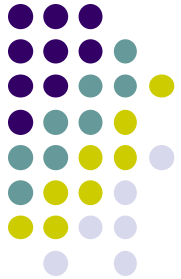
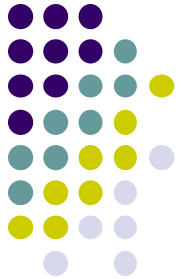
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА



Математическая статистика -

это раздел математики, посвящённый математическим методам сбора, систематизации, обработки, анализа и использования экспериментальных данных.

Вариационные ряды и их характеристики



Пусть требуется изучить некоторую совокупность объектов относительно некоторого количественного или качественного признака.

Иногда проводят сплошное обследование.

На практике чаще всего делают **выборку**, т.е. отбирают часть объектов совокупности.

- Количество объектов в выборке называется её **объёмом** (n).

Наблюдаемые значения элементов выборки называются **вариантами** $x_1, x_2, \dots, x_n$ (варианты могут повторяться).

Количество раз сколько варианта x_i встретилась в выборке, называется **частотой** это варианты (n_i).

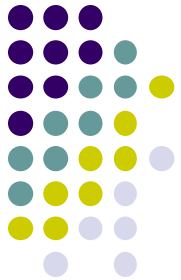
$$\sum n_i = n.$$

Величина $w_i = \frac{n_i}{n}$ называется **частостью** варианты x_i.

Накопленная (кумулятивная частота)

$$n_{i+1}^{нак} = n_i^{нак} + n_{i+1}, \text{ где } n_1^{нак} = n_1$$

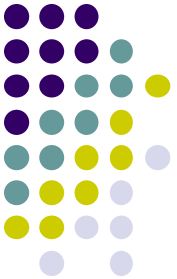
Дискретным вариационным рядом ***называется ранжированный в порядке возрастания ряд вариантов с соответствующими им частотами или частостями.***



Варианта	x_1	x_2	...	x_m
Частота варианты	n_1	n_2	...	n_m
Относительная частота варианты (частость)	$w_1 = \frac{n_1}{n}$	$w_2 = \frac{n_2}{n}$	...	$w_m = \frac{n_m}{n}$

Пример. 20 студентов на экзамене по психологии получили такие оценки (по пятибалльной системе): 5, 4, 4, 3, 3, 5, 2, 3, 4, 3, 3, 4, 4, 4, 3, 5, 4, 4, 3 ,5. Составить дискретный вариационный ряд.

Дискретный вариационный ряд:



Варианта	2	3	4	5
Частота варианты	1	7	8	4

Пример. Имеются данные о заработной плате сотрудников предприятия за сентябрь 2012 г., тыс. руб.:

10,11, 21,13,14,15,14,15,15,19,15,15,16,16,
16,19,16,16,19,19,19,16,19,17,17,17,18,18,18,
18,18,18,18,18,19,15,19,20,21,20,21,21,20,12,
23,23, 12,22, 12,21,21,21, 17,17,17

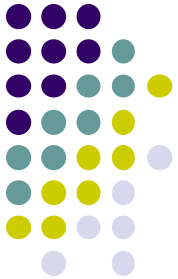
x_i	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Итого
Частота n_i	1	1	3	1											55
Накопленные частоты $n_i^{\text{нак}}$	1	2	5	6											55

Если число различных значений признака в выборке велико, или признак является непрерывным (т.е. может принять любое значение в некотором интервале), то **составляют интервальный вариационный ряд:** нужно весь промежуток изменения значений выборки (от минимального до максимального) разбить на интервалы, а затем подсчитать число значений из выборки, попадающих в каждый интервал (частоты).

