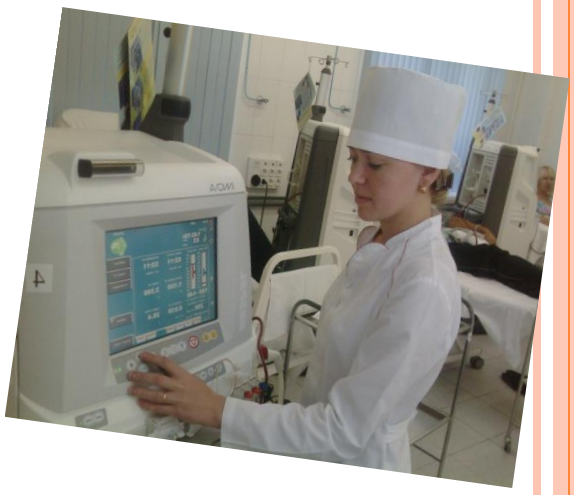


Лекция 7
Методы анализа данных в Excel.

Составитель: Космачева И.М.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТИ И МАТСТАТИСТИКИ

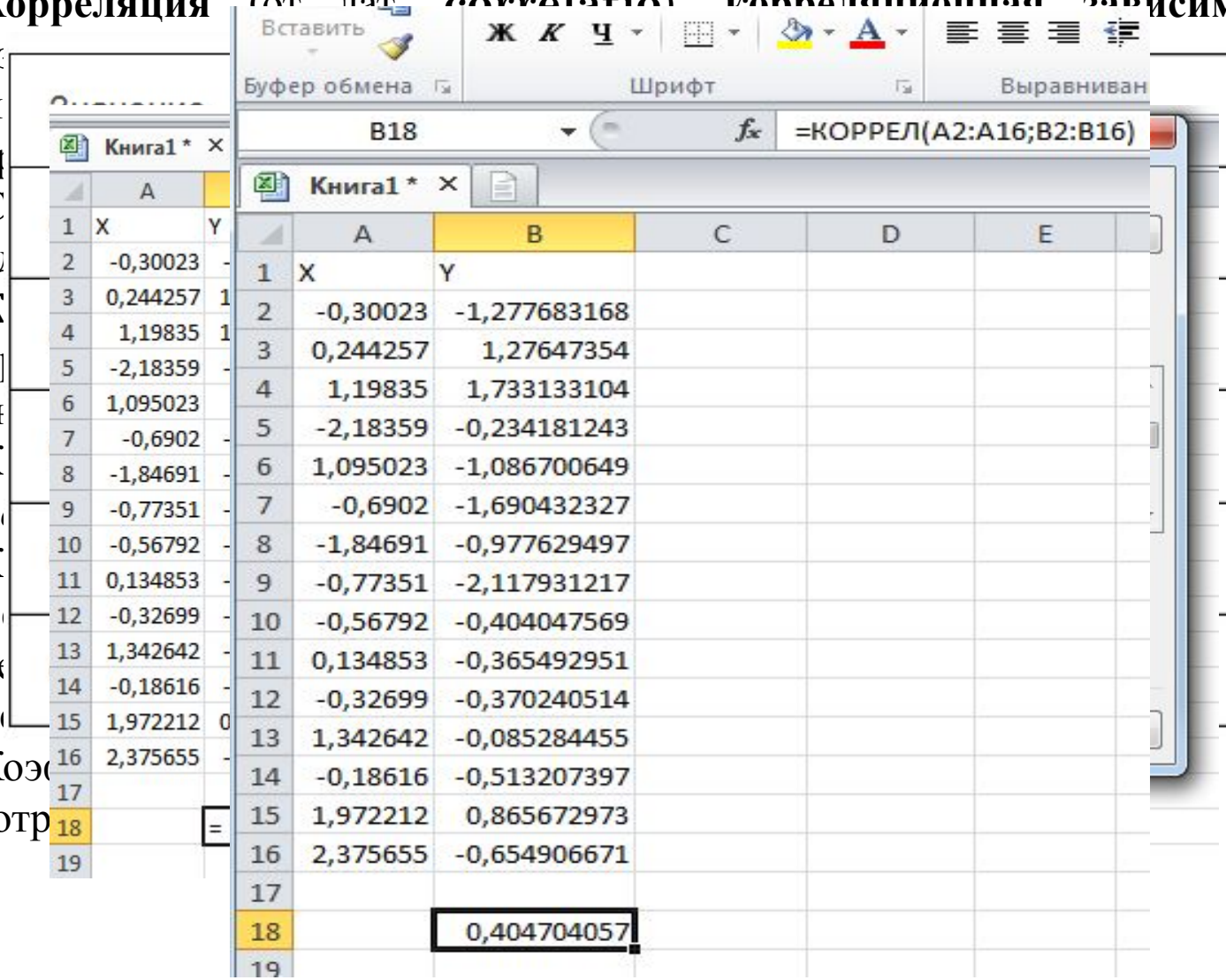
- Любое значение параметра, вычисленное на основе ограниченного числа наблюдений, непременно содержит элемент случайности. Результат эксперимента - случайная величина.
- Такое приближенное, случайное значение называется **оценкой параметра**.
- **Оценкой параметра** называют функцию результатов наблюдений над случайной величиной (статистику), с помощью которой судят о значении параметра .
- **$\tilde{a}(N)$** – статистическая оценка параметра **a** по данным **N** опытов (прогонов).
- Генеральная совокупность характеризуется одним или несколькими параметрами**: μ, σ^2, σ** и т.д.

