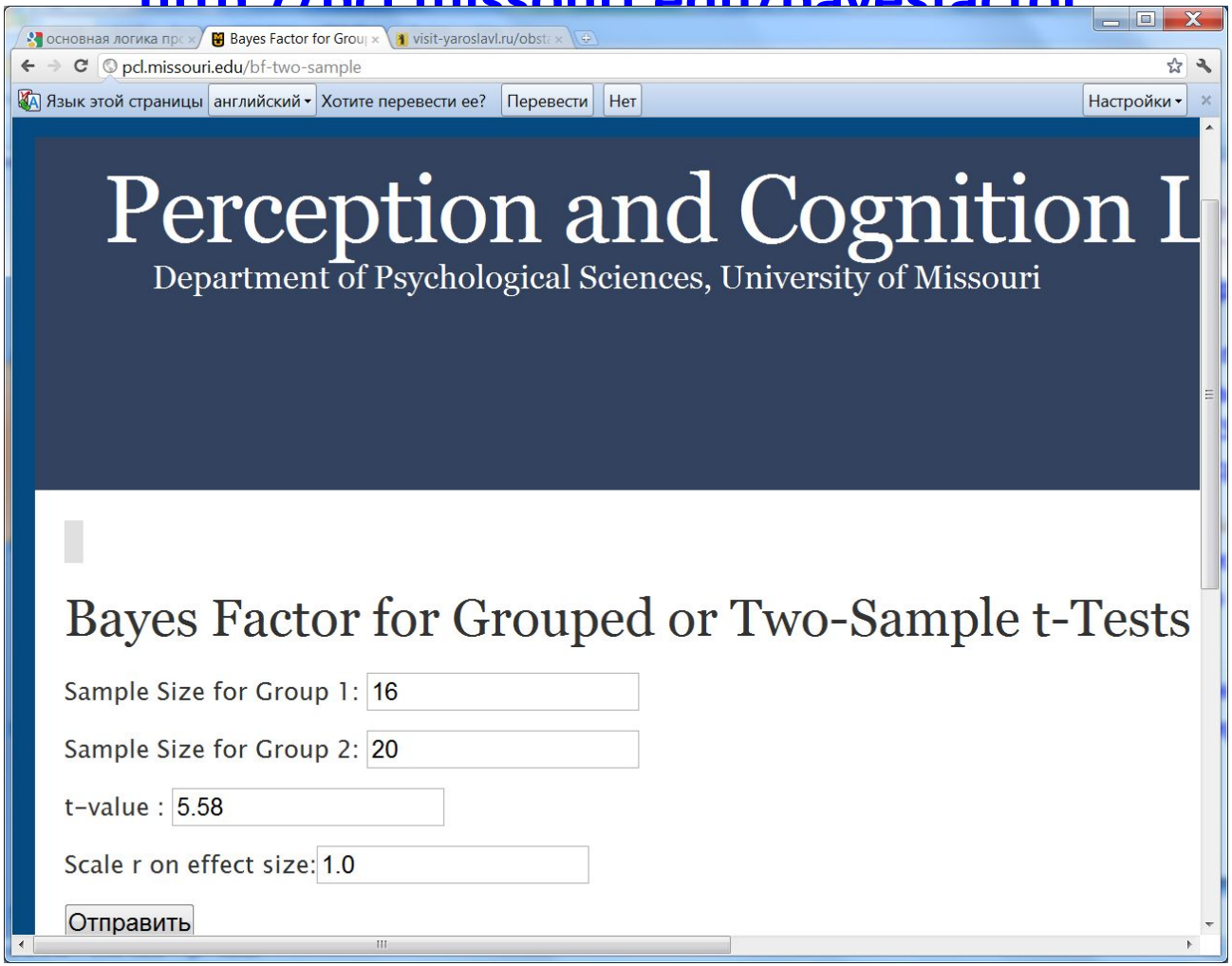
- $BF_{01} = P(D|H_0) / P(D|H_1)$
- $BF_{10} = 1 / BF_{01} = P(D|H_1) / P(D|H_0)$

