Пакет «Анализ данных» в Excel

1. Гистограмма

Для этого выполним последовательность действий: Пакет Анализа → Генерация случайных чисел → в открывшемся диалоговом окне установим параметры для генерации выборки из нормального распределения (рис. 1). Параметр Случайное рассеивание устанавливать не нужно, он предназначен для возможности повторить генерацию тех же самых значений.

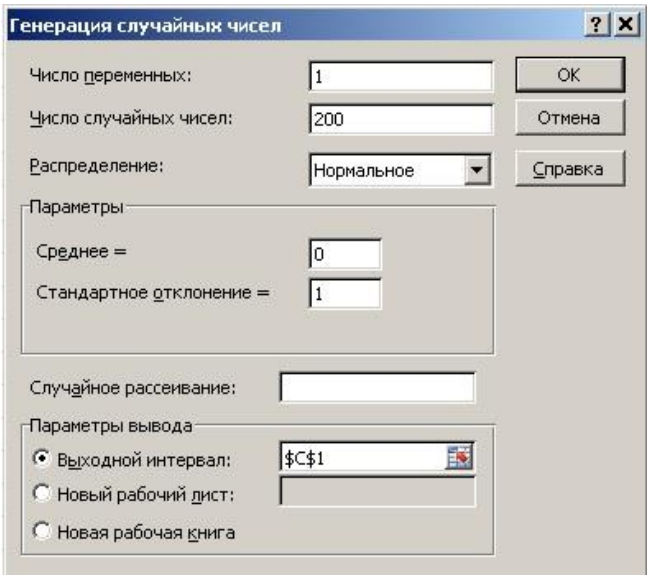


Рис 1. Параметры для генерации нормального распределения

Для полученных данных выполним **Описательную статистку**. Эта процедура уже была описана выше.

Для построения гистограммы выполним **Пакет Анализа → Гистограмма →** в открывшемся диалоговом окне установим параметры, задав **Входной интервал** и установив **Вывод графика**. Результаты разместим на **Новый рабочий лист** (рис. 2). Интервал карманов можно не устанавливать, в этом случае он сформируется автоматически.

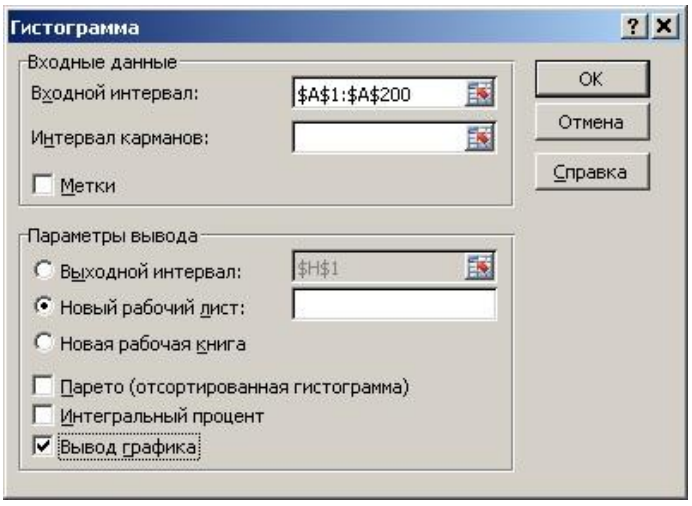


Рис 2. Параметры для построения гистограммы

В результате, на новый рабочий будут помещены таблица и гистограмма (рис. 3) для сгенерированного нормального распределения. Первый столбец таблицы задает границы карманов, а второй частоту, то есть количество элементов выборки, попавших в указанный карман.

Карман	Частота
-3,297	1
-2,909	0
-2,521	2
-2,133	4
-1,745	5
-1,357	8
-0,969	14
-0,580	21
-0,192	24
0,196	33
0,584	28
0,972	23
1,360	20
1,748	10
Еще	7

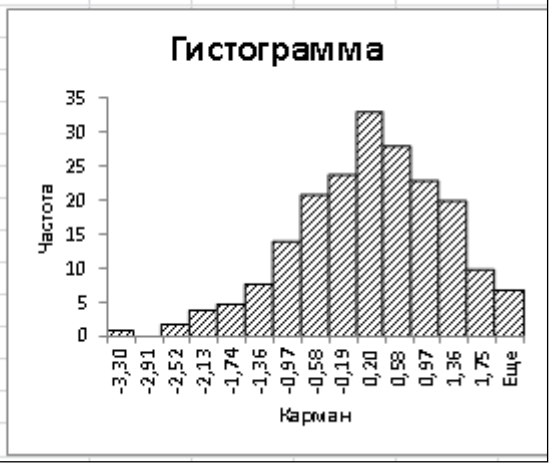


Рис 3. Таблица частот и гистограмма для сгенерированного нормального распределения

Гистограмма — это столбиковая диаграмма для отображения распределения частот по диапазонам значений переменной. Горизонтальная ось соответствует значениям переменной, а вертикальная частотам. Построенную гистограмму называют также гистограммой частот, в отличии от гистограммы относительных частот.

Аналогичные действия выполним для генерации равномерного распределения. На рис. 4 изображены таблица частот и гистограмма для сгенерированного равномерного распределения.

Карман	Частота
0,00	1
0,07	8
0,14	15
0,21	16
0,28	13
0,35	19
0,42	12
0,49	14
0,55	16
0,62	18
0,69	18
0,76	8
0,83	17
0,90	14
Еще	11

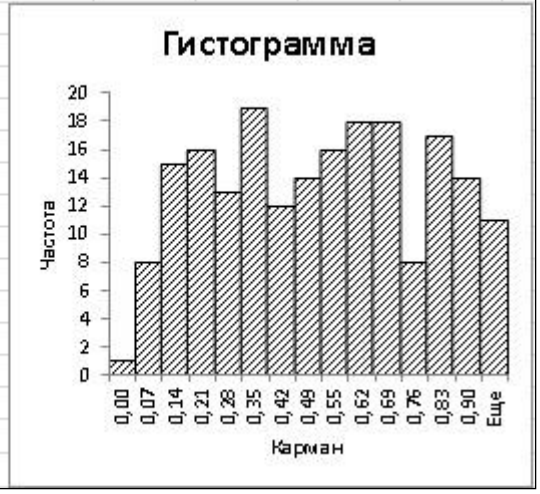


Рис 4. Таблица частот и гистограмма для сгенерированного равномерного распределения

Возьмем еще один произвольный набор данных, тип распределения которых неизвестен, например, данные о выручке магазина за три месяца. Для него тоже посчитаем описательные статистики и постоим гистограмму. На рис. 5 изображены таблица частот и гистограмма, полученные для данных о выручке магазина.

