

# **Доверительные интервалы**

# План

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

# Закон больших чисел (ЗБЧ)

# Как устроен мир



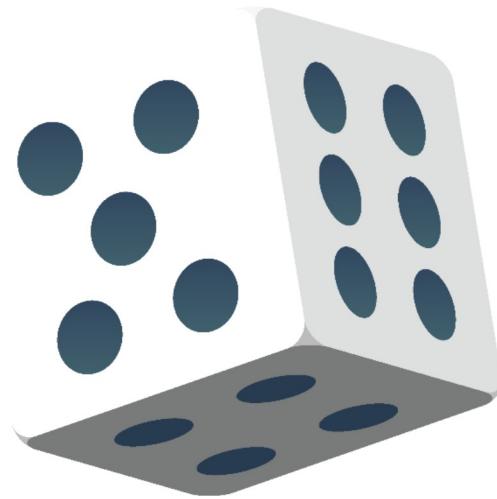
X

- Теория вероятностей изучает различные процессы порождения данных (некоторый сундук). В реальности мы не наблюдаем эти процессы.
- Однако эти процессы порождают **выборки**. Математическая статистика изучает их и пытается восстановить их структуру.

# Закон больших чисел (ЗБЧ)

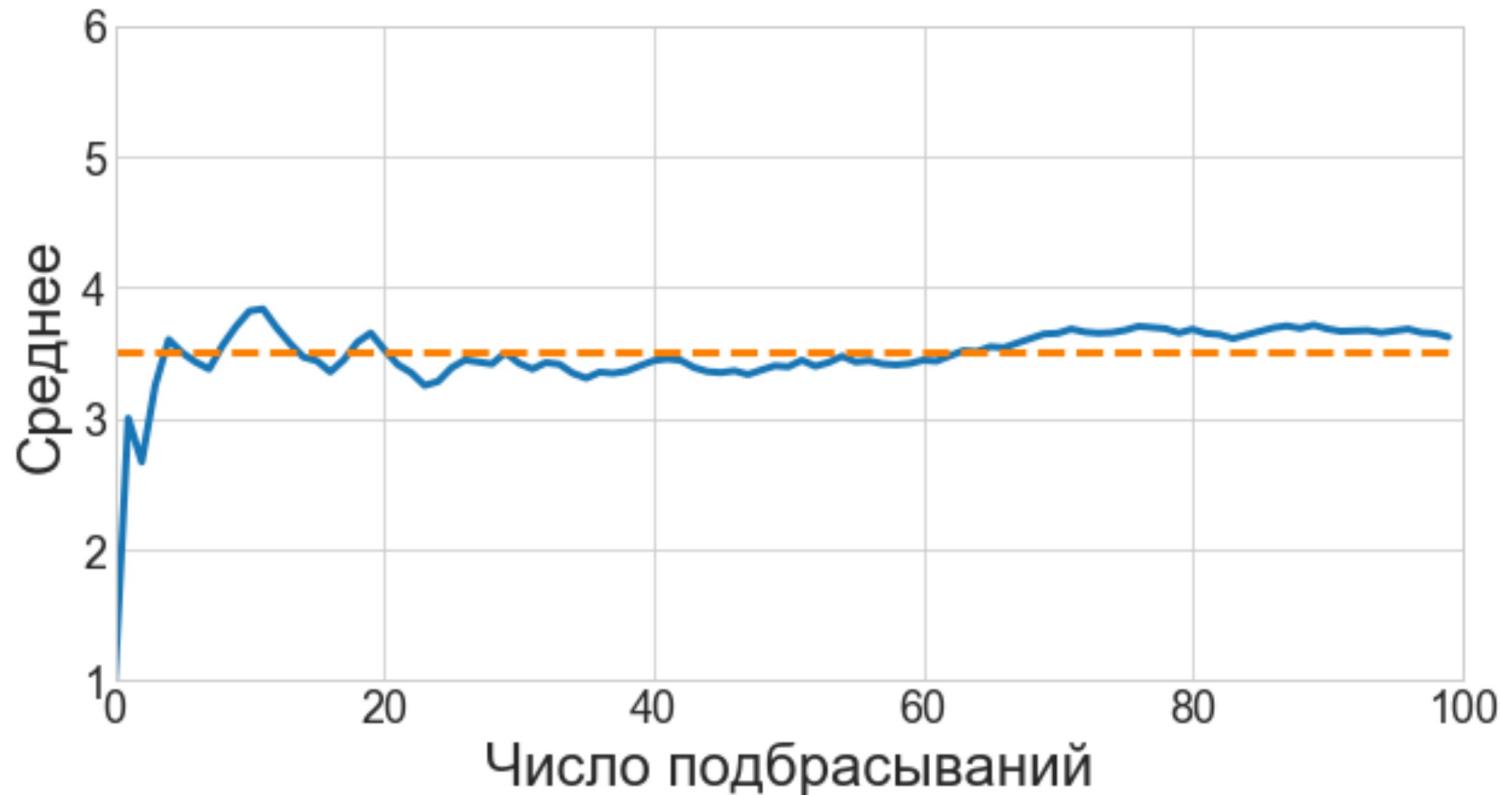
ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа

