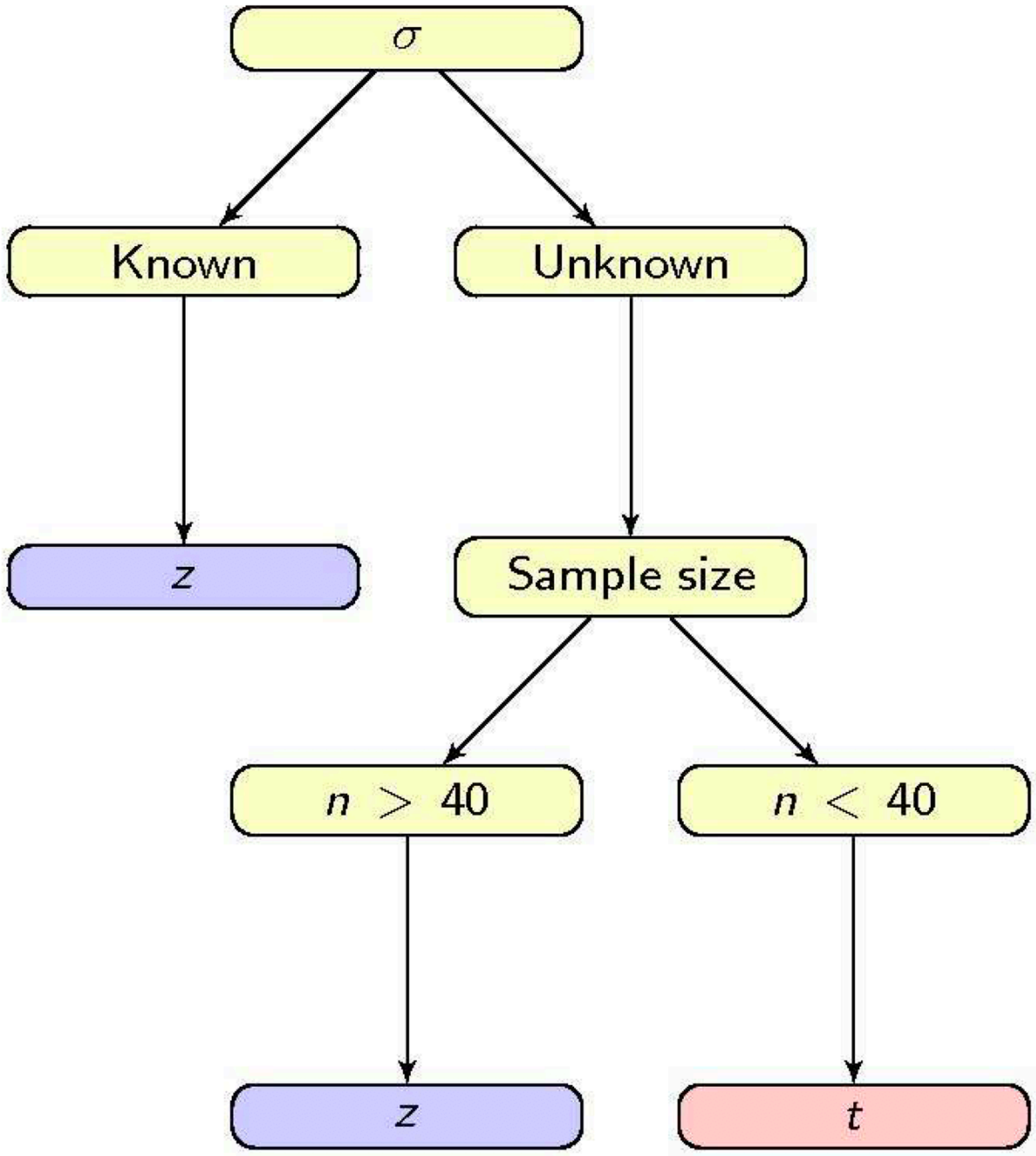
Статистика 2

Выбор распределения

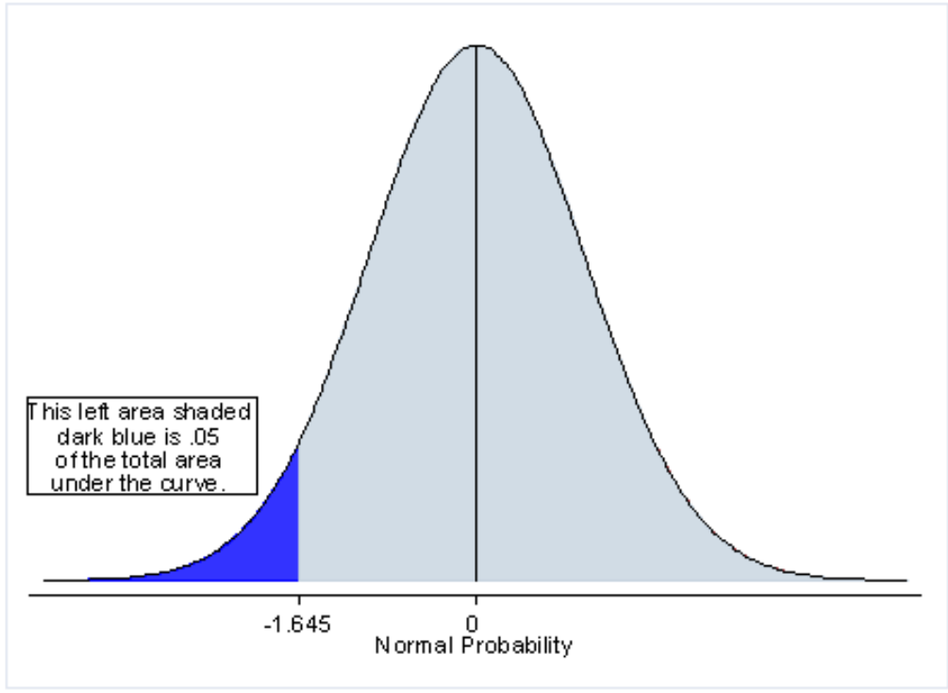


Вычисление p-value

А) Для односторонних тестов

1) Левосторонний тест

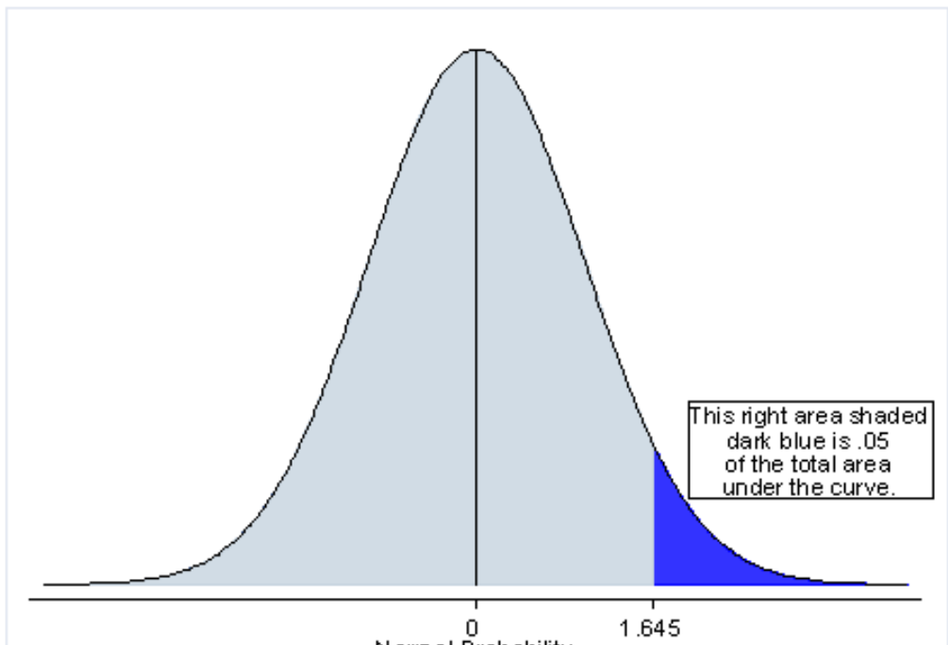
$$P(x < test\_value) = p\_value$$



2) Правосторонний тест

$$P(x > test\_value) = p\_value$$

$$1 - P(x < test\_value) = p\_value$$

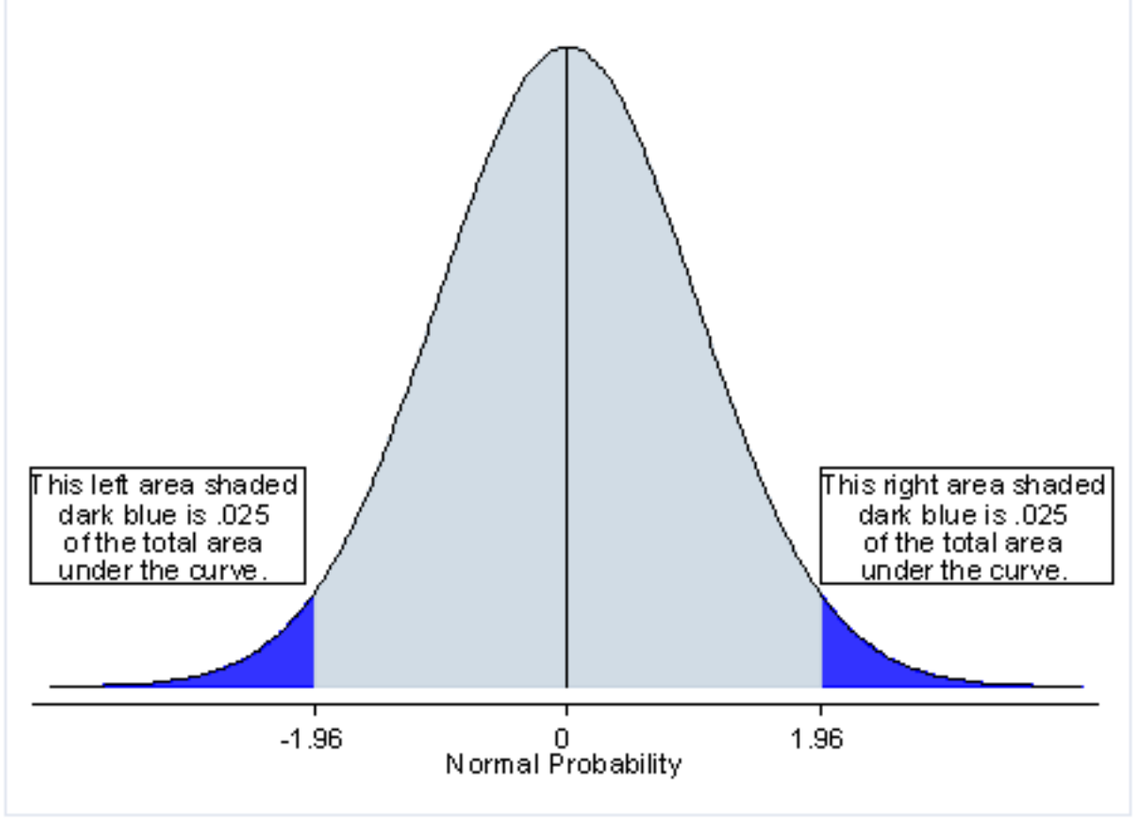


Вычисление p-value

Б) Для двусторонних тестов

$$P(x < -abs(test\_value)) + P(x > abs(test\_value)) = p\_value$$

$$P(x < -abs(test\_value)) + 1 - P(x < abs(test\_value)) = p\_value$$



Вычисление критического значения

Для нормального распределения

$$z\_value = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Для распределения Стьюдента

$$t\_value = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Вычисление стандартного отклонения в случае двух выборок

В общем случае

$$SE(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Если дисперсии равны

$$SE(\overline{X} - \overline{Y}) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Если дисперсии равны, но неизвестны

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

$$SE(\overline{X} - \overline{Y}) = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Задача

Четвероклассник Федя является сторонником теории, что из двух девочек он нравится больше той, которая дольше думает, прежде чем ударить его после того, как он дернул ее за косичку. К несчастью для двух его одноклассниц, Ани и Светы, Федя, кроме того, прочитал книгу “Статистика для самых маленьких”, поэтому он решил получить