Представление числовых данных в виде таблиц и диаграмм

Распределение частот[1]

При увеличении объема выборки ни упорядоченный массив, ни диаграмма «ствол и листья» уже не позволяют легко представлять, анализировать и интерпретировать результаты. Для больших наборов данных следует создавать сводные таблицы, распределяя данные по группам (или категориям). Такой способ представления данных называется распределением частот. *Распределение частот* представляет собой сводную таблицу, в которой данные распределены по группам или категориям. Если данные сгруппированы в виде распределения частот, процесс их анализа и интерпретации становится более управляемым и осмысленным. При распределении частот следует внимательно выбирать интервал группирования, или размах групп, а также вычислять границы каждой группы, не допуская их перекрытия.

Выбор количества групп

Количество групп, выбранных для группировки данных, непосредственно зависит от объема исходной выборки. Чем больше элементов содержит выборка, тем больше групп можно создать. Как правило, распределение частот должно содержать не менее 5 и не более 15 групп. Если групп слишком мало или слишком много, новую информацию получить сложно. Выделение групп процесс творческий, и я бы рекомендовал в качестве первого подхода использовать формулу Стерджесса:

$$(1)\ k = 1 + \log_2 n$$

где k – число групп, n – объем выборки; далее визуально определить по графику, насколько удачным получилось разбиение и, если требуется, скорректировать число групп на величину ± 1.

Вычисление интервала группирования

Каждая группа, образующая распределение частот, должна иметь одинаковый размах. Чтобы определить ширину интервала группирования, диапазон изменения данных делят на заданное количество групп.

$$(2)\ \text{Ширина интервала группирования} = \text{Диапазон} / \text{Количество групп}$$

В нашем примере (см. первое упоминание в заметке Как упорядочить массив данных) имеются данные о 158 фондах (рис. 1).

-6,1	-2,8	-1,2	-0,7	0,5	1,8	1,9	2,5	2,8	3,3
3,5	3,8	3,8	4,0	4,2	4,3	4,5	4,6	5,0	5,1
5,2	5,4	5,5	5,8	5,9	6,0	6,2	6,3	6,5	6,5
7,0	7,1	7,1	7,2	7,2	7,3	7,5	7,6	7,6	7,8
7,8	7,8	7,9	8,1	8,1	8,2	8,3	8,3	8,4	8,5
8,5	8,5	8,6	8,8	8,8	8,8	9,0	9,0	9,1	9,1
9,1	9,2	9,3	9,3	9,5	9,5	9,5	9,5	9,6	9,6
9,7	9,8	9,9	9,9	9,9	9,9	10,1	10,1	10,1	10,1
10,2	10,3	10,3	10,4	10,5	10,5	10,5	10,5	10,5	10,5
10,6	10,7	10,7	10,8	10,9	11,0	11,0	11,1	11,1	11,1
11,2	11,2	11,3	11,3	11,3	11,3	11,4	11,5	11,5	11,5
11,6	11,7	11,7	11,9	11,9	12,2	12,2	12,3	12,3	12,4
12,5	12,7	12,9	12,9	12,9	13,0	13,1	13,2	13,4	13,4
13,7	13,7	13,9	14,1	14,7	14,8	14,9	15,0	15,7	15,8
15,8	16,0	16,9	17,0	17,0	17,6	17,8	18,1	18,1	18,2
18,5	18,5	18,7	18,9	21,4	22,0	22,9	26,3		

Рис. 1. Упорядоченный массив, содержащий данные о пятилетней среднегодовой доходности 158 фондов, ориентированных на быстрый рост капитала, за период с 1 января 1997 до 31 декабря 2001

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 105–124

Для такого массива достаточно создать восемь групп ($k = 1 + log_2 158 = 8,3$). Диапазон значений массива вычисляется по формуле 26,3 – (–6,1) = 32,4. С учетом формулы (2) ширина интервала группирования = 32,4 / 8 = 4,05. Для удобства округляем до 5 (в меньшую сторону округлять нельзя, так как какие-то значения выпадут из рассмотрения).

