Лекция 3

Теория вероятностей (продолжение). Ковариация и корреляция

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

11 апреля

В предыдущих сериях

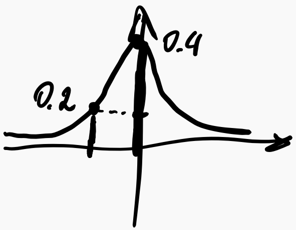
- Примеры задач теории вероятностей, статистики и машинного обучения.
- Генеральная совокупность и выборка.
- Типы переменных.
- Основные понятия ТВ.
- Функция распределения и функция вероятности.

$$F_X(x) = \underbrace{\mathbb{P}\{X \le x\}}\qquad \underbrace{P\{X = x\}}$$

Функция плотности

Функция плотности
Некоторая функция $f(x)$, описывающая непрерывную случайную величину X, для которой верно:

$$\boxed{F_X(a) = \mathbb{P}(X \leqslant a) = \int_{-\infty}^{a} f(x)dx}$$



- Сами значения показывают «относительную вероятность», но обычно не интерпретируются.
- Площадь под ней равна 1.
- Самое важное свойство:

$$\boxed{\mathbb{P}(a \leqslant X \leqslant b) = \int_{a}^{b} f(x)dx}$$

Плотность как относительная вероятность

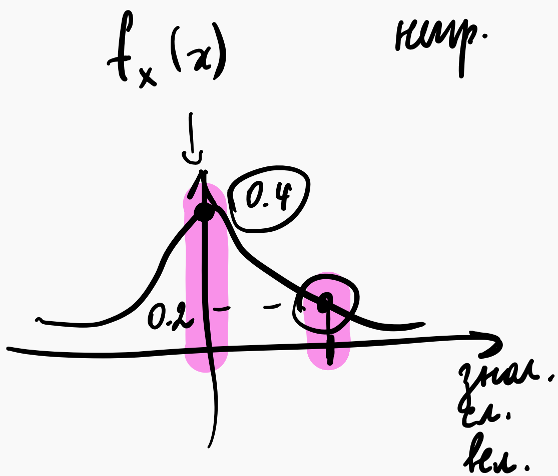
	Дискретная	Непрерывная
$\mathbb{P}\{X = a\}$	Ф. вероятности	0, но ф. плотности
$\mathbb{P}\{X \leqslant a\}$	Ф. распр.	Ф. распр.

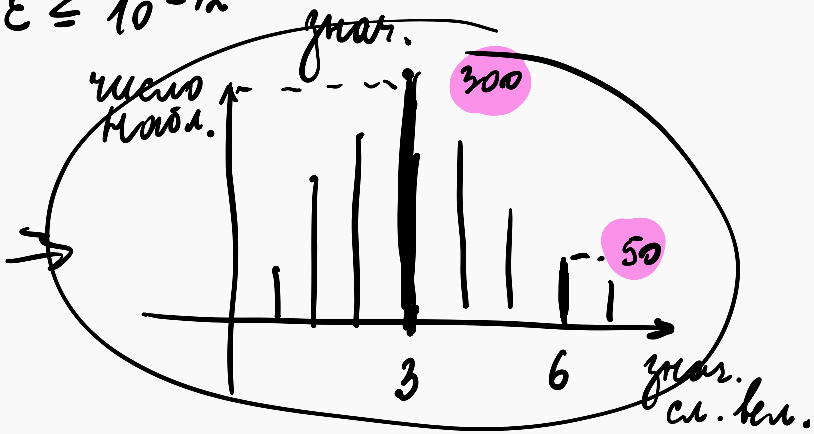
$$\mathbb{P}\{X = a\} \approx \int_{a-\varepsilon}^{a+\varepsilon} f_X(x)\,dx$$

непр.

мал. пром.

