Доверительные интервалы

План

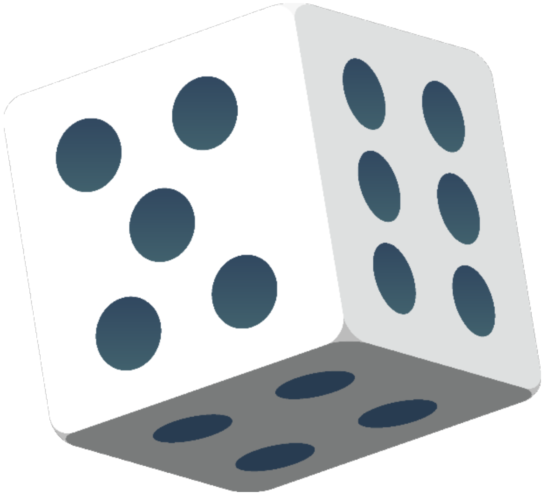
- Закон больших чисел
- Центральная предельная теорема
- Асимптотические доверительные интервалы
  - для мат. ожидания
  - для доли

Закон больших чисел (ЗБЧ)

Закон больших чисел (ЗБЧ)

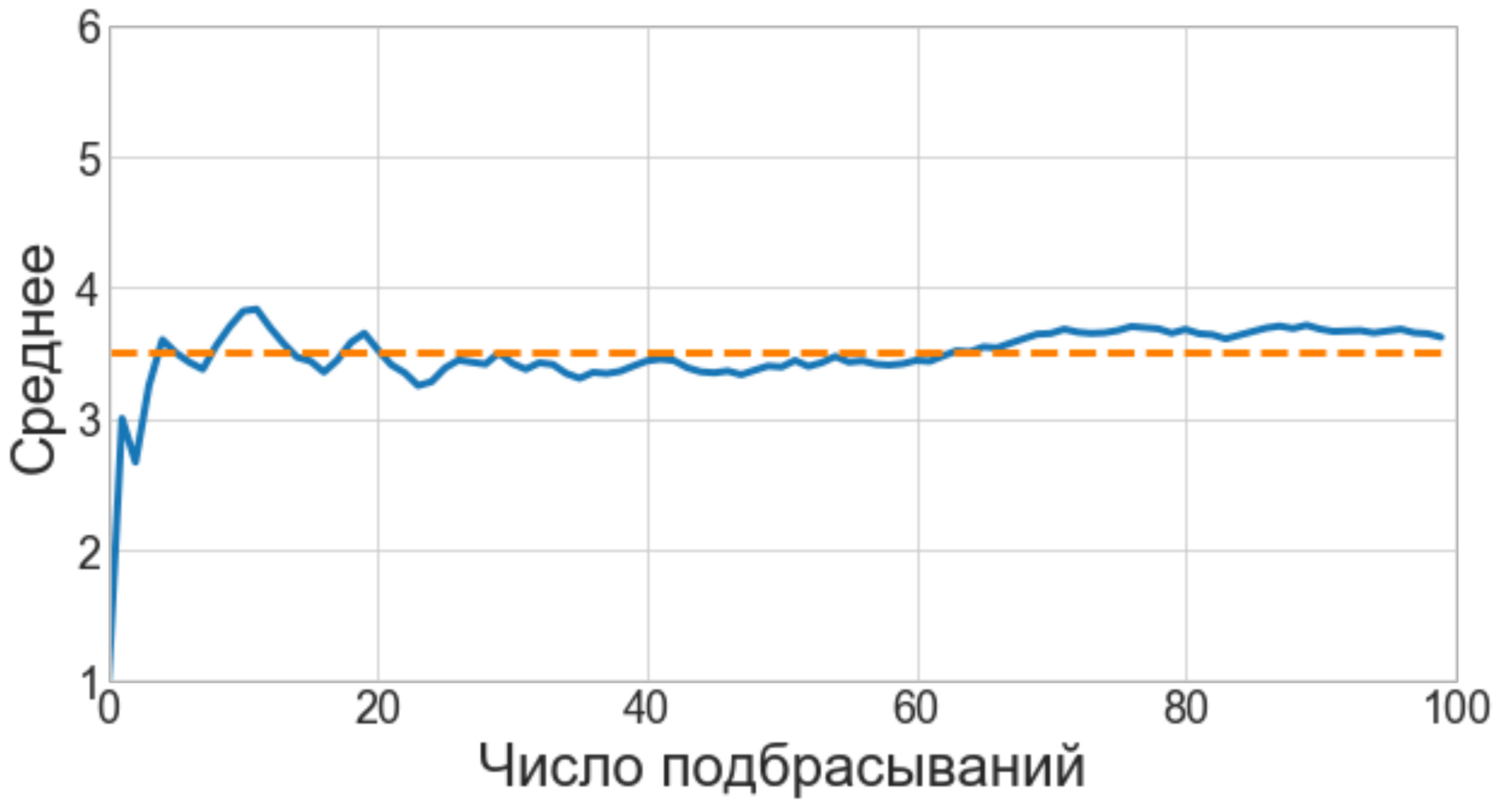
ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа

Пример: Игральная кость



Закон больших чисел (ЗБЧ)

ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа



Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть $X_1, \dots, X_n$ попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $Var(X_1) < \infty$ тогда:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

Среднее сходится по вероятности к математическому ожиданию при $n \to \infty$

Слабая форма ЗБЧ (Чебышёв)

Простым языком:

- Среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа
- Среднее для бесконечного числа случайных величин неслучайно
- Если у нас есть страховая фирма, мы можем заработать немного денег (самая простая формулировка)

Страховка

Вероятность того, что на машину во дворе упадёт дерево составляет 0.01. Страховка в год стоит 100 рублей. В случае падения клиенту выплачивается 11000 рублей. Какой будет средняя прибыль компании с одной страховки?

