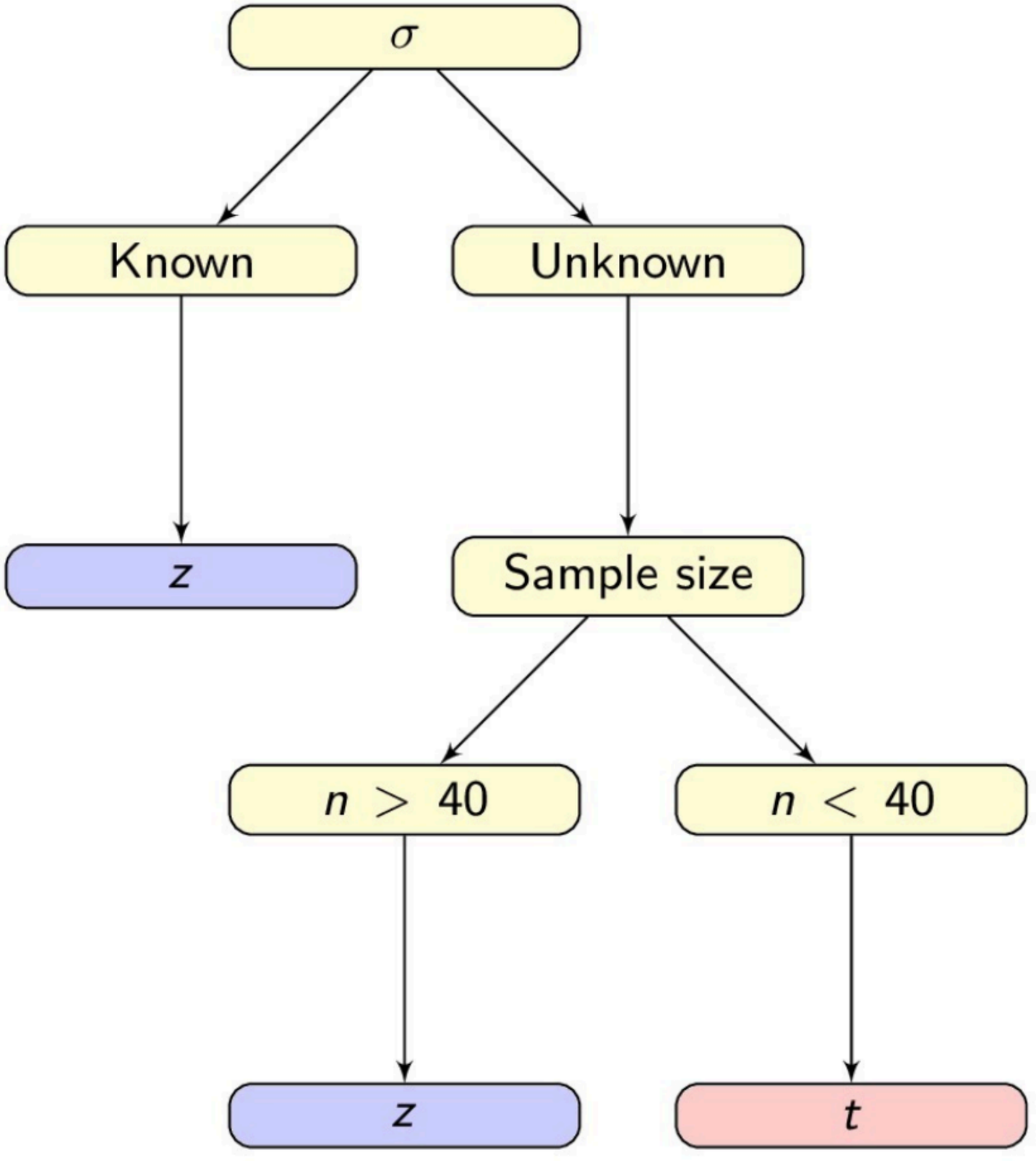

σ
Known
Unknown
z
Sample size
$n > 40$
$n < 40$
z
t

z-test

$$\frac{\overline{X} - (\mu_x)}{\frac{\sigma_x}{\sqrt{n_x}}}$$

$$\frac{\overline{X - Y} - (\mu_{x-y})}{\frac{\sigma_{x-y}}{\sqrt{n}}}$$

t-test

$$\frac{\overline{X} - (\mu_x)}{\frac{s_x}{\sqrt{n_x}}}$$

$$\frac{\overline{X - Y} - (\mu_{x-y})}{\frac{s_{x-y}}{\sqrt{n}}}$$

z-test

t-test

Одновыборочный

$$\frac{\overline{X} - (\mu_x)}{\frac{\sigma_x}{\sqrt{n_x}}}$$

$$\frac{\overline{X} - (\mu_x)}{\frac{s_x}{\sqrt{n_x}}}$$

Парный

$$\frac{\overline{X - Y} - (\mu_{x-y})}{\frac{\sigma_{x-y}}{\sqrt{n}}}$$

$$\frac{\overline{X - Y} - (\mu_{x-y})}{\frac{s_{x-y}}{\sqrt{n}}}$$

Двувыборочные

z-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$$\sigma_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sigma_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

t-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

$$s_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{s_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Двувыборочные

