CS Space

Вероятностное моделирование

лекция 1

Артур Игнатьев
CS Space, МКН СПбГУ, ИТМО

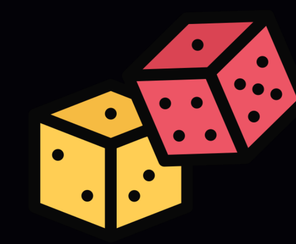
Владимир Евменов
CS Space, Huawei

Эксперимент

Конечное число исходов $\{\omega_1, \omega_2, \ldots, \omega_N\}$

- Однократное подбрасывание монетки
  $\Omega = \{$орел, решка$\}$

- Подбрасывание игральной кости
  $\Omega = \{1,2,3,4,5,6\}$



- n-кратное подбрасывание монетки
  $\Omega = \{\omega : \omega = (a_1, a_2, \ldots, a_n)\}$
- Урна с шарами. Есть 5 красных шара и 5 зеленых
  $\Omega = \{\omega : \omega = (a_1, a_2, a_3),$ где $a_i \in [10]\}$
- Колода из 36 карт

Вероятностная модель

Пространство элементарных событий $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$

- $A \subset \Omega$ — событие
- 3 раза подбросили монетку
- A — выпало два орла $= \{\text{ООР}, \text{ОРО}, \text{ООР}\}$
- B — не выпал орел $= \{\text{РРР}\}$
- $A \cup B$ — выпало четное число орлов

Вероятностная пространство

Пространство элементарных событий $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$

- $p_1, p_2, \ldots, p_N \geq 0$
- $p_1 + p_2 + \ldots + p_N = 1$
- $P(A) = \sum_{k:\omega_k \in A} p_k$ – вероятность события

Свойства.

- $P(\varnothing) = 0,\ P(\Omega) = 1$
- $0 \leq P(A) \leq 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cup B) \leq P(A) + P(B)$
- $P(\overline{A}) = 1 - P(A)$
- $P(\cup_i A) \leq \sum_i P(A_i)$

Условная вероятность

- $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$

Свойства.

- $P(A \mid A) = 1$
- Если $A \subset B$, то $P(B \mid A) = 1$
- $P(\varnothing \mid A) = 0$
- Если $B_1 \cap B_2 = \varnothing$, то $P(B_1 \cup B_2 \mid A) = P(B_1 \mid A) + P(B_2 \mid A)$
- $P(B \mid A) + P(\overline{B} \mid A) = 1$

Формула полной вероятности

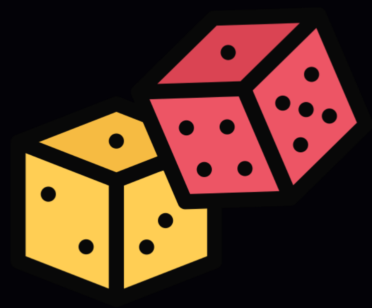
- $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n, A_i \cap A_j = \varnothing$ и $P(A_i) > 0$
- $B = B \cap \cup_i A_i = \cup_i B \cap A_i$
- $P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i) P(A_i)$
- Если $0 < P(A) < 1$, то $P(B) = P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})$
- **Задача.** Из полного набора костей домино взята одна кость. Найдите вероятность того, что наудачу взятую вторую кость можно приставить к первой по правилам домино.

Формула Байеса

- Если для событий A и B верно $P(A) > 0$ и $P(B) > 0$
- $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$
- $\Omega = A_1 \cup A_2 \cup \ldots \cup A_n, A_i \cap A_j = \varnothing$ и $P(A_i) > 0, P(B)$
- $P(A_i \mid B) = \dfrac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)}$

Независимые события

- События A и B независимые, если $P(A \cap B) = P(A)P(B)$
- **Пример.** Бросили кубик



- A — выпало четное число, $P(A) = \frac{1}{2}$
- B — выпало число, кратное трем, $P(B) = \frac{1}{3}$
- C — выпало число не менее 4, $P(C) = \frac{1}{2}$
- $P(A \cap B) = \frac{1}{6} = P(A)P(B)$
- $P(A \cap C) = \frac{1}{3} \neq \frac{1}{4} = P(A)P(C)$
- $P(B \cap C) = \frac{1}{6} = P(B)P(C)$

