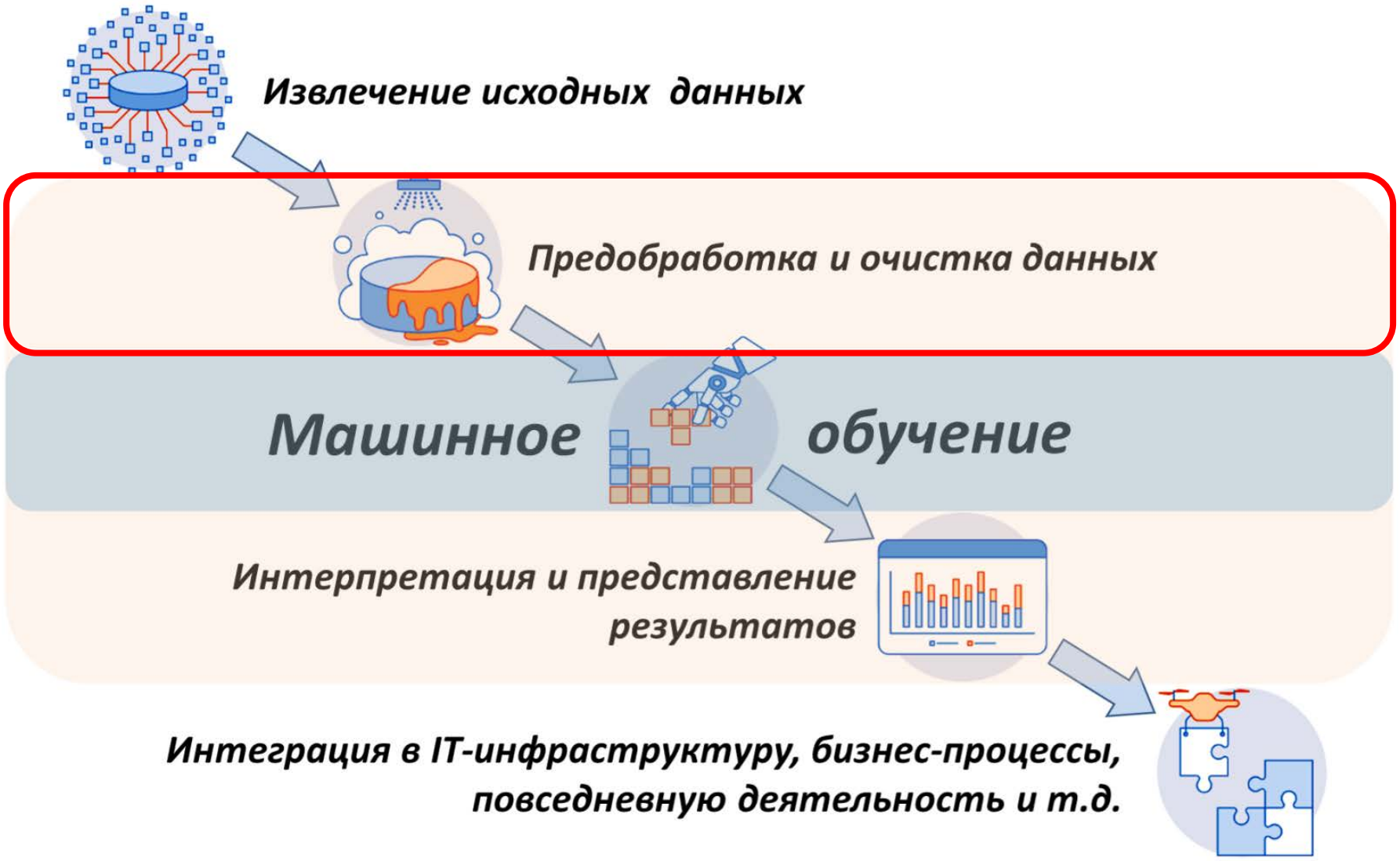
Методы машинного обучения

Лекция 2

Исходные данные

Предобработка и очистка данных



Возможные проблемы исходных данных

- Малый объем обучающей выборки;
- Некорректность входных данных;
  - Неполные
  - Неточные
- Противоречивость данных;
- Разнородность признаков;
- Неструктурированные/отсутствует разметка.

Риски, связанные с постановкой задачи:

- «грязные» данные – заказчик не обеспечивает качество данных;
- Неясные критерии качества модели – заказчик не определился с целями или индикаторами.