$\varepsilon \leq 10^{-12}$





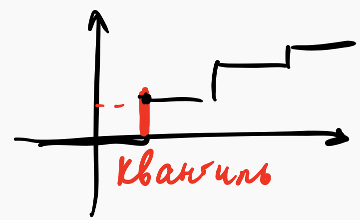
Математическое ожидание
Среднее (средневзвешенное по вероятностям) значение случайной величины.

$$\mathbb{E}(X) = \mathbb{P}(X = x_1)x_1 + \mathbb{P}(X = x_2)x_2 + \dots$$

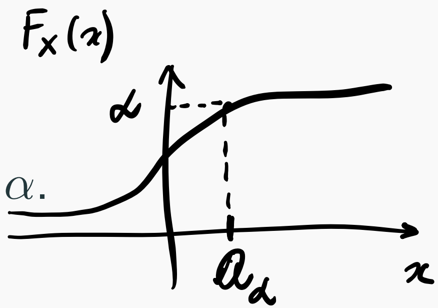
$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

Кв : 0.3
Перц : 30%

Квантиль (перцентиль)
Значение, которое случайная величина не превышает с фиксированной вероятностью (если вероятность в процентах, то называется перцентиль).

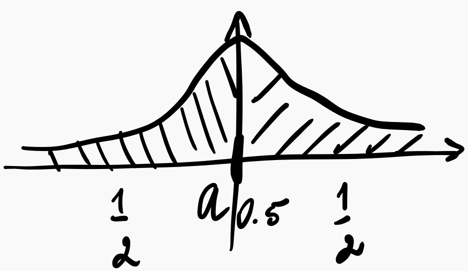


$$\begin{cases} \mathbb{P}(X \leqslant a_\alpha) \geqslant \alpha, \\ \mathbb{P}(X \geqslant a_\alpha) \geqslant 1 - \alpha. \end{cases}$$



Для непрерывных распределений: $F_X(a_\alpha) = \alpha$.

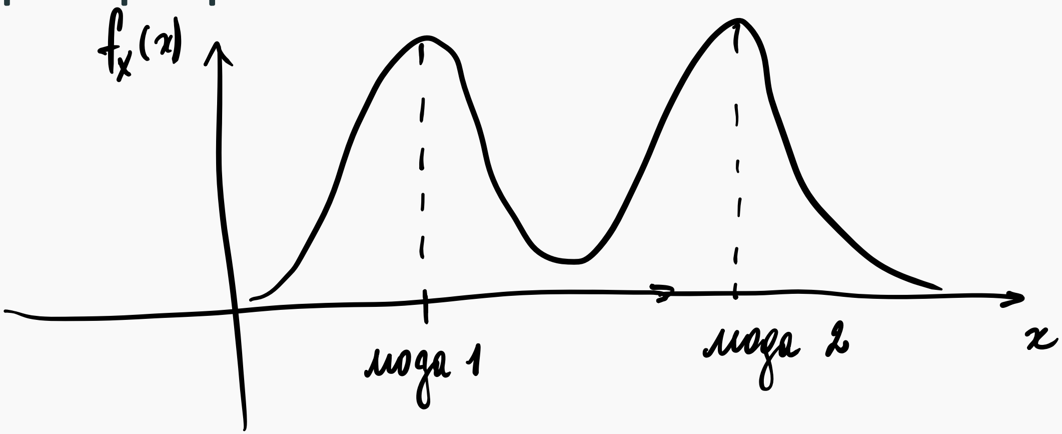
Медиана
Квантиль порядка 0.5.



Меры центральной тенденции

Мода (неформально)
Наиболее вероятное значение случайной величины.

Пример с ростом собак



грань	1	2	3	4	5	6
$\mathbb{P}\{X=x\}$	*	$\frac{1}{18}$	$\frac{5}{6}$	$\frac{1}{18}$	*	*

Меры разброса

Диспресия
Среднеквадратичное отклонение случайной величины относительно её математического ожидания.