X_i – прибыль с одного человека

$\bar{X}$ – средняя прибыль компании

X_i	100	-10900
$\mathbb{P}(X_i = k)$	0.99	0.01

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1) = 100 \cdot 0.99 - 10900 \cdot 0.01 = -10$$

Вопрос про больницы

- Есть две больницы: большая и маленькая.
- В обеих принимают роды. Выяснилось, что в одной из них оценка вероятности появления мальчика составила 0.7.
- В какой больнице это скорее всего произошло и почему?

Вопрос про больницы

Скорее всего это произошло в маленькой больнице. При малых объёмах выборки вероятность отклониться от 0.5 больше. Именно об этом говорит нам ЗБЧ.

depositphotos.com

Некорректная работа при малых числах

- Данные часто поступают на обработку в агрегированной форме (по городам, по людям, по статьям из газет)
- Для субъектов с маленьким числом наблюдений ЗБЧ не работает (города с маленьким населением)
- Среднее значение при маленьких выборках плохо отражает фактическое математическое ожидание

➤ *http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf*

Резюме

ЗБЧ говорит, что при больших выборках и отсутствии аномалий среднее, рассчитанное по выборке, оказывается близким к теоретическому математическому ожиданию

Сходимость по вероятности

Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть $X_1, \dots, X_n$ попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $Var(X_1) < \infty$ тогда:

$$\bar{X} = \frac{X_1 + \ \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

Среднее сходится по вероятности к математическому ожиданию при $n \to \infty$

Сходимость по вероятности

Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ **сходится по вероятности** к случайной величине X, если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \to 1 \text{ при } n \to \infty$$

Сходимость по вероятности

Последовательность случайных величин $X_1, \ldots, X_n, \ldots$ **сходится по вероятности** к случайной величине X, если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \to 1 \text{ при } n \to \infty$$

То есть:

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

Обычно пишут:

$$X_n \xrightarrow{p} X \text{ при } n \to \infty \quad \text{либо} \quad \operatorname*{plim}_{n \to \infty} X_n = X$$

Резюме

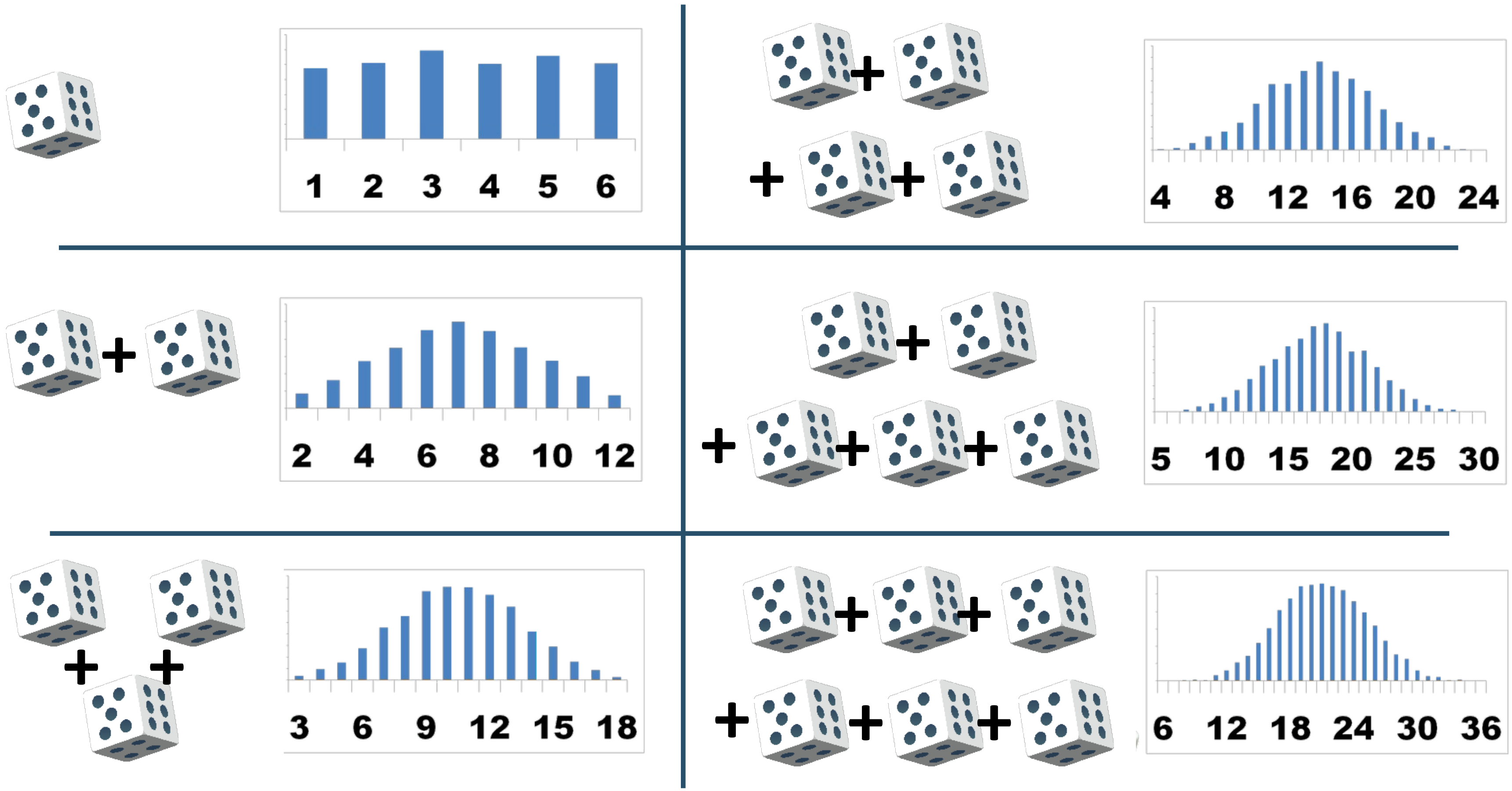
В слабой форме ЗБЧ среднее сходится к математическому ожиданию по вероятности

