Введение в компьютерный и интеллектуальный анализ данных

2. Описательная статистика
Общие сведения.

- ***Статистика*** — это наука, включающая разветвленную систему научных дисциплин, изучающих количественную сторону массовых явлений и процессов в неразрывной связи с их качественной стороной.
- ***Предметом*** статистики служат массовые явления и процессы, а также складывающиеся в них количественные закономерности. Например, производство товаров, экспорт, импорт, уровень жизни и т.д.
- ***Статистический метод*** включает в себя следующие составные элементы:
  - научно организованный сбор первичной статистической информации;
  - сводка, обработка и группировка статистической информации;
  - обобщение и интерпретация статистической информации (на этом этапе определяются закономерности развития явления, даются прогнозные оценки).

2. Описательная статистика
Общие сведения.

- ***Описательная статистика*** позволяет с помощью специальных методов осуществить удобное представление эмпирических данных для последующего анализа в виде частотных распределений, графических изображений и различных характеристик (средних, ранговых показателей);

- ***Математическая статистика*** – теория принятия статистических решений, позволяющая с помощью специальных методов обработки данных дать их правильную интерпретацию.

2. Описательная статистика
Генеральная совокупность. Выборка.

- **Генеральной совокупностью** называется совокупность объектов или наблюдений, все элементы которой подлежат изучению при статистическом анализе.
- Часть объектов генеральной совокупности, используемая для исследования, называется выборочной совокупностью или **выборкой**.

2. Описательная статистика
Объекты, признаки, наблюдения, шкалы измерений.

В задачах анализа данных исследуемый объект A характеризуется некоторым набором признаков X_1, X_2, …, X_N. В процессе наблюдения за объектом осуществляются эксперименты, связанные с измерением (регистрацией, фиксацией и т.п.) значений признаков. Результатом измерения признака X_i в эксперименте t является численное значение признака x_{it}, которое называется наблюдением.

- Совокупность наблюдений $\{x_{it}\}$, i-1,2,…, N, t=1,2,…,n над некоторым объектом A называется выборкой наблюдений объема n пространстве N признаков.

2. Описательная статистика
Объекты, признаки, наблюдения, шкалы измерений.

1. Номинальная – состоит из названий, имен или категорий для сортировки или классификации объектов по некоторому признаку.

A=B, A≠B

2. Порядковая – числа присваиваются объектам, чтобы обозначить относительную позицию объектов, но не величину между ними.

A=B, A≠B, A>B, A<B

3. Интервальная – позволяет классифицировать и упорядочивать объекты, а также количественно описать различия между свойствами объектов. Для задания такой шкалы устанавливают единицу измерения и произвольную точку отсчета.

A=B, A≠B, A>B, A<B, A+B, A-B

4. Относительная (шкала отношений) – к этой шкале относятся все интервальные переменные, которые имеют абсолютную нулевую точку. Поэтому переменные относящиеся к интервальной шкале, как правило, имеют и шкалу отношений.

A=B, A≠B, A>B, A<B, A+B, A-B, A*B, A/B

2. Описательная статистика
Объекты, признаки, наблюдения, шкалы измерений.

возраст	сфера деятельности	объем кредита	категория кредитоспособности	пол
39	1	1520	1	м
42	2	1000	2	ж
23	2	850	1	ж
41	3	6475	1	ж
37	1	2356	2	м
21	1	500	3	м

Номинальная: сфера деятельности, пол
Порядковая: категория кредитоспособности
Интервальная: возраст
Относительная: возраст, объём кредита

2. Описательная статистика
Выборка.

- Пусть некоторый признак генеральной совокупность описывается некоторой случайной величиной X. Рассмотрим выборку (x_1, x_2, …, x_n) объема n. Элементы этой выборки представляют собой значения случайной величины X.
- На первом этапе статистической обработки производится ранжирование выборки, т.е. упорядочивание чисел x_1, x_2, …, x_n по возрастанию.
- Различные элементы выборки называются вариантами.
- Частотой варианты x_i называется число m_i, показывающее, сколько раз эта варианта встречается в выборке. Относительной частотой (частостью, долей) называется число $\omega_i = \frac{m_i}{n}$.