ОСНОВНЫЕ СТАТИСТИКИ

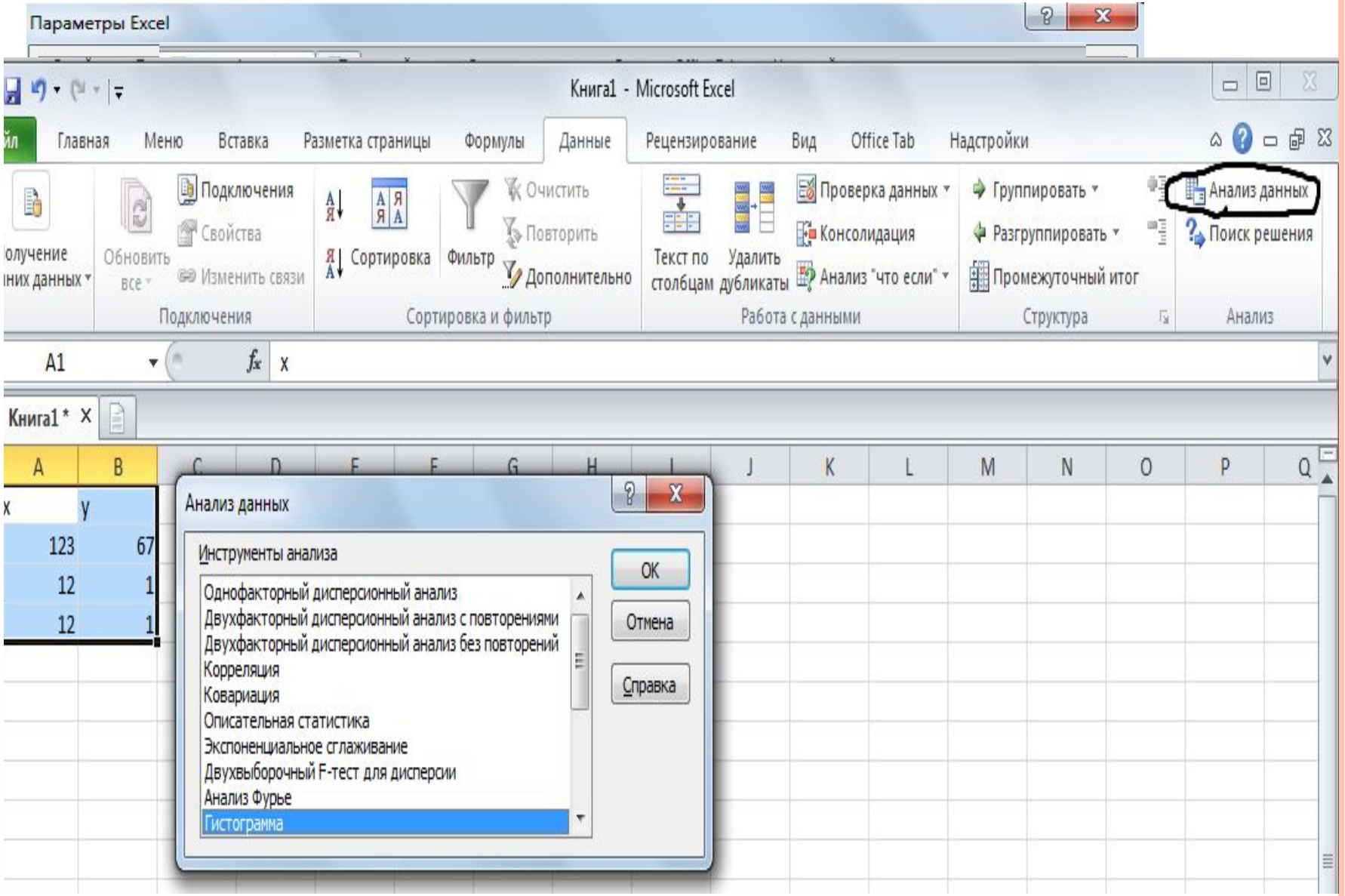
- ***Выборочное среднее x*** *– оценка математического ожидания, среднее арифметическое элементов выборки.*
- ***Выборочная дисперсия*** S^2 *– среднее квадратов отклонения элементов выборки от выборочного среднего, является оценкой дисперсии, характеризует разброс выборочных значений.*
- ***Стандартное отклонение S*** *– корень из дисперсии.*
- ***Коэффициент вариации*** *– отношение выборочного среднего квадратического отклонения к выборочной средней, характеризует рассеяние вне зависимости от размерности вариант .*
- ***Размах варьирования-*** *разность между наибольшей и наименьшей вариантами.*
- ***Медиана Ме.***
- ***Мода Мо.***
- ***Коэффициент эксцесса Е.***
- ***Коэффициент асимметрии А.***
- ***Процентиль.***