# Интерпретация убедительности Бейзовых факторов, $BF_{10}$ и $BF_{01}$

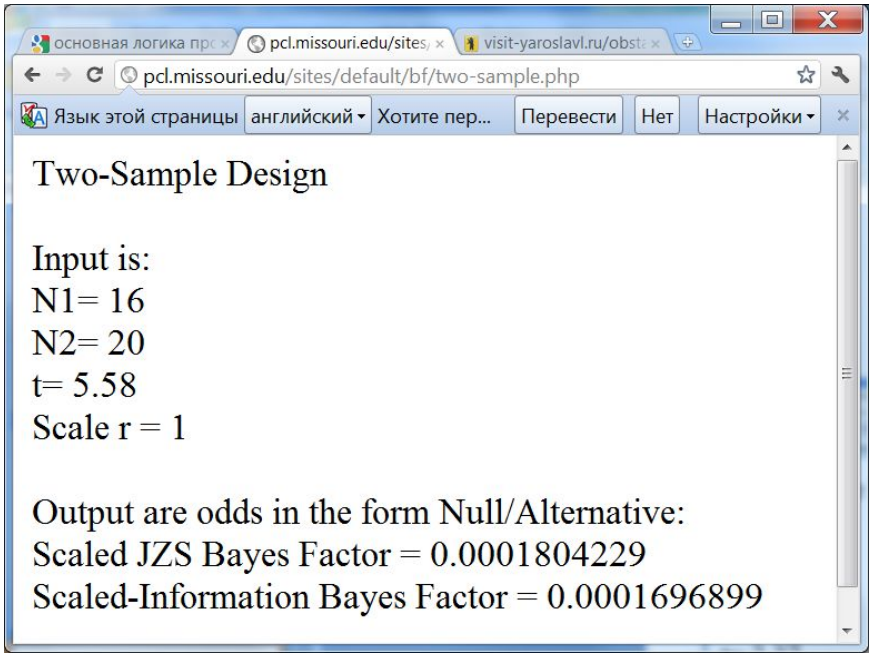
| $BF_{01}$ | *Свидетельство в пользу гипотезы $H_0$ против гипотезы $H_1$* |
|---|---|
| >100 | Убедительное |
| 30 – 100 | Очень сильное |
| 10 – 30 | Сильное |
| 3 – 10 | Умеренное (слабое) |
| 1 – 3 | Пренебрежимо малое |
| $BF_{10}$ | *Свидетельство в пользу гипотезы $H_1$ против гипотезы $H_0$* |

# Бейзов фактор, програма Bayes Factor Calculators

http://pcl.missouri.edu/bayesfactor



# Вывод результатов (output)




- В 5555 раз (1/0,00018) более правдоподобно получить наблюдаемое различие
- (*ES* = 52,1 у.е.) между сравниваемыми группами при условии, что верна гипотеза $H_1$: $ES \neq 0$, нежели при условии, что верна гипотеза $H_0$: $ES = 0$.
- Такое значение $BF_{01}$ принято интерпретировать как чрезвычайно убедительное свидетельство против нулевой гипотезы $H_0$: $ES = 0$ в пользу альтернативной гипотезы $H_1$: $ES \neq 0$.

- **Достаточно малое значение *P* заставляет думать, что произошло нечто неожиданное.**
- **И обычно это интерпретируется как неверность нулевой гипотезы.**
- **Однако, если для этих же данных бейзов фактор $BF_{01}$ не мал, то причину таких неожиданностей следует искать не в том, что неверна научная нулевая гипотеза.**
- **Возможны иные причины этого, такие как экспериментальное смещение или неверная модель.**
- **Для исследования иных причин, нужны другие альтернативные гипотезы.**

# Статистические предсказания и воспроизводимость

# Значение вероятностной *P*-величины

- **Значение *P* есть наблюдаемое значение (реализация) соответствующей случайной величины**

$$\tilde{P}$$

- **Всякий раз мы наблюдаем одно из ее возможных значений.**

- **Отсюда следует, что, строго говоря, на основе всего лишь одного изолированного исследования нельзя делать определенные выводы.**
- **Любое научное исследование должно повторяться многократно, и должна исследоваться воспроизводимость результатов.**

