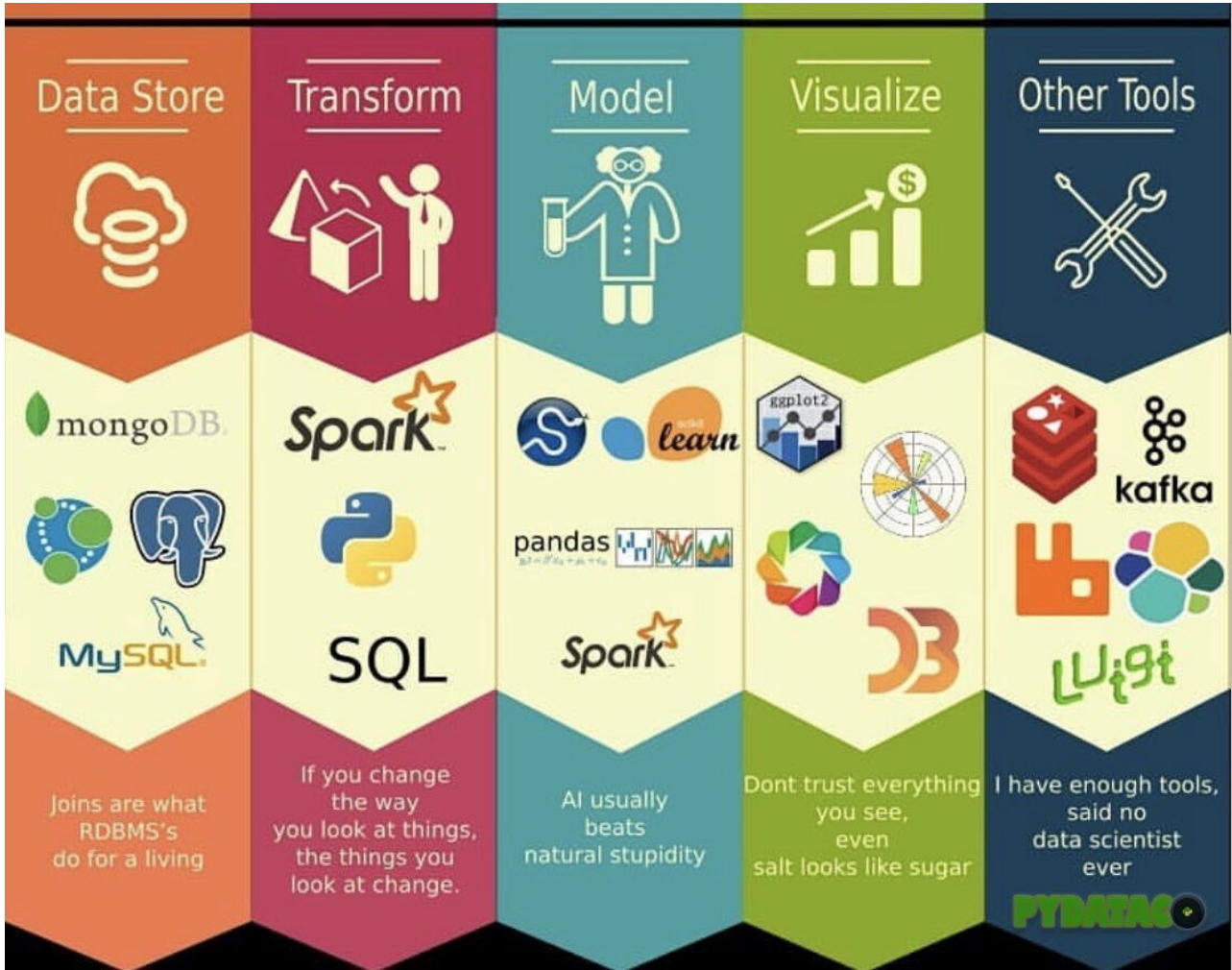
Разведочный анализ данных

Майнор: Статистический анализ и его применение. 2020

I. Модели и первичный анализ статистических данных

Лекция 1

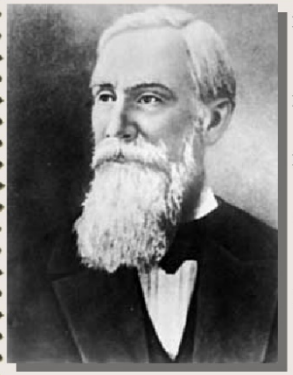


Немного истории и Ключевые понятия

Reverend Thomas Bayes [beɪz]) **– Преподобный Томас Бейз (Байес, Бейес)** c. 1702 – 17.04.1761

Пьер Симон Лаплас (1749-1827)

Пафнутий Львович Чебышев (1821-1894)



Пафнутий Львович начальное образование получил дома. В 1837 поступил в Московский университет, в 1846 защитил магистерскую диссертацию *Опыт элементарного анализа теории вероятностей.*

В 1847 был приглашен в Петербургский университет на кафедру математики, где читал лекции по алгебре и теории чисел.