Вычисление границ групп

Для вычисления распределения частот необходимо так определить *границы групп*, чтобы они не пересекались. Перекрытие групп не допускается. Поскольку размах каждой группы, построенной на основе данных о пятилетней среднегодовой доходности фондов, равен 5%, границы групп должны быть установлены так, чтобы учесть все данные. По возможности эти границы должны быть достаточно наглядными. Например, величины из первой группы должны изменяться в диапазоне от –10% до –5% и так далее, пока не будут сформированы 8 неперекрывающихся групп, ширина каждой из которых равна 5% (рис. 2).

Пятилетняя среднегодовая доходность				Количество фондов
от	–10,0	до	–5,0	1
от	–5,0	до	0,0	3
от	0,0	до	5,0	14
от	5,0	до	10,0	58
от	10,0	до	15,0	61
от	15,0	до	20,0	17
от	20,0	до	25,0	3
от	25,0	до	30,0	1

Рис. 2. Распределение частот для пятилетней среднегодовой доходности 158 фондов

Главным преимуществом этой таблицы является возможность легко вычислять основные характеристики данных. Например, таблица демонстрирует, что диапазон среднегодовой доходности 158 фондов ограничен числами –10% и 30%, причем показатели в основном группируются в диапазоне 5…15%.

С другой стороны, эта сводная таблица имеет недостаток: по ней невозможно определить, как распределены индивидуальные данные внутри групп. Например, доходность трех фондов из представленных в таблице, изменяется в диапазоне 20–25%, но определить, вокруг какого значения они сконцентрированы (20 или 25%), невозможно. Для представления средней доходности этих трех фондов выбирается срединная точка (22,5%). Срединной точкой интервала –10…–5%, является значение –7,5% и т.д.

Субъективность при выборе границ групп

Выбор границ групп при вычислении распределения частот является субъективным. Если наборы данных невелики, одинаковый выбор границ групп для разных выборок может привести к разным результатам. Например, если при вычислении распределения частот для показателей пятилетней среднегодовой доходности ширину интервалов группирования установить равной 4%, а не 5%, возникнет смещение распределения. Особенно сильно этот эффект проявляется при работе с малыми выборками.

Смещение распределения возникает не только в результате изменения границ групп. Например, ширину интервала группирования можно оставить равной 5%, изменив границы первой и последней групп. Эта манипуляция также приводит к смещению распределения, особенно, если объем выборки невелик. К счастью, по мере увеличения объема выборки этот эффект становится менее выраженным.

Распределение относительных частот и процентное распределение

Для более углубленного анализа распределения частот можно построить либо распределение относительных частот (долей) либо процентное распределение. Выбор распределения зависит от того, с какими данными желает работать пользователь: с долями или процентами (рис. 3).

Пятилетняя среднегодовая доходность				Доля фондов	Процент фондов
от	−10,0	до	−5,0	0,006	0,6
от	−5,0	до	0,0	0,019	1,9
от	0,0	до	5,0	0,089	8,9
от	5,0	до	10,0	0,367	36,7
от	10,0	до	15,0	0,386	38,6
от	15,0	до	20,0	0,108	10,8
от	20,0	до	25,0	0,019	1,9
от	25,0	до	30,0	0,006	0,6
Итого				1,000	100,0

Рис. 3. Распределение относительных частот и процентное распределение для пятилетней среднегодовой доходности 158 фондов

Таким образом, доля фондов, ориентированных на быстрый рост капитала, среднегодовая доходность которых изменяется от 10 до 15%, равна 0,386, а процент — 38,6%. Работать с долями или процентами удобнее, чем с количеством элементов в группе. Распределение относительных частот, как и процентное распределение, позволяет сравнивать даже наборы данных, имеющие разные объемы.

Функция распределения

Часто оказывается полезной таблица интегральных процентов, которую также называют распределением интегральных процентов. Функция распределения позволяет обнаружить информацию, которая ускользает от распределения частот (рис. 4). (Для построения распределение интегральных процентов были использованы данные, приведенные на рис. 3.)