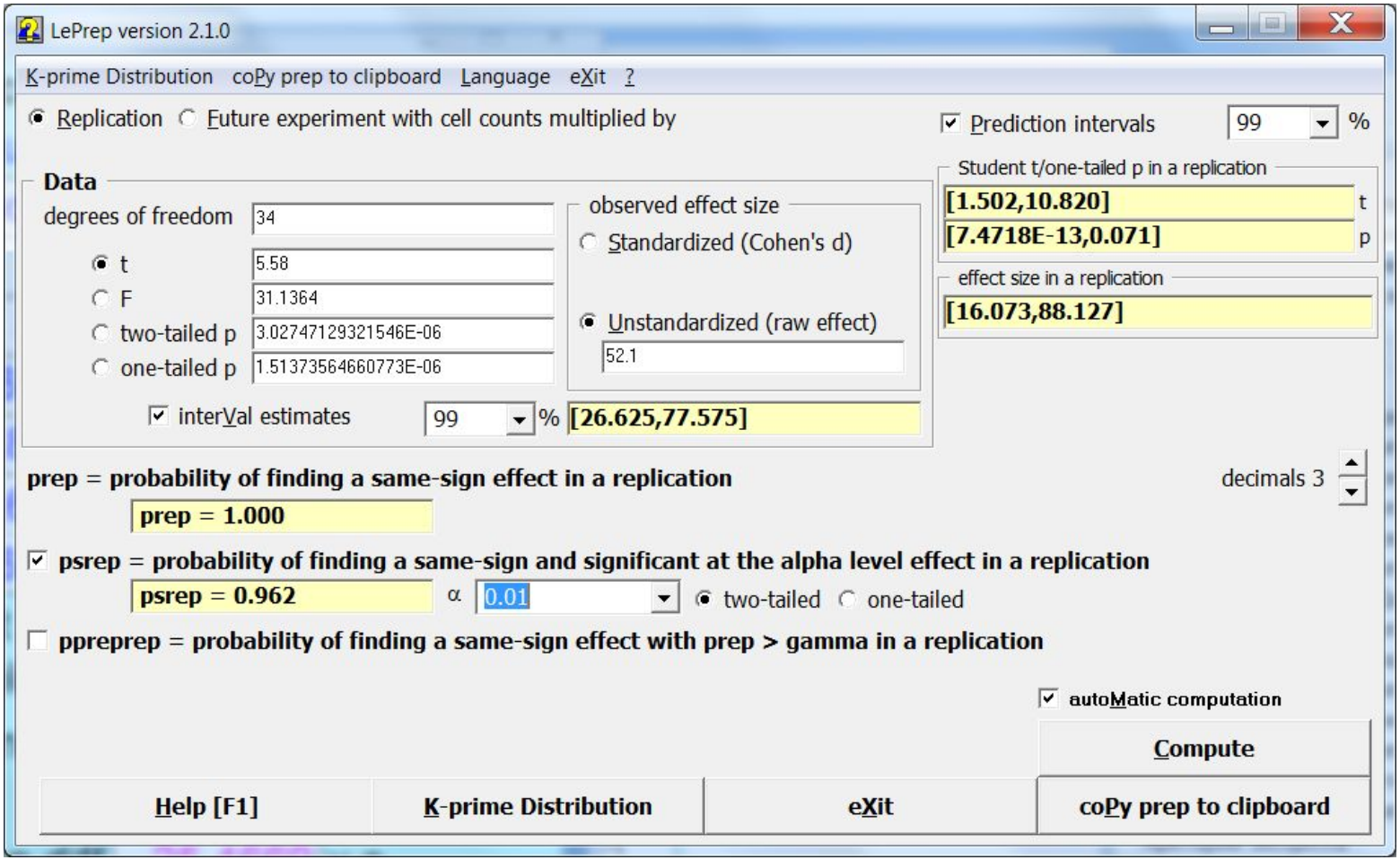
# Доверяя, повторяй

- **Часто считается, что если получен «статистически значимый» результат, то это исключает необходимость повторить исследование.**
- **Повторность (воспроизведение) часто рассматривается как нечто суетное и мирское.**

