Chapter 42

Summarizing data by using descriptive statistics

Questions answered in this chapter:

- What defines a typical value for a data set?
- How can I measure how much a data set spreads from its typical value?
- Together, what do the mean and standard deviation of a data set tell me about the data?
- How can I use descriptive statistics to compare data sets?
- For a given data point, can I easily find its percentile ranking within the data set? For example, how can I find the ninetieth percentile of a data set?
- How can I easily find the second largest or second smallest number in a data set?
- How can I rank numbers in a data set?
- What is the trimmed mean of a data set?
- When I select a range of cells, is there an easy way to get a variety of statistics that describe the data in those cells?
- Why do financial analysts often use the geometric mean to summarize the average return on a stock?
- How can I use boxplots to summarize and compare datasets?

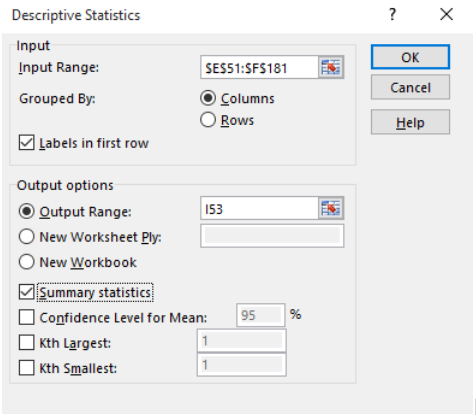
In Chapter 41, “Summarizing data by using histograms and Pareto charts,” I showed how you can describe data sets by using histograms. In this chapter, I show how to describe a data set by using particular characteristics of the data, such as the *mean*, *median*, *standard deviation*, and *variance*—measures that Microsoft Excel 2016 groups together as descriptive statistics. You can obtain the descriptive statistics for a set of data by clicking Data Analysis in the Analysis group on the Data tab (available as an add-in) and then selecting the Descriptive Statistics option. After you enter the relevant data and click OK, all the descriptive statistics of your data are displayed. You can also obtain descriptive statistics by using Excel functions. At the end of this chapter, I’ll show how boxplots can be used to summarize and compare data sets.

Answers to this chapter's questions

What defines a typical value for a data set?

To illustrate the use of descriptive statistics, let's return to the Cisco and GM monthly stock return data in the file Stock.xlsx. To create a set of descriptive statistics for this data, click Data Analysis in the Analysis group on the Data tab, select Descriptive Statistics, and then click OK. Fill in the Descriptive Statistics dialog box as shown in Figure 42-1.

> **Note** If the Data Analysis command doesn't appear on the Data tab, click the File tab, click Options, and then click Add-Ins in the left pane. In the Manage list in the Excel Options dialog box, click Excel Add-Ins, and then click Go. In the Add-Ins dialog box, select Analysis ToolPak (the first choice, not Analysis ToolPak - VBA), and then click OK.



F42xx01

FIGURE 42-1 The Descriptive Statistics dialog box.

F42xx01: This figure shows the settings needed to obtain descriptive statistics for GM and Cisco stock.

The input range I entered is the monthly Cisco and GM returns, located in the range E51:F181 (including the labels in row 51). I filled in the remainder of the Descriptive Statistics dialog box as shown in Figure 42-1 for the following reasons:

- I selected Columns in the Grouped By options because each data set is listed in a different column.
- I selected Labels In First Row because the first row of the data range contains labels and not data.
- I selected Output Range in the Output Options section, and I selected cell I53 of the current worksheet as the first cell in the output range.
- By selecting Summary Statistics, I ensured that I get the most commonly used descriptive-

statistics measures for both the GM and Cisco monthly returns.

When you click OK, Excel calculates the descriptive statistics, as shown in Figure 42-2.

	I	J	K	L
53	*GM*		*CSCO*	
54				
55	Mean	0.009	Mean	0.056
56	Standard Error	0.008	Standard Error	0.011
57	Median	-0.005	Median	0.050
58	Mode	#N/A	Mode	0.051
59	Standard Deviation	0.090	Standard Deviation	0.122
60	Sample Variance	0.008	Sample Variance	0.015
61	Kurtosis	0.475	Kurtosis	-0.320
62	Skewness	0.224	Skewness	0.105
63	Range	0.517	Range	0.541
64	Minimum	-0.240	Minimum	-0.203
65	Maximum	0.277	Maximum	0.339
66	Sum	1.206	Sum	7.224
67	Count	130.000	Count	130.000

F42xx02

FIGURE 42-2 The descriptive statistics results for Cisco and GM stocks.

F42xx02: This figure shows descriptive statistics for Cisco and GM.

Now let's interpret the descriptive statistics that define a typical value (or a central location) for Cisco's monthly stock returns. The descriptive statistics output contains three measures of central location: *mean* (or *average*), *median*, and *mode*.

- **Mean** The mean of a data set is written as x and is simply the average of all observations in the sample. If the data values were x_1, x_2,…, x_n, then the following equation calculates the mean:

$$\bar{x} = \frac{1}{n}\sum . x_i$$ G42xx01

  Here, n equals the number of observations in the sample, and x_i is the ith observation in the sample. We find that Cisco's mean monthly return is 5.6 percent per month.