Для сходимости по вероятности верны такие же арифметический свойства, как и для обычных пределов

Центральная предельная теорема (ЦПТ)

Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



Центральная предельная теорема

Теорема:

Пусть $X_1, \ldots, X_n$ попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\mathrm{Var}(X_1) < \infty$ тогда:

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$

Иногда пишут: либо:

$$\frac{\bar{X}_n - \mathbb{E}(X_1)}{\sqrt{\frac{Var(X_1)}{n}}} \xrightarrow{d} N(0,1) \qquad \sqrt{n} \cdot \frac{\bar{X}_n - \mathbb{E}(X_1)}{sd(X_1)} \xrightarrow{d} N(0,1)$$

Сходимость по распределению

Центральная предельная теорема

Теорема:

Пусть $X_1, \ldots, X_n$ попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\mathrm{Var}(X_1) < \infty$ тогда:

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$

✔ Буква d над стрелкой означает сходимость по распределению

Сходимость по распределению

Последовательность случайных величин $X_1, \dots, X_n, \dots$ **сходится по распределению** к случайной величине X, если

$$\lim_{n\to\infty} \mathrm{F}_{X_n}(x) = F_X(x),$$

то есть последовательность функций распределения $\mathrm{F}_{X_n}(x)$ сходится к функции $\mathrm{F}_X(x)$ во всех точках x, где $\mathrm{F}_X(x)$ непрерывна.

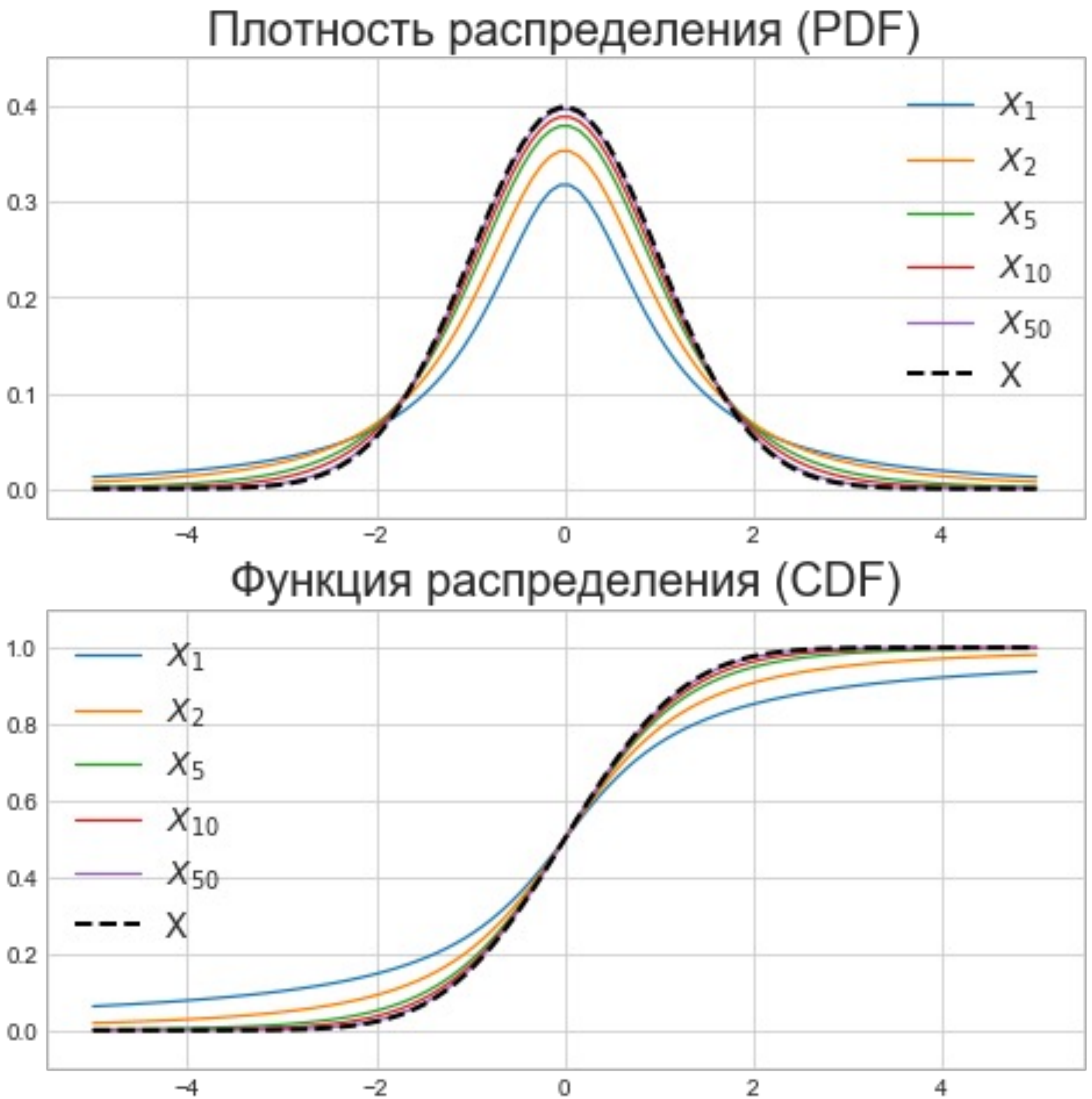
Обычно пишут:

$$X_n \xrightarrow{d} X \text{ при } n \to \infty$$

либо:

$$X_n \xrightarrow{F} X \text{ при } n \to \infty$$

Сходимость по распределению



Центральная предельная теорема

Простым языком:

- Сумма достаточно большого числа случайных величин имеет распределение близкое к нормальному
- Есть очень большое количество формулировок ЦПТ с разными условиями
- Главное, чтобы случайные величины были похожи друг на друга и не было такого, что одна из них резко выделяется на фоне остальных

Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

X_2 – готовил завтрак, убежало молоко, задержался убрать

X_3 – автобус приехал пораньше

X_4 – из-за аварии попали в пробку

…

$X =$  $+$  $+$ 

$+$  $+ \ldots$

X — время прихода Миши на первую пару

hotpng.com, seekpng.com

Центральная предельная теорема

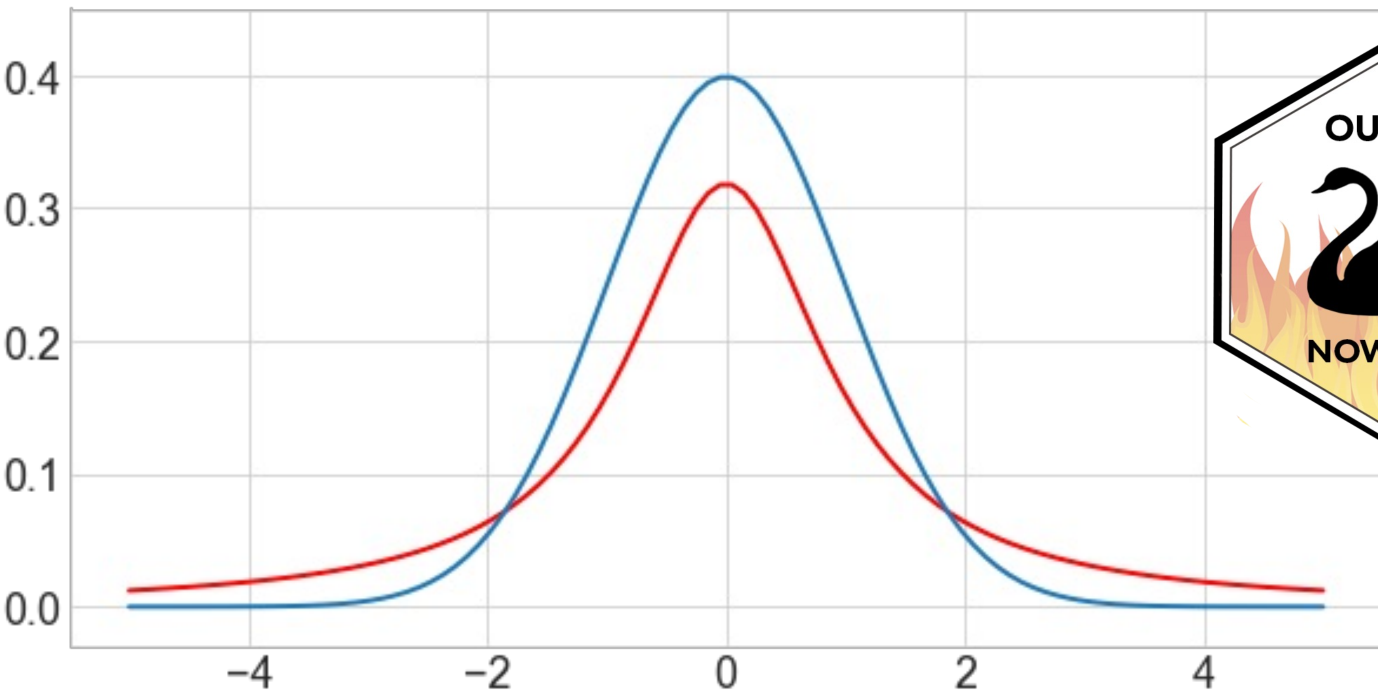
- X – время прихода Миши на первую пару
- Распределение близко к нормальному
- Если одна из случайных величин резко выделяется на фоне остальных, нормальность ломается, появляются **тяжёлые хвосты**

