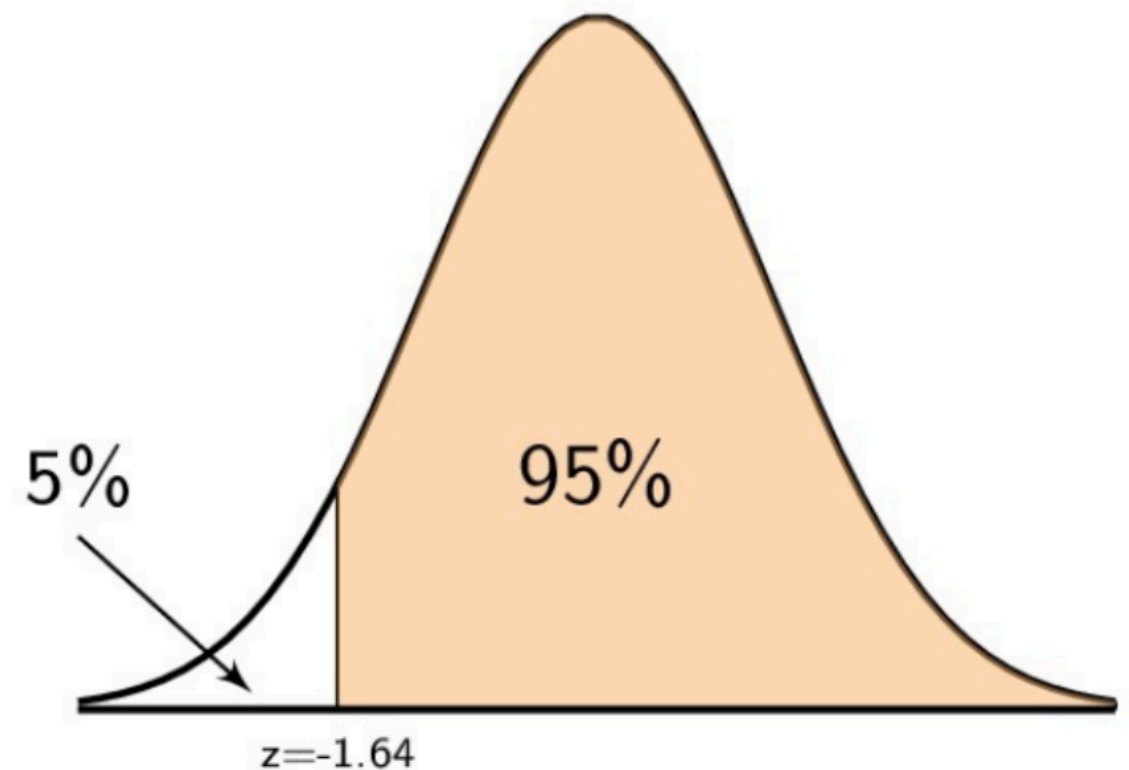
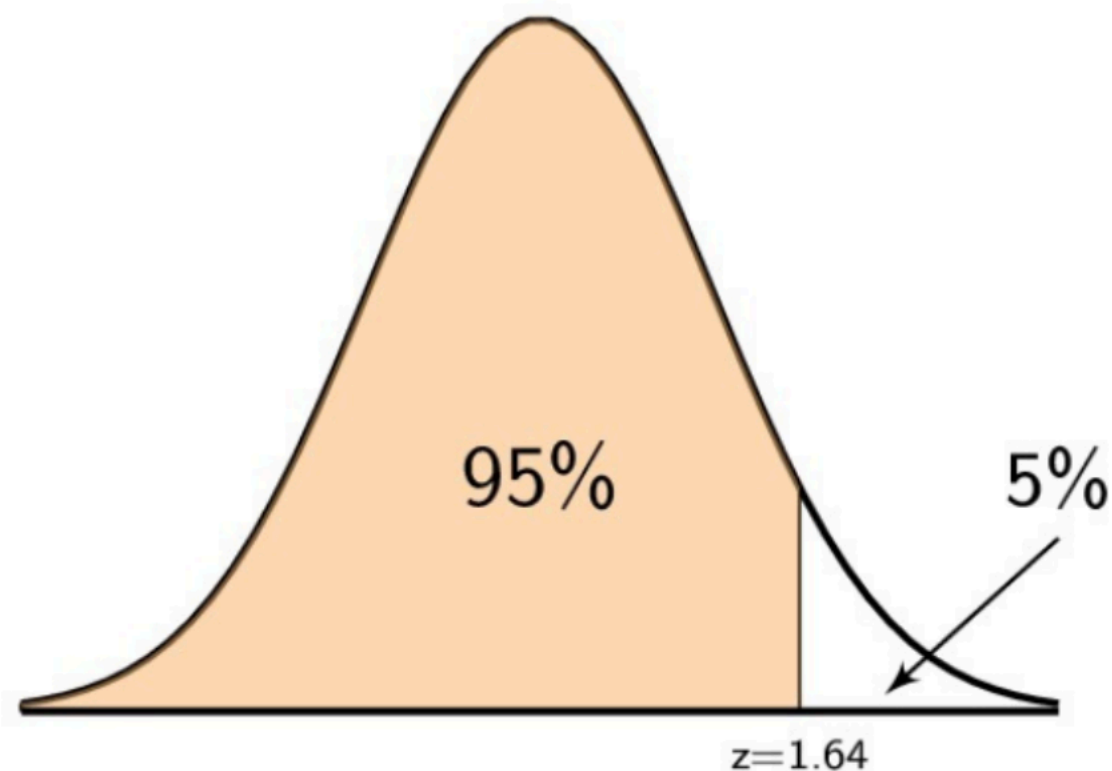


Односторонний доверительный интервал



Смотрим только с одной стороны

$$\mu < \bar{X} + |z_{\alpha}| \cdot SE$$

$$\mu > \bar{X} - |z_{\alpha}| \cdot SE$$

$$\mu_x - \mu_y < \bar{X} - \bar{Y} + |z_{\alpha}| \cdot SE$$

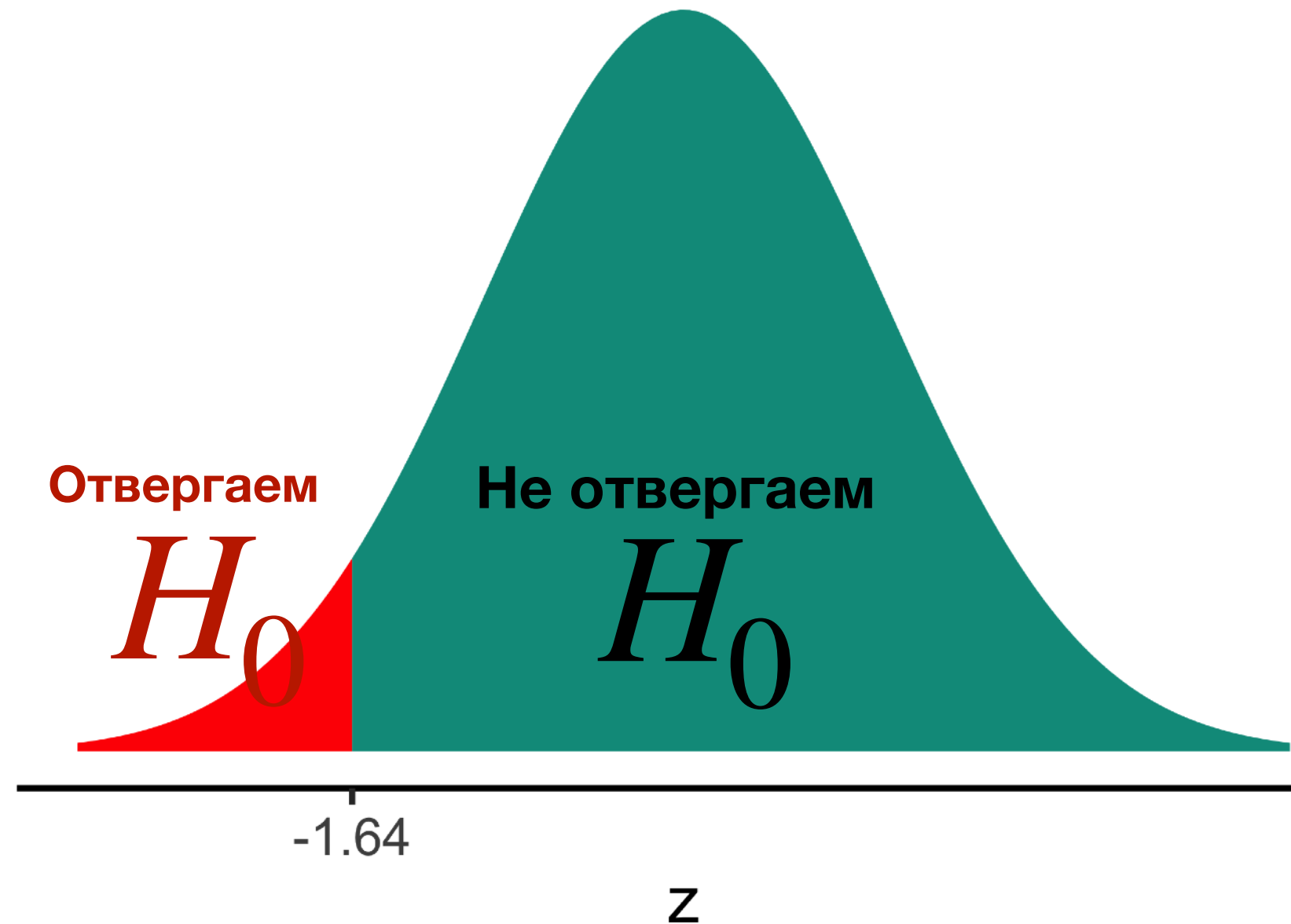
$$\mu_x - \mu_y > \bar{X} - \bar{Y} - |z_{\alpha}| \cdot SE$$