  It is always true that the sum of the deviations of all values from the mean equals 0. Thus, you can think of a data set's mean as a *balancing point* for the data. Of course, without using the Descriptive Statistics option, you can obtain a sample's mean in Excel by applying the AVERAGE function to the appropriate cell range.

- **Median** The median of a sample is the *middle* observation when the data is listed from smallest to largest. If a sample contains an odd number of observations, the median is the observation that has as many observations below it as above it. Thus, for a sample of nine, the median would be the fifth smallest (or fifth largest) observation. When a sample includes an even number of observations, you can simply average the two middle observations. Essentially, the median is the fiftieth percentile of the data. For example, the median monthly return on Cisco's stock is 5 percent. You could also obtain this information by using the MEDIAN function.

- **Mode** The mode is the most frequently occurring value in the sample. If no value occurs more than once, the mode does not exist. For GM, no monthly return occurred more than once for the years 1990–2000, so the mode does not exist. For Cisco, the mode was approximately 5.14 percent. In versions of Excel prior to Excel 2010, you could also use the MODE function to compute the mode. If no data value occurred more than once, the MODE function returned #NA.

The problem is that a data set can have more than one mode, and the MODE function simply returned the first mode it found. For this reason, Excel 2010 introduced two functions: MODE.SNGL and MODE.MULT. (See the file Modefunctions.xlsx.)

MODE.SNGL performs exactly as the MODE function performed in earlier Excel versions. Earlier versions of Excel, however, do not recognize the MODE.SNGL function.

MODE.MULT is an *array function*. You'll learn more about array functions in Chapter 88, "Array formulas and functions." To use the MODE.MULT function (or any other array function), you must first select the range of cells to which the function will return values. Next, enter the function or formula. Finally, do not just press Enter; you must hold down the Ctrl key followed by the Shift key and then press Enter. (Or you can press Tab.) The problem with the MODE.MULT function is that you do not know in advance how many modes a data set has, so you do not know the correct size of the range that you need to select.

The file Modefunctions.xlsx (see Figure 42-3) shows the use of all three functions involving modes.

	C	D	E	F	G
3		3 and 5 are both modes			
4					
5		3			
6		4			
7		5			
8		3			
9		2			
10		1			
11		5			
12					
13			3	=MODE(D5:D11)	
14	3 cells selected		3	=MODE.SNGL(d5:d11)	
15	Mode.Mult				
16	3		3	Mode.Mult	
17	5		5	two cells selected	
18	#N/A				
19			Mode.mult one cell selected		

F42xx03

FIGURE 42-3 Examples of Excel's MODE functions.

F42xx03: This figure shows examples of Excel's various MODE functions.

The data set displayed in Figure 42-3 has two modes (3 and 5). In cell E13, I entered the old-school MODE function with the formula =MODE(D5:D11). Excel returned the first mode it found (3). In cell E14, the formula =MODE.SNGL(D5:D11) duplicates this result.

After selecting the cell range E16:E17, I array-entered the formula =MODE.MULT(D5:D11), and Excel returned both the modes (3 and 5). After selecting the cell range C16:C18 (a three-cell range), I array-entered the same formula, and C18 was filled with an #N/A because the data set had no third mode to fill the cell. Finally, I selected the single cell range E20 and array-entered the same formula. Because I selected a range containing a single cell, Excel returned only the first mode it found (3).

The mode is rarely used as a measure of central location. It is interesting to note, however, that for a symmetric data set, the mean, median, and mode are equal.

A natural question is whether the mean or median is a better measure of central location. Essentially, the mean is the best measure of central location if the data set does not exhibit an excessive skew. Otherwise, you should use the median as the measure of central location. If a data set is highly skewed, extreme values distort the mean. In this case, the median is a better measure of a typical data set value. For example, the US government reports median family income instead of mean family income because family income is highly positively skewed.

The *skewness measure* reported by the descriptive statistics output indicates whether a data set is highly skewed, in the following ways:

- A skew greater than +1 indicates a high degree of positive skew.
- A skew less than –1 indicates a high degree of negative skew.
- A skew between –1 and +1 inclusive indicates a relatively symmetric data set.

Thus, monthly returns of GM and Cisco exhibit a slight degree of positive skewness. Because the skewness measure for each data set is less than +1, the mean is a better measure of a typical return than the median. You can also use the SKEW function to compute the skew of a data set.

By the way, *kurtosis*, which sounds like a disease, is not a very important measure, although you can see that it is one of the descriptive statistics results listed in Figure 42-2. Kurtosis near 0 means a data set exhibits *peakedness* close to the normal (or standard bell-shaped) curve. (I'll discuss the normal curve in Chapter 70, "The normal random variable and Z-scores.") *Positive kurtosis* means that a data set is more peaked than a normal random variable, whereas *negative kurtosis* means that data is less peaked than a normal random variable. GM monthly returns are more peaked than a normal curve, whereas Cisco monthly returns are less peaked than a normal curve.

How can I measure how much a data set spreads from its typical value?

Let's consider two investments. Each yields an average of 20 percent per year. Before deciding which investment you prefer, you'd like to know about the spread, or riskiness, of the investment. The most important measures of the spread (or dispersion) of a data set from its mean are *sample variance*, *sample standard deviation*, and *range*.

