Элементы математической статистики

Ахмеджанова Т.Д.

«статистика»

- происходит от латинского слова status - состояние, положение вещей. Первоначально оно употреблялось в значении «политическое состояние».
- В научный обиход это слово вошло в XVIII в. и первоначально употреблялось в значении «государствоведение».

- Математическая статистика возникла и развивалась параллельно с теорией вероятностей (XVII в.).
- Дальнейшее развитие математической статистики (вторая половина XIX — начало XX в.) обязано П. Л. Чебышеву, А. А. Маркову, А. М. Ляпунову, К. Гауссу, А. Кетле, Ф. Гальтону, К.Пирсону и др.

В XX в. наиболее существенный вклад в математическую статистику был сделан советскими :

В. И. Романовский, Е. Е. Слуцкий, А. Н. Колмогоров, Н. В. Смирнов;

английскими:

Стьюдент, Р. Фишер, Э. Пирсон;

американскими математиками:

Ю. Нейман, А. Вальд.

Математическая статистика

– раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов. Такое определение сформулировано математиками А.Н. Колмогоровым и Ю.В. Прохоровым.

Математическая статистика исходит из предположения, что наблюдаемая изменчивость окружающего мира имеет два источника:

- действие известных причин и факторов. Они порождают изменчивость, закономерно объяснимую.
- действие случайных причин и факторов.

Большинство природных и общественных явлений обнаруживают изменчивость, которая не может быть целиком объяснена закономерными причинами. В таком случае прибегают к концепции случайной изменчивости.

Выражение «случайный» в данном контексте означает «подчиняющийся законам теории вероятностей».

Проверка различных научных гипотез и моделей является случайным событием, так как результаты исследования определяются большим количеством заранее непредсказуемых факторов.

Определенные закономерности можно выявить только в случае массовых наблюдений вследствие закона больших чисел.

Закон больших чисел – это объективный математический закон, согласно которому совместное действие большого числа случайных факторов приводит к результату, почти не зависящему от случая.

Статистический подход

– выявление закономерной изменчивости на фоне случайных факторов и причин.

Методы математической статистики позволяют оценить параметры имеющихся закономерностей, проверить те или иные гипотезы об этих закономерностях.

Аппарат математической статистики

является инструментом для отсеивания закономерностей от случайностей.

Задача исследователя

- накапливать информацию об окружающем мире, пытаясь выделить закономерности из случайностей.

- В **теории вероятностей** рассматриваются случайные величины с заданным распределением или случайные эксперименты, свойства которых целиком известны. Предмет теории вероятностей – свойства и взаимосвязи этих величин (распределений).
- **Математическая статистика** опирается на методы и понятия теории вероятностей, но решает в каком-то смысле обратные задачи.

Характеристика областей применения аппарата

Теория вероятностей

- Модель, описывающая изучаемое явление или объект, известна априори (до опыта). Есть сведения обо всей генеральной совокупности, описывающей исследуемое явление.
- Используемый математический аппарат не зависит от предметной области.
- Выводы о поведении исследуемого объекта или явления делаются по всей генеральной совокупности.

Математическая статистика

- Модель, описывающая исследуемое явление, априори неизвестна.
- Для определения модели можно проводить пробные испытания (сформировать выборку из генеральной совокупности).
- Иногда модель может быть задана априори с точностью до неизвестных параметров.
- Значения неизвестных параметров модели могут быть приближенно получены по выборке из генеральной совокупности.
- Выводы о поведении объекта или явления делаются по выборке ограниченного объема и распространяются на всю генеральную совокупность.

Предмет исследования в математической статистике

- совокупность объектов, однородных относительно некоторых признаков.

Например,

- дети 10 лет г. Братска;
- пловцы-мастера спорта России.

Допустим, повторением одного и того же случайного эксперимента в одинаковых условиях получен набор числовых результатов. При этом у исследователя возникают вопросы:

- Если мы наблюдаем одну случайную величину – как по набору ее значений в нескольких опытах сделать как можно более точный вывод о ее распределении?
- Если мы наблюдаем одновременно проявление двух (или более) признаков, т.е. имеем набор значений нескольких случайных величин — что можно сказать об их зависимости? Есть она или нет? А если есть, то какова эта зависимость?

Если сделать предположения о распределении или о его свойствах до эксперимента, то по опытным данным обычно требуется подтвердить или опровергнуть эти гипотезы с определенной степенью достоверности.

Наиболее благоприятной для исследования оказывается ситуация, когда можно уверенно утверждать о некоторых свойствах наблюдаемого эксперимента – например, о наличии функциональной зависимости между наблюдаемыми величинами, о нормальности распределения, о его симметричности, о наличии у распределения плотности или о его дискретном характере, и т.д.

Пусть каждому i объекту соответствует значение x_i, $i = \overline{1, N}$ где N - количество всех исследуемых объектов. Совокупность всех возможных значений (теоретически домысливаемых) N объектов называется ***генеральной совокупностью***, а N – ***объемом генеральной совокупности***.

Генеральная совокупность может быть конечной или бесконечной.

Например, изучение физической подготовленности детей 10 лет г. Братска.

- Пусть количество реально наблюдаемых объектов из N равно *n*. Тогда x_i, – ***выборка из генеральной совокупности***, *n* – ***объем выборки***.

Выборка из генеральной совокупности должна обладать следующими свойствами:

- каждый элемент x_i выбран случайно;
- все x_i имеют одинаковую вероятность попасть в выборку;
- n должно быть настолько велико, насколько это позволяет решать задачу с требуемым качеством (выборка должна быть репрезентативной, представительной).

Формы представления выборки из генеральной совокупности.

1. Представление выборки из генеральной совокупности в **негруппированном виде**. Этот ряд называется **простым статистическим рядом.**

   Такая форма связана с наличием сведений о каждом элементе выборки.

Пример:

- измерена масса тела 10 девочек 6 лет. Полученные данные образуют простой статистический ряд:

  24 22 23 26 24 23 25 27 25 25

Отдельные значения статистического ряда называются **вариантами**. Если варианта x_i появилась m раз, то число m называют **частотой**, а ее отношение к объему выборки m/n – **относительной частотой (частостью)**.

