Прикладные вопросы теории вероятностей и математической статистики в моделировании социально-экономических процессов

Цель работы - ознакомиться с вероятностно-статистическими методами моделирования процессов и систем.

Основные понятия

Исходные данные для моделирования и его результаты часто представляют собой массивы случайных чисел. Это относится как к составляющим вектора внешних сил (снеговая нагрузка, ветровая нагрузка и др.), так и к выходным данным, например, к результатам многократных повторений машинного эксперимента (прогонов модели). Такие массивы должны быть упорядочены с целью получения данных, удобных для моделирования или разработки практических рекомендаций по результатам моделирования. Обработку массивов случайных чисел производят по правилам математической статистики.

Практическое значение вероятностных методов состоит в том, что они позволяют по известным характеристикам простых случайных явлений прогнозировать характеристики более сложных явлений. В данной работе будут рассмотрены примеры решения типовых задач статистической обработки выборки, вопросы графического представления выборки, вычисления ее числовых характеристик и проверки близости эмпирической и теоретической функций распределения.

Порядок выполнения работы

По результатам наблюдений над случайной величиной требуется:

Задание 1. Вариационные ряды. Эмпирическая функция распределения

1. Найти дискретный вариационный ряд, выборочную (эмпирическую) функцию распределения данной выборки и построить ее график.
2. Найти интервальный вариационный ряд, выборочную (эмпирическую) функцию распределения данной выборки и построить ее график

Задание 2. Графическое представление выборочных данных

Для дискретного и интервального вариационных рядов построить гистограмму, полигон в Excel.

Задание 3. Расчет числовых характеристик выборки

Найти числовые характеристики выборки с помощью встроенных функций Excel.

Задание 4. Подобрать подходящее теоретическое распределение и проверить гипотезу о согласии эмпирического распределения с теоретическим распределением с помощью критериев Пирсона

Варианты заданий

Вариант 1. X - средняя прочность бетона (в МПа). Приведены результаты измерения средней прочности бетона у 100 железобетонных изделий.

22,0 21,5 23,5 23,0 21,0 21,5 21,0 22,5 21,0 20,0 19,0 20,5 19,0 20,5
21,5 21,0 21,0 21,5 21,0 21,5 21,0 21,5 20,0 22,0 20,0 19,5 22,0 22,5
21,5 21,0 21,0 21,0 22,5 21,5 21,0 20,0 20,5 21,0 21,0 22,0 20,0 20,5
22,5 23,5 20,0 22,5 23,5 19,5 21,0 23,0 20,0 21,5 22,5 22,0 19,5 20,5
21,0 21,5 21,5 21,0 18,5 20,5 19,5 22,5 22,0 20,5 22,0 23,0 21,5 21,0
21,5 20,5 18,5 23,5 19,0 21,0 21,0 21,5 20,5 23,5 22,5 22,0 21,0 22,5
19,0 22,5 24,5 24,5 21,0 21,5 21,0 21,5 21,5 24,0 20,0 25,5 20,0 18,0
22,5 20,5

Вариант 2. С целью определения оптимального количества цемента для укрепления грунта испытано 100 образцов цементо-грунта. X - количество цемента в % к массе грунта. Приведены результаты испытаний.

3,0 4,0 4,1 3,9 3,8 4,3 4,3 3,7 3,1 3,6 4,2 3,8 3,9 3,2 3,6 3,8 3,9 2,9 3,7
3,4 4,0 4,7 3,6 3,2 3,2 3,7 3,8 3,6 3,3 3,1 4,2 4,6 4,3 4,1 3,9 3,7 3,3 3,4
3,7 3,7 4,3 3,6 4,1 4,2 4,1 3,8 4,1 3,5 3,8 3,5 4,0 3,7 3,9 3,6 3,6 3,7 3,4
3,5 3,7 3,5 3,0 4,0 3,7 3,9 3,7 3,4 3,6 3,8 4,8 3,4 3,3 3,8 3,3 3,9 4,0 3,8
3,7 4,1 3,4 4,0 3,2 3,7 4,4 3,7 4,1 4,2 3,8 3,9 4,1 3,9 3,6 3,8 3,7 3,6 3,5
3,9 4,6 3,5 4,4 4,0

Вариант 3. Прочность бетона при его твердении со временем возрастает. Для анализа кинетики твердения бетона произвели испытания 100 стандартных образцов. X - время твердения (в сутках). Приведены результаты испытаний.

17 17 13 16 7 8 10 9 10 12 7 10 16 15 12 14 15 9 14 11 18 13 13 10 21 11
10 11 13 11 5 14 19 15 9 15 11 22 16 14 13 14 13 20 12 3 17 20 18 14 13
17 12 12 6 8 9 13 8 20 16 15 12 14 18 11 15 5 17 18 11 13 13 11 13 13 11
14 13 15 15 16 13 7 13 19 17 19 9 11 8 14 7 9 16 14 14 15 18 17

Вариант 4. X - количество бракованных железно-бетонных изделий в смену (в %). Приведены результаты оценок брака за 100 смен.

3,0 4,0 3,7 4,1 3,6 3,8 4,9 3,4 3,3 3,3 3,9 4,0 3,8 3,7 4,1 3,4 4,0 2,9 3,7
4,4 3,8 4,1 3,8 3,9 4,1 4,1 3,6 3,8 3,7 3,6 3,5 3,8 3,9 4,6 3,5 4,4 4,0 3,5
4,3 3,8 2,8 3,2 3,1 4,0 3,7 3,6 3,5 3,6 3,5 3,7 3,4 3,8 4,1 3,7 4,0 3,8 3,0
4,0 4,1 3,9 3,8 3,6 4,0 3,7 3,1 3,6 4,2 3,7 3,8 3,0 3,6 3,8 3,9 3,4 3,7 3,4
4,0 4,7 3,6 3,5 3,2 3,7 3,8 3,6 3,3 3,7 4,2 4,6 4,3 4,1 3,9 3,3 3,4 3,7 3,7
4,3 3,9 3,7

Вариант 5. X - предел текучести стали (в кг / мм). Приведены результаты испытаний 100 различных марок стали.

51 42 68 53 49 79 35 63 55 29 42 42 17 45 38 56 29 25 41 37 52 40 68 47 46 51 38 47 60 53 67 41 26 47 90 63 34 57 45 72 40 76 75 15 35 28 71 60 56 43 52 63 75 30 61 68 64 18 65 48 66 18 87 51 48 36 32 31 46 67 60 78 41 54 66 54 21 39 74 24 39 35 50 35 72 78 65 44 53 71 65 33 52 49 30 59 80 20 26 36

Вариант 6. X - количество бракованных труб в смену (в м). Приведены результаты оценок брака за 100 смен.

13 13 11 13 13 11 14 13 15 15 16 22 7 13 19 17 19 16 11 8 14 7 9 14 16 14 14 15 18 12 8 10 9 10 11 5 15 14 20 12 8 14 18 11 11 13 9 19 11 15 5 17 18 15 20 16 15 12 14 9 14 11 18 13 13 21 16 15 12 14 15 11 10 9 10 12 7 17 17 13 16 7 11 12 6 8 9 13 13 18 14 13 17 17 13 20 12 3 18 10

Вариант 7. X - средняя прочность бетона (в МПа). Приведены результаты измерения средней прочности бетона у 100 железобетонных изделий.

18,5 20,5 19,5 22,0 22,5 20,5 22,0 23,0 21,5 21,0 21,0 21,5 20,5 18,5 23,5 21,0 19,0 2,5 20,5 23,5 22,5 21,0 22,0 22,5 19,0 22,5 24,5 21,0 24,5 21,5 21,0 21,5 21,5 20,0 24,0 25,0 20,0 18,0 22,5 22,0 20,5 21,5 23,0 23,0 21,0 21,0 21,5 22,5 21,0 20,0 19,0 19,0 20,5 20,5 21,5 21,0 21,0 21,0 21,5 21,5 21,0 21,5 20,0 20,0 22,0 19,5 22,0 22,5 21,5 21,0 21,0 21,0 21,5 21,0 21,5 20,0 21,0 20,5 21,0 22,0 21,5 21,5 21,0 20,0 21,5 22,5 22,0 20,5 19,5 21,0 23,0 20,0 23,5 22,5 19,5 20,0 20,5 23,5 22,5 21,0

Вариант 8. X - отклонение диаметра трубы от нормативного вследствие коррозии (в мм). Приведены результаты исследования 100 труб одинакового диаметра.

0,62 0,69 0,80 0,63 1,02 1,10 0,72 0,96 0,80 0,88 0,63 0,84 0,58 0,80 0,60 0,76 0,87 0,96 0,72 0,82 0,95 0,82 1,03 0,95 0,67 1,06 0,90 0,91 0,75 0,96 0,73 0,97 0,70 0,69 0,69 0,61 1,04 0,78 0,98 0,93 0,90 0,83 0,79 0,71 0,61 0,70 0,81 0,56 0,80 0,88 0,89 1,10 0,83 0,58 0,85 0,57 0,95 0,76 0,78 0,97 0,55 0,55 0,94 0,90 0,86 0,81 0,79 0,74 0,89 1,01 0,63 1,02 0,98 0,65 0,95 0,93 0,86 0,72 0,89 0,80 0,94 1,03 0,63 0,92 1,05 0,89 0,89 0,65 0,77 0,84 0,58 0,82 0,73 1,09 0,78 0,58 0,92 0,82 1,08 0,85

Вариант 9. X - количество бракованных железно-бетонных изделий в смену (в %). Приведены результаты оценок брака за 100 смен.

3,3 3,4 3,7 3,7 4,3 4,1 4,1 4,2 4,1 3,8 4,1 4,2 3,8 3,5 4,0 3,7 3,9 3,8 3,6 3,7 3,4 3,5 3,7 3,9 3,0 4,0 4,1 3,9 3,8 4,1 4,3 3,7 3,1 3,6 4,2 3,4 3,9 3,2 3,6 3,8 3,9 4,0 3,7 3,4 4,0 4,7 3,6 2,9 3,2 3,7 3,8 3,6 3,3 3,7 4,2 4,6 4,3 4,1 3,9 4,4 3,0 4,0 3,7 3,9 3,7 3,8 3,6 3,8 4,8 3,4 3,3 3,9 3,4 3,9 4,1 3,7 4,0 4,0 3,7 3,6 3,5 3,6 3,5 3,8 4,3 3,8 2,8 3,2 3,1 3,7 3,9 4,7 3,5 4,4 4,1 3,6 3,9 3,6 3,7 3,4

Вариант 10. Прочность бетона при его твердении со временем возрастает. Для анализа кинетики твердения бетона произвели испытания 100 стандартных образцов. X - время твердения (в сутках). Приведены результаты испытаний.

21 11 10 11 13 5 14 19 15 9 11 22 16 14 12 13 20 12 3 18 18 14 13 17 17
6 8 9 13 17 17 13 16 7 10 9 10 12 7 16 16 12 14 15 14 11 18 13 13 20 16
15 12 14 11 15 5 17 18 18 11 11 13 9 15 14 20 12 8 8 16 14 14 15 18 11 8
14 7 9 7 13 19 17 19 14 13 15 16 13 13 11 13 13 10 9 10 11 16 12

Контрольные вопросы

1. Случайные величины, законы их распределения.
2. Основные виды теоретических распределений дискретной случайной величины.
3. Основные виды теоретических распределений непрерывной случайной величины.
4. Точечные оценки.
5. Доверительные интервалы. Надежность. Точность.
6. Статистическая проверка статистических гипотез.
7. Подбор подходящего теоретического распределения. Критерии согласия.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Задание 1. Вариационные ряды. Эмпирическая функция распределения

Краткая теория

Для решения задач, связанных с анализом данных при наличии случайных непредсказуемых воздействий, разработан математический аппарат – математическая статистика, что позволяет выявлять закономерности на основе случайностей, делать на их основе обоснованные выводы и прогнозы.

Важнейшими понятиями математической статистики являются понятия генеральной совокупности и выборки.

Генеральной совокупностью наблюдаемого признака (случайной величины) X называют множество всевозможных значений, принимаемых наблюдаемым признаком X.

Часть отобранных объектов из генеральной совокупности называется **выборочной совокупностью**, или **выборкой**. Результаты измерений изучаемого признака n объектов выборочной совокупности порождают n значений $x_1, x_2, \dots, x_n$ случайной величины X. Число n называется **объемом выборки**.

Выборку можно рассматривать двояко:

а) как случайный вектор длины n, каждая компонента которого имеет такое же распределение, как и наблюдаемый признак;

б) как на результаты измерений, т.е. набор nчисел.

Случайная величина X называется **дискретной случайной величиной**, если она принимает свое значение из некоторого конечного фиксированного набора, например, случайная величина X – число появления шестерки при двух бросках игрального кубика

$$X: 0,1,2\ .$$

Случайная величина X называется **непрерывной случайной величиной**, если она принимает любое значение из некоторого интервала (в том числе $-\infty$ и $+\infty$), например, рост человека.

После получения выборки имеем данные, которые представляют собой множество чисел, расположенных в беспорядке. Анализ таких данных весьма затруднителен, и для изучения скрытых закономерностей их подвергают определенной обработке.

Простейшая операция – **ранжирование** опытных данных, результатом которого являются значения, расположенные в порядке не убывания. Если среди элементов встречаются одинаковые, то они объединяются в одну группу. Значение случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется **вариантом**, а изменение этого значения – **варьированием**. Варианты будем обозначать строчными буквами с соответствующими порядковому номеру группы индексами $x^{(1)}, x^{(2)}, ..., x^{(N)}$, где N– число групп. При этом $x^{(1)} < x^{(2)} < ... < x^{(N)}$.

Численность отдельной группы сгруппированного ряда данных называется частотой n_i, где i– индекс варианта, а отношение частоты данного варианта к общей сумме частот называется частностью (или относительной частотой) и обозначается $\omega_i, i = 1, ..., N$, т.е.

$$\omega_i = \frac{n_i}{\sum_{j=1}^{N} n_j},$$

при этом $\sum_{j=1}^{N} n_j = n$ – объему выборки.

