

Prepared By

Handaru Jati, Ph.D

Universitas Negeri Yogyakarta

handaru@uny.ac.id

Chapter 7



Statistical Analysis with Excel



Chapter Overview

- 7.1 Introduction
- 7.2 Understanding Data
 - 7.2.1 Descriptive Statistics
 - 7.2.2 Histograms
- 7.3 Relationships in Data
 - 7.3.1 Trend Curves
 - 7.3.2 Regression
- 7.4 Distributions
- 7.5 Summary
- 7.6 Exercises

7.1 Introduction

This chapter illustrates the tools available in Excel for performing statistical analysis. These tools include new functions, the data analysis toolpack, and some new chart features. This chapter is not intended to teach the statistical concepts which can be used in Excel's analysis, but rather demonstrate to the reader that several tools are available in Excel to perform these statistical functions. Statistical analysis is used often in DSS applications for analyzing input and displaying conclusive output. These tools will be especially used in applications involving simulation. We discuss the application of statistical analysis in simulation in Chapter 9 and again in Chapter 20 with VBA. We have several DSS applications which use statistical analysis tools, such as Queuing Simulation.

7.2 Understanding Data

Statistical analysis provides an understanding of a set of data. Using statistics, we can determine an average value, a variation of the data from this average, a range of data values, and perform other interesting analysis. We begin this analysis by using statistical Excel functions.

One of the basic statistical calculations to perform is finding the *mean* of a set of numbers; the mean is simply the average, which we learned how to calculate with the **AVERAGE** function in Chapter 4:

`=AVERAGE(range or range_name)`

Figure 7.1 displays a table of family incomes for a given year. We first name this range of data, cells B4:B31, as "FamIncome." We can now find the average, or *mean*, family income for that year using the AVERAGE function as follows (see Figure 7.2):

`=AVERAGE(FamIncome)`

	A	B	C
1	Yearly Family Income		
2			
3		Income	
4		\$33,628	
5		\$38,808	
6		\$36,399	
7		\$34,611	
8		\$42,532	
9		\$35,996	
10		\$49,255	
11		\$36,469	
12		\$45,460	
13		\$40,161	
14		\$29,800	
15		\$45,006	
16		\$30,165	
17		\$32,193	
18		\$25,095	
19		\$45,272	
20		\$46,612	
21		\$28,794	
22		\$47,941	
23		\$25,703	
24		\$42,606	
25		\$37,253	
26		\$40,766	
27		\$39,331	
28		\$27,912	
29		\$28,525	
30		\$48,599	
31		\$45,548	

Figure 7.1 Family incomes for a given year.

D6		fx =AVERAGE(FamIncome)			
	A	B	C	D	E
1	Yearly Family Income				
2					
3		Income			
4		\$33,628			
5		\$38,808			
6		\$36,399		Mean (Average)	
7		\$34,611		\$37,873	
8		\$42,532			
9		\$35,996			
10		\$49,255			
11		\$36,469			

Figure 7.2 Calculating the mean, or average, of all family incomes using the AVERAGE function.

Similar to the mean, the **median** can also be considered the “middle” value of a set of numbers. The median is the middle number in a list of sorted data. To find the median, we use the **MEDIAN** function, which takes a range of data as its parameter:

$$=MEDIAN(\text{range or range_name})$$

To determine the median of the above family incomes, we enter the MEDIAN function as follows:

$$=MEDIAN(\text{FamIncome})$$

We can check whether or not this function has returned the correct result by sorting the data and finding the middle number. Since there are an even number of family incomes recorded in the table, we must average the two middle numbers. The result is the same (see Figure 7.3).

	A	B	C	D	E	F	G
1	Yearly Family Income						
2							
3		Income					
4		\$33,628					\$25,095
5		\$38,808		Mean (Average)			\$25,703
6		\$36,399		\$37,873			\$27,912
7		\$34,611					\$28,525
8		\$42,532		Median			\$28,794
9		\$35,996		\$38,031			\$29,800
10		\$49,255					\$30,165
11		\$36,469					\$32,193
12		\$45,460					\$33,628
13		\$40,161					\$34,611
14		\$29,800					\$35,996
15		\$45,006					\$36,399
16		\$30,165					\$36,469
17		\$32,193					\$37,253
18		\$25,095					\$38,808
19		\$45,272					\$39,331
20		\$46,612					\$40,161
21		\$28,794					\$40,766
22		\$47,941					\$42,532
23		\$25,703					\$42,606
24		\$42,606					\$45,006
25		\$37,253					\$45,272
26		\$40,766					\$45,460
27		\$39,331					\$45,548
28		\$27,912					\$46,612
29		\$28,525					\$47,941
30		\$48,599					\$48,599
31		\$45,548					\$49,255

Figure 7.3 Using the MEDIAN function and verifying the result by sorting the data and finding the middle value.

Another important value, **standard deviation**, is the square root of the **variance**, which measures the difference between the mean of the data set and the individual values. Finding the standard deviation is simple with the **STDEV** function. The parameter for this function is also just the range of data for which we are calculating the standard deviation:

$$=STDEV(\text{range or range_name})$$

In Figure 7.4, we calculate the standard deviation of the family income data using the following function:

$$=STDEV(\text{FamIncome})$$

	A	B	C	D	E
1	Yearly Family Income				
2					
3		Income			
4		\$33,628			
5		\$38,808		Mean (Average)	
6		\$36,399		\$37,873	
7		\$34,611			
8		\$42,532		Median	
9		\$35,996		\$38,031	
10		\$49,255			
11		\$36,469		Standard Deviation	
12		\$45,460		\$7,401	
13		\$40,161			
14		\$29,800			
15		\$45,000			

Figure 7.4 Using the STDEV function.

 <p>Summary</p>	Statistical Functions	
	AVERAGE	Finds the mean of a set of data.
	MEDIAN	Finds the median of a set of data.
	STDEV	Finds the standard deviation of a set of data.

The **Analysis Toolpack** provides an additional method by which to perform statistical analysis. This Excel Add-In includes statistical analysis techniques such as *Descriptive Statistics*, *Histograms*, *Exponential Smoothing*, *Correlation*, *Covariance*, *Moving Average*, and others (see Figure 7.5). These tools automate a sequence of calculations that require much data manipulation if only Excel functions are being used. We will now discuss how to use *Descriptive Statistics* and *Histograms* in the *Analysis Toolpack*.

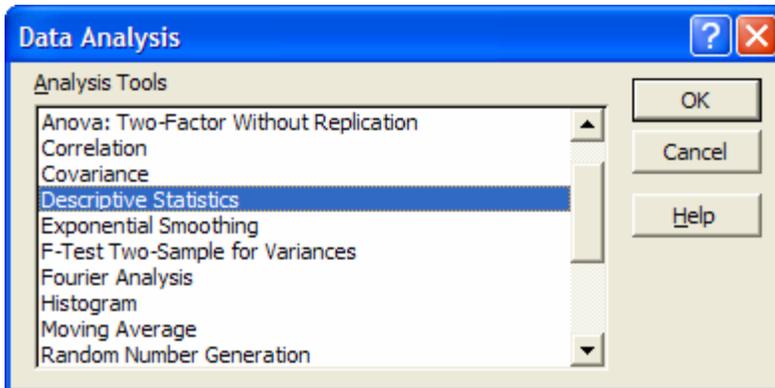


Figure 7.5 The Data Analysis dialog box provides a list of analytical tools.

(Note: Before using the *Analysis Toolpack*, you must ensure that it is an active Add-in. To do so, choose *Tools > Add-ins* from the Excel menu and select *Analysis Toolpack* from the list. If you do not see it on the list, you may need to re-install Excel on your computer. After you have checked *Analysis Toolpack* on the *Add-ins* list, you should find the *Data Analysis* option under the *Tools* menu option.)

7.2.1 Descriptive Statistics

The *Descriptive Statistics* option provides a list of statistical information about our data set, including the mean, median, standard deviation, and variance. To use *Descriptive Statistics*, we go to *Tools > Data Analysis > Descriptive Statistics*. Choosing the *Descriptive Statistics* option from the *Data Analysis* window (shown in Figure 7.5) displays a new window (shown in Figure 7.6).

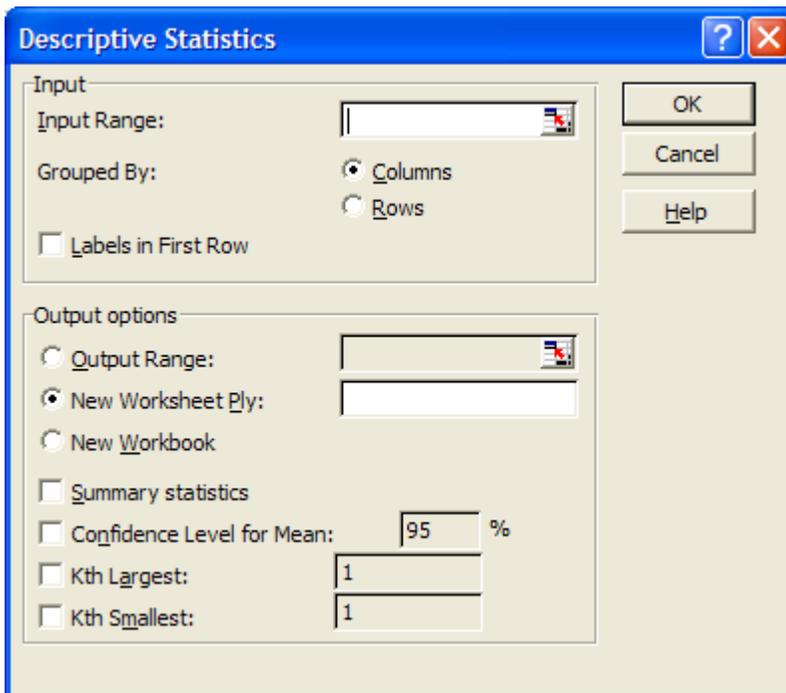


Figure 7.6 The Descriptive Statistics dialog box appears after it is chosen from the Data Analysis list.

The *Input Range* refers to the location of the data set. We can check whether our data is *Grouped By Columns* or *Rows*. If there are labels in the first row of each column of data, then we check the *Labels in First Row* box. The *Output Range* refers to where we want the results of the analysis to be displayed in the current worksheet. We could also place the analysis output in a new worksheet or a new workbook. The *Summary Statistics* box calculates the most commonly used statistics from our data. We will discuss the last three options, *Confidence Level for Mean*, *Kth Largest*, and *Kth Smallest*, later in the chapter.

Let's now consider an example in order to appreciate the benefit of this tool. In Figure 7.7 below, there is a table containing quarterly stock returns for three different companies. We want to determine the average stock return, the variability of stock returns, and which quarters had the highest and lowest stock returns for each company. This information could be very useful for selecting a company in which to invest.

	A	B	C	D
1	Quarterly Stock Returns between 1995 and 2000			
2				
3	Date	MSFT	GE	INTEL
4	Q1 1995	0.04	-0.06	-0.11
5	Q2 1995	0.14	0.19	-0.06
6	Q3 1995	0.08	-0.02	0.12
7	Q4 1995	0.03	0.04	0.17
8	Q1 1996	-0.04	-0.03	-0.02
9	Q2 1996	-0.07	0.01	0.13
10	Q3 1996	0.10	0.00	-0.04
11	Q4 1996	-0.13	0.03	0.11
12	Q1 1997	0.04	-0.02	-0.05
13	Q2 1997	0.02	-0.03	-0.16
14	Q3 1997	0.08	0.07	-0.07
15	Q4 1997	0.10	-0.02	-0.04
16	Q1 1998	0.05	0.08	0.06
17	Q2 1998	-0.08	0.03	0.00
18	Q3 1998	0.01	0.01	0.17
19	Q4 1998	-0.04	-0.02	-0.01
20	Q1 1999	0.11	0.07	0.01
21	Q2 1999	-0.08	0.02	0.09
22	Q3 1999	0.08	0.02	-0.01
23	Q4 1999	-0.05	0.04	0.11
24	Q1 2000	-0.16	0.03	0.09
25	Q2 2000	0.02	0.00	-0.03
26	Q3 2000	0.10	-0.02	-0.11
27	Q4 2000	-0.03	0.01	0.14
28				

Figure 7.7 Quarterly stock returns for three companies.

We use the Descriptive Statistics tool to answer these questions. In the Descriptive Statistics dialog box (see Figure 7.8) we enter the range *B3:D27* for the *Input Range*. (Notice that we do not select the first column, *Date*, since we are not interested in a

First, let's become familiar with the *Mean*, *Median*, and *Mode*. As already mentioned, the *Mean* is simply the average of all values in a data set, or all observations in a sample. We have already observed that without the Analysis Toolpack, the mean value can be found with the AVERAGE function in Excel. The *Median* is the "middle" observation when the data is sorted in ascending order. If there is an odd number of values, then the median is truly the middle value. If there is an even number of values, then it is the average of the two middle values.

The *Mode* is the most frequently occurring value. If there is no repeated value in the data set, then there is no *Mode* value, as in this example. The *Mean* is usually considered the best measure of the central data value if the data is fairly symmetric; otherwise the *Median* is more appropriate. In this example, we can observe that the *Mean* and *Median* values for each company differ slightly; however, we use the *Mean* value to compare the average stock returns for this company. This analysis alone implies that GE and INTEL have higher stock returns, on average, than MSFT. But these values are still very close, so we need more information to make a better comparative analysis.

Now, let's consider the *Standard Error*, *Standard Deviation*, and *Sample Variance*. All of these values measure the spread of the data from the mean. The *Sample Variance* is the average squared distance from the mean to each data point. The *Standard Deviation* is the square root of the *Sample Variance* and is more frequently used. Looking at these values for the example data, we can observe that INTEL has a highly varied stock return, while GE's is more stable. Therefore, even though they have the same *Mean* value, this difference in the *Standard Deviation* makes GE a more favorable stock in which to invest.

We will discuss *Standard Error*, which is used in connection with trends and trendlines, in more detail later.

The *Standard Deviation*, usually referred to as s , is an important value in understanding variation in data. Most data, 68% of a Normal distribution, lies between $+s$ and $-s$ from the mean. Almost all of the data, 95% of a Normal distribution, lies between $+2s$ and $-2s$ from the mean. Any values in the data set that lie more than $\pm 2s$ from the mean are called **outliers**. Note that outliers do not have to be defined based on $\pm 2s$, but can be measured by any other multiplier of standard deviation or any other set deviation from the mean value. Outliers can provide insightful information about a data set.

For example, if we create a chart of the GE data, we can observe that the second data value is an outlier since it is $\pm 2s = \pm 2 * 0.05 = \pm 0.1$ from the mean (0.02); in other words, any value above 0.12 or below -0.08 is an outlier. The second data value for GE is +0.19 (see Figure 7.10). This figure may imply that something significant happened to GE as a company during Q2 1995, that something affected the national economy, or that they faced any number of (un)predictable situations. However, since the second data value is the only outlier in the last five years of quarterly data for GE, it seems that the mean and standard deviation are accurate measures of the behavior of GE stock returns.

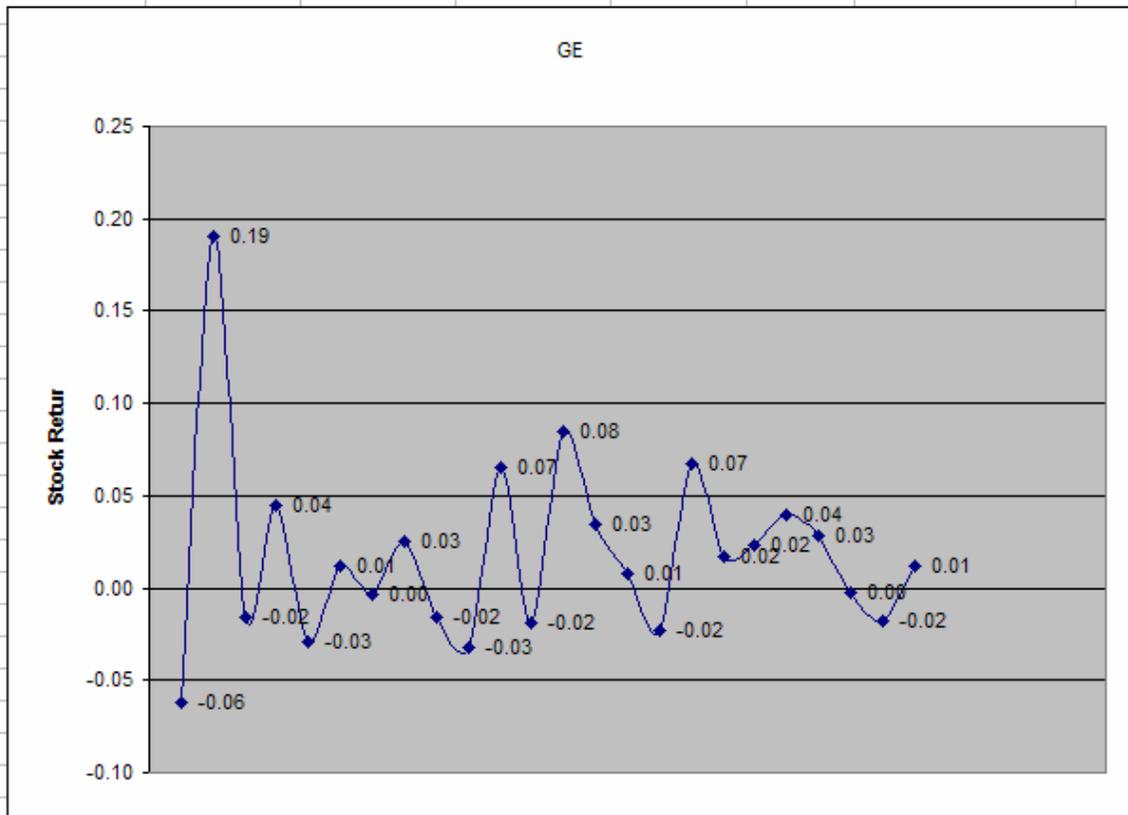


Figure 7.10 The second data point is an outlier since it is greater than $2s$ from the mean.

