

# Множественное тестирование

ПМИ ФКН ВШЭ, 14 октября 2019 г.

Денис Деркач

ФКН ВШЭ

Денис Деркач

# Оглавление

Мотивация

Групповая вероятность ошибки первого рода

False Discovery Rate

Теорема Байеса и FDR

Перестановочные тесты

Множественные доверительные интервалы

Look Elsewhere Effect

If you torture your data long enough, they will tell you whatever you want to hear.

MILLS (1993)

Мотивация

# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

- › Что надо сделать?

# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

- › Что надо сделать?
- › Провести экзамен по самому важному предмету.

# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

- › Что надо сделать?
- › Провести экзамен по самому важному предмету.
- › А если предметов несколько?

# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

- › Что надо сделать?
- › Провести экзамен по самому важному предмету.
- › А если предметов несколько?
- › Провести несколько финальных экзаменов, проверить отклонения.



# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

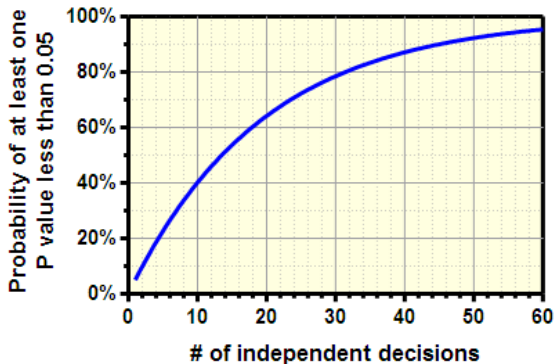
- › Что надо сделать?
- › Провести экзамен по самому важному предмету.
- › А если предметов несколько?
- › Провести несколько финальных экзаменов, проверить отклонения.
- › Какой будет в этом случае вероятность найти различие?

# Мотивация: тестирование образовательных программ

Есть две разные образовательные программы, Вы хотите понять какая из них лучше.

- › Что надо сделать?
- › Провести экзамен по самому важному предмету.
- › А если предметов несколько?
- › Провести несколько финальных экзаменов, проверить отклонения.
- › Какой будет в этом случае вероятность найти различие?
- ›  $\mathbb{P}(\text{significant}) = \mathbb{P}(\text{at least one significant result}) = 1 - \mathbb{P}(\text{no significant results}) = 1 - (1 - \alpha)^k$ . Например, для  $k = 20$  и  $\alpha = 0.05$ ,  $\mathbb{P} = 0.64$ .

# Зависимость от числа тестов



Вероятность найти хотя бы одно отклонения быстро увеличивается с ростом количества тестирований.

# Мотивация: Пример электричество из Швеции

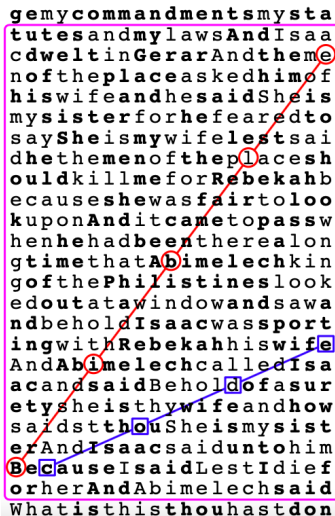
- › В 1992 году в Швеции исследование о влиянии линий электропередач на здоровье.
- › 25 лет наблюдали всех людей, которые жили вблизи линий электропередач.
- › Вывод: есть связь между линией электропередач и возникновением лейкемии.

# Мотивация: Пример электричество из Швеции

- › В 1992 году в Швеции исследование о влиянии линий электропередач на здоровье.
- › 25 лет наблюдали всех людей, которые жили вблизи линий электропередач.
- › Вывод: есть связь между линией электропередач и возникновением лейкемии.
- › Проблема: тестировали сразу 800 болезней. Некоторые из них случайным образом дали корреляцию.