ОСНОВНЫЕ СТАТИСТИКИ

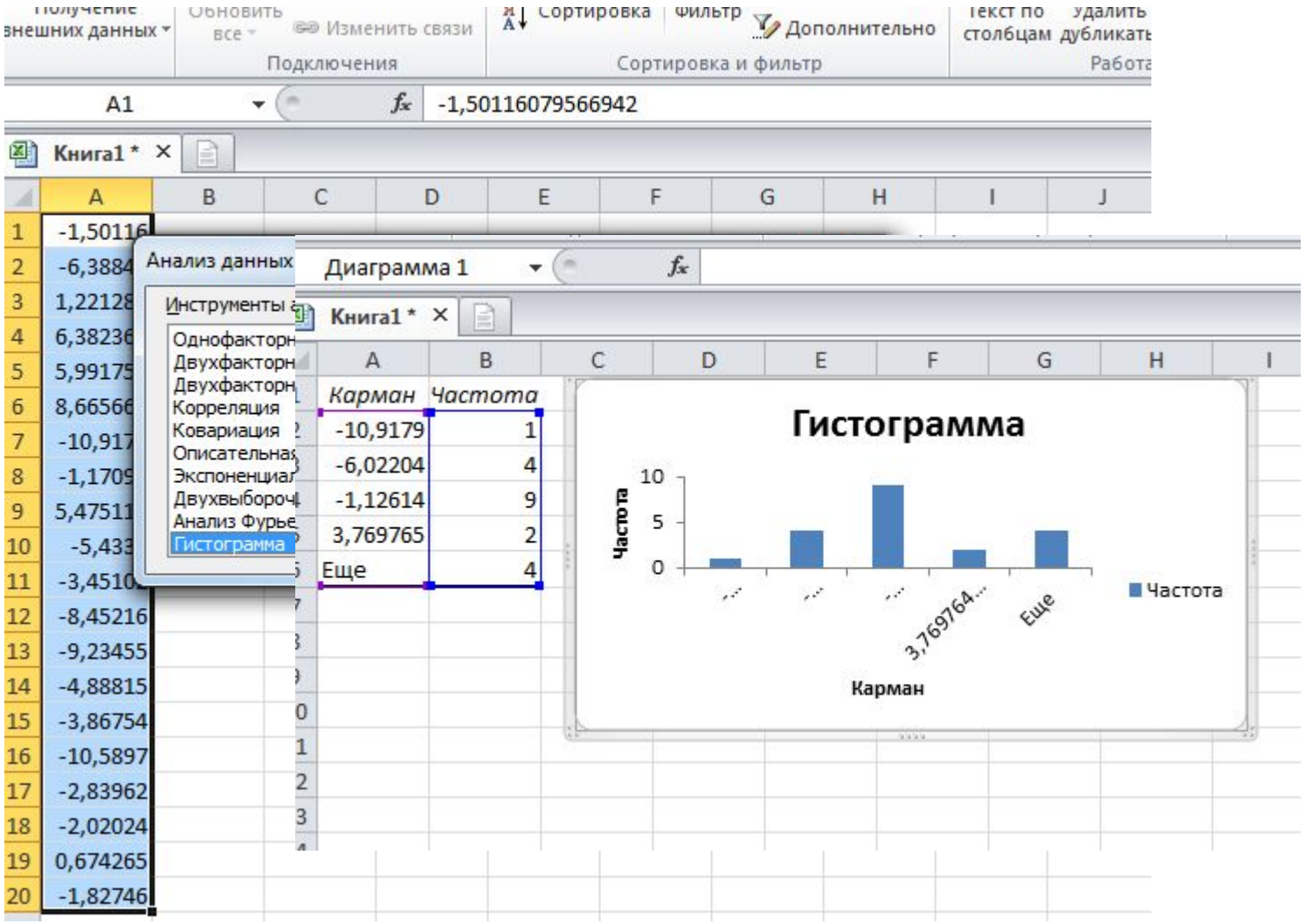
- **Корреляция** (от лат. correlatio) корреляционная зависимость -
- **К**
- **К**
- **К**
- **К**
- Коэ от **-1**



ОСНОВНЫЕ СТАТИСТИКИ



СТАТИСТИКА В EXCEL



ФУНКЦИИ В EXCEL

	A	B	C	D	E	F	G
1							
2	Выборочные значения	Границы интервалов	Частоты	Частности			
3	20,3	10	0	0			
4	15,3	12	2	2/55			
5	14,3	14	4	4/55	=С4/СЧЁТ(A$3:A$57)		
6	19,3	16	8	8/55			
7	10,1	18	12	12/55			
8	13,9	20	15	3/11			
9	19,5	22	11	1/5			
10	17,8	24	3	3/55			
11	15,4		0	0			
12	16,8		55	1			
13	20,1					=СУММ(D3:D11)	
14	17,8			=СУММ(C3:C11)			
15	21,1						
16	19,8	={ЧАСТОТА(A3:A57;B3:B10)}					
17	17,2						
18	13,5						
19	17,2						
20	13,2						

ОСНОВНЫЕ СТАТИСТИКИ

- При анализе результатов исследования необходимо представить их в обобщенной форме. Самым распространенным методом обобщения данных является их описание с помощью какой-либо меры центральной тенденции и какой-либо оценки вариабельности.
- Оценка вариабельности показывает, насколько хорошо среднее значение отражает свойства рассматриваемой выборки результатов.
- **Среднее квадратическое отклонение** не только характеризует **разброс результатов**, но также позволяет рассчитать **процентили**, с помощью которых можно судить **о степени исключительности конкретного результата**.
- При этом предполагается, что **данные распределяются** по **нормальному закону.** *Это условие соблюдается в большинстве случаев, с которыми обычно сталкиваются исследователи, однако не во всех.*