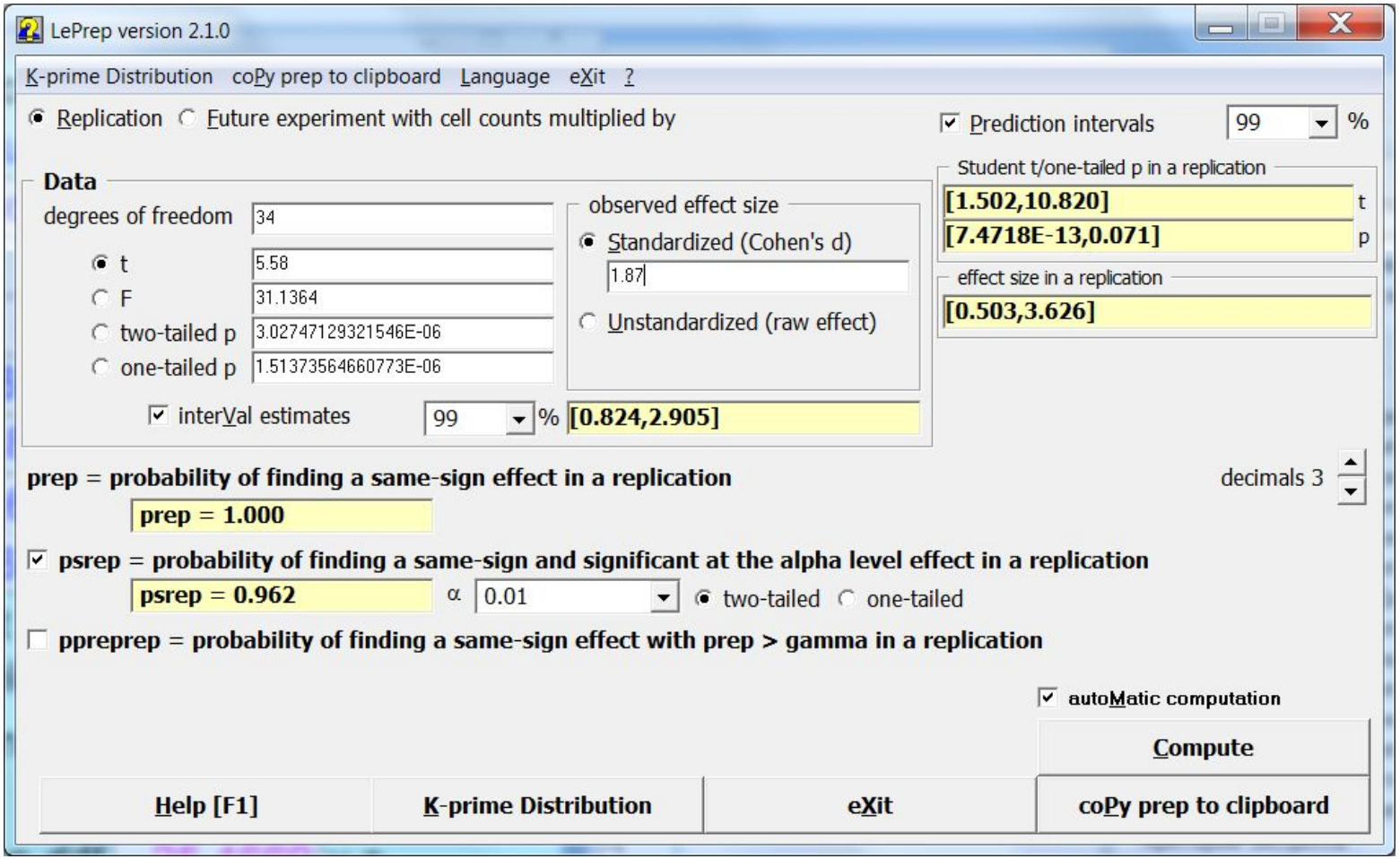
- ***«Проверка нулевой гипотезы есть метод обнаружения маловероятных событий, которые заслуживают дальнейшего изучения»*** **(Fisher).**

# Воспроизводимость и предсказания абсолютного размера эффекта для групп матерей здоровых детей и детей с ЗВУР. Программа LePrep

http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/PAC.htm



# Воспроизводимость и предсказания стандартизированного размера эффекта по Коуэну (Cohen) $d_c$



# Воспроизводимость и предсказания размеров эффекта *ES* и $d_c$ для групп матерей здоровых детей и детей с ЗВУР

| Показатель | *ES* | $d_c$ |
|---|---|---|
| 99%-е предсказательные интервалы (ПИ) для размеров эффекта | [16,1; 88,1] | [0,50; 3,63] |
| 99%-й предсказательный интервал (ПИ) для $P_{val}$ | $[7 \cdot 10^{-13}$; 0,071] | |
| $P_{srep}$ - вероятность воспроизведения эффекта с тем же знаком и значимого на уровне $\alpha = 0,01$ | 0,96 | |