Для t-test аналогично

Связь доверительного интервала и принятия/непринятия гипотезы

$$H_0 : \mu_x - \mu_y = a$$

$$H_1 : \mu_x - \mu_y < a$$



равносильно

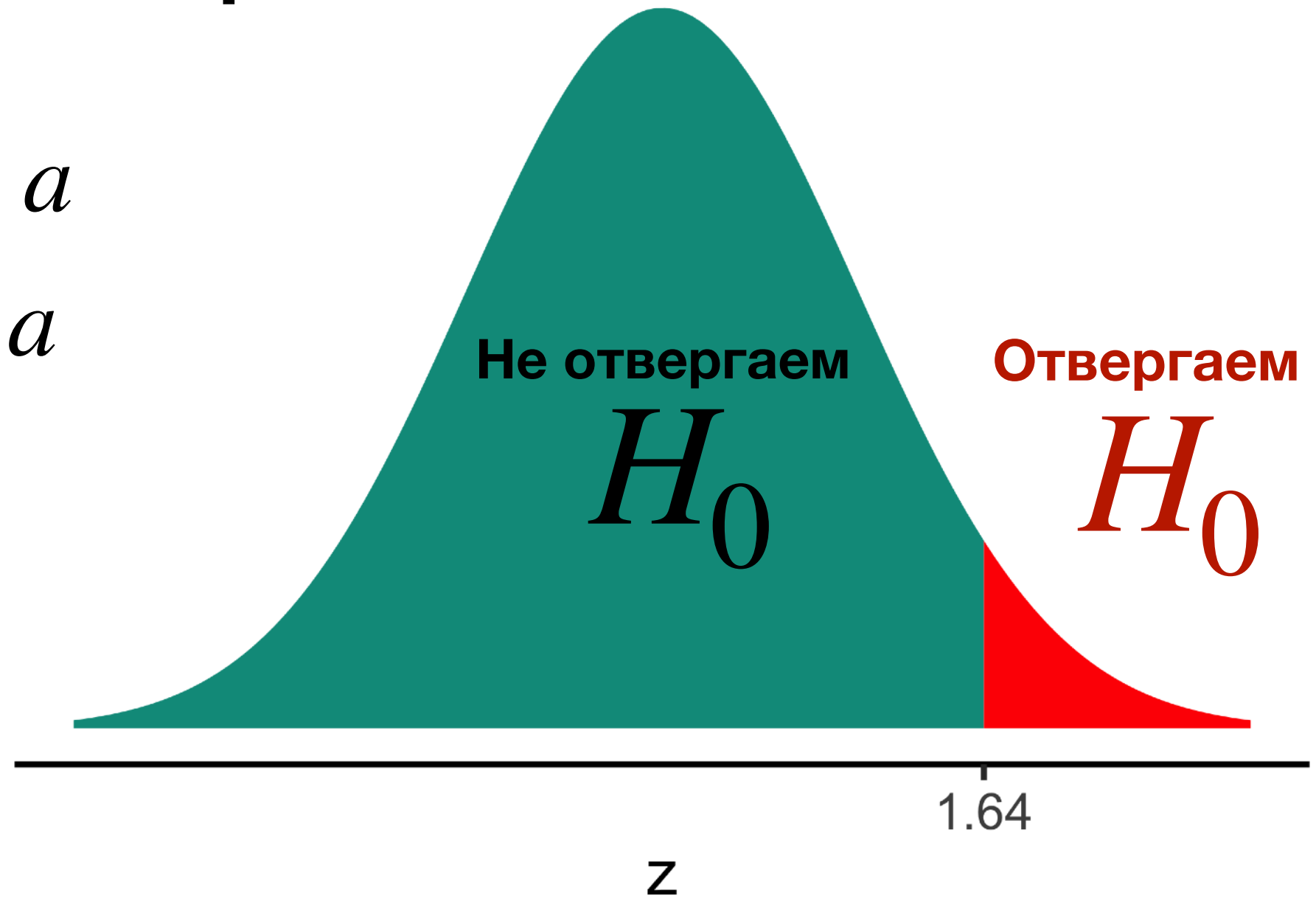
$$a \in [\bar{X} - \bar{Y} - z_\alpha \cdot SE, +\infty) \quad H_0 \text{ не отвергаем}$$

$$a \notin [\bar{X} - \bar{Y} - z_\alpha \cdot SE, +\infty) \quad H_0 \text{ отвергаем}$$

Связь доверительного интервала и принятия/непринятия гипотезы

$$H_0 : \mu_x - \mu_y = a$$

$$H_1 : \mu_x - \mu_y > a$$



равносильно

$$a \in (-\infty, \bar{X} - \bar{Y} + z_\alpha \cdot SE] \quad H_0 \text{ не отвергаем}$$

$$a \notin (-\infty, \bar{X} - \bar{Y} + z_\alpha \cdot SE] \quad H_0 \text{ отвергаем}$$

Задача

Копьеметатель утверждает, что средняя длина его броска не меньше 98.45м.

Копьеметатель при подготовке к турниру сделал серию из 12 бросков.

Среднее расстояние броска - 96.72 м. Из прошлых бросков известна дисперсия его бросков - 4.72 м.

С помощью построения доверительного интервала проверьте гипотезу о том, что

- 1) Постройте 99% односторонний доверительный интервал для длины броска.
- 2) Есть ли 1% значимости основание не доверять его заявлению?

Доверительный интервал для теста пропорций

$$SE = \sqrt{\frac{pq}{n}}$$

Двусторонний доверительный интервал

$$p \in \hat{p} \pm z_{\alpha/2} \cdot SE$$

$$\Delta p \in \hat{p} \pm z_{\alpha/2} \cdot SE$$

Односторонний доверительный интервал

$$p \in [\hat{p} - z_{\alpha} \cdot SE, +\infty)$$

$$p \in (-\infty, \hat{p} + z_{\alpha} \cdot SE]$$

$$\Delta p \in [\hat{p}_x - \hat{p}_y - z_{\alpha} \cdot SE, +\infty)$$

$$\Delta p \in (-\infty, \hat{p}_x - \hat{p}_y + z_{\alpha} \cdot SE]$$

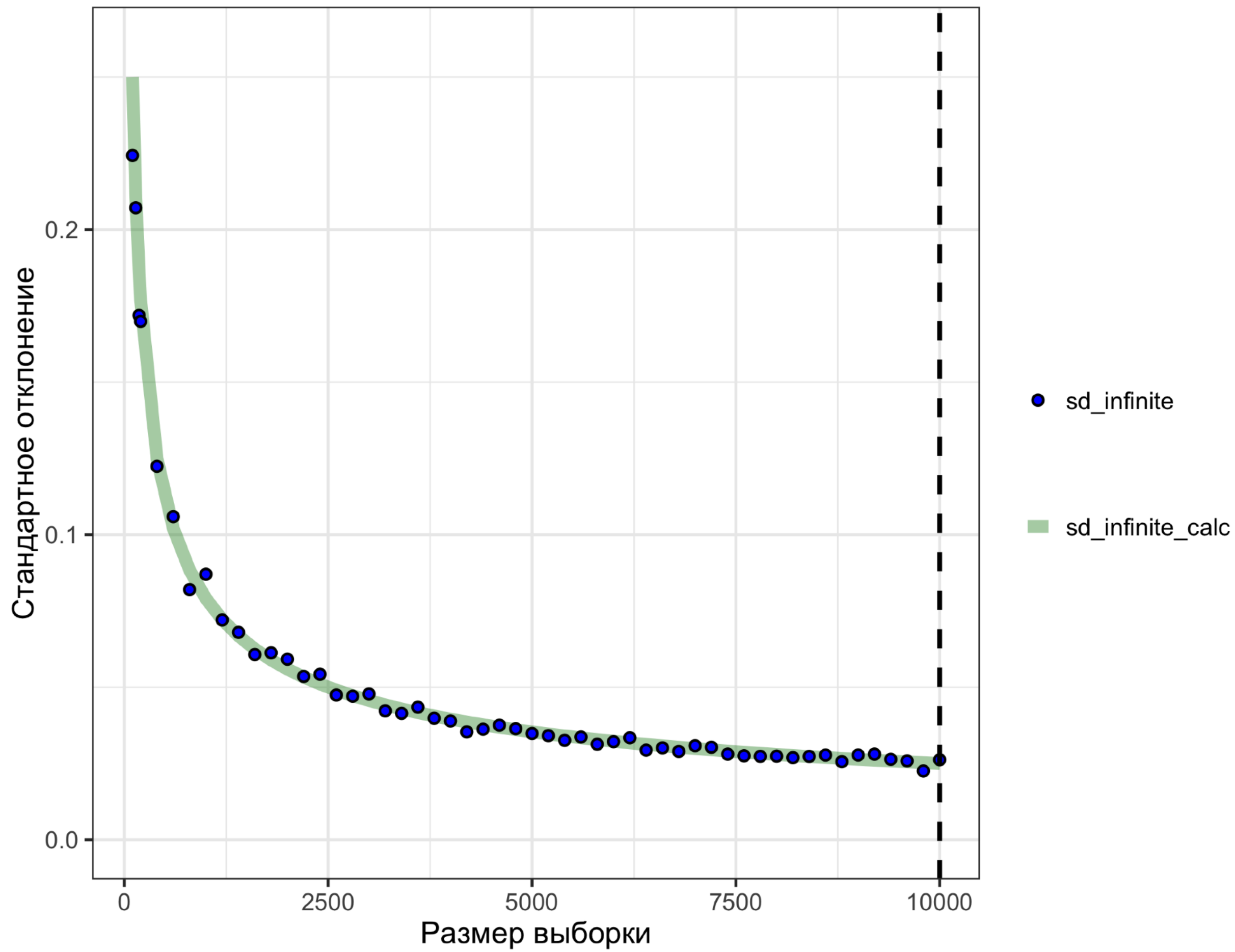
Задача

Была проведен опрос населения мегаполиса N. Согласно опросу, доля курящих - 11%. Было опрошено 100 человек.