We can identify outliers by looking at a chart of data, or we can actually locate values in the data set that are greater than $+2s$ and smaller than $-2s$. To do so, we can place the following formula in an adjacent column to the data:

$$=IF(ABS(data_value - mean_value) > 2*s, "outlier",)$$

This formula states that if the absolute value of the difference between the data value and the mean is greater than $2s$, then the word “outlier” will appear in the cell. We reference the mean and standard deviation values from the results of the Descriptive Statistics analysis. We can now easily identify outliers by looking for the word “outlier” in the adjacent column. Using just the column of GE data and this formula, we can observe that we have identified the same outlier point for GE (see Figure 7.11). (Another formula could have been used with the IF and OR functions as well.)

E5		fx =IF(ABS(C5- \$J\$5) >2*\$J\$9, "outlier", "")				
	A	C	E	F	I	J
1	Quarterly Stock Returns between 1995 and 2000					
2						
3	Date	GE			GE	
4	Q1 1995	-0.06				
5	Q2 1995	0.19	outlier		Mean	0.02
6	Q3 1995	-0.02			Standard Error	0.01
7	Q4 1995	0.04			Median	0.01
8	Q1 1996	-0.03			Mode	#N/A
9	Q2 1996	0.01			Standard Deviation	0.05
10	Q3 1996	0.00			Sample Variance	0.00
11	Q4 1996	0.03			Kurtosis	4.82
12	Q1 1997	-0.02			Skewness	1.69
13	Q2 1997	-0.03			Range	0.25
14	Q3 1997	0.07			Minimum	-0.06
15	Q4 1997	-0.02			Maximum	0.19
16	Q1 1998	0.08			Sum	0.43
17	Q2 1998	0.03			Count	24
18	Q3 1998	0.01				
19	Q4 1998	-0.02				
20	Q1 1999	0.07				
21	Q2 1999	0.02				
22	Q3 1999	0.02				
23	Q4 1999	0.04				
24	Q1 2000	0.03				
25	Q2 2000	0.00				
26	Q3 2000	-0.02				
27	Q4 2000	0.01				

Figure 7.11 Identifying the outlier by using a formula with the IF and ABS functions.

Another way to discover outliers is by using *Conditional Formatting* with the *Formula Is* option. With the formula below, we can simply select the column of values in our data set and fill in the *Conditional Formatting* dialog box to highlight outlier points:

$$=ABS(data_value - mean_value) > 2*s$$

Again, concerning the GE data, we can apply *Conditional Formatting* to identify the outliers as cells highlighted in red. In Figure 7.12, we demonstrate how we applied the *Formula Is* option.

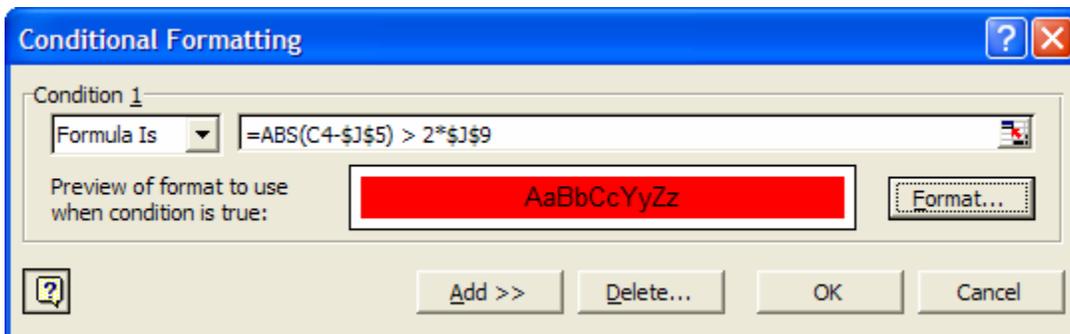


Figure 7.12 Applying the Formula Is option to the example data.

In Figure 7.13, we can observe that the same outlier point has been formatted.

	A	C	E	F	I	J
1	Quarterly Stock Returns between 1995 and 2000					
2						
3	Date	GE			GE	
4	Q1 1995	-0.06				
5	Q2 1995	0.19	outlier		Mean	0.02
6	Q3 1995	-0.02			Standard Error	0.01
7	Q4 1995	0.04			Median	0.01
8	Q1 1996	-0.03			Mode	#N/A
9	Q2 1996	0.01			Standard Deviation	0.05
10	Q3 1996	0.00			Sample Variance	0.00
11	Q4 1996	0.03			Kurtosis	4.82
12	Q1 1997	-0.02			Skewness	1.69
13	Q2 1997	-0.03			Range	0.25
14	Q3 1997	0.07			Minimum	-0.06
15	Q4 1997	-0.02			Maximum	0.19
16	Q1 1998	0.08			Sum	0.43
17	Q2 1998	0.03			Count	24
18	Q3 1998	0.01				
19	Q4 1998	-0.02				
20	Q1 1999	0.07				
21	Q2 1999	0.02				
22	Q3 1999	0.02				
23	Q4 1999	0.04				
24	Q1 2000	0.03				
25	Q2 2000	0.00				
26	Q3 2000	-0.02				
27	Q4 2000	0.01				

Figure 7.13 The outlier point is highlighted.

Let's now return to the *Descriptive Statistics* results to understand the remaining analysis values. *Kurtosis* is a measure of peakedness in the data. It compares the data peak to that of a Normal curve (which we will discuss in more detail in a later section). The *Skewness* is a measure of how symmetric or asymmetric data is. A *Skewness* value greater than +1 is the degree to which the data is skewed in the positive direction; likewise, a value less than -1 is the degree to which the data is skewed in the negative direction. A *Skewness* value between -1 and +1 implies symmetry. The *Skewness* values for MSFT and INTEL imply that their data is fairly symmetric; however, the *Skewness* value for GE is 1.69, which implies that it is skewed positively. That is, there is a peak early on in the data and then the data is stable (as we saw in the GE data graph in Figure 7.10).

The *Range* is the difference between the minimum and maximum value in the data set. The smaller this value is, the less variable is the data and therefore, the more desirable. The *Minimum*, *Maximum*, and *Sum* values are self-explanatory. The *Count* number reveals the quantity of values in the data set.

The last three options in the *Descriptive Statistics* dialog box, *Confidence Level for Mean*, *Kth Largest*, and *Kth Smallest*, can provide some extra information about our data. The *Confidence Level for Mean* calculates the mean value in the *Descriptive Statistics* report constrained to a specified confidence level. The mean is calculated using the specified confidence level (for example, 95% or 99%), the standard deviation, and the size of the sample data. The confidence level and the calculated mean are then added to the analysis report; we can compare the actual mean to this calculated mean based on the specified confidence level.

The *Kth Largest* and *Kth Smallest* options provide the respectively ranked data value for a specified value of k . For example, for $k = 1$, the *Kth Largest* returns the maximum data value and the *Kth Smallest* returns the minimum data value. The value of k can range from 1 to the number of data points in the input.

For example, let's create the same Descriptive Statistics report, but this time check the last three options of the initial dialog box (see Figure 7.14). We have specified a confidence interval of 95%, and we want to know the 2nd largest value and 2nd smallest value.

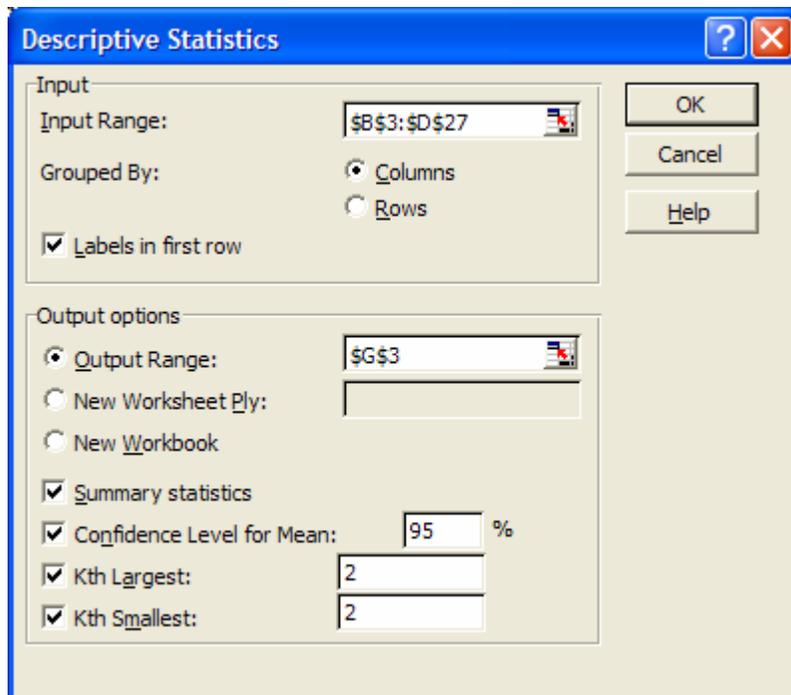


Figure 7.14 This time, specifying the last three options in the Descriptive Statistics dialog box.

In the new results, there are three additional rows of data (see Figure 7.15). Notice that there is a slight difference in the mean calculation with the enforced confidence level.

`=PERCENTRANK(B4:B27, 0.01)`

The result is that 0.388, or about 39 percent of the data, is less than the mean. These Excel functions, along with the others mentioned above, when combined with the Descriptive Statistics analysis tool, can help determine much constructive information about data.

 Summary	Descriptive Statistics	
	Outliers	Any values in the data set that lie more than $\pm 2s$ from the mean.
	PERCENTILE	A function that returns a value for which a desired percentile <i>k</i> of the specified <i>data_set</i> falls below.
	PERCENTRANK	A function that returns the percentile of the <i>data_set</i> that falls below a given <i>value</i> .

7.2.2 Histograms

Histograms calculate the number of occurrences, or frequency, with which values in a data set fall into various intervals. To create a histogram in Excel, we choose the *Histogram* option from the *Analysis Toolpack* list. A dialog box in which we will specify four main parameters then appears. These four parameters are: input, bins, output, and charts options (see Figure 7.16).

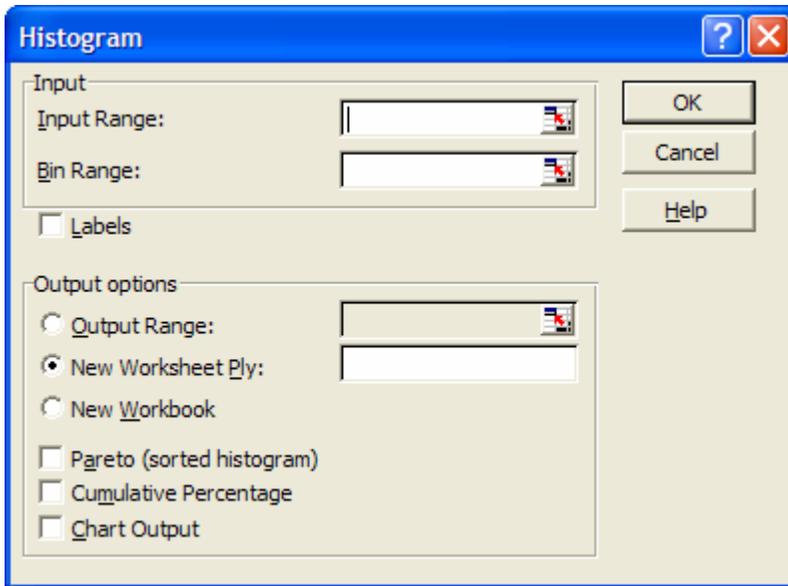


Figure 7.16 The Histogram dialog box.

The *Input Range* is the range of the data set. The *Bin Range* specifies the location of the bin values. **Bins** are the intervals into which values can fall; they can be defined by the user or can be evenly distributed among the data by Excel. If we specify our own bins, or intervals, then we must place them in a column on our worksheet. The bin values are specified by their upper bounds; for example, the intervals (0-10), (10-15), and (15-20)

are written as 10, 15, and 20. The Output Range is the location of the output, or the frequency calculations, for each bin. This location can be in the current worksheet or in a new worksheet or a new workbook. The chart options include a simple *Chart Output* (the actual histogram), a *Cumulative Percentage* for each bin value, and a *Pareto* organization of the chart. (Pareto sorts the columns from largest to smallest.)

Let's look at the MSFT stock return data from the examples above. We may want to determine how often the stock returns are at various levels. To do so, we go to *Tools > Data Analysis > Histogram* and specify the parameters of the *Histogram* dialog box (see Figure 7.17). Our *Input Range* is the column of MSFT data, including the "MSFT" label in the first row. For now, we leave the *Bin Range* blank and let Excel create the bins, or intervals. We check *Labels* since we have included a label for our selected data. We pick a cell in the current worksheet as our *Output Range* and then select *Chart Output*. The resulting histogram and frequency values are shown in Figure 7.18.

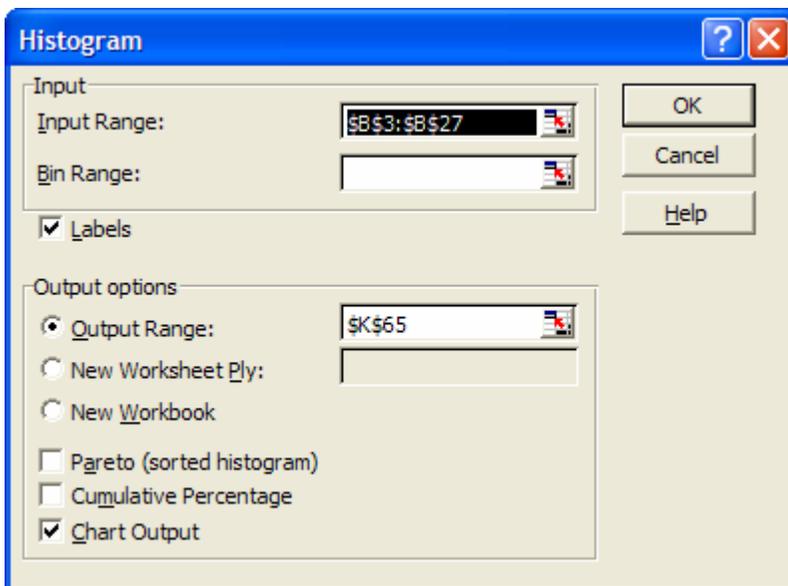


Figure 7.17 Entering data into the Histogram dialog box.

