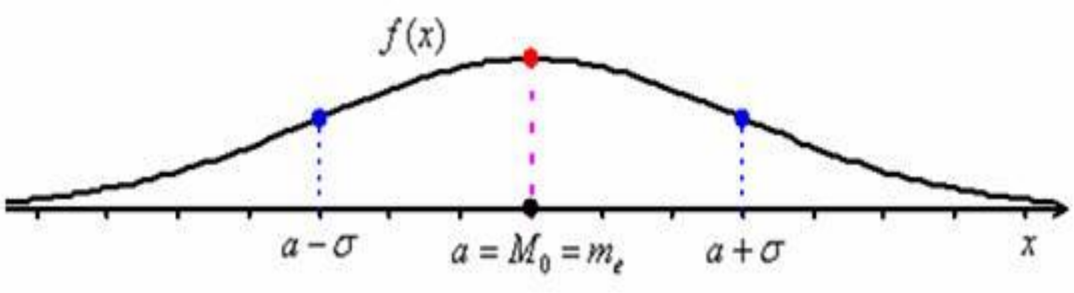
Практическое занятие 7: Асимметрия и эксцесс эмпирического распределения

Цель занятия: приобретение навыков вычисления асимметрии и эксцесса эмпирического распределения.

Асимметрия и эксцесс - это показатели, характеризующие геометрическую форму распределения. **Асимметрия** характеризует меру скошенности графика влево / вправо, а **эксцесс** – меру его высоты.

Данные показатели рассчитываются как для эмпирических, так и для теоретических распределений и за «эталон» симметрии принято нормальное распределение:



Очевидно, что любое нормальное распределение строго симметрично относительно своего центра, следовательно, его асимметрия равна нулю. Поэтому эксцесс нормального распределения (любого) принимают за «отправную» нулевую точку.

В теории вероятностей существуют строгие формулы для вычисления коэффициентов асимметрии $A_з$ и эксцесса $E_к$.

<u>Пример</u>

По результатам выборочного исследования у n =25 рабочих цеха были установлены их квалификационные разряды: 4, 5, 6, 4, 4, 2, 3, 5, 4, 4, 5, 2, 3, 3, 4, 5, 5, 2, 3, 6, 5, 4, 6, 4, 3.

Требуется: вычислить выборочную среднюю, моду и медиану.

Если даны *первичные данные* (исходные необработанные значения), то их можно тупо просуммировать и разделить результат на объём выборки:

$\bar{x}_в = \frac{4+5+6+...+4+3}{25} = 101/25 \approx 4{,}04 \approx 4$ – среднестатистический квалификационный разряд рабочих цеха.

Но во многих задачах требуется составить вариационный ряд.

x_i	n_i
2	3
3	5
4	8
5	6
6	3
Σ	25

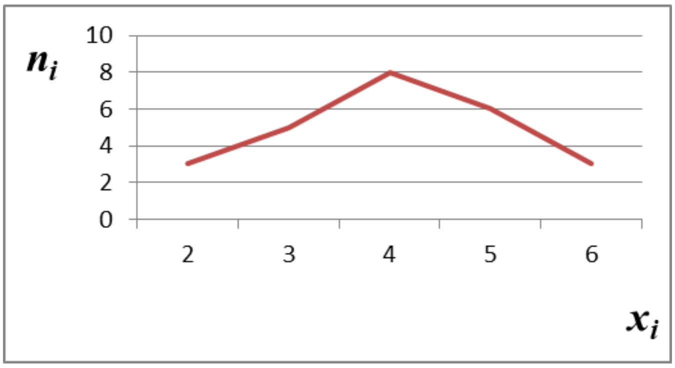
Тогда

$$\bar{x}_в == \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{2*3+3*5+4*8+5*6+6*3}{25} = 101/25 = 4{,}04$$

Мода

Мода $\boldsymbol{M_0}$ дискретного вариационного ряда – это *варианта* с максимальной частотой. В данном случае $\boldsymbol{Mo}=\boldsymbol{x_3}=4$.