We can discuss *sample variance* and *sample standard deviation* together. The sample variance s^2 is defined by the following formula:

$\frac{1}{n-1}\sum(x_i-\overline{x})^2$ G42xx02

You can think of the sample variance as the average squared deviation of the data from its mean. Intuitively, it seems like you should divide by *n* to compute a true average squared deviation, but for technical reasons you need to divide by *n–1*.

Dividing the sum of the squared deviations by *n–1* ensures that the sample variance is an unbiased measure of the true variance of the population from which the sampled data is drawn.

The sample standard deviation *s* is just the square root of s^2.

The following is an example of these computations for the three numbers 1, 3, and 5:

$s^2=\frac{1}{3-1}\ [(1-3)^2+(3-3)^2+(5-3)^2]=4.$ G42xx03

We find that these computations yield the following result:

$s=\sqrt{4}=2.$ G42xx04

In the stock example, the sample standard deviation of monthly returns for Cisco is 12.2 percent with a sample variance of $0.015\%^2$. Naturally, $\%^2$ is hard to interpret, so you usually look at the sample standard deviation. For GM, the sample standard deviation is 8.97 percent.

In Excel 2007 or earlier, the sample variance of a data set was computed with the VAR function, and the sample standard deviation was computed with the STDEV function. You can still use these functions in Excel 2016, but Excel 2010 added the equivalent functions VAR.S and STDEV.S. (The *S* stands for *sample*.) The relatively new functions VAR.P and STDEV.P compute the population variance and population standard deviation. To compute a population variance or standard deviation, simply replace *n-1* in the denominator of the definition of s^2 by n.

The *range* of a data set is the largest number in the data set minus the smallest number. Here, the range in the monthly Cisco returns is equal to 54 percent, and the range for GM monthly returns is 52 percent.

Together, what do the mean and standard deviation of a data set tell me about the data?

Assuming that a histogram follows a *Gaussian*, or *normal* population, the *rule of thumb* (set of related math rules) tells us the following:

- Approximately 68 percent of all observations are between *x–s* and *x+s*.
- Approximately 95 percent of all observations are between *x–2s* and *x+2s*.
- Approximately 99.7 percent of all observations are between *x–3s* and *x+3s*.

For example, you would expect that approximately 95 percent of all Cisco monthly returns are from –19 percent through 30 percent, as shown here:

Mean–2s = .056-2(.122) = –19% and Mean+2s = .056+2*(.122) = 30%*

Any observation more than two standard deviations away from the mean is called an *outlier*. For the Cisco data, 9 of 130 observations (or roughly 7 percent of all returns) are outliers. In general, the rule of thumb is less accurate for highly skewed data sets but is usually very accurate for relatively symmetric data sets, even if the data does not come from a normal population.

Many valuable insights can be obtained by finding causes of outliers. Companies should try to ensure that the causes of "good outliers" occur more frequently and the causes of "bad outliers" occur less frequently.

Using conditional formatting to highlight outliers

You'll find that it is often useful to highlight all outliers in a data set. An example is shown in Figure 42-4. (See the Stockprices worksheet in the Stock.xlsx file.)

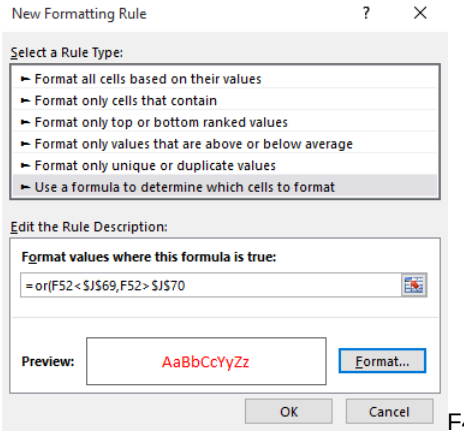
	D	E	F
48			
49	min	-0.240	-0.203
50	max	0.277	0.339
51	INTC	GM	CSCO
52	0.037	0.022	0.011
53	-0.054	-0.035	0.011
54	0.222	0.116	0.042
55	-0.026	-0.021	0.071
56	-0.053	-0.021	-0.038
57	-0.250	-0.131	-0.029
58	-0.004	-0.088	-0.091
59	0.008	0.014	0.311
60	0.119	0.014	0.339
61	0.027	-0.058	0.136
62	0.188	0.055	0.304

F42xx04

FIGURE 42-4 Highlighting the outliers for Cisco with conditional formatting.

F42xx04: This figure shows the outliers for monthly Cisco returns highlighted via conditional formatting.

For example, to highlight the outliers for the Cisco data, you first compute the lower cutoff for an outlier (*mean–2s*) in cell J69 and the upper cutoff for an outlier (*mean+2s*) in cell J70. Next, select the entire range of Cisco returns (cells F52:F181). Then go to the first cell in the range (F52), select Conditional Formatting on the Home tab (in the Styles group), and select New Rule. In the New Formatting Rule dialog box, select Use A Formula To Determine Which Cells To Format, and then fill in the rest of the dialog box as shown in Figure 42-5: in the Format Values Where This Formula Is True box, enter the formula =OR(F52<$J$69,F52>$J$70).



F42xx05

FIGURE 42-5 Conditional formatting rules to select outliers, as shown in the New Formatting Rule dialog box.