Учет пропусков

- Исключение объектов/признаков, имеющих неполные сведения (удаление строк/столбцов);
- Заполнить при помощи интерполяции;
- Найти в других источниках и тем самым дополнить данные;
- Закодировать пропуски специальным значением;
- Привлечь эксперта в соответствующей предметной области:
  - использование специфичных математических моделей
  - генерация псевдослучайных значений, подчиняющиеся некому распределению с учетом других известных признаков;
  - генерация синтетических данных.

Увеличение информативности примеров для повышения скорости и эффективности обучения

Еще одной целью предобработки данных, является увеличение информативности примеров для повышения скорости и эффективности обучения. Чем больше бит информации принесет каждый пример, тем лучше используются имеющиеся данные. Среднее количество информации, приносимой каждым примером *x*, равно энтропии распределения значений этой компоненты *H(x)*. Если эти значения сосредоточены в относительно небольшой области единичного интервала, информационное содержание такой компоненты мало и когда все значения переменной совпадают, эта переменная не несет никакой информации. Напротив, если значения переменной *x* равномерно распределены в единичном интервале, информация такой переменной максимальна.

Таким образом, общий принцип предобработки данных состоит в таком кодировании и нормировке непротиворечивых данных, чтобы добиться максимизации энтропии входов и выходов.

Кодирование входов-выходов

Качественные данные можно разделить на две группы:

- **упорядоченные (ординальные - от англ. order);**
- **неупорядоченные (категориальные)**.

В обоих случаях переменная относится к одному из дискретного набора классов $\{c_1,\ldots,c_n\}$. Но в первом случае эти классы упорядочены, т.е. можно сказать, что $c_1<\ldots<c_n$, тогда как во втором такая упорядоченность отсутствует.

Кодирование - Ординальные переменные

Ординальные переменные близки к числовой форме и для их кодирования достаточно, просто пронумеровать имеющиеся значения переменных числами, таким образом, который бы сохранял существующую упорядоченность. Самым простым способом является установка в соответствие каждому классу своего целого номера, отличающегося на 1 от соседних номеров. Так, например, классу c_1 соответствует номер 1, а классу c_n - номер n.

Но может иметься проблема неравномерности выборки, а нам необходимо стремиться к тому, чтобы максимизировать энтропию закодированных данных, что достигается использованием равномерного распределения.

Исходя из этих соображений, единичный отрезок разбивается на n отрезков, что соответствует числу классов, с длинами пропорциональными числу примеров каждого класса в обучающей выборке: $\Delta x_k = \frac{P_k}{P}$, где P_k - число примеров класса k, а P - общее число примеров. Центр каждого такого отрезка будет являться численным значением для соответствующего ординального класса.

Кодирование – Неупорядоченные (категориальные) переменные

Неупорядоченные (категориальные/номинальные) переменные являются простым обозначением классов.

- Двоичное кодирование (иногда называют one-hot-кодированием). Каждому значению ставится в соответствие собственный атрибут, который может принимать значение 0 или 1, т.е. при наличии значения соответствующий ему атрибут устанавливается равным 1 или в противном случае 0.

  Первый класс кодируется как (1,0,…,0), второй – (0,1,0,…,0) и т.д.

  Если кодируемый параметр может принимать значения *n* классов, то размерность входного вектора увеличится на *n-1* элемент.

- Кодирование *n* классов неупорядоченных переменных *m*-битным двоичным кодом.

  Пример: c_1=(0,0), c_2=(0,1), c_3=(1,0), c_4=(1,1)

  Размерность входного вектора увеличится на $log_2 n$ элементов.

Кодирование выходных значений

Кодирование выходных значений направлено на увеличение эффективности обучения и упрощение интерпретации результатов работы модели.

- Метод кодирования выходных значений с помощью двоичного вектора, при использовании которого количество выходных значений равняется количеству классов, где *i*-ая компонента вектора соответствует *i*-ому классу;
- номер кластера, записанный в двоичной форме, т.е. кодировать *n* классов *m*-битным двоичным кодом;
- Разбиении задачи с *n* классами на *k* подзадач с двумя классами каждая. $k = A_n^2 = n(n-1)/2$

Выходы	Возможные классы
1	1-2
2	1-3
3	1-4
4	2-3
5	2-4
6	3-4

Класс	Содержится в выходе
1	1,2,3
2	1,4,5
3	2,4,6
4	3,5,6

Нормировка данных

Стандартизация данных – это процесс приведения вектора каждого признака к такому виду, что его математическое ожидание станет нулевым, а дисперсия – единичной.