ОСНОВНЫЕ СТАТИСТИКИ

- **Коэффициент эксцесса Е** - характеризует «островерхость» гистограммы или полигона по сравнению с кривой Гаусса нормального распределения.
- **Коэффициент асимметрии А** - характеризует степень симметричности гистограммы или полигона по сравнению с кривой Гаусса. Если коэффициенты асимметрии и эксцесса близки к нулю, то форму распределения можно считать близкой нормальному.
- Если значения переменной распределены несимметрично относительно центра, то группы лучше описывать с помощью **медианы и квантилей (процентилей, квартилей, децилей).**

ОСНОВНЫЕ СТАТИСТИКИ

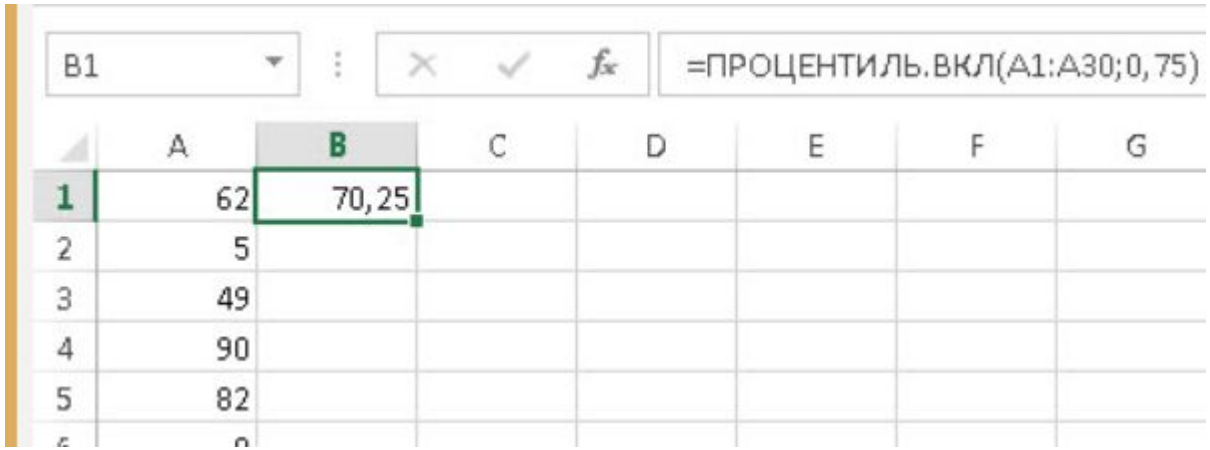
- **Квантилью x_p** (p-квантилью, квантилью уровня p) случайной величины, имеющей функцию распределения $F_x(x)$, называют решение x_p уравнения $F_x(x) = p$. **Для некоторых p** уравнение $F_x(x) = p$ может иметь **несколько решений**, для некоторых - **ни одного**.

Квантили, наиболее часто встречающиеся в практических задачах, имеют свои названия:

- ***медиана*** - квантиль уровня 0.5;
- ***нижняя квартиль*** - квантиль уровня 0.25;
- ***верхняя квартиль*** - квантиль уровня 0.75;
- ***децили*** - квантили уровней 0.1, 0.2, …, 0.9;
- ***процентили*** - квантили уровней 0.01, 0.02, …, 0.99.
- **Процентиль на уровне P - это** такое значение, ниже которого расположено P процентов наблюдений данной переменной. Например, значение 50-й процентили указывает, что 50% значений располагается ниже этого уровня.

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННОГО РЯДА

- **Процентиль** можно посчитать используя excel. Пусть значения лежат в диапазон от A1:A30. Надо ввести данную формулу **=ПРОЦЕНТИЛЬ.ВКЛ(A1:A30;0,75).**
- 75 процентиль ряда чисел равен 70,25, т.е. 75 % значений лежат ниже 70,25, на у остальные 25% лежат выше 70,25



ОСНОВНЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННОГО РЯДА

- **Медиана** - это такое значение признака, которое делит **упорядоченное** (ранжированное) множество данных пополам так, что одна половина всех значений оказывается меньше медианы, а другая - больше.