z-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Нет равенства дисперсий

t-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

$n_x + n_y - 2 > 40$

Есть равенство дисперсий

$n_x + n_y - 2 < 40$

$$\sigma_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

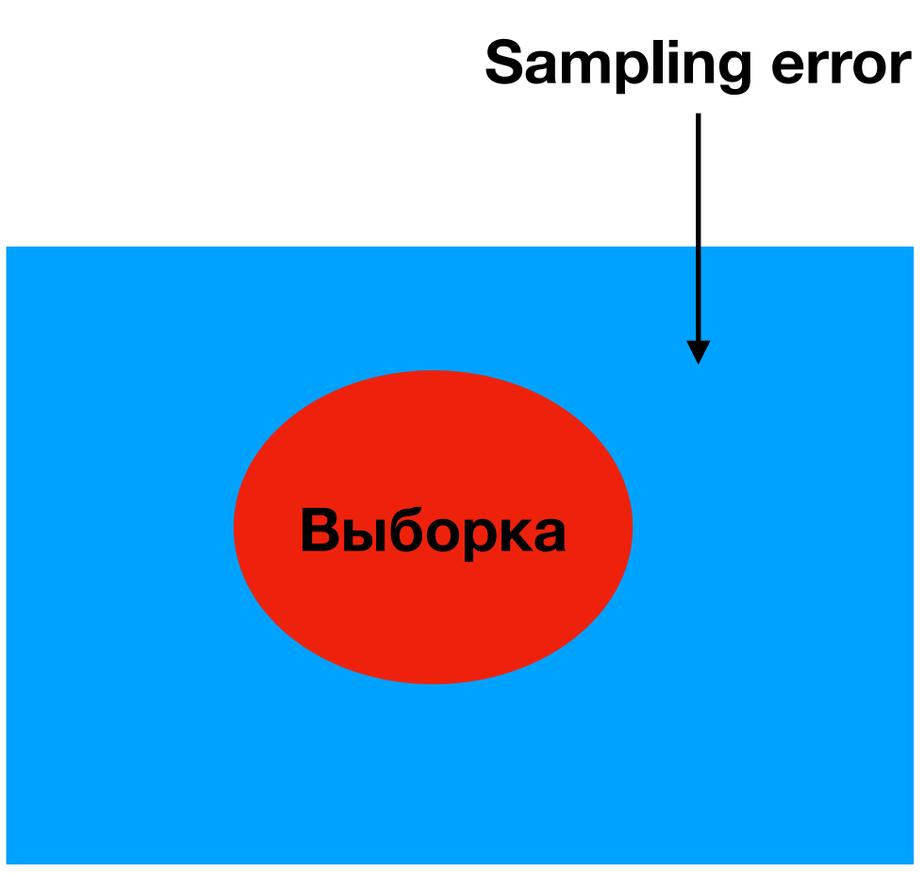
$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sigma_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

$$s_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{s_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

SE - sampling error

Ошибка в оценке параметра генеральной совокупности по выборке, а не по всей генеральной совокупности



z-test

t-test

Одновыборочный

$$\frac{\overline{X} - (\mu_x)}{SE = \frac{\sigma_x}{\sqrt{n_x}}}$$

$$\frac{\overline{X} - (\mu_x)}{SE = \frac{s_x}{\sqrt{n_x}}}$$

Парный

$$\frac{\overline{X - Y} - (\mu_{x-y})}{SE = \frac{\sigma_{x-y}}{\sqrt{n}}}$$

$$\frac{\overline{X - Y} - (\mu_{x-y})}{SE = \frac{s_{x-y}}{\sqrt{n}}}$$

Двувыборочные

z-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{SE = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

Нет равенства дисперсий

t-test

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{SE = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

Есть равенство дисперсий

$$\sigma_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{SE = \sigma_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

$$s_{xy}^2 = \frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}$$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{SE = s_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Какова вероятность этой последовательности быть написанной на этом слайде?

ATGCAGAGGAGGGGCAGCAAGG

На утреннике давали коробки с ланчами. В коробку клали два фрукта. Каждый фрукт - нектарин или яблоко. Фрукты клали независимо, число яблок равно числу нектаринов.

Какая вероятность, что мне достанется ланч с двумя нектаринами?

На утреннике давали коробки с ланчами. В коробку клали два фрукта. Каждый фрукт - нектарин или яблоко. Фрукты клали независимо, число яблок равно числу нектаринов.

Какая вероятность, что мне достанется ланч с двумя нектаринами?

1/4

На утреннике давали коробки с ланчами. В коробку клали два фрукта. Каждый фрукт - нектарин или яблоко. Фрукты клали независимо, число яблок равно числу нектаринов.

Какая вероятность, что мне достанется ланч с двумя нектаринами?

1/4

Я взял ланч. Какова вероятность того, что в коробке, что я взял - два нектарина?

На утреннике давали коробки с ланчами. В коробку клали два фрукта. Каждый фрукт - нектарин или яблоко. Фрукты клали независимо, число яблок равно числу нектаринов.

Какая вероятность, что мне достанется ланч с двумя нектаринами?

1/4

Я взял ланч. Какова вероятность того, что в коробке, что я взял - два нектарина?

Они либо там лежат, либо нет..

Точечная оценка vs Доверительный интервал

Точечная оценка (например, выборочное среднее) не содержит информации о том, насколько мы в ней уверены. Условно говоря, если мы сказали, что среднее равно 3.5, то насколько мы уверены, что среднее не равно 3.4.