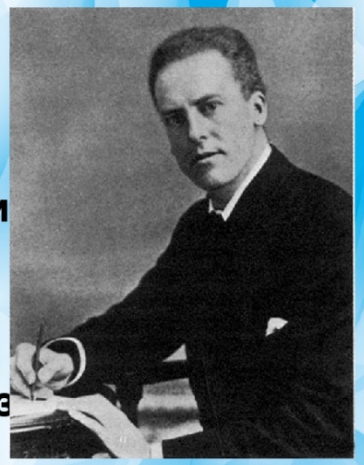
В 1849 вышла книга Чебышева *Теория сравнений*, по которой он в том же году защитил докторскую диссертацию в Петербургском университете.

Чебышев сумел создать новые направления в разных областях: теории вероятностей, теории приближения функций многочленами, интегральном исчислении, теории чисел и т.д.

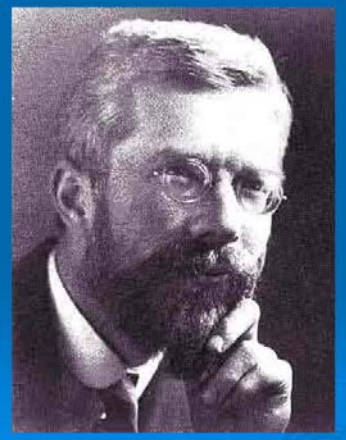
Известны работы ученого в области математического анализа. Среди прикладных задач, которыми занимался Чебышев, — построение точных географических карт, вопросы деформации поверхностей, вопросы теоретической и практической механики.

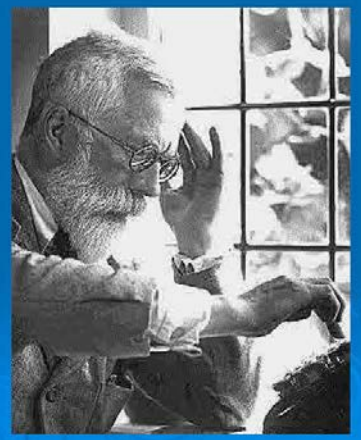
В 1878 Чебышев изобрел счетную машину нового типа (хранится в Музее искусств и ремесел во Франции).

Карл Пирсон - английский математик, статистик, биолог и философ; основатель математической статистики, один из основоположников биометрики.



Рональд Фишер





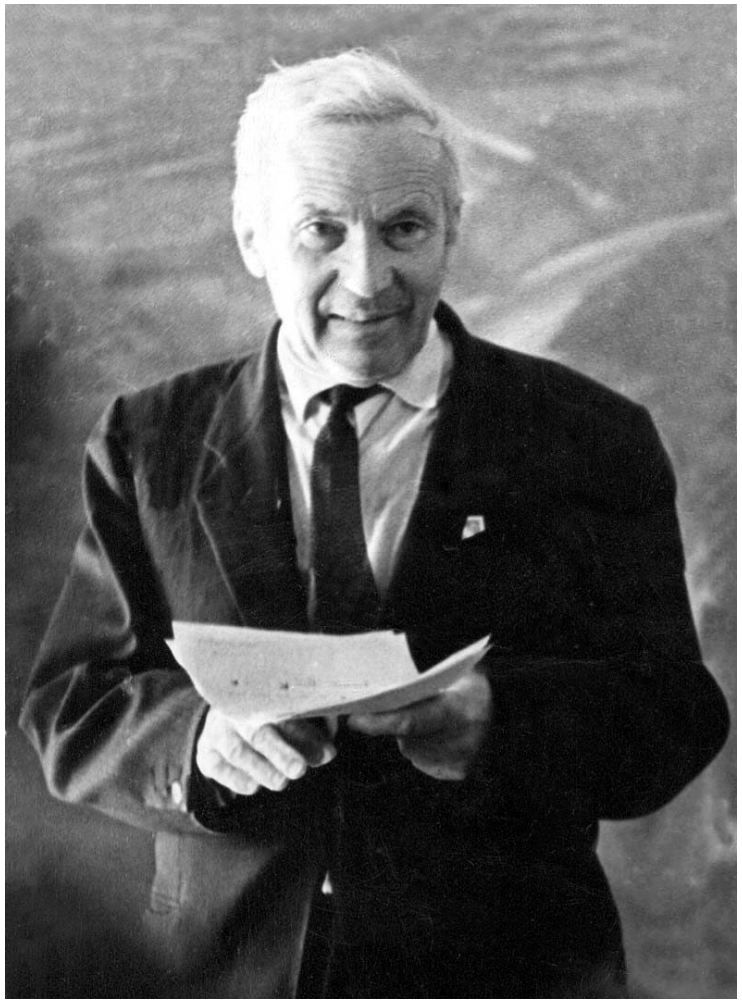
Марков А.А. (1856-1922)

Расширил область применения закона больших чисел и центральной предельной теоремы, распространив их на зависимые опыты.

Назад

Джон Тьюки

Андрей Николаевич Колмогоров



Элементы структуированных данных

Типы данных

Непрерывные данные (continuous)

Данные, которые могут принимать любое значение в интервале.

Синонимы: интервал, число с плавающей точкой, числовое значение.

Дискретные данные (discrete)

Данные, которые могут принимать только целочисленные значения, такие как количественные значения.

Синонимы: целое число, количество.

Категориальные данные (categorical)

Данные, которые могут принимать только определенный набор значений, в частности набор возможных категорий.

Синонимы: перечисления, перечислимые данные, факторы, именованные данные, полихотомические данные.

