Введение в компьютерный и интеллектуальный анализ данных

Воротницкая Т.И.

2. Описательная статистика
Числовые характеристики одномерных признаков. Средние величины и степенные средние.

- **Средняя величина** — это обобщающая характеристика изучаемого признака в исследуемой совокупности.
- Средняя, являясь обобщающей характеристикой всей совокупности, должна ориентироваться на определённую величину, связанную со всеми единицами этой совокупности. Эту величину можно представить в виде функции $f(x_1, x_2, \ldots, x_n)$.
- Если в приведенной выше функции все величины $x_1, x_2, \ldots, x_n$ заменить их средней величиной $\bar{x}$, то значение этой функции должно остаться прежним $f(x_1, x_2, \ldots, x_n)=f(\bar{x}, \bar{x}, \ldots, \bar{x})$.
- из данного равенства и определяется суммарная средняя

$$\sum_{i=1}^{n} f(x_i) = nf(\bar{x})$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Средние величины и степенные средние.

Основные принципы расчета средней величины:

- Необходим обоснованный выбор статистической совокупности, для которой определяется средняя величина.
- При определении средней величины исходят из качественного содержания статистических величин, учитывая возможную взаимосвязь изучаемых признаков.
- Средняя величина должна рассчитываться по однородной совокупности, которая позволяет применять метод группировки, предполагающий расчет системы обобщающих показателей.
- Общая средняя величина должна подкрепляться и поясняться групповыми средними величинами.

2. Описательная статистика
Числовые характеристики одномерных признаков. Средние величины и степенные средние.

Средние величины

Степенные средние

- **арифметическое**
- **гармоническое**
- **геометрическое**
- **квадратическое**
- **кубическое**

Структурные средние

мода

медиана

2. Описательная статистика
Числовые характеристики одномерных признаков. Средние величины и степенные средние.

- Общая формула для степенной средней $f(x) = x^p$

$$x_1^p + x_2^p + \ldots + x_n^p = \bar{x}_p^p + \bar{x}_p^p + \ldots + \bar{x}_p^p \Leftrightarrow$$

$$\sum_{i=1}^{n} x_i^p = n\left(\bar{x}_p\right)^p \Rightarrow \bar{x}_p = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{1/p}$$

где $\bar{x}_p$ - средняя порядка p для выборки $x_1, x_2, \ldots, x_n$

- Взвешенная средняя величина рассчитывается по сгруппированным статистическим величинам с использованием следующей формулы

$$\bar{x}_p = \left(\frac{\sum_{i=1}^{k} x_{(i)}^p m_i}{n}\right)^{1/p}$$

$x_{(i)}^p$ - значения отдельных статистических величин, по которым проводилась группировка (середина интервалов);

m_i — частота каждой варианты в группе

2. Описательная статистика
Числовые характеристики одномерных признаков. Средние величины и степенные средние.

степень	простая	взвешенная	название
p=-1	$\bar{x}_{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	$\bar{x}_{-1} = \frac{n}{\sum_{i=1}^{k} \frac{m_i}{x_{(i)}}}$	среднее гармоническое
p⟶0	$\bar{x}_0 = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$	$\sqrt[m_i]{x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}} = \sqrt[n]{\prod_{i=1}^{n} x_i^{m_i}}$	среднее геометрическое
p=1	$\bar{x}_1 = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x}_1 = \frac{\sum_{i=1}^{k} x_{(i)} m_i}{n}$	среднее арифметическое
p=2	$\bar{x}_2 = \left(\frac{\sum_{i=1}^{n} x_i^2}{n} \right)^{1/2}$	$\bar{x}_2 = \left(\frac{\sum_{i=1}^{k} x_{(i)}^2 m_i}{n} \right)^{1/2}$	среднее квадратическое

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее арифметическое.

- Под **средним арифметическим** понимается такое значение признака, которое бы имела каждая единица совокупности, если бы общий итог всех значений признака был распределён равномерно между всеми единицами совокупности.

- Пример для несгруппированных данных. Требуется вычислить средний стаж работы 12 работников рекламного агентства. При этом известны индивидуальные значения признака (стажа) в годах: 6, 4, 5, 3, 3, 5, 5, 6, 3, 7, 4, 5.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6+4+5+3+3+5+5+6+3+7+4+5}{12} = \frac{56}{12} = 4{,}7$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее арифметическое.

- Пример для сгруппированных данных.

x_i	3	4	5	6	7
m_i	3	2	4	2	1

- В данном случае применяют взвешенное среднее

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i m_i}{n} = \frac{3*3+4*2+5*4+6*2+7*1}{12} = 4{,}7$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее арифметическое.

- Пример для интервального вариационного ряда. В таблице приведены данные об урожайности ржи на различных участках поля:

урожайность ржи, ц/га	[9 - 12)	[12 - 15)	[15 - 18)	[18 - 21)	[21 - 24)	[24 - 27)
доля участка в общей площади, %	6	12	33	22	19	8

$$\bar{x} = \frac{\sum_{i=1}^{k} c_i m_i}{n} =$$

$$= \frac{10.5*6+13.5*12+14.5*33+19.5*22+22.5*19+25.5*8}{100} = 18,3$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Свойства среднего арифметического.