2. Описательная статистика
Вариационный ряд.

- Ряд вариант, расположенных в порядке возрастания их значений, с соответствующими этим элементам частотами называется вариационным рядом. Вариационные ряды бывают дискретными и интервальными (соответственно для дискретной и непрерывной случайной величины).

Варианты x_i	x_1	x_2	…	x_n
Частоты m_i	m_1	m_2	…	m_n

2. Описательная статистика
Дискретный вариационный ряд.

- **Пример.** В магазине продана мужская обувь следующих размеров: 36, 38, 37, 41, 37, 41, 38, 42, 39, 39, 42, 42, 42, 39, 42, 39, 40, 40, 40, 39, 39. Построить дискретный вариационный ряд для проданной обуви.
- *Решение.* Признаком этой статистической совокупности является размер обуви. Значениями признака служат числа: 36, 37, 38, 39, 40, 41, 42. Повторяемость этих величин в совокупности соответственно равна 1, 2, 2, 6, 3, 2, 5.

Размер обуви (варианты)	36	37	38	39	40	41	42
Число проданных пар (частоты)	1	2	2	6	3	2	5

2. Описательная статистика
Интервальный вариационный ряд.

- Для построения интервального вариационного ряда разбивают множество значений вариант на интервалы $[a_i, a_{i+1})$ т.е. производят их группировку.
- Рекомендуется количество интервалов выбирать по формуле Стёрджесса $k = 1 + [\log_2 N]$, где N – число единиц совокупности. Длина интервала равна $\Delta = \frac{x_{max} - x_{min}}{k}$.
- Выражение в числителе называют размахом интервала.
- Посчитывая число вариант, попавших в интервал $[a_i, a_{i+1})$, получим значение частот m_i, $i = \overline{1,k}$.
- Если варианта находится на границе интервала, то ее присоединяют к правому интервалу.

Варианты x_i	$[a_1, a_2)$	$[a_2, a_3)$	…	$[a_k, a_{k+1})$
Частоты m_i	m_1	m_2	…	m_n

2. Описательная статистика
Интервальный вариационный ряд.

- **Пример.** Построить интервальный вариационный ряд по первичным данным о размерах прибыли 20 коммерческих банков за год (млрд. денежных единиц).

3.7 4.3 6.7 5.6 5.1 8.1 4.6 5.7 6.4 5.9 5.2 6.2 6.3 7.2 7.9
5.8 4.9 7.6 7.0 6.9

- *Решение.* Упорядочиваем ряд:

3.7 4.3 4.6 4.9 5.1 5.2 5.6 5.7 5.8 5.9 6.2 6.3 6.4 6.7 6.9
7.0 7.2 7.6 7.9 8.1

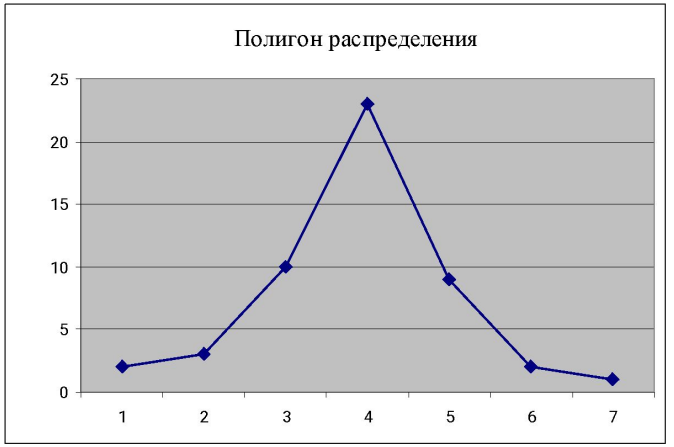
- Вычисляем размах: 8.1 - 3.7 = 4.4.
- по формуле Стёрджесса вычисляем число групп: k=5.
- Вычисляем величину интервала: h=4.4 / 5 = 0.88. Округляем до 0.9.
- Вычисляем границы интервалов и посчитываем количество вариант, попавших в каждый интервал

X	[3.7; 4.6)	[4.6; 5.5)	[5.5; 6.4)	[6.4; 7,3)	[7.3; 8.2)
m_i	2	4	6	5	3

2. Описательная статистика
Графические изображения вариационных рядов. Полигон.

