Сбор и анализ данных в Python

Seara

2021

1 Week 1

1.1 Основные понятия теории вероятностей

Пакт. X, Y, Z - случайные величины
x, y, z - какие-то конкретные значения
A, B, C - какие-то события
$\mathbb{P}$ - вероятность
$\mathbb{E}(X)$ - мат. ожидание
$Var(X)$ - дисперсия
$Cov(X, Y), \rho(X, Y)$ - ковариация и корреляция

1.2 Дискретные случайные величины

Случайная величина и её распределение. Случайные величины бывают:

- Дискретные(Множество значений конечно или счетно)
- Непрерывные(Принимают бесконечное, континуальное число значений)

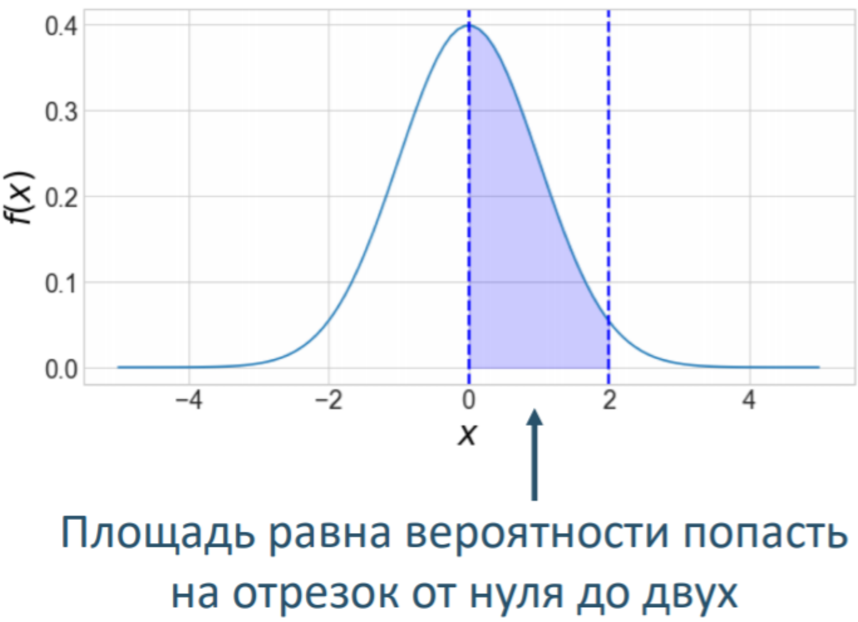
Распределение дискретной случайной величины - таблица, которая описывает, какие значения принимает случайная величина с какой вероятностью. Сумма вероятностей должна быть равна 1, каждая вероятность лежит между 0 и 1.
Функция распределения - функция, которая определяет вероятность события $X \le x$, то есть

$$F(X) = \mathbb{P}(X \le x) = \sum \mathbb{P}(X = k) \cdot [X \le x]$$

$$[X \le x] = \begin{cases} 1 & [X \le x] \\ 0 & otherwise \end{cases}$$

1.3 Непрерывные случайные величины

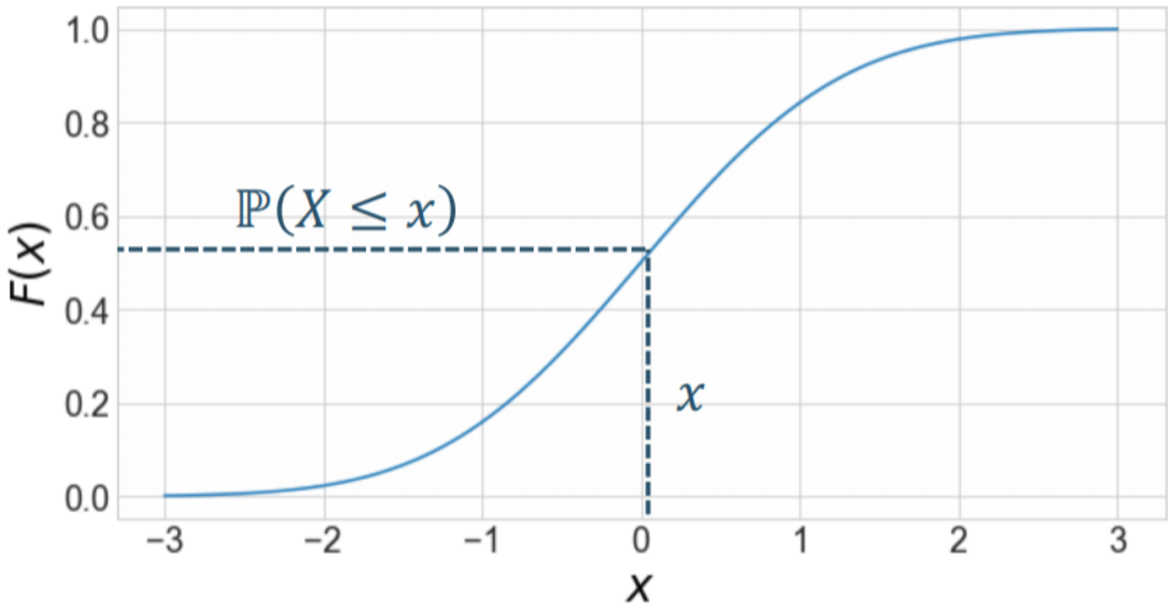
Плотность распределения. Распределение непрерывной случайной величины описывается плотностью распределения вероятностей.



Площадь под всей плотностью должна быть равна 1.