**Пример:** Игровая 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,  $\text{Var}(X_1) < \infty$  тогда:

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

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

# Слабая форма ЗБЧ (Чебышёв)

## Простым языком:

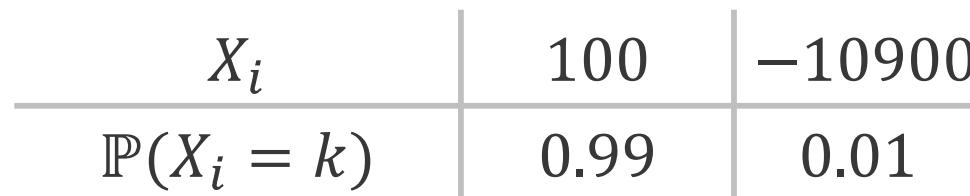
- Среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа
- Среднее для бесконечного числа случайных величин неслучайно
- Если у нас есть страховая фирма, мы можем заработать немного денег (самая простая формулировка)

# 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



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

# 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 объемах выборки вероятность отклониться  
от 0.5 больше. Именно об этом говорит нам ЗБЧ.



# 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,  $\text{Var}(X_1) < \infty$  тогда:

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

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

# Сходимость по вероятности

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

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



# Сходимость по вероятности

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

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

То есть:

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



Обычно пишут:

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

## Резюме

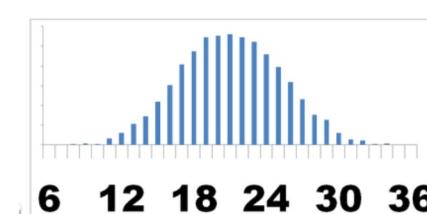
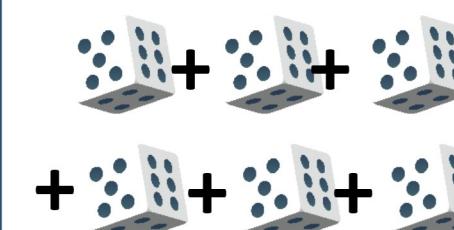
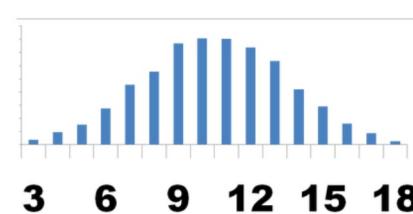
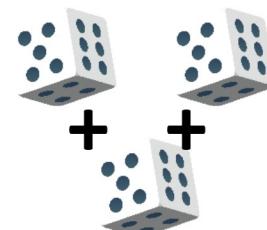
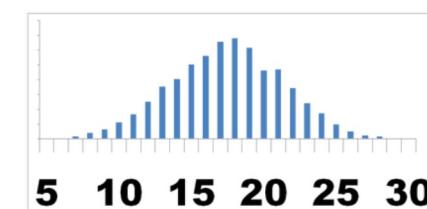
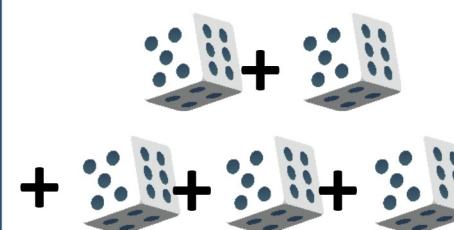
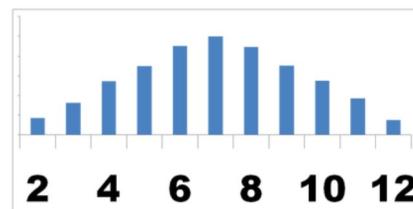
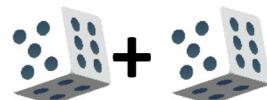
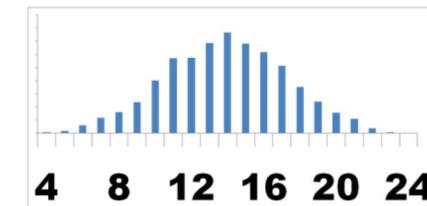
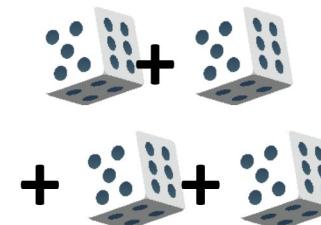
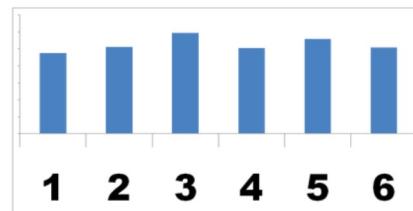
В слабой форме ЗБЧ среднее сходится к математическому ожиданию по вероятности