Если данные содержат нечетное число значений (8, 9, **10**, 13, 15), то медиана есть центральное значение;

Если данные содержат четное число значений (5, **8, 9**, 11), то медиана есть точка, лежащая посередине между двумя центральными значениями.

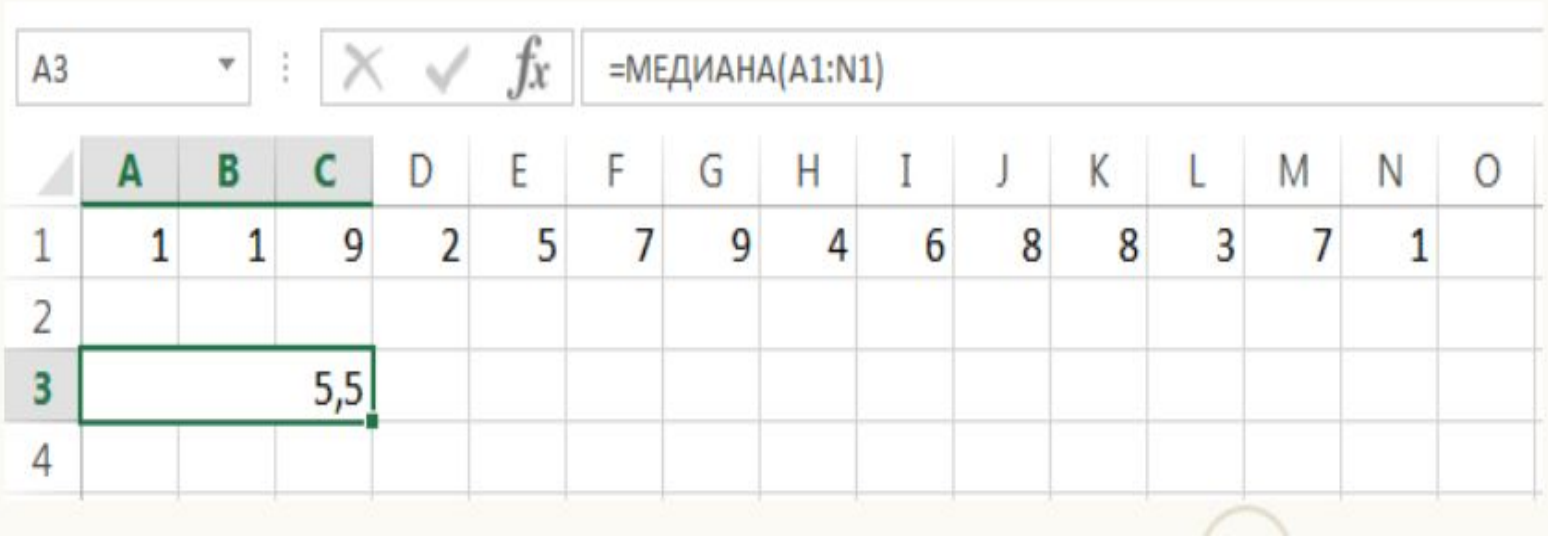
- **Мода** - это такое значение из множества измерений, которое встречается наиболее часто. Когда два соседних значения встречаются одинаково часто и чаще, чем любое другое значение, мода есть среднее этих двух значений.

ФУНКЦИИ В EXCEL

МЕДИАНА()

Статистическая функция **МЕДИАНА** возвращает медиану из заданного массива числовых данных. Медианой называют число, которое является серединой числового множества. Если в списке нечетное количество значений, то функция возвращает то, что находится ровно по середине. Если же количество значений четное, то функция возвращает среднее для двух чисел.

Например, на рисунке ниже формула возвращает медиану для списка, состоящего из 14 чисел.



