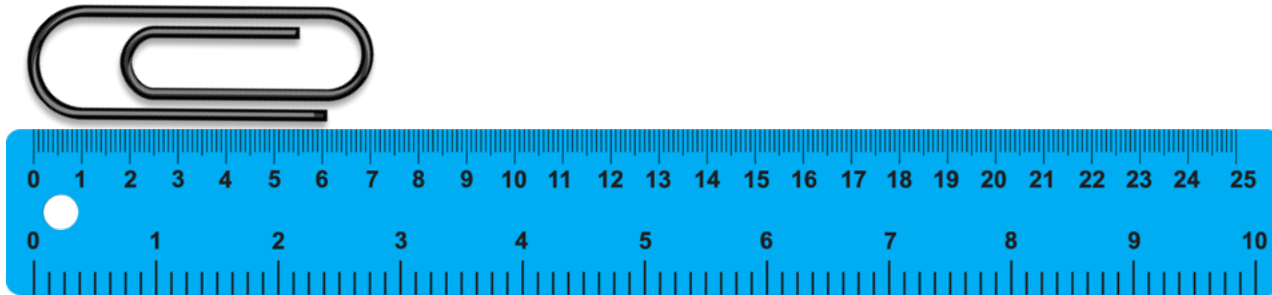


Доверительные интервалы

Что такое доверительный интервал

Зачем нужны доверительные интервалы

Надо измерить длину скрепки. Её длина 7 см, но мы не знаем наверняка, так как деления на линейке недостаточно точны



- Измерение делается с точностью, которую допускает линейка
- Длина скрепки 7 ± 0.1 см
- При дальнейших расчётах мы должны учитывать погрешность измерения

Предсказательный интервал

- Случайная величина $X \sim F(x)$
- Предсказательный интервал порядка $1 - \alpha$:

$$\mathbb{P}\left(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Для $X \sim N(\mu, \sigma^2)$ предсказательным интервалом будет

$$\mathbb{P}\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \sigma\right) = 1 - \alpha$$

! Границы предсказательного интервала – константы, случайная величина лежит между ними

Предсказательный интервал

- $\bar{x} \sim N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \Rightarrow$ предсказательный интервал для \bar{X} :

$$\mathbb{P}\left(\mu - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

- Доверительный интервал для μ :

$$\mathbb{P}\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right) = 1 - \alpha$$

❗ Границы доверительного интервала – случайные величины, мы пытаемся получить их по выборке


Доверительный интервал

Интервал $[\theta_L; \theta_U]$ называется **доверительным интервалом** для параметра θ , с уровнем доверия $1 - \alpha$, если при бесконечном повторении эксперимента в $100 \cdot (1 - \alpha)$ процентах случаев этот интервал будет покрывать истинное значение параметра θ

Величину α называют **уровнем значимости**

- ❗ Если мы много раз измеряем скрепку, то с вероятностью $1 - \alpha$ наш доверительный интервал покрывает её истинную длину

Зачем нужны доверительные интервалы

- Точечная оценка делается по случайной выборке \Rightarrow неопределённость
 - Нужно делать выводы в каком-то диапазоне
 - Доверительный интервал показывают, насколько мы уверены в точечной оценке
-  На практике пытаются построить наиболее короткий доверительный интервал

Зачем нужны доверительные интервалы

Антон:

С вероятностью 95% среднее лежит между 1 и 20

Ширина: 19

Наташа:

С вероятностью 95% среднее лежит между 17 и 23

Ширина: 6

- ❗ У обоих интервалов надёжность 95% (ошибка в 5% случаев), но разная точность. Наташин интервал уже, то есть точнее.

Зачем нужны доверительные интервалы

Многие метрики, интересные бизнесу, считаются по случайным выборкам, хочется знать, в каком диапазоне они изменяются.

Запасы полезных ископаемых оценивают по образцам пород (случайная выборка). Инвесторам хочется знать объём запасов в лучшем и в худшем случаях, а не только в среднем.

Обычно доверительные интервалы строят для прогнозов.

Асимптотические доверительные интервалы

Асимптотический интервал для среднего

- ЦПТ позволяет построить доверительный интервал для любого среднего
- Наблюдаем X_1, \dots, X_n
- **Предполагаем:** X_i независимы и одинаково распределены, число наблюдений n велико, нет выбросов

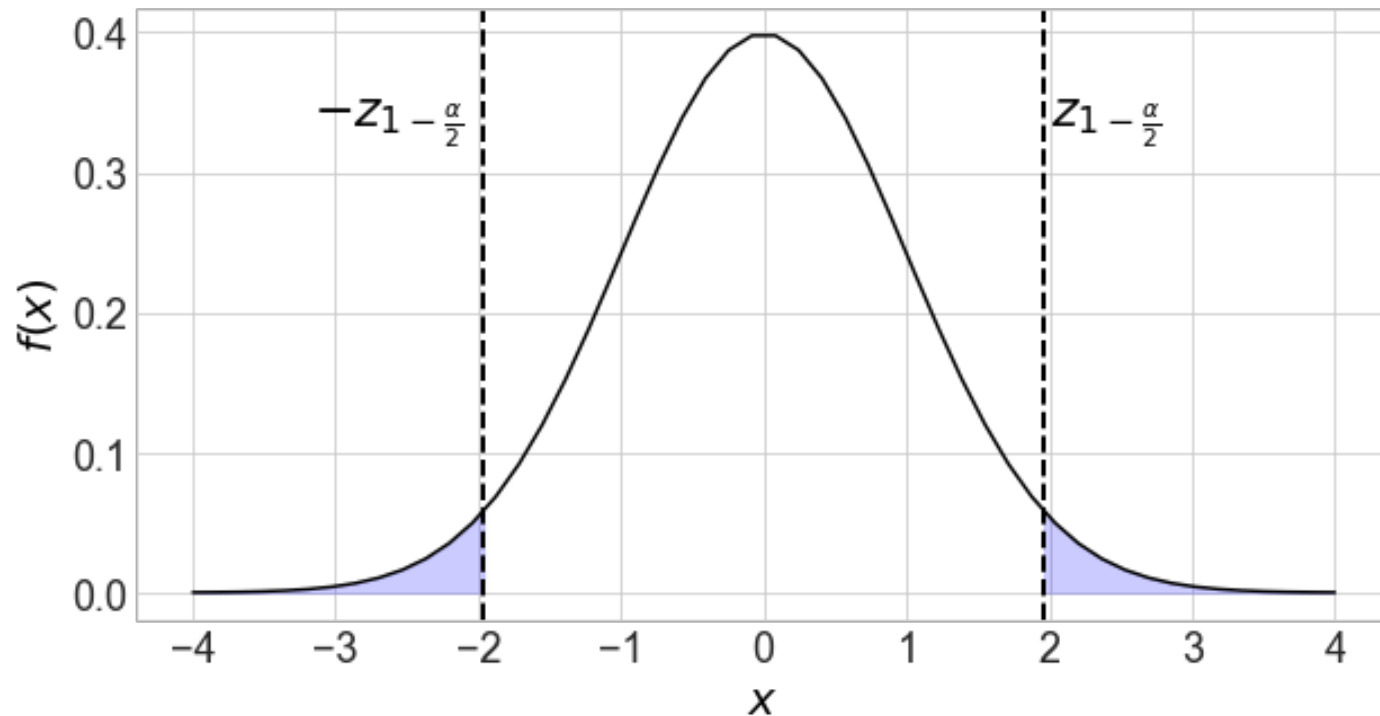
$$\bar{x} \stackrel{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \bar{x} - \mu \stackrel{asy}{\sim} N\left(0, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \stackrel{asy}{\sim} N(0, 1)$$

центрирование нормирование

Асимптотический интервал для среднего

Можно зафиксировать любую надежность $1 - \alpha$
и построить **доверительный интервал**:

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$



Асимптотический интервал для среднего

$$\mathbb{P} \left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{\sqrt{\hat{\sigma}^2/n}} \leq z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$\mathbb{P} \left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

$$\mathbb{P} \left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right) = 1 - \alpha$$

Почему можно заменить σ на $\hat{\sigma}$?

Почему можно заменить σ на $\hat{\sigma}$

По ЦПТ:
$$\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0,1) \text{ при } n \rightarrow \infty$$

$$\boxed{\frac{\sqrt{\frac{\hat{\sigma}^2}{n}}}{\sqrt{\frac{\sigma^2}{n}}}} \cdot \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \xrightarrow{d} N(0,1) \text{ при } n \rightarrow \infty$$

$$\overset{p}{\rightarrow} 1 \quad \overset{d}{\rightarrow} 1$$

Так как $\hat{\sigma}^2$ состоятельная оценка для σ^2 , $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$

Почему можно заменить σ на $\hat{\sigma}$

По ЦПТ: $\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0,1) \text{ при } n \rightarrow \infty$

$$1 \cdot \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \xrightarrow{d} N(0,1) \text{ при } n \rightarrow \infty$$

Получается, что при замене дисперсии на её оценку, предельное распределение не меняется.

$$\mathbb{P} \left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \right) = 1 - \alpha$$

Мощь средних

Длина интервала:

$$\Delta = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

При росте n длина интервала падает

При росте дисперсии длина интервала увеличивается

При росте надёжности $1 - \alpha$ длина увеличивается

Дельта-метод

Если:

$$X_1, \dots, X_n \sim iid, \quad \mathbb{E}(X_1) = \mu, \text{Var}(X_1) = \sigma^2$$

$g(t)$ – дифференцируемая функция

Тогда:

$$g(\bar{x}) \sim N\left(g(\mu), \frac{\sigma^2}{n} \cdot g'(\mu)^2\right)$$

Обобщение ЦПТ на случай функции от среднего.

