Understanding Frequency Distributions

In addition to charts that show two variables—such as numbers broken down by categories in a Column chart, or the relationship between two numeric variables in an XY chart—there is another sort of Excel chart that deals with one variable only. It's the visual representation of a *frequency distribution*, a concept that's absolutely fundamental to intermediate and advanced statistical methods.

A frequency distribution is intended to show how many instances there are of each value of a variable. For example:

- The number of people who weigh 100 pounds, 101 pounds, 102 pounds, and so on
- The number of cars that get 18 miles per gallon (mpg), 19 mpg, 20 mpg, and so on
- The number of houses that cost between $200,001 and $205,000, between $205,001 and $210,000, and so on

Because we usually round measurements to some convenient level of precision, a frequency distribution tends to group individual measurements into classes. Using the examples just given, two people who weigh 100.2 and 100.4 pounds might each be classed as 100 pounds; two cars that get 18.8 and 19.2 mpg might be grouped together at 19 mpg; and any number of houses that cost between $220,001 and $225,000 would be treated as in the same price level.

As it's usually shown, the chart of a frequency distribution puts the variable's values on its horizontal axis and the count of instances on the vertical axis. Figure 1.10 shows a typical frequency distribution.

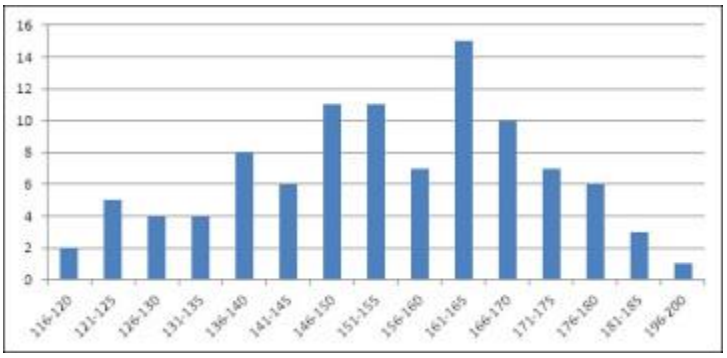


Figure 1.10 Typically, most records cluster toward the center of a frequency distribution.

You can tell quite a bit about a variable by looking at a chart of its frequency distribution. For example, Figure 1.10 shows the weights of a sample of 100 people. Most of them are between 140 and 180 pounds. In this sample, there are about as many people who weigh a lot (say, over 175 pounds) as there are whose weight is relatively low (say, up to 130). The range of weights—that is, the difference between the lightest and the heaviest weights—is about 85 pounds, from 116 to 200.

There are lots of ways that a different sample of people might provide different weights than those shown in Figure 1.10. For example, Figure 1.11 shows a sample of 100 vegans. (Notice that the distribution of their weights is shifted down the scale somewhat from the sample of the general population shown in Figure 1.10.)

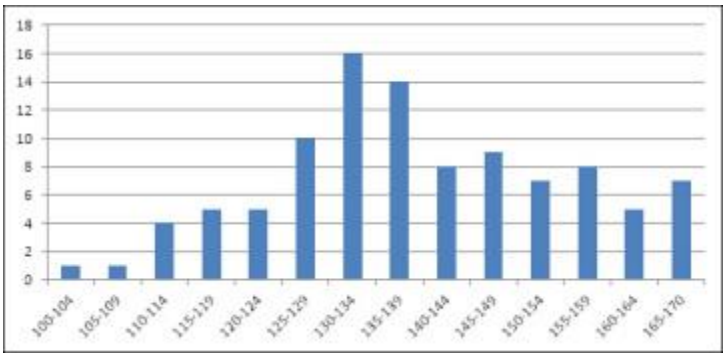


Figure 1.11 Compared to Figure 1.10, the location of the frequency distribution has shifted to the left.

The frequency distributions in Figures 1.10 and 1.11 are relatively symmetric. Their general shapes are not far from the idealized normal "bell" curve, which depicts the distribution of many variables that describe living beings. This book has much more to say in later chapters about the normal curve, partly because it describes so many variables of interest, but also because Excel has so many ways of dealing with the normal curve.