Нормализация данных – это процесс масштабирования вектора каждого признака, то есть приведение его к такому виду, что вектор будет иметь единичную норму (при этом есть разные способы оценки\подсчета нормы).

Линейное преобразование:

- $\tilde{x}_i = \frac{x_i - x_i^{min}}{x_i^{max} - x_i^{min}}$ в единичный отрезок: $\tilde{x}_i \in [0,1]$
- $\tilde{x}_i = 2\frac{x_i - x_i^{min}}{x_i^{max} - x_i^{min}} - 1$ для отображения данных в интервал [-1,1]

L1 норма: $x_i' = \frac{x_i}{\|x\|_1} = \frac{x_i}{\sum_j |x_j|}$, где $\|x\|_1$ и есть L1 норма, а вся формула целиком отображает процесс нормализации вектора x.

L2 норма: $x_i' = \frac{x_i}{\|x\|_2} = \frac{x_i}{\sqrt{\sum_j x_j^2}}$, где $\|x\|_2$ и есть L2 норма, а вся формула целиком отображает процесс нормализации вектора x.

Нормировка данных на основании статистических характеристик

Другой формой масштабирования является вычисление для каждого признака среднего значения и среднеквадратичного отклонения, т.е. статистических характеристик.

Выборочное среднее (несмещённая оценка математического ожидания E [X]):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Выборочная дисперсия в математической статистике — это оценка теоретической дисперсии распределения, рассчитанная на основе данных выборки. Пусть $X = \{x_1 \dots x_n\}$ – выборка из распределения вероятности.

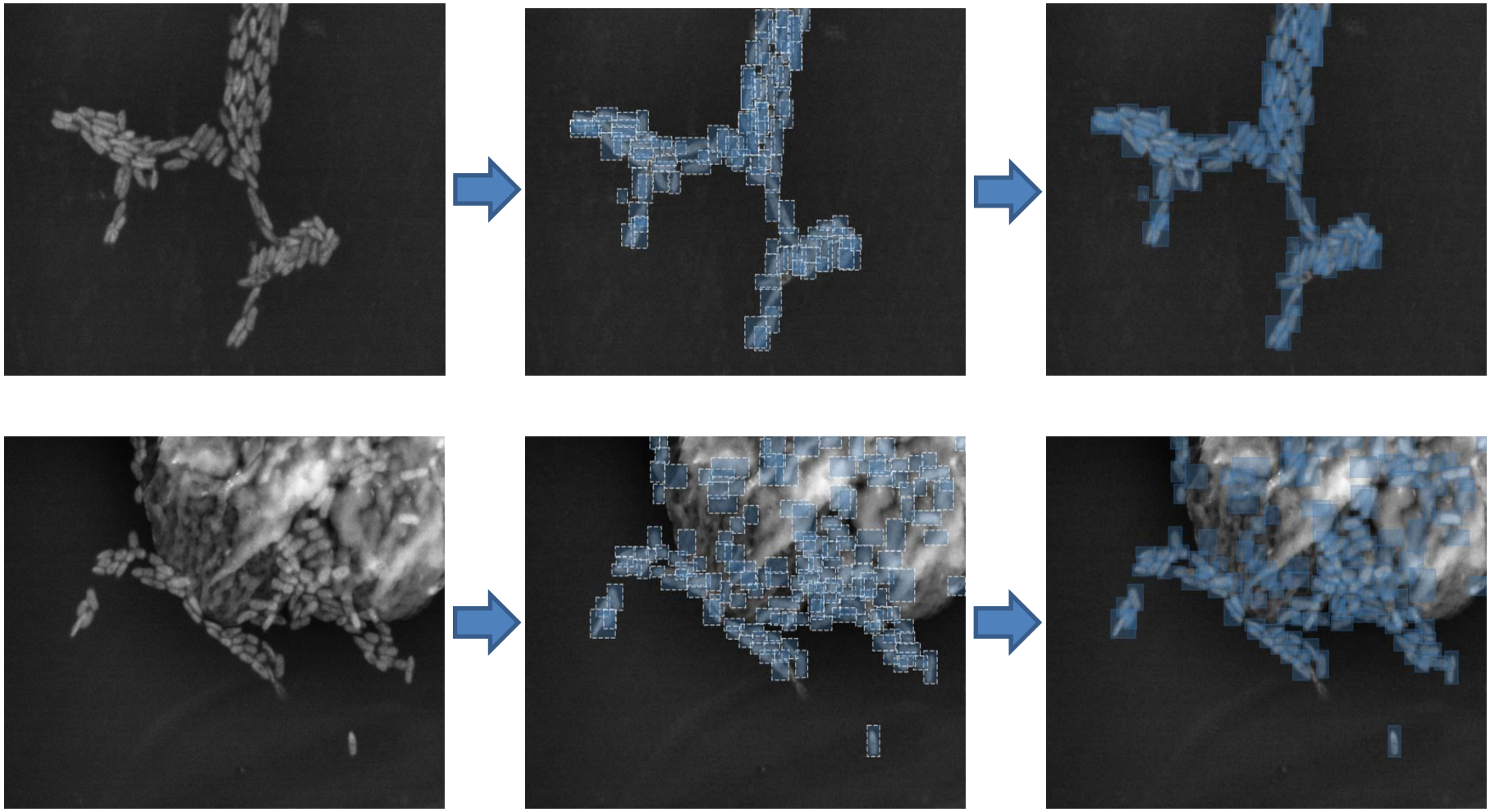
Виды выборочных дисперсий:

- Смещённая: $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Несмещённая, или исправленная: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

Значения признака масштабируются с помощью вычитания среднего и деления результата на значение среднеквадратичного отклонения для данного признака.

$$\tilde{X}_i = \frac{X_i - \bar{X}}{S^2}$$

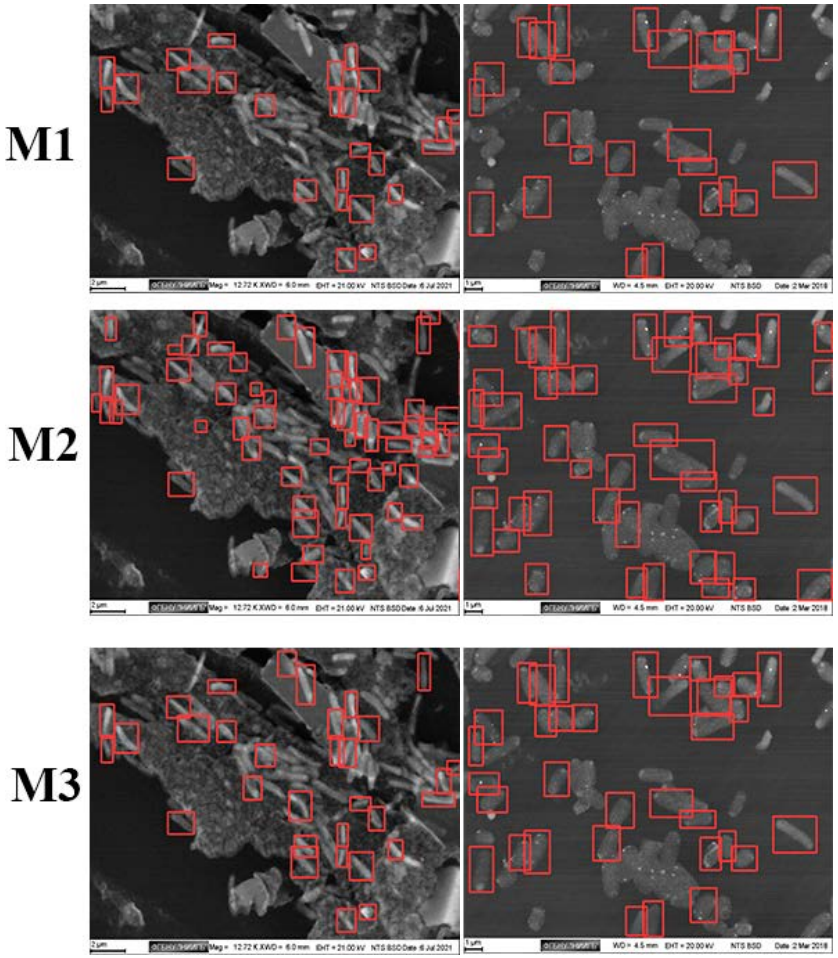
Разметка данных / изображений

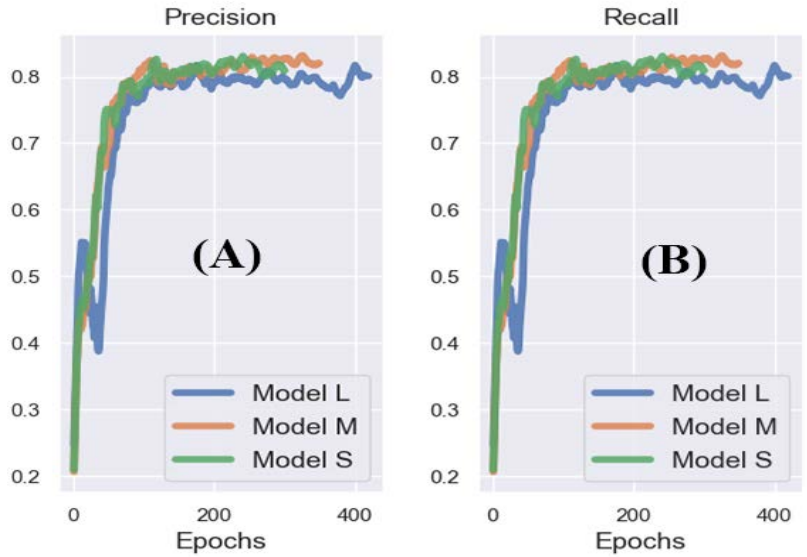


Автоматизированная разметка изображений

Обучение моделей для автоматизированной разметки изображений

Результаты разметки бактерий ранга Pseudomonas aeruginosa и Escherichia_coli





Метрики качества модели	Модели		
	S	M	L
Точность	0.911130	0.920790	0.909750
Полнота	0.836680	0.829020	0.809520
Номер лучшей эпохи	217	257	327
Всего эпох обучения	309	358	428

Метрики качества модели	Модели		
	M1	M2	M3
Точность	0.883990	0.909750	0.911380
Полнота	0.869170	0.809520	0.817360
Номер лучшей эпохи	245	327	9
Всего эпох обучения	328	428	107

Теория вероятностей и математическая статистика

Теория вероятностей – раздел математики, в котором изучаются случайные величины и события, их свойства и возможные операции над ними. Таким образом, ключевое понятие, лежащее в основе этой дисциплины, – вероятность события *Pi*.

Математическая статистика (Mathematical statistics) – это наука о математических методах анализа данных, полученных на основе большого числа наблюдений (измерений, опытов). Большая часть математической статистики основана на **вероятностных моделях**.

Основными задачами математической статистики являются оценивание и проверка гипотез.

Методы теории вероятностей