2. Представление выборки в виде **вариационного ряда**

(в упорядоченном виде):

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(i)} \le \ldots \le x_{(n)} .$$

В этом случае $x_{(i)}$ – член вариационного ряда, или **варианта**. Часто $x_{(i)}$ называют **порядковой статистикой**.

Пример:

Вариационный ряд:

22 23 23 24 24 25 25 25 26 27

- Таблица, в первой строке которой записаны все значения величины (варианты), во второй –- соответствующие им частоты, называется также **вариационным рядом по значениям**.

Пример:

x_i	22	23	24	25	26	27
n_i	1	2	2	3	1	1

Понятие **репрезентативная выборка** не всегда можно связать с её объемом *n*. Чаще это зависит от реально исследуемого объекта или явления, объема генеральной совокупности, трудоёмкости и стоимости получения наблюдений или измерений для формирования выборки.

Форма представления выборки из генеральной совокупности в виде вариационного ряда не приводит к потере информации о каждом элементе выборки, но искажает информацию, устанавливая зависимость между соседними элементами выборки.

Необходимо помнить! Члены вариационного ряда, в отличие от элементов исходной выборки, уже не являются взаимно независимыми (по причине их предварительной упорядоченности).

Представление выборки в **группированном виде**.

Такая форма представления выборки из генеральной совокупности связана с разбиением области задания случайной величины X на L интервалов группирования. При этом известно только количество элементов выборки n_j, , попавших в j интервал и последовательность границ интервалов разбиения.

Для определения числа *L* интервалов искусственного группирования пользуются формулой Старджеса

$$L = 1 + 3.322 \lg n$$

Иногда *L* может быть задано природой исследуемого явления или условиями проведения эксперимента. В этом случае ширина каждого интервала может быть отличной от других (**неравноточное группирование**).

На некоторых этапах статистического анализа необходимо исходную выборку представлять в группированном виде.

Последовательность процедуры группирования неупорядоченной выборки из генеральной совокупности

1. Формирование вариационного ряда.
2. Выделение минимального и максимального элементов выборки

$$x_{min} = x_{(1)},$$
$$x_{max} = x_{(n)}.$$

3. Определение числа интервалов группирования осуществляется из соображения точности и устанавливается эмпирическим путем в зависимости от объема выборки, либо по формуле Старджеса, либо определяется природой явления или условиями проведения эксперимента. Округление при нахождении L осуществляется до ближайшего целого числа.

4. Определение ширины интервалов гистограммы (при равноточном группировании)

$$h = \frac{x_{(n)} - x_{(1)}}{L}$$

Если при вычислении h необходимо округлить результат, следует помнить, что последний интервал группирования будет меньше ширины h при округлении в большую сторону и больше h - при округлении в меньшую сторону.

5. Формирование последовательности границ интервалов разбиения.
Образуемый вариационный ряд границ интервалов группирования будет выглядеть как
$x_{(1)}, x_{(1)} + h, x_{(1)} + 2h, \ldots, x_{(1)} + (L-1) \times h, x_{(n)}$.

- Иногда, для того чтобы $x_{(1)}$ и $x_{(n)}$ попали внутрь соответственно 1-го и L-го интервалов группирования, границы $x_{(1)}$ и x(n) корректируют следующим образом:

$$x'_{(1)} = x_{(1)} - h/2,$$

$$x'_{(n)} = x_{(n)} + h/2.$$

- Следовательно, число интервалов разбиения увеличивается на 1

$$L' = L + 1.$$

- При этом последовательность границ интервалов разбиения будет представлена в виде

  $x'_{(1)}, x'_{(1)} + h, x'_{(1)} + 2h, \ldots, x'_{(1)} + L \times h, x'_{(n)}$

6. *Определение количества элементов выборки n_j, попавших в каждый j интервал.*

Пример

Даны объемы ежедневной выработки в течение месяц (в тыс. руб.) пятидесяти продавцов молочных изделий, работающих в разных районах города

