Множественное тестирование и корреляции

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

Какое число генов мы обнаружить?

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

Какое число генов мы обнаружить?

Бог его знает.

Мы не знаем долю генов, для которых H0 верна

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

Болезнь – ампутированная 10 лет назад нога.

Какое число генов мы обнаружить?

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

Болезнь – ампутированная 10 лет назад нога.

Какое число генов мы обнаружить?

В этой ситуации мы подозреваем, что H0 верна для всех генов

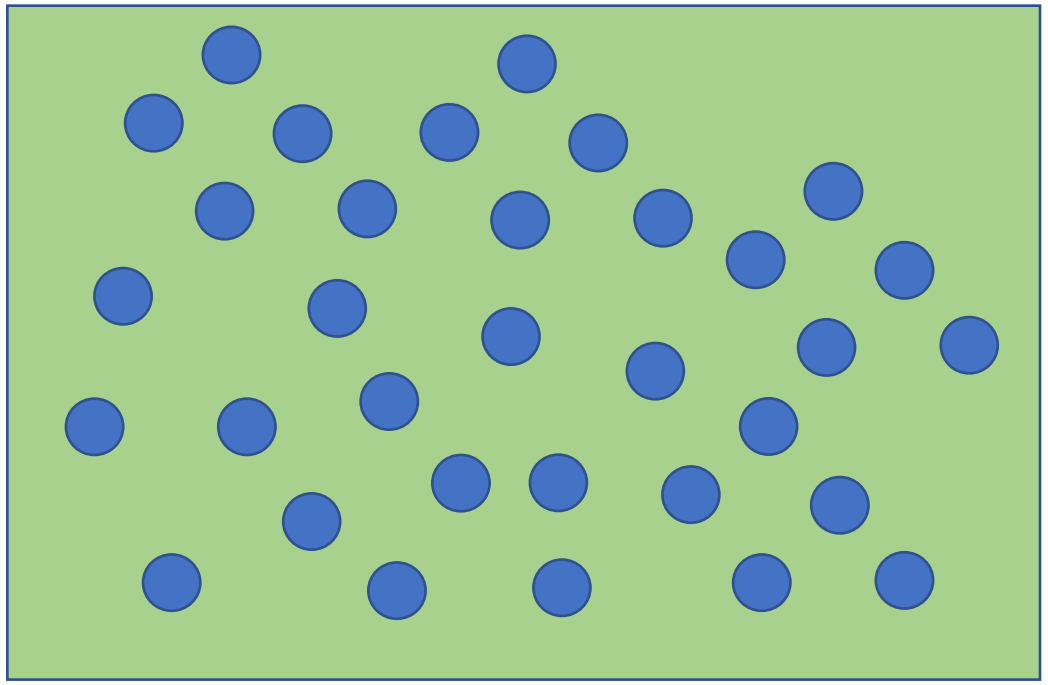
Вероятность ошибки первого рода

Пусть мы ввели некую процедуру T, которая на основе выборки отвергает или не отвергает H_0.

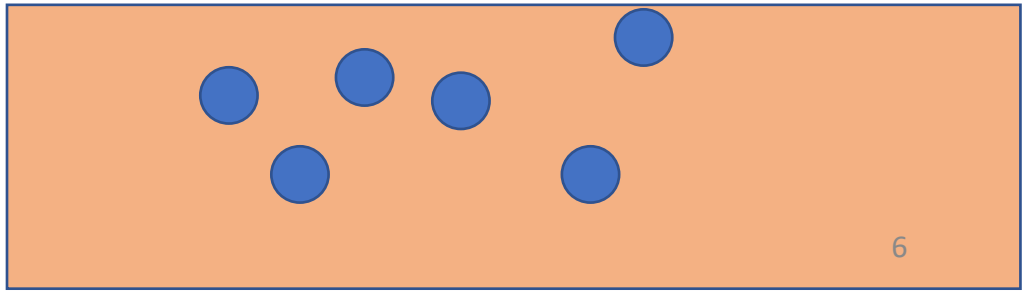
Допустим, мы провели бесчисленное множество экспериментов, для которых H_0 - верна, и оказалось, что наша процедура ошибочно отвергает H_0 в доле случаев α.

Тогда мы говорим, что вероятность ошибки первого рода равна α

$\boldsymbol{H_0}$ не отвергнута ошибочно



$\boldsymbol{H_0}$ ошибочно отвергнута



P-value и вероятность ошибки первого рода

Процедура T, которая состоит в том, что мы считаем $p - value$ нашего наблюдения и если

$$p - value \leq \alpha$$

отвергаем гипотезу H_0, будет поддерживать ошибку первого рода на уровне α (уровень значимости α)

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

Болезнь – ампутированная 10 лет назад нога.

Какое число генов мы обнаружить?

В этой ситуации мы подозреваем, что H0 верна для всех генов

Мы ожидаем обнаружить 30000 * 0.05 = 1500 генов

Проблема множественного тестирования

Представим себе ситуацию, что у нас есть датасет с 30000 генами, экспрессии которых померены у здоровых и больных людей.

Для каждого из этих генов проведем t-test и, если p-value меньше 0.05, будем считать ген влиящим на болезнь.

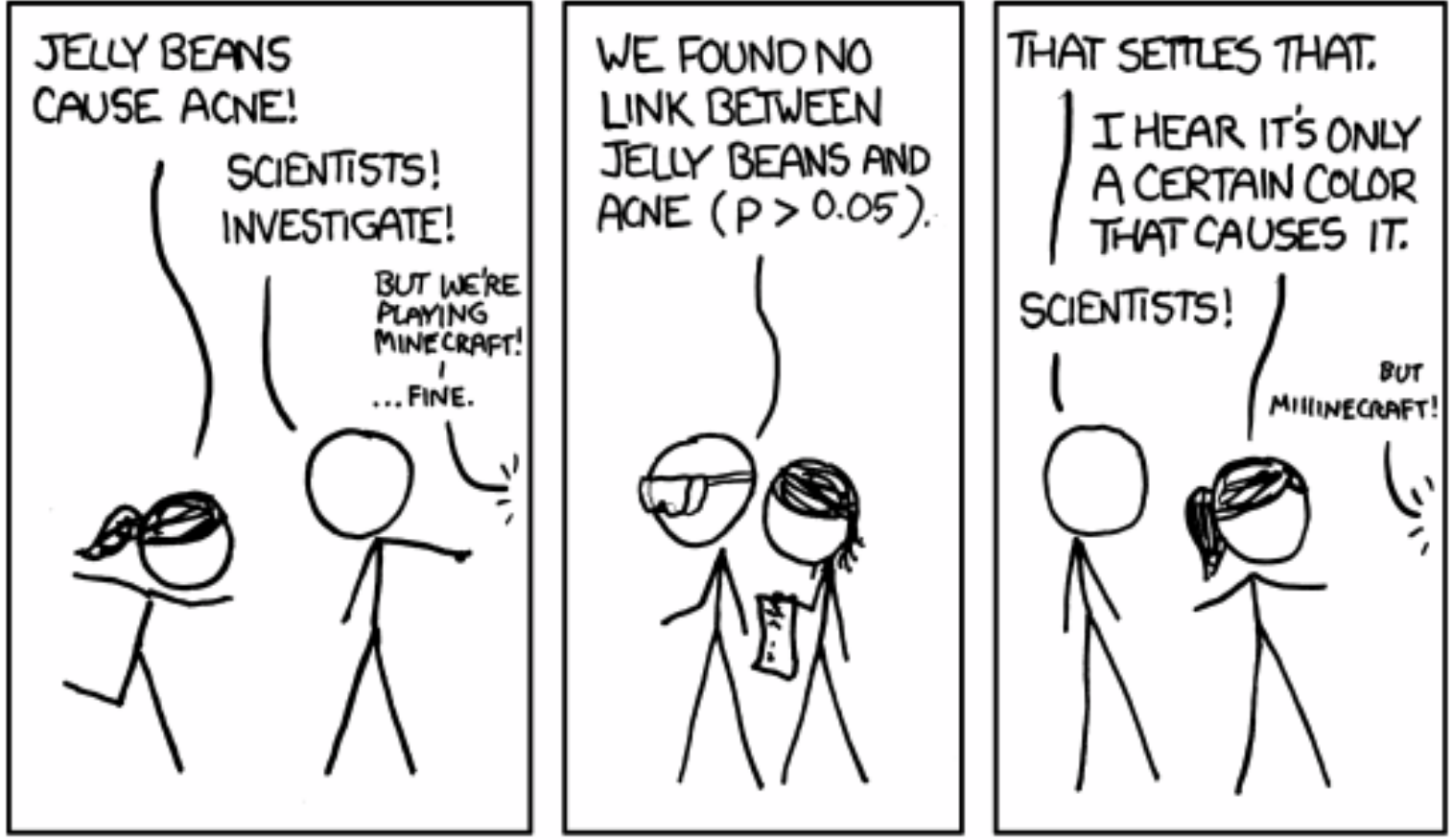
Болезнь – ампутированная 10 лет назад нога.

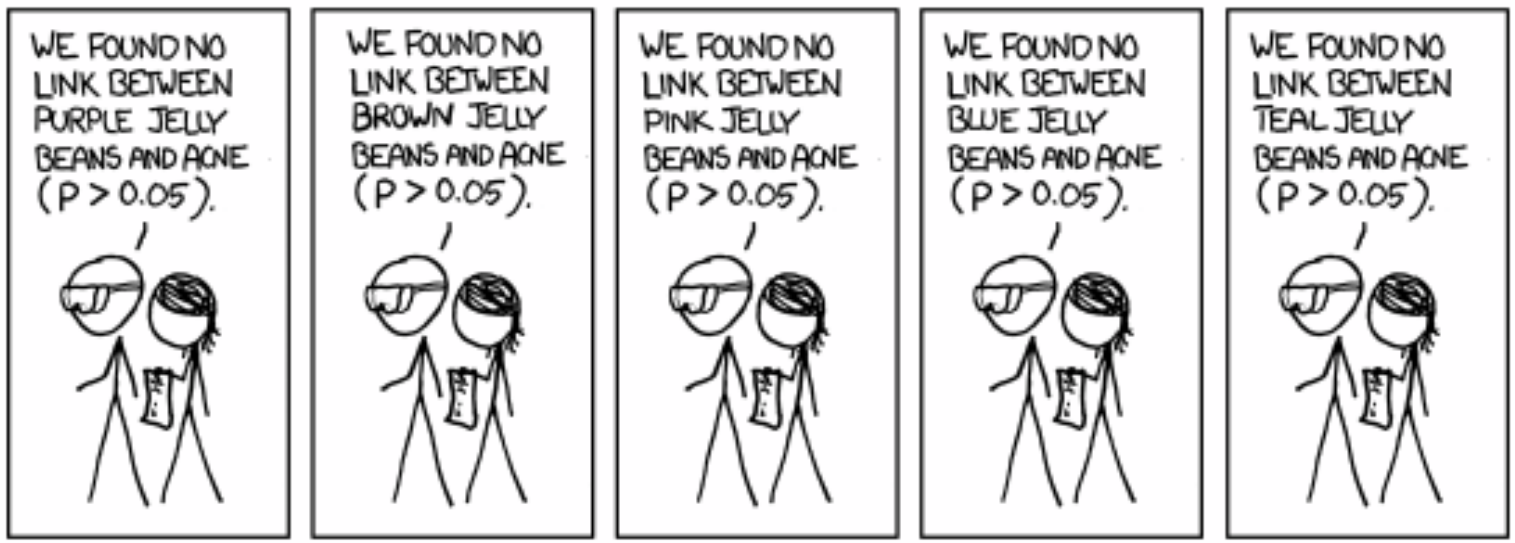
Какое число генов мы обнаружить?