Для сходимости по вероятности верны такие же арифметические 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  $X_1, \dots, X_n$  попарно независимые и одинаково распределённые случайные величины с конечной дисперсией,  $\text{Var}(X_1) < \infty$  тогда:

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



Иногда пишут: либо:

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

# Сходимость по распределению

# Центральная предельная теорема

**Теорема:**

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

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

- ✓ 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

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

то есть последовательность функций распределения  
 $F_{X_n}(x)$  сходится к функции  $F_X(x)$  во всех точках  $x$ , где  $F_X(x)$   
непрерывна.



Обычно пишут:

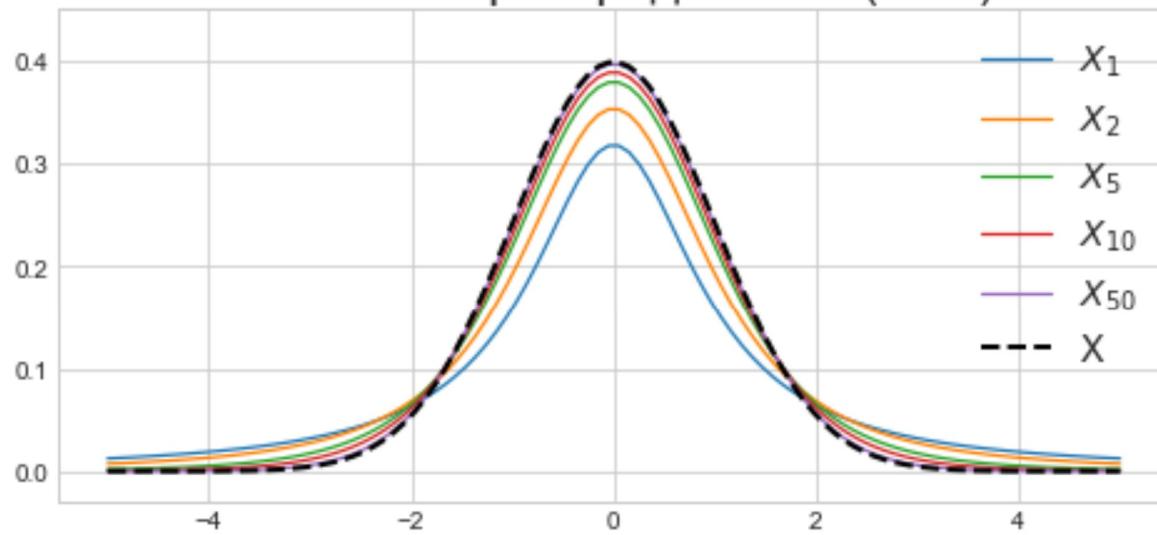
$$X_n \xrightarrow{d} X \text{ при } n \rightarrow \infty$$

либо:

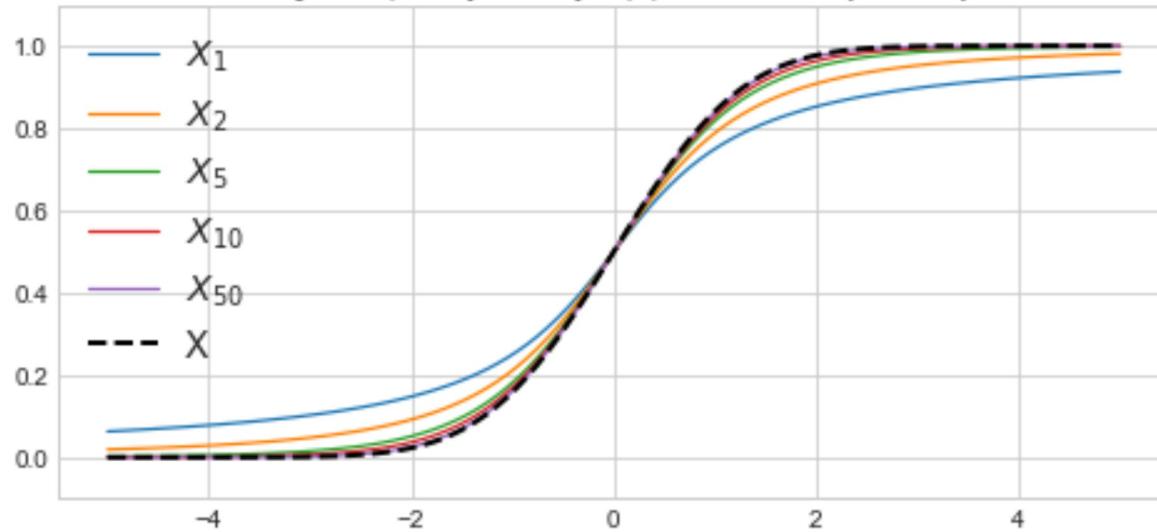
$$X_n \xrightarrow{F} X \text{ при } n \rightarrow \infty$$

# Сходимость по распределению

Плотность распределения (PDF)



Функция распределения (CDF)



# 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

...