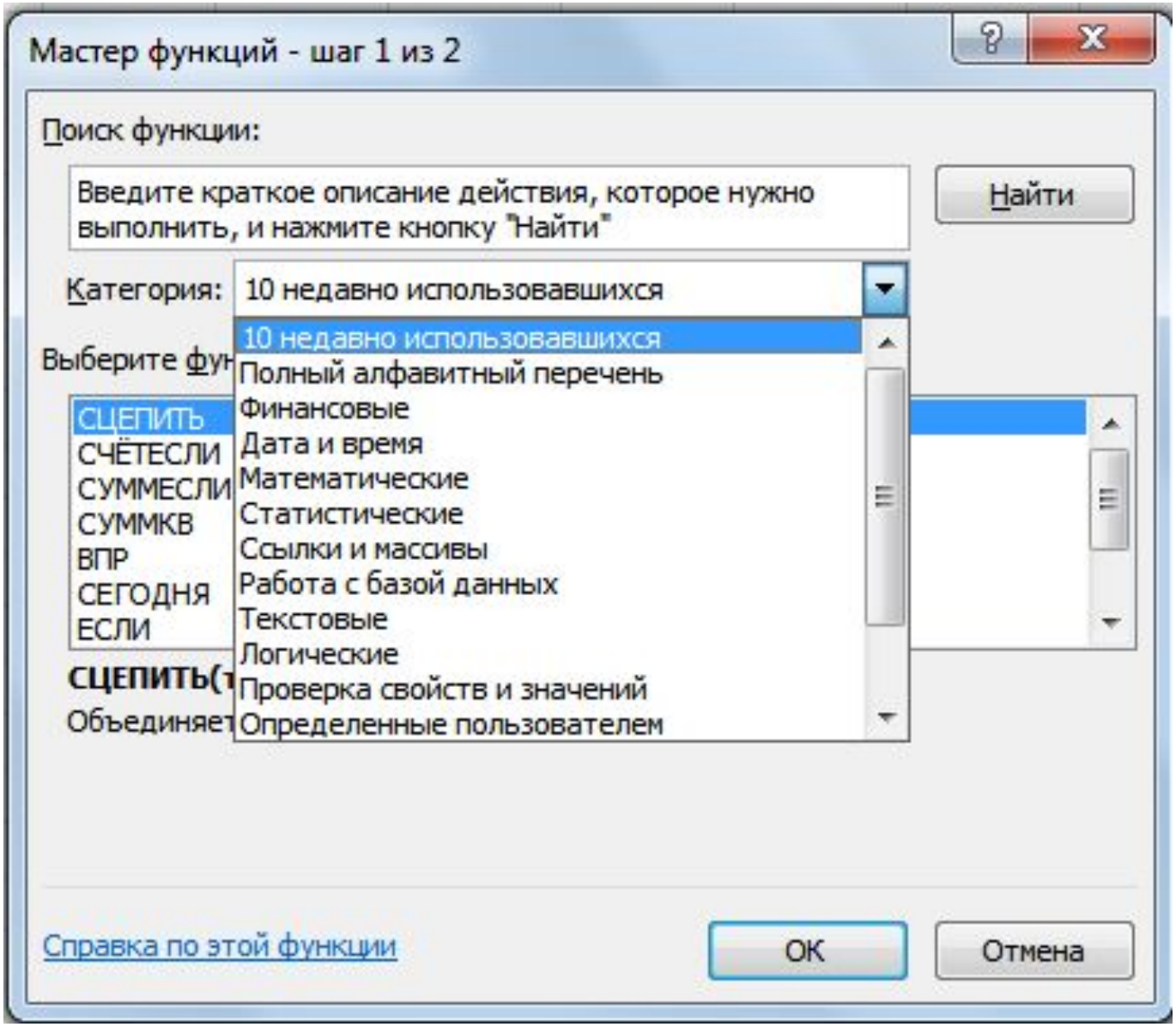
ФУНКЦИИ В EXCEL

Характеристика	**Функция**
Объем выборки	СЧЁТ(массив данных)
Выборочное среднее	СРЗНАЧ(массив данных)
Дисперсия	ДИСПВ(массив данных)
Стандартное отклонение	СТАНДОТКЛОН(массив данных)
Медиана	МЕДИАНА(массив данных)
Мода	МОДА(массив данных)
Коэффициент эксцесса	ЭКСЦЕСС(массив данных)
Коэффициент асимметрии	СКОС(массив данных)
Процентиль	ПРОЦЕНТИЛЬ (массив данных; k)
Квартиль	КВАРТИЛЬ(массив данных;часть)

