

Доверительные интервалы

План

- Что такое доверительный интервал
- Асимптотические доверительные интервалы
- Точные доверительные интервалы для нормальных выборок
- Как построить точный доверительный интервал для любого распределения

Обозначения

- Внимание: в этой презентации будут тонко использоваться греческие буквы с крышечкой и без крышечки. Это традиция в статистике
- Когда они **без** крышечки, речь идет о **параметрах** некоторого распределения (можно воспринимать их как неизвестные константы)

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



- Когда они **с** крышечкой, речь идет о некоторых **статистиках**, посчитанных по выборке из случайных величин, а значит тоже о случайных величинах с каким-то распределением. Этими статистиками мы будем оценивать **параметры**, которые обозначаются той же греческой буквой, но **с** крышечки.

$$\hat{\mu} = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right) \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$(n-1) \cdot \frac{\hat{\sigma}^2}{\sigma^2} = \frac{n-1}{\sigma^2} \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1)$$

Что такое доверительный интервал

Схема математической статистики

Выборка: X_1, \dots, X_n Параметр: θ



Точность
оценки,
прогнозов

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Союзники

Асимптотические
(при большом n)

- ЦПТ
- Дельта-метод

доверительные
интервалы

Ответы на
вопросы
проверка
гипотез

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Точные

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

Схема математической статистики

Выборка: X_1, \dots, X_n Параметр: θ



Точность
оценки,
прогнозов

доверительные
интервалы

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

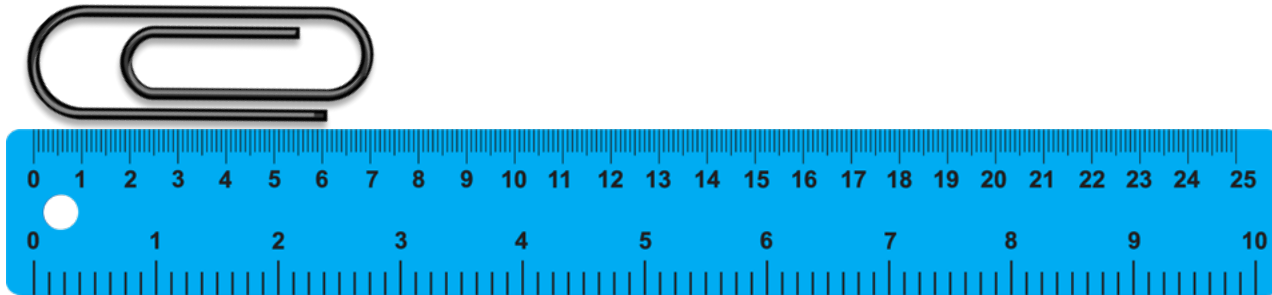
- Несмещенная
- Состоятельная
- Эффективная

Ответы на
вопросы

проверка
гипотез

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

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

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

- Доверительный интервал для μ :

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

❗ 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

Схема математической статистики

Выборка: X_1, \dots, X_n Параметр: θ

$$\hat{\theta}$$

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

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

$\hat{\theta}$



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



Точность
оценки,
прогнозов



доверительные
интервалы



Ответы на
вопросы
проверка
гипотез

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники


Асимптотические
(при большом n)

- ЦПТ
- Дельта-метод

Точные

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

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

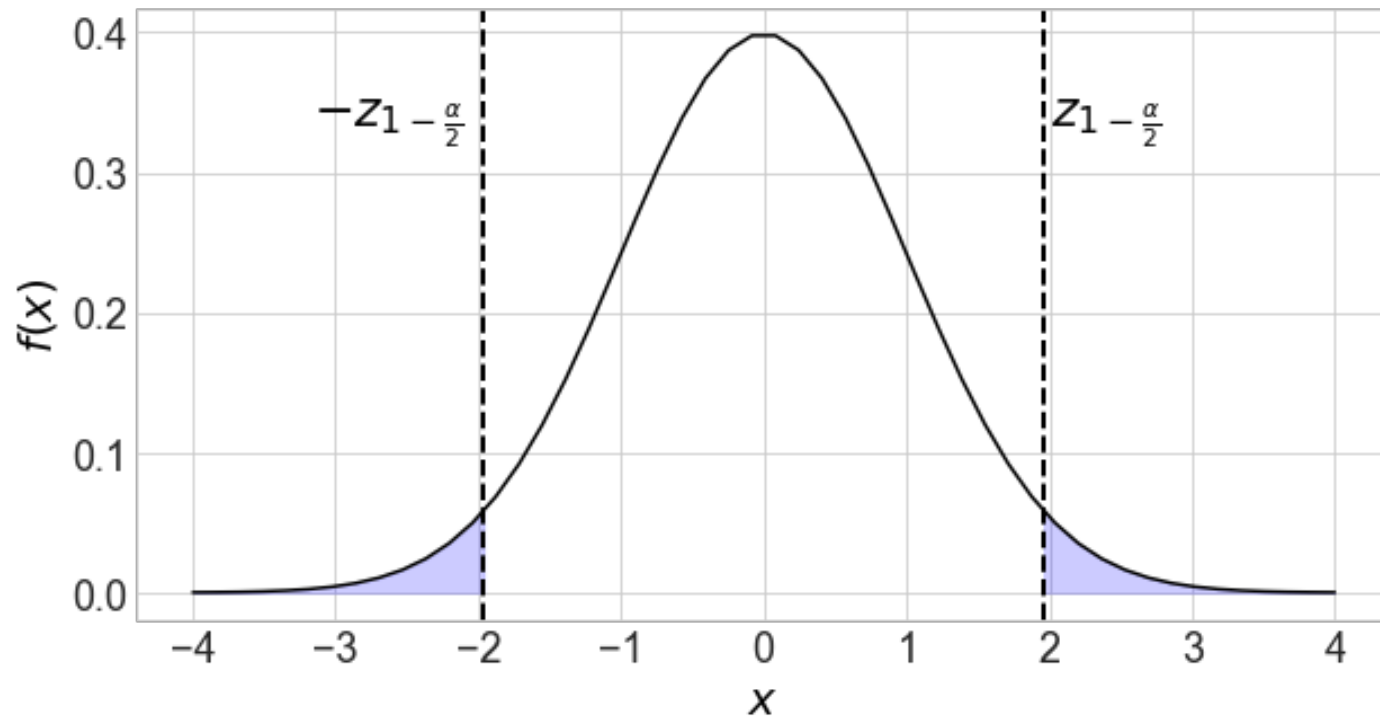
$$\bar{X} \stackrel{asy}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \bar{X} - \mu \stackrel{asy}{\sim} N\left(0, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \stackrel{asy}{\sim} N(0, 1)$$

центрирование стандартизация

Асимптотический интервал для среднего

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

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



Асимптотический интервал для среднего

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

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

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

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 , $\hat{\sigma}^2 \overset{p}{\rightarrow} \sigma^2$

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.