Двоичные данные (binary)

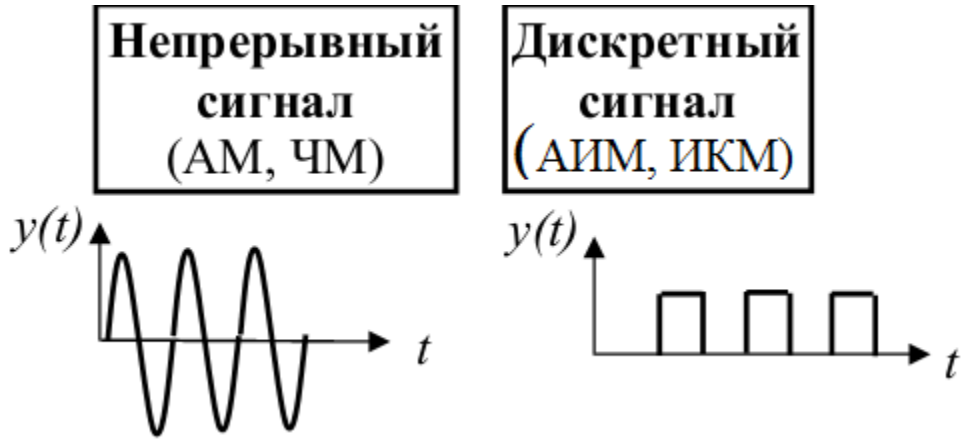
Особый случай категориальных данных всего с двумя категориями значений (0/1, истина/ложь).

Синонимы: дихотомический, логический, флаг, индикатор, булево значение.

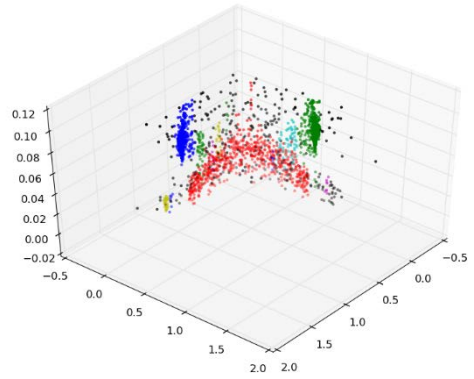
Порядковые данные (ordinal)

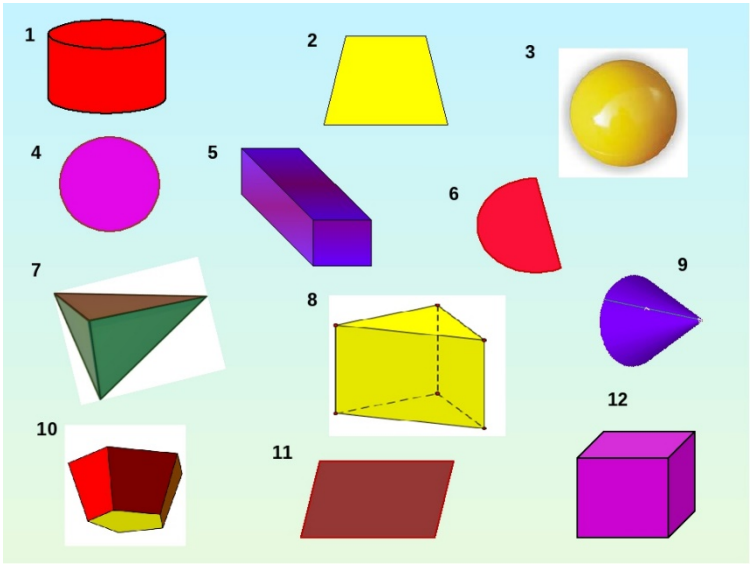
Категориальные данные с явно выраженной упорядоченностью.

Синонимы: порядковый фактор.



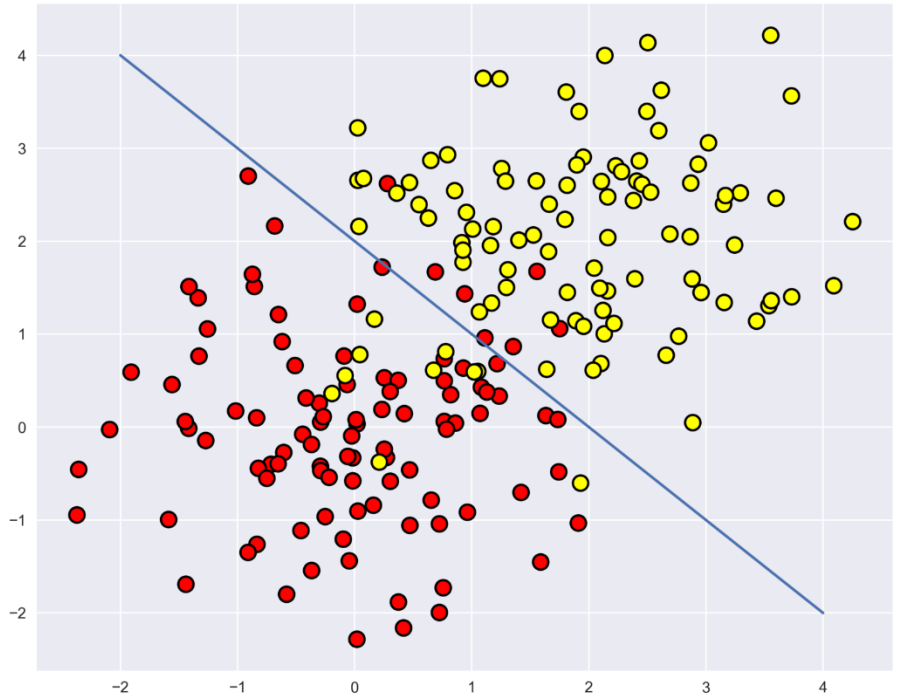
Категориальные данные





Порядковые данные

Бинарные данные



Категория персональных данных:

- **Указываются основные категории персональных данных,** ***например:***

Ст. 3 ФЗ № 152. ПЕРСОНАЛЬНЫЕ ДАННЫЕ:

- Фамилия, имя, отчество
- Дата, месяц, год рождения
- Адрес
- Доходы
- Сведения об образовании
- Сведения о профессии
- Социальное положение
- Сведения о семейном положении
- Сведения об имущественном положении
- Другая информация (см. слайд 9)

Ст. 10 ФЗ № 152. СПЕЦИАЛЬНЫЕ КАТЕГОРИИ ПЕРСОНАЛЬНЫХ ДАННЫХ:

- Расовая принадлежность
- Национальная принадлежность
- Политические взгляды
- Философские убеждения
- Религиозные убеждения
- Сведения о состоянии здоровья

Ст. 11 ФЗ № ФЗ 152. БИОМЕТРИЧЕСКИЕ ДАННЫЕ:

(сведения, которые характеризуют физиологические особенности человека: сведения дактилоскопической регистрации: отпечатки пальцев, ладони; результаты анализа ДНК; образ лица; сетчатка глаза; особенности строения тела, отдельных органов и тканей; отклонения в развитии; атавизмы; психическое состояние здоровья)

(http://www.r-tutor.com/r-introduction/basic-data-types).

- Дополнительная информация типам данных

Кадр данных (data frame)

Прямоугольные данные (как в электронной таблице) — это типичная структура данных для статистических и машинно-обучаемых моделей.

Признак (feature)

Столбец в таблице принято называть *признаком.*

Синонимы: атрибут, вход, предиктор, переменная.

Исход (outcome)

Многие проекты науки о данных сопряжены с предсказанием *исхода* — нередко в формате да/нет (например, в табл. 1.1 это ответ на вопрос "были ли торги состязательными или нет?"). Для предсказания *исхода* в эксперименте или статистическом исследовании иногда используются *признаки.*

Синонимы: зависимая переменная, отклик, цель, выход.

Записи (records)

Строку в таблице принято называть *записью*.

Синонимы: случай, образец, прецедент, экземпляр, наблюдение, шаблон, паттерн, выборка.

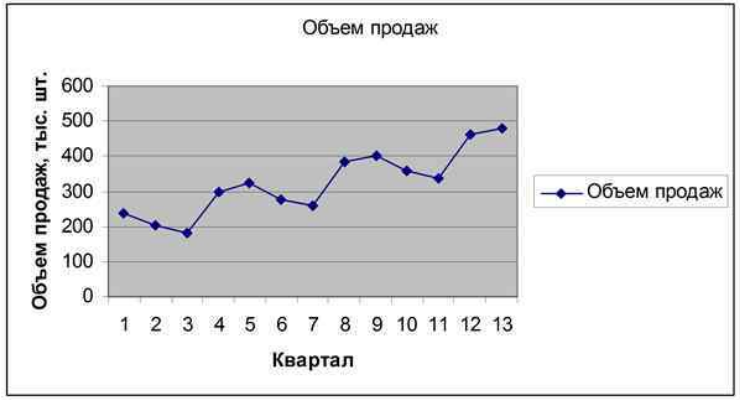
Таблица 1.1. *Типичный формат данных*

Категория	Валюта	Рейтинг продавца	Длитель-ность	Дань закры-тия	Цена закры-тия	Цена откры-тия	Конкурентно-способность?
Music/Movie/ Game	US	3249	5	Mon	0,01	0,01	0
Music/Movie/ Game	US	3249	5	Mon	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	1
Automotive	US	3115	7	Tue	0,01	0,01	1