Независимость в совокупности

- События $A_1, A_2, \ldots, A_n$ независимы в совокупности, если для любых $1 \leq i_1 < i_2 < \ldots < i_k \leq n$
  $P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \ldots P(A_{i_k})$
- Независимость дает независимость в совокупности?
- **Пример.** На гранях тетраэдра числа 1, 2, 3, 4. Бросили тетраэдр.
- $A = \{1,2\}, B = \{1,3\}, C = \{1,4\}$

- $P(A) = P(B) = P(C) = \dfrac{1}{2}$
- $P(A \cap B) = P(B \cap C) = P(A \cap C) = \dfrac{1}{4}$
- $P(A \cap B \cap C) = \dfrac{1}{4} \neq \dfrac{1}{8} = P(A)P(B)P(C)$
- В другую сторону?

Случайная величина

- Пространство элементарных событий $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$
- $\xi : \Omega \to \mathbb{R}$ — случайная величина
- **Пример.** Бросили кубик два раза. $\Omega = \{(i, j) \,|\, i, j \in [6]\}$
- $\xi = i + j$ — случайная величина
- Распределение случайно величины — набор вероятностей $P(\xi = a)$
- Независимые с.в. $P(\xi = a, \eta = b) = P(\xi = a)P(\eta = b)$ при всех a и b

Характеристики с.в.

- Математическое ожидание
$$\mathbb{E}\xi = \sum_{\omega\in\Omega} \xi(\omega)P(\omega) = \sum_{x\in\xi(\Omega)} xP(\xi = x)$$
- Медиана a с.в. ξ, если $P(\xi \le a) \ge \frac{1}{2}$ и $P(\xi \ge a) \ge \frac{1}{2}$

Свойства.

- $\xi \ge 0 \implies \mathbb{E}\xi \ge 0$
- $\mathbb{E}(a\xi + b\eta) = a\mathbb{E}\xi + b\mathbb{E}\eta$
- $\xi \ge \eta \implies \mathbb{E}\xi \ge \mathbb{E}\eta$
- $|\mathbb{E}\xi| \le \mathbb{E}|\xi|$
- $(\mathbb{E}|\xi\eta|)^2 \le \mathbb{E}\xi^2\mathbb{E}\eta^2$
- Если ξ и η независимы, то $\mathbb{E}(\xi\eta) = \mathbb{E}\xi\mathbb{E}\eta$

Характеристики с.в.

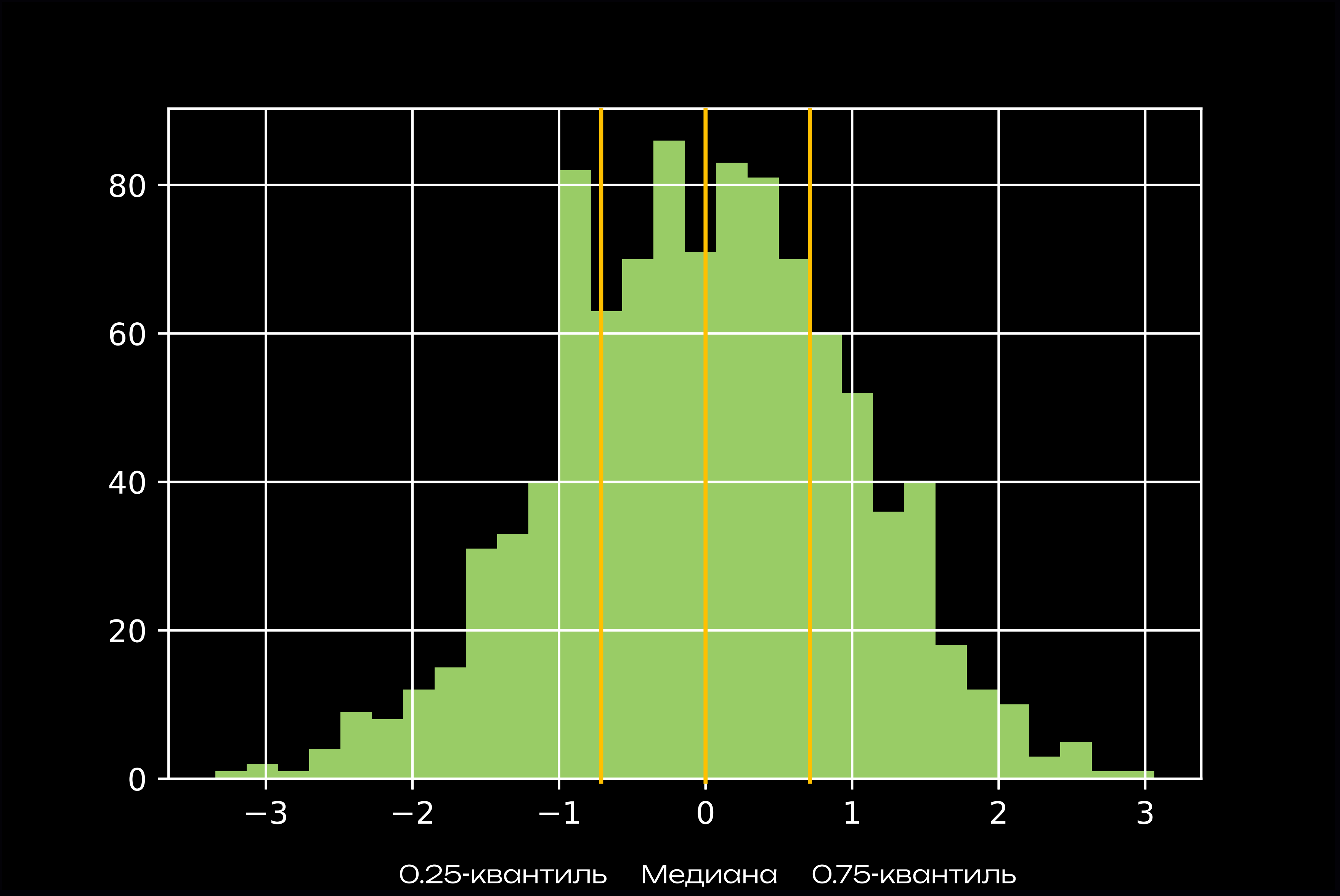
- Дисперсия
  $\mathbb{D}\xi = \mathbb{E}((\xi - \mathbb{E}\xi)^2)$

