

Множественное тестирование

Рассмотрим датасет с 30000 генов, в котором нет ни одного дифференциально экспрессирующегося гена

Проведем t-test для каждого гена. Будем считать ген дифференциально экспрессируемым если $p < 0.05$.

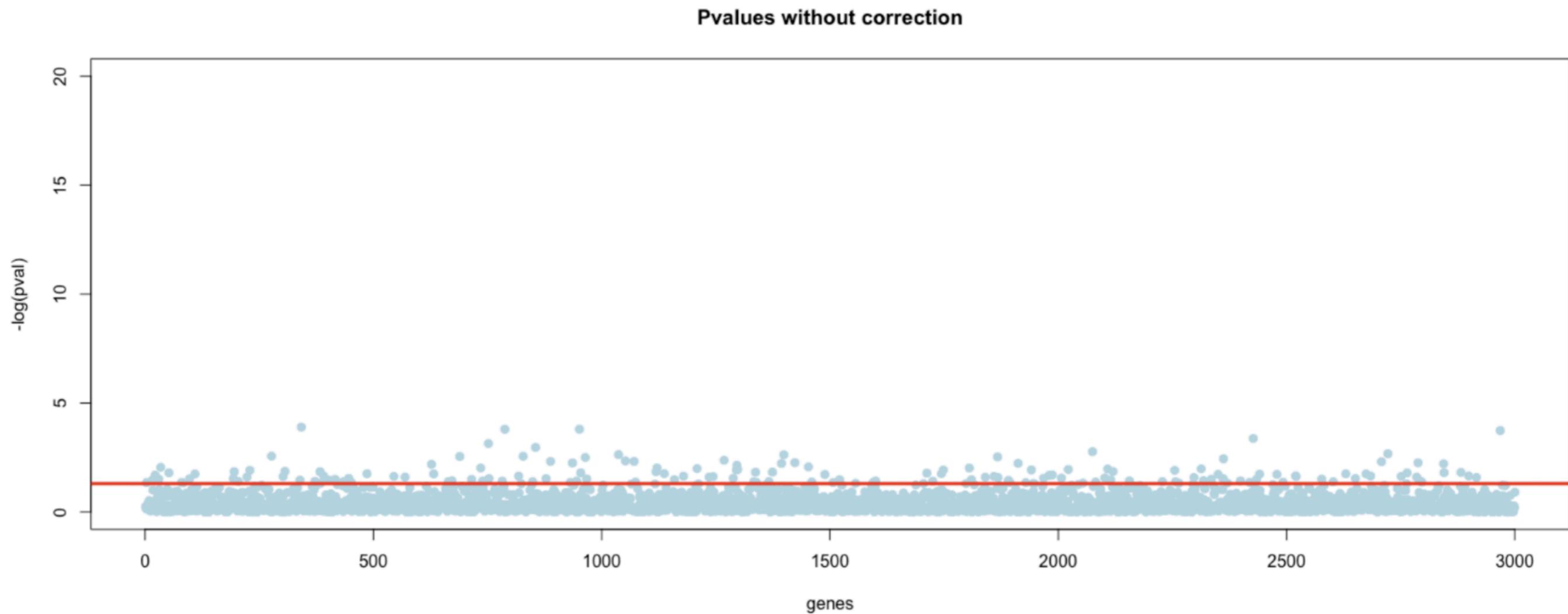
Какова вероятность, что ни один ген не будет помечен как дифференциально экспрессируемый?

Сколько в среднем генов будет помечено как дифференциально экспрессируемые?

Эксперимент 1

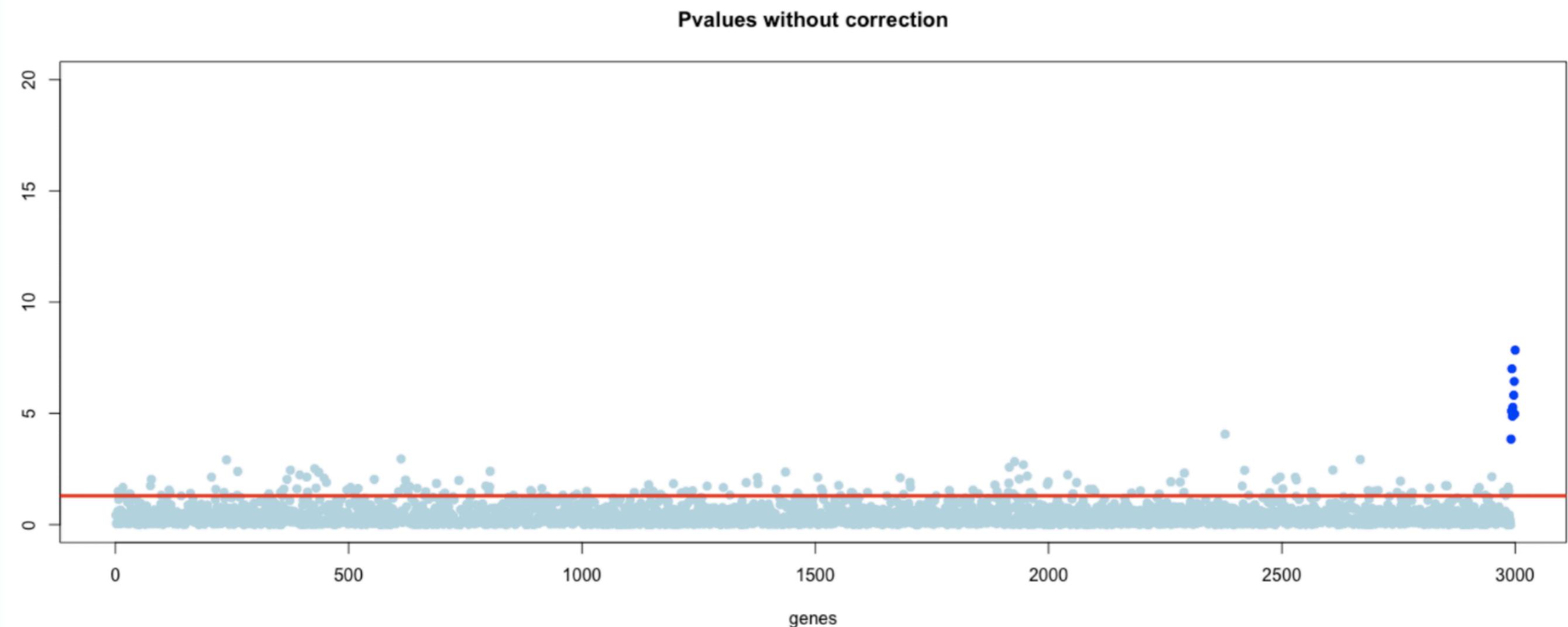
Возьмем и симулируем набор из 50 пациентов с 3000 генов, которые не меняют свою экспрессию значимо по ходу эксперимента (имеем результаты до и после).

Значения “экспрессии” генов будет брать из нормального распределения.



Эксперимент 2

Возьмем и симулируем набор из 50 пациентов с 2990 генами, которые не меняют свою экспрессию значимо по ходу эксперимента (имеем результаты до и после) и 10 генами, что ее меняют. Значения “экспрессии” генов будет брать из нормального распределения.



Поправки

- FWER (Family-Wise Error Rate) - вероятность, что среди отобранных генов хотя бы один ложно-положительный ген меньше заданного порога (0.05, например)
- FDR (False Discovery Rate) - процент ложно-положительных генов среди отобранных не больше, например, 20%

Смысл alpha **разный для двух подходов**

FWER

test	p-value	test	k	p-value
test1	p-value1	test1'	1	p-value1'
test2	p-value2	test2'	2	p-value2'
...
testN	p-valueN	testM'	M	p-valueN'

Наша изначальная таблица

Очень боимся
ошибиться даже в
одном случае!