Чтобы решить этот вопрос используют **доверительный интервал**

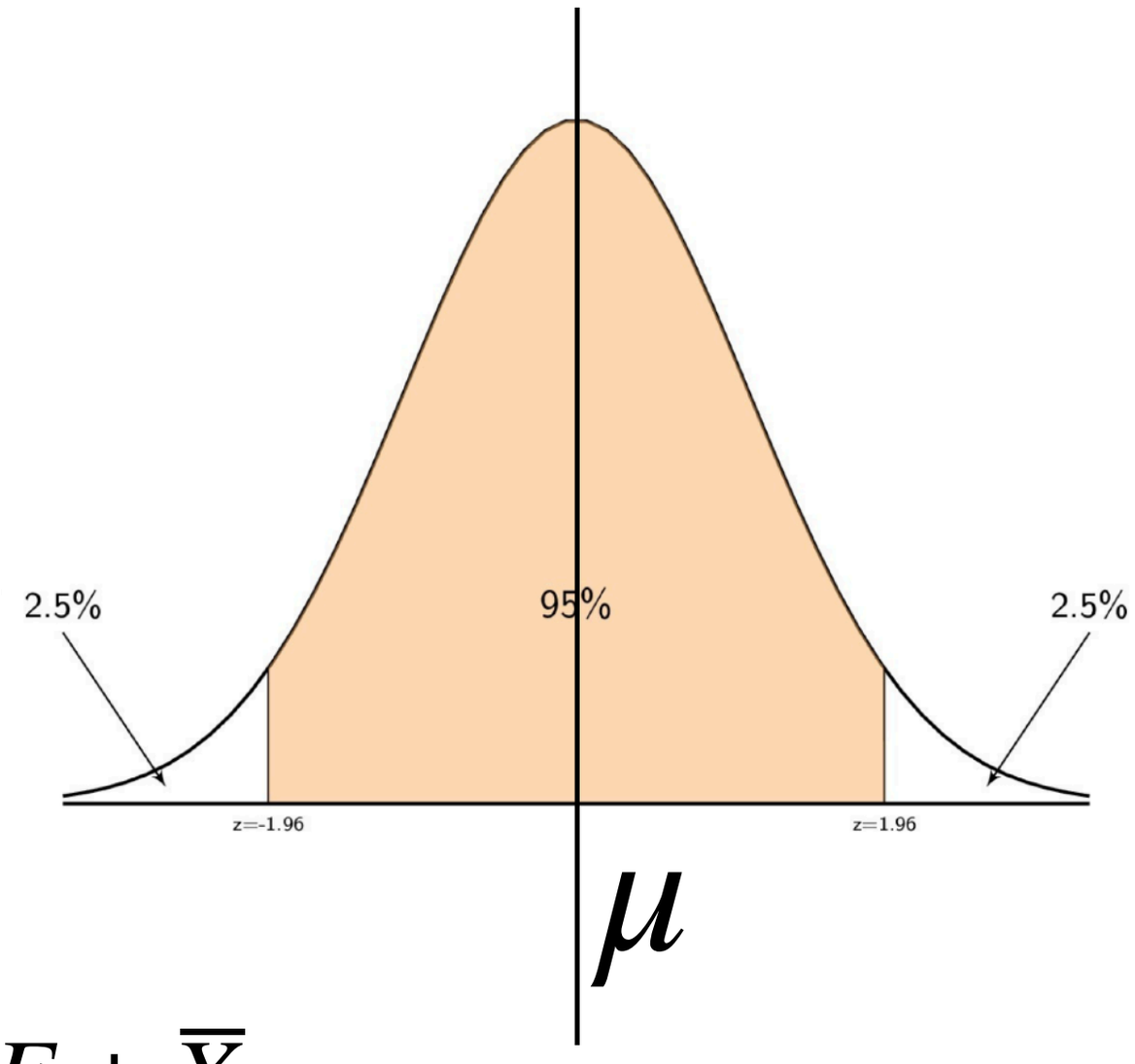
Доверительный интервал для среднего

$$\overline{X} \sim N(\mu, SE)$$ **(если выполняются условия z-test)**

Исходя из такого распределения среднего,

$$\frac{\overline{X} - \mu}{SE} \sim N(0,1)$$

$$z\_score = \frac{\overline{X} - \mu}{SE}$$



Если много раз считать выборочное среднее, то в 95% случаев z-score будет находиться в пределах от -1.96 до +1.96

$$-1.96 \leq \frac{\overline{X} - \mu}{SE} \leq 1.96$$

$$-1.96 \cdot SE + \overline{X} \leq \mu \leq 1.96 \cdot SE + \overline{X}$$

95% доверительный интервал для среднего

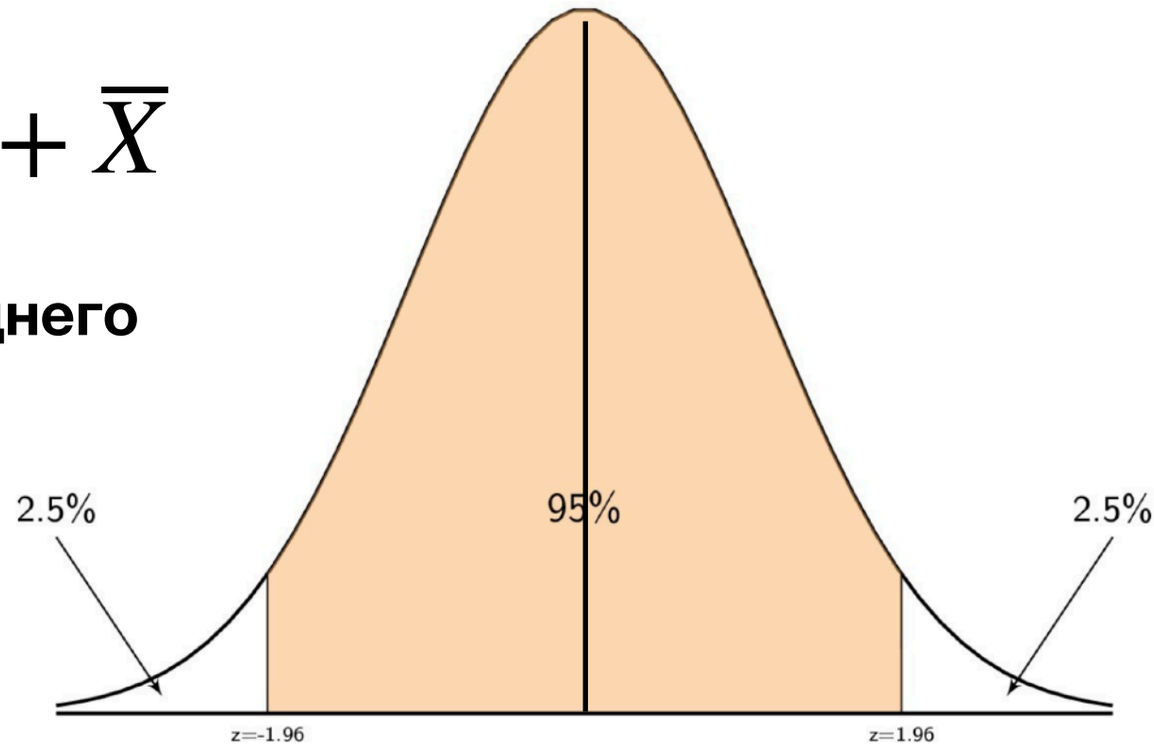
Доверительный интервал для среднего

$$\overline{X} \sim N(\mu, SE)$$

(если выполняются условия z-test)

$$-1.96 \cdot SE + \overline{X} \leq \mu \leq 1.96 \cdot SE + \overline{X}$$

95% доверительный интервал для среднего



В общем случае для (100-alpha) доверительного интервала

$$\overline{X} - |z_{\alpha/2}| \cdot SE \leq \mu \leq \overline{X} + |z_{\alpha/2}| \cdot SE$$