Асимптотический интервал для дисперсии

Выборочную дисперсию можно выразить через средние

$$s^2 = \frac{n}{n-1} \cdot \hat{\sigma}^2 = \frac{n}{n-1} (\overline{x^2} - \bar{x}^2)$$

Немного поупражнявшись с ЦПТ и сходимостями можно получить асимптотическое распределение для выборочной дисперсии:

$$s^2 \sim N\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right), \quad \mu_4 = \mathbb{E}[(X_i - \mu)^4]$$

Оно может быть использовано для строительства доверительных интервалов

Резюме

- Доверительный интервал помогает понять, насколько надёжной получилась точечная оценка
- При большой выборке без выбросов ЦПТ помогает построить асимптотический доверительный интервал для любой функции от среднего
- Если наблюдений мало, нужны другие союзники

**Асимптотический доверительный
интервал для разницы средних**

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid$$

$$Y_1, \dots, Y_m \sim iid$$

$$\bar{x} \stackrel{asy}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

$$\bar{y} \stackrel{asy}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Разность нормальных случайных величин – нормальная случайная величина:

$$\mathbb{E}(\bar{x} - \bar{y}) = \mathbb{E}(\bar{x}) - \mathbb{E}(\bar{y}) = \mu_1 - \mu_2$$

$$Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

$$\bar{x} - \bar{y} \stackrel{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid$$

$$Y_1, \dots, Y_m \sim iid$$

$$\bar{x} - \bar{y} \stackrel{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

Асимптотический доверительный интервал для $\mu_1 - \mu_2$:

$$(\bar{x} - \bar{y}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}$$

Асимптотические доверительные интервалы для долей

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$X_1, \dots, X_n \sim iid \quad X_i = \begin{cases} 1, & \text{если любит кофе} \\ 0, & \text{если не любит кофе} \end{cases}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

Из-за того, что X_i принимают значение либо 0, либо 1, для оценки доли можно посчитать среднее

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x}$$

X_i	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	p

Найдём математическое ожидание и дисперсию оценки, а потом воспользуемся ЦПТ

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

$$\hat{p} = \frac{X_1 + \dots + X_n}{n} = \bar{x} \quad \begin{array}{c|c|c} X_i & 0 & 1 \\ \hline \mathbb{P}(X_i = k) & 1-p & p \end{array}$$

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n} \cdot n \cdot \mathbb{E}(X_1) = p$$

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} \cdot n \cdot \text{Var}(X_1) = \frac{p(1-p)}{n}$$

$$\bar{x} \overset{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \hat{p} = \bar{x} \overset{asy}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Мощь долей

Получаем доверительный интервал для доли:

$$\bar{x} \stackrel{asy}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \quad \hat{p} = \bar{x} \stackrel{asy}{\sim} N\left(p, \frac{\hat{p}(1 - \hat{p})}{n}\right)$$

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Мощь долей

Получаем доверительный интервал для разности долей:

$$\bar{x} - \bar{y} \stackrel{asy}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

$$\hat{p}_1 - \hat{p}_2 \stackrel{asy}{\sim} N\left(p_1 - p_2, \frac{\hat{p}_1(1 - \hat{p}_1)}{n} + \frac{\hat{p}_2(1 - \hat{p}_2)}{m}\right)$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n} + \frac{\hat{p}_2(1 - \hat{p}_2)}{m}}$$

Число наблюдений

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Можно определить число наблюдений, чтобы длина доверительного интервала не превышала заранее выбранный диапазон

$$\Delta = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$n = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \hat{p}(1-\hat{p})}{\Delta^2}$$

Число наблюдений

$$n = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \hat{p}(1 - \hat{p})}{\Delta^2}$$

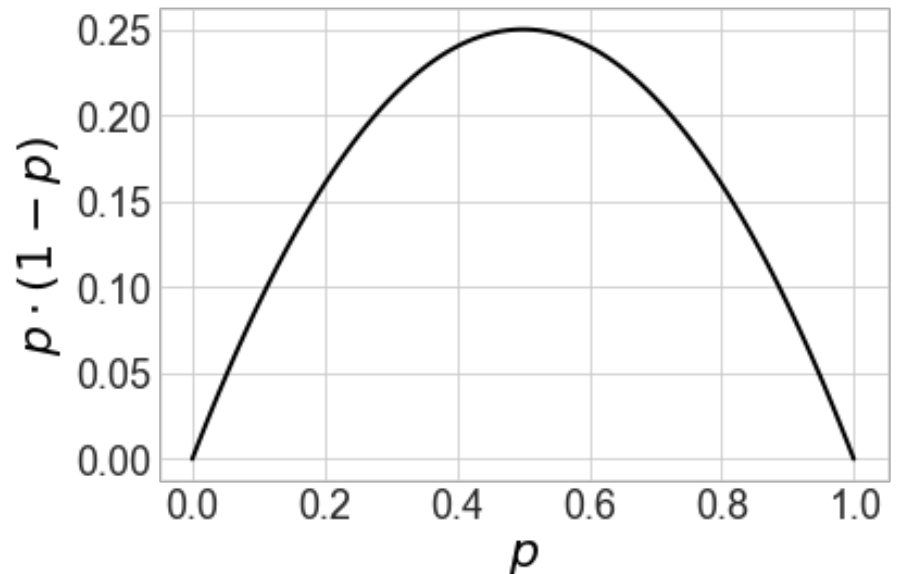
До начала испытаний мы не знаем \hat{p} , но мы знаем, что величина $\hat{p}(1 - \hat{p})$ никогда не будет превышать **0.25**

$$f(p) = p \cdot (1 - p) = p - p^2$$

$$f'(p) = 1 - 2p = 0$$

$$\Rightarrow p = 0.5$$

$$f(p) = 0.25$$



Число наблюдений

$$n = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \hat{p}(1 - \hat{p})}{\Delta^2}$$

До начала испытаний мы не знаем \hat{p} , но мы знаем, что величина $\hat{p}(1 - \hat{p})$ никогда не будет превышать 0.25

Эту оценку сверху мы можем использовать для поиска необходимого значения n :

$$n = \frac{4 \cdot z_{1-\frac{\alpha}{2}}^2 \cdot \hat{p}(1 - \hat{p})}{\Delta^2} \leq \frac{z_{1-\frac{\alpha}{2}}^2}{\Delta^2}$$

Резюме

- Доля – это среднее, посчитанное по выборке из нулей и единиц
- С помощью ЦПТ можно построить доверительные интервалы для долей
- Из-за того, что вероятность принимает значения на отрезке от нуля до единицы, мы можем оценить, сколько наблюдений нам нужно собрать для определённой ширины интервала

Точные доверительные интервалы для нормальных выборок

Схема математической статистики

Выборка: x_1, \dots, x_n Параметр: θ

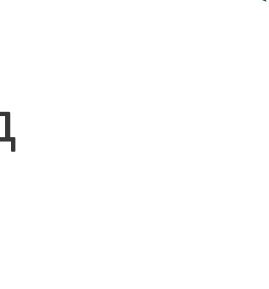
$\hat{\theta}$



Точность
оценки,
прогнозов



доверительные
интервалы



Ответы на
вопросы
проверка
гипотез

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Союзники

Асимптотические
(при большом n)

- ЦПТ
- Дельта-метод

Точные

- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Схема математической статистики

Выборка: x_1, \dots, x_n Параметр: θ

$\hat{\theta}$

$f_{\hat{\theta}}(t)$

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники

Асимптотические
(при большом n)