**При независимом повторении эксперимента эффект может не воспроизвестись и оказаться статистически незначимым (нижняя граница 99%-го ПИ для $P_{val} = 0,071 > 0,05$) и размер эффекта по Коуэну может оказаться малым, достигая нижней границы 99%-го ПИ для него: 0,5.**

# Ошибки I и II рода и мощность статистического критерия

## $H_0$: есть беременность; $H_1$: нет беременности

**Истинный позитив, верна $H_0$**



**Ложный позитив, ошибка I рода, ложная тревога**

**Ложный негатив, ошибка II рода, халатная беспечность**



**Истинный негатив, верна $H_1$**

# Судебные ошибки

| Вердикт: подозреваемый | Действительность: подозреваемый | |
|---|---|---|
| | $H_0$: виновен | $H_1$: невиновен |
| Виновен | Верное решение | Неверное решение (Ошибка *первого* рода, ложное осуждение) |
| Невиновен | Неверное решение (Ошибка *второго* рода, ложное оправдание) | Верное решение |

# Диагностика

| Тест / Болезнь | Есть болезнь (D = 1) | Нет болезни (D = 0) |
|---|---|---|
| Положи-тельный |  Чувствительность | ✗ Ложный (+) |
| Отрица-тельный | ✗ Ложный (-) |  Специфичность |

# Теория Неймана-Пирсона: Ошибки I и II рода и мощность критерия

| Критерий / Действи-тельность | $H_0$ не отклонена | $H_0$ отклонена |
|---|---|---|
| Верна $H_0$, нет различия ($D = 0$) |  Верное решение |  Ошибка I рода с вероятностью **α** |
| Верна $H_1$, есть различие ($D \neq 0$) |  Ошибка II рода с вероятностью **β** |  Мощность 1 - **β**; Верное решение |

# Ошибки I и II рода

- **Ошибка I рода: отклонение верной нулевой гипотезы;**
- **Аналитик решает (берет на себя смелость) отклонить нулевую гипотезу, когда в действительности она верна.**
- **Вероятность ошибки I рода традиционно обозначается $\alpha$.**

- **Ошибка II рода: принятие неверной (ложной) нулевой гипотезы;**
- **Аналитик решает (берет на себя смелость) принять нулевую гипотезу, когда в действительности она неверна.**
- **Вероятность ошибки II рода традиционно обозначается $\beta$.**

# Ошибки I и II рода

| Результат применения статистического критерия | Верная гипотеза | |
|---|---|---|
| | $H_0$ | $H_1$ |
| Решено принять $H_0$ и отклонить $H_1$ | $H_0$ верно принята<br>$H_1$ верно отклонена<br>Вероятность (1 – β) – мощность | $H_1$ неверно принята<br>$H_0$ неверно отклонена,<br>(Ошибка *первого* рода, ложная тревога)<br>Вероятность α – уровень значимости |
| Решено принять $H_1$ и отклонить $H_0$ | $H_1$ неверно принята<br>$H_0$ неверно отклонена,<br>(Ошибка *второго* рода,недостаточная бдительность)<br>Вероятность β | $H_1$ верно принята,<br>$H_0$ верно отклонена<br>Вероятность (1 – α) |

# Компромисс

- **Например, в случае металлодетектора. $H_0$ – обнаружен нейтральный предмет.**
- **повышение чувствительности прибора приведёт к увеличению риска *ошибки первого рода* (ложная тревога), а**
- **понижение чувствительности - к увеличению риска *ошибки второго рода* (пропуск запрещённого предмета).**

# Мощность статистического критерия

- Мощность статистического критерия есть вероятность того, что критерий правильно отклонит ложную нулевую гипотезу (правильно примет верную альтернативную гипотезу).
- Традиционно ее обозначают $(1 - \beta)$, где $\beta$ - вероятность ошибки II рода.
- Чем больше мощность критерия, тем меньше вероятность совершить ошибку II рода.
- Мощность статистического критерия измеряет способность критерия выявлять истинные различия (эффекты).
- Ее можно интерпретировать как чувствительность статистического критерия к отклонениям от условий нулевой гипотезы.