Функция распределения. Это функция, которая определяет вероятность события $X \le x$, то есть

$$F(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t)dt, f(t) - \ density\ function$$



Важные свойства.

- Плотность определениа только для непрерывных случайных величин. Для дискретных табличка.
- $f(x) = F'(x)$
- $\int_{-\infty}^{+\infty} f(t)dt = 1, f(t) \geq 0 \; \forall t$
- $F(X)$ не убывает и лежит между 0 и 1
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f(t)dt = F(b) - F(a)$
- Вероятность того, что непрерывная случайная величина попадет в точку, равна 0.

1.4 Характеристики случайных величин

Математическое ожидание. Математическое ожидание - среднее значение случайной величины.

$$\mathbb{E}(X) = \sum_{k=1}^{n} k \cdot \mathbb{P}(X = k)$$

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} t \cdot f(t)dt$$

Свойства мат ожидания.

- $\mathbb{E}(a) = a$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(a \cdot X) = a \cdot \mathbb{E}(X)$
- $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$, если независимы
- Математическое ожидание случайной величины не случайно.
- $\mathbb{E}(X - \mathbb{E}(X)) = 0$

Дисперсия. Дисперсия - мера разброса случайной величины вокруг её среднего.

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \sum_{k=1}^{n}(k - \mathbb{E}(X))^2 \cdot \mathbb{P}(X = k)$$

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \int_{-\infty}^{+\infty}(t - \mathbb{E}(X))^2 \cdot f(t)dt$$

Более удобное её искать по формуле

$$Var(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

Среднеквадратическое отклонение.

$$\sigma(X) = \sqrt{Var(X)}$$

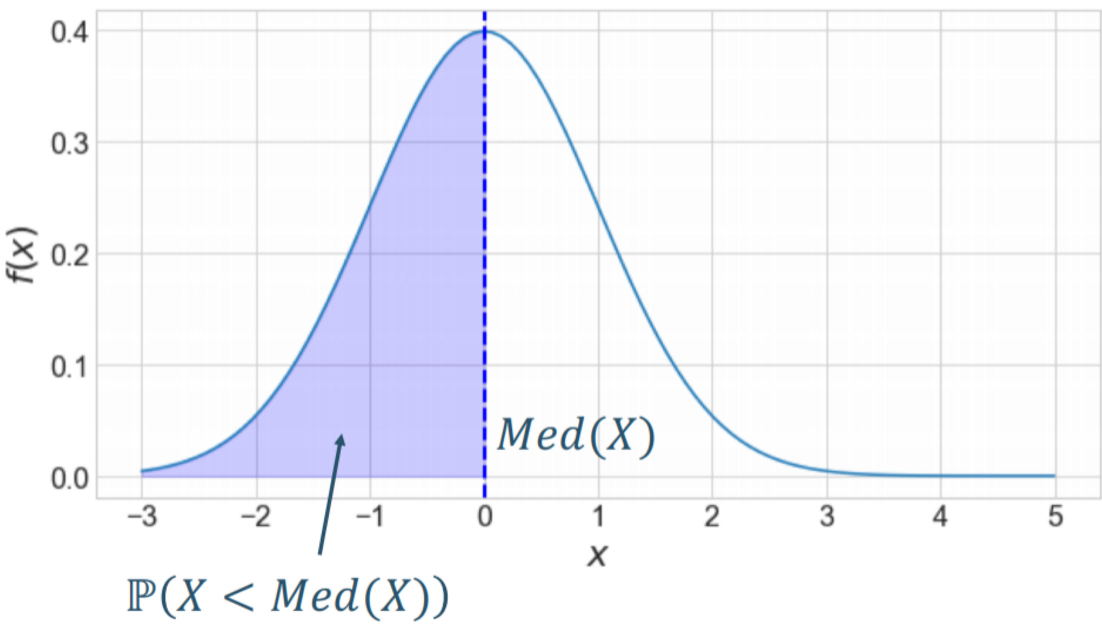
Свойства дисперсии.

- $Var(a) = 0$
- $Var(X + Y) = Var(X) + Var(Y)$, если независимы
- $Var(a \cdot X) = a^2 \cdot Var(X)$
- $Var(X - Y) = Var(X) + Var(Y)$, если независимы
- Дисперсия случайной величины не случайна

Мода случайной величины. Мода случайной величины - значение, которому соответствует наибольшая вероятность(для дискретной величины) и локальный максимум плотности распределения(для непрерывной случайной величны). По-другому это значение, которое встречается во множестве наблюдений наиболее часто.

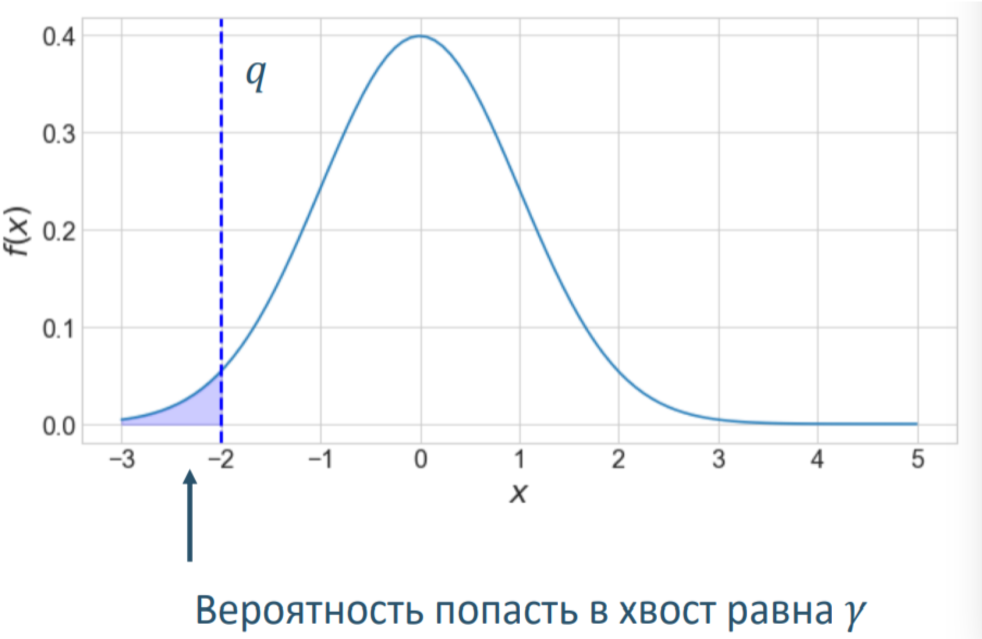
Медиана случайной величины. Медиана случайной величины - такое её значение, что

