БМС –Биомедстатистика

Никита Николаевич Хромов-Борисов
Кафедра физики, математики и информатики СПбГМУ им. акад. И.П. Павлова
Nikita.KhromovBorisov@gmail.com
(812) 234-18-40 – дом.
(812) 234-66-55 – раб.
8-952-204-89-49 – моб.

Лекция 2. Гармонизация статистических доказательств и предсказаний

- **Эпидемиологи смотрят на мир сквозь решетку таблицы 2×2.**
- **При этом надо помнить, что результат обследования является бинарным (дихотомическим):**
- **либо положительным, либо отрицательным, т.е. без промежуточных градаций.**
- **Дихотомическое деление привлекательно своей простотой.**
- **Однако такое упрощение является серьезным ограничением, поскольку результаты подобных обследований зачастую являются мерными.**

Два основных типа Статистических Данных и их моделей

- **Счетные Данные**
- ***Счетные Данные*** **получают путем подсчета объектов, предметов.**
- **Моделью для них являются *Дискретные Случайные Величины* и, соответственно, *Дискретные Распределения***

- **Мерные Данные**
- ***Мерные Данные*** **получаются путем измерения признаков.**
- **Моделью для них являются *Непрерывные Случайные Величины* и, соответственно, *Непрерывные Распределения.***

- **Счетные данные подсчитываются.**
- **Мерные данные измеряются.**

Пример: каковы признаки этой собаки?

- Качественные:
- Ее окрас - коричневый с черным
- У нее длинная шерсть
- Она энергичная
- Количественные:
- счетные:
  - У нее 4 ноги
  - У нее два брата
- мерные:
  - Ее вес – 25,5 кг
  - Ее рост (в холке) 56,5 см

Цитокины и диагностика синдрома задержки развития плода (СЗРП)

Королева Л.И.

СЗРП

- **Термин Синдром задержки развития плода (СЗРП) используется для описания плода, масса которого гораздо меньше ожидаемой для данного гестационного возраста.**
- **Плод/ребенок, масса тела которого попадает в нижние 10% распределения нормальной популяции данного гестационного возраста, рассматривается как имеющий СЗРП.**
- **Оценка базируется на стандартизованных таблицах соотношения массы тела и гестационного возраста.**
- **По данным отечественных авторов СЗРП в акушерской практике встречается с частотой от 5% до 17,6%.**
- **Согласно последним отечественным данным частота (распространенность) СЗРП на протяжении последних 10 лет находилась в пределах 3,5 – 4,6%.**

СЗРП

- **Плод с задержкой внутриутробного развития подвержен повышенному риску внутриутробной гибели или неонатальной смерти, асфиксии до или во время родов.**
- **Сразу после рождения ему угрожает аспирация мекония, гипогликемия, гипотермия, РДС и множество других состояний.**
- **Частота перинатальной смертности при СЗРП повышена в 7-10 раз, очень велика и перинатальная заболеваемость.**
- **Перечисленные отрицательные обстоятельства показывают, как важно выявлять СЗРП еще до родов, оптимизировать условия внутриутробного развития плода, планировать и проводить роды, используя наиболее безопасные средства, и обеспечивать наилучший уход в послеродовом периоде.**

Содержание цитокина у 16 здоровых матерей и у 20 матерей с СЗРП

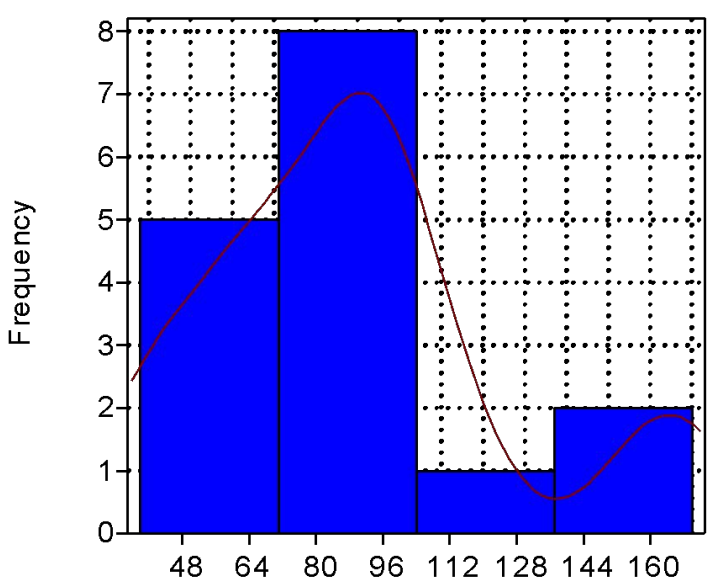
Здоровые				СЗРП			
№	Цитокин, у.е.	№	Цитокин, у.е.	№	Цитокин, у.е.	№	Цитокин, у.е.
1	38	9	92	1	104	11	144
2	42	10	93	2	121	12	146
3	58	11	94	3	123	13	147
4	59	12	101	4	123	14	149
5	70	13	103	5	127	15	151
6	71	14	115	6	130	16	153
7	81	15	159	7	132	17	162
8	86	16	170	8	134	18	168
				9	134	19	171
			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

Гистограмма

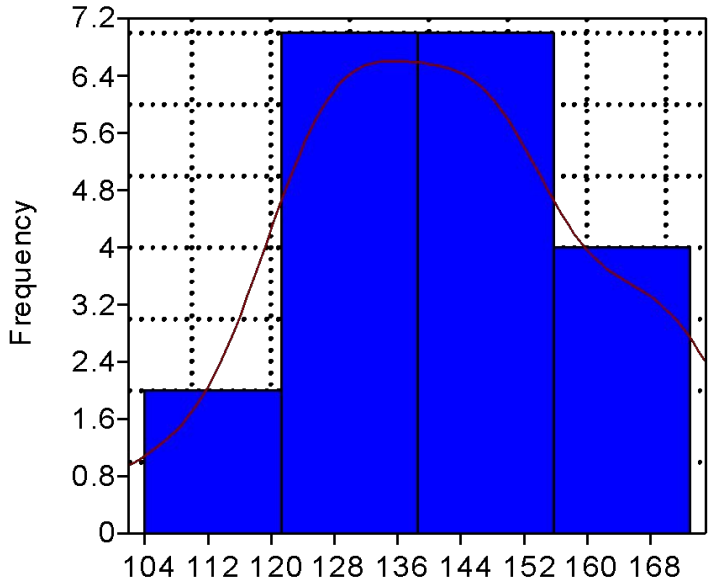
- **Гистограмма**
- **(от др.-греч. ἱστός — столб + γράμμα — черта, буква, написание)**
- **— столбиковая диаграмма**
- **— способ графического представления табличных данных.**

Гистограммы содержания цитокина у матерей здоровых детей и детей с СЗРП

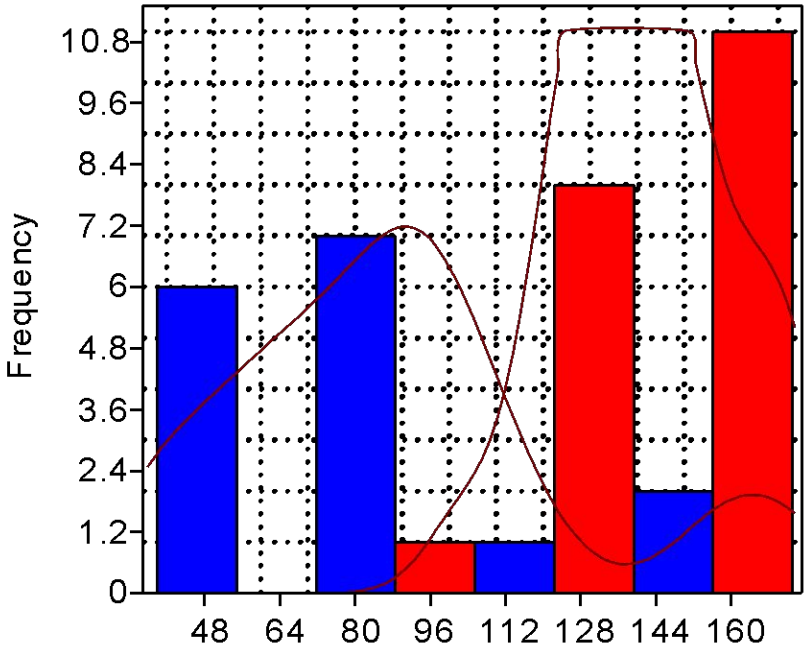
Здоровые



СЗРП

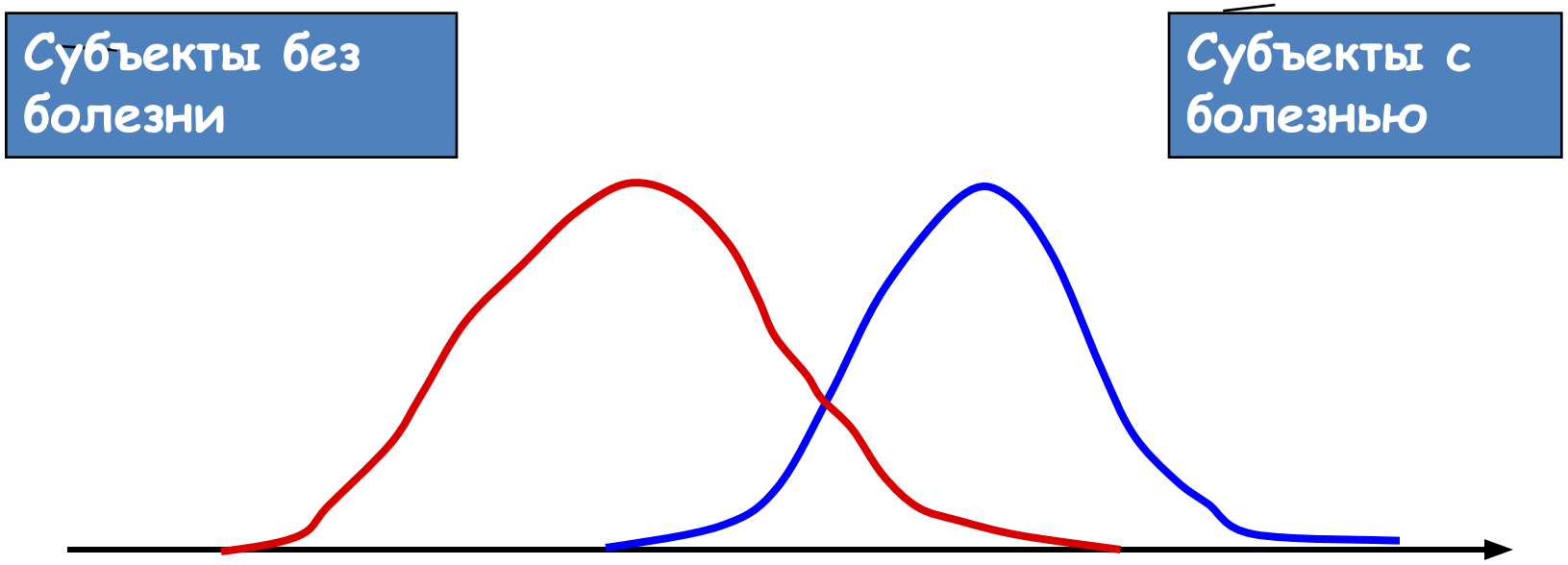


Сопоставление гистограмм содержания цитокина у матерей здоровых детей и детей с СЗРП

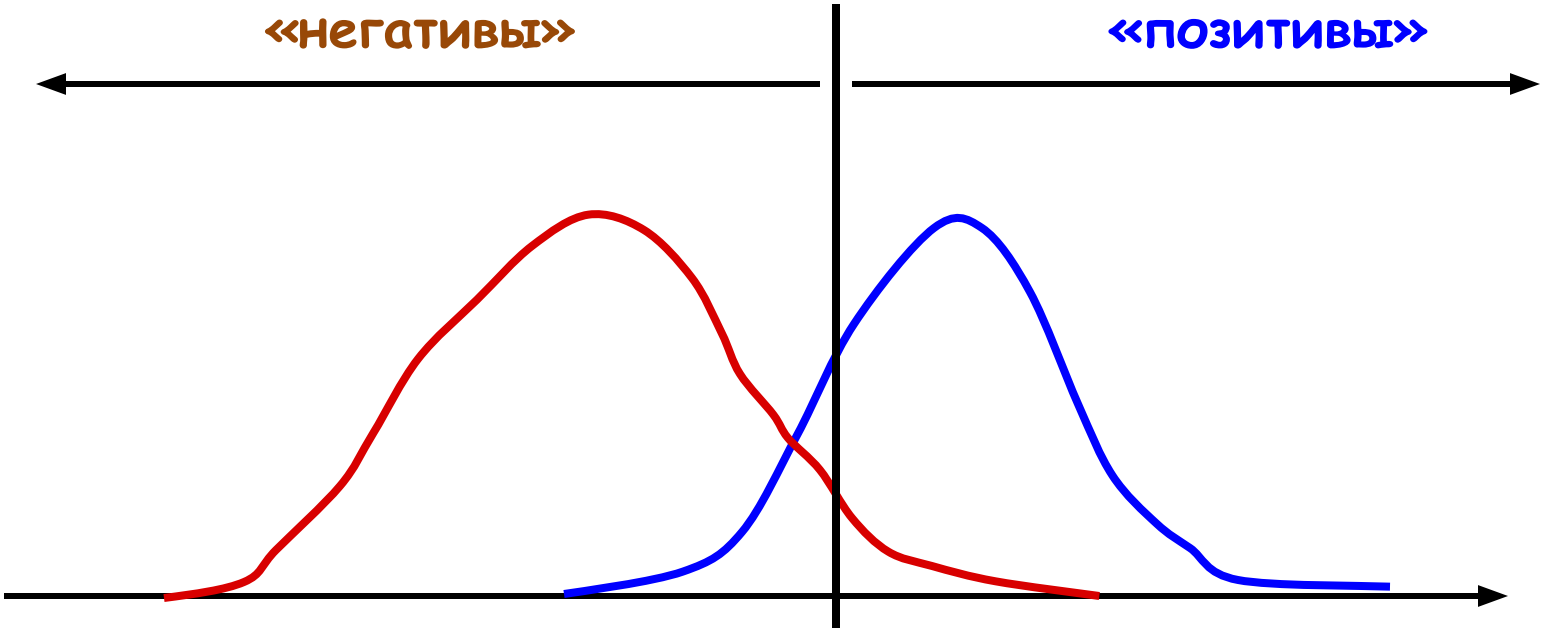


ROC-анализ: удобный инструмент для оценки качества диагностических исследований на основе мерных признаков

