Статистика 3

Задача

Петя заявляет, что может в среднем бросить стандартный теннисный мяч на расстояние 20м. В ходе серии из 45 опытов среднее расстояние, на которое он смог бросить мяч, равно 19.4. Стандартное отклонение расстояний равно 3.5
Что можно утверждать о Петином заявлении на уровне значимости 0.05?

Решение

H0: Петя в среднем бросает мяч на 20 м.
H1: Петя в среднем бросает мяч меньше, чем на 20 м

Экспериментов больше, чем 40, потому в анализе можем пользоваться нормальным распределением

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{45}} = 0.52$$

$$z\_value = \frac{\overline{X} - 20}{SE} = \frac{19.4 - 20}{0.52} = -1.15$$

$$p\_value = P(z < z\_value) = 0.125$$

На уровне значимости 0.05 мы не имеем достаточно оснований для отвердения H0

Задача

Петя заявляет, что может в среднем бросить стандартный теннисный мяч на расстояние 20м. В ходе серии из 45 опытов среднее расстояние, на которое он смог бросить мяч, равно 19.4. Стандартное отклонение расстояний равно 3.5 Построить 95% двусторонний доверительный интервал для среднего расстояния, на которое может бросить Петя.

Экспериментов больше, чем 40, потому в анализе можем пользоваться нормальным распределением

$$\mu = \overline{X} \pm z_{\alpha/2} \cdot SE$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3.5}{\sqrt{45}} = 0.52$$

$$z_{\alpha/2} = z_{(1-0.95)/2} = z_{0.025} = 1.96$$

$$\mu = \overline{X} \pm z_{\alpha/2} \cdot SE = 19.4 \pm 1.96 \cdot 0.52$$

$$\mu \in [18.3,\ 20.42]$$

Задача

Биоинформатик Федя и биоинформатик Витя любят перекидывать друг другу задачи. Происходит это следующим образом - каждый из них какое-то время делает малую часть задачи, а затем посылает все остальное второму. Известно, что чем задача скучнее, тем быстрее происходит ее перекидывание. Имеется два долгосрочных проекта. Для первого среднее время нахождения проекта у одного из биоинформатиков 8.4 дня, а для второго - 10.5. Стандартные отклонения равны 1.7 и 2.2. Количество перекидывай для первого проекта - 14, для второго - 11. Можно ли утверждать, что первая задача более скучная на уровне значимости 0.01? Предполагаем равные дисперсии.

Решение

H0: Задачи одинаково скучные
H1: Первая задача более скучная

Мы не знаем, реальных дисперсий для наших выборок, но знаем, что они равны. Потому нам необходимо для начала его подсчитать.

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} = 1.93$$

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.93 \cdot \sqrt{\frac{1}{14} + \frac{1}{11}} = 0.77$$

Кроме того, мы теперь не можем использовать нормальное распределение, так данных недостаточно и мы используем выборочное стандартное отклонение.

$$t\_value = \frac{8.4 - 12.7}{0.77} = -2.73$$

Решение

Мы не знаем, реальных дисперсий для наших выборок, но знаем, что они равны. Потому нам необходимо для начала его подсчитать.

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} = 1.93$$

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.93 \cdot \sqrt{\frac{1}{14} + \frac{1}{11}} = 0.77$$

Кроме того, мы теперь не можем использовать нормальное распределение, так данных недостаточно и мы используем выборочное стандартное отклонение.

$$t\_value = \frac{8.4 - 12.7}{0.77} = -2.73$$

$$df = n_1 + n_2 - 2 = 14 + 11 - 2 = 23$$

$$p\_value = P(t < -2.73) = 0.0059$$

Значит мы имеем достаточно оснований для отвержения гипотезы H0 на уровне значимости 0.01

Множественное тестирование

FDR

FDR (False discovery rate) - метод коррекции на множественное тестирование, когда нам гарантируется, что доля ложноположительных результатов будет не больше определенного порога alpha

Benjamini-Hochberg correction

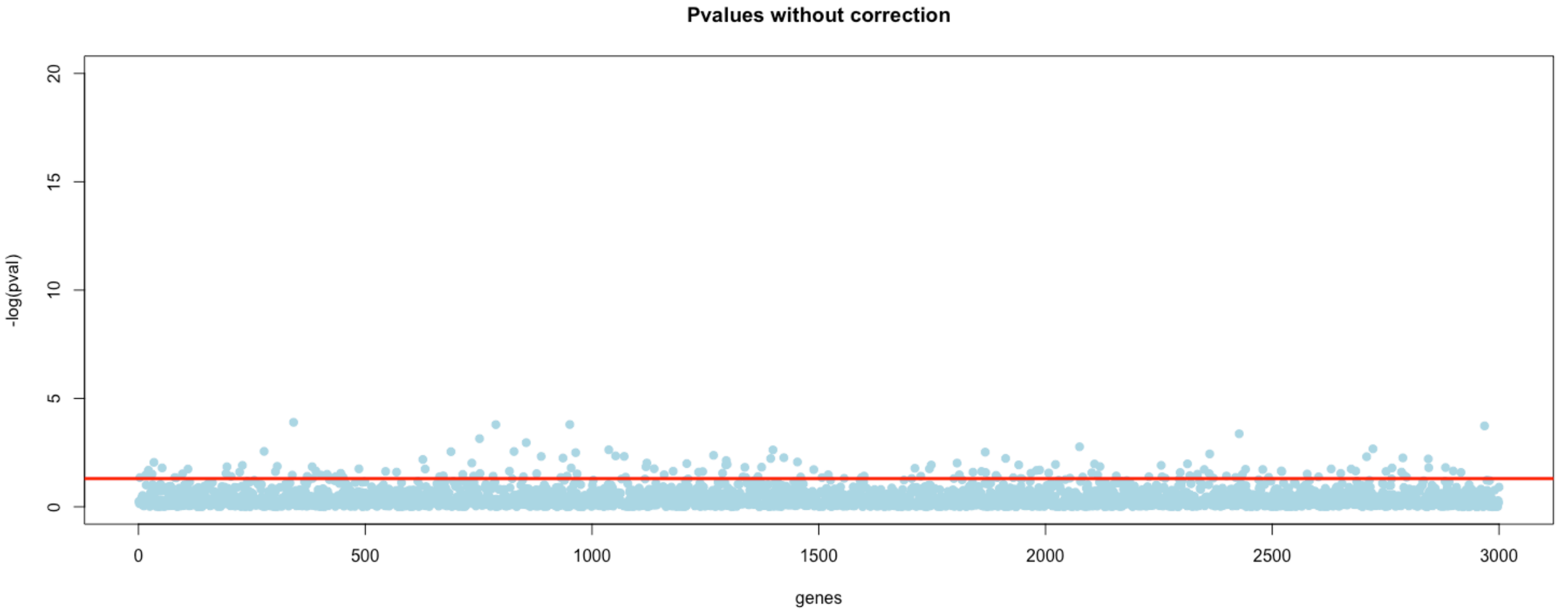
Предполагает независимость проводимых тестов и их результатов (p-value), либо их положительную зависимость (если один тест имеет низкое p-value, то остальные так же имеют тенденцию иметь низкое p-value)