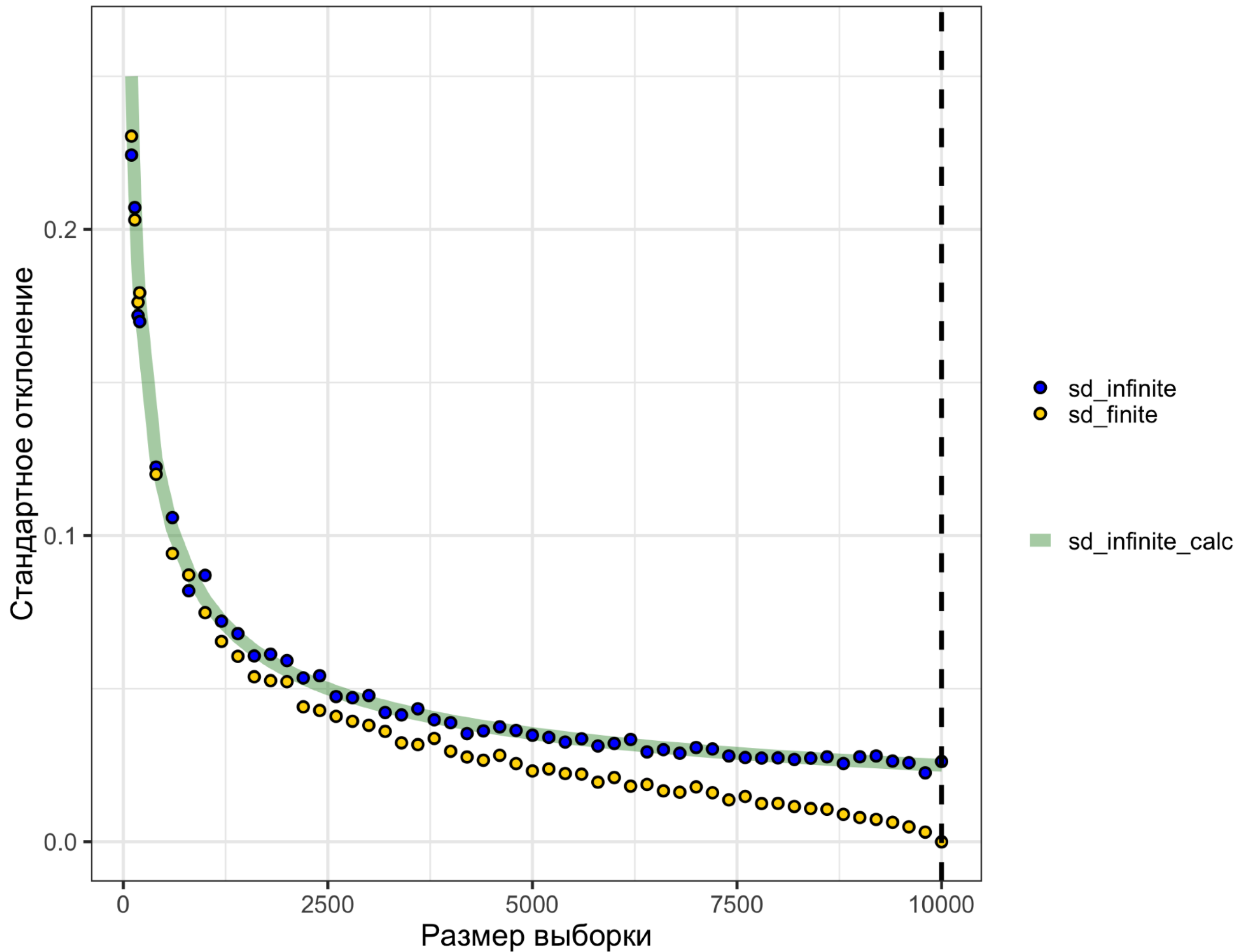
Для мегаполиса P., согласно опросу, доля курящих - 12%. Было опрошено 120 человек.

Постройте 95% доверительный интервал для разности пропорций

Поправка на конечность выборки



Поправка на конечность выборки



Поправка на конечность выборки

Первое объяснение (правильное)

Если генеральная совокупность конечная, то когда мы набираем из нее объекты для анализа,

они получаются зависимыми.

Среднему на зависимость между наблюдениями все равно, потому среднее выборочного среднего остается прежним

Но вот дисперсия выборочного среднего от этого меняется. Нужен поправочный фактор

Можно вывести (аналогично тому, как вы выводили среднее и дисперсию для гипергеометрического распределения на теорвере) формулу этой поправки

$$SE = \sqrt{\frac{N - n}{N - 1}} \cdot SE_{uncorrected}$$

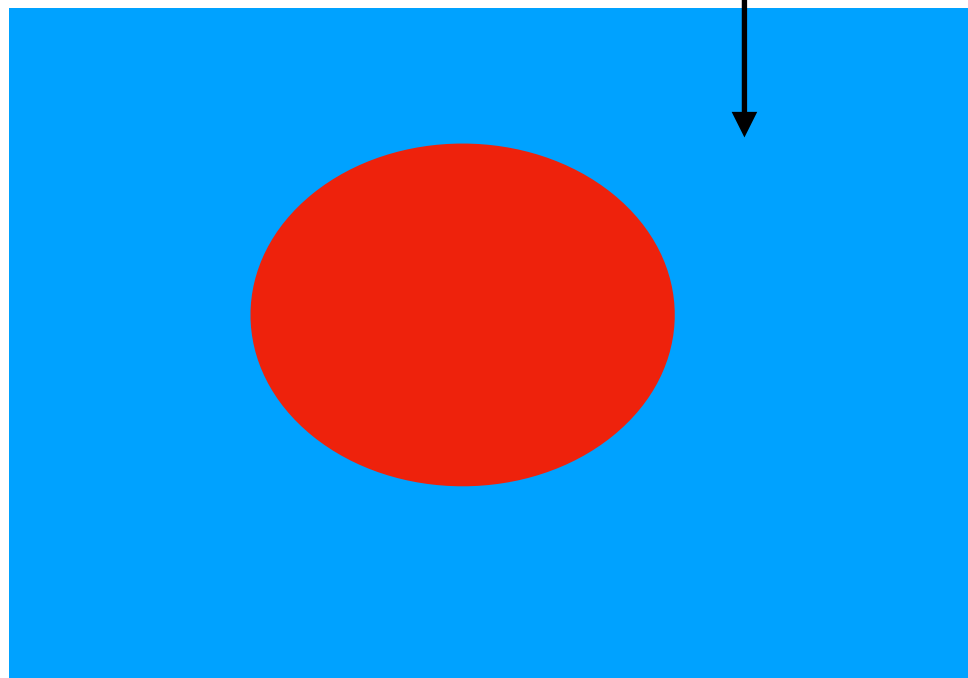
Второе объяснение (на пальцах)

Если генеральная совокупность конечная, то в принципе мы можем ее всю взять. В этом случае мы **точно** оценим ее среднее $\rightarrow SE=0$
Согласно же формуле, которую мы использовали раньше

$$SE(\hat{X}) = \frac{\sigma}{\sqrt{n}}$$

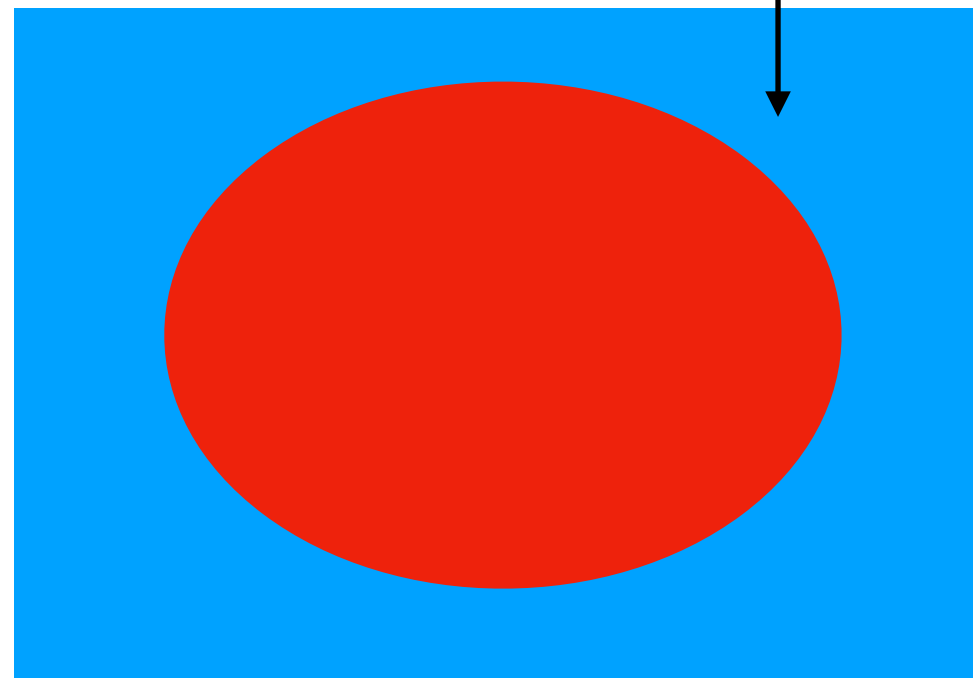
Даже если возьмем всю генсовокупность - по этой формуле ошибка есть. А она должна быть 0.
Аналогично, чем наша выборка по размеру ближе к генсовокупности, тем больше будет ошибаться эта формула. Потому нужна поправка