Распределения мерного диагностического признака у субъектов с болезнью и без нее

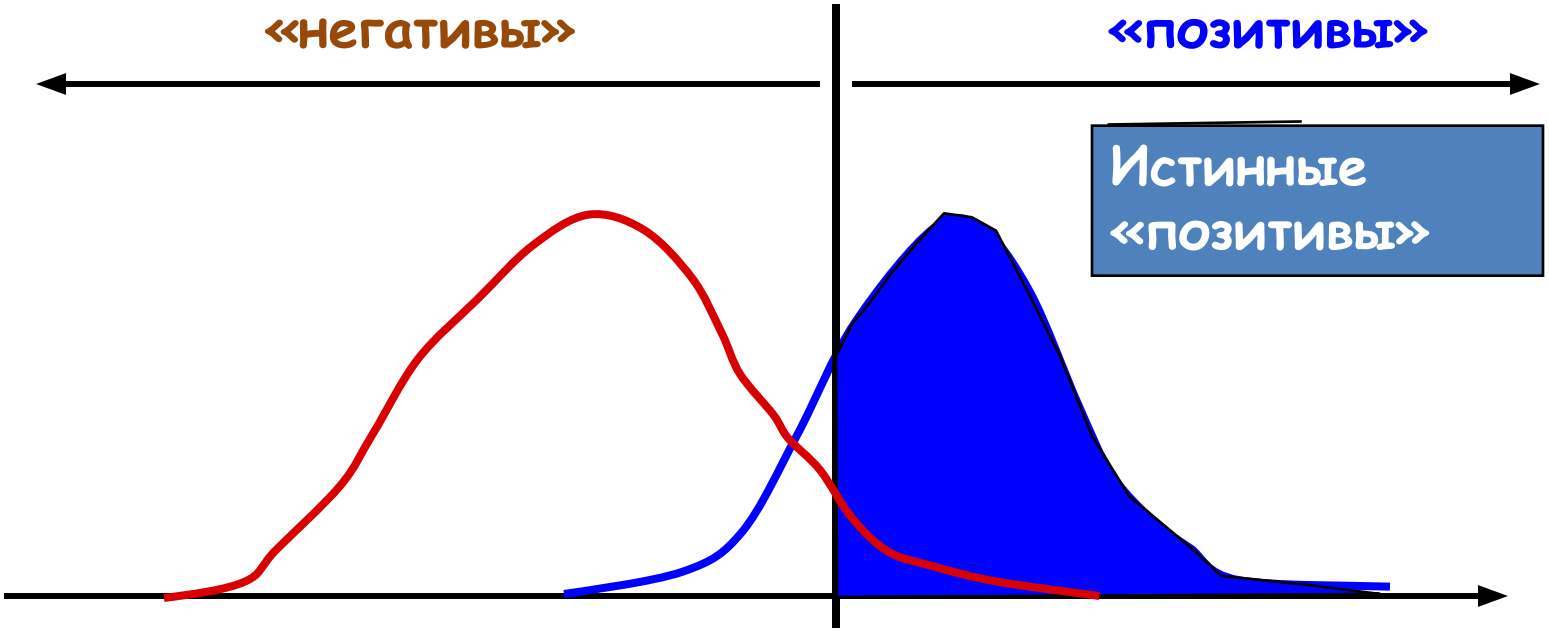


Пороговое отсекающее значение



Значения мерного диагностического признака

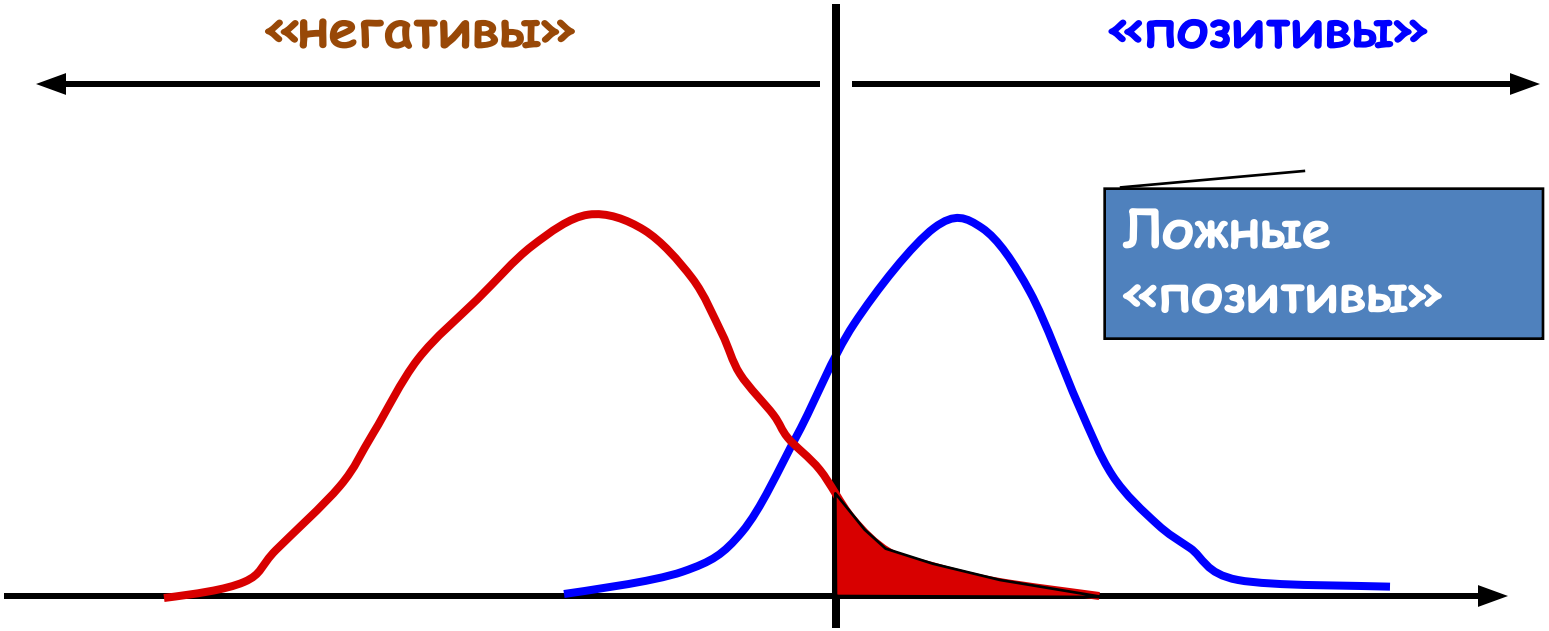
Истинные «позитивы»



Значения мерного диагностического признака

Субъекты без болезни **Субъекты с болезнью**

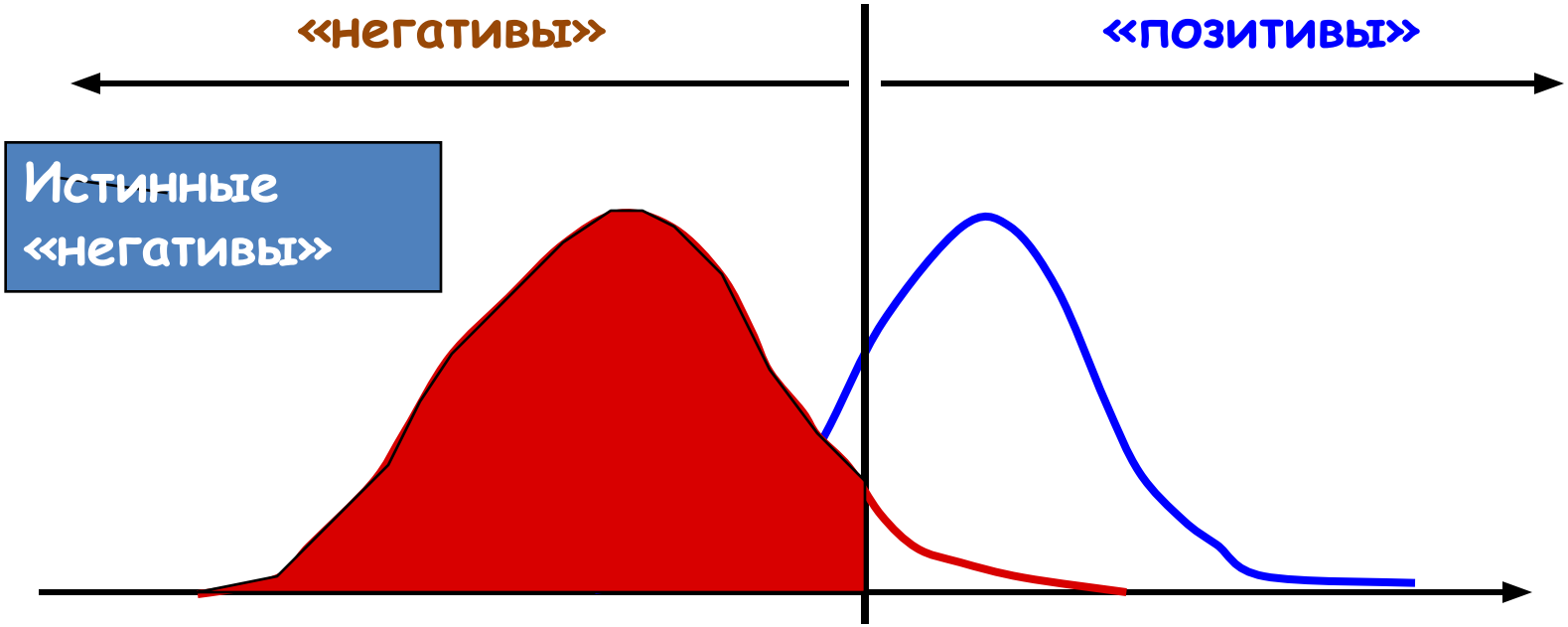
Ложные «позитивы»



Значения мерного диагностического признака

Субъекты без болезни **Субъекты с болезнью**

Истинные «негативы»

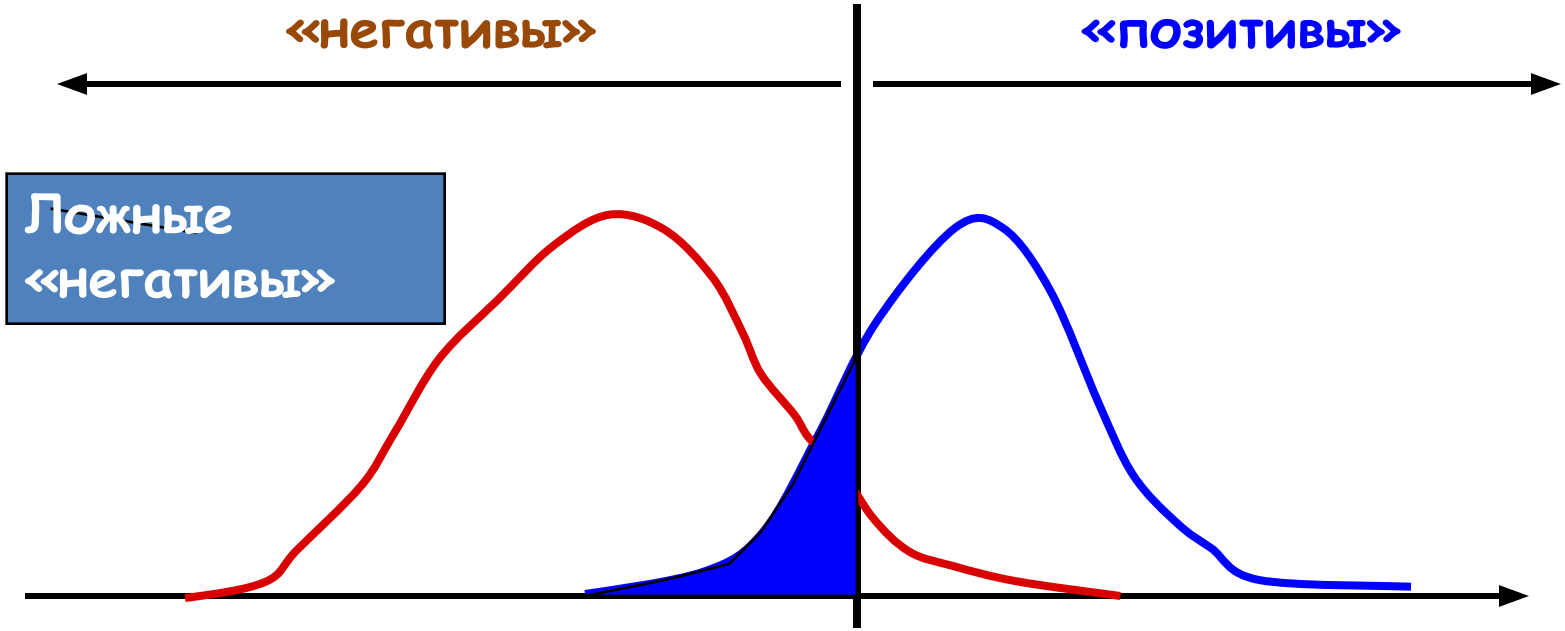


Значения мерного диагностического признака

Субъекты без болезни

Субъекты с болезнью

Ложные «негативы»



Значения мерного диагностического признака

Субъекты без болезни

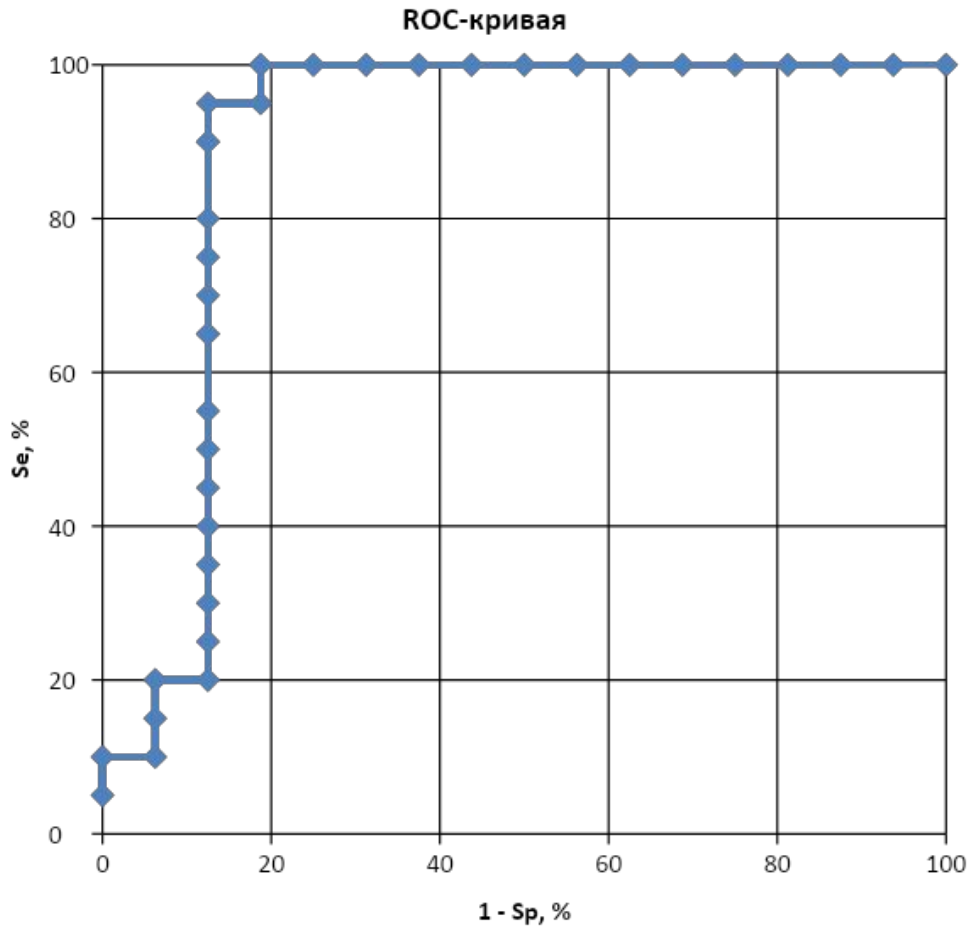
Субъекты с болезнью

Операционная характеристика приёмника

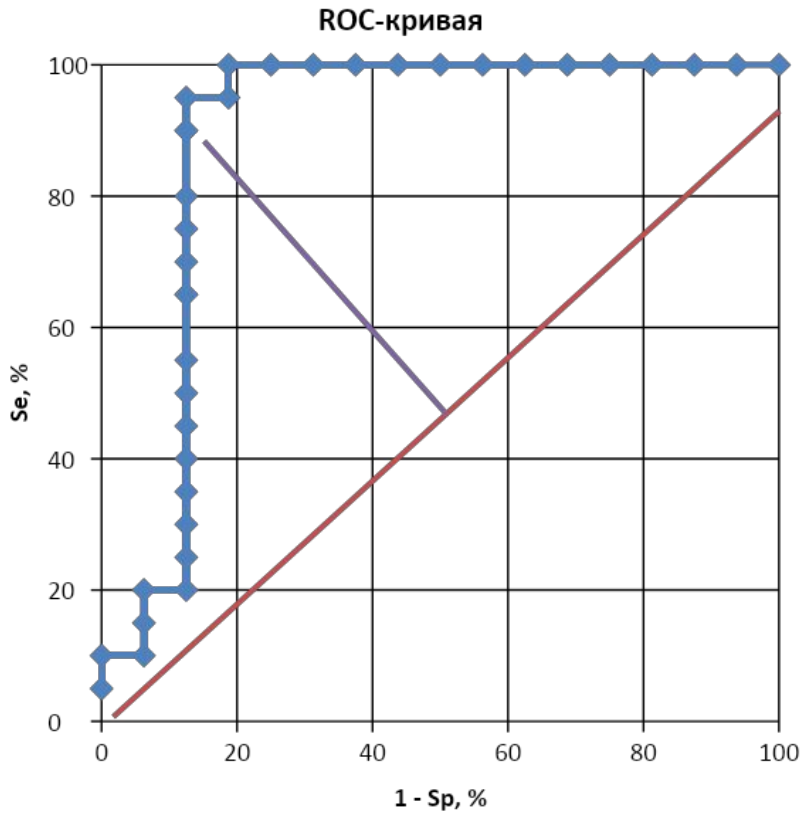
- **Термин *операционная характеристика приёмника* (Receiver Operating Characteristic, ROC) пришёл из теории обработки сигналов,**
- **Эту характеристику впервые ввели во время II мировой войны, после поражения американского военного флота в Пёрл Харборе в 1941 году, когда была осознана проблема повышения точности распознавания самолётов противника по радиолокационному сигналу.**
- **Позже нашлись и другие применения: медицинская диагностика, приёмочный контроль качества, кредитный скоринг, предсказание лояльности клиентов, и т.д.**