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

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 $\hat{\sigma}^2$ можно выразить через смещенную выборочную дисперсию \hat{s}^2 ,

а \hat{s}^2 – через средние

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

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

$$= \frac{n}{n-1} \cdot \hat{s}^2$$

$$= \frac{n}{n-1} (\overline{X^2} - \bar{X}^2)$$

► <https://www.stat.umn.edu/geyer/s06/5102/notes/ci.pdf>

Асимптотический интервал для дисперсии

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

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

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

Точные доверительные интервалы для нормальных выборок

Схема математической статистики

Выборка: x_1, \dots, x_n Параметр: θ



Точность
оценки,
прогнозов

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Союзники

Асимптотические
(при большом n)

- ЦПТ
- Дельта-метод

доверительные
интервалы

Ответы на
вопросы
проверка
гипотез

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Точные

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

Схема математической статистики

Выборка: X_1, \dots, X_n Параметр: θ



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

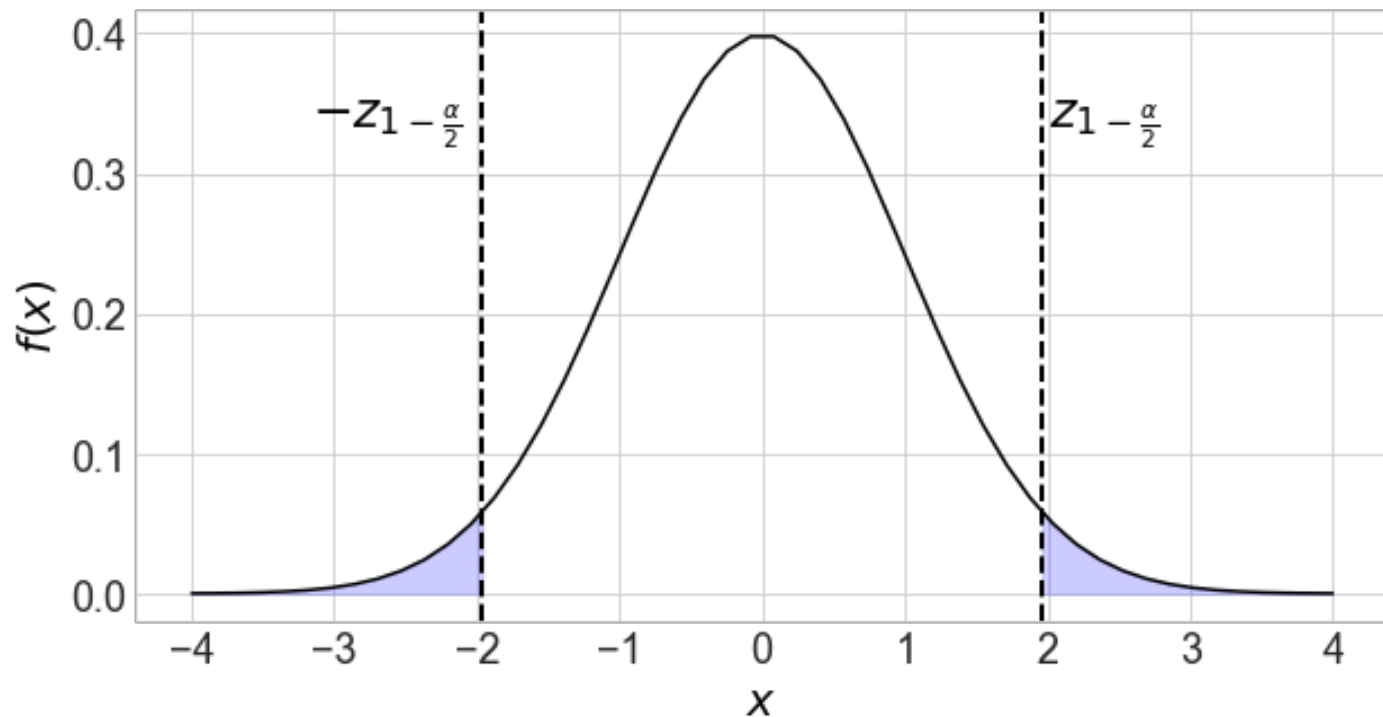
$$\hat{\mu} = \bar{X} = \frac{X_1 + \dots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Распределение точное, сумма нормальных случайных величин – нормальна.

Дисперсия известна

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

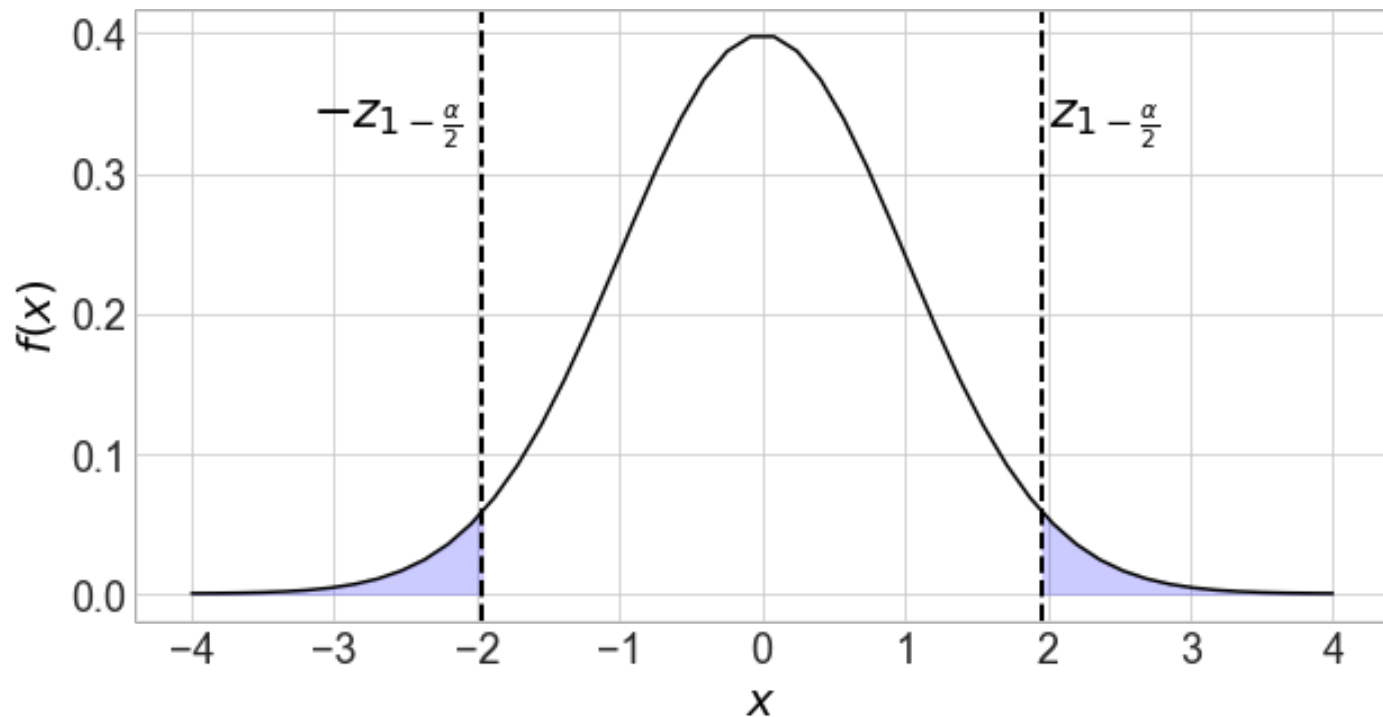
■ $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$