Sampling error



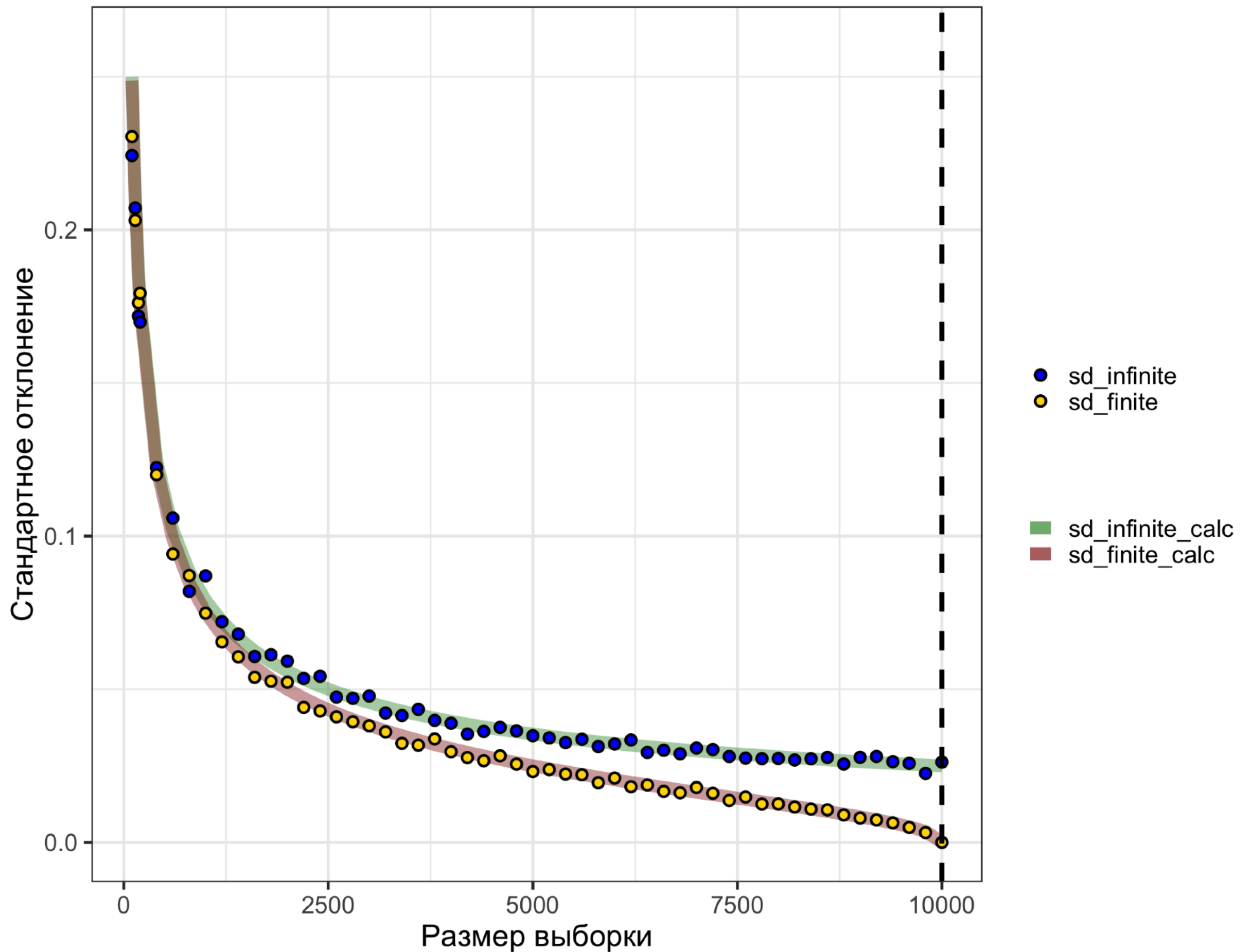
Популяция

Sampling error

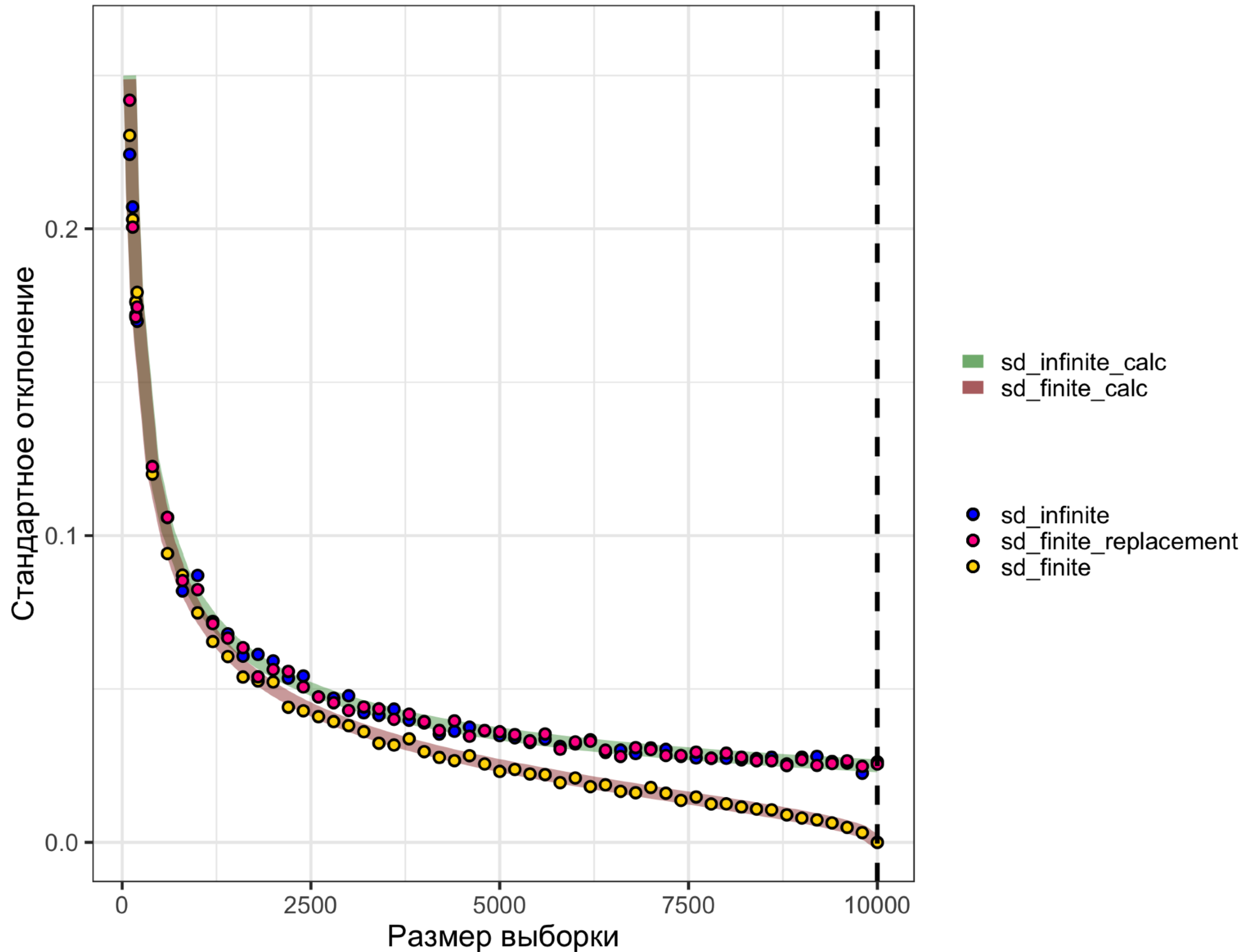


Популяция

Поправка на конечность выборки



Поправка на конечность выборки



Задача

Телекоммуникационная компания опрашивает жителей маленькой деревни, в которой проживает 100 семей. Случайная выборка из 12 семей имеет среднее значение интернет трафика 1.34 Гб со стандартным отклонением 0.60 Гб. Постройте 95% доверительный интервал для среднего интернет трафика семей, проживающих в данном городе, учитывая, что выборка была сделана без возвращения.

Для теста пропорций

Если берем из конечной выборки - число успехов имеет как раз гипергеометрическое распределение

$$E(k) = np$$

$$D(k) = npq \cdot \frac{N - n}{N - 1}$$

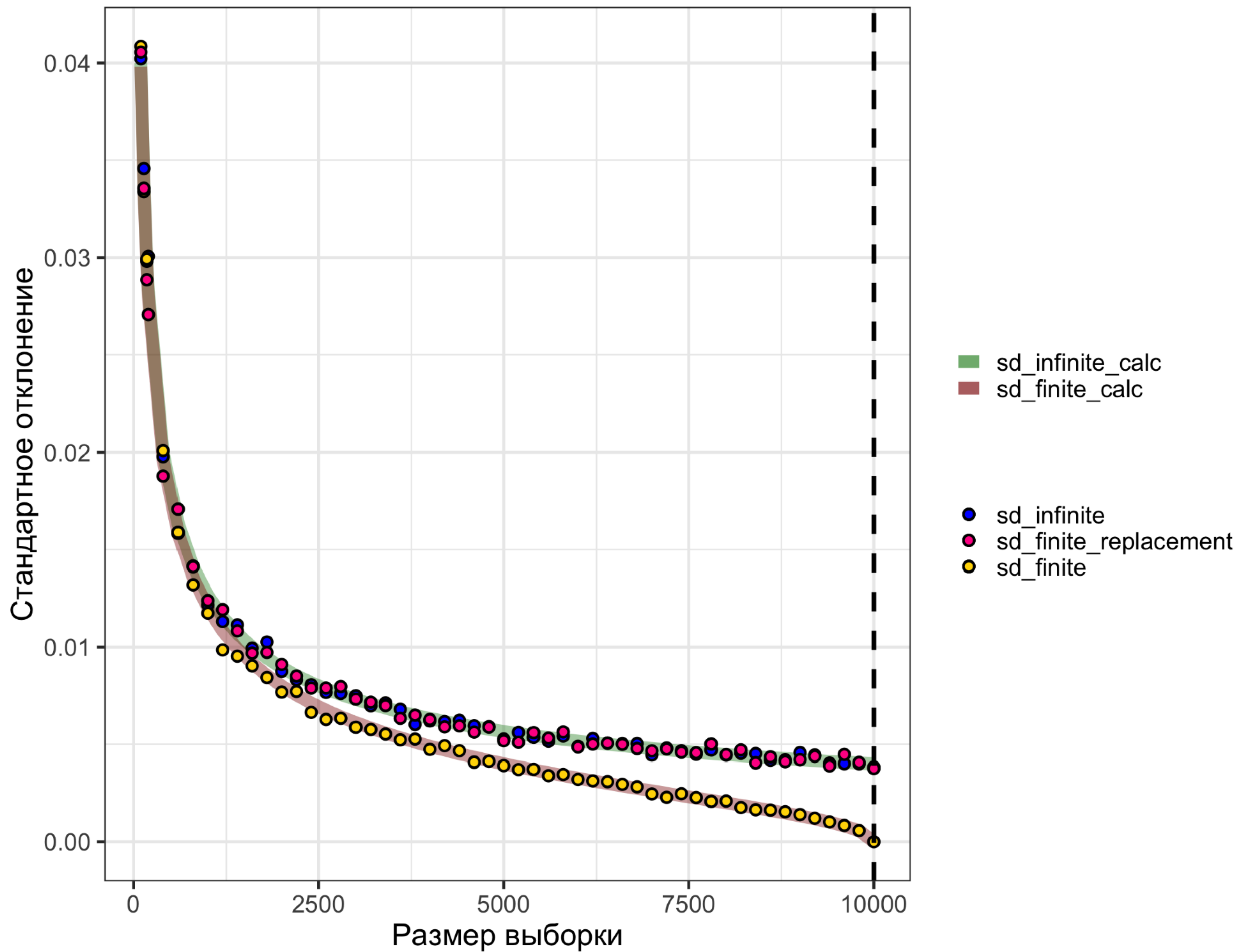
Тогда для доли:

$$E(\hat{p}) = p$$

$$D(\hat{p}) = \frac{pq}{n} \cdot \frac{N - n}{N - 1} \quad \Rightarrow \quad SE = \sqrt{\frac{N - n}{N - 1}} \sqrt{\frac{pq}{n}}$$

$$SE = \sqrt{\frac{N - n}{N - 1}} \cdot SE_{uncorrected}$$

Поправка на конечность выборки



Задача

Была проведен опрос населения города N (всего 3000 человек). Согласно опросу, доля курящих - 11%. Было опрошено 250 человек. Подсчитайте 95% доверительный интервал для доли курящих в городе N.