$$\mathbb{P}(X < Med(X)) = \mathbb{P}(X > Med(X)) = 0.5$$



Квантиль уровня γ. Квантиль уровня γ - это такое число q, что

$$\mathbb{P}(X \le q) = \gamma$$



1.5 Типы случайных величин

Распределение Бернулли. Случайная величина в нем принимает два значения 0 или 1. Например пол родившегося ребёнка. Задается вероятностью успеха p

$$X \sim Bern(p)$$

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

$$Var(X) = p \cdot (1 - p)$$

Биномиальное распределение. Число попаданий в баскетбольную корзину. Задается n - число испытаний и p - вероятность успеха.

$$X \sim Bin(p, n)$$

$$\mathbb{E}(X) = n \cdot p$$

$$Var(X) = n \cdot p \cdot (1 - p)$$

$$\mathbb{P}(X = k) = \binom{n}{k} \cdot p^k (1 - p)^{n-k}$$

Геометрическое распределение. Номер броска, когда произошло первое попадание в корзину. Задается p - вероятность успеха.

$$X \sim Geom(P)$$

$$\mathbb{E}(X) = \frac{1}{p}$$

$$Var(X) = \frac{1 - p}{p^2}$$

$$\mathbb{P}(X = k) = p \cdot (1 - p)^{k-1}$$

Произвольное дискретное распределние. Подбрасывание игральной кости. Задаётся таблицей.

Распределение Пуассона. Число автобусов, проехавших за час мимо остановки. Задается λ - интенсивность потока событий.

$$X \sim Poiss(\lambda)$$

$$\mathbb{P}(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

$$Var(X) = \lambda$$

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

Экспоненциальное распределение. Время до события. Нету памяти, время которое осталось ждать не зависит от того, сколько уже прошло.

$$X \sim Exp(\lambda)$$

$$f(x) = \lambda \cdot e^{-\lambda \cdot x},\ x \geq 0$$

$$F(x) = 1 - e^{-\lambda \cdot x},\ x \geq 0$$

$$\mathbb{E}(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

Равномерное распределение. Время рождения ребёнка

$$X \sim U[a; b]$$

$$f(x) = \frac{1}{b - a}, x \in [a; b]$$

$$F(x) = \frac{x - a}{b - a}, x \in [a; b]$$

$$\mathbb{E}(X) = \frac{a + b}{2}$$

$$Var(X) = \frac{(b - a)^2}{12}$$

Нормальное распределение. Погрешность весов.

$$X \sim N(\mu, \sigma^2)$$

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

$$Var(X) = \sigma^2$$

Перцентиль. Перцентиль порядка k - это такое число, что $k\%$ мерьше этого числа. Квартили - перцентили с шагом 0.25. Интерквартильный размах

$$IQR = x_{0.75} - x_{0.25}$$

1.6 Эмпирическая тема

Эмпирическая функция распределения. Эмпирическая функция распределения - функция, которая определяет для каждого x частоту события $X \leq x$. Похоже на функцию распределения дискретной случайной величины.

Гистограмма. Гистограмма - эмпирическая оценка плотности распределения. По оси x откладывают значения. По оси y частоты. Область возможных значений обычно дробят на отрезки, бины. Чем короче бины тем детальнее рисуется гистограмма.

2 Week 2

2.1 Независимость

Говорят, что случайные величины X и Y независимы, если

$$F(x, y) = F_X(x) \cdot F_Y(y)$$

В терминах плотностей:

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

Ковариация. Ковариация - мера линейной зависимости двух случайных величин

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

По аналогии с дисперсией, более простая формула:

$$Cov(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Свойства ковариации.

- $Cov(X, Y) = Cov(Y, X)$
- $Cov(a, b) = 0$
- $Cov(a \cdot X, Y) = a \cdot Cov(X, Y)$
- $Cov(X + a, Y) = Cov(X, Y)$
- $Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)$
- $Cov(X, X) = Var(X)$
- Если случайные величины независимы: $Cov(X, Y) = 0$
- Обратное неверно
- Если X и Y зависимы, то $\mathbb{E}(X, Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y) + Cov(X, Y)$
- $Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$

Корреляция Пирсона.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X) \cdot \sigma(Y)}$$

Коэффициент корреляции характеризует тесноту и направленность линейной связи между случайными величинами и принимает значение от -1 до 1. Корреляция Пирсона улавливает только линейную взаимосвязь и чувствительна к выбросам.

Корреляция Спирмена. Корреляция Спирмена - мера силы монотонной взаимосвязи. Вычисляется как корреляция Пирсона между рангами наблюдений.