Пятилетняя среднегодовая доходность, %				Процент фондов в группе	Процент фондов, доходность которых не превышает верхней границы группы
от	−10,0	до	−5,0	0,6	0,0
от	−5,0	до	0,0	1,9	0,6
от	0,0	до	5,0	8,9	2,5 = 0,6 +1,9
от	5,0	до	10,0	36,7	11,4 = 0,6 + 1,9 + 8,9
от	10,0	до	15,0	38,6	48,1 = 0,6 + 1,9 + 8,9 + 36,7
от	15,0	до	20,0	10,8	86,7 = 0,6 + 1,9 + 8,9 + 36,7 + 38,6
от	20,0	до	25,0	1,9	97,5 = 0,6 + 1,9 + 8,9 + 36,7 + 38,6 + 10,8
от	25,0	до	30,0	0,6	99,4 = 0,6 + 1,9 + 8,9 + 36,7 + 38,6 + 10,8 + 1,9
от	30,0	до	35,0	0,0	100,0 = 0,6 + 1,9 + 8,9 + 36,7 + 38,6 + 10,8 + +1,9 + 0,6

Рис. 4. Распределение интегральных процентов

Для вычисления распределения частот можно воспользоваться командой *Данные → Анализ данных* (рис. 5).

Рис. 5. Анализ данных

Если надстройка *Анализ данных* не отражается, ее нужно предварительно установить. Выберите меню *Файл → Параметры* (рис. 6). В открывшемся окне *Параметры Excel*, выберите меню *Надстройки → Пакет анализа* и кликните на кнопке *Перейти*.

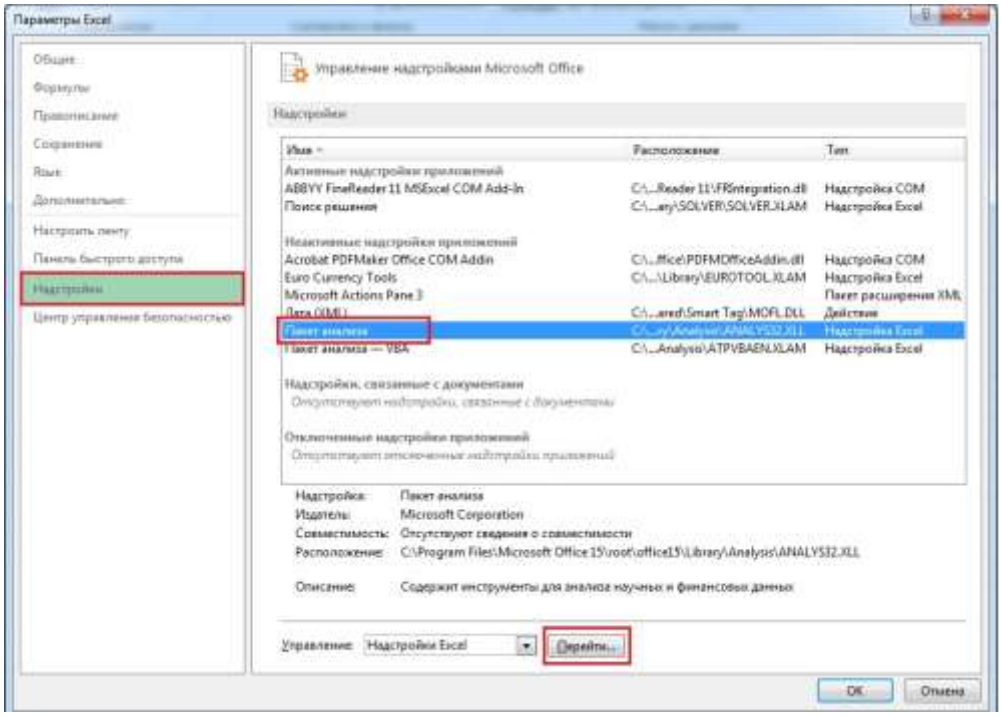


Рис. 6. Параметры Excel

В открывшемся окне Надстройки поставьте галочку на опции Пакет анализа и кликните Ok (рис. 7).