Оценка необходимого числа объектов

Часто еще до самого исследования требуется оценить число объектов, которое потребуется для того, чтобы получить результаты с заданной точностью.

Задача

Господин К. хочет оценить долю людей, смотрящих телеканал “Домашний” в мегаполисе N с точностью $\pm 5\%$ на уровне значимости $\alpha=10\%$. Какое минимальное число людей необходимо для этого опросить?

Задача

Господин К. хочет оценить средние траты людей на докторскую колбасу, смотрящих телеканал “Домашний” в городе N (население - 3000 тыс человек) с точностью $\pm 5\%$ на уровне значимости $\alpha=10\%$. Какое минимальное число людей необходимо для этого опросить?

Тест на равенство дисперсии

$$s^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

<https://medium.com/bluekiri/the-role-of-the-t-student-distribution-29259010d0fe>

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2$$

$$H_0 : \sigma^2 = a^2$$

$$H_1 : \sigma^2 \neq a^2 \quad H_1 : \sigma^2 < a^2 \quad H_1 : \sigma^2 > a^2$$

Тест на равенство дисперсии

$$s^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

<https://medium.com/bluekiri/the-role-of-the-t-student-distribution-29259010d0fe>

$$\frac{s^2(n-1)}{\sigma^2} \sim \chi_{n-1}^2$$

При условии выполнения H_0

$$H_0 : \sigma^2 = a^2$$

$$\frac{s^2 \cdot (n-1)}{a^2} = \frac{s^2 \cdot (n-1)}{\sigma^2} \sim \chi_{n-1}^2$$

Задача

Производитель шестеренок утверждает, что стандартное отклонение диаметров шестеренок менее 0.1 мм. Случайная выборка из 15 шестеренок имеет стандартное отклонение 0.12 мм. Достаточно ли оснований для того, чтобы не доверять заявлению производителя на уровне значимости 0.05?

Задача

Производитель шестеренок утверждает, что стандартное отклонение диаметров шестеренок менее 0.1 мм. Случайная выборка из 15 шестеренок имеет выборочное стандартное отклонение 0.12 мм. Постройте 98% доверительный интервал для стандартного отклонения шестеренок

Задача

Производитель шестеренок утверждает, что стандартное отклонение диаметров шестеренок менее 0.1 мм. Случайная выборка из 15 шестеренок имеет выборочное стандартное отклонение 0.12 мм. Постройте 98% доверительный интервал для стандартного отклонения шестеренок

$$\frac{s^2 \cdot (n - 1)}{\sigma^2} \sim \chi^2_{n-1}$$

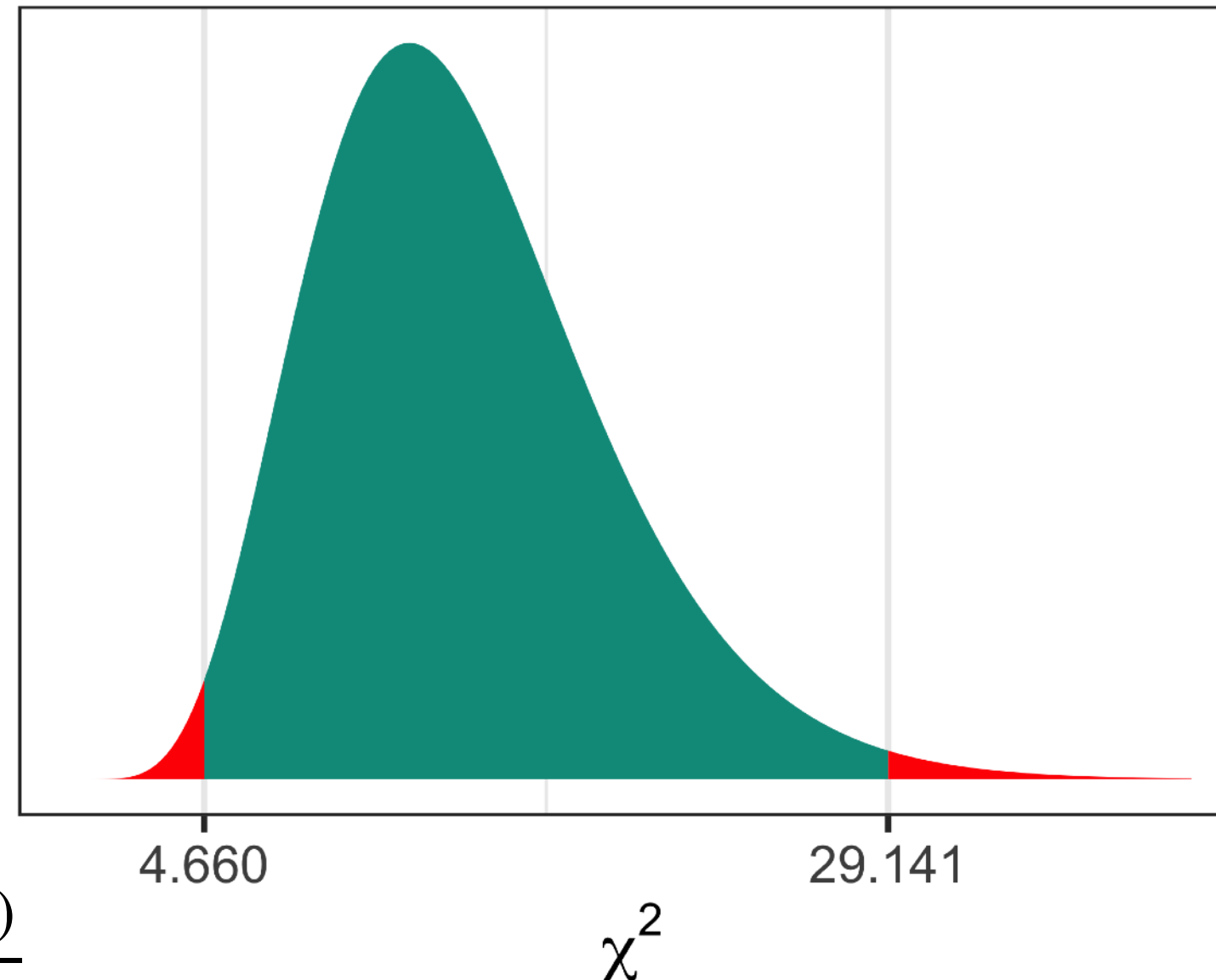
$$P(4.660 < \chi^2(14) < 29.141) = 0.98$$

$$4.660 < \frac{0.12^2 \cdot (15 - 1)}{\sigma^2} < 29.141$$

$$\frac{1}{29.141} < \frac{\sigma^2}{0.12^2 \cdot (15 - 1)} < \frac{1}{4.660}$$

$$\frac{0.12^2 \cdot (15 - 1)}{29.141} < \sigma^2 < \frac{0.12^2 \cdot (15 - 1)}{4.660}$$

$$0.08 < \sigma < 0.21$$



Тест Фишера (тест на равенство дисперсий)

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

Если

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

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

Тест Фишера (тест на равенство дисперсий)

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \Rightarrow \quad F(n_x - 1, n_y - 1)$$

$$H_1 : \sigma_x^2 \neq \sigma_y^2 \quad H_1 : \sigma_x^2 < \sigma_y^2 \quad H_1 : \sigma_x^2 > \sigma_y^2$$

Задача

В ходе в опыта с 10 марокканскими обезьянами в лаборатории 1, выборочное отклонение времени их реакции на раздражитель отклонение равно 1.7с.

В лаборатории 2 в ходе опыта с 7 марокканскими обезьянами выборочное отклонение равно 2.7. Есть ли основания предполагать равенство дисперсий на уровне значимости 0.01?

Проблема теста Фишера

Тест Фишера обычно иногда применяют, чтобы проверить разумность предположение о равенстве дисперсий. Чтобы затем применить t-test, предполагающий равенство дисперсий. Насколько это правильно?

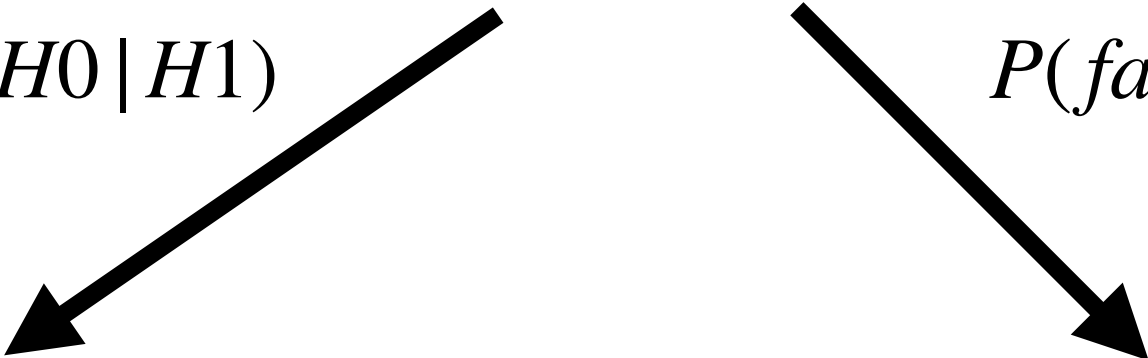
Часть людей считает, что ОК.