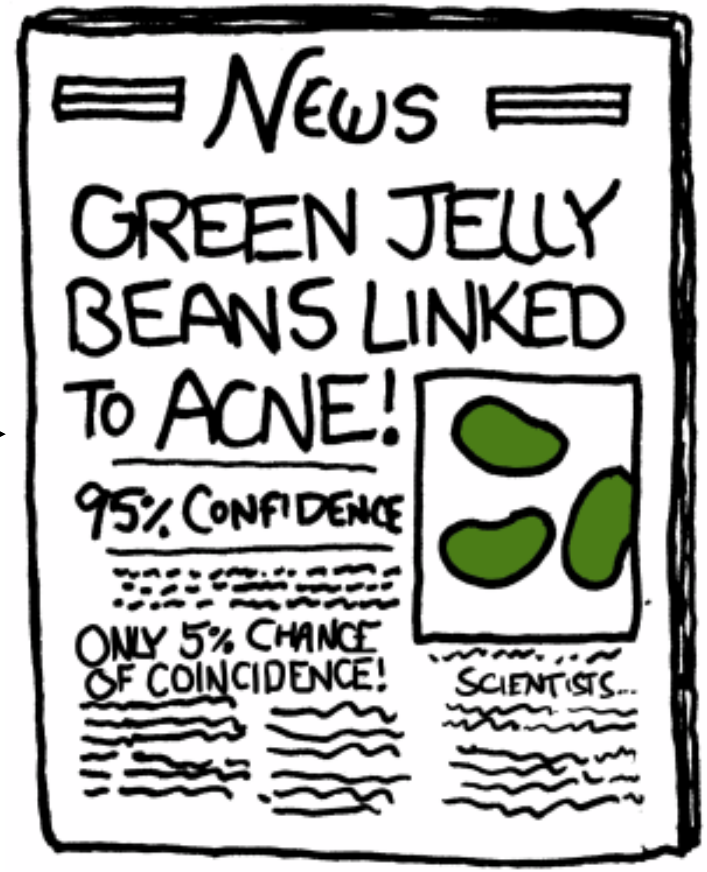
В этой ситуации мы подозреваем, что H0 верна для всех генов

Мы ожидаем обнаружить 30000 * 0.05 = 1500 генов

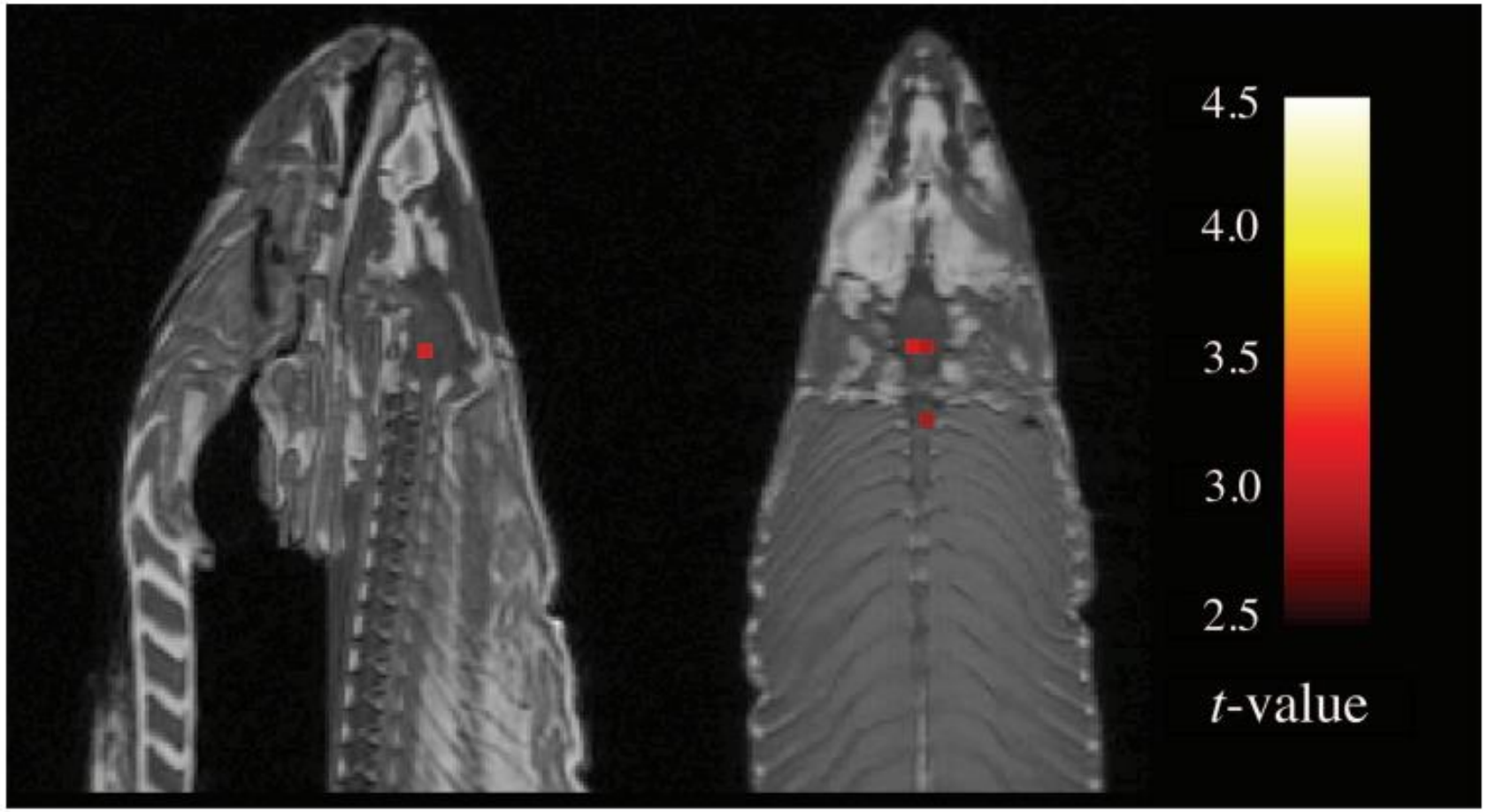
(в реальности нет ни одного такого гена)


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

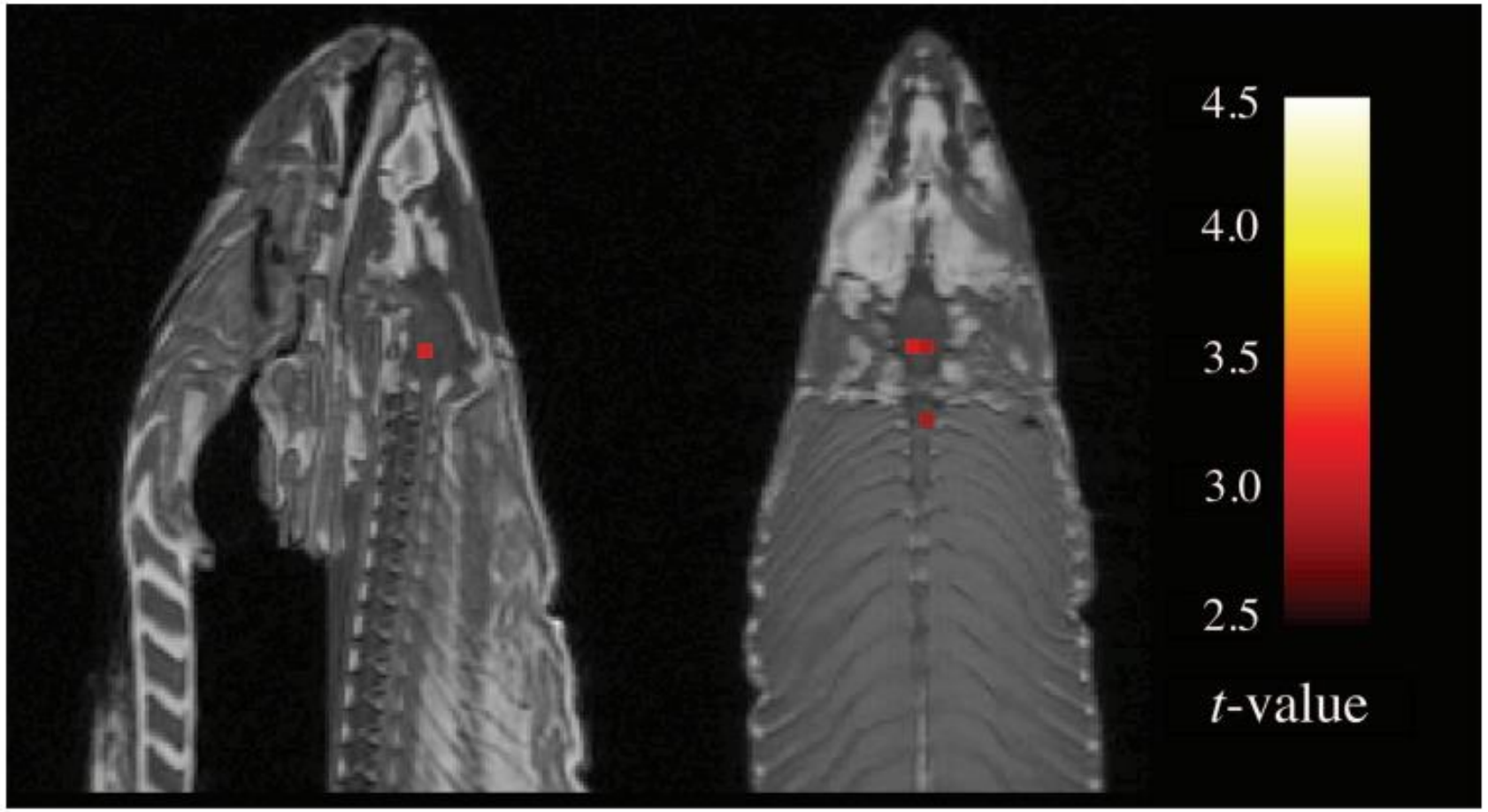

WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Лосось умеет определять эмоции людей