Benjamini-Hochberg-Yekutieli correction

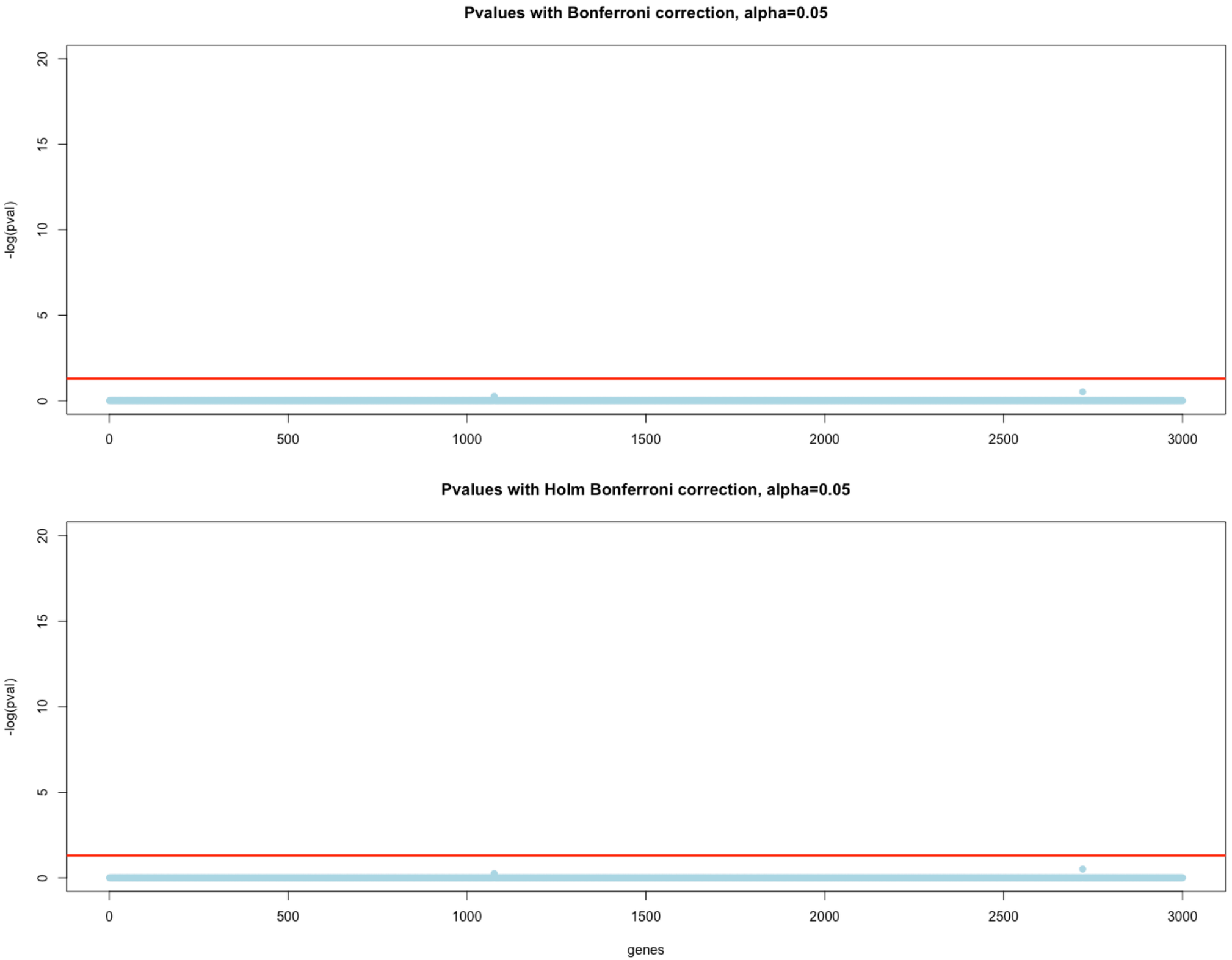
Мы снимаем ограничение на независимость тестов, но взамен наш тест имеет меньшую мощность (вероятность ошибки 2 рода выше)

Эксперимент 1

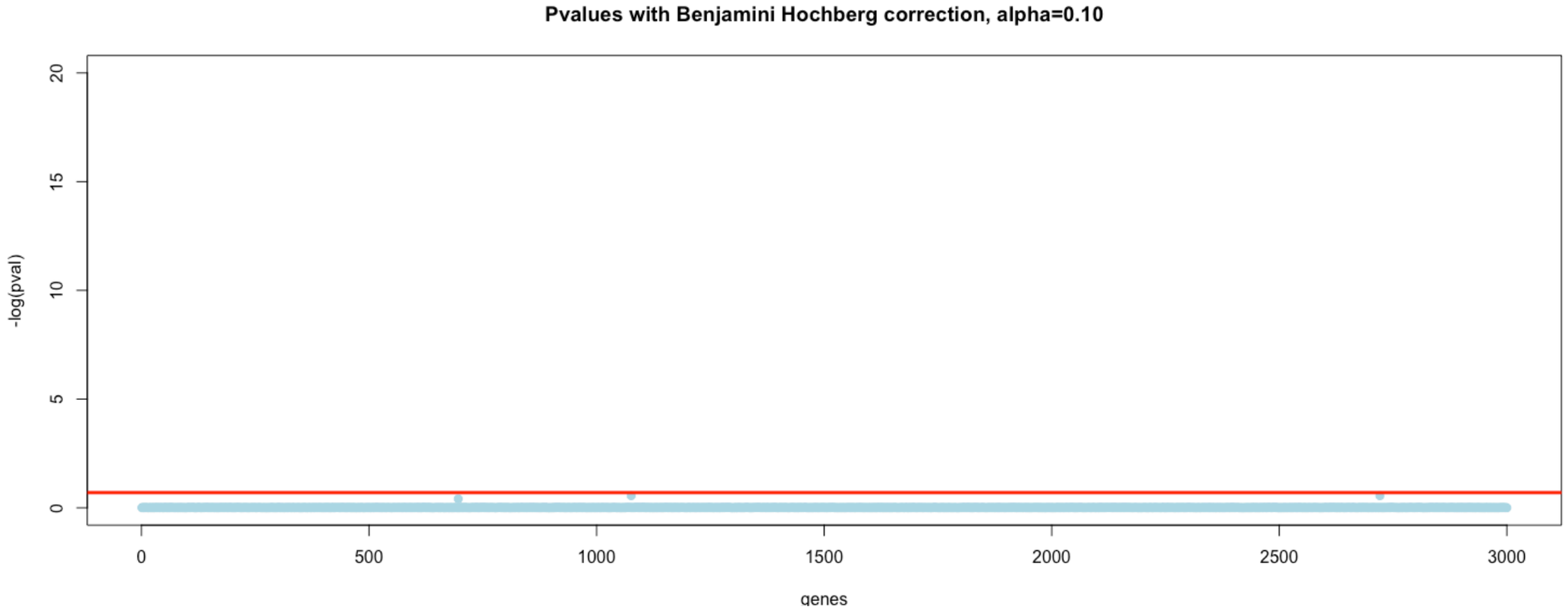
Возьмем и симулируем набор из 50 пациентов с 3000 генов, которые не меняют свою экспрессию значимо по ходу эксперимента (имеем результаты до и после). Значения “экспрессии” генов будет брать из нормального распределения.