$$X = \text{} + \text{} + \text{} + \text{} + \dots$$

$X$  – время прихода Миши на первую пару

# 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**

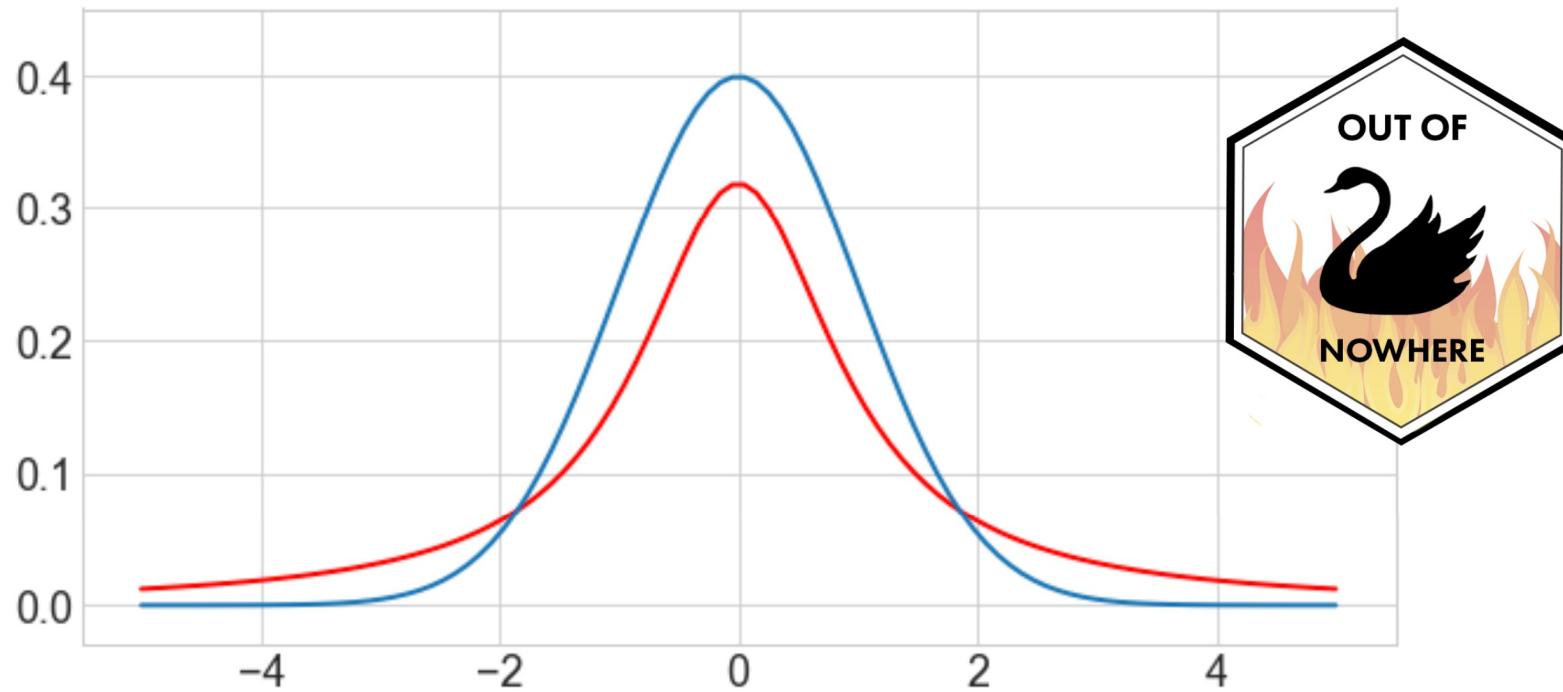
# Крайнеземье и средиземье



ЦПТ и ЗБЧ работают в Средиземье



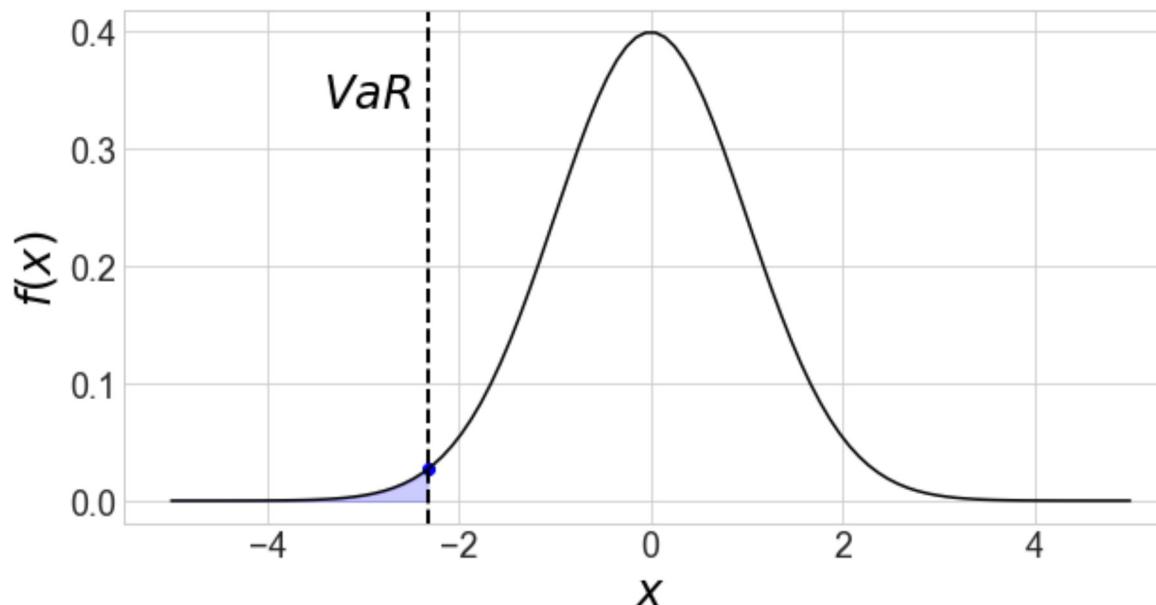
# Крайнеземье и средиземье



- 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:**  $\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$

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

**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:  $X_1, \dots, X_n$  Параметр:  $\theta$

$$\hat{\theta}$$

Как оценить

- Метод моментов
- Метод максимального правдоподобия

Хорошие свойства

- Несмешенная
- Состоятельная
- Эффективная

Союзники

Асимптотические  
(при большом  $n$ )

- ЦПТ
- Дельта-метод

Точные

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

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

Точность  
оценки,  
прогнозов

доверительные  
интервалы

Ответы на  
вопросы  
проверка  
гипотез

# Что такое доверительный интервал

# Схема математической статистики

Выборка:  $X_1, \dots, X_n$  Параметр:  $\theta$

$\hat{\theta}$

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

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!