Мертвый лосось умеет определять эмоции людей



Авторы проверяли много-много регионов мозга, для каждого делали тест

https://www.wired.com/2009/09/fmrisalmon/

Проблема множественного тестирования

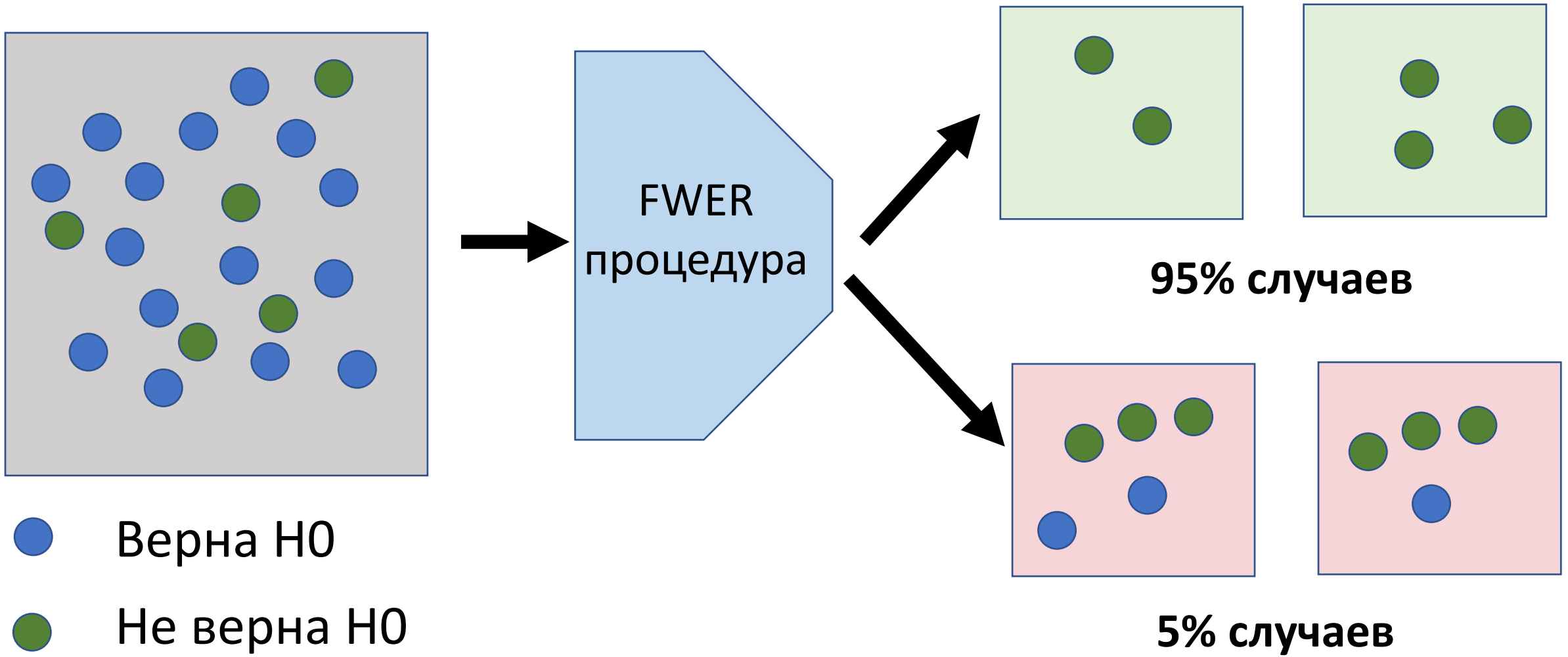
Если наше исследование состоит из большого количества тестов, то, контролируя уровень ошибки первого рода для каждого теста отдельно на уровне 0.05, мы все равно получаем большое количество **ложноположительных результатов**.

Проблема множественного тестирования

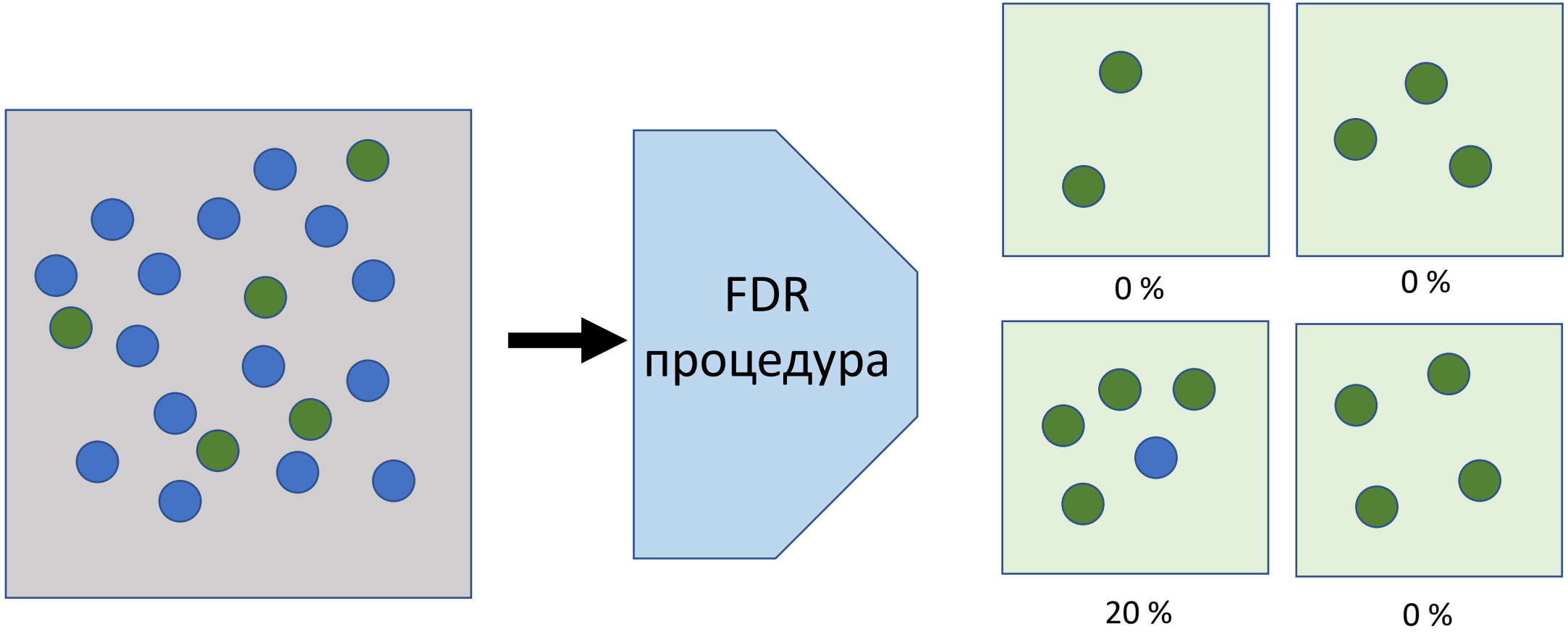
Если наше исследование состоит из большого количества тестов, то, контролируя уровень ошибки первого рода для каждого теста отдельно на уровне 0.05, мы все равно получаем большое количество **ложноположительных результатов**.

Хотим контролировать уровень ошибки не для каждого теста в отдельности, а для группы тестов

FWER
(Family-wise error rate)



FDR
(False discovery rate)



Верна H0

Не верна H0

В среднем доля неверно отобранных тестов – 5%

Контроль уровня ошибки для группы тестов (поправка на множественное тестирование)

Вероятность того, что мы хотя бы один ген обозначили значимым неверно – 0.05

FWER
(Family-wise error rate)

Среди тех генов, которые мы обозначили значимыми, в среднем не более 5% обозначенных неверно

FDR
(False discovery rate)

Известные FWER-поправки

1. Поправка Шидака - предполагает независимость тестов. Очень сильное требование
2. Поправка Бонферонни – независимость неважна
3. Поправка Холма-Бонферонни – независимость неважна, а еще при прочих равных найдем больше реально значимых генов (выше мощность)

Известные FWER-поправки

1. Поправка Шидака - предполагает независимость тестов. Очень сильное требование — Не ваш бро(?)
2. Поправка Бонферонни – независимость неважна — Не ваш бро!!!
3. Поправка Холма-Бонферонни – независимость неважна, а еще при прочих равных найдем больше реально значимых генов (выше мощность) — Ваш бро
4. Бывают поправки, учитывающие природу тестов, то, как именно они проводятся и т.д. Тоже дают бОльшую мощность — Ваш бро

Известные FDR-поправки

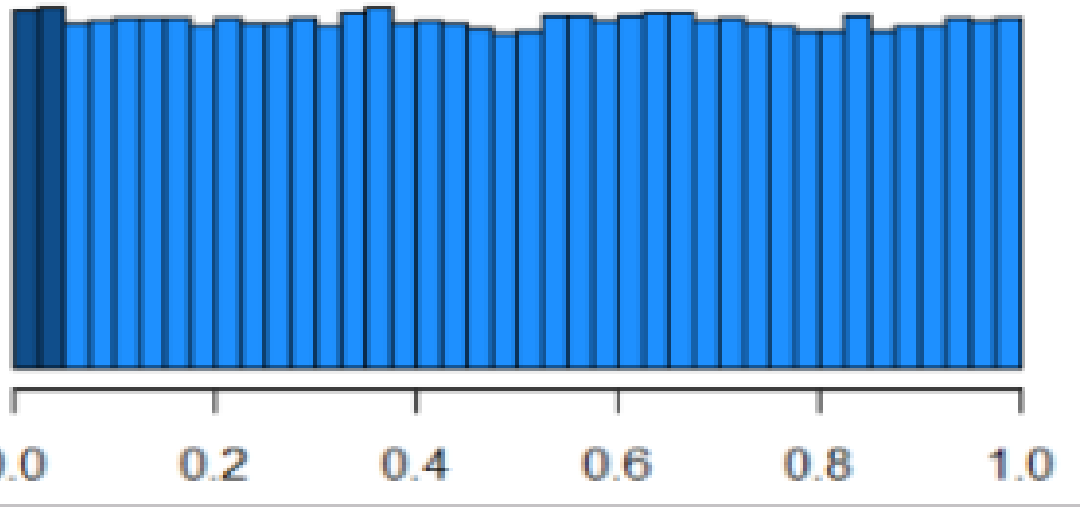
1. Поправка Benjamini-Hochberg - предполагает то, что тесты положительно зависимы. Часто выполняемое требование;
2. Benjamini-Hochberg-Yekutieli – независимость неважна. Менее мощный

Ваш бро

Ваш бро(?)

Положительная зависимость

В ситуации с jelly beans из изначального примера, цвета бобов независимы. Как и их способность вызывать акне (они его не вызывают). В этом случае если тестировать бесконечно много цветов и смотреть распределение полученных p-value – оно будет равномерным



Положительная зависимость

Теперь представим себе, что мы делаем попарные сравнения между группами.

Если среднее группа A значимо отличается от среднего группы B, то у нас есть основания БОЛЬШЕ надеяться на то, что оно и значимо отличается от среднего групп C, и D, и т.д.

Таким образом, малое p-value одной гипотезы говорит о более вероятном малом p-value – это **положительная зависимость**

Группа	Среднее
A	17.2
B	19.5
C	22.0
D	23.0
E	22.0

Негативная зависимость

Реальный пример придумать сложно – она и реально встречается достаточно редко.

Пусть мы проверяем следующие гипотезы:

A: в государстве больше мужчин, чем женщин

B: в государстве больше наборов XY, чем XX

C: в государстве меньше наборов XY, чем XX

Представим себе, что мы знаем, что в государстве мужчин больше, чем женщин

Тогда p-value для A будет маленьким, для B – маленьким, а для C – большим. Чем p-value A меньше – чем больше p-value C

A и C – **негативно зависимы**

Какую зависимость логичнее предположить для генов?