Крайнеземье и средиземье

ЦПТ и ЗБЧ работают в Средиземье

freepngimg.com

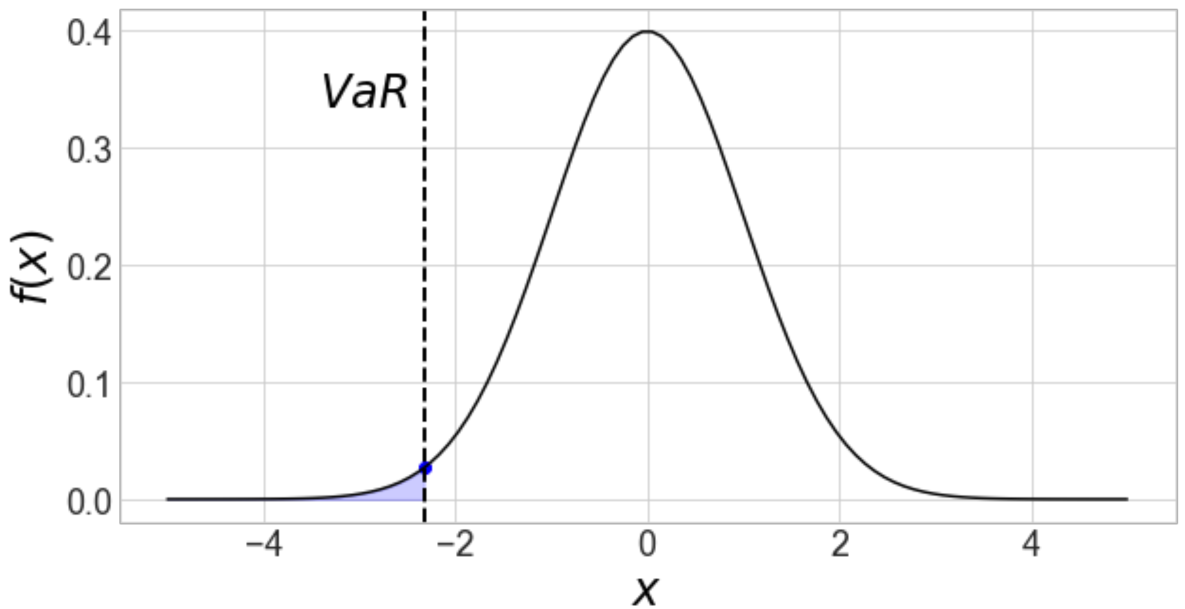
Крайнеземье и средиземье



- Хвосты красного распределения тяжёлые
- Под ними сосредоточена большая вероятностная масса
- Статистика недооценивает тяжесть хвостов из-за того, события из них встречаются редко

Тяжёлые хвосты и финансы

- Важно понимать, сколько денег мы потеряем в самом плохом случае
- Пытаются смоделировать 5% квантиль распределения доходностей, VaR – Value at risk.
- Не нужно уметь хорошо моделировать всё распределение доходностей, достаточно уметь моделировать левый хвост



Тяжёлые хвосты и финансы

- Распределение доходностей чаще всего отличается от нормального, его хвосты оказываются тяжёлыми
- Сложно набрать достаточное количество статистики, чтобы адекватно оценить с какой вероятностью произойдёт катастрофа (катастрофы очень редки)
- Оценки всегда занижены
- Нужны специальные методы для работы с Крайнеземьем и тяжёлыми хвостами

ЗБЧ vs ЦПТ (две теоремы о среднем)

ЗБЧ: $$\frac{X_1 + \ \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

ЦПТ: $$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$

ЗБЧ: одно среднее, посчитанное по выборке размера n. При росте n среднее стабилизируется около математического ожидания

ЦПТ: много средних, посчитанных по разным выборкам размера n. При росте n распределение всё больше похоже на нормальное, оно всё компактнее вокруг математического ожидания

Резюме

ЦПТ говорит, что при больших выборках и отсутствии аномалий мы можем аппроксимировать распределение среднего нормальным распределением

В случае, если какие-то случайные величины сильно выделяются на фоне остальных, мы имеем дело с тяжёлыми хвостами

Тяжёлые хвосты часто встречаются в финансах и требуют к себе отдельного статистического подхода

Схема математической статистики

Выборка: $X_1, \ldots, X_n$ Параметр: θ

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

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
- $\chi_n^2, t_n, F_{n,k}$
- Ещё союзники!

Метод моментов

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

- $\boxed{\text{Метод моментов}}$
- 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!

Параметрическое оценивание

$X_1, \dots, X_n$ одинаково независимо распределены $(\ iid\ )$

Строя различные модели, мы будем иногда предполагать, что выборка имеет некоторое **определенное** распределение

Любое распределение характеризуется некоторыми **параметрами**, которые мы **не знаем**

Простейший способ **оценки** неизвестных параметров – это метод моментов

Метод моментов

$X_1, \dots, X_n$ одинаково независимо распределены $(\ iid\ )$

Момент $\mathbb{E}\left(X_i^k\right)$ зависит от неизвестного параметра θ:

$$\mathbb{E}\left(X_i^k\right) = f(\theta)$$

Теоретический момент должен совпадать с выборочным

$$\mathbb{E}\left(X_i^k\right) = f(\theta) \approx \overline{X^k} = \frac{\sum x_i^k}{n}$$

Решим уравнение и получим **оценку метода моментов**

$$\hat{\theta}_{MM} = f^{-1}(\overline{X^k})$$

Чаще всего хватает первого момента и берут $k = 1$,
то есть решают уравнение:

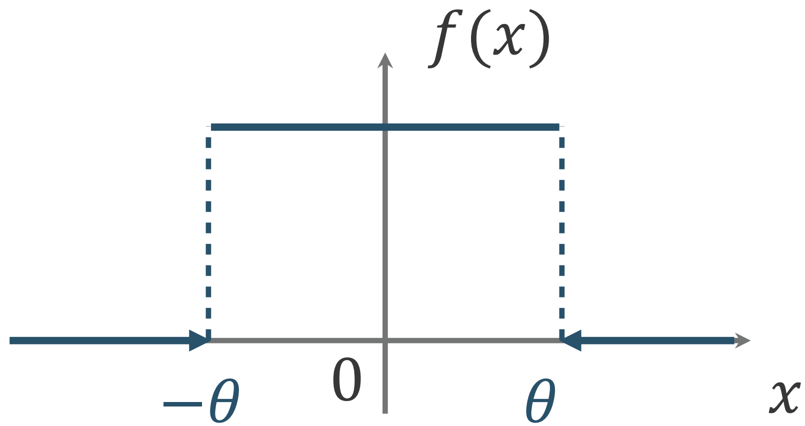
$$\mathbb{E}(X_i) \approx \frac{\sum x_i}{n}$$