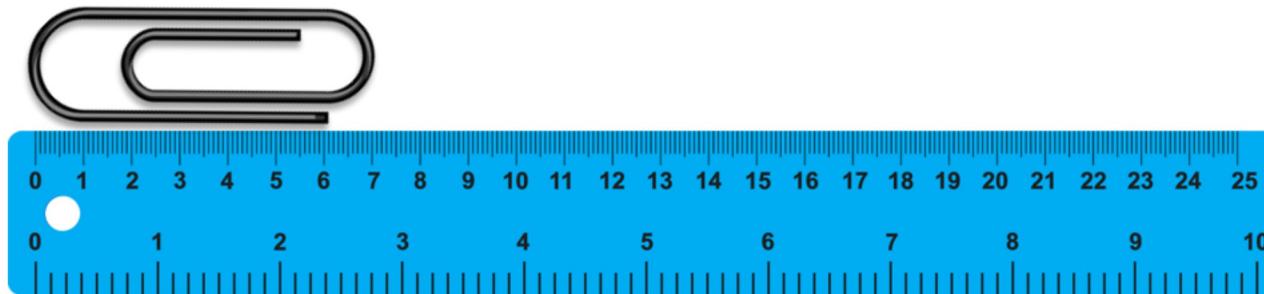
Точность  
оценки,  
прогнозов

доверительные  
интервалы

Ответы на  
вопросы  
проверка  
гипотез

# 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  $7 \pm 0.1$  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:  $X_1, \dots, X_n$  Параметр:  $\theta$

$\hat{\theta}$

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

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

# 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:

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

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

$\alpha$  – 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:

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

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

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

центрирование

нормирование

# Асимптотический интервал для мат. ожидания

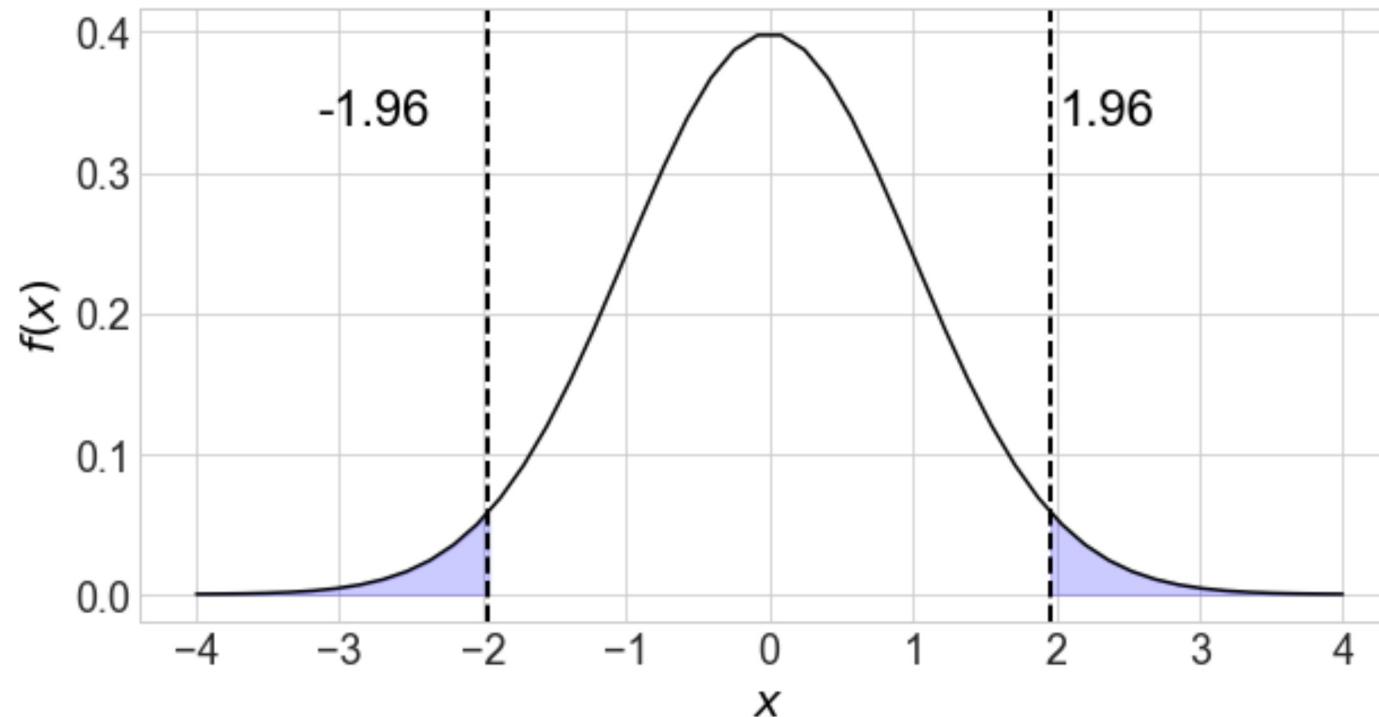
- ЦПТ позволяет построить доверительный интервал для любого мат. ожидания
  - Наблюдаем  $X_1, \dots, X_n \sim iid$
  - Предполагаем:  $X_i$  независимы и одинаково распределены, число наблюдений  $n$  велико, нет выбросов