- Если $x_i=c$, i=1,2,…,n, c- const, то $\bar{x}=c$.
- Сумма отклонений индивидуальных значений признака от среднего арифметического равна нулю: $\sum_{i=1}^{n}(x_i-\bar{x})=0$
- Сумма квадратов отклонений индивидуальных значений признака от средней арифметической меньше, чем сумма квадратов их отклонений от любой другой произвольной величины C

$$\sum_{i=1}^{n}(x_i-\bar{x})^2<\sum_{i=1}^{n}(x_i-C)^2$$

- Если все усредняемые варианты уменьшить или увеличить на постоянное число A, то и со средней арифметической произойдут аналогичные изменения

$$\frac{\sum_{i=1}^{n}(x_i\pm A)}{n}=\bar{x}\pm A$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Свойства среднего арифметического.

- Если все варианты значений признака уменьшить или увеличить в А раз, то среднее также соответственно увеличится или уменьшится в А раз:

$$\frac{\sum_{i=1}^{n}\frac{x_i}{A}}{n}=\frac{\frac{1}{A}\sum_{i=1}^{n}x_i}{n}=\frac{1}{A}\bar{x}$$

- Если веса уменьшить или увеличить в А раз, то среднее арифметическое от этого не изменится

$$\frac{\sum_{i=1}^{k}x_{(i)}\frac{m_i}{A}}{\sum_{i=1}^{k}\frac{m_i}{A}}=\frac{\frac{1}{A}\sum_{i=1}^{k}x_{(i)}m_i}{\frac{1}{A}\sum_{i=1}^{k}m_i}=\bar{x}$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее гармоническое.

- В статистике среднее гармоническое применяется в случае, когда наблюдения, для которых требуется получить среднее арифметическое, заданы обратными значениями.

$$\bar{x}_p = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{1/p}, p = -1$$

- Среднее гармоническое взвешенное используется, когда известен числитель исходного соотношения среднего, но не известен его знаменатель
- Среднее гармоническое невзвешенное может использоваться вместо взвешенного в тех случаях, когда значения m_i для единиц совокупности равны.

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее гармоническое взвешенное.

Пример: Рассмотрим расчет средней урожайности. В таблице приведен валовой сбор и урожайность сельскохозяйственной культуры «X» по районам области:

Район	Валовый сбор, тыс.т	Урожайность, ц/га
A	52	10
B	40	14
C	31	15
D	67	8

- Средняя урожайность$=\frac{\text{общий валовый сбор (тыс.ц)}}{\text{общая посевная площадь (тыс.га)}}$
- $\bar{x}=\frac{520+400+310+670}{\frac{520}{10}+\frac{400}{14}+\frac{310}{15}+\frac{670}{8}}$=10,3(ц/га)

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее гармоническое невзвешенное.

- $\bar{x} = \frac{n}{\sum \frac{1}{x_i}}$

Пример: в фирме, специализирующейся на торговле по почте на основе предварительных заказов, упаковкой и отправкой товаров занимаются два работника. Первый на обработку одного заказа затрачивает 5 минут, второй – 15. Каковы средние затраты времени на 1 заказ, если общая продолжительность дня у обоих работников равна?

$$\bar{x} = \frac{60 + 60}{\frac{60}{5} + \frac{60}{15}} = \frac{2}{0{,}2 + 0{,}067} = 7{,}5(\text{мин})$$

$$\frac{60}{7{,}5} + \frac{60}{7{,}5} = 16(\text{заказов})$$

(в примере используется невзвешенное среднее, т.к. рабочий день у сотрудников одинаковый)

2. Описательная статистика
Числовые характеристики одномерных признаков. Среднее геометрическое.

- Среднее геометрическое используется в анализе динамики для определения среднего темпа роста
- Пример: известны данные о темпах роста производства продукции:

Год	1998	1999	2000	2001
Темп роста	1,24	1,39	1,31	1,15

среднегодовой темп роста равен

$$\overline{T} = \sqrt[4]{1{,}24 * 1{,}39 * 1{,}31 * 1{,}15} = 1{,}27$$

2. Описательная статистика
Робастные показатели вариационных рядов.

- Средние величины, которые сохраняют точность представления выборки при наличии аномальных наблюдений, называются **робастными**.

1. **Усечённая средняя арифметическая порядка** $\alpha, 0 \le \alpha < \frac{1}{2}$

$$\bar{x}_\alpha = \frac{1}{n-2m}\left(x_{(m+1)} + \ldots + x_{(n-m)}\right)$$

$\alpha = \frac{m}{n}$ - доля «ненадежных» вариант в выборке, m – количество «выбросов»

2. **среднее по Винздору порядка** $\alpha, 0 \le \alpha < \frac{1}{2}$

$$\bar{x}_\alpha = \frac{1}{n}\left(mx_{(m+1)} + x_{(m+1)} + \ldots + x_{(n-m)} + mx_{(n-m)}\right)$$

- Пример: X=(0,12; 0,96; 0,97; 1,00; 1,01; 1,02; 1,04; **10,52**)

$$\bar{x} = \frac{0{,}12 + \ldots + 1{,}04 + 10{,}52}{8} = 2{,}08$$

- 1. m=1, α=1/8, $\bar{x}_{0,125} = \frac{1}{8-2}(0{,}96 + \ldots + 1{,}04) = 1{,}00$

- 2. $\bar{x}_{1/8} = \frac{1}{8}\left(2(0{,}96 + 1{,}04) + 0{,}97 + \ldots + 1{,}02\right) = 1{,}00$

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

Мода - это варианта, которая имеет наибольшую частоту. Она соответствует определенному значению признака.