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



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

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

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

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 

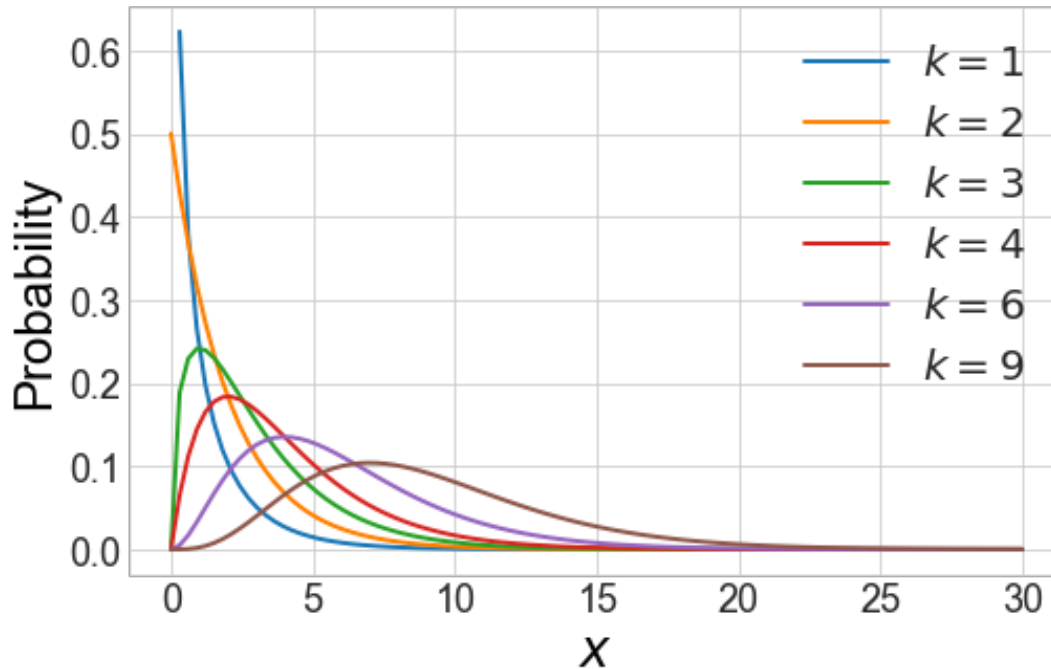


Когда возникает
на практике:

$$\hat{s}^2 = \overline{X^2} - \bar{X}^2$$


- 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:

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

$$\text{Var}(Y) = 2k$$

Союзники: распределение Стюдента

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

Тогда случайная величина

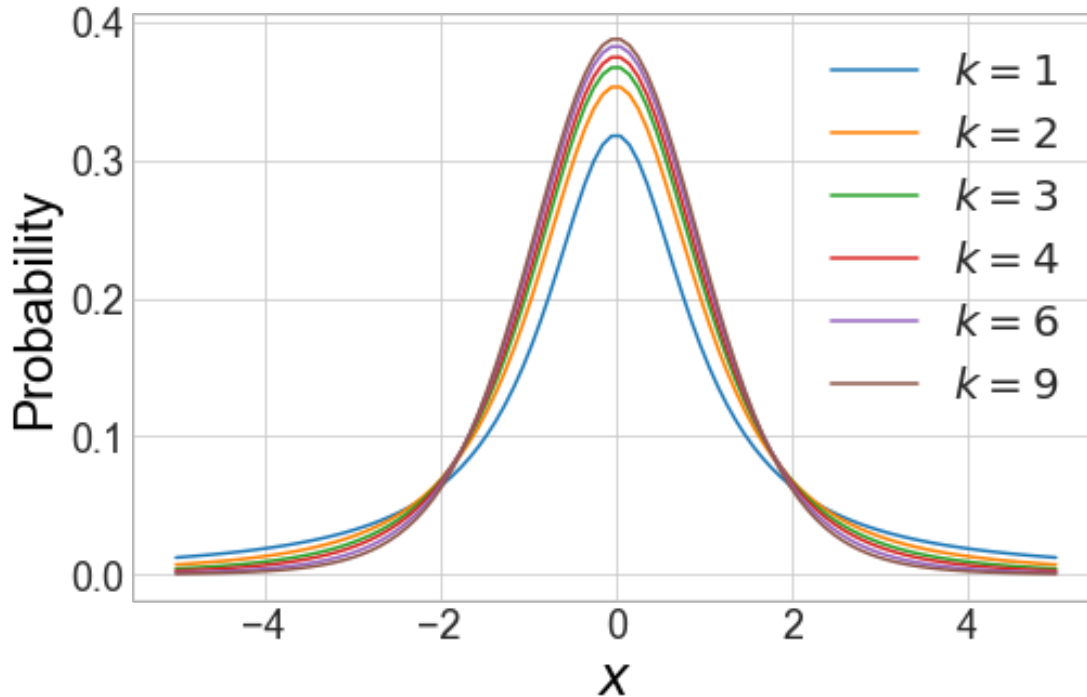
$$t = \frac{X_0}{\sqrt{Y/k}} \sim t(k) \quad \square$$

имеет распределение Стюдента 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 Стюдента

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:

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

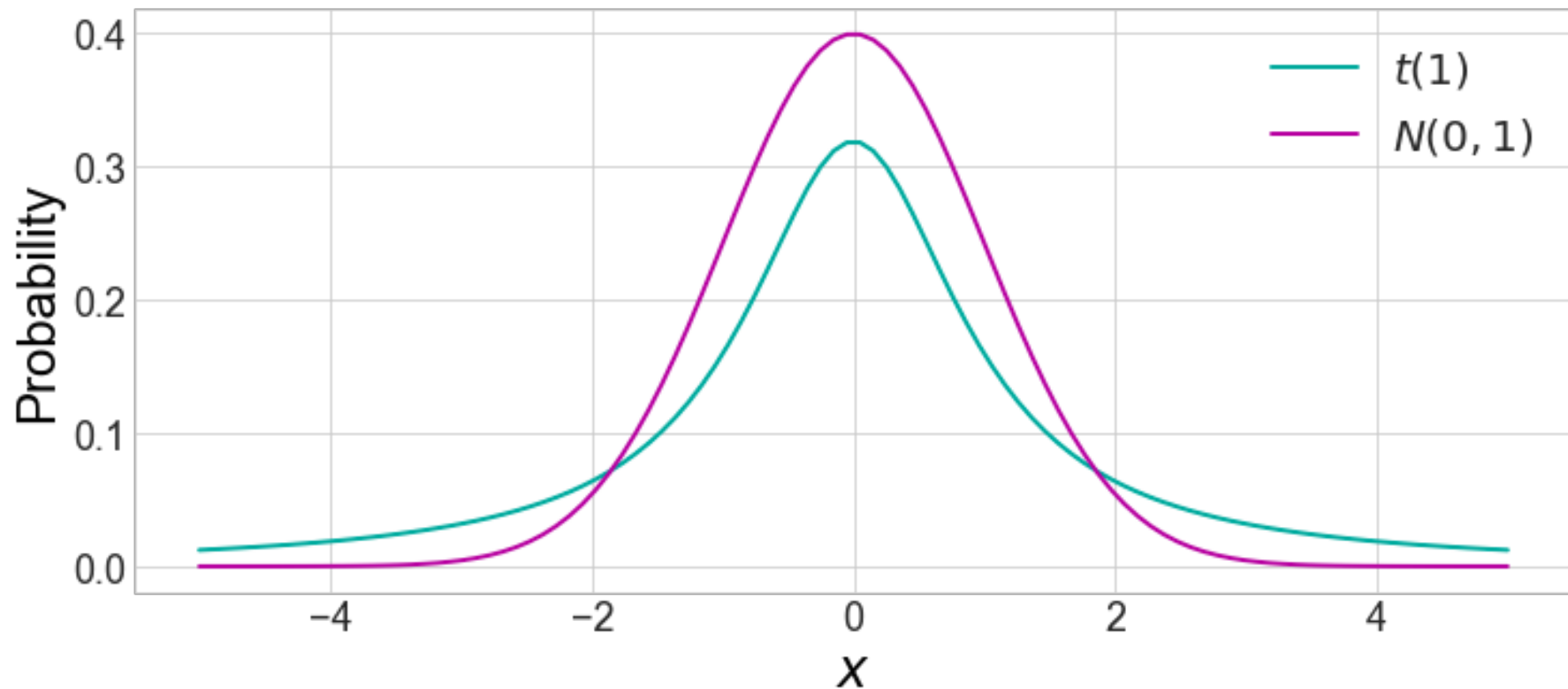
Характеристики:

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

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

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 $\hat{\sigma}^2$ независимы
2. $\frac{(n-1) \cdot \hat{\sigma}^2}{\sigma^2}$ имеет χ^2 – распределение с $n - 1$ степенью свободы 

Дисперсия неизвестна

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

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



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

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

$N(0, 1)$?



Дисперсия неизвестна

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

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

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

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

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

Дисперсия неизвестна

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

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

$N(0, 1)$

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

Дисперсия неизвестна

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

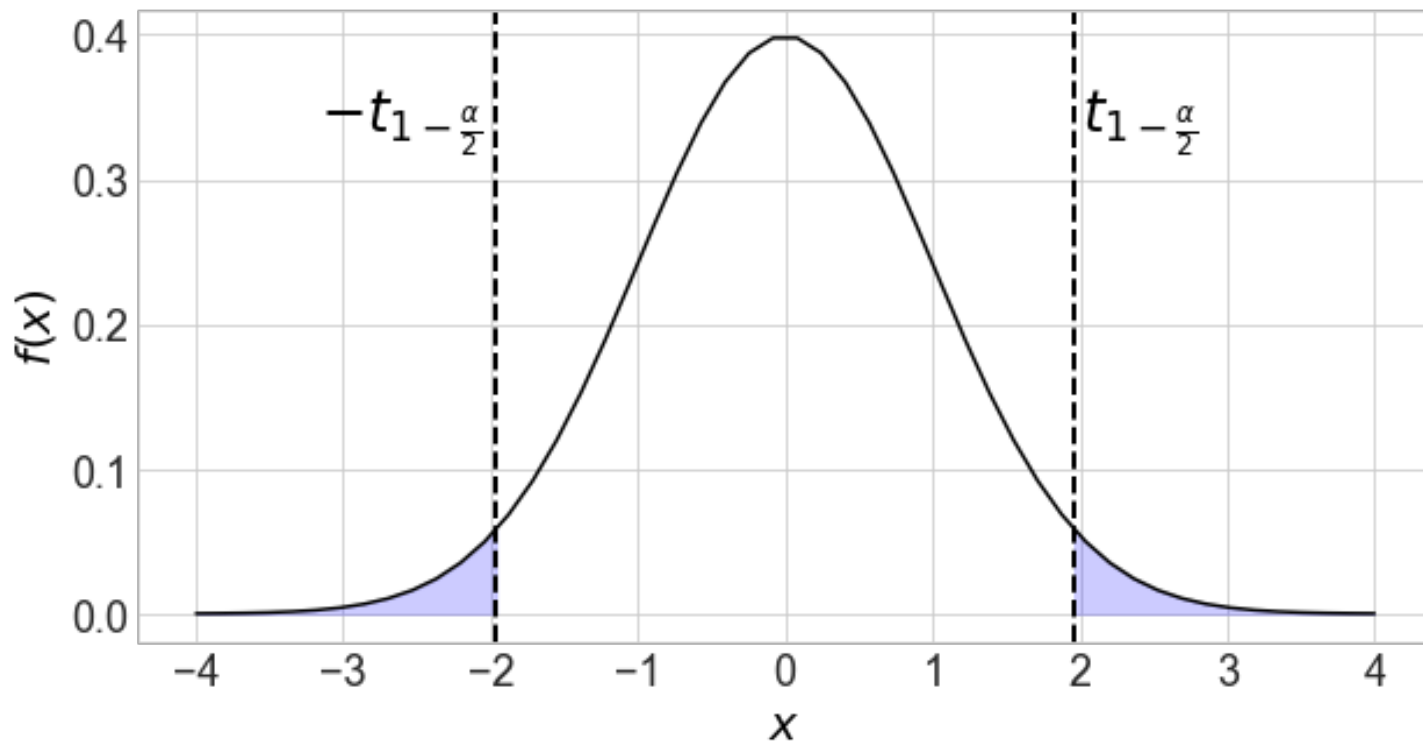
$$\frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{\sigma^2}{(n-1)}}} = \boxed{\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \cdot \frac{\sqrt{\frac{\sigma^2}{(n-1)}}}{\sqrt{\frac{\hat{\sigma}^2}{(n-1)}}}$$