# Мотивация: Библейский код

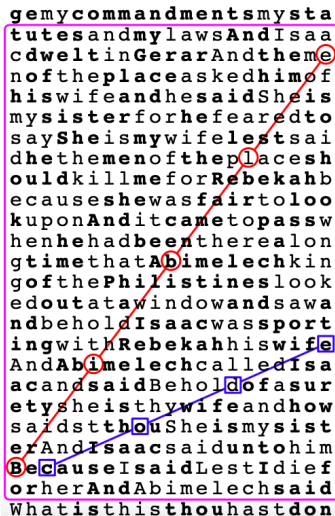
- › Идея: в Библии есть зашифрованная информация, которую необходимо найти, например, читать только буквы, расположенные на диагоналях через 4.
- › действительно, в Библии было найдено множество «предсказаний».



gemycommandmentsmysta  
tutesandmylawsAndIsaa  
cdweltinGerarAndtheme  
noftheplaceaskedhimof  
hiswifeandhesaidSheis  
mysisterforhefearedto  
saySheismywifelestsai  
dhethemenoftheplacesh  
ouldkillmeforRebekahb  
ecauseshewasfairtoloo  
kuponAnditcametopassw  
henhehadbeentherealon  
gtimethatAbimelechkin  
gofthePhilistineslook  
edoutatawindowandsawa  
ndbeholdIsaacwassport  
ingwithRebekahhiswife  
AndAbimelechcalledisa  
acandsaidBeholdofasur  
etysheisthywifeandhow  
saidstthouSheismysist  
erAndIsaacsaiduntohim  
BecauseIsaidLestIdief  
orherAndAbimelechsaid  
Whatisthisthouhastdon

# Мотивация: Библейский код

- › Идея: в Библии есть зашифрованная информация, которую необходимо найти, например, читать только буквы, расположенные на диагоналях через 4.
- › действительно, в Библии было найдено множество «предсказаний».
- › Проблема: Библия большая, всегда можно найти нужное правило и прочитать несколько «скрытых сообщений».



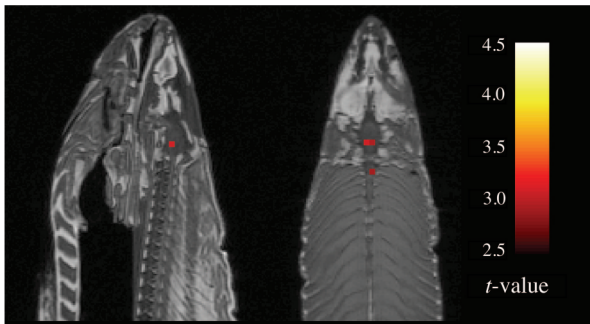
gemycommandmentsmysta  
tutesandmylawsAndIsaa  
cdweltinGerarAndtheme  
noftheplaceaskedhimof  
hiswifeandhesaidSheis  
mysisterforhefearedto  
saySheismywifelestsai  
dhethemenoftheplacesh  
ouldkillmeforRebekahb  
ecauseshewasfairtoloo  
kuponAnditcametopassw  
henhehadbeentherealon  
gtimethatAbimelechkin  
gofthePhilistineslook  
edoutatawindowandsawa  
ndbeholdIsaacwassport  
ingwithRebekahhiswife  
AndAbimelechcalledisa  
acandsaidBeholdofasur  
etysheisthywifeandhow  
saidstthouSheismysist  
erAndIsaacsaiduntohim  
BecauseIsaidLestIdief  
orherAndAbimelechsaid  
Whatisthisthouhastdon

# Множественные тесты при сравнении геномов

Обычный эксперимент включает в себя попарные сравнения, то есть до 10000 тестов проходят одновременно. Таким образом около 500 тестов могут дать "значительное" различие.



# Мозговая активность у мёртвого лосося



The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive.

Шнобелевская премия 2011

# Как учесть эффект в исследовании?

- › Не учитывать, но полностью описывать процесс (включая отсутствие коррекций).
- › Изначально выделять одну интересную область.
- › Изначально планировать эксперимент, не предусматривающий множественные сравнения.

NB: все планы экспериментов делаются заранее.

# Исходы экспериментов

Рассмотрим  $m$  различных проверок гипотез:

$$H_{0i} \text{ vs. } H_{1i}, \quad i = 1, \dots, m.$$

	$H_0$ не отклонена	$H_0$ отклонена	$\sum$
$H_0$ верна	U	V	$m_0$
$H_0$ неверна	T	S	$m_1$
$\sum$	$m - R$	R	$m$

- ›  $m_0$  - количество верных нулевых гипотез.
- ›  $R$  - количество отклонённых нулевых гипотез.