- Если все варианты наблюдаются с одинаковой частотой, то говорят, что вариационный ряд не имеет моды.
- Если две или более соседние варианты имеют наибольшие частоты, равные между собой, то мода равна средней арифметической этих вариант.
- Если равные варианты, имеющие наибольшие частоты, расположены не по соседству, то принято говорить, что признак имеет две и более моды (бимодальный, полимодальный признаки и т.д.)

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

- Пример: Распределение проданной обуви по размерам характеризуется следующими показателями:

размер обуви	36	37	38	39	40	**41**	42	43	44	45	и выше
число пар, в % к итогу	—	1	6	8	22	30	20	11	1	1	—

- Для интервальных рядов распределения с равными интервалами мода определяется по формуле:

$$M_o = x_{Mo} + h_{Mo} \frac{m_{Mo} - m_{Mo-1}}{(m_{Mo} - m_{Mo-1}) + (m_{Mo} - m_{Mo+1})}$$

x_{Mo} - нижняя граница модального интервала (интервала с наибольшей частотой) начальное значение интервала, содержащего моду;

h_{Mo} - величина модального интервала

m_{Mo} - частота модального интервала

m_{Mo-1}- частота предмодального интервала

m_{Mo+1}- частота послемодального интервала

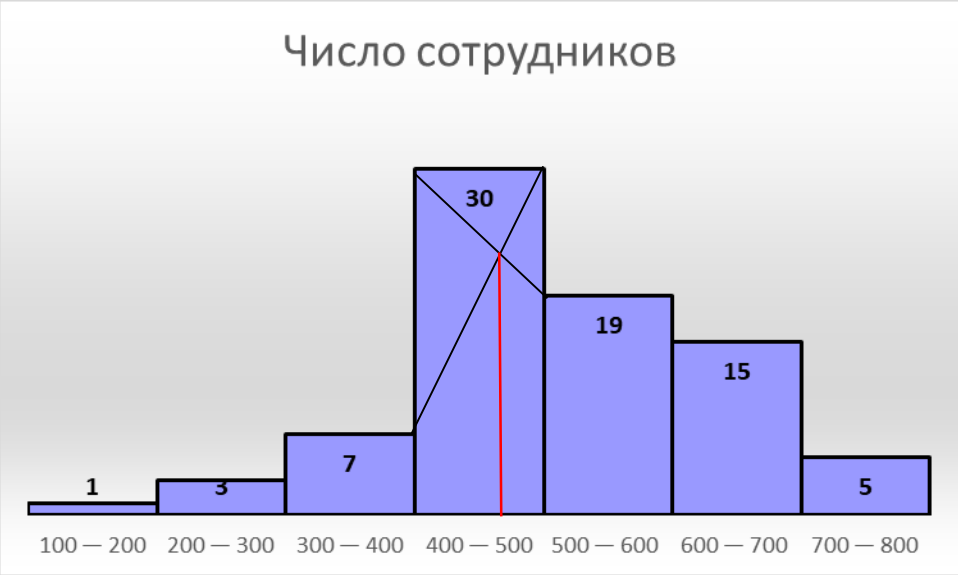
2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

- Пример: распределение заработной платы персонала некоторого предприятия характеризуется следующими данными

Заработная плата ден. ед.	Число сотрудников
100 — 200	1
200 — 300	3
300 — 400	7
400 — 500	30
500 — 600	19
600 — 700	15
700 — 800	5
ИТОГО	80

$$M_o = x_{Mo} + h_{Mo} \frac{m_{Mo} - m_{Mo-1}}{(m_{Mo} - m_{Mo-1}) + (m_{Mo} - m_{Mo+1})}$$

$$M_o = 400 + 100 \frac{30 - 7}{(30 - 7) + (30 - 19)} = 467{,}7$$



2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

Медиана — это варианта, которая расположена по середине ранжированного вариационного ряда.

- Для ранжированного дискретного (не сгруппированного) вариационного ряда с нечётным числом вариант, медиана расположена в центре ряда.
- Для ранжированного дискретного вариационного ряда с чётным числом членов ряда n=2k, медианой будет среднее арифметическое из двух смежных вариант
- Для сгруппированного дискретного ряда $$Me = \min_{1 \le i \le k} \left\{ x_i : m_i^C \ge \frac{n}{2} \right\}$$

Месячная з/п, руб.	Число рабочих	Сумма накопленных частот
110	2	2
130	6	8 (2+6)
160	12	20 (8+12)
190	16	—
220	4	—
	40	

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

- Для интервального вариационного ряда подсчитываем накопленные частоты, определяем медиальный интервал по полусумме частот интервального ряда.

$$Me = x_{Me} + h\frac{\frac{n}{2} - m^{C}_{Me-1}}{m_{Me}}$$

x_{Me} - нижняя граница интервала, содержащего медиану;

h - величина медианного интервала;