При этом оптимальное количество интервалов определяется по формуле $k = 1 + 3{,}322 \cdot \lg n$

а длина интервала $h = \frac{X_{\max} - X_{\min}}{1+3{,}322 \cdot \lg n}$

Интервальный вариационный ряд



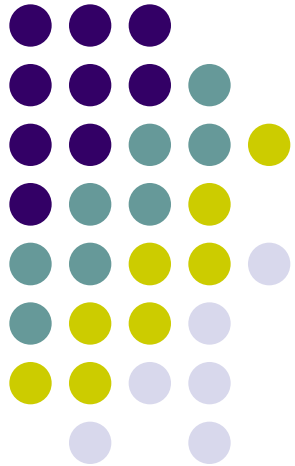
$[x_1 ; x_2)$	$[x_2; x_3)$	...	$[x_m ; x_{m+1}]$
n_1	n_2	...	n_m
$w_1 = \frac{n_1}{n}$	$w_2 = \frac{n_2}{n}$	...	$w_m = \frac{n_m}{n}$

$$k = 1 + 3,322 \cdot \lg 55 \approx 6,8$$

$$h = \frac{23-10}{6,8} \approx 2$$

Границы интервала, x_i-x_{i+1}	[10,12)	[12,14)	[14,16)	[16,18)	[18,20)	[20,22)	[22,24)
Середина интервала	11	13	15	17	19	21	**23**
Частота, n_i	x_i^* 2	4	8	12	16	10	**3**
$n_i^{нак}$ Накопленная частота,	**2**	**6**	**14**	**26**	**42**	**52**	**55**

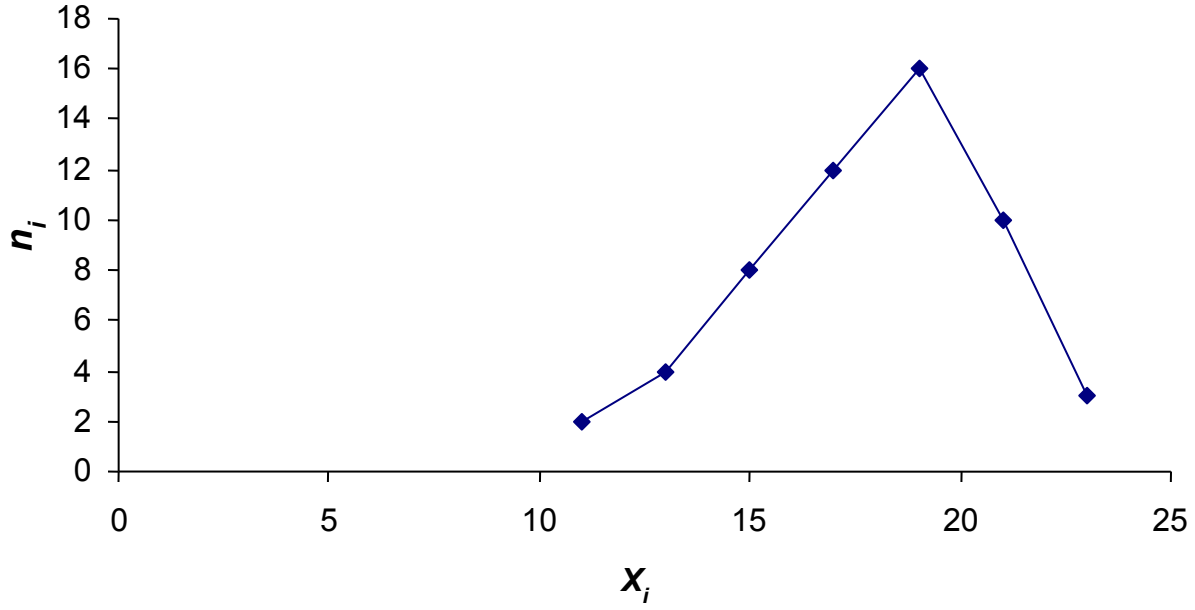
Графическое изображение вариационных рядов



Полигоном частот (относительных частот) интервального ряда называется ломаная с вершинами в точках (в точках $(x_i^*, n_i), i = \overline{1,k}$) (- середины (x_i^*, p_i) x_i^* интервалов).