Свойства.

- $\mathbb{D}\xi \geq 0$
- $\mathbb{D}\xi = \mathbb{E}(\xi^2) - (\mathbb{E}\xi)^2$
- $\mathbb{D}(\xi + a) = \mathbb{D}\xi$
- $\mathbb{D}(c\xi) = c^2\mathbb{D}\xi$

Квантиль с.в.

- Пусть $\alpha \in (0,1)$, α-квантиль называется число x_α: $P(\xi \le x_\alpha) \ge \alpha$ и $P(\xi \ge x_\alpha) \ge 1 - \alpha$



Оценка характеристик по модели

Пример. Барбершоп:

- 3 мастера
- первый клиент приходит в 9 часов + $Exp(1/20)$ минут, второй приходит через $Exp(1/20)$ минут после первого и т.п.
- если есть свободный мастер, то клиент идет к нему, иначе ждет в очереди
- мастер обслуживает клиента в течение $U(15,45)$ минут
- в 16:00 вход закрывается, а барбершоп закрывается после обслуживания последнего уже зашедшего клиента
- Какова медианная доля клиентов, которым приходится сидеть в очереди? Каков 0.95 квантиль времени сидения в очереди? Во сколько в среднем закрывается барбершоп?

Метод Монте-Карло

Когда теоретически оценить характеристику нелегко, можно использовать метод монтекарловских симуляций.

Вкратце идея такая:

1. с помощью модели генерируем N независимых реализаций данных
2. по полученным реализациям считаем нужную характеристику

В случае с барбершопом нужно будет сгенерировать N дней, для i-го дня посчитать долю ждущих r_i, среднее время сидения в очереди w_i и время закрытия c_i, после чего

- оценить медиану долей как центральный элемент в отсортированном наборе r_i
- оценить 0.95 квантиль времени ожидания как 0.95N-й элемент в отсортированном наборе w_i
- оценить среднее время закрытия как среднее набора c_i

Метод Монте-Карло

- Тонкости...
- Как понять, что нужно сделать с набором значений, чтобы оценить искомую характеристику?
- Какой N взять, чтобы было хорошо?
- Посчитанные значения наверняка будут неточными. Как посчитать погрешность?
- Всегда ли посчитанная оценка будет реально оценивать то, что нужно?

Выборка

- Набор $x_{[n]} = [x_1, \ldots, x_n]$ из n независимых реализаций случайного элемента X называется выборкой объема n. Распределение $\mathscr{P}_X$ называется генеральной совокупностью.