Дискретным вариационным рядом называется ранжированная совокупность вариантов $x^{(i)}$ с соответствующими им частотами n_i или частностями ω_i.

Если число возможных значений дискретной случайной величины достаточно велико или наблюдаемая случайная величина является непрерывной, то строят **интервальный вариационный ряд**, под которым понимают упорядоченную совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частностями попаданий в каждый из них значений случайной величины.

Как правило, частичные интервалы, на которые разбивается весь интервал варьирования, имеют одинаковую длину Δ, которая может быть вычислена по следующей формуле

$$\Delta = \frac{R}{N} = \frac{x_{max} - x_{min}}{N}.$$

где R – размах варьирования (изменения) случайной величины;

x_{max} , x_{min} – наибольшее и наименьшее значения исследуемой случайной величины;

N – число частичных интервалов группировки.

Некоторые авторы рекомендуют пользоваться следующими эмпирическими формулами для определения числа интервалов:

$$N = \sqrt{n} \quad ,N = 5 \cdot lg(n) \, ,$$

$N = 1 + 3{,}322 \cdot lg(n)$ – формула Стерджеса.

В рекомендациях по стандартизации Р 50.1.033-2001 "Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат" рекомендует следующие значения N в зависимости от объема выборки n:

Объем выборки n	Число интервалов группировки N
40 – 100	7 – 9
100 – 500	8 – 12
500 – 1000	10 – 16
1000 – 10000	12 – 22

В теории вероятностей для характеристики распределения случайной величины X служит функция распределения

$$F(x) = P(X < x) \, ,$$

определяющую для каждого значения x вероятность того, что случайная величина X примет значение, меньшее x, т.е. равная вероятности события $A = \{X < x\}$, где x – любое действительное число.

Одной из основных характеристик выборки является **выборочная (эмпирическая) функция распределения**

$$F_n^*(x) = \frac{n_x}{n} \, ,$$

где n_x – количество элементов выборки, меньших чем X. Другими словами, $F_n^*(x)$ есть относительная частота появления события $A = \{X < x\}$ в n независимых испытаниях. Главное различие между $F(x)$ и $F_n^*(x)$ состоит в том, что $F(x)$ определяет вероятность события A, а

выборочная функция распределения $F_n^*(x)$ – относительную частоту этого события.

Свойства функции $F_n^*(x)$:

1. $0 \le F_n^*(x) \le 1$.

2. $F_n^*(x)$ – неубывающая функция.

3. $F_n^*(-\infty) = 0; \quad F_n^*(+\infty) = 1$.

Функция $F_n^*(x)$ является "ступенчатой", имеются разрывы в точках, которым соответствуют наблюдаемые значения вариантов. Величина скачка равна относительной частоте варианта.

Аналитически $F_n^*(x)$ задается следующим соотношением:

$$F_n^*(x) = \begin{cases} 0 \text{ при } x \le x^{(1)}; \\ \sum_{j=1}^{i-1} \omega_j \text{ при } x^{(i-1)} < x \le x^{(i)}, \qquad i = 2,3 \dots, N; \\ 1 \text{ при } x > x^{(N)}, \end{cases}$$

где ω_i – соответствующие относительные частоты;

$x^{(i)}$ – элементы вариационного ряда (варианты).

Замечание. В случае интервального вариационного ряда под $x^{(i)}$ понимается середина i-го частичного интервала. Эмпирическую функцию распределения непрерывной случайной величины так же называют «накопленная частота».

Перед вычислением $F_n^*(x)$ полезно построить дискретный или интервальный вариационный ряд.

Пример выполнения

Постановка задачи 1. На телефонной станции проводились наблюдения над числом неправильных соединений в минуту. Наблюдения в течение 30 минут дали следующие результаты (табл. 1).

3	0	1	5	1	2	4	5	3	4
2	4	2	0	2	3	1	3	2	1
4	3	0	2	1	0	4	2	3	2

Требуется найти дискретный вариационный ряд, выборочную (эмпирическую) функцию распределения данной выборки и построить ее график.

Решение.

Очевидно, что число X является дискретной случайной величиной, а полученные данные есть значения этой случайной величины.

В результате выполнения операций ранжирования и группировки были получены шесть значений случайной величины (варианты): 0; 1; 2; 3; 4; 5. При этом значение 0 в этой группе встречается 4 раза, значение 1 – 5 раз, значение 2 – 8 раз, значение 3 – 6 раз, значение 4 – 5 раз, значение 5 – 2 раза. n=4+5+8+6+5+2=30

Вычисленные значения частот и частностей приведены в табл. 2.

Таблица 2.

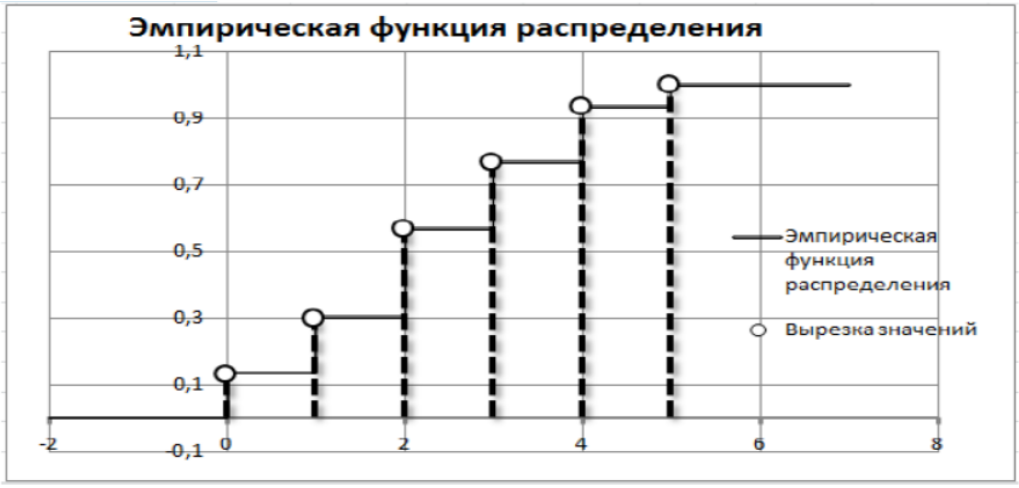
Индекс	i	1	2	3	4	5	6
Варианта	x_i	0	1	2	3	4	5
Частота	n_i	4	5	8	6	5	2
Частность	ω_i	4/30	5/30	8/30	6/30	5/30	2/30

Используя данный дискретный вариационный ряд (см. табл. 2), вычислим значения $F_n^*(x)$ по формуле, приведенной выше, и занесем их в табл. 3.

Таблица 3

x	$F_{30}^*(x)$
$x\le 0$	0
$0 < x \le 1$	$\omega_1 = \frac{4}{30}$
$1 < x \le 2$	$\omega_1 + \omega_2 = \frac{4}{30} + \frac{5}{30} = \frac{9}{30}$
$2 < x \le 3$	$\omega_1 + \omega_2 + \omega_3 = \frac{4}{30} + \frac{5}{30} + \frac{8}{30} = \frac{17}{30}$
$3 < x \le 4$	$\omega_1 + \omega_2 + \omega_3 + \omega_4 = \frac{4}{30} + \frac{5}{30} + \frac{8}{30} + \frac{6}{30} = \frac{23}{30}$
$4 < x \le 5$	$\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 = \frac{4}{30} + \frac{5}{30} + \frac{8}{30} + \frac{6}{30} + \frac{5}{30} = \frac{28}{30}$
$x > 5$	$\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 = \frac{28}{30} + \frac{2}{30} = \frac{30}{30} = 1$

По данным таблицы 3 построим график эмпирической функции распределения (рисуем в ручную и фото вставляем в файл excel).



Решение задачи в Excel.

Переименуйте Лист 1 в 1_Дискретный. Наберите массив 30 значений исходных данных выборки.

K4 | 2

	A	B	C	D	E	F	G	H	I	J	K
1		Выборка X									
2		3	0	1	5	1	2	4	5	3	4
3		2	4	2	0	2	3	1	3	2	1
4		4	3	0	2	1	0	4	2	3	2

3. Найдите величины x_{max}, x_{min}, n, используя встроенные функции Excel *МАКС*, *МИН* и *СЧЕТ*.

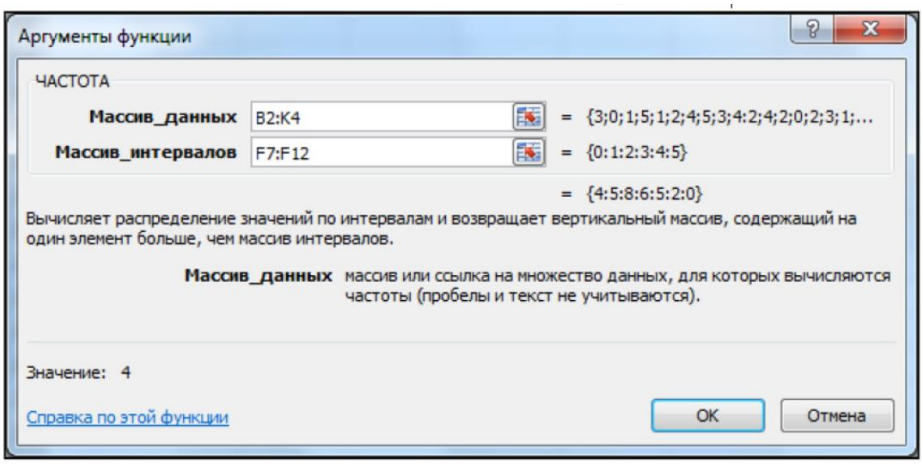
C8 | =СЧЁТ(B2:K4)

	A	B	C	D	E
1		Выборка X			
2		3	0	1	5
3		2	4	2	0
4		4	3	0	2
5					
6		x_{max} =	5		
7		x_{min} =	0		
8		n =	30		

4. Сформируйте столбец вариант x_i от 0 до 5 и с помощью функции *ЧАСТОТА* найдите частоту появления значений случайной величины X в данном интервале.

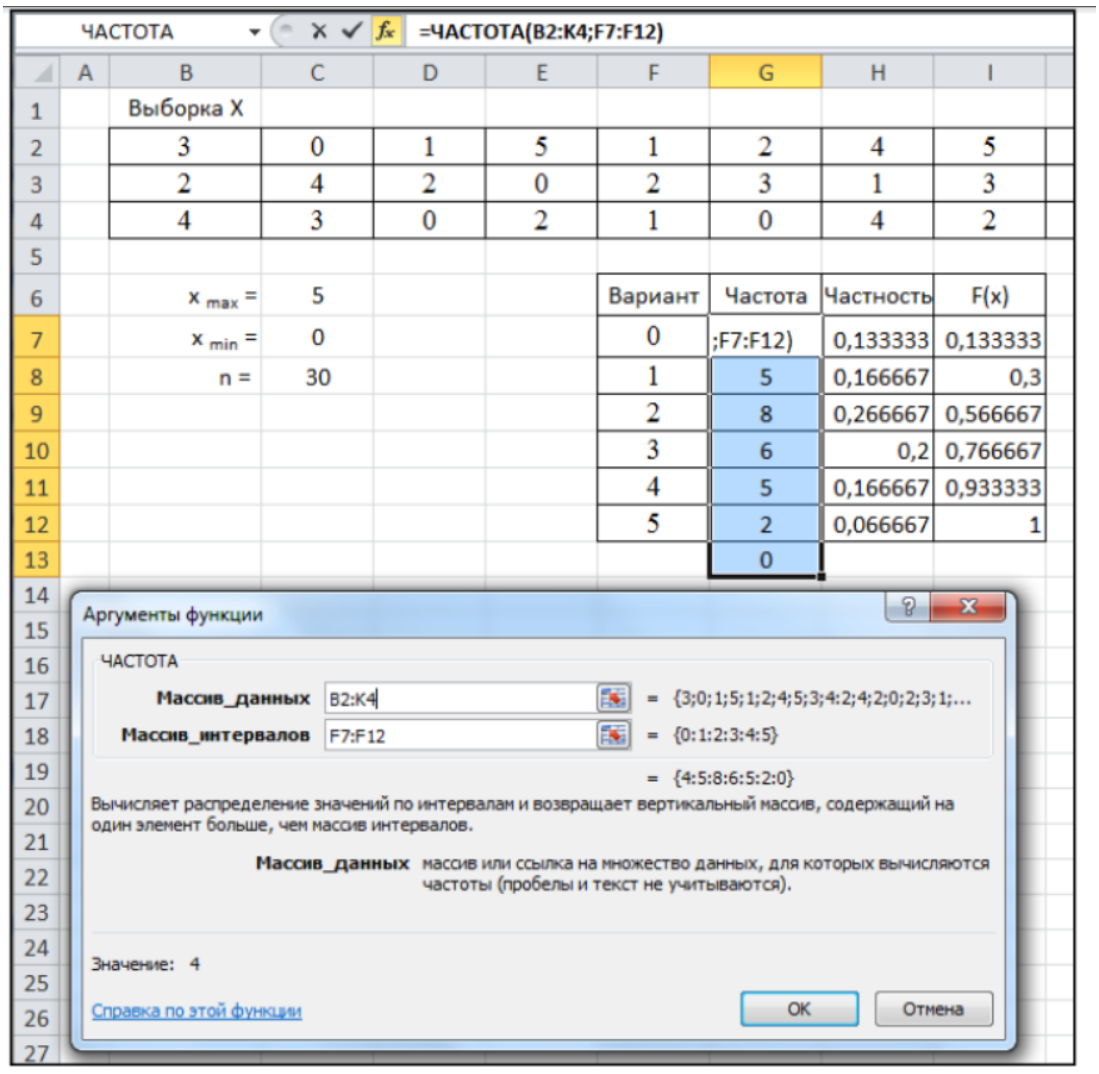
Синтаксис функции:

ЧАСТОТА(массивданных;массивинтервалов).



Массив данных – массив или ссылка на множество данных, для которых вычисляются частоты. В нашем случае это диапазон B2:K2.Если массив данных не содержит значений, то функция *ЧАСТОТА* возвращает массив нулей.