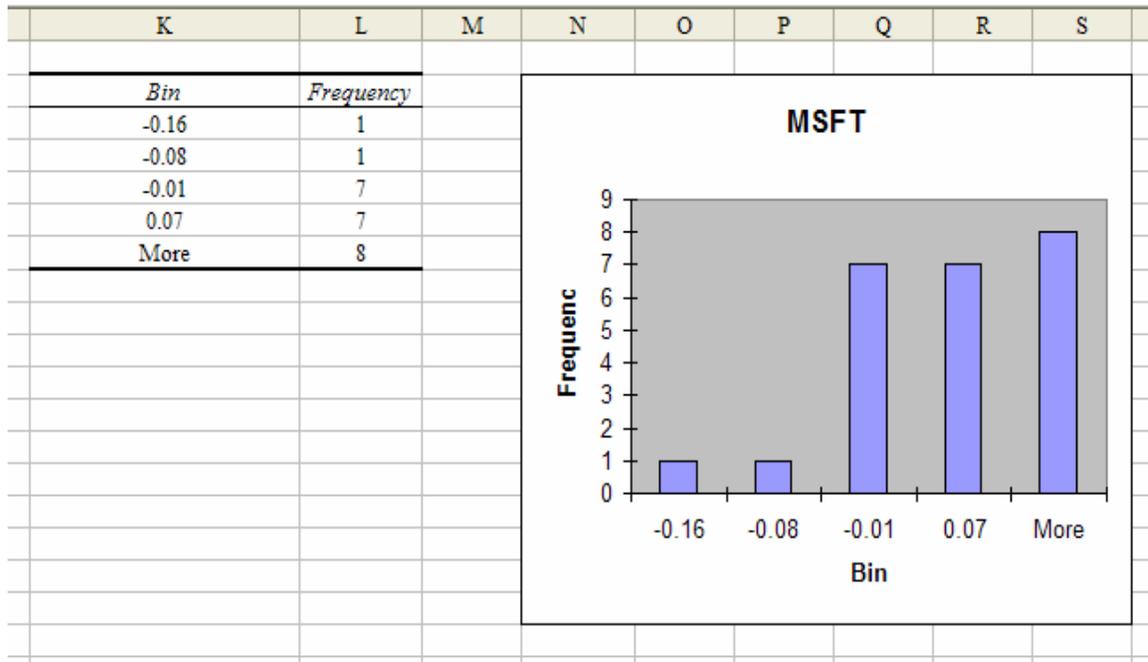


Figure 7.18 The resulting histogram and frequencies for the example data.

First, let's discuss the *Bin* values. Remember that each bin value is an upper bound on an interval; that is, the intervals that Excel has created for this example are (below -0.16), (-0.16, -0.08), (-0.08, -0.01), (-0.01, 0.07), and (above 0.07). We can deduce that most of our data values fall in the last three intervals. It may have been more useful to use intervals relative to the mean and standard deviation of the MSFT data. In other words, we could create the intervals (below -2s), (-2s, -s), (-s, mean), (mean, s), (s, 2s), and (above 2s). To enforce these intervals, we create our own *Bin Range*. In a new column, we list the upper bounds of these intervals using the mean and standard deviation values from the *Descriptive Statistics* results for the MSFT data. We also create a title for this column to include in the *Bin Range* (see Figure 7.19).

I
<i>MSFT Bins</i>
-0.15
-0.07
0.01
0.09
0.17

Figure 7.19 Creating the Bin Range for the example data.

We now choose *Tools > Data Analysis > Histogram* from the menu again and this time add the *Bin Range* (see Figure 7.20).

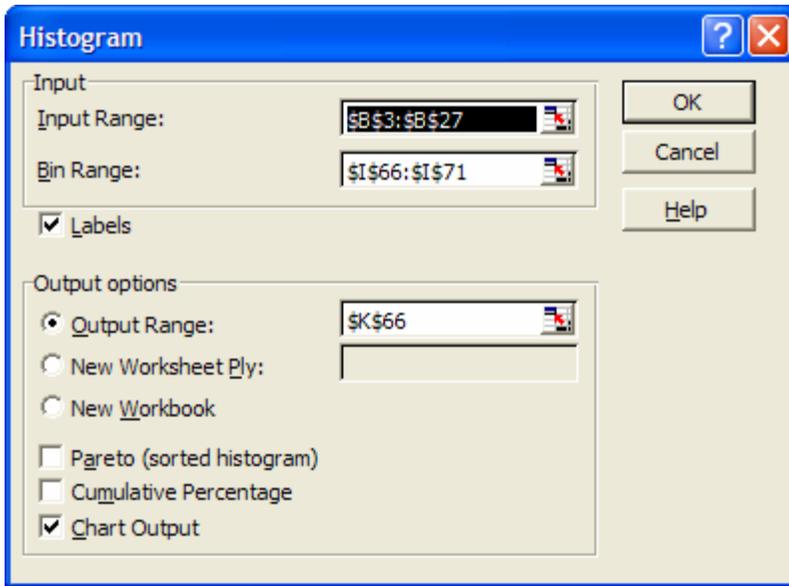


Figure 7.20 The *Histogram* dialog box now has a specified *Bin Range*.

Our Bin Range now calculates the frequencies and creates the histogram (see Figure 7.21). We can analyze this data to determine that the majority of our data lies above the mean (15 points above the mean versus 9 points below the mean). This conclusion validates the result of the PERCENTRANK function, as discussed in the previous section where we learned that 39 percent of the data values are below the mean; therefore 61 percent, or the majority, of our data is above the mean. We can also observe from this histogram result that there is one outlier; in other words, there is one data point that falls below -2s. We will perform some more analysis with these histogram results later in the chapter.

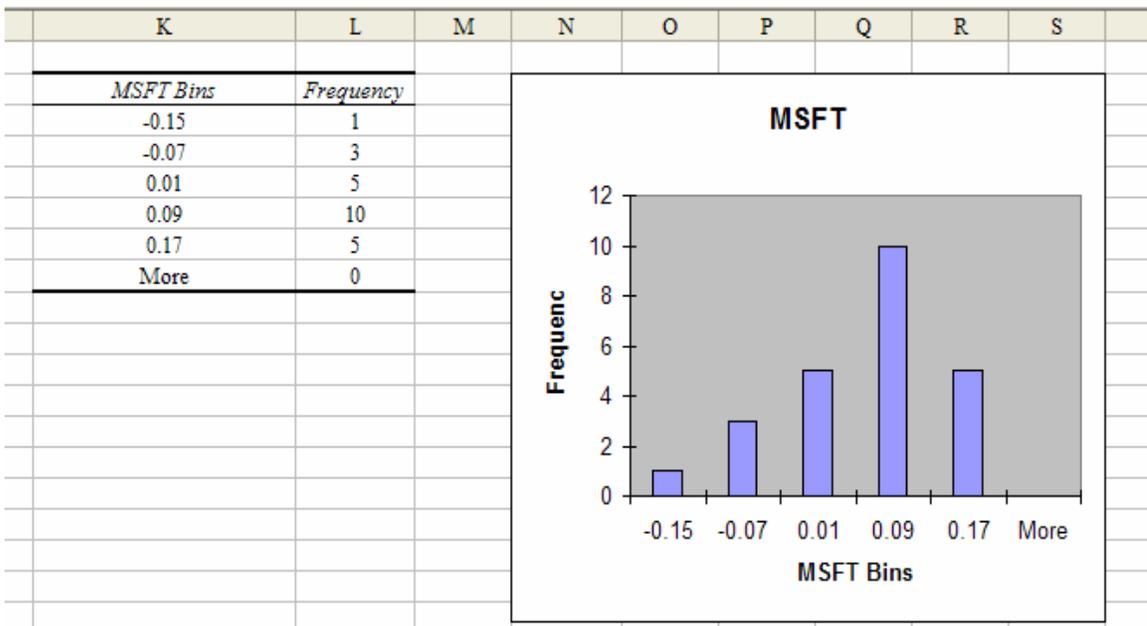


Figure 7.21 The resulting histogram uses the specified *Bin Range*.

A histogram can also be formatted. As with any chart, we right-click on the histogram and change the *Chart Options* or other parameters. For example, we have removed the *Legend* from the histograms shown above. If desired, we can also modify the font of the axis labels by right-clicking on the axis and choosing *Format Axis*.

We can also remove the gaps between the bars in the histogram to better recognize possible common distributions of the data. To remove these gaps, we right-click on a bar in the graph and select *Format Data Series* from the list of drop-down options. Then, we select *Options* and set the *Gap Width* to 0 (see Figure 7.22).

The histogram results can now be easily outlined to identify common distributions or other analyses (see Figure 7.23). We will discuss distributions later, but for now, let's define some common histogram shapes.

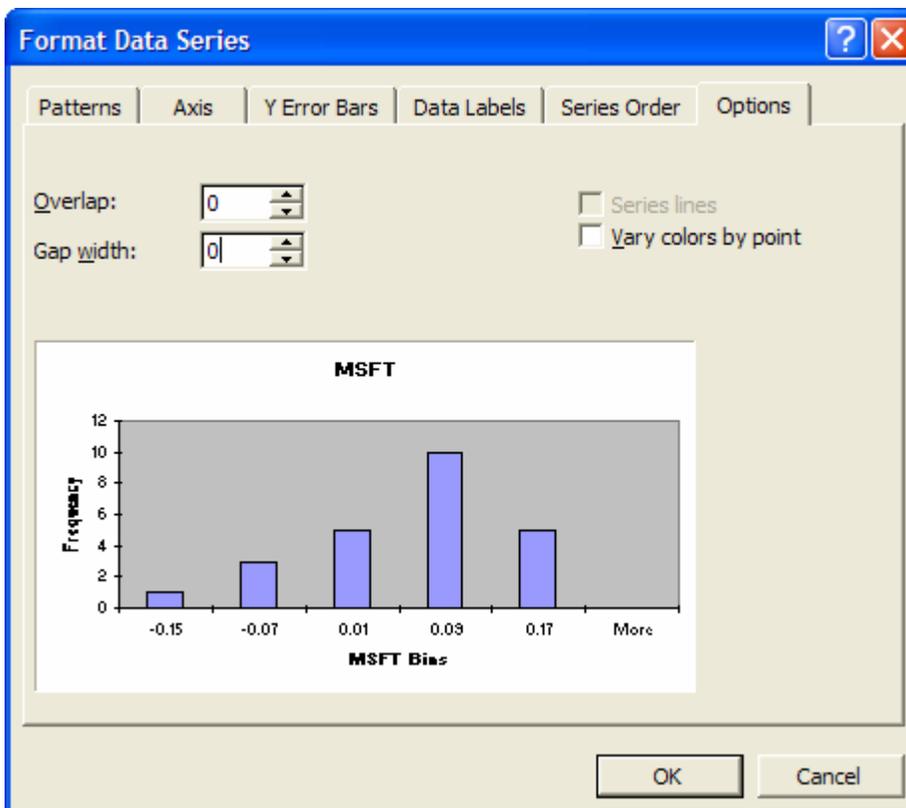


Figure 7.22 Removing the gaps by right-clicking on the bars, choosing *Format Data Series*, and setting the *Gap Width* to 0.

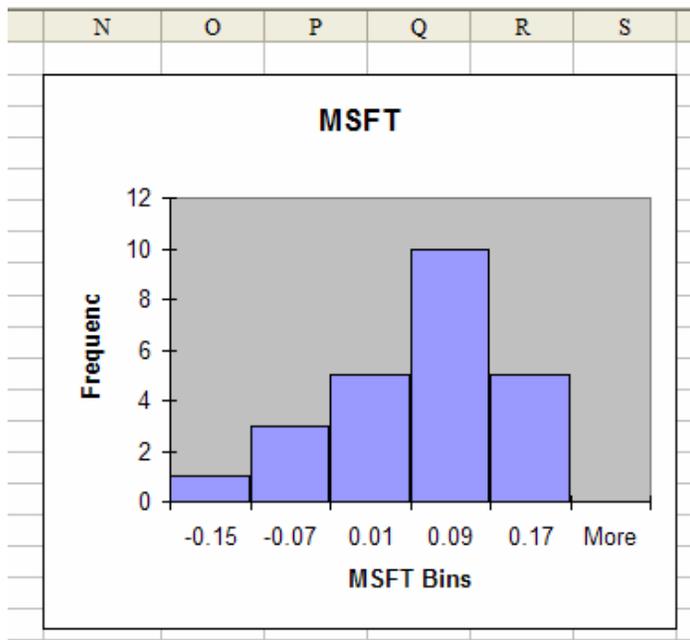


Figure 7.23 The histogram without gaps.

The histogram's four basic shapes are ***symmetric***, ***positively skewed***, ***negatively skewed***, and ***multiple peaks***. A histogram is symmetric if it has only one peak; that is, if there is a central high part and almost equal lower parts to the left and right of the peak. For example, test scores are commonly symmetric; they are sometimes referred to as a bell curve because of their symmetric shape.

A skewed histogram also only has one peak; however, the peak is not central, but far to the right with many lower points on the left, or far to the left with many lower points on the right. A positively skewed histogram has a peak on the left and many lower points (stretching) to the right. A negatively skewed histogram has a peak on the right and many lower points (stretching) to the left. Most economic data sets have skewed histograms.

Multiple peaks imply that more than one source, or population, of data is being evaluated.

In our example, the MSFT stock returns seem to be fairly symmetric. Remember, the Skewness value from the Descriptive Statistics analysis was also between -1 and 1. However, we can also observe that there is some negative skewness.

 Summary	Histograms	
	Bins	The intervals of values for which frequencies are calculated.
	Symmetric	A histogram with only one peak: a central high part with almost equal lower parts to the left and right of this peak.
	Positively Skewed	A histogram with a peak on the left and many lower points (stretching) to the right.
	Negatively Skewed	A histogram with a peak on the right and many lower points (stretching) to the left.
	Multiple Peaks	A histogram with multiple peaks suggests that more than one source, or population, of data is being evaluated.

7.3 Relationships in Data

It is often helpful to determine if any relationship exists among data. This calculation is usually accomplished by comparing data relative to other data. Some examples include analyzing product sales in relation to particular months, production rates in relation to the number of employees working, and advertising costs in relation to sales.

Relationships in data are usually identified by comparing two variables: the **dependent variable** and the **independent variable**. The dependent variable is the variable that we are most interested in. We may be trying to predict values for this variable by understanding its current behavior in order to better predict its future behavior. The independent variable is the variable that we use as the comparison in order to make the prediction. There may be various independent variables with known values that we can use to analyze the relationship against the dependent variable. However, only one independent variable provides the most accurate understanding of the dependent variable's behavior.

Note that since our prediction of the dependent variable relies on the independent variable, we can only predict values corresponding to known independent variable values. In other words, we cannot predict beyond the scope of available independent variable data, nor can we predict the independent variable itself.

We can graph this data (with the *XY Scatter* chart type) by placing the independent variable on the x-axis and the dependent variable on the y-axis and then using a tool in Excel called a **trend curve** to determine if any relationship exists between these variables.

 Summary	Dependent variable	The variable that a user is trying to predict or understand.
	Independent variable	The variable used to make predictions.
	Trend Curve	The curve on a graph of data, with the independent variable on the x-axis and the dependent variable on the y-axis; it estimates the behavior of the dependent variable.

7.3.1 Trend Curves

To add a trend curve to our chart, we right-click on the data points in our *XY Scatter* chart and choose *Add Trendline* from the drop-down list of options. There are five basic trend curves that Excel can model: **Linear**, **Exponential**, **Power**, **Moving Average**, and **Logarithmic**. Each of these curves is illustrated in the *Add Trendline* dialog box, which appears in Figure 7.24.

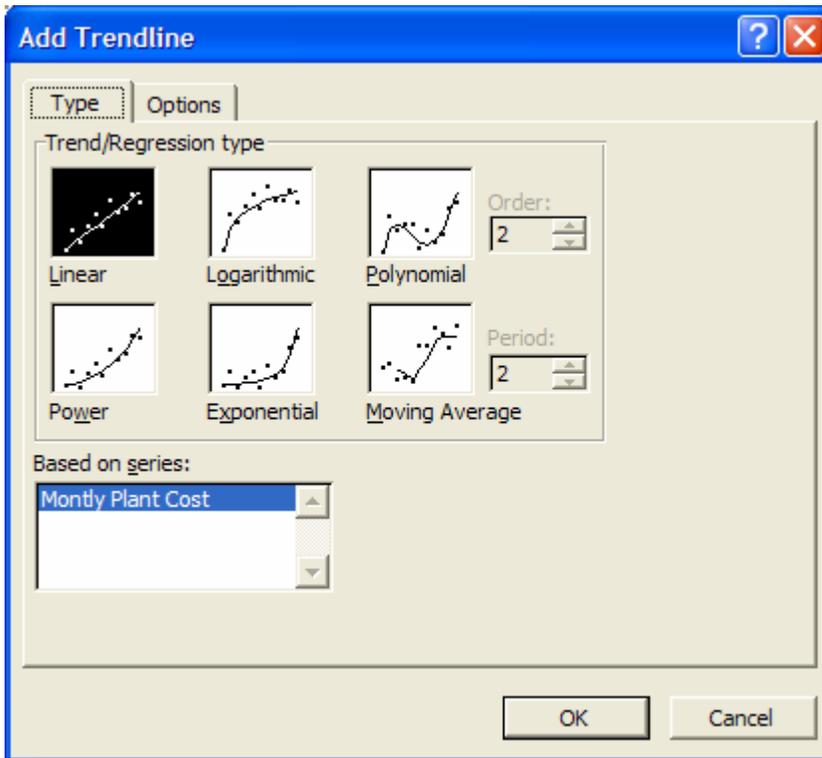


Figure 7.24 The five trend curves that Excel can fit to data.