F42xx05: This figure shows the settings needed to highlight Cisco outliers.

This condition ensures that if cell F52 is either more than *2s* above or below the mean monthly Cisco return, the format you select (a red font color in this case) will be applied to cell F52. This formatting condition is automatically copied to the selected range, and all outliers show up in red.

How can I use descriptive statistics to compare data sets?

You can use descriptive statistics to summarize the differences between data sets; for example, between Cisco and GM monthly returns. Looking at the shape and the measures and spread of a typical value, you can conclude the following:

- Typically (looking at either the mean or the median), Cisco monthly returns are higher than GM.
- Cisco monthly returns are more variable (looking at standard deviation, variance, and range) than monthly GM returns.
- Both Cisco and GM monthly returns exhibit slightly positive skews. GM monthly returns are more peaked than a normal curve, whereas Cisco monthly returns are less peaked than a normal curve.

Later in this chapter, I will show how boxplots make it easy to visually compare data sets.

For a given data point, can I easily find its percentile ranking within the data set? For example, how can I find the ninetieth percentile of a data set?

Before Excel 2010, the PERCENTILE and PERCENTRANK functions were useful when you wanted to determine an observation's relative position in a data set. Four related functions were added in Excel 2010: PERCENTILE.INC, PERCENTILE.EXC, PERCENTRANK.INC, and PERCENTRANK.EXC. The functions PERCENTILE.INC and PERCENTRANK.INC give results

identical to the old PERCENTILE and PERCENTRANK functions. Take note that previous versions of Excel do not recognize these functions. Examples of how all these functions work are in the file Percentile.xlsx, shown in Figure 42-6.

	C	D	E	F	G	H
2		RANK	RANK			
3	Data	EXC	INC			
4	10	0.062	0	Percentile	EXC	INC
5	20	0.125	0.071	0.1	16	24
6	30	0.187	0.142	0.2	32	38
7	40	0.25	0.214	0.3	48	52
8	50	0.312	0.285	0.4	64	66
9	60	0.375	0.357	0.5	80	80
10	70	0.437	0.428	0.6	96	94
11	80	0.5	0.5	0.7	112	108
12	90	0.562	0.571	0.8	128	122
13	100	0.625	0.642	0.9	204	136
14	110	0.687	0.714			
15	120	0.75	0.785			
16	130	0.812	0.857			
17	140	0.875	0.928			
18	300	0.937	1			

F42xx06

FIGURE 42-6 Examples of the PERCENTILE and PERCENTRANK functions.

F42xx06: This figure shows examples of Excel functions used to compute percentiles and the percentage rank of data.

The PERCENTILE, PERCENTILE.INC, AND PERCENTILE.EXC functions return the percentile of a data set that you specify. The syntax of the these functions takes the form PERCENTILE.INC(data,k), which returns the *k*th percentile of the information in the cell range specified by data.

Consider a data set consisting of *n* pieces of data. The PERCENTILE and PERCENTILE.INC functions returned the *p*th percentile (0<p<1) as the 1+(n-1)p ranked item in the data set. For example, in H13, the formula PERCENTILE.INC(C4:C18,F13) computes the ninetieth percentile of the data in C4:C18 as 1+(15-1).9, which equals the 13.6 ranked item. That is, assuming the data is sorted in ascending order, Excel computes a number 60 percent of the way between the thirteenth data point (130) and the fourteenth data point (140). This yields 136.

The PERCENTILE.EXC function computes the *k*th percentile as the (n+1)p ranked item in the data set. PERCENTILE.EXC computes the ninetieth percentile of the data as (15+1)(.9), which is the 14.4 ranked item. That is, the ninetieth percentile (assuming again data is sorted in ascending order) is computed to be 40 percent of the way between the fourteenth data point (140) and the fifteenth data point (300). This yields (0.60)(140) + (0.40)(300) =204. You can see that the two functions return drastically different answers. If you consider the data to have been drawn by sampling from a large set of data, you might assume that given the data you've seen, there is much more than a 10 percent chance that a piece of data would be more than 136. After all, two of the 15 data points are more than 130, so it does not seem reasonable to say that the ninetieth percentile of the data is only 136.

Therefore, saying the ninetieth percentile is 204 seems more reasonable. I strongly recommend using the .EXC function instead of the .INC function. Note that the .EXC function does not compute a percentile for 0 and 1. The *.EXC* extension stands for the fact that the PERCENTILE.EXC function excludes the zero and one hundredth percentiles.

The PERCENTRANK, PERCENTRANK.INC, and PERCENTRANK.EXC functions return the ranking of an observation relative to all values in a data set. The syntax of the PERCENTRANK.EXC function, for example, is PERCENTRANK.EXC(data,value). The PERCENTRANK and PERCENTRANK.INC functions both calculate the percentile rank of the *k*th smallest number in the data set as (k-1)/(n-1). Thus, as shown in cell E4, the PERCENTILE or PERCENTILE.INC function yields a rank of 0 for 10, because *k* =1 for this data point. The PERCENTRANK.EXC function computes the rank of the *k*th smallest data point as k/(n+1). In cell D4, the PERCENTRANK.EXC function returns a rank of 1/16 = 0.0625. A percentile ranking of 6.25 percent seems more realistic than a ranking of 0 percent, because there is little reason to think that a value of 10 is the smallest data point in a larger data set from which this data was sampled.