Непрямоугольные структуры данных

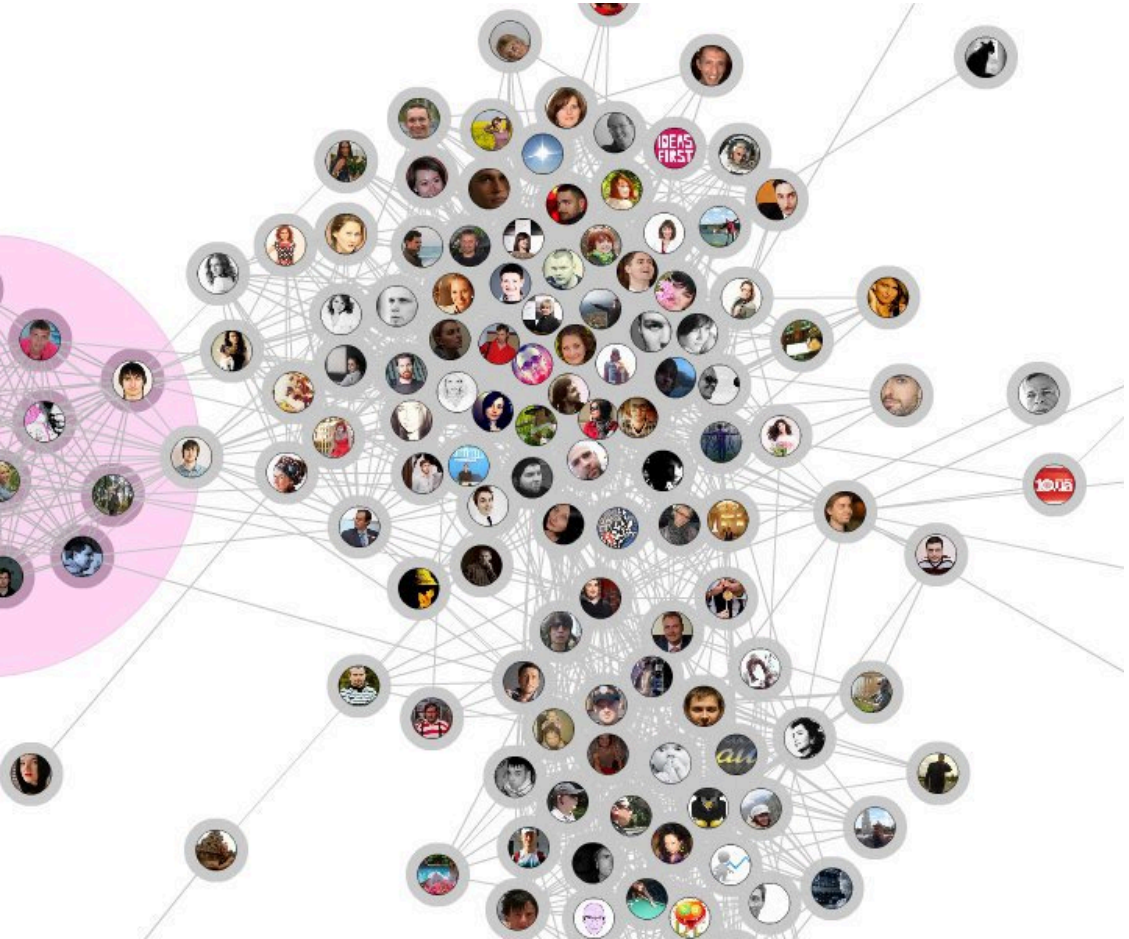
- Временной ряд

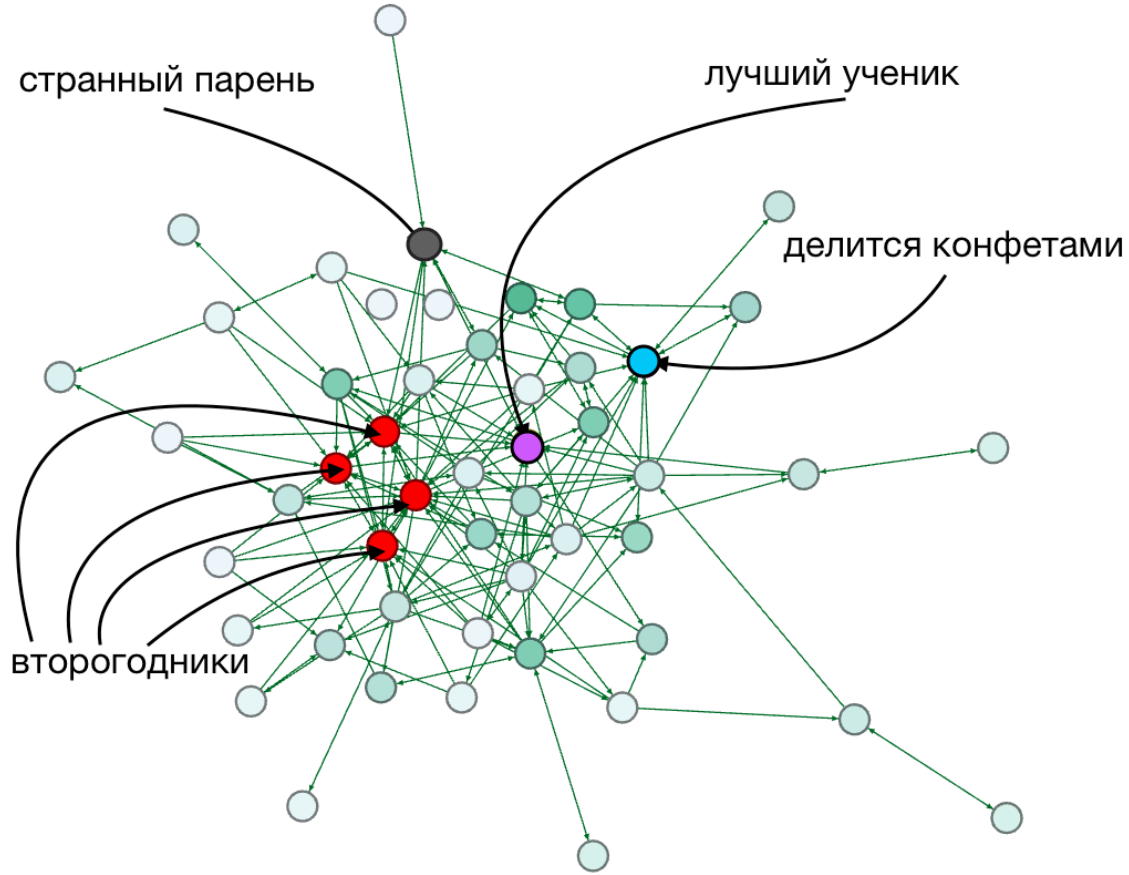
Временно́й ряд содержит последовательные данные измерений одной и той же переменной. Эти данные представляют собой сырой материал для статистических методов предсказания, и они также являются ключевым компонентом данных, производимых устройствами — Интернет вещей.