We will discuss how to identify linear, exponential, and power curves in a chart. If a graph looks like a straight line would run closely through the data points, then a linear curve is best. If the dependent variable (on the y-axis) appears to increase at an increasing rate, then the exponential curve is more favorable. Similar to the exponential curve is the power curve; however, the power curve has a slower rate of increase in terms of the dependent variable.

Depending on which curve we select, Excel fits this type of trend curve to our data and creates a **trendline** in the chart. There are different equations for each trend curve used to create the trendline based on our data. We will discuss this in more detail later. For *Linear* trend curves, Excel produces the “best fitting” trendline of the selected trend curve by minimizing the sum of the squared vertical distances from each data point to the trendline. This vertical distance is called the error, or **residual**. A positive error implies that a point lies above the line, and a negative error implies that a point lies below the line. This trendline is therefore referred to as the **least squares line**.

After we select the curve that we feel best fits our data, we click on the *Options* tab (see Figure 7.25). The first option to set is the trendline's name; we can either use the automatic name (default) or create a custom name. The next option is to specify a period forward or backward for which we want to predict the behavior of our dependent variable. This period is in units of our independent variable. This is a very useful tool since it is one of the main motivations for using trend curves. The last set of options allows us to specify an intercept for the trendline and to display the trendline equation and the R-squared value on the chart. We will usually not check to Set Intercept; however, we always recommend checking to Display Equation and Display R-Squared Value. We will discuss the equation and the R-squared value for each trend curve in more detail later.

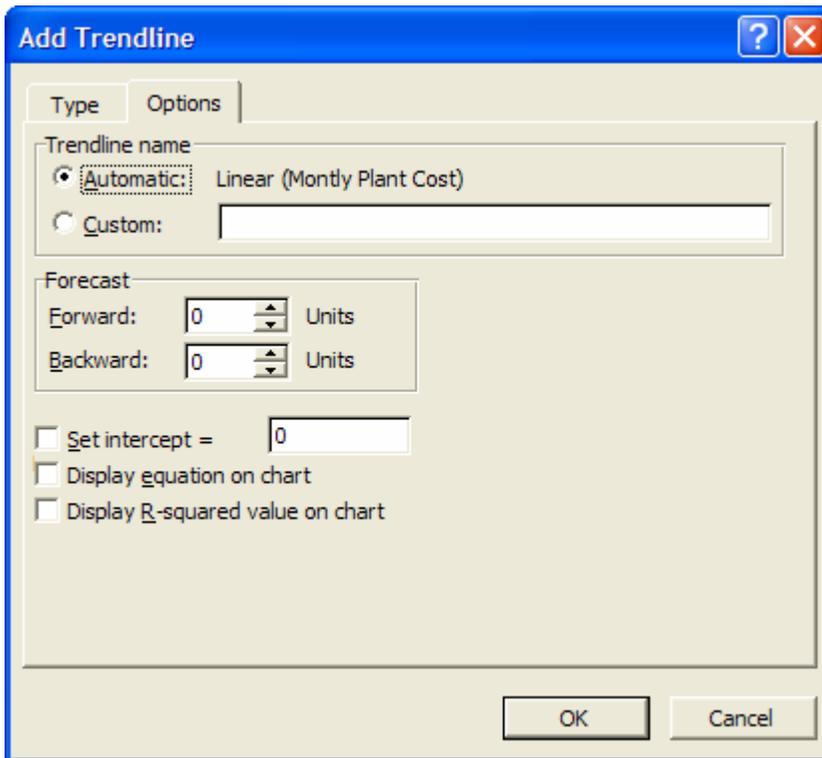


Figure 7.25 The Options tab of the Add Trendline dialog box.

We can also right-click on any trendline after it has been created and choose *Format Trendline* from the list of options. This selection allows us to modify the *Type* and *Options* initially specified as well as to change any *Patterns* on the trendline (see Figure 7.26).

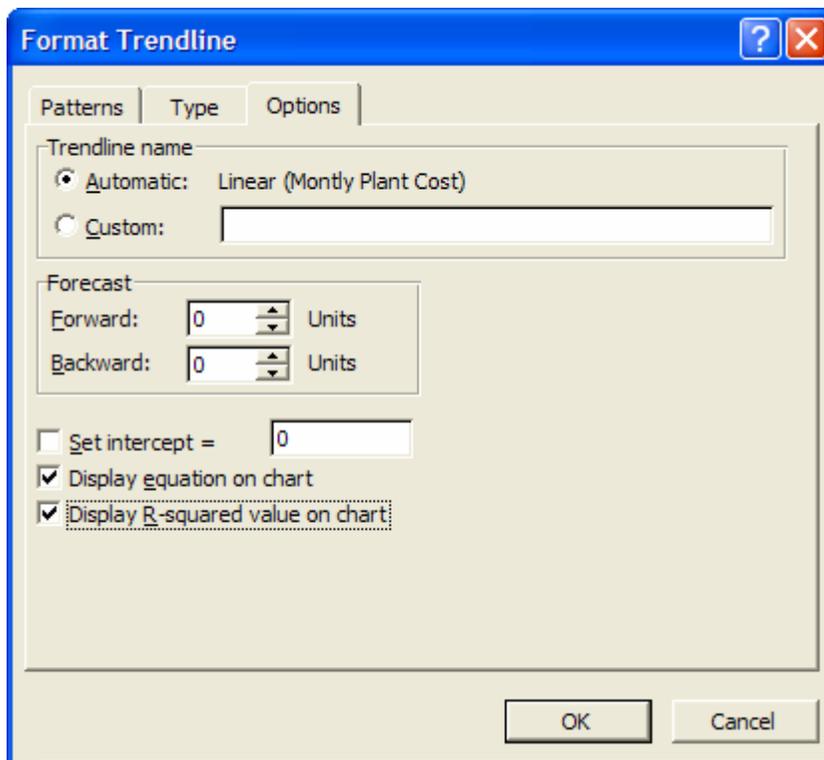


Figure 7.26 Right-clicking on a trendline to format it or change Type or Options.

Let's compare some examples of these three different trend curves. We will begin with *Linear* curves. Suppose a company has recorded the number of "Units Produced" each month and the corresponding "Monthly Plant Cost" (see Figure 7.27). The company may be able to accurately determine how much they will produce each month; however, they want to be able to estimate their plant costs based on this production amount. They will therefore need to determine, first of all, if there is a relationship between "Units Produced" and "Monthly Plant Cost." If so, then they need to establish what type of relationship it is in order to accurately predict future monthly plant costs based on future unit production.

The dependent variable is therefore the "Monthly Plant Cost" and the independent variable is the "Units Produced." We begin this analysis by making an XY Scatter chart of the data (with the dependent variable on the y-axis and the independent variable on the x-axis). Figure 7.28 displays this chart of "Monthly Plant Cost per Units Produced."

	A	B	C
1	Production Cost		
2			
3	Month	Units Produced	Montly Plant Cost
4	1	1260	\$99,850
5	2	1007	\$58,096
6	3	1096	\$96,360
7	4	873	\$65,675
8	5	532	\$51,870
9	6	476	\$27,462
10	7	482	\$27,808
11	8	1173	\$110,118
12	9	692	\$67,470
13	10	690	\$39,808
14	11	564	\$32,538
15	12	470	\$45,825

Figure 7.27 A record of the “Units Produced” and the “Monthly Plant Cost” for twelve months.

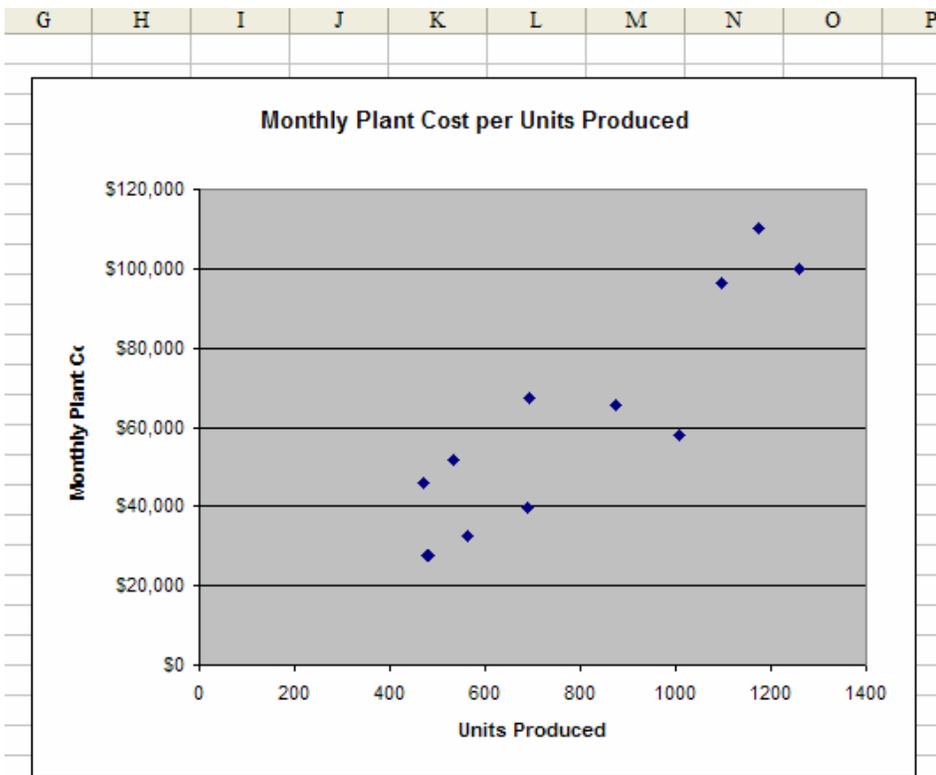


Figure 7.28 The XY Scatter Chart for the “Monthly Plant Cost per Units Produced.”

We can now right-click on any of the data points and choose *Add Trendline* from the list of drop-down options (see Figure 7.29). The *Linear* trend curve seems to fit this data best. (You might also think the *Power* trend curve fits well. It is okay to try different trend curves to evaluate which gives you the most accurate relationship for predictions.) We select *Linear* from the *Type* tab and then select *Display Equation on Chart* from the *Options* tab (see Figure 7.30).

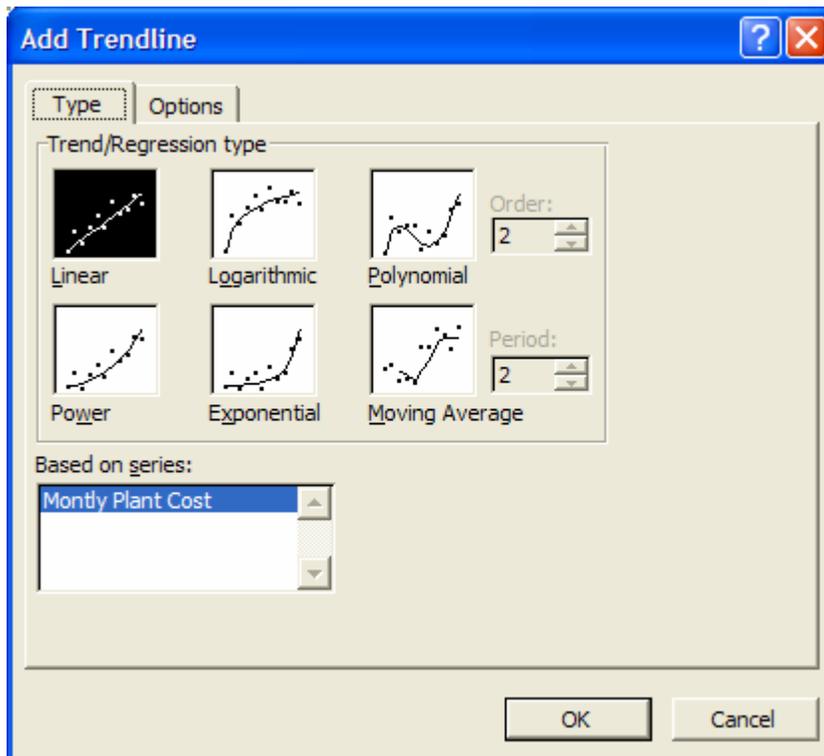


Figure 7.29 Selecting the Linear trend curve from the *Type* tab.

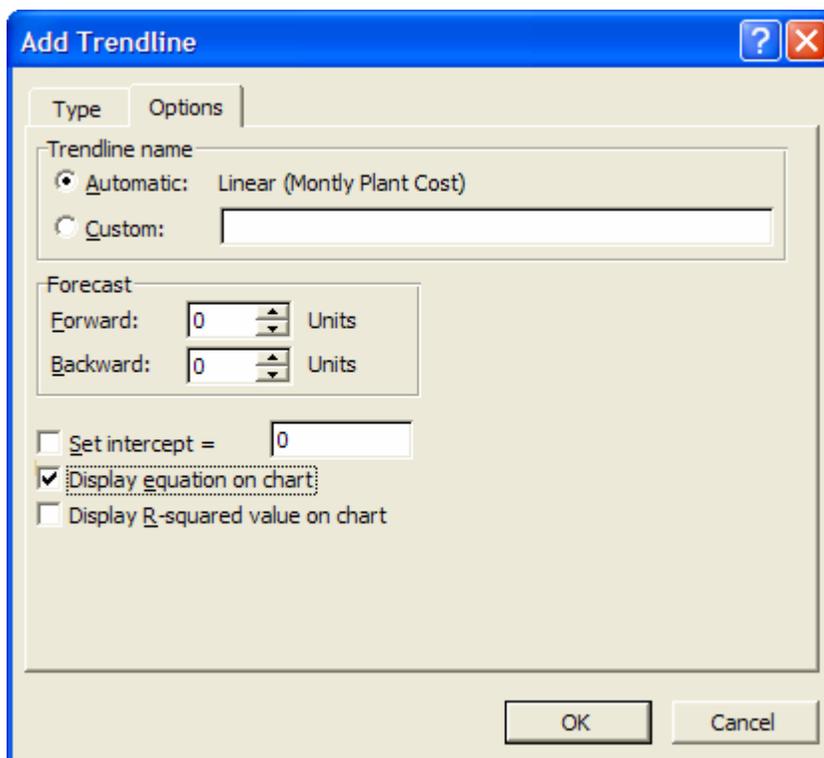


Figure 7.30 Checking the Display Equation on the Chart option.

The trendline and the equation are then added to our chart, as illustrated in Figure 7.31.

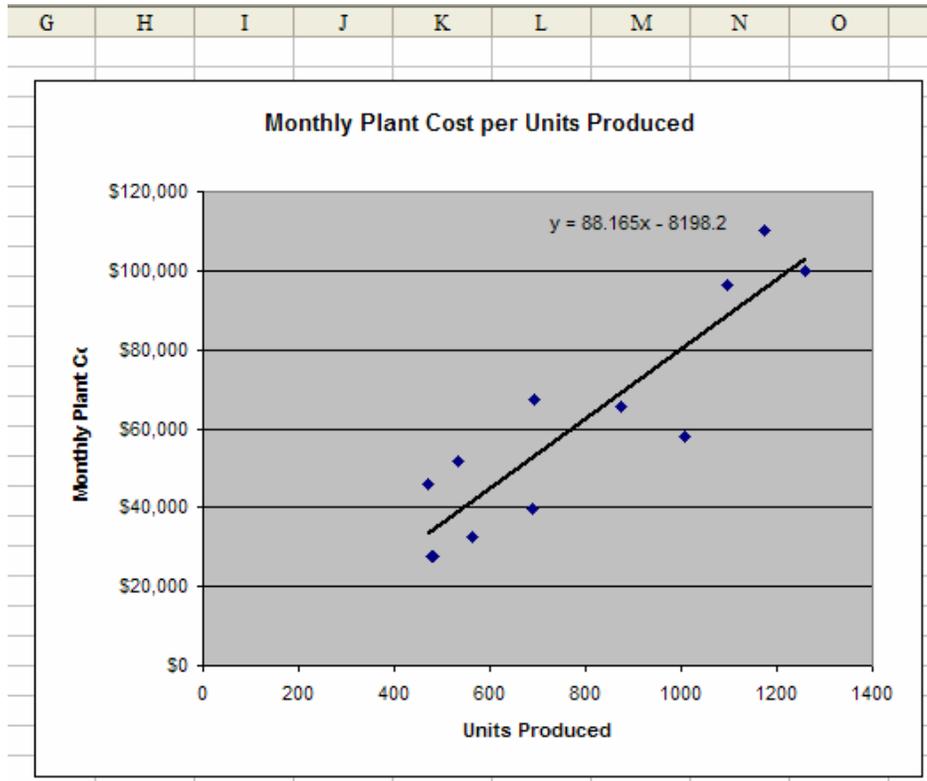


Figure 7.31 Adding the Linear trendline to the chart.