Карман	Частота
17,9	1
21,9	5
25,8	12
29,8	23
33,8	24
37,7	11
41,7	9
45,7	3
49,6	1
Еще	1

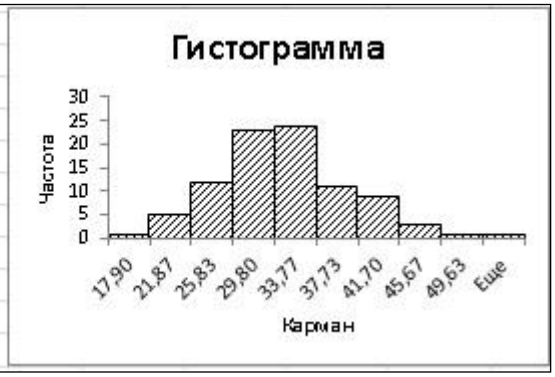


Рис 5. Таблица частот и гистограмма для данных о выручке магазина

Сравним описательные статистики для рассмотренных распределений (рис. 6). В столбце 1 расположены данные для выборки из нормального распределения, в столбце 2 — для выборки из равномерного, и в столбце 3 — для выборки с неизвестным распределением.

	A	B	D	E	G	H
1	*Столбец 1*		*Столбец 2*		*Столбец 3*	
2						
3	Среднее	0,002	Среднее	0,487	Среднее	30,68
4	Стандартная ошибка	0,073	Стандартная ошибка	0,019	Стандартная ошибка	0,71
5	Медиана	0,107	Медиана	0,501	Медиана	30,25
6	Мода	#Н/Д	Мода	#Н/Д	Мода	26,10
7	Стандартное отклон	1,036	Стандартное отклон	0,263	Стандартное отклон	6,73
8	Дисперсия выборки	1,072	Дисперсия выборки	0,069	Дисперсия выборки	45,30
9	Эксцесс	-0,001	Эксцесс	-1,107	Эксцесс	0,82
10	Асимметричность	-0,412	Асимметричность	0,010	Асимметричность	0,65
11	Интервал	5,433	Интервал	0,964	Интервал	35,70
12	Минимум	-3,297	Минимум	0,004	Минимум	17,90
13	Максимум	2,136	Максимум	0,968	Максимум	53,60
14	Сумма	0,320	Сумма	97,357	Сумма	2761,10
15	Счет	200	Счет	200	Счет	90

Рис 6. Описательная статистика для различных выборок

Предположим, нам ничего не известно о типе этих распределений, есть только описательные статистики и гистограммы. Сможем ли мы сделать некоторые выводы на основе этих данных. Какое из распределений отнести к нормальному?

По виду гистограммы можно сразу исключить второй случай. Оставшиеся два имеют небольшую асимметрию, причем для первого набора она отрицательна, а для второго положительна.

Эксцесс первого распределения довольно близок к нулю, что дает основания предположить, что данное распределение близко к нормальному. Однако это предположение следует проверить, используя критерии согласия.

Что касается третьего распределения, оно является островершинным, при этом среднее и медиана близки по значению. По виду гистограммы можно сказать, что оно унимодальное, рассчитанное описательное значение моды не имеет содержательного смысла, так как исходные данные являются непрерывными. Данное распределение может быть близким к нормальному, однако это обязательно нужно проверить, используя различные статистические методы.

2. Доверительные интервалы

Описательная статистика надстройки **Анализ данных** вычисляет доверительные интервалы. Для расчета доверительных интервалов в Excel есть еще три встроенные функции, относящиеся к разделу **Статистические**. Функция **ДОВЕРИТ.СТЬЮДЕНТ** возвращает доверительный интервал для среднего генеральной совокупности, используя распределение Стьюдента, функция **ДОВЕРИТ.НОРМ** возвращает доверительный интервал для среднего генеральной совокупности, используя нормальное распределение, функция **ДОВЕРИТ** является аналогом функции **ДОВЕРИТ.НОРМ** и сохранена в пакетах последних версий для совместимости с более ранними версиями.

Все эти функции имеют три обязательных аргумента (**альфа; станд_откл; размер**):

- **альфа** задает уровень значимости, используемый при вычислении доверительного интервала. Он связан с **Уровнем надежности** в **Описательной статистике** формулой **Уровень надежности = (1 – альфа)*100%.** Например, значение аргумента **альфа** равное 0,05, означает 95-процентный **Уровень надежности**, который связан с доверительной

вероятностью формулой **Уровень надежности =** Доверительная вероятность***100%**;

- **станд_откл** — выборочное стандартное отклонение для диапазона данных;
- **размер** — объем выборки.

Описательная статистика надстройки **Анализ данных** при вычислении доверительных интервалов использует встроенную функцию **ДОВЕРИТ.СТЬЮДЕНТ**.

Для рассмотренного выше примера, **Описательная статистика** выдает значение 1,7542149. Сравним: ДОВЕРИТ.СТЬЮДЕНТ(0,05;N8;N16)= 1,7542149.

3. Примеры применения описательной статистики

Пример 2. В нашем распоряжении имеются выборочные данные о температуре воздуха, собранные на некоторой опытной станции за три года на небольшом острове. Нас интересует климат этого острова и насколько он комфортен для проживания. Исходные данные приведены в табл. 1.

Таблица 1

Выборка

10,1	-4,3	-2,2	10,6	8,9	9,4	16,6	7,3	4,7	-2,7
15,2	17,1	-4,5	8,2	15,2	6,2	19,7	11,5	1,3	15,5
4,5	8,8	9,8	13,8	4,0	4,7	4,7	8,1	-3,1	14,8
13,5	20,1	12,0	7,4	8,3	6,7	-0,6	18,5	7,0	16,5
6,5	9,4	2,6	3,3	6,8	4,7	7,9	-1,0	6,6	17,2
-1,1	16,5	-9,3	2,7	10,6	9,5	11,5	-2,0	2,3	7,8
-3,0	12,3	5,7	0,5	3,8	5,0	2,3	7,9	20,1	-7,5
7,1	4,9	9,3	8,7	7,6	18,1	1,7	20,6	11,3	15,7
8,0	14,3	15,8	3,4	16,5	2,6	13,4	9,6	15,5	6,7
-2,8	-1,6	9,5	19,2	16,2	3,3	8,2	9,4	1,0	13,0

Для проведения расчетов данные необходимо поместить в один столбец. Пусть она занимают диапазон **A1:A100**. Объем выборки равен n=100.

Чтобы получить некоторые предварительные данные об изучаемой величине, воспользуемся описательной статистикой **Пакета анализа** (рис. 7). Исходный массив был отсортирован в порядке возрастания.