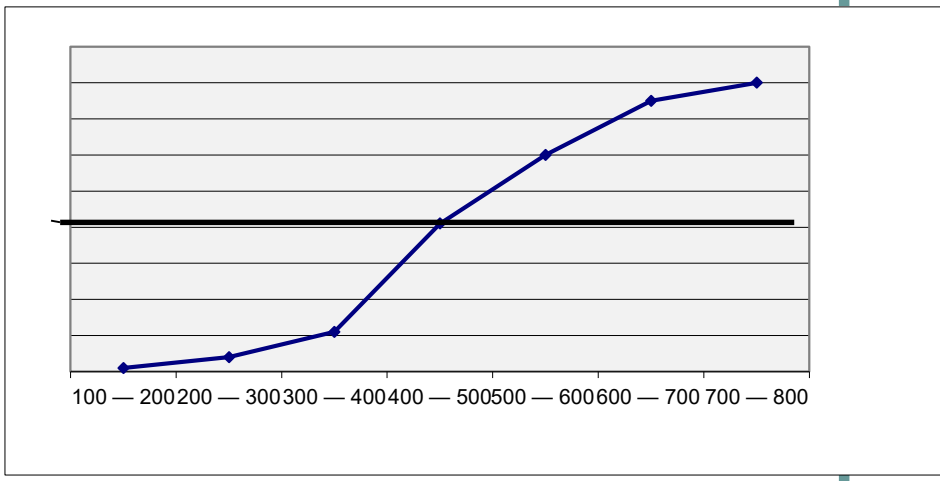
n - количество вариант в ряду;

m^{C}_{Me-1} - сумма накопленных частот, предшествующих медианному интервалу;

m_{Me} - частота медианного интервала.

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

Заработная плата	Число сотрудников	Сумма накопленных частот
100 — 200	1	1
200 — 300	3	4
300 — 400	7	11
400 — 500	30	41
500 — 600	19	60
600 — 700	15	75
700 — 800	5	80
ИТОГО	80	1



$x_{Me} = 400;\ h = 100;\ n = 80;\ m^{C}_{Me-1} = 11;\ m_{Me} = 30$

$$Me = 400 + 100\frac{0,5*80-11}{30} = 400 + 96,66 = 496,66$$

Для графического определения медианы последнюю ординату кумуляты делят пополам. Через полученную точку проводят прямую, параллельную оси *x* до пересечения ее с кумулятой. Абсцисса точки пересечения является медианой представленного на графике распределения.

Числовые характеристики одномерных признаков. Структурные средние величины.

Свойство медианы — сумма абсолютных отклонений членов ряда от медианы есть величина наименьшая.

$$\sum_{i=1}^{n} |x_i - Me| = \min$$

Медиана используется при проектировании оптимального расположения остановок общественного транспорта, стендов и пр.

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

Пример. На прямолинейном шоссе длиной 100 км расположены 10 складов с некоторым товаром. Менеджеру необходимо выбрать такое место для строительства супермаркета, чтобы общий пробег транспорта по доставке товаров к супермаркету был минимален.

местоположение склада, км	7	24	28	37	40	46	60	78	86	92
m (число доставок товара)	10	15	5	20	5	25	15	30	10	65
накопленные частоты	10	25	30	50	55	80	95	125	135	200

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

местоположение склада, км	7	24	28	37	40	46	60	78	86	92
m (число доставок товара)	10	15	5	20	5	25	15	30	10	65
накопленные частоты	10	25	30	50	55	80	95	125	135	200

$$\bar{x} = 63.7$$

$$\sum_i \left| x_{(i)} - \bar{x} \right| m_i = 4983$$

$$Me = 78$$

$$\sum_i \left| x_{(i)} - Me \right| m_i = 4840$$

$$Mo = 92$$

$$\sum_i \left| x_{(i)} - Mo \right| m_i = 5660$$

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины.

Квантили — порядковые характеристики, занимающие определённое место в ранжированном вариационном ряду:

Квантилем порядка q вариационного ряда называется значение признака X, которое делит всё наблюдаемое множество значения признака в пропорции $\frac{q}{1-q}, 0 < q < 1$

Частные случаи:

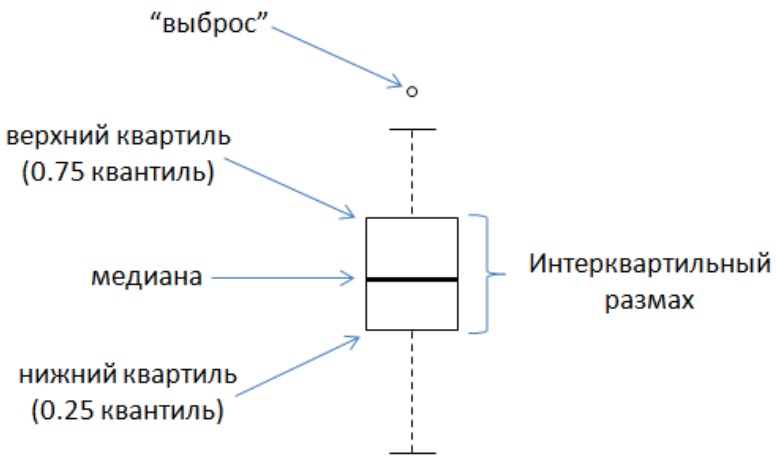
- Медиана
- **Квартили, которые делят вариационный ряд на 4 части**
  - нижняя квартиль $x_{1/4}$
  - средняя квартиль (медиана)
  - верхняя квартиль $x_{3/4}$
- Децили (1/10 всей совокупности)
- Перцентили (квантиль уровня 1/100)

Интерквартильным размахом называется разность между третьим и первым квартилями. Интерквартильный размах является характеристикой разброса распределения.