Note The PERCENTILE and PERCENTRANK functions are easily confused. To simplify, PERCENTILE yields a possible data value, whereas PERCENTRANK yields a percentage.

How can I easily find the second largest or second smallest number in a data set?

The formula =LARGE(range,k) returns the *k*th largest number in a cell range. The formula =SMALL(range,k) returns the *k*th smallest number in a cell range. For example, in the file Trimmean.xlsx, in cell F1, the formula =LARGE(C4:C62,2) returns the second largest number in the cell range C4:C62 (99), whereas in cell F2, the formula =SMALL(C4:C62,2) returns the second smallest number in the cell range C4:C62 (80). (See Figure 42-7.)

	C	D	E	G	H
1				2nd highest score	99
2				2nd lowest score	80
3	Scores	Rank(ties)	Average Rank	10% trimmed mean	90.04
4	93	20	21.5	5% trimmed mean	90.02
5	84	48	49		
6	88	38	39		
7	100	1	1		
8	86	45	45.5		
9	86	45	45.5	93 is 20th-23rd	
10	95	12	14	94 17th-19th	
11	92	24	24.5		
12	88	38	39		
13	94	17	18		
14	97	5	6.5		
15	91	26	27		
16	92	24	24.5		
17	95	12	14		
18	93	20	21.5		
19	80	56	57.5		
20	89	32	34.5		
21	98	3	3.5		

F42xx07

FIGURE 42-7 Examples of the LARGE and SMALL functions, the RANK and RANK.AVG functions, and the trimmed mean.

F42xx07: This figure shows examples of the LARGE and SMALL functions, as well as various RANK functions.

How can I rank numbers in a data set?

The RANK function ranks numbers in a data set. The syntax of the RANK function is RANK(number,array,0). Excel 2010 introduced a function, RANK.EQ, which returns results identical to the RANK function. This formula yields the rank of a number in a given array, where the largest number in the array is assigned rank 1, the second largest number is rank 2, and so on. The syntax RANK(number,array,1) or RANK.EQ(number,array,1) results in assigning a rank of 1 to the smallest number in the array, a rank of 2 to the second smallest number, and so on. In the file Trimmean.xlsx (see Figure 42-7), copying the formula =RANK.EQ(C4,$C$4,$C$62,0) from cell D4 to D5:D62 returns the rank of each test score. For example, the score of 100 in cell C7 is the highest score, whereas the scores of 98 in cells C21 and C22 tied for the third highest score. Note that the RANK function returned a 3 for both scores of 98.

The Excel function RANK.AVG has the same syntax as the other RANK functions, but in the case of ties, RANK.AVG returns the average rank for all the tied data points. For example, since the two scores of 98 ranked third and fourth, RANK.AVG returns 3.5 for each. I generated the average ranks by copying the formula =RANK.AVG(C4,$C$4:$C$62,0) from E4 to E5:E62.

What is the trimmed mean of a data set?

Extreme skewness in a data set can distort the mean of the data set. In these situations, people usually use the median as a measure of the data set's typical value. The median, however, is unaffected by many changes in the data. For example, compare the following two data sets:

```
Set 1: -5, -3, 0, 1, 3, 5, 7, 9, 11, 13, 15

Set 2: -20, -18, -15, -10, -8, 5, 6, 7, 8, 9, 10
```

These data sets have the same median (5), but the second data set should have a lower "typical" value than the first. The *trimmed mean* trims off data points from the top and bottom of the data set. The Excel TRIMMEAN function is less distorted by extreme values than the AVERAGE function, but it is more influenced by extreme values than the median. The formula =TRIMMEAN(range,percent) computes the mean of a data set, after deleting the data points at the top percent divided by 2 and bottom percent divided by 2. For example, applying the TRIMMEAN function with percent=10% converts the mean after deleting the top 5 percent and bottom 5 percent of the data. In cell H3 of the file Trimmean.xlsx, the formula =TRIMMEAN(C4:C62,0.1) computes the mean of the scores in C4:C62 after deleting the three highest and three lowest scores. (The result is 90.04.) In cell H4, the formula =TRIMMEAN(C4:C62,0.05) computes the mean of the scores in C4:C62 after deleting the top and bottom scores. This calculation occurs because 0.05*59=2.95 would indicate the deletion of 1.48 of the largest observations and 1.48 of the smallest. Rounding off 1.48 results in deleting only the top and bottom observations. (See Figure 42-7.)

When I select a range of cells, is there an easy way to get a variety of statistics that describes the data in those cells?

To see the solution to this question, select the cell range C4:C36 in the file Trimmean.xlsx. In the lower-right corner of your screen, the Excel status bar displays a cornucopia of statistics describing the numbers in the selected cell range. Figure 42-8 shows several of them. If you right-click the status bar, you can change the displayed set of statistics. I selected Minimum and Maximum to see those values in addition to the default statistics shown in Figure 42-8. For the cell range C4:C36, the mean is 90.39, there are 33 numbers, the smallest value is 80, and the largest value is 100.

F42xx08

FIGURE 42-8 The statistics shown on the status bar.

F42xx08: This figure shows the statistics displayed by the status bar.