	A	B	C	D
1	-9,3		*Столбец1*	
2	-7,5			
3	-4,5		Среднее	7,89
4	-4,3		Стандартная	0,67
5	-3,1		Медиана	7,95
6	-3,0		Мода	4,70
7	-2,8		Стандартное	6,73
8	-2,7		Дисперсия вы	45,35
9	-2,2		Эксцесс	-0,46
10	-2,0		Асимметричн	-0,17
11	-1,6		Интервал	29,90
12	-1,1		Минимум	-9,30
13	-1,0		Максимум	20,60
14	-0,6		Сумма	789,10
15	0,5		Счет	100,00

Рис 7. Описательная статистика

Средняя температура не радует, меньше чем в Сочи, где среднегодовая температура по поверхности России равна +14.2 °C , но больше чем в Оймяконе, где всего −15.5 °C.

Построим гистограмму. Воспользуемся для этого средством **Гистограмма.**

Интервалы карманов желательно предварительно вычислить. Для этого разбивают исходный ряд на интервалы. Обычно количество интервалов *k* задают, используя формулу Стерджесса $k = [1 + 3{,}321 \cdot \lg n]$, здесь $\lg n$ - десятичный логарифм, n — объем выборки. Квадратные скобки означают целую часть от числа, то есть округление до ближайшего целого в меньшую сторону. В Excel для вычисления *k* можно использовать формулу =ЦЕЛОЕ(1+3,321*LOG10(*n*)). При *n*=100 оптимальное число интервалов равно7.Длина интервалов вычисляется по формуле $h = (x_{max} - x_{min})/k$, где x_{max} — максимальное значение выборки, а x_{min} — минимальное, h=4,27. Для удобства округлим значение длины интервала, примем h=4,3.

Для построения интервального ряда воспользуемся инструментом **Гистограмма** из **Пакета анализа**, который посчитает частоты, и построит графические характеристики распределения — гистограмму и кумуляту. Подготовим для этого интервал карманов, диапазон **I4:I9**, формулы для расчетов приведены на рис. 16. Интервал карманов отмечен серым фоном. Результаты вычислений приведены на рис. 17. Здесь мы воспользовались некоторыми данными описательной статистики.

	F	G	H	I	J
1					
2	Интервал	=D11		Начало	Конец
3	Минимум	=D12		=G3	=I3+$G$7
4	Максимум	=D13		=I3+$G$7	=I4+$G$7
5	Счет	=D15		=I4+$G$7	=I5+$G$7
6	Кол. инт.	=ЦЕЛОЕ(1+3,321*LOG10(G5))		=I5+$G$7	=I6+$G$7
7	Длина	=ОКРУГЛ(G2/G6;1)		=I6+$G$7	=I7+$G$7
8				=I7+$G$7	=I8+$G$7
9				=I8+$G$7	=I9+$G$7

Рис 16. Формулы для расчета интервала карманов

	F	G	H	I	J	K	L
1							
2	Интервал	29,90		Начало	Конец	Середина	Частота
3	Минимум	-9,30		-9,30	-5,00	-7,15	2
4	Максимум	20,60		-5,00	-0,70	-2,85	11
5	Счет	100,00		-0,70	3,60	1,45	13
6	Кол. инт.	7		3,60	7,90	5,75	22
7	Длина интервала	4,30		7,90	12,20	10,05	25
8				12,20	16,50	14,35	17
9				16,50	20,80	18,65	10

Рис 17. Результаты расчетов интервала карманов

Выполним последовательность действий: **Данные → Анализ данных → Гистограмма**. В открывшемся диалогом окне укажем входной интервал, интервал карманов, установим галочки **Интегральный процент** и **Вывод графика**, для построения гистограммы и кумуляты. Выберем вывод результатов на новый рабочий лист. Установка параметров показана на рис. 18.

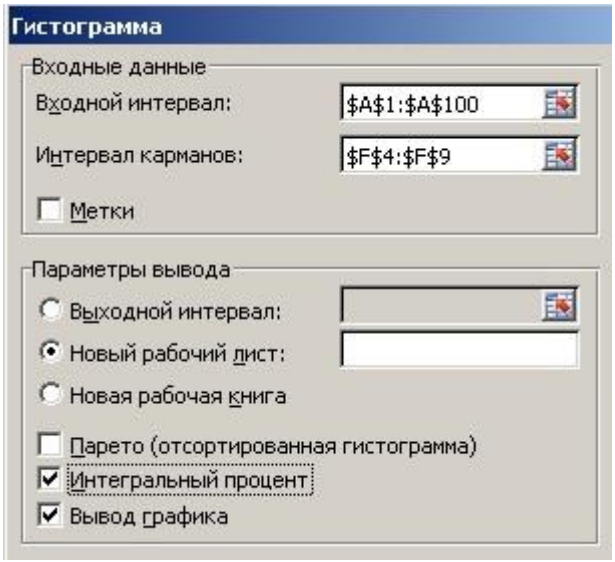


*Рис 18. Диалоговое окно инструмента **Гистограмма***

В результате на новый рабочий лист выведется таблица расчета частот (рис. 19) и графики (рис. 20).

	A	B	C
1	*Карман*	*Частота*	*Интегральный %*
2	-5,00	2	2,00%
3	-0,70	11	13,00%
4	3,60	13	26,00%
5	7,90	22	48,00%
6	12,20	25	73,00%
7	16,50	17	90,00%
8	Еще	10	100,00%

Рис 19. Результаты расчета частот

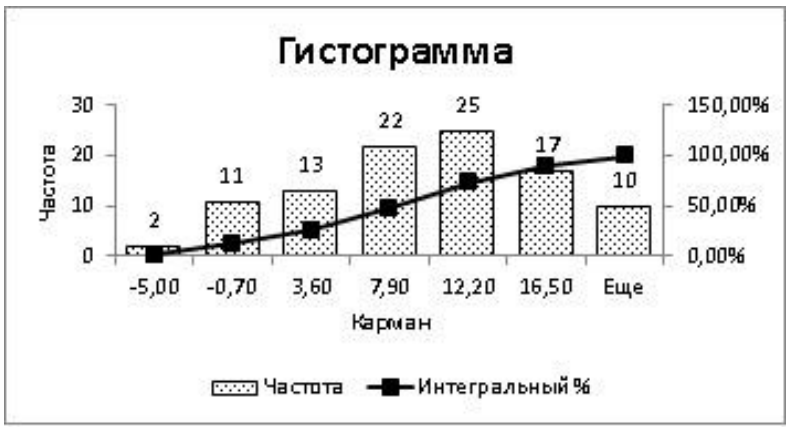


Рис 20. Гистограмма и кумулята

Карман указывает верхнюю границу интервала. Интегральный процент отображает процентное содержание накопленных частот, которые используются для построения кумуляты. Графики, приведенные на рис.20 немного отредактированы, в частности, добавлены для наглядности подписи данных и значения частот.

На основании полученных расчетов можно построить интервальный ряд. *Интервальным* статистическим рядом называется упорядоченная совокупность интервалов варьирования значений случайной величины с соответствующими частотами попаданий в каждый из них значений величины.