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины. Ящик с усами.

Диаграммы размахов (коробчатые диаграммы) или "ящики с усами" (box-whisker plots) -показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.

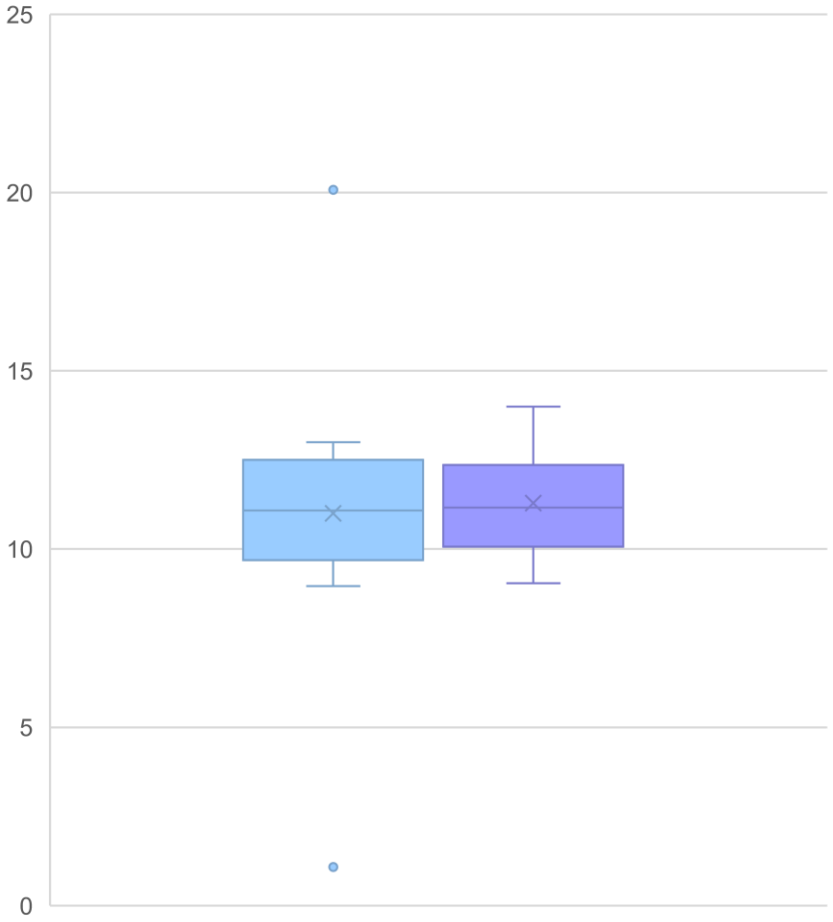
Диаграммы размаха можно использовать для визуальной экспресс-оценки разницы между двумя и более группами (например, между датами отбора проб, экспериментальными группами, участками пространства, и т.п.).



В R для построения диаграмм размахов служит функция boxplot().

2. Описательная статистика
Числовые характеристики одномерных признаков. Структурные средние величины. Ящик с усами.

ряд 1	ряд 2
12,5	12,58
9,32	9,4
12,26	12,34
20,08	12,36
10,74	10,82
8,96	9,04
1,08	11,12
13	14
12	12,08
12,56	12,64
9,69	9,77
10,23	10,31
11,56	11,64
11,08	11,16
9,99	10,07



длина «усов»:
Q1-1,5(Q3-Q1)
Q3+1,5(Q3-Q1)
Q1-нижний квартиль
Q3-верхний квартиль

2. Описательная статистика
Показатели вариации.

- **вариация признака** - различие индивидуальных значений признака внутри изучаемой совокупности

- Под **вариацией** в статистике понимают такие количественные изменения величины исследуемого признака в пределах однородной совокупности, которые обусловлены перекрещивающимся влиянием действия различных факторов.

- Показатели вариации: абсолютные и относительные.
- Абсолютные: размах вариации, среднее линейное отклонение, дисперсия, среднее квадратическое отклонение.
- Относительные: коэффициент вариации, относительное линейное отклонение и др.

2. Описательная статистика
Показатели вариации. Абсолютные.

- ***Размах вариации*** - это разность между наибольшим и наименьшим значениями вариант. (устанавливает только крайние отклонения и не отражает отклонений всех вариант в ряду)
- ***Среднее линейное отклонение*** определяется как средняя арифметическая из отклонений индивидуальных значений от средней, без учета знака этих отклонений (учитывает различие всех единиц изучаемой совокупности)

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad \bar{d} = \frac{\sum_{i=1}^{k} |x_{(i)} - \bar{x}| m_i}{\sum_{i=1}^{k} m_i}$$

- ***Дисперсия*** - это средняя арифметическая квадратов отклонений каждого значения признака от общей средней.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad \sigma^2 = \frac{\sum_{i=1}^{k} m_i (x_{(i)} - \bar{x})^2}{\sum_{i=1}^{k} m_i}$$

- Чем меньше среднее квадратическое отклонение, тем лучше средняя арифметическая отражает собой всю представляемую совокупность.