статистически достоверную разницу. Он планировал дернуть каждую девочку по 12 раз и засекать время отклика. К сожалению, на момент, когда было собрано 11 значений для Светы и 10 значений для Ани, Аня ударила его портфелем по голове, в связи с чем эксперимент было решено досрочно завершить. Для Светы были получены следующие значения (сек)

[5, 10, 7, 3, 6, 10, 5, 10, 8, 9, 10]

Для Ани:

[11, 5, 4, 7, 3, 5, 2, 6, 5, 4]

1. Есть ли статистическая разница между временами отклика двух девочек, если считать, что стандартное отклонение времени ответа на подобные раздражители у детей известна и равна 2.5 и распределения нормальные?
2. Если гипотеза верна, то какой вывод может сделать Федя кроме того, что статистика может быть опасна для жизни?

Решение

1) Сформулируем гипотезы H0 и H1.
H0 : девочки реагируют одинаково
H1: Есть значимая разница с уровнем значимости alpha = 0.05

Так как распределения по условию нормальные и нам известно стандартное отклонение, то будем использовать нормальный тест.
Посчитаем z-value:

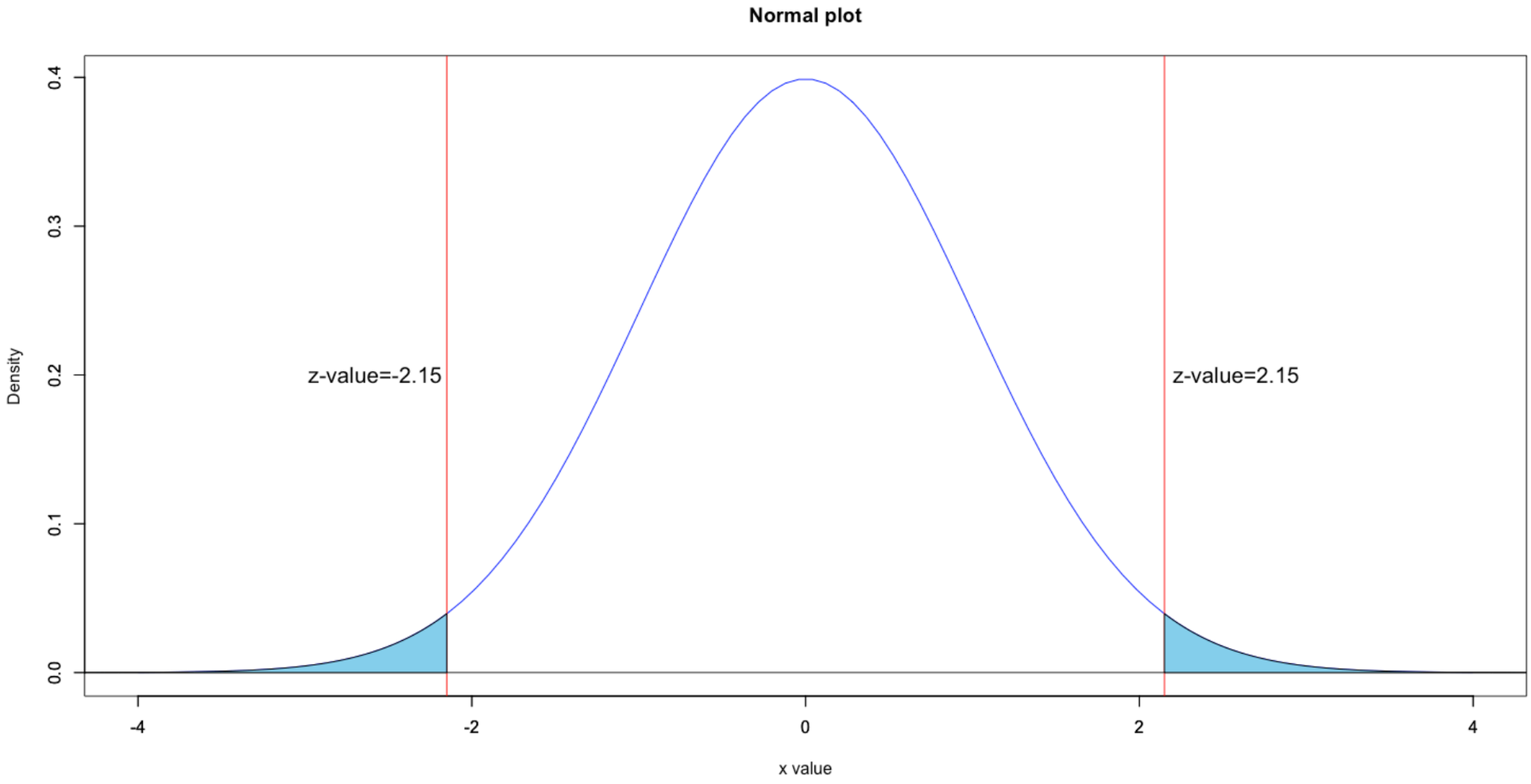
$$z\_value = \frac{\overline{X} - \overline{Y}}{SE}$$

$$SE = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.5 \cdot \sqrt{\frac{1}{11} + \frac{1}{10}} = 1.09 \qquad \overline{X} = 7.545 \qquad \overline{Y} = 5.2$$

$$z\_value = \frac{7.55 - 5.20}{1.09} = 2.16$$

Решение

В данном случае нам нужно проверить двухстороннюю гипотезу.



$$p\_value = 1 - P(x < 2.16) + P(x < -2.16) = 2 \cdot P(x < -2.16) = 0.03 < 0.05$$

В последнем равенстве мы использовали факт того, что нормальное распределение симметрично.