На рис.21 приведена таблица, в которой указаны границы интервалов, середины интервалов и частоты. На этом же рисунке приведен полученный интервальный ряд, где x_i — значения, а n_i — частоты интервального ряда.

I	J	K	L	M	N	O	P
Начало	Конец	Середина	Частота				
-9,30	-5,00	-7,15	2				
-5,00	-0,70	-2,85	11				
-0,70	3,60	1,45	13				
3,60	7,90	5,75	22				
7,90	12,20	10,05	25				
12,20	16,50	14,35	17				
16,50	20,80	18,65	10				
xi	-7,15	-2,85	1,45	5,75	10,05	14,35	18,65
ni	2	11	13	22	25	17	10

Рис 21. Интервальный ряд

Таким образом, исходный диапазон выборки был разбит на интервалы, эти интервалы являются полуоткрытыми. Все выборочные значения, попавшие в интервал, заменяются серединой интервала. Полученный интервальный ряд может понадобиться для дальнейшего статистического анализа.

Пример3. На рис.22 представлены данные о количестве голов, забитых в матчах чемпионата России по футболу в 2013/2014. Проведем анализ, полученных данных.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		**АМК**	**АНЖ**	**ВОЛ**	**ДИН**	**ЗЕН**	**КРА**	**КРЫ**	**КУБ**	**ЛОК**	**РОС**	**РУБ**	**СПА**	**ТЕР**	**ТОМ**	**УРА**	**ЦСК**
2	**АМК**		1	6	3	3	4	0	4	0	1	0	3	1	2	2	4
3	**АНЖ**	4		0	4	3	3	1	0	4	1	1	1	3	2	1	3
4	**ВОЛ**	2	3		5	4	1	3	1	3	3	3	1	1	1	3	3
5	**ДИН**	2	3	4		2	3	2	4	4	2	0	5	1	1	3	6
6	**ЗЕН**	2	3	2	3		5	3	2	3	2	8	6	2	0	3	2
7	**КРА**	3	1	3	2	3		2	3	4	2	1	4	5	4	1	1
8	**КРЫ**	4	2	4	3	5	1		0	4	2	4	3	2	1	2	4
9	**КУБ**	3	2	4	2	5	4	4		4	4	2	4	4	2	5	4
10	**ЛОК**	4	0	3	1	2	4	3	1		5	0	0	3	0	3	3
11	**РОС**	6	2	4	5	4	4	3	0	2		0	1	3	3	2	0
12	**РУБ**	3	6	4	4	3	1	2	2	3	3		3	2	3	1	0
13	**СПА**	1	4	7	5	6	5	1	2	4	2	0		0	3	1	3
14	**ТЕР**	2	2	2	1	2	1	1	3	1	3	0	1		2	2	2
15	**ТОМ**	0	4	1	4	3	2	2	3	2	5	1	3	0		3	3
16	**УРА**	0	3	3	5	3	2	2	3	3	5	3	2	3	0		4
17	**ЦСК**	3	0	3	2	1	6	3	1	1	1	3	1	5	2	1	

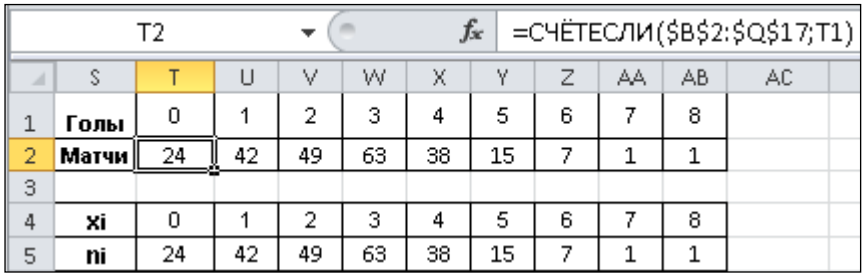
Рис. 22. Таблица голов

Объем данных n в данном случае равен количеству проведенных матчей, n=240. Изучаемая переменная X — количество голов забитых в матче. В данном примере, исследуемая переменная принимает небольшое количество различных значений. Чисто визуально можно определить, что наименьшее количество голов равно нулю, а наибольшее восьми. Применять здесь описательную статистику пакета анализа

не очень удобно из-за большого объема. Данные можно сгруппировать и построить дискретный статистический ряд.

Для построения статистического ряда необходимо указать значения, которые принимает переменная X, и посчитать частоты. Так как исходный дискретный ряд принимает конечное количество значений, для нахождения частоты повторяемости признака, воспользуемся функцией **СЧЁТЕСЛИ(диапазон;критерий)**, которая подсчитывает количество непустых ячеек в указанном диапазоне, удовлетворяющих заданному критерию.

Подготовим таблицу для расчета частот (рис. 23). В строке 1 поместим количество забитых голов в порядке возрастания. В ячейку **T2** введем формулу, которая отобразилась в строке формул. Напомним, что для ввода ссылок на диапазоны в формулах пользуйтесь мышью. Первый аргумент функции **СЧЁТЕСЛИ**, диапазон **$B$2:$Q$17** — ссылка на исходную таблицу данных. Второй аргумент функции ссылается на ячейку с критерием. Для диапазона используется абсолютная адресация, тогда эту формулу можно будет поместить в соседние ячейки копированием, или «протягиванием». Таким образом, получен сгруппированный ряд, в первой строке указано количество голов, а во втором — соответствующее количество матчей. Обозначим x_i — значение переменной, или варианта, а n_i — частоту встречаемости этой переменной.



	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	**Голы**	0	1	2	3	4	5	6	7	8	
2	**Матчи**	24	42	49	63	38	15	7	1	1	
3											
4	**xi**	0	1	2	3	4	5	6	7	8	
5	**ni**	24	42	49	63	38	15	7	1	1	

Рис. 23. Построение дискретного ряда

Добавим **pi** — относительные частоты, получим дискретный статистический ряд (табл. 2).

Таблица 2

Дискретный статистический ряд

xi	0	1	2	3	4	5	6	7	8
ni	24	42	49	63	38	15	7	1	1
pi	0,100	0,175	0,204	0,263	0,158	0,063	0,029	0,004	0,004

Относительные частоты вычислены как отношение частоты к объему выборки: $p_i = n_i / n = n_i / 240$. Заметим, что $\sum n_i = 240$, $\sum p_i = 1$.

Мода найдется как значение, имеющую наибольшую частоту, в данном случае 3, то есть чаще всего в матчах было забито три гола. Так как $n = 240$, четное число, медиана равна полусумме значений, стоящих в вариационном ряду на 120 и 121 местах.

Расчеты приведены на рис. 24, а формулы для расчетов на рис. 25.