- Случайной выборкой называется набор $X_{[n]} = [X_1, \ldots, X_n]$, $X_i \sim X$ независимых одинаково распределенных случайных элементов. Выборка это реализация (одна) случайной выборки.

- Предположение о том, что некоторый набор данных является выборкой — это сильное предположение; всегда нужно подумать о том, адекватно ли оно!

Эмпирическое распределение

- Для выборки $x_{[n]}$ рассмотрим эмпирическое распределение $\mathscr{P}_n^\star$:

	x_1	x_2	$\dots$	x_n
P	$\frac{1}{n}$	$\frac{1}{n}$	$\dots$	$\frac{1}{n}$

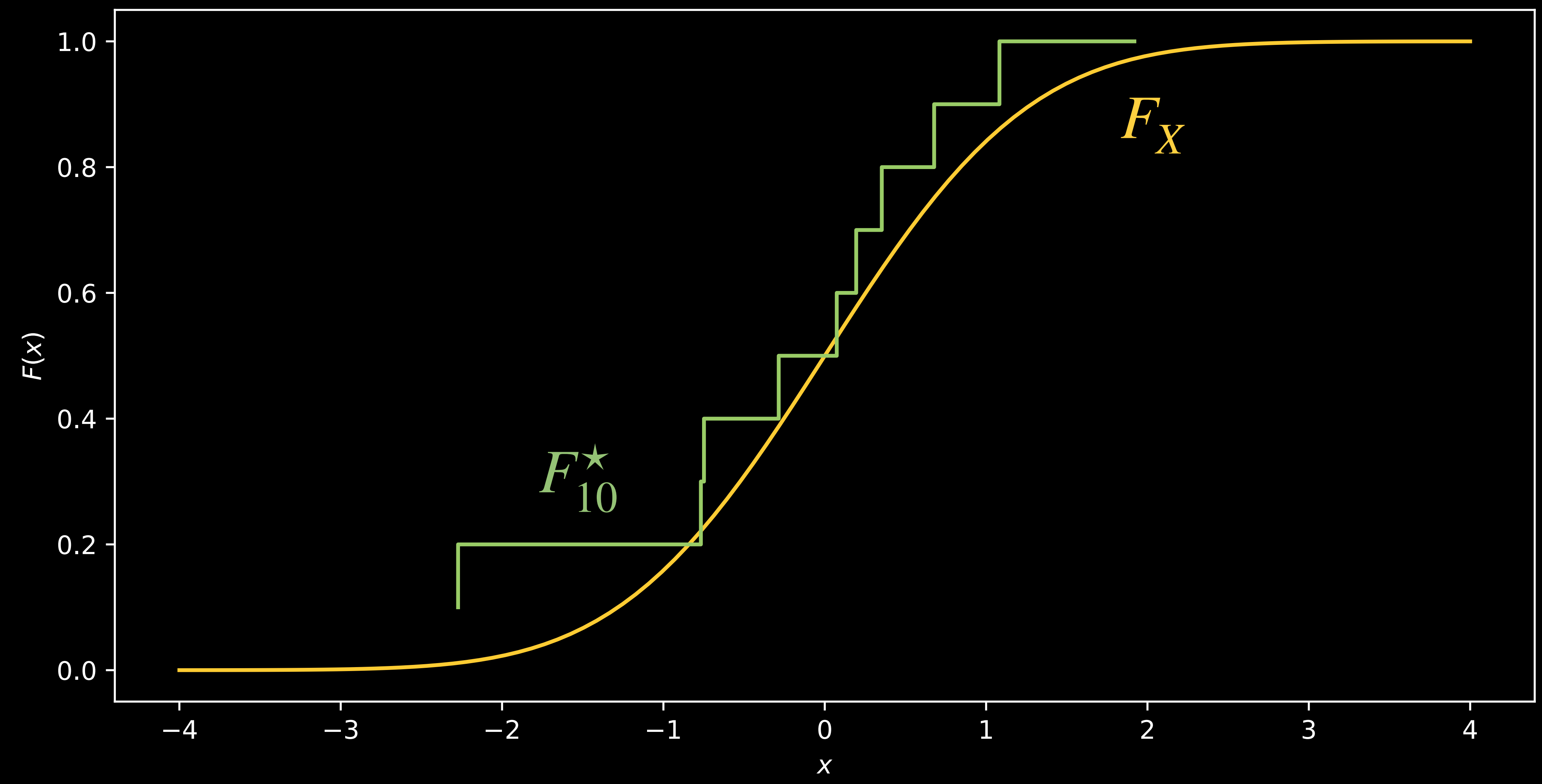
- Дискретное распределение. Вероятность события A равна
$$P(A) = \frac{1}{n} \sum_{i=1}^{n} [x_i \in A]$$