Тесты, для которых
мы отвергаем H_0 .

Гарантируем, что вероятность того, что
во всей отобранный таблице встретится
хотя бы один тест, для которого мы
ошибочно отвергли H_0 - α

FWER

- 1) Sidak**
- 2) Bonferroni**
- 3) Holm-Bonferroni**

One-step procedure

Сравниваем р-value каждого тестов с одним и тем же порогом, если р-value меньше порога, то отвергаем H_0

test	p-value	Порог
test1	p-value1	thres
test2	p-value2	thres
...	...	thres
testN	p-valueN	thres

Sidak

$$p < 1 - \sqrt[n]{1 - alpha} = thres$$

Предполагаем независимость p-value

Используем one-step procedure

Почти никогда не используется

Bonferroni

$$p < \frac{\alpha}{N} = thres$$

p-value могут быть зависимы
Используем one-step procedure

Задача

При проверке десяти однотипных нулевых гипотез были получены следующие p-value:

0.0067, 0.1574, 0.0515, 0.0018, 0.0085, 0.0012, 0.0664, 0.0231, 0.0008, 0.0093

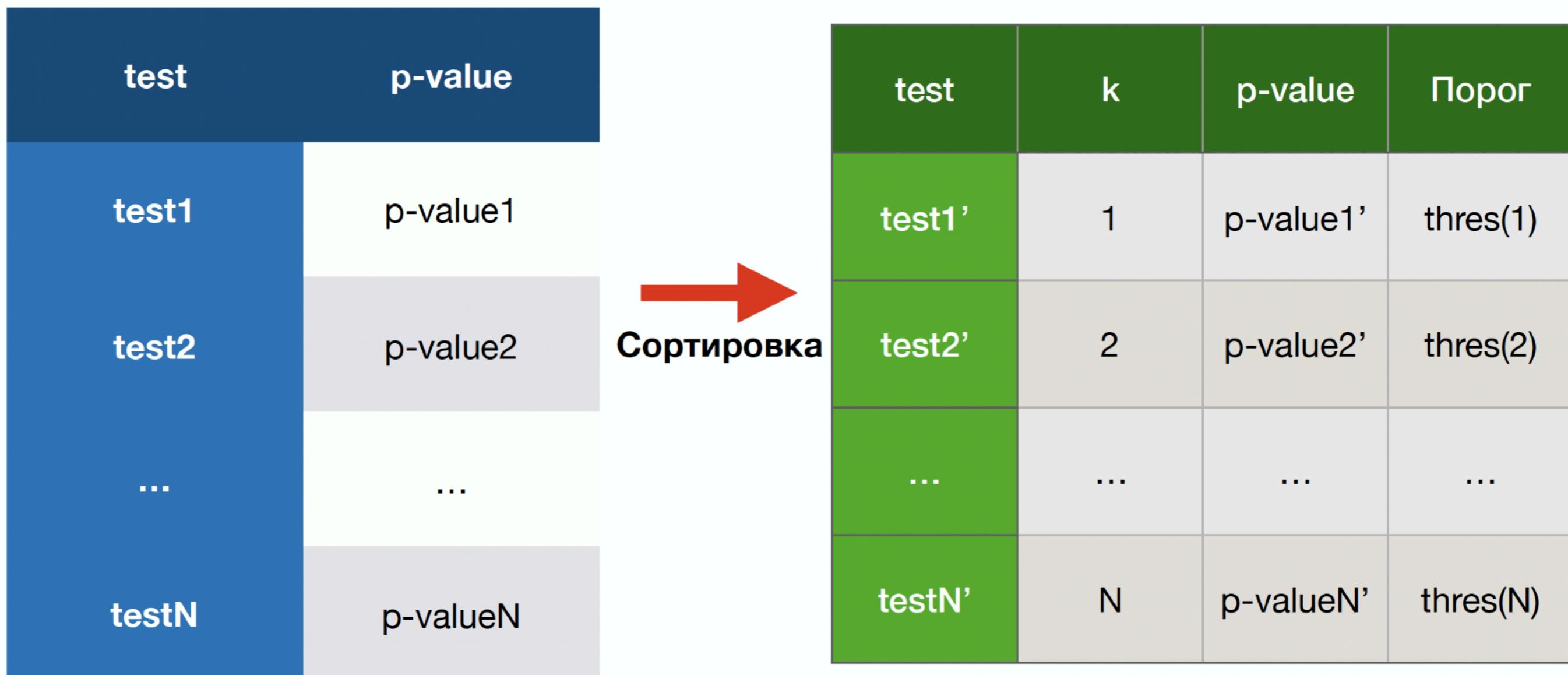
Если использовать поправку Бонферони, сколько раз нулевая гипотеза будет отвергнута с групповой вероятностью ошибки первого рода 5%?

Задача

k	p-value	Порог
1	0.0008	
2	0.0012	
3	0.0018	
4	0.0067	
5	0.0085	
6	0.0093	
7	0.0231	
8	0.0515	
9	0.0664	
10	0.1574	

Step-down procedure

1) Сортируем наши тесты по p-value от меньшего p-value к большему



Step-down procedure

- 2) Порог зависит от k
- 3) Идем от $k=1$ до первого $k=j$, для которого $p_value' \geq thres$ (то есть сверху вниз)

test	k	p-value	Порог
test1'	1	p-value1'	thres(1)
test2'	2	p-value2'	thres(2)
...
testj'	j	p-valuej'	thres(j)
...
testN'	N	p-valueN'	thres(N)

4) Для $k \geq j$
принимаем H_0 ,
для $k < j$ -
отвергаем

Holm-Bonferroni

$$p_k < \frac{\alpha}{N - k + 1} = \text{thres}(k)$$

p-value могут быть зависимы
Используем step-down procedure

Тест имеет те же гарантии, что и Bonferroni, но теоретически позволяет отобрать больше генов.
Потому использовать Bonferroni только тогда, когда вас обязывают это сделать

Задача

При проверке десяти однотипных нулевых гипотез были получены следующие p-value:

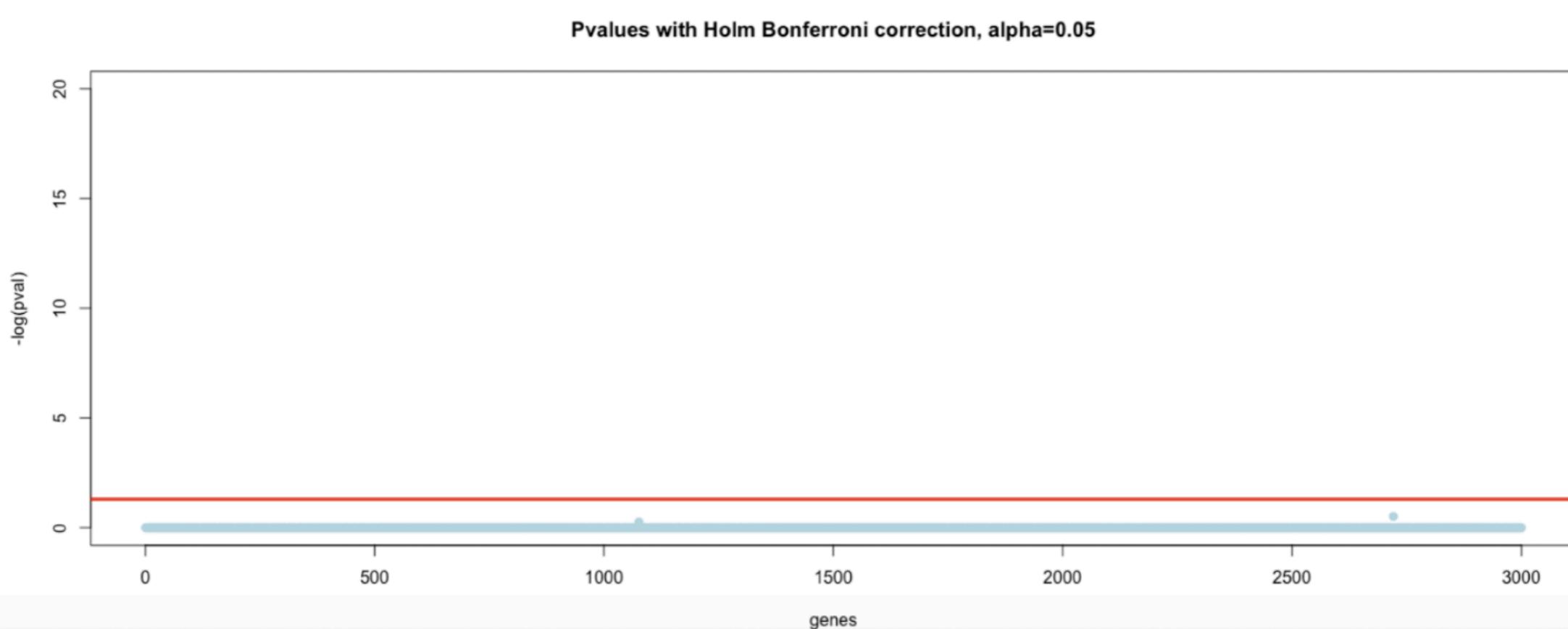
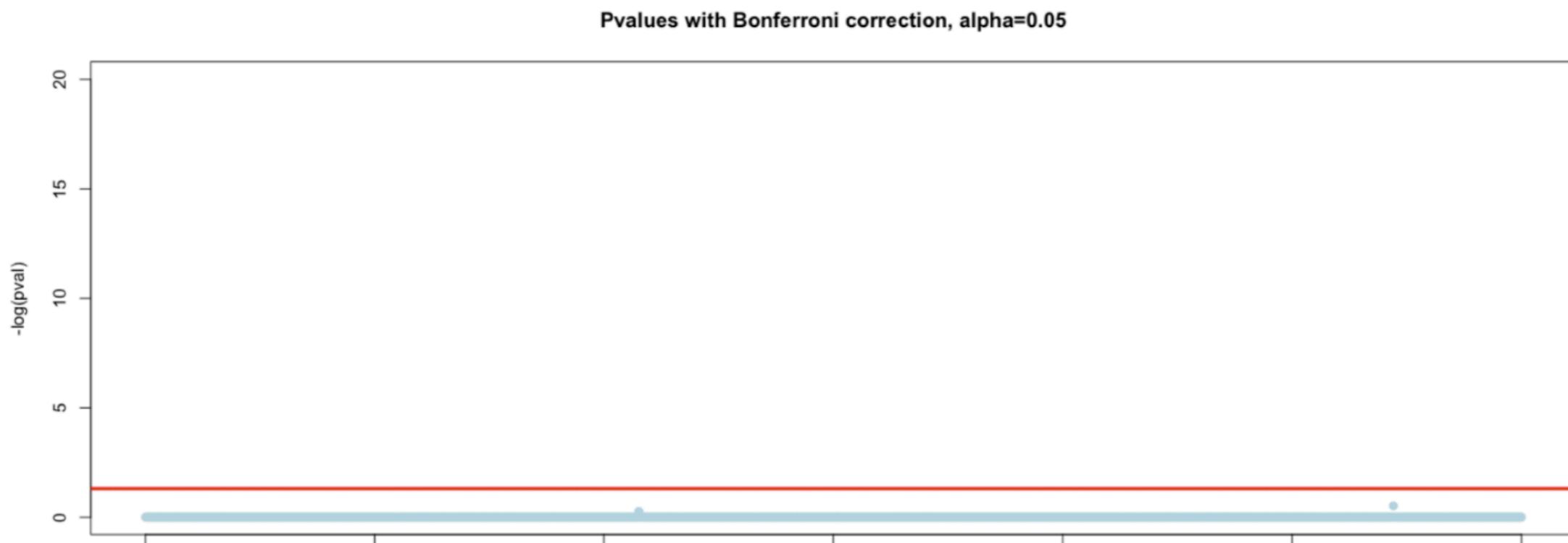
0.0067, 0.1574, 0.0515, 0.0018, 0.0085, 0.0012, 0.0664, 0.0231, 0.0008, 0.0093

Если использовать поправку Холма-Бонферони, сколько раз нулевая гипотеза будет отвергнута с групповой вероятностью ошибки первого рода 5%?

Задача

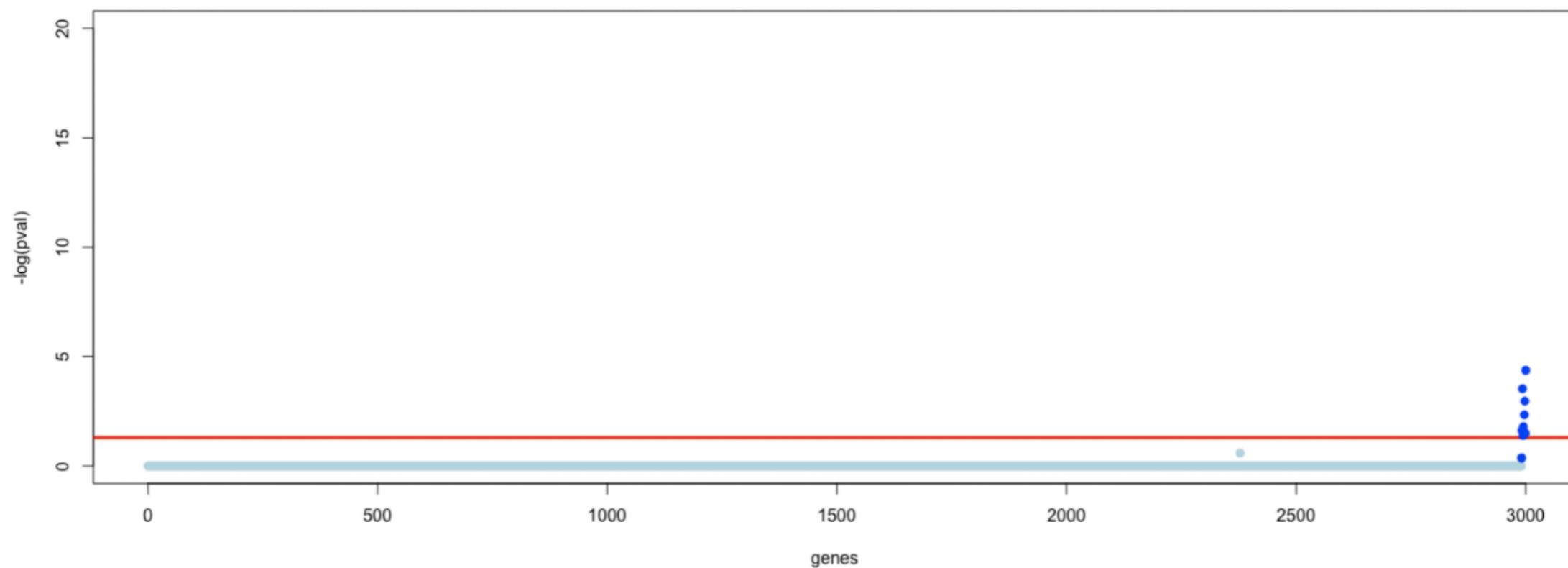
k	p-value	Порог
1	0.0008	
2	0.0012	
3	0.0018	
4	0.0067	
5	0.0085	
6	0.0093	
7	0.0231	
8	0.0515	
9	0.0664	
10	0.1574	