- ЦПТ
- Дельта-метод

Точные

- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Точность
оценки,
прогнозов

доверительные
интервалы

Ответы на
вопросы
проверка
гипотез

**Точные доверительные интервалы
для нормальных выборок: средние**

Доверительные интервалы для нормального

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2)$$



Строим
доверительный
интервал для μ :
 σ^2 известна
 σ^2 неизвестна



Строим доверительный
интервал для σ^2 :
 μ известно
 μ неизвестно

Дисперсия известна

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \sigma^2 \text{ известна}$$

Известно, что распределение точное, ЦПТ использовать не нужно

Пример: Измеряем что-то, знаем погрешность прибора

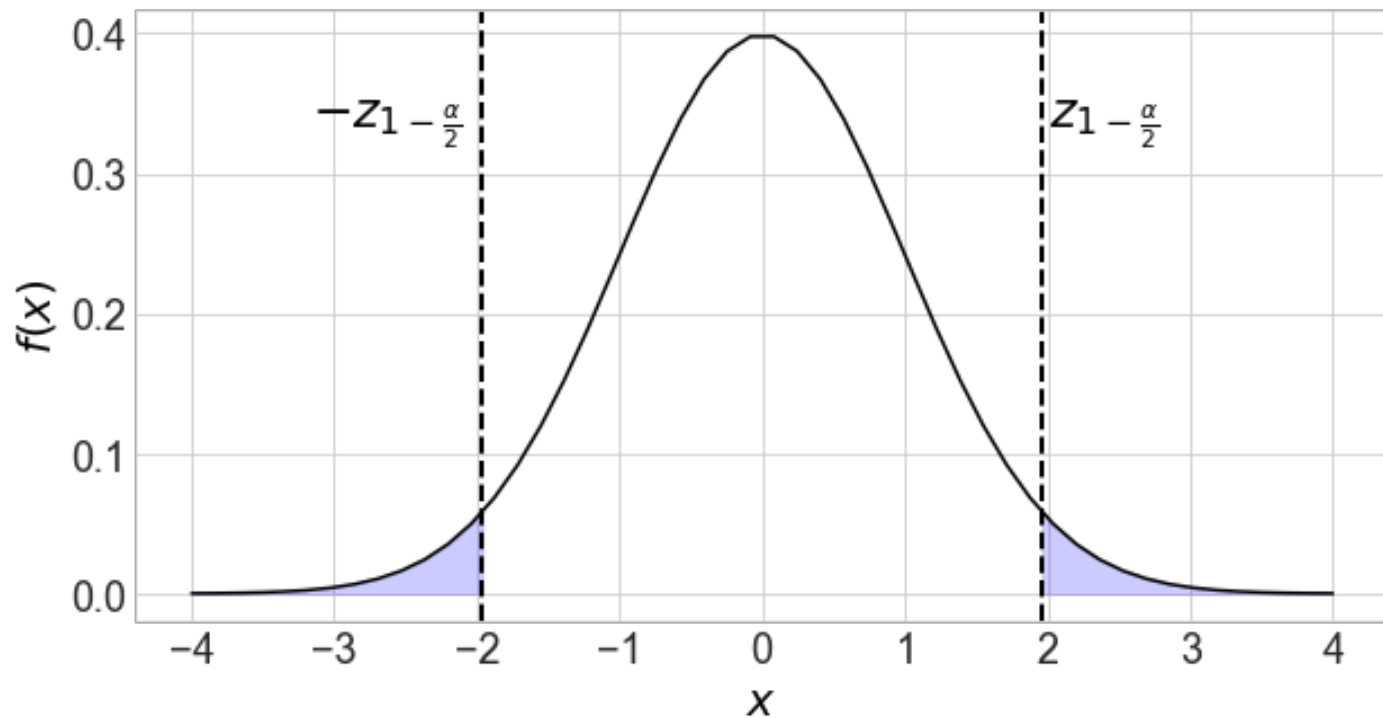
$$\hat{\mu} = \bar{x} = \frac{X_1 + \dots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Распределение точное, сумма нормальных случайных величин – нормальна.

Дисперсия известна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \sigma^2$ известна

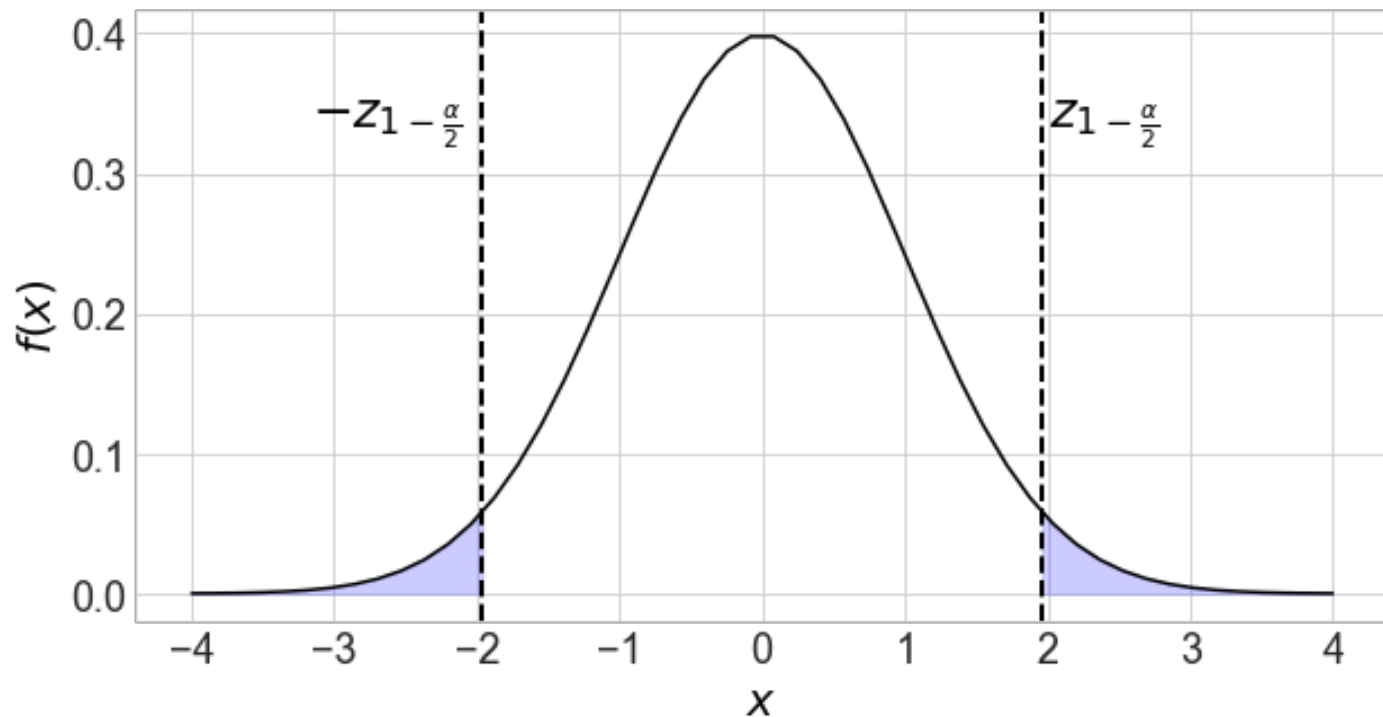
$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$



Дисперсия известна

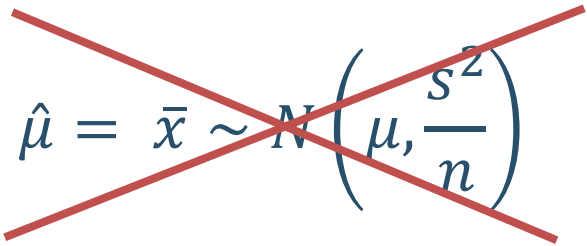
Доверительный интервал строится по аналогии с асимптотикой, но является точным:

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2),$ σ^2 **не**известна

$$\hat{\mu} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\hat{\mu} = \bar{x} \sim N\left(\mu, \frac{s^2}{n}\right)$$

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim ???$$

Союзники: распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы

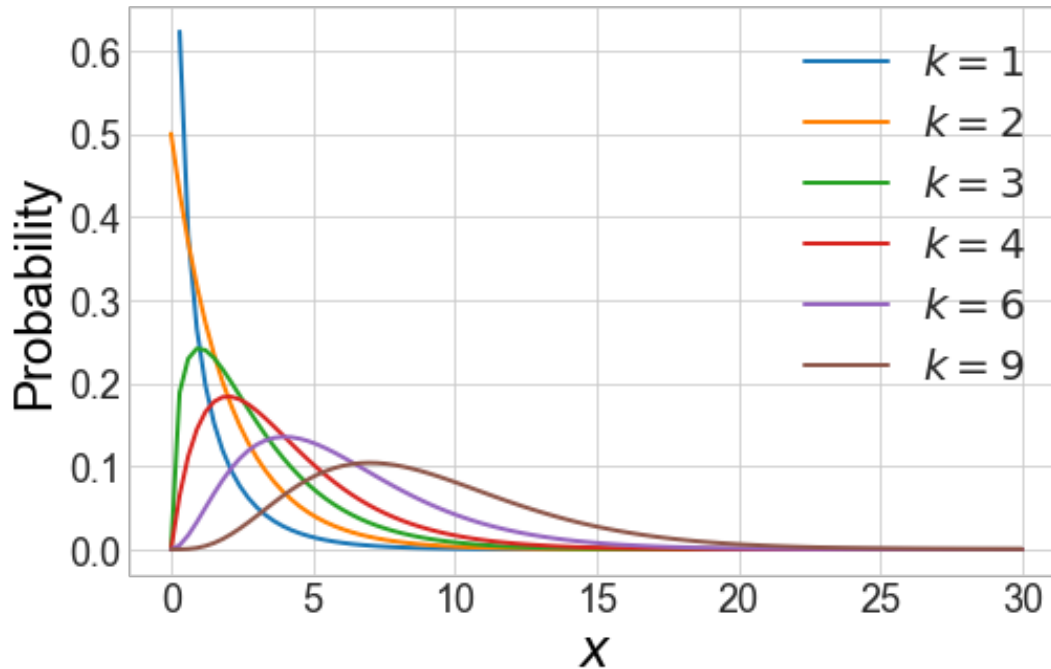


Когда возникает
на практике:

$$\hat{\sigma}^2 = \overline{x^2} - \bar{x}^2$$

- Если выборка пришла из $N(0,1)$, величина $\overline{x^2}$ будет иметь “хи-квадрат” распределение
- Для выборочной дисперсии тоже можно получить “хи-квадрат” распределение

Союзники: распределение хи-квадрат



$$X_1, \dots, X_k \sim iid N(0,1)$$

$$Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$$

Из-за квадратов
принимает только
положительные
значения

Плотность:

$$f(x) = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, x \geq 0$$

Характеристики:

$$\mathbb{E}(X) = k$$

$$Var(X) = 2k$$

Союзники: распределение Стьюдента

Независимые случайные величины $X_0 \sim N(0,1)$, $Y \sim \chi_k^2$.