	A	B	C	D	E	F	G	H
1	xi	ni	pi	xi × ni	xi-xcp	(xi-xcp)^2	ni (xi-xcp)^2	
2	0	24	0,100	0	0	0	0	
3	1	42	0,175	42	1	1	42	
4	2	49	0,204	98	2	4	196	
5	3	63	0,263	189	3	9	567	
6	4	38	0,158	152	4	16	608	
7	5	15	0,063	75	5	25	375	
8	6	7	0,029	42	6	36	252	
9	7	1	0,004	7	7	49	49	
10	8	1	0,004	8	8	64	64	
11	**36**	**240**	**1**	**613**	**36**	**204**	**2153**	Сумма
12								
13						Среднее	2,5541667	
14						Дисперсия	8,9708333	
15						СКВО	2,9951349	
16						Мода	3	
17						Медиана	3	

Рис 24. Результаты расчетов основных числовых характеристик

	A	B	C	D	E	F	G	H
1	xi	ni	pi	xi × ni	xi-xcp	(xi-xcp)^2	ni (xi-xcp)^2	
2	0	24	=B2/$B$11	=A2*B2	=A2-$A$12	=E2*E2	=B2*F2	
3	1	42	=B3/$B$11	=A3*B3	=A3-$A$12	=E3*E3	=B3*F3	
4	2	49	=B4/$B$11	=A4*B4	=A4-$A$12	=E4*E4	=B4*F4	
5	3	63	=B5/$B$11	=A5*B5	=A5-$A$12	=E5*E5	=B5*F5	
6	4	38	=B6/$B$11	=A6*B6	=A6-$A$12	=E6*E6	=B6*F6	
7	5	15	=B7/$B$11	=A7*B7	=A7-$A$12	=E7*E7	=B7*F7	
8	6	7	=B8/$B$11	=A8*B8	=A8-$A$12	=E8*E8	=B8*F8	
9	7	1	=B9/$B$11	=A9*B9	=A9-$A$12	=E9*E9	=B9*F9	
10	8	1	=B10/$B$11	=A10*B10	=A10-$A$12	=E10*E10	=B10*F10	
11	**=СУММ**	**=СУММ**	**=СУММ(C2:C10)**	**=СУММ(D2:D10)**	**=СУММ(E2:E**	**=СУММ(F2:F10)**	**=СУММ(G2:G10)**	Сумма
12								
13						Среднее	=D11/B11	
14						Дисперсия	=G11/B11	
15						СКВО	=КОРЕНЬ(G14)	
16						Мода	3	
17						Медиана	3	

Рис 25 Формулы для расчетов основных числовых характеристик

В среднем за матч забивалось 2-3 гола. Так как медиана равна 3, то в среднем в половине матчей было забито не менее трех голов, а в половине не более трех.

Наглядность распределению значений данных придают графические характеристики. Для дискретных рядов строят полигон частот или относительных частот. Для сгруппированного дискретного ряда строят также гистограмму частот или относительных частот.

Полигон относительных частот, который называют также многоугольником распределения, представляет собой замкнутую ломаную линию, соединяющую точки с координатами (x_i, p_i). Для построения полигона обычно используют **График** с маркерами. Выделите диапазон с относительными частотами **C2:C10 → Вставка → Диаграмма → График с маркерами**. Построенную диаграмму нужно подкорректировать. Измените подписи данных горизонтальной оси, для этого выделите подписи оси мышью → **Конструктор → Выбрать данные → Изменить подписи горизонтальной оси** → введите диапазон **A2:A10**, в котором находятся значения **xi**. Поменяйте параметры оси, выбрав **Положение оси: по делениям**. Добавьте название диаграмме и удалите легенду (рис. 26).

Рис 26. График полигона относительных частот

Многоугольник распределения является одной из форм закона распределения. Если отбросить две последние точки, которые являются выбросами, можно сказать, что значения симметрично распределяются около среднего значения.

Еще одна графическая характеристика — гистограмма распределения. Построенная на частотах (рис. 27) или относительных частотах сгруппированного дискретного или интервального статистического ряда она дает наглядное представление о типе распределения.

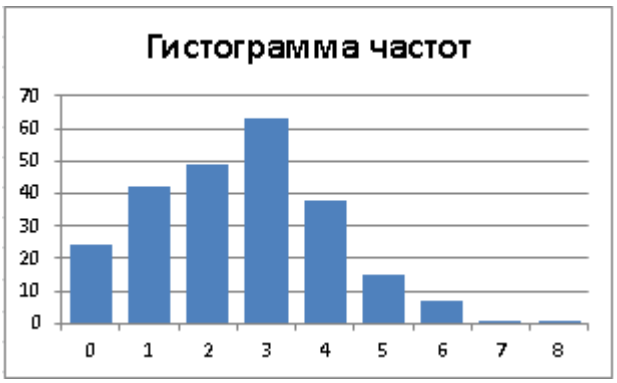


Рис 27. Гистограмма частот

Пример3. На рис. 28 приведены выборочные данные о росте футболистов российских клубов уже предварительно отсортированные по возрастанию исследуемой переменной. На рисунке они размещены в прямоугольном диапазоне только для наглядности.

	A	B	C	D	E
1	167,8	173,0	177,1	180,3	186,0
2	168,2	173,2	177,5	181,2	186,1
3	168,5	173,7	178,1	181,3	186,9
4	168,8	173,9	178,2	181,5	187,5
5	169,1	174,2	178,8	181,7	187,9
6	170,0	174,7	179,0	182,8	189,6
7	171,8	174,9	179,3	183,2	190,0
8	172,4	175,2	179,7	184,4	190,4
9	172,8	176,3	179,9	184,9	190,5
10	172,9	176,5	179,9	185,9	191,1

Рис. 28. Выборка

Если исследуется непрерывный ряд данных, в этом случае значения изучаемой переменной X могут отличаться друг от друга на сколь угодно малую величину. Если к тому же объем данных велик, то сначала строится интервальный статистический, или вариационный, ряд. Затем уже его подвергают статистическому анализу. Непрерывные величины, обычно имеют единицу измерения: времени, длины, массы, объема и т.д.

Объем выборки n=50. Изучаемая переменная X — рост футболистов российских клубов. Исследуемая переменная является непрерывной. По этим данным можно построить интервальный статистический ряд.

Обычно исходные данные не упорядочены. Чтобы получить ранжированный ряд в Excel, необходимо сначала, обязательно, поместить все данные в один столбец и отсортировать.

Разобьем исходный ряд на интервалы. Зададим количество интервалов *k*, используя формулу Стерджесса. При *n*=50 оптимальное число интервалов равно 6.

Для расчета частот есть две альтернативы: воспользоваться встроенной функцией **ЧАСТОТА** или средством **Гистограмма** в **Пакете анализа**. Рассмотрим здесь первый вариант, так как второй вариант был рассмотрен выше.