Массив интервалов – массив или ссылка на множество интервалов, в которые группируются значения аргумента массив данных. В нашем случае это диапазон F7:F12. Если массив интервалов не содержит значений, то функция *ЧАСТОТА* возвращает количество элементов в аргументе Массив данных.



Функция *ЧАСТОТА* вводится как формула массива после выделения интервала смежных ячеек, в которые нужно вернуть полученный массив частот.

Количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве интервалов. Дополнительный элемент в возвращаемом массиве содержит количество значений, больших, чем максимальное значение в интервалах, т.е. больше 5 в нашем случае.

Поскольку данная функция возвращает массив, она должна задаваться в качестве формулы массива и работа с ней завершается трехклавишной комбинацией CTRL+SHIFT+ENTER.

Функция *ЧАСТОТА* игнорирует пустые ячейки и тексты.

5. Сформируйте столбец частностей, вычислив значения ω_i, i = 1, ..., 6 по формуле

$$\omega_i = \frac{n_i}{n}$$

ЧАСТОТА | =G7/$C$8

	A	B	C	D	E	F	G	H	I
1		Выборка X							
2		3	0	1	5	1	2	4	5
3		2	4	2	0	2	3	1	3
4		4	3	0	2	1	0	4	2
5									
6		x_{max} =	5			Вариант	Частота	Частность	F(x)
7		x_{min} =	0			0	4	=G7/$C$8	0,133333
8		n =	30			1	5	0,166667	0,3
9						2	8	0,266667	0,566667
10						3	6	0,2	0,766667
11						4	5	0,166667	0,933333
12						5	2	0,066667	1
13							0		

6. Сформируйте столбец значений выборочной функции распределения $F_n^*(x)$. При этом первое значение в ячейке I7 просто копируется из ячейки H7. Следующее значение вычисляется как накопленная сумма предыдущего значения ω_1 из ячейки I7 и текущего значения ω_2 из ячейки H8:

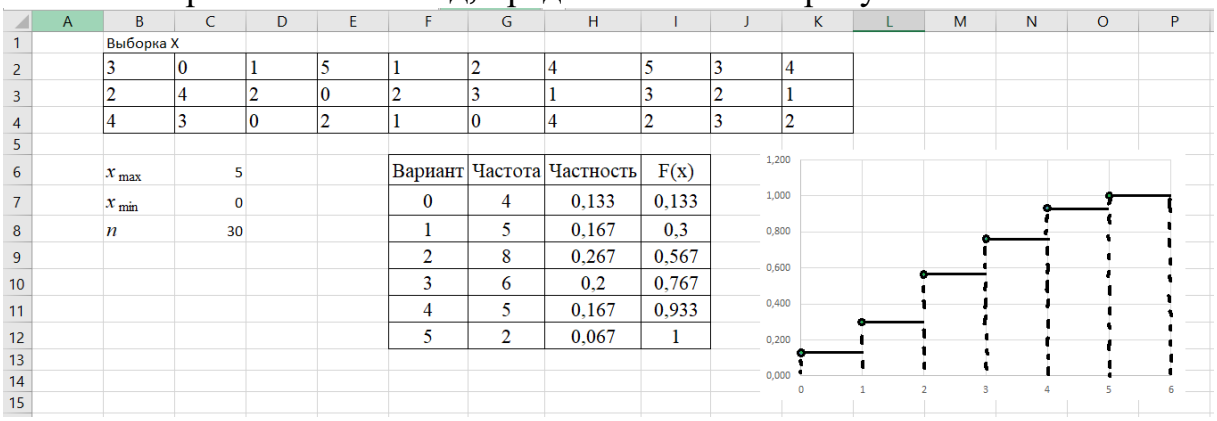
=I7+H8.

Вариант	Частота	Частность	F(x)
0	4	0,133333	=H7
1	5	0,166667	0,3
2	8	0,266667	0,566667
3	6	0,2	0,766667
4	5	0,166667	0,933333
5	2	0,066667	1
	0		

Вариант	Частота	Частность	F(x)
0	4	0,133333	0,133333
1	5	0,166667	=I7+H8
2	8	0,266667	0,566667
3	6	0,2	0,766667
4	5	0,166667	0,933333
5	2	0,066667	1
	0		

Затем данная формула копируется автозаполнением в остальные ячейки диапазона, с выходом на значение, равное 1.

Лист Excel работы имеет вид, представленный на рисунке:



Выборка X										
3	0	1	5	1	2	4	5	3	4	
2	4	2	0	2	3	1	3	2	1	
4	3	0	2	1	0	4	2	3	2	

x_{max}	5
x_{min}	0
n	30

Вариант	Частота	Частность	F(x)
0	4	0,133	0,133
1	5	0,167	0,3
2	8	0,267	0,567
3	6	0,2	0,767
4	5	0,167	0,933
5	2	0,067	1

Постановка задачи 2. Исследуется рост учащихся (в сантиметрах) в студенческой группе из 25 человек. Получена выборка (см. табл. 4) из следующих 25 значений.

Таблица 4.

184	182	182	180	177
179	173	179	192	173
190	163	177	186	170
178	185	173	179	165
179	173	179	166	170

Требуется: найти интервальный вариационный ряд, выборочную (эмпирическую) функцию распределения данной выборки и построить ее график.

Решение.

Найдем максимальное и минимальное значения в исследуемой выборке

$$x_{max} = 192,\ x_{min} = 163.$$

Вычислим размах варьирования R исследуемого признака по формуле

$$R = x_{max} - x_{min} = 192 - 163 = 29.$$

Для нахождения числа интервалов группировки N воспользуемся формулой

$$N \approx \sqrt{n} = \sqrt{25} = 5.$$

Далее следует группировка выборки. При этом интервал варьирования признака $[x_{min}, x_{max}]$ разбивается на N интервалов группировки одинаковой длины Δ, а затем подсчитывается число попаданий признака в i-й интервал группировки – n_i, i=1, …, N.

$$\Delta = \frac{R}{N} = \frac{x_{max} - x_{min}}{N} = \frac{29}{5} \approx 5{,}8 = 6.$$

При этом каждый интервал группировки $\Delta_i=(a_i;b_i)$ характеризуется своим правым и левым концом, числом n_i – попаданием признака в этот интервал. Иногда интервал характеризуют не границами, а его средним значением.

Дальнейшие вычисления удобно представить в табл. 5.

Таблица 5.

i	Интервал группировки Δ_i	Кол-во попаданий в интервал	Частоты n_i	Относительные частоты $\omega_i = \frac{n_i}{n}$	Накопленные частоты
1	162,5-168,5	| | |	3	3/25	3/25
2	168,5-174,5	| | | | | |	6	6/25	9/25
3	174,5-180,5	| | | | | | | | |	9	9/25	18/25
4	180,5-186,5	| | | | |	5	5/25	23/25
5	186,5-192,5	| |	2	2/25	252/25 = 1
∑			25	1	

Чтобы значение исследуемого признака не попадало на границы интервала группировки, примем минимальное значение признака не 163, а 162,5 и от этого значения начнем строить интервалы длиной $\Delta = 6$ (см. второй столбец табл. 5).

Откладывая по оси абсцисс средние значения интервалов группировки, а по оси ординат – значения накопленных частот, строим график эмпирической функции распределения.

Решение задачи в Excel.

1. Переименуйте Лист 2 в 1_Непрерывный. Наберите массив 25 значений исходных данных выборки.

2. Найдите величины x_{max}, x_{min}, n, N, $\Delta_{округл}$ используя встроенные функции Excel*МАКС*, *МИН*, *СЧЕТ*, *КОРЕНЬ* и *ОКРУГЛ*.

3. Сформируйте столбец интервалов варьирования от значения 162,5 с шагом $\Delta = 6$. Первое значение набираем с клавиатуры, а второе вычисляем с помощью формулы

=E9+$C$13.

Остальные значения получим копированием с помощью Автозаполнения.

4. Сформируйте столбец Частота и с помощью функции *ЧАСТОТА* найдите частоту появления значений исследуемой случайной величины X в каждом из интервалов.

5. Заполните столбец относительных частот, рассчитав значение в ячейке G9 по формуле

=F9/$C$10.

Остальные значения получим копированием формулы с помощью Автозаполнения.

6. Вычислите середины интервалов группировки, рассчитав значение в ячейке H9 по формуле

=(E9+E10)/2.

Остальные значения в диапазоне H10:H13 получим копированием формулы с помощью Автозаполнения.

7. Заполните столбец накопленных частот. При этом, значение в ячейке I9 получим, копируя значение ячейки G10по формуле

=G10.

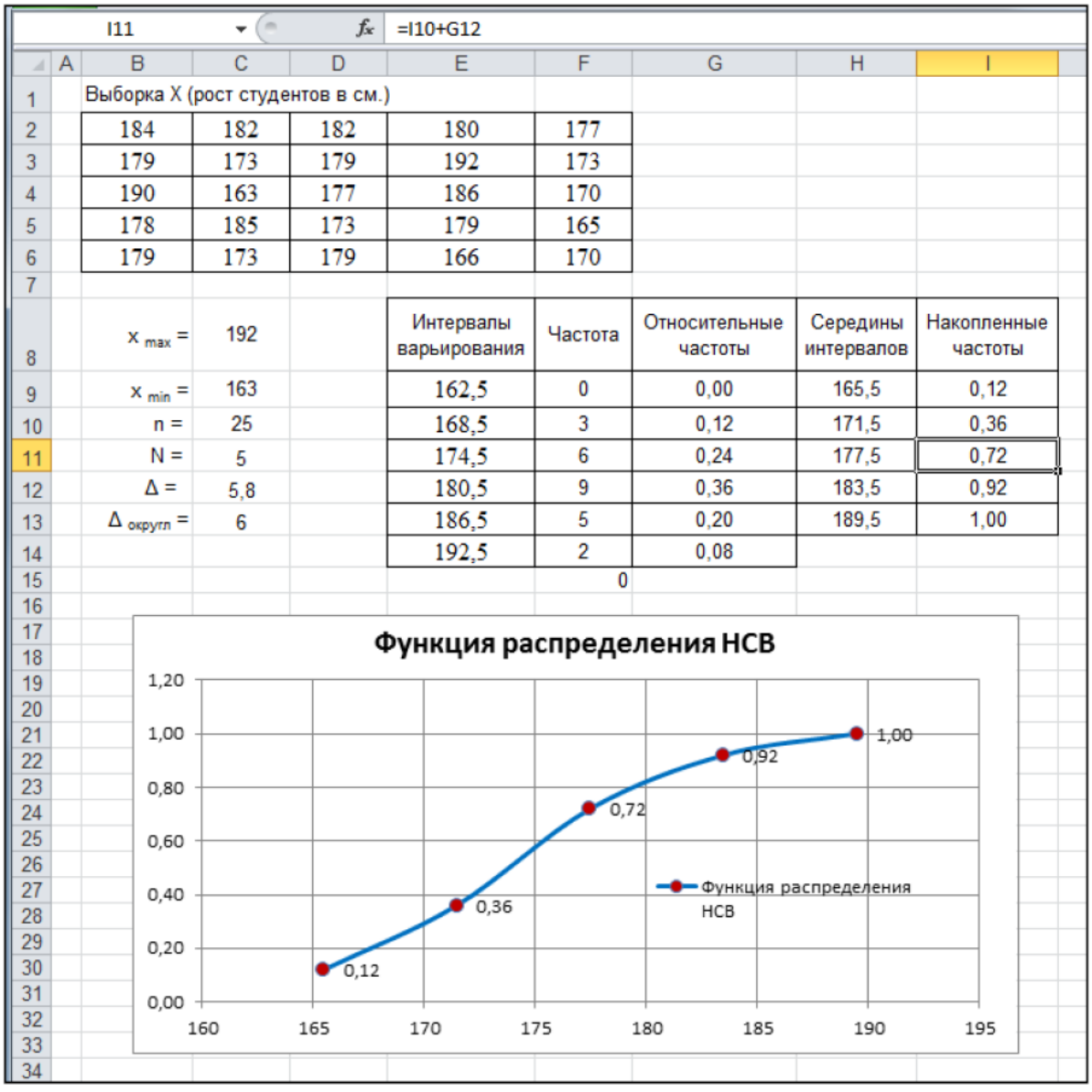
Значение в ячейке I10 получим по формуле

=I9+G11.

Остальные значения в диапазоне I11:I13 получим, копируя формулу с помощью Автозаполнения.

8. По данным двух последних столбцов построим график эмпирической функции распределения.

Лист Excel работы имеет вид, представленный на рисунке.



Задание 2. Графическое представление выборочных данных

Краткая теория

Существует три основных метода графического представления выборочных данных – гистограмма (столбчатая диаграмма), полигон частот и сглаженная кривая (огива).

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною Δ, а высоты равны отношению n_i / Δ (плотность частоты).

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс, на расстоянии n_i / Δ.

Площадь i-го частичного прямоугольника равна $\Delta \cdot (n_i / \Delta) = n_i$ – сумме частот вариант i-го интервала; следовательно, площадь гистограммы частот равна сумме всех частот, то есть объему выборки n.

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длинною Δ, а высоты равны отношению ω_i / Δ (плотность относительной частоты).

Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии ω_i / Δ. Площадь i-го частичного прямоугольника равна $\Delta \cdot (\omega_i / \Delta) = \omega_i$ – относительной частоте вариант, попавших в i-й интервал. Следовательно, площадь гистограммы относительных частот равна сумме всех относительных частот, то есть единице.

Полигоном частот называют ломаную, отрезки которой соединяют точки (x_1, n_1), (x_2, n_2), …, (x_N, n_N).

Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки (x_i, n_i) соединяют отрезками прямых и получают полигон частот.

Полигоном относительных частот называют ломаную, отрезки которой соединяют точки (x_1, ω_1), (x_2, ω_2), …, (x_N, ω_N).

Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат ω_i. Точки (x_i, ω_i) соединяют отрезками прямых и получают полигон относительных частот.