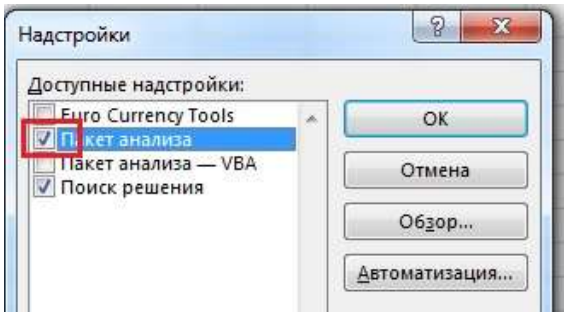


Рис. 7. Надстройки

Теперь нужно подготовить исходные данные. Расположим числа доходности 158 фондов (как на рис. 1) в столбце *A* (рис. 8). Вообще говоря, это не обязательно. Можно расположить данные и в виде двумерного массива, как на рис. 1. В столбце *C* размещаем упорядоченный массив верхних границ диапазонов. Именно этот массив и будет чуть позже введен в поле *Интервал карманов*. Здесь есть маленькая тонкость. Excel включит верхнюю границу в диапазон. Например, интервал, для которого указана верхняя граница 10, фактически является интервалом 5,00001…10. Именно к этому интервалу будет относиться число 10, а не к следующему интервалу 10…15. Можно сказать и иначе: нижняя граница не входит в интервал, а верхняя – входит. Запускаем надстройку *Анализ данных*, из списка *Инструменты анализа* выбираем пункт *Гистограмма*, жмем *ОК*.

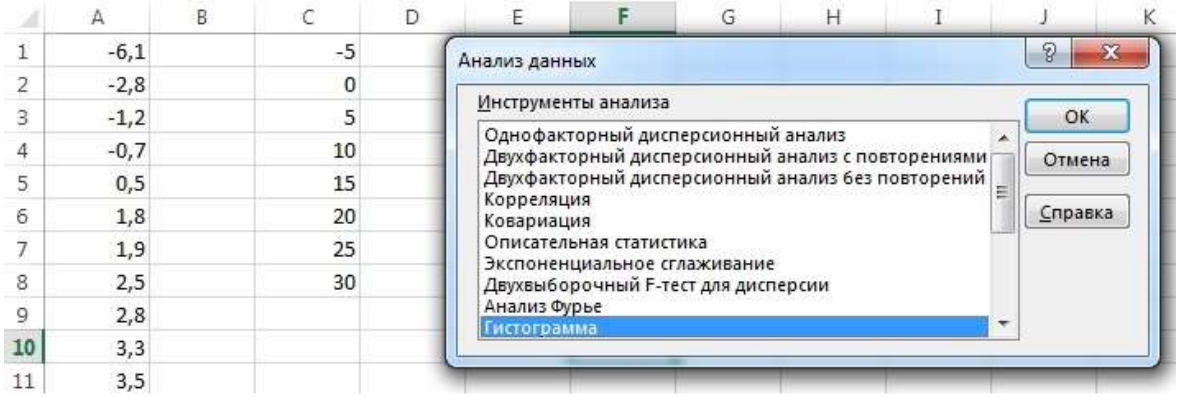


Рис. 8. Подготовка исходных данных и запуск надстройки

В диалоговом окне *Гистограмма* (рис. 9) в поле *Входной интервал* выбираем наш массив A1:A158, в поле *Интервал карманов* интервал C1:C8, переключатель *Параметры вывода* ставим в положение *Новый рабочий лист*, включаем *Интегральный процент* и *Вывод графика*, жмем ОК.