Моду легко отыскать по таблице, и ещё легче на **полигоне частот** – это абсцисса самой высокой точки:



Поиск моды в среде MS Excel (без предварительного упорядочения):

4	5	6	4	4	2	3	5	4	4	5	2	3	3	4	5	5	2	3	6	5	4	6	4	3		4

=МОДА.ОДН(A1:Y1)

Иногда таковых значений несколько (с одинаковой максимальной частотой), и тогда модой считают каждое из них.

Медиана

Медиана $\boldsymbol{m_e}$ вариационного ряда – это значение, которая делит его на две равные части (по количеству вариант). Н*е важно,* ***дискретного*** *или* ***интервального****, генеральной совокупности или выборочной*.

Медиану можно отыскать несколькими способами.

Если даны первичные данные, то сортируем их по возрастанию либо убыванию и находим середину ранжированного ряда (для примера 8): $\boldsymbol{m_e=x_{13}}=4$. Почему именно 13-е число? Потому что перед ним находится 12 чисел и после него тоже 12 чисел, таким образом, значение $\boldsymbol{x_{13}}=4$ разделило ряд на две равные части, а значит, является медианой.

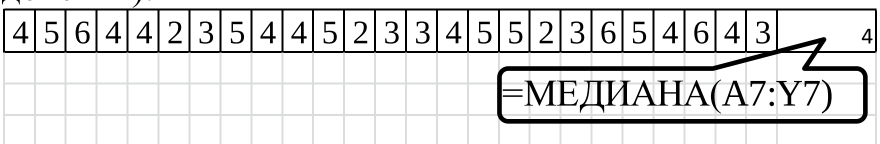
Этот номер можно найти аналитически:

- если совокупность содержит **нечётное** количество чисел, то делим её объём пополам: n/2=25/2=12,5 и округляем полученное значение в большую сторону: 13 – получая тем самым срединный номер.
- если совокупность содержит **чётное** количество чисел, например, 20, то делаем то же самое: n/2=20/2=10, и медианное значение здесь рассчитывается как среднее арифметическое 10-го и следующего числа:

$\boldsymbol{m_e=(x_{10}+x_{11})/2}$.

Изложенная инструкция работает для упорядоченного (по возрастанию либо убыванию) ряда.

Поиск медианы в среде MS Excel (без предварительного упорядочения):



Полученные значения близки друг к другу, и это говорит о симметрии вариационного ряда относительно центра, что хорошо видно по полигону частот. И с высокой вероятностью можно утверждать, что примерно так же распределена и вся генеральная совокупность (все рабочие цеха).

И тут возникает следующий закономерный вопрос: а зачем вообще нужна *мода* с *медианой*? – ведь есть *средняя*.

А дело в том, что в ряде случаев среднее значение неудовлетворительно характеризует центральную тенденцию статистической совокупности:

4	5	6	4	4	2	3	5	4	4	5	2	3	3	4	5	5	2	3	6	5	4	6	4	3	4	4,04

=МОДА.ОДН(A1:Y1)

=СРЗНАЧ((A1:Y1))

4	5	6	4	4	2	3	5	4	4	5	2	3	3	4	5	5	2	3	6	5	4	6	4	3	4	4,04

=МЕДИАНА(A7:Y7)

100	5	6	100	4	2	3	5	100	100	5	100	3	3	100	5	5	2	3	6	5	100	6	4	3	100	31,00

=МОДА.ОДН(A12:Y12)