# Как можно исправить

- › Per comparison error rate (PCER): ожидаемая частота ошибок первого рода  $PCER = \mathbb{E}(V)/m$ .
- › Per-family error rate (PFER): ожидаемое количество ошибок первого рода.  $PFE = E(V)$ .
- › Family-wise error rate (FWER): групповая вероятность ошибки первого рода  $FWER = \mathbb{P}(V \geq 1)$ .
- › False discovery rate (FDR): ожидаемое отношение ошибок первого рода к общему количеству отклонений  
 $FDR = \mathbb{E}(V/R | R > 0) \mathbb{P}(R > 0)$ .
- › Positive false discovery rate (pFDR):  $pFDR = E(V/R | R > 0)$ .

Групповая  
вероятность ошибки  
первого рода

# Групповая вероятность ошибки первого рода

Самая популярная величина для контроля: Family-Wise Error Rate (вероятность по крайней мере одной ошибки первого рода):

$$\mathbb{P}(V \geq 1) = 1 - \mathbb{P}(V = 0)$$

Обычно выделяют два типа контроля FWER :

- › Одношаговое: одновременно изменить все  $p$ -значения.
- › Последовательная: адаптивная процедура

# Метод Бонферрони

## Определение (Метод Бонферрони)

Обозначим через  $p_1, \dots, p_m$  величины  $p$ -значений этих проверок. Для заданных  $p$ -значений  $P_1, \dots, P_m$  основная гипотеза  $H_{0i}$  отклоняется, если

$$p_i < \frac{\alpha}{m}.$$

Фактически, вводит новые  $p$  значения:

$$\tilde{p}_i = \min(m p_i, 1)$$

Пример: если мы делаем 100 тестов и хотим получить  $\alpha = 0.05$ , то нам придётся получить  $p_i = 0.05/100 = 5 \cdot 10^{-4}$  для теста, чтобы сказать, что разница значимая.

# Оценка метода

- › Неинтуитивен: интерпретация результатов зависит от количество других выполненных тестов.
- › Групповую нулевая гипотеза редко представляет интерес у исследователей.
- › Высокая вероятность ошибок 2 рода, т. е. не отклонения групповой нулевой гипотезы, когда существуют важные эффекты.
- › Быстрое снижение мощности теста при при росте  $m$ .



# Метод Холма

Можно предложить последовательные коррекции.

- › отсортируем реальные  $p$ -величины по возрастанию:

$$p_{(1)} \geq \dots \geq p_{(m)};$$

- › скорректируем следующим образом:

$$\tilde{p}_i = \min((m - i + 1)p_{(i)}, 1);$$

- › будем использовать новые  $p$ -значения для проверки гипотез, начиная с  $\tilde{p}_1$ .

Этот метод обеспечивает  $FWER \geq \alpha$ , причём он равномерно мощнее метода Бонферрони.

Например, для 1000 тестов:  $\tilde{p}_1 = 1000p_1$ ,  $\tilde{p}_2 = 999p_2$  и тд.

# Другие методы контроля FWER

- › метод Шидака: последовательная коррекция  $\alpha_m = (1 - \alpha)^{1/m}$ ;
- › метод Хохберга: аналогично методу Холма, но проверка начинается с  $\tilde{p}_m$ ;
- › метод Хоммеля: проверка наличия  $j$  такого, что  $p_{n-j+k} < \frac{k\alpha}{j}$ , где  $k = 1, \dots, j$ , по нахождению максимального  $j$  тестируем  $p_i \leq \alpha/j$ .

False Discovery Rate

# False Discovery Rate, FDR

	$\mathbb{H}_0$ не отклонена	$\mathbb{H}_0$ отклонена	$\Sigma$
$\mathbb{H}_0$ верна	U	V	$m_0$
$\mathbb{H}_0$ неверна	T	S	$m_1$
$\Sigma$	$m - R$	R	$m$

## Определение

Доля (FDP) и частота (FDR) появления ложных отклонений (среди отклонений вообще):

$$FDP = \frac{V}{R} \mathbb{P}(R > 0)$$

$$FDR = \mathbb{E}(FDP).$$

# FDR и FWER

- › FWER используется в случае, если нам очень важно контролировать вероятность ошибки первого рода;
- › в некоторых случаях, более целесообразно ослабить этот подход и контролировать вероятность получить  $k$  ошибок первого рода (kFWER) или вообще контролировать частоту ложных срабатываний (FDP, FDR).
- › при этом стоит отметить, что мы не делаем ничего сверхестественного:

$$\frac{\mathbb{E}(V)}{m} \leq FDR \leq FWER \leq \mathbb{E}(V)$$

# Метод Бенджамини-Хохберга

1. Пусть  $P_{(1)} < \dots < P_{(m)}$  -величины  $p$ -value, отсортированные по возрастанию.
2. Пусть  $C_m = 1$  в случае, если  $P_{(1)}, \dots, P_{(m)}$  независимы, в противном случае  $C_m = \sum_{i=1}^m \frac{1}{i}$ .
3. Определим:  $I_i = \frac{i\alpha}{C_m m}$ ,  $R = \max\{i : P_{(i)} < I_i\}$ .
4.  $T = P_{(R)}$  - пороговое значение метода.
5. Отклоняются такие  $H_{0i}$ , для которых  $P_i \leq T$ .