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

Дисперсия неизвестна

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

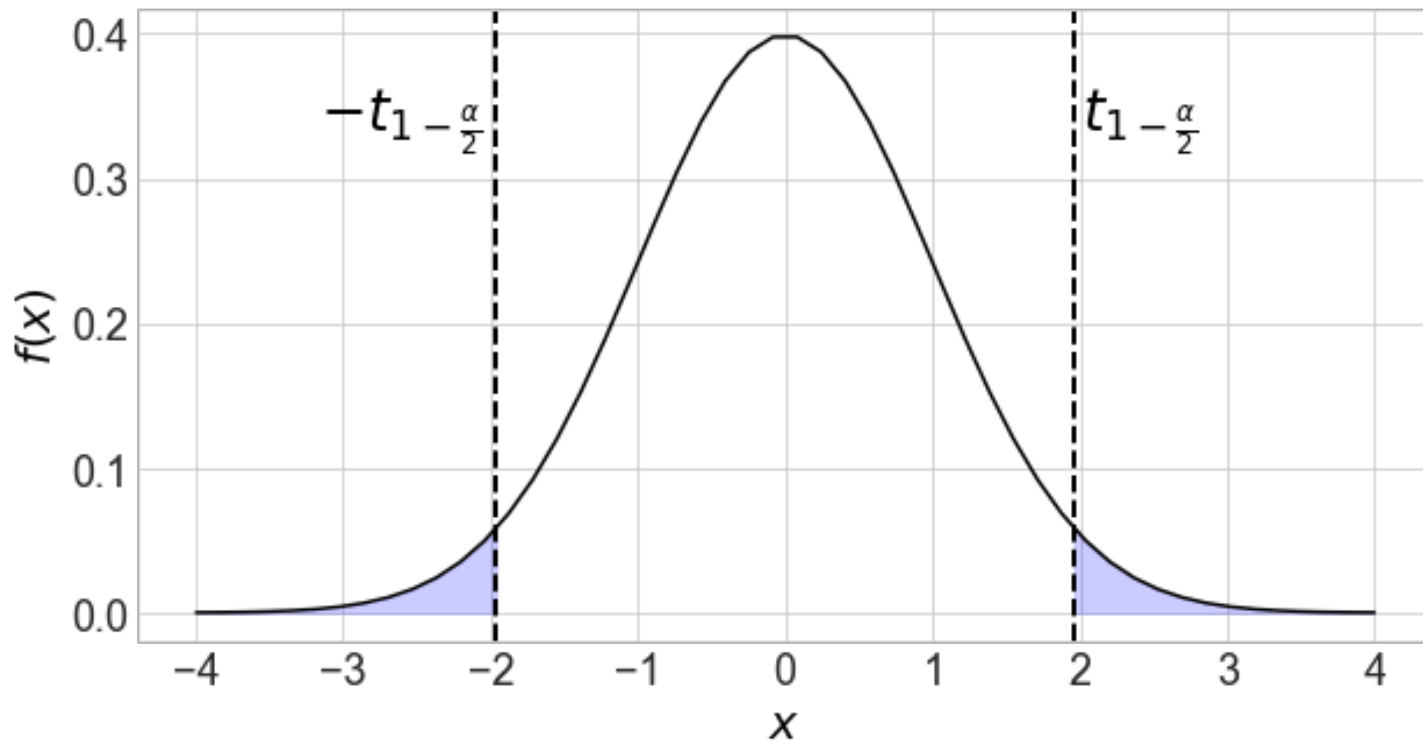
$$\frac{\bar{X} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \sim t(n-1) \quad \text{■}$$



Дисперсия неизвестна

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

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



Точный vs Асимптотический

Асимптотический

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

Точный

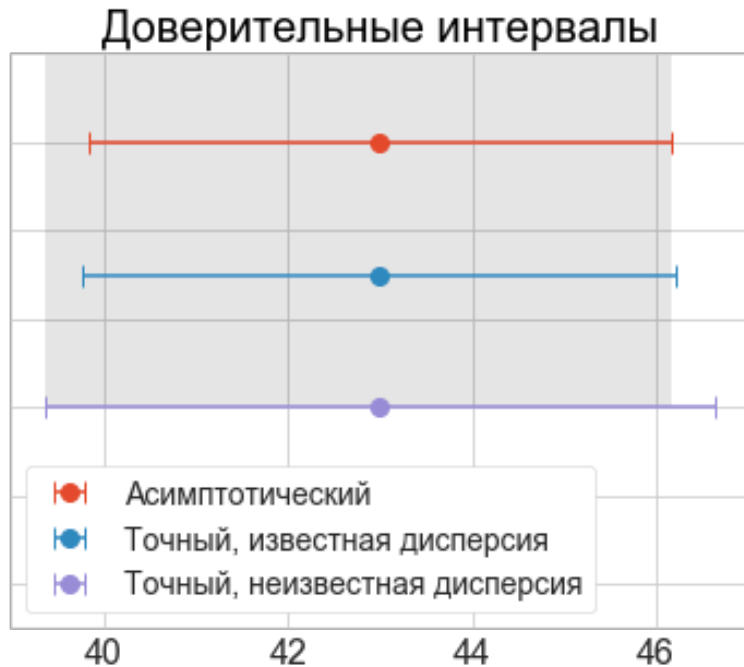
- Союзники: теорема Фишера, t -распределение
- Работает при любом n
- Выборка независимая из нормального распределения

Пример

Измерили зарплаты: $\bar{x} = 43$ тыс. и $\hat{\sigma} = 5.1$ тыс.

В выборку попало $n = 10$ наблюдений.

В реальности $\sigma = 5.2$ тыс. (знаем из переписи населения)