Полигона для вариационного ряда

Середина интервала	11	13	15	17	19	21	23
Частота, n_i	2	4	8	12	16	10	3



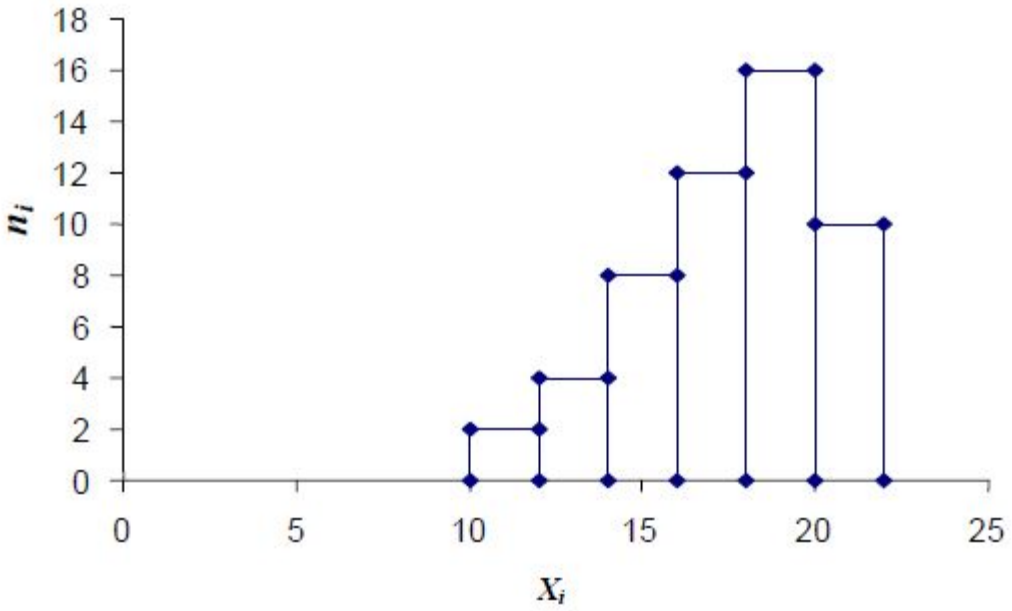
Гистограмма

Применяется для изображения только интервальных вариационных рядов и представляет собой ступенчатую фигуру из прямоугольников с основаниями, равными интервалам значений признака и высотами, равными частотам (частостям) интервалов.

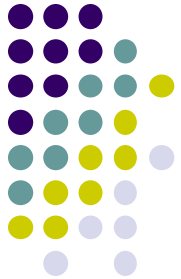
При этом по оси абсцисс откладываются интервалы, а по оси ординат – частоты (или частости) в случае равенства интервалов, или плотности распределения частот (или частостей) в случае неравенства интервалов.

Гистограмма

Границы интервала, x_i-x_{i+1}	[10,12)	[12,14)	[14,16)	[16,18)	[18,20)	[20,22)	[22,24)
Частота, n_i	2	4	8	12	16	10	3



Кумулята



Кумулятивная кривая (*кумулята*) – кривая накопленных частот.