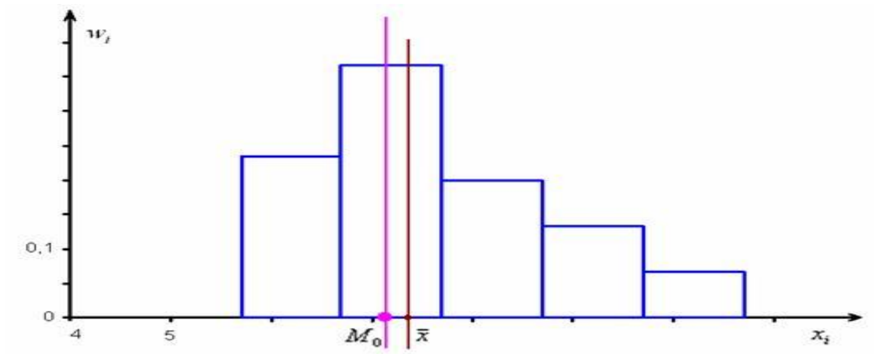
100	5	6	100	4	2	3	5	100	100	5	100	3	3	100	5	5	2	3	6	5	100	6	4	3	5	31,00

=МЕДИАНА(A16:Y16)

Простейшим критерием симметрии является ***равенство* средней, моды** и **медианы**: но в жизни такого идеального совпадения, конечно, не бывает (даже тело человека немного асимметрично), и поэтому у «почти симметричных» распределений эти показатели должны располагаться очень близко друг к другу. И в самом деле, как было вычислено в Примере:

$\bar{x}_e = 4{,}04$, $\boldsymbol{M_o}$=4, $\boldsymbol{m_e}$=4.

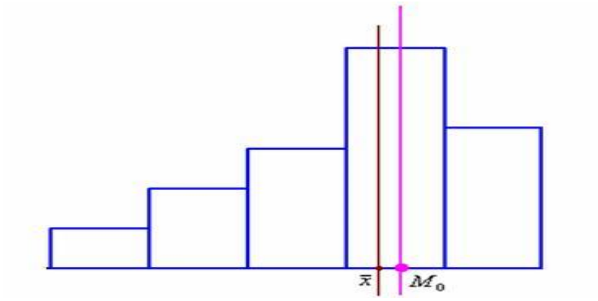
Правосторонняя асимметрия характеризуется удлинённым правым «хвостом».



Простейшим признаком правосторонней асимметрии является тот факт, что $\bar{x} > M_o$, и это неудивительно – ведь справа находится значительное количество *вариант*, и поэтому *средняя* $\bar{x}$ смещена вправо. И поэтому английский статистик **Карл Пирсон** предложил следующую формулу для расчёта ***коэффициента асимметрии***:

$A_з = \dfrac{\overline{x} - M_o}{\sigma}$, где $\boldsymbol{\sigma}$ – **среднее квадратическое отклонение** статистической совокупности. Что тоже логично, ведь у разных распределений – разный «разброс» значений и разные представления о мере асимметрии.

Левостороння асимметрия, наоборот, характеризуются удлинённым левый «хвостом» и неравенством $\overline{x} < M_o$.



Из формулы $A_з = \dfrac{\overline{x} - M_o}{\sigma}$ следует, что в левостороннем случае коэффициент асимметрии отрицателен (т.к. $\overline{x} < M_o$), а в правостороннем – положителен ($\overline{x} > M_o$), и чем больше $\boldsymbol{A_з}$ **по модулю** – тем сильнее скос распределения.

Недостаток формулы Пирсона состоит в том, что она описывает лишь центральную часть распределения и практически не учитывает «периферию».

Формула, которая охватывает все варианты для **выборочной совокупности** объёма $\boldsymbol{n}$: $A_з = \dfrac{m_з}{\sigma_в^{\ 3}}$,

где $\sigma_в^{\ 3}$ – куб **стандартного выборочного отклонения**, а $\boldsymbol{m_з}$ – так называемый *центральный эмпирический момент третьего порядка*.

Для несгруппированной статической совокупности:

$$m_з = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^3}{n}$$

Для сформированного вариационного ряда:

$$m_з = \frac{\sum_{i=1}^{k}(x_i - \overline{x}_в)^3 n_i}{n},$$

где $\boldsymbol{x_i}$ – *варианты* **дискретного ряда** или середины *частичных интервалов* **интервального ряда**, а $\boldsymbol{n_i}$ – соответствующие частоты.

Смысл знаков тот же самый: если $\boldsymbol{A_з}>0$, то распределение скошено вправо, если $\boldsymbol{A_з}<0$ – то влево.