Let’s now decipher what the trendline equation is. The x variable is the independent variable, in this example, the “Units Produced.” The y variable is the dependent variable, in this example, the “Monthly Plant Cost.” This equation suggests that for any given value of x, we can compute y. That is, for any given value of “Units Produced,” we can calculate the expected “Monthly Plant Cost.” We can therefore transfer this equation into a formula in our spreadsheet and create a column of “Predicted Cost” relative to the values from the “Units Produced” column.

In Figure 7.32 the following formula operates in the “Predicted Cost” column:

$$=88.165*B4 - 8198.2$$

We copy this formula for the entire “Predicted Cost” column using relative referencing for each value in the “Units Produced” column. We then create an “Error” column, which simply subtracts the “Predicted Cost” values from the actual “Monthly Plant Cost” values. As the figure suggests, there is always some error since the actual data does not lie on a straight line. (Again, you could try calculating the “Predicted Costs” using a *Power* trend curve to compare the “Error” values.)

E4		fx =88.165*B4 - 8198.2				
	A	B	C	D	E	F
1	Production Cost					
2						
3	Month	Units Produced	Monthly Plant Cost		Predicted Cost	Error
4	1	1260	\$99,850		\$102,890	\$3,040
5	2	1007	\$58,096		\$80,584	\$22,488
6	3	1096	\$96,360		\$88,431	-\$7,929
7	4	873	\$65,675		\$68,770	\$3,095
8	5	532	\$51,870		\$38,706	-\$13,164
9	6	476	\$27,462		\$33,768	\$6,307
10	7	482	\$27,808		\$34,297	\$6,490
11	8	1173	\$110,118		\$95,219	-\$14,898
12	9	692	\$67,470		\$52,812	-\$14,658
13	10	690	\$39,808		\$52,636	\$12,828
14	11	564	\$32,538		\$41,527	\$8,988
15	12	470	\$45,825		\$33,239	-\$12,586
16						

Figure 7.32 Adding the “Predicted Cost” and “Error” columns to the table using the Linear trendline equation.

Now we have enough information to address the initial problem for this example: predicting future “Monthly Plant Costs” based on planned production amounts. In Figure 7.33, we have added “Units Produced” values for three more months. Copying the formula for “Predicted Cost” to these three new rows gives us the predicted monthly costs.

	A	B	C	D	E
1	Production Cost				
2					
3	Month	Units Produced	Monthly Plant Cost		Predicted Cost
4	1	1260	\$99,850		\$102,890
5	2	1007	\$58,096		\$80,584
6	3	1096	\$96,360		\$88,431
7	4	873	\$65,675		\$68,770
8	5	532	\$51,870		\$38,706
9	6	476	\$27,462		\$33,768
10	7	482	\$27,808		\$34,297
11	8	1173	\$110,118		\$95,219
12	9	692	\$67,470		\$52,812
13	10	690	\$39,808		\$52,636
14	11	564	\$32,538		\$41,527
15	12	470	\$45,825		\$33,239
16	13	520			\$37,648
17	14	670			\$50,872
18	15	642			\$48,404
19					

Figure 7.33 Calculating the “Predicted Cost” for the next three months.

(Note that the independent variable must be known for the time period for which we want to predict the dependent variable values. That is, we are only able to predict the monthly

costs for this example because we assume that the production amount is known for the next three months.)

Now let's discuss *Exponential* trend curves. In Figure 7.34, we have "Sales" data for ten years. If we want to be able to predict sales for the next few years, we must determine what relationship exists between these two variables. So, our independent variable is "Years" and our dependent variable is "Sales."

	A	B	C
1	Sales Data		
2			
3	Year	Sales	
4	1	70	
5	2	183	
6	3	340	
7	4	649	
8	5	1243	
9	6	1979	
10	7	4096	
11	8	6440	
12	9	8459	
13	10	12154	

Figure 7.34 Sales per year.

After creating the *XY Scatter* chart of this data (x-axis as "Year," y-axis as "Sales"), we right-click on a data point to add the trendline (see Figure 7.35). This time, we choose an *Exponential* curve to fit our data. (Again the *Power* curve seems like another possible fit that we could test.) We also choose to display the trendline equation on the chart. Figure 7.36 displays the resulting chart with the trendline.

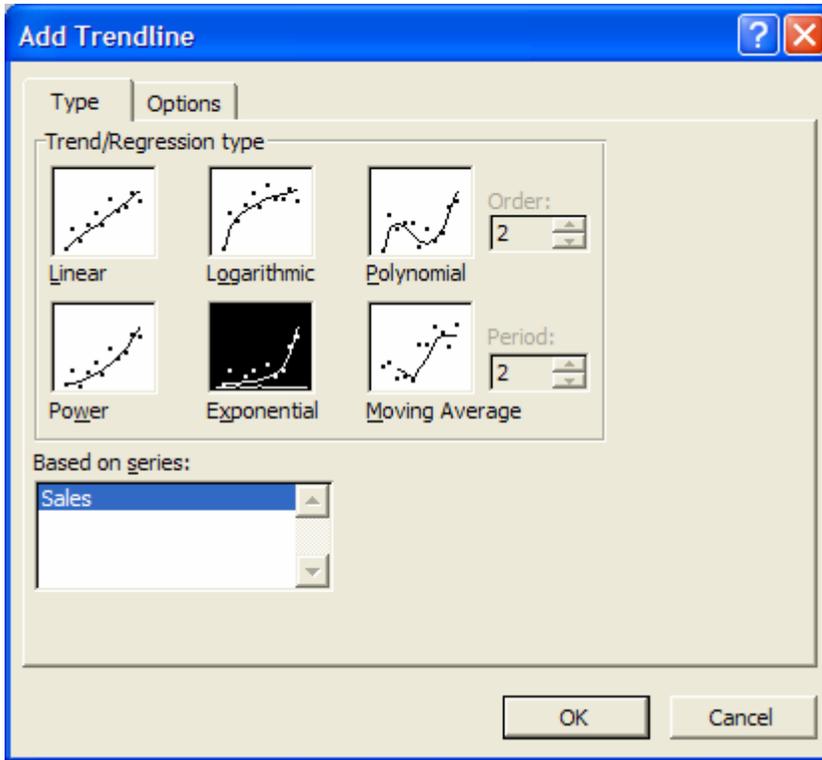


Figure 7.35 Choosing the Exponential trend curve.

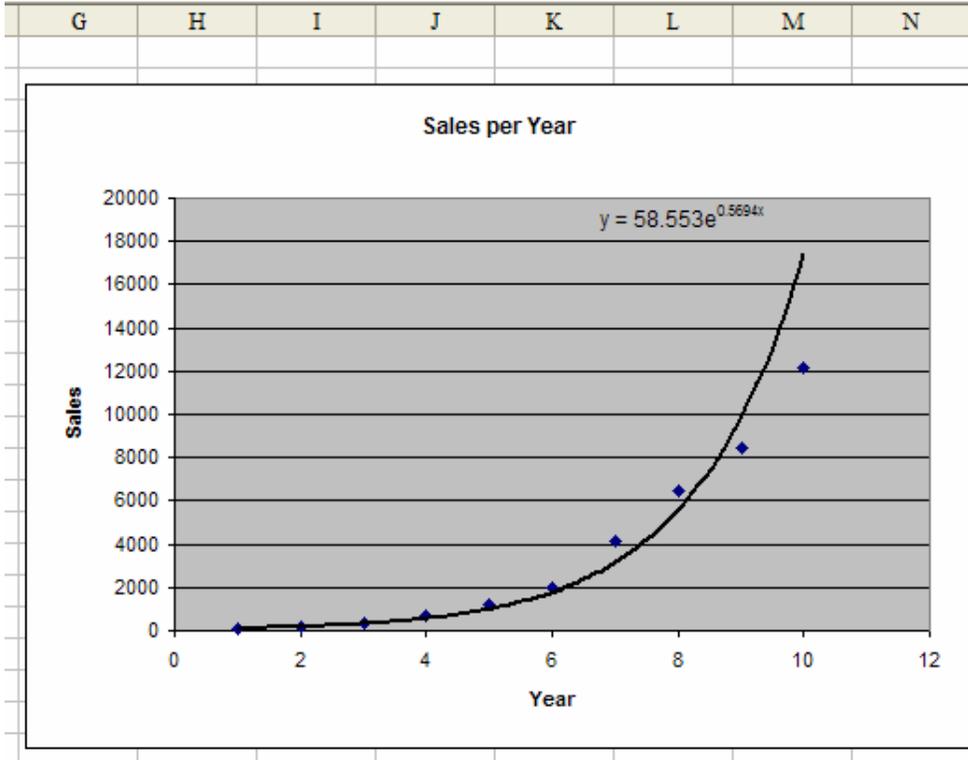


Figure 7.36 Adding the Exponential trendline to the charted data.

Let's analyze the equation provided on the chart. Again, the *y* variable represents the dependent variable, in this example, "Sales." The *x* variable represents the independent variable, in this example, "Year." We can therefore transform this equation into a formula in our spreadsheet and create a "Prediction" column in which we estimate sales based on the year. In Figure 7.37, we have done so using the following formula:

$$=58.553*EXP(0.5694*A4)$$

The EXP function raises *e* to the power in parentheses. We have copied this formula for all of the years provided in order to compare our estimated values to the actual values. Notice that there are some larger "Error" values as the years increase.

D4		fx =58.553*EXP(0.5694*A4)			
	A	B	C	D	E
1	Sales Data				
2					
3	Year	Sales		Prediction	Error
4	1	70		103.48	33.48
5	2	183		182.86	-0.14
6	3	340		323.16	-16.84
7	4	649		571.08	-77.92
8	5	1243		1009.22	-233.78
9	6	1979		1783.50	-195.50
10	7	4096		3151.81	-944.19
11	8	6440		5569.90	-870.10
12	9	8459		9843.16	1384.16
13	10	12154		17394.90	5240.90

Figure 7.37 Calculating the "Prediction" values with the *Exponential* trendline equation.

We can now use this formula to predict sales values for future years. However, the *Exponential* trend curve has a sharply increasing slope that may not be accurate for many situations. For example, in six years from our current data, year 16, we have estimated about 530,000 sales using the *Exponential* trendline equation. This amount seems a highly unlikely number given previous historical data (see Figure 7.38). Even though the *Exponential* trend curve increases rapidly towards infinity, it is unlikely that sales will do the same. Therefore, for predicting values much further in the future, we may consider using a different trend curve (perhaps the *Power* curve).

D15		fx =58.553*EXP(0.5694*A15)			
	A	B	C	D	E
1	Sales Data				
2					
3	Year	Sales		Prediction	Error
4	1	70		103.48	33.48
5	2	183		182.86	-0.14
6	3	340		323.16	-16.84
7	4	649		571.08	-77.92
8	5	1243		1009.22	-233.78
9	6	1979		1783.50	-195.50
10	7	4096		3151.81	-944.19
11	8	6440		5569.90	-870.10
12	9	8459		9843.16	1384.16
13	10	12154		17394.90	5240.90
14	...				
15	16			529841.05	
16					

Figure 7.38 Using the Exponential trendline equation to predict sales for year 16.

Now let’s consider an example of a *Power* trend curve. In Figure 7.39, we are presented with yearly “Production” and the yearly “Unit Cost” of production. We want to determine the relationship between “Unit Cost” and “Production” in order to be able to predict future “Unit Costs.”

	A	B	C	D
1	Production Data			
2				
3	Year	Production	Unit Cost	
4	1982	64000	\$ 3,700.00	
5	1983	70000	\$ 3,416.00	
6	1984	100000	\$ 3,125.00	
7	1985	150000	\$ 2,583.00	
8	1986	175000	\$ 2,166.00	
9	1987	400000	\$ 1,833.00	
10	1988	785000	\$ 1,788.00	

Figure 7.39 Yearly Production and Unit Costs.

We begin by creating the XY Scatter chart and then right-clicking on a data point to add a trendline. This time we choose a *Power* curve to fit the data (see Figure 7.40). (*Exponential* may also be an appropriate fit for this data, but the slope of the recorded data points does not seem to be that steep.) Even though our data is decreasing, not increasing, it is the slope of the data points that we are observing in order to find a suitable fit. Again, we choose to display the trendline equation with the *Options* tab. Figure 7.41 demonstrates the resulting trendline with the charted data points.

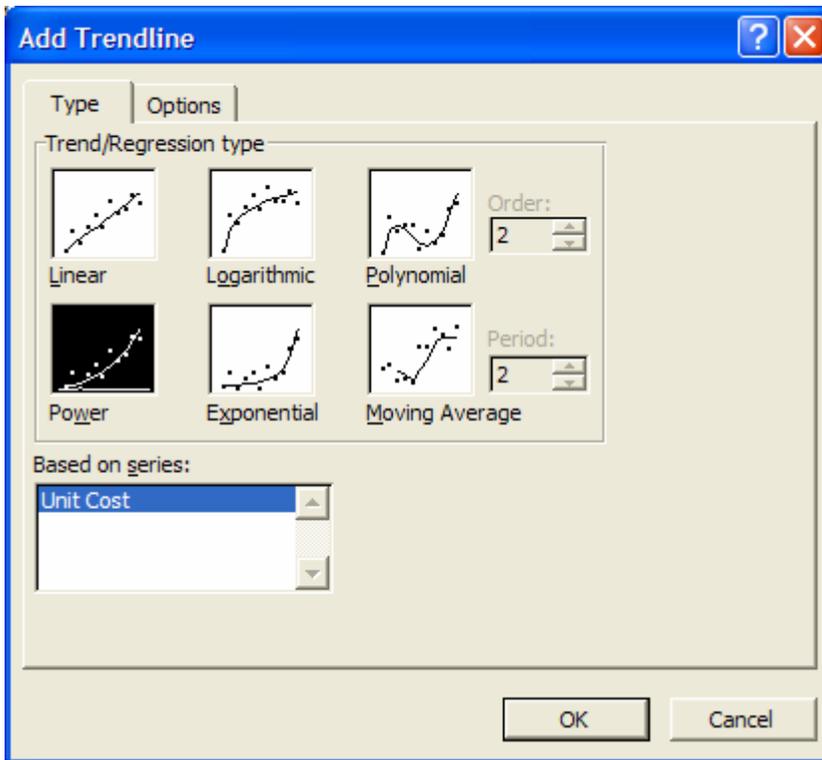


Figure 7.40 Choosing the *Power* curve.

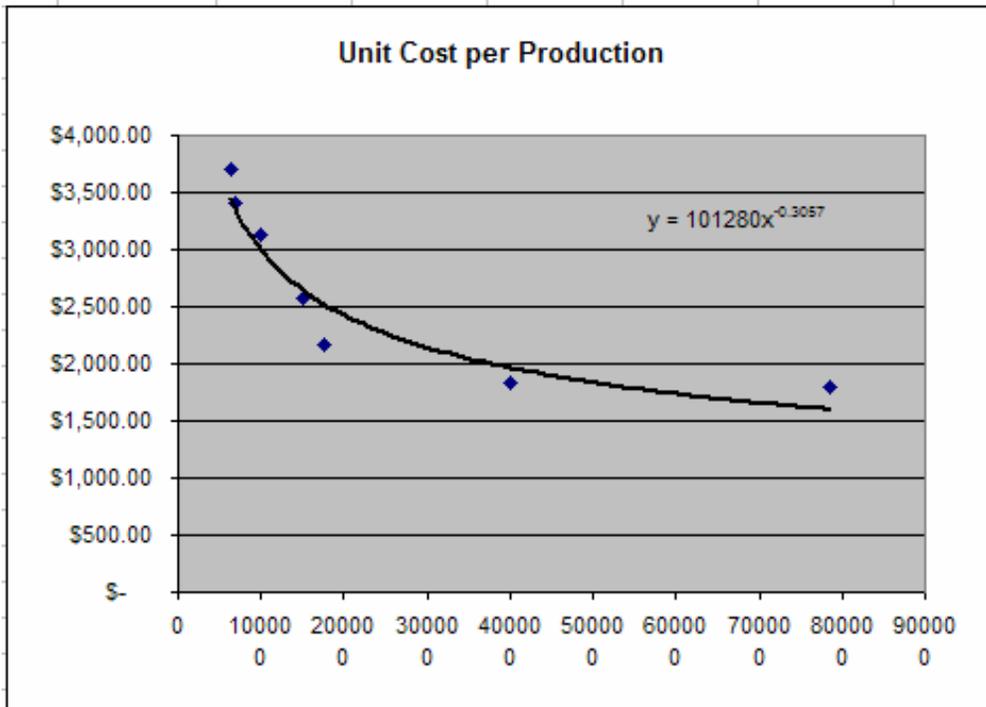


Figure 7.41 Fitting the *Power* curve to the “Unit Cost per Cumulative Production” chart.