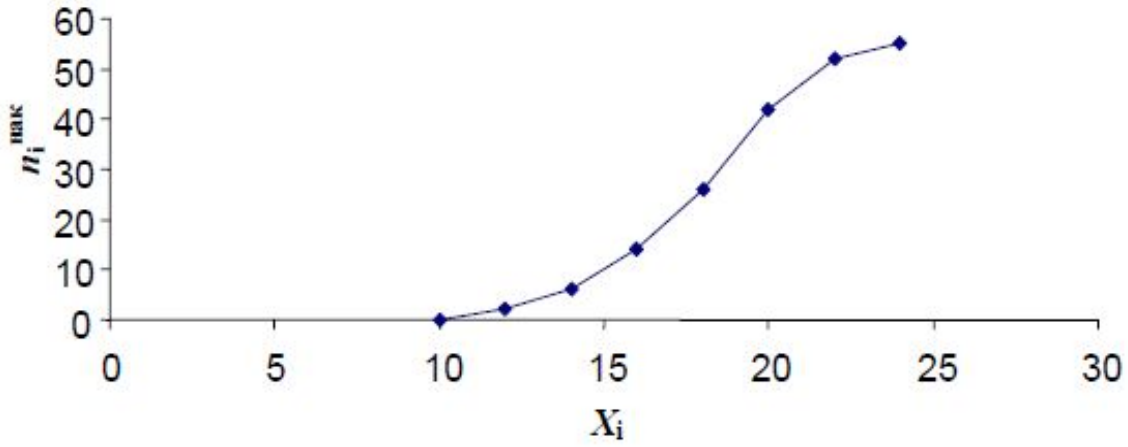
Представляет ломаную, соединяющую точки

$$(x_i ; n_i^{нак})$$

Для *интервального вариационного ряда* ломаная начинается с точки $(x_{нач} ; 0)$.

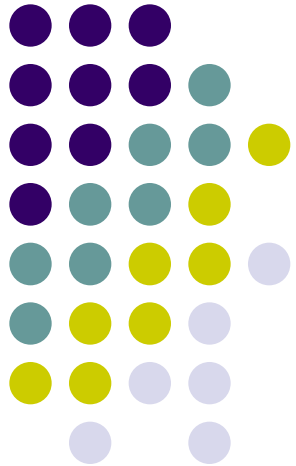
Кумулята

Границы интервала, x_i-x_{i+1}	[10,12)	[12,14)	[14,16)	[16,18)	[18,20)	[20,22)	[22,24)
Накопленная частота,	2	6	14	26	42	52	55

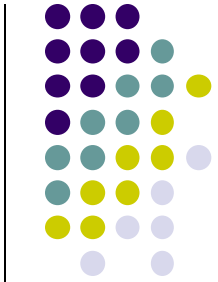


Числовые характеристики вариационного ряда:

- Средние величины
- Показатели вариации



Средние величины



Средние величины характеризуют значение признака, вокруг которого концентрируются наблюдения или, как говорят, *центральную тенденцию распределения.*

К ним относят:

среднюю арифметическую,

Моду,

медиану.

Средней арифметической *вариационного ряда* называется сумма произведений всех вариантов на соответствующие частоты, деленная на сумму частот:

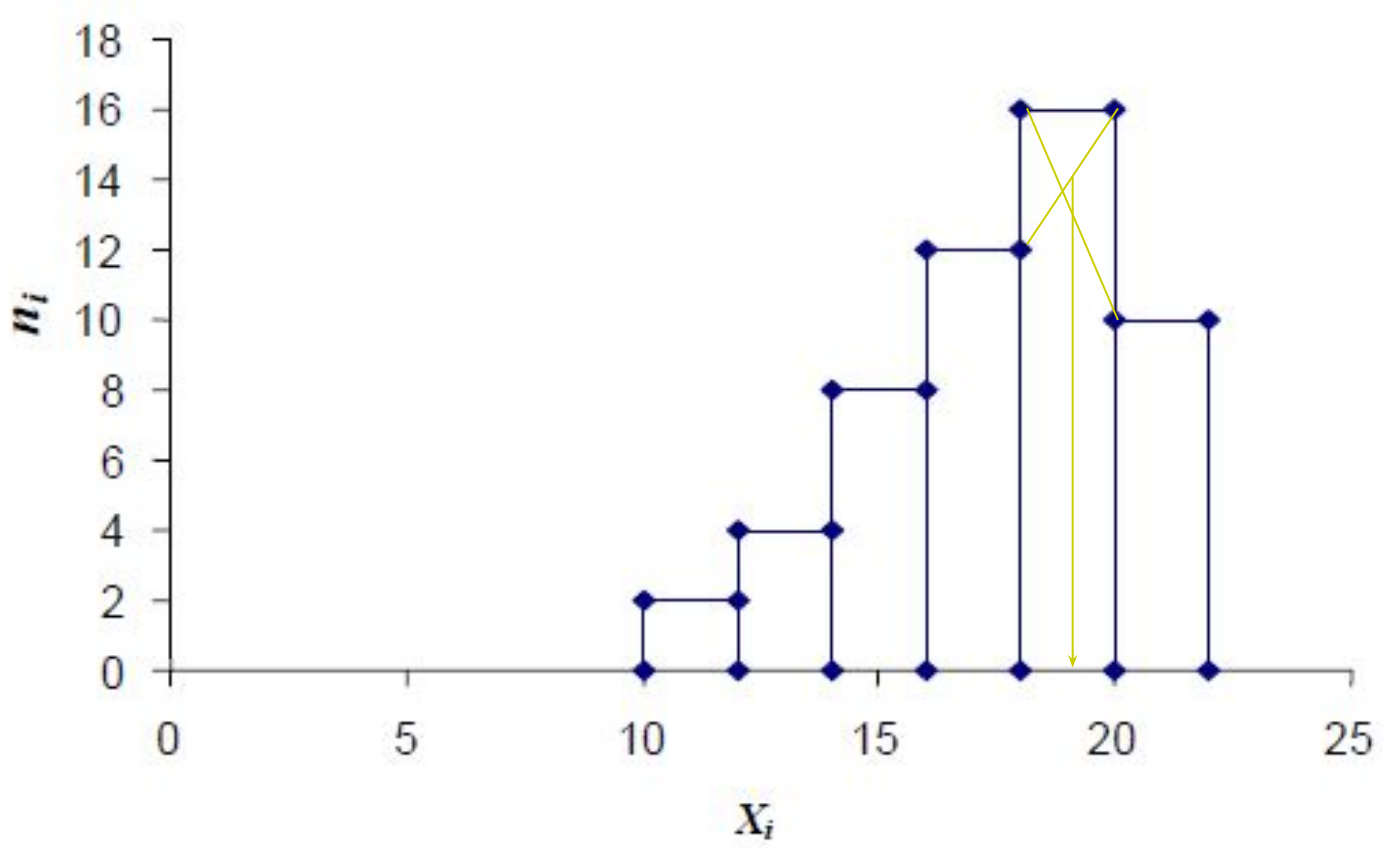
$$\overline{x} = \frac{\sum_{i=1}^{m} x_i \cdot n_i}{n},$$

где x_i *- варианты дискретного ряда или середины интервалов интервального вариационного ряда;*

n_i *- соответствующие им частоты.*

Мода (Mo) - *это значение, которое встречается в выборке наиболее часто.*

Мода может быть приближённа найдена по гистограмме: выбираем самую высокую ступеньку, её вершины крест-накрест соединяем с вершинами предшествующей и следующей за ней степеньками, из точки пересечения опускаем перпендикуляд на ось OX, это и есть мода.


18
16
14
12
10
8
6
4
2
0
n_i
0
5
10
15
20
25
X_i

Моду можно найти по формуле

$$Mo = X_{mo} + b \cdot \frac{n_{mo} - n_{mo-1}}{(n_{mo} - n_{mo-1}) + (n_{mo} - n_{mo+1})}, \text{ где}$$

X_{mo} - нижняя граница модального интервала (модальный интервал – интервал, имеющий наибольшую частоту);

n_{mo} - частота модального интервала;

n_{mo-1} - частота интервала, предшествующего модальному;

n_{mo+1} - частота интервала, последующего за модальным.

$$Mo = 18 + 2 \cdot \frac{16 - 12}{(16 - 12) + (16 - 10)} = 18{,}8, \text{ т.к.}$$

X_{mo} - нижняя граница модального интервала (18);

n_{mo} - частота модального интервала (16);