- **ROC-кривая**
- **– графическая характеристика качества диагностического теста,**
- **зависимость доли истинных позитивов среди субъектов с болезнью:**
  - $Se = f(T+|D+) = f(T+,D+)/f(D+)$
- **от доли ложных позитивов среди субъектов с болезнью:**
  - $(1 - Sp) = f(T+|D-) = f(T+,D-)/f(D+)$
- **при варьировании порога отсечения для распознавания наличия или отсутствия болезни.**

ROC-кривая для данных о содержании цитокина у матерей здоровых детей и детей с СЗРП. Программа AtteStat http://attestatsoft.narod.ru/



Графическая интерпретация порога отсечения на ROC-кривой для данных о содержании цитокина у матерей здоровых детей и детей с СЗРП



- **Порог отсечения *Tr* есть такое значение мерного диагностического признака, для которого расстояние от диагонали на ROC-кривой является максимальным.**
- **В данном случае это точка, для которой**
- ***Se* = 0,95 и *Sp* = 0,88**

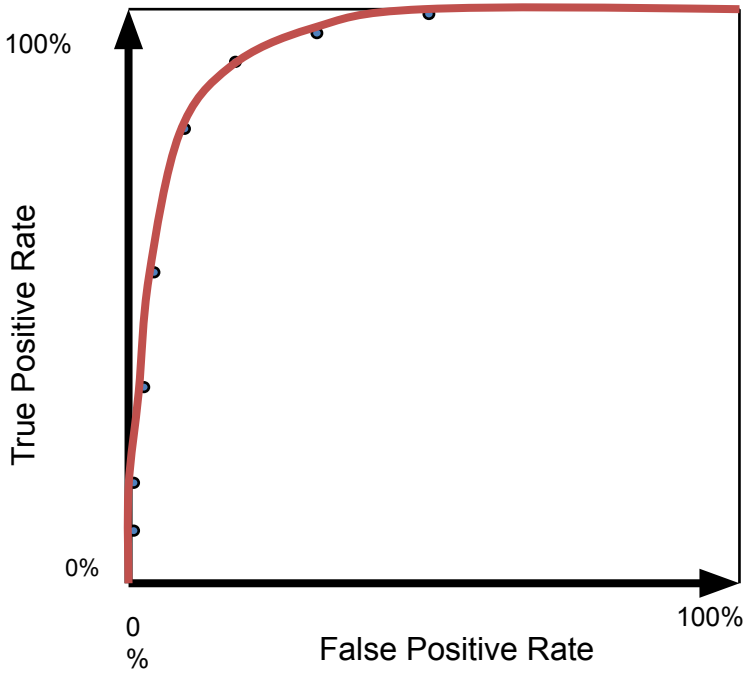
Нахождение оптимального порога отсечения, $Tr_0 = 121$

Порог, *Tr*	*Se + Sp*	Порог, *Tr*	*Se + Sp*	Порог, *Tr*	*Se + Sp*
173	1,05	140	1,43	101	1,69
171	1,1	134	1,30	94	1,63
170	1,04	134	1,53	93	1,56
168	1,09	132	1,58	92	1,50
162	1,14	130	1,30	86	1,44
159	1,08	127	1,68	81	1,38
153	1,13	123	1,68	71	1,31
151	1,18	123	1,79	70	1,25
149	1,23	$Tr_0 = 121$	1,83	59	1,19
147	1,28	115	1,76	58	1,13
146	1,33	104	1,81	42	1,06
144	1,38	103	1,75	38	1,00

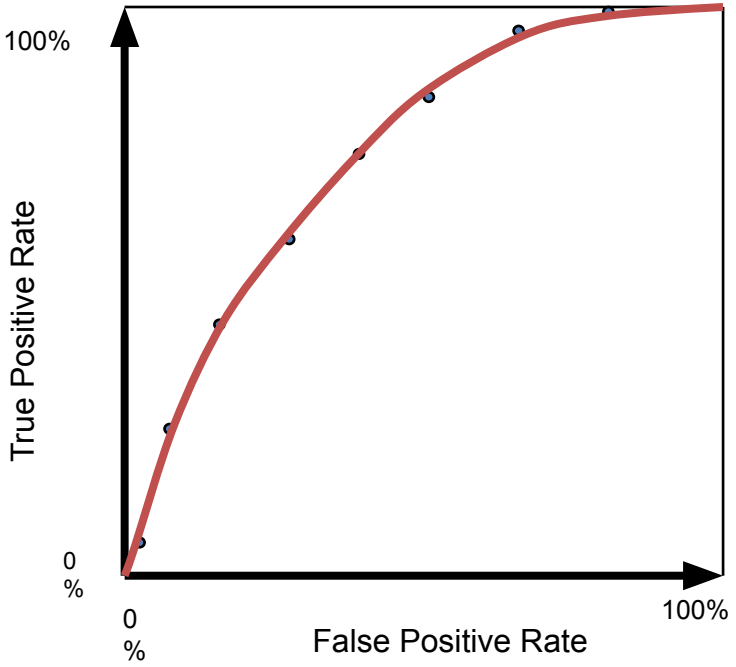
- **Решающее правило:**
- **Значения признака, превышающие порог Tr_0 = 121 или равные ему, принимаются за положительный результат диагностического теста.**
- **Значения признака ниже порога Tr_0 = 121 принимаются за отрицательный результат диагностического теста.**

Сравнение ROC-кривых

Хороший тест:

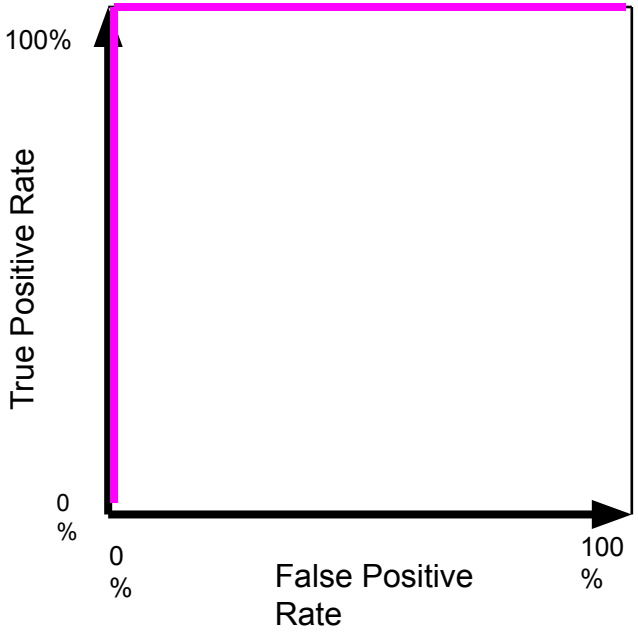


Посредственный тест:



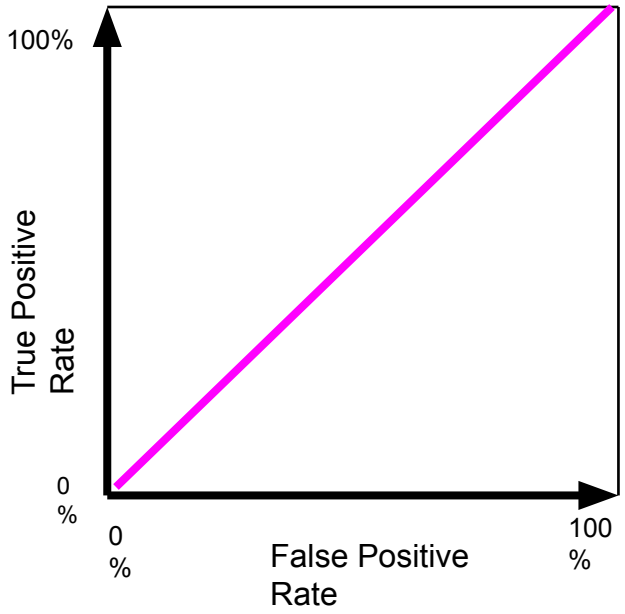
Предельные варианты ROC-кривых

Наилучший тест:



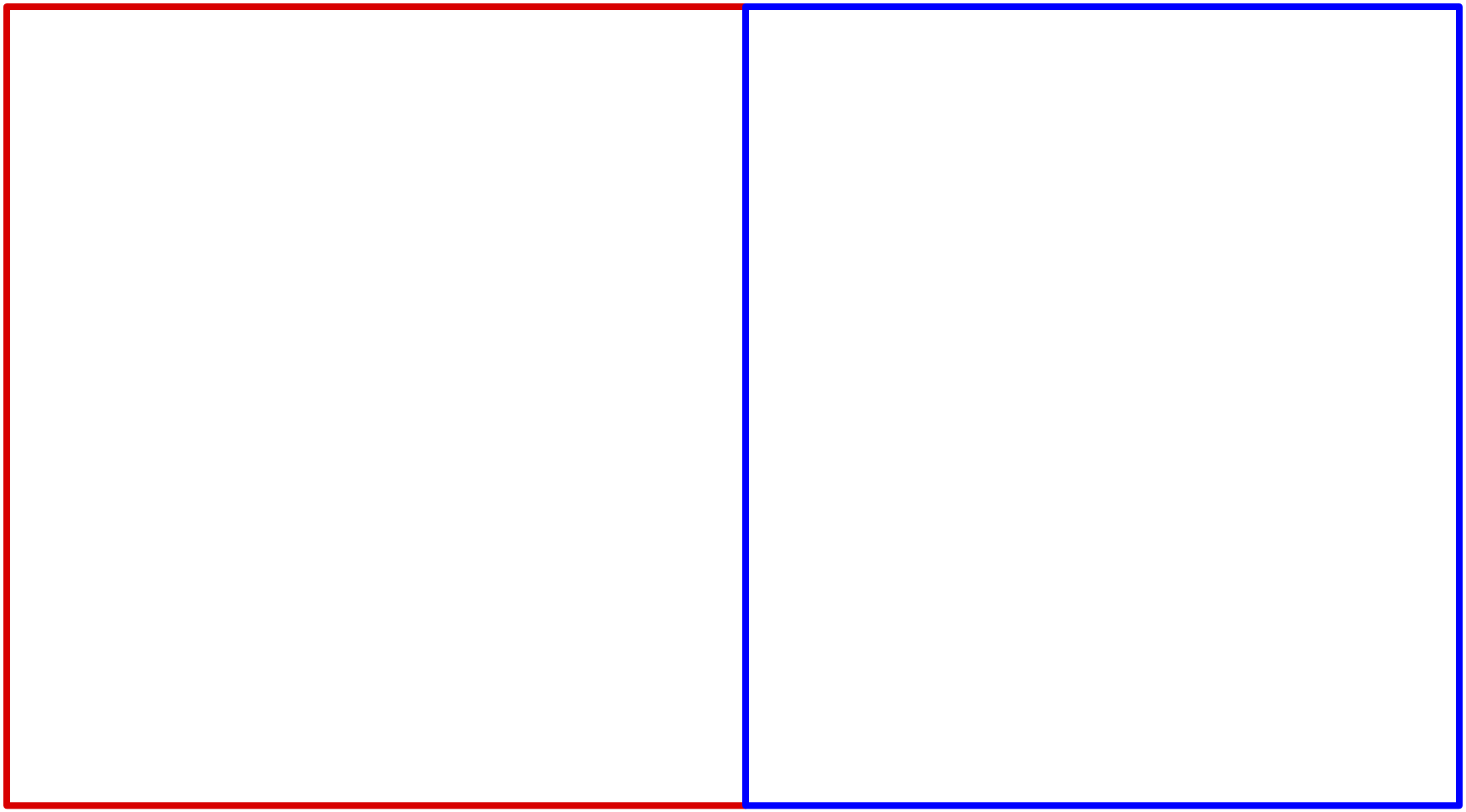
Распределения значений мерного признака не пересекаются вовсе

Наихудший тест:

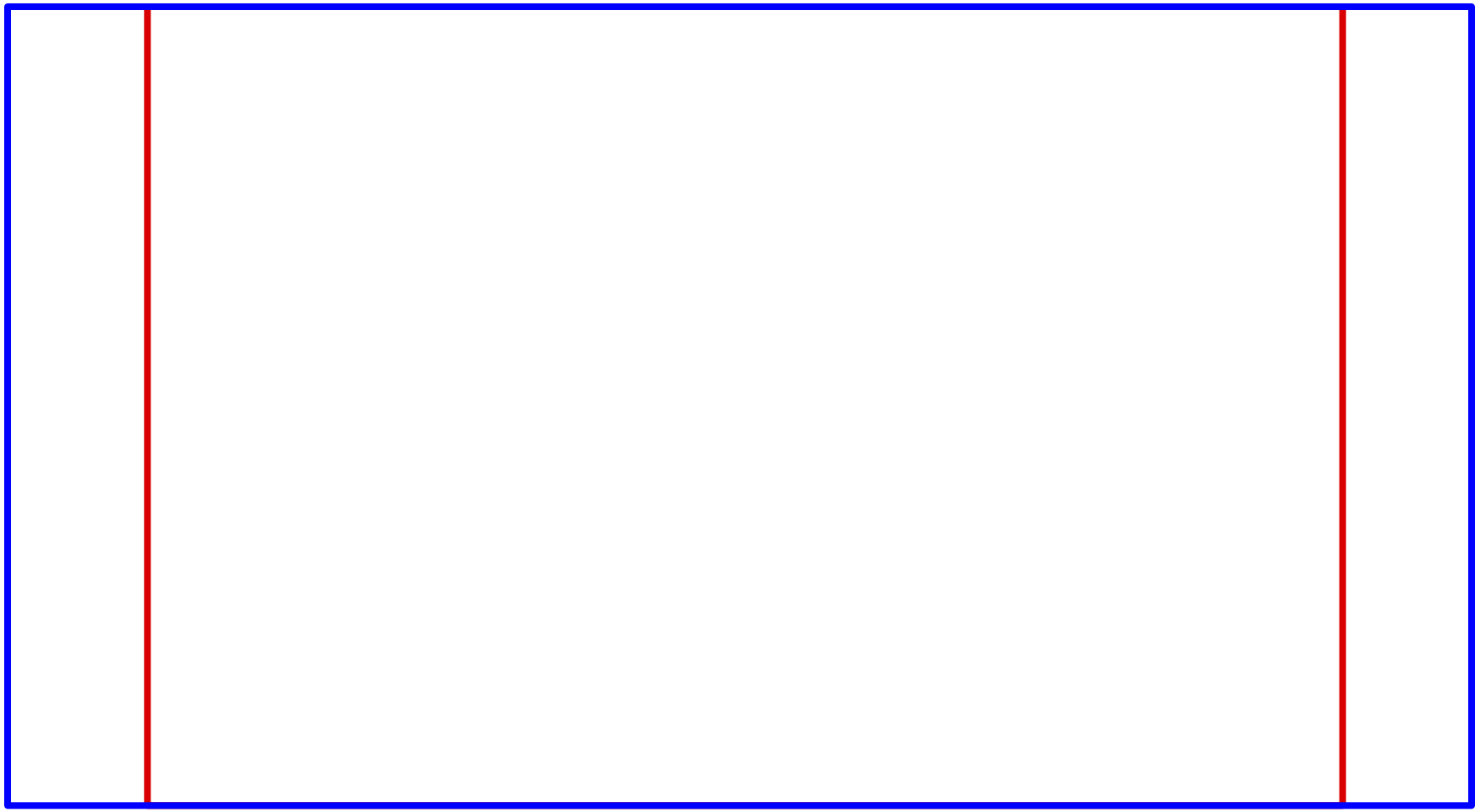


Распределения значений мерного признака полностью совпадают

Наилучший тест: распределения значений мерного диагностического признака в двух группах не перекрываются



Наихудший тест: распределения значений мерного диагностического признака в двух группах полностью перекрываются



Результаты ROC-анализа

- **Оптимальный порог отсечения: *Tr* = 121**
- ***AUC* = $_{0,75}0,89_{1,00}$**
- **Указаны границы 99%-го ДИ для *AUC*.**

- **Чувствительность: *Se* = 0,95**
- **Специфичность: *Sp* = 0,88**

«Площадь под кривой»

- *AUC* (Area Under Curve)
- - площадь под ROC-кривой - полезный обобщенный показатель качества диагностического теста.
- Чем больше значение *AUC*, тем «лучше» способность диагностического теста распознавать наличие и отсутствие болезни,
- Кроме того, данный показатель удобно использовать для сравнительного анализа нескольких методов диагностики.