Эксперимент 1

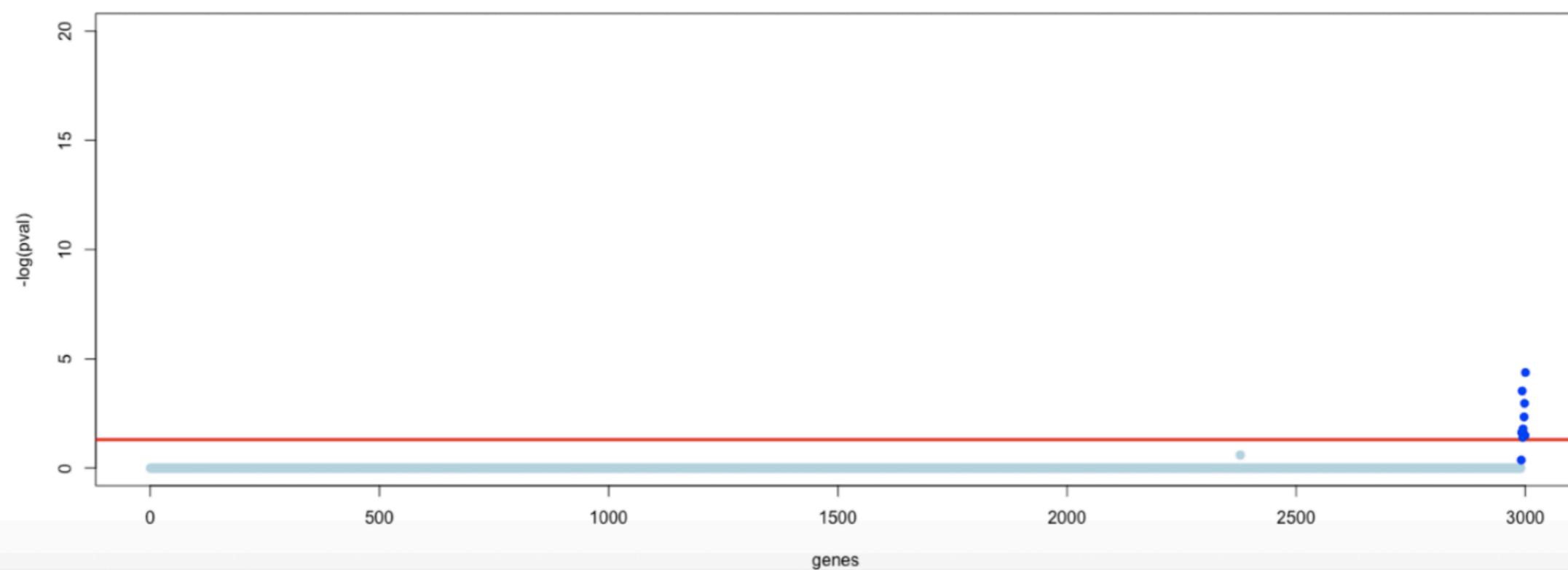


Эксперимент 2

Pvalues with Bonferroni correction, alpha=0.05



Pvalues with Holm Bonferroni correction, alpha=0.05



FDR

test	p-value
test1	p-value1
test2	p-value2
...	...
testN	p-valueN



test	k	p-value
test1'	1	p-value1'
test2'	2	p-value2'
...
testM'	M	p-valueN'

Наша изначальная таблица

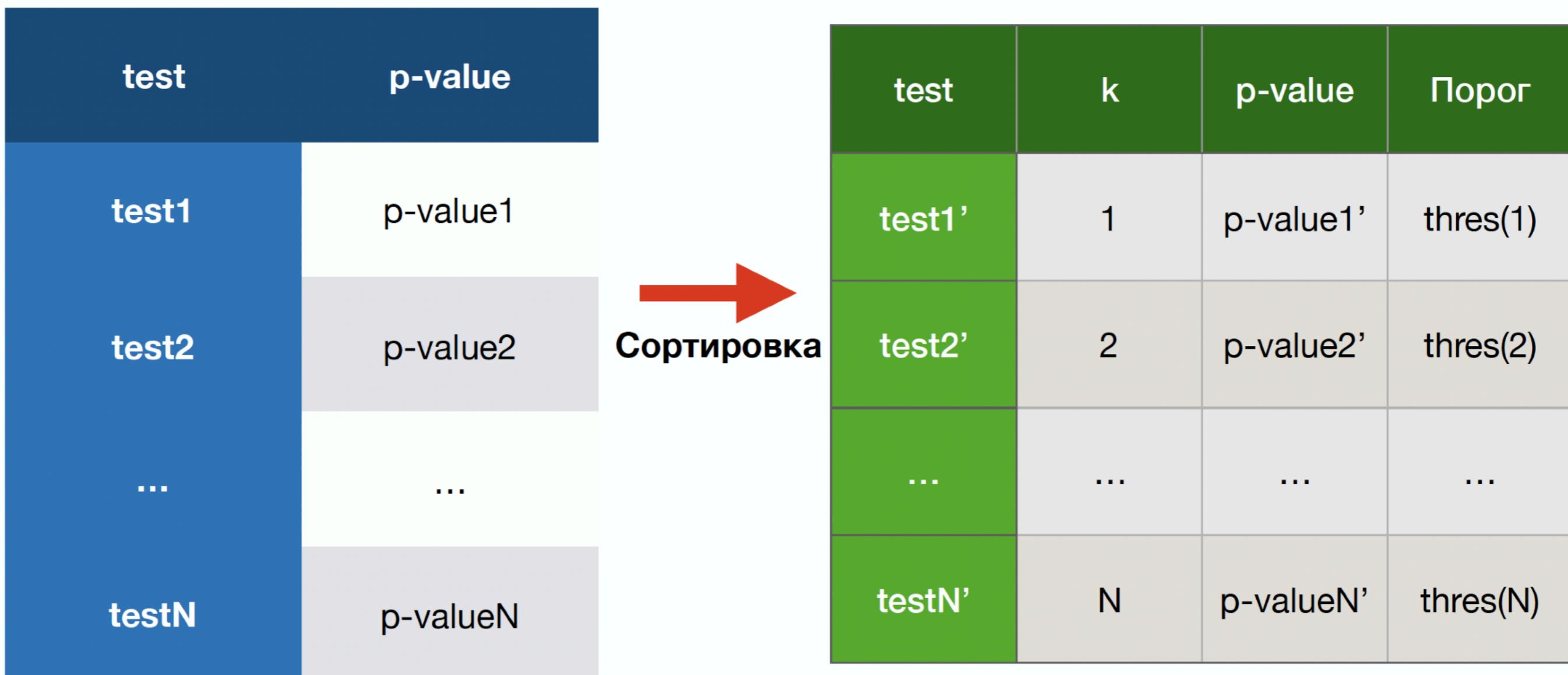
**Хотим набрать
данных с каким-то
уровнем шума для
 дальнейших
исследований**

**Тесты, для которых
мы отвергаем H_0 .**

**Гарантируем, что доля генов, для
которых мы ошибочно отвергли H_0 -
alpha**

Step-up procedure

1) Сортируем наши тесты по p-value от меньшего p-value к большему



Step-up procedure

- 2) Порог зависит от k
- 3) Идем от $k=N$ до первого $k=j$, для которого $p_value' < thres$ (то есть снизу вверх)

test	k	p-value	Порог
test1'	1	p-value1'	thres(1)
test2'	2	p-value2'	thres(2)
...
testj'	j	p-valuej'	thres(j)
...
testN'	N	p-valueN'	thres(N)

4) Для $k > j$
принимаем H_0 ,
для $k \leq j$ -
отвергаем H_0

Benjamini-Hochberg correction

Предполагает независимость проводимых тестов и их результатов (p-value), либо их положительную зависимость (если один тест имеет низкое p-value, то остальные так же имеют тенденцию иметь низкое p-value)

$$p_k < \frac{k}{N}\alpha = \text{thres}(k)$$

p-value могут быть зависимы
Используем step-up procedure

Задача

При проверке десяти однотипных нулевых гипотез были получены следующие p-value:

0.008, 0.0012, 0.017, 0.026, 0.0027, 0.032, 0.033, 0.048, 0.054, 0.062

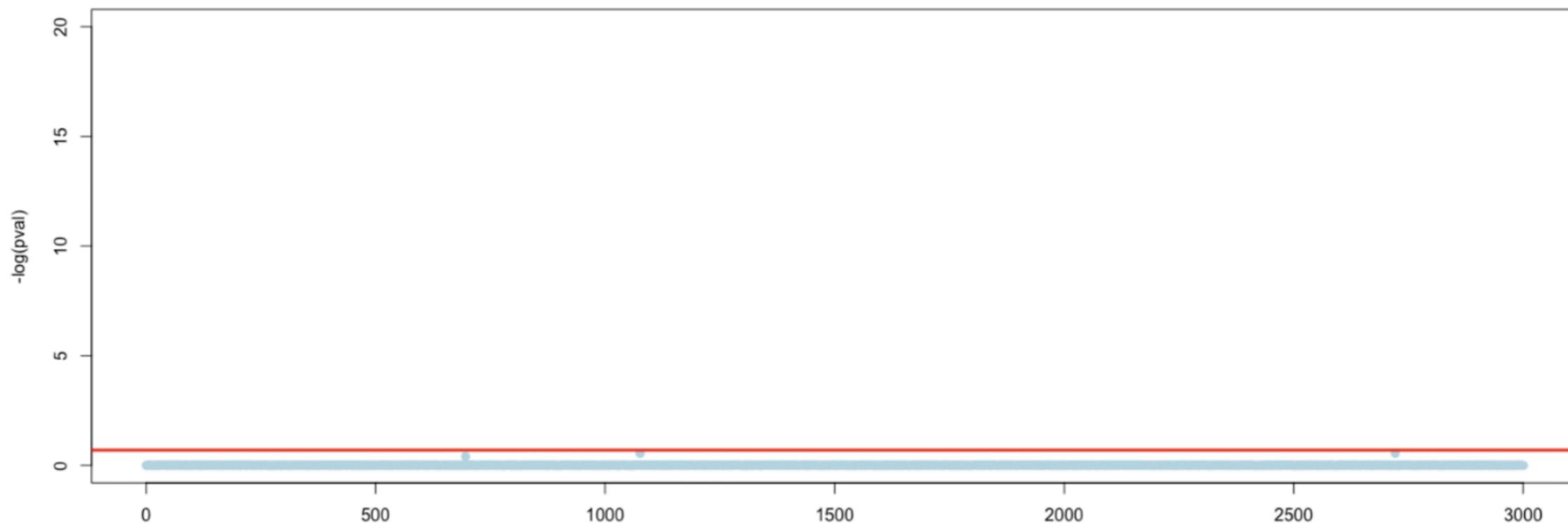
Если использовать поправку Бонферони, сколько раз нулевая гипотеза будет отвергнута с групповой вероятностью ошибки первого рода 5%?

Задача

k	p-value	Порог
1	0.0008	
2	0.0012	
3	0.0017	
4	0.0026	
5	0.0027	
6	0.032	
7	0.033	
8	0.048	
9	0.054	
10	0.062	

Эксперимент 1

Pvalues with Benjamini Hochberg correction, alpha=0.10



Adjusted p-values

Все активно использующиеся сравнения имеют похожий вид

$$p < \frac{\alpha}{C}$$

Вводим такую величину:

$$P_{adjusted} = p \cdot \alpha$$

Можем напрямую сравнивать с порогом

$$P_{adjusted} < \alpha$$

Adjusted p-value (случай one-step procedure)

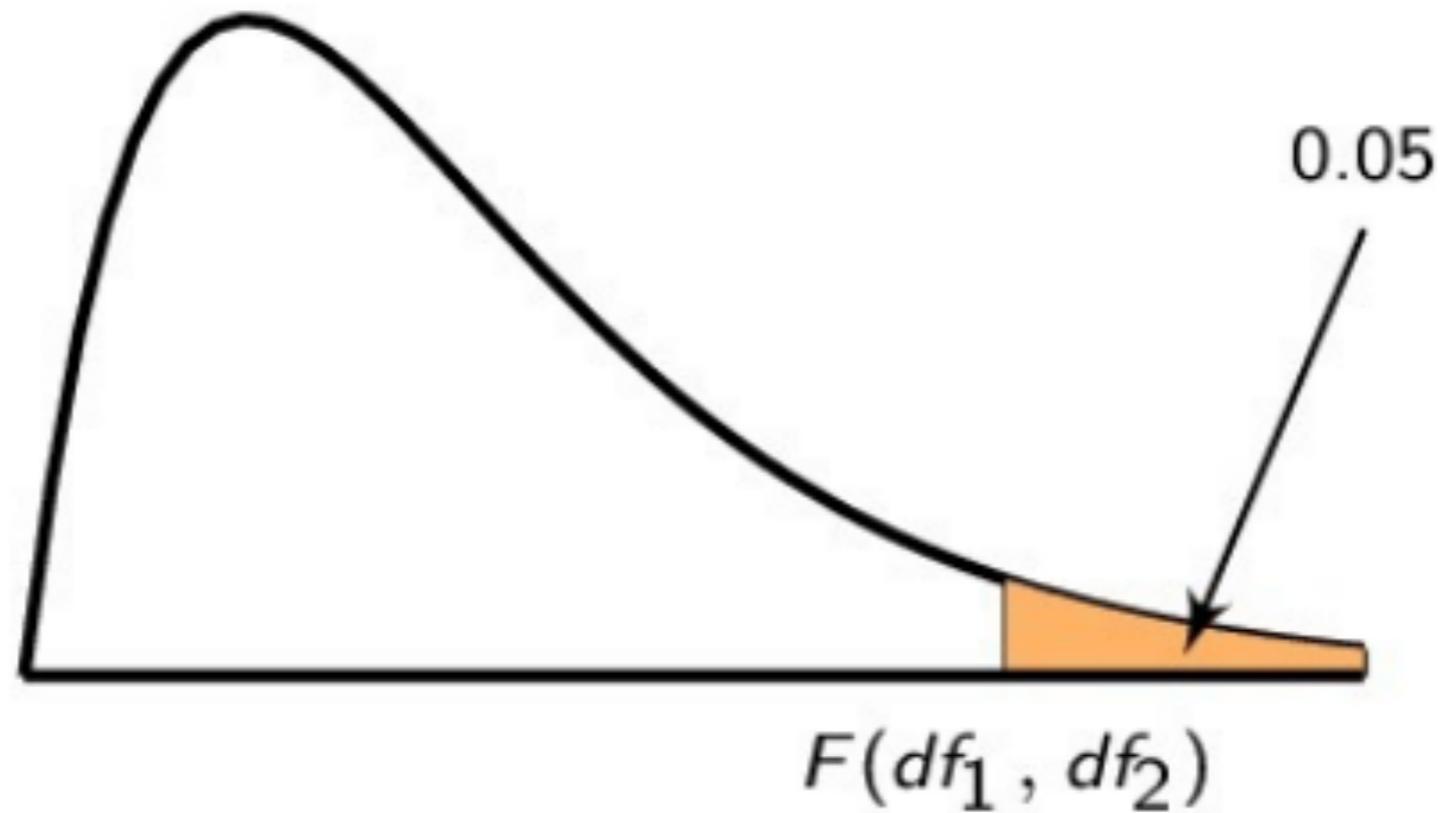
$$p < \frac{\alpha}{C}$$

$$P_{adjusted} = p \cdot C$$

$$P_{adjusted} < \alpha$$

Для не one-step procedure сложнее, но тоже можно

Тест Фишера



Тест Фишера (тест на равенство дисперсий)