Функция **ЧАСТОТА(массив_данных;массив_интервалов)** вычисляет распределение значений по интервалам и возвращает вертикальный массив чисел, содержащий на один элемент больше, чем массив интервалов. Для использования этой функции нужно подготовить массив интервалов.

Разместим исходные данные в столбце **A**, это будет диапазон **A2:A51.** Если данные не упорядочены, выполним сортировку столбца **A** по возрастанию. На рис. 4.29 отображены первые 10 значений исходных данных.

Сформируем массив интервалов в столбце **B** следующим образом. В первую ячейку массива интервалов **B2** поместим минимальное выборочное значение, в ячейку **B3** введем формулу и протянем ее вниз, сформировав таким образом границы интервалов. Формулы приведены на рис.4.30.

	A	B	C	D	E	F	G	H
1	**Массив данных**	**Массив интервалов**	**Массив частот**					
2	167,80	167,80	1		≤167,80		k=	6
3	168,20	171,68	5		(167,80 ; 171,68]		h=	3,88
4	168,50	175,57	12		(171,68 ; 175,57]			
5	168,80	179,45	9		(175,57 ; 179,45]			
6	169,10	183,33	10		(179,45 ; 188,33]			
7	170,00	187,22	6		(188,33 ; 187,22]			
8	171,80	191,10	6		(187,22 ; 191,10]			
9	172,40		1		> 191,10			
10	172,80							

Рис. 29. Расчет частот

	A	B	C	D	F	G	H
1	**Массив данных**	**Массив интервалов**	**Массив частот**				
2	167,8	=МИН(A2:A51)	=ЧАСТОТА(A2:A51;B2:B8)			k=	=ЦЕЛОЕ(1+3,321*LOG10(50))
3	168,2	=B2+$H$3	=ЧАСТОТА(A2:A51;B2:B8)			h=	=(МАКС(A2:A51)-МИН(A2:A51))/H2
4	168,5	=B3+$H$3	=ЧАСТОТА(A2:A51;B2:B8)				
5	168,8	=B4+$H$3	=ЧАСТОТА(A2:A51;B2:B8)				
6	169,1	=B5+$H$3	=ЧАСТОТА(A2:A51;B2:B8)				
7	170	=B6+$H$3	=ЧАСТОТА(A2:A51;B2:B8)				
8	171,8	=B7+$H$3	=ЧАСТОТА(A2:A51;B2:B8)				
9	172,4		=ЧАСТОТА(A2:A51;B2:B8)				
10	172,8						

Рис. 30. Расчет частот, формулы

Теперь заполним массив частот. Функция **ЧАСТОТА** здесь должна быть введена как формула массива, поскольку данная функция возвращает массив значений. Для этого выполните следующие шаги:

- выделите мышью диапазон **C2:C9**, который содержит на 1 ячейку больше, чем массив интервалов;
- введите формулу **=ЧАСТОТА(A2:A15;B2:B8)**, она автоматически будет вводиться в **C2**, первую ячейку выделенного диапазона;
- нажмите одновременно клавиши **CTRL+SHIFT+ENTER**.

В ячейки **C2:C9** будет введена одна и та же формула массива (см. рис. 30). Признаком формулы массива являются фигурные скобки, в режиме **Показать формулы** фигурные скобки не отображаются. Результаты расчетов частот приведены на рис. 29. Там же, в столбце **E**, указаны для наглядности границы интервалов, соответствующие частотам.

Обратите внимание, что в первом и последнем интервале оказалось по единице. Добавим эти единицы к прилегающим интервалам. Единица в последнем интервале появилась из-за приближенных вычислений, так как в ячейке **B8** на самом деле распложено значение чуть меньшее, чем 191,10. С этим можно, конечно, побороться. Например, поместить в ячейку **B8** максимальное значение выборки, или при расчете длины интервала вычислить его, используя округление вверх.

Функцию **ЧАСТОТА** относится к **Статистическим** функциям, ее лучше вводить, используя мастера функций. На рис. 31 приведено диалоговое окно функции.

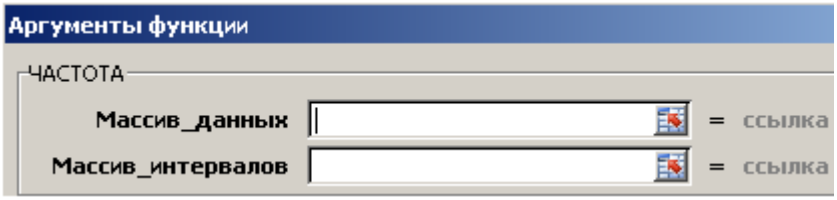


Рис 31. Диалоговое окно функции ЧАСТОТА

Данные для интервального ряда вычислены. Окончательный вид интервального ряда приведен на рис. 32.

	A	B	C	D	E	F	G	H
1	№		Границы				Середина	Частота
2	интервала		интервала				интервала	
3	1	[	167,80	;	171,68	]	169,74	6
4	2	(	171,68	;	175,57	]	173,63	12
5	3	(	175,57	;	179,45	]	177,51	9
6	4	(	179,45	;	183,33	]	181,39	10
7	5	(	183,33	;	187,22	]	185,28	6
8	6	(	187,22	;	191,10	]	189,16	7

Рис 32. Интервальный ряд

Для дальнейшего статистического анализа в качестве значений будут использоваться середины полученных интервалов. Напомним, что середина интервала находится как полусумма значений его концов.

Объем выборки не большой, равен 50, поэтому данный ряд можно было не разбивать на интервалы, а применить описательную статистику к исходному ряду данных, используя все имеющиеся данные. Выделим диапазон с данными **A2:A51** и проведем описательную статистику из **Пакетом анализа**. Результаты приведены в табл. 3.

Описательная статистика *Таблица 3*

Столбец1	
Среднее	178,972
Стандартная ошибка	0,93684
Медиана	178,9
Мода	179,9
Стандартное отклонение	6,62445
Дисперсия выборки	43,8833
Эксцесс	-0,9176
Асимметричность	0,1543
Интервал	23,3
Минимум	167,8
Максимум	191,1
Сумма	8948,6
Счет	50

Как видим, все посчитано, даже мода, которая в данном случае не является содержательной характеристикой. Значение средней, моды и медианы близки, что является признаком симметричного распределения. В данном случае интервальный ряд (табл. 4) может дать более содержательные характеристики. Здесь **fi** — частоты, **fi*** — накопленные частоты.