2. Описательная статистика
Показатели вариации. Относительные.

- Коэффициент вариации
- $V_\sigma = \frac{\sigma}{\bar{x}} 100\%$

- Показатели вариации дают характеристику однородности совокупности. Совокупность вариации считается однородной, если коэффициент вариации не превышает 33%.

- Важная функция обобщающих показателей вариации – это оценка надёжности средних. Чем меньше $\bar{d}, \boldsymbol{\sigma^2}, \boldsymbol{V_\sigma}$, тем однородней полученная совокупность явлений и надежнее полученная средняя.

2. Описательная статистика
Показатели вариации. Относительные.

- Пример: покажем расчет среднего квадратического отклонения по данным дискретного ряда распределения студентов одного факультета по возрасту

группы студентов по возрасту (x)	**число студентов (f)**	$x_i f_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 f$
17	20	340	-3,9	15,21	304,2
18	80	1440	-2,9	8,41	672,8
19	90	1710	-1,9	3,61	324,9
20	110	2200	-0,9	0,81	89,1
21	130	2730	0,1	0,01	1,3
22	170	3740	1,1	1,21	205,7
23	90	2070	2,1	4,41	396,9
24	60	1440	3,1	9,61	576,6
итого	***750***	***15670***			***2571,5***

2. Описательная статистика
Показатели вариации. Относительные.

группы студентов по возрасту (x)	число студентов (f)	$x_i f_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 f$
17	20	340	-3,9	15,21	304,2
18	80	1440	-2,9	8,41	672,8
19	90	1710	-1,9	3,61	324,9
20	110	2200	-0,9	0,81	89,1
21	130	2730	0,1	0,01	1,3
22	170	3740	1,1	1,21	205,7
23	90	2070	2,1	4,41	396,9
24	60	1440	3,1	9,61	576,6
итого	***750***	***15670***			***2571,5***

$$\overline{x} = \frac{15670}{750} = 20,9$$

$$\sigma^2 = \frac{2571,5}{750} = 3,43$$

$$\sigma = \sqrt{3,43} = 1,85$$

$$V_\sigma = \frac{1,85}{20,9}100\% = 8,9\%$$

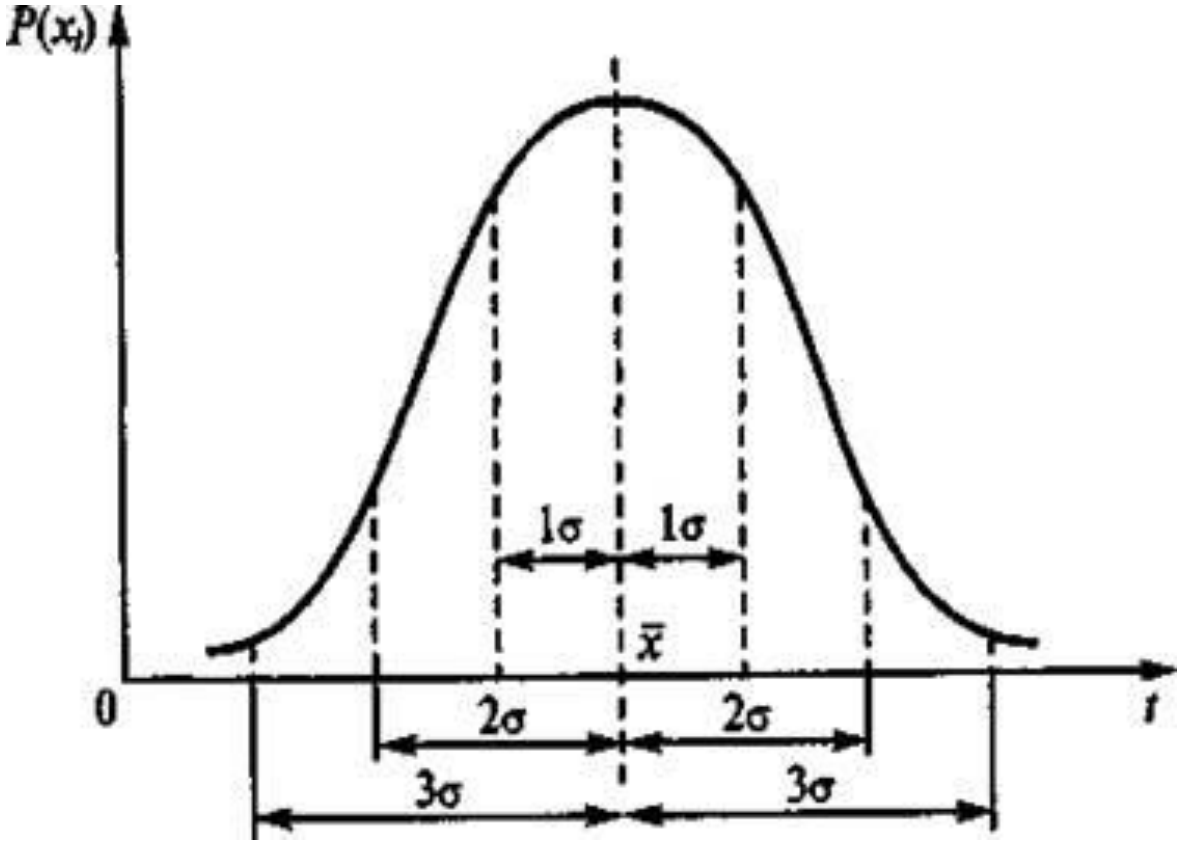
Совокупность студентов по возрасту однородна по своему составу

2. Описательная статистика
Показатели вариации. Правило 3-х сигм.