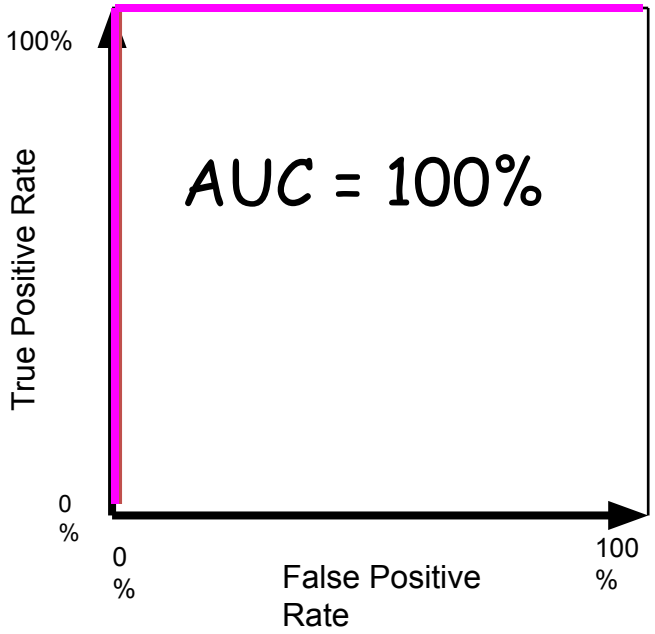
Идеальный, бесполезный и абсурдный тесты в терминах *AUC*

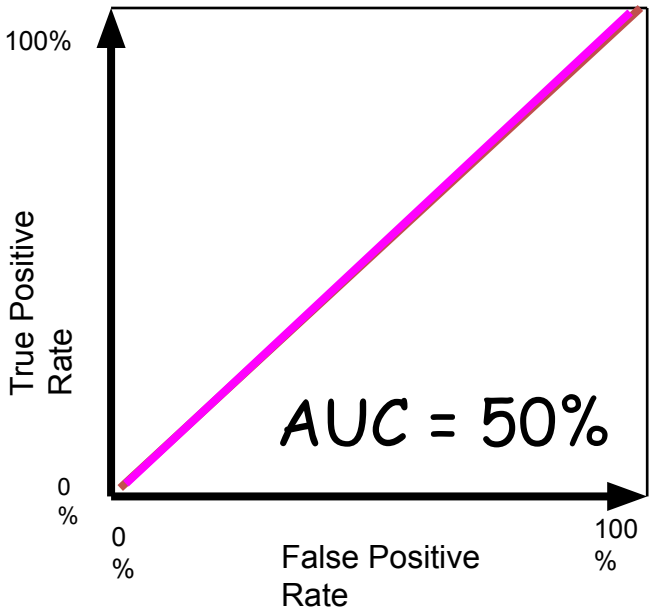
- Если тест идеальный, то
  - *AUC* = 1.
- Если
  - *AUC* = 0,5,
- то тест бесполезен.
- Если
  - *AUC* < 0,5,
- то тест следует признать абсурдным или даже «вредным».

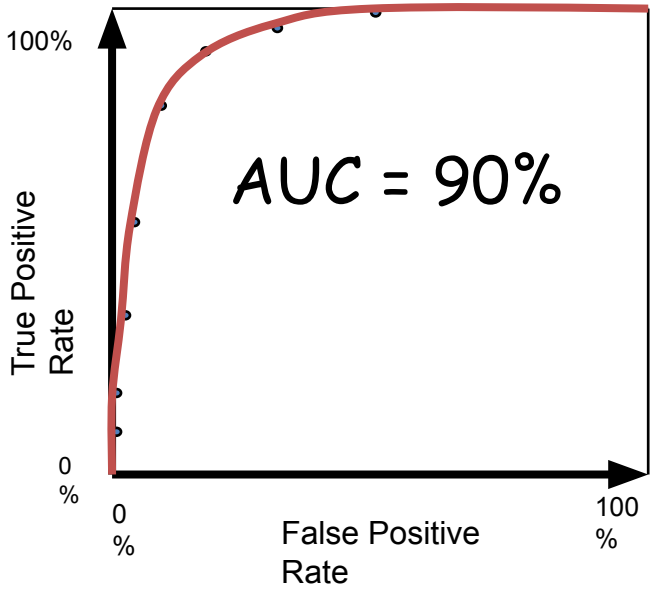
Словесные интерпретации для градаций *AUC*

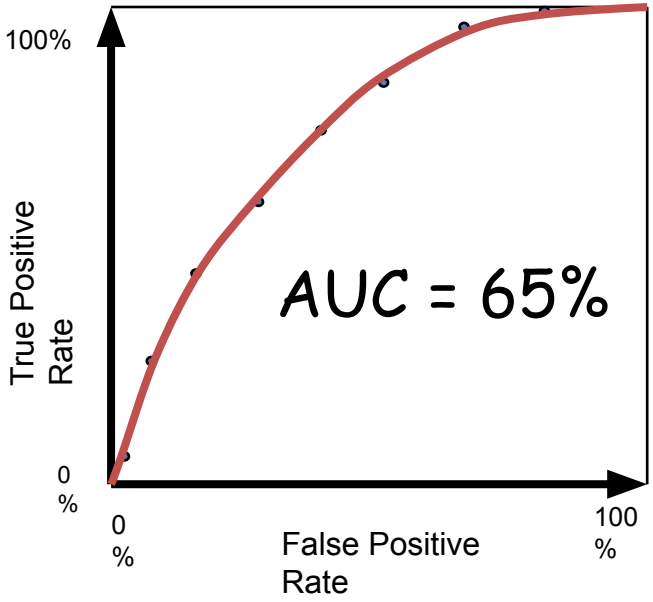
Интервал *AUC*	Способность диагностического теста распознавать наличие или отсутствие болезни
1,0 – 0,9	Отличная
0,8 – 0,9	Хорошая
0,7 – 0,8	Удовлетворительная
0,6 – 0,7	Посредственная
0,5 – 0,6	Неудовлетворительная
< 0,5	Абсурдная («вредная»)

AUC для ROC-кривых









Обсуждение результатов

- 99%-й ДИ для *AUC* = $_{0,75}0,89_{1,00}$ не накрывает неинформативное значение *AUC* = 0,50.
- Следовательно, оцениваемое значение *AUC* статистически значимо отличается от бесполезного (неинформативного) значения 0,5 на уровне значимости $\alpha = 0,01$.
- Однако с практической точки зрения способность проверяемого диагностического теста распознавать наличие или отсутствие болезни следует признать всего лишь удовлетворительной, поскольку нижняя граница 99%-го ДИ для $AUC_L = 0,75$ не выходит за границы соответствующего интервала (0,7 – 0,8).

Результирующая таблица 2×2

Тест: цитокин, у.е.	СЗРП		Всего
	есть	нет	
≥ 121	19	2	21
< 121	1	14	15
Всего	20	16	36

Обсуждение результатов

- $Se = {}_{0,78}0{,}95_{0,99}$
- $Sp = {}_{0,66}0{,}88_{0,93}$

- 99%-ые ДИ и для *Se* и для *Sp* не накрывают неинформативные значения *Se* = 0,5 и *Sp* = 0,5.
- Следовательно, оцениваемые значения этих параметров статистически значимо отличаются от указанных неинформативных значений.
- Поскольку нижняя граница 99%-го ДИ для *Se п*ревышает значение 0,7, то чувствительность проверяемого диагностического теста следует признать удовлетворительной.
- Для *Sp* нижняя граница 99%-х ДИ не превышает значение 0,7.
- Поэтому специфичность проверяемого диагностического теста следует признать посредственной.

Обсуждение результатов

- $LR[+] = {}_{1,4}7,6_{42}$
- $LR[-] = {}_{0,005}0,057_{0,71}$

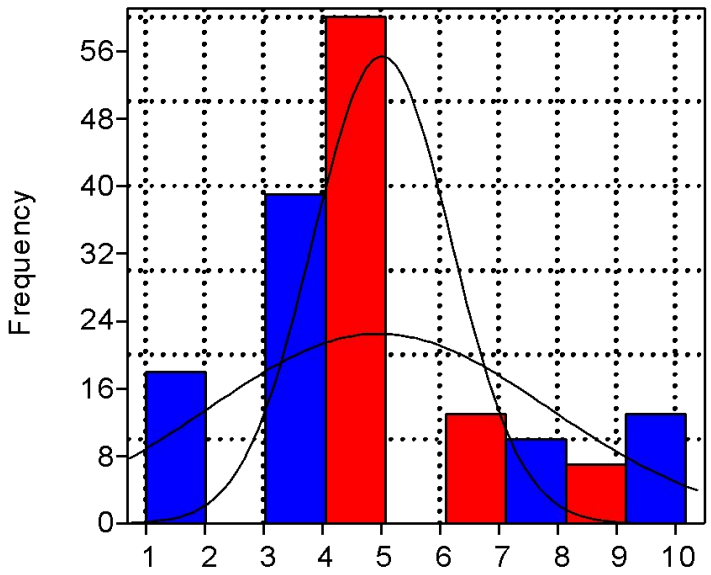
- 99%-ые ДИ и для *LR*[+] и для *LR*[-] не накрывают неинформативные значения *LR*[+] = 1,0 и *LR*[-] = 1,0.
- Следовательно, оцениваемые значения этих параметров статистически значимо отличаются от указанных неинформативных значений.
- Однако нижняя граница 99%-го ДИ для *LR*[+] не превышает значение 3,0, а верхняя граница 99%-го для *LR*[-] превышает значение 0,3.
- Поэтому способность как положительных, так и отрицательных результатов данного диагностического теста распознавать как наличие, так и отсутствие болезни следует признать неудовлетворительными.

Предостережение

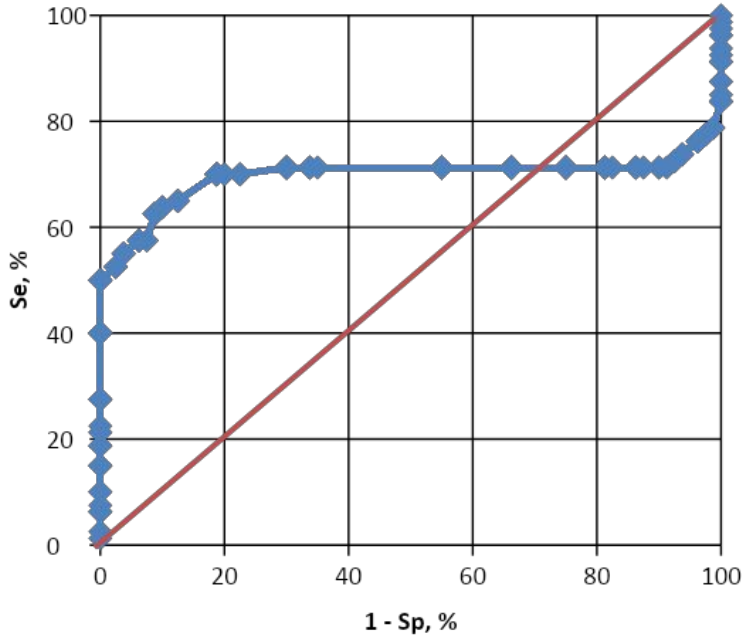
- Подобные исследования следует рассматривать как сугубо предварительные
- (пилотные, разведочные, обучающие).
- Об этом свидетельствуют в частности чрезвычайно широкие доверительные интервалы (ДИ) для оцениваемых параметров.
- Поэтому такие исследования надо обязательно повторить с выборками гораздо большего объема и удостовериться, воспроизводятся ли результаты.