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

# Мощь средних

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

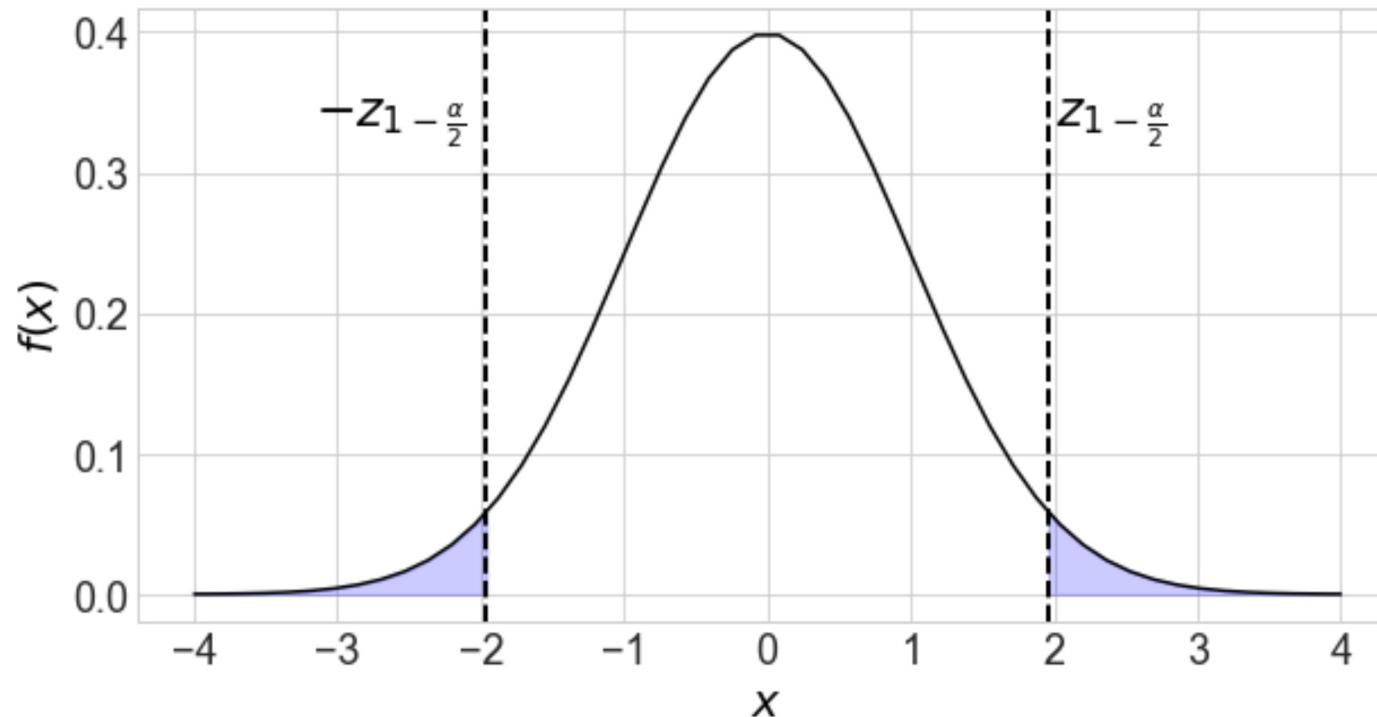
$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \stackrel{\text{цпт}}{\sim} N(0, 1)$$