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

Задача

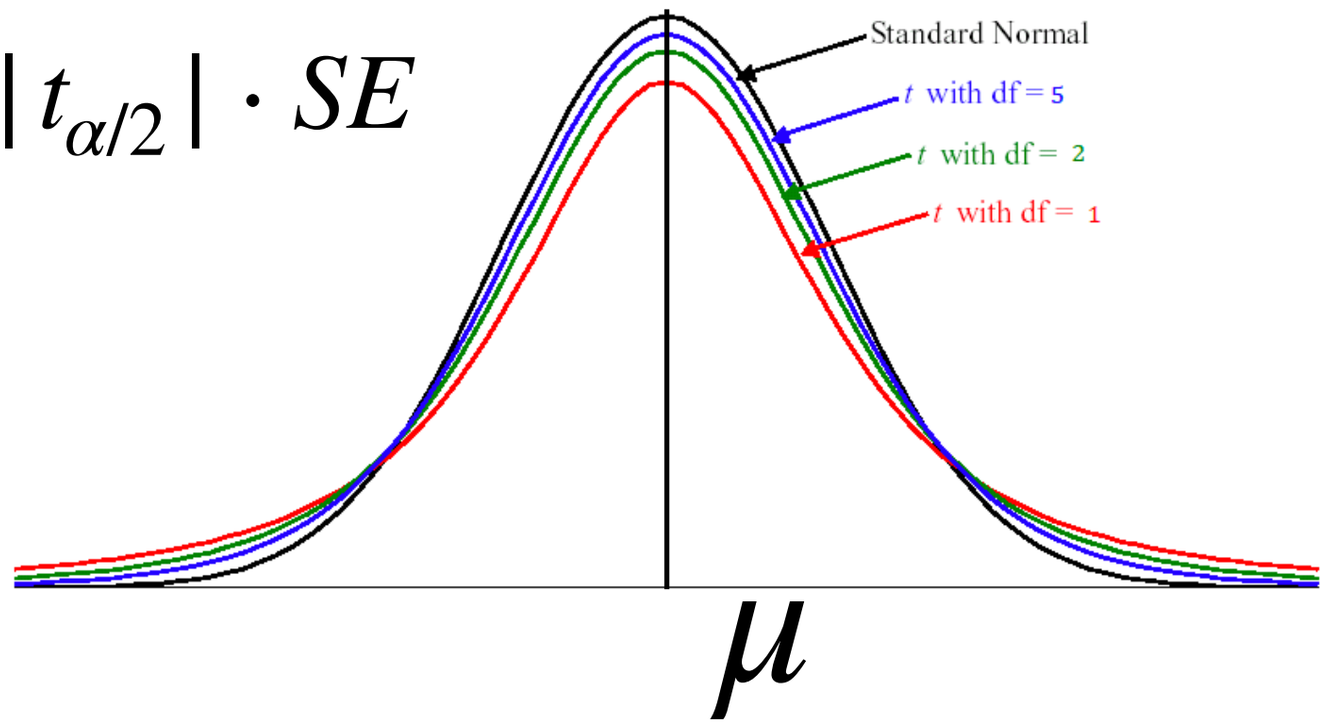
Ученый изучает то, на какое расстояние способны кинуть обезьяны банан под воздействием мельдония. Он взял группу из 47 обезьян и замерил то, на какое расстояние они бросают до курса мельдония. После этого все обезьяны прошли курс мельдония. После этого было опять измерена длина их броска. Средняя длина броска до курса мелькания - 11.5 м. После курса мельдония - 13.2. Выборочная дисперсия разниц длин бросков одних и тех же обезьян до и после курса мелькания - 1.7м. Постройте 98% доверительный интервал для разницы средних.

Доверительный интервал для среднего

А если выполняются условия только t-test ?

$$\overline{X} - |t_{\alpha/2}| \cdot SE \le \mu \le \overline{X} + |t_{\alpha/2}| \cdot SE$$

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



Более тяжелые хвосты!

Задача

Ученый изучает то, на какое расстояние способны кинуть обезьяны банан под воздействием мельдония. Он взял группу из 30 обезьян и замерил то, на какое расстояние они бросают до курса мельдония. После этого все обезьяны прошли курс мельдония. После этого было опять измерена длина их броска. Средняя длина броска до курса мелькания - 11.5 м. После курса мельдония - 13.2. Выборочная дисперсия разниц длин бросков одних и тех же обезьян до и после курса мелькания - 1.7м. Постройте 98% доверительный интервал для разницы средних.