Why do financial analysts often use the geometric mean to summarize the average return on a stock? The file Geommean.xlsx contains the annual returns of two fictitious stocks. (See Figure 42-9.)

	B	C	D
1			
2			
3			
4		Stock 1	Stock 2
5	Year 1	0.05	-0.5
6	Year 2	0.05	0.7
7	Year 3	0.05	-0.5
8	Year 4	0.05	0.7
9	Average	0.05	0.1
10			
11		1+return	
12		1.05	0.5
13		1.05	1.7
14		1.05	0.5
15		1.05	1.7
16	geometric means	0.05	-0.07805

F42xx09

FIGURE 42-9 A geometric mean.

F42xx09: This figure illustrates how to compute the geometric mean of a set of annual stock returns.

Cell C9 indicates that the average annual return on Stock 1 is 5 percent and the average annual return on Stock 2 is 10 percent. This would seem to indicate that Stock 2 is a better investment. If you think about it, however, what will probably happen with Stock 2 is that one year you will lose 50 percent and the next gain 70 percent. This means that every two years $1.00 becomes 1(1.7)(.5)=.85. Because Stock 1 never loses money, you know that it is clearly the better investment. Using the geometric mean as a measure of average annual return helps to correctly conclude that Stock 1 is the better investment. The *geometric mean* of *n* numbers is the *n*th root of the product of the numbers (the central number in a geometric progression, where you multiply the numbers together and then take the squared root if there are two numbers, the cubed root if there are three numbers, and so on). For example, the geometric mean of 1 and 4 is the square root of 4 (2), whereas the geometric mean of 1, 2, and 4 is the cubed root of 8 (also 2). To use the geometric mean to calculate an average annual return on an investment, you add 1 to each annual return and take the geometric mean of the

resulting numbers. Then subtract 1 from this result to obtain an estimate of the stock's average annual return.

The formula =GEOMMEAN(range) finds the geometric mean of numbers in a range. So, to estimate the average annual return on each stock, you proceed as follows:

1. Compute 1 + each annual return by copying from C12 to C12:D15 the formula =1+C5.
2. Copy from C16 to D16 the formula =GEOMEAN(C12:C15)–1.

The annual average return on Stock 1 is estimated to be 5 percent, and the annual average return on Stock 2 is –7.8 percent. Note that if Stock 2 yields the mean return of –7.8 percent during two consecutive years, \$1 becomes $1*(1–0.078)^2=0.85$, which agrees with common sense.

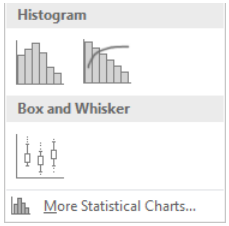
How can I use boxplots to summarize and compare datasets?

Recall from Chapter 41 that a *histogram* is a chart that shows the frequencies of a data set in various ranges. A *boxplot* is a chart that graphically displays five important descriptive values for a data set. These values include the following:

- The minimum value
- The maximum value
- The first quartile (the 25th percentile)
- The median (the 50th percentile)
- The third quartile (the 75th percentile)

An example of a boxplot is shown in Figure 42-11. The length of the box in the boxplot is the *interquartile range* (IQR) = 75th percentile – 25th percentile.

Using the data in the file Boxplottemp.xlsx (in the Templates folder), we create a boxplot to analyze a classic data set: the military draft numbers from the 1969 draft lottery. In the 1969 lottery, a container was filled with the numbers 1–366, which were then supposedly thoroughly mixed. Then a ball was drawn for January 1 (number 305), next a ball was drawn for January 2 (number 159), and so on. Men with their birth date selected as number 1 were drafted first, then men with draft number 2, and so on. Lower draft numbers made it more likely that a man would be drafted. The cell range A7:B373 contains the draft-lottery number for each birth date (in column B) and the month of the year (in column A). After selecting this range, from the Insert tab, selected Insert Statistic Chart (in the Charts group) and then choose the Box And Whisker option shown in Figure 42-10.

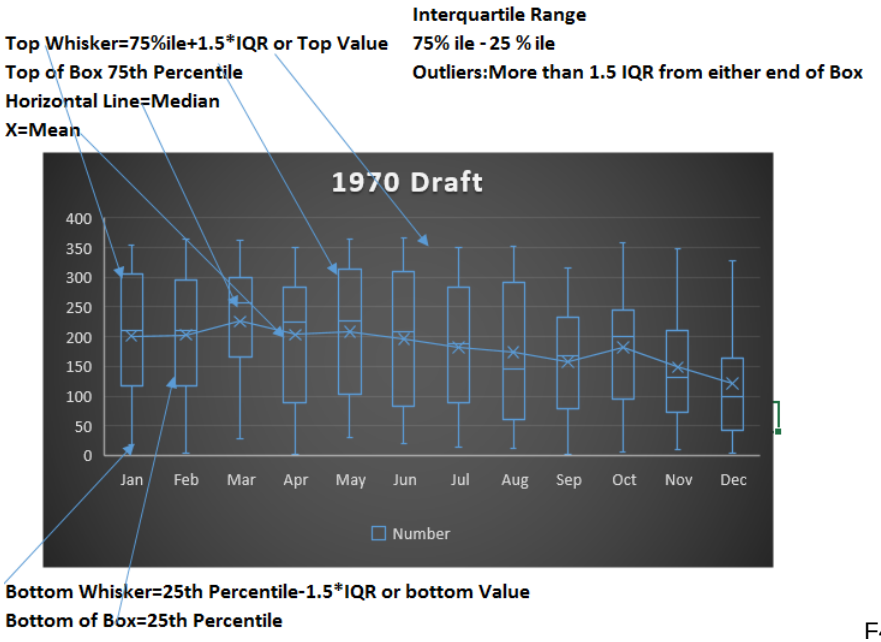


F42xx10

FIGURE 42-10 Selecting a Box And Whisker chart.