Метод моментов

Если оказывается, что $\mathbb{E}(X_i) = 0$, тогда используют моменты более высоких порядков:

$$X_1, \dots, X_n \sim iid\ U[-\theta; \theta]$$

$$\mathbb{E}(X_i) = 0 \Rightarrow \bar{x} = 0$$



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

❗ Используя первый момент, нельзя получить оценку

$$\mathbb{E}(X_i^2) = \frac{\theta^2}{3} \Rightarrow \overline{x^2} = \frac{\theta^2}{3}$$

$$\Rightarrow \hat{\theta}_{MM} = \left(3\overline{x^2}\right)^{0.5}$$

Метод моментов

Если у распределения несколько параметров, используют несколько моментов:

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

Нужно оценить два параметра: дисперсию
и математическое ожидание, используем два момента:

$$\begin{cases} \mathbb{E}(X_i) \approx \bar{x} \\ \mathbb{E}\left(X_i^2\right) \approx \overline{x^2} \end{cases} \Leftrightarrow \begin{cases} \mu = \bar{x} \\ \sigma^2 + \mu^2 = \overline{x^2} \end{cases} \Leftrightarrow \begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \overline{x^2} - \bar{x}^2 \end{cases}$$

Что такое доверительный интервал

Схема математической статистики

Выборка: $X_1, \ldots, X_n$ Параметр: θ

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

Точность оценки, прогнозов

доверительные интервалы

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмещенная
- Состоятельная
- Эффективная

Союзники

Асимптотические (при большом n)

- ЦПТ
- Дельта-метод

Точные

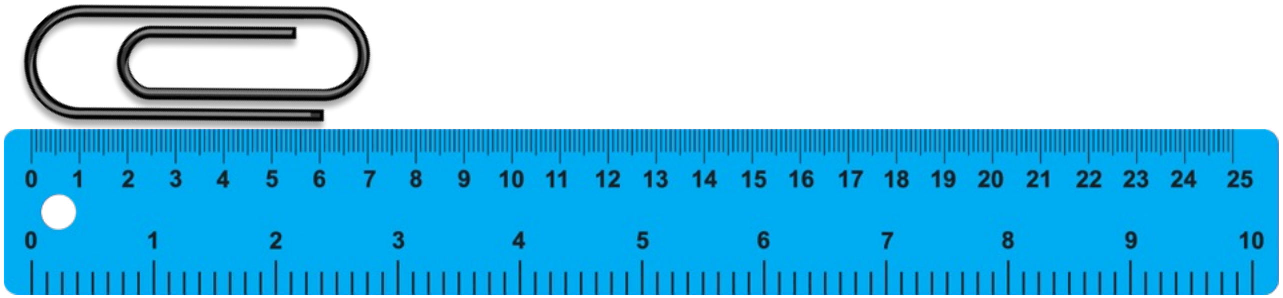
- Теорема Фишера
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

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

Зачем нужны доверительные интервалы

- Точечная оценка делается по случайной выборке ⇒ 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.

Схема математической статистики

Выборка: $X_1, \ldots, X_n$ 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
- $\chi^2_n, t_n, F_{n,k}$
- Ещё союзники!

Мощь средних

ЗБЧ даёт нам возможность с помощью метода моментов построить оценку $\hat{\theta}_{MM}$

ЦПТ даёт нам информацию о распределении $\hat{\theta}_{MM}$,
мы можем построить доверительный интервал:

$$\mathbb{P}\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R\right) = 1 - \alpha$$

$$\mathbb{P}\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R\right) = 0.95$$

α – уровень значимости

Если мы 100 раз попытаемся сесть на поезд на уровне значимости 0.05, в среднем мы будем опаздывать 5 раз

Мощь средних

ЗБЧ даёт нам возможность с помощью метода моментов построить оценку $\hat{\theta}_{MM}$

ЦПТ даёт нам информацию о распределении $\hat{\theta}_{MM}$,
мы можем построить доверительный интервал:

$$X_1, \ldots, X_n \sim iid$$

$$\mathbb{E}(X_i) = \mu, \; Var(X_i) = \sigma^2$$

$$\hat{\mu} = \bar{x} \overset{\text{ЦПТ}}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \Leftrightarrow \bar{x} - \mu \overset{\text{ЦПТ}}{\sim} N\left(0, \frac{\hat{\sigma}^2}{n}\right) \Leftrightarrow \frac{\bar{x} - \mu}{\sqrt{\frac{\hat{\sigma}^2}{n}}} \overset{\text{ЦПТ}}{\sim} N(0, 1)$$

центрирование нормирование

Асимптотический интервал для мат. ожидания

- ЦПТ позволяет построить доверительный интервал для любого мат. ожидания
- Наблюдаем $X_1, \ldots, X_n \sim iid$
- **Предполагаем:** X_i независимы и одинаково распределены, число наблюдений n велико, нет выбросов