Таблица 4

Интервальный ряд

xi	169,74	173,63	177,51	181,39	185,28	189,16
fi	6	12	9	10	6	7
fi*	6	18	27	37	43	50

Мода и медиана для интервального ряда считаются по специальным формулам. Для расчета модального значения определим модальный интервал, это интервал с максимальной частотой. Наибольшая частота равна 12. Следовательно, второй интервал (171,68; 175,57] — модальный. Для интервального ряда вычисляется приближенное значение моды по формуле:

$$M_{\text{о}} = x_{M\text{о}} + h \cdot \frac{f_{M\text{о}} - f_{M\text{о}-1}}{\left(f_{M\text{о}} - f_{M\text{о}-1}\right) + \left(f_{M\text{о}} - f_{M\text{о}+1}\right)}.$$

Здесь $x_{\text{Мо}}$ — нижняя граница модального интервала; h — величина модального интервала; $f_{\text{Мо}}$ — частота модального интервала; $f_{\text{Мо-1}}$ — частота интервала, предшествующего модальному; $f_{\text{Мо+1}}$ — частота интервала, следующего за модальным:

$$M_{\text{о}} = 171{,}68 + 3{,}88 \cdot \frac{12-6}{(12-6)+(12-9)} \approx 174{,}27.$$

Рост примерно 174 см является наиболее распространенным среди футболистов.

Для расчета медианы определим медианный интервал, это интервал в котором содержать 25-26 элементы ряда, то есть срединные элементы, так как объем выборки равен 50. Следовательно, медианным является третий интервал (175,57 ; 179,45], так как накопленная частота этого интервала равна 27. Для интервального ряда вычисляется приближенное значение медианы по формуле:

$$Me = x_{M\text{е}} + h \cdot \frac{0{,}5\sum f - S_{M\text{е}-1}}{f_{M\text{е}}}.$$

Здесь x_{Me} — нижняя граница медианного интервала; h — величина медианного интервала; Σf — сумма всех частот ; S_{Me-1} — накопленная частота интервала, предшествующего медианному; f_{Me} — частота медианного интервала:

$$Me = 175{,}57 + 3{,}88 \cdot \frac{25-18}{9} \approx 178{,}58.$$

Половина спортсменов имеют рост не более 178,58, а другая половина — не менее 178,58.

Графические характеристики интервального ряда — гистограмма относительных частот, как эмпирический аналог функции распределения и кумулята, как эмпирический аналог функции распределения, если исследуемая случайная величина является непрерывной.

В табл. 5. приведены результаты расчетов относительных частот **pi** и накопленных относительных частот **pi*** для построения гистограммы (рис.33) и кумуляты (рис. 34).

Таблица 5

Интервальный ряд

xi	169,74	173,63	177,51	181,39	185,28	189,16
fi	6	12	9	10	6	7
pi	0,12	0,24	0,18	0,20	0,12	0,14
pi*	0,12	0,36	0,54	0,74	0,86	1

Рис 33. Гистограмма относительных частот

Рис 34. Кумулята относительных частот

3 Выравнивание временных рядов

Для выравнивания временных рядов **Анализ данных** имеет два инструмента: **Скользящее среднее** и **Экспоненциальное сглаживание**. Во временных рядах фактором, от которого зависит переменная прогнозирования, является время.

3.1. Скользящее среднее

Метод *скользящего среднего* является одним из наиболее используемых способов выравнивания значений временного ряда. Обычно используется для сглаживания краткосрочных колебаний и выделения основных тенденций, или тренда. Метод заключается в том, что значения исходного ряда заменяются прогнозными значениями, которые считаются по формуле $\tilde{x}_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}$, где k — количество наблюдений по которым будет вычисляться усреднение. Прогнозных значений будет меньше на (k-1) единицу. Заметим, что эта формула используется инструментом **Скользящее среднее**. На практике для усреднения применяют и другие формулы.

Пример. Пусть имеются данные за три года о продаже соков некоторой торговой компанией (табл. 7). Требуется выявить тенденцию временного ряда, используя скользящее среднее.

Таблица 7

Данные о продажах

Год	2010				2011				2012			
Квартал	1	2	3	4	1	2	3	4	1	2	3	4
Объем продаж (тыс. литров)	96	144	192	64	104	160	200	76	108	168	212	88

Визуальный анализ показывает, что к третьему кварталу продажи увеличиваются. Данный временной ряд содержит сезонные колебания периодичностью 4. Применим четырехточечную скользящую среднюю.

Выполним последовательность действий: **Данные → Анализ данных → Скользящее среднее**. В открывшемся диалогом окне укажем **Входной интервал A3:M3**, установим галочку **Метки в первой строке**, так как мы захватили заголовок ряда данных.

Для параметра **Интервал** укажем 4, так как усреднение будем проводить по четырем точкам. По умолчанию для вычисления скользящего среднего используется три точки.

Укажем **Выходной интервал**, здесь достаточно указать одну, начальную ячейку, для вывода данных. Исходные данные могут находиться в строке или столбце. Так как исходный интервал расположен в строке, выходные данные тоже будут расположены в строке.

Выберем дополнительно **Вывод графика и Стандартные погрешности**. Установка параметров показана на рис. 35.

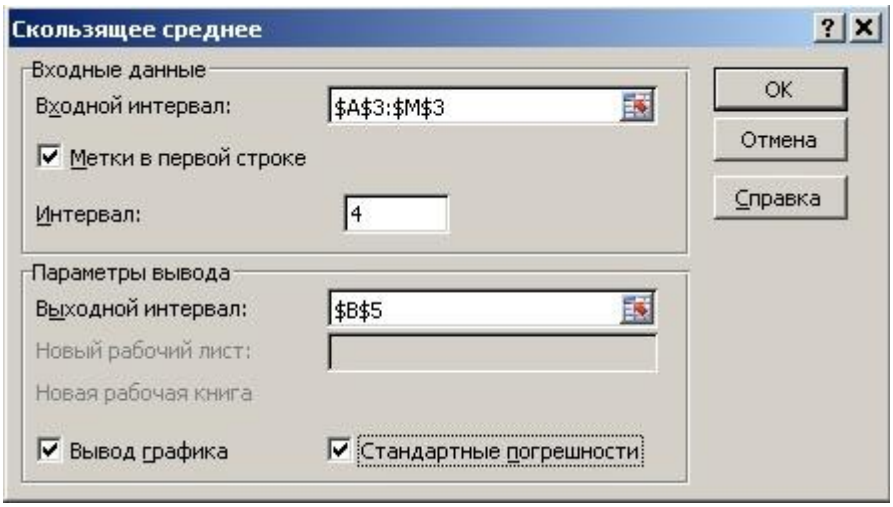


*Рис 35. Диалоговое окно инструмента **Скользящее среднее***

В результате на рабочем листе, начиная с указанной ячейки **B5**, будут выведены прогнозные значения, рассчитанные по методу скользящего среднего, и значения стандартных погрешностей (рис. 36), а также графики фактических и прогнозных значений (рис. 37).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Год	2010				2011				2012			
2	Квартал	1	2	3	4	1	2	3	4	1	2	3	4
3	Объем продаж (тыс. литров)	96	144	192	64	104	160	200	76	108	168	212	88
4													
5	Прогноз	#Н/Д	#Н/Д	#Н/Д	124	126	130	132	135	136	138	141	144
6	Погрешности	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	49,010	48,706	49,470	49,470	50,512	49,651

Рис 36. Результаты расчетов с применением ***Скользящее среднего***

Как мы видим, прогнозных значений будет на три меньше. Эти значения легко посчитать вручную. Например, для первой прогнозной точки получим (96+144+192+64)/4=124. Погрешность вычисляется корень из суммы квадратов разностей между исходными и расчетными *k* значениями, деленной на число *k*. То есть в ячейку **H6**, например, автоматически помещается формула =КОРЕНЬ(СУММКВРАЗН(E3:H3;E5:H5)/4).

