



## 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:  $\mu$ ,  $\sigma^2$ ,  $\sigma$  и т.д.



# ОСНОВНЫЕ СТАТИСТИКИ

- ▣ *Выборочное среднее  $\bar{x}$  – 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  $Me$ .*
- ▣ *Мода  $Mo$ .*
- ▣ *Коэффициент эксцесса  $E$ .*
- ▣ *Коэффициент асимметрии  $A$ .*
- ▣ *Процентиль.*



# ОСНОВНЫЕ СТАТИСТИКИ

- Корреляция (от лат. *correlatio*) — корреляционная зависимость

В

М

Т

С

О

К

П

З

К

н

К

П

В

Н

Коэ

(отр

	X	Y
1		
2	-0,30023	-
3	0,244257	1
4	1,19835	1
5	-2,18359	-
6	1,095023	-
7	-0,6902	-
8	-1,84691	-
9	-0,77351	-
10	-0,56792	-
11	0,134853	-
12	-0,32699	-
13	1,342642	-
14	-0,18616	-
15	1,972212	0
16	2,375655	-
17		
18		
19		

Вставить

Буфер обмена

Шрифт

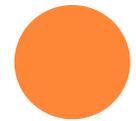
Выравниван

В18

$f_x$  =КОРРЕЛ(A2:A16;B2:B16)

	A	B	C	D	E
1	X	Y			
2	-0,30023	-1,277683168			
3	0,244257	1,27647354			
4	1,19835	1,733133104			
5	-2,18359	-0,234181243			
6	1,095023	-1,086700649			
7	-0,6902	-1,690432327			
8	-1,84691	-0,977629497			
9	-0,77351	-2,117931217			
10	-0,56792	-0,404047569			
11	0,134853	-0,365492951			
12	-0,32699	-0,370240514			
13	1,342642	-0,085284455			
14	-0,18616	-0,513207397			
15	1,972212	0,865672973			
16	2,375655	-0,654906671			
17					
18		0,404704057			
19					

от -1



# ОСНОВНЫЕ СТАТИСТИКИ

Параметры Excel

Книга1 - Microsoft Excel

Главная Меню Вставка Разметка страницы Формулы **Данные** Рецензирование Вид Office Tab Настройки

Подключения: Подключения, Свойства, Изменить связи

Сортировка и фильтр: Сортировка, Фильтр, Дополнительно

Работа с данными: Текст по столбцам, Удалить дубликаты, Проверка данных, Консолидация, Анализ "что если"

Структура: Группировать, Разгруппировать, Промежуточный итог

Анализ: **Анализ данных**, Поиск решения

А1 fx x

Книга1 \* x

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
x	y															
123	67															
12	1															
12	1															

Анализ данных

Инструменты анализа

- Однофакторный дисперсионный анализ
- Двухфакторный дисперсионный анализ с повторениями
- Двухфакторный дисперсионный анализ без повторений
- Корреляция
- Ковариация
- Описательная статистика
- Экспоненциальное сглаживание
- Двухвыборочный F-тест для дисперсии
- Анализ Фурье
- Гистограмма**

OK Отмена Справка

# СТАТИСТИКА В EXCEL

получение внешних данных | Обновить все | Изменить связи | Подключения | Сортировка и фильтр | Фильтр | Дополнительно | текст по столбцам | удалить дубликаты | Работа

A1 | fx | -1,50116079566942

Книга1 \* x

	A	B	C	D	E	F	G	H	I	J
1	-1,50116									
2	-6,3884									
3	1,22128									
4	6,38236									
5	5,99175									
6	8,66566									
7	-10,917									
8	-1,1709									
9	5,47511									
10	-5,433									
11	-3,4510									
12	-8,45216									
13	-9,23455									
14	-4,88815									
15	-3,86754									
16	-10,5897									
17	-2,83962									
18	-2,02024									
19	0,674265									
20	-1,82746									

Анализ данных | Диаграмма 1 | fx

Инструменты

- Однофакторн
- Двухфакторн
- Двухфакторн
- Корреляция
- Ковариация
- Описательная
- Экспоненциал
- Двухвыбороч
- Анализ Фурье
- Гистограмма**

	A	B	C	D	E	F	G	H	I
	Карман	Частота							
	-10,9179	1							
	-6,02204	4							
	-1,12614	9							
	3,769765	2							
	Еще	4							

Гистограмма

Частота

Частота

Карман

Еще

3,769764...