15 19 6 18 21 16 20 17 15 10

16 20 7 19 22 17 21 19 16 11

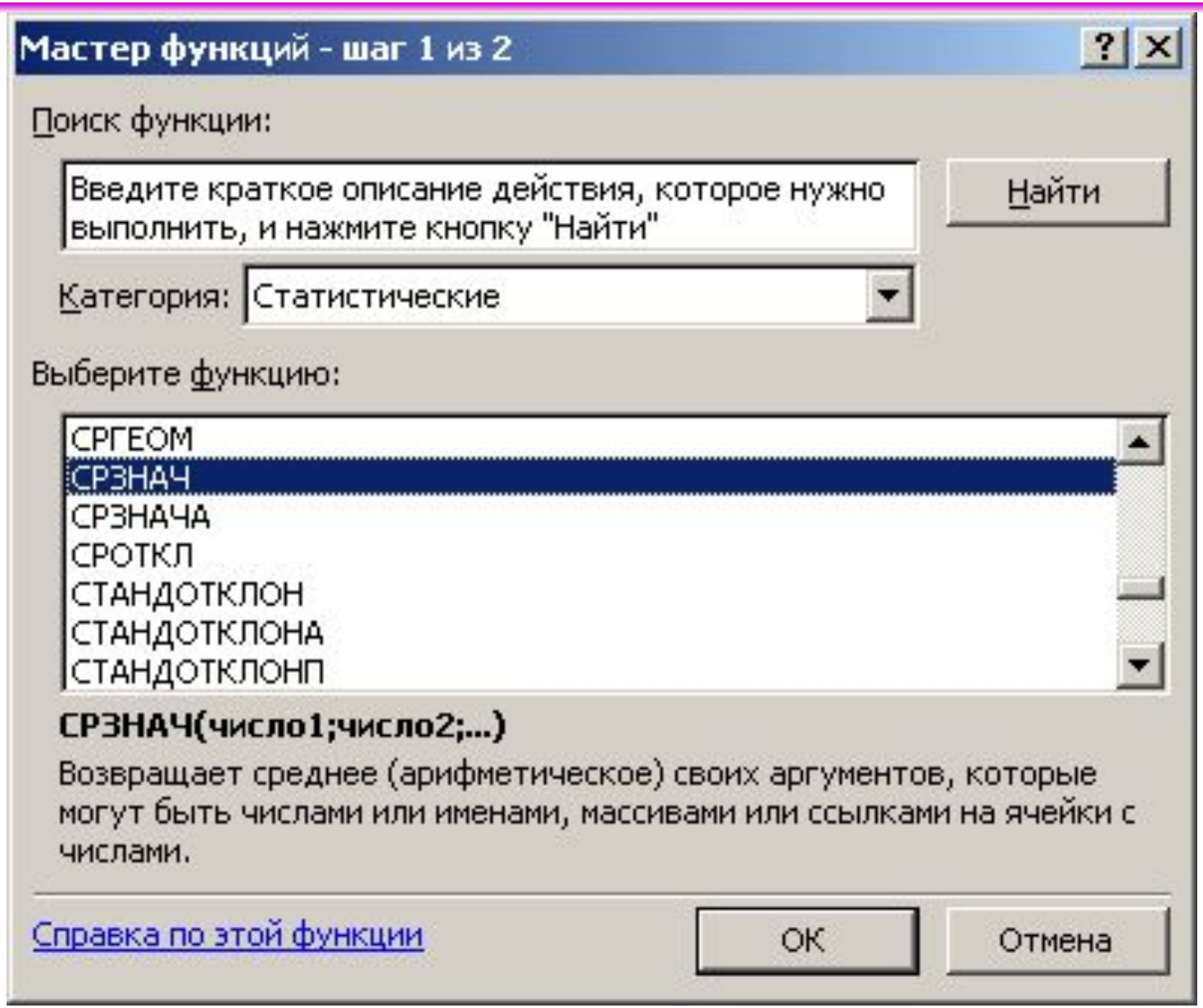
19 10 8 18 20 8 18 16 20 12

16 21 21 9 19 19 14 18 19 19

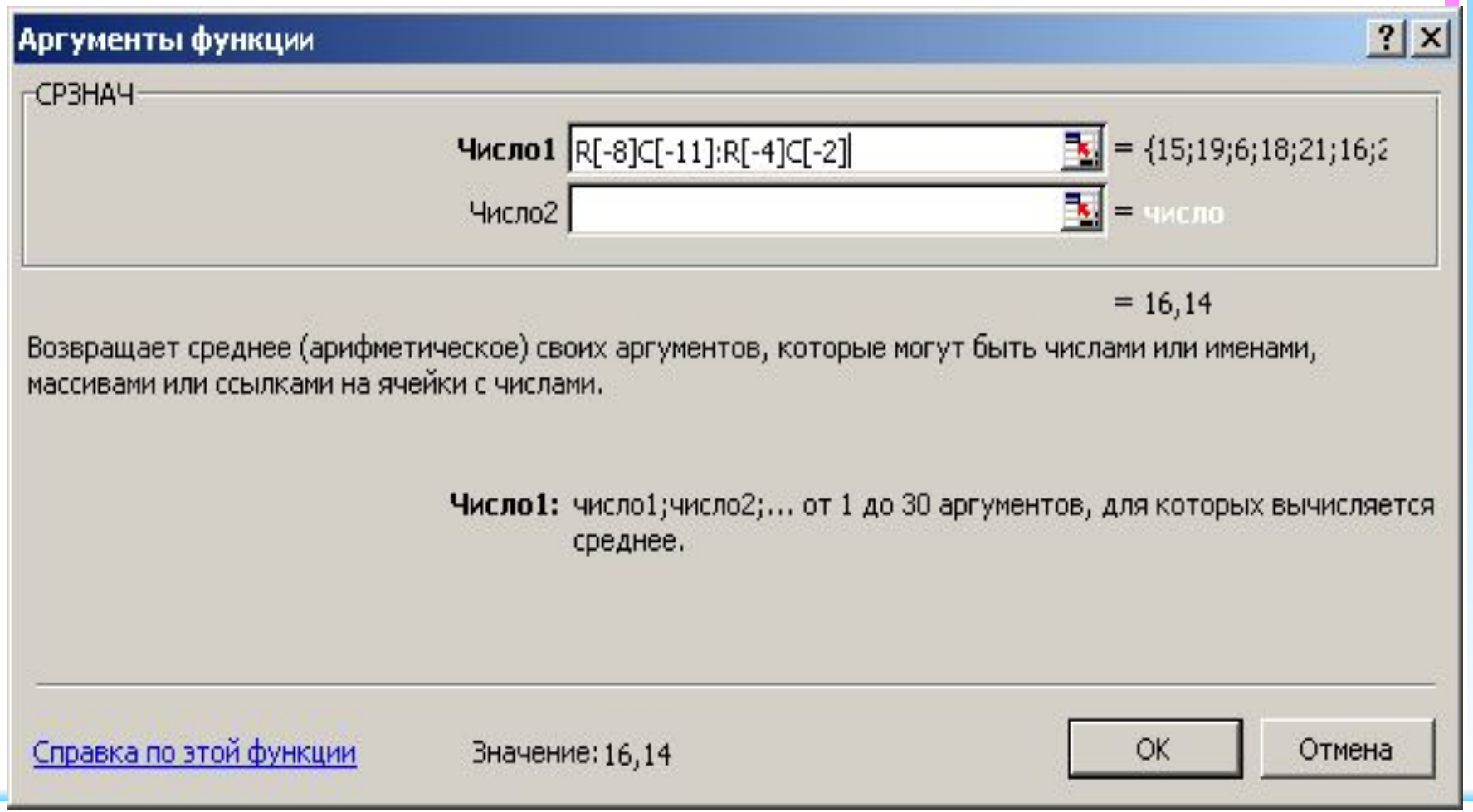
12 20 20 8 13 10 18 17 22 18.