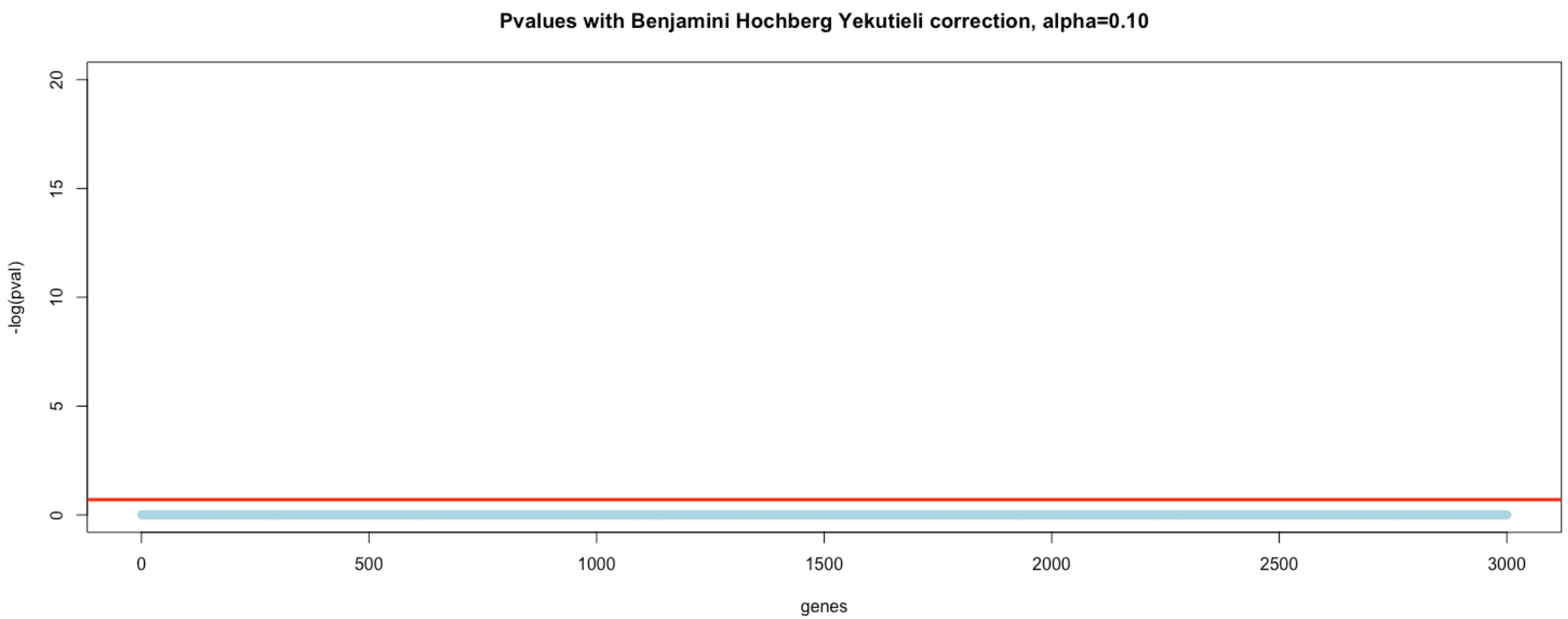


Эксперимент 1



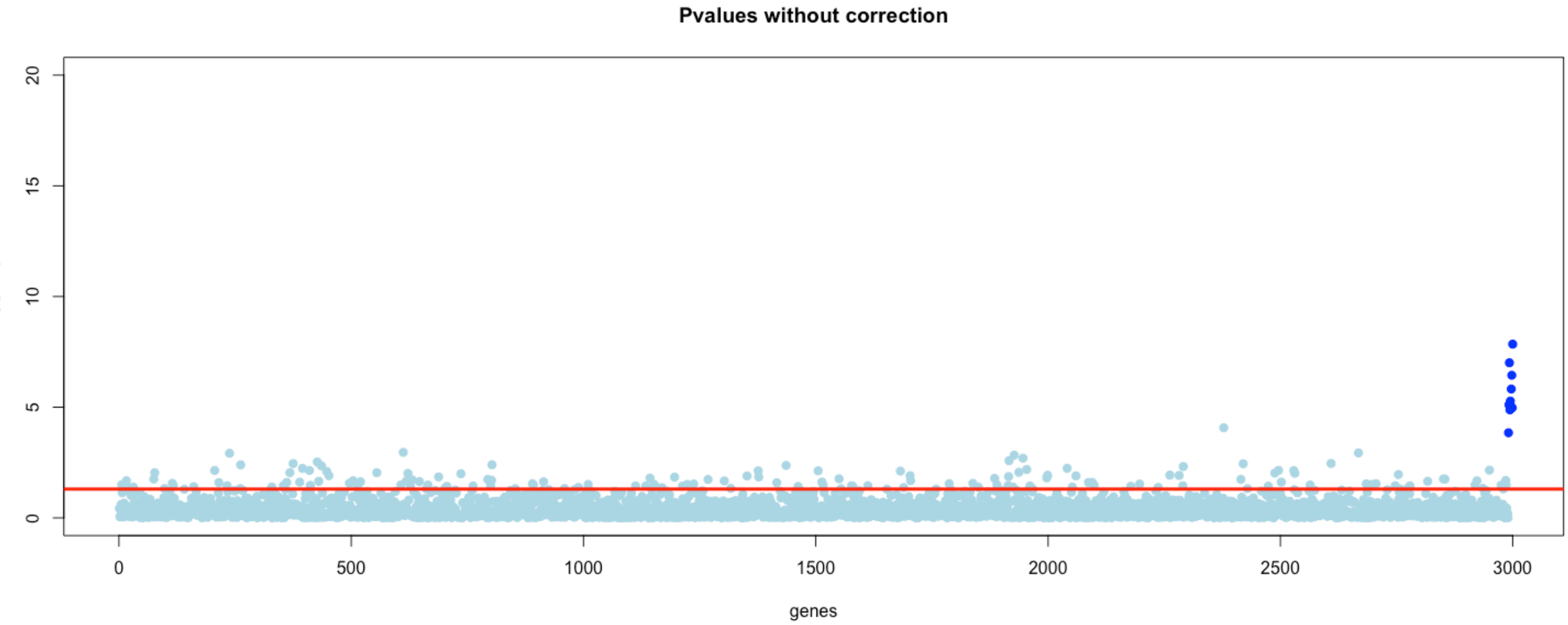
Эксперимент 1