# Метод Стори

формально, мы можем убрать вероятность  $R > 0$  из определения FDR:

$$pFDR = \mathbb{E} \frac{V}{R},$$

Тогда метод Бенджамини-Хохберга становится методом Стори.

# Теорема Байеса и FDR



# $Q$ -значение

По аналогии с  $p$ -значением мы можем ввести  $Q$ -значение для критической области  $C$ :

## Определение

Для некоторого значения статистики  $T = t$   $Q$ -значение будет считаться:

$$Q(t) = \inf_{\{C:t \in C\}} pFDR(C)$$

$Q$ -значение — функция  $p$ -значений для этого набора тестов. Имеет смысл аналогичный  $p$ -значению для одного теста.

# FDR и теорема Байеса

Пусть проводится  $m$  идентичных тестов с независимыми статистиками  $T_1, \dots, T_m$  для которых определена критическая область  $C$ . Нулевая гипотеза верна с априорной вероятностью  $\pi_0 = \mathbb{P}(\mathbb{H}_0 \text{ is true})$ . Тогда

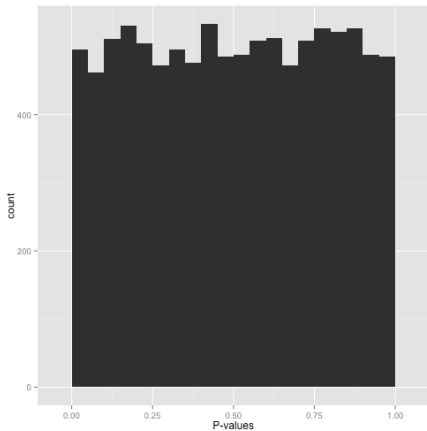
$$pFDR(C) = \mathbb{P}(\mathbb{H}_0 \text{ is true} | T \in C).$$

По теореме Байеса:

$$pFDR(C) = \frac{\pi_0 \mathbb{P}(T \in C | \mathbb{H}_0 \text{ is true})}{\mathbb{P}(T \in C)}.$$

# Распределения $p$ -value

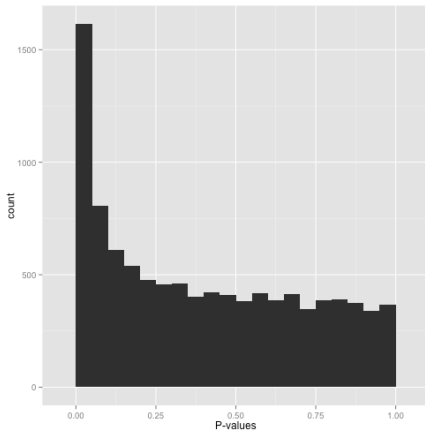
- › Если нулевая гипотеза верна, распределение будет равномерным.



figs from VarianceExplained

# Распределения $p$ -value

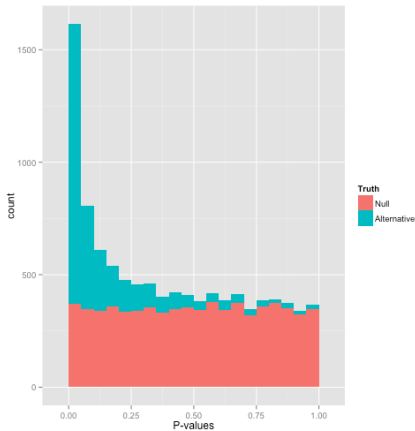
- › Если нулевая гипотеза верна, распределение будет равномерным.
- › Если нулевая гипотеза неверна, распределение будет сосредоточенно около 0.