Замечание. В случае интервального вариационного ряда под x_i понимается середина i-го частичного интервала.

Постановка задачи 1. На телефонной станции проводились наблюдения над числом неправильных соединений в минуту. Наблюдения в течение 30 минут дали следующие результаты (табл. 1).

3	0	1	5	1	2	4	5	3	4
2	4	2	0	2	3	1	3	2	1
4	3	0	2	1	0	4	2	3	2

Требуется построить гистограмму, полигон в Excel.

Решение задачи в Excel.

Переименуйте Лист 3 в 2_Дискретный. Наберите массив 30 значений исходных данных выборки.

Найдите величины x_{max}, x_{min}, n, Δ, используя встроенные функции Excel *МАКС*, *МИН* и *СЧЕТ*.

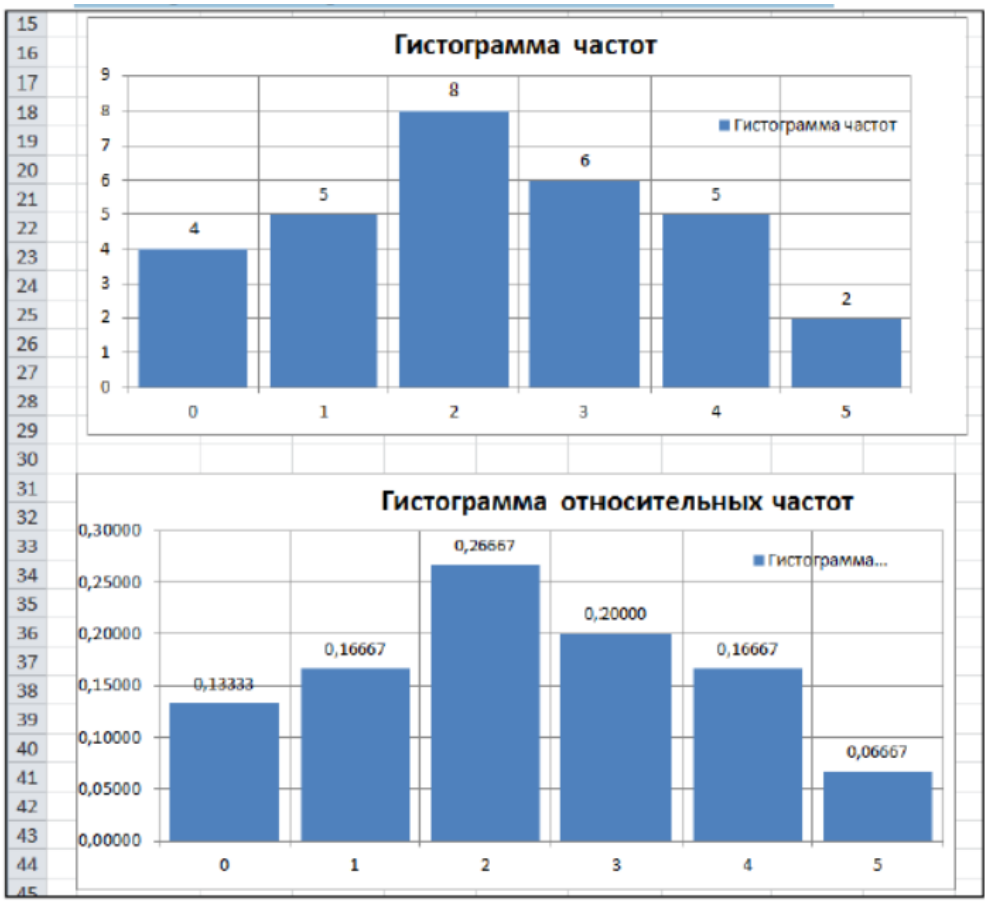
3. Сформируйте столбец вариант от 0 до 5 и с помощью функции *ЧАСТОТА* найдите частоту появления значений случайной величины X в данном интервале.

4. Вычислите столбцы значений n_i / Δ (плотность частоты) и ω_i / Δ (плотность относительной частоты).

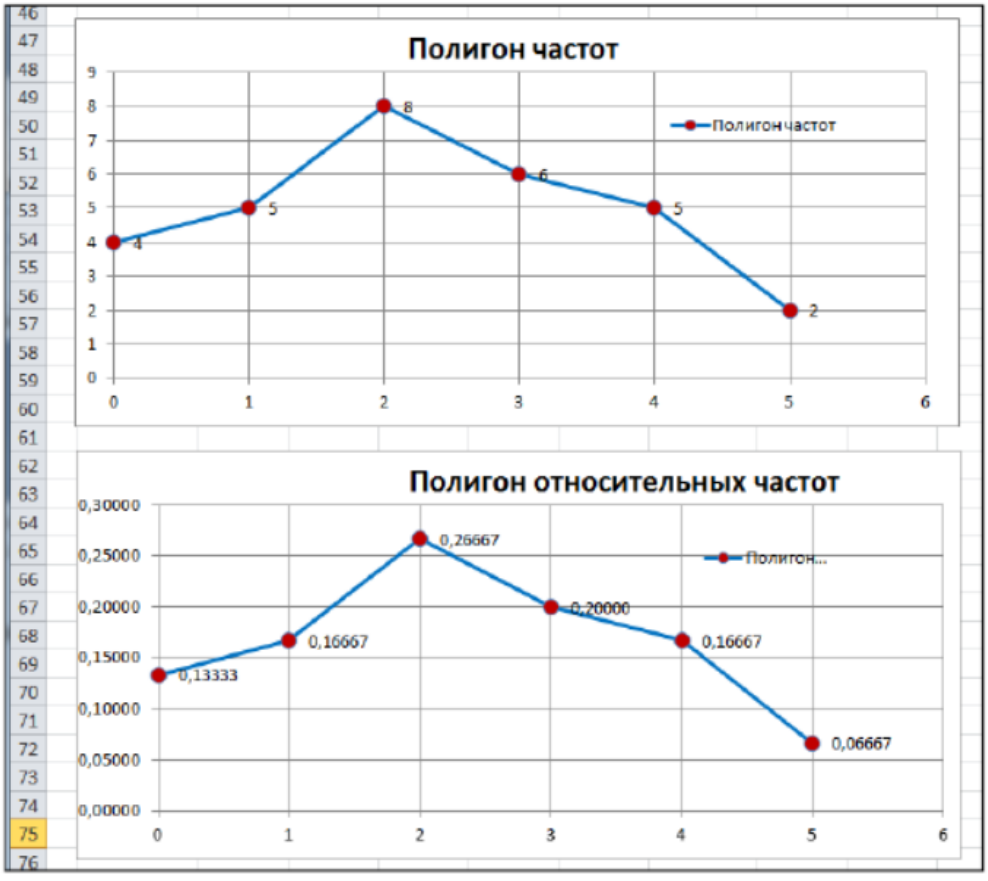
I7 =G7/($C$9*$C$8)

	A	B	C	D	E	F	G	H	I	J	K
1		Выборка X									
2		3	0	1	5	1	2	4	5	3	4
3		2	4	2	0	2	3	1	3	2	1
4		4	3	0	2	1	0	4	2	3	2
5											
6		x_{max} =	5			Вариант	Частота	Плотность n_i / Δ	$n_i /(\Delta \cdot n)$		
7		x_{min} =	0			0	4	4	0,13333		
8		n =	30			1	5	5	0,16667		
9		Δ =	1			2	8	8	0,26667		
10						3	6	6	0,20000		
11						4	5	5	0,16667		
12						5	2	2	0,06667		
13							0	**30**	**1**		

5. Построим гистограммы частот и относительных частот.



6. Вычислите столбец значений ω_i –относительных частот выборки и по данным столбцов 1, 2 и 5 постройте графики полигона частот и полигона относительных частот.



Постановка задачи 2. Исследуется рост учащихся (в сантиметрах) в студенческой группе из 25 человек. Получена выборка (см. табл. 4) из следующих 25 значений.

Таблица 4.

184	182	182	180	177
179	173	179	192	173
190	163	177	186	170
178	185	173	179	165
179	173	179	166	170

Требуется: построить гистограмму, полигон в Excel

Решение.

1. Переименуйте Лист 4 в 2_Непрерывный. Наберите массив 25 значений исходных данных выборки.
2. Оформите лист как показано на рисунке

H9 | =(E9+E10)/2

	A	B	C	D	E	F	G	H
1		Выборка X (рост студентов в см.)						
2		184	182	182	180	177		
3		179	173	179	192	173		
4		190	163	177	186	170		
5		178	185	173	179	165		
6		179	173	179	166	170		
7								
8		x_{max} =	192		Интервалы варьирования	Частота	Относительные частоты	Середины интервалов
9		x_{min} =	163		162,5	0	0,00	165,5
10		n =	25		168,5	3	0,12	171,5
11		N =	5		174,5	6	0,24	177,5
12		Δ =	5,8		180,5	9	0,36	183,5
13		$Δ_{округл}$ =	6		186,5	5	0,20	189,5
14					192,5	2	0,08	
15						0		

3. Сформируйте столбец частот *n*i и скопируйте в него не нулевые данные столбца частот, полученные с помощью встроенной функции *ЧАСТОТА*. Используйте контекстное меню команды Вставка: Параметры вставки → Значения 

.

4. Вычислите плотности частот в ячейке J9 по формуле

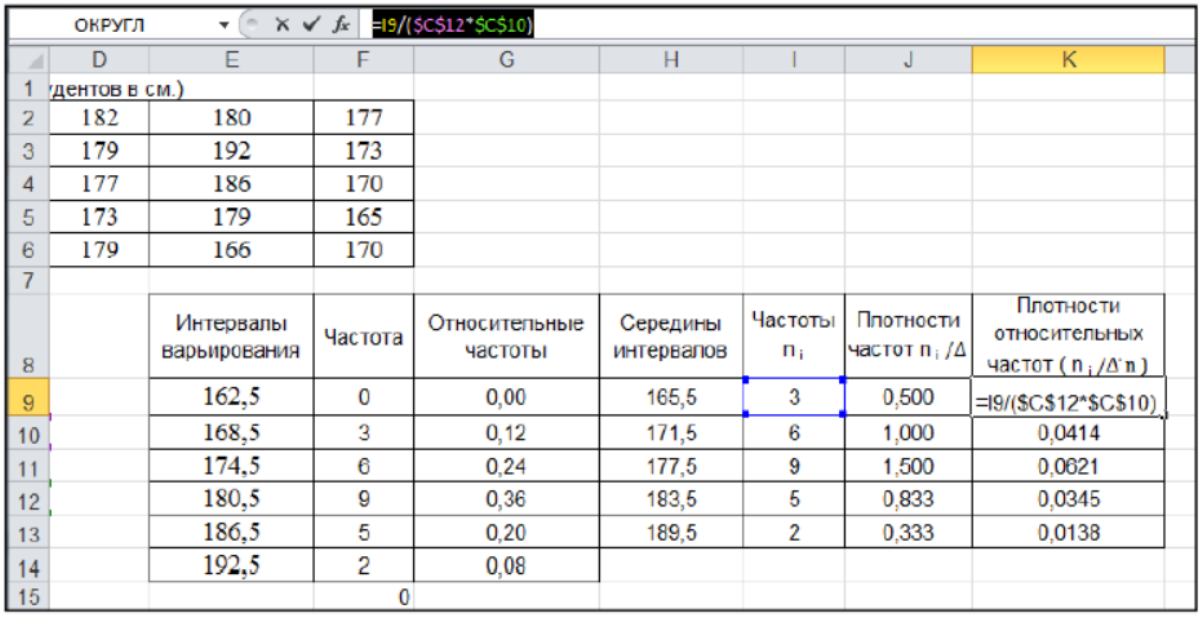
=I9/$C$13.

Остальные значения получим копированием с помощью Автозаполнения.

5. Вычислите плотности относительных частот в ячейке K9 по формуле

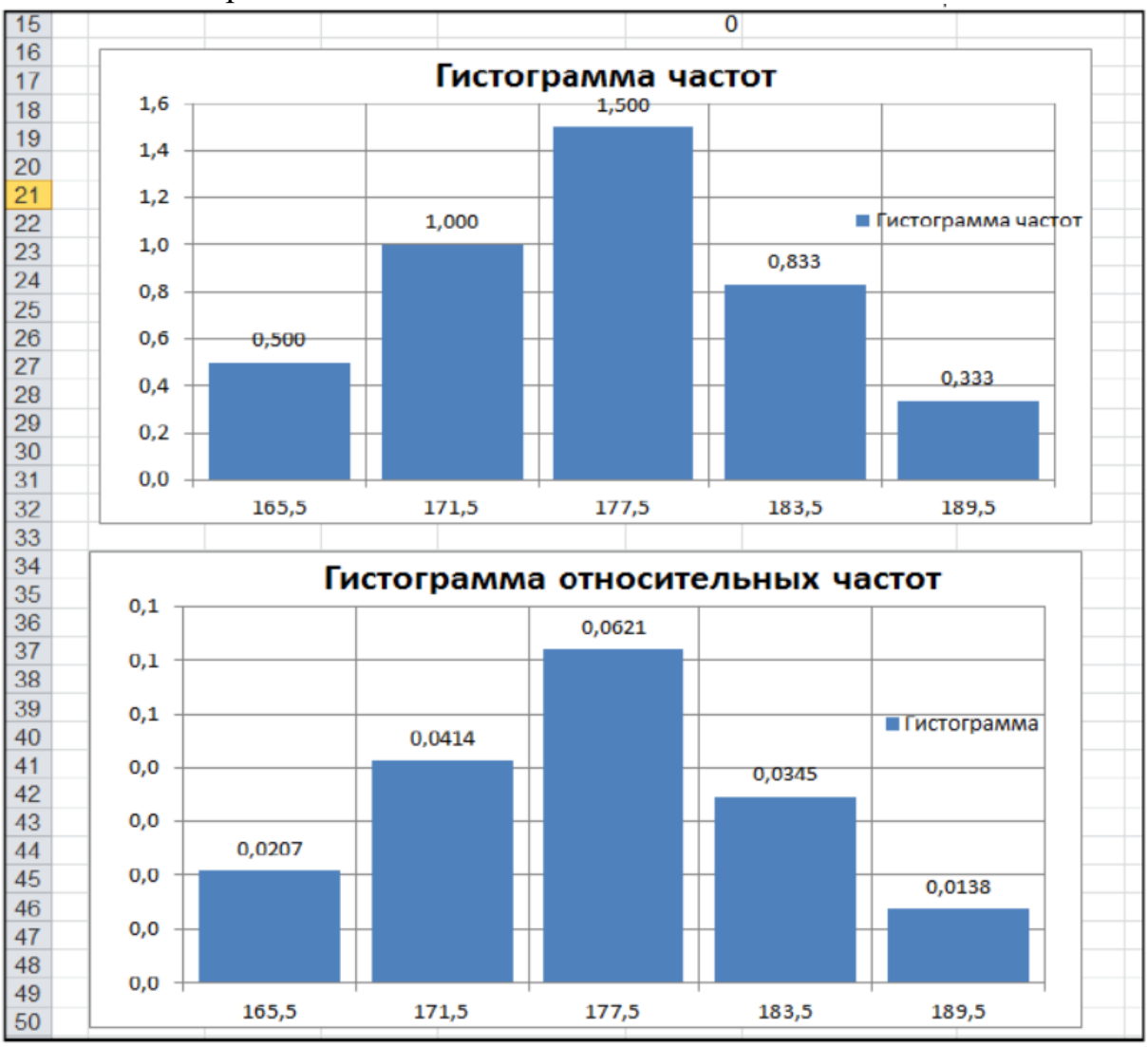
=I9/($C$12*$C$10).