Одно распределение «вложено» в другое: ROC-анализ неприменим

Гистограмма

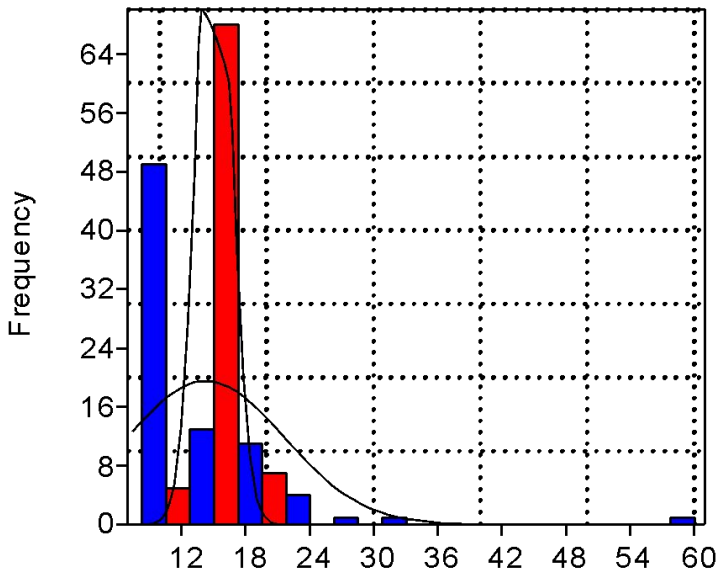


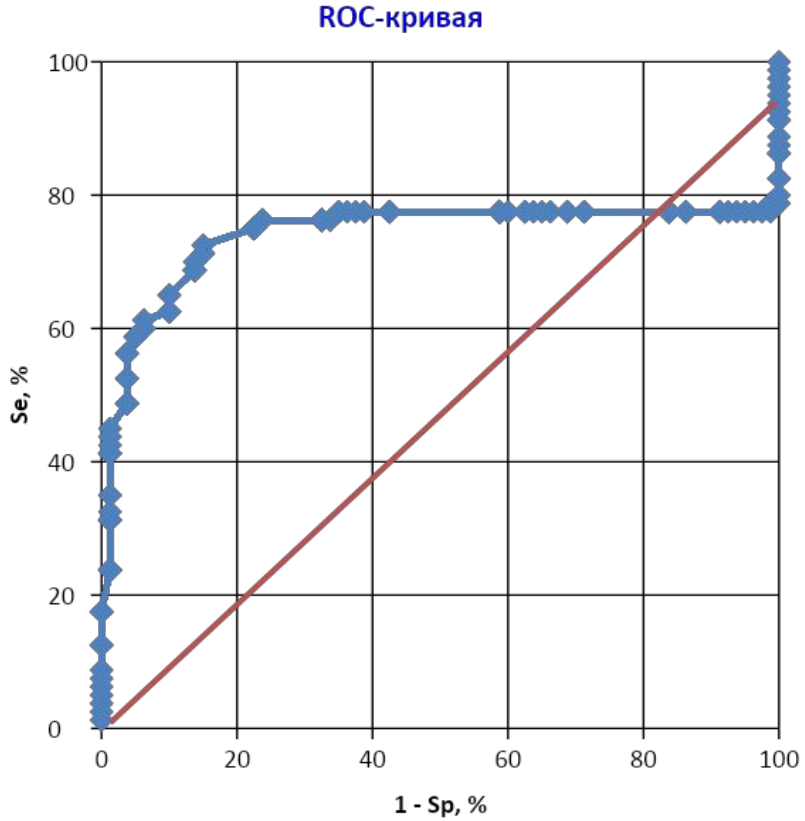
ROC-кривая



Еще пример, когда ROC-анализ неприменим

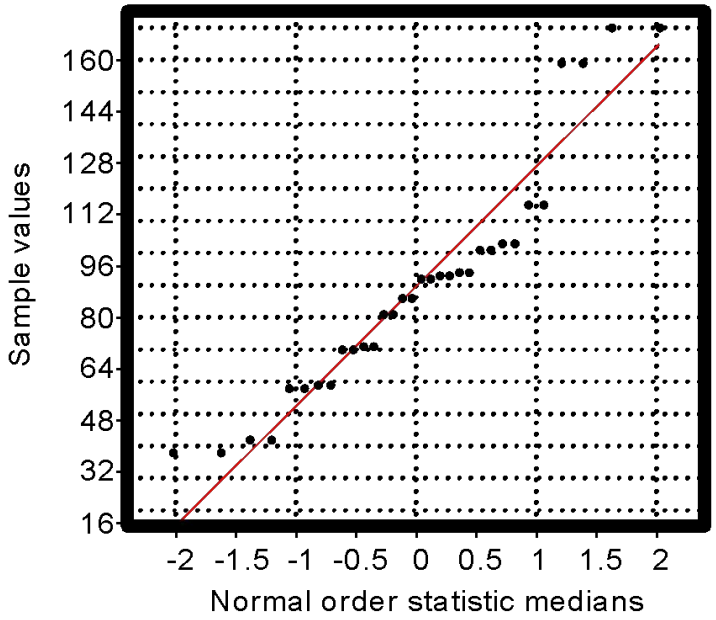
Гистограмма



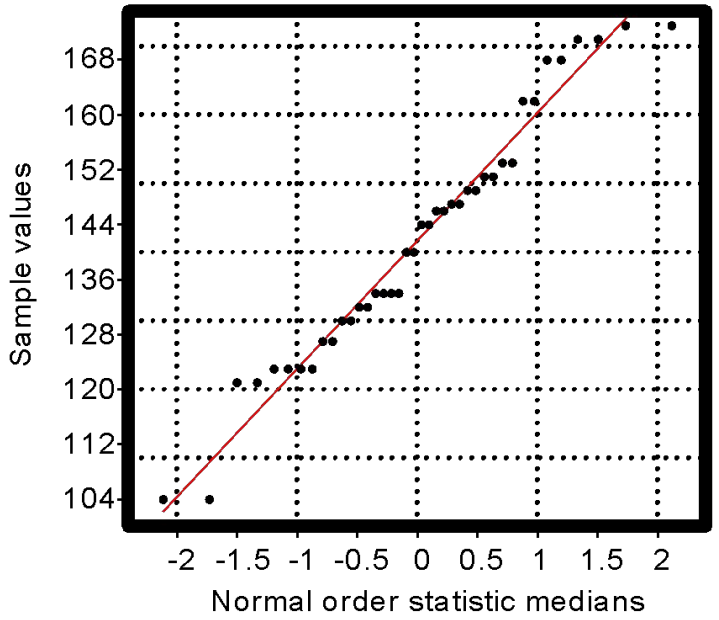


Нормальные вероятностные графики

Здоровые



СЗРП



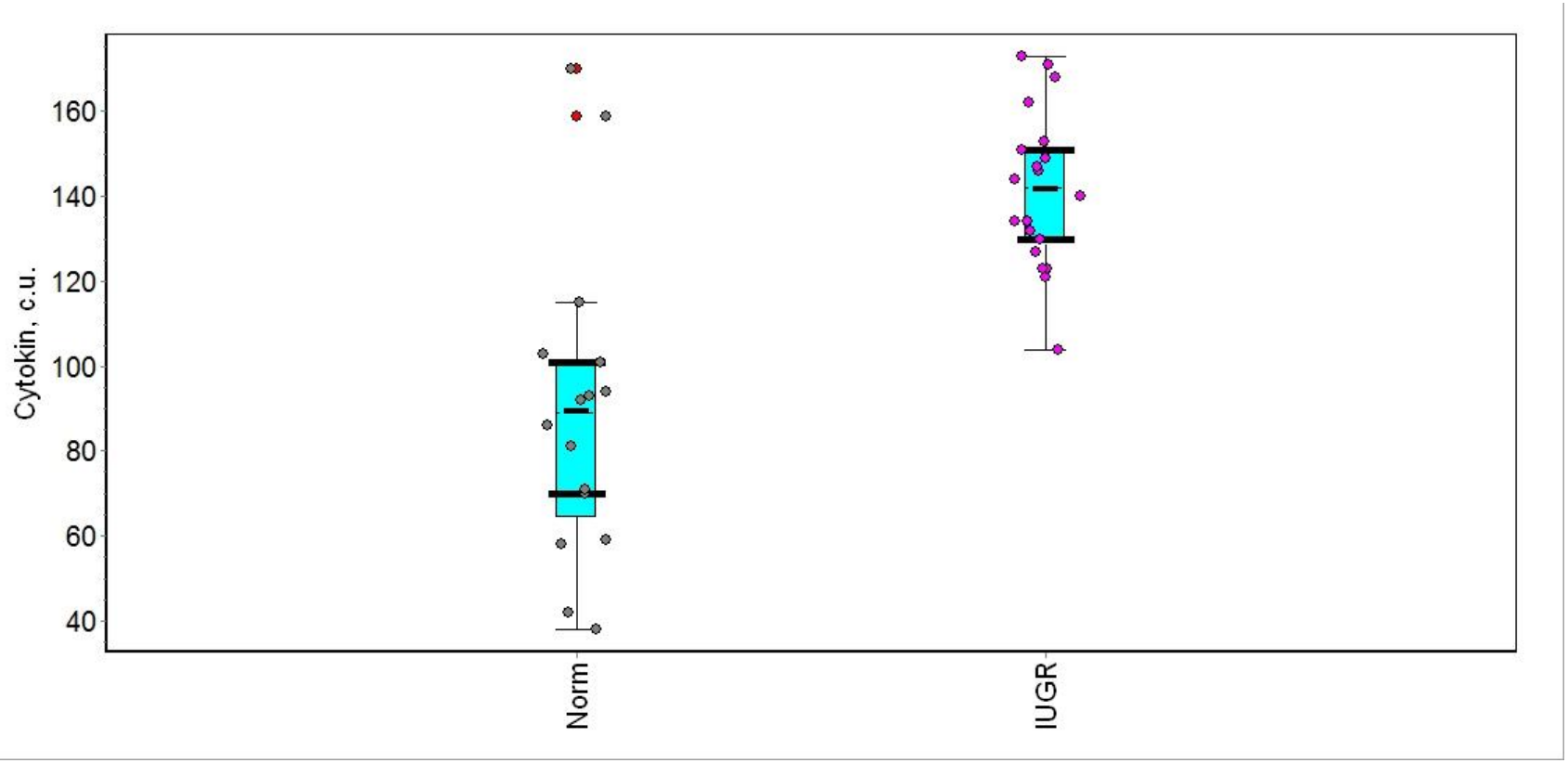
Проверка нормальности (гауссовости) распределения у матерей здоровых детей и детей с СЗРП

Статистический критерий	Наблюдаемые *P*-значения, P_{val}	
	Здоровые	СЗРП
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

Все *P*-значения превышают пороговое значение 0,05. Следовательно у нас нет оснований сомневаться в гипотезе о нормальности распределения, порождающего наблюдаемые данные.

Графики (диаграммы) «короб с усами», программа Instat+

http://www.rdg.ac.uk/ssc/software/instat/instat.html



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, вызывающих сомнения.**
- **В математической статистике, термин «гипотеза» означает предположение, которое не только вызывает сомнения, но и которое мы собираемся в данный момент проверить.**
- **Проверка статистической гипотезы состоит в выяснении того, насколько совместима эта гипотеза с имеющимися данными.**

Проверяемая гипотеза

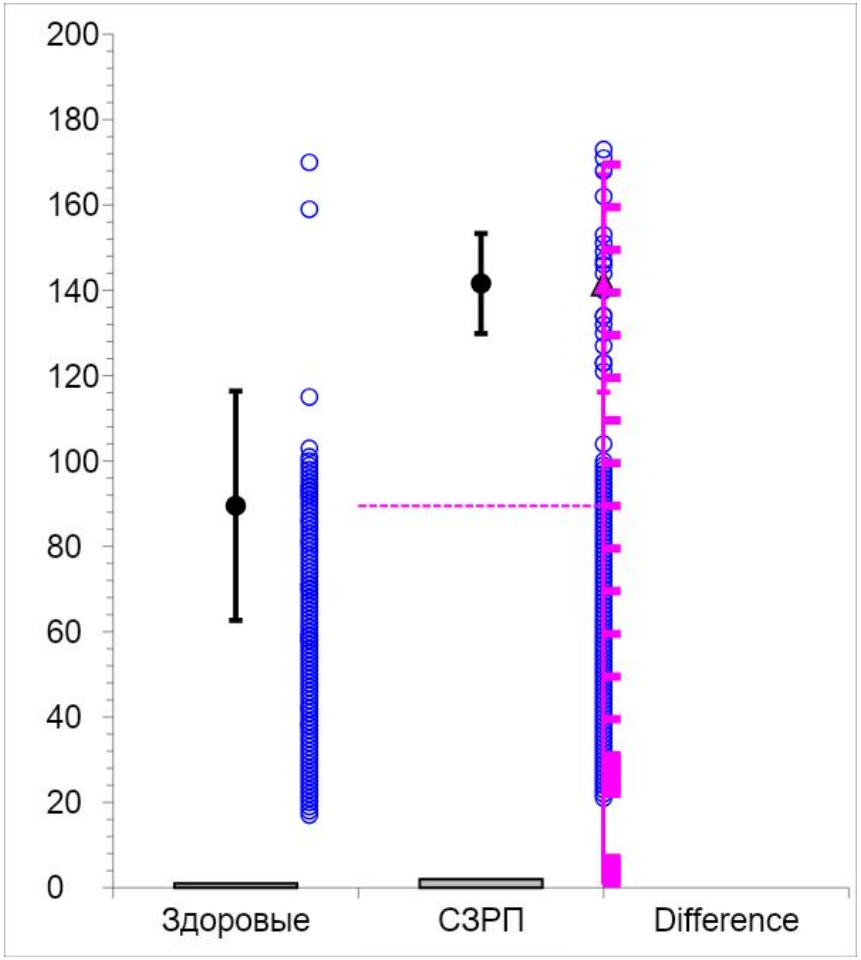
- **В подавляющем большинстве реальных ситуаций проверяемая статистическая гипотеза является гипотезой об отсутствии того или иного эффекта:**
- **об отсутствии различий, например, о равенстве нулю разности средних;**
- **об отсутствии тех или иных эффектов, связей, соответствий, зависимостей и т.п.**
- **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 α.

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

Графическое представление результатов статистического сравнения групп матерей здоровых детей и детей с СЗРП, 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 оцениваемая разность 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*.**

P-значение

- **Для проверки нулевых гипотез с помощью статистических критериев основным приемом является вычисление значения вероятности, которое называется *P*-значением.**

P-значение

- ***P*-значение есть условная вероятность, а именно:**
- **Вероятность получить наблюдаемое значение $t_{\text{набл.}}$ 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:**
  - $P_{val} = \Pr[|T| \geq |t_{\text{набл.}}| \mid H_0]$.
- **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 в рамках выбранной нами (нулевой) модели.**

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

- Когда наблюдаемое *P*-значение 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.

Андрей Николаевич Колмогоров

(урождённый *Катаев*, 12(25).04.1903 — 20.10.1987)

- **Пророк в своем отечестве**

Колмогоров А. Н. Вероятность. ВиМСЭ (1951). С. 97: http://ru.science.wikia.com/wiki/Вероятность_(в_теории_вероятностей)

- **При практическом употреблении вычисленных значений вероятности мы неизбежно приходим к вопросу о том, сколь малыми значениями вероятностей мы можем пренебречь.**
- ***В математической статистике вероятность, которой решено пренебрегать в данном исследовании, называют уровнем значимости.***
- **На практике этот вопрос решается каждый раз по-разному, в зависимости от того, насколько велика необходимость быстрого перехода от накопления надежных данных к их действительному употреблению.**

Колмогоров, 1951, 1956

- *«Норма в 0,05 для серьезных научных исследований явно недостаточна» (1956).*
- *«Хотя в статистике обычно рекомендуют пользоваться уровнями значимости от 0,05 при предварительных ориентировочных исследованиях и до 0,001 при окончательных серьезных выводах, часто достижима значительно большая достоверность [статистическая значимость – НХ] вероятностных выводов.*
- *Например, основные выводы статистической физики основаны на пренебрежении лишь вероятности порядка меньшего 0, 000 000 000 1 ($<10^{-10}$)» (1951).*
- **Воспроизведено в: Колмогоров А. Н. В кн.: Вероятность и математическая статистика. Энциклопедия / Гл. ред. Ю. В. Прохоров. — М.: Изд-во «Большая Российская Энциклопедия», 1999. – с. 97 и 975.**

- **В модных ныне изысканиях различного рода генетических предрасположенностей, когда проверяются миллионы аллелей различных генов, исследователи ориентируются на *P*-значения порядка**
- **10^{-7}.**
- **При таком уровне значимости приходится обследовать сотни тысяч людей.**
- **Но даже при столь суровой требовательности результаты далеко не всегда воспроизводятся в повторных проверочных исследованиях.**

«Фильтруйте базар»: Sterne J.A.C., Davey Smith G. Sifting the evidence – what's wrong with significance tests? *BMJ,* 2001. – Vol. 322. – P. 227-231.

- **В наши дни Колмогорову вторят зарубежные авторы:**
- ***P-значение* близкое к 0,05 не является сильным свидетельством против нулевой гипотезы.**
- **Сильными свидетельствами против H_0 следует признавать значения $P < 0{,}001$.**
- **В публикациях надо представлять точные *P*-значения без соотнесения их с какими-либо пороговыми (критическими) значениями (типа 0,05).**
- **Наравне с *P*-значениями нужно указывать доверительные интервалы.**

Традиционная интерпретация *P*-значений (шкала Michelin)

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

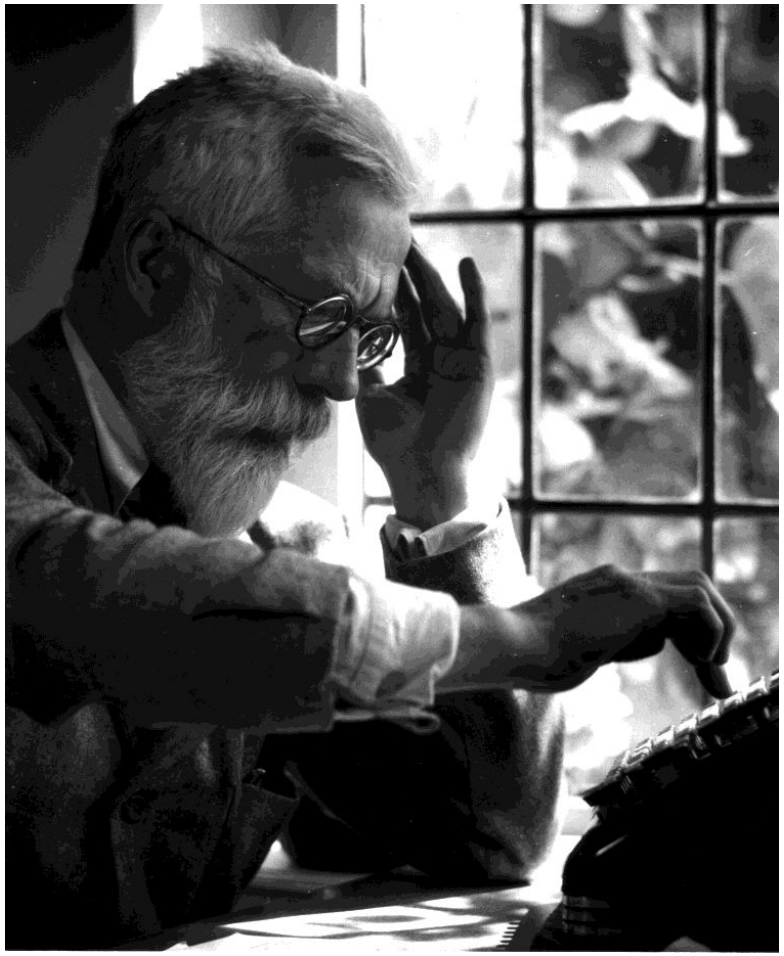
Глотов Н.В., Животовский Л.А., Хованов Н.В., Хромов-Борисов Н.Н. Биометрия, Л.: Изд-во ЛГУ, 1982. – 264 с.