- Вообще $P(x_i)$ не равно $\dfrac{1}{n}$, так как в выборке может быть несколько одинаковых значений

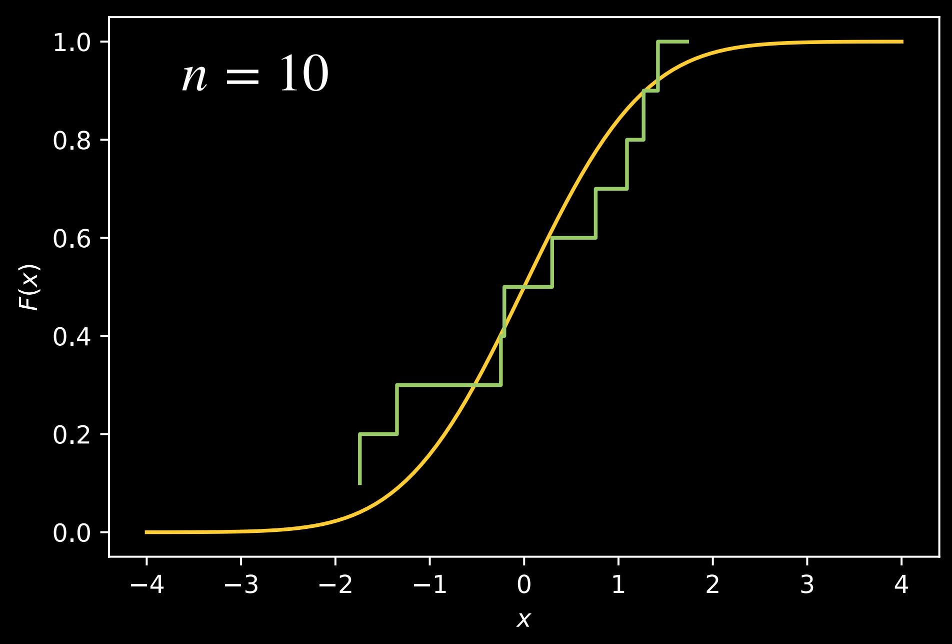
- Эмпирическая случайная величина $X^\star$, имеющая распределение $\mathscr{P}_n^\star$. Реализацией $X^\star$ является случайный элемент выборки.

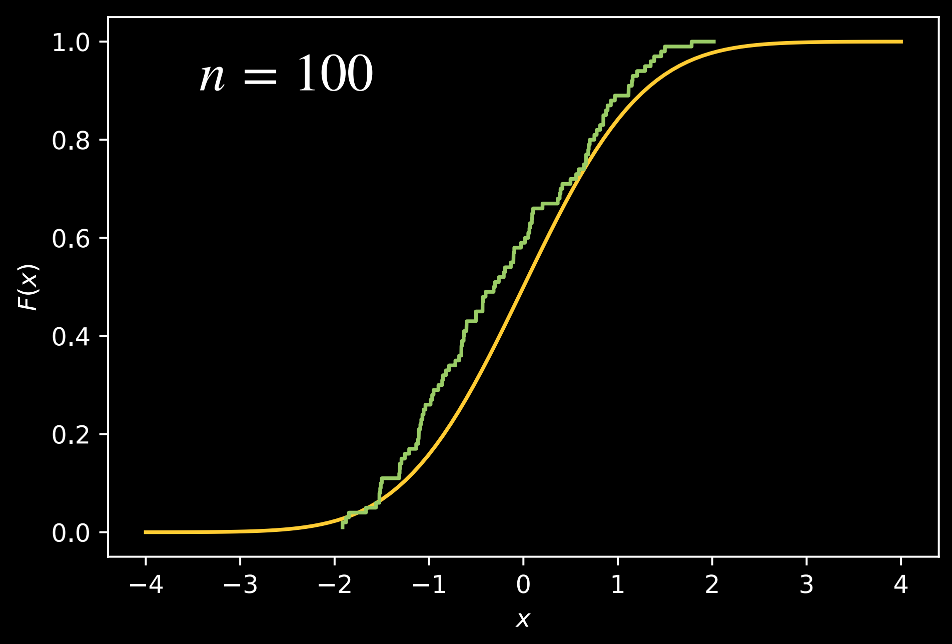
Эмпирическая ф.р.