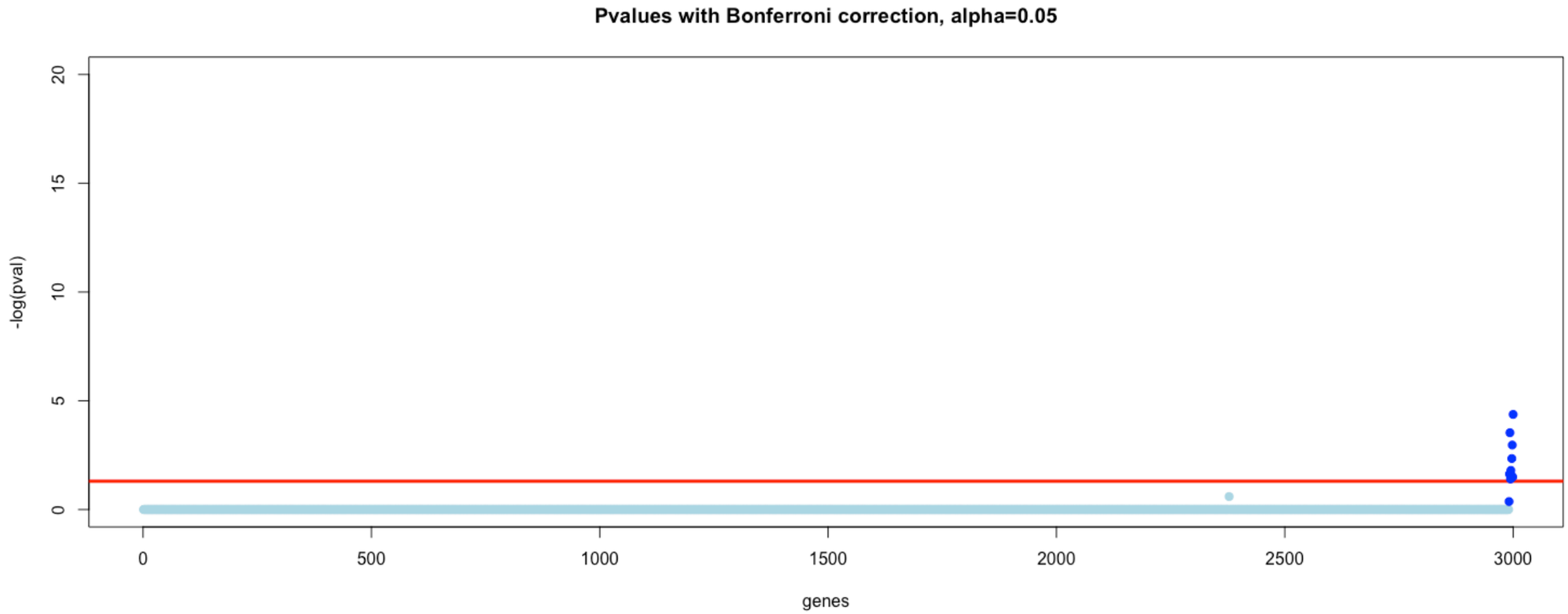


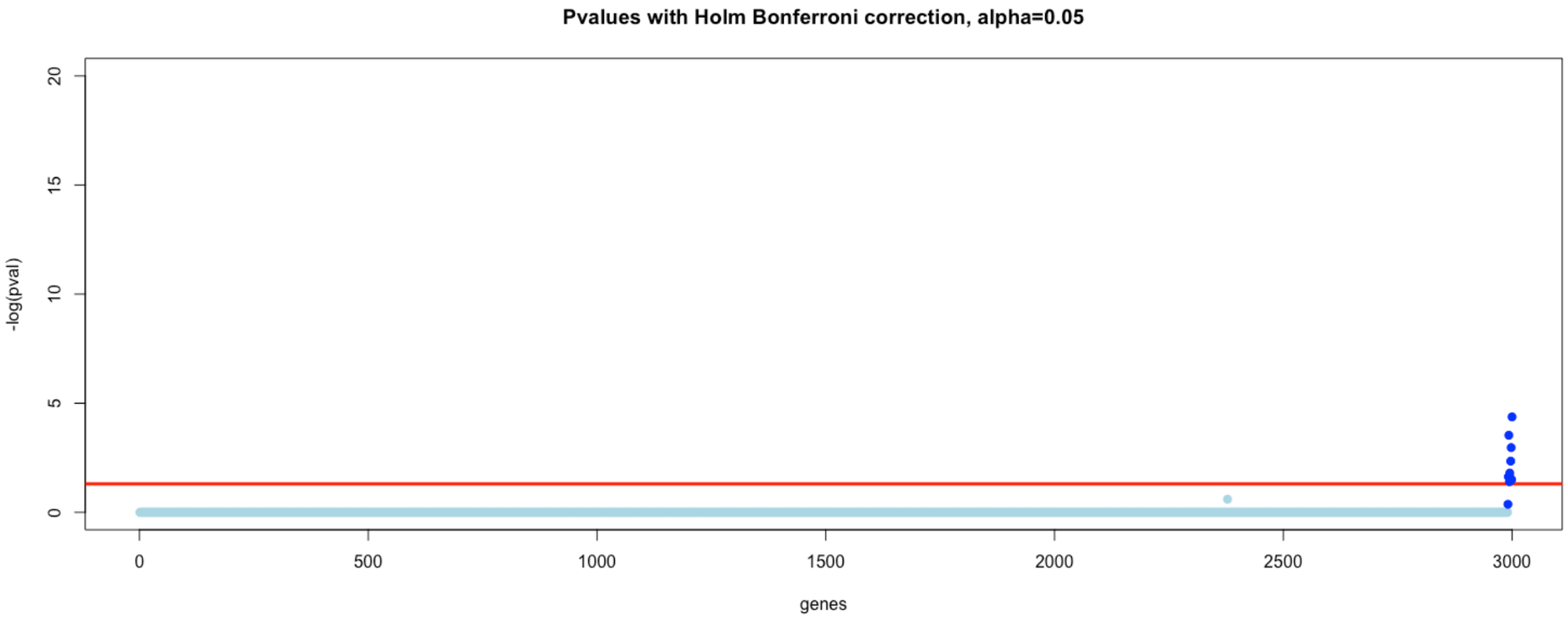
Эксперимент 2

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.