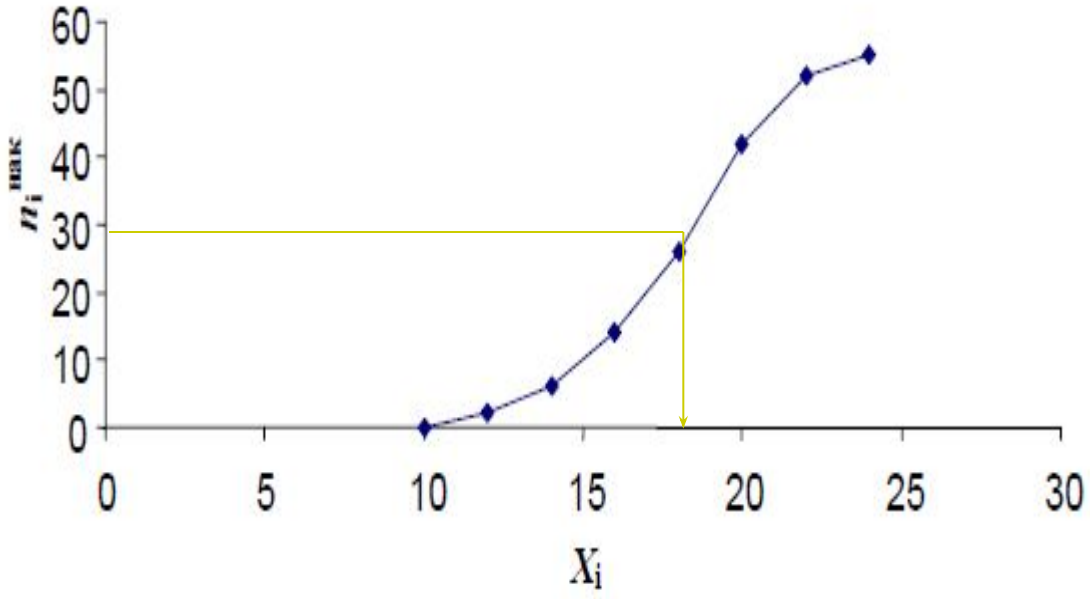
n_{mo-1} - частота интервала, предшествующего модальному (12);

n_{mo+1} - частота интервала, последующего за модальным (10).

Чаще всего встречающийся размер заработной платы среди работников предприятия примерно равен 18,8 тыс.руб.(*Mo*=18,8);

Медиана ***Me*** *- это значение, которое делит вариационный ряд пополам.*

Медиана может быть приближенно найдена с помощью кумуляты как значение признака, для которого $n_x^{нак} = \frac{n}{2}$



Медиану можно найти по формуле

$$Me = X_{me} + b \cdot \frac{\frac{n}{2} - n_{me-1}^{нак}}{n_{me}}, \text{ где}$$

X_{me} - нижняя граница медианного интервала (медианный интервал – интервал, в который попадает варианта, делящая ряд пополам);

n_{me} - частота медианного интервала;

$n_{me}^{нак}$ - накопленная частота интервала, предшествующего медианному.

$$Me = 18 + 2 \cdot \frac{\frac{55}{2} - 26}{16} = 18{,}1875, \text{ т.к.}$$

X_{me} - нижняя граница медианного интервала (18);

n_{me} - частота медианного интервала (16);

$n_{me}^{нак}$ - накопленная частота интервала, предшествующего медианному (26).

Примерно половина работников предприятия имеют заработную плату больше, чем 18,2 тыс.руб., а примерно половина работников

Показатели вариации

- ***Дисперсией*** *вариационного ряда называется средняя арифметическая квадратов отклонений вариантов от их средней арифметической*:

$$D = \frac{\sum_{i=1}^{m} \left(x_i - \overline{x} \right)^2 \cdot n_i}{n}$$

- ***Среднее квадратическое отклонение***

$$\sigma = \sqrt{D}$$

Показатели вариации

Определение. ***Коэффициент вариации*** представляет собой выраженное в процентах отношение среднего квадратического отклонения к средней арифметической:

$$V = \frac{\sigma}{\overline{X}} \cdot 100\%.$$

Замечание. Коэффициент вариации используют как характеристику однородности совокупности. Совокупность является однородной, если коэффициент вариации не превышает 33%, а средняя величина вычисленная в этой совокупности надежна или типична.