При этом принята следующая условная градация: если полученное значение **по модулю** меньше, чем 0,25, то асимметрия незначительна, если 0,25<|$A_з$|<0,5, то умеренная, и если |$A_з$|>0,5, то существенная.

И чем МЕНЬШЕ по модулю $A_з$**,** тем рассматриваемое эмпирическое распределение **БЛИЖЕ к нормальному распределению** с параметрами $a = \bar{x}_в$. $\sigma = \sigma_в$.

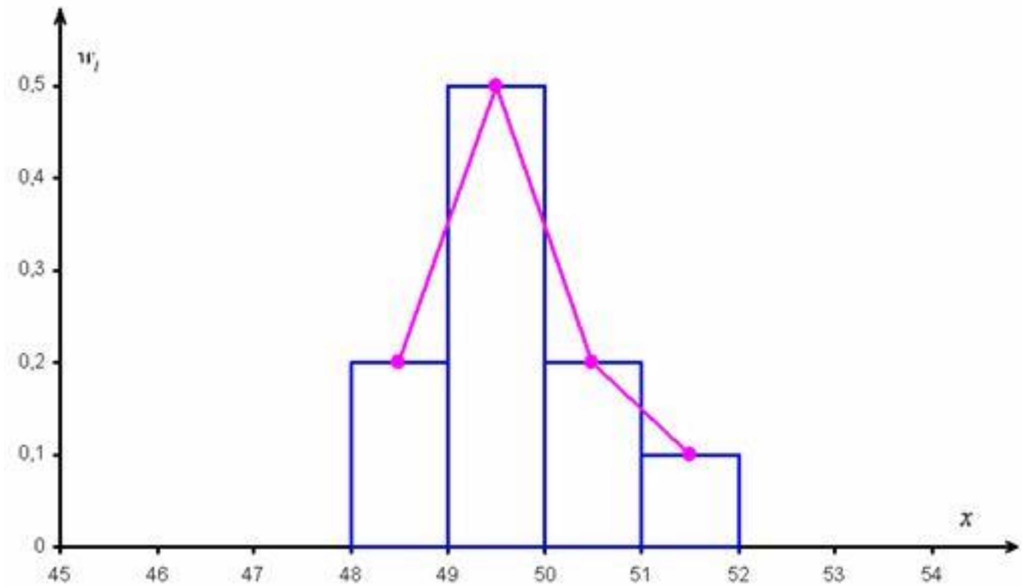
Справочно формулы из **теории вероятностей**:

асимметрия случайной величины рассчитывается по «родственной» формуле $A_з = \dfrac{\mu_3}{\delta^3}$, где σ – среднее квадратическое отклонение, а $\mu_3 = M[(X - M(X))^3]$ – *центральный теоретический момент 3-го порядка*.

Для **дискретной случайной величины** $\mu_3 = \sum (x_i - M(X))^3 p_i$

Для **непрерывной** – через интеграл: $\mu_3 = \int_{-\infty}^{\infty} (x_i - M(X))^3 f(x)dx$.

Посмотрим на чертёж *Примера* из практического занятия об **интервальном вариационном ряде**:



Видно, что гистограмма и полигон серьёзно вытянуты вверх. Но это только кажется. Дело в том, что **стандартное отклонение** $\sigma_в$ этого распределения невелико, и для сего небольшого рассеяния такая высота ДАЖЕ МАЛА. мала – по сравнению с «эталонным» нормальным распределением с параметрами $a = \bar{x}_в$. $\sigma = \sigma_в$.

Коэффициент эксцесса эмпирического распределения рассчитывается по формуле:

$$E_k = \frac{m^4}{\sigma_в{}^4} - 3,$$

где $\boldsymbol{m_4}$ - центральный эмпирический момент четвёртого порядка:

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_в)^4}{n}$$ – для несгруппированных данных

$$m_4 = \frac{\sum_{i=1}^{k}(x_i - \overline{x}_в)^4 n_i}{n}$$ – для сформированного вариационного ряда.