Значит на уровне значимости 0.05 мы отвергаем H0 и можем утверждать, что имеется значимая разница между временем реакций девушек

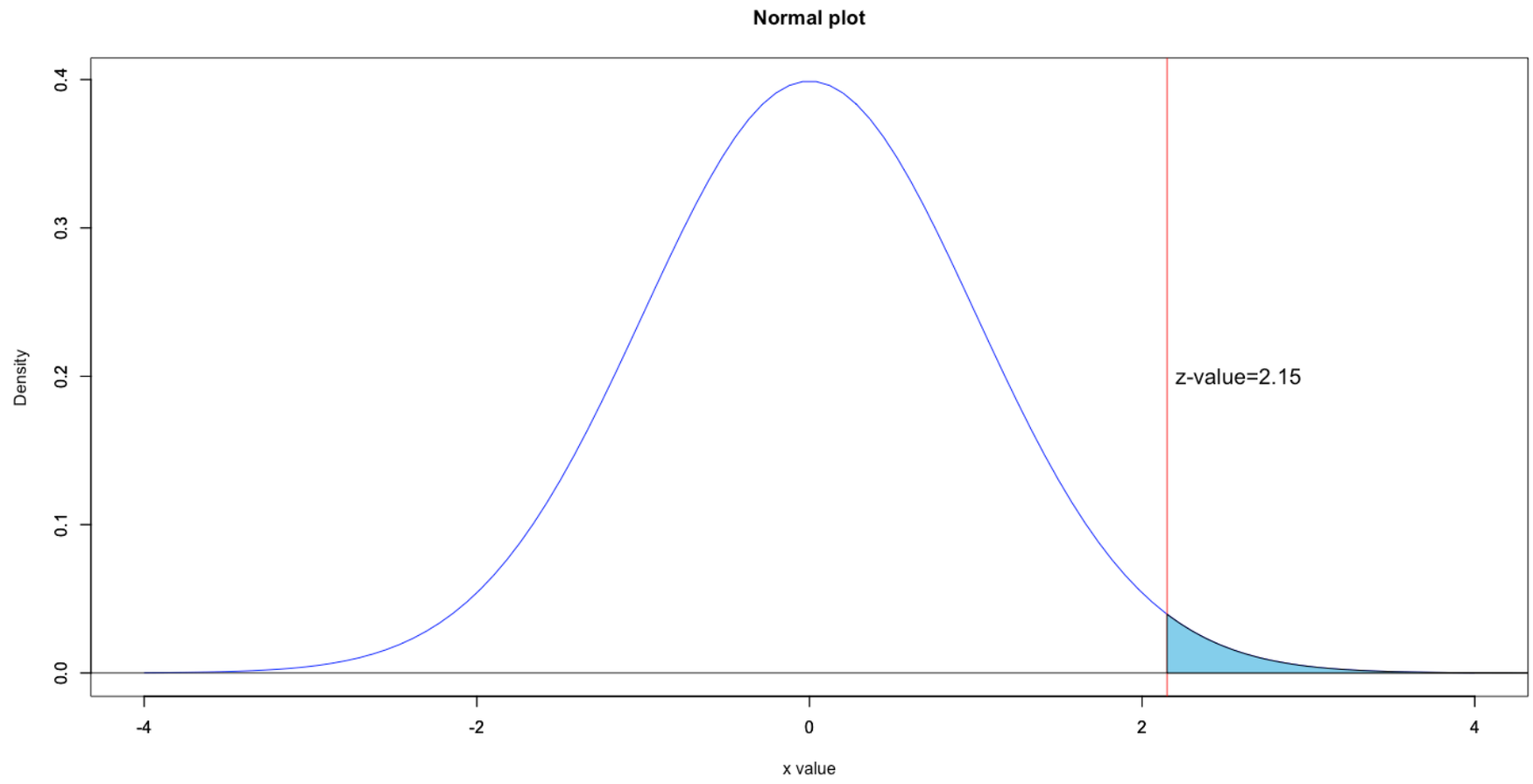
Решение

2) Сформулируем гипотезы H0 и H1.

H0 : девочки реагируют одинаково

H1: Света думает дольше, прежде чем ударить Федю

В данном случае нам нужно проверить одностороннюю гипотезу, причем правостороннюю, т.к мы утверждаем, что Света думает дольше



$$p\_value = 1 - P(x < 2.16) = 0.015$$

Значит на уровне значимости 0.05 мы отвергаем H0 и можем утверждать, что Света думает дольше, прежде чем ударить, и, по теории Феди, он нравится ей больше

Тест Стьюдента и стандартное отклонение

В общем случае

$$SE(\overline{X} - \overline{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

Если дисперсии равны

$$SE(\overline{X} - \overline{Y}) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$df = n_1 + n_2 - 2$$

Если выборка парная

$$SE(\overline{X} - \overline{Y}) = SE(X - Y)$$

$$df = n - 1$$

Задача

Четвероклассник Ваня, друг Феди, решил провести свой эксперимент, так как он считает наоборот, девочке, которая реагирует быстрее, он нравится больше. Он точно знает, что нравится Соне. С другой стороны, ему точно известно, что Кате нравится Миша, а не он. Он планировал провести эксперимент так же, как и Федя. Он успел собрать 12 наблюдений для Сони и только 9 для Кати, так как на этапе получение 10-го значения, ему дал в глаз Миша, нарушив этим чистоту эксперимента. После этого Ваня решил, что собранных данных достаточно.

Значения для Сони:

[6.33, 2.33, 4.98, 3.55, 6.61, 1.18, 2.21, 7.49, 7.53, 7.21, 9.29, 7.37]

Значения для Кати:

[10.52, 10.55, 10.06, 6.50, 9.45, 6.70, 9.25, 11.61, 10.23]

Ваня не верит данным из предыдущей задачи о значении стандартного отклонения для реакции детей, но считает это значение равно для всех детей.

Подтверждается ли гипотеза Вани при условии нормальности распределений?

Решение

Сформулируем гипотезы H0 и H1.
H0 : девочки реагируют одинаково
H1: Соня значимо реагирует быстрее с уровнем значимости alpha = 0.05

В отличии от предыдущего раза, мы не знаем, реальных дисперсий для наших выборок, но знаем, что они равны.
Потому нам необходимо для начала его подсчитать.

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} \qquad s_x = 2.60 \qquad s_y = 1.74$$

$$s = \sqrt{\frac{(12-1)s_x^2 + (9-1)s_Y^2}{12+9-2}} = 2.27$$

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.27 \cdot \sqrt{\frac{1}{12} + \frac{1}{9}} = 1.0$$

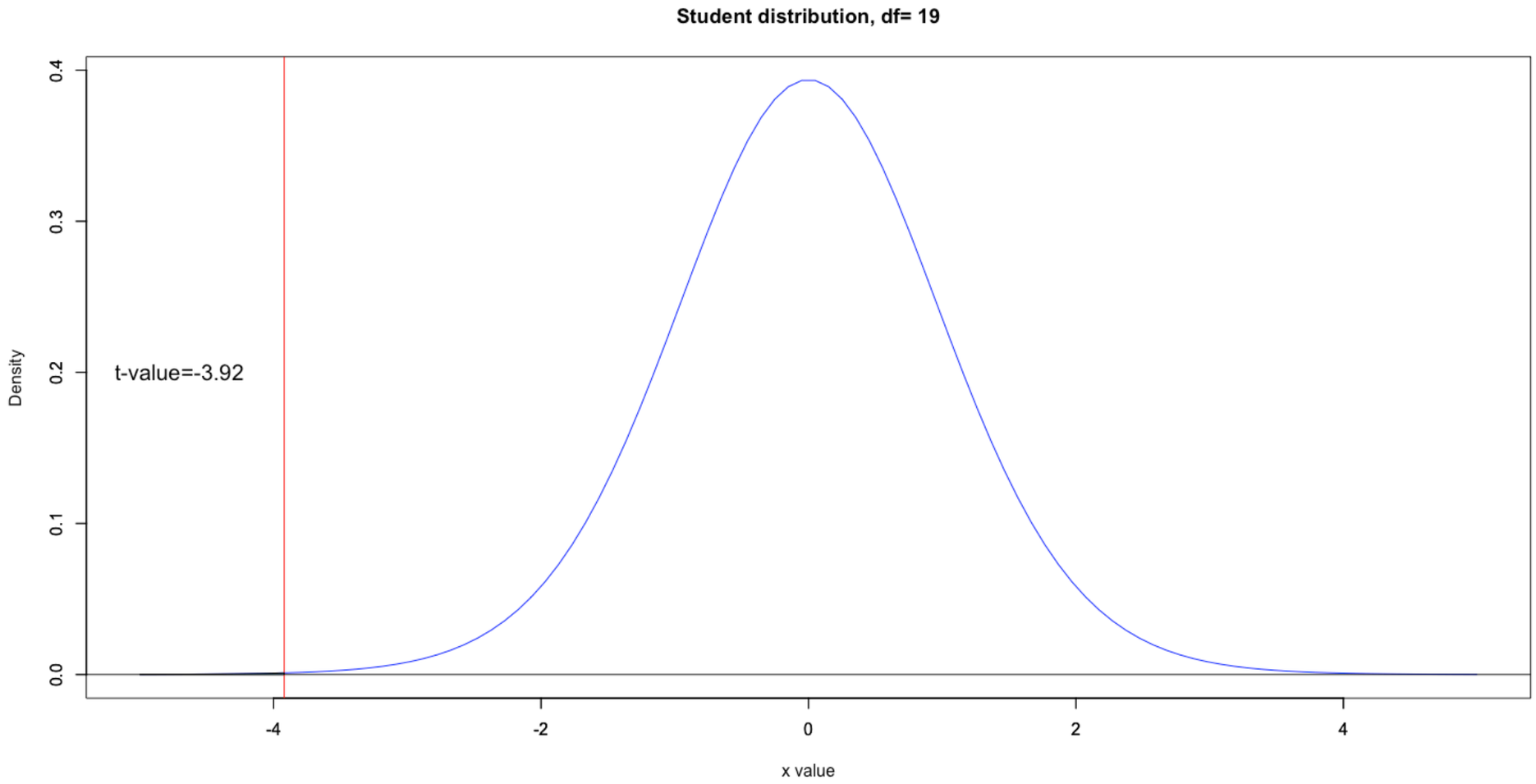
Кроме того, мы теперь не можем использовать нормальное распределение, так данных недостаточно и мы используем выборочное стандартное отклонение.