ФУНКЦИИ В EXCEL

	A	B	C	D	E
1					
2	Выборочные значения				
3	20,3	Программирование			
4	15,3				
5	14,3	*17,907*			
6	19,3		=СУММ(A3:A57)/55		
7	10,1		=КВАДРОТКЛ(A3:A57)/55		
8	13,9	*8,601*			
9	19,5				
10	17,8				
11	15,4	Стандартные функции Excel			
12	16,8				
13	20,1		=СРЗНАЧ(A3:A57)		
14	17,8	*17,907*			
15	21,1	*8,601*			
16	19,8		=ДИСПР(A3:A57)		
17	17,2				
18	13,5				

ФУНКЦИИ В EXCEL



ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

- ***Интервальной оценкой*** параметра θ называется числовой интервал (a,b) который с заданной вероятностью p ***(надежностью)*** покрывает неизвестное значение параметра θ.
- *Величина доверительного интервала зависит от объема выборки(уменьшается с ростом n) и надежности p (увеличивается с ростом p).*
- Такой интервал (a,b) называется ***доверительным***, а вероятность p ***доверительной вероятностью.*** Вместо нее часто задают величину ***$\alpha=1-p$, называемую уровнем значимости.***

p: 0,95; 0,99;0,999 α: 0,05; 0,01;0,001

ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

	F	G	H	I
1		Уровень значимости		0,05
2		Интервал	Левая граница	Правая граница
3		Матожидание		
4		Дисперсия		

=СРЗНАЧ(A1:A25)-ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)
=СРЗНАЧ(A1:A25)+ДОВЕРИТ(I1;СТАНДОТКЛОН(A1:A25);25)

ФУНКЦИИ В EXCEL

- **МИН(Число1;Число2;)** – вычисление наименьшего значения из списка аргументов, логические и текстовые значения игнорируются.
- **МАКС(Число1;Число2;)** – вычисление наибольшего значения из списка аргументов, логические и текстовые значения игнорируются.
- **СЧЁТ(Значение1;Значение2;)** – подсчитывает количество ячеек в диапазоне, которые содержат числа. ***СЧЁТ(70;50;100;«масса») →3***
- **СЧЁТЗ(Значение1;Значение2;)** – подсчитывает количество непустых ячеек в указанном диапазоне.

ФУНКЦИИ В EXCEL

- **СЧЁТЕСЛИ(Диапазон;Критерий)** – подсчитывает количество ячеек в диапазоне, удовлетворяющих заданному условию.
- ***СЧЁТЕСЛИ(B:B; «Грипп")*** – количество ячеек в столбце **B**, содержащих слово **Грипп**.
- ***СЧЁТЕСЛИ(D:D;">13.10.2010")*** – количество ячеек в столбце **D** с датой посещения после **13.10.2010**.
- **СРЗНАЧЕСЛИ(Диапазон;Условие; Диапазон_усреднения)** – подсчитывает среднее арифметическое для ячеек, удовлетворяющих заданному условию.

Функции в EXCEL

- **ЕСЛИ(Лог_выражение;Значение_если_истина; Значение_если_ложь)**

Лог_выражение [Logical_test] – выражение, относительно которого можно судить: истина

это или ложь. Необходимо задать условие, используя ссылки на адреса ячеек: >, >=, <, <=,

<>, =. Можно использовать функции: **И** [AND], **ИЛИ** [OR].

- **СЕГОДНЯ()-**вставка текущей даты в формате даты
- **РАБДЕНЬ(Нач_дата;Число_дней;Праздники)** – определение даты, отстоящей на заданноес число рабочих дней вперед или назад от начальной даты.
- **ЧИСТРАБДНИ(Нач_дата;Кон_дата;Праздники)** – определение полных рабочих дней между двумя указанными датами.
- **ОКРУГЛ(Число;Число_разрядов)** – округляет число до указанного количества десятичных разрядов (по правилам математики).

СПАСИБО ЗА ВНИМАНИЕ.