- **Полигон**, как правило служит для изображения дискретного вариационного ряда и представляет собой ломаную, соединяющие точки плоскости с координатами (x_i, m_i), $i = \overline{1,n}$. Для интервального ряда также строится полигон, только его ломаная проходит через точки (c_i, m_i), где $c_i=(a_i+a_{i+1})/2$, $i = \overline{1,k}$.
- **Пример.** Распределение квартир жилого фонда по числу проживающих в них.

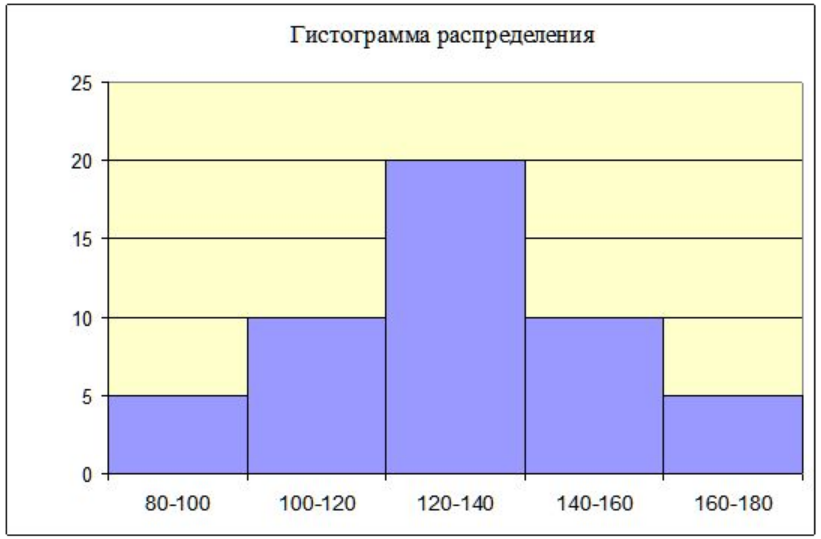
Число живущих в квартире	1	2	3	4	5	6	7	всего
Число квартир	2	3	10	23	9	2	1	50



2. Описательная статистика
Графические изображения вариационных рядов. Гистограмма.

- **Гистограмма** служит только для представления интервальных вариационных рядов и имеет вид ступенчатой фигуры из прямоугольников с основаниями, равными длине интервалов Δ, и высотами, равными частотам m_i интервалов.
- Пример. Распределение продавцов магазина по выработке.

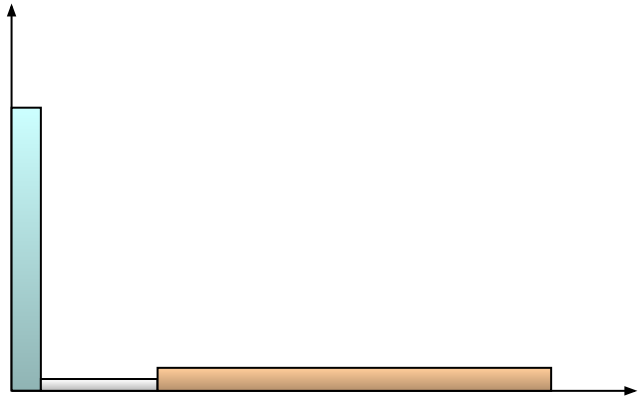
Выработка продавца, тыс. руб.	Число продавцов	% к итогу (частость)
80-100	5	10
100-120	10	20
120-140	20	40
140-160	10	20
160-180	5	10
Итого	50	100



2. Описательная статистика
Графические изображения вариационных рядов. Гистограмма.