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

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

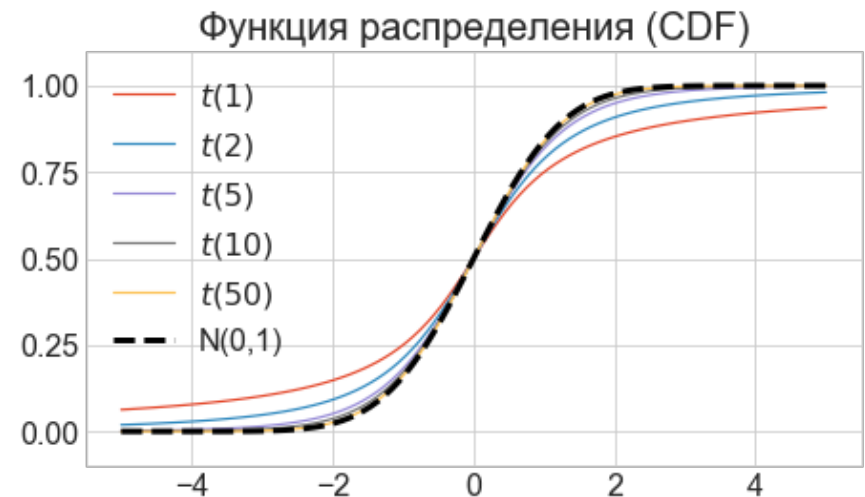
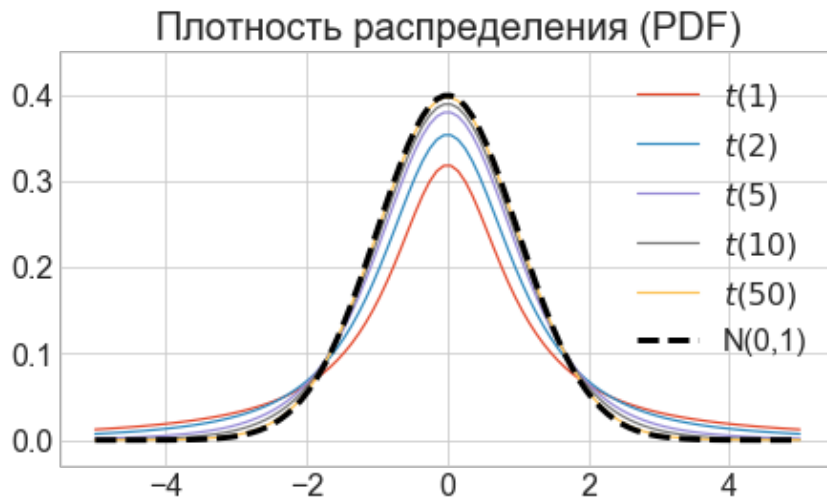
$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad 43 \pm 2.26 \cdot \frac{5.1}{\sqrt{10}}$$

- ✔ 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

$$\bar{X} \stackrel{asy}{\sim} N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right) \quad \bar{Y} \stackrel{asy}{\sim} N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$$

$$\bar{X} - \bar{Y} \stackrel{asy}{\sim} N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

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

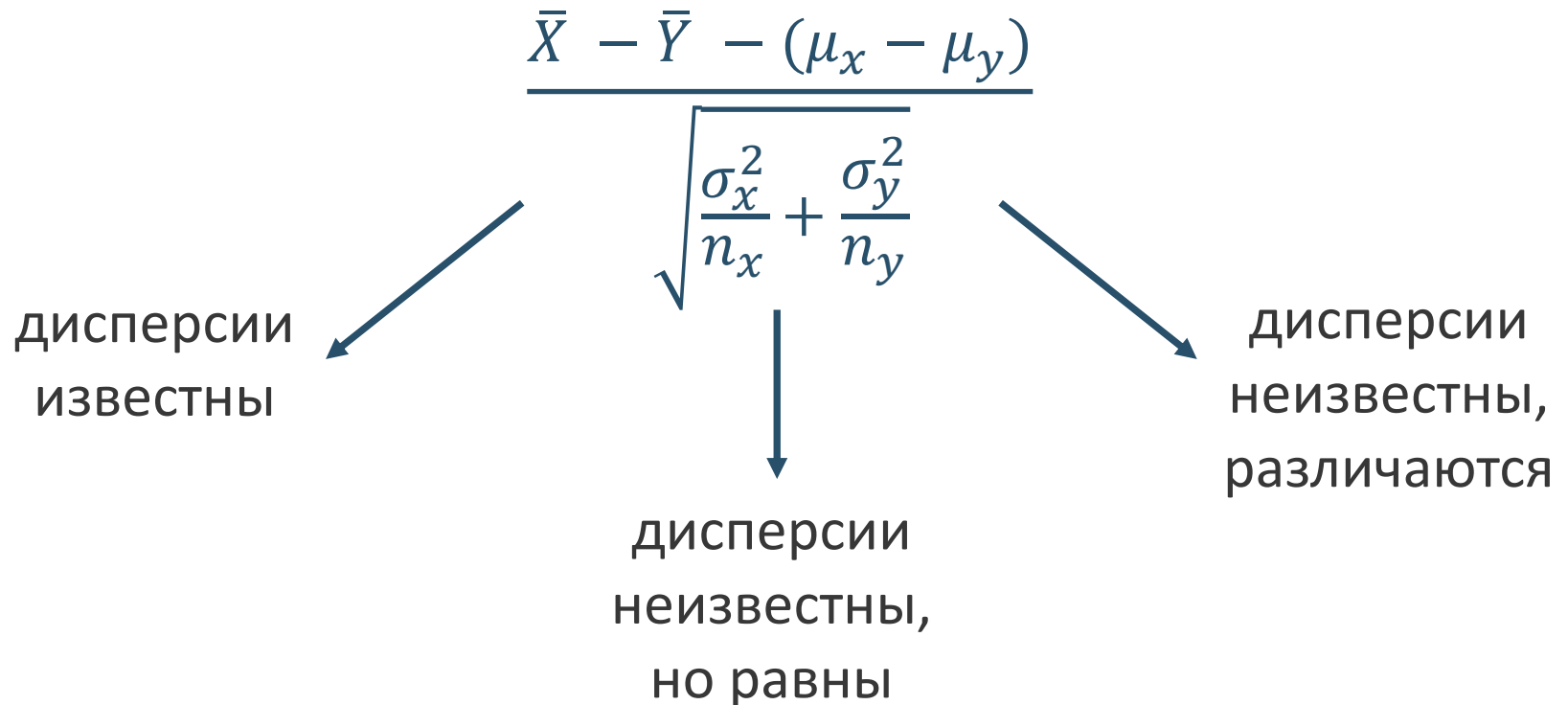
$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \stackrel{asy}{\sim} N(0,1)$$
$$(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

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:



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:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}^2}{n_x} + \frac{\hat{\sigma}^2}{n_y}}}$$



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:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_{pooled}^2}{n_x} + \frac{\hat{\sigma}_{pooled}^2}{n_y}}} \sim t(n_x + n_y - 2)$$



Объединённая оценка
дисперсии:


$$\hat{\sigma}_{pooled}^2 = \frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}$$

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:


$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \sim t(v)$$

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:

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}} \sim t(v)$$



Распределение
приближенное
(распределение
Уэлча)



$$v = \frac{\left(\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y} \right)^2}{\frac{\hat{\sigma}_x^4}{n_x^2(n_x - 1)} + \frac{\hat{\sigma}_y^4}{n_y^2(n_y - 1)}}$$