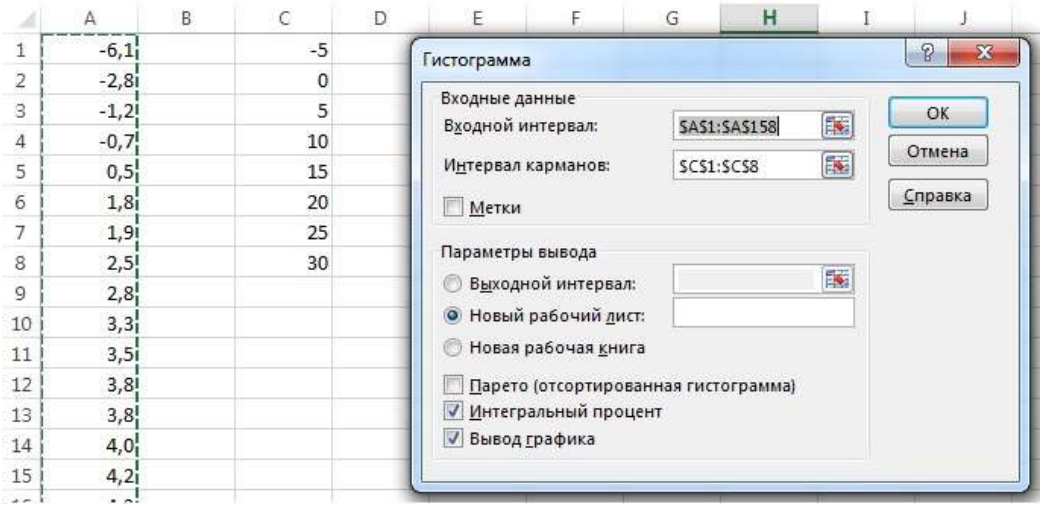
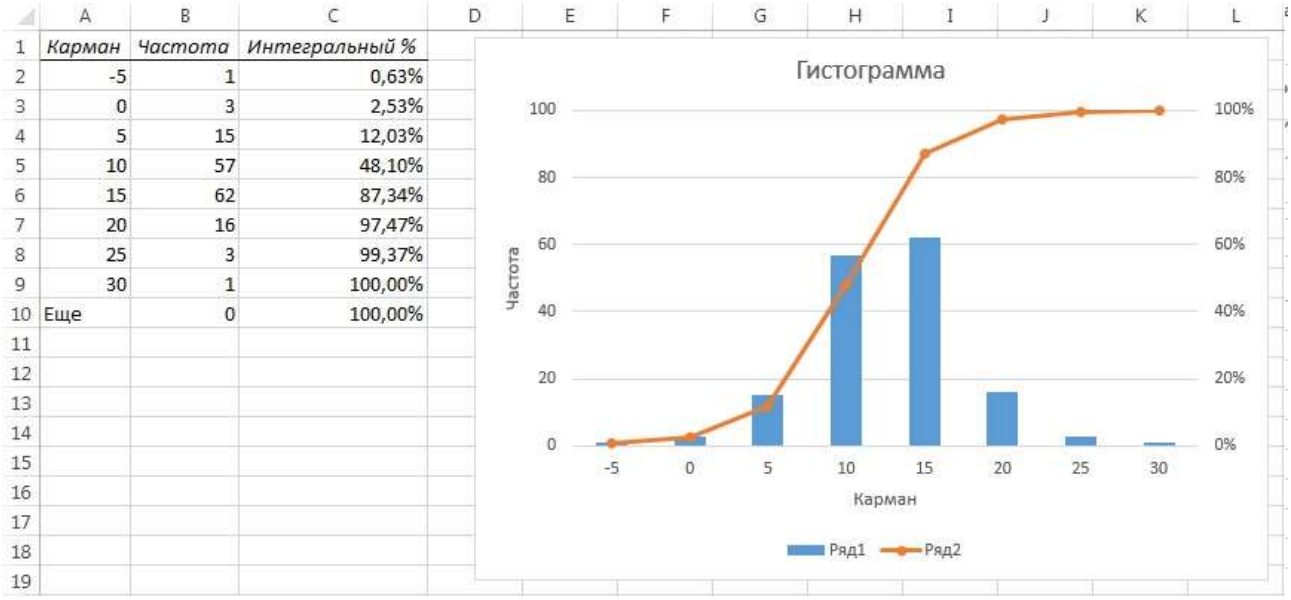


Рис. 9. Настройка гистограммы

На отдельном листе выводится таблица, аналогичная рис. 3, а также график (гистограмма) с функцией распределения по диапазонам и интегральным процентом (рис. 10). В столбце *Карман* таблицы и на оси абсцисс гистограммы указаны верхние границы диапазонов. При этом в первый диапазон попадают все значения меньше первой указанной границы, то есть все значения меньше «минус 5» (включая -5), а в диапазон под названием *Еще* – все значения превышающие самую большую границу, то есть больше 30. Вы можете «поиграть» значениями *Карманов*, чтобы почувствовать, как они работают.



Карман	Частота	Интегральный %
-5	1	0,63%
0	3	2,53%
5	15	12,03%
10	57	48,10%
15	62	87,34%
20	16	97,47%
25	3	99,37%
30	1	100,00%
Еще	0	100,00%

Рис. 10. Гистограмма с функцией распределения по диапазонам и интегральным процентом

Гистограмма

Следуя принципу «лучше один раз увидеть, чем сто раз услышать», для анализа статистических данных часто используют графические изображения, а не таблицы. Например, с помощью гистограммы описывают числовые данные, сгруппированные по частоте, относительной частоте или процентной доле. *Гистограмма* — это диаграмма, на которой изображены столбики, границы которых совпадают с границами групп. При построении гистограмм исследуемая случайная величина откладывается по оси X, а количество элементов в соответствующих группах, их относительная частота или процентная доля — по оси Y. На рис. 10 изображена гистограмма, построенная на основе данных о пятилетней среднегодовой доходности 158 фондов, ориентированных на быстрый рост капитала. По оси Y отложено количество элементов в группах.