Если $\boldsymbol{E_k}>0$, то эмпирическое распределение является более высоким («островершинным») – относительно «эталонного» нормального распределения с параметрами $\boldsymbol{a}= \overline{x}_в$. $\boldsymbol{\sigma = \sigma_в}$. Если же $\boldsymbol{E_k}<0$ – то более низким и пологим. И чем больше $\boldsymbol{E_k}$ **по модулю**, тем «аномальнее» высота в ту или иную сторону.

Задание 1

Выборочная проверка партии чая, поступившего в торговую сеть, дала следующие результаты:

Вес, грамм, ***x***	48-49	49-50	50-51	51-52
Количество пачек, n_i	20	50	20	10

Требуется вычислить коэффициенты асимметрии и эксцесса

Методика выполнения

Интервалы		x_i	n_i	$x_i n_i$	$(x_i - x_в)^2 n_i$	$(x_i - x_в)^3 n_i$	$(x_i - x_в)^4 n_i$
48	49	48.5	20	970	28.8	-34.56	41.472
49	50	49.5	50	2475	2	-0.4	0.08
50	51	50.5	20	1010	12.8	10.24	8.192
51	52	51.5	10	515	32.4	58.32	104.976
Суммы			100	4970	76	33.6	154.72

$\bar{x}_в =$ **49,7**

$D_в =$ **0,76**

$\delta =$ **0,8717798**

Коэффициент асимметрии:

$m_3 =$ **0,336**

$m_4 =$ **1,5472**

$$A_3 = \frac{m_3}{\delta_в^3} = 0{,}5071$$

Коэффициент эксцесса:

$$E_k = \frac{m_4}{\delta_в^4} - 3 = -0{,}32133$$

Вычисленный **коэффициент асимметрии**: $A_3 \approx 0{,}51 > 0$, то есть, распределение обладает существенной правосторонней асимметрией.

Вычисленный **коэффициент эксцесса**: $E_к \approx -0{,}32$ показывает, что распределение заметно ниже, чем нормальное распределение с параметрами $\boldsymbol{a} = \bar{x}_в = 49{,}7$, $\boldsymbol{\sigma} = \boldsymbol{\sigma_в} \approx 0{,}872$.

Помимо геометрических форм, эти коэффициенты позволяют «прикинуть», насколько близка к нормальному распределению не только выборочная, но и вся генеральная совокупность.

Варианты задания 1

№ вар.	Вес, грамм, X_i				
	47-48	48-49	49-50	51-52	52-53
	Количество пачек, ni				
1	15	19	31	41	30
2	21	39	42	18	27
3	47	20	29	25	33
4	21	28	17	50	33
5	32	21	34	27	35
6	25	27	22	18	43
7	35	12	30	30	40
8	19	43	30	18	18
9	27	35	50	43	36
10	45	31	22	42	34
11	15	10	23	10	43
12	30	39	34	19	31
13	42	41	37	18	25

№ вар.	Вес, грамм, X_i				
	47-48	48-49	49-50	51-52	52-53
	Количество пачек, ni				
14	10	43	10	12	22
15	31	49	25	26	23
16	40	18	26	48	25
17	24	46	28	42	46
18	13	38	49	13	13
19	29	11	24	43	36
20	15	29	20	31	15
21	34	49	39	46	49
22	39	36	25	25	22
23	27	10	27	28	36
24	44	43	32	44	49
25	32	26	39	33	13
26	34	21	45	41	20

Задание 2

В результате эксперимента получены данные, записанные в виде статистического ряда:

17,1	21,4	15,9	19,1	22,4	20,7	17,9	18,6	21,8	16,1
19,1	20,5	14,2	16,9	17,8	18,1	19,1	15,8	18,8	17,2
16,2	17,3	22,5	19,9	21,1	15,1	17,7	19,8	14,9	20,5
17,5	19,2	18,5	15,7	14	18,6	21,2	16,8	19,3	17,8
18,8	14,3	17,1	19,5	16,3	20,3	17,9	23	17,2	15,2
15,6	17,4	21,3	22,1	20,1	14,5	19,3	18,4	16,7	18,2
16,4	18,7	14,3	18,2	19,1	15,3	21,5	17,2	22,6	20,4
22,8	17,5	20,2	15,5	21,6	18,1	20,5	14	18,9	16,5
20,8	16,6	18,3	21,7	17,4	23	21,1	19,8	15,4	18,1
18,9	14,7	19,5	20,9	15,8	20,2	21,8	18,2	21,2	20,1

В условии речь идёт о результатах эксперимента, а значит, перед нами выборочная совокупность, т.к. теоретически опыты можно повторять бесконечное количество раз.

Задание

1) Составить интервальный вариационный ряд, состоящий из 9 равных интервалов.

2) Построить гистограмму относительных частот и эмпирическую функцию распределения.

3) Найти моду и медиану.

4) Вычислить выборочную среднюю, дисперсию, среднее квадратическое отклонение и коэффициент вариации.

5) Вычислить коэффициенты асимметрии и эксцесса, сделать выводы.

Методика выполнения

1) Составление интервального вариационного ряда, состоящего из 9 равных интервалов

Минимальное значение $x_{\min} =$ **14**

Максимальное значение $x_{\max} =$ **23**

Размах вариации $R = x_{\max} - x_{\min} =$ **9**

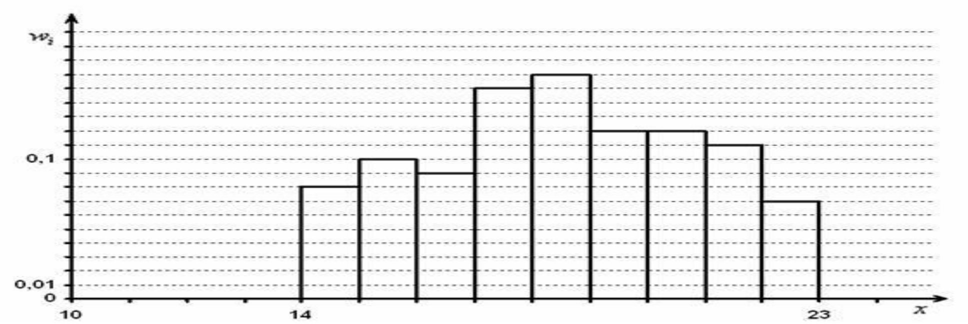
Количество частичных интервалов $k =$ **9** *(по условию)*

Длина интервала $h = \dfrac{x_{\max} - x_{\min}}{k} =$ **1**

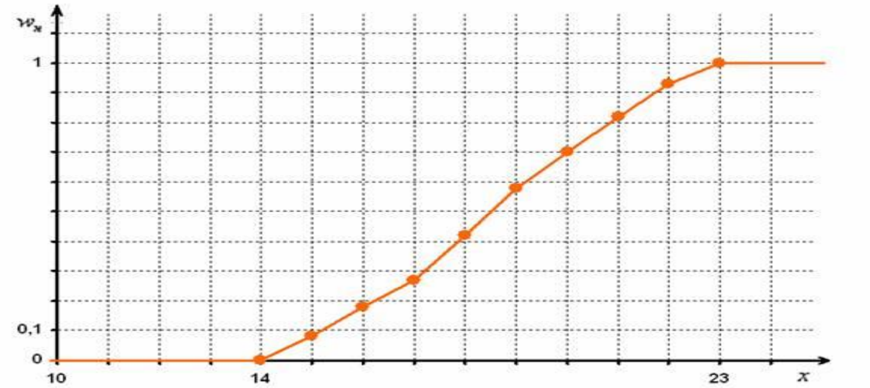
Разметить интервалы и подсчитать частоту по каждому интервалу,
вычислить относительные частоты и относительные накопленные частоты:

Интервалы		n_i	w_i	$w_н$
14	15	8	0,08	0,08
15	16	10	0,1	0,18
16	17	9	0,09	0,27
17	18	15	0,15	0,42
18	19	16	0,16	0,58
19	20	12	0,12	0,7
20	21	12	0,12	0,82
21	22	11	0,11	0,93
22	23	7	0,07	1
	Суммы:	100	1	

2) Строим гистограмму относительных частот:



и эмпирическую функцию распределения:



3) Вычисление моды и медианы

x $_0$ =18 - нижняя граница модального интервала

h =1 - длнна модального интервала

n $_M$ =16 - частота модального интервала

n $_{M-1}$ =15 - частота предыдущего интервала

n $_{M+1}$ =16 - частота следующего интервала

Мода $M_0 = x_0 + \dfrac{n_M - n_{M-1}}{(n_M - n_{M-1}) + (n_M - n_{M+1})} \cdot h =$ **18,2** ед.

Медиана $m_e = x_0 + \dfrac{0{,}5n - n^{н}_{m-1}}{n_m} \cdot h =$ 18+(0,5*100-42)/16= **18,5** ед.

n =100 =- объём выборочной совокупности

половину вариант содержит интервал (18; 19) и ***x*** $_0$ =18 - его нижняя граница

h =1 - длина мелианного интервала

n $_m$ =16 - частота медианного интервала

n $^{н}_{m-1}$ = 8+10+9+15 = 42 - накопленная частота предыдущего интервала

4) Вычисление выборочной средней, дисперсии, среднего квадратического отклонения и коэффициента вариации.

x_i	n_i	$x_i n_i$	$(x_i - \bar{x}_в)^2 n_i$	$(x_i - \bar{x}_в)^3 n_i$	$(x_i - \bar{x}_в)^4 n_i$
14,5	8	116	129,2832	-519,718464	2089,268225
15,5	10	155	91,204	-275,43608	831,8169616
16,5	9	148,5	36,7236	-74,181672	149,8469774
17,5	15	262,5	15,606	-15,91812	16,2364824
18,5	16	296	0,0064	-0,000128	2,56E-06
19,5	12	234	11,5248	11,294304	11,06841792
20,5	12	246	47,0448	93,148704	184,4344339
21,5	11	236,5	97,6844	291,099512	867,4765458
22,5	7	157,5	110,8828	441,313544	1756,427905
Суммы	100	1852	539,96	-48,3984	5906,575952

$\bar{x}_в = \dfrac{\sum x_i n_i}{n} =$ **18,5**

$D_в = \dfrac{\sum (x_i - \bar{x}_в)^2 n_i}{n} =$ **5,3996**

$\sigma_в = \sqrt{D_в} =$ **2,3237**

$V = \dfrac{\sigma_в}{\bar{x}_в} \cdot 100\% =$ ***12,55*** %

5) Вычисление коэффициентов асимметрии и эксцесса.

$$m_3 = \frac{\sum (x_i - \bar{x}_в)^3 n_i}{n} = \textbf{-0,484} \qquad m_4 = \frac{\sum (x_i - \bar{x}_в)^4 \cdot n_i}{n} = \textbf{59,07}$$

$$A_s = \frac{m_3}{\sigma_в^3} = \textbf{-0,039} \qquad E_k = \frac{m_4}{\sigma_в^4} - 3 = \textbf{-0,97}$$

Таким образом, выборочная совокупность практически симметрична, но несколько ниже, чем нормальное распределение с параметрами $\boldsymbol{a} = \bar{x}_в = 18{,}52$, $\boldsymbol{\sigma} = \boldsymbol{\sigma_в} \approx 2{,}3237$.