Доверительный интервал для разницы средних

Если выполняются условия z-test

$$\mu \in \overline{X} - \overline{Y} \pm |z_{\alpha/2}| \cdot SE$$

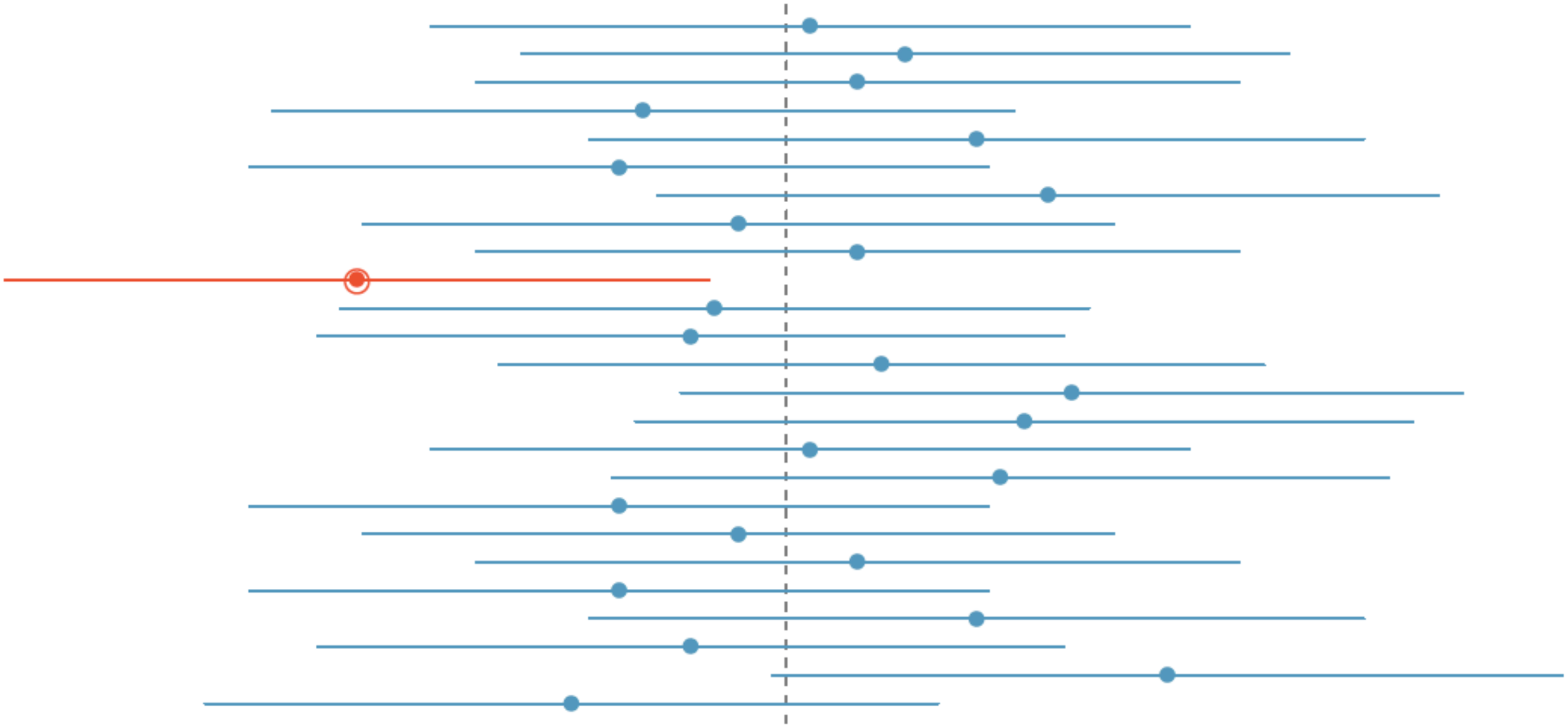
Если выполняются условия только t-test

$$\mu \in \overline{X} - \overline{Y} \pm |t_{\alpha/2}| \cdot SE$$

Задача

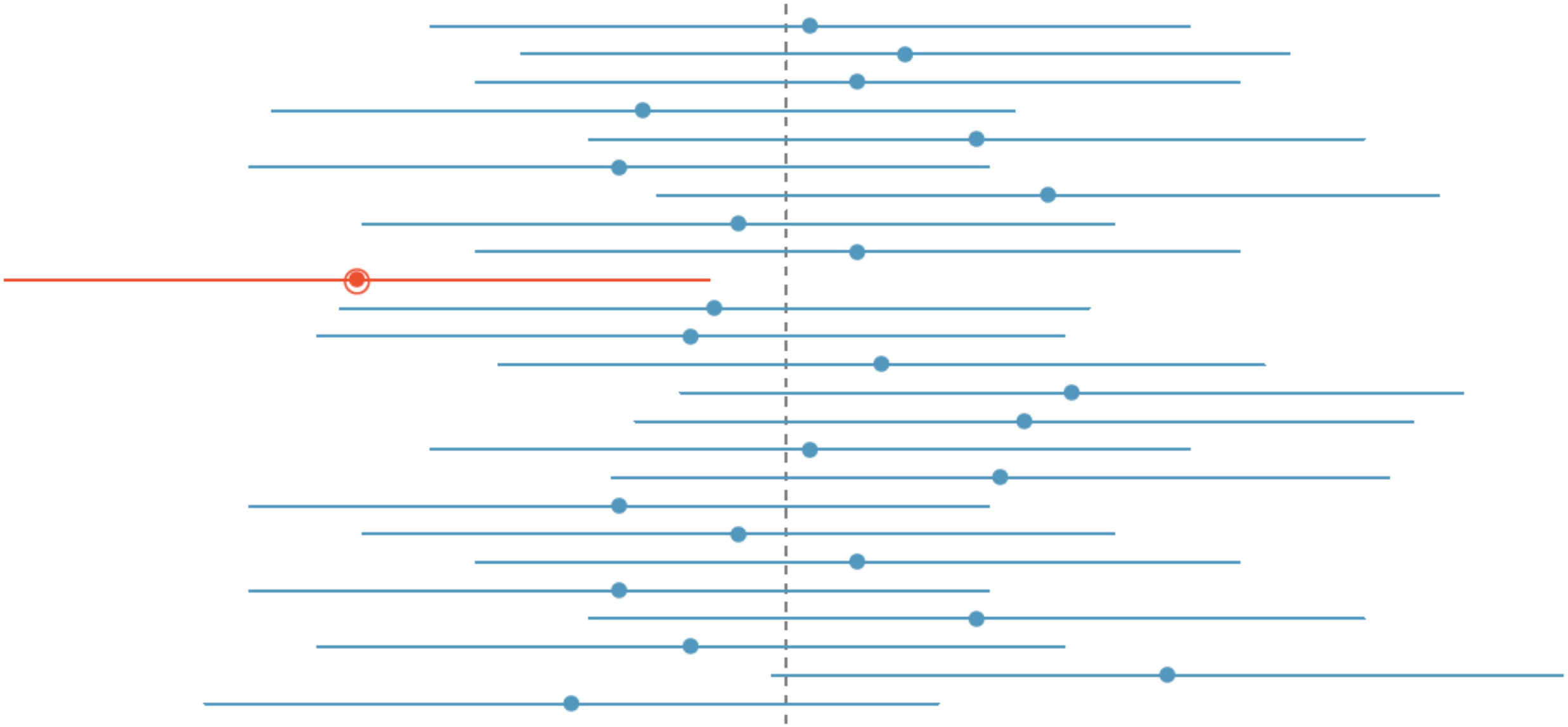
Молодой немецкий фермер Ганс хочет проверить два новых фосфорорганических удобрения под торговыми марками “ГенФос” и “Кузнечик”. В результате испытаний на двух группах по 9 и 7 растений соответственно он получил, что среднее количество плодов с растений первой группы - 10.4, а со второй - 12. Выборочные 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, постройте 99% доверительный интервал для разницы среднего числа плодов.