# ФУНКЦИИ В EXCEL

	A	B	C	D	E	F	G
1							
2	<b>Выборочные значения</b>	<b>Границы интервалов</b>	<b>Частоты</b>	<b>Частности</b>			
3	20,3	10	0	0			
4	15,3	12	2	2/55			
5	14,3	14	4	4/55	=C4/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 variability.
  - Оценка variability 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  $E$**  - характеризует «островерхость» гистограммы или полигона по сравнению с кривой Гаусса нормального распределения.
- **Коэффициент асимметрии  $A$**  - 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

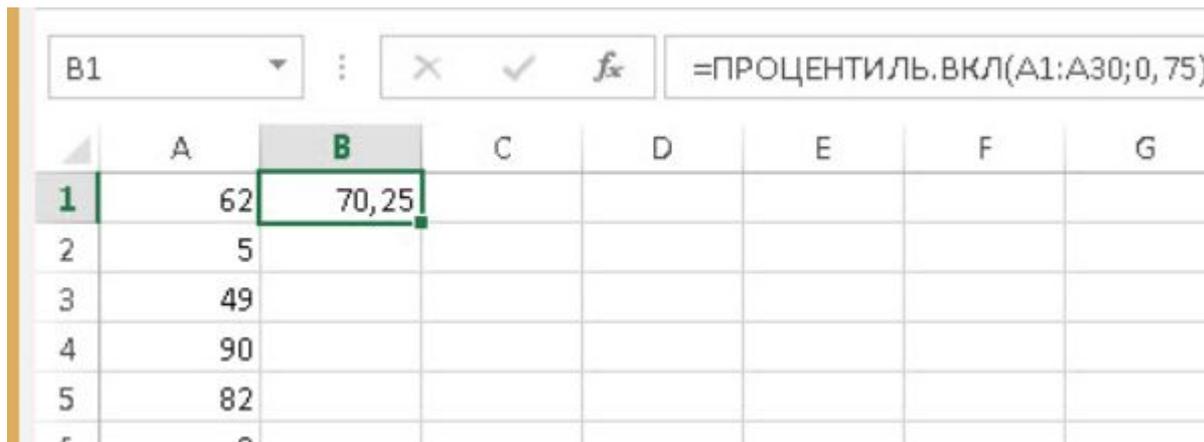
- **Квантилью**  $x_p$  ( $p$ -квантилью, квантилью уровня  $p$ ) случайной величины, имеющей функцию распределения  $F_x(x)$ , называют решение  $x_p$  уравнения  $F_x(x) = p$ . Для некоторых  $p$  уравнение  $F_x(x) = p$  может иметь **несколько решений**, для некоторых - **ни одного**.

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



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
1	62	70,25					
2	5						
3	49						
4	90						
5	82						

The formula bar at the top shows the formula: `=ПРОЦЕНТИЛЬ.ВКЛ(A1:A30;0,75)`



# 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.

A3	:	  	=МЕДИАНА(A1:N1)												
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	1	1	9	2	5	7	9	4	6	8	8	3	7	1	
2															
3			5,5												
4															