Looking at the *Power* trendline equation, we again identify x to be the independent variable, in this case, “Production,” and y to be the dependent variable, in this case, the “Unit Cost.” We transform this equation into a formula on the spreadsheet in a “Forecast” column to compare our estimated values with the actual costs. We copy the following formula for all of the given years:

$$=101280*B4^{-0.3057}$$

Figure 7.42 displays these forecasted cost values and the “Error” calculated between the forecasted and actual data. The error values seem to be fairly stable, therefore implying a reliable fit.

E4		fx =101280*B4^-0.3057				
	A	B	C	D	E	F
1	Production Data					
2						
3	Year	Production	Unit Cost		Forecast	Error
4	1982	64000	\$ 3,700.00		\$3,437.75	\$ 262.25
5	1983	70000	\$ 3,416.00		\$3,344.85	\$ 71.15
6	1984	100000	\$ 3,125.00		\$2,999.33	\$ 125.67
7	1985	150000	\$ 2,583.00		\$2,649.67	\$ 66.67
8	1986	175000	\$ 2,166.00		\$2,527.71	\$ 361.71
9	1987	400000	\$ 1,833.00		\$1,963.24	\$ 130.24
10	1988	785000	\$ 1,788.00		\$1,597.58	\$ 190.42
11						

Figure 7.42 Creating the “Forecast” and “Error” columns with the *Power* trendline equation.

We would now like to make a note about using data with dates (for example the “Year” in the above example). If dates are employed as an independent variable, we must convert them into a simple numerical list. For example, if we had chosen to assign the “Year” column in the above example as an independent variable for predicting the “Unit Cost,” we would have had to renumber the years from 1 to 7, 1 being the first year, 2 the second, etc., in which the data was collected. Using actual dates may yield inaccurate calculations.

 <p>Summary</p>	Trend Curves	
	Linear Curve	$y = a*x - b$
	Exponential Curve	$y = a*e^{(b*x)}$ or $y = a*EXP(b*x)$
	Power Curve	$y = a*x^b$
	Residual	The vertical distance, or error, between the trendline and the data points.
	Least Squares Line	The trendline with the minimum residual.

7.3.2 Regression

Another more accurate way to ensure that the relationships we have chosen for our data are reliable fits is by using regression analysis parameters. These parameters include the **R-Squared value**, **standard error**, **slope** and **intercept**.

The R-Squared value measures the amount of influence that the independent variable has on the dependent variable. The closer the R-Squared value is to 1, the stronger the linear relationship between the independent and dependent variables is. If the R-Squared value is closer to 0, then there may not be a relationship between them. We can then draw on multiple regression and other tools to determine a better independent variable to predict the dependent variable.

To determine the R-Squared value of a regression, or a trendline, we can use the *Add Trendline* dialog box on a chart of data and specify to *Display R-Squared Value on Chart* in the *Options* tab (see Figure 7.43).

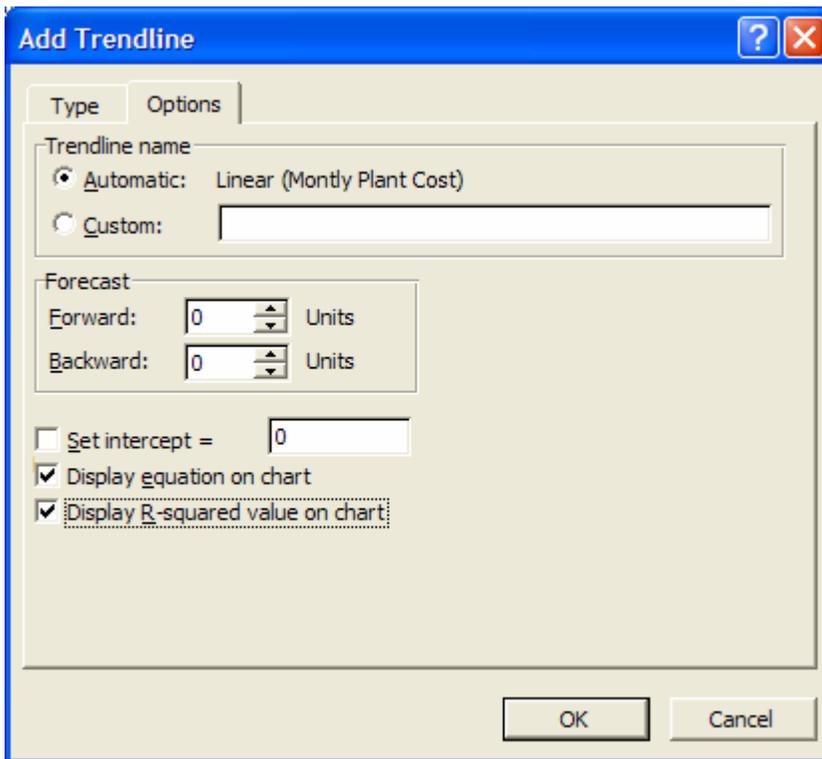


Figure 7.43 Checking the *Display R-Squared Value on Chart* option.

Let's review the previous three examples to discover their R-Squared values. We have gone back to our charts and added the R-Squared display option by right-clicking on the trendline previously created. We then *Format Trendline* to revisit the *Options* tab and specify this new option.

For the first example, we fit a *Linear* trendline to the “Monthly Plant Cost per Units Produced” chart (see Figure 7.44). The R-Squared value is 0.8137, which is fairly close to 1. We could try other trend curves and compare the R-Squared values to determine which fit is the best.

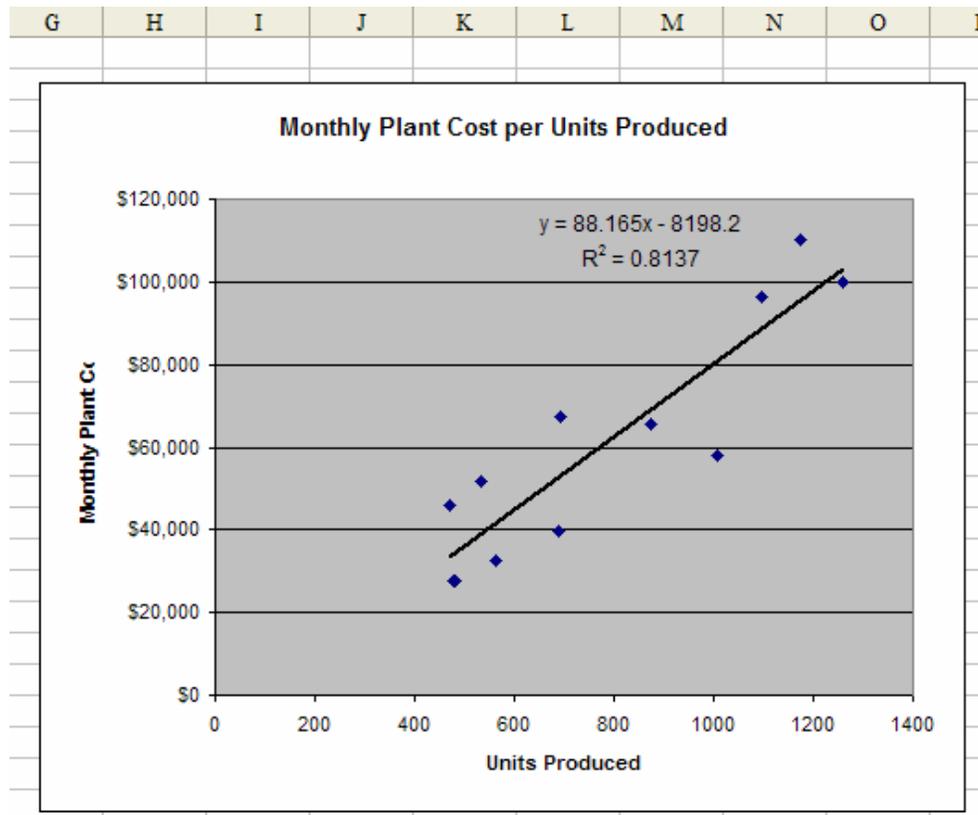


Figure 7.44 The R-Squared value on the Linear trendline.

In the following example, we fit an *Exponential* trendline to the “Sales per Year” chart (see Figure 7.45). The R-Squared value for this data is 0.9828. This value is very close to 1 and therefore a sound fit. Again, it is wise to compare the R-Squared values for *Exponential* and *Power* curves on a set of data with an increasing slope.

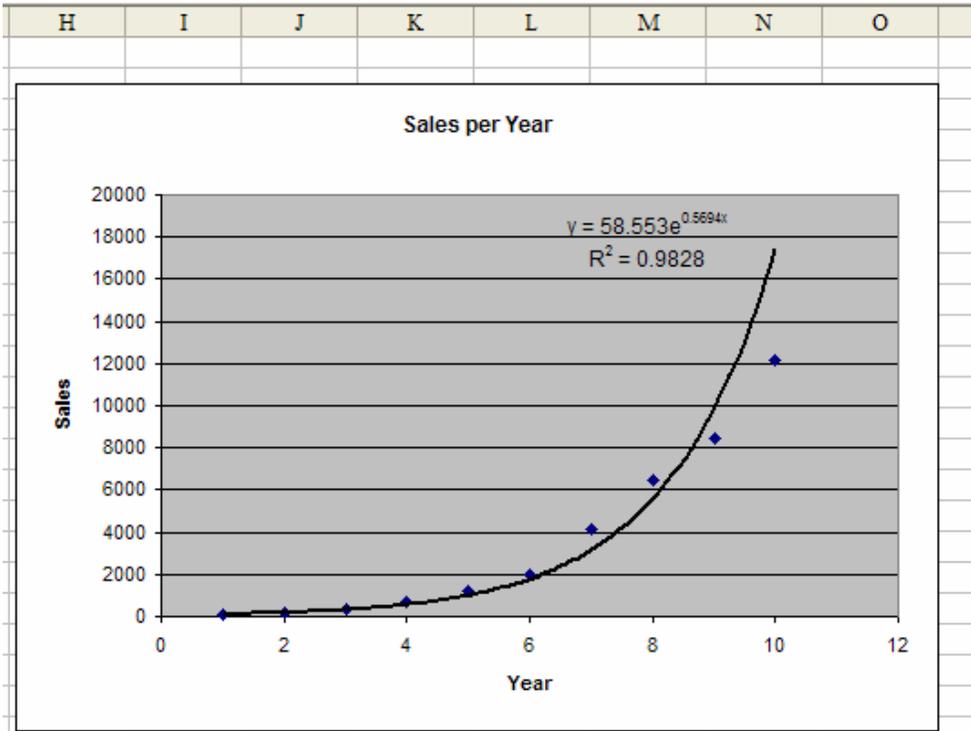


Figure 7.45 The R-Squared value for the Exponential trendline.

In the last example, we fit a *Power* trendline to the “Unit Cost per Cumulative Production” chart (see Figure 7.46). The R-Squared value is 0.9485, which is also very close to 1 and therefore an indication of a good fit.

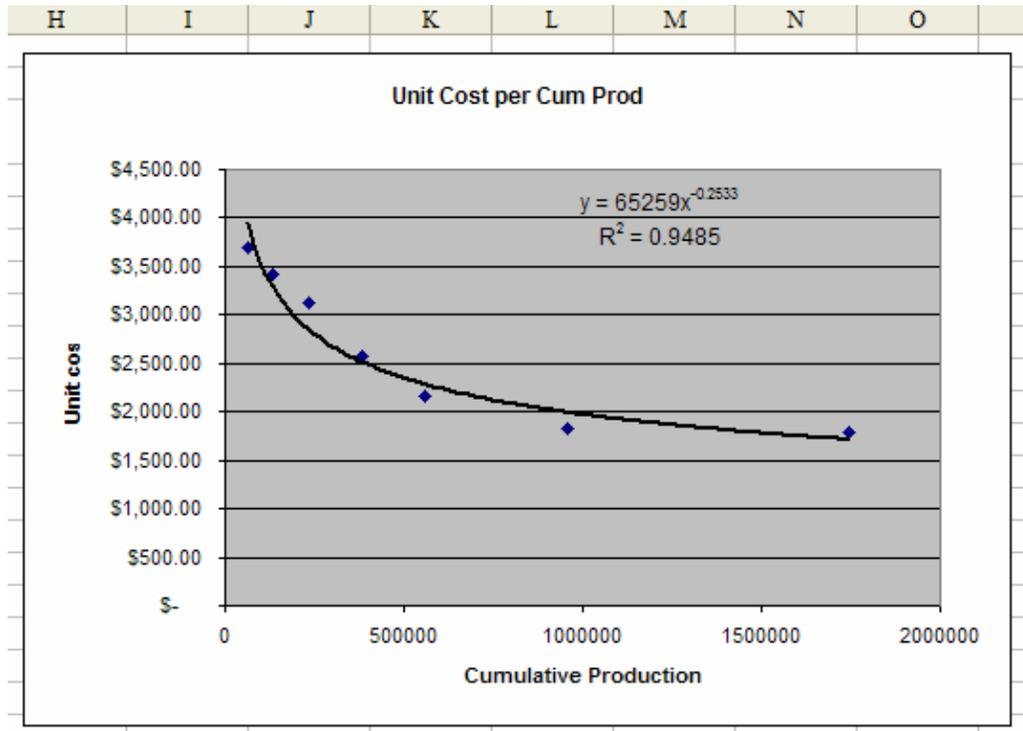


Figure 7.46 The R-Squared value with the Power trendline.

Excel’s **RSQ** function can calculate an R-squared value from a set of data using the *Linear* trendline. The format of the RSQ function is:

$$=RSQ(y_range, x_range)$$

Note that this function only works with *Linear* trend curves. We must also make sure that we have entered the *y_range*, or the dependent variable data, before the *x_range*, or the independent variable data. In Figure 7.47, we have employed the RSQ function with the first example from above to measure the accuracy of a *Linear* trendline as applied to the “Monthly Plant Cost per Units Produced” data. We can verify that the result of this function is the same as the one attained with the R-Squared value.

C21		fx =RSQ(C4:C15,B4:B15)	
	A	B	C
1	Production Cost		
2			
3	Month	Units Produced	Monthly Plant Cost
4	1	1260	\$99,850
5	2	1007	\$58,096
6	3	1096	\$96,360
7	4	873	\$65,675
8	5	532	\$51,870
9	6	476	\$27,462
10	7	482	\$27,808
11	8	1173	\$110,118
12	9	692	\$67,470
13	10	690	\$39,808
14	11	564	\$32,538
15	12	470	\$45,825
19			
20			
21		RSQ	0.8137

Figure 7.47 Using the RSQ function to calculate the R-Squared value of the *Linear* trendline.

The standard error measures the accuracy of any predictions made. In other words, it measures the “spread” around the least squares line, or the trendline. We have learned previously that this value can be found using *Descriptive Statistics*. It can also be calculated in Excel with the **STEYX** function. The format of this function is:

$$=STEYX(y_range, x_range)$$

Again, this function can only be used for *Linear* trend curves. In the example above, we have calculated the standard error using the STEYX function (see Figure 7.48). We can now use this value to check for outliers as we did using the standard deviation value in the previous sections. These outliers reveal how accurate our fit is with a *Linear* trendline.

C22		fx =STEYX(C4:C15,B4:B15)	
	A	B	C
1	Production Cost		
2			
3	Month	Units Produced	Monthly Plant Cost
4	1	1260	\$99,850
5	2	1007	\$58,096
6	3	1096	\$96,360
7	4	873	\$65,675
8	5	532	\$51,870
9	6	476	\$27,462
10	7	482	\$27,808
11	8	1173	\$110,118
12	9	692	\$67,470
13	10	690	\$39,808
14	11	564	\$32,538
15	12	470	\$45,825
19			
20			
21		RSQ	0.8137
22		STEYX	\$12,974

Figure 7.48 Using the STEYX function to calculate the standard error.

Two other Excel functions that can be applied to a linear regression line of a collection of data are **SLOPE** and **INTERCEPT**. The SLOPE function's format is:

$$=SLOPE(y_range, x_range)$$

Similarly, the intercept of the linear regression line of the data can be determined with the INTERCEPT function. The format of this function is:

$$=INTERCEPT(y_range, x_range)$$

In Figure 7.49, we are finding the slope and intercept of the linear regression line of the "Monthly Plant Cost per Units Produced" data.