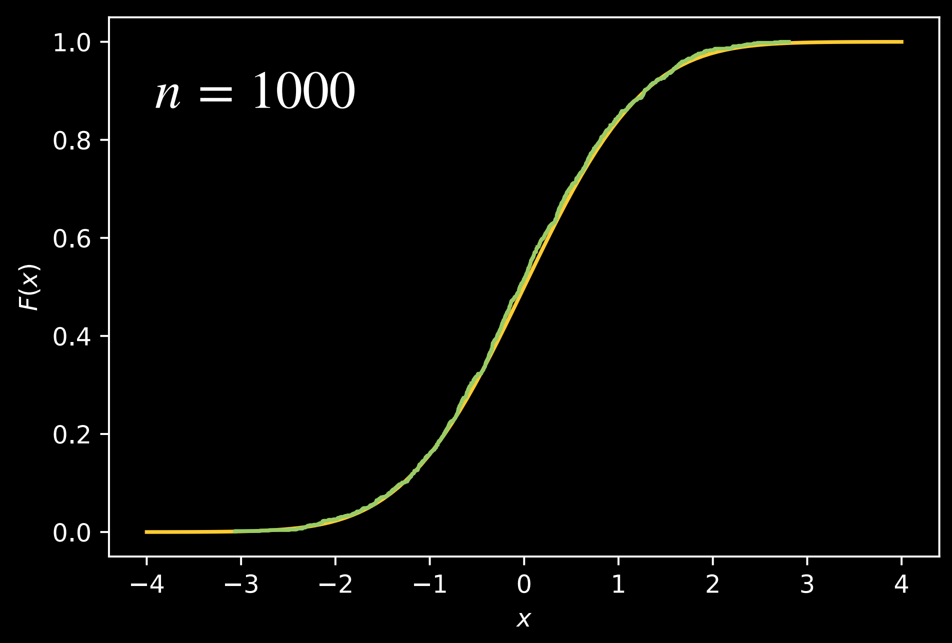
- Для выборки $x_{[n]}$ с одномерным $\mathscr{P}_X$ рассмотрим $F_n^\star = \dfrac{1}{n}\sum_{i=1}^{n}[x_i < x]$ — эмпирическую функцию распределения.