F42xx10: This figure shows how to select a Box and Whisker chart.

To create the boxplot in Figure 42-11, from the Design tab, I chose the fourth option in the Chart Styles group. Then I hovered over the plot area of the chart that shows the diagrams, and when the ScreenTip displayed Series “Number”, I right-clicked to select Format Data Series. In the Format Data Series pane, I chose the Show Outlier Points check box because any point more than 1.5*(IQR) from either end of box is declared an outlier. I then chose Show Mean Markers (the Xs on the chart) and Show Mean Line (the line connecting the Xs).



F42xx11

FIGURE 42-11 A boxplot for the draft-lottery data.

F42xx11: This figure shows a boxplot for the draft-lottery data.

For each month’s draft numbers, the following information is shown. (Figure 42-12 shows the calculations for January.)

- The top of the box (305) is the 75th percentile.
- The bottom of the box (118) is the 25th percentile.
- The horizontal line (211) is the median, or 50th percentile.
- The top of the upper whisker (355) is the minimum (largest data point, 75th percentile +1.5IQR).
- The bottom of the lower whisker (17) is the maximum (smallest data point, 25th percentile – 1.5*IQR).

	E	F	G
9	Jan Median	211	=MEDIAN(Jan)
10	Jan Mean	201.161	=AVERAGE(Jan)
11	Jan 25%ile	118	=QUARTILE.EXC(Jan,1)
12	Jan 75%ile	305	=QUARTILE.EXC(Jan,3)
13	Jan max	355	=MAX(Jan)
14	Jan min	17	=MIN(Jan)
15			
16	IQR	187	=F12-F11
17	1.5*IQR	280.5	=1.5*F16
18	Upper Outlier Cutoff	585.5	=F12+F17
19	Lower Outlier Cutoff	-162.5	=F11-F17

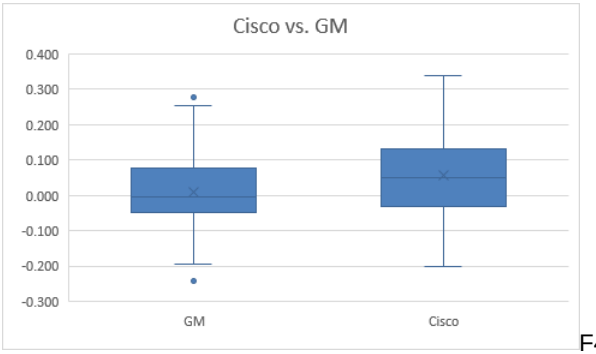
F42xx12

FIGURE 42-12 The computations for the January boxplot data.

F42xx12: This figure shows the calculations underlying the January portion of the boxplot.

The main takeaway from the boxplot is that the means and medians appear to be decreasing during the calendar year. This indicates that later dates tend to have lower draft numbers, and the drawing was not random. More advanced statistical methods, such as resampling (see Chapter 77) confirm that the lottery exhibited a lack of randomness. The most frequent explanation advanced for the failure of randomness is that lower numbered balls were put in first, and the balls were not properly mixed. This led to later dates tending to have lower draft numbers. In 1970, the balls were more thoroughly mixed, and no evidence of nonrandomness was found.

As a second example of the power of boxplots, Figure 42-13 (see the file Stocksandboxplots.xlsx) shows a boxplot comparing the Cisco and GM stock returns that were discussed earlier in this chapter.



F42xx13

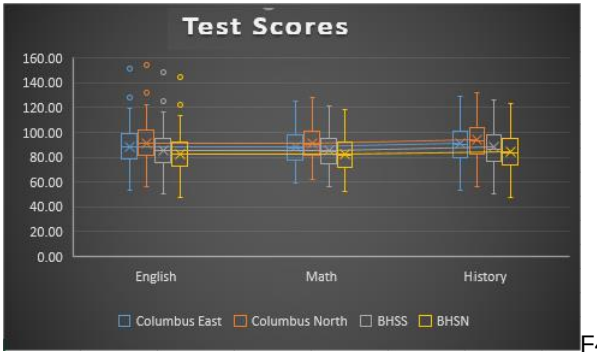
FIGURE 42-13 A boxplot of Cisco and GM monthly returns.

F42xx13: This figure shows a boxplot contrasting the Cisco and GM monthly returns.

We immediately draw the following three conclusions from the boxplot:

- The Cisco box is higher than the GM box, so on average Cisco did better than GM.
- The Cisco box is longer (taller) than the GM box and the Cisco whiskers are longer than the GM whiskers, so Cisco exhibits more variability than GM.
- The top and bottom whiskers for each stock are roughly of the same length, and for each stock the mean and median are virtually identical. This indicates that the GM and Cisco data sets exhibit symmetry.