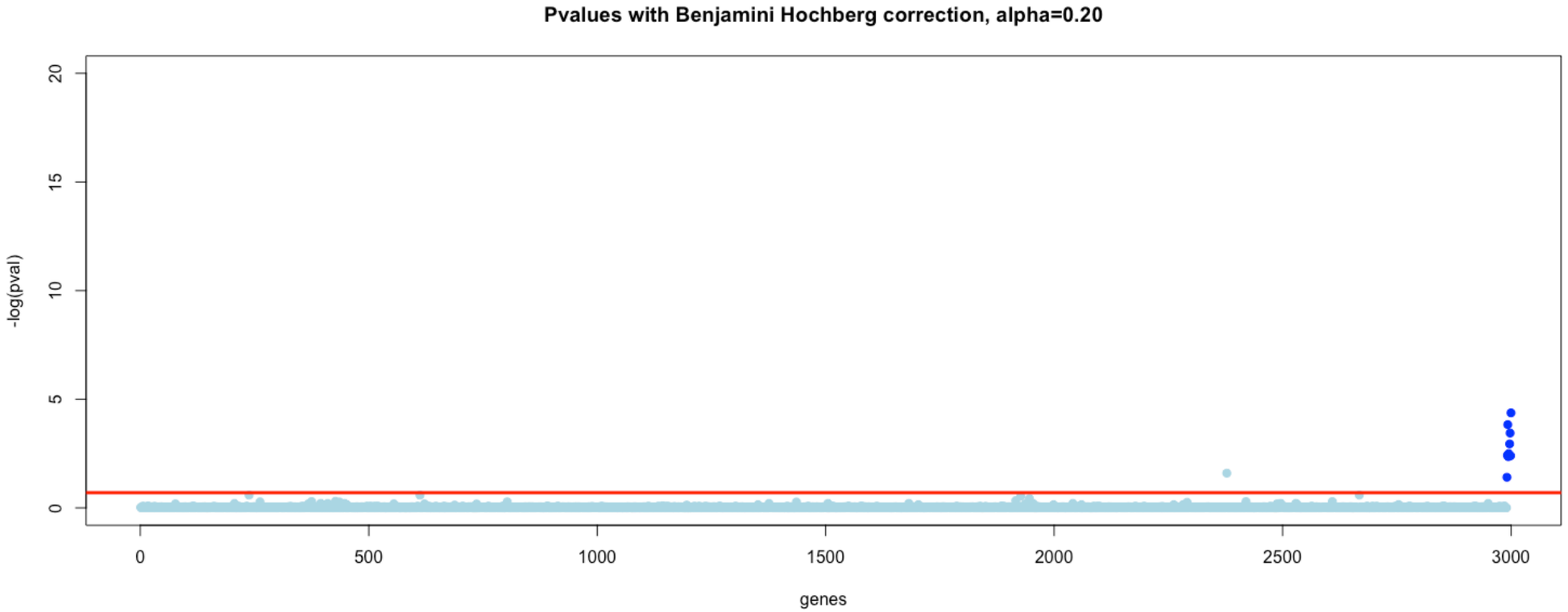


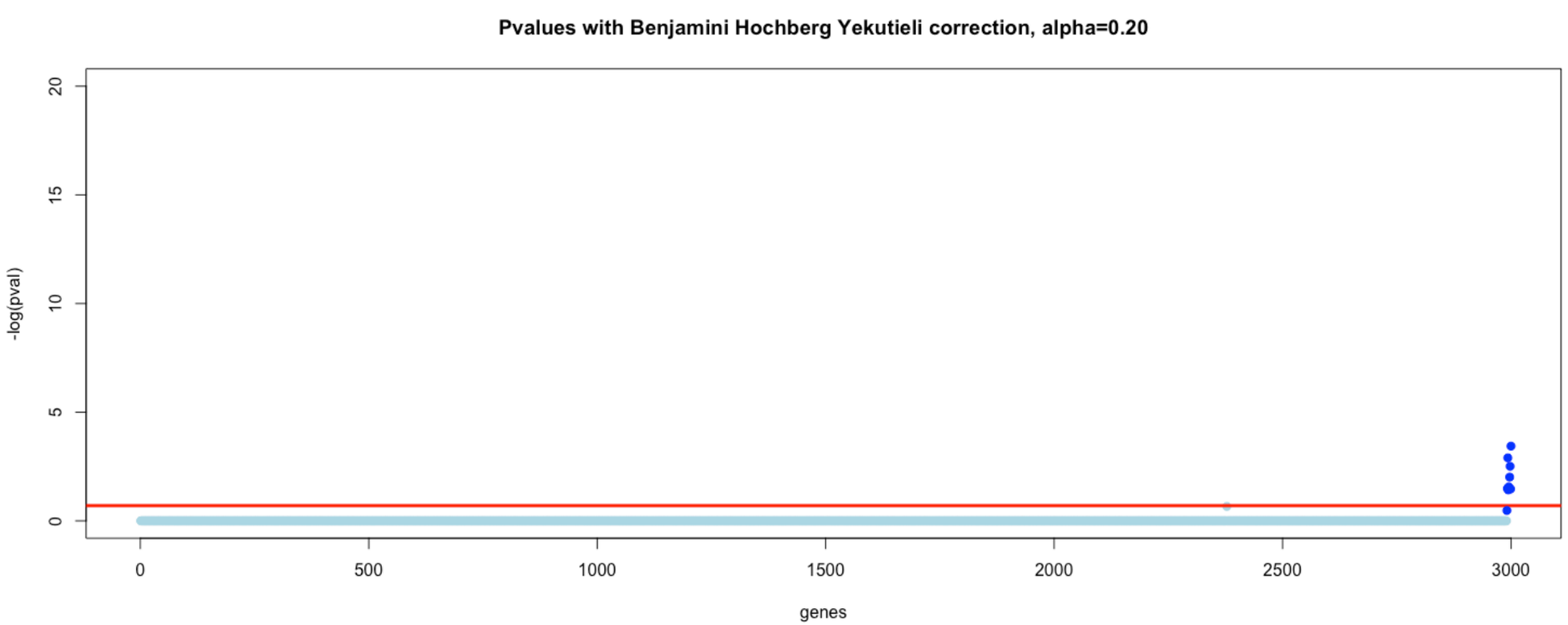
Эксперимент 2





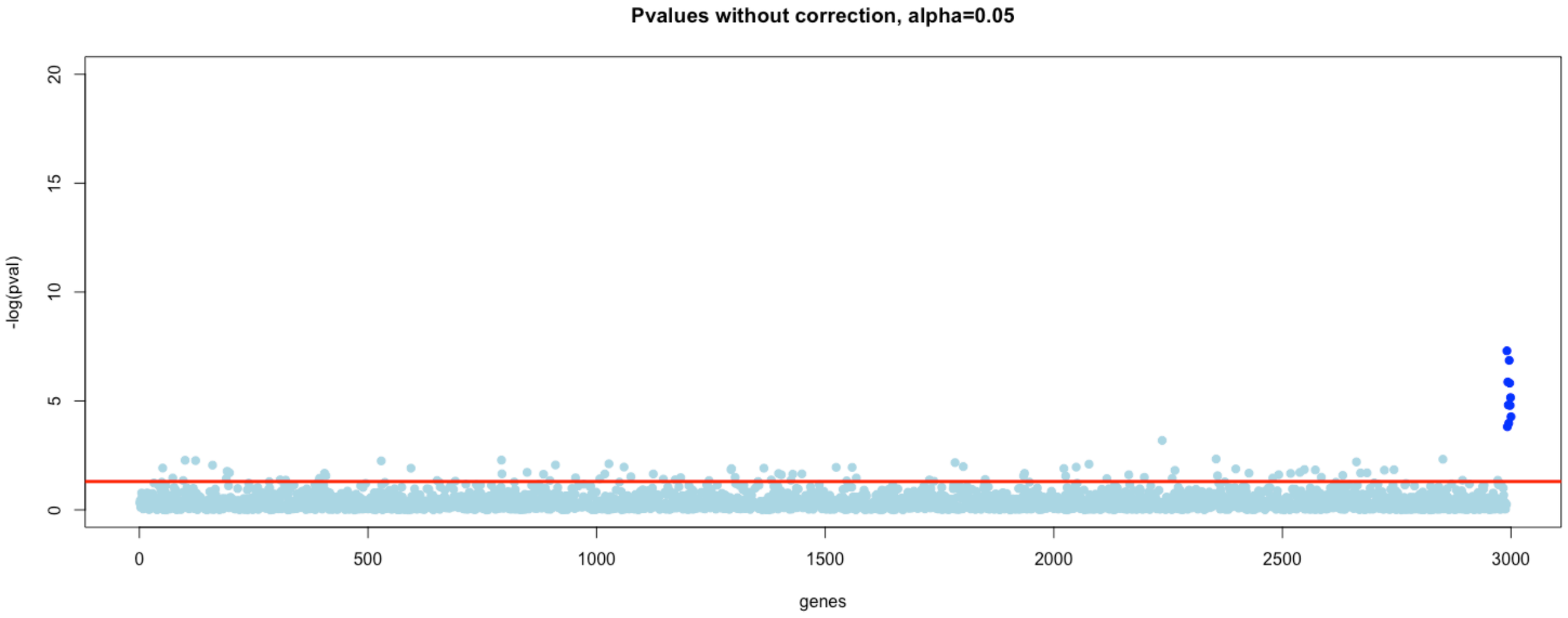
Эксперимент 2



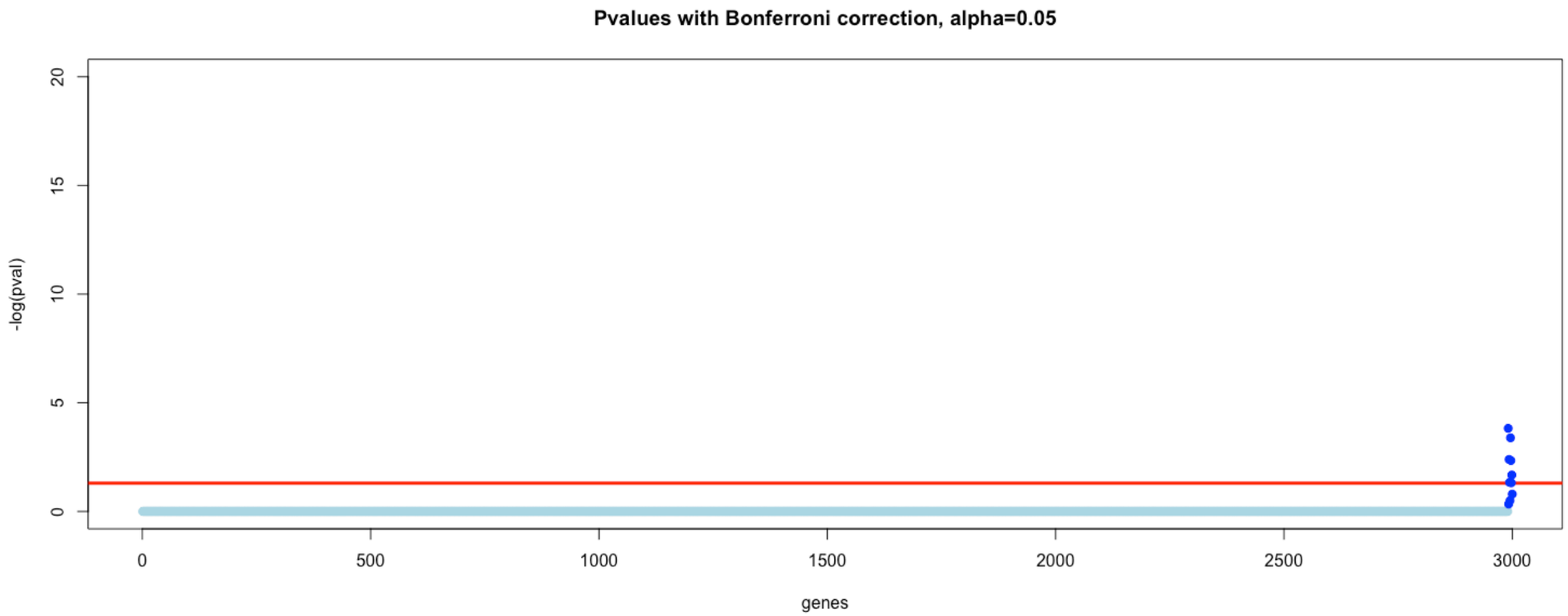


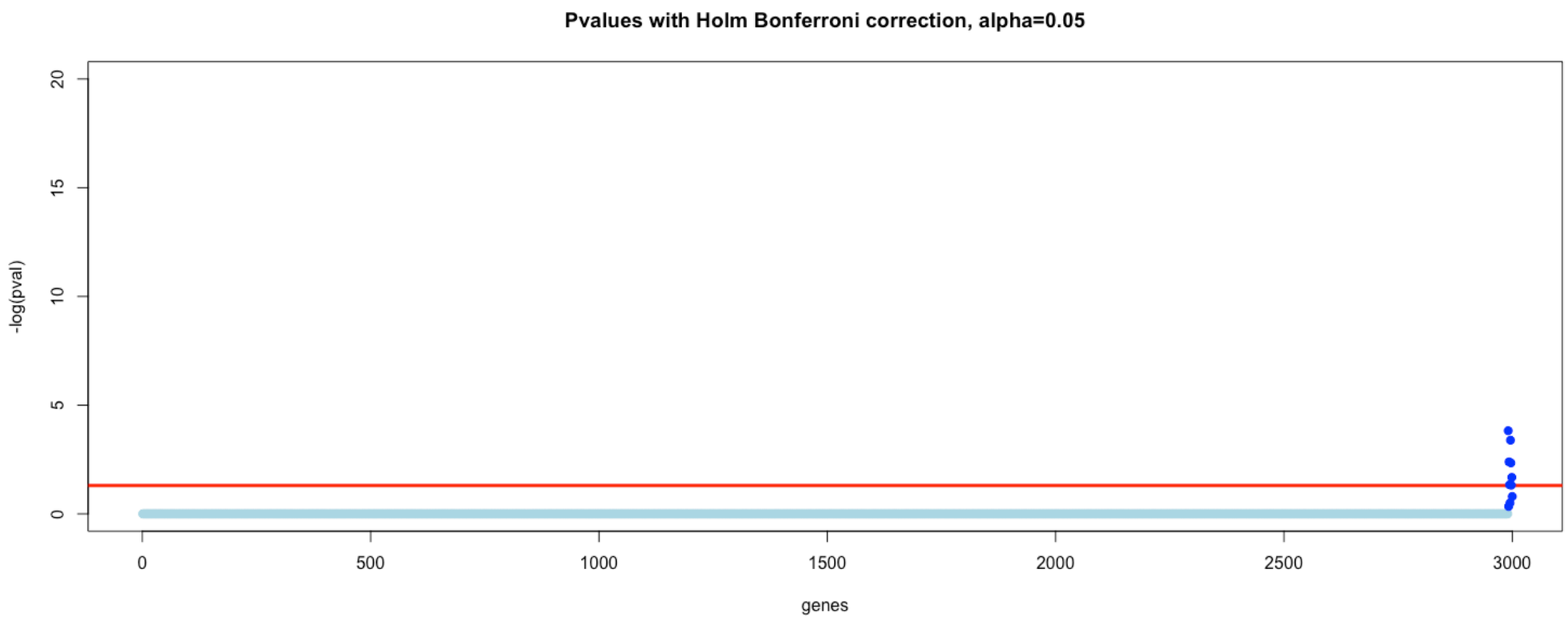
Эксперимент 3

Возьмем реальные данные о весе мышек, не сидевших ни на каких диетах. Будем отбирать случайным образом часть мышей в “контроль”, а другую часть - в “эксперимент” и проводить t-test. Сделаем такую процедуру 3000 раз, 10 из которых будем дополнительно смещать “эксперимент”, обеспечивая значимость изменения в этих случаях