Проблема теста Фишера

На самом деле пусть в нашей выборке неравные дисперсии

F-test

$P(\text{reject_}H_0 | H_1)$



Тест Вэлча

Все работает правильно,
 $\alpha = 0.05$

$P(\text{fail_to_reject_}H_0 | H_1)$

**Тест Стьюдента, предполагающий
равенство дисперсий**

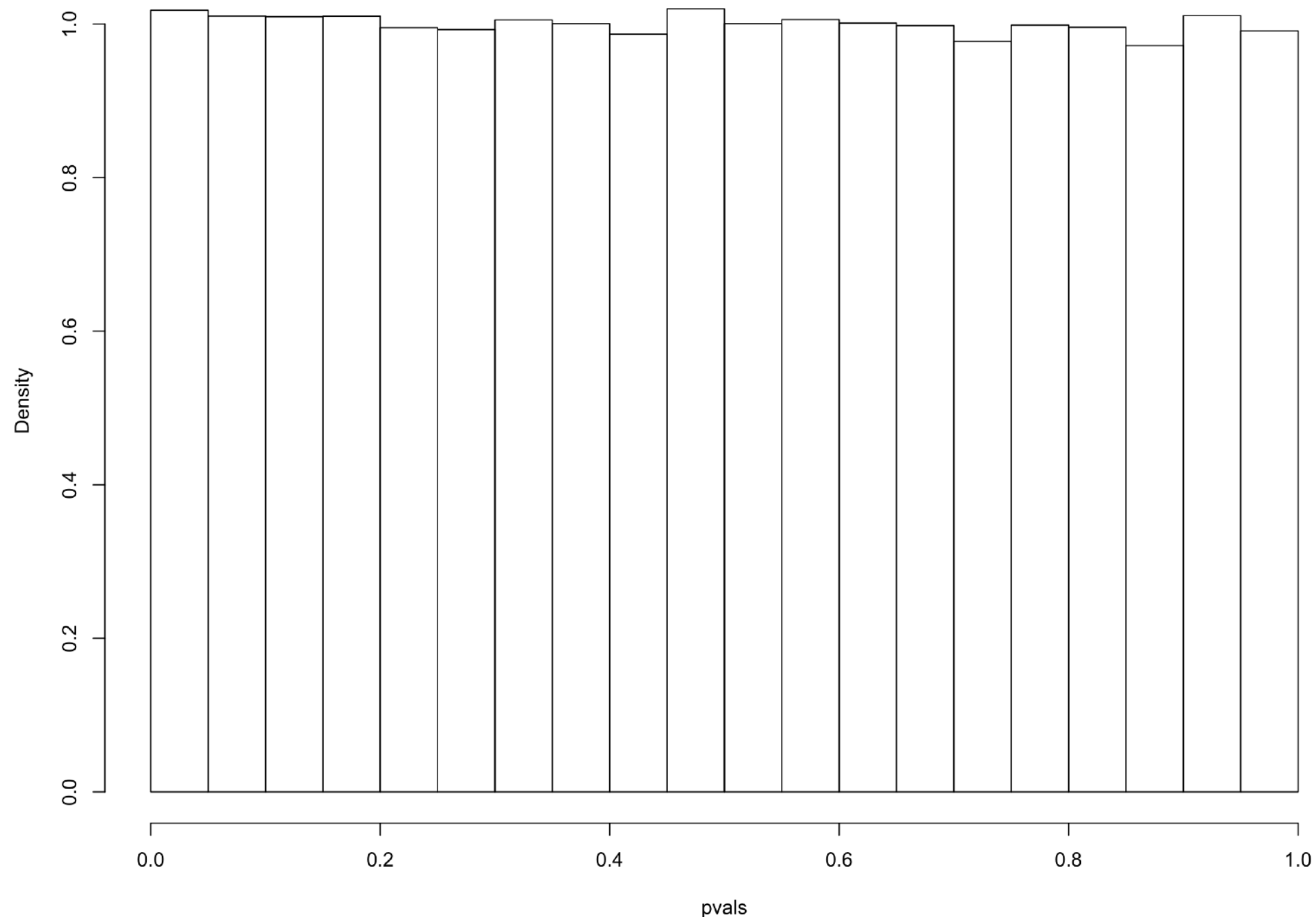
Все работает не знаем как,
не знаем α . Может как его понизить,
так и повысить.

Зависит от того: больше ли
выборка с большей дисперсией
Или наоборот, выборка с большей дисперсией
меньше

Проблема теста Фишера

Как должно выглядеть распределение p-value при условии верности H_0 ?

Равномерно на отрезке [0,1]



Проблема теста Фишера

Как должно выглядеть распределение p-value для непрерывного распределения при условии верности H_0 ?

Проблема теста Фишера

Как должно выглядеть распределение p-value для непрерывного распределения при условии верности H_0 ?

В случае левосторонней альтернативы p-value - это просто функция распределения. Функция распределения функции распределения - равномерно распределенная величина

$$F(F(x)) = U[0,1]$$

Поэтому **равномерно** на отрезке $[0,1]$

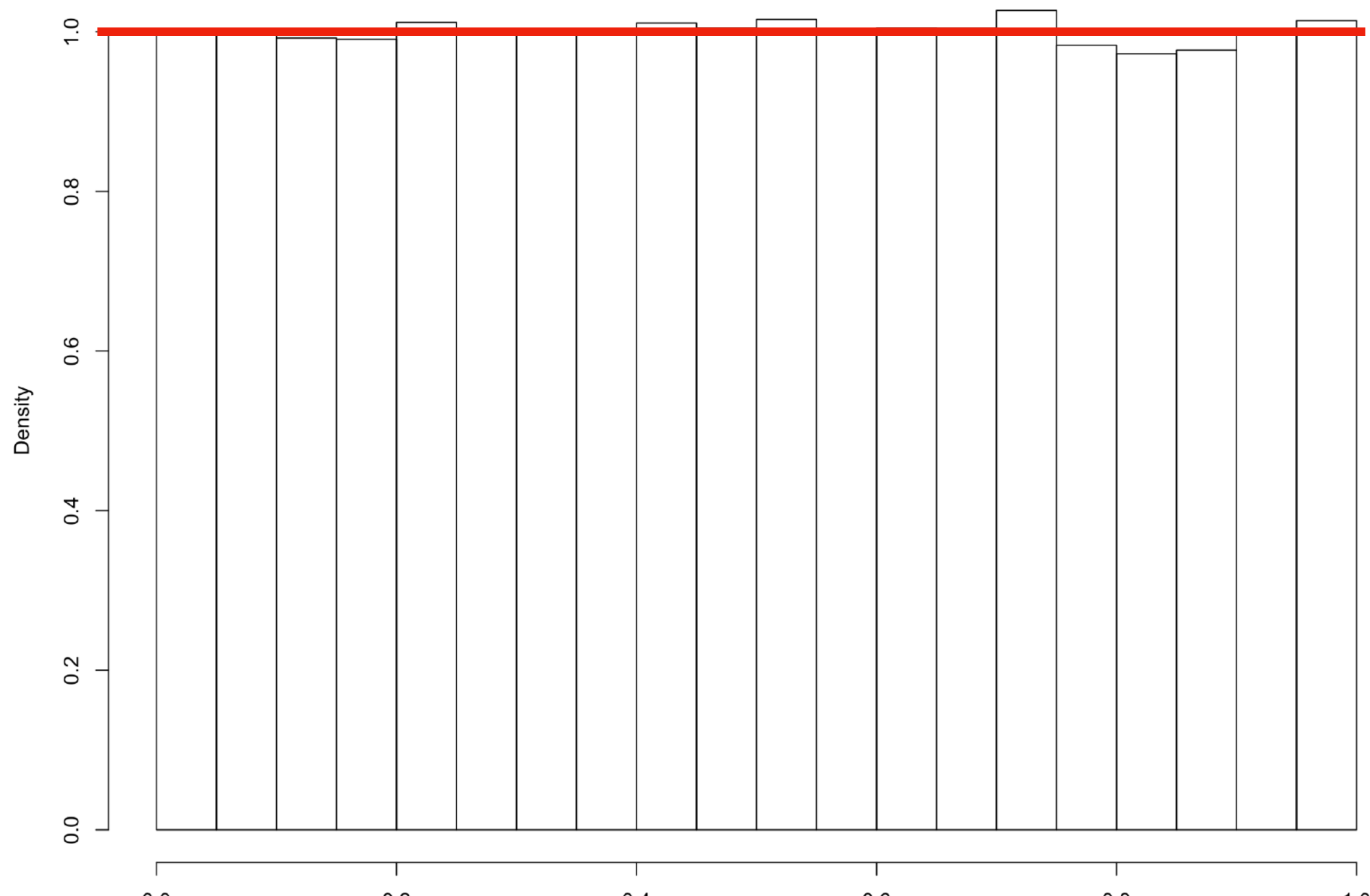
Аналогично можно вывести для других альтернатив.

<https://www.mathworks.com/matlabcentral/answers/413172-why-p-values-are-uniformly-distributed-when-the-null-hypothesis-is-true>

<https://stats.stackexchange.com/questions/10613/why-are-p-values-uniformly-distributed-under-the-null-hypothesis>