Доверительный интервал



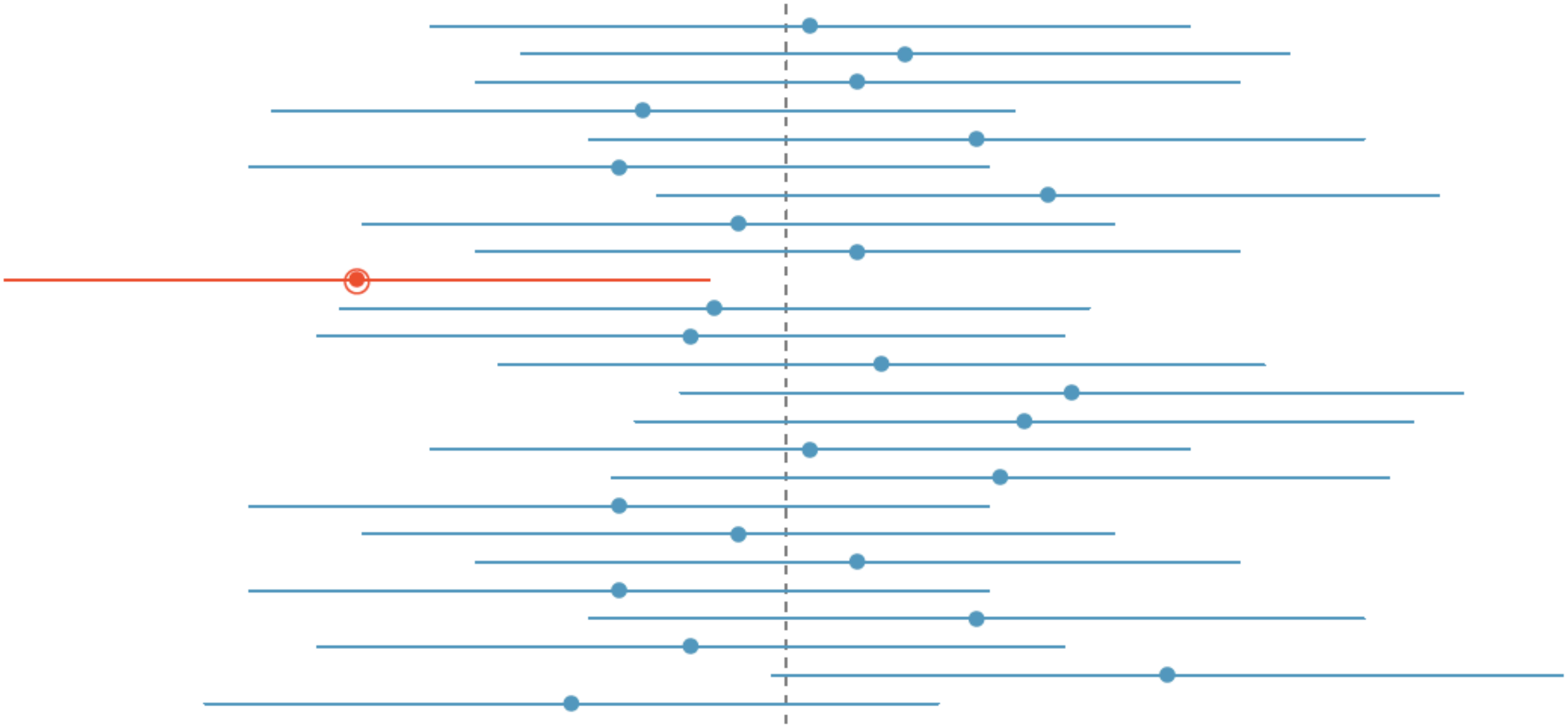
95%-доверительный интервал - 95% построенных по такой же процедуре интервалов будут содержать истинное среднее

Доверительный интервал



Говорить, что построенный по данной выборке 95%-доверительный интервал содержит истинное среднее с вероятностью 95% неверно. Выборка нам уже дана. Доверительный интервал либо содержит истинное среднее, либо нет.

Доверительный интервал



Для каждой выборки размера N границы 95% доверительного интервала могут получаться разными

Связь доверительного интервала и принятия/непринятия гипотезы

?