$$t\_value = \frac{5.51 - 9.43}{1.00} = -3.92 \qquad \overline{X} = 5.51 \qquad \overline{Y} = 9.43$$

Решение

В данном случае число степеней свободы равно

$$df = n_1 + n_2 - 2 = 19$$



Рассматриваем левостороннюю альтернативу

$$p\_value = P(x < -3.92) = 0.0005$$

Значит на уровне значимости 0.05 мы отвергаем H0 и можем утверждать, что Соня реагирует быстрее

Задача

Молодой немецкий фермер Ганс хочет проверить два новых фосфорорганических удобрения под торговыми марками “ГенФос” и “Кузнечик”. В результате испытаний на двух группах по 9 и 7 растений соответственно он получил, что среднее количество плодов с растений первой группы - 10.4, а со второй - 12. Выборочные стандартные отклонения - 1.7 и 2.1. Принимая во внимание предположение о том, что дисперсии в обеих группах были одинаковы и распределение количества плодов нормальное, можно ли утверждать, что какое-то из удобрений лучше?

Решение

Сформулируем гипотезы H0 и H1.
H0 : Удобрения одинаковые
H1: Второе удобрение работает лучше

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

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} = 1.88$$

$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.88 \cdot \sqrt{\frac{1}{9} + \frac{1}{7}} = 0.95$$

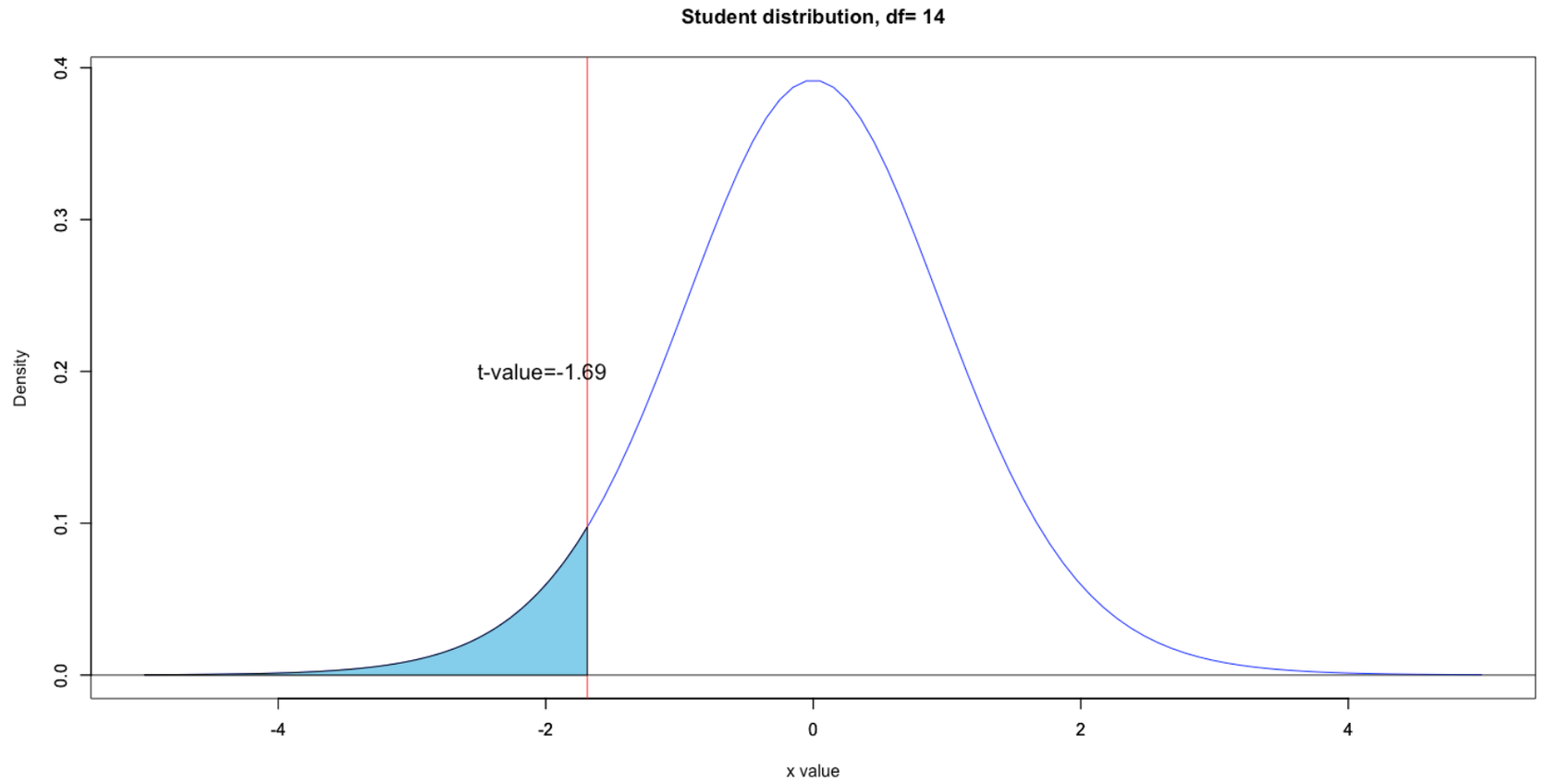
Кроме того, мы теперь не можем использовать нормальное распределение, так данных недостаточно и мы используем выборочное стандартное отклонение.

$$t\_value = \frac{10.4 - 12}{0.95} = -1.69$$

Решение

Число степеней свободы в данном случае равно, альтернатива левосторонняя

$$df = n_1 + n_2 - 2 = 7 + 9 - 2 = 14$$



$$p\_value = P(x < -1.69) = 0.056$$

Значит на уровне значимости 0.05 мы не имеем достаточных оснований для отвержения H0

Заметим, что если бы по ошибке мы использовали в данном примере нормальное распределение, то мы получили

$$p\_value\_wrong = P(x < -1.69) = 0.046$$

и ошиблись в решении задачи

Задача

Молодой немецкий фермер Ганс хочет проверить два новых фосфорорганических удобрения под торговыми марками “ГенФос” и “V-Product”. В результате испытаний на двух группах по 42 и 45 растений соответственно он получил, что среднее количество плодов с растений первой группы - 10.5, а со второй - 13. Выборочные стандартные отклонения - 1.1 и 1.2. Также стоит отметить, что второе удобрение дороже, потому Ганс хотел бы его использовать только в том случае, если разница в продукции статистически достоверно больше 2 плодов. Принимая во внимание предположение о том, что дисперсии в обеих группах были одинаковы, какое из удобрений стоит выбрать Гансу на уровне значимости 0.05?