- Вероятность того, что случайная величина отклонится от своего математического ожидания на большую величину, чем утроенное среднее квадратическое отклонение, практически равна нулю.
- Правило справедливо только для случайных величин, распределенных по нормальному закону.

Пример. Пусть имеется выборка наблюдений за ежедневными продажами в магазине. Значения их распределены по нормальному закону с математическим ожиданием 150 000 руб. и среднеквадратическим отклонением 20 000 руб. Тогда в соответствии с правилом 3-х сигм продажи ниже, чем 150 000 - 20 000 х 3 = 90 000, и выше, чем 150 000 + 20 000 х 3 = 210 000, являются практически невозможными событиями. Фактически это означает, что рассматривать данные объемы продаж как потенциально возможные не имеет смысла.

2. Описательная статистика
Показатели вариации. Правило 3-х сигм.



2. Описательная статистика
Зависимость статистических признаков. Регрессия.

- Признаки по их значению для изучения взаимосвязи делятся на два класса: факторные - обуславливающие изменения других, связанных с ними, признаков, и результативные – признаки, изменяющиеся под действием факторных признаков.
- Связь между признаками может быть функциональной и стохастической.
- **Функциональная связь** – связь при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.
- Если причинная зависимость появляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений, то такая зависимость называется **стохастической**.
- **Корреляционная связь**, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков.

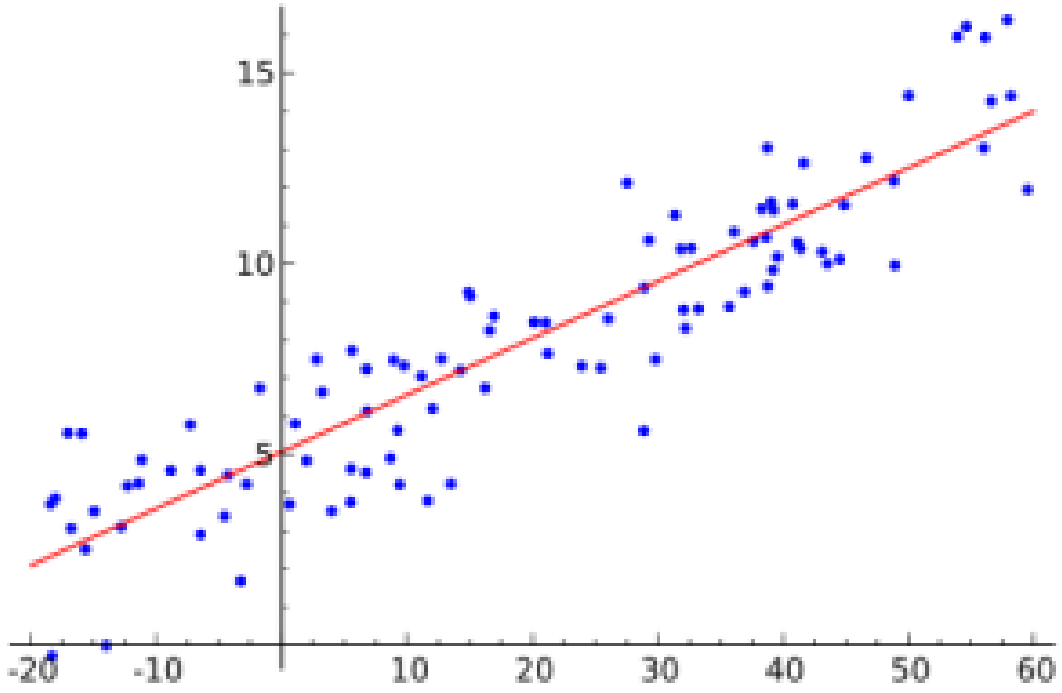
2. Описательная статистика
Зависимость статистических признаков. Регрессия.

- Регрессионный анализ заключается в определении аналитического выражения связи, в котором изменение одной величины (зависимой(результативный признак)), обусловлено влиянием одной или нескольких независимых величин (факторных признаков).

- **Y = a+b*X**, где X – это предиктор или исходные данные, а Y – это предсказываемая величина

- По оси абсцисс мы отмечаем значения предиктора, а по оси ординат значения предсказываемой величины. Тогда простая линейная регрессия это прямая, проведенная таким образом, чтобы минимизировать расхождение между истинными значениями предсказываемой величины и точками на линии, соответствующими значениям предикторов.

2. Описательная статистика
Зависимость статистических признаков. Регрессия.



2. Описательная статистика
Зависимость статистических признаков. Регрессия.