Пространственные структуры данных, которые используются в картографической и геопространственной аналитике, более сложны и вариативны, чем прямоугольные структуры данных. В их *объектном* представлении центральной частью данных являются объект (например, дом) и его пространственные координаты. В *полевой* проекции, в отличие от него, основное внимание уделяется небольшим единицам пространства и значению соответствующего метрического показателя (яркости пиксела, например).

Графовые (или сетевые) структуры данных используются для представления физических, социальных и абстрактных связей. Например, граф социальной сети, такой как Facebook или LinkedIn, может представлять связи между людьми в сети. Соединенные дорогами центры распределения являются примером физической сети. Графовые структуры широко применяются в определенных типах задач, таких как оптимизация сети и рекомендательные системы.





Среднее (mean)

Сумма всех значений, деленная на количество значений.

Синоним: среднее арифметическое.

Среднее взвешенное (weighted mean)

Сумма произведений всех значений на их веса, деленная на сумму весов.

Синоним: среднее арифметическое взвешенное.

Медиана (median)

Такое значение, при котором половина сортированных данных находится выше и ниже данного значения.

Синоним: 50-й процентиль.

Медиана взвешенная (weighted median)

Такое значение, при котором половина суммы весов находится выше и ниже сортированных данных.

Среднее усеченное (trimmed mean)

Среднее число всех значений после отбрасывания фиксированного числа предельных значений.

Синоним: обрезанное среднее.

Робастный (robust)

Не чувствительный к предельным значениям.

Синоним: устойчивый.

Выброс (outlier)

Значение данных, которое сильно отличается от большинства данных.

Синоним: предельное значение.

Среднее

Самой элементарной оценкой центрального положения является среднее значение, или *среднее арифметическое*. Среднее — это сумма всех значений, деленная на число значений. Рассмотрим следующий ряд чисел: {3, 5, 1, 2}. Среднее составит $(3+5+1+2)/4 = 11/4 = 2,75$. Вы часто будете встречать символ $\bar{x}$ (произносится "x с чертой"), который обозначает среднее значение выборки из популяции, или генеральной совокупности. Формула среднего значения для ряда из n значений $x_1, x_2, ..., x_n$ следующая:

$$\text{Среднее} = \bar{x} = \frac{\sum_i x_i}{n}.$$

Разновидностью среднего является *среднее усеченное*, которое вычисляется путем отбрасывания фиксированного числа сортированных значений с каждого конца последовательности и затем взятия среднего арифметического оставшихся значений. Если представить сортированные значения как $x_{(1)}, x_{(2)}, ..., x_{(n)}$, где $x_{(1)}$ — самое маленькое значение, а $x_{(n)}$ — самое большое, то формула для вычисления усеченного среднего с пропуском p самых малых и самых больших значений будет следующей:

$$\text{Среднее усеченное} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}.$$

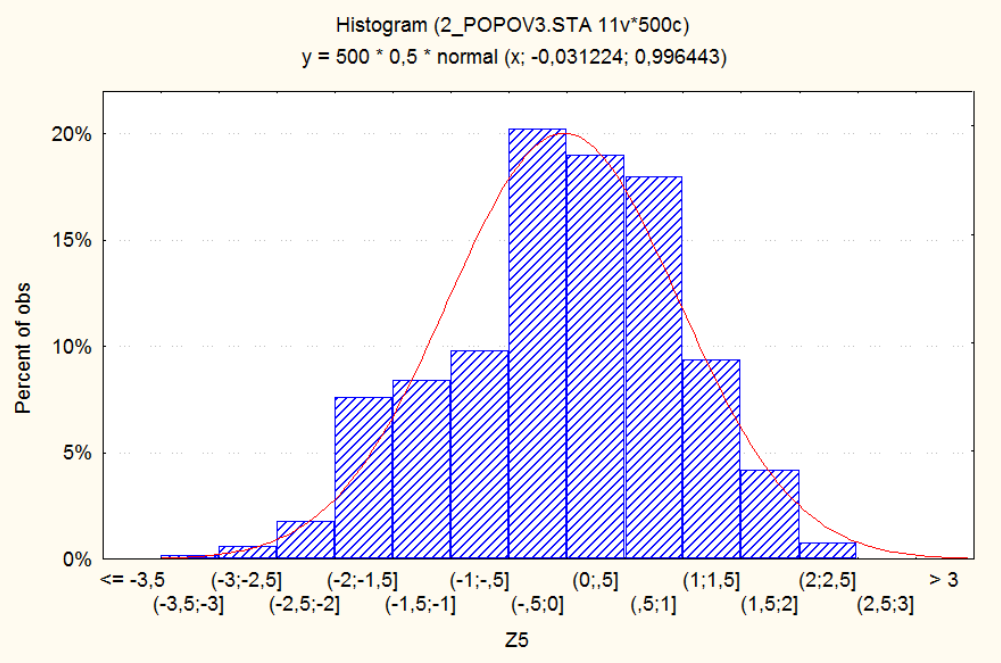
$$\text{Среднее взвешенное} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_i w_i}.$$