Пример:

	X	Y
Выборка:	10, 8, 6, 7, 4, 10, 9, 5	9, 9, 4, 5, 6, 8, 10, 7
Порядок:	7, 5, 3, 4, 1, 8, 6, 2	6, 7, 1, 2, 3, 5, 8, 4
Ранг:	7.5, 5, 3, 4, 1, 7.5, 6, 2	6.5, 6.5, 1, 2, 3, 5, 8, 4
	r_x	r_y

$$\hat{\rho}_s(X,Y) = \hat{\rho}_p(r_x, r_y) \approx 0.645$$

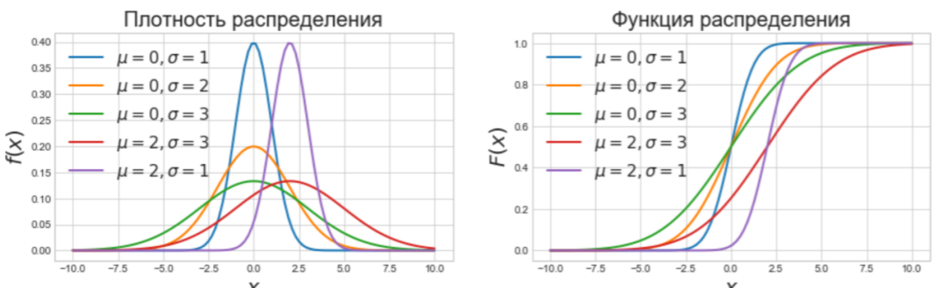
Корреляция Спирмана пытается уловить в данных монотонность.

Нормальное распределение. Нормальная случайная величина

Нормальная случайная величина: $X \sim N(\mu, \sigma^2)$

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

Функцию распределения нельзя найти в аналитическом виде, интеграл не берётся



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad F(x) = \int_{-\infty}^{x} f(x)\ dx$$

Свойства нормального распределения.

- Распределение симметрично относительно точки $\mathbb{E}(X)$
- Параметр μ не влияет на форму кривой и отвечает за её сдвиг кривой вдоль оси x, параметр σ определяет степень размытости кривой.
- $X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$
- $X + a \sim N(\mu_x + a, \sigma_x^2)$

- $a \cdot X \sim N(a \cdot \mu_x, a^2 \cdot \sigma_x^2)$

Центрирование и нормирование.

$$X \sim N(\mu, \sigma^2)$$

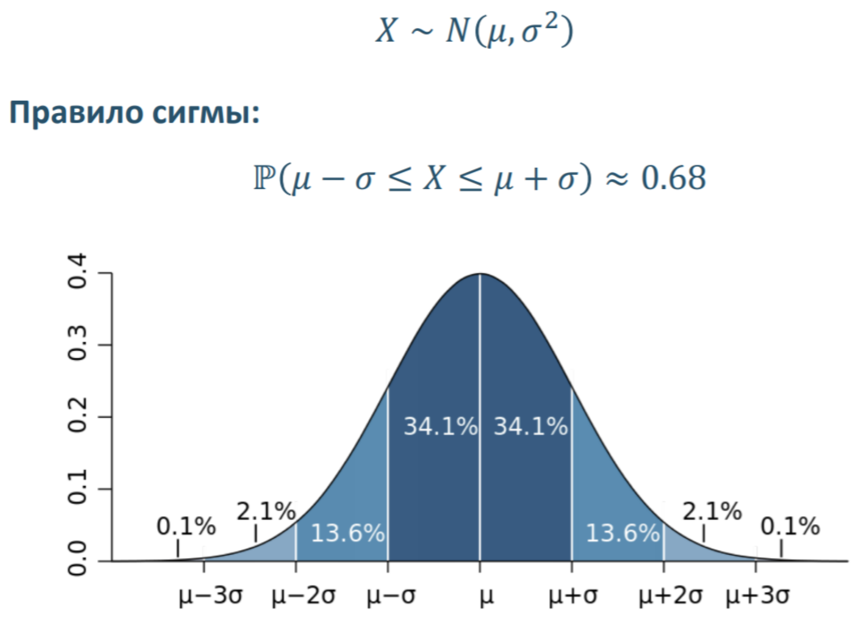
Центрирование

$$X - \mu \sim N(0, \sigma^2)$$

Нормирование

$$\frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

Распределение $N(0,1)$ называется **стандартным нормальным распределением.** Для функции распределения случайной величины $N(0,1)$ составлены таблицы.
Правила сигм.



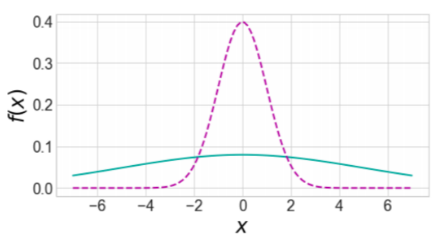
Эксцесс и куртосис. Куртосис

$$\frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4}$$

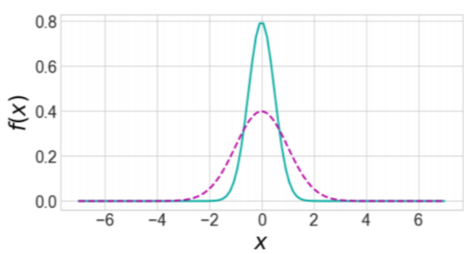
Эксцессом случайной величины X называют величину

$$\beta_x = \frac{\mathbb{E}[(X - \mathbb{E}(X))^4]}{\sigma^4} - 3$$

Число 3 вычитается из курстосиса, чтобы эксцесс нормального распределения был равен 0. Если хвосты распределения легче, а пик острее, чем у нормального распределения, тогда $\beta_X > 0$. Если хвосты распределения тяжелее, а пик более приплюснутый, тогда $\beta < 0$.