- **Выбор уровня значимости определяется важностью биологических выводов, которые должен сделать экспериментатор.**
- **В настоящее время многие биометрики склоняются к следующему правилу:**
- **а) если $P > 0{,}05$, то принимается нулевая гипотеза;**
- **б) если $P < 0{,}01$, то нулевая гипотеза отклоняется и принимается конкурирующая;**
- **в) если $0{,}01 < P < 0{,}05$, то результат считается неопределенным.**

[0,05; 0,01] – «серая зона»

P-значение	Статистическая значимость	Шкала Мишлена
> 0,05	Незначимо	
От 0,05 до 0,01	Неопределенно	*
От 0,01 до 0,001	Значимо	**
< 0,001	Высоко значимо	***

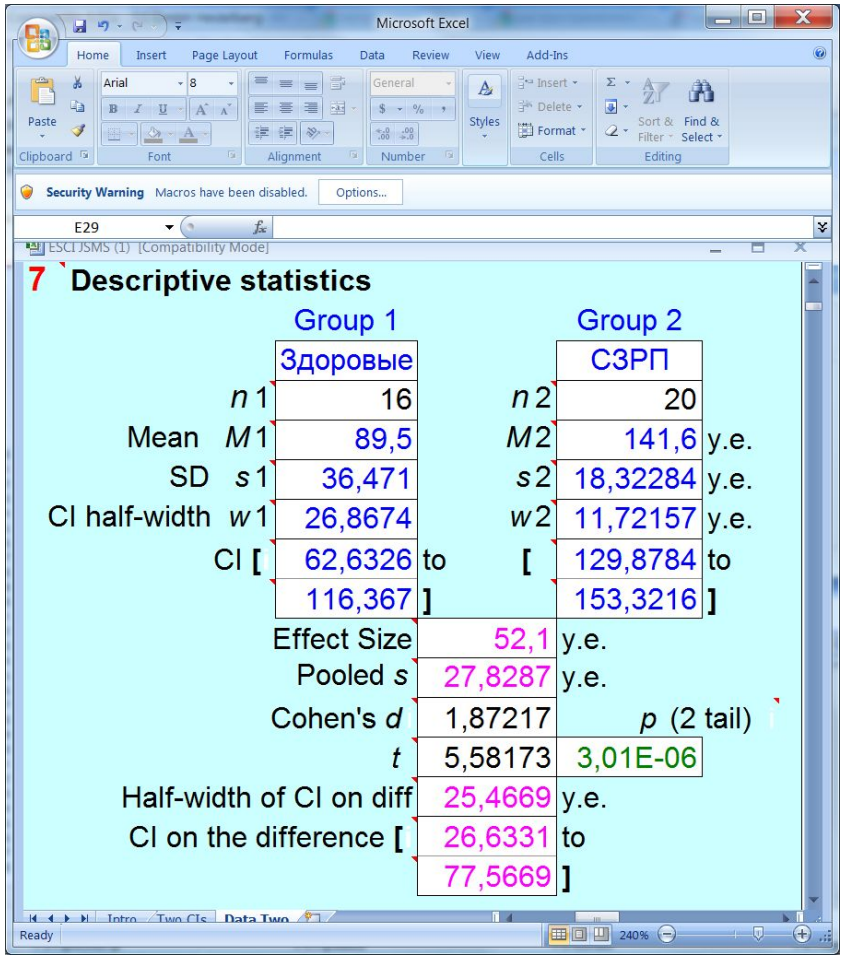
Sir Ronald Aylmer Fisher
17.02.1890 – 29.07.1962



Пожелание: «гибкие» *P*-значения

- ***«В действительности ни один исследователь не пользуется фиксированным уровнем значимости с которым из года в год и при любых обстоятельствах он отвергает нулевые гипотезы.***
- ***Он больше доверяет своему уму и каждый конкретный случай рассматривает в свете совокупности имеющихся доказательств и своих идей и представлений».***
- ***R. A. Fisher R. A. Statistical Methods and Scientific Inference, 1956***

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



- **Основная логика использования *P*-значений состоит в том, что если оно мало́, 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_0.**
- **В данном случае $P_{val} = 3 \cdot 10^{-6}$.**
- **Вывод: различие в содержании цитокина у матерей здоровых детей и детей с СЗРП статистически высоко значимо; во второй группе оно выше, чем в первой.**

Акт интеллектуальной смелости

- Когда *P*-значение 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

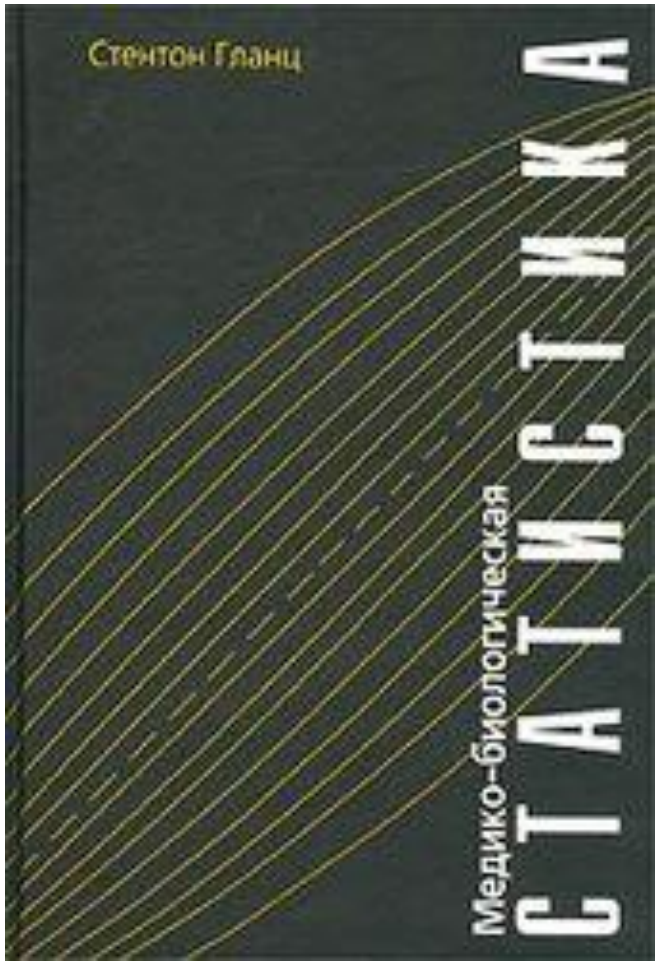
- **Квинтэссенцию традиционных (частотнических) заключений при проверке статистических гипотез принято интерпретировать так:**
- ***чем меньше P-значение, тем весомее доводы против нулевой гипотезы H_0, которые предоставляют нам имеющиеся данные; тем больше у нас оснований сомневаться в H_0.***
- **Отсюда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

Распространенное заблуждение

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

P-значение не есть вероятность нулевой гипотезы!

- **К сожалению, даже в известной книге С.Гланца можно встретить утверждение:**
- **«Упрощая, можно сказать, что *P — это вероятность справедливости нулевой гипотезы*»**
- **Гланц С. Медико-биологическая статистика. — М.: Практика, 1998. — с. 119.**
- **Это мнение глубоко ошибочно и чревато пагубными последствиями.**
- **К чести автора, в последующих (у нас не переведенных) изданиях этой его книги оно отсутствует.**



Калибровка *P*-значения

- Sellke T., Bayarri M.J., Berger J.O.
- Calibration of p Values for Testing Precise Null Hypotheses
- *The American Statistician*, Vol. 55, No. 1. (2001), pp. 62-71.
- При

$$p < 1/e$$

$$P(H_0|D) \geq \left[1 + \frac{1}{-e\,p \ln p}\right]^{-1}$$

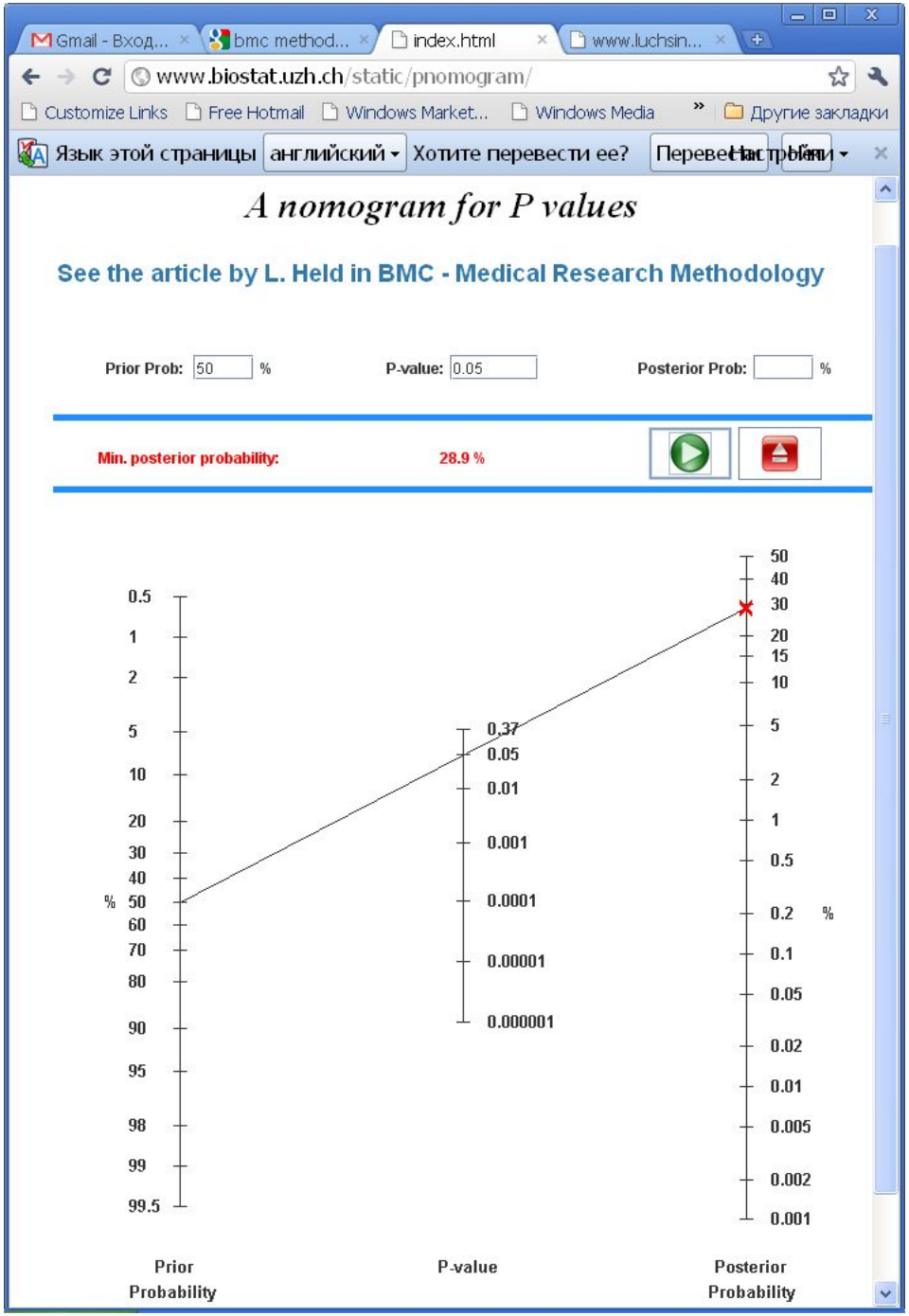
Калибровка *P*-значений

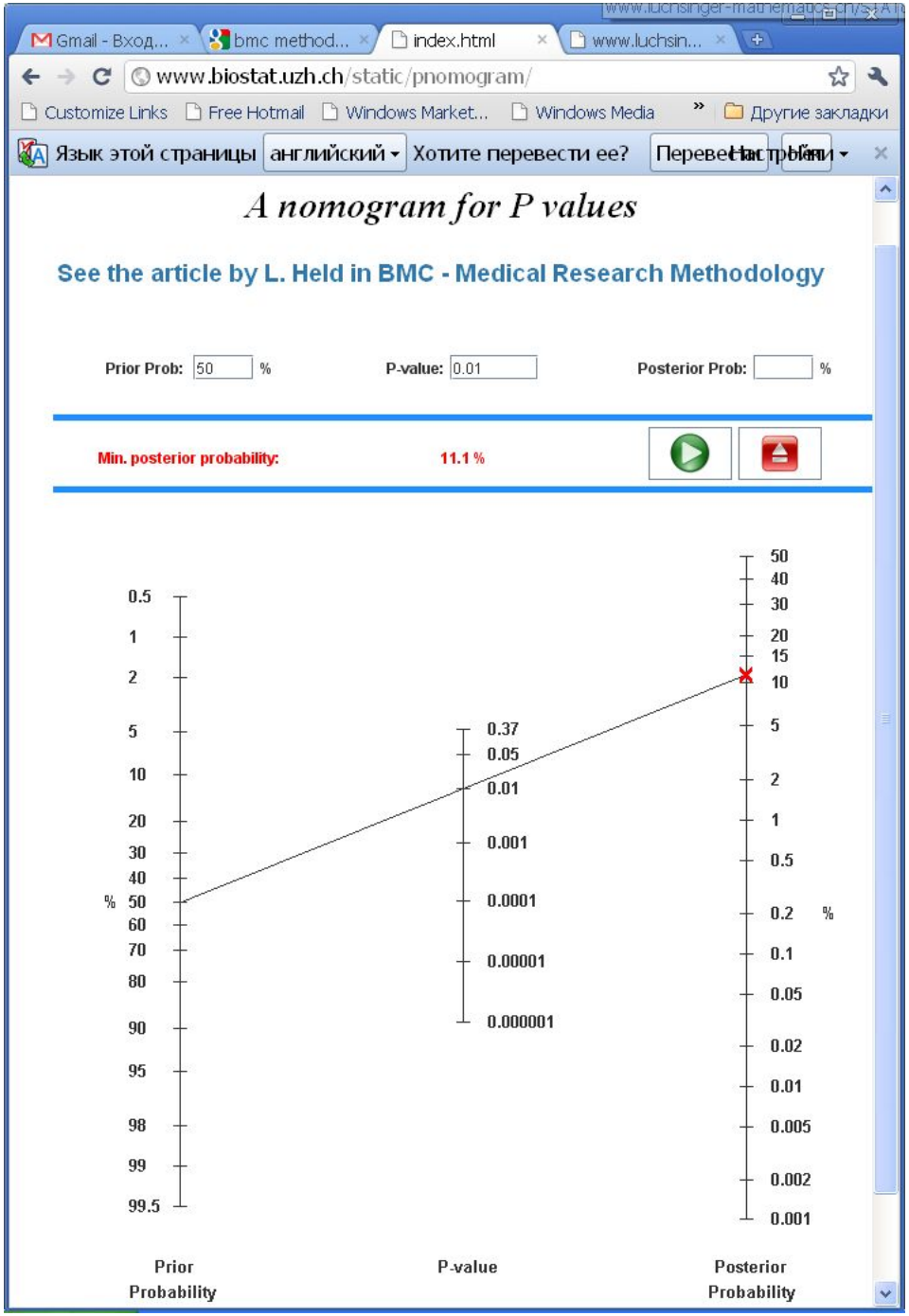
Held L. A nomogram for P values.

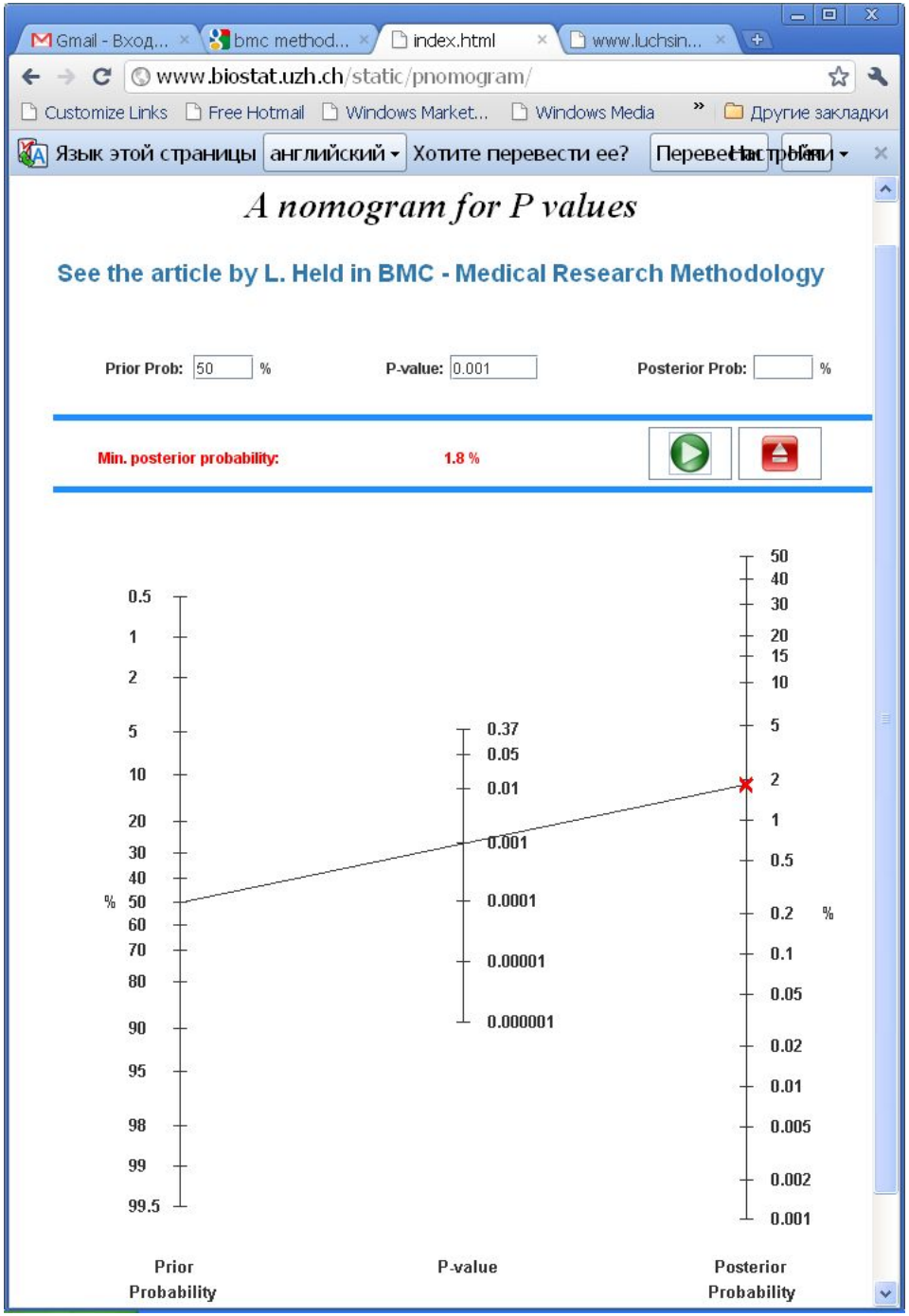
***BMC Medical Research Methodology* 2010, 10:21**

doi:10.1186/1471-2288-10-21

http://www.biostat.uzh.ch/static/pnomogram/


Gmail - Вход...
bmc method...
index.html
www.luchsin...
www.biostat.uzh.ch/static/pnomogram/
Customize Links
Free Hotmail
Windows Market...
Windows Media
Другие закладки
Язык этой страницы
английский
Хотите перевести ее?
A nomogram for P values
See the article by L. Held in BMC - Medical Research Methodology
Prior Prob: 50 %
P-value: 0.05
Posterior Prob: %
Min. posterior probability: 28.9 %
0.5
1
2
5
10
20
30
40
% 50
60
70
80
90
95
98
99
99.5
0.37
0.05
0.01
0.001
0.0001
0.00001
0.000001
50
40
30
20
15
10
5
2
1
0.5
0.2 %
0.1
0.05
0.02
0.01
0.005
0.002
0.001
Prior Probability
P-value
Posterior Probability