C25		fx =INTERCEPT(C4:C15,B4:B15)	
	A	B	C
1	Production Cost		
2			
3	Month	Units Produced	Montly Plant Cost
4	1	1260	\$99,850
5	2	1007	\$58,096
6	3	1096	\$96,360
7	4	873	\$65,675
8	5	532	\$51,870
9	6	476	\$27,462
10	7	482	\$27,808
11	8	1173	\$110,118
12	9	692	\$67,470
13	10	690	\$39,808
14	11	564	\$32,538
15	12	470	\$45,825
19			
20			
21		RSQ	0.8137
22		STEYX	\$12,974
23			
24		SLOPE	\$88
25		INTERCEPT	-\$8,198

Figure 7.49 Finding the slope and intercept with the SLOPE and INTERCEPT functions.

 <p>Summary</p>	<p>Regression</p>	<p>Measures the amount of influence that the independent variable has on the dependent variable.</p>
	<p>R-Squared Value</p>	<p>Measures the accuracy of any predictions made.</p>
	<p>Standard Error</p>	<p>Measures the accuracy of any predictions made.</p>
	<p>More Statistical Functions</p>	
	<p>RSQ</p>	<p>Finds the R-squared value of a set of data.</p>
	<p>STEYX</p>	<p>Finds the standard error of regression for a set of data.</p>
	<p>SLOPE</p>	<p>Finds the slope of a set of data.</p>
<p>INTERCEPT</p>	<p>Finds the intercept of a set of data.</p>	

7.4 Distributions

We will now discuss some of the more common distributions that can be recognized when performing a statistical analysis of data. These are the **Normal**, **Exponential**, **Uniform**, **Binomial**, **Poisson**, **Beta**, and **Weibull** distributions. The Normal, Exponential and Uniform distributions are those most often used in practice. The Binomial and Poisson are also common distributions.

Most of these distributions have Excel functions associated with them. These functions are basically equivalent to using distribution tables. In other words, given certain parameters of a set of data for a particular distribution, we can look at a distribution table to find the corresponding area from the distribution curve. These Excel functions perform this task for us.

Let's begin with the Normal distribution. The parameters for this distribution are simply the value that we are interested in finding the probability for, and the mean and standard deviation of the set of data. The function that we apply with the Normal distribution is **NORMDIST**, and with these parameters, the format for this function is:

$$=NORMDIST(x, \text{mean}, \text{std_dev}, \text{cumulative})$$

We will use the *cumulative* parameter in many Excel distribution functions. This parameter takes the values *True* and *False* to determine if we want the value returned from the **cumulative distribution function** or the **probability density function**, respectively. A general difference between these two functions is that the *cumulative distribution function (cdf)* determines the probability that a value in the data set is less than or equal to *x*, while the *probability density function (pdf)* determines the probability that a value is exactly equal to *x*. We will employ this general definition to understand the *cumulative* parameter of other distribution functions as well.

For example, suppose annual drug sales at a local drugstore are distributed Normally with a mean of 40,000 and standard deviation of 10,000. What is the probability that the actual sales for the year are 42,000? To answer this, we use the NORMDIST function:

$$=NORMDIST(42000, 40000, 10000, \text{True})$$

This function returns a 0.58 probability, or 58% chance, that given this mean and standard deviation for the Normal distribution, annual drug sales will be 42,000 (see Figure 7.50).

C8		fx =NORMDIST(A8,\$B\$4,\$B\$5,TRUE)		
	A	B	C	
1	Normal			
2				
3	Drug demand			
4	Mean	40000		
5	Std Dev	10000		
6				
7	Prob that demand is:			
8	42000		0.58	
9				

Figure 7.50 Using the NORMDIST with the cumulative distribution function.

The cumulative distribution can also determine the probability that a value will lie in a given interval. Using the same example data, what is the probability that annual sales will be between 35,000 and 49,000? To find this value, we subtract the cdf values for these two bounds:

$$=NORMDIST(49000, 40000, 10000, True) - NORMDIST(35000, 40000, 10000, True)$$

This function returns a 0.51 probability, or 51% chance, that annual sales will be between 35,000 and 49,000 (see Figure 7.51).

C14		fx =NORMDIST(A12,B4,B5,TRUE) - NORMDIST(A11,B4,B5,TRUE)				
	A	B	C	D	E	F
1	Normal					
2						
3	Drug demand					
4	Mean	40000				
5	Std Dev	10000				
6						
7	Prob that demand is:					
8	42000		0.58			
9						
10	Prob that demand is between 35000 and 49000					
11	35000		0.31			
12	49000		0.82			
13	between		0.51			
14			0.51			

Figure 7.51 Using the NORMDIST function with an interval of x values.

Related to the Normal distribution is the Standard Normal distribution. If the mean of our data is 0 and the standard deviation is 1, then placing these values in the NORMDIST function with the *cumulative* parameter as *True* determines the resulting value from the Standard Normal distribution. There are also two other functions that determine the Standard Normal distribution value: **STANDARDIZE** and **NORMSDIST**.

STANDARDIZE converts the x value from a data set of a mean not equal to 0 and a standard deviation not equal to 1 into a value that does assume a mean of 0 and a standard deviation of 1. The format of this function is:

$$=STANDARDIZE(x, mean, std_dev)$$

The resulting standardized value is then used as the main parameter in the NORMSDIST function:

$$=NORMSDIST(standardized_x)$$

This function then finds the corresponding value from the Standard Normal distribution. These functions are valuable as they relieve much manual work in converting a Normal x value into a Standard Normal x value.

Let's now consider the same example as above to determine the probability that a drugstore's annual sales are 42,000. We standardize this using the following function:

$$=STANDARDIZE(42000, 40000, 10000)$$

The result of this function is 0.2. We can then use this value in the NORMSDIST function to compute the probability:

$$=NORMSDIST(0.2)$$

This function again returns a probability of 0.58 that the sales will reach 42,000 (see Figure 7.52).

E8		fx =STANDARDIZE(A8,B4,B5)					
	A	B	C	D	E	F	G
1	Normal						
2							
3	Drug demand						
4	Mean	40000					
5	Std Dev	10000					
6							
7	Prob that demand is:				Standardized value		
8	42000		0.58		0.2		0.58

Figure 7.52 Using the STANDARDIZE and NORMSDIST functions.

The Uniform distribution does not actually have a corresponding Excel function; however, there is a simple formula that models the Uniform distribution. This formula, or pdf, is:

$$1 / (b - a)$$

Given that a value x is Uniformly distributed between a and b , we can use this formula to determine the probability that x will have an integer value in this interval. To apply this formula in Excel, we recommend creating three columns: one for possible a values, one for possible b values, and one for the result of the Uniform formula (see Figure 7.50). Then, we just enter the Uniform formula by referencing the a and b value cells.

C19		fx = 1 / (B19 - A19)	
	A	B	C
15			
16	Uniform		
17			
18	a value	b value	Uniform cdf value
19	2	5	0.333

Figure 7.53 Using the Uniform distribution formula for various values of a and b .

The Poisson distribution has only the mean as its parameter. The function we use for this distribution is **POISSON** and the format is:

$$=POISSON(x, mean, cumulative)$$

(Note that for the Poisson distribution, the mean may be in terms of $lambda*time$.) The Poisson distribution value is the probability that the number of events that occur is either between 0 and x (cdf) or equal to x (pdf).

For example, consider a bakery that serves an average of 20 customers per hour. Find the probability that, at the most, 35 customers will be served in the next two hours. To do so, we use the POISSON function with a mean value of $\lambda \cdot \text{time} = 20 \cdot 2$.

$$=\text{POISSON}(35, 20 \cdot 2, \text{True})$$

This function returns a 0.24 probability value that 35 customers will be served in the next two hours (see Figure 7.54).

A29		fx =POISSON(B27,A24*B26,TRUE)		
	A	B	C	
20				
21	Poisson			
22				
23	Customer Service Rate			
24	20	cust/hr		
25				
26	In Time:	2	hrs	
27	Numb of Cust	35		
28				
29	0.24			
30				

Figure 7.54 Using the POISSON function with the service time.

The Exponential distribution has only one parameter: λ . The function we use for this distribution is **EXPONDIST** and its format is:

$$=\text{EXPONDIST}(x, \lambda, \text{cumulative})$$

(Note that the λ value is equivalent to $1/\text{mean}$.) The *cumulative* parameter is the same as described above. The x value is what we are interested in finding the distribution value for, and λ is the distribution parameter.

A common application of the Exponential distribution is for modeling interarrival times. Let's use the bakery example from above. If we are told that, on average, 20 customers are served per hour and we assume that each customer is served as soon as he or she arrives, then the arrival rate is said to be 20 customers per hour. This arrival rate can be converted into the interarrival mean by inverting this value; the interarrival mean, or the Exponential mean, is therefore $1/20$ hours per customer arrival. Therefore, if we want to determine the probability that a customer arrives in 10 minutes, we set $x = 10/60 = 0.17$ hour and $\lambda = 1/(1/20) = 20$ hours in the EXPONDIST function:

$$=\text{EXPONDIST}(0.17, 20, \text{True})$$

This function returns a probability value of 0.96 that a customer will arrive within 10 minutes (see Figure 7.55).

C36		fx =EXPONDIST(A36,1/B33,TRUE)	
	A	B	C
30			
31	Exponential		
32			
33	Interarrival rate	0.05	hr/cust
34			
35	Interarrival Time		
36	0.17	hours	0.96

Figure 7.55 Using the EXPONDIST function with the interarrival time.

The Binomial distribution has the following parameters: the number of trials and the probability of success. We are trying to determine the probability that the number of successes is less than (using *cdf*) or equal to (*pdf*) some *x* value. The function for this distribution is **BINOMDIST** and its format is:

$$=BINOMDIST(x, trials, prob_success, cumulative)$$

(Note that the values of *x* and *trials* should be integers.) For example, suppose a marketing group is conducting a survey to find out if people are more influenced by newspaper or television ads. Assuming, from historical data, that 40 percent of people pay more attention to ads in the newspaper, and 60 percent pay more attention to ads on television, what is the probability that out of 100 people surveyed, 50 of them respond more to ads on television? To determine this, we use the BINOMDIST function with the *prob_success* value equal to 0.60.

$$=BINOMDIST(50, 100, 0.60, True)$$

This function returns a value of 0.03 that 50 out of 100 people will report that they respond more to television ads than newspaper ads (see Figure 7.56).

A46		fx =BINOMDIST(B44,B40,B42,TRUE)	
	A	B	C
37			
38	Binomial		
39			
40	Number of Trials	100	
41	Prob Newspaper	0.40	
42	Prob Television	0.60	
43			
44	Prefer Television	50	
45			
46	0.03		
47			

Figure 7.56 Using the BINOMDIST function with the survey data.

The Beta distribution has the following parameters: *alpha*, *beta*, *A*, and *B*. *Alpha* and *beta* are determined from the data set; *A* and *B* are optional bounds on the *x* value for

which we want the Beta distribution value. The function for this distribution is **BETADIST** and its format is:

$$=BETADIST(x, \alpha, \beta, A, B)$$

If *A* and *B* are omitted, then a standard cumulative distribution is assumed and they are assigned the values 0 and 1, respectively.

For example, suppose a management team is trying to complete a big project by an upcoming deadline. They want to determine the probability that they can complete the project in 10 days. They estimate the total time needed to be one to two weeks based on previous projects that they have worked on together; these estimates will be the bound values, or the *A* and *B* parameters. They can also determine a mean and standard deviation from this past data to be 12 and 3 days, respectively. We can use this mean and standard deviation to compute the alpha and beta parameters; we do so using some complex transformation equations (shown in Figure 7.57), resulting in alpha = 0.08 and beta = 0.03. (Note that usually alpha and beta can be found in a resource table for the Beta distribution.) We can then use the BETADIST function as follows:

$$=BETADIST(10, 0.08, 0.03, 7, 14)$$

The result reveals that there is a 0.28 probability that they can finish the project in 10 days (see Figure 7.57).

	A	B	C	D	E	F
47						
48	Beta					
49						
50	Mean	12				
51	Std Dev	3				
52						
53	A (lower bound)	7				
54	B (upper bound)	14				
55						
56	transformed mean	0.71	(mean - lower) / (upper - lower)			
57	transformed std dev	0.18	stdev^2 / (upper - lower)^2			
58						
59	alpha	0.08	trmean*((trmean*(1-trmean)/trstdev)-1)			
60	beta	0.03	(1-trmean)*((trmean*(1-trmean)/trstdev)-1)			
61						
62	Project time	10	(days)			
63						
64		0.28				
65						

Figure 7.57 Using BETADIST and calculating the *alpha* and *beta* values.

The Weibull distribution has the parameters *alpha* and *beta*. The function we use for this distribution is **WEIBULL** and its format is:

$$=WEIBULL(x, \alpha, \beta, \text{cumulative})$$

(Note that if *alpha* is equal to 1, then this distribution becomes equivalent to the Exponential distribution with *lambda* equal to $1/\beta$.) The Weibull distribution is most commonly employed to determine reliability functions. Consider the inspection of 50 light bulbs. Past data reveals that on average, a light bulb lasts 1200 hours, with a standard deviation of 100 hours. We can use these values to calculate alpha and beta to be 14.71 and 1243.44, respectively. (Note that usually alpha and beta can be located in a resource table for the Beta distribution.) We can now use the WEIBULL distribution to determine the probability that a light bulb will be reliable for 55 days = 1320 hours.

$$=WEIBULL(1320, 14.71, 1243.44, True)$$

The result is a 0.91 probability that a light bulb will last up to 1320 hours, or 55 days (see Figure 7.58).

B76		fx =WEIBULL(B74,B71,B72,TRUE)	
	A	B	C
65			
66	Weibull		
67			
68	Mean	1200	
69	Std Dev	100	
70			
71	alpha	14.71	
72	beta	1243.44	
73			
74	Reliability	1320	(hours)
75			
76		0.91	
77			

Figure 7.58 Using the WEIBULL function to determine the reliability of a light bulb.

 <p>Summary</p>	Distribution Function	Parameters
	NORMDIST	<i>x, mean, std_dev, cumulative</i>
	EXPONDIS	<i>x, lamda, cumulative</i>
	Uniform	<i>a, b</i>
	BINOMDIST	<i>x, trials, prob_success, cumulative</i>
	POISSON	<i>x, mean, cumulative</i>
	BETADIST	<i>x, alpha, beta, A, B</i>
WEIBULL	<i>x, alpha, beta, cumulative</i>	
Other Distribution Functions	FDIST, GAMMADIST, HYPGEOMDIST, LOGNORMDIST, NEGBINOMDIST	

7.5 Summary

- Some of Excel's basic statistical functions are: AVERAGE to find the mean, MEDIAN to find the median, and STDEV to find the standard deviation of a set of data.
- The Analysis Toolpack is an Excel Add-In that includes statistical analysis techniques such as *Descriptive Statistics*, *Histograms*, *Exponential Smoothing*, *Correlation*, *Covariance*, *Moving Average*, and others.
- The *Descriptive Statistics* option provides a list of statistical information about a data set, including the mean, median, standard deviation, and variance.
- The *Mean* is the average of all values in a data set, or all observations in a sample. The *Median* is the "middle" observation when data is sorted in ascending order. The *Mode* is the most frequently occurring value.
- The *Sample Variance* is the average squared distance from the mean to each data point. The *Standard Deviation*, s , is the square root of the *Sample Variance*. Any values in the data set that lie more than $\pm 2s$ from the mean are called *outliers*. Excel functions such as IF, ABS, and OR can identify outliers. Conditional Formatting can also be used.
- *Kurtosis* is a measure of data peakedness. *Skewness* is a measure of how symmetric or asymmetric data is.
- The *Confidence Level for Mean* constrains the mean calculation to a specified confidence level. The *Kth Largest* and *Kth Smallest* options provide the respectively ranked data value for a specified value of k .
- Similar to the *Kth Largest* and *Kth Smallest* options with *Descriptive Statistics* are the two Excel functions PERCENTILE and PERCENTRANK.
- Histograms calculate the number of occurrences, or frequency, which values in a data set fall into various intervals. Bins are the intervals into which values can fall; they can be defined by a user or can be evenly distributed among the data by Excel. The bin values are specified by their upper bounds.
- There are four basic shapes to a histogram: *symmetric*, *positively skewed*, *negatively skewed*, and *multiple peaks*.
- Relationships in data are usually identified by comparing the *dependent variable* and the *independent variable*. The dependent variable is a variable that the user tries to predict values for; the independent variable is the variable that the user employs as the comparison in order to make the prediction.
- You can graph this data (with the *XY Scatter* chart type) by placing the independent variable on the x-axis and the dependent variable on the y-axis and then using a *trend curve* to determine if any relationship exists between these variables. There are five basic trend curves that Excel can model: *Linear*, *Exponential*, *Power*, *Moving Average*, and *Logarithmic*.
- With Linear curves, there are two values that measure the accuracy of the relationship between the dependent and independent variables. The R-Squared value measures the amount of influence that the independent variable has on the dependent variable. It can be calculated from the trendline chart or with the RSQ function. The standard error also measures the accuracy of any predictions made from this relationship. This value can be determined using the STEYX function.
- The SLOPE and INTERCEPT functions also analyze a Linear trend curve.
- Some of the more common distributions that can be recognized when performing a statistical analysis of data are the *Normal*, *Exponential*, *Uniform*, *Binomial*, *Poisson*, *Beta*, and *Weibull* distributions. Most of these distributions have Excel functions associated directly with them and are basically equivalent to using distribution tables.