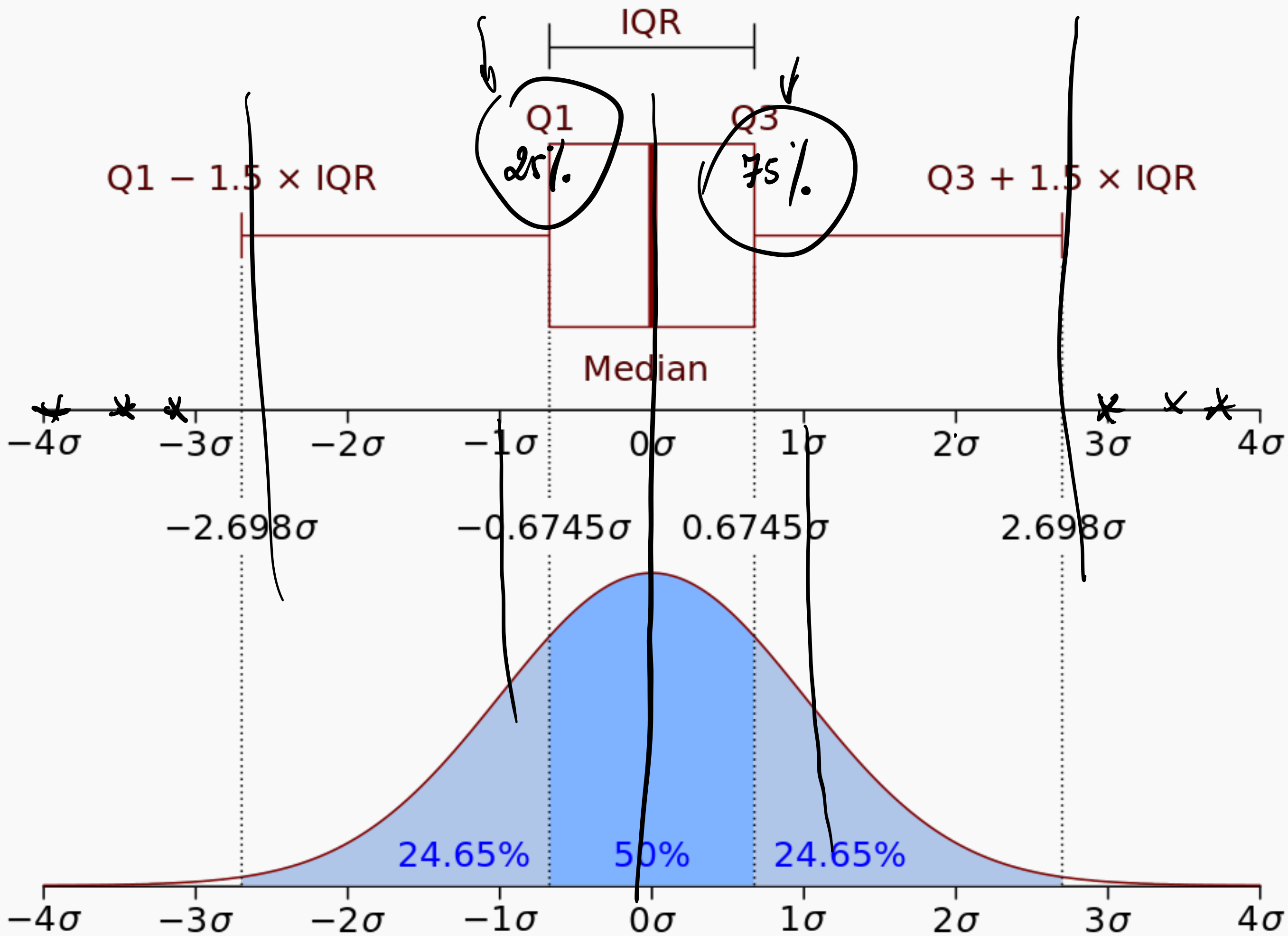
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

Стандартное отклонение
Корень из диспресии.

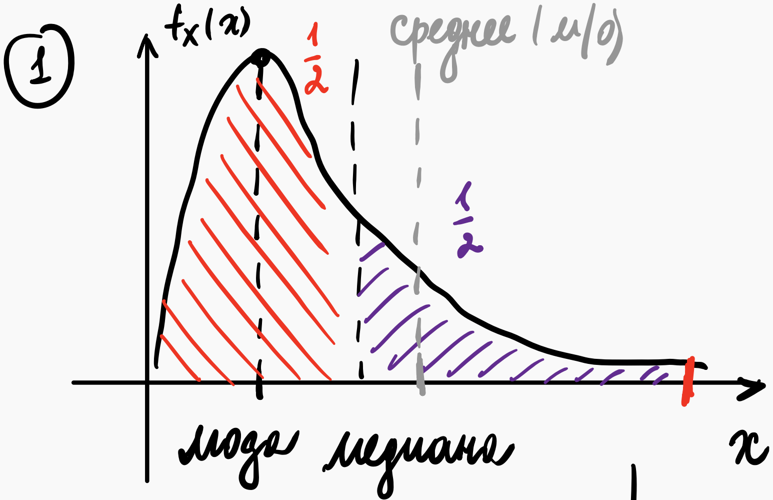
$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