Gmail - Вход...
bmc method...
index.html
www.luchsin...
www.biostat.uzh.ch/static/pnomogram/
Customize Links
Free Hotmail
Windows Market...
Windows Media
Другие закладки
Язык этой страницы
английский
Хотите перевести ее?
A nomogram for P values
See the article by L. Held in BMC - Medical Research Methodology
Prior Prob: 50 %
P-value: 0.01
Posterior Prob: %
Min. posterior probability: 11.1 %
0.5
1
2
5
10
20
30
40
% 50
60
70
80
90
95
98
99
99.5
0.37
0.05
0.01
0.001
0.0001
0.00001
0.000001
50
40
30
20
15
10
5
2
1
0.5
0.2 %
0.1
0.05
0.02
0.01
0.005
0.002
0.001
Prior Probability
P-value
Posterior Probability


Gmail - Вход...
bmc method...
index.html
www.luchsin...
www.biostat.uzh.ch/static/pnomogram/
Customize Links
Free Hotmail
Windows Market...
Windows Media
Другие закладки
Язык этой страницы
английский
Хотите перевести ее?
A nomogram for P values
See the article by L. Held in BMC - Medical Research Methodology
Prior Prob: 50 %
P-value: 0.001
Posterior Prob: %
Min. posterior probability: 1.8 %
0.5
1
2
5
10
20
30
40
% 50
60
70
80
90
95
98
99
99.5
0.37
0.05
0.01
0.001
0.0001
0.00001
0.000001
50
40
30
20
15
10
5
2
1
0.5
0.2 %
0.1
0.05
0.02
0.01
0.005
0.002
0.001
Prior Probability
P-value
Posterior Probability

«Цена» *P*-значения

P-значение	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%.

Chow SC, Shao J, and Wang, H. Sample Size Calculations in Clinical Research. Second edition, Chapman Hall/CRC Press, Taylor & Francis, New York, New York. P. 6, Table 1.1.2.

Бейзовская интерпретация *P*-значения

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

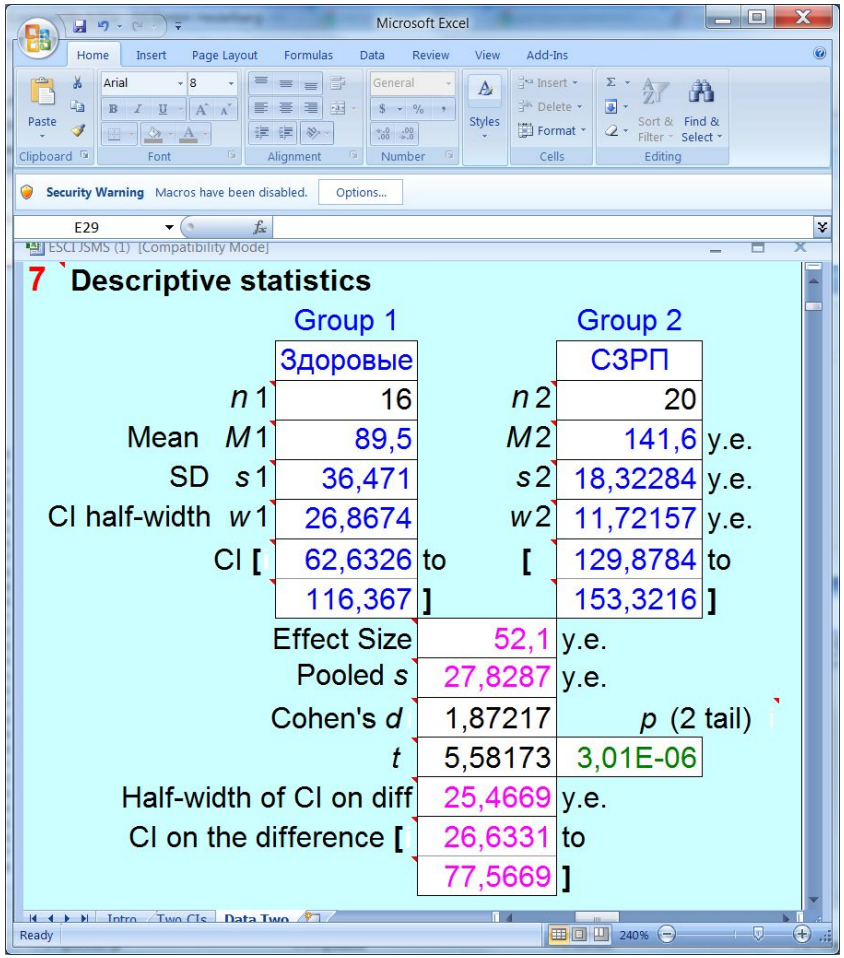
$$d_C = \frac{M_1 - M_2}{s_{pooled}}$$

Интерпретация стандартизированного размера эффекта d_c

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

Размер эффекта, d_c	Градация эффекта
0 – 0,2	Ничтожный
0,2 – 0,6	Малый (слабый)
0,6 – 1,2	Умеренный
1,2 – 2,0	Большой (сильный)
2,0 – 4,0	Очень большой
4,0 - ∞	Абсолютный

Результаты статистического сравнения групп матерей здоровых детей и детей с СЗРП, (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:**
  - $ES = {}_{26,6}52{,}1_{77,6}$ **у.е.**
- **Стандартизированный размер эффекта по Коуэну:**
- $d_c = 1{,}87$
- **Его можно интерпретировать как сильный (большой).**

Бейзов фактор, *BF*

- **Бейзов фактор** – это показатель того, насколько хорошо две гипотезы могут предсказать данные.
- Гипотеза, которая предсказывает наблюдаемые данные лучше – это та из них, которая имеет больше свидетельств в свою пользу.
- Бейзов фактор *BF* принципиально отличается от *P*-значения.
- Бейзов фактор не является вероятностью сам по себе, а является отношением вероятностей, и он может варьировать от нуля до бесконечности.
- Он требует две гипотезы, тем самым четко указывая, что если есть свидетельства против нулевой гипотезы, то должны существовать свидетельства и в пользу альтернативной гипотезы.

$$BF_{01} = P\{D|H_0\} / P\{D|H_1\}$$

$$BF_{10} = P\{D|H_1\} / P\{D|H_0\}$$

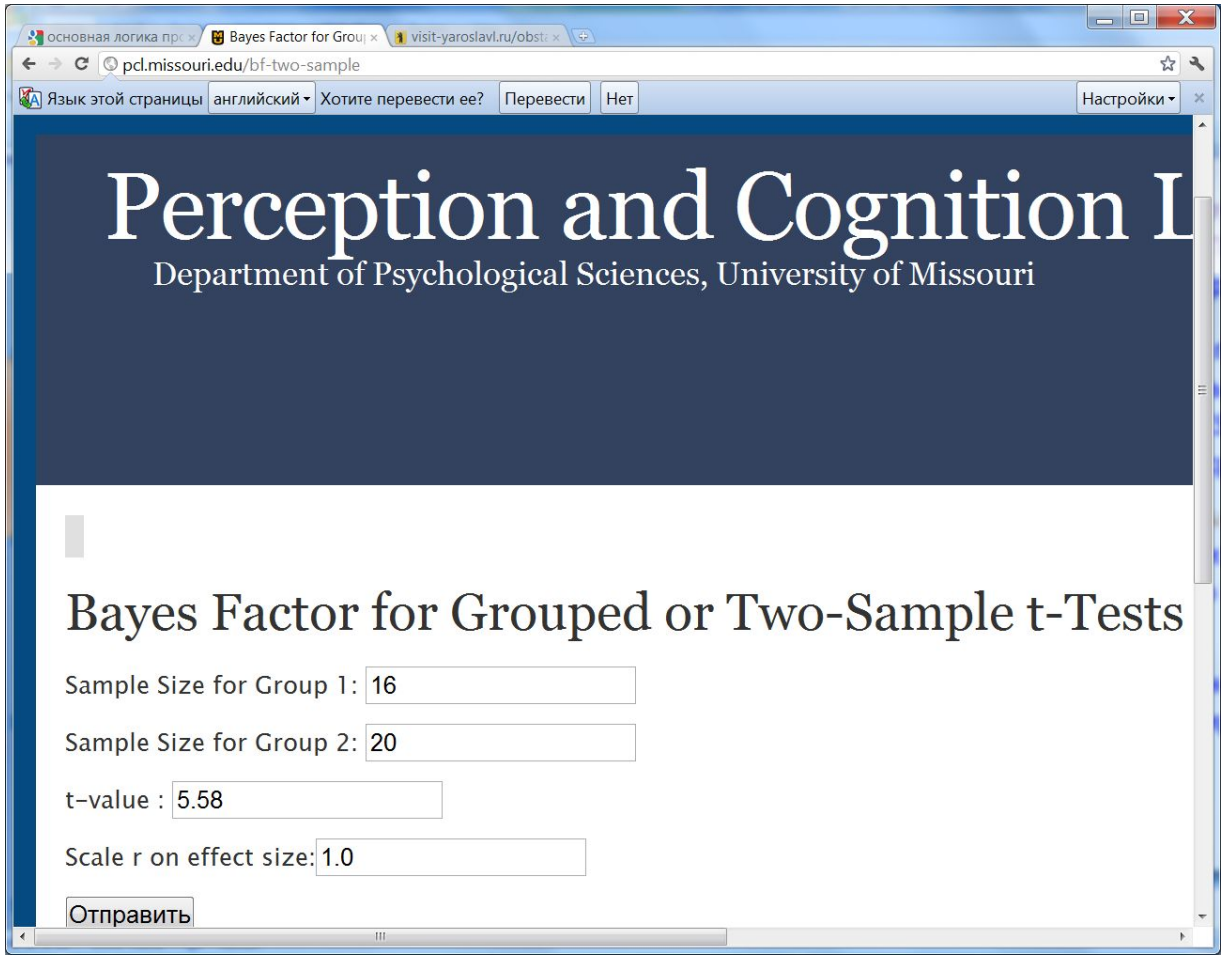
Интерпретация убедительности Бейзовых факторов, BF_{10} и BF_{01}

BF_{10}	*Свидетельство в пользу гипотезы H_1 против гипотезы H_0*
>100	Убедительное
30 – 100	Очень сильное
10 – 30	Сильное
3 – 10	Умеренное
1 – 3	Пренебрежимо малое

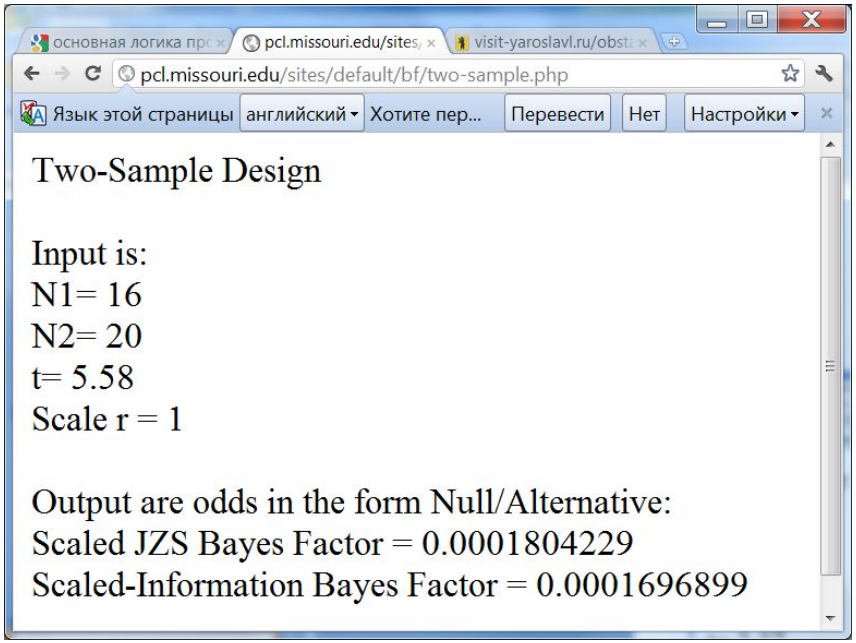
BF_{01}	*Свидетельство в пользу гипотезы H_0 против гипотезы H_1*
<0,01	Убедительное
0,01 – 0,03	Очень сильное
0,03 – 0,1	Сильное
0,1 – 0,3	Умеренное
0,3 - 1	Пренебрежимо малое

Бейзов фактор, программа Bayes Factor Calculators

http://pcl.missouri.edu/bayesfactor



Вывод результатов (output)

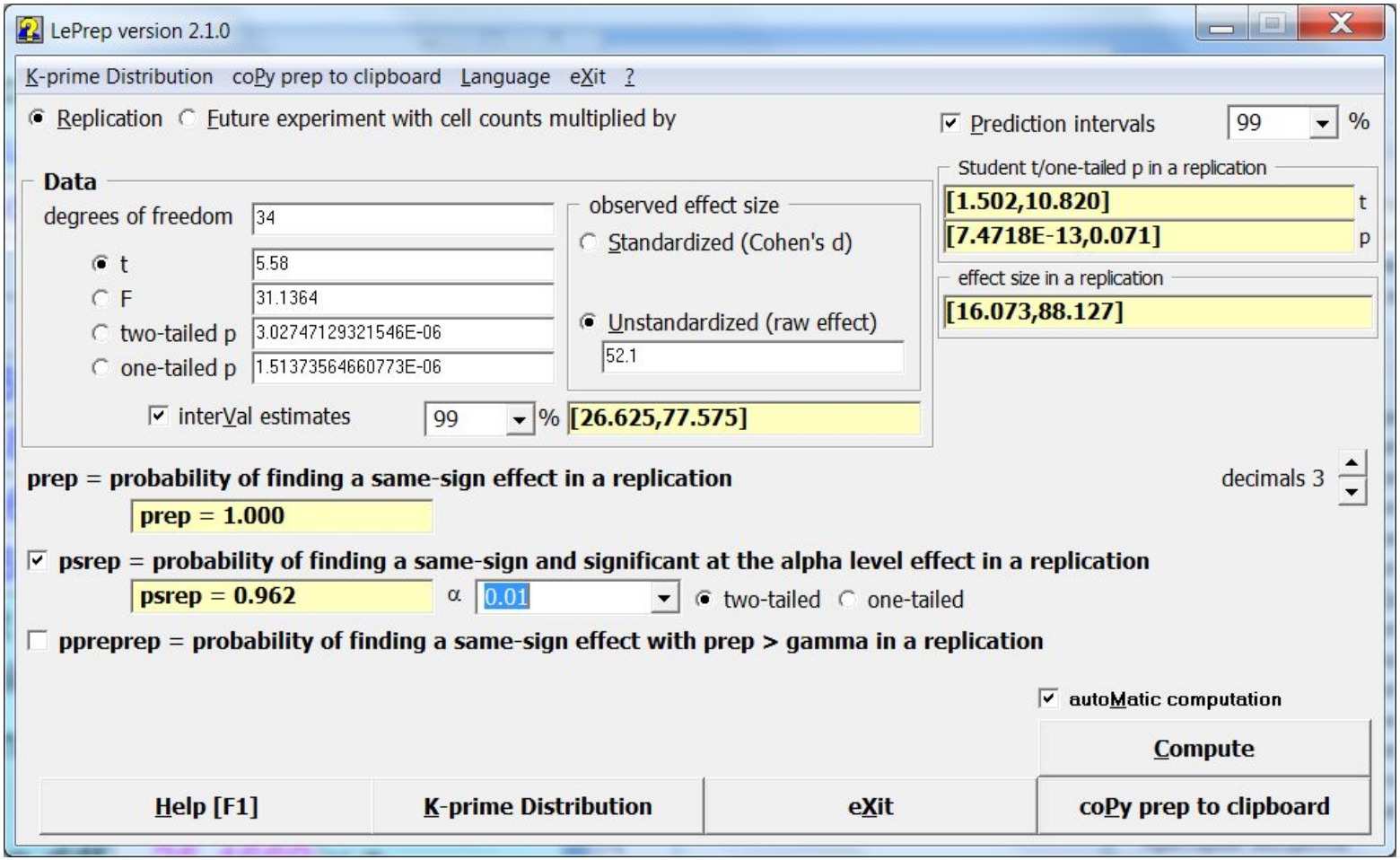


- В 5555 раз (1/0,00018) более правдоподобно получить наблюдаемые различия
- (*ES* = 52,1 у.е.) между сравниваемыми группами при условии, что верна гипотеза H_1: $ES \neq 0$, нежели при условии, что верна гипотеза H_0: $ES = 0$.
- Такое значение BF_{01} принято интерпретировать как чрезвычайно убедительное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