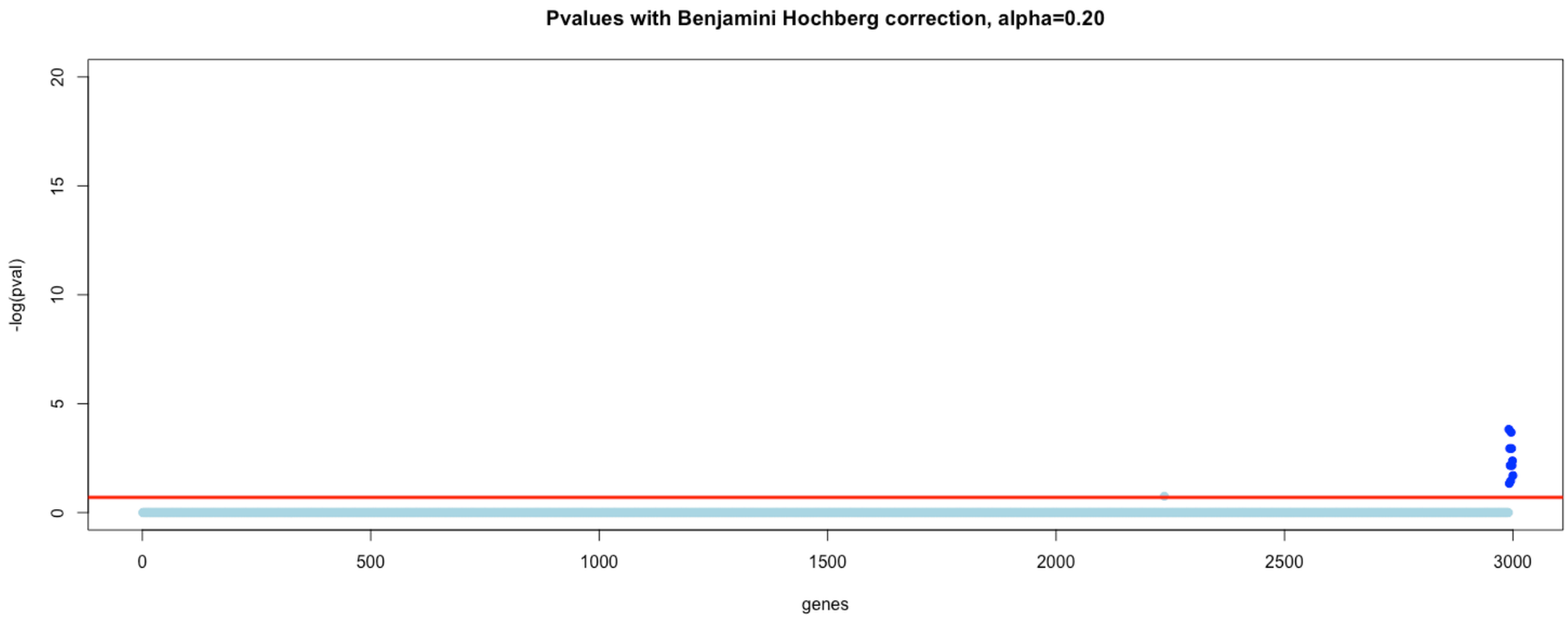


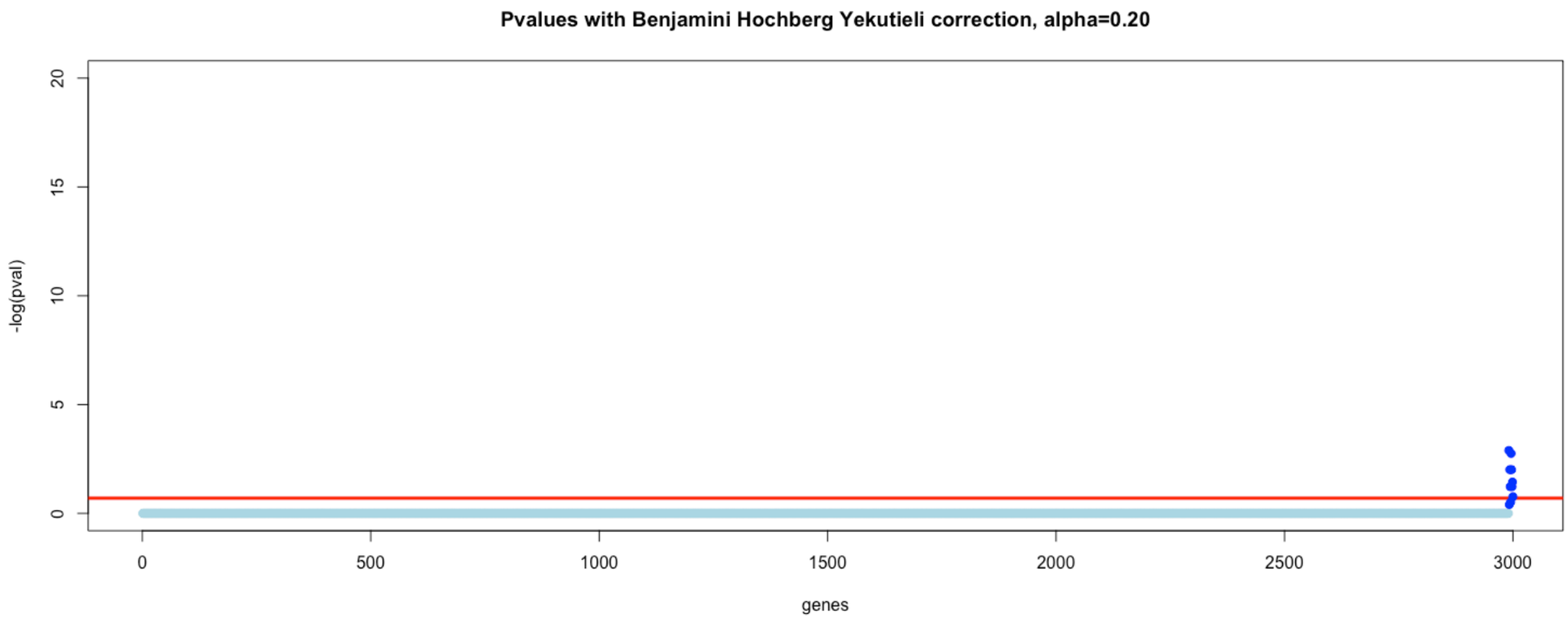
Эксперимент 3





Эксперимент 3





Важно!