Показатели вариации

Определение. ***Коэффициент асимметрии*** представляет собой отношение центрального момента третьего порядка к кубу среднего квадратического отклонения:

$$A_s = \frac{\sum (x_i^* - \bar{x})^3 n_i}{n\sigma^3}.$$

Замечание. Коэффициент асимметрии положителен (т.е. имеет место правосторонняя асимметрия), если «длинная часть» кривой распределения расположена справа от математического ожидания (или моды), и отрицателен (т.е. имеет место левосторонняя асимметрия), если «длинная часть» кривой распределения расположена слева от математического ожидания (или моды).; если $A_s=0$, то асимметрии нет.

Показатели вариации

Определение. ***Коэффициент эксцесса*** представляет собой разность отношения центрального момента четвертого порядка к четвёртой степени среднего квадратического отклонения и 3:

$$E_k = \frac{\sum (x_i^* - \bar{x})^4 n_i}{n\sigma^4} - 3 .$$

Замечание. Коэффициент эксцесса положителен, если кривая распределения имеет более высокую и «острую» вершину, чем нормальная кривая, и отрицателен, если кривая распределения имеет более низкую и «пологую» вершину, чем нормальная кривая; если E_k=0, то распределение является нормальным.

Расчётная таблица

x_i^*	n_i	$x_i^* n_i$	$x_i^* - \bar{x}$	$(x_i^* - \bar{x})^2 n_i$	$(x_i^* - \bar{x})^3$	$(x_i^* - \bar{x})^3 n_i$	$(x_i^* - \bar{x})^4$	$(x_i^* - \bar{x})^4 n_i$
11	2	22	-6,8364	93,47	-319,5	-639,01	2184,24	4368,48
13	4	52	-4,8364	93,56	-113,12	-452,5	547,11	2188,45
15	8	120	-2,8364	64,36	-22,818	-182,55	64,72	517,77
17	12	204	-0,8364	8,394	-0,585	-7,0205	0,49	5,87
19	16	304	1,1636	21,66	1,57562	25,21	1,83	29,34
21	10	210	3,1636	100,1	31,6636	316,64	100,17	1001,72
23	3	69	5,1636	79,99	137,679	413,04	710,92	2132,77
Σ		981		461,53		-526,19		10244,4

Среднее значение:

$$\bar{x} = \frac{981}{55} \approx 17{,}84$$

Средний размер заработной платы работников предприятия равен 17,8 тыс.руб. ($\bar{x} \approx 17{,}84$);

Дисперсия:

$$D = \frac{461{,}53}{55} \approx 8{,}39$$

Среднее квадратическое отклонение:

$$\sigma = \sqrt{8{,}39} \approx 2{,}9$$

Коэффициент вариации:

$$V = \frac{2{,}9}{17{,}84} \cdot 100\% \approx 16{,}3\%$$

Выборка является однородной ($V \approx 16{,}3\% < 33\%$);

Коэффициент асимметрии:

$$A_s = \frac{-526{,}19}{55 \cdot 2{,}9^3} \approx -0{,}4$$

Присутствует несущественная левосторонняя асимметрия ($A_s \approx -0{,}4 < 0$);

Коэффициент эксцесса:

$$E_k = \frac{10244{,}4}{55 \cdot 2{,}9^4} - 3 \approx -0{,}36$$

Присутствует несущественное отклонение от нормального распределения, а именно, кривая распределения имеет более низкую и «пологую» вершину, чем нормальная кривая ($E_k \approx -0{,}36 < 0$).