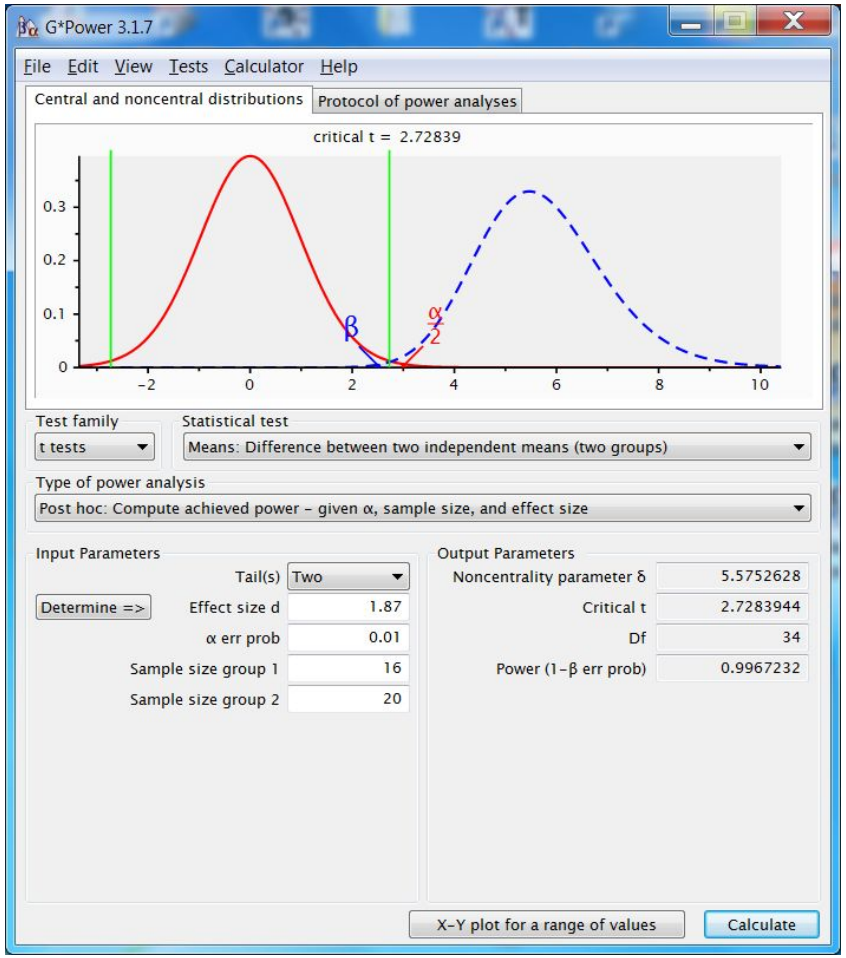
- **Мощность отвечает на вопрос:**
- **Если эффект (определенного размера) действительно существует, то какова вероятность того, что эксперимент с выборкой определенного размера даст «статистически значимый» результат?**

# Анализ мощности *a priori* или *post-hoc*

- **Анализ мощности можно проводить либо *a priori*, т. е. до получения данных, либо *post hoc*, т.е. после получения данных.**
- ***A priori* анализ мощности обычно используется для оценки объема выборки *N*, необходимого для достижения приемлемой мощности.**
- ***Post hoc* анализ мощности используется для оценки достигнутой мощности.**
- **В этом случае предполагается, что наблюдаемый эффект и его варьирование равны истинным значениям параметров.**

# Оценка достигнутой мощности (post hoc). Программа G*Power

http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/




- **Достигнутая мощность проведенного исследования составила**
- **(1 – β) = 0,9967**