Остальные значения получим копированием с помощью Автозаполнения. Полученная таблица имеет вид:

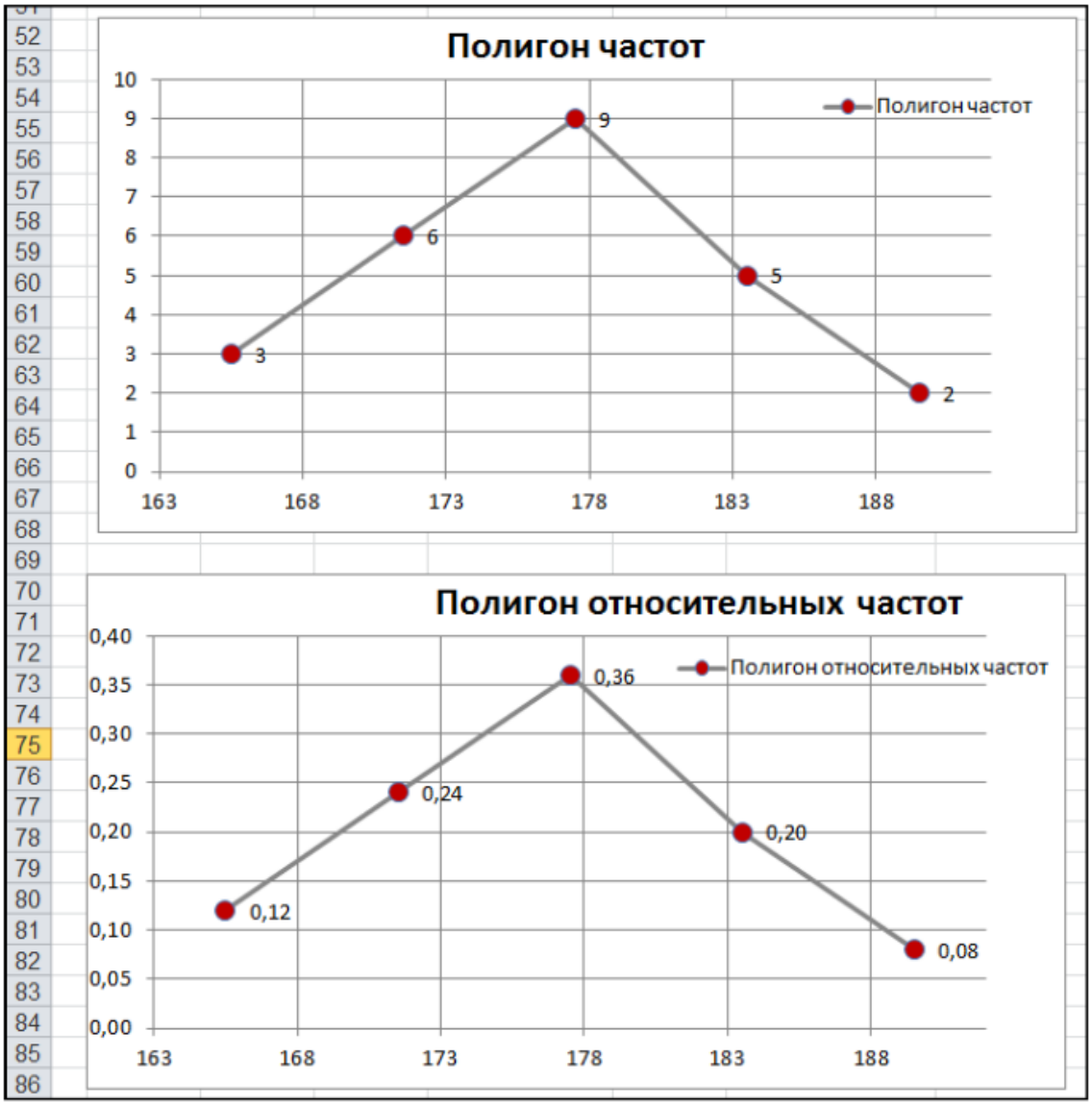


	D	E	F	G	H	I	J	K
1	дентов в см.)							
2	182	180	177					
3	179	192	173					
4	177	186	170					
5	173	179	165					
6	179	166	170					
7								
8		Интервалы варьирования	Частота	Относительные частоты	Середины интервалов	Частоты n_i	Плотности частот n_i/Δ	Плотности относительных частот $(n_i/\Delta \cdot n)$
9		162,5	0	0,00	165,5	3	0,500	=I9/($C$12*$C$10)
10		168,5	3	0,12	171,5	6	1,000	0,0414
11		174,5	6	0,24	177,5	9	1,500	0,0621
12		180,5	9	0,36	183,5	5	0,833	0,0345
13		186,5	5	0,20	189,5	2	0,333	0,0138
14		192,5	2	0,08				
15			0					

6. По данным двух последних столбцов построим графики гистограммы частот и гистограммы относительных частот.



7. Постройте графики полигона частот и полигона относительных частот.



Задание 3. Расчет числовых характеристик выборки

Краткая теория

Первый шаг к осмыслению скрытых в выборке закономерностей − это ее графическое представление, то есть построение гистограммы, полигона частот и эмпирической функции распределения. Однако выборки, имеющие похожие графические изображения, могут различаться своими числовыми характеристиками. Числовые характеристики вариационных рядов вычисляют по данным, полученным в результате наблюдений (статистическим данным), поэтому их называют также статистическими характеристиками или оценками.

Выборочные характеристики являются оценками соответствующих характеристик генеральной совокупности. Эти оценки должны удовлетворять определенным требованиям. В соответствии с важнейшими требованиями, оценки должны быть:

- *несмещенными*, то есть стремиться к истинному значению характеристики генеральной совокупности при неограниченном увеличении количества испытаний;
- *состоятельными*, то есть с ростом размера выборки оценка должна стремиться к значению соответствующего параметра генеральной совокупности с вероятностью, приближающейся к 1;
- *эффективными*, то есть для выборок равного объема используемая оценка должна иметь минимальную дисперсию.

Выборка может характеризоваться следующими числовыми характеристиками.

1. *Характеристики положения*.

Самой известной и наиболее употребляемой характеристикой любого вариационного ряда является его средняя арифметическая, называемая также *выборочным средним*. Средняя арифметическая характеризует значения признака, вокруг которого концентрируются наблюдения, т.е. центральную тенденцию распределения. При статистическом анализе выборки, кроме средней арифметической, широко применяют структурные, или порядковые, средние, к которым относятся *медиана* и *мода*.

Выборочное среднее рассчитывается по формуле

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

Если же анализируемые данные представлены в виде вариационного ряда, то для вычисления выборочного среднего применяется одно из следующих соотношений:

- для дискретного вариационного ряда

$$\overline{x} = \frac{\sum_{i=1}^{N} x^{(i)} \cdot n_i}{n} = \sum_{i=1}^{N} x^{(i)} \cdot \omega_i ;$$

- для интервального вариационного ряда

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

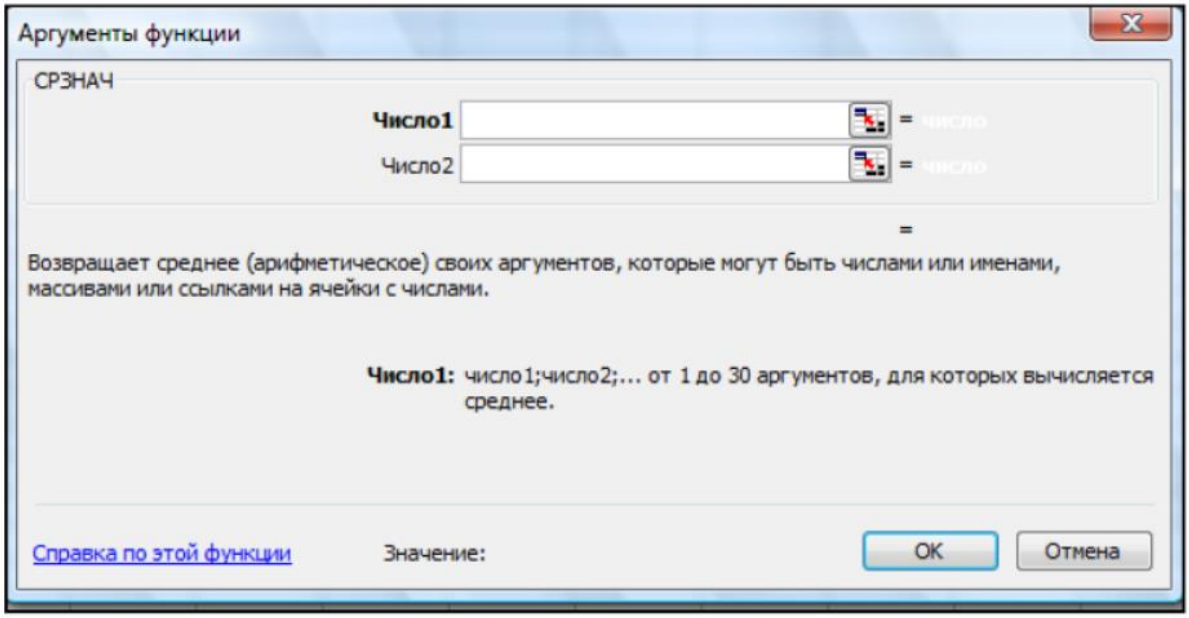
где i

ω_i – частность (относительная частота), соответствующая i-й варианте ил и i-му частичному интервалу; x_i^* – середина i-го частичного интервала группировки.

В Excel среднее значение находится с помощью функцией *СРЗНАЧ* .

Синтаксис функции:

СРЗНАЧ(число1; число2; ...).



Число1, число2, .. − это от 1 до 30 аргументов, для которых вычисляется среднее.

Аргументы должны быть либо числами, либо именами, массивами или ссылками, содержащими числа.

Достоинство *медианы* как меры центральной тенденции заключается в том, что на нее не влияет изменение крайних членов вариационного ряда, если любой из них, меньший медианы, остается меньше ее, а любой, больший медианы, продолжает быть большее ее.

Медиана предпочтительнее средней арифметической для ряда, у которого крайние варианты по сравнению с остальными оказались чрезмерно большими или малыми. Особенность *моды* как меры центральной тенденции заключается в том, что она также не изменяется при изменении крайних членов ряда, т.е. обладает определенной устойчивостью к вариации признака.

Выборочная медиана разбивает выборку пополам: слева и справа от нее оказывается одинаковое число элементов выборки. Если число элементов выборки четно, $n = 2k$, то выборочную медиану определяют по формуле

$$Me = (x_k + x_{k+1})/2,$$

где x_k и x_{k+1} –k-е и (k +1) -е выборочные значения из вариационного ряда.

При нечетном $n = 2k + 1$объеме выборки медиану находят по формуле

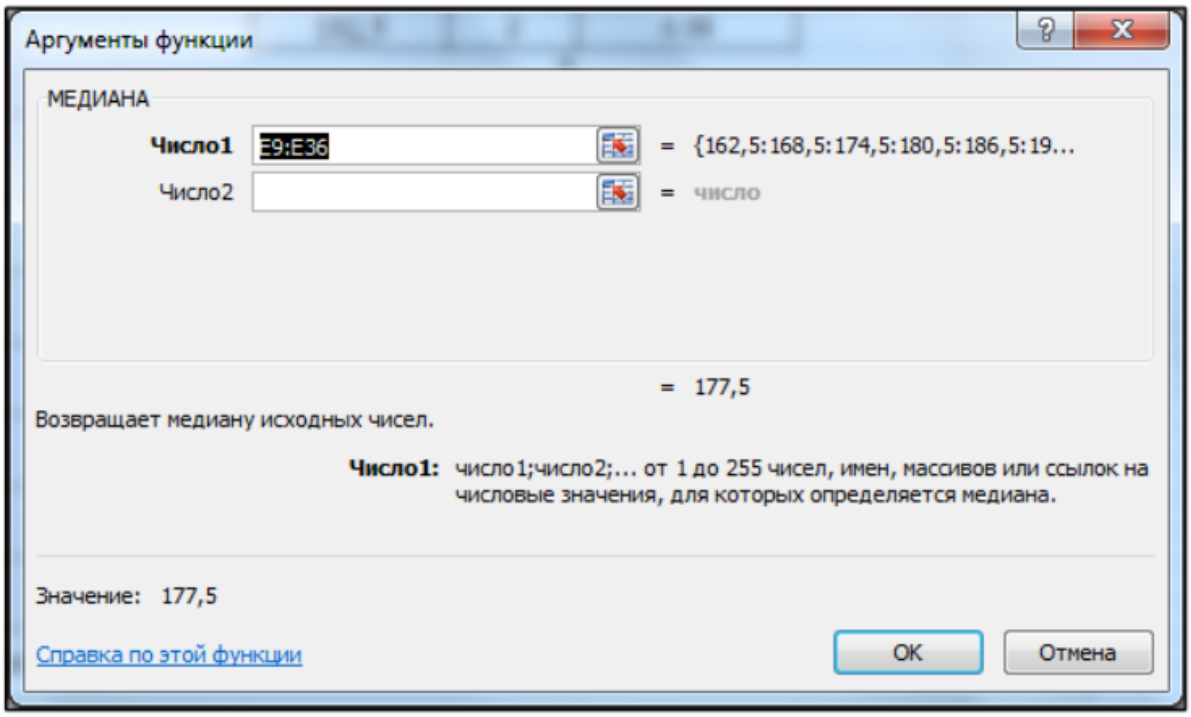
$$Me = x_{k+1},$$

т.е. за значение медианы принимают величинуx_{k+1}.