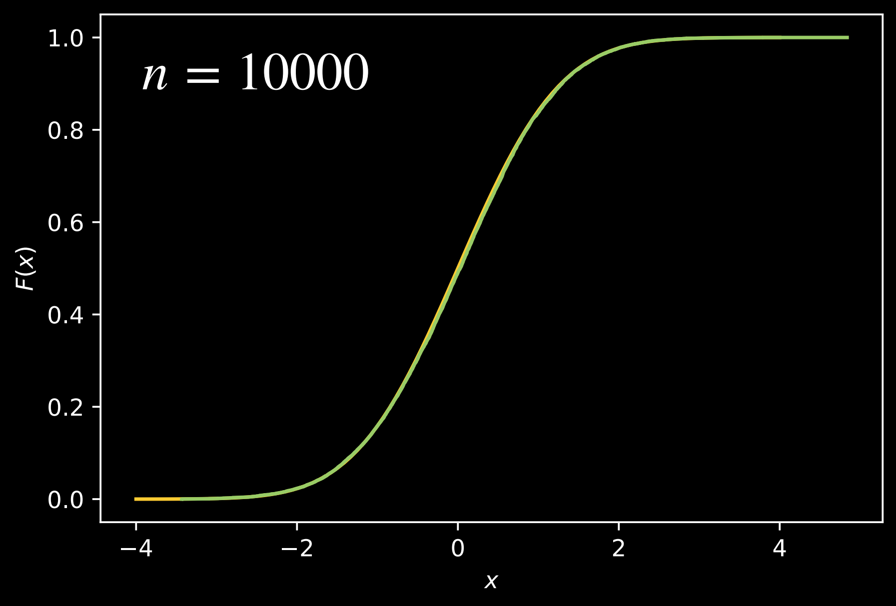


Количество и качество

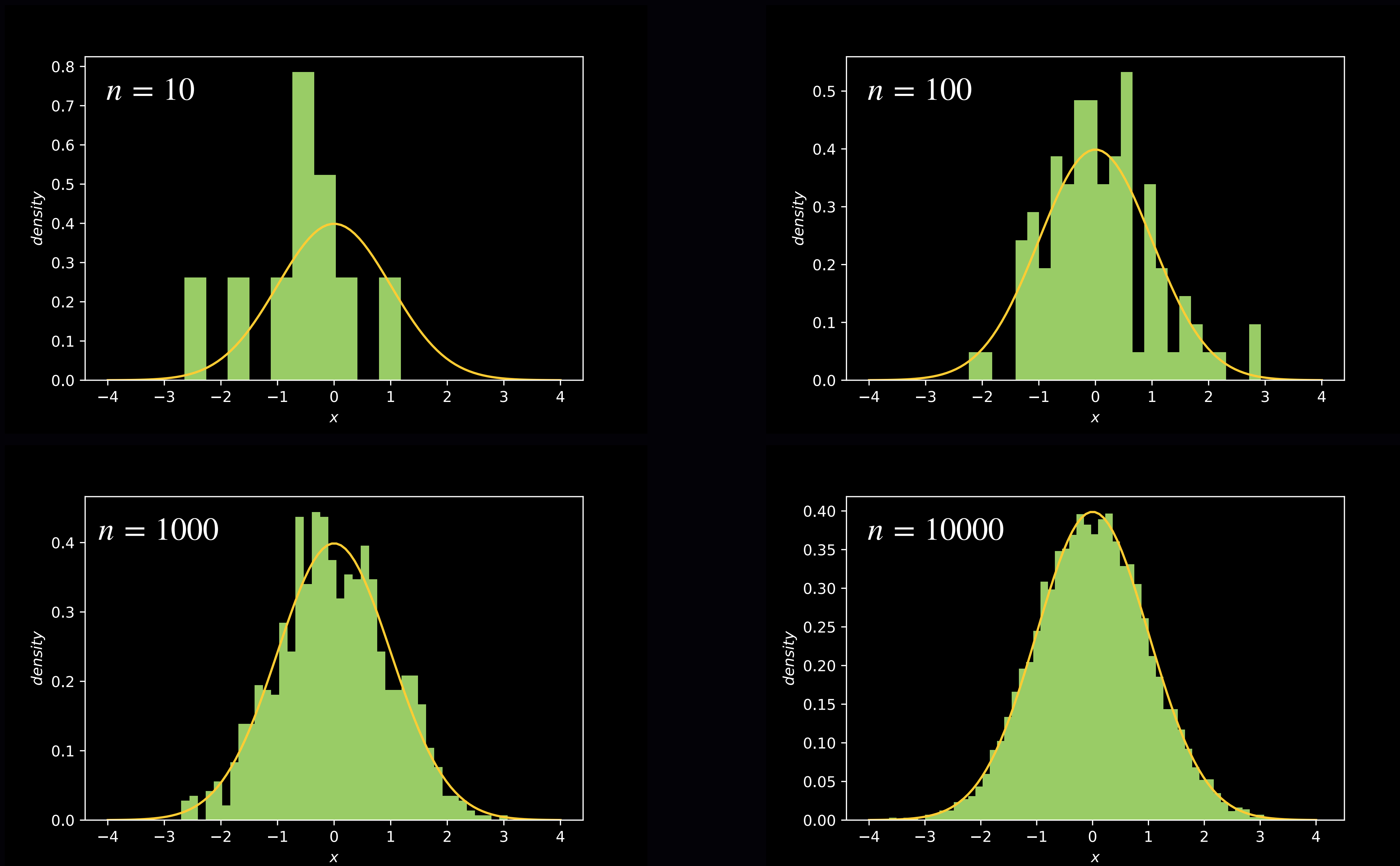








Количество и качество



Почему все работает?

- **Закон больших чисел.**
  ξ_n — независимые одинаково распределенные случайные величины. Тогда $\dfrac{\xi_1 + \ldots + \xi_n}{n} \underset{n\to\infty}{\to} a$ почти наверно $\iff \mathbb{E}\xi_n = a.$

- **Теорема Гливенко-Кантелли.**
  $\sup\limits_{x\in\mathbb{R}} | F_n^\star(x) - F_X(x) | \to 0$ почти наверное.