Тогда случайная величина

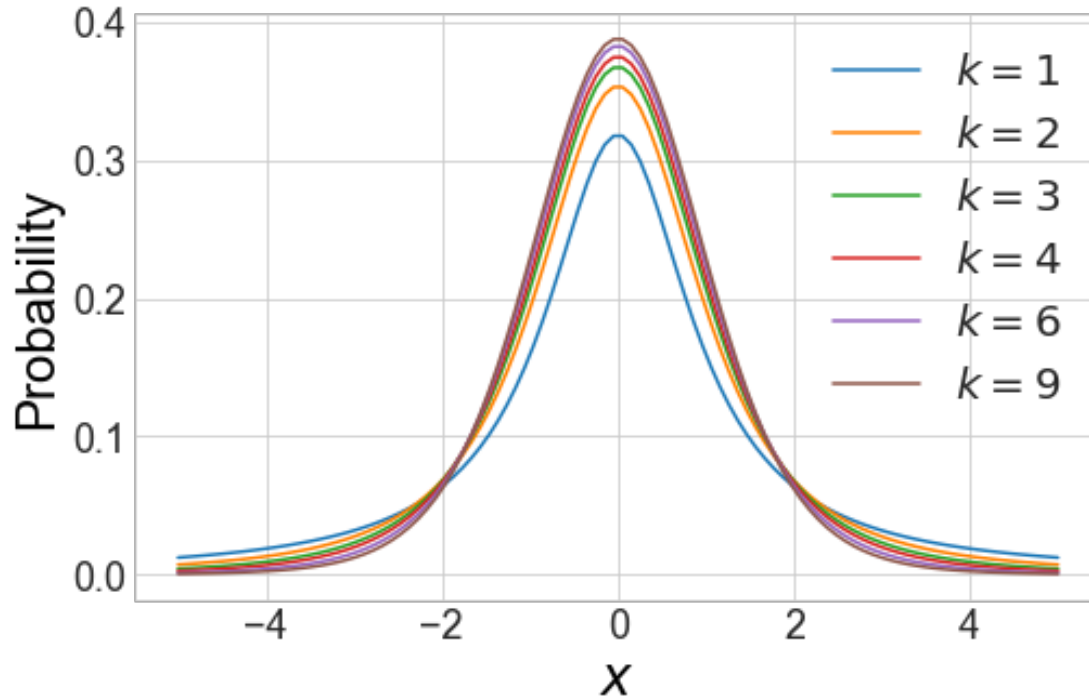
$$t = \frac{X_0}{\sqrt{Y/k}} \sim t(k)$$

имеет распределение Стьюдента с k степенями свободы.

✓ Когда возникает на практике:

Мы будем часто встречаться с выражением $\frac{\bar{x}}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$,
имеющим распределение Стьюдента

Союзники: распределение Стьюдента



$$X_0 \sim N(0,1), Y \sim \chi_k^2,$$

$$t = \frac{X_0}{\sqrt{Y/k}} \sim t(k)$$

Плотность:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

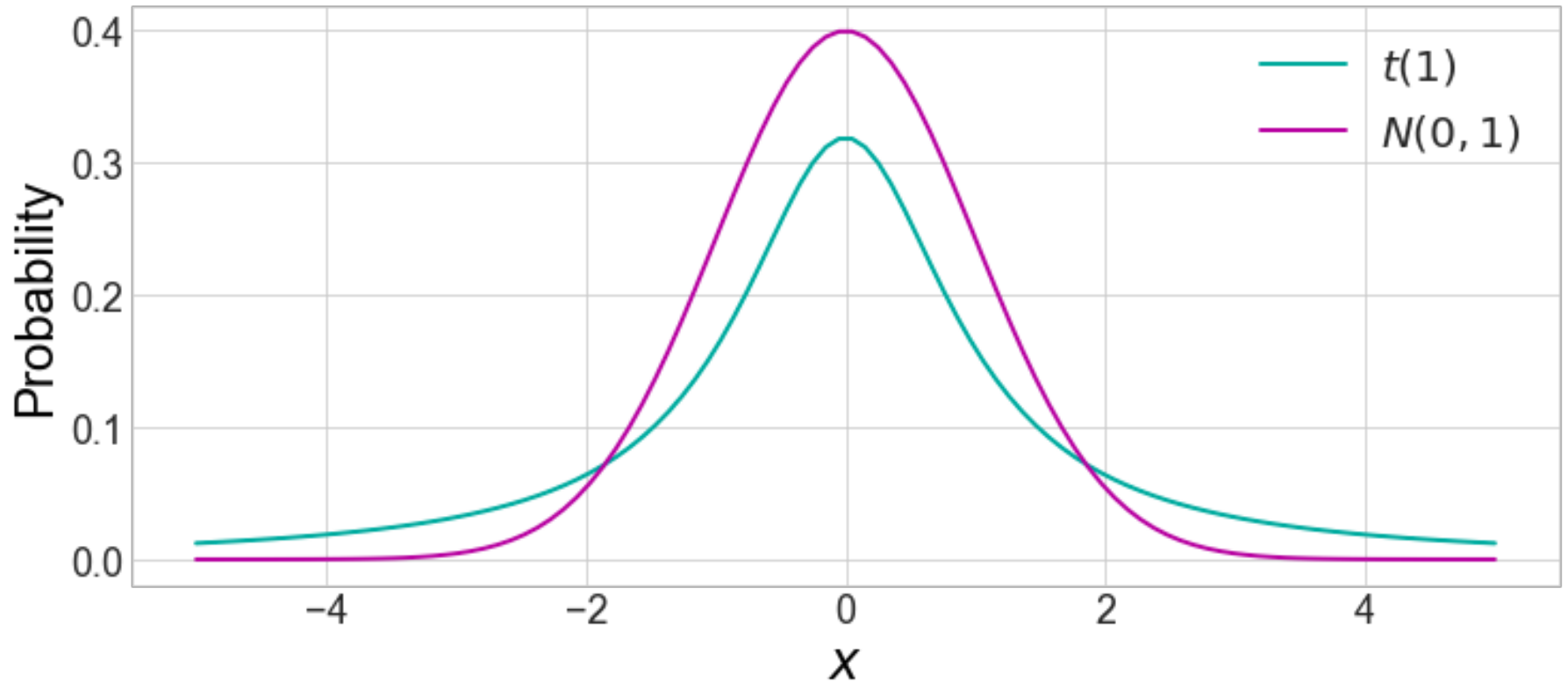
Характеристики:

$$\mathbb{E}(t) = 0$$

$$\text{Var}(t) = \frac{k}{k-2}, k > 2$$

Тяжёлые хвосты

Распределение Стьюдента обладает более тяжёлыми хвостами, нежели нормальное



Союзники: теорема Фишера


Теорема:

Пусть $X_1, \dots, X_n \sim iid N(0,1)$, тогда

1. Выборочное среднее \bar{x} и дисперсия s^2 независимы
2. $\frac{(n-1) \cdot s^2}{\sigma^2}$ имеет χ^2 – распределение с $n - 1$ степенью свободы

Дисперсия неизвестна

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \sigma^2 \text{ не известна}$$

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$


Надо заменить на σ^2 , чтобы получить нормальное

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2),$ σ^2 **не**известна

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{\sigma^2}{(n-1)}}} = \boxed{\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}} \cdot \boxed{\frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{s^2}{(n-1)}}}}$$

$N(0, 1)$?