# 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  
 $\mu$  в  $100 \cdot (1 - \alpha)\%$  случаев

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

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

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

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

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

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

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

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

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

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



Иногда кратко пишут:

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

# 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

# Асимптотический интервал для дисперсии

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

а  $\hat{s}^2$  – через средние

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

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

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

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

► <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

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

# 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:

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

$$\bar{X} \stackrel{\text{ЦПТ}}{\sim} N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

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

$$\bar{Y} \stackrel{\text{ЦПТ}}{\sim} N\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

Разность нормальных случайных величин – нормальная  
случайная величина:

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

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

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

# Разность средних

Цены на недвижимость в двух районах города:

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

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

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

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:

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

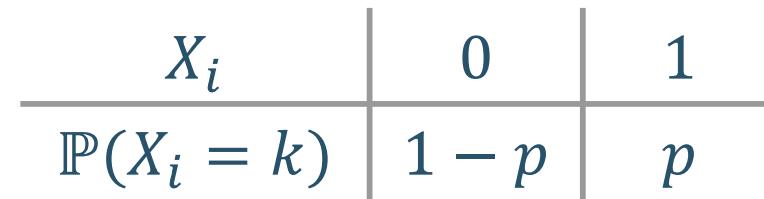
$X_i$	0	1
$\mathbb{P}(X_i = k)$	$1 - p$	$p$

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

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

# Мощь долей

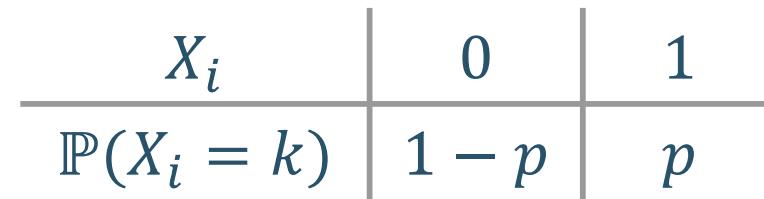
По аналогии можно построить асимптотические доверительные интервалы для долей:

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


Найдём математическое ожидание и дисперсию оценки, а потом воспользуемся ЦПТ

# Мощь долей

По аналогии можно построить асимптотические доверительные интервалы для долей:

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


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

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

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

# Мощь долей

Получаем доверительный интервал для доли:

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

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

# Мощь долей

Получаем доверительный интервал для разности долей:

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

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

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

# 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

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

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

# Число наблюдений

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

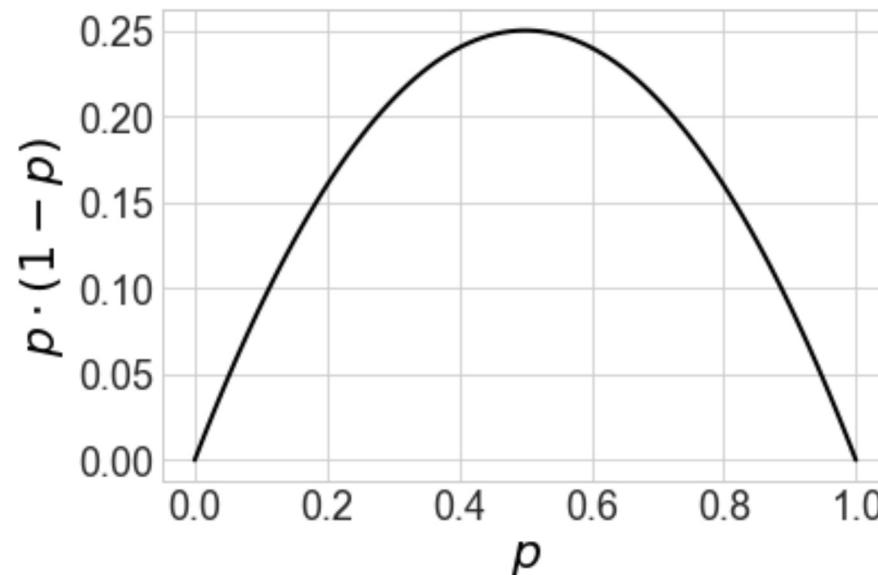
До начала испытаний мы не знаем  $\hat{p}$ , но мы знаем, что величина  $\hat{p}(1 - \hat{p})$  никогда не будет превышать 0.25

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

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

$$\Rightarrow p = 0.5$$

$$f(p) = 0.25$$



# 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