Решение

Сформулируем гипотезы H0 и H1.
H0 : Разница между применением двух удобрений не больше 2 плодов
H1: Второе удобрение увеличивает продукцию плодов на не менее чем 2 плода больше, чем первое

В данном случае мы имеем достаточно наблюдений для того, чтобы использовать нормальное распределение для тестирования наших гипотез. Аналогично предыдущим случаям посчитаем s и SE

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} = 1.15$$

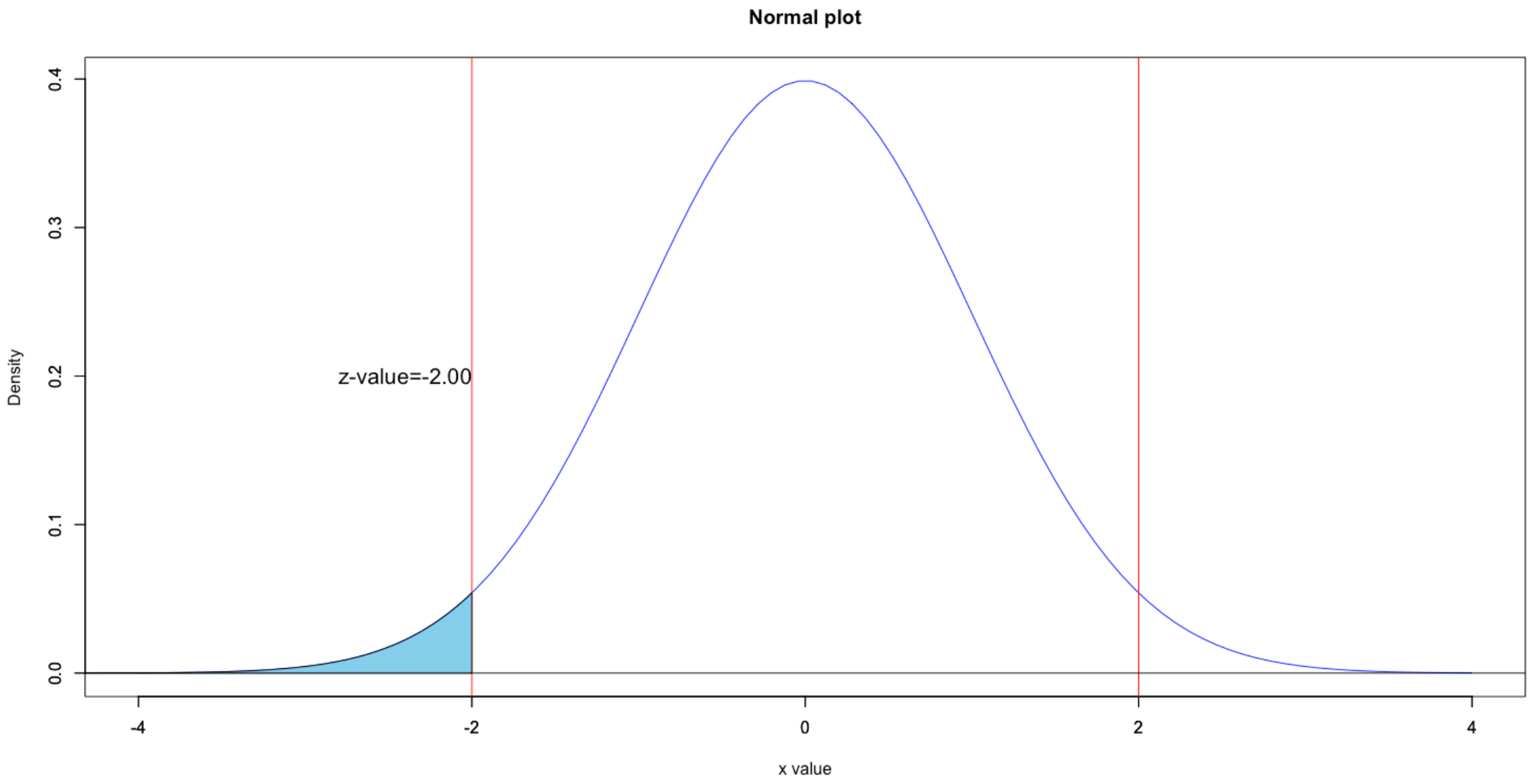
$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.15 \cdot \sqrt{\frac{1}{42} + \frac{1}{45}} = 0.25$$

Теперь мы считаем z-value, но прибавляем к разнице средних 2, таким образом учитывая требование, что второе удобрение в результатах должно быть лучше не меньше чем на 2 плода

$$z\_value = \frac{\overline{X} - \overline{Y} + \mu}{s} = \frac{(10.5 - 13) + 2}{0.25} = -2$$

Решение

Наша альтернатива опять является левосторонней.



$$p\_value = P(x < -2) = 0.02$$

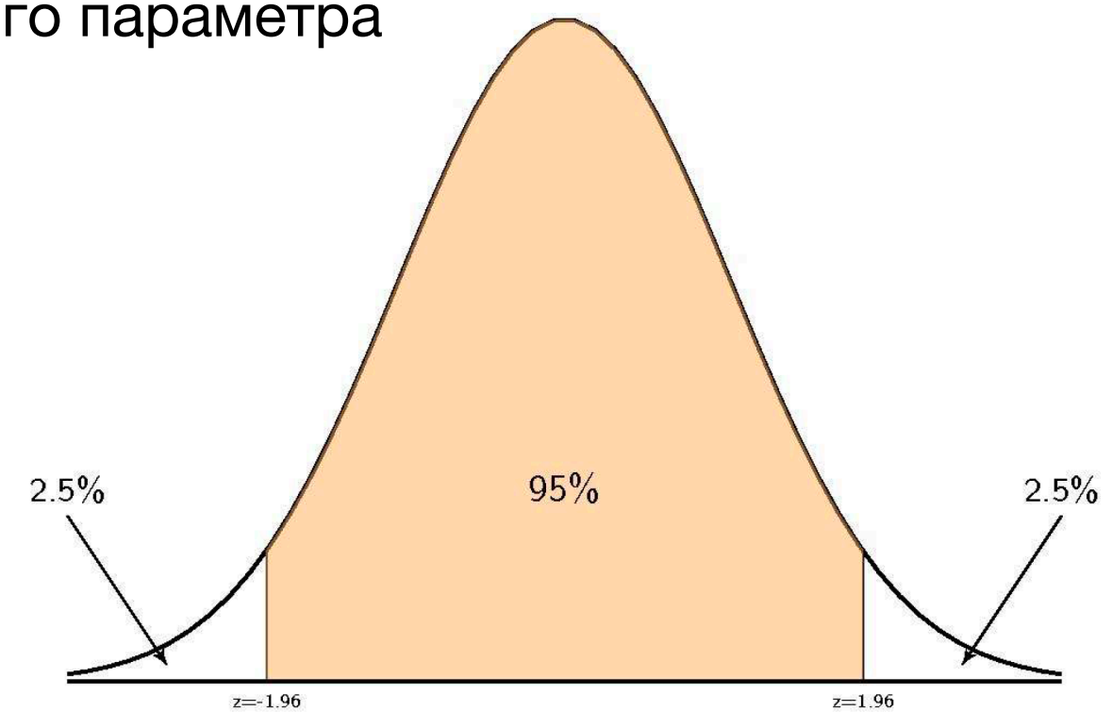
Значит на уровне значимости 0.05 мы отвергаем H0, второе удобрение удовлетворяет условиям Ганса

Доверительный интервал

Часто нас интересует интервал, в котором лежит оцениваемый параметр.
Мы можем оценить, в каком количестве экспериментов с аналогичным построением определенного интервала значений параметра, этот интервал будет содержать оцениваемый параметр. Такой интервал называется **доверительным интервалом.**

Еще возможны интерпретации:

1) Мы на 95% уверены, что наш параметр лежит в заданном интервале
2) Если мы верим, что наша выборка “типичная”, то есть 95% выборок такие же, то тогда построенный доверительный интервал содержит истинное значение оцениваемого параметра



Это важно!