Так, например, если в диапазоне записаны значения 1, 2, 3, 4, 5, то функция МЕДИАНА вернет значение, равное 3, а если диапазон 1, 2, 3, 4, то найденное значение 2,5.

Синтаксис функции:

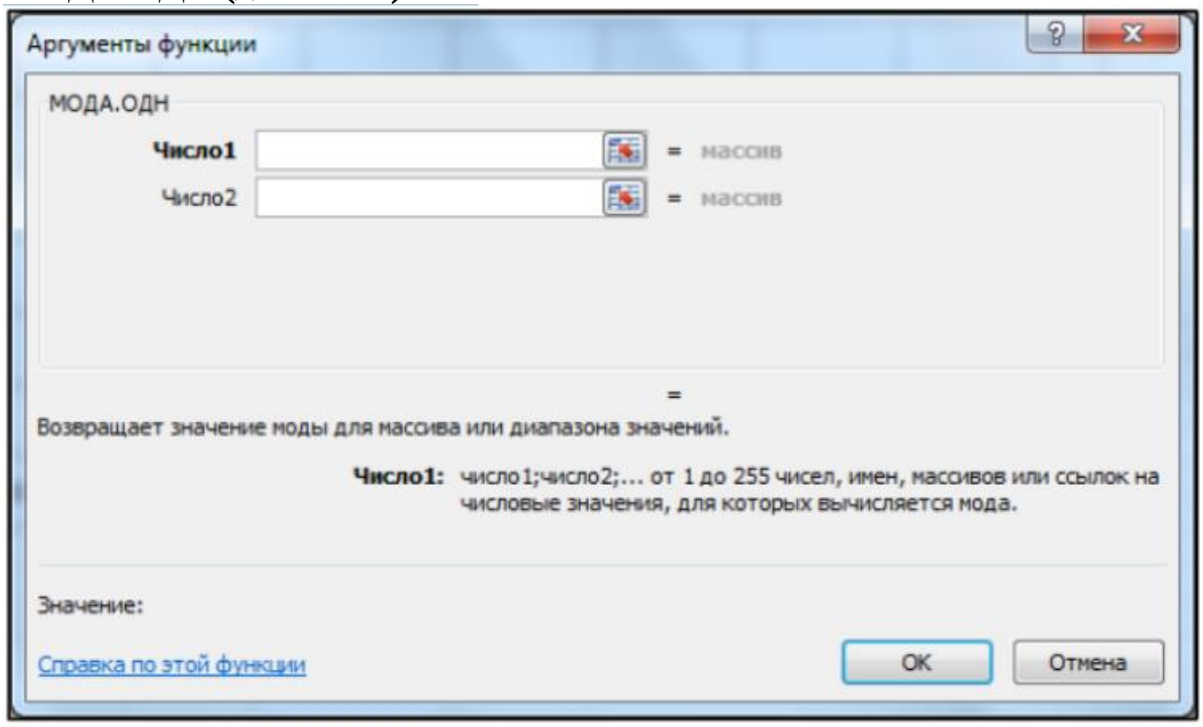
МЕДИАНА (диапазон).



Мода *Mo* используется для нахождения наиболее часто встречающегося в выборке значения.

Синтаксис функции:

МОДА.ОДН (диапазон).



При поиске игнорируются пустые ячейки, текстовые и логические значения. Если использовать функцию для нахождения *Mo* выборки 1, 2, 3, 4, 4, то функция даст 4. Если значения в выборке не повторяются, то функция выдаст сообщение об ошибке #Н/Д.

2. *Характеристики рассеяния.*

Для получения полного представления о вариационном ряде (определив центральную тенденцию распределения с помощью характеристик

положения) далее оценивают рассеяние (вариацию, изменчивость) исследуемого признака вокруг этих величин.

Простейшим и, весьма приближенным показателем вариации (изменчивости), является вариационный *размах*. Размах вариации наиболее полезен, если нужен быстрый и общий взгляд на изменчивость при сравнении большого количества выборок. *Размах* выборки вычисляется по формуле

$$R = x_{\max} - x_{\min}.$$

Но наибольший интерес представляют меры вариации (рассеяния) наблюдений вокруг средних величин, в частности, вокруг средней арифметической. К таким оценкам относятся *выборочная дисперсия* и *среднее квадратичное отклонение*.

Дисперсия выборки – это параметр, характеризующий степень разброса элементов выборки относительно среднего значения *x*. Чем больше дисперсия, тем дальше отклоняются значения элементов выборки от среднего значения.

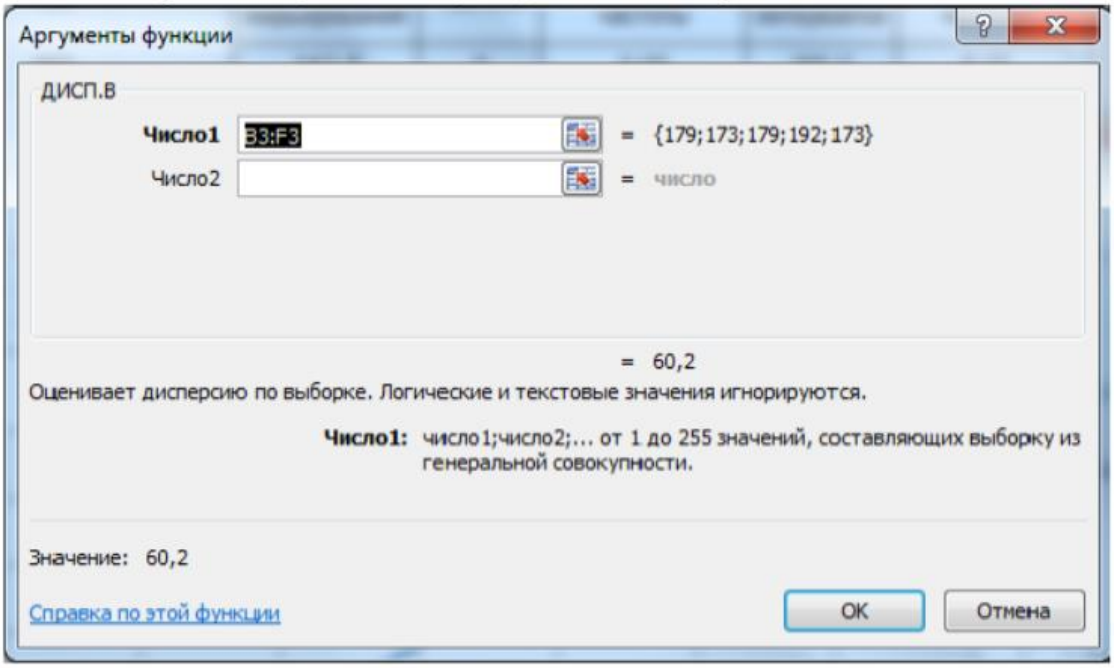
Выборочная дисперсия находится по формуле

$$D = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

Для вычисления выборочной дисперсии с помощью Excel используется функция *ДИСП.В*.

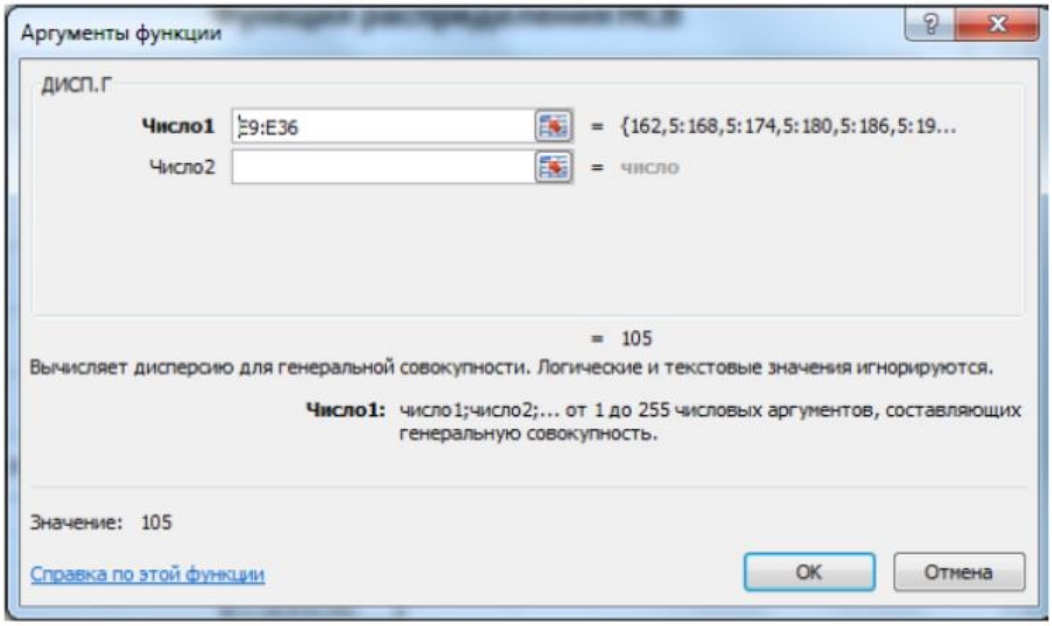
Синтаксис функции:

ДИСП.В(*число1; число2; …; число255*).



где *число1; число2; …; число255*– числа или адреса ячеек, содержащих числовые данные. Ячейки, содержащие текстовые, логические данные или пустые, при вычислении выборочной дисперсии игнорируются.

Для вычисления дисперсии генеральной совокупности в Excel используется функция *ДИСП.Г*.
Синтаксис функции:
ДИСП.Г(*число1; число2; …; число255*).



Если данные представлены в виде вариационного ряда, то целесообразно для вычисления D вместо приведенной выше формулы использовать соотношения:
- для дискретного вариационного ряда

$$D = \frac{\sum_{i=1}^{N} (x^{(i)} - \bar{x})^2 \cdot n_i}{n} = \sum_{i=1}^{N} (x^{(i)} - \bar{x})^2 \cdot \omega_i ;$$

- для интервального вариационного ряда

$$D = \frac{\sum_{i=1}^{N} (x_i^* - \bar{x})^2 \cdot n_i}{n} = \sum_{i=1}^{N} (x_i^* - \bar{x})^2 \cdot \omega_i$$

Выборочная дисперсия обладает одним существенным недостатком: если среднее арифметическое выражается в тех же единицах, что и значения случайной величины, то, согласно определению, дисперсия выражается уже в квадратных единицах. Этого недостатка можно избежать, если использовать в качестве меры вариации признака *среднее квадратичное отклонение*

$$S = \sqrt{D} .$$

При малых объемах выборки дисперсия является смещенной оценкой, поэтому при объемах $n > 30$ используют *исправленную дисперсию* и *исправленное среднее квадратичное отклонение*.

Среднее квадратичное отклонение *S*, полученное при выборке *n<30*, носит название смещенного и его среднее значение занижено по сравнению со средним квадратичным отклонением для генеральной совокупности.

При числе испытаний n < 30

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad .$$

При числе испытаний 30≤n<50

$$S_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \quad ,$$

$$S = M_k \cdot S_1 ,$$

где $M_к$ – коэффициент, зависящий от числа испытаний.

Значения $M_к$ приведены в табл. 1 для$K=n-1$

Таблица 1

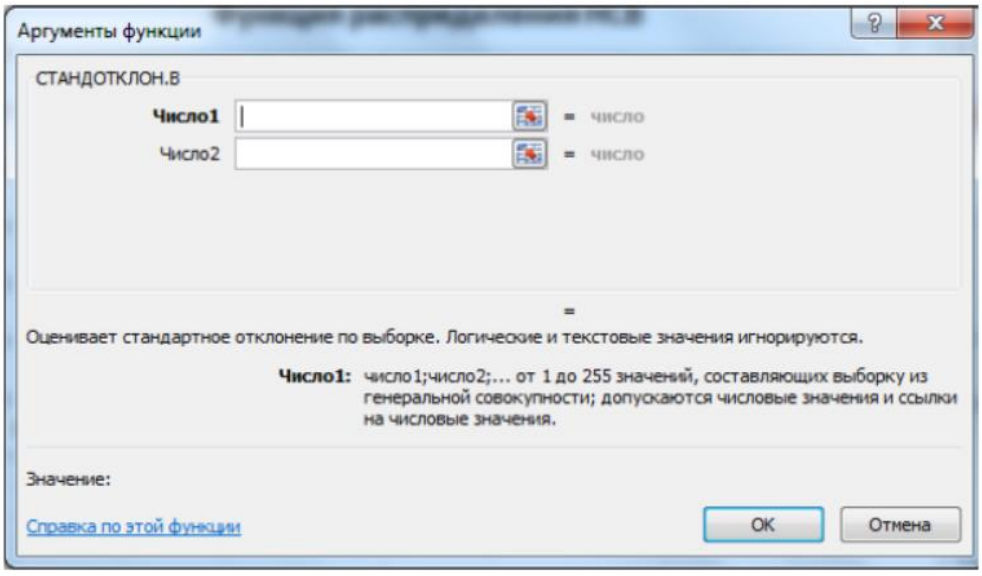
К	2	3	4	9	19	30	50	60
$M_к$	1,128	1,085	1,064	1,028	1,013	1,008	1,005	1,004

При $n > 60$ значение коэффициента$M_к \sim 1$.