figs from VarianceExplained

# Распределения $p$ -value

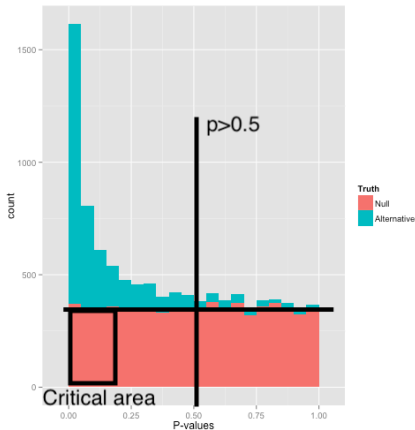
- › Если нулевая гипотеза верна, распределение будет равномерным.
- › Если нулевая гипотеза неверна, распределение будет сосредоточенно около 0.
- › Останется только разделить вклады.



figs from VarianceExplained

# Распределения $p$ -value

- › Если нулевая гипотеза верна, распределение будет равномерным.
- › Если нулевая гипотеза неверна, распределение будет сосредоточенно около 0.
- › Останется только разделить вклады.
- › Например, на основании поведения  $p$ -значений в правой стороне



# Перестановочные тесты

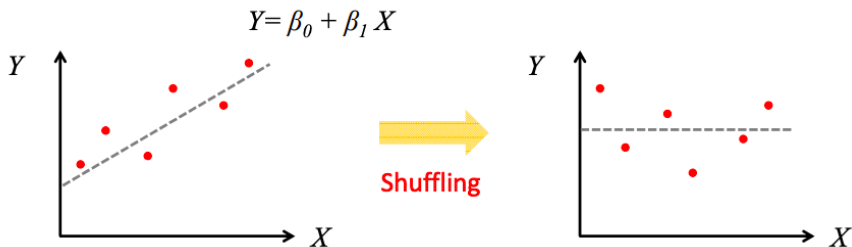
# $p$ -значения

- › во всех процедурах тестирования подразумевается, что распределение -значений ”правильное”;
- › часто, такое предположение неверно, например, если  $p$ -значение было получено ассимптотическими методами;
- › из-за этого мы можем применить критерий перестановок.



# Пример в линейной регрессии

Если мы строим регрессию, и если точки  $Y$  случайно связаны с  $X$  и указывают на нулевую гипотезу, что истинный наклон ноль, мы можем перемешать  $Y$ , связывая их с  $X$  случайно, каждый раз проверяя наклон:



# Критерий перестановок

## Определение

Критерий перестановок применяется для проверки того, отличаются ли распределения.

Пусть  $X_1, \dots, X_m \sim F_X$  и  $Y_1, \dots, Y_n \sim F_Y$  - две независимые выборки. Требуется решить:

$$\mathbb{H}_0 : F_X = F_Y \text{ vs. } \mathbb{H}_1 : F_X \neq F_Y$$

Критерий перестановок не использует предположения об асимптотической сходимости к нормальному распределению.

# Критерий перестановок:

1. Обозначим через  $T(x_1, \dots, x_m, y_1, \dots, y_n)$  некоторую тестовую статистику, например,  $T(X_1, \dots, X_m, Y_1, \dots, Y_n) = |\bar{X}_m - \bar{Y}_n|$ .
2. Положим  $N = m + n$  и рассмотрим все  $N!$  перестановок объединенной выборки  $X_1, \dots, X_m, Y_1, \dots, Y_n$ .
3. Для каждой из перестановок подсчитаем значение статистики  $T$ .
4. Обозначим эти значения  $T_1, \dots, T_{N!}$ .

Если  $\mathbb{H}_0$  верна, то при фиксированных упорядоченных значениях  $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$  значение статистики  $T$  распределены равномерно на множестве  $T_1, \dots, T_{N!}$ .

# $p$ -значение для перестановок

Обозначим как перестановочное распределение статистики  $T$  такое, согласно которому:

$$P_0(T = T_i) = \frac{1}{N!}, \quad i = 1, \dots, N!$$

Пусть  $t_{obs}$  - значение статистики, которое было получено в опыте.  
Тогда:

$$Q\text{-value} = \mathbb{P}(T > t_{obs} | f) = \frac{1}{N!} \sum_{j=1}^{N!} \mathbb{I}(T_j > t_{obs}), \quad f \in \mathcal{F}_0$$

## Пример

Допустим, что  $(X_1, X_2, Y_1) = (1, 9, 3)$ . Пусть  $T(X_1, X_2, Y_1) = |\bar{X} - \bar{Y}| = 2$ , тогда

Перестановка	Значение $T$	Вероятность
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

$$Q\text{-value} = \mathbb{P}(T > 2) = 4/6$$

# Множественные доверительные интервалы

# Доверительные интервалы

В литературе описаны способы коррекции в случае множественных одновременных построений доверительных интервалов (Бенджамини Йекутели, 2012).