Опираясь на понятие вероятности события *Pi*, можно дать определение **полной группы событий**, которая определяется как система случайных событий и обладает следующими свойствами:

- В результате случайного эксперимента непременно произойдет одно и только одно из составляющих ее событий;
- Сумма вероятностей всех событий полной группы равна 1.

Среди базовых операций по анализу данных можно выделить подсчеты вероятностных характеристик выборки: медианы, математического ожидания, дисперсии, а также величины среднеквадратического отклонения.

Вероятностные характеристики выборки

$X = \{x_1 \dots x_n\}$ – изучаемая выборка

- **Медиана** – число, характеризующее выборку по среднему из ее значений. То есть, если все данные выборки различны, и она упорядочена по возрастанию, то ровно половина из элементов выборки будет меньше медианы, и ровно половина – больше. x_k – медиана, если $x_j < x_k$, при $j \in [1, k)$ и $x_k < x_i$, при $i \in (k, n]$ и *k=n/2*.

- **Математическое ожидание** – среднее значение вероятностных элементов выборки. Формально она рассчитывается так:

$$M|X| = \sum x_i p_i$$

- **Выборочное среднее** (несмещённая оценка мат. ожидания E [X]).

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- **Дисперсия** – мера разброса элементов выборки относительно ее математического ожидания. Рассчитывается данная величина по формуле

$$D|X| = \sum (x_i - M|X|)^2 \, p_i$$

- **Выборочная дисперсия** и **среднеквадратическое отклонение** – величина, характеризующая рассеивание значений выборки относительно ее математического ожидания. Формула расчета

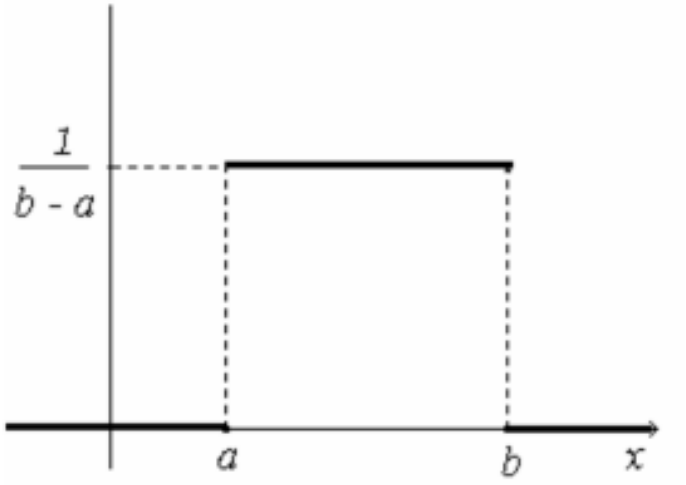
$$\sigma = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