При сравнении *нескольких* наборов данных бывает довольно сложно создавать диаграммы «ствол и листья» и гистограммы. Например, иногда трудно правильно интерпретировать разницу между высотами соответствующих столбцов разных гистограмм. Для нескольких наборов данных предпочтительными оказываются полигоны, построенные по относительным частотам или процентным долям.

Полигон

Как и при построении гистограмм, величина исследуемой переменной откладывается вдоль горизонтальной оси. По вертикальной оси откладывается количество элементов в каждой группе, их относительная доля или процент. *Процентный полигон* представляет собой график, построенный путем соединения средних точек, соответствующих процентной доле каждой группы (рис. 11). Надстройка *Анализ данных* не умеет строить полигоны; с методом, использованным при построении графика на рис. 11 можно ознакомиться на соответствующем листе Excel-файла.

Рис. 11. Процентный полигон для пятилетней доходности

Полигон интегральных процентов, или кривая распределения, является графическим изображением распределения суммарных процентов (накопительным итогом).

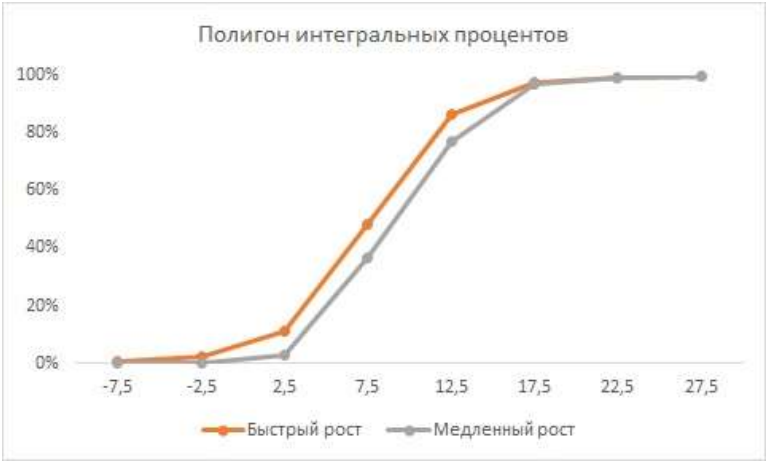


Рис. 12. Полигон интегральных процентов

На рис. 12 изображены полигоны интегральных процентов (метод построения см. Excel-файл) на основе пятилетней среднегодовой доходности 158 фондов, ориентированных на быстрый рост капитала, и 101 фонда, ориентированного на медленный рост капитала. На оси X отложены средние значения диапазонов. Видно, что среднегодовая доходность 48,1% фондов, ориентированных на быстрый рост капитала, не превышает 10%, в то время как доля фондов, ориентированных на медленный рост капитала, в этом интервале равна 36,7%. Обратите внимание на то, что в интервале до 20% кривая распределения среднегодовой доходности фондов, ориентированных на быстрый рост капитала, расположена слева от кривой распределения доходности фондов, ориентированных на медленный рост капитала. В то же время количество фондов, ориентированных на быстрый и медленный рост капитала, доходность которых не превышает 20,0%, приблизительно одинаково.

Изображение двумерных числовых данных

Выше мы рассмотрели гистограммы, полигоны, кривые распределений и полигоны накопленных частот, представляющие собой удобные графические инструменты для анализа числовых *одномерных* данных. Для анализа *двумерных* числовых величин используется иной вид графического представления – диаграмма разброса. В программе Excel эта диаграмма называется точечной, а в научной литературе — корреляционной. Такие диаграммы оказываются полезными в разных областях деловой активности. Например, специалисты по маркетингу с помощью таких диаграмм могут исследовать эффективность рекламной кампании, сравнивая объемы недельных продаж и расходы на рекламу, а менеджеры по кадрам — изучать систему оплаты труда в компании, сравнивая трудовой стаж сотрудников и их текущую зарплату.