7.6 Exercises

7.6.1 Review Questions

1. What function calculates the mean of a data set?
2. What is the difference between the mean, median, and mode of a set of data?
3. List some of the Analysis Toolpack's useful tools.
4. What statistical analysis values does the Descriptive Statistics tool provide?
5. From what value is the standard deviation derived?
6. How is an outlier identified?
7. Write an alternate formula for identifying an outlier using the IF and OR functions.
8. What is Skewness? What is an appropriate value of Skewness for a symmetric data set?
9. What is the difference between the result of the PERCENTILE and PERCENTRANK functions?
10. What are the bins of a histogram? How are they created?
11. What is a Pareto organization of a chart?
12. What are the four basic shapes of a histogram?
13. What is an example of a negatively skewed histogram?
14. Give an example of a dependent and independent variable relationship.
15. Can a trendline be fitted to any type of chart created in Excel?
16. What are the three most common trend curves?
17. What two values measure the accuracy of a Linear trendline?
18. What are the parameters of the Binomial distribution function?
19. What relationship is there between the Weibull and Exponential distribution functions?
20. How do you convert a Normal x value into a Standard Normal x value?

7.6.2 Hands-On Exercises

1. The table below provides a sample of the starting salaries of all geography graduates from a state university this year. What is your best estimate of a "typical" starting salary for a geography graduate?

Student	Salary
1	\$22,259.00
2	\$20,931.00
3	\$26,271.00
4	\$26,034.00
5	\$26,632.00
6	\$25,038.00
7	\$26,338.00
8	\$25,489.00
9	\$26,670.00
10	\$29,952.00
11	\$21,237.00
12	\$27,469.00
13	\$27,568.00
14	\$22,341.00

15	\$26,967.00
16	\$25,381.00
17	\$22,075.00
18	\$28,211.00
19	\$20,748.00
20	\$27,978.00

2. A quality assurance expert at a soft drink bottling plant has been assigned to develop a plan to reduce the number of defective bottles that the plant produces. To find the cause of the defects, she plans to analyze factors associated with the bottling lines and the types of bottles being produced. The expert has randomly sampled sets of bottles from different bottling lines and counted the number of defective bottles in the sample. She records the bottling line, the size of the sample, and the number of nonconforming bottles. She then computes the fraction of nonconforming bottles. The table below contains her results. Create this table in an Excel spreadsheet and make the following modifications:
 - a. Fill in the values for the “Fraction Nonconforming” column by dividing the number of nonconforming bottles in the sample by the sample size. Display the results as a percentage.
 - b. Compute the mean and standard deviation of the fraction of nonconforming bottles found in the samples and record the results in the bottom right-hand corner of the spreadsheet.

Sample No.	Line Number	Sample Size	No. Nonconforming	Fraction Nonconforming
1	1	100	9	
2	1	100	5	
3	1	90	7	
4	2	75	4	
5	2	100	11	
6	2	90	8	
7	2	80	8	
8	3	80	6	
9	3	100	10	
10	3	90	7	
			Mean=	
			Standard Deviation=	

3. In New York, Electro produces voltage that regulates equipment and then ships the equipment to Chicago. The voltage held is measured in NY before each unit is shipped to Chicago. The voltage held by each unit is also measured when the unit arrives in Chicago. A sample of voltage measurements at each city is provided below.

Equip Num	NY	Chicago
1	84	87
2	38	30
3	46	17
4	57	57
5	40	29
6	27	63
7	21	62
8	30	66
9	71	49
10	25	60

A voltage regulator is considered acceptable if it can hold a voltage of between 25 and 75 volts.

- Using Descriptive Statistics, comment on what you can observe about the voltages held by units before shipment and after shipment.
 - What percentage of units is acceptable before and after shipping?
 - Do you have any suggestions about how to improve the quality of Electro's regulators?
 - 10% of all NY regulators have a voltage exceeding what value?
 - 5% of all NY regulators have a voltage less than or equal to what value?
4. Use the data below to create a histogram for annual returns on stocks, bills, and bonds. Which investment has the highest average return?

Year	Stocks	T. Bills	T. Bonds
1928	43.81%	3.08%	0.84%
1929	-8.30%	3.16%	4.20%
1930	-25.12%	4.55%	4.54%
1931	-43.84%	2.31%	-2.56%
1932	-8.64%	1.07%	8.79%
1933	49.98%	0.96%	1.86%
1934	-1.19%	0.30%	7.96%
1997	31.86%	4.91%	9.94%
1998	28.34%	5.16%	14.92%
1999	20.89%	4.39%	-8.25%
2000	-9.03%	5.37%	16.66%
2001	-11.85%	5.73%	5.57%

- Using the above data, describe the type of histogram for each investment option: symmetric, positively skewed, negatively skewed, and multiple peaks.
- The table below records monthly returns on the S and P stock index and Dell stock.

Date	S and P return	Dell return
1-Jan-99	0.042	0.366351677

1-Feb-99	-0.031	-0.19875
1-Mar-99	0.04	0.020280811
1-Apr-99	0.039	0.00764526
1-May-99	-0.024	-0.163884674
1-Jun-99	0.056	0.074410163
1-Jul-99	-0.031	0.10472973
1-Aug-99	-0.005	0.194189602
1-Sep-99	-0.027	-0.14340589
1-Oct-99	0.063	-0.040358744
1-Nov-99	0.02	0.07165109
1-Dec-99	0.059	0.186046512
1-Jan-00	-0.05	-0.246323529
1-Feb-00	-0.019	0.061788618
1-Mar-00	0.098	0.321592649
1-Apr-00	-0.03	-0.070683662
1-May-00	-0.021	-0.139650873
1-Jun-00	0.025	0.143478261
1-Jul-00	-0.016	-0.108998733
1-Aug-00	0.062	-0.007112376
1-Sep-00	-0.053	-0.293696275
1-Oct-00	-0.004	-0.042596349

Find the following information:

- The slope of the least squares line of the Dell stock and S and P.
- The R-Squared value of the Dell stock and S and P.
- Which seems like a better investment, and why.

7. The following table lists the square footage and sales price for several houses.

Square Footage	Price
4016	\$ 453,696.00
2762	\$ 286,676.00
2262	\$ 239,414.00
4626	\$ 499,226.00
2550	\$ 249,500.00
2957	\$ 297,958.00
2868	\$ 318,876.00
3356	\$ 375,736.00
2536	\$ 278,672.00
3451	\$ 348,943.00
2807	\$ 305,121.00
2891	\$ 281,972.00
2984	\$ 294,496.00
3506	\$ 363,588.00
2694	\$ 293,482.00
2455	\$ 249,225.00
2732	\$ 313,788.00

2436	\$ 272,652.00
3894	\$ 384,248.00
4093	\$ 435,207.00
1548	\$ 182,280.00

- a. If you build a 400 square foot addition to your house, by how much do you feel you will increase its value?
 - b. What percentage of the variation in home values is explained by variation in house size?
 - c. A 2500 square foot house is selling for \$470,000. Is this price out of line with typical home values? Explain.
8. Given additional information on the number of bedrooms and bathrooms for the above house data, which factor (“Square Footage,” “Bedrooms,” or “Bathrooms”) has the strongest relationship with the sales price?

Square Footage	Bathrooms	Bedrooms	Price
4016	4	5	\$453,696.00
2762	2	3	\$286,676.00
2262	2	4	\$239,414.00
4626	4	6	\$499,226.00
2550	2	4	\$249,500.00
2957	2	4	\$297,958.00
2868	2	2	\$318,876.00
3356	4	3	\$375,736.00
2536	2	4	\$278,672.00
3451	4	5	\$348,943.00
2807	2	3	\$305,121.00
2891	2	3	\$281,972.00
2984	3	2	\$294,496.00
3506	4	3	\$363,588.00
2694	2	3	\$293,482.00
2455	2	3	\$249,225.00
2732	2	3	\$313,788.00
2436	2	2	\$272,652.00
3894	3	5	\$384,248.00
4093	5	5	\$435,207.00
1548	2	2	\$182,280.00

9. Given the yearly revenues (in millions) of the companies in the table below, determine the following:
- a. Which company’s revenues best fit an Exponential trend curve.
 - b. The annual percentage growth rate for revenues.
 - c. Predicted 2003 revenues.

(all revenues in millions)

Year	Staples	Wal-Mart	INTEL
2001	\$10,874	\$191,329	\$33,726
2000	\$8,842	\$185,013	\$29,389
1999	\$7,123	\$157,634	\$28,273
1998	\$5,181	\$117,958	\$25,070
1997	\$3,968	\$104,859	\$20,847

10. A marketing manager estimates total sales as a function of price, as seen below.
- Estimate the relationship between price and demand.
 - Predict the demand for the \$69 price.
 - By how much will a 1 percent increase in price reduce the demand?

Price	Demand
\$45.00	350
\$55.00	310
\$65.00	276
\$75.00	203
\$85.00	140
\$95.00	98

11. The manager of the sales department of a leading magazine publication has recorded the number of subscriptions sold for various numbers of sales calls.
- If he were to make 75,000 sales calls next month, how many subscriptions could he estimate selling?
 - If he wanted to sell 80,000 subscriptions, how many sales calls would he have to make?

Sales Calls	Number of Subscriptions Sold
30,000	35,297
40,000	50,687
50,000	63,053
100,000	75,033
150,000	84,377
200,000	89,579

12. A human resources manager wants to examine the relationship between annual salaries and the number of years that employees have worked at the company. A sample of collected data is below:

Years Worked	Annual Salary
1	\$45,000
2	\$48,000
3	\$50,000
4	\$53,000
5	\$55,000
10	\$60,000
15	\$65,000

20	\$70,000
30	\$75,000

- a. Which should be the independent variable and which should be the dependent variable?
 - b. Estimate the relationship between these two variables and interpret the least squares line.
 - c. How well does this line fit the data?
13. Consider the relationship between the size of the population and the average household income level for several small towns. A sample of this data is below:

Town Population	Average Household Income
100	\$25,000
250	\$28,000
760	\$30,000
1100	\$33,000
1600	\$35,000
2000	\$40,000
5000	\$45,000
10000	\$50,000
14,000	\$55,000

- a. Which should be the independent variable and which should be the dependent variable?
 - b. Estimate the relationship between these two variables and interpret the least squares line.
 - c. How well does this line fit the data?
14. A bank is trying to prove that they do not practice gender discrimination. They have a record of the education level, age, gender, and salary of each employee (provided below).

Employee ID	Education Level	Age	Gender	Salary
1	MBA	55	M	\$74,712
2	MBA	47	F	\$70,730
3	MS	44	F	\$60,289
4	MBA	46	M	\$69,537
5	BS	39	M	\$48,558
6	BS	38	M	\$51,621
7	MS	42	F	\$57,667
8	BA	28	F	\$54,923
9	BA	41	M	\$55,749

Determine which factor has the strongest relationship with the salary of the employees.

15. An electric company produces different quantities of electricity each month, depending on demand. The table below lists the number of units of electricity produced and the total cost of producing each quantity.

Month	Units	Cost
1	763	\$48,498
2	538	\$37,944
3	384	\$30,759
4	525	\$36,649
5	489	\$33,322
6	722	\$47,521
7	657	\$47,045
8	770	\$48,906
9	606	\$46,141
10	489	\$36,496
11	795	\$49,947
12	780	\$49,599

- Which trend curve fits the data better, a Linear, Exponential, or Power curve?
 - What are the R-Squared values of each curve?
 - How much cost can they expect if they produce 800 units?
16. A new industrial production company wants to analyze their production time to determine if they have improved productivity after gaining a few months of experience. The table below records the times to produce each batch of products.

Batch	Time
1	125
2	110
3	105
4	103
5	99
6	98
7	94
8	93
9	92
10	87
11	89
12	85
13	84
14	78
15	81
16	79
17	76

18	75
19	76
20	74

- a. Which curve best fits this data?
 - b. If this data follows a learning curve, then how much time can the company expect to spend producing the next batch?
17. Suppose that car sales follow a Normal distribution with a mean of 50,000 cars and a standard deviation of 14,000 cars.
- a. There is a 1 percent chance that the car sales will be how many cars next year?
 - b. What is the probability that they will sell less than or equal to 2.7 million cars during the next year?
18. Given that the weight of a typical American male follows a Normal distribution with a mean of 180 lb and standard deviation of 30 lbs, what fraction of American males weigh more than 225 lbs?
19. If a financial report shows an average income of \$45,000 with a standard deviation of \$12,000, what percentage of people on this report make more than \$60,000, assuming this data follows a Normal distribution? Convert this into a Standard Normal distribution and answer the same question.
20. Assume that the monthly sales of a toys store follow an Exponential distribution with mean 560. What is the probability that sales will be over 600 in January?
21. The annual number of accidents occurring in a particular manufacturing plant follows a Poisson distribution with mean 15.
- a. What is the probability of observing exactly 15 accidents at this plant?
 - b. What is the probability of observing less than 15 accidents?
 - c. You can be 99 percent sure that less than how many accidents will occur?
22. Using the Binomial distribution, assume that on average 95 percent of airline passengers show up for a flight. If a plane can seat 200 passengers, how many tickets should be sold to make the change of an overbooked flight less than or equal to 5 percent?
23. A professor gives his students a 20-question True or False exam. Each correct answer is worth 5 points. Consider a student who randomly guesses on each question.
- a. If no points are deducted for incorrect answers, what is the probability that the student will score at least 60 points?
 - b. If 5 points are deducted for each incorrect answer what is the probability that the student will score at least 60 points?
24. Suppose that the interarrival time between customers at a bank are Exponentially distributed with a mean of 45 seconds. If you just observed an arrival, what is the

probably that you will need to wait more than a minute before observing the next arrival? What is the probability that you will need to wait at least 2 minutes?

25. The following table presents the weekly sales of floppy disk drives in a local computer dealer.

Week	1	2	3	4	5	6	7	8	9	10	11	12
Sales	86	75	72	83	132	65	110	90	67	92	98	73

- a. Find the trendline that fits the data best (linear, exponential, etc).
 - b. Present the R-square for each trendline considered in part a.
 - c. What are the expected sales for weeks 13 and 14?
26. The length of an injection-molded plastic case that holds magnetic tape is normally distributed with mean 80.3 millimeters and standard deviation 0.2 millimeters.
- a. What is the probability that a part is longer than 80.5 millimeters or shorter than 80 millimeters?
 - b. Assuming that the cases will continue to be produced using the current process, up to what length will a part be 99% of the time?
27. The weight of a Coca-Cola bottle is normally distributed with a mean of 12 ounces and a standard deviation of 0.5 ounces.
- a. What is the probability that the bottle weights more than 13 ounces?
 - b. What is the probability that the bottle weights no more than 13 ounces and no less than 11 ounces
 - c. What must the standard deviation of weight be in order for the company to state 99.9% of its bottles weight less than 13 ounces?
 - d. If the standard deviation remains 0.5 ounce, what must the mean be in order for the company to state that 99.9% of the bottles produced are less than 13 ounces?
28. The length of time (in seconds) that a user views a page on a Web site before moving to another page is lognormal random variable with parameters $\theta = 0.5$ and $\omega^2 = 1$.
- a. What is the probability that a page is viewed for more than 10 seconds?
 - b. What is the length of time that 50% of users view the page?
 - c. Plot the density function of this distribution. Change the value of θ to 1 and plot the density function again.
29. The lifetime of a semiconductor laser follows a Weibull distribution with parameters $\alpha = 2$ and $\beta = 700$ hours.
- a. Determine the probability that a semiconductor laser lasts at least 600 hours.

- b. Determine the probability that a semiconductor laser fails before 400 hours.
- c. Plot the density function of this distribution.