В EXCEL

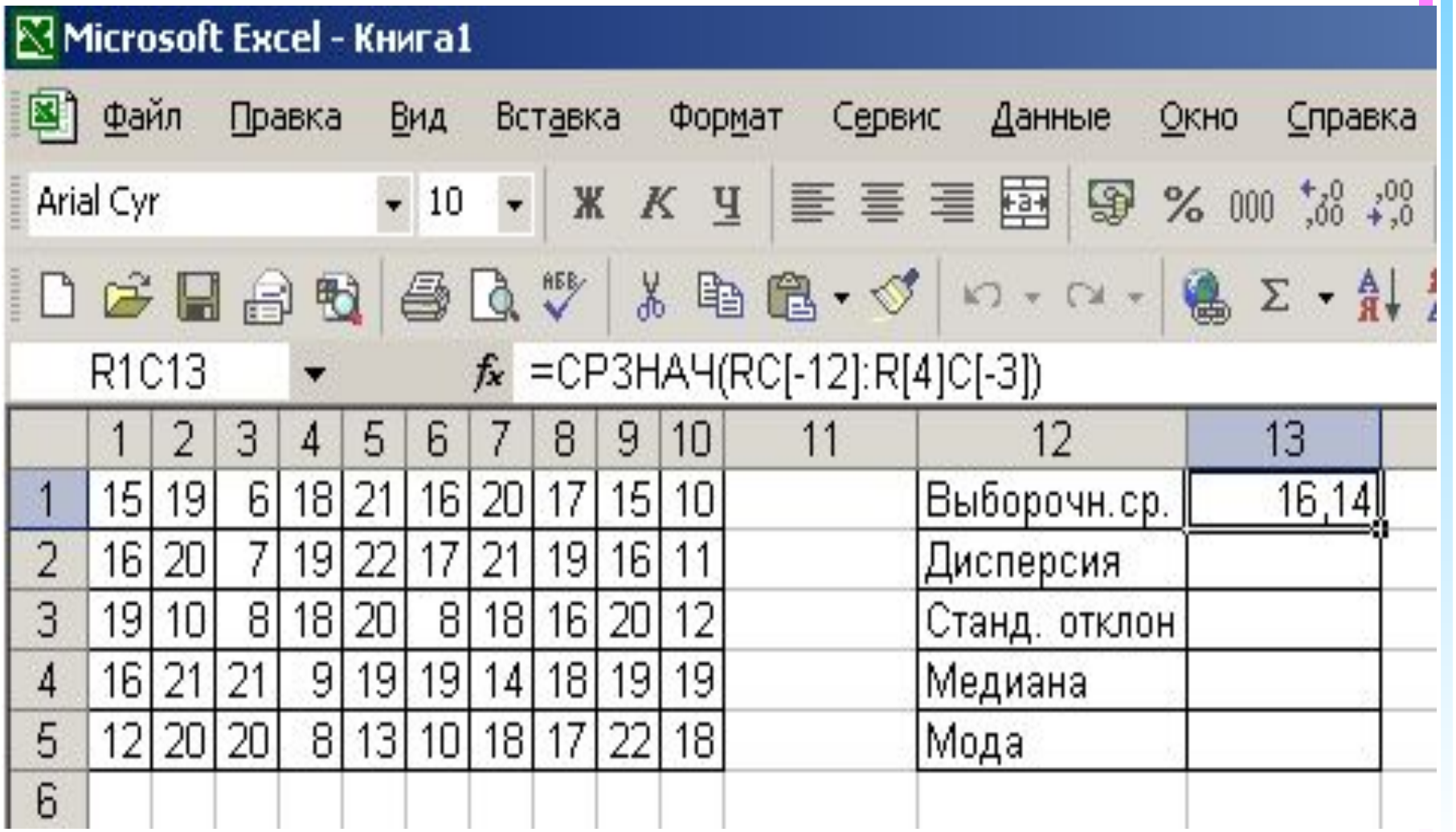
Находим основные числовые характеристики выборки: выборочную среднюю, выборочную дисперсию, стандартное отклонение, моду, медиану. Для этого в Excel в отдельные ячейки вводим данные выборки, устанавливаем курсор в желаемой ячейке, выбираем «мастер функций» «статистические», «СРЗНАЧ», нажимаем ОК:


Мастер функций - шаг 1 из 2
Поиск функции:
Введите краткое описание действия, которое нужно выполнить, и нажмите кнопку "Найти"
Найти
Категория: Статистические
Выберите функцию:
СРГЕОМ
СРЗНАЧ
СРЗНАЧА
СРОТКЛ
СТАНДОТКЛОН
СТАНДОТКЛОНА
СТАНДОТКЛОНП
СРЗНАЧ(число1;число2;...)
Возвращает среднее (арифметическое) своих аргументов, которые могут быть числами или именами, массивами или ссылками на ячейки с числами.
Справка по этой функции
ОК
Отмена

В «Число1» ставим курсор и выделяем весь диапазон, в котором находится выборка, нажимаем ОК:

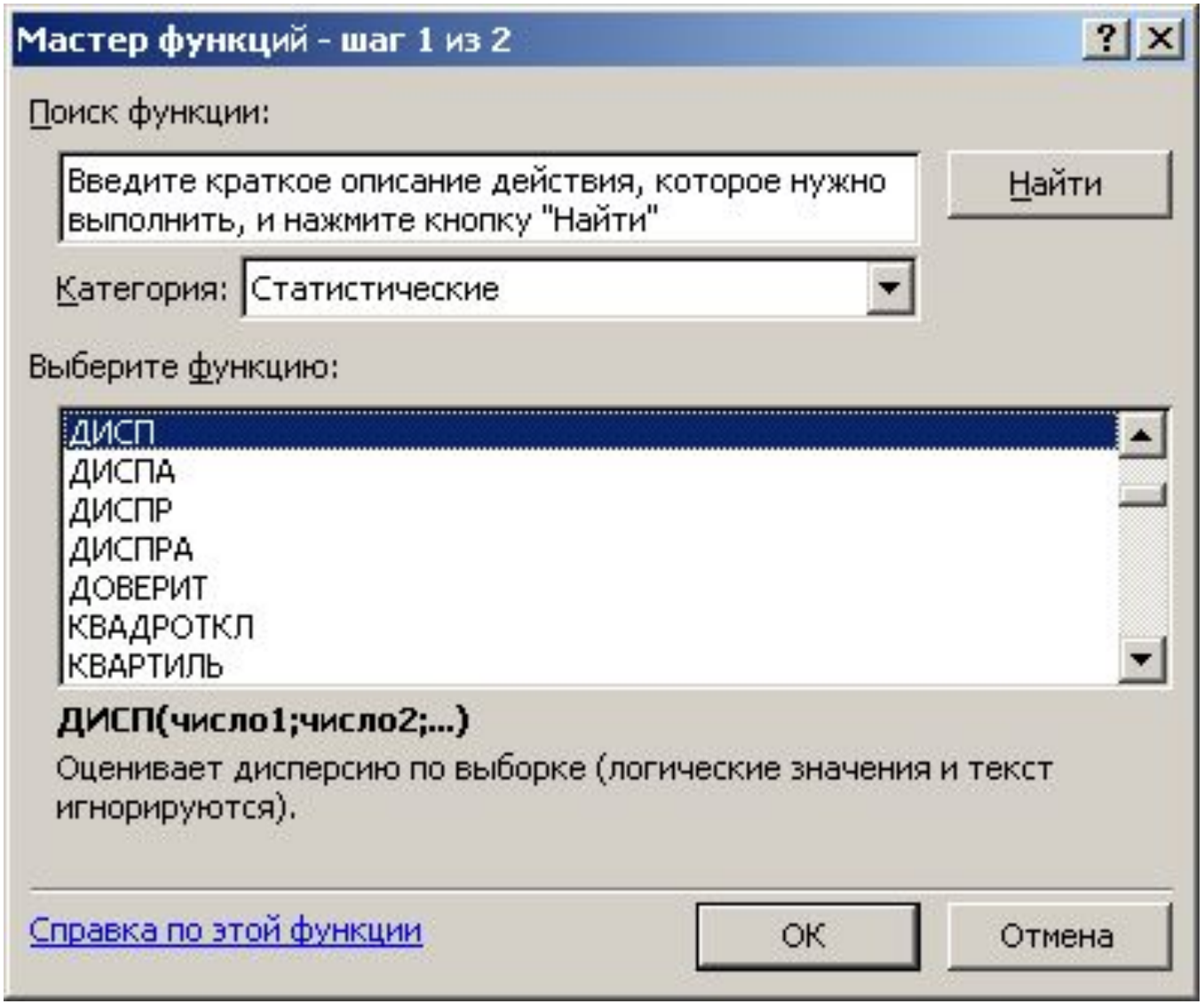


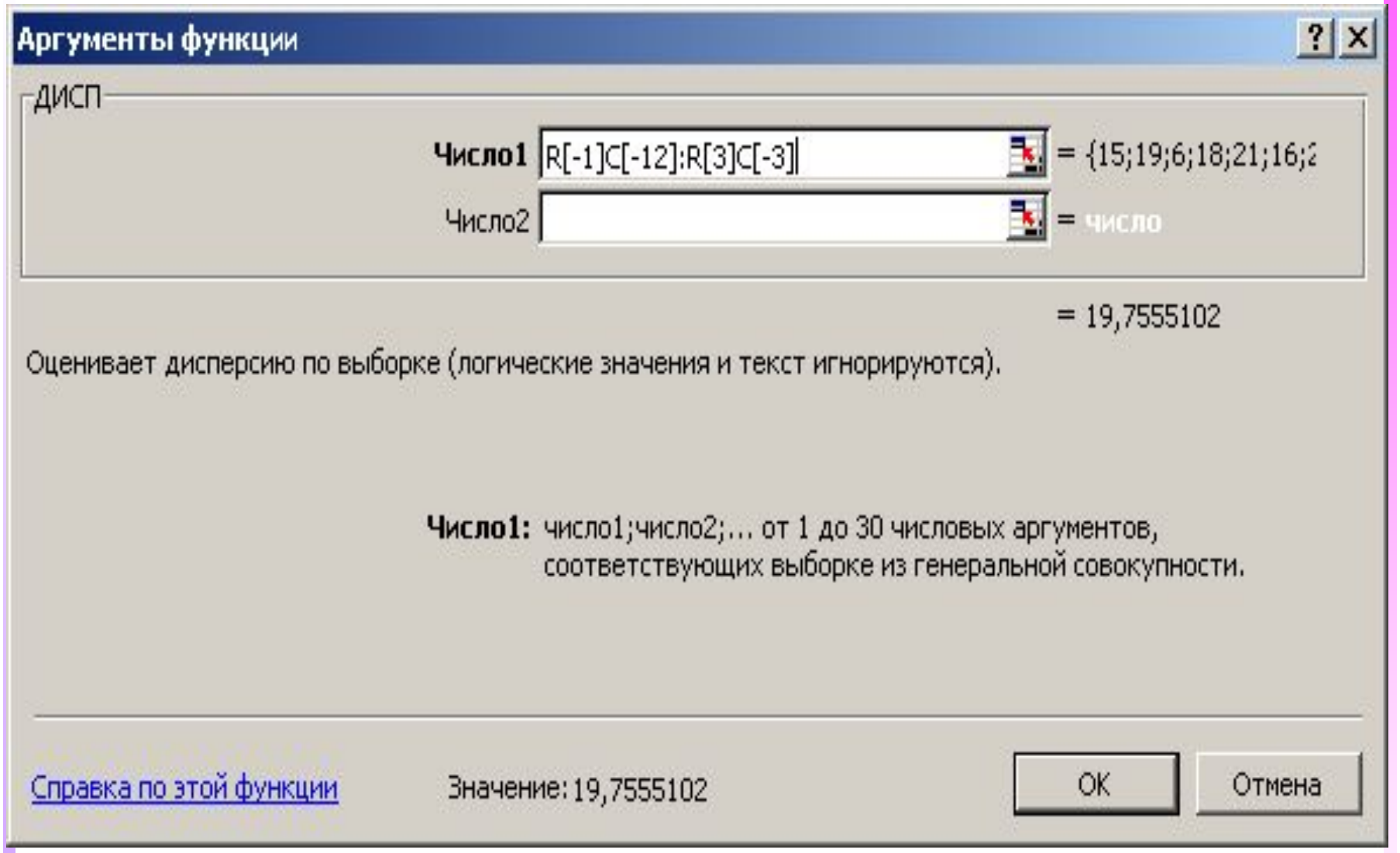
Получаем в соответствующей ячейке искомое значение:



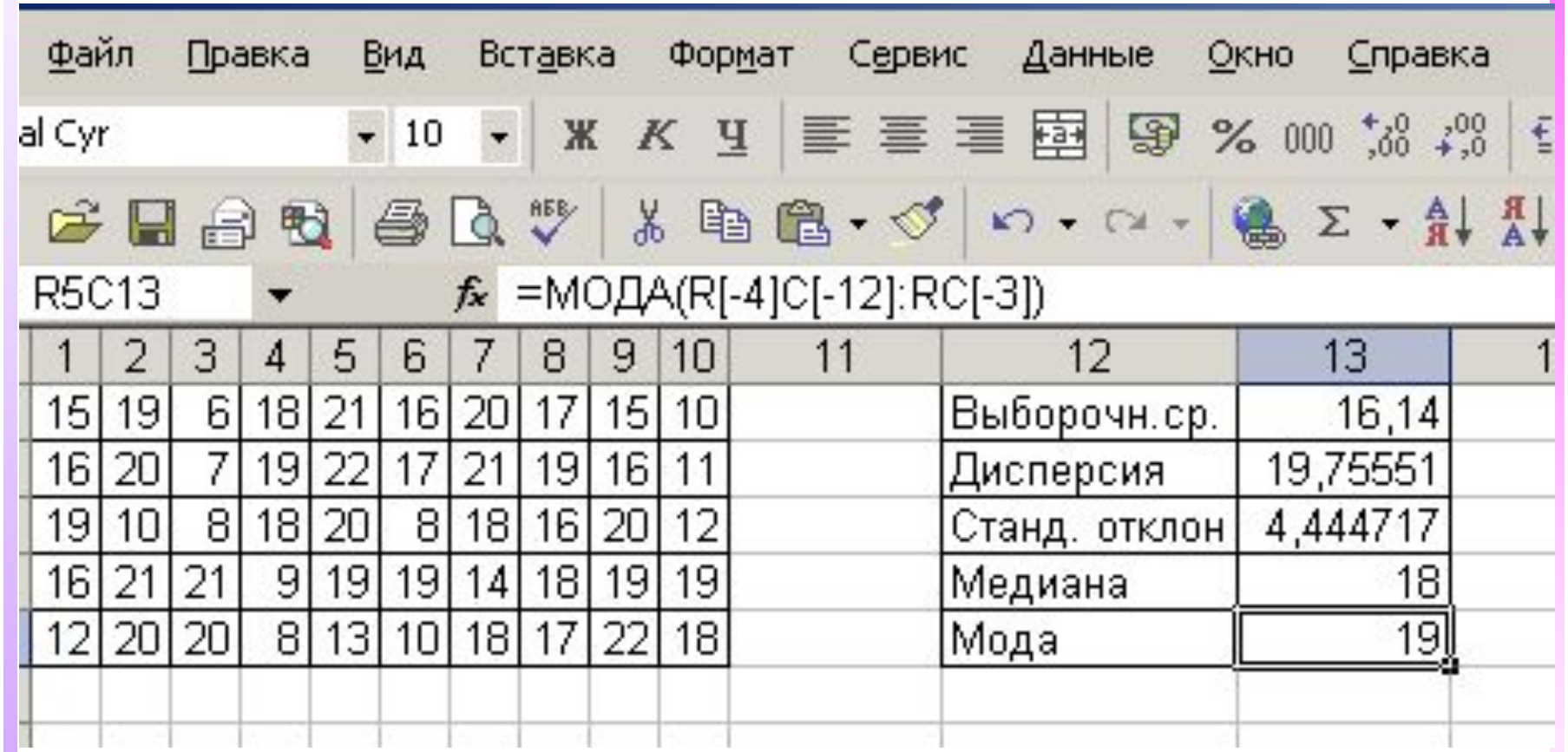
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	15	19	6	18	21	16	20	17	15	10		Выборочн.ср.	16,14
2	16	20	7	19	22	17	21	19	16	11		Дисперсия	
3	19	10	8	18	20	8	18	16	20	12		Станд. отклон	
4	16	21	21	9	19	19	14	18	19	19		Медиана	
5	12	20	20	8	13	10	18	17	22	18		Мода	
6													

Далее действуем аналогично:




Аргументы функции
ДИСП
Число1
R[-1]C[-12]:R[3]C[-3]
= {15;19;6;18;21;16;
Число2
= число
= 19,7555102
Оценивает дисперсию по выборке (логические значения и текст игнорируются).
Число1: число1;число2;... от 1 до 30 числовых аргументов, соответствующих выборке из генеральной совокупности.
Справка по этой функции
Значение: 19,7555102
OK
Отмена

Так получаем основные числовые характеристики:



Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

al Cyr 10 Ж К Ч

R5C13 =МОДА(R[-4]C[-12]:RC[-3])

1	2	3	4	5	6	7	8	9	10	11	12	13	1
15	19	6	18	21	16	20	17	15	10		Выборочн.ср.	16,14	
16	20	7	19	22	17	21	19	16	11		Дисперсия	19,75551	
19	10	8	18	20	8	18	16	20	12		Станд. отклон	4,444717	
16	21	21	9	19	19	14	18	19	19		Медиана	18	
12	20	20	8	13	10	18	17	22	18		Мода	19	

Представим выборку в группированном виде.