Для вычислений среднего квадратичного отклонения выборки применяется функция *СТАНДОТКЛОН.В*.

Синтаксис функции:

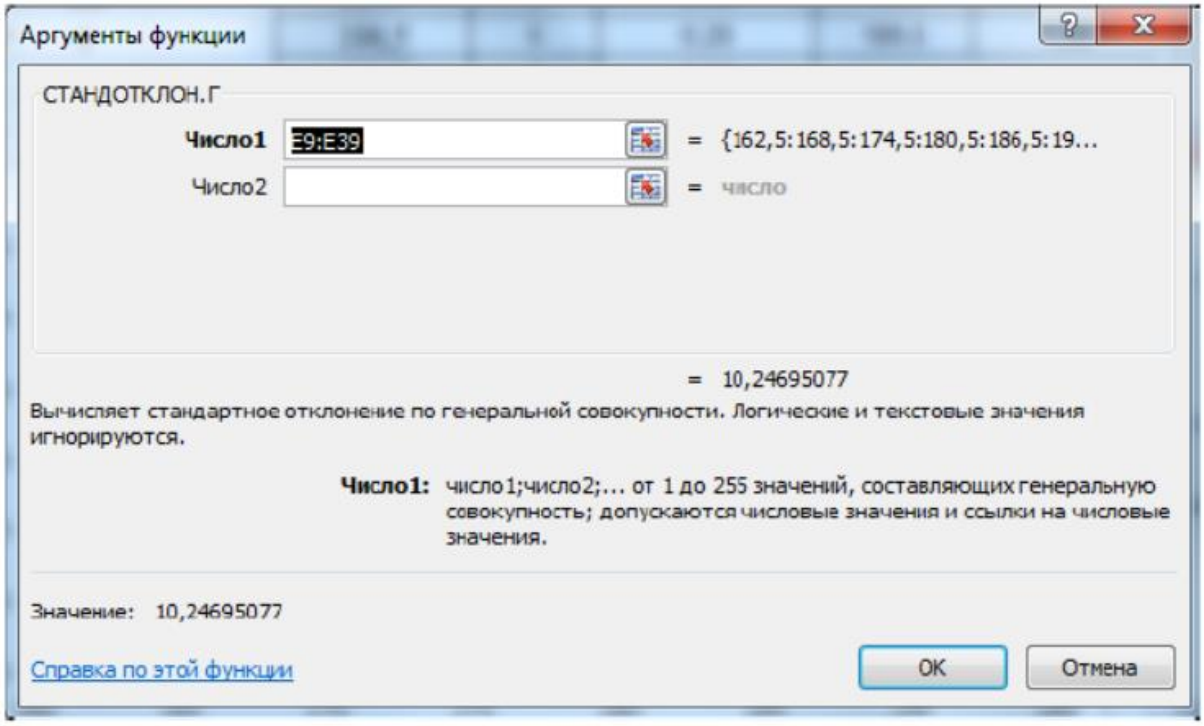
СТАНДОТКЛОН.В(число1; число2; ...).



Число1, число2,... −от 1 до 255 числовых аргументов, соответствующих выборке из генеральной совокупности. Вместо аргументов, разделенных точкой с запятой, можно также использовать массив или ссылку на массив.

Функция *СТАНДОТКЛОН.В* оценивает среднее квадратичное отклонение (стандартное отклонение) по выборке. Стандартное отклонение − это мера того, насколько широко разбросаны точки данных относительно их среднего. *СТАНДОТКЛОН.В* предполагает, что аргументы являются только выборкой из генеральной совокупности.

Если данные представляют всю генеральную совокупность, то стандартное отклонение следует вычислять с помощью функции *СТАНДОТКЛОН.Г*.

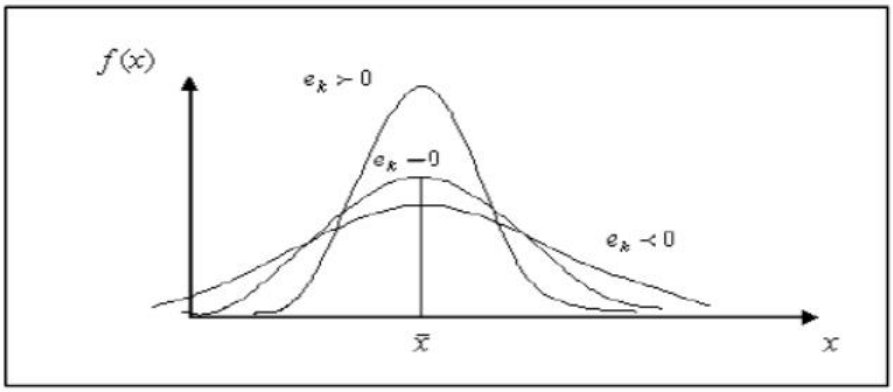


3. *Характеристики формы*.

К характеристикам формы относят коэффициент асимметрии и эксцесс. *Выборочный эксцесс* характеризует островершинность эмпирического распределения относительно стандартного нормального.

Эксцесс стандартного нормального распределения равен трем. Если эксцесс положителен ($e_k > 0$), то полигон вариационного ряда имеет более крутую вершину. Это говорит о скоплении значений признака в центральной зоне ряда распределения, т.е. о преимущественном появлении в данных значений, близких к средней величине.

Если эксцесс отрицателен ($e_k < 0$), то полигон имеет более пологую вершину по сравнению с нормальной кривой. Это означает, что значения признака не концентрируются в центральной части ряда, а достаточно равномерно рассеяны по всему диапазону от минимального до максимального значения. Чем больше абсолютная величина эксцесса, тем существеннее распределение отличается от нормального, смотри рисунок.



Выборочный эксцесс может быть найден по формуле

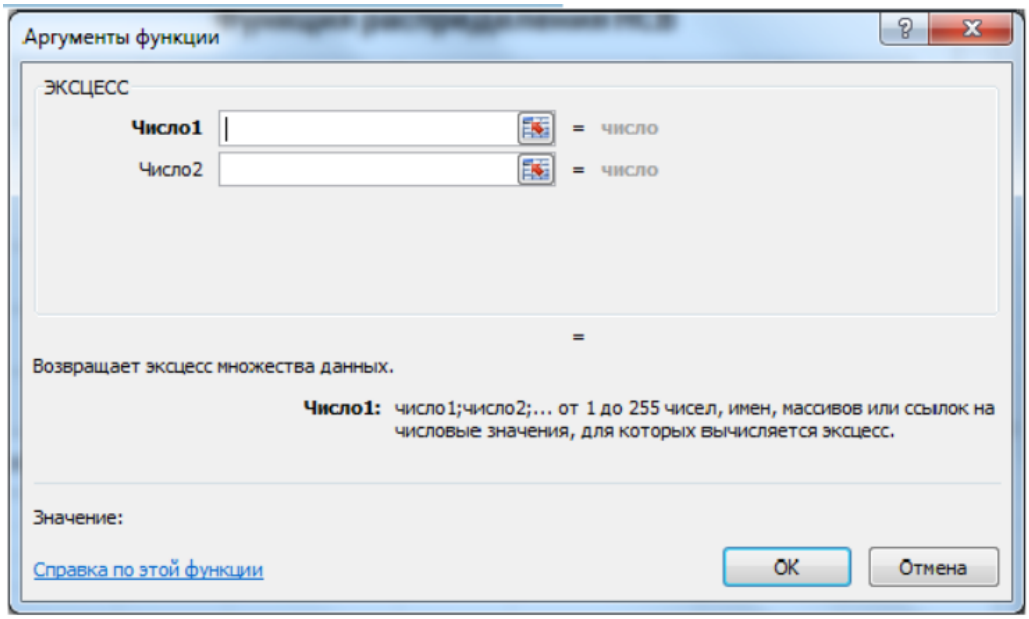
$$e_k = \frac{\mu_4}{S^4} - 3,$$

где $\mu_4 = \sum_{i=1}^{n} (x_i - \bar{x})^4$.

Для вычислений выборочного эксцесса выборки применяется функция Excel *ЭКСЦЕСС*.

Синтаксис функции:

ЭКСЦЕСС (число1; число2; ...).

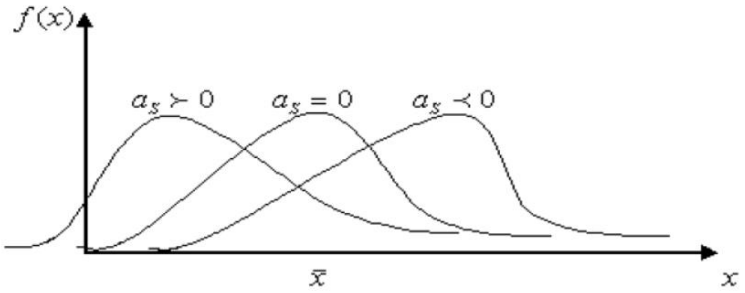


Коэффициент асимметрии характеризует симметрию распределения выборочных данных около центра выборки $\bar{x}$, для стандартного нормального распределения коэффициент асимметрии равен 0 (a_s=0).

Если распределение асимметрично, одна из ветвей построенного полигона частот имеет более пологий спуск, чем другая.

Если правая ветвь графика более пологая то это означает преимущественное появление в распределении более высоких значений

признака, при этом коэффициент асимметрии $a_s>0$.В противном случае $a_s<0$, при этом в распределении чаще встречаются более низкие значения (смотри рисунок).



Чем больше значение коэффициента асимметрии, тем более асимметрично распределение (до 0,25 асимметрия незначительная; от 0,25 до 0,5 умеренная; свыше 0,5 – существенная).

Коэффициент асимметрии вычисляется по формуле

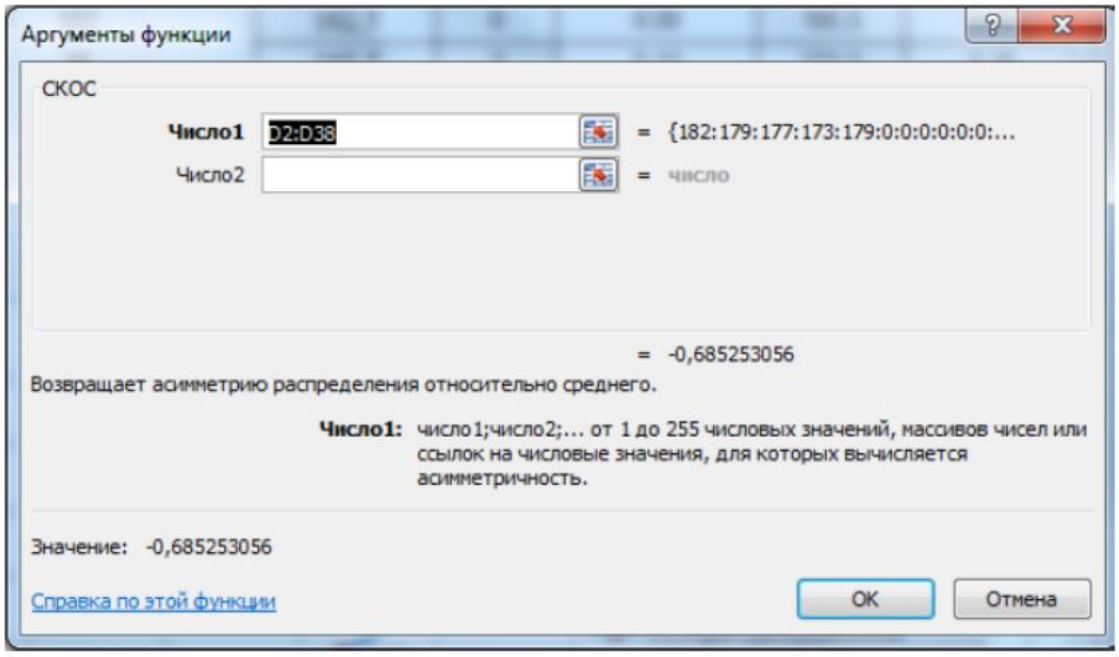
$$a_s = \frac{\mu_3}{S^3},$$

где $\mu_3 = \sum_{i=1}^{n}(x_i - \bar{x})^3$.

Для вычисления коэффициента асимметрии выборки применяется функция *СКОС*.

Синтаксис функции:

СКОС(число1; число2; ...).



Пример выполнения

Постановка задачи. Приведены размеры месячных зарплат (в тыс. руб.) 27 швей-мотористок, работающих по сдельно-премиальной системе оплаты труда (табл. 1).