Непрерывное равномерное распределение

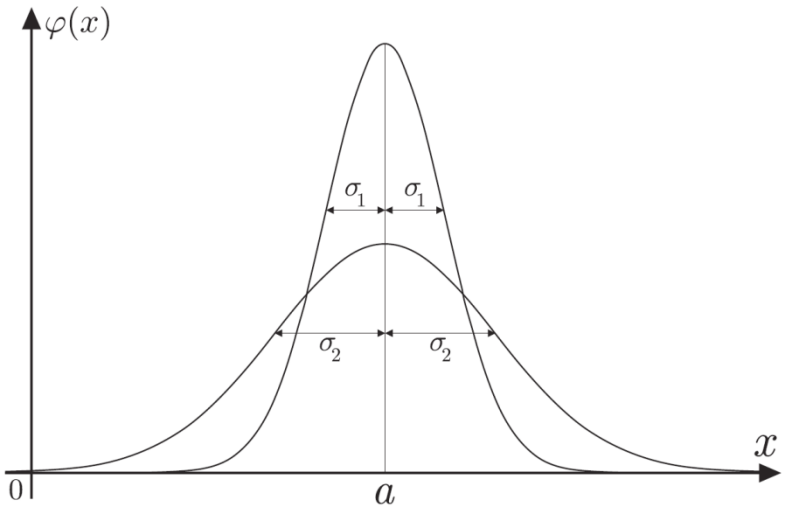
Непрерывное равномерное распределение в теории вероятностей — распределение случайной вещественной величины, принимающей значения, принадлежащие некоторому промежутку конечной длины, характеризующееся тем, что плотность вероятности на этом промежутке почти всюду постоянна.

Говорят, что случайная величина имеет непрерывное равномерное распределение на отрезке [a,b], где $a, b \in \mathbb{R}$, если ее плотность $f_X(x)$ имеет вид:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$$



Нормальное распределение



Определение. Случайная величина ξ имеет нормальное распределение вероятностей с параметрами μ и σ^2, если ее Закон плотности распределения вероятности задается формулой:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

где σ – среднеквадратическое отклонение;

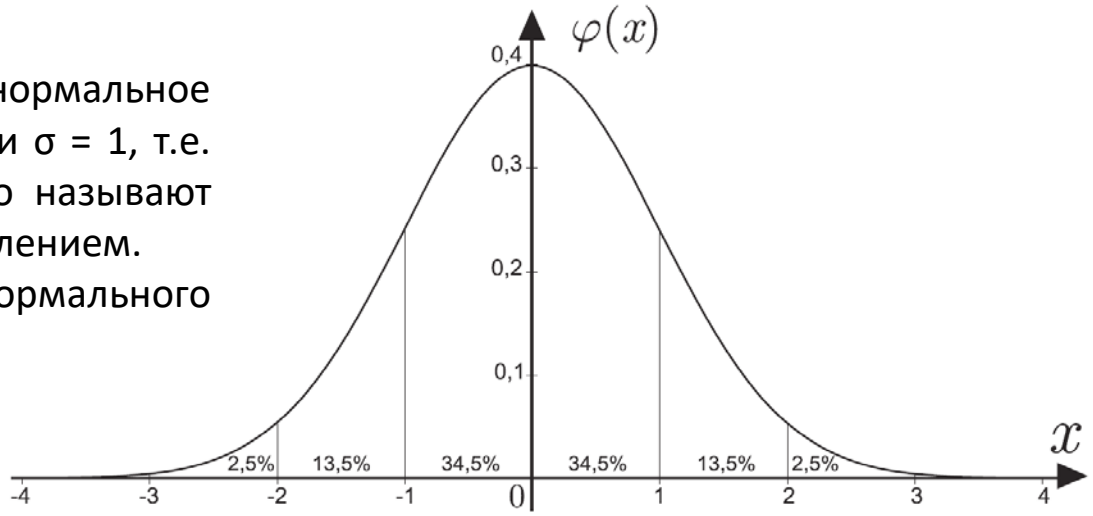
μ – математическое ожидание.

Краткое обозначение: $\xi \sim N(\mu, \sigma^2)$

Особую роль играет нормальное распределение с параметрами $\mu = 0$ и σ = 1, т.е. распределение N(0, 1), которое часто называют *стандартным* нормальным распределением.