Положительную – если меняется значимо экспрессия одного гена, то гены, с ним связанные – будут менять экспрессию значимо тоже.
Несвязанные же с ним будут менять экспрессию независимо – нам тоже подходит

Какую зависимость логичнее предположить для генов?

Положительную – если меняется значимо экспрессия одного гена, то гены, с ним связанные – будут менять экспрессию значимо тоже.

Поэтому в случае FDR чаще всего мы слышим именно про поправку Benjamini-Hochberg

Когда использовать FWER, а когда FDR?

1. Если последний этап обработки данных и нам ВАЖНО не получать ни одного false-positive – то использовать FWER
2. Иначе – использовать FDR :

   1) мы готовы к большому числу false-positive

   2) есть последующие стадии, которые уберут false-positive, при этом им полезнее иметь много шумной информации, чем мало нешумной

Пример: поиск ингибитора

Мы готовы экспериментально протестировать тысячи веществ, чтобы найти ОДНО с нужным эффектом.

Потому нет никакой необходимости в использовании FWER, используем FDR.

Пример: поиск причин болезни

Мы можем набрать в начале много генов, среди которых 10% или 20% шума – на следующих этапах мы отсеем их за счет того, что будем смотреть не на гены, а на семейства.

Случайные гены дадут случайные семейства – их отсеем.

Порог для поправок

1. Для FWER принят 0.05 – хотим **вероятность хотя бы одного ложноположительного гена не более 0.05**
2. Для FDR нет общепринятого порога. Все зависит от того, на какую **долю неправильно определенных генов** вы согласны. Хоть 90%.

Ковариация

Пусть у нас есть выборка X, посчитаем ее среднее

$$\bar{X} = \frac{1}{n}\sum X_i$$

Так же мы можем посчитать для каждого наблюдения его отклонение от среднего выборочного

$$\Delta X_i = X_i - \bar{X}$$

Ковариация

Пусть у нас есть выборка X, посчитаем ее среднее

$$\bar{X} = \frac{1}{n}\sum X_i$$

Так же мы можем посчитать для каждого наблюдения его отклонение от среднего выборочного

$$\Delta X_i = X_i - \bar{X}$$

Ковариация

Пусть у нас есть выборка Y, посчитаем ее среднее

$$\bar{Y} = \frac{1}{n}\sum Y_i$$

Так же мы можем посчитать для каждого наблюдения его отклонение от среднего выборочного

$$\Delta Y_i = Y_i - \bar{Y}$$

Ковариация

Давайте теперь посмотрим, на произведение соответствующих отклонений

$$\Delta X_{\mathrm{i}} \Delta Y_i = (Y_i - \bar{Y})(X_i - \bar{X})$$

Очевидно, что это произведение больше 0, когда оба отклонения имеют одинаковый знак и меньше 0, когда разный

Ковариация

Введем ковариацию - выборочное среднее произведений соответствующих отклонений. Делим на n-1 по причине схожей с той, по которой выборочная дисперсия считается с коэффициентом 1/(n-1), а не 1/n

$$\operatorname{cov}(X, Y) = \frac{1}{n-1} \sum_i (Y_i - \bar{Y})(X_i - \bar{X})$$

Ковариация

Эта величина больше 0, если положительным отклонениям от среднего в X соответствуют положительные отклонения от среднего в Y и наоборот

Эта величина меньше 0, если положительным отклонениям от среднего в X соответствуют отрицательные отклонения от среднего в Y и наоборот

$$\mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i}(Y_i - \bar{Y})(X_i - \bar{X})$$

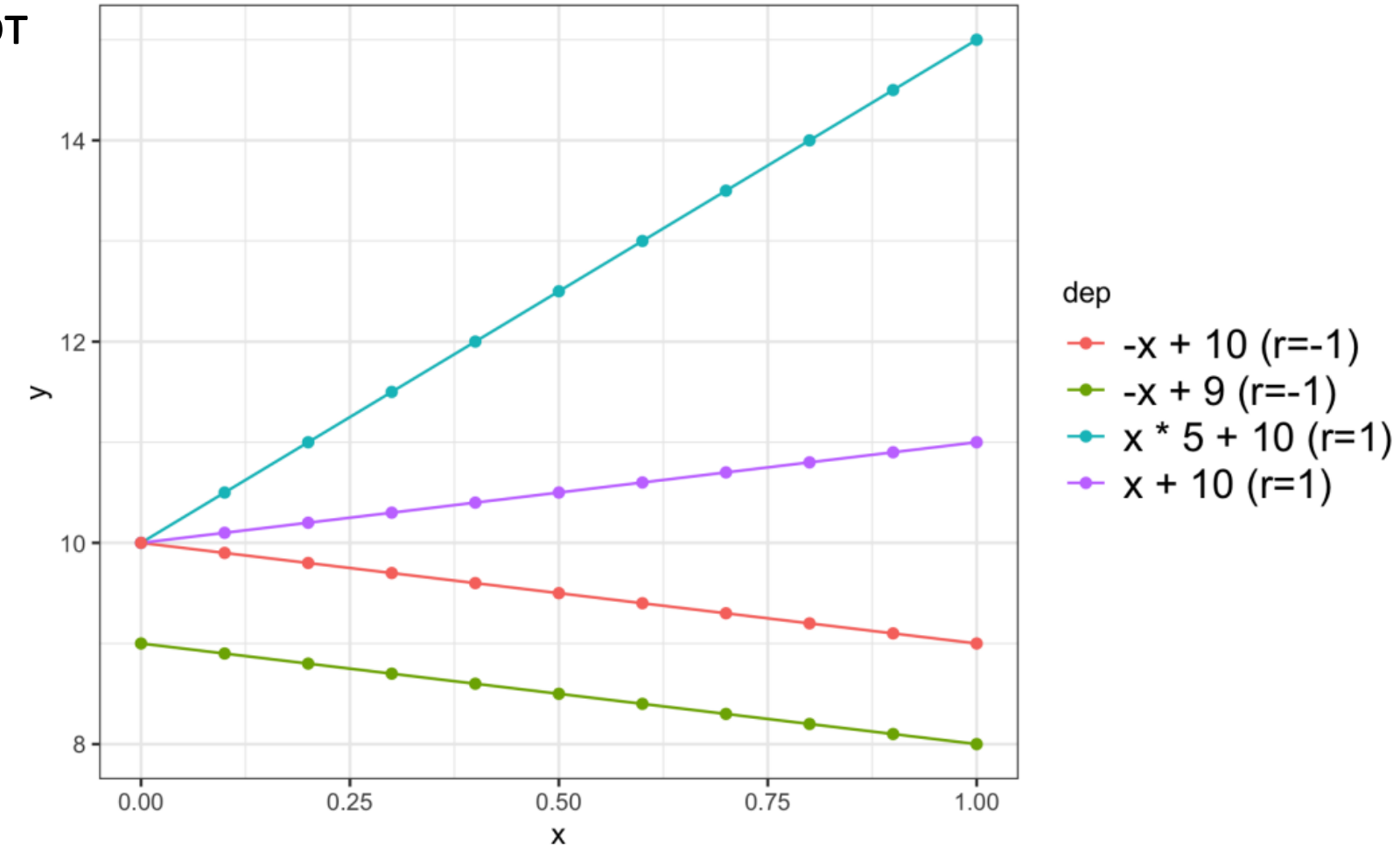
Корреляция

Проблема ковариации в том, что она зависит от масштаба измерямых величин. Потому разделим ее на стандартные отклонения этих величин. Получим корреляцию

$$\mathrm{cor}(X,Y) = \frac{\frac{1}{n-1}\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sigma_x \sigma_y}$$

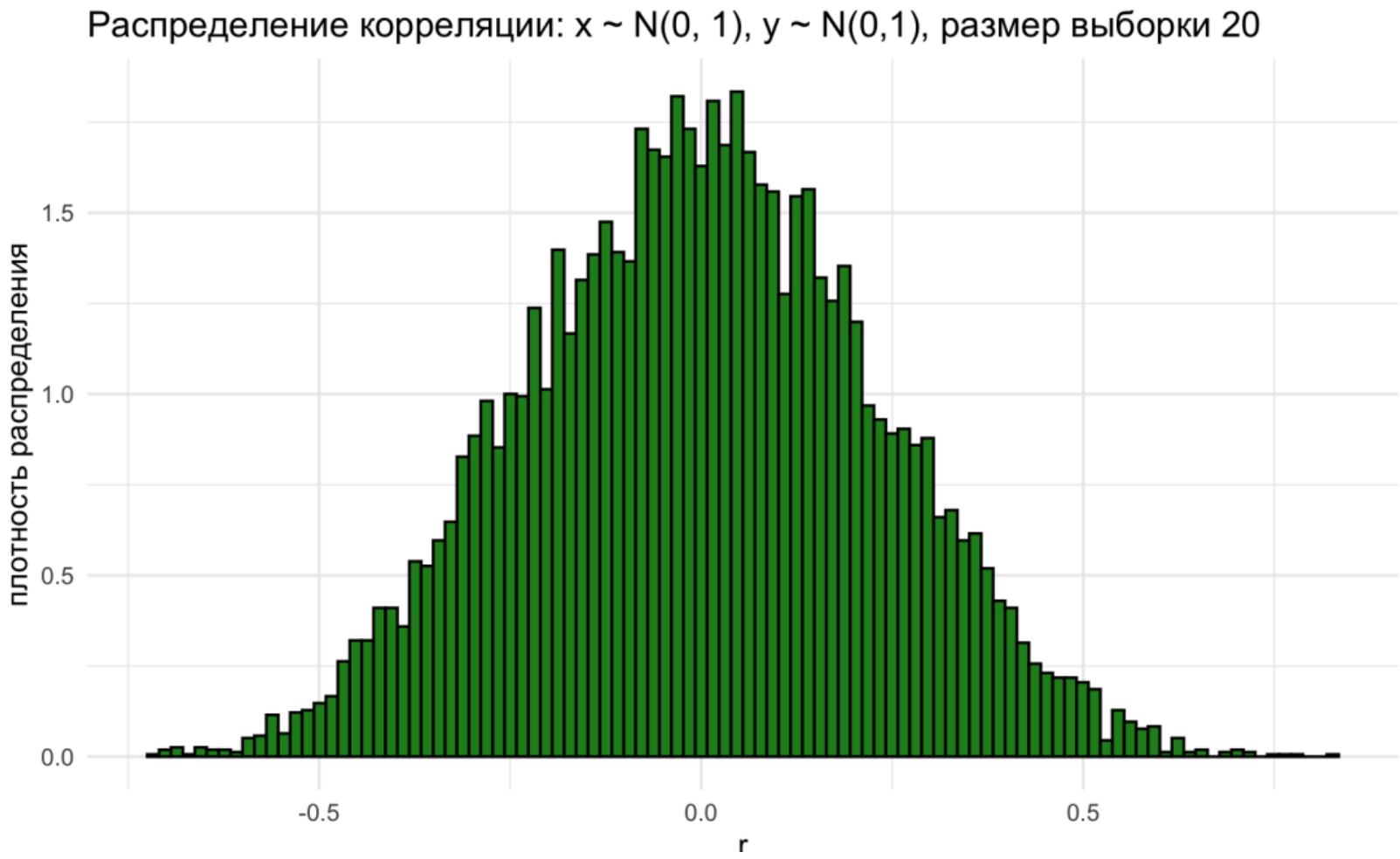
Корреляция

1. Всегда лежит в пределах от -1 до 1
2. Если к x или y прибавить одно и то же число, то значение корреляции не изменится
3. Если x или y умножить на одно и то же положительное число, то значение корреляции не изменится, если на отрицательное - поменяется на противоположное