# 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)

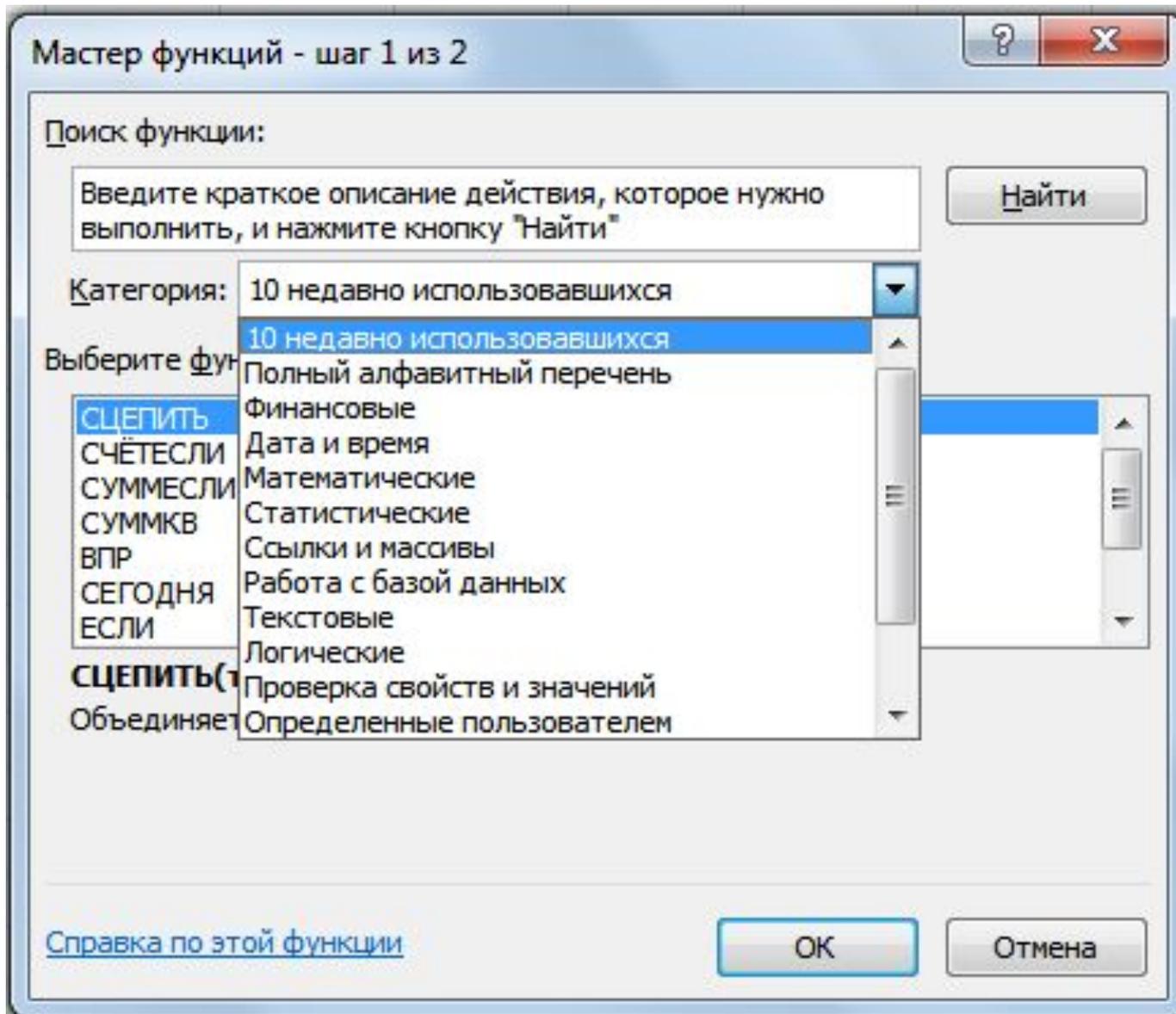


# ФУНКЦИИ В EXCEL

	А	В	С	Д	Е
1					
2	Выборочные значения				
3	20,3	Программирование			
4	15,3				
5	14,3	<b>17,907</b>	←		
6	19,3		=СУММ(А3:А57)/55		
7	10,1		=КВАДРОТКЛ(А3:А57)/55		
8	13,9	<b>8,601</b>	←		
9	19,5				
10	17,8				
11	15,4	Стандартные функции Excel			
12	16,8				
13	20,1		=СРЗНАЧ(А3:А57)		
14	17,8	<b>17,907</b>	←		
15	21,1	<b>8,601</b>	←		
16	19,8		=ДИСПР(А3:А57)		
17	17,2				
18	13,5				



# ФУНКЦИИ В EXCEL



# ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

- *Интервальной оценкой* параметра  $\theta$  называется числовой интервал  $(a, b)$  который с заданной вероятностью  $p$  (*надежностью*) покрывает неизвестное значение параметра  $\theta$ .
- *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

$\alpha$ : 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			
5					

**=СРЗНАЧ(А1:А25)-ДОВЕРИТ(І1;СТАНДОТКЛОН(А1:А25);25)**  
**=СРЗНАЧ(А1:А25)+ДОВЕРИТ(І1;СТАНДОТКЛОН(А1:А25);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(В:В; «Грипп»)** – количество ячеек в столбце **В**, содержащих слово **Грипп**.
- ▣ **СЧЁТЕСЛИ(Д:Д; ">13.10.2010")** – количество ячеек в столбце **Д** 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 заданное 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.**