Плотность стандартного нормального распределения:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Распределения, связанные с нормальным

Область применения. При операциях с нормальными случайными величинами, которые приходится проводить при анализе данных, возникает несколько новых видов распределений (и соответствующих им случайных величин). В первую очередь, это распределение **Стьюдента, χ^2 и *F*-распределения**. Эти распределения играют очень важную роль в прикладном и теоретическом анализе. Так, при выяснении точности и достоверности статистических оценок используются процентные точки распределений Стьюдента и χ^2 . Распределение статистик многих критериев, использующихся для проверки различных предположений, хорошо приближается этими распределениями.

Распределение хи-квадрат

Определение. Пусть случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ – независимы, и каждое из них имеет стандартное нормальное распределение N(0,1). Говорят, что случайная величина χ_n^2 определенная как:

$$\chi_n^2 = \xi_1^2 + \cdots + \xi_n^2$$

имеет распределение хи-квадрат с n степенями свободы.

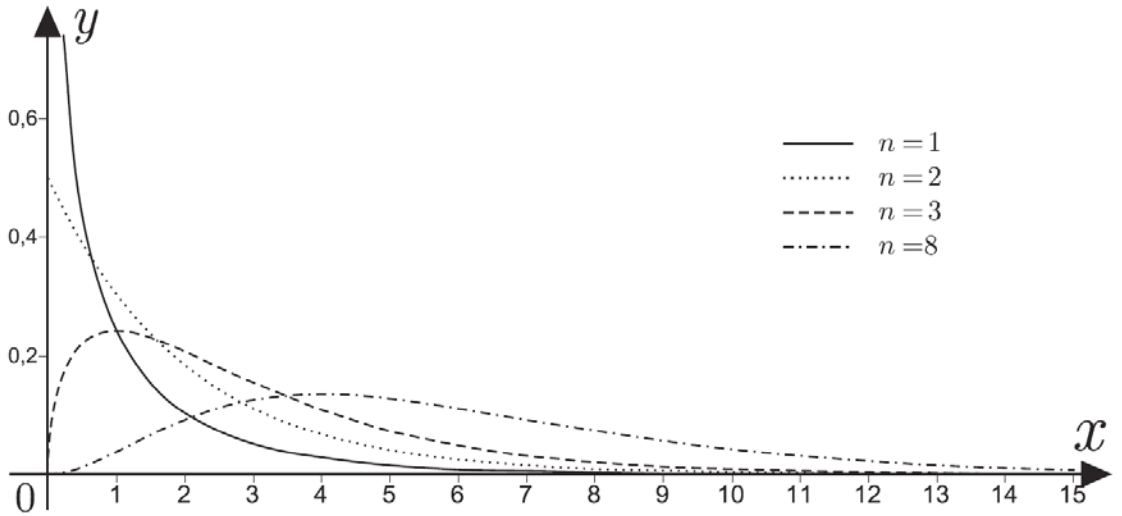
Функция плотности χ_n^2 в точке x (x>0) равна:

$$\frac{1}{2^{n/2}} \frac{1}{\Gamma(n/2)} x^{\frac{n}{2}-1} e^{x/2}$$

где $\Gamma(\cdot)$ есть гамма функция. На практике эта плотность распределения непосредственно используется редко.

Свойства. Математическое ожидание и дисперсия случайной величины χ_n^2 равны:

$$M\chi_n^2 = n,\ D\chi_n^2 = 2n.$$



Распределение Стьюдента

Определение. Пусть случайные величины $\xi_0, \xi_1, \dots, \xi_n$ – независимы, и каждое из них имеет стандартное нормальное распределение N(0,1).

Введем случайную величину: $t_n = \dfrac{\xi_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \xi_i^2}}$