Связь доверительного интервала и принятия/непринятия гипотезы

$$H_0 : \mu = a$$
$$H_1 : \mu \neq a$$

равносильно

$a \in \overline{X} \pm z_{\alpha/2} \cdot SE$ **=> не отвергаем H0**

$a \notin \overline{X} \pm z_{\alpha/2} \cdot SE$ **=> отвергаем H0**

Связь доверительного интервала и принятия/непринятия гипотезы

$$H_0 : \mu_x - \mu_y = a$$
$$H_1 : \mu_x - \mu_y \neq a$$

равносильно

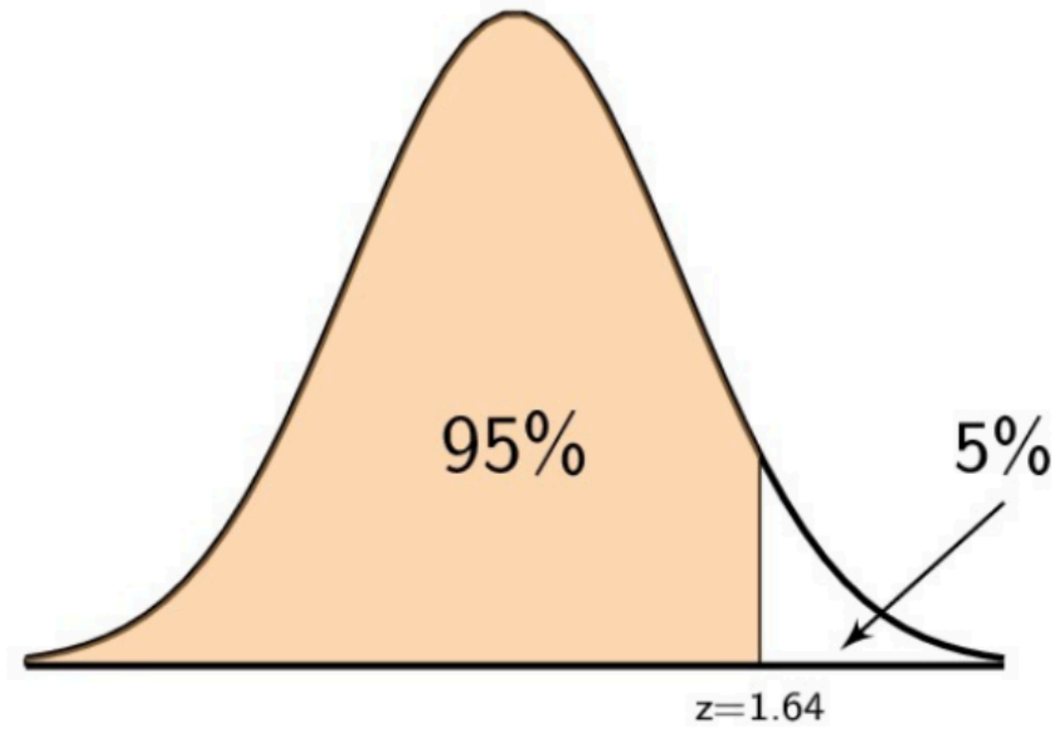
$a \in \overline{X} - \overline{Y} \pm z_{\alpha/2} \cdot SE$ **=> не отвергаем H0**

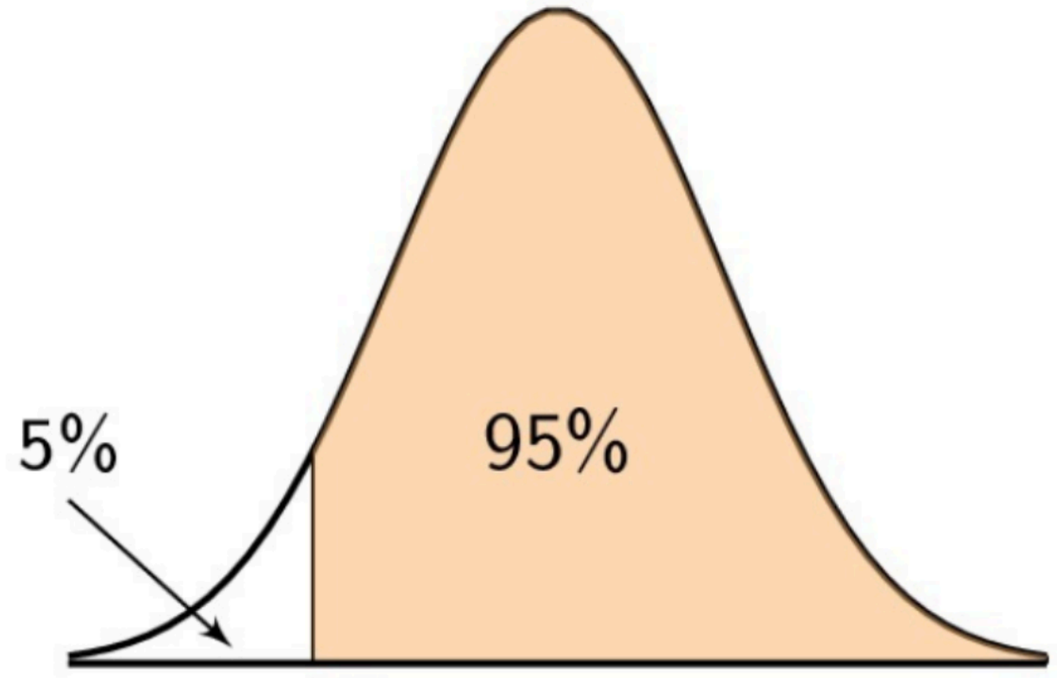
$a \notin \overline{X} - \overline{Y} \pm z_{\alpha/2} \cdot SE$ **=> отвергаем H0**

Задача

Программист Петя решил изучить Haskell. Его товарищ по работе Иван сказал, что тот стал материться с иной частотой, 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 Haskel