- При построении гистограммы распределения вариационного ряда с неравными интервалами: на оси абсцисс откладываются отрезки, которые соответствуют величине интервалов вариационного ряда, а по оси ординат откладывают не частоты, а плотности распределения, т.е. частоты, рассчитанные на единицу ширины интервала, т.е. сколько единиц в группе приходится на единицу величины интервала.
- Пример. Распределение активов коммерческого банка по степени риска определяется следующими данными:

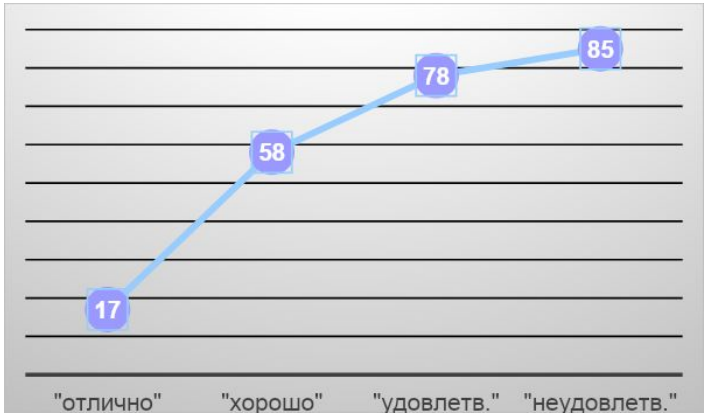
Группы активов по степени риска в %	Структура активов, в %	Высота на графике
0 – 10	61	61:10=6,1
10 – 25	4	4:15=0,27
25 – 100	35	35:75=0,47
всего	100	



2. Описательная статистика
Графические изображения вариационных рядов. Кумулятивная кривая.

- **Кумулята** или **кумулятивная кривая** (частотная функция распределения) в отличие от полигона строится по накопленным частотам или частостям. При этом на оси абсцисс помещают значения признака, а на оси ординат - накопленные частоты или частости.
- Накопленные частоты показывают сколько единиц совокупности имеют значение признака не больше, чем рассматриваемое, и определяются последовательным суммированием частот интервалов.

Категории	Абсолютная частота	Накопленная абсолютная частота	Относительная частота	Накопленная относительная частота
"отлично"	17	17	0,2	0,2
"хорошо"	41	58	0,482352941	0,682352941
"удовлетв."	20	78	0,235294118	0,917647059
"неудовлетв."	7	85	0,082352941	1
	85			



2. Описательная статистика
Графические изображения вариационных рядов. Кривая Лоренца (кривая концентрации доходов).

- Четыре индивида (A, B, C и D) получают суммарный доход в 10 000 руб. в месяц, который распределяется между ними в соответствии с данными в таблице:

	Получаемый доход, руб.	Удельный вес дохода индивида в общем доходе, %	Кумулятивный ряд доходов (накопленные частоты), %	Удельный вес каждого индивида в их общем числе, %	Кумулятивный ряд численности, %
A	1 000	10	10	25	25
B	1 500	15	25	25	50
C	3 000	30	55	25	75
D	4 500	45	100	25	100
Всего	10 000	100	—	100	—

- Для построения кривой Лоренца отложим по оси абсцисс последовательно просуммированные удельные веса индивидов в их общем числе, учитывая, что удельный вес каждого из них составляет одну четверть, или 25 %, а по оси ординат — кумулятивные доли доходов этих людей. Соединив все точки, получим кривую Лоренца

2. Описательная статистика
Кривая Лоренца.

2. Описательная статистика
Кривая Лоренца.

- Неравенство доходов характеризуется степенью отклонения кривой Лоренца от биссектрисы 1-го координатного угла.
- Это отклонение можно измерить через отношение площади заштрихованной фигуры между кривой Лоренца и прямой OD к площади всего треугольника ODK.
- В результате получим показатель, который в литературе называется **коэффициентом концентрации** (коэффициентом Джини).
- Значение коэффициента концентрации для нашего примера будет равно
- Чем ближе значение этого коэффициента к единице, тем выше дифференциация доходов, и, наоборот, чем ближе его значение к нулю, тем более равномерным является распределение доходов