Обратите внимание, что мы использовали разные alpha для FWER и FDR

Почему?

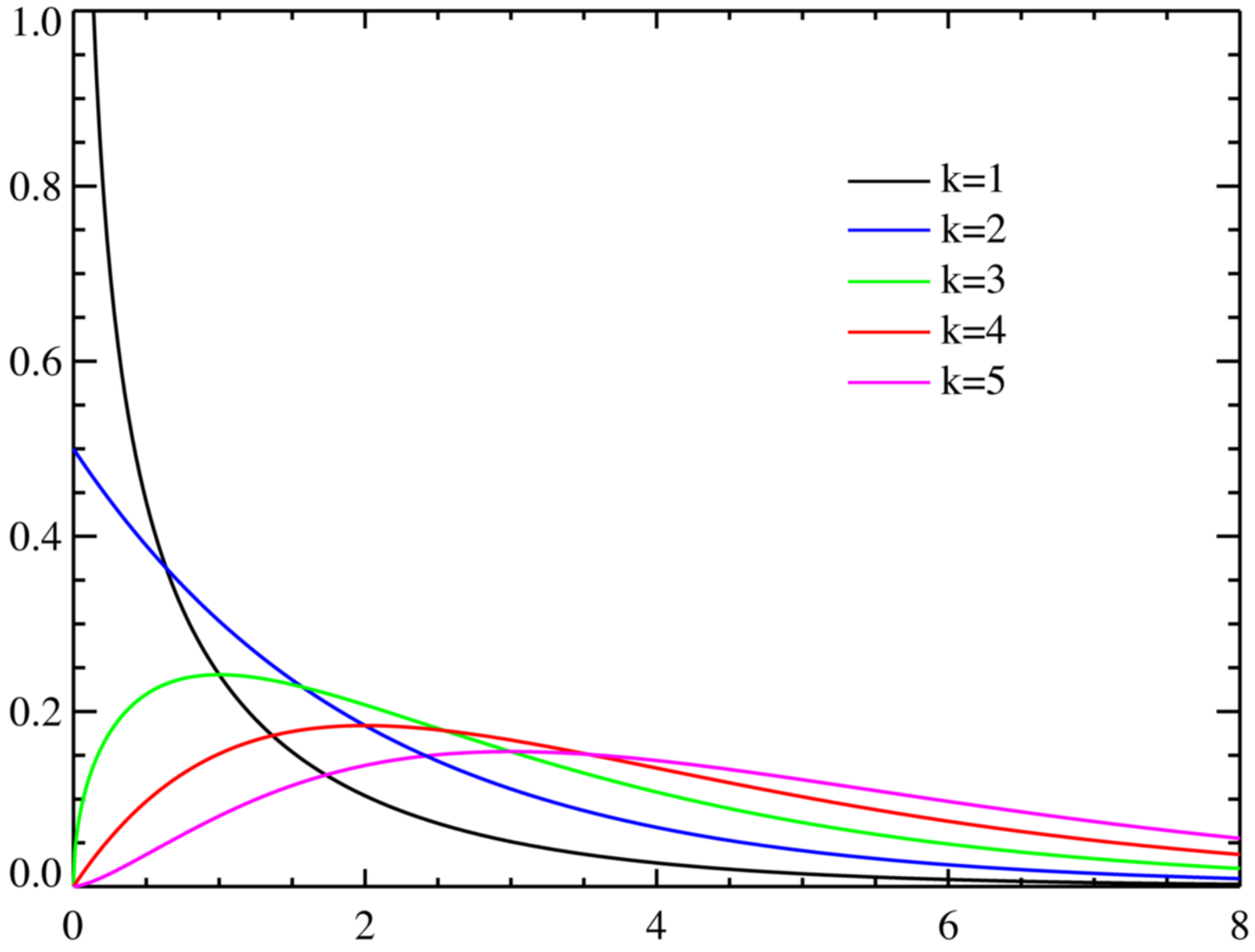
Важно!

Обратите внимание, что мы использовали разные alpha для FWER и FDR

Почему?

Мы так сделали, так как это разные числа! В случае FWER это вероятность, что хоть один ген из отобранных будет ложноположительным, в случае FDR - это максимальная доля ложноположительных генов среди отобранных

Распределение Хи-квадрат



Сумма k независимых стандартных нормальных случайных величин

Критерий Хи-квадрат

Ваш любимый…
Существует в двух вариациях

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

Критерий Хи-квадрат

Тест на Goodness of fit

Насколько ваша модель распределения данной переменной описывает реально наблюдаемые значения

H0: модель верна

H1: Модель неверна

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

$$df = n - 1$$

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 шо поделать” “ненавижу”. Проверить гипотезу о том, что категории независимы. Уровень значимости принять равным 0.01

Отношение/ Категория	Школьники	Студенты	Программист, < 5 лет	Программист, > 5 лет
Хороший язык	40	22	17	12
Ну а шо 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 отношения одинаково, P(хороший язык|категория) = P(хороший язык) * P(категория). И так далее.

Отношение /Категория	Школьники	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	40	22	17	12	91	0.35
Ну а шо 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	Школьники	Студенты	Программист, < 5 лет	Программист, > 5 лет	Сумма	Вероятность
Хороший язык	31,85	19,11	20,93	20,02	91	0,35
Ну а шо 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 H0 о независимости

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.

Решение

Гипотеза H0: Все языки получили равное число лайков, распределение лайков равномерное
Гипотеза H1: 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 H0 о том, что распределение лайков равномерное

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

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 (двусторонний тест)

Точный тест Фишера

Левый хвост, сложить вероятности всех таблиц здесь

Все хорошо

Правый хвост, сложить вероятности всех таблиц здесь

Таблица, перекошенная, как наша, но в другую сторону

Наша таблица

Таблицы с еще более перекошенной в другую сторону связью

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

	Исход есть	Исход а нет	Всего
Фактор есть	A	B	A+ B
Фактора 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 H0: Юноши и девушки сидят на диетах одинаково
Гипотеза H1: Девушки сидят на диетах чае

$$p(table) = ?$$

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

$$p(table_1) = ?$$

$$Pvalue = ?$$

Понятия, которые надо знать

- Генеральная совокупность, выборка, репрезентативность выборки
- H0 и H1 гипотезы
- Левосторонняя альтернатива и правосторонняя альтернатива
- P-value, критическое значение, уровень значимости
- Доверительный интервал
- Нормальное распределение, распределение Стьюдента, распределение Хи-квадрат
- Множественное тестирование, FWER, FDR
- В чем отличие alpha в FWER и FDR
- Почему надо применять на Бонферрони, а Холм-Бонферонни?
- В чем отличия критерия Хи-квадрат от Точного теста Фишера
- Какие проблемы сопряжены с Точным тестом Фишера?