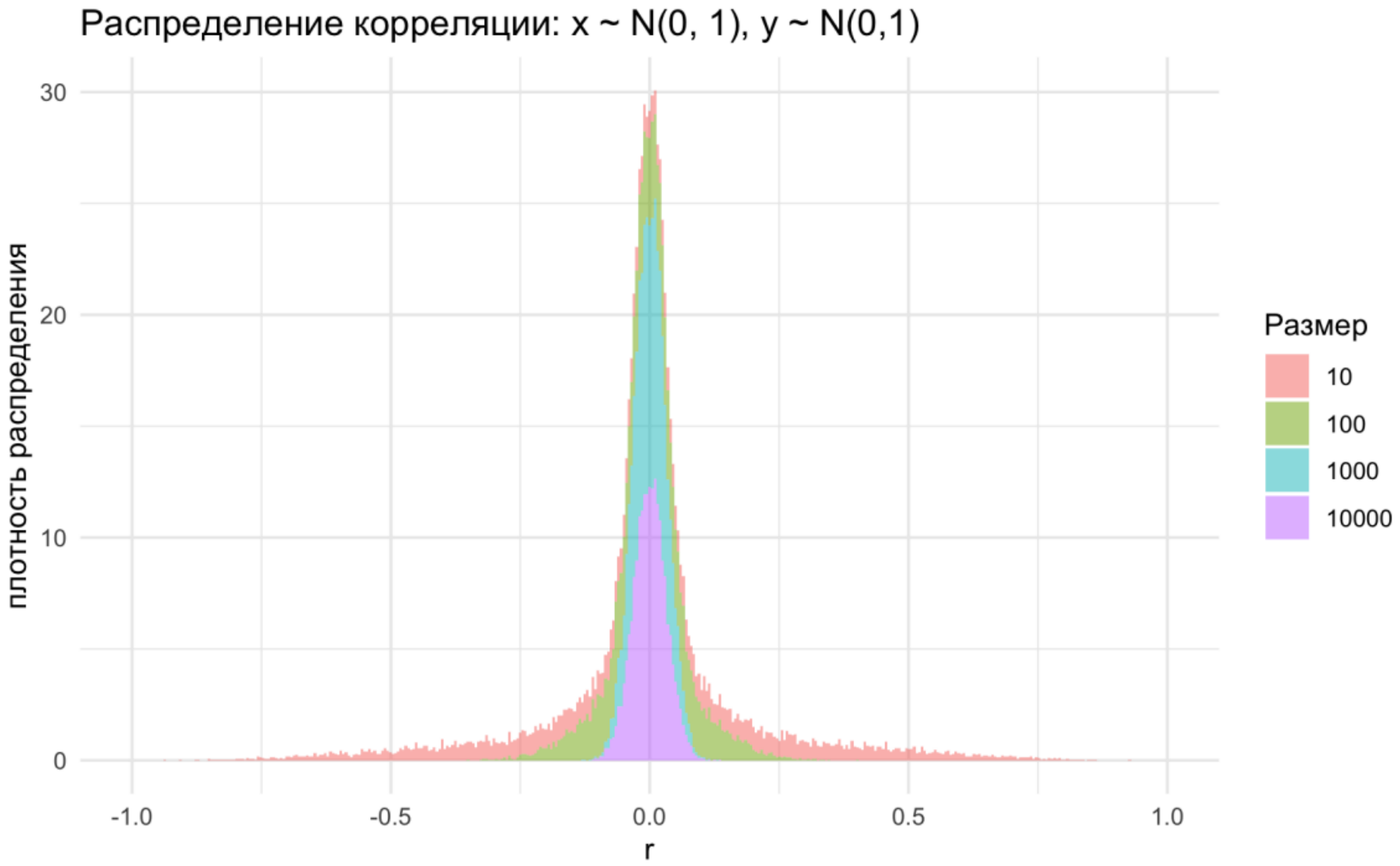


Корреляция

4. Корреляция выборок - случайная величина, оценка корреляции генеральных совокупностей



Корреляция


Распределение корреляции: x ~ N(0, 1), y ~ N(0,1)
плотность распределения
30
20
10
0
-1.0
-0.5
0.0
0.5
1.0
r
Размер
10
100
1000
10000

Корреляция

Если мы предполагаем, что настоящий коэффициент корреляции (посчитанный для генеральных совокупностей) равен 0 (наше H0), то вот такая величина распределена по Стьюденту

$$t = \frac{r}{1 - r^2}\sqrt{n - 2} \sim t_{n-2}$$

Тест на значимость корреляции в R - cor.test

Корреляция

5. Если x и y нормально распределены, то из равенства корреляции 0 следует независимость переменных

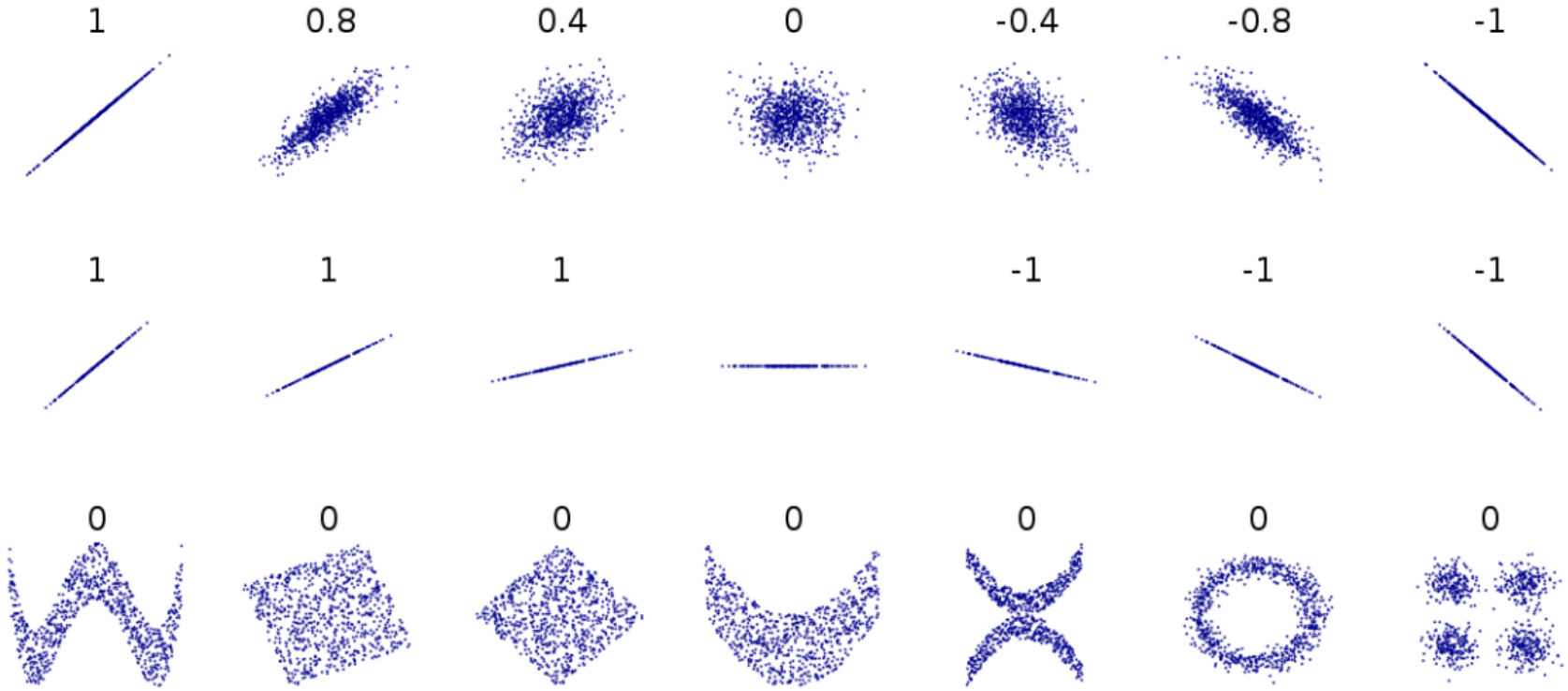
Но применять корреляцию можно не только к нормально распределенным данным

```
x <- runif(1000, 0, 1)
y <- 5 * x + rnorm(1000, sd=0.01)
cor(x, y)
```

```
## [1] 0.9999773
```