$$\mathbb{E}(X_i) = \mu,\ Var(X_i) = \sigma^2$$

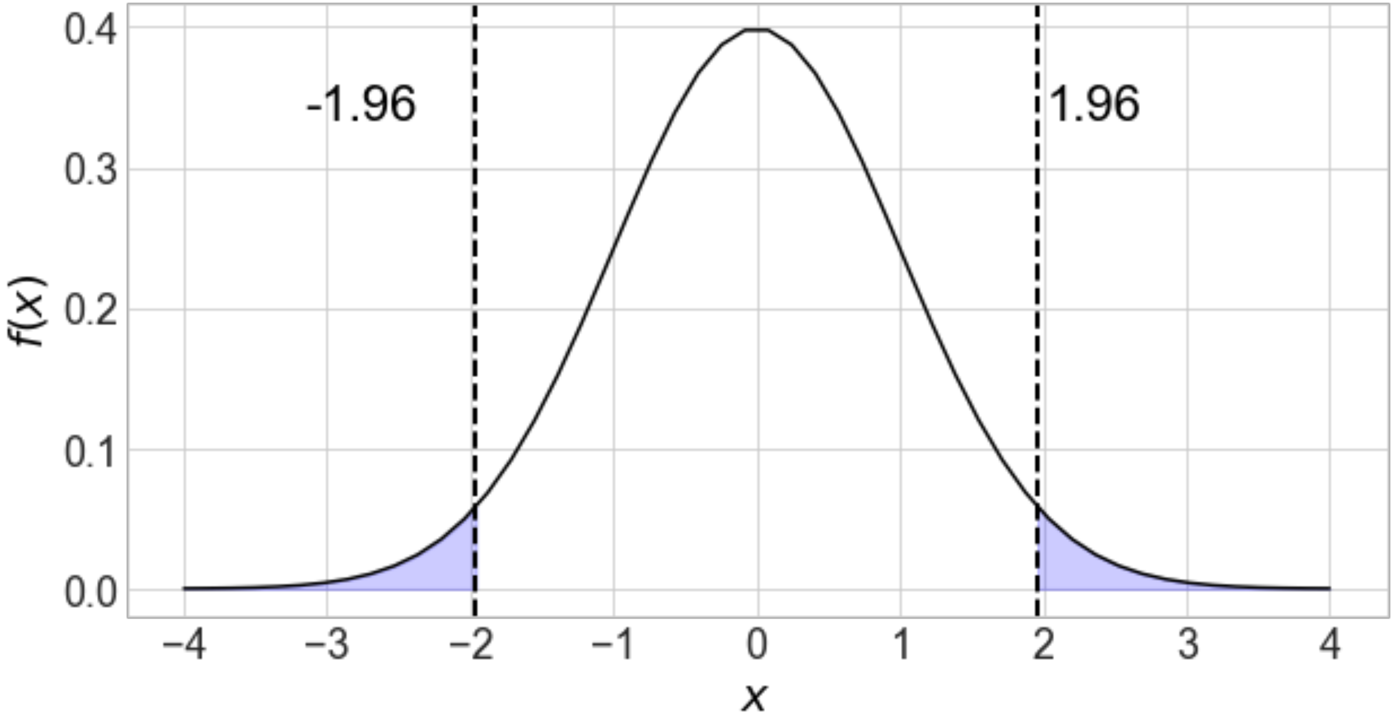
$$\bar{X} \overset{\text{ЦПТ}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \bar{X} - \mu \overset{\text{ЦПТ}}{\sim} N\left(0, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \overset{\text{ЦПТ}}{\sim} N(0,1)$$

центрирование　　　　стандартизация

Мощь средних

Вероятность того, что наша случайная величина окажется между -1.96 и 1.96 равна 0.95

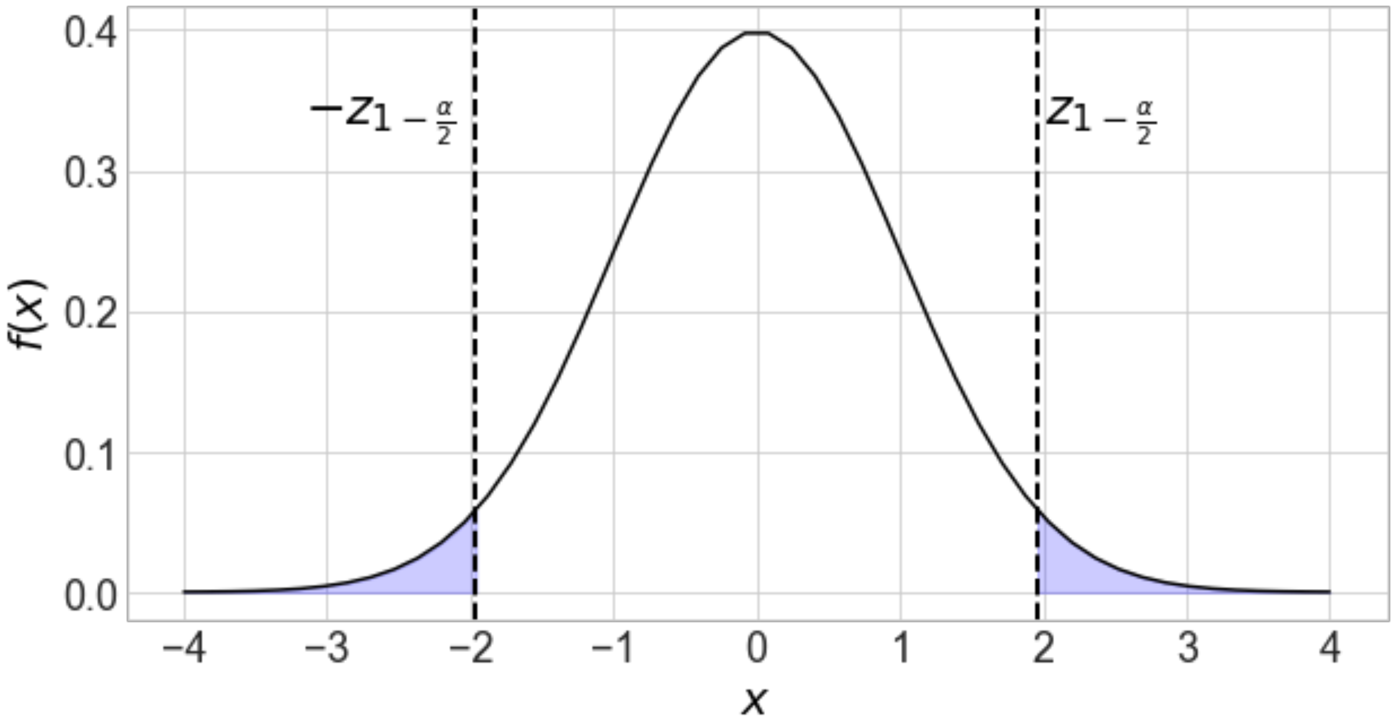
$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \overset{\text{ЦПТ}}{\sim} N(0,1)$$