Односторонний доверительный интервал





Смотрим только с одной стороны

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

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

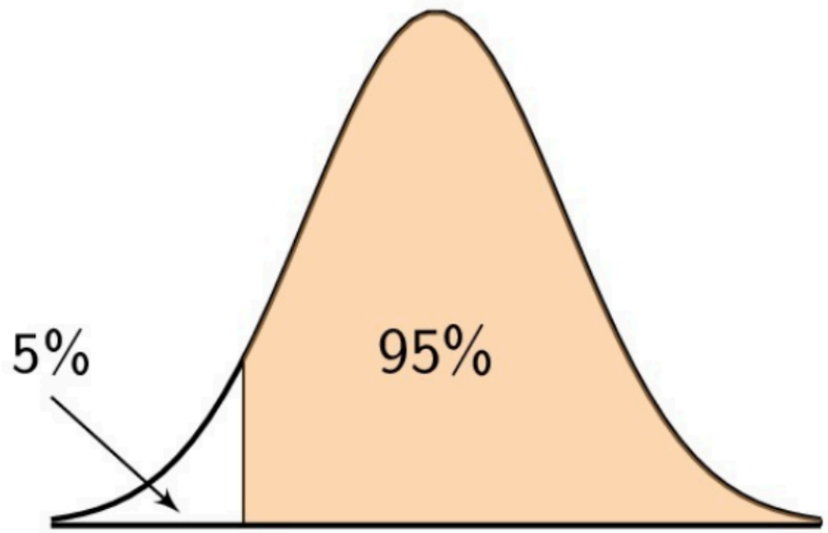
Для t-test аналогично

Связь доверительного интервала и принятия/непринятия гипотезы

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

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

равносильно



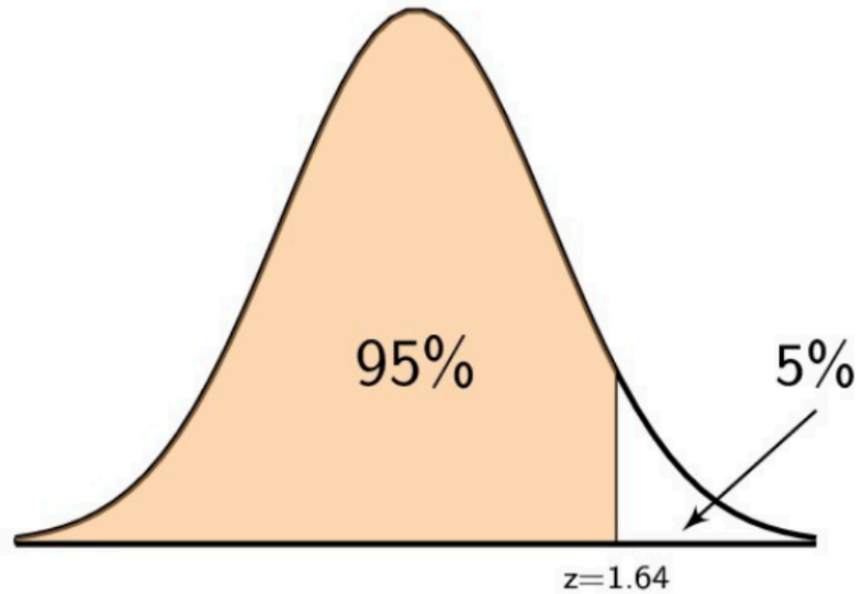
$$a \in (-\infty, \overline{X} - \overline{Y} + z_\alpha \cdot SE]$$ **H0 не отвергаем**

$$a \notin (-\infty, \overline{X} - \overline{Y} + z_\alpha \cdot SE]$$ **H0 отвергаем**

Связь доверительного интервала и принятия/непринятия гипотезы

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

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



равносильно

$a \in [\overline{X} - \overline{Y} - z_\alpha \cdot SE, +\infty)$ **H0 не отвергаем**

$a \notin [\overline{X} - \overline{Y} - z_\alpha \cdot SE, +\infty)$ **H0 отвергаем**

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 **односторонний** 95% доверительный интервал для разницы между числом произносимых матных слов Петей до и после начала изучения Haskel

Тест пропорций

Хотим оценить долю носителей p какого-то признака в популяции.
Или хотим сравнить доли носителей признака в разных популяциях
Предполагам, что признак бинарный - либо есть или нет.
Набираем выборку объектов, смотрим, у скольких наличествует признак

Как распределено число объектов с признаком?

$$X \sim C_n^k p^k (1-p)^{n-k}$$

$$E(X) = np$$

$$D(X) = npq$$

Тест пропорций

Хотим оценить долю носителей p какого-то признака в популяции.
Или хотим сравнить доли носителей признака в разных популяциях
Предполагам, что признак бинарный - либо есть или нет.
Набираем выборку объектов, смотрим, у скольких наличествует признак

Как распределено число объектов с признаком?

$$X \sim C_n^k p^k (1-p)^{n-k}$$

$$E(X) = np$$

$$D(X) = npq$$

Объектов берем много, тогда..

Тест пропорций

Хотим оценить долю носителей p какого-то признака в популяции.
Или хотим сравнить доли носителей признака в разных популяциях
Предполагам, что признак бинарный - либо есть или нет.
Набираем выборку объектов, смотрим, у скольких наличествует признак

Как распределено число объектов с признаком?

$$X \sim N(np, \sqrt{npq})$$

Можно делать поправку на непрерывность. В этом курсе не делаем

Тест пропорций

Одновыборочный

$$H_0 : p = a$$

$$H_1 : p \neq a \qquad H_1 : p < a \qquad H_1 : p > a$$

$$\overline{X} = \hat{p} \sim N(p, SE = \sqrt{\frac{pq}{n}})$$

$$\frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Задача

В городе N. решили оценить число людей, которые будут голосовать за кандидата Тыкву. Считается, что доля людей не сильно изменилась с прошлого года, когда за него голосовало 70% жителей города. Проверьте на уровне значимости 0.01 гипотезу, что доля действительно не изменилась, при условии, что было опрошено 700 жителей и среди них голосовать за кандидата хотят 65% процентов людей.

Тест пропорций

Двувыборочный

$$H_0 : p_x - p_y = a$$

$$H_1 : p_x - p_y \neq a \quad H_1 : p_x - p_y < a \quad H_1 : p_x - p_y > a$$

$$\hat{p}_x \sim N(p_x, \sqrt{\frac{p_x q_x}{n_x}}) \qquad \hat{p}_y \sim N(p_y, \sqrt{\frac{p_y q_y}{n_y}})$$

$$\hat{p}_x - \hat{p}_y \sim N(p_x - p_y, SE = \sqrt{\frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}})$$

Тест пропорций

Двувыборочный

$$H_0 : p_x - p_y = a$$

$$\hat{p}_x - \hat{p}_y \sim N(p_x - p_y, \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}})$$

Задача

В городах N. и M. решили сравнить долю людей, которые будут голосовать за кандидата Тыкву. Предполагается, что в городе N. Поддержка Тыквы на 10% больше.. Проверьте гипотезу на уровне значимости 0.03. В городе N. Было опрошено 500 жителей и среди них доля голосующих за Тыкву - 59%, в городе M. - 550 жителей и доля голосующих за Тыкву - 47%.

Тест пропорций

Двувыборочный, частный случай

$$H_0 : p_x - p_y = 0 \quad \Rightarrow \quad p_x = p_y = p$$

Можем считать общую оценку для p. И подставлять ее в формулу для SE

$$\hat{p} = \frac{n_x p_x + n_y p_y}{n_x + n_y}$$

$$\hat{p}_x - \hat{p}_y \sim N(p_x - p_y, \sqrt{\frac{\hat{p}(1-\hat{p})}{n_x} + \frac{\hat{p}(1-\hat{p})}{n_y}})$$

$$\hat{p}_x - \hat{p}_y \sim N(p_x - p_y, \sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}})$$

Задача

В городах N. и M. решили сравнить долю людей, которые будут голосовать за кандидата Тыкву. Предполагается, что доли не отличаются. Проверьте гипотезу на уровне значимости 0.03. В городе N. Было опрошено 500 жителей и среди них доля голосующих за Тыкву - 59%, в городе M. - 550 жителей и доля голосующих за Тыкву - 54%.