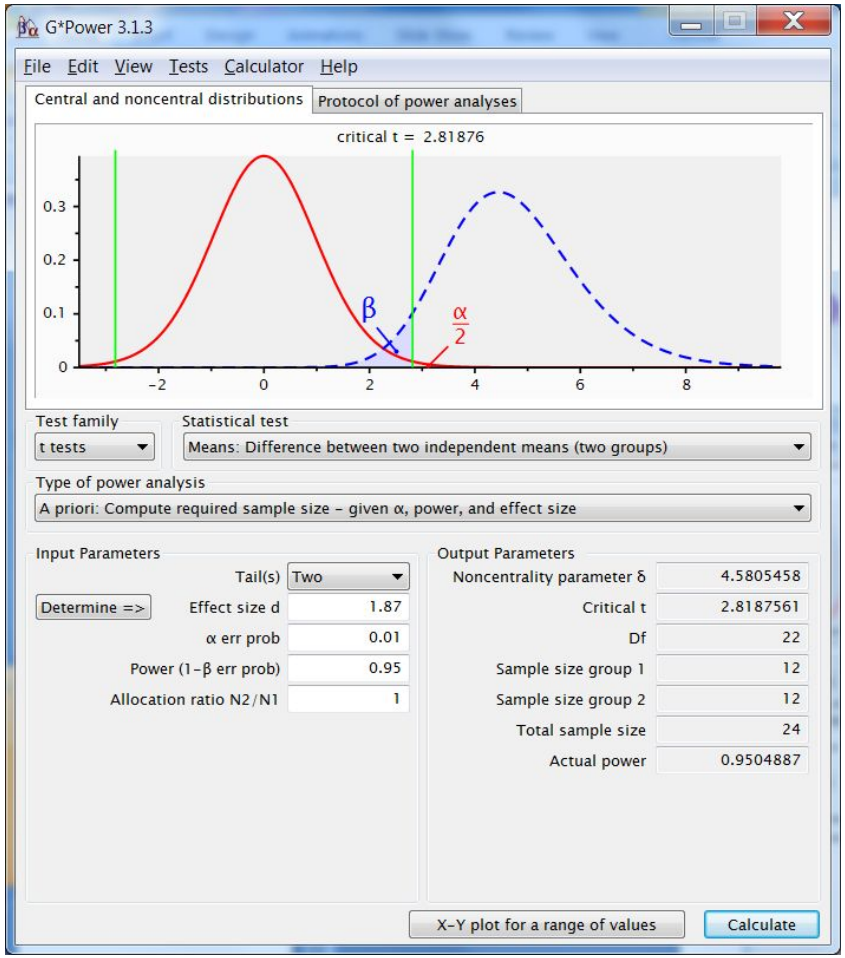
# Элементы планирования эксперимента

# Программа G*Power

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3

- **Оценка *a priori* минимально необходимого объема выборки *N* для достижения статистически значимого отличия наблюдаемой доли от ожидаемого значения при заданных уровне значимости *α* и мощности (1 – β).**

# Оценка необходимых объемов выборок (a priori)




- **Для достижения приемлемой статистической мощности**
- **(1 – β) = 0,95**
- **достаточно было иметь группы по 12 человек.**

# Научный метод

- Ни один уважающий себя ученый не ограничится в своих исследованиях одним-единственным экспериментом, хотя бы ради того, чтобы исключить неизбежные ошибки наблюдения, измерений, подсчетов и т. д.
- Законы Менделя стали законами только после того, как их справедливость была продемонстрирована для всех диплоидных организмов, размножающихся половым путем – от растений до человека.
- Смешно было бы, если Майкельсон и Морли провели бы всего лишь одно измерение скорости света и на основании такого этого единственного измерения утверждали бы, что скорость света постоянна (в пределах точности измерения, которую и оценить-то невозможно, если измерение одно).

# Культ одиночного изолированного исследования

- **Чрезмерное «увлечение» анализом одиночных наборов данных пронизывает почти всю статистическую литературу и является серьезной болезнью статистического образования.**
- **Конечно же, не всегда возможно собрать больше данных, и некоторые научные эксперименты столь дорогостоящи, что правомочно извлекать из данных как только возможно больше информации.**
- **Однако, во многих других ситуациях *можно и нужно* собирать как можно больше данных, и это представляется благоразумным.**
- **Наука не дается малой кровью.**

## **Джон Уайлдер Тьюки** (*John Wilder Tukey*, 16.04.1915 — 26.07.2000)


- **Исследования должны быть как минимум двухэтапными.**
- **Первый этап – разведочное (пилотное, порождающее гипотезы) исследование.**
- **Второй этап – проверочное (подтверждающее или опровергающее) исследование.**
- **Оно планируется на основе результатов разведочного исследования.**