1. Формируем вариационный ряд

6 9 12 15 16 18 19 19 20 21

7 10 12 16 17 18 19 19 20 21

8 10 13 16 17 18 19 19 20 21

8 10 14 16 17 18 19 20 20 21

8 11 15 16 18 18 19 20 21 22.

Находим $x_{(1)} = 6$, $x_{(n)} = 22$.

3. Определяем число интервалов разбиения по формуле Старджеса

$$L = 1 + 3{,}322 \lg 50 = 6.6 \text{ , } L = 7.$$

4. Находим ширину интервала разбиения h

$$h = (22 - 6) / 7 = 2.2857.$$

Ограничимся двумя знаками после запятой и получим h = 2.28. Так как h округлено в сторону уменьшения, последний интервал будет шире предыдущих.

5. Строим вариационный ряд границ интервалов группирования (без корректировки границ первого и последнего интервалов):

   [6; 8.28), [8.28; 10.56), [10.56; 12.84), [12.84; 15.12), [15.12; 17.4), [17.4; 19.68), [19.68; 22].

6. Находим количество элементов выборки n_j, попавших в j интервал:

j	1	2	3	4	5	6	7
n_j	5	4	3	4	8	14	12

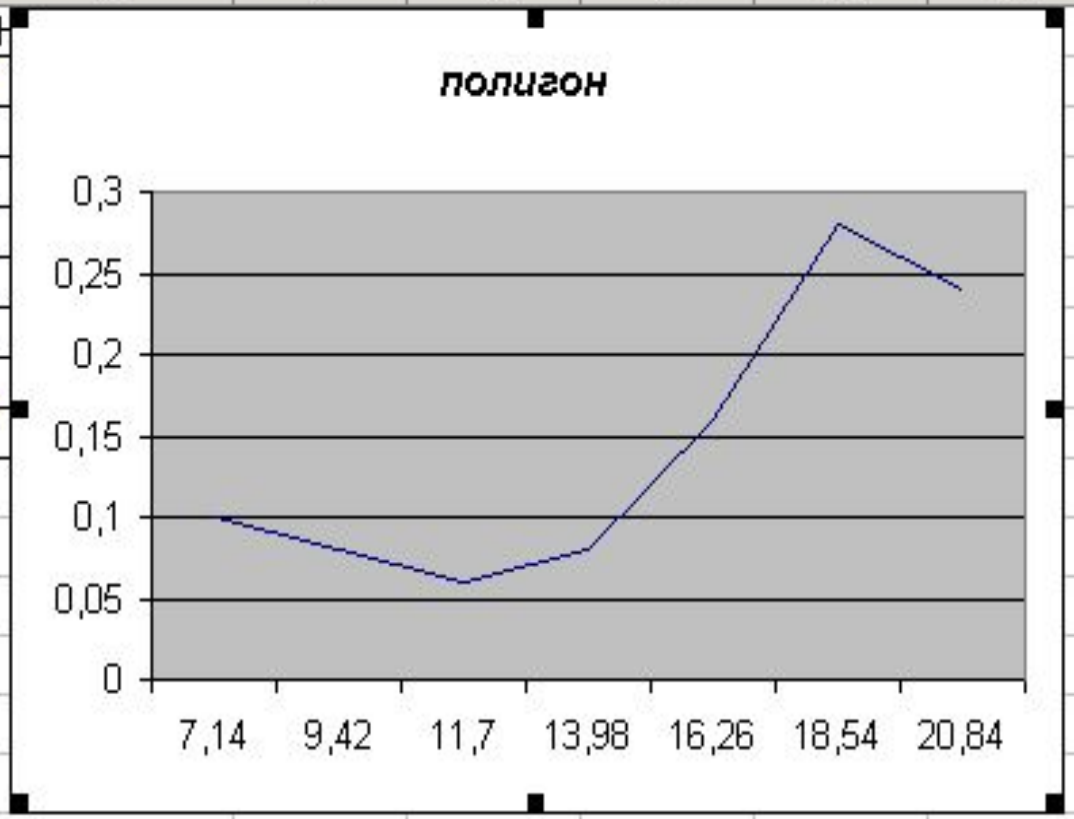
- Группированная форма представления случайной величины не содержит информации о каждом элементе выборки.
- При этом часто в качестве значения случайной величины на интервале принимается его середина.

Используя полученные результаты и с помощью стандартных функций Excel получаем таблицу:

Интервал	Серед.интерв.	Частота	Отн.частота	Накопл. Част.
[6; 8.28)	7,14	5	0,1	0,1
[8.28; 10.56)	9,42	4	0,08	0,18
[10.56; 12.84)	11,7	3	0,06	0,24
[12.84; 15.12)	13,98	4	0,08	0,32
[15.12; 17.4)	16,26	8	0,16	0,48
[17.4; 19.68)	18,54	14	0,28	0,76
[19.68; 22]	20,84	12	0,24	1
Сумма		50	1	