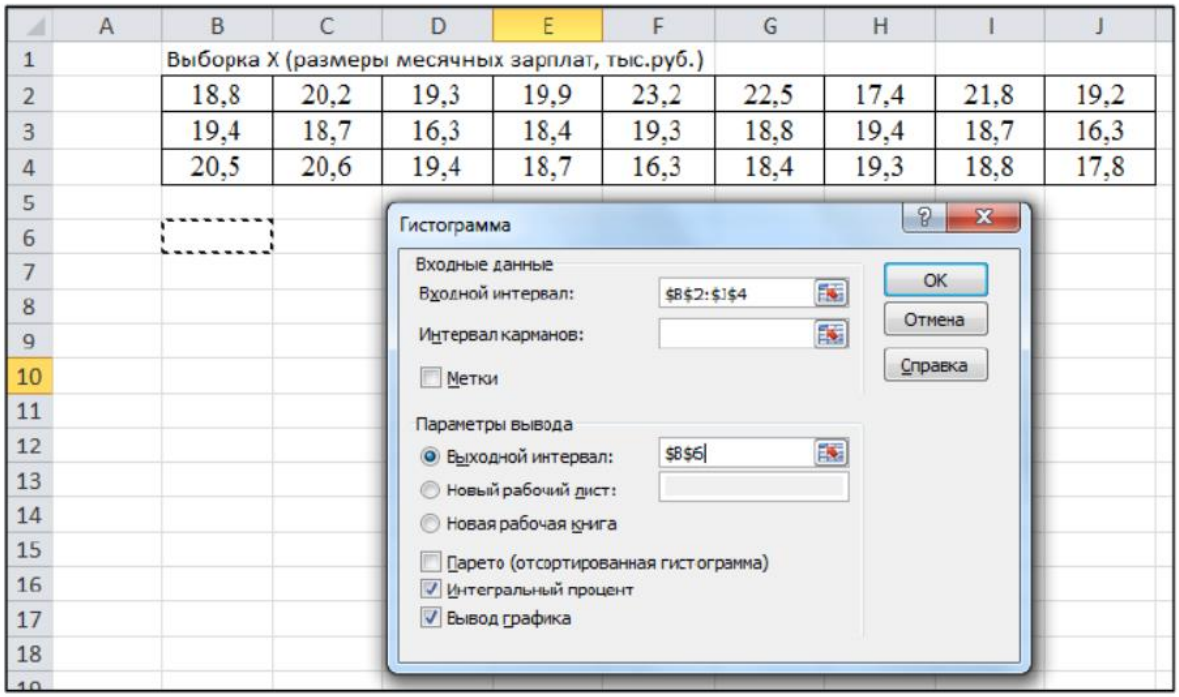
Таблица 1.

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

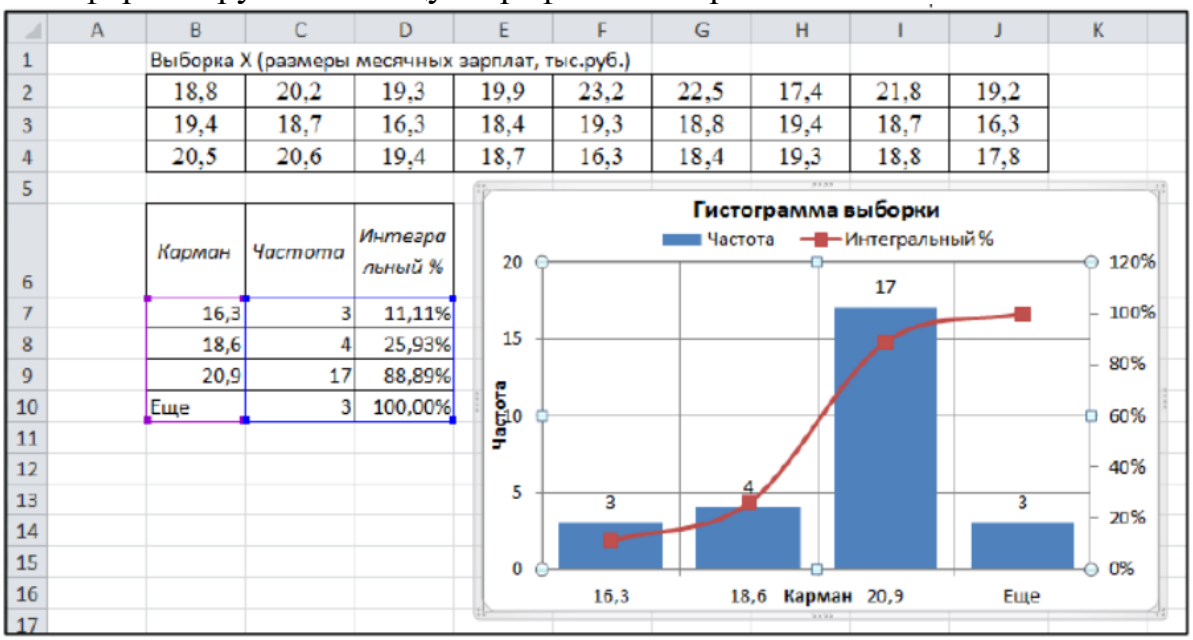
Требуется: найти числовые характеристики выборки с помощью встроенных функций Excel.

Решение задачи в Excel.

Переименуйте Лист 5 в Задание 3 и наберите таблицу исходных данных. Постройте гистограмму выборки с помощью Надстройки Пакет анализа, не указывая интервалы группировки выборки.



4. Отформатируйте таблицу и график гистограммы.



5. Найдите числовые характеристики выборки с помощью встроенных функций Excel, рассмотренных выше.

18	1. Характеристики положения:			
19		Выборочное среднее	19,163	
20		Медиана	19,200	
21		Мода	18,800	
22	2. Характеристики рассеяния:			
23		Наибольшее	23,2	
24		Наименьшее	16,3	
25		Размах выборки	6,9	
26		Дисперсия выборки		2,763
27		Дисперсия ген.совокупности		2,661
28		Среднее квадратич. отклонение		1,662
29		СКО генеральной совокупности		1,631
30	3. Характеристики формы:			
31		Выборочный эксцесс	0,8245	
32		Коэффиц. асимметрии	0,4384	
33				

Задание 4. Проверить гипотезу о согласии эмпирического распределения с теоретическим распределением с помощью критерия Пирсона

Краткая теория

При наличии числовых характеристик случайной величины (математического ожидания, дисперсии, коэффициента вариации) законы ее распределения могут быть определены в первом приближении по таблице 1.

Таблица 1

Законы распределения случайной величины в зависимости от коэффициента вариации

Пределы изменения коэффициента вариации V_x	Закон распределения случайной величины X
$V_x \le 0,3$	Нормальный
$0,3 \le V_x < 0,4$	Гамма-распределение
$0,4 \le V_x < 1$	Вейбулла
$V_x = 1$	Экспоненциальный, Пуассона

Чтобы подобрать подходящее теоретическое распределение, необходимо построить кривую плотности распределения, после этого выбрать похожую из известных типов распределений. Если есть основания отдать предпочтение тому или иному распределению, то кривую строить нет необходимости. Затем выдвигают гипотезу о соответствии экспериментального и теоретического распределений, проверяют её на

заданном уровне значимости, используя критерии согласия. Существуют несколько критериев.

Критерий Пирсона (хи-квадрат) применим только к сгруппированным данным. Рекомендуется, чтобы объем выборки был больше 100 и численность интервалов (групп), была не менее 5. Исходные данные разбивают на m интервалов и вычисляют для каждого:

- *экспериментальные частоты* $p_i^* = n_i / n$, n_i – количество данных попавших в i – й интервал, n – объём выборки;
- *теоретические частоты* $p_i = F(x_{i+1}) - F(x_i)$, найденные по таблицам и формулам для выбранного типа теоретического распределения;
- *экспериментальную величину*

$$(\chi^2)^* = n\sum_{i=1}^{m}\frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{m}\frac{(n_i - np_i)^2}{np_i}.$$

По таблицам квантилей распределения χ^2 при заданном уровне значимости β (обычно 5%) и известном числе степеней свободы f находят теоретическое значение χ^2. f равно количеству интервалов минус число независимых условий, наложенных на экспериментальные частоты p_i^*. Примерами таких условий могут быть: равенство единице суммы всех частот, совпадение статистического среднего с гипотетическим, совпадение дисперсий и т.п. Следовательно:

$$f=m-r-1,$$

где m – число интервалов, r – число параметров, определяемых из опытных данных.

Пример. Если предполагаемое распределение – нормальное, то оценивают два параметра – математическое ожидание и среднее квадратическое отклонение, тогда $f=m-r-1=m-2-1=m-3$.

Если $(\chi^2)^*<\chi^2$, то функция распределения при заданном уровне значимости (β=5%) согласуется с экспериментальными данными.

Пример. Пользуясь критерием Пирсона, подобрать теоретический закон распределения для часовой выработки автомобилей КамАЗ-5511, статистическое распределение которой приведено в таблице 2.

Таблица 2

Вариационный ряд часовой выработки автомобиля

Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05

В MS Excel:

	A	B	C	D	E	F	G	H	I
1	Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
2	Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05
3									
4	середина интервала	4,75	6,25	7,75	9,25	10,75	12,25	13,75	15,25
5	Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05

По форме гистограммы рис.1 можно предположить, что часовая выработка автомобиля подчиняется нормальному закону.

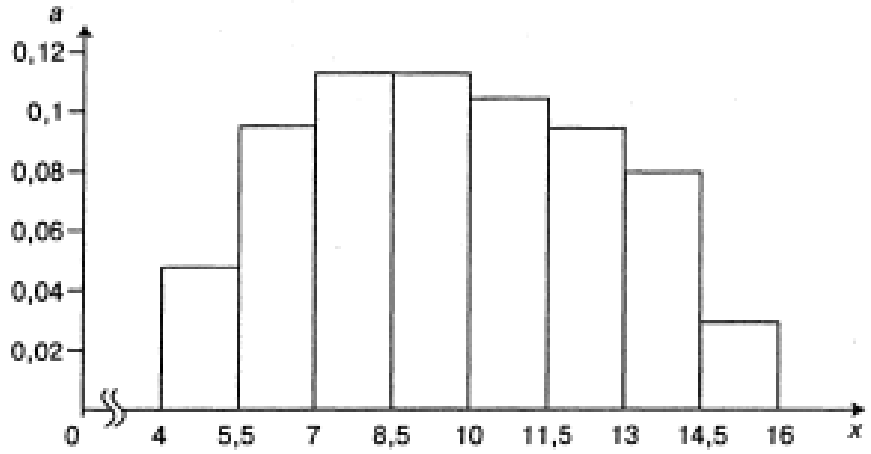


Рис. 1. Гистограмма часовой выработки автомобиля

Для оценки числовых характеристик нормального распределения вычислим:

математическое ожидание в ячейке L4 =СУММПРОИЗВ(B4:I4;B5:I5)

$$m_x = \sum \bar{x}_i p_i^* = 4,75 \cdot 0,07 + 6,25 \cdot 0,14 + 7,75 \cdot 0,17 + 9,25 \cdot 0,17 +$$
$$+10,75 \cdot 0,15 + 12,25 \cdot 0,14 + 13,75 \cdot 0,11 + 15,25 \cdot 0,05 = 9,7;$$

дисперсию в ячейке L8 =СУММПРОИЗВ(B7:I7;B5:I5)

	A	B	C	D	E	F	G	H	I
1	Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
2	Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05
3									
4	середина интервала	4,75	6,25	7,75	9,25	10,75	12,25	13,75	15,25
5	Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05
6									
7	$(m_x - \bar{x}_i)^2$	24,5025	11,9025	3,8025	0,2025	1,1025	6,5025	16,4025	30,8025
8	Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05

$$D_x = \sum_{i=1}^{k} (m_x - \bar{x}_i)^2 p_i^* =$$
$$= (9,7-4,75)^2 \cdot 0,07 + (9,7-6,25)^2 \cdot 0,14 + \ldots + (9,7-15,25)^2 \cdot 0,05 = 8,48$$

среднее квадратическое отклонение в ячейке L9

$$\sigma_x = \sqrt{D_x} = \sqrt{8,48} = 2,91;$$

коэффициент вариации

$$V_x = \frac{\sigma_x}{m_x} = \frac{2,91}{9,7} = 0,3.$$

Величина V_x=0,3 свидетельствует о том, что теоретическое распределение близко к нормальному закону распределения. Проверим данную гипотезу, воспользовавшись критерием согласия Пирсона.

Определим теоретическую вероятность попадания значений часовой выработки автомобиля в заданные интервалы, используя формулу:

$$p_i = F(x_{i+1}) - F(x_i) = \Phi\left(\frac{x_{i+1} - m_x}{\sigma_x}\right) - \Phi\left(\frac{x_i - m_x}{\sigma_x}\right),$$

где x_i, x_{i+1} – границы i-го интервала,

$\Phi(u)$ – функция Лапласа.

Составим расчетную таблицу.

	K	L	M	N	O	P	Q	R	S	T	U
1				x_i	x_{i+1}	$\Phi(u_i)$	$\Phi(u_{i+1})$	p_i	np_i	n_i	$(n_i - np_i)^2/np_i$
2				4	5,5	-0,5	-0,4254	0,075	7,5	7	0,03
3				5,5	7	-0,4254	-0,3230	0,102	10,2	14	1,39
4	мат.ожид	9,7		7	8,5	-0,3230	-0,1598	0,163	16,3	17	0,03
5				8,5	10	-0,1598	0,0410	0,201	20,1	17	0,47
6				10	11,5	0,0410	0,2317	0,191	19,1	15	0,87
7				11,5	13	0,2317	0,3714	0,140	14,0	14	0,00
8	дисперсия	8,48		13	14,5	0,3714	0,4503	0,079	7,9	11	1,22
9	ср.квад.отк	2,91		14,5	16	0,4503	0,5	0,050	5,0	5	0,00
10									χ^2 критическое		4,01
11									χ^2 теоретическое		11,1

Ячейка	Формула
P2	-0,5
P3	=НОРМ.РАСП(N3;$L$4;$L$9;1)-0,5 Копируем вниз до ячейки P9
Q2	=P3 Копируем до ячейки Q8
Q9	0,5
R2	=P2-Q2 Копируем вниз до ячейки R9
S2	=P2*100 Копируем вниз до ячейки S9
T2-T9	Частоты=относительные частоты в ячейках B8-I8, умноженные на 100
U2	=(T2-S2)^2/S2 Копируем вниз до ячейки U9

Вычислим значение меры расхождения по формуле

$$(\chi^2)^* = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} = 4{,}01$$ (в ячейке U10 =СУММ(U2:U9)).

Определим число степеней свободы $f=m-r-1=8-2-1=5$.

χ^2 теоретическое найдем, используя функцию ХИ2.ОБР категории «Статистические». Выделим ячейку U11 и в строке формул введем =ХИ2.ОБР(0,95;5). В этой ячейке получим теоретическое значение критерия, $\chi^2=11{,}1$.

Так как $(\chi^2)^*<\chi^2$, то функция распределения при заданном уровне значимости (β=5%) согласуется с экспериментальными данными и гипотезу о том, что часовая выработка автомобиля распределена по нормальному закону, можно считать правдоподобной.