метод наименьших квадратов (МНК)

- неизвестные параметры модели выбираются таким образом, чтобы сумма квадратов отклонений эмпирических значений от модельных была минимальной:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2 \to min.$$

$$\begin{cases} \frac{\partial RSS}{\partial a} = \sum_{i=1}^{n} (a + bx_i - y_i) = 0 \\ \frac{\partial RSS}{\partial b} = \sum_{i=1}^{n} (a + bx_i - y_i)\, x_i = 0, \end{cases} \qquad \begin{cases} \frac{\partial RSS}{\partial a} = an + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\ \frac{\partial RSS}{\partial b} = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i x_i = 0, \end{cases}$$

$$\begin{cases} a + b\bar{x} - \bar{y} = 0 \\ a\bar{x} + b\overline{x^2} - \overline{xy} = 0, \text{ где} \end{cases}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x^2{}_i, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

2. Описательная статистика
Зависимость статистических признаков. Регрессия.

Зависимость между размером чистого дохода и объёмом акционеров коммерческих банков какого-то региона какой-то страны

Банк	Чистый доход, млрд., y_i	Объём вложений акционеров, млрд., x_i
1	0,1	8,8
2	1,3	9,4
3	0,1	10
4	2,6	10,6
5	0,1	11
6	0,3	11,9
7	4,6	12,7
итого	17,1	120,1

2. Описательная статистика
Зависимость статистических признаков. Регрессия.

Банк	Чистый доход, млрд., y_i	Объём вложений акционеров, млрд., x_i	x^2	xy
1	0,1	8,8	77,44	0,88
2	1,3	9,4	88,36	12,22
3	0,1	10	100	1
4	2,6	10,6	112,36	27,56
5	0,1	11	121	1,1
6	0,3	11,9	141,61	3,57
7	4,6	12,7	161,29	58,42
итого	17,1	120,1	802,06	104,75

$$\begin{cases} 7a + 74{,}4b = 9{,}1 \\ 74{,}4a + 802{,}06b = 104{,}75 \end{cases}$$, a=-9,178, b=0,982

$$\begin{cases} x = 13 \\ y = 3{,}58 \end{cases}$$

2. Описательная статистика
Регрессия в R.

Пример. Цены акций нефтедобывающих компаний зависят от цен на нефть. В нулевом приближении, считаем, что

Цена Акции = K * Цена барреля нефти + Некая константа

Если определить константу и коэффициент, то можно по цене на нефть предсказывать цену акции.

В R проведем линейную регрессию цен акций компании по цене на нефть. Исходные данные в файле *neft.txt*.

```
>data <- read.csv('C:\\ neft.txt', sep='\t') # читает данные из файла
>fit <- lm(data$ROSN ~ data$BRN ) # строит линейную регрессию,
```

~ указывает зависимость

```
>summary(fit) # выводит статистический отчет
```

2. Описательная статистика
Регрессия в R.

```
Call:
lm(formula = data$ROSN ~ data$BRN)

Residuals:
    Min      1Q  Median      3Q     Max
-39.534 -13.287  -7.255  17.163  39.107

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.4052    19.1136   3.213   0.0036 **
data$BRN      1.9440     0.2446   7.947 2.65e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.3 on 25 degrees of freedom
Multiple R-squared: 0.7164,     Adjusted R-squared: 0.705
F-statistic: 63.15 on 1 and 25 DF,  p-value: 2.654e-08

> |
```

(1)

(2)

2. Описательная статистика
Зависимость статистических признаков. Линейный коэффициент корреляции.

Линейный коэффициент корреляции характеризует тесноту и направление связи между двумя коррелируемыми признаками в случае наличия между ними линейной зависимости.

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (y)^2]}}$$

Линейный коэффициент корреляции изменяется в пределах от -1 до1.

Оценка линейного коэффициента корреляции

значение	Характеристика связи	Интерпретация связи
r = 0	Отсутствует	-
0 < r < 1	Прямая	с увеличением x увеличивается y
-1 < r < 0	Обратная	с увеличением x уменьшается y и наоборот
r = 1	функциональная	Каждому значению факторного признака соответствует одно значения результативного признака

2. Описательная статистика
Зависимость статистических признаков. Линейный коэффициент корреляции.

Пример: на основе выборочных данных о деятельности пяти предприятий одной из отраслей промышленности какой-то республики оценить тесноту связи между трудоемкостью продукции предприятия (X, чел.-час.) и объемом ее производства (Y, млн руб.)

№	объём произведенной продукции, млн руб.	затраты на 1000 изделий, чел.-час.	yx	y^2	x^2
1	33,2	0,15	4,98	1102,24	0,0225
2	121	0,12	14,52	14641	0,0144
3	99,5	0,11	10,945	9900,25	0,0121
4	59,8	0,09	5,382	3576,04	0,0081
5	80,3	0,08	6,424	6448,09	0,0064
сумма	***393,8***	***0,55***	***42,251***	***35667,62***	***0,0635***
среднее	***78,76***	***0,11***	***8,45***	***7133,524***	***0,0127***

$$\sigma_x = \overline{x^2} - (\bar{x})^2 = 0{,}0127 - (0{,}11)^2 = 930{,}3864,$$
$$\sigma_y = \overline{y^2} - (\bar{y})^2 = 7133{,}524 - (78{,}76)^2 = 0{,}0006$$

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} = \frac{8{,}45 - 78{,}76 * 0{,}11}{\sqrt{930{,}3864 * 0{,}0006}} = -0{,}3$$

R – статистические функции

- mean(x),
- median(x),
- quantile(x, probs = seq(0, 1, 0.25)),
- var(x),
- sd(x),
- boxplot().