Доверительный интервал не является интервалом, содержащим оцениваемый параметр с заданной вероятностью

Доверительный интервал

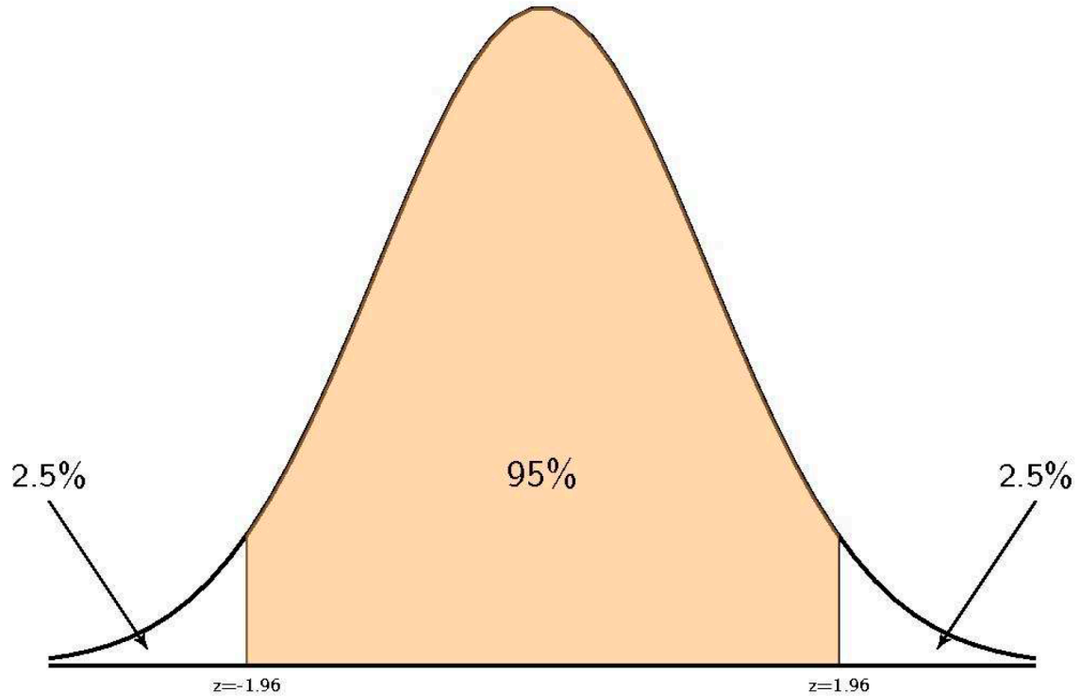
$$Parameter = Estimate \pm Critical * SE$$

$$SE = Standard\ error$$

$$Critical = critical\ value$$

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

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



Для любого симметричного распределения d

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

Задача

Программист Петя решил изучить Haskell. Его товарищ по работе Иван сказал, что тот стал материться чаще, чем раньше, когда программировал только на C++. Петя не поверил данному заявлению, но Иван предоставил информацию, согласно которой в месяц, когда Петя еще не учил этот новый язык, он произносил в среднем 10.5 матных слова за рабочий день, а за этот месяц среднее число слов равно 15.5.
Число дней в обоих приведенных месяцах 31. Известно, что стандартное отклонение числа матных слов, произносимых Петей осталось постоянным и равно 4. Построить 95% доверительный интервал для разницы между числом произносимых матных слов Петей до и после начала изучения Haskell.

Решение

Так как нам известно настоящее стандартное отклонение, будем использовать нормальное распределение для оценки интервала

$$delta = \overline{X} - \overline{Y} \pm z_{\alpha/2} \cdot SE$$

$$SE = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4 \cdot \sqrt{\frac{1}{31} + \frac{1}{31}} = 1.0$$

$$delta = \overline{X} - \overline{Y} \pm z_{\alpha/2} \cdot SE = 15.5 - 10.5 \pm z_{0.025} \cdot 1.00 = 5 \pm 2$$

Таким образом изменение количества матных слов в день, произносимых Петей, увеличилось на значение от 3 до 7 слов на уровне значимости 0.05.

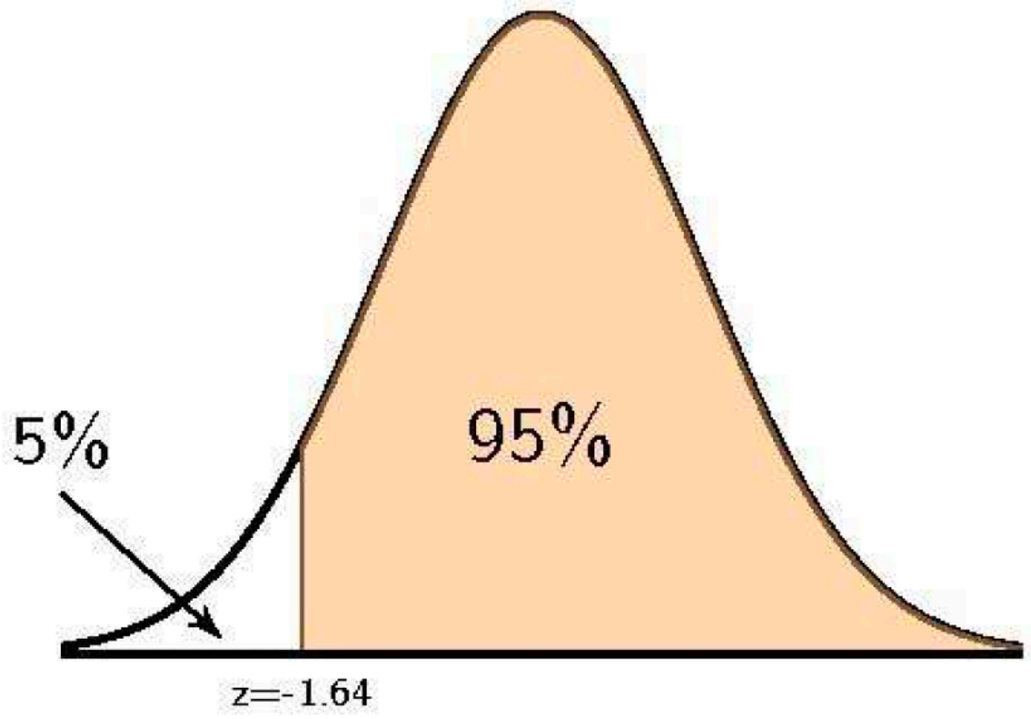
$$delta \in [3, 7]$$

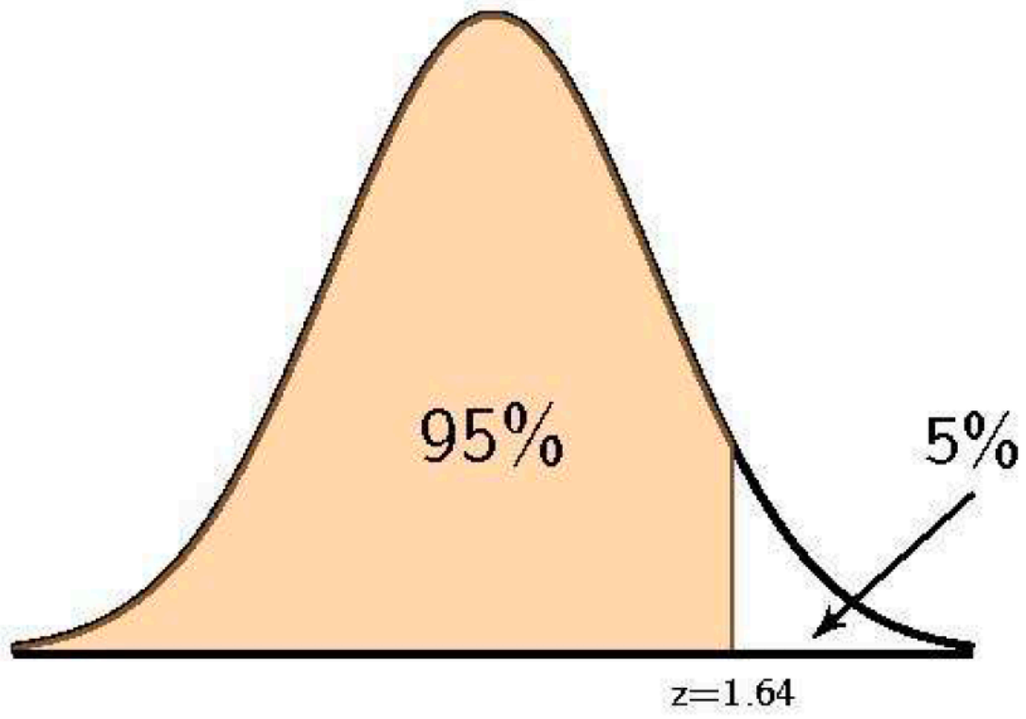
Заметим, что этот доверительный указывает так же и на то, что увеличение статистически значимо отличается от 0

Односторонний доверительный интервал

$$\mu > \overline{X} - z_\alpha \cdot SE \qquad \mu < \overline{X} + z_\alpha \cdot SE$$

$$\mu > \overline{X} - t_\alpha \cdot SE \qquad \mu < \overline{X} + t_\alpha \cdot SE$$





Задача

Пациент заявляет, что он потребляет только 2000 калорий в день, но диетолог подозревает, что на самом деле это больше. Он проверил меню пациента в течении 30 дней и определил, что среднее за это время больше чем 2100 калорий. Постройте односторонний 95% односторонний доверительный интервал для количества калорий в диете пациента. Считайте, что стандартное отклонение равно 350 калориям в день

Решение

Так как нам известно стандартное отклонение, будем использовать нормальное распределение для оценки параметров

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

$$SE = \sigma\sqrt{\frac{1}{n}} = 63.9$$

$$\mu = \overline{X} - z_{\alpha} \cdot SE = 2100 - 63.9 \cdot z_{0.05} = 2100 - 105 = 1995$$

Таким образом, среднее число потребляемых пациентом калорий лежит в интервале от 1995 до +бесконечности

$$\mu \in [1995, +\infty]$$

Задача

Молодой немецкий фермер Ганс хочет проверить два новых фосфорорганических удобрения под торговыми марками “ГенФос” и “V-Product”. В результате испытаний на двух группах по 42 и 45 растений соответственно он получил, что среднее количество плодов с растений первой группы - 10.5, а со второй - 13. Выборочные стандартные отклонения - 1.1 и 1.2. Также стоит отметить, что второе удобрение дороже, потому Ганс хотел бы его использовать только в том случае, если разница в продукции статистически достоверно больше 2 плодов. Принимая во внимание предположение о том, что дисперсии в обеих группах были одинаковы, какое из удобрений стоит выбрать Гансу на уровне значимости 0.05?