Отрицательный эксцесс

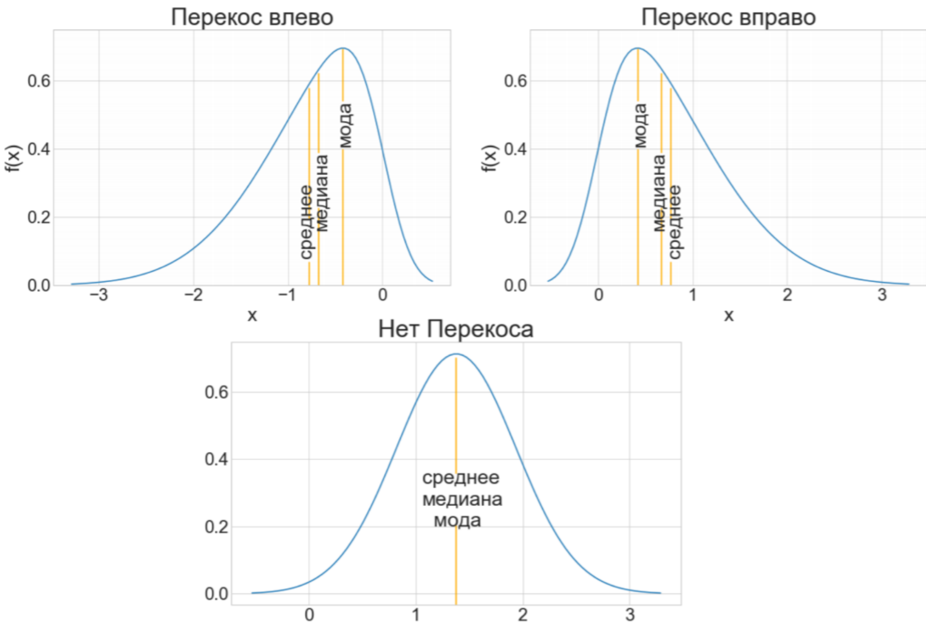


Положительный эксцесс

Коэффициент асимметрии (skewness). Коэффициентом ассиметрии случайной величины X назвают величину

$$A_x = \frac{\mathbb{E}[(X - \mathbb{E}(X))^3]}{\sigma^3}$$

Если плотность распределения симметрична, то $A_X = 0$. Если левый хвост тяжелее, то $A_X > 0$. Если праый хвост тяжелее, то $A_X < 0$



Многомерное нормальное распределение.

- По аналогии можно определить нормальное распределение для любой размерности

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N(\mu, \Sigma)$$

$$\mu = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \mathbb{E}(X_3) \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) \end{pmatrix}$$

2.2 Ядерные оценки плотности.

Гистограмма. Гистограмма простейший непараметрический способ получить оценку плотности распределения.

$$f_n(x) = \frac{1}{n \cdot h} \cdot \sum [z_k < x_i \leq z_{k+1}]$$

Размер бина(длина отрезка)

$$h = z_{k+1} - z_k$$

Добавим скользящие границы

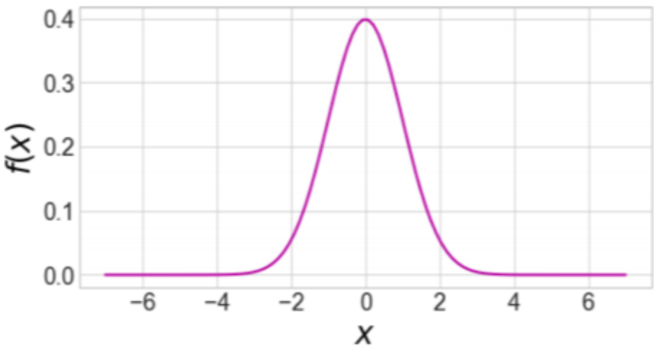
$$f_n(x) = \frac{1}{n \cdot h} \cdot \sum [x - \frac{h}{2} < x_i \le x + \frac{h}{2}]$$

Перепишем в более удобном виде

$$f_n(x) = \frac{1}{n \cdot h} \cdot \sum K(\frac{x - x_i}{h}) \quad K(z) = [-\frac{1}{2} < z \le \frac{1}{2}]$$

Такая функция придаёт каждому наблюдению вес 0 или 1. Функция $K(z)$ называется ядерная функция. Она должна быть неотрицательной и $\int K(Z)dz = 1$ (сумма всех весов равна 1). Ядерные функции бывают разные, чаще всего используют Гауссовское ядро:

$$K(z) = \frac{1}{\sqrt{2\,\pi}} \cdot e^{-\frac{x^2}{2}}$$

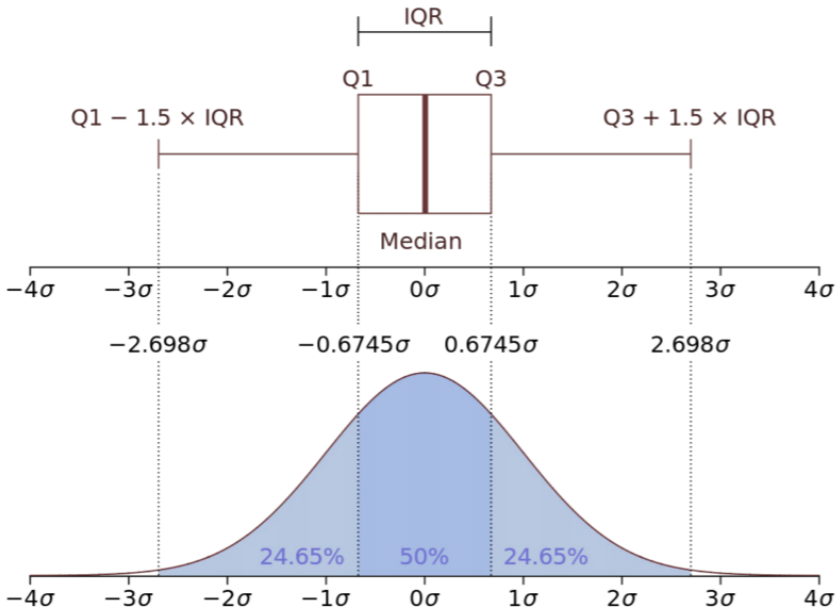


2.3 Пропуски в данных.

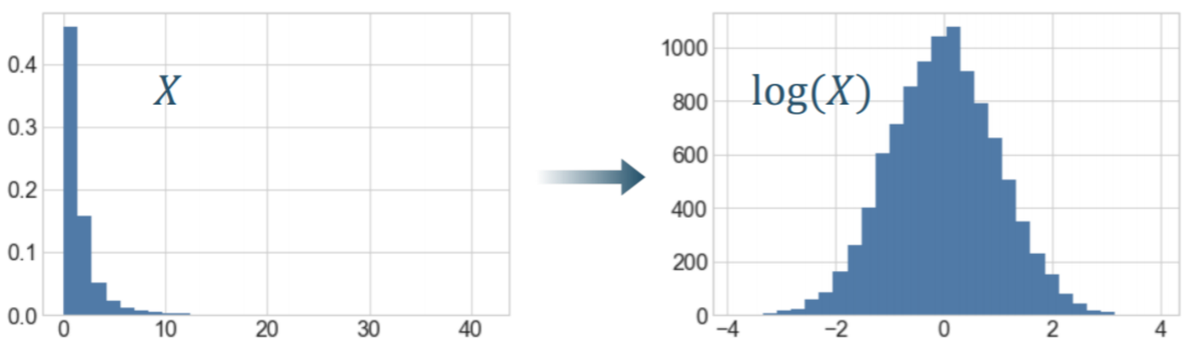
Борьба с пропусками

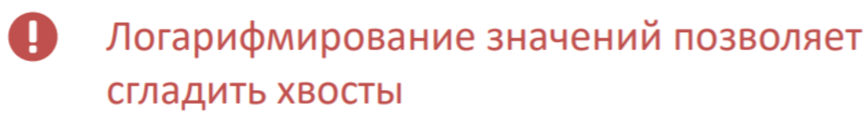
- Простые методы: замена средним, медианой, модой
- Сложные методы: основаны на машинном обучении, смотрят на другие примеры и пытаются предсказать что было пропущено.

Выбросы. Выброс - результат измерений, который сильно выделяется на общем фоне. Многие алгоритмы чувствительны к выбросам и переобучаются под них. Можно воспользоваться **правилом трёх сигм** и избавиться от всех выбросов, что выходят за границы. Так же существоет правило 1.5 интерквартильных размахов (IQR): Если наблюдения оказались за пределами выделенного интервала, они выбросы. Иногда используют 3 IQR.

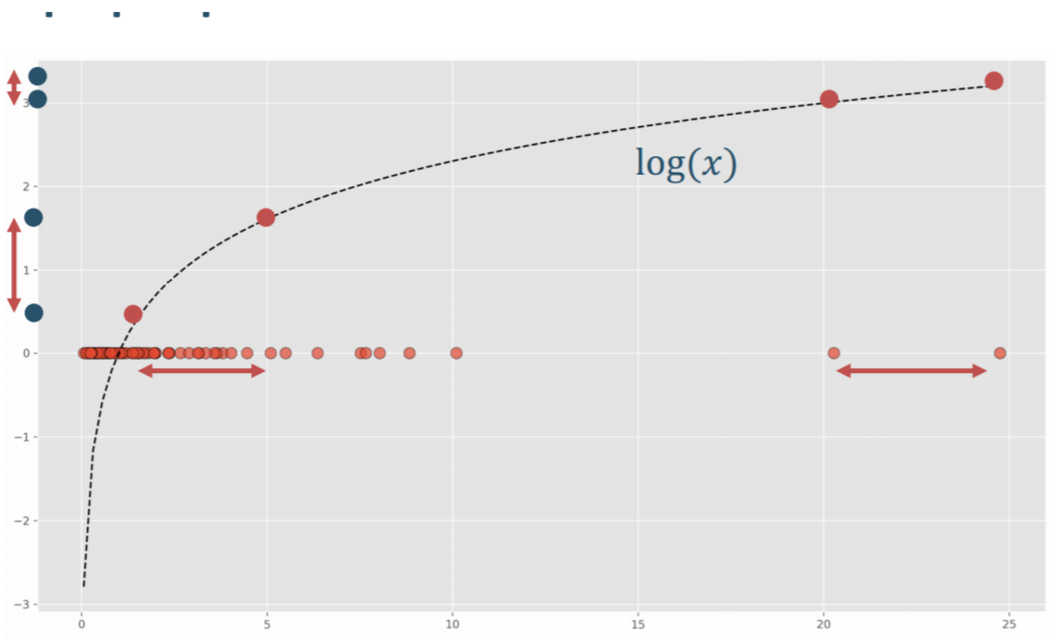


Преобразование данных. Логарифмирование





Логирафм обладает затухающим свойством, поэтому разница больших величин не такая большая как разница маленьких.