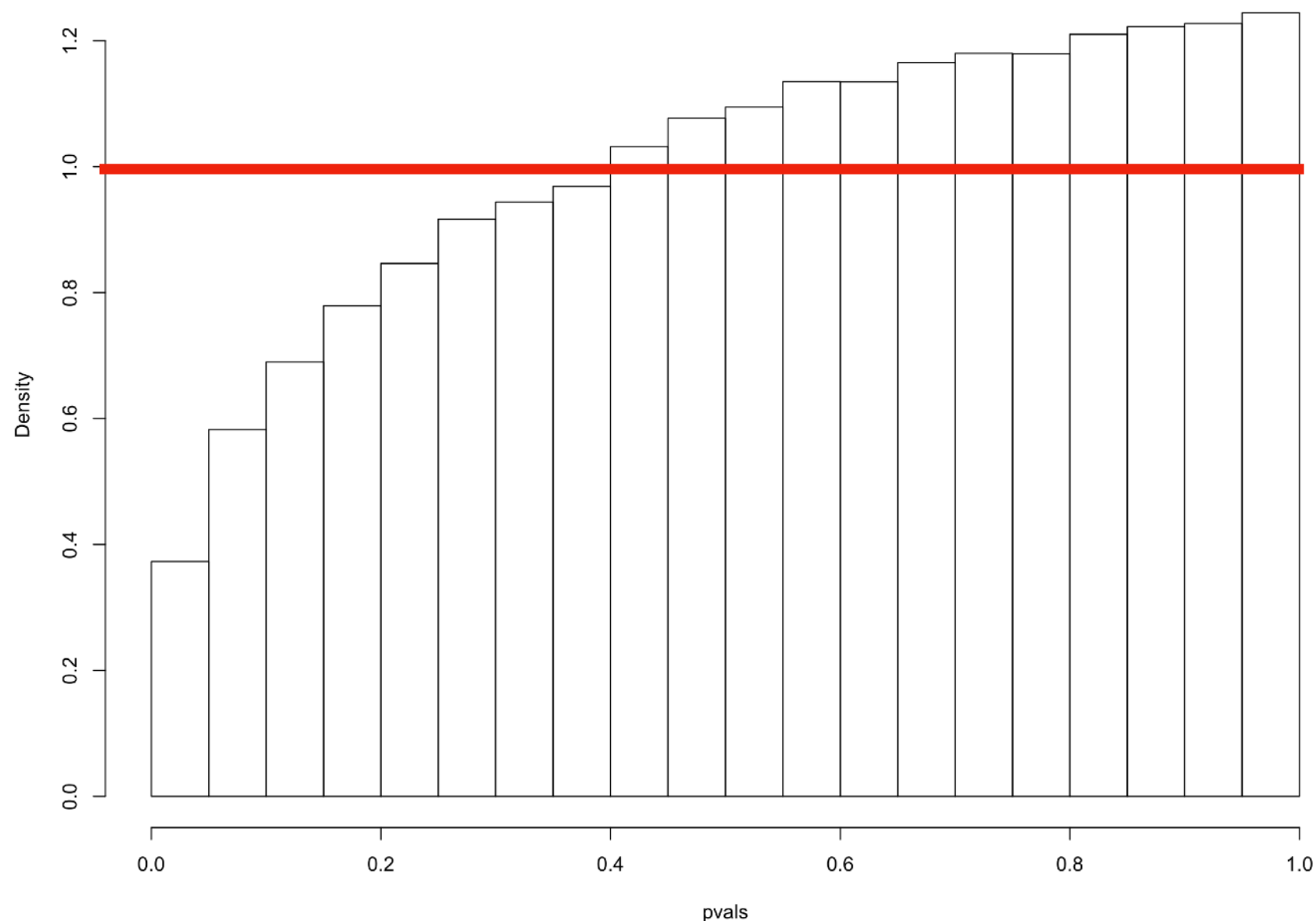
Проблема теста Фишера

Тест Стьюдента, предполагающий равенство дисперсий. Все выполняется,
все хорошо



Проблема теста Фишера

Тест Стьюдента, предполагающий равенство дисперсий. Выборка, соответствующая генеральной совокупности с большей дисперсией, **больше**.

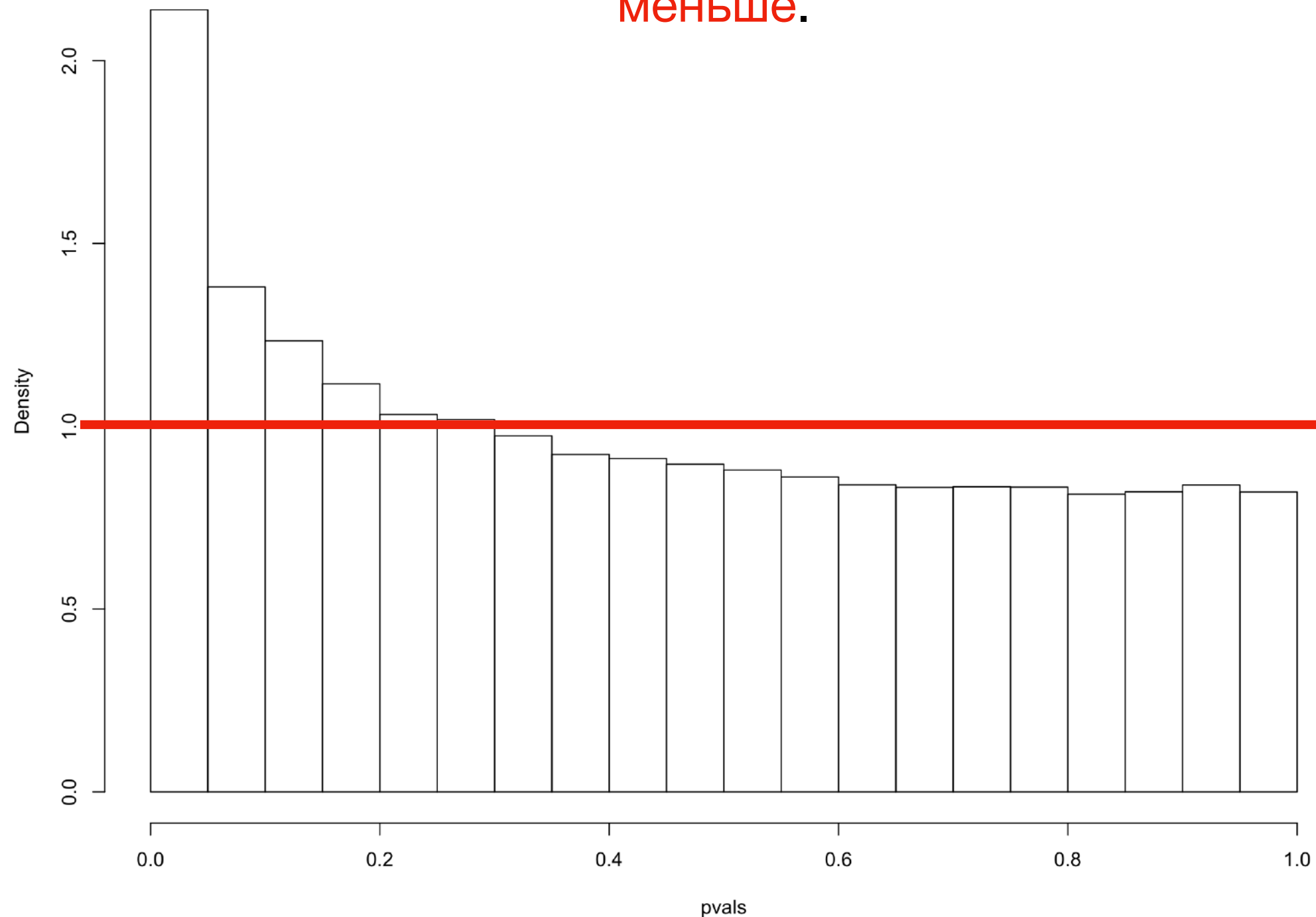


Чаще ошибочно не отвергаем H_0

Проблема теста Фишера

Тест Стьюдента, предполагающий равенство дисперсий. Выборка, соответствующая генеральной совокупности с большей дисперсией,

меньше.



Чаще ошибочно отвергаем H_0

Проблема теста Фишера

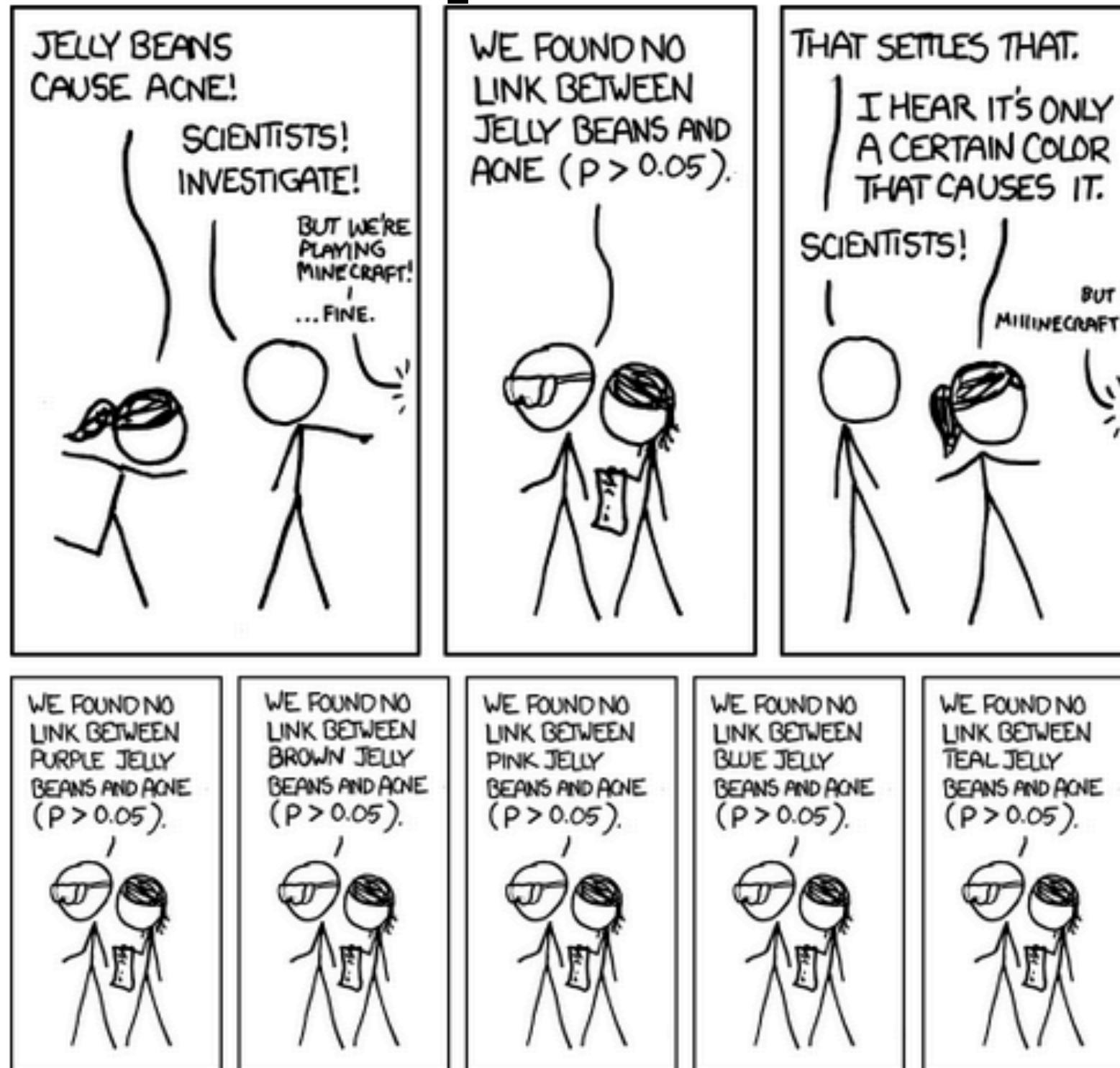
На дом:

Тест Стьюдента, предполагающий равенство дисперсий. Выборки имеют разные дисперсии, но одинаковый размер. Какое распределение p-value?