$$\frac{s_x^2(n_x - 1)}{\sigma_x^2} \sim \chi_{n_x - 1}^2$$

$$\frac{s_y^2(n_y - 1)}{\sigma_y^2} \sim \chi_{n_y - 1}^2$$

Если

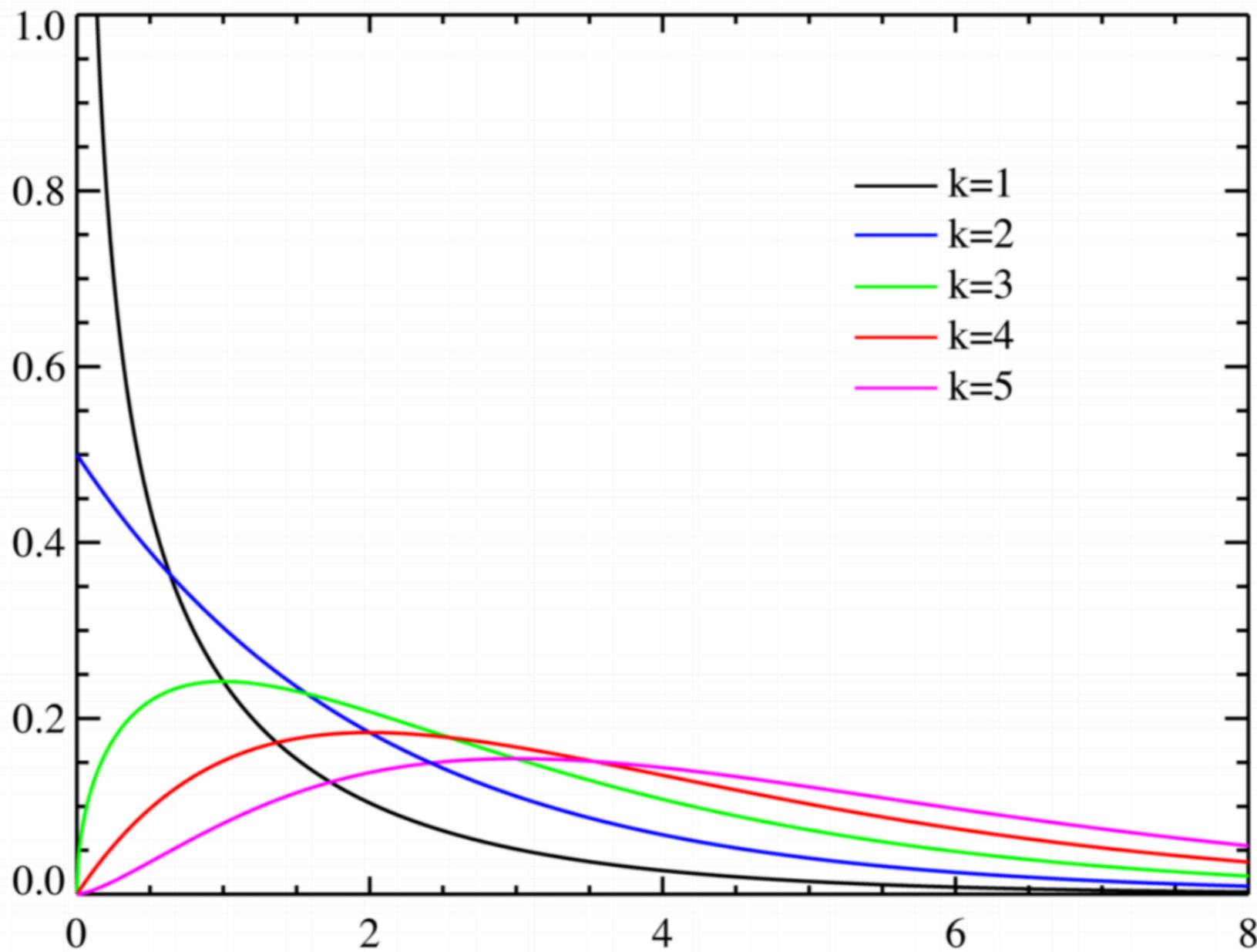
$$\sigma_x^2 = \sigma_y^2$$

$$\frac{\frac{s_x^2}{\sigma^2}}{\frac{s_y^2}{\sigma^2}} = \frac{\frac{s_x^2 \cdot (n_x - 1)}{\sigma_x^2} \cdot \frac{1}{n_x - 1}}{\frac{s_y^2 \cdot (n_y - 1)}{\sigma_y^2} \cdot \frac{1}{n_y - 1}} \sim \frac{\frac{\chi_{n_x - 1}^2}{n_x - 1}}{\frac{\chi_{n_y - 1}^2}{n_y - 1}} = F(n_x - 1, n_y - 1)$$

Problem 2.1 (Exercise laboratory problem revisited)

A hospital exercise laboratory technician notes the resting pulse rates of five joggers to be 60, 58, 59, 61, and 67, respectively, while the resting pulse rates of seven non-exercisers are 83, 60, 75, 71, 91, 82, and 84, respectively. The means and standard deviations for these samples are 61, 78, 3.54, and 10.23, respectively. Is equal variances assumption reasonable in this case?