Still, many variables follow a different sort of frequency distribution. Some are skewed right (see Figure 1.12) and others left (see Figure 1.13).

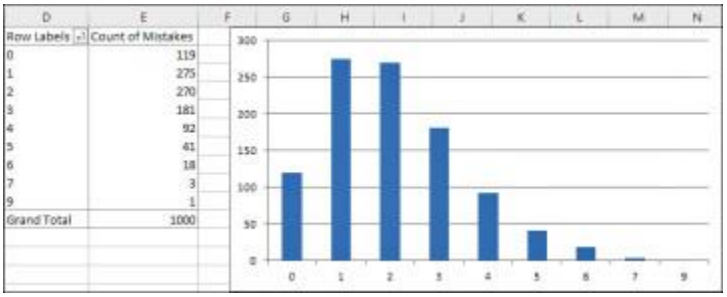


Figure 1.12 A frequency distribution that stretches out to the right is called positively skewed.

Row Labels	Count of Life Expectancy (Decade)
0's	1
20's	3
30's	18
40's	41
50's	92
60's	181
70's	270
80's	275
90's	119
Grand Total	1000

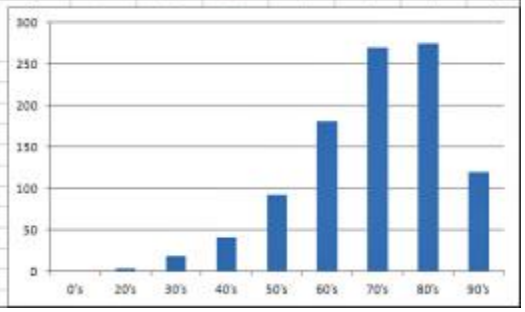


Figure 1.13 Negatively skewed distributions are not as common as positively skewed distributions.

Figure 1.12 shows counts of the number of mistakes on individual federal tax forms. It's normal to make a few mistakes (say, one or two), and it's abnormal to make several (say, five or more). This distribution is positively skewed.

Another variable, home prices, tends to be positively skewed, because although there's a real lower limit (a house cannot cost less than $0) there is no theoretical upper limit to the price of a house. House prices therefore tend to bunch up between $100,000 and $300,000, with fewer between $300,000 and $400,000, and fewer still as you go up the scale.

A quality control engineer might sample 100 ceramic tiles from a production run of 10,000 and count the number of defects on each tile. Most would have zero, one, or two defects, several would have three or four, and a very few would have five or six. This is another positively skewed distribution—quite a common situation in manufacturing process control.

Because true lower limits are more common than true upper limits, you tend to encounter more positively skewed frequency distributions than negatively skewed. But negative skews certainly occur. Figure 1.13 might represent personal longevity: Relatively few people die in their twenties, thirties and forties, compared to the numbers who die in their fifties through their eighties.

Using Frequency Distributions

It's helpful to use frequency distributions in statistical analysis for two broad reasons. One concerns visualizing how a variable is distributed across people or objects. The other concerns how to make inferences about a population of people or objects on the basis of a sample.

Those two reasons help define the two general branches of statistics: *descriptive* statistics and *inferential* statistics. Along with descriptive statistics such as averages, ranges of values, and percentages or counts, the chart of a frequency distribution puts you in a stronger position to understand a set of people or things because it helps you visualize how a variable behaves across its range of possible values.

In the area of inferential statistics, frequency distributions based on samples help you determine the type of analysis you should use to make inferences about the population. As you'll see in later chapters, frequency distributions also help you visualize the results of certain choices that you must make—choices such as the probability of coming to the wrong conclusion.

Visualizing the Distribution: Descriptive Statistics

It's usually much easier to understand a variable—how it behaves in different groups, how it may change over time, and even just what it looks like—when you see it in a chart. For example, here's the formula that defines the normal distribution:

- $u = (1 / ((2\pi)^{.5})\ \sigma)\ e\ \hat{}\ (-(X - \mu)^2 / 2\ \sigma^2)$

And Figure 1.14 shows the normal distribution in chart form.