Сгустки в начале оси абсцисс стали распределены более равномерно из-за того, что там логарифм растет быстрее. Расстояние между точками с большими значениями стало меньше, так как там логарифм растет медленее.

Преобразование Бокса-Кокса. Иногда скорость роста дефолтного логарифма не подходит. Нам поможет преобразование Бокса-Кокса. Параметр p надо подобрать.

$$x_i^* = \begin{cases} \log\,(x), & p = 0 \\ \dfrac{x^p - 1}{p}, & 0 \le p \le 1 \end{cases}$$

- Параметр p можно выбрать, максимизируя корреляцию между квантилями нормального распределения и x^*
- Если в выборке есть отрицательные значения, можно сдвинуть её в положительную область

$$x_i^* = \begin{cases} \log\,(x+\alpha), & p = 0 \\ \dfrac{(x+\alpha)^p - 1}{p}, & 0 \le p \le 1 \end{cases}$$

2.4 Масштабироваание и категориальные переменные

Способы масштабирования.

Нормализация (Standard Scaler):

$$x_i^* = \frac{x_i - \bar{x}}{\hat{\sigma}}$$

Масштабирование на отрезок [0; 1] (Minmax Scaler):

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Устойчивая к выбросам нормализация (Robust Scaler):

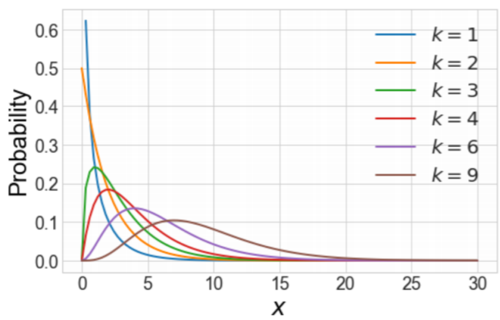
$$x_i^* = \frac{x_i - med(X)}{Q_3 - Q_1}$$

Категориальные переменные. One-hot encoding(бинарное кодирование) топ, главное не попасть в **Dummy-ловушку**. Dummy-ловушка - ситуация, когда набор закодированных столбцов в сумме даёт колонку из единиц. Возникает линейная зависимость и методы плохо работают. Поэтому надо всегда $drop\_first = True$. Количество столбцов должно быть $k-1$ от количества уникальных значений категориального признака.

3 Week 5

3.1 Распределение хи-квадрат

Случайные величины $X_1, ..., X_k \sim iid\ N(0,1)$.
Случайная величина $Y = X_1^2 + ... + X_k^2 \sim \chi_k^2$ имеет 'хи-квадрат' распределение с k степенями свободы. (iid - identically independently distributed)



$X_1, \dots, X_k \sim iid\ N(0,1)$

$Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$

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:

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

$Var(X) = 2k$

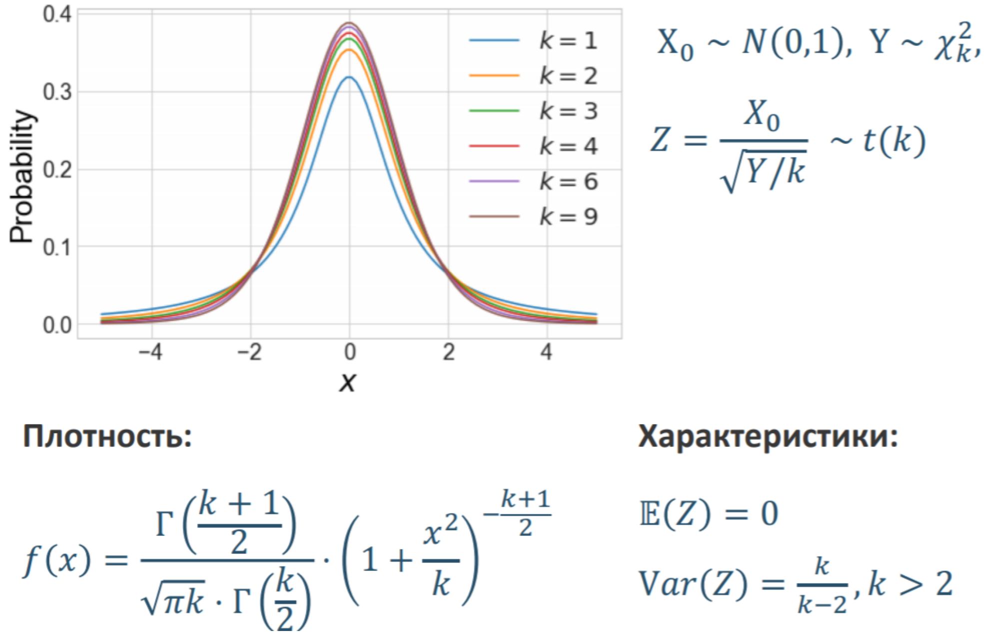
Число степеней свободы. Число степеней свободы - количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.

3.2 Распределение Стьдента

Независимые случайные величины $X_0 \sim N(0,1),\ Y \sim \chi^2_k$. Тогда случайная величина

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

имеет распределение Стьюдента с k степенями свободы.



Распределение Стьюдента обладает более тяжелыми хвостами, нежели нормальное. По мере возрастания k, кривая функции плотности все больше напоминает стандартное нормальное распределение.