Мощь средних

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

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



Мощь средних

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

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

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

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

При бесконечном повторении эксперимента интервал будет накрывать истинное значение параметра μ в $100 \cdot (1 - \alpha)\%$ случаев

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

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

$\hat{\sigma}^2$– состоятельная оценка для σ^2, то есть $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$

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

$$\xrightarrow{p} 1 \Rightarrow \xrightarrow{d} 1$$

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

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

$\hat{\sigma}^2$– состоятельная оценка для σ^2, то есть $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$

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

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

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

- Иногда кратко пишут:

$$\mu \in \{ \bar{X} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \}$$

Мощь средних

Длина интервала:

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

При росте n длина интервала падает

При росте дисперсии длина интервала увеличивается

При росте надёжности $1 - \alpha$ длина увеличивается

Дельта-метод

Если:

$$X_1, \ldots, X_n \sim iid, \qquad \mathbb{E}(X_1) = \mu, Var(X_1) = \sigma^2$$

$g(t)$ – 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.

Резюме

- Центральная предельная теорема позволяет построить для среднего асимптотический доверительный интервал
- Доверительный интервал позволяет описать степень неуверенности в полученной оценке
- Такой доверительный интервал верен при большом количестве наблюдений, если в выборке нет аномалий

Асимптотический доверительный интервал для разницы средних

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid \qquad Y_1, \dots, Y_m \sim iid$$

$$\bar{X} \overset{\text{ЦПТ}}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n}\right) \qquad \bar{Y} \overset{\text{ЦПТ}}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Разность нормальных случайных величин – нормальная случайная величина:

$$\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_1 - \mu_2$$

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

$$\bar{X} - \bar{Y} \overset{\text{ЦПТ}}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

Разность средних

Цены на недвижимость в двух районах города:

$$X_1, \dots, X_n \sim iid \qquad\qquad Y_1, \dots, Y_m \sim iid$$

$$\bar{X} - \bar{Y} \overset{\text{ЦПТ}}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

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

$$(\mu_1 - \mu_2) \in \{(\bar{X} - \bar{Y}) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}\}$$

Асимптотические доверительные интервалы для долей

Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

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

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

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n} = \bar{X}$$

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:

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

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

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

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

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

Мощь долей

Получаем доверительный интервал для доли:

$$\bar{X} \overset{\text{ЦПТ}}{\sim} N\left(\mu, \frac{\hat{\sigma}^2}{n}\right) \qquad \hat{p} = \bar{X} \overset{\text{ЦПТ}}{\sim} N\left(p, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

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

Мощь долей

Получаем доверительный интервал для разности долей:

$$\bar{X} - \bar{Y} \overset{\text{ЦПТ}}{\sim} N\left(\mu_1 - \mu_2, \frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}\right)$$

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

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

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

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

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

Число наблюдений

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

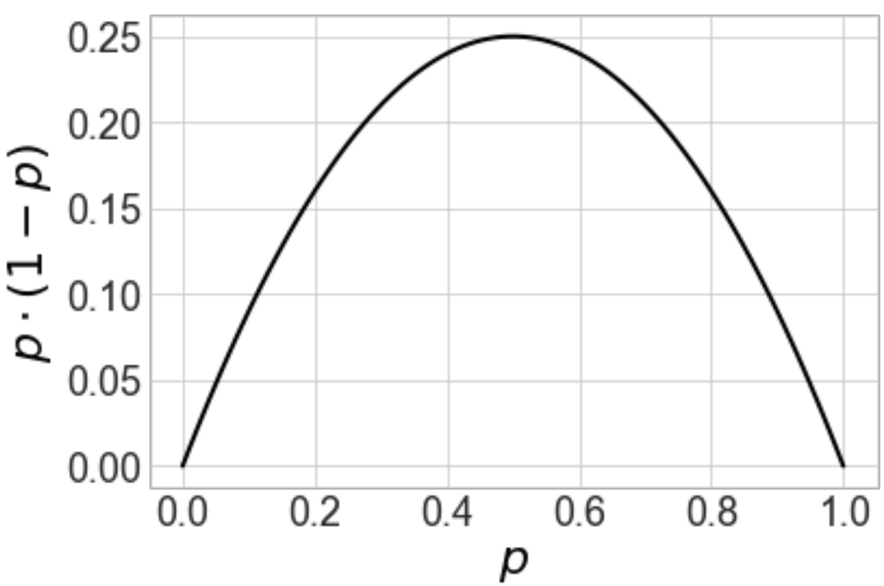
До начала испытаний мы не знаем $\hat{p}$, но мы знаем, что величина $\hat{p}(1-\hat{p})$ никогда не будет превышать 0.25

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

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

$$\Rightarrow p = 0.5$$

$$f(p) = 0.25$$



Число наблюдений

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

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:

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

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

Асимптотические доверительные интервалы для дисперсии

Асимптотический интервал для дисперсии

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

а $\hat{s}^2$ – через средние

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

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

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

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

➤ *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:

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

Оно может быть использовано для строительства доверительных интервалов

$$\hat{\sigma}^2 \in \{ \bar{X} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \} \qquad s = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

➤ *https://www.stat.umn.edu/geyer/s06/5102/notes/ci.pdf*

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