Медиана и робастная оценка

Медиана — это число, расположенное в сортированном списке данных ровно посередине. Если число данных четное, срединным значением является то, которое не находится в наборе данных фактически, а является средним арифметическим двух значений, которые делят сортированные данные на верхнюю и нижнюю половины.

Выбросы

- Выброс — это любое значение, которое сильно удалено от других значений в наборе данных.



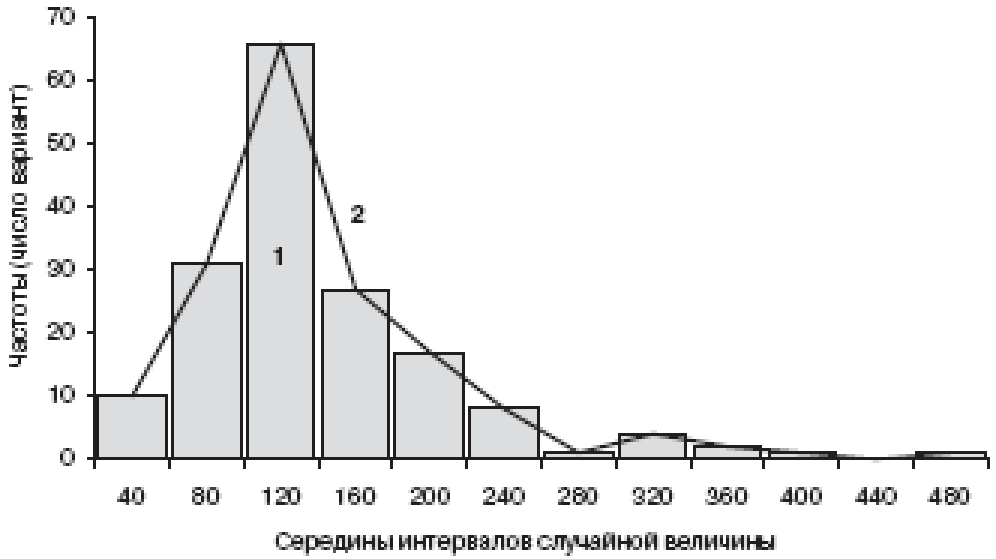


Рис. 1.1. Гистограмма (1) и полигон (2) частот

Пример

численности населения и уровня убийств

В табл. 1.2 показаны первые несколько строк из набора данных, содержащего данные о численности населения и уровне убийств (в единицах убийств на 100 тыс. человек в год) по каждому штату.

Таблица 1.2. *Несколько строк данных* `data.frame` *о численности населения и уровне убийств по штатам*

№	Штат	Население	Уровень убийств
1	Alabama	4 779 736	5,7
2	Alaska	710 231	5,6
3	Arizona	6 392 017	4,7
4	Arkansas	2 915 918	5,6
5	California	37 253 956	4,4
6	Colorado	5 029 196	2,8
7	Connecticut	3 574 097	2,4
8	Delaware	897 934	5,8

Вычислим среднее, среднее усеченное и медиану численности населения, используя R:

```
> state <- read.csv(file="/Users/andrewbrucel/book/state.csv")
> mean(state[["Population"]])
[1] 6162876
> mean(state[["Population"]], trim=0.1)
[1] 4783697
> median(state[["Population"]])
[1] 4436370
```

```
> weighted.mean(state[["Murder.Rate"]], w=state[["Population"]])
[1] 4.445834
> library("matrixStats")
> weightedMedian(state[["Murder.Rate"]], w=state[["Population"]])
[1] 4.4
```

Дисперсность

Центральное положение — это всего одна из размерностей в обобщении признака. Вторая размерность, *вариабельность*, именуемая также *дисперсностью*, показывает, сгруппированы ли значения данных плотно, или же они разбросаны. В основе статистики лежит вариабельность: ее измерение, уменьшение, различение произвольной вариабельности от реальной, идентификация разных источников реальной вариабельности и принятие решений в условиях ее присутствия.

Отклонения (deviations)

Разница между наблюдаемыми значениями и оценкой центрального положения.

Синонимы: ошибки, остатки.

Дисперсия (variance)

Сумма квадратических отклонений от среднего, деленная на $n-1$, где n — число значений данных.

Синонимы: среднеквадратическое отклонение, среднеквадратическая ошибка.

Стандартное отклонение (standard deviation)

Квадратный корень из дисперсии.

Синонимы: норма l_2, евклидова норма.

Среднее абсолютное отклонение (mean absolute deviation)

Среднее абсолютных значений отклонений от среднего[2].

Синонимы: норма l_1, манхэттенская норма.

Медианное абсолютное отклонение от медианы (median absolute deviation from the median)

Медиана абсолютных значений отклонений от медианы.