3.3 Распределение Фишера

Независимые случайные величины $X \sim \chi^2_k, X \sim \chi^2_m$. Случайная величина

$$Z = \frac{X/k}{Y/m} \sim F(k,m)$$

имеет распределение Фишера с k, m степенями свободы.

Вывод. Распределения хи-квадрат, Стьюдента, Фишера часто встречаются на практике при анализе нормально распределённых выборок.

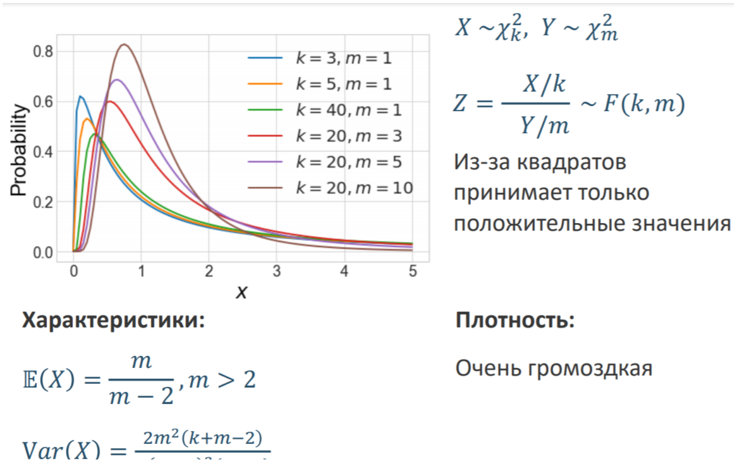
4 Закон больших чисел

ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин 'стабилизируется' с ростом их числа.

Слабая форма ЗБЧ. Пусть $X_1, ..., X_n$ попарно независимые и одинакого распределённые случайные величины с конечной дисперсией $Var(X_1) < \infty$ тогда:

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

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



Сходимость по вероятности. Последовательность случайных величин $X_1, ..., X_n, ...$ сходится по вероятности к случайной величине X, если

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

То есть:

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

Свойства сходимости по вероятности

Можно выносить константу за знак предела:

$$\operatorname*{plim}_{n\to\infty}(c \cdot X_n) = c \cdot \operatorname*{plim}_{n\to\infty} X_n, \qquad c \in \mathbb{R}$$

Предел суммы – сумма пределов:

$$\operatorname*{plim}_{n\to\infty}(X_n + Y_n) = \operatorname*{plim}_{n\to\infty} X_n + \operatorname*{plim}_{n\to\infty} Y_n$$

Предел произведения – произведение пределов:

$$\operatorname*{plim}_{n\to\infty}(X_n \cdot Y_n) = \operatorname*{plim}_{n\to\infty} X_n \cdot \operatorname*{plim}_{n\to\infty} Y_n$$

Сходимость не портится из-за непрерывных функций

$$\operatorname*{plim}_{n\to\infty} g(X_n) = g(\operatorname*{plim}_{n\to\infty} X_n), g(t) - \text{непрерывная}$$

4.1 Центральная предельная теорема

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному.
Пусть $X1, ..., X_n$ попарно независимые и одинакого распределённые случайные величины с конечной дисперсией $Var(X_1) < \infty$ тогда:

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

d над стрелкой означает сходимость по распределению.

Сходимость по распределению. Последовательность случайных величин $X_1, ..., X_n, ...$ сходится по распределению к случайной величине X, если

$$\lim_{n\to\infty} F_{X_n}(x) = F_x(x),$$

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

ЗБЧ vs. ЦПТ

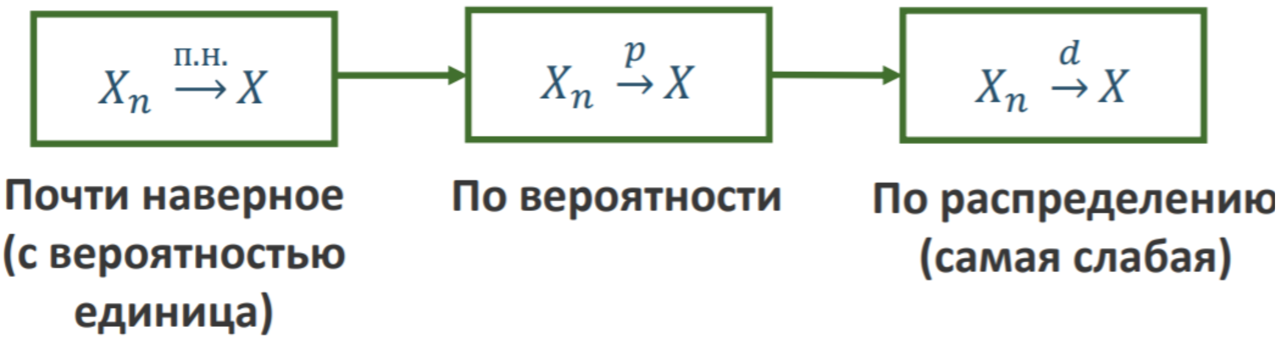
ЗБЧ: $$\frac{X_1 + \ldots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

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

4.2 Виды сходимостей



Сходимость «почти наверное» самая сильная из трёх, сходимость «по распределению» самая слабая

Сходимость почти наверное. Последовательность случайных величин $X_1, ..., X_n, ...$ сходится почти наверное (с вероятностью единица) к случайной величине X, если

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

то есть у последовательности есть предел с вероятностью 1.

Сильная форма ЗБЧ. Пусть $X_1, ..., X_n$ последовательность независимых и одинакого распределённых случайных величин с $\mathbb{E}(|X_1|) < \infty$ тогда:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow{a.s.} \mathbb{E}(X_1)$$

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