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \sigma^2$ **не**известна

$$\frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{s^2}{(n-1)}}} = \frac{1}{\sqrt{\frac{(n-1) \cdot s^2}{(n-1) \cdot \sigma^2}}} = \frac{1}{\sqrt{\frac{(n-1) \cdot s^2}{\sigma^2}} / (n-1)}$$

χ_{n-1}^2

По теореме Фишера
(работает только для
нормальных выборок)

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2),$ σ^2 **не**известна

$$\frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{s^2}{(n-1)}}} = \frac{1}{\sqrt{\frac{(n-1) \cdot s^2}{(n-1) \cdot \sigma^2}}} = \boxed{\frac{1}{\sqrt{\frac{(n-1) \cdot s^2}{\sigma^2} / (n-1)}}}$$

$$\sqrt{\frac{1}{\frac{\chi_{n-1}^2}{n-1}}}$$

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2),$ σ^2 **не**известна

$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{\sigma^2}{(n-1)}}} = \boxed{\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}} \cdot \boxed{\frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{s^2}{(n-1)}}}}$$

$N(0, 1)$

$$\sqrt{\frac{1}{\frac{\chi_{n-1}^2}{n-1}}}$$

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2),$ σ^2 **не**известна

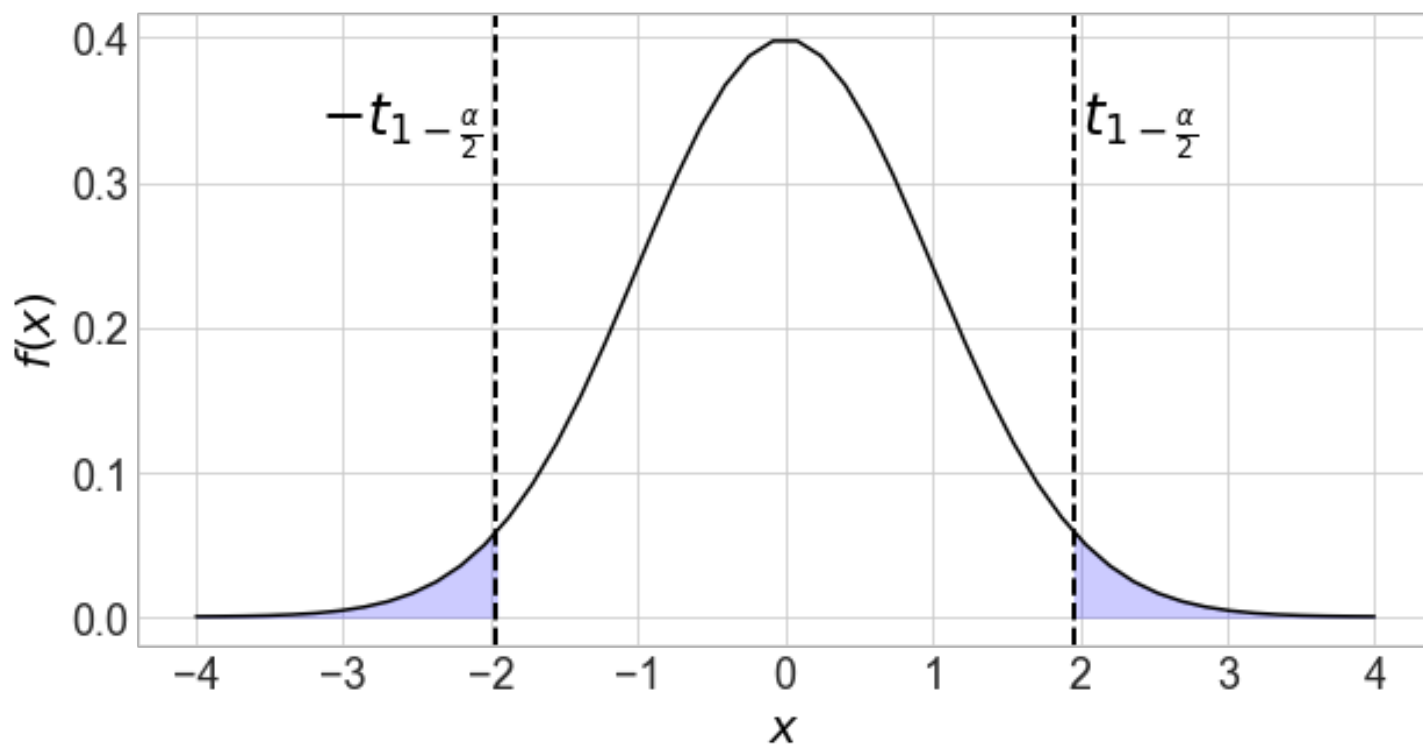
$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{\sigma^2}{(n-1)}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{s^2}{(n-1)}}}$$

$$\frac{N(0, 1)}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}} = t(n-1)$$

Дисперсия неизвестна

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \sigma^2$ **не**известна

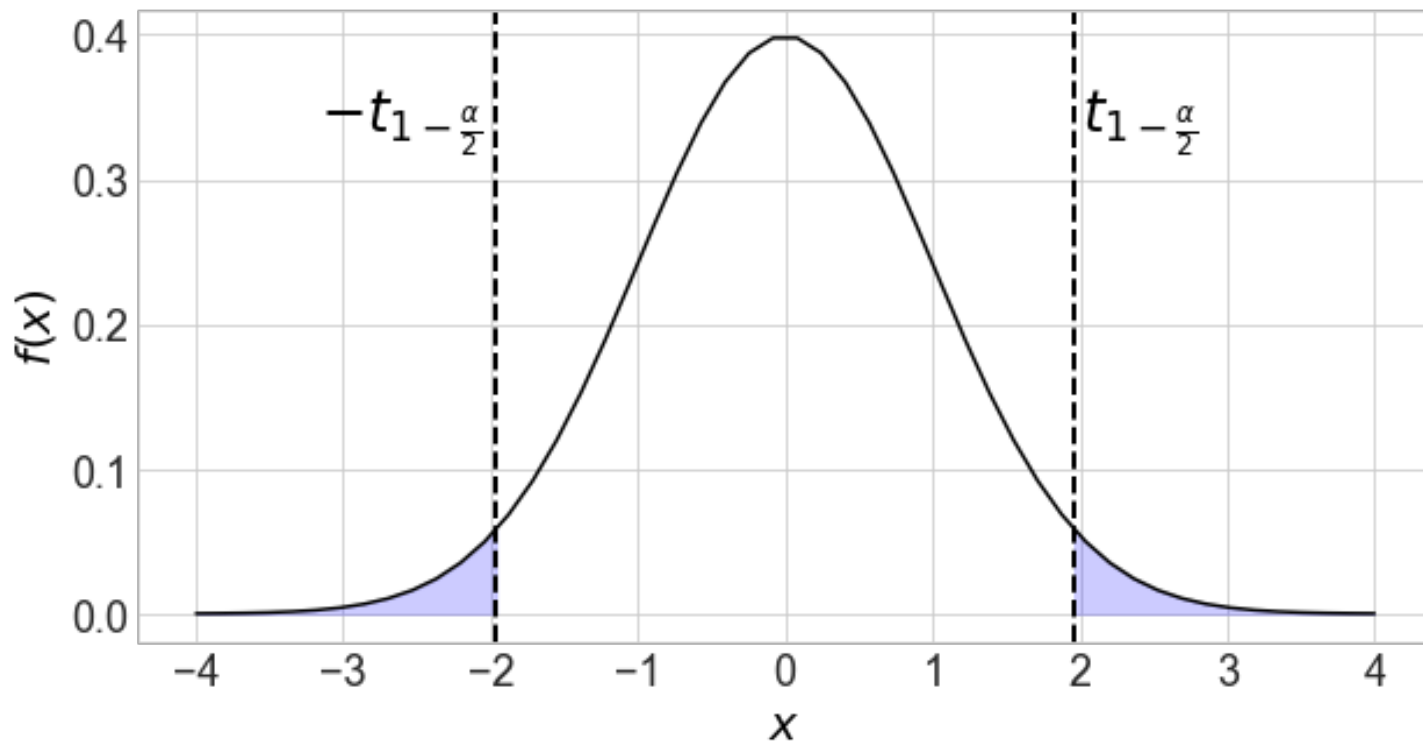
$$\frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t(n-1)$$



Дисперсия неизвестна

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \sigma^2 \text{ не известна}$$

$$P\left(\bar{x} - t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$



Точный vs Асимптотический

Асимптотический

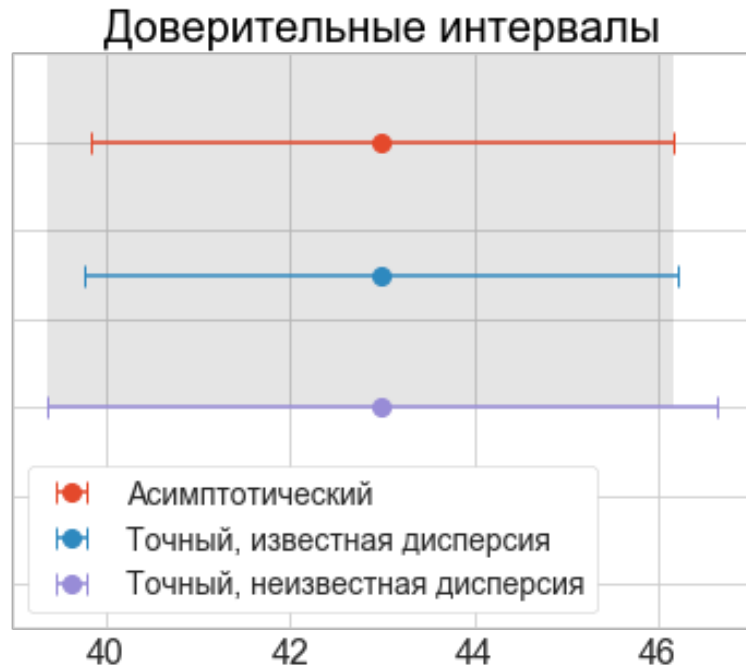
- Союзник: ЦПТ
- Работает при большом n
- Выборка независимая, без аномалий