Размах (range)

Разница между самым большим и самым малым значениями в наборе данных.

Порядковые статистики (order statistics)

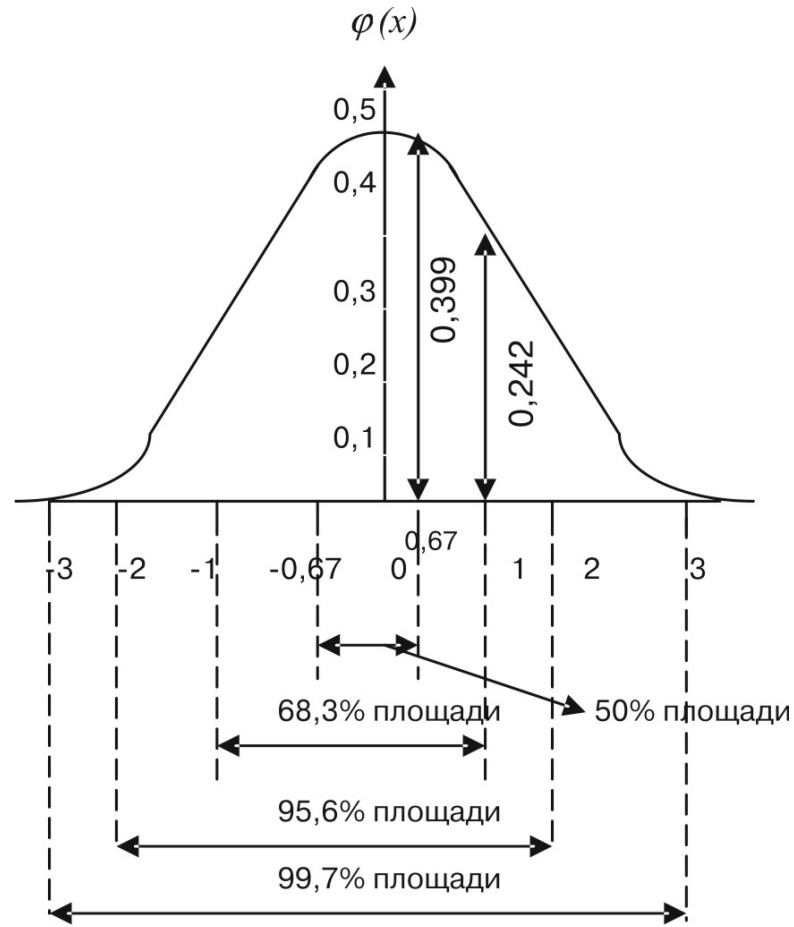
Метрические показатели на основе значений данных, отсортированных от самых малых до самых больших.

Синоним: ранг.

Процентиль (percentile)

Такое значение, что P процентов значений принимает данное значение или меньшее и $(100-P)$ процентов значений принимает данное значение или большее.

Синоним: квантиль.



Межквартильный размах (interquartile range)

Разница между 75-м и 25-м процентилями.

Синонимы: МКР, IQR.

Стандартное отклонение и связанные с ним оценки

Наиболее широко используемые оценки вариабельности основаны на разницах, или *отклонениях*, между оценкой центрального положения и наблюдаемыми данными. Для набора данных {1, 4, 4}, среднее равняется 3, и медиана — 4. Отклонения от среднего представляют собой разницы: $1-3=-2$, $4-3=1$, $4-3=1$. Эти отклонения говорят о том, насколько данные разбросаны вокруг центрального значения.

Один из способов измерить вариабельность состоит в том, чтобы оценить типичное значение этих отклонений. Усреднение самих отклонений мало, поэтому отрицательные отклонения нейтрализуют положительные. Фактически сумма отклонений от среднего как раз равна нулю. Вместо этого простой подход заключается в том, чтобы взять среднее абсолютных значений отклонений от среднего значения. В предыдущем примере абсолютное значение отклонений равно {2, 1, 1}, а их среднее — $(2+1+1)/3=1{,}33$. Это и есть среднее абсолютное отклонение, которое вычисляется по следующей формуле:

$$\text{Среднее абсолютное отклонение} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n},$$

где $\bar{x}$ — среднее значение в выборке, или выборочное среднее.

Самыми известными оценками вариабельности являются *дисперсия* и *стандартное отклонение*, которые основаны на квадратических отклонениях. Дисперсия — это среднее квадратических отклонений, а стандартное отклонение — квадратный корень из дисперсии.

$$\text{Дисперсия} = s^2 = \frac{\sum (x-\bar{x})^2}{n-1};$$

$$\text{Стандартное отклонение} = s = \sqrt{\text{Дисперсия}}.$$

Степени свободы и *n* или *n* – 1?

В книгах по статистике всегда так или иначе обсуждается вопрос, почему в формуле дисперсии у нас в знаменателе $n-1$, вместо n, который приводит к понятию *степеней свободы*. Это различие не является важным, поскольку n обычно настолько велико, что уже не имеет большого значения, будет ли деление выполняться на n или $n-1$.

Если в формуле дисперсии применить интуитивно понятный знаменатель n, то истинное значение дисперсии и стандартного отклонения в популяции будет недооценено. Это называется *смещенной* оценкой. Однако если поделить на $n-1$ вместо n, то стандартное отклонение становится *несмещенной* оценкой.