Варианты задания 2

№ вар.	Данные экспериментов, x_i															
1	16,1	21,3	18,7	19,6	20,8	19,3	14,5	19,0	16,2	17,9	15,3	18,8	14,7	18,8	23,6	21,8	22,5
2	23,5	13,1	17,6	19,7	23,5	17,7	21,0	16,6	23,4	17,9	19,3	21,6	16,2	23,0	14,6	17,3	18,9
3	21,1	18,6	21,9	16,5	23,0	16,5	22,4	18,8	23,8	15,0	23,5	19,1	18,2	15,8	19,2	14,3	20,3
4	17,5	13,1	16,8	19,3	20,5	14,0	23,6	24,0	15,4	16,0	15,0	21,6	19,3	18,7	17,4	17,9	16,2
5	23,0	19,1	18,1	13,3	19,3	17,9	23,7	17,1	21,9	17,1	17,4	22,7	16,3	14,9	23,0	14,5	20,7
6	19,4	23,0	22,2	17,5	23,5	17,6	15,5	21,4	14,9	19,7	18,2	22,1	20,3	15,4	13,8	18,1	16,5
7	17,6	14,3	18,8	16,0	17,3	21,1	18,6	14,2	13,9	22,2	16,9	21,6	21,9	17,9	16,1	19,8	21,0
8	17,7	14,3	22,8	23,8	18,5	23,1	16,5	15,1	21,5	18,9	18,0	13,2	23,3	21,2	21,2	19,9	13,7
9	21,4	20,5	19,2	20,7	15,0	14,3	22,0	17,5	19,5	16,0	18,7	18,5	23,6	21,5	23,0	18,5	18,4
10	15,0	21,5	23,3	19,1	22,0	20,7	21,6	20,4	15,7	18,9	22,5	19,6	18,7	18,1	16,8	17,2	22,0
11	15,7	22,4	17,2	18,0	16,2	17,6	13,8	15,6	15,2	15,5	18,3	16,3	13,2	19,1	20,8	22,8	14,8
12	20,4	16,1	18,9	15,4	16,5	20,9	20,3	15,8	17,1	13,7	14,3	14,9	18,2	21,9	20,4	18,0	14,5
13	20,7	19,1	14,5	16,5	23,2	13,1	17,6	17,0	16,7	23,2	17,5	14,1	20,3	22,5	20,4	13,8	13,2
14	13,3	18,7	16,7	14,0	16,4	22,0	15,9	13,4	21,8	17,6	16,2	16,3	16,8	18,8	23,3	17,5	22,3
15	21,8	21,4	23,5	16,5	22,3	16,2	19,7	13,6	21,6	13,9	19,4	14,2	16,2	22,2	19,1	13,4	19,5
16	17,8	19,2	17,4	21,2	17,3	17,6	13,7	16,1	18,2	20,4	16,0	16,4	14,0	14,1	13,9	22,0	19,0
17	14,5	21,7	14,8	14,6	20,9	17,9	13,9	19,4	16,8	14,3	23,6	18,8	15,6	23,0	13,3	18,7	19,3
18	15,9	18,1	21,8	15,1	16,8	18,5	18,6	18,9	20,0	15,8	21,9	18,8	22,6	15,0	13,4	16,8	19,8
19	13,8	20,4	20,4	22,9	13,7	14,5	20,3	17,7	21,8	18,5	19,3	19,6	23,2	18,6	14,6	19,5	22,3
20	22,1	18,9	18,2	19,3	15,7	19,9	18,7	18,8	14,7	19,2	17,0	13,6	19,6	14,5	20,5	18,3	14,9
21	15,0	13,4	15,4	23,3	14,4	20,3	17,1	14,3	18,6	23,3	18,5	14,1	23,1	20,0	15,9	18,1	18,6
22	14,3	18,6	13,2	23,8	15,2	17,1	22,4	17,1	15,7	19,8	22,6	18,8	13,7	14,9	16,3	20,9	21,2
23	15,2	23,8	16,9	13,7	16,4	16,7	18,5	13,9	13,5	15,8	14,3	21,2	19,3	22,8	14,9	22,1	23,5
24	15,8	22,0	23,7	22,5	14,9	23,6	23,4	20,9	16,1	13,0	15,2	19,2	17,9	16,3	21,9	17,8	23,4
25	18,4	13,9	19,4	17,6	17,1	15,1	17,6	23,8	16,8	14,5	21,6	21,3	20,9	20,5	17,3	13,6	18,4