Точный

- Союзники: теорема Фишера, t -распределение
- Работает при любом n
- Выборка независимая из нормального распределения

Пример

Измерили зарплаты: $\bar{x} = 43$ тыс. и $s = 5.1$ тыс. В выборку попало 10 наблюдений. В реальности $\sigma = 5.2$ тыс. (знаем из переписи населения)



$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad 43 \pm 1.96 \cdot \frac{5.1}{\sqrt{10}}$$

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \quad 43 \pm 1.96 \cdot \frac{5.2}{\sqrt{10}}$$

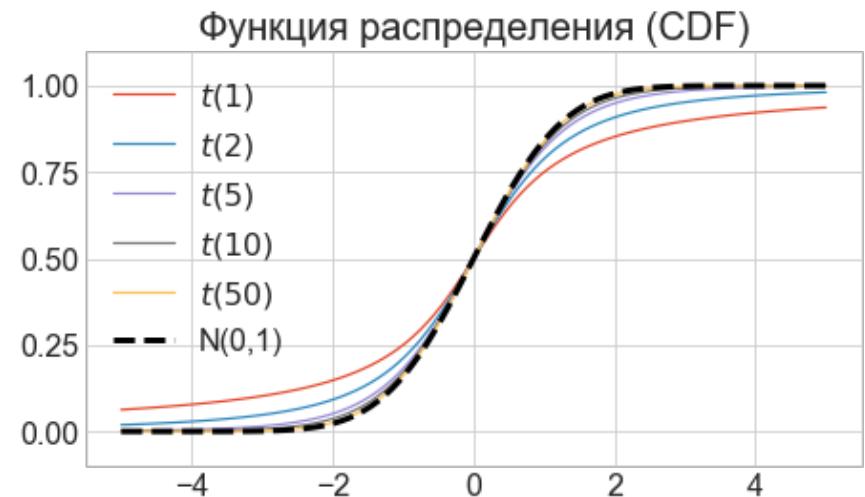
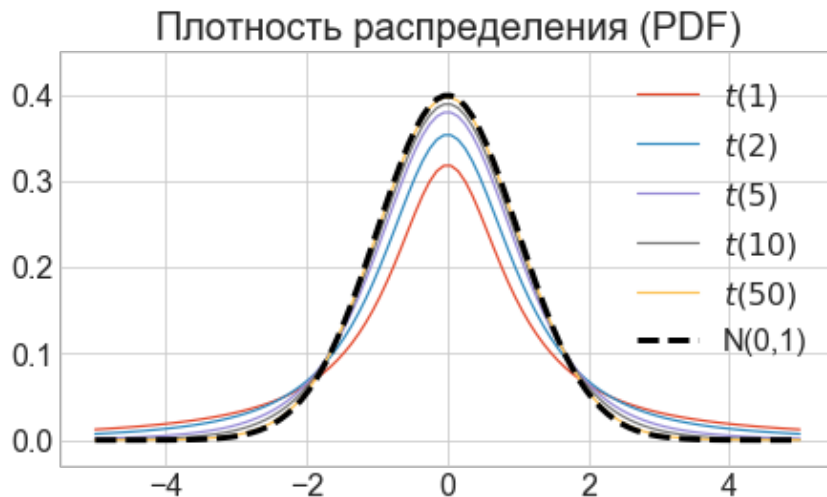
$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad 43 \pm 2.26 \cdot \frac{5.1}{\sqrt{10}}$$

- ✔ Точные доверительные интервалы часто оказываются шире асимптотических

Когда начинаются большие n

Распределение Стьюдента сходится к нормальному по распределению при росте числа степеней свободы:

$$t(n) \xrightarrow{d} N(0,1) \text{ при } n \rightarrow \infty$$



- ✓ При больших выборках разница между точным и асимптотическим интервалами минимальна

Резюме

Если известно распределение, можно строить точные доверительные интервалы

Для нормальных выборок при неизвестной дисперсии в этом помогает распределение Стьюдента

Из-за того, что распределение Стьюдента обладает более тяжёлыми хвостами, чем нормальное, точные доверительные интервалы обычно оказываются шире

**Точные доверительные интервалы
для нормальных выборок:
разность средних**

Асимптотический интервал для разности средних

- ЦПТ позволяет построить доверительный интервал для любого среднего
- Наблюдаем X_1, \dots, X_{n_x} и Y_1, \dots, Y_{n_y}
- **Предполагаем:** X_i, Y_i независимы и одинаково распределены, число наблюдений велико, нет выбросов, выборки независимы друг от друга

$$\bar{x} \overset{asy}{\approx} N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right) \qquad \bar{y} \overset{asy}{\approx} N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$$

$$\bar{x} - \bar{y} \overset{asy}{\approx} N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

Асимптотический интервал для разности средних

- ЦПТ позволяет построить доверительный интервал для любого среднего
- Наблюдаем X_1, \dots, X_{n_x} и Y_1, \dots, Y_{n_y}
- **Предполагаем:** X_i, Y_i независимы и одинаково распределены, число наблюдений велико, нет выбросов, выборки независимы друг от друга



Теперь хотим
построить точный
интервал

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \stackrel{asy}{\approx} N(0,1)$$
$$(\bar{x} - \bar{y}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

The diagram illustrates the three cases for the denominator of the z-test formula based on the knowledge of variances:

- Left branch: дисперсии известны (variances are known)
- Bottom branch: дисперсии неизвестны, но равны (variances are unknown, but equal)
- Right branch: дисперсии неизвестны, различаются (variances are unknown, but different)

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0, 1)$$

дисперсии
известны



Можем строить
точный интервал

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s^2}{n_x} + \frac{s^2}{n_y}}}$$

дисперсии
неизвестны,
но равны

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s^2}{n_x} + \frac{s^2}{n_y}}} \sim t(n_x + n_y - 2)$$

Объединённая оценка
дисперсии:

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

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim t(v)$$

дисперсии
неизвестны,
различаются

Разность средних (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_{n_x} \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_{n_y} \sim iid N(\mu_y, \sigma_y^2)$$

Нас интересует случайная величина:

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim t(v)$$



Распределение
приближенное
(распределение
Уэлча)

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right)^2}{\frac{s_x^4}{n_x^2(n_x - 1)} + \frac{s_y^4}{n_y^2(n_y - 1)}}$$

Проблема Беренца-Фишера

Не существует точного распределения для статистики

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

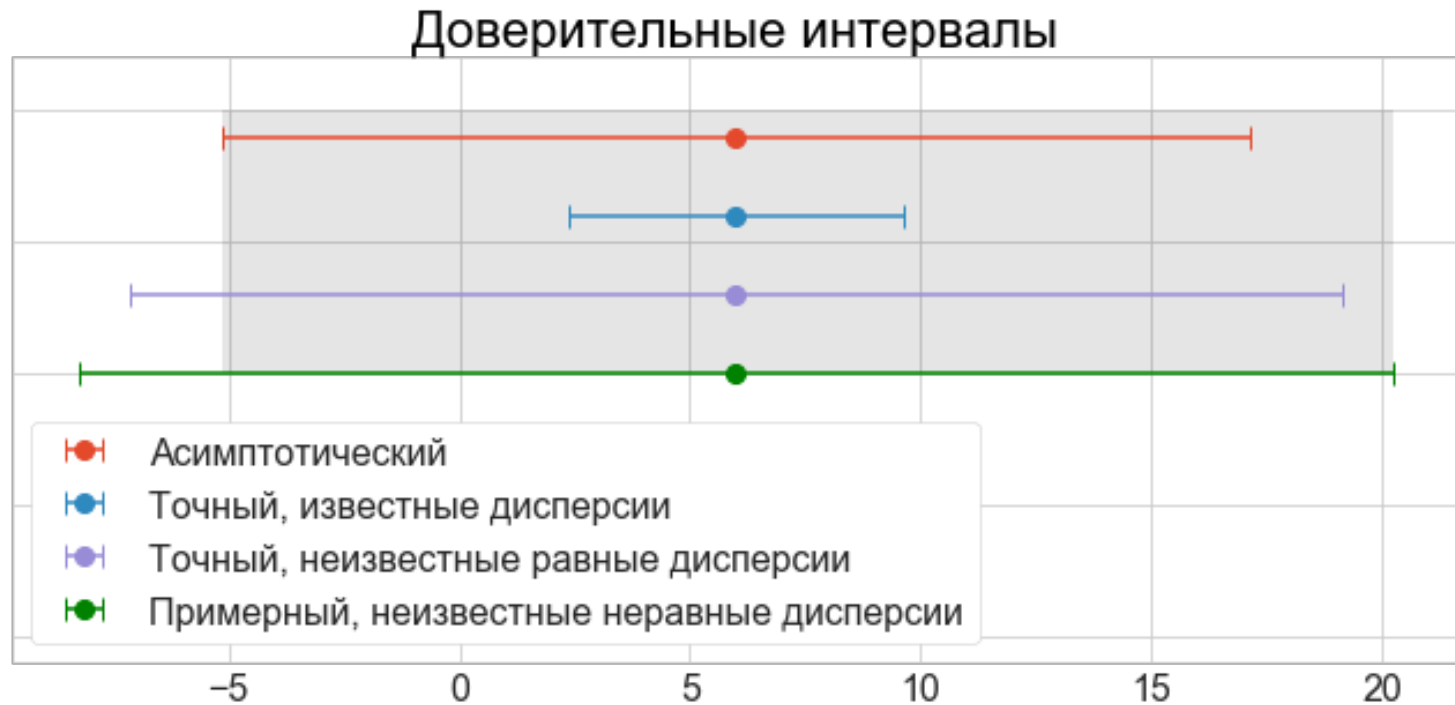
Невозможно точно сравнить средние двух независимых выборок, дисперсии которых неизвестны.