Используя диаграмму разброса, менеджер по логистике может анализировать вклад таможенного сбора в суммарные логистические затраты (рис. 13). Диаграммы разброса играют важную роль при изучении коэффициента корреляции, а также в регрессионном анализе.

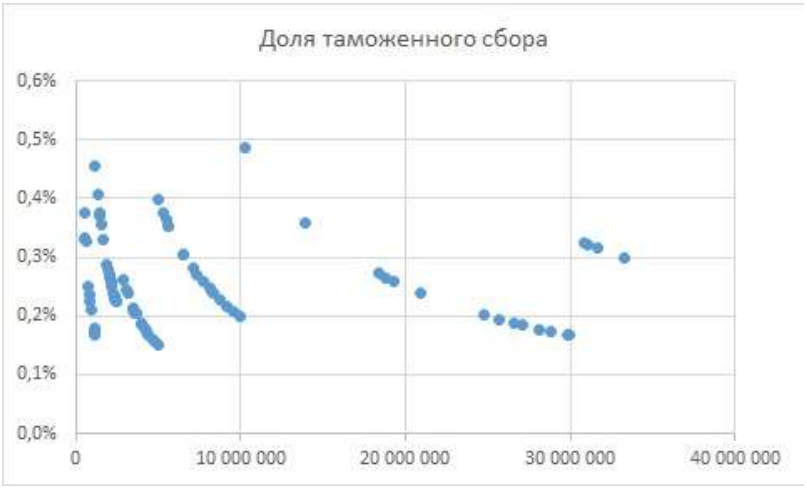
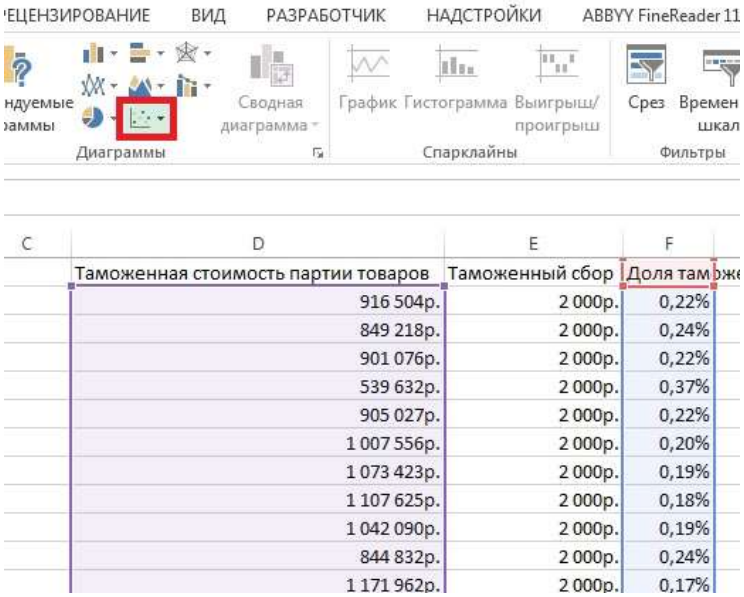


Рис. 13. Корреляция таможенного сбора (как процентной доли таможенной стоимости, ось Y) и таможенной стоимости, ось X

Для построения диаграммы разброса выберите два столбца и кликните на типе диаграммы *Точечная* (рис. 14). Обратите внимание на то, что *Мастер диаграмм* по умолчанию считает, что переменная X находится в первом столбце диапазона. Если данные на вашем листе расположены иначе, поменяйте столбцы местами.



C	D	E	F
	Таможенная стоимость партии товаров	Таможенный сбор	Доля там
	916 504р.	2 000р.	0,22%
	849 218р.	2 000р.	0,24%
	901 076р.	2 000р.	0,22%
	539 632р.	2 000р.	0,37%
	905 027р.	2 000р.	0,22%
	1 007 556р.	2 000р.	0,20%
	1 073 423р.	2 000р.	0,19%
	1 107 625р.	2 000р.	0,18%
	1 042 090р.	2 000р.	0,19%
	844 832р.	2 000р.	0,24%
	1 171 962р.	2 000р.	0,17%

Рис. 14. Построение точечной диаграммы

Предыдущая заметка Как упорядочить массив данных

К оглавлению Статистика для менеджеров с использованием Microsoft Excel