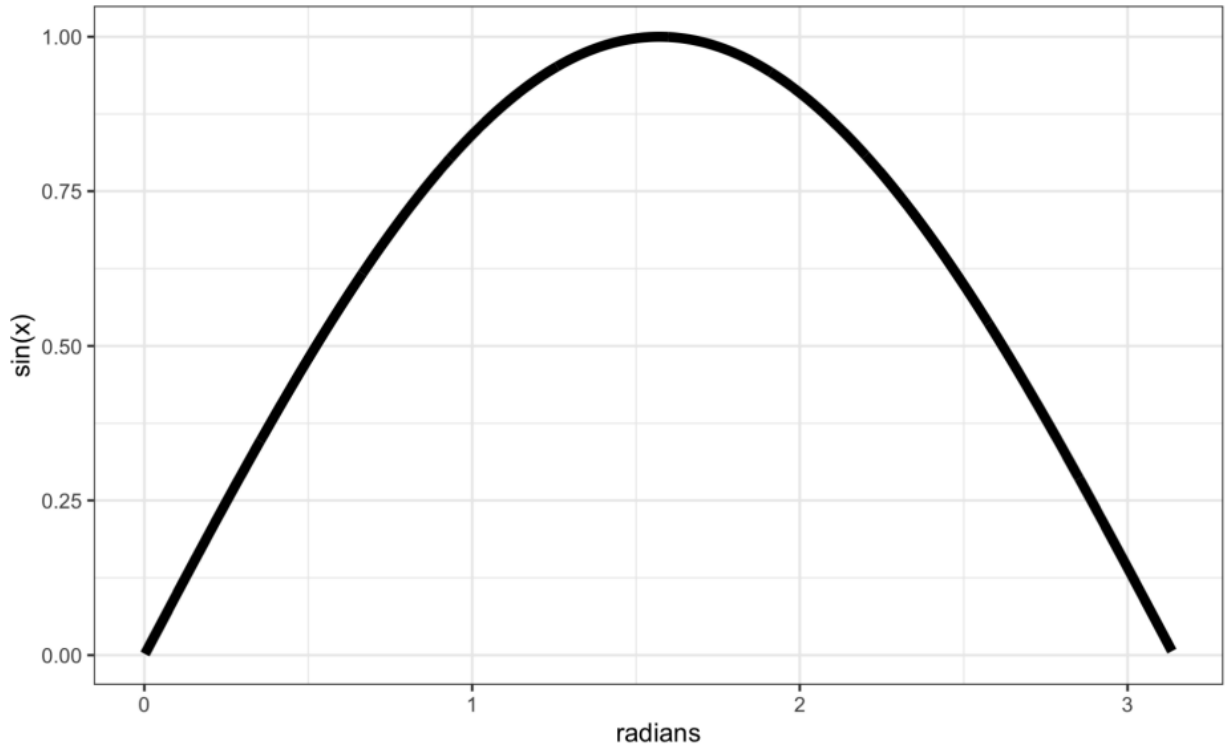
Корреляция

6. В общем случае из равенства корреляции 0 не следует независимость переменных



Корреляция

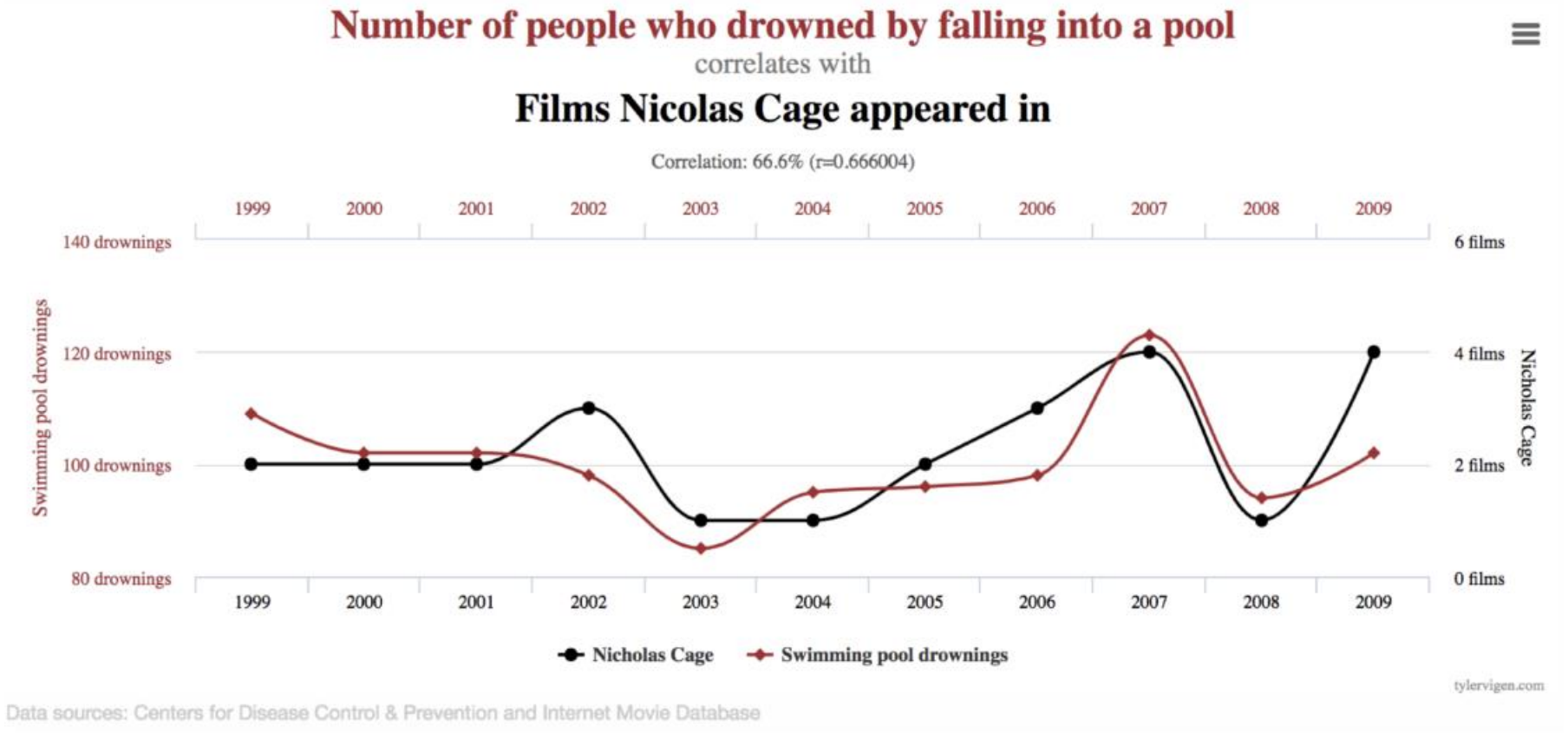
7. В общем случае из равенства корреляции 0 не следует независимость переменных



Берем x равномерно на отрезке [0; pi]. Корреляция в среднем - 0

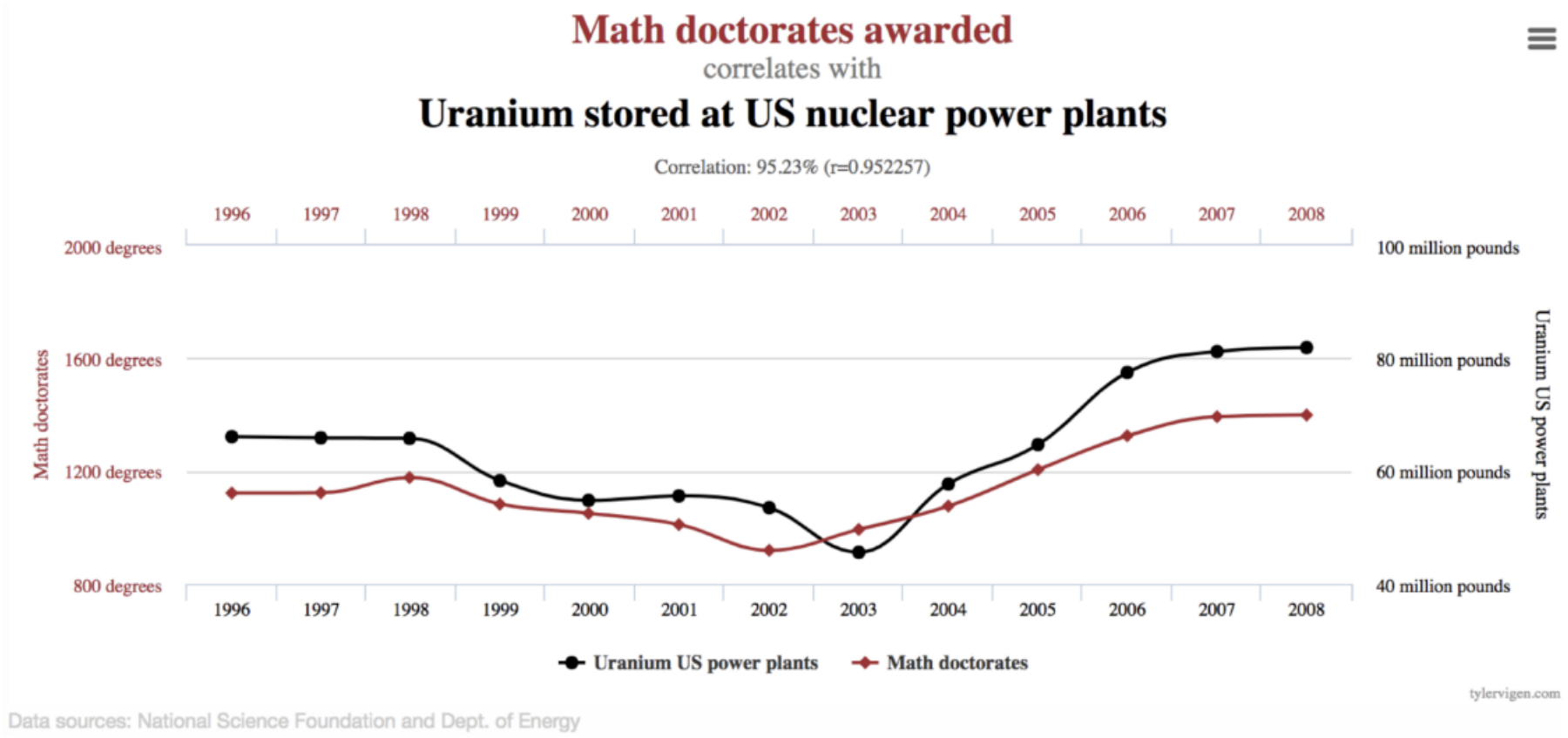
Корреляция

8. Из значимого значения коэффициента корреляции не следует причинно-следственной связи



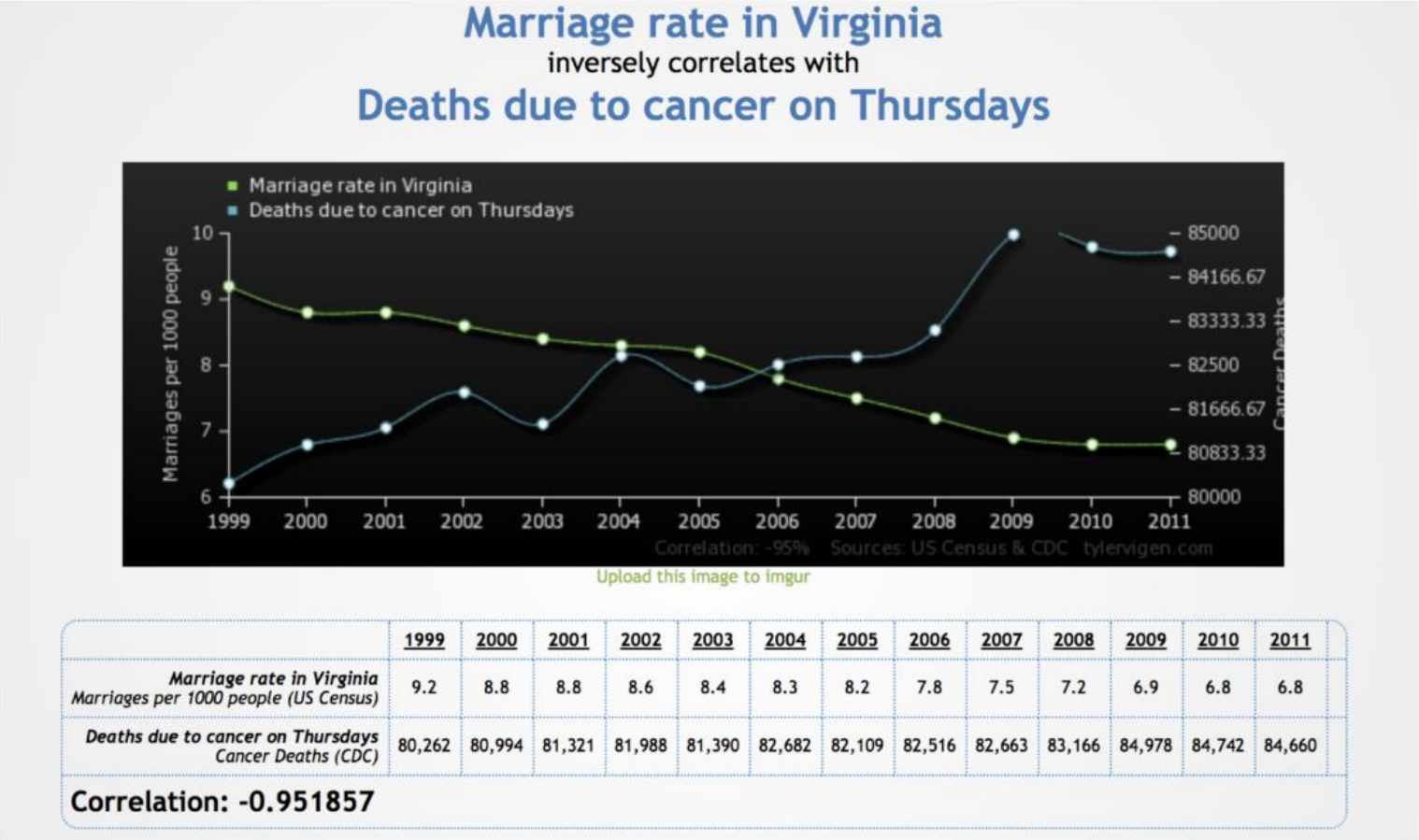
Корреляция

8. Из значимого значения коэффициента корреляции не следует причинно-следственной связи