Коррекция покрытия пропорциональна количеству используемых параметров в задаче.

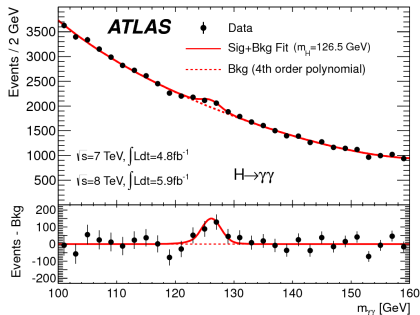
Особое внимание следует уделить утере корреляций в случае простой коррекции.

Look Elsewhere Effect



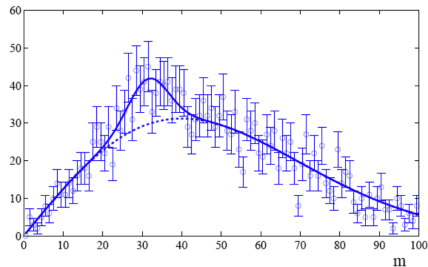
# Мотивация: поиск бозона Хиггса

- › При поиске новой частицы не знают точное положение её пика.
- › Потому сканируют все возможные значения.
- › В итоге, надо сказать является ли пик, который мы увидели настоящим пиком или нет.



# Поиск сигнальных пиков

- › теоретическая модель позволяет увидеть наличие пика в любом месте  $m$ ;
- › данные показывают наличие гауссоподобного эффекта в точке  $m_0$  с амплитудой  $\mu$ ;
- › как оценить значимость эффекта?



# Локальное $p$ -значение

- › Предположим, что мы ожидали пик в точке  $m_0$ ;
- › тогда мы можем использовать тест отношения правдоподобий для нулевой гипотезы отсутствия сигнала  $\mu = 0$ :

$$t_{fix} = -2 \ln \frac{\mathcal{L}(0, m_0)}{\mathcal{L}(\hat{\mu}, m_0)};$$

- › отсюда мы можем найти  $p$ -значение:

$$p_{local} = \int_{t_{fix}}^{\infty} f(t_{fix}|0) dt_{fix};$$

- › это будет локальное  $p$ -значение для конкретной точки  $m_0$ .

# Глобальное $p$ -значение

- › Предположи, что мы не знаем где пик;
- › тогда мы можем использовать тест отношения правдоподобий для нулевой гипотезы отсутствия сигнала  $\mu = 0$ :

$$t_{float} = -2 \ln \frac{\mathcal{L}(0)}{\mathcal{L}(\hat{\mu}, m_0)};$$

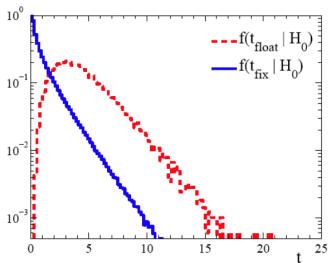
- › но правдоподобие нулевой гипотезы не зависит от  $m$ !
- › Мы можем найти  $p$ -значение:

$$p_{global} = \int_{t_{float}}^{\infty} f(t_{float}|0) dt_{float};$$

- › это будет  $p$ -значение для неизвестной массы.

# Распределения $t_{fix}$ , $t_{float}$

- › Для достаточно большого объёма данных,  $t_{fix} \sim \chi^2$  (согласно теореме Вилкса)
- › А вот для  $t_{float}$  всё неочевидно: теорема Вилкса работать не будет потому что количество мешающих параметров не фиксировано.



Мы можем получить распределение симуляцией.

# Примерное правило пересчёта

- › Мы хотим получить коррекцию на тот факт, что мы не знаем, где находится пик (и при этом хотим не использовать ресурсы).
- › Оказалось, что можно вывести коррекцию:

$$p_{global} = p_{local} + \langle N(c) \rangle,$$

где  $\langle N(c) \rangle$  - среднее количество пересечений  $t_{fix}$ , основанного на наблюдении в точке  $m_0$ :

$$c = t_{fix,obs} = Z_{local}^2,$$

где  $Z_{local} = \Phi^{-1}(1 - p_{local})$ .

- › Таким образом, мы либо симулируем все возможные положения сигнала, либо корректируем локальное  $p$ -значение.

# Скан