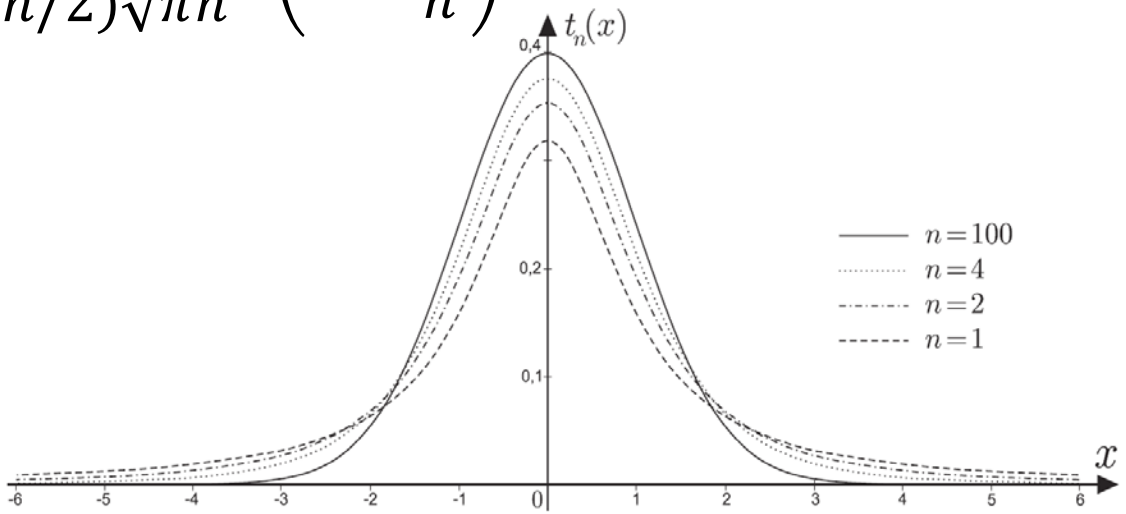
Ее распределение называют распределением Стьюдента. Саму случайную величину часто называют стьюдентовской дробью или стьюдентовым отношениеми. Число n = 1, 2, ... называют числом степеней свободы распределения Стьюдента.

Плотность распределения Стьюдента в точке x равна:

$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

Свойства:

$Mt_n = 0,\ Dt_n = \dfrac{n}{n-2}.$



F-распределение

Определение. Пусть $\eta_1, \dots, \eta_m; \xi_1, \dots, \xi_n$ (где *m,n* – натуральные числа) обозначают независимые случайные величины, каждое из которых распределено по стандартному нормальному закону распределения *N(0,1)*. Говорят, что случайная величина $F_{m,n}$, определенная как

$$F_{m,n} = \frac{\frac{1}{m}(\eta_1^2 + \dots + \eta_m^2)}{\frac{1}{n}(\xi_1^2 + \dots + \xi_n^2)}$$

имеет *F*-распределение с параметрами *m* и *n*. Натуральные числа *m,n* называют числами степеней свободы. *F*-распределение иногда называют еще распределением дисперсионного отношения.

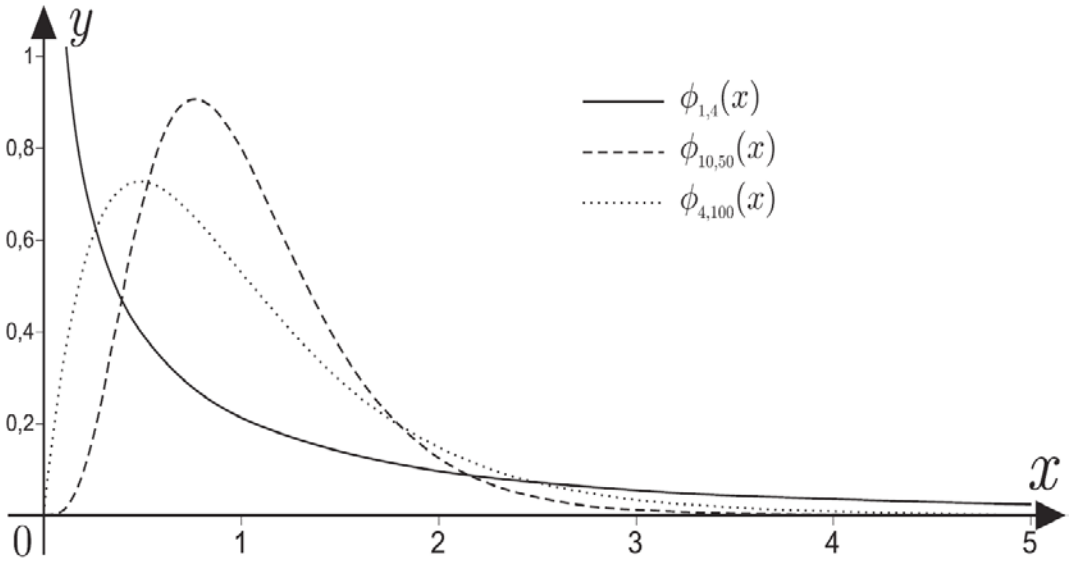
Свойства.

Математическое ожидание для $n > 2$:

$$MF_{m,n} = \frac{n}{n-2}$$

Дисперсия для $n > 4$:

$$DF_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$



Спасибо за внимание