Проблема Беренца-Фишера

Не существует точного распределения для статистики



$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

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

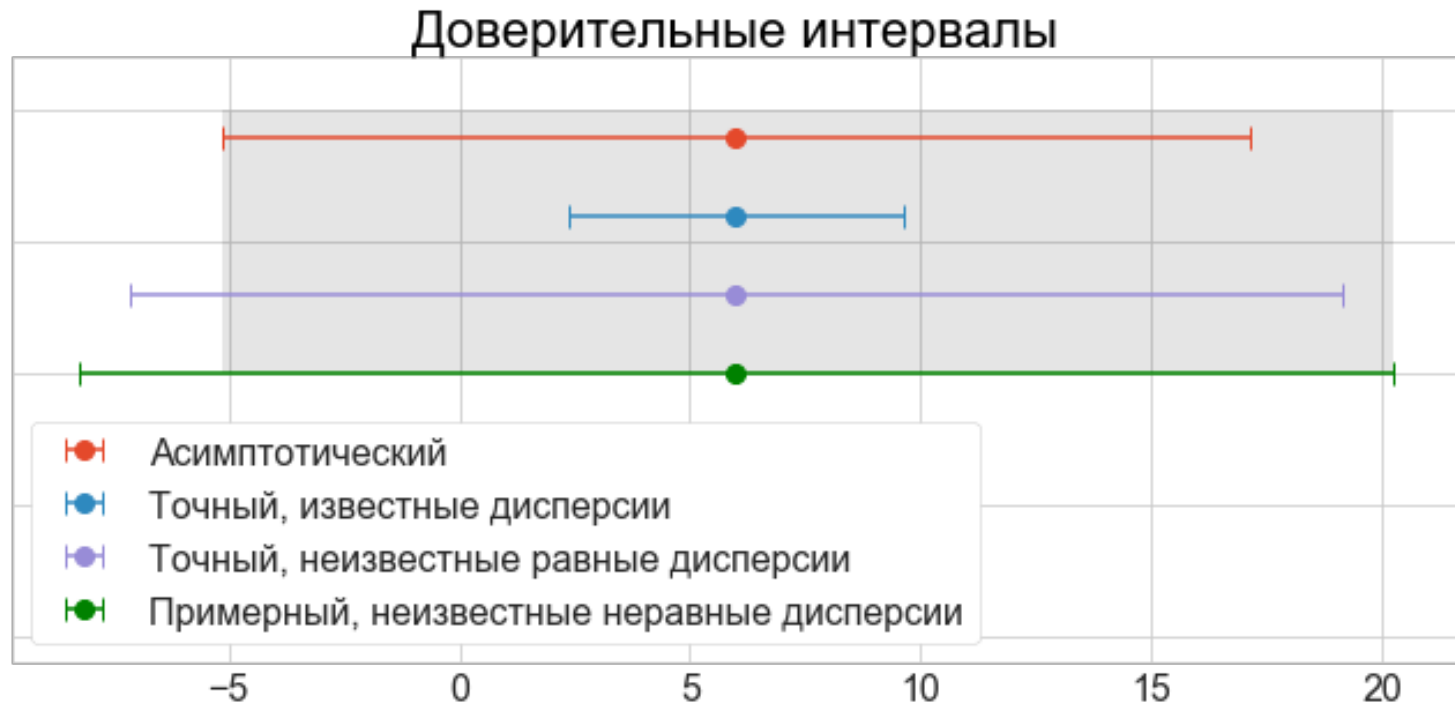
Аппроксимация с предыдущего слайда хорошо работает, если $n_x = n_y$ либо знак неравенства между n_x и n_y такой же как между σ_x и σ_y

Пример 1

Измерили зарплаты мужчин и женщин в тысячах рублей: $\bar{x} = 43$, $\hat{\sigma}_x = 5.1$, $\bar{y} = 37$, $\hat{\sigma}_y = 11.7$.

В обеих выборках было по 10 наблюдений.

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



Пример 1

Неизвестны (асимптотика):

$$\bar{X} - \bar{Y} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

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

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):

$$\bar{X} - \bar{Y} \pm t(n_x + n_y - 2)_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n_x} + \frac{\hat{\sigma}^2}{n_y}}$$

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

Неизвестны, не равны (примерный):

$$\bar{X} - \bar{Y} \pm t(v)_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}$$

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

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

$$d_i = X_i - Y_i$$

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

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

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

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

Точный, неизвестная дисперсия:

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

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 $\hat{\sigma}^2$ независимы
2. $\frac{(n-1) \cdot \hat{\sigma}^2}{\sigma^2}$ имеет χ^2 – распределение с $n - 1$ степенью свободы

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

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

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



Надо как-то привести к χ_n^2

Математическое ожидание известно

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

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

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

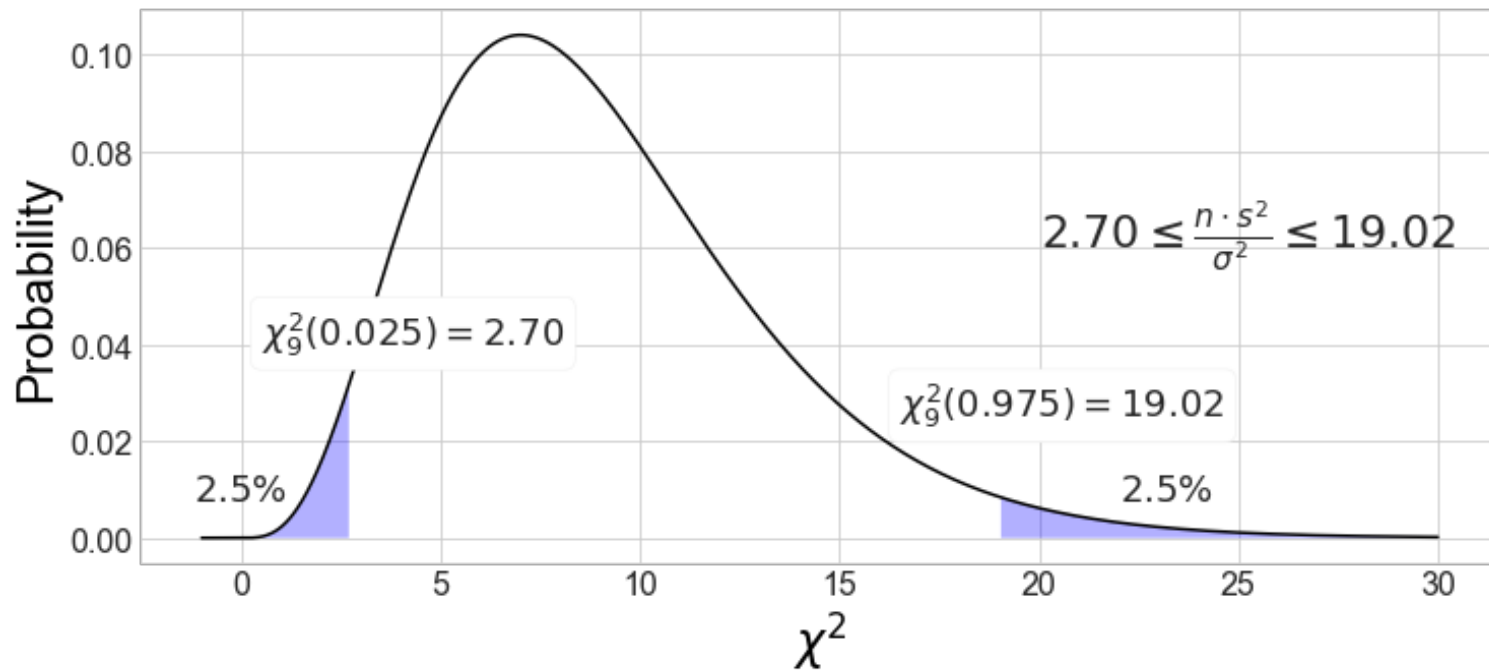
$$\frac{n \cdot \hat{s}^2}{\sigma^2} = \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