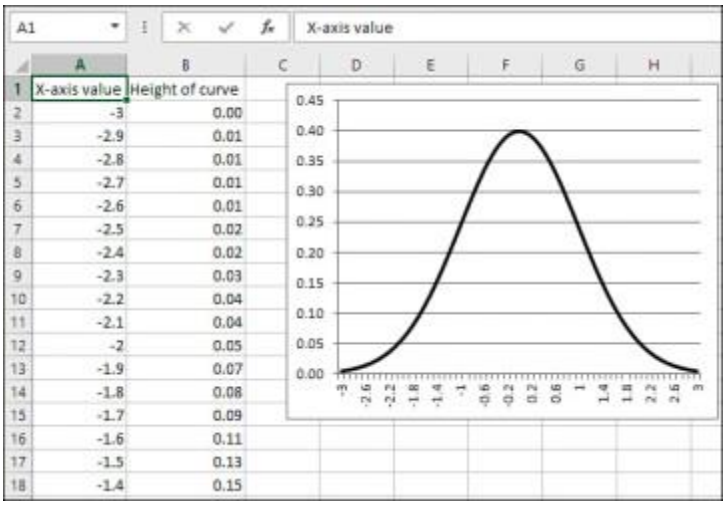


Figure 1.14 The familiar normal curve is just a frequency distribution.

The formula itself is indispensable, but it doesn't convey understanding. In contrast, the chart informs you that the frequency distribution of the normal curve is symmetric and that most of the records cluster around the center of the horizontal axis.

NOTE

The formula was developed by a seventeenth-century French mathematician named Abraham De Moivre. Excel simplifies it to this:

=NORMDIST(1,0,1,FALSE)

In Excel 2010 and 2013, it's this:

=NORM.S.DIST(1,FALSE)

Those are *major* simplifications.

Again, personal longevity tends to bulge in the higher levels of its range (and therefore skews left as in Figure 1.13). Home prices tend to bulge in the lower levels of their range (and therefore skew right). The height of human beings creates a bulge in the center of the range, and is therefore symmetric and *not* skewed.

Some statistical analyses assume that the data comes from a normal distribution, and in some statistical analyses that assumption is an important one. This book does not explore the topic in great detail because it comes up infrequently. Be aware, though, that if you want to analyze a skewed distribution there are ways to normalize it and therefore comply with the requirements of the analysis. Very generally, you can use Excel's SQRT() and LOG() functions to help normalize a negatively skewed distribution, and an exponentiation operator (for example, =A2^2 to square the value in A2) to help normalize a positively skewed distribution.

NOTE

Finding just the right transformation for a particular data set can be a matter of trial and error, however, and the Excel Solver add-in can help in conjunction with Excel's SKEW() function. See Chapter 2, "How Values Cluster Together," for information on Solver, and Chapter 7, "Using Excel with the Normal Distribution," for information on SKEW(). The basic idea is to use SKEW() to calculate the skewness of your transformed data and to have Solver find the exponent that brings the result of SKEW() closest to zero.

Visualizing the Population: Inferential Statistics

The other general rationale for examining frequency distributions has to do with making an inference about a population, using the information you get from a sample as a basis. This is the field of inferential statistics. In later chapters of this book, you will see how to use Excel's tools—in particular, its functions and its charts—to infer a population's characteristics from a sample's frequency distribution.

A familiar example is the political survey. When a pollster announces that 53% of those who were asked preferred Smith, he is reporting a descriptive statistic. Fifty-three percent of the sample preferred Smith, and no inference is needed.

But when another pollster reports that the margin of error around that 53% statistic is plus or minus 3%, she is reporting an inferential statistic. She is extrapolating from the sample to the larger population and inferring, with some specified degree of confidence, that between 50% and 56% of all voters prefer Smith.

The size of the reported margin of error, six percentage points, depends heavily on how confident the pollster wants to be. In general, the greater degree of confidence you want in your extrapolation, the greater the margin of error that you allow. If you're on an archery range and you want to be virtually certain of hitting your target, you make the target as large as necessary.

Similarly, if the pollster wants to be 99.9% confident of her projection into the population, the margin might be so great as to be useless—say, plus or minus 20%. And although it's not headline material to report that somewhere between 33% and 73% of the voters prefer Smith, the pollster can be confident that the projection is accurate.