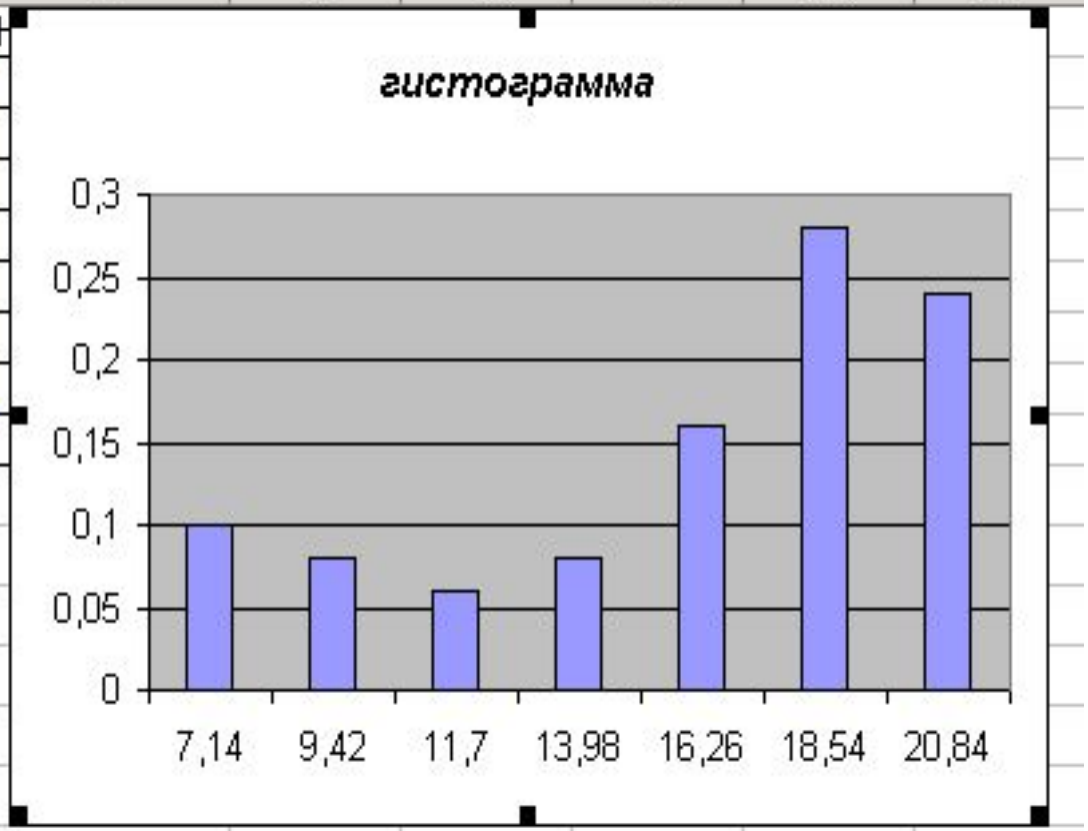
Строим соответствующие графики: полигон

Серед. интерв.	Частота	Отн. частота
7,14	5	0,1
9,42	4	0,08
11,7	3	0,06
13,98	4	0,08
16,26	8	0,16
18,54	14	0,28
20,84	12	0,24
	50	1



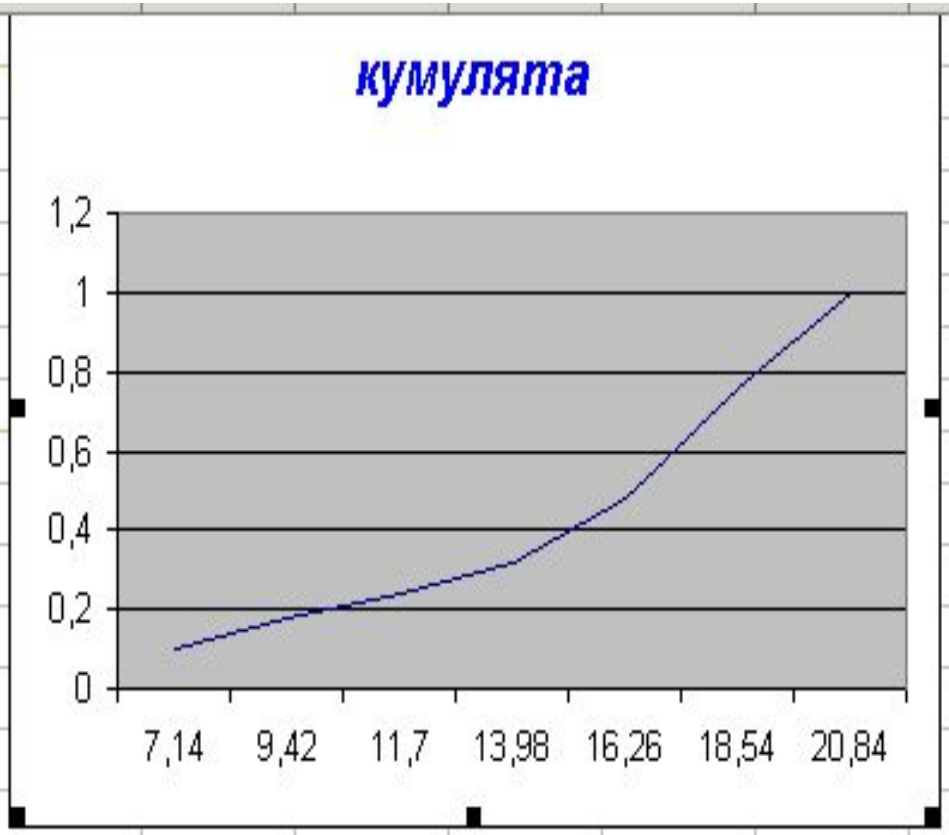
гистограмма

Серед.интерв.	Частота	Отн.частота
7,14	5	0,1
9,42	4	0,08
11,7	3	0,06
13,98	4	0,08
16,26	8	0,16
18,54	14	0,28
20,84	12	0,24
	50	1



кумулята:

Серед.интерв.	Частота	Отн.частота	Накопл. Част.
7,14	5	0,1	0,1
9,42	4	0,08	0,18
11,7	3	0,06	0,24
13,98	4	0,08	0,32
16,26	8	0,16	0,48
18,54	14	0,28	0,76
20,84	12	0,24	1
	50	1	



Это важно!

От негруппированной выборки всегда можно перейти к группированной, но не наоборот. Переход к группированной форме представления выборки сопряжен с потерей информации об исследуемом объекте, процессе или явлении.

Характеристики случайной величины, полученные по выборке из генеральной совокупности, называются ***выборочными*** или ***эмпирическими характеристиками***, а характеристики, полученные по генеральной совокупности, – ***теоретическими*** или ***генеральными характеристиками.***

Все методы математической статистики можно разделить на ***параметрические методы***, основанные на использовании знаний о вероятностной модели, и ***непараметрические***, когда априорных представлений о виде модели нет, или она не используется.