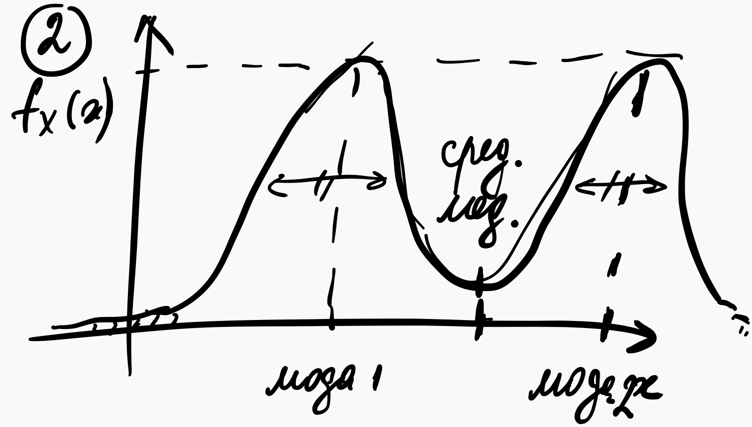
Размах
Разность между наибольшим и наименьшим значением случайной величины.

Обобщающая картинка (Wikimedia Commons)



Два примера, которые спрашивают на собеседованиях



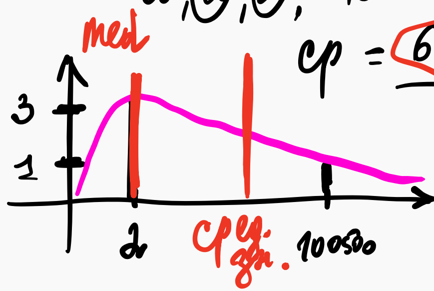


мода?
медиана?
мат. ож.?

$$med = \frac{2+2}{2} = 2$$

2, 2, 2, 100500

$$ср = \frac{6 + 100500}{4} > 2$$



Многообразие оценок

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

$$\frac{X_1 + \ldots + X_n}{n}$$

$$\frac{X_1 + X_n}{2}$$

$$\frac{X_2 + X_4 + \ldots + X_{2n}}{\#\,\text{кол-во}}$$

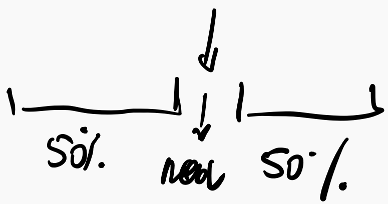
«Хорошие» оценки

$X = [X_1, X_2, \ldots, X_n]$, X_i – нез., один. распр. сл. вел.

$$\mathbb{E}(X) \longrightarrow \bar{X} = \frac{\sum_i X_i}{n}$$

$$Var(X) \longrightarrow \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

$\mathrm{med}(X) \Longrightarrow$ Вариац. ряд (X_i, упоряд. по возр.)



$$\hat{\mathbb{E}}(X) = \bar{X}$$

Пример: среднее число посетителей

- В магазин ежедневно приходят посетители. Требуется оценить некоторую величину, характеризующую среднее число посетителей за день.
- Для оценки будем записывать число посетителей в течение 100 дней:

$$\boxed{X_1, X_2, X_3 \ldots X_{100}.}$$

Пример: среднее число посетителей (стандартное решение)

- Очевидный вариант – математическое ожидание.
- В качестве оценки математического ожидания рассчитаем среднее арифметическое по всем наблюдениям:

$$\hat{\mathbb{E}}(X) = \frac{X_1 + X_2 + \ldots X_{100}}{100}.$$

- Оценка математического ожидания равна среднему. Можно показать, что при некоторых условиях $\hat{\mathbb{E}}$ является «хорошей» оценкой $\mathbb{E}$.

Пример: среднее число посетителей (нестандартное решение)

- А что если $X_1, \ldots, X_{97}$ все меньше 3, а $X_{98}, \ldots, X_{100}$ все больше 40 (дни распродаж)?
- В качестве оценки средней величины разумнее взять медиану.

$$\mathrm{med}(X_1 \ldots X_{100}) < 3$$

Ковариация и выборочная ковариация

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)] = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

$$\mathrm{sCov} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i = \bar{Y})$$