Математическое ожидание известно

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



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



Математическое ожидание известно

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

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

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

Математическое ожидание неизвестно

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

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

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



Оценка ломает всю логику
Нужен новый союзник

Математическое ожидание неизвестно

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

Теорема Фишера:

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

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:

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

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

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

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 $\hat{\sigma}^2 = 0.05$.

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

$$\frac{(10 - 1) \cdot 0.05}{19.02} \leq \sigma^2 \leq \frac{(10 - 1) \cdot 0.05}{2.70}$$

$$0.023 \leq \sigma^2 \leq 0.166 \quad \text{■}$$

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 $\hat{\sigma}^2 = 0.05$.

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

$$\frac{10 \cdot 0.04}{\chi_{10}^2(0.975)} \leq \sigma^2 \leq \frac{10 \cdot 0.04}{\chi_{10}^2(0.025)}$$

$$\frac{10 \cdot 0.04}{20.48} \leq \sigma^2 \leq \frac{10 \cdot 0.04}{3.24}$$

$$0.017 \leq \sigma^2 \leq 0.111$$

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:

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

Нас интересует случайная величина:

$$\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \sim ?$$

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

Распределение Фишера

Независимые случайные величины χ_k^2 и χ_m^2 .

Случайная величина

$$F = \frac{\chi_k^2/k}{\chi_m^2/m} \sim F(k, m)$$

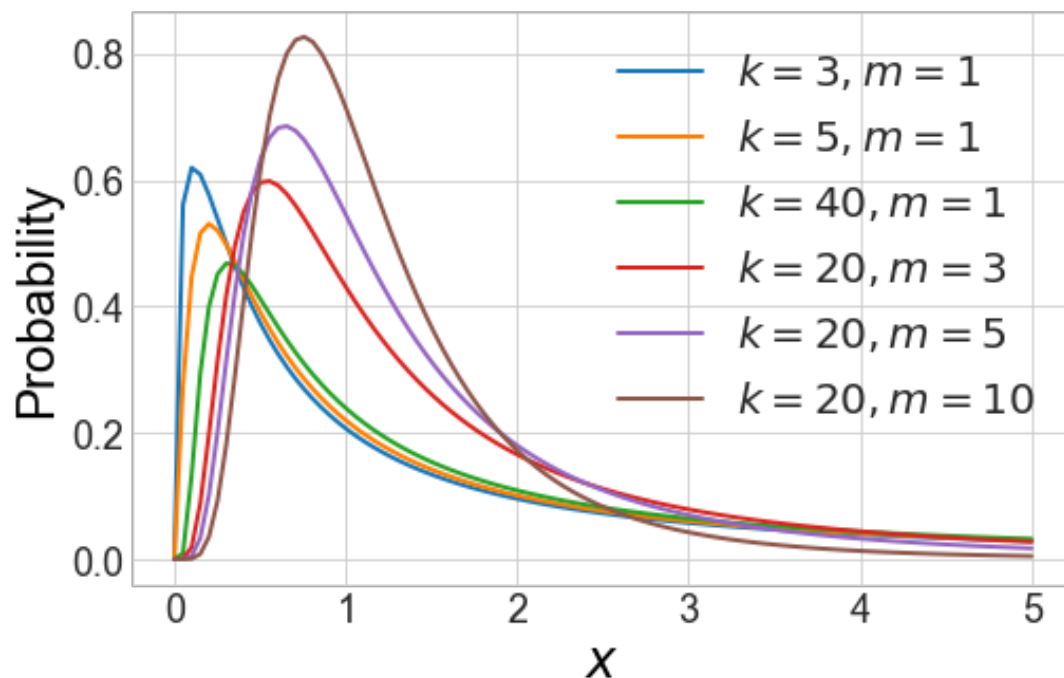
имеет распределение Фишера с k, m степенями свободы.



Когда возникает на практике:

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

Распределение Фишера



$$F = \frac{\chi_k^2/k}{\chi_m^2/m} \sim F(k, m)$$

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

Отношение дисперсий (независимые выборки)

Выборки не зависят друг от друга:

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

Теорема Фишера:

$$\frac{(n_x - 1) \cdot \hat{\sigma}_x^2}{\sigma_x^2} \sim \chi_{n_x-1}^2 \quad \frac{(n_y - 1) \cdot \hat{\sigma}_y^2}{\sigma_y^2} \sim \chi_{n_y-1}^2$$
$$\frac{\frac{(n_x-1) \cdot \hat{\sigma}_x^2}{\sigma_x^2}}{\frac{(n_y-1) \cdot \hat{\sigma}_y^2}{\sigma_y^2}} = \frac{\frac{\chi_{n_x-1}^2}{n_x-1}}{\frac{\chi_{n_y-1}^2}{n_y-1}} \sim F_{n_x-1, n_y-1}$$

Распределение Фишера

Сократим число степеней свободы

$$\frac{\hat{\sigma}_x^2 / \sigma_x^2}{\hat{\sigma}_y^2 / \sigma_y^2} \sim F_{n_x-1, n_y-1}$$

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:

$$\frac{\hat{\sigma}_x^2 / \sigma_x^2}{\hat{\sigma}_y^2 / \sigma_y^2} \sim F_{n_x-1, n_y-1}$$

Итоговый интервал:

$$\frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \cdot F_{n_x-1, n_y-1} \left(\frac{\alpha}{2} \right) \leq \frac{\sigma_x^2}{\sigma_y^2} \leq \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2} \cdot F_{n_x-1, n_y-1} \left(1 - \frac{\alpha}{2} \right)$$

Пример

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

$$\frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} \cdot F_{9,9}(0.025) \leq \frac{\sigma_A^2}{\sigma_B^2} \leq \frac{\hat{\sigma}_A^2}{\hat{\sigma}_B^2} \cdot F_{9,9}(0.975)$$

$$0.31 \leq \frac{\sigma_A^2}{\sigma_B^2} \leq 5 \quad \text{■}$$

В интервал попала единица, неясно какая дисперсия больше

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