Аппроксимация с предыдущего слайда хорошо работает, если $n_x = n_y$ либо знак неравенства между n_x и n_y такой же как между σ_x и σ_y

Пример 1

Измерили зарплаты мужчин и женщин в тысячах рублей: $\bar{x} = 43$, $s_x = 5.1$, $\bar{y} = 37$, $s_y = 11.7$. В обеих выборках было по 10 наблюдений.

Из переписи известно, что $\sigma_x = 5.2$, $\sigma_y = 12$



Пример 1

Неизвестны (асимптотика):

$$\bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$43 - 37 \pm 1.96 \cdot \sqrt{\frac{5.1^2}{10} + \frac{11^2}{10}}$$

Известны (точный):

$$\bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

$$43 - 37 \pm 1.96 \cdot \sqrt{\frac{5.2^2}{10} + \frac{12^2}{10}}$$

Неизвестны, равны (точный):

$$\bar{x} - \bar{y} \pm t(n_x + n_y - 2)_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n_x} + \frac{s^2}{n_y}}$$

$$43 - 37 \pm 2.3 \cdot \sqrt{\frac{81}{10} + \frac{81}{10}}$$

Неизвестны, не равны (примерный):

$$\bar{x} - \bar{y} \pm t(v)_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$43 - 37 \pm 2.51 \cdot \sqrt{\frac{5.1^2}{10} + \frac{11^2}{10}}$$

Разность средних (зависимые выборки)

Выборки зависят друг от друга:

$$X_1, \dots, X_n \sim iid N(\mu_x, \sigma_x^2) \quad Y_1, \dots, Y_n \sim iid N(\mu_y, \sigma_y^2)$$

- Измерения делаются на одних и тех же объектах
- Можем посмотреть прирост на отдельных объектах

$$d_i = X_i - Y_i$$

$$\bar{x} - \bar{y} = \overline{x - y}$$

- Получаем ситуацию с распределением Стьюдента, дисперсию считаем по формуле:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$$

Пример 2

Измерили зарплаты в 2020 и 2021 годах.

Измеряли для одних и тех же людей.

2020	50	40	45	45	35
2021	60	30	30	35	30
d_i	10	-10	-15	-10	-5

$$\bar{d} = \frac{1}{5} \sum_{i=1}^5 d_i = -6 \qquad s^2 = \frac{1}{5-1} \sum_{i=1}^5 (d_i - \bar{d})^2 = 92.5$$

Точный, неизвестная дисперсия:

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad -6 \pm 2.78 \cdot \frac{9.62}{\sqrt{5}}$$

Резюме

В зависимости от того, что мы знаем о дисперсии, для разности средних из независимых нормальных выборок мы получаем разные виды доверительных интервалов

Для средних из зависимых выборок (наблюдаем изменения на одних и тех же объектах) работают те же самые доверительные интервалы, что и для одновыборочных средних

**Точные доверительные интервалы
для нормальных выборок: дисперсии**

Зачем оценивать интервалы для дисперсий

Станок упаковывает чай по 100 грамм с какой-то заданной дисперсией. Если настройки станка расшатываются и погрешность становится слишком большой, получаем много бракованных партий.

Любая ценная бумага оценивается через среднюю доходность. Чем больше риск, тем выше доходность. Инвестору при формировании портфеля важно знать, в каком диапазоне для бумаги могут меняться обе характеристики. Один из способов посчитать риск – оценка дисперсии.

Союзники: теорема Фишера

Теорема:

Пусть $X_1, \dots, X_n \sim iid N(0,1)$, тогда

1. Выборочное среднее \bar{x} и дисперсия s^2 независимы
2. $\frac{(n-1) \cdot s^2}{\sigma^2}$ имеет χ^2 – распределение с $n - 1$ степенью свободы

Распределение Фишера

Независимые случайные величины $X \sim \chi_k^2$, $Y \sim \chi_m^2$.

Случайная величина

$$F = \frac{\sqrt{X/k}}{\sqrt{Y/m}} \sim F(k, m)$$

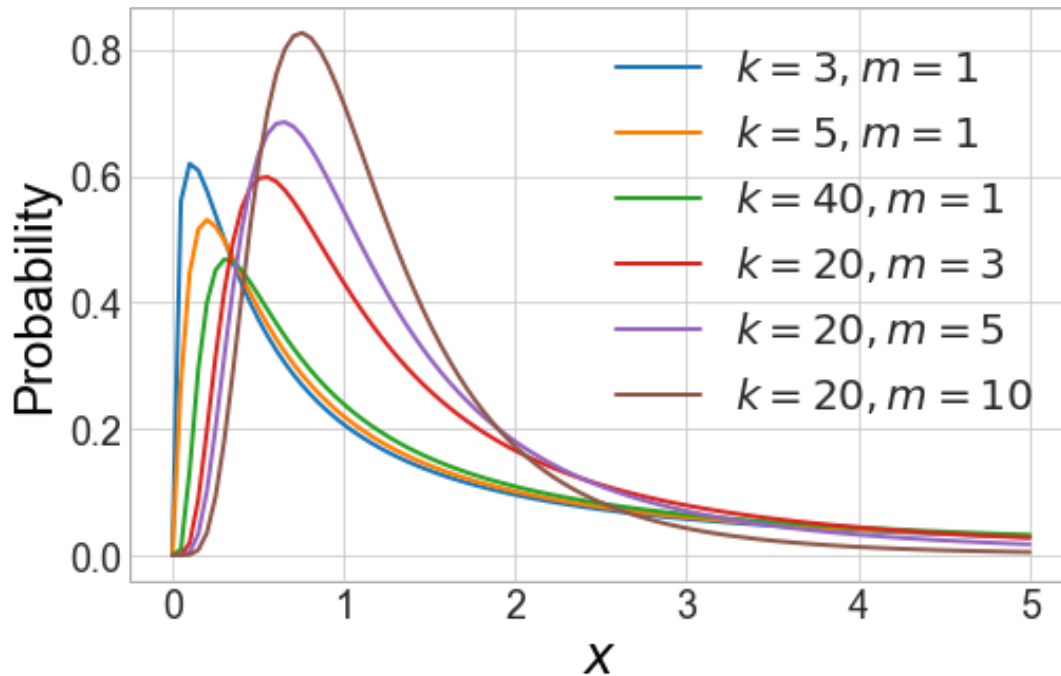
имеет распределение Фишера с k, m степенями свободы.



Когда возникает на практике:

Встречается при сравнении дисперсий.
Чтобы сравнить их между собой, одну дисперсию делят на вторую.