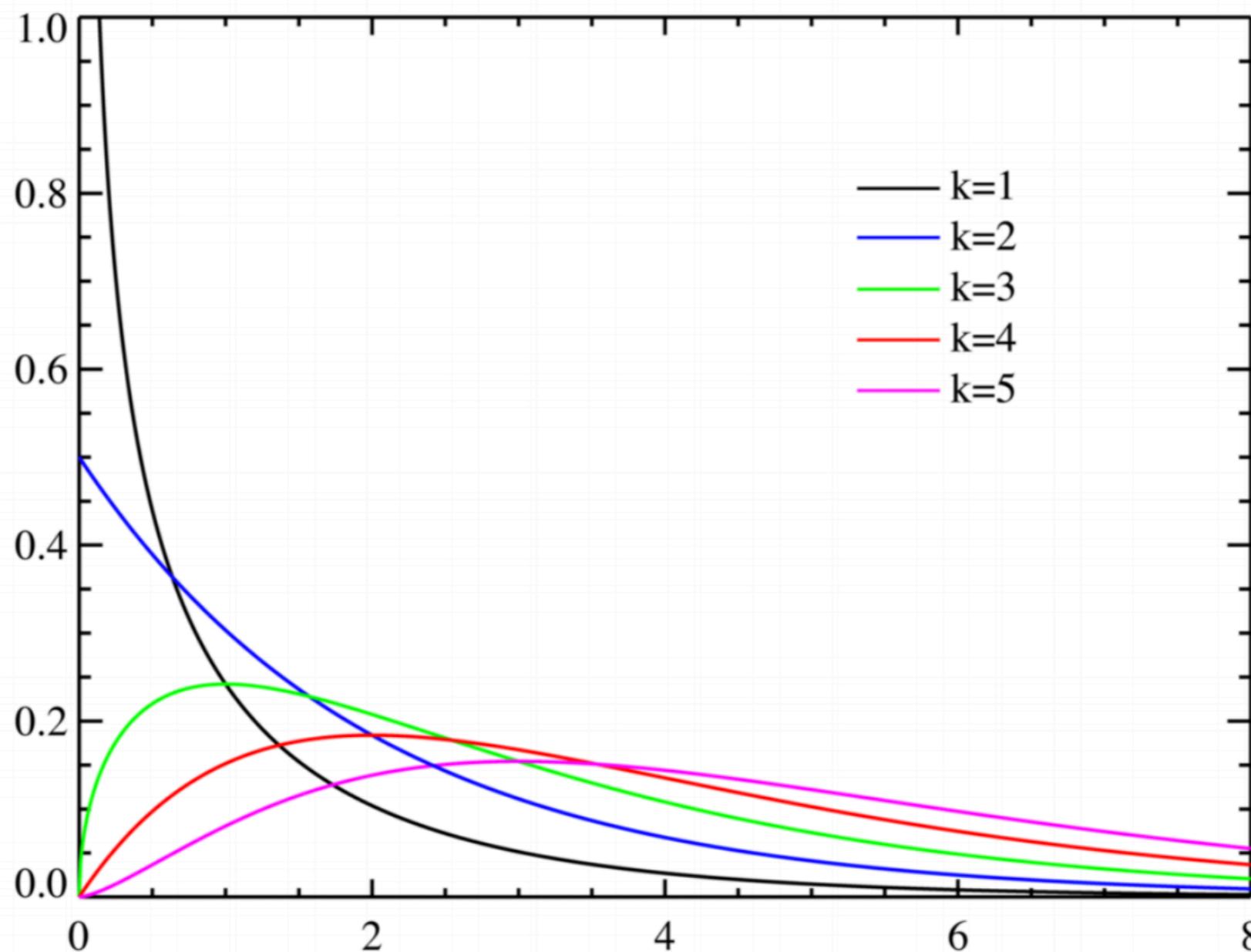
Распределение Хи-квадрат



Сумма к независимых стандартных нормальных
случайных величин

Правильно?

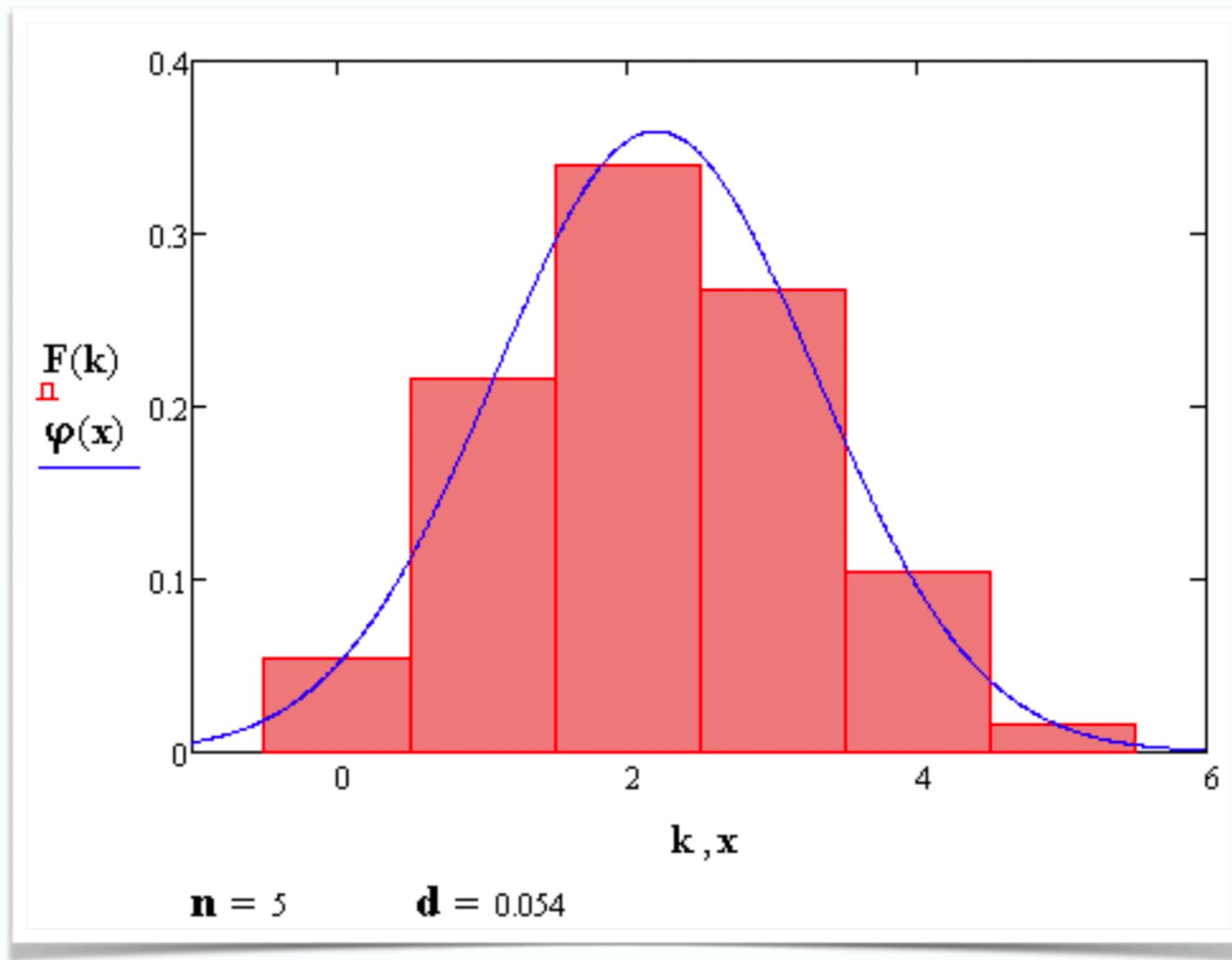
Распределение Хи-квадрат



Сумма квадратов k независимых случайных
стандартных нормальных величин

**Откуда они берутся в хи-
тесте ?**

Откуда они берутся в хи-тесте ?



Откуда они берутся в хи-тесте ?

Любой хи-квадрат тест - проверка того, что ваши данные распределены по мультиномиальному закону

Значение	val_0	val_1	val_2	val_3	val_4	...
Вероятность	p0	p1	p2	p3	p4	...

Критерий Хи-квадрат

Тест на Goodness of fit

Насколько ваша модель распределения данной переменной описывает реально наблюдаемые значения

H0: модель верна

H1: Модель неверна

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$df = n - 1$$

n - число ячеек

Задача

Программист Петя считает, что количество лайков, которые соберут посты с шутками на тему неприятных особенностей языка, одинаковы. Для теста были выбраны языки C++, Python, Javascript, Java и R. Количество лайков для постов про эти языки составило соответственно:

17, 23, 72, 44, 65

Прав ли Петя? Уровень значимости 0.001, так он не хочет никого в случае чего обидеть незаслуженно.

Задача

Программист Петя считает, что количество лайков, которые соберут посты с шутками на тему неприятных особенностей языка, одинаковы. Для теста были выбраны языки C++, Python, Javascript, Java и R. Количество лайков для постов про эти языки составило соответственно:

17, 23, 72, 44, 65

Прав ли Петя? Уровень значимости 0.001, так он не хочет никого в случае чего обидеть незаслуженно.

	C++	Python	Javascript	Java	R
Число лайков	17	23	72	44	65
Вероятность при условии H_0	0.20	0.20	0.20	0.20	0.20

Решение

Гипотеза Н0: Все языки получили равное число лайков, распределение лайков равномерное

Гипотеза Н1: Языки получили значимо разное число лайков

Если распределение лайков равномерное, то ожидаемое число лайков для каждого языка:

$$E = 221/5 = 44.2$$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} = 56.2$$

$$df = n - 1 = 4$$

$$P(\chi^2(4) > 56.2) = 1e - 11 < 0.001$$

На уровне значимости 0.01 мы отвергаем гипотезу Н0 о том, что распределение лайков равномерное

Goodness of fit

- Всегда ли число степеней свободы $n - 1$?

Goodness of fit

- Всегда ли число степеней свободы $n - 1$?

Нет

Оно зависит от того, сколько условий вы накладываете

Goodness of fit

- Всегда ли число степеней свободы $n - 1$?

Нет

Оно зависит от того, сколько условий вы накладываете на ваши наблюдения/величин из них считаете непосредственно до теста

В предыдущем случае есть только одно условие
- сумма всех наблюдений равна n .