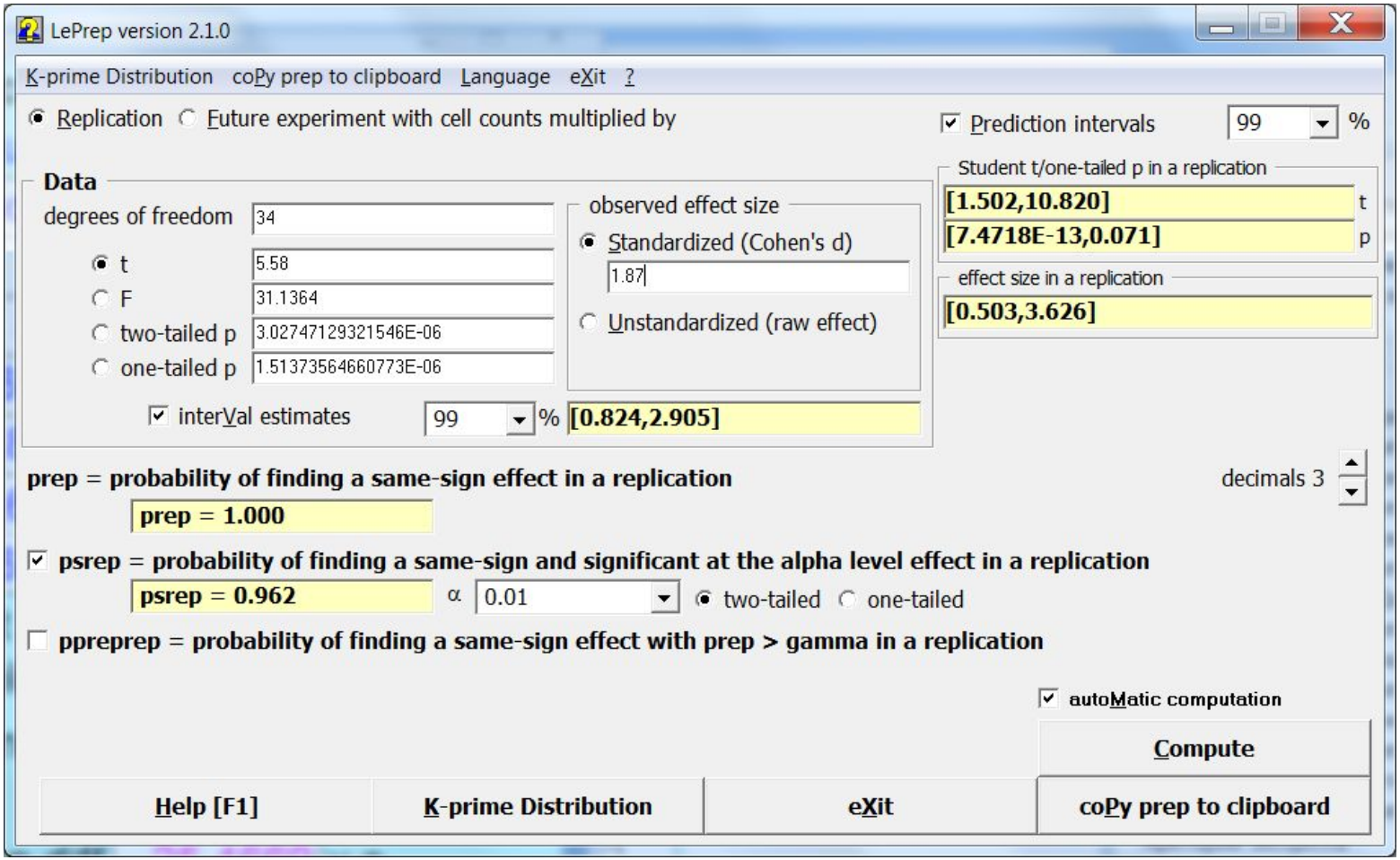
Статистические предсказания и воспроизводимость

Воспроизводимость и предсказания абсолютного размера эффекта для групп матерей здоровых детей и детей с СЗРП. 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 СЗРП

Показатель	*ES*	d_c
Предсказательные интервалы (ПИ) для размеров эффекта	[16,1; 88,1]	[0,50; 3,63]
Предсказательные интервалы (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 ПИ для $P_{val} < 0,05$) и размер эффекта по Коуэну может оказаться малым, достигая нижней границы ПИ для него: 0,5.

Ошибки I и II рода и мощность статистического критерия

Диагностика

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

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

Компромисс

- **Например, в случае металлодетектора**
- **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 β - вероятность ошибки II рода.**
- **Чем больше мощность критерия, тем меньше вероятность совершить ошибку II рода.**

Мощность статистического критерия

- **Мощность статистического критерия измеряет способность критерия выявлять истинные различия (эффекты).**
- **Ее можно интерпретировать как чувствительность статистического критерия к отклонениям от условий нулевой гипотезы.**

Доверяя, повторяй

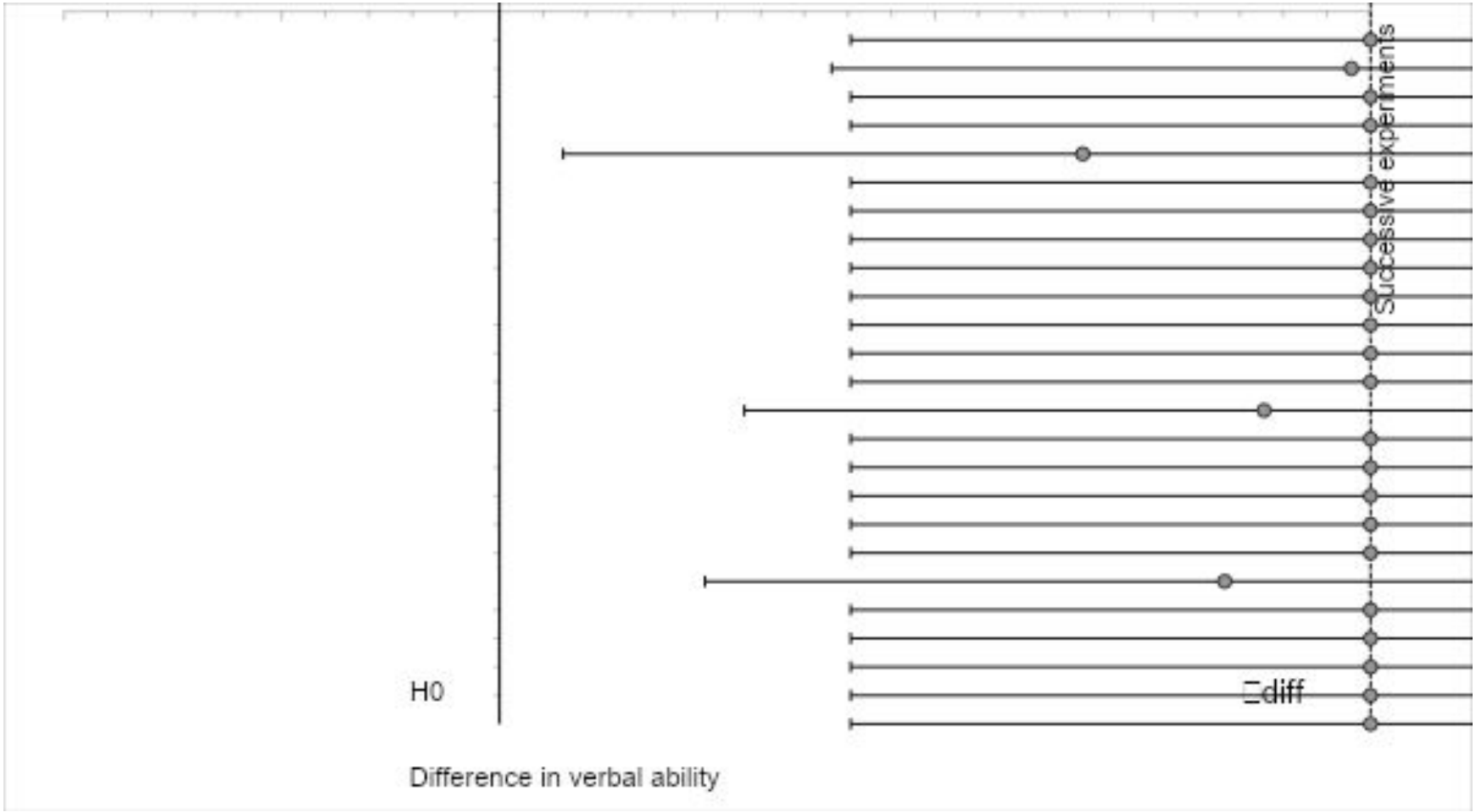
- **Часто считается, что если получен «статистически значимый» результат, то это исключает необходимость повторить исследование.**
- **Повторность (воспроизведение) часто рассматривается как нечто суетное и мирское.**

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

Воспроизводимость *P*-значений и ДИ

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspectives on Psychological Science, 3, 286-300.

Программа ESCI PPS p intervals http://www.latrobe.edu.au/psy/esci/

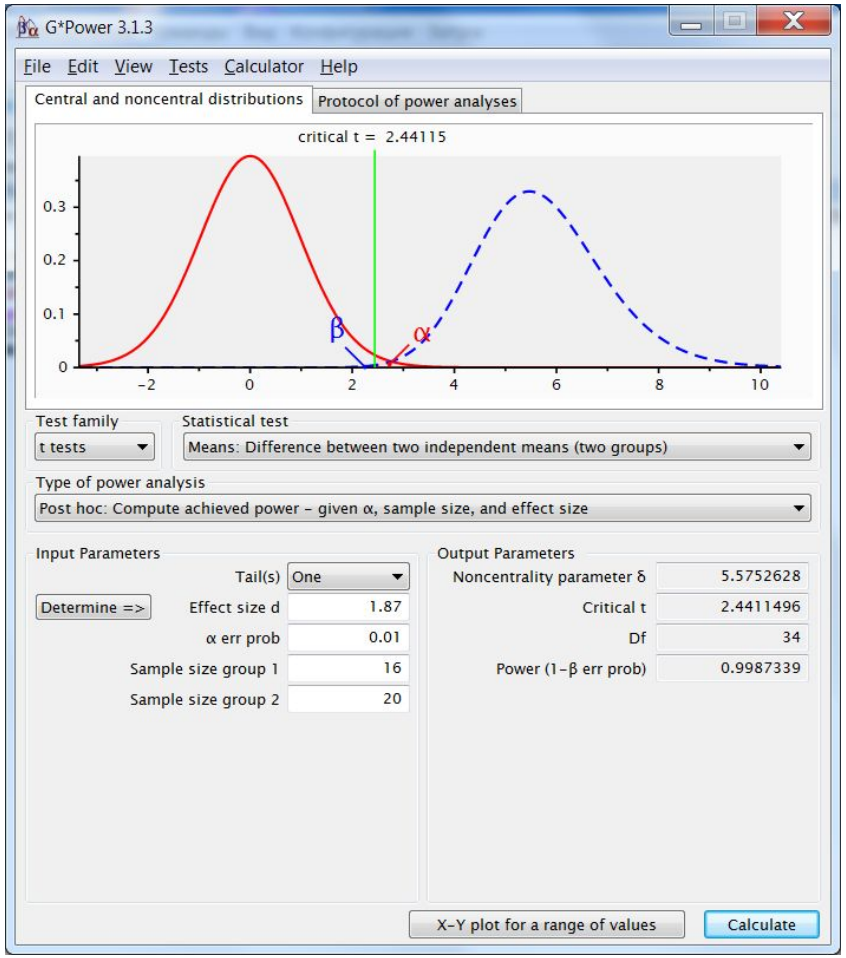


Анализ мощности *a priori* или *post-hoc*

- **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) = 0,9987**

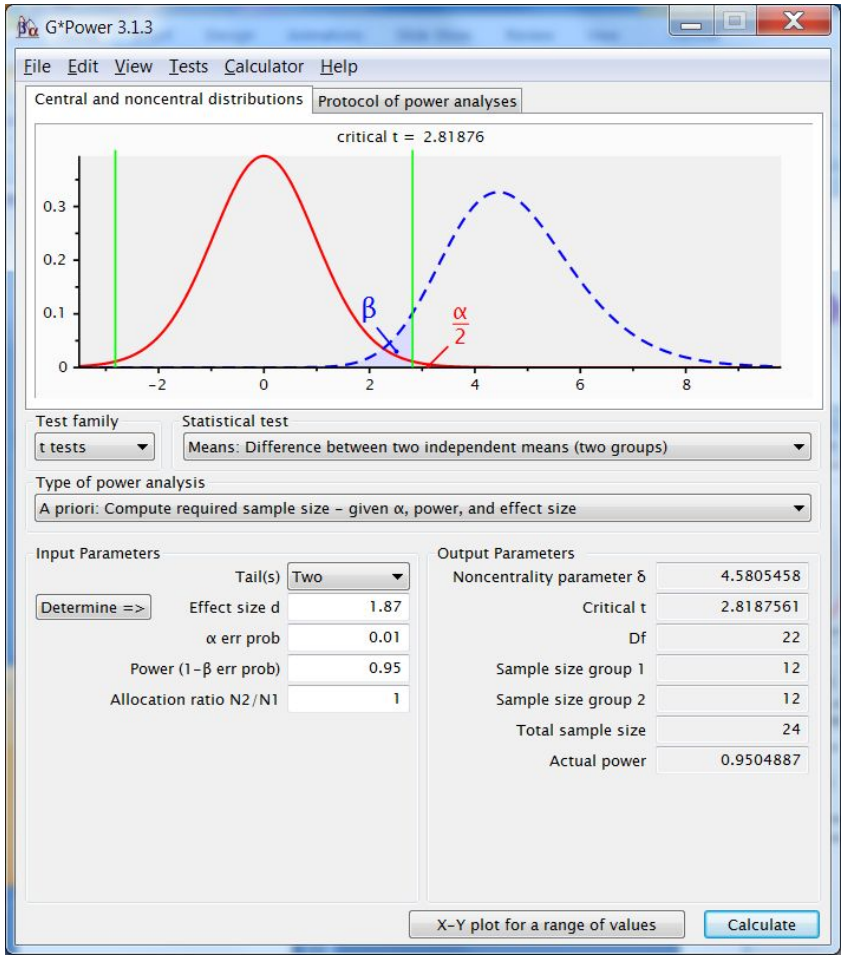
Элементы планирования эксперимента

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.**

Значение вероятностной *P*-величины

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

$$\tilde{P}$$

- **Всякий раз мы наблюдаем одно из ее возможных значений.**
- **Когда H_0 верна, то P_{val} имеет непрерывное равномерное распределение на отрезке**
  - **[0; 1].**

- **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.**

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 Мйкельсон 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.**

Повторение – мать познания

- **Повторение составляет суть науки:**
- **ученый должен всегда задумываться о том, что произойдет, если он или другой ученый повторят его эксперимент (Guttman, 1977).**

- ***Ученые разработали метод определения надежности (валидности) своих результатов.***
- ***Они научились задавать вопрос: воспроизводимы ли они?* (Scherr, 1983).**

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.**

Спасибо за внимание!
Слайды доступны для всех

Никита Николаевич Хромов-Борисов

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

Nikita.KhromovBorisov@gmail.com

(812) 234-18-40 – дом.

(812) 234-66-55 – раб.

8-952-204-89-49 – моб.