Распределение Фишера



$$X \sim \chi_k^2, Y \sim \chi_m^2$$

$$F = \frac{\sqrt{X/k}}{\sqrt{Y/m}} \sim F(k, m)$$

Из-за квадратов
принимает только
положительные
значения

Характеристики:

$$\mathbb{E}(F) = \frac{m}{m-2}, m > 2$$

$$\text{Var}(F) = \frac{2m^2(k+m-2)}{n(m-2)^2(m-4)}$$

Плотность:

Очень громоздкая

Доверительные интервалы для нормального

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2)$$



Строим
доверительный
интервал для μ :
 σ^2 известна
 σ^2 неизвестна



Строим доверительный
интервал для σ^2 :
 μ известно
 μ неизвестно

Математическое ожидание известно

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \mu$ известно

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

$[N(0, \sigma^2)]^2$

Надо как-то привести к χ_n^2

Математическое ожидание известно

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \mu$ известно

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 = \frac{\sigma^2}{n} \sum_{i=1}^n \boxed{\frac{(X_i - \mu)^2}{\sigma^2}} = \frac{\sigma^2}{n} \cdot \chi_n^2$$

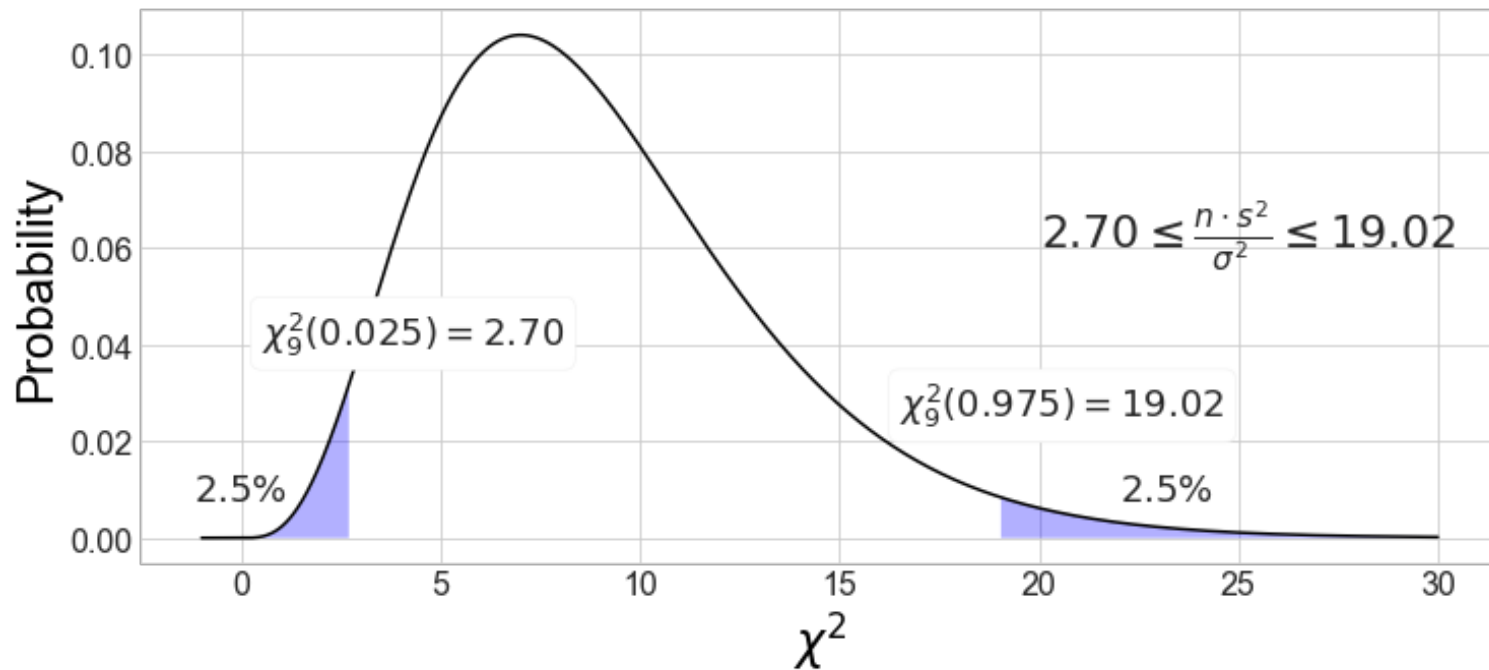
$[N(0, 1)]^2$

$$\frac{n \cdot s^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

Математическое ожидание известно

$$\frac{n \cdot s^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

$$P\left(\chi_n^2\left(\frac{\alpha}{2}\right) \leq \frac{n \cdot s^2}{\sigma^2} \leq \chi_n^2\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$



Математическое ожидание известно

$$\frac{n \cdot s^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

$$P\left(\chi_n^2\left(\frac{\alpha}{2}\right) \leq \frac{n \cdot s^2}{\sigma^2} \leq \chi_n^2\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$


$$P\left(\frac{n \cdot s^2}{\chi_n^2\left(1 - \frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{n \cdot s^2}{\chi_n^2\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

Математическое ожидание неизвестно

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \mu$ **не**известно

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

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



Оценка ломает всю логику
Нужен новый союзник

Математическое ожидание неизвестно

$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \mu \text{ неизвестно}$

Теорема Фишера:

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

В ситуации, когда математическое ожидание известно, у статистики n степеней свободы

Когда оно неизвестно, у статистики $n - 1$ степень свободы

Интуиция: одна степень свободы используется для оценки математического ожидания

Математическое ожидание неизвестно

$$X_1, \dots, X_n \sim iid N(\mu, \sigma^2), \quad \mu \text{ не известно}$$

Теорема Фишера:

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

$$P\left(\chi_{n-1}^2\left(\frac{\alpha}{2}\right) \leq \frac{(n-1) \cdot s^2}{\sigma^2} \leq \chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$P\left(\frac{(n-1) \cdot s^2}{\chi_{n-1}^2\left(1 - \frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1) \cdot s^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

Пример

Джордан считает, что вложения в бумаги с высокой дисперсией доходности рискованно, и хочет знать, в каком диапазоне колеблется дисперсия для одной из его акций. За последние 10 лет для бумаги $s^2 = 0.05$.

$$\frac{(10 - 1) \cdot 0.05}{\chi_9^2(0.975)} \leq \sigma^2 \leq \frac{(10 - 1) \cdot 0.05}{\chi_9^2(0.025)}$$

$$0.016 \leq \sigma^2 \leq 0.038$$

Пример

Джордан считает, что вложения в бумаги с высокой дисперсией доходности рискованно, и хочет знать, в каком диапазоне колеблется дисперсия для одной из его акций. За последние 10 лет для бумаги $s^2 = 0.05$.

Джордан инсайдер и знает доходность бумаги (это каким инсайдером надо быть!). Получилось, что $s^2 = 0.04$.

$$\frac{10 \cdot 0.05}{\chi_{10}^2(0.975)} \leq \sigma^2 \leq \frac{10 \cdot 0.05}{\chi_{10}^2(0.025)}$$

$$0.015 \leq \sigma^2 \leq 0.039$$

Резюме

Если известно распределение, можно строить точные доверительные интервалы не только для математических ожиданий, но и для дисперсий

Для нормальных выборок в этом помогают теорема Фишера и распределение “Хи-квадрат”

**Точные доверительные интервалы
для нормальных выборок:
отношение дисперсий**

Отношение дисперсий (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_n \sim iid N(\mu_1, \sigma_1^2) \quad Y_1, \dots, Y_m \sim iid N(\mu_2, \sigma_2^2)$$

Нас интересует случайная величина:

$$\frac{s_n^2}{s_m^2} \sim ?$$

Из-за квадратов разность оказывается плохой мерой для различия в дисперсиях

Отношение дисперсий (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_n \sim iid N(\mu_1, \sigma_1^2) \quad Y_1, \dots, Y_m \sim iid N(\mu_2, \sigma_2^2)$$

Нас интересует случайная величина:

$$\frac{s_n^2}{s_m^2} \sim ?$$

$$\frac{s_n^2 \cdot \sigma_m^2}{s_m^2 \cdot \sigma_n^2}$$

Теорема Фишера:

$$\frac{(n-1) \cdot s_n^2}{\sigma_n^2} \sim \chi_{n-1}^2$$

$$\boxed{\frac{\frac{(n-1) \cdot s_n^2}{\sigma_n^2}}{n-1} / \frac{\frac{(m-1) \cdot s_m^2}{\sigma_m^2}}{m-1}} = \boxed{\frac{\chi_{n-1}^2}{n-1} / \frac{\chi_{m-1}^2}{m-1}}$$

$$\frac{(m-1) \cdot s_m^2}{\sigma_m^2} \sim \chi_{m-1}^2$$

$$F_{n-1, m-1}$$

Отношение дисперсий (независимые выборки)

Выборки не зависят друг от друга:

$$X_1, \dots, X_n \sim iid N(\mu_1, \sigma_1^2) \quad Y_1, \dots, Y_m \sim iid N(\mu_2, \sigma_2^2)$$

Нас интересует случайная величина:

$$\frac{s_n^2 \cdot \sigma_m^2}{s_m^2 \cdot \sigma_n^2} \sim F_{n-1, m-1}$$

Итоговый интервал:

$$\frac{s_m^2}{s_n^2} \cdot F_{n-1, m-1} \left(\frac{\alpha}{2} \right) \leq \frac{\sigma_m^2}{\sigma_n^2} \leq \frac{s_m^2}{s_n^2} \cdot F_{n-1, m-1} \left(1 - \frac{\alpha}{2} \right)$$

Пример

У Джордана есть две бумаги. Он хочет посмотреть, насколько сильно они различались по уровню риска за последние 10 лет, $s_A^2 = 0.05$, $s_B^2 = 0.04$

$$\frac{s_A^2}{s_B^2} \cdot F_{9,9}(0.025) \leq \frac{\sigma_m^2}{\sigma_n^2} \leq \frac{s_A^2}{s_B^2} \cdot F_{9,9}(0.975)$$

$$0.31 \leq \frac{\sigma_m^2}{\sigma_n^2} \leq 5$$

Резюме

Для того, чтобы посмотреть, насколько дисперсии двух независимых выборок различаются между собой, используется отношение дисперсий

Для нормальных выборок в этом помогают теорема Фишера и распределение Фишера