In the file Boxplotmultiple.xlsx (see Figure 42-14), you can see how boxplots can be used to compare several populations on multiple variables.



F42xx14

FIGURE 42-14 A boxplot comparing high-school test scores.

F42xx14: This figure shows a boxplot comparing high-school test scores in multiple subjects.

For each high school, we are given in a different column English, math, and history test scores. From the boxplot we quickly see the following:

- Each high school had two outliers (on the high side) on English test scores.
- Columbus North students tend to perform the best on each test, and Bloomington North students tend to perform the worst on each test.
- For each test, it appears that all the schools have boxes and whiskers of similar length, indicating the variability of scores on each test is consistent across schools.

Problems

1. Use the data in the file Stock.xlsx to generate descriptive statistics for Intel and GE stock.
2. Use your answer to problem 1 to compare the monthly returns on Intel and GE stock.
3. City Power & Light produces voltage-regulating equipment in New York and ships the equipment to Chicago. A voltage regulator is considered acceptable if it can hold a voltage of 25–75 volts. The voltage held by each unit is measured in New York before each unit is shipped. The voltage is measured again when the unit arrives in Chicago. A sample of voltage measurements from each city is given in the file Citypower.xlsx.
    - Using descriptive statistics, comment on what you have learned about the voltage held by units before 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 City Power & Light's regulators?
    - Ten percent of all New York regulators have a voltage exceeding what value?
    - Five percent of all New York regulators have a voltage less than or equal to what value?
4. In the file Decadeincome.xlsx, you are given a sample of incomes (in thousands of 1980 dollars) for a set of families sampled in 1980 and 1990. Assume that these families are representative of the whole United States. Some Republicans claim that the country was better off in 1990 than in 1980 because the average income increased. Do you agree?
5. Use descriptive statistics to compare the annual returns on stocks, T-bills, and corporate bonds. Use the data contained in the file Historicalinvest.xlsx.
6. In 1969 and 1970, eligibility for the US armed-services draft was determined on the basis of a draft-lottery number. The number was determined by birth date. A total of 366 balls, one for each possible birth date, were placed in a container and shaken. The first ball selected was given the number 1 in the lottery, and so on. Men whose birthdays corresponded to the lowest numbers were drafted first. The file Draftlottery.xlsx contains the actual results of the 1969 and 1970 drawings. For example, in the 1969 drawing, January 1 received the number 305. Use

descriptive statistics to demonstrate that the 1969 draft lottery was not random and the 1970 lottery was random. Hint: Use the AVERAGE and MEDIAN functions to compute the mean and median lottery number for each month.

7. The file Jordan.xlsx gives the starting salaries (hypothetical) of all 1984 geography graduates from the University of North Carolina (UNC). What is your best estimate of a "typical" starting salary for a geography major? In reality, the major at UNC with the highest average starting salary in 1984 was geography, because the great basketball player Michael Jordan was a geography major!

8. Use the LARGE or SMALL function to sort the annual stock returns in the file Historicalinvest.xlsx. What advantage does this method of sorting have over clicking the Sort button?

9. Compare the mean, median, and trimmed mean (trimming 10 percent of the data) of the annual returns on stocks, T-bills, and corporate bonds given in the file Historicalinvest.xlsx.

10. Use the geometric mean to estimate the mean annual return on stocks, bonds, and T-bills in the file Historicalinvest.xlsx.

11. The file Dow.xlsx contains monthly returns on the 30 Dow stocks during the last 20 years. Use this data to determine the three stocks with the largest mean monthly returns.

12. Using the Dow.xlsx data again, determine the three stocks with the most risk or variability.

13. Using the Dow.xlsx data, determine the three stocks with the highest skew.

14. Using the Dow.xlsx data, how do the trimmed-mean returns (trim off 10 percent of the returns) differ from the overall mean returns?

15. The file Incomedata.xlsx contains incomes of a representative sample of Americans in the years 1975, 1985, 1995, and 2005. Describe how US personal income has changed over this time period.

16. The file Coltsdata.xlsx contains yards gained by the 2006 Indianapolis Colts on each rushing and passing play. Describe how the outcomes of rushing plays and passing plays differ.

17. In the file Problem17datat.xlsx, you are given daily returns on Facebook stock. Use this data to answer the following questions:

    - Do Facebook stock returns exhibit significant skewness?
    - Identify all the outliers (using the rule of thumb). Is the number of outliers consistent with the rule of thumb?
    - There is a 1 percent chance that the daily return on Facebook will exceed ____.

18. One theory in brain science states that the level of dopamine in a person's nervous system

determines whether someone will exhibit psychotic behavior. In the file Problem18data.xlsx, you are given the dopamine levels for 10 psychotic and 14 nonpsychotic adults. Use descriptive statistics and a boxplot to compare and contrast the distribution of dopamine in psychotic and nonpsychotic people.

19. Participants in a group of 508 people were asked to estimate the percentage of African nations that are members of the United Nations. A wheel of fortune containing the numbers 1–100 was spun before people answered the question. Unknown to the people, the wheel was rigged to show either 25 or 65. In Problem19data.xlsx, you are given the responses of the participants in the experiment. Describe how the wheel result influenced the responses given by the subjects.