Решение

Покажем, что эту задачу можно решить при помощи подсчета доверительного интервала

$$s = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_Y^2}{n+m-2}} = 1.15$$

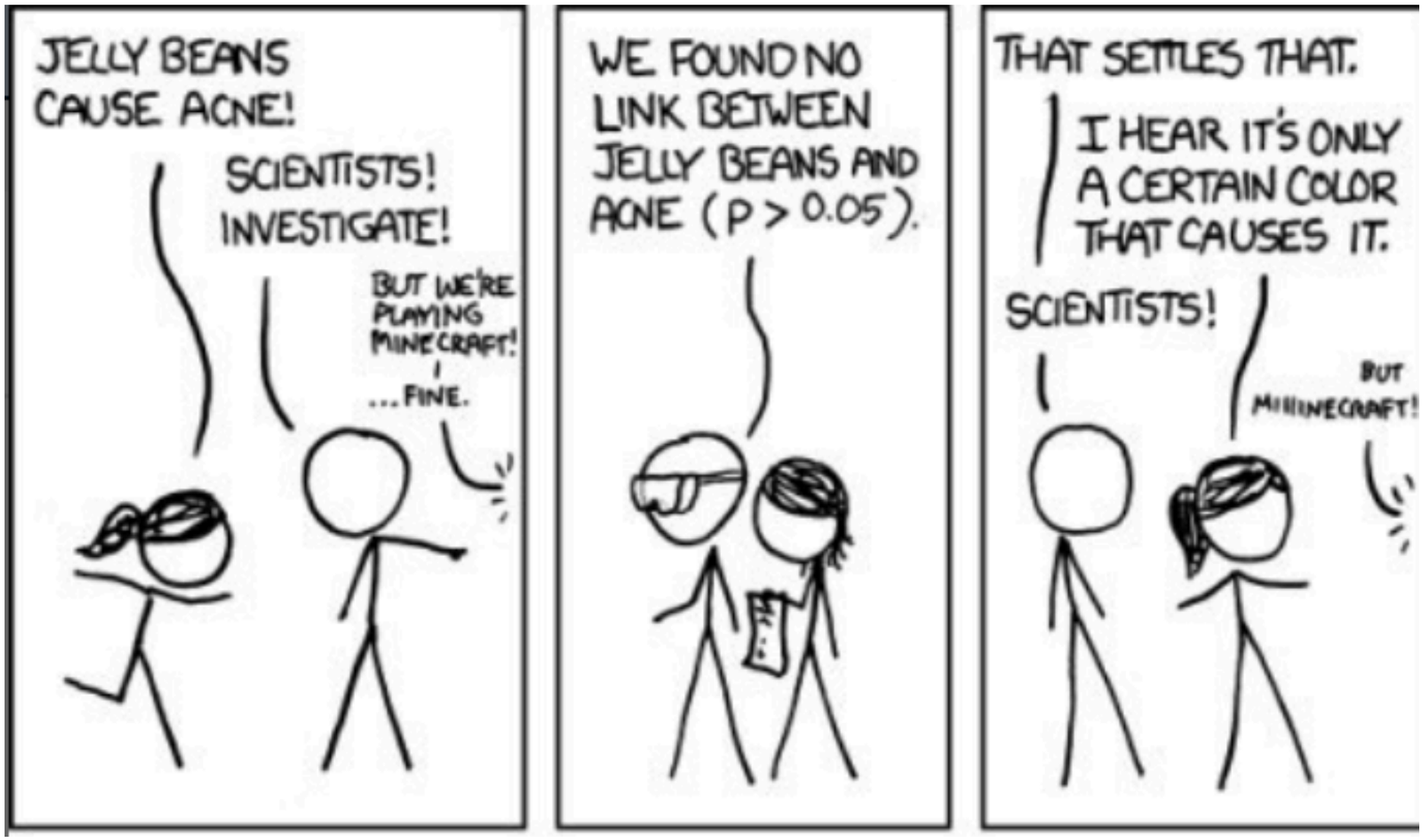
$$SE = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.15 \cdot \sqrt{\frac{1}{42} + \frac{1}{45}} = 0.25$$

$$delta = \overline{X} - \overline{Y} - z_{\alpha/2} \cdot SE = 13 - 10.5 \pm z_{0.05} \cdot 0.25 = 2.5 - 0.25 \cdot 1.64 = 2.5 - 0.41 = 2.09$$

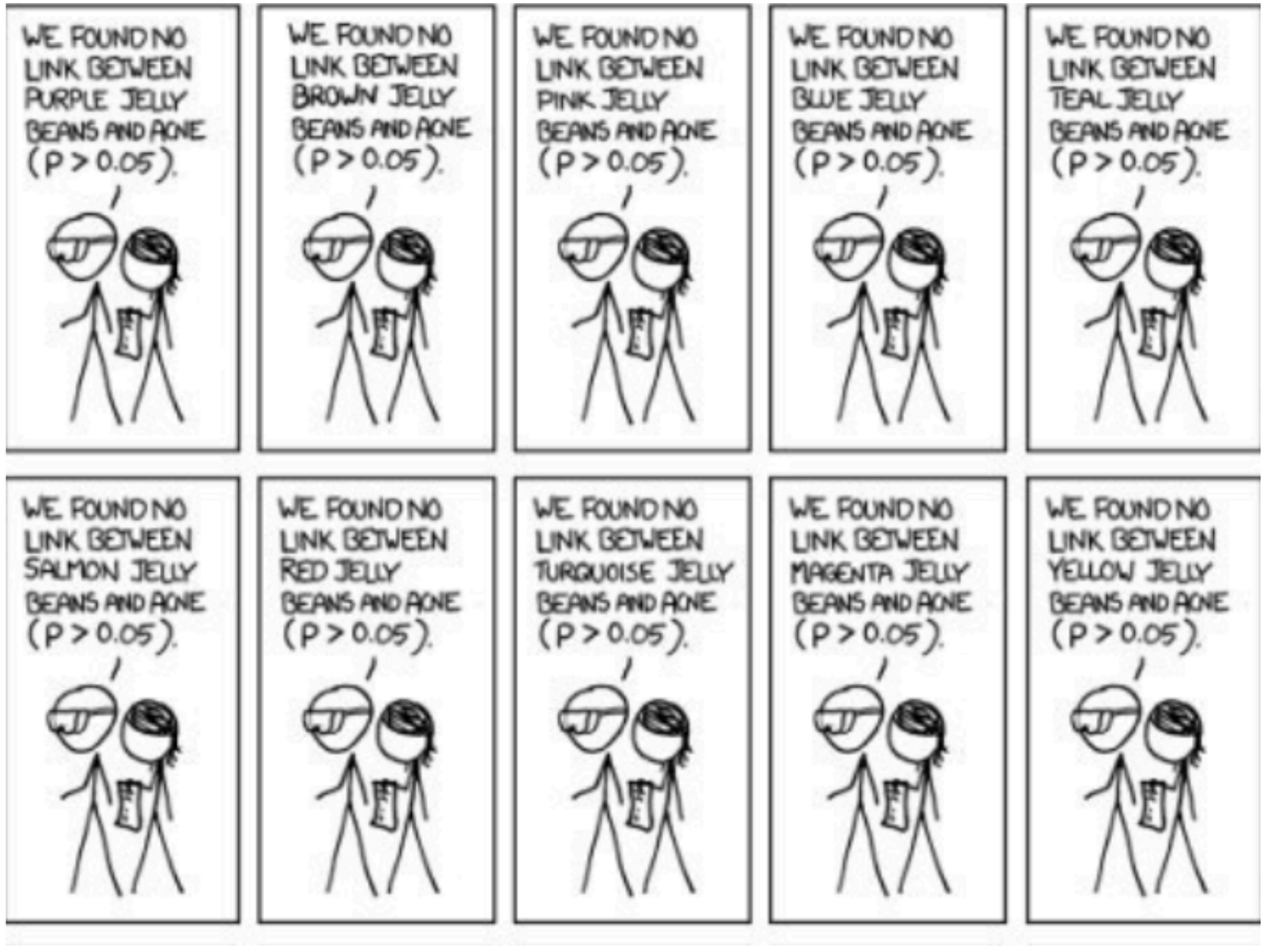
$$delta \in [2.09, +\infty]$$

Так как значения меньше 2 не входят в данный интервал, второе удобрение подходит требованиям Ганса