Корреляция

8. Из значимого значения коэффициента корреляции не следует причинно-следственной связи

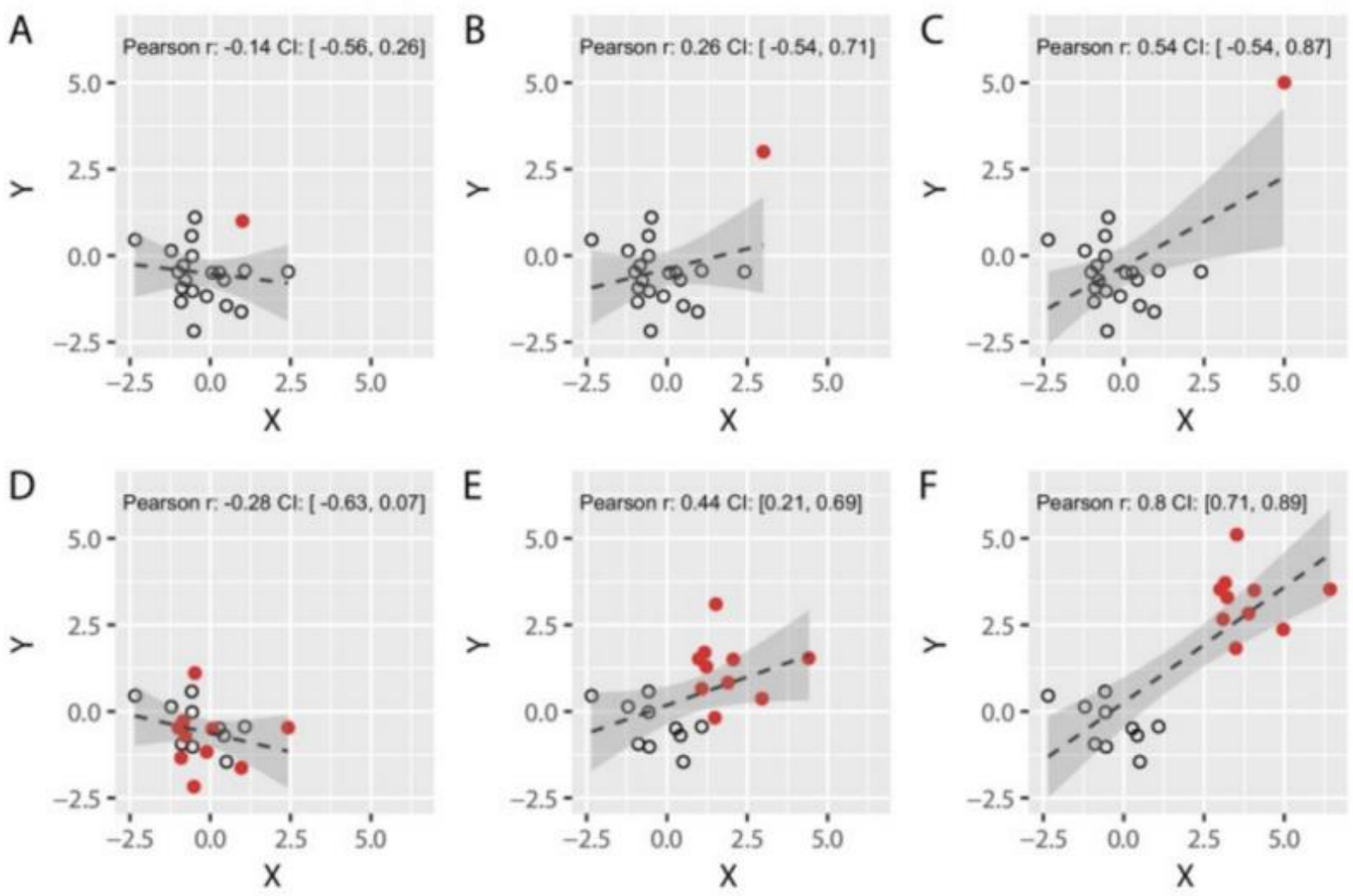


	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Marriage rate in Virginia *Marriages per 1000 people (US Census)*	9.2	8.8	8.8	8.6	8.4	8.3	8.2	7.8	7.5	7.2	6.9	6.8	6.8
Deaths due to cancer on Thursdays *Cancer Deaths (CDC)*	80,262	80,994	81,321	81,988	81,390	82,682	82,109	82,516	82,663	83,166	84,978	84,742	84,660