Потому из числа наблюдений (n) мы и вычитаем 1

Задача

Наблюдается число студентов, опаздывающих на 0 минут, минуту, две, три, четыре и 5 минут и более

Значение	0	1	2	3	4	5 и более
Студентов	14	30	33	14	6	3

Проверьте гипотезу о том, что число студентов распределено по Пуассону

Решение

Наблюдается число студентов, опаздывающих на 0 минут, минуту, две, три, четыре и 5 минут (больше не опаздывают)

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3

Если число студентов распределено по Пуассону, то $\lambda = (14 * 0 + 30 * 1 + 2 * 33 + 3 * 14 + 4 * 6 + 5 * 3) / 100 = 1.77$

Можно подсчитать (по формуле или с использованием функции dpoiss R вероятность значения попасть в каждую из ячеек)

Решение

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3
p	0.17	0.30	0.27	0.16	0.07	0.02

Остается подсчитать ожидаемое число студентов

Значение	0	1	2	3	4	5
Студентов	14	30	33	14	6	3
p	0.17	0.30	0.27	0.16	0.07	0.02
Ожидаемое	17	30	27	16	7	2

Решение

Теперь можно подсчитать значение статистики, оно равно 2.76

В этом случае у нас было условие на то, что наблюдений суммарно $N = 100$ и на то, что λ нашего Пуассоновского распределения равна 1.77 (мы ее считали из наших наблюдений)

Потому суммарно получаем число степеней свободы равным $n - 2 = 6 - 2 = 4$.

Получаем, что $p\text{-value}$ близко к 1, то есть у нас нет оснований отвергать гипотезу о том, что наблюдения распределены по Пуассону.

Решение

Сколько степеней свободы надо в случае:

- Проверить гипотезу о том, что наши наблюдения распределены по нормальному закону?
- Проверить гипотезу о том, что наши наблюдения распределены по равномерному закону?
- Проверить гипотезу о том, что наши наблюдения распределены по Пуассону с параметром 2?

Решение

Сколько степеней свободы надо в случае:

- Проверить гипотезу о том, что наши наблюдения распределены по нормальному закону?
 $n - 1 - 2 = n - 3$, считаем среднее и дисперсию
- Проверить гипотезу о том, что наши наблюдения распределены по равномерному закону?
 $n - 1$, мы это делаем по-умолчанию
- Проверить гипотезу о том, что наши наблюдения распределены по Пуассону с параметром 2?
 $n - 1$, мы взяли параметр не из наблюдений

Производитель шоколада заявляет, что его шоколадные "яйца" содержат игрушки трех типов в соотношении 1:2:3. Случайная выборка из 90 яиц содержит 19, 33 и 38 игрушек каждого типа, соответственно. Есть ли основания недоверять заявлению производителя на 5% уровне доверия?

- A. Нет, поскольку пи-велью ниже 5%
- B. Нет, поскольку пи-велью между 5% и 10%
- C. Нет, поскольку пи-велью выше 10%
- D. Да, поскольку пи-велью ниже 5%
- E. Да, поскольку пи-велью между 5% and 10%

Критерий Хи-квадрат

Тест на независимость

Используется как на то, есть ли значимая ассоциация между двумя факторными переменными

H0: факторы независимы

H1: факторы зависимы

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$df = (n - 1) \cdot (m - 1)$$

Где df - число степеней свободы, n - число разных значений первой переменной, m - число разных значений второй

Задача

Для четырех категорий людей - школьников, студентов, программистов (закончивших учебу со стажем < 5 лет и программистов (закончивших учебу) со стажем больше 5 лет имеются данные о их отношении к PHP. Отношение может быть “хороший язык”, “ну а что поделать” “ненавижу”. Проверить гипотезу о том, что категории независимы. Уровень значимости принять равным 0.01

Отношение/ Категория	Школьники	Студенты	Программис т, < 5 лет	Программис т, > 5 лет
Хороший язык	40	22	17	12
Ну а что поделать	15	12	20	35
Ненавижу	35	20	22	10

Решение

Гипотеза H0: Отношение не зависит от категории

Гипотеза H1: Отношение зависит от категории

Если отношение не зависит от категории, то $P(\text{хороший язык, категория}) = P(\text{хороший язык}) * P(\text{категория})$. То есть вероятность объекта оказаться в ячейке - произведение вероятностей в соответствующих столбце и строке

Отношение /Категория	Школьник и	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	40	22	17	12	91	0.35
Ну а что поделать	15	12	20	35	82	0.32
Ненавижу	35	20	22	10	87	0.33
Сумма	90	54	59	57	260	
Вероятность	0.35	0.21	0.23	0.22		-

Решение

Тогда ожидаемые нами числа:

Отношение /Категория	Школьник и	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	31,85	19,11	20,93	20,02	91	0,35
Ну а что поделать	29,12	17,472	19,136	18,304	82	0,32
Ненавижу	30,03	18,018	19,734	18,876	87	0,33
Сумма	90	54	59	57	260	
Вероятность	0,35	0,21	0,23	0,22		-

Решение

Посчитаем значение критерия хи-квадрат

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi^2 = 35.8$$

$$df = (4 - 1) \cdot (3 - 1) = 6$$

$$P(\chi^2(6) > 33.8) = 0.0004 < 0.01$$

На уровне значимости 0.01 мы отвергаем гипотезу H_0 о независимости

Проблемы с критерием Хи-квадрат

Критерий Хи-квадрат можно применять только тогда, когда ожидаемое число наблюдений в каждой клетке больше 5.

Иначе необходимо использовать точный тест Фишера

Менеджер компании хочет узнать есть ли взаимосвязь между географическим регионом проживания и наличием компьютера Макинтош. Она опрашивает 100 человек и получает следующие данные:

	Есть Мак	Нет Мака	Всего
Северо-восточный	12	14	26
Юго-западный	26	13	39
Средний запад	17	18	35
Total	55	45	100

Для проверки гипотезы о независимости этих двух факторов, чему равны тестовая статистика и соответствующее критическое значение на 5% уровне значимости?

Точный тест Фишера

	Исход есть	Исхода нет	Всего
Фактор есть	A	B	A + B
Фактора нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

$$p(table) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

В чем проблема?:

Точный тест Фишера

	Исход есть	Исхода нет	Всего
Фактор есть	A	B	A + B
Фактора нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

$$p(table) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

В чем проблема?:?

Мы получили точечную оценку. Для получения p-value нам надо посчитать весь хвост (односторонний тест) или оба хвоста (двустронний тест)

Точный тест Фишера

**Левый хвост, сложить
вероятности всех
таблиц здесь**

**Правый хвост,
сложить вероятности
всех таблиц здесь**

Все хорошо

**Таблица, перекошенная, как
наша, но в другую сторону**

**Таблицы с
еще более
перекошенно
й в другую
сторону
связью**

	Исход есть	Исход а нет	Всего
Факто р есть	A	B	A + B
Факто ра нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

Наша таблица

	Исход есть	Исход а нет	Всего
Факто р есть	A	B	A + B
Факто ра нет	C	D	C + D
Всего	A + C	B + D	A + B + C + D

**Таблицы с еще
более
перекошенной
в нашу сторону
связью**

Пример

	Юноши	Девушки	Всего
На диете	1	9	10
Без диеты	11	3	14
Всего	12	12	24

Гипотеза Н0: Юноши и девушки сидят на диетах одинаково

Гипотеза Н1: Девушки сидят на диетах чае

$$p(\text{table}) = ?$$

Пример

Для вычисление p-value нам надо посчитать еще все таблицы, которые критичнее нашей, в данном случае она одна..

	Юноши	Девушки	Всего
На диете	0	10	10
Без диеты	12	2	14
Всего	12	12	24

$$p(table_1) = ?$$

$$Pvalue = ?$$