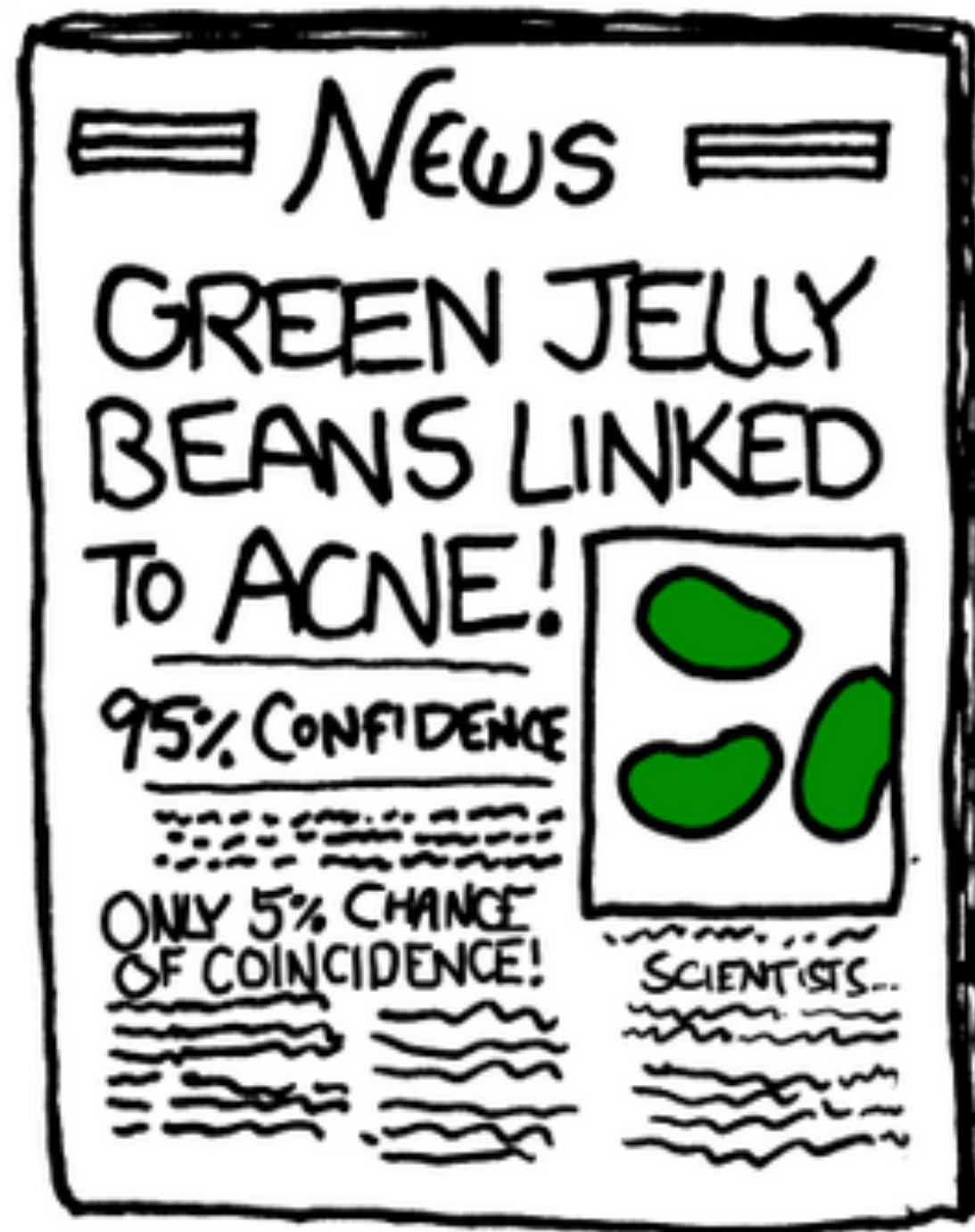
Множественное тестирование

**Не ищите того, чего нет. А то в конечном итоге -
найдете. А это будет ошибка первого рода**
@Игорь

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 $p < 0.05$.

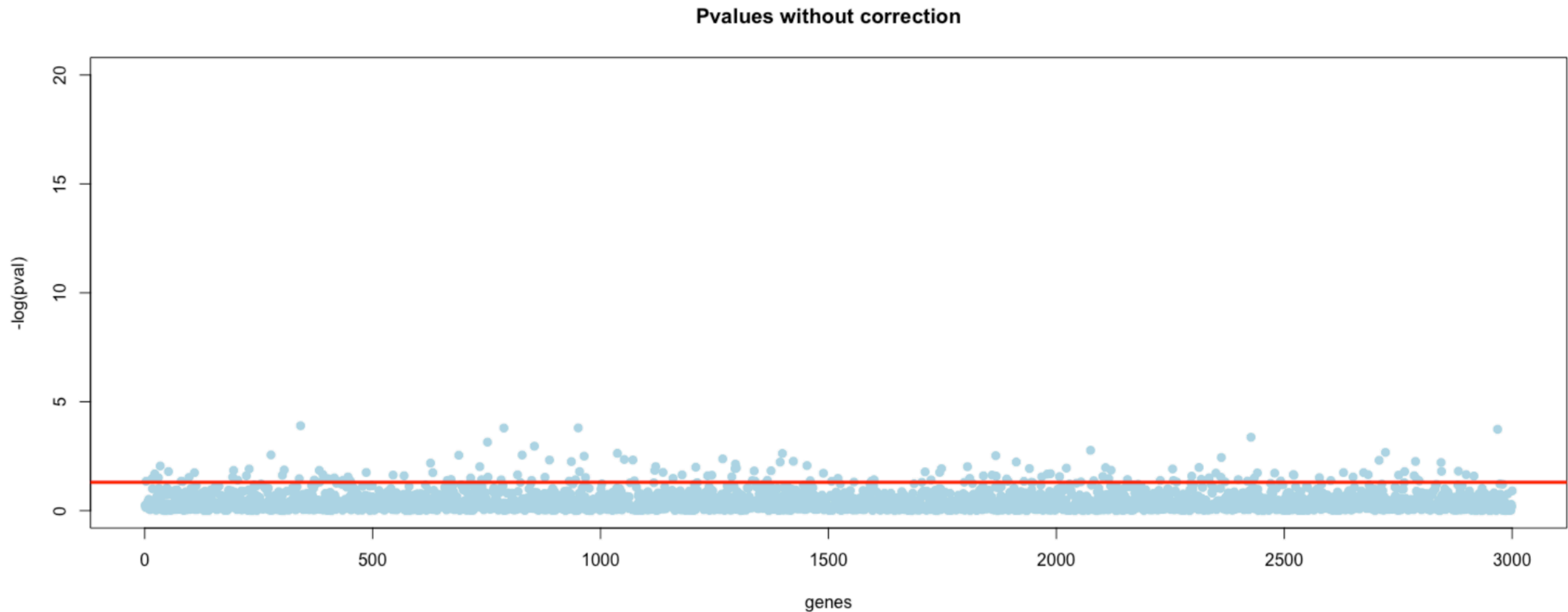
Какова вероятность, что ни один ген не будет помечен как дифференциально экспрессируемый?

Сколько в среднем генов будет помечено как дифференциально экспрессируемые?

Эксперимент 1

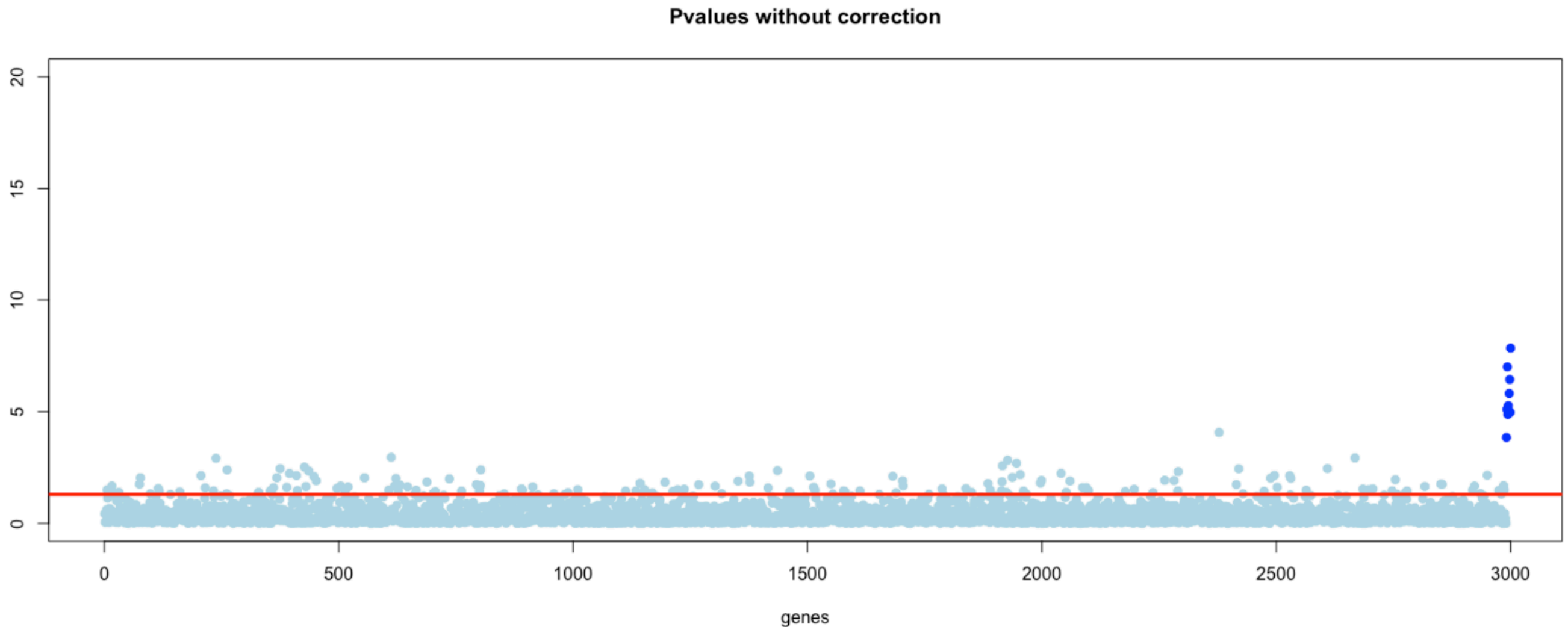
Возьмем и симулируем набор из 50 пациентов с 3000 генов, которые не меняют свою экспрессию значимо по ходу эксперимента (имеем результаты до и после).

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 α **разный** для двух подходов