Рис 37. График ***Скользящее среднее***

3.2. Экспоненциальное сглаживание

Экспоненциальное сглаживание также применяется к временным рядам. В этом методе сглаживания учитывается «старение» данных – в процессе сглаживания больший вес имеют последние данные. Сглаженные значения вычисляются по формуле $\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t$, где α — коэффициент сглаживания, 0< α <1, $\hat{y}_{t+1}$ — прогнозное значение за период (t+1), $\hat{y}_t$ — прогнозное значение за период t, y_t — фактическое значение да период t, причем $\hat{y}_1 = y_1$.

Выберем теперь **Экспоненциальное сглаживание** и установим параметры в диалоговом окне (рис. 38). Укажем **Входной интервал** и **Выходной интервал,**

установим **Фактор затухания** равным 0,5, отметим **Выбор графика** и **Стандартные погрешности**.

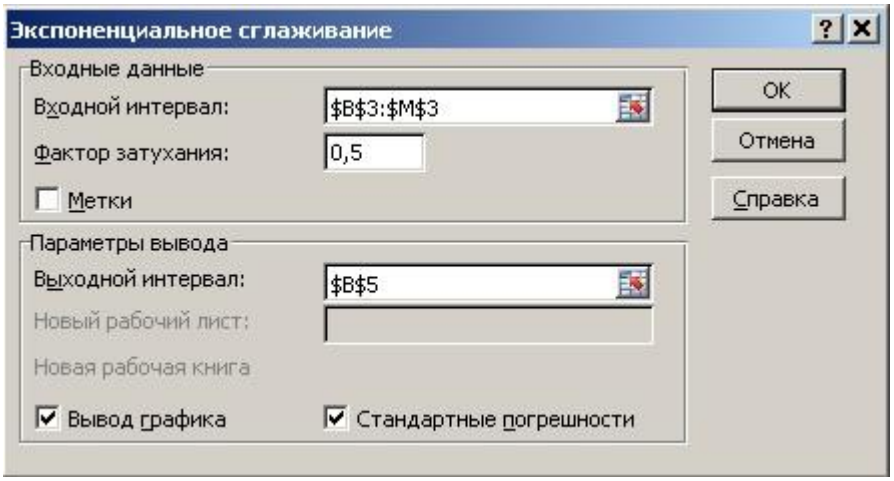


*Рис 38. Диалоговое окно **Экспоненциальное сглаживание***

Выполним еще раз **Экспоненциальное сглаживание**, теперь с **Фактором затухания** 0.8. Результаты сглаживания и исходные данные приведены на рис. Соответствующие графики изображены на рис. 39 и рис. 40.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Год	2010				2011				2012			
2	Квартал	1	2	3	4	1	2	3	4	1	2	3	4
3	Объем продаж (тыс. литров)	96	144	192	64	104	160	200	76	108	168	212	88
4	α=0,5												
5		#Н/Д	96	120	156	110,00	107,00	133,50	166,75	121,38	114,69	141,34	176,67
6		#Н/Д	#Н/Д	#Н/Д	#Н/Д	72,92	67,54	61,40	49,22	71,80	65,41	61,26	51,68
7	α=0,8												
8		#Н/Д	96	106	123	111,1	109,7	119,7	135,8	123,8	120,7	130,1	146,5
9		#Н/Д	#Н/Д	#Н/Д	#Н/Д	66,42	60,5	44,9	54,84	64,67	58,5	44,97	55,36

*Рис 39. Результаты расчетов с применением **Экспоненциального сглаживания***

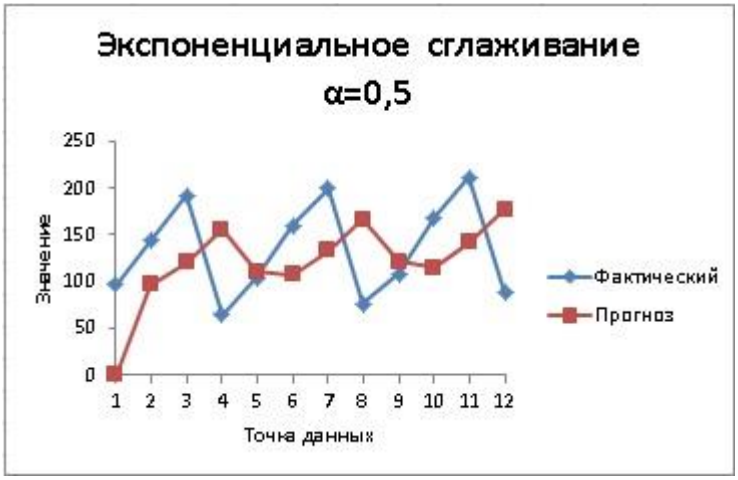


*Рис 40. График **Экспоненциальное сглаживание***

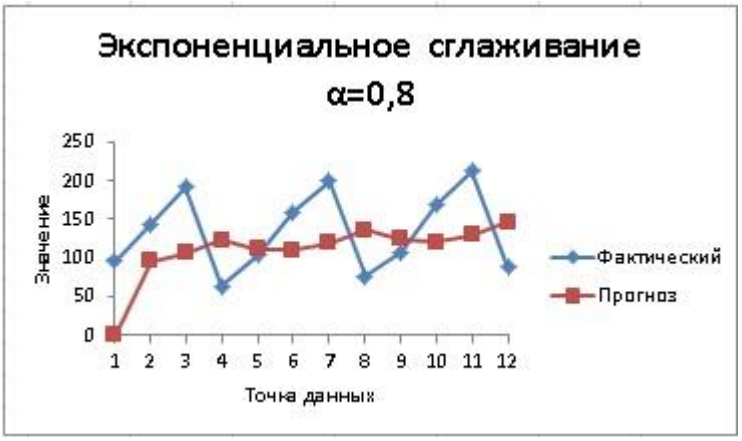


Рис 41. График ***Экспоненциальное сглаживание***