Correlation: -0.951857

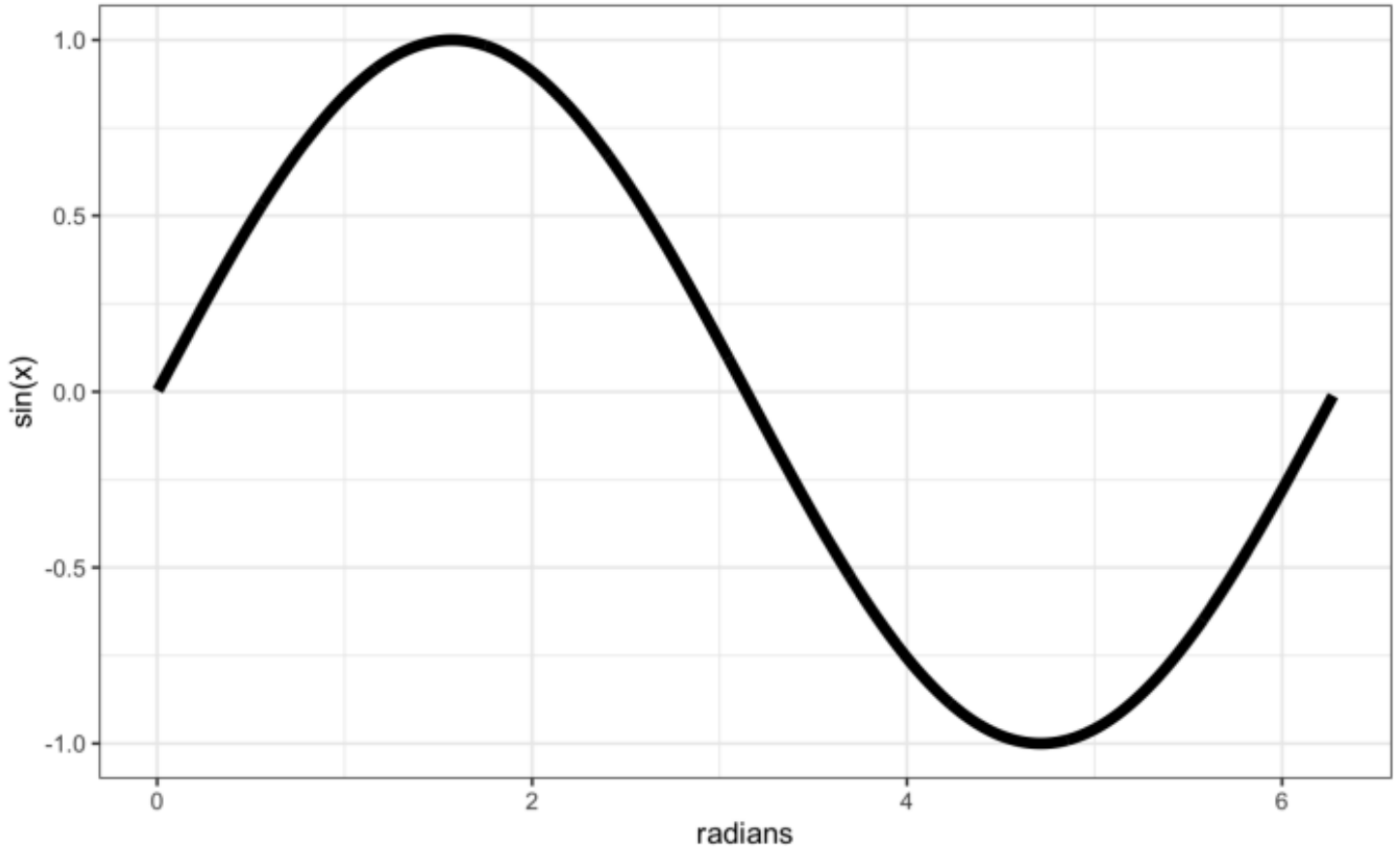
Корреляция

9. Всегда нужно строить график, для которого считаете корреляцию



Корреляция

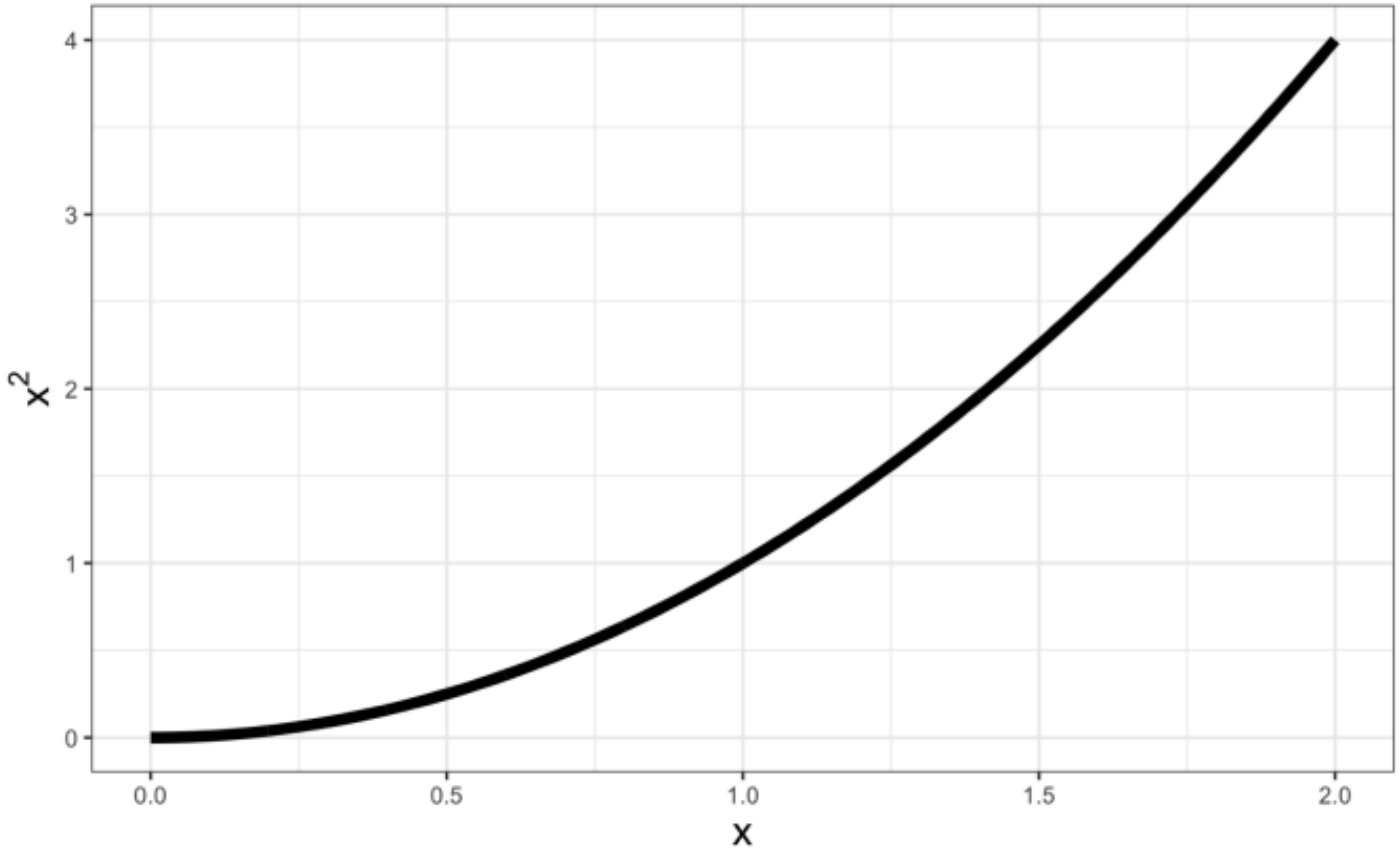
9. Всегда нужно строить график, для которого считаете корреляцию



Берем x равномерно на отрезке [0; 2pi]. Корреляция в среднем - -0.74

Корреляция (Пирсона)

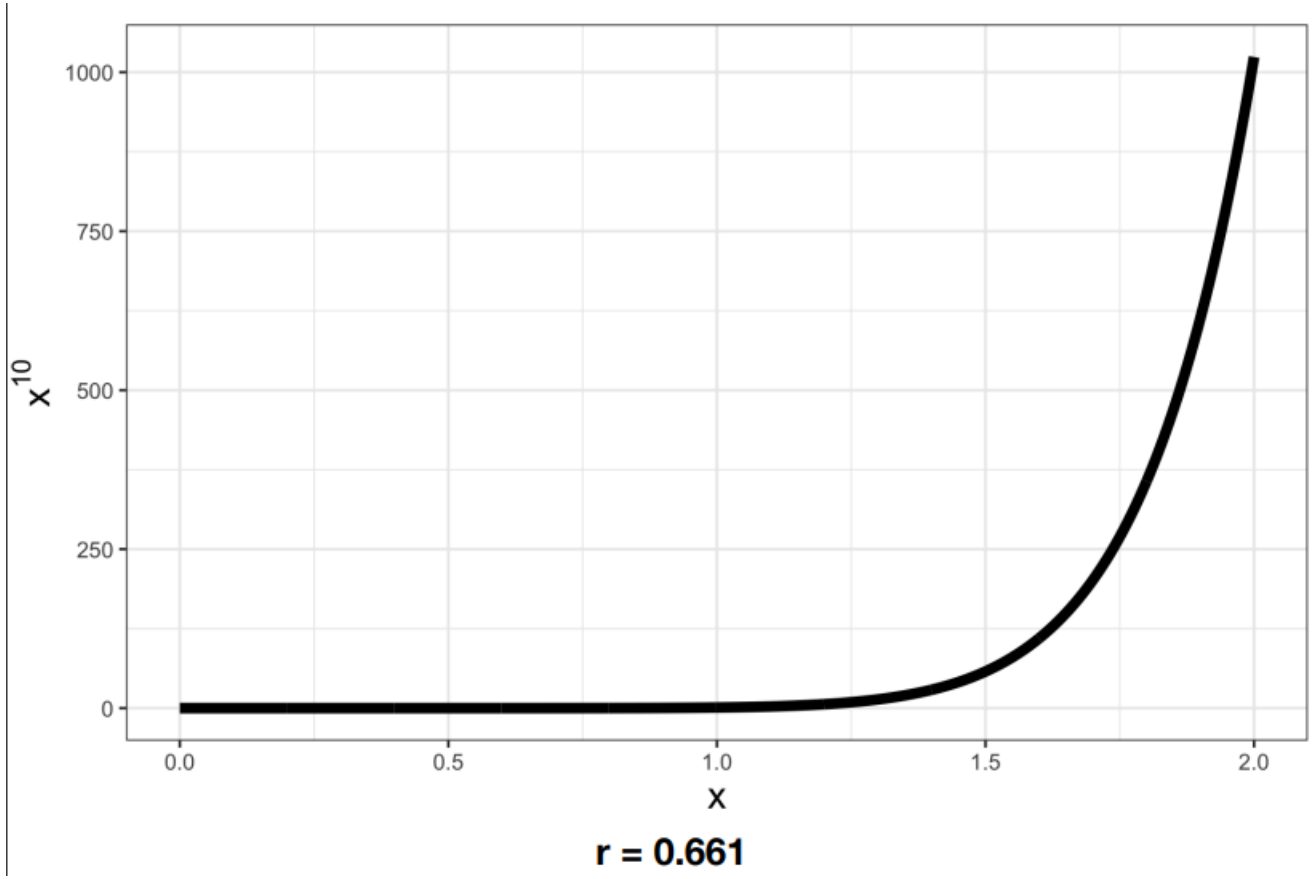
10. Плохо учитывает не линейные зависимости



r = 0.916

Корреляция (Пирсона)

10. Плохо учитывает не линейные зависимости

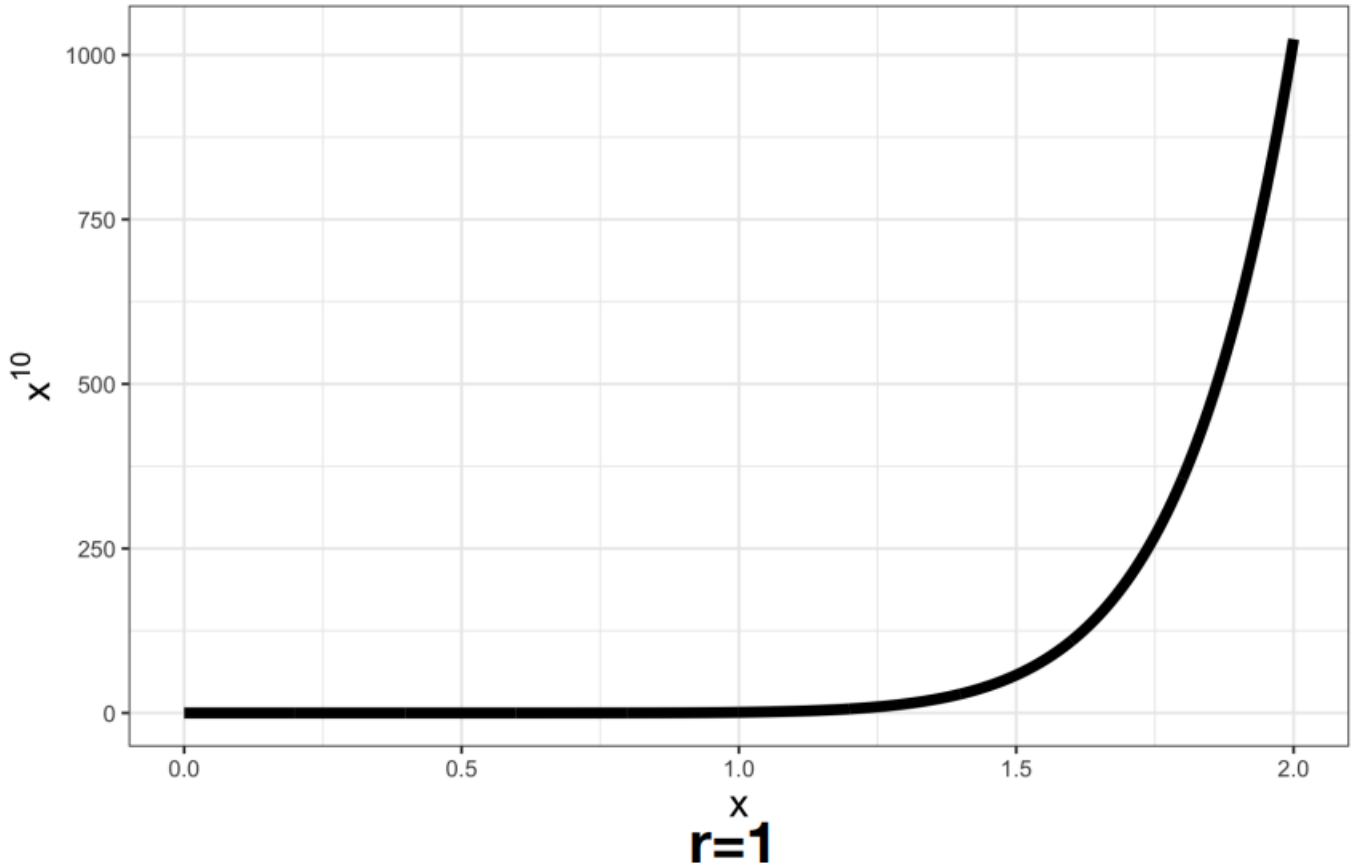


Корреляция Спирмена

Считаем корреляции между рангами наблюдений – учитывает любую монотонность

Корреляция Спирмена

Считаем корреляции между рангами наблюдений – учитывает любую монотонность



Корреляция Тау-Кенделла

То же самое, но более устойчива к выбросам

У меня есть набор данных, 1000 измеренных величин и y, с которой я хочу найти связь.
Я считаю корреляцию между ней и каждым признаком. Нужна ли поправка на множественное тестирование?

У меня есть набор данных, 1000 измеренных величин и y, с которой я хочу найти связь.
Я считаю корреляцию между ней и каждым признаком. Нужна ли поправка на множественное тестирование?

Да

```
m <- matrix(rnorm(1001 * 100, mean = 0, sd=1), ncol=1001)
pvals <- sapply(1:1000, function(x){cor.test(m[, x], m[, 1001])$p.value })
sum(pvals < 0.05)
```

```
## [1] 47
```

Множественное тестирование=multiplicity problem

Проблема множественного тестирования касается любой ситуации, когда вы работаете со некой случайной величиной много раз.

Можно получить любую корреляцию, любое качество и т.д, если просто попробовать достаточно большое число раз.

Попробовали много подходов – множественное тестирование.

Попробовали много датасетов – множественное тестирование

И т.д.