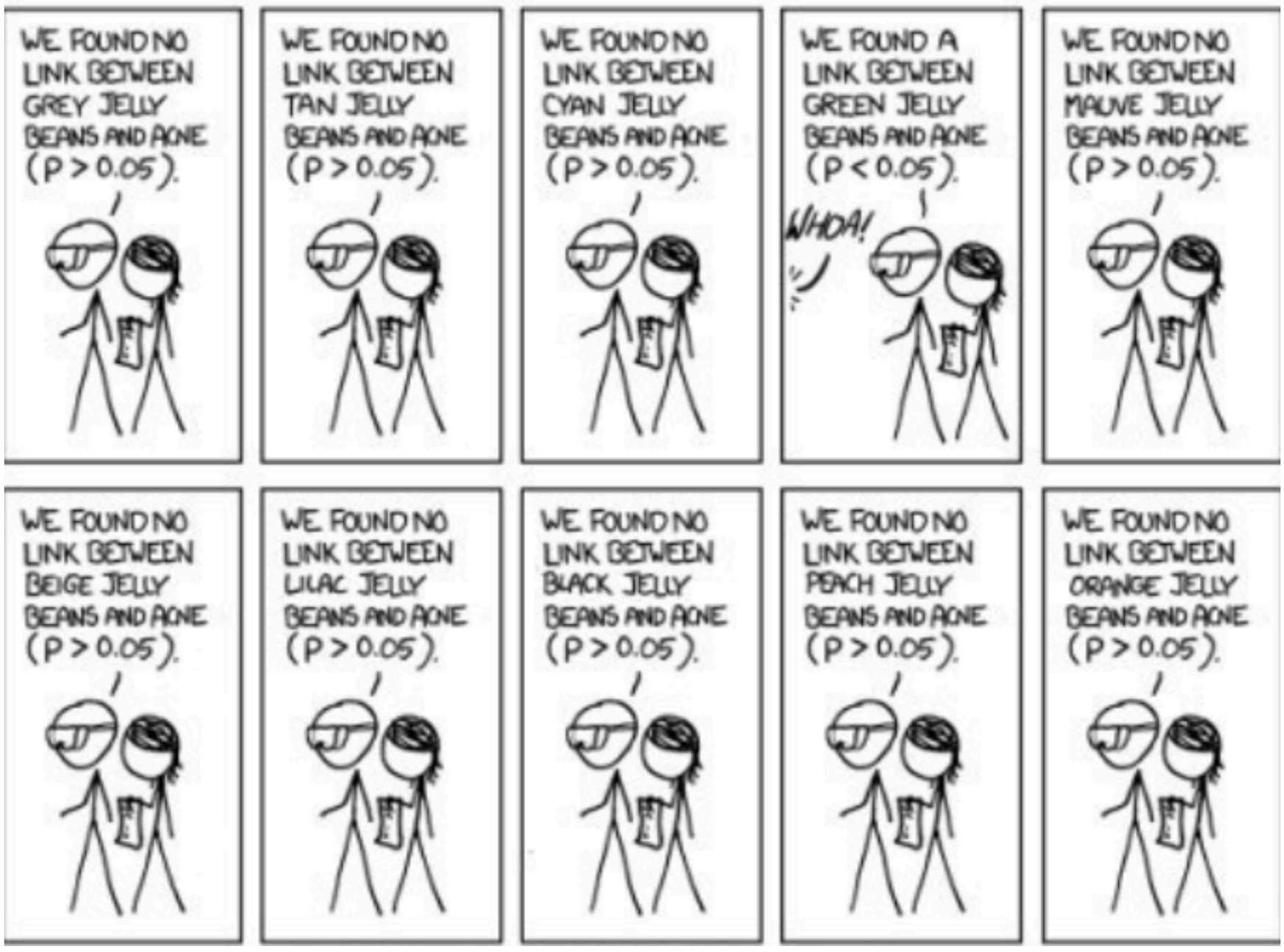
Множественное тестирование



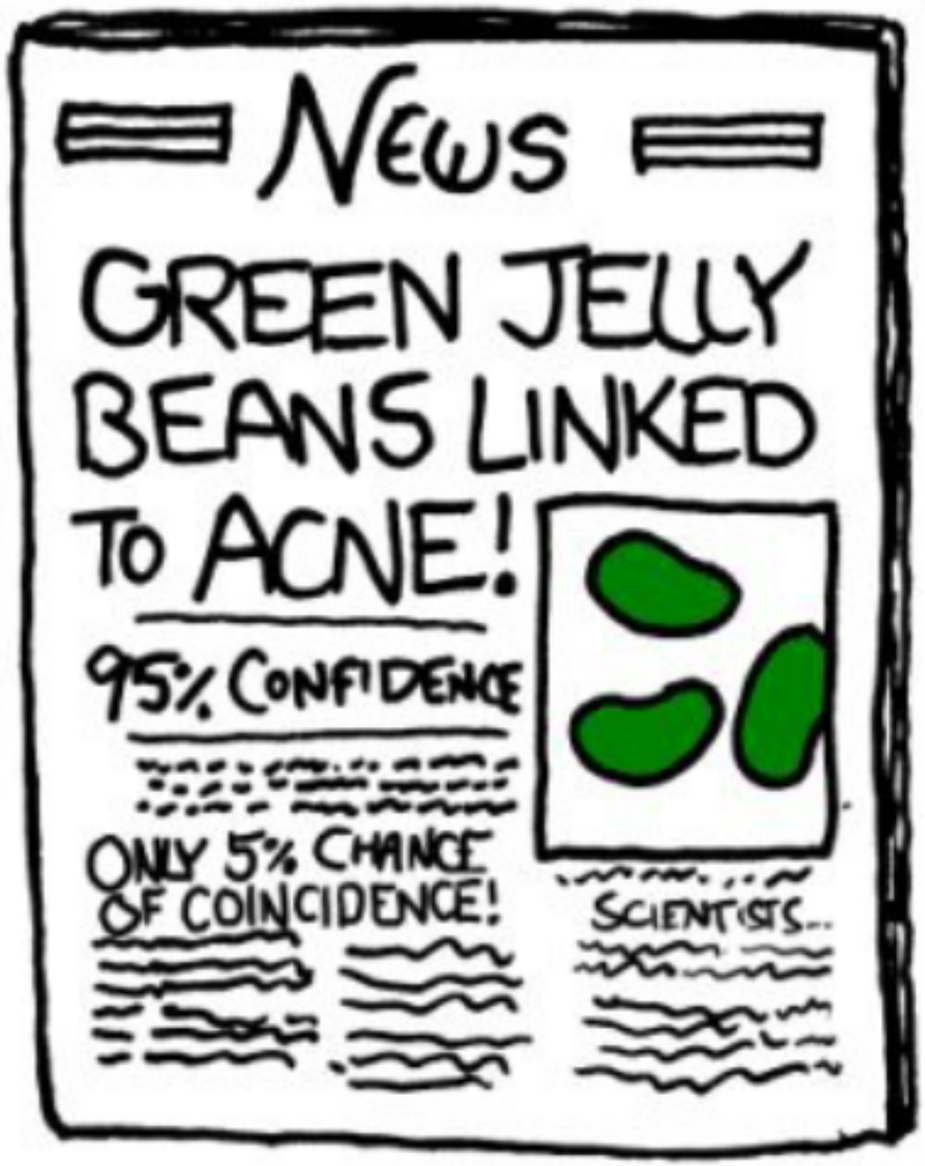
Множественное тестирование



Множественное тестирование



Множественное тестирование



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?

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%

FWER

One-step procedures:

1) Sidak correction
2) Bonferonni correction

Step-down procedures:

1) Holm-Sidak correction
2) Holm-Bonferonni correction

Step-up procedures:

Hochberg correction

Sidak

$$p < 1 - \sqrt[n]{1 - 0.05}$$

P-value генов независимы

Bonferonni

$$p < \frac{0.05}{N}$$

P-value генов могут быть зависимы

Holm-Sidak

$$p < 1 - \sqrt[N+1-k]{1 - 0.05}$$

P-value генов независимы

Holm-Bonferonni

$$p < \frac{0.05}{N + 1 - k}$$

P-value генов могут быть зависимы

FDR

- Benjamini-Hochberg correction

- Benjamini-Yekutieli correction