But the size of the margin of error also depends on certain aspects of the frequency distribution in the sample of the variable. In this particular (and relatively straightforward) case, the accuracy of the projection from the sample to the population depends in part on the level of confidence desired (as just briefly discussed), in part on the size of the sample, and in part on the percent favoring Smith in the sample. The latter two issues, sample size and percent in favor, are both aspects of the frequency distribution you determine by examining the sample's responses.

Of course, it's not just political polling that depends on sample frequency distributions to make inferences about populations. Here are some other typical questions posed by empirical researchers:

- What percent of the nation's existing houses were resold last quarter?
- What is the incidence of cardiovascular disease today among diabetics who took the drug Avandia before questions about its side effects arose in 2007? Is that incidence reliably different from the incidence of cardiovascular disease among those who never took the drug?
- A sample of 100 cars from a particular manufacturer, made during 2013, had average highway gas mileage of 26.5 mpg. How likely is it that the average highway mpg, for all that manufacturer's cars made during that year, is greater than 26.0 mpg?
- Your company manufactures custom glassware. Your contract with a customer calls for no more than 2% defective items in a production lot. You sample 100 units from your latest production run and find 5 that are defective. What is the likelihood that the entire production run of 1,000 units has a maximum of 20 that are defective?

In each of these four cases, the specific statistical procedures to use—and therefore the specific Excel tools—would be different. But the basic approach would be the same: Using the characteristics of a frequency distribution from a sample, compare the sample to a population whose frequency distribution is either known or founded in good theoretical work. Use the numeric functions in Excel to estimate how likely it is that your sample accurately represents the population you're interested in.

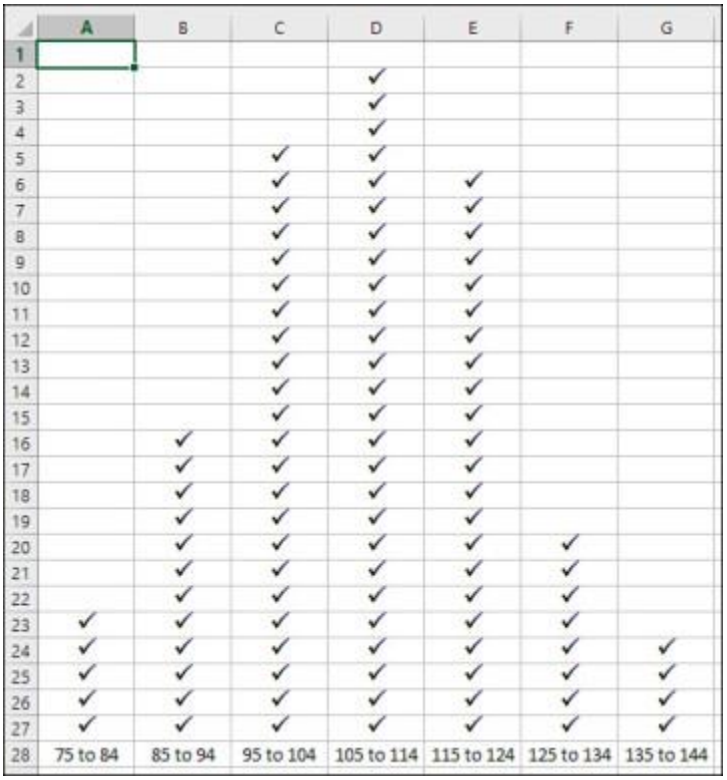
Building a Frequency Distribution from a Sample

Conceptually, it's easy to build a frequency distribution. Take a sample of people or things and measure each member of the sample on the variable that interests you. Your next step depends on how much sophistication you want to bring to the project.

Tallying a Sample

One straightforward approach continues by dividing the relevant range of the variable into manageable groups. For example, suppose that you obtained the weight in pounds of each of 100

people. You might decide that it's reasonable and feasible to assign each person to a weight class that is ten pounds wide: 75 to 84, 85 to 94, 95 to 104, and so on. Then, on a sheet of graph paper, make a tally in the appropriate column for each person, as suggested in Figure 1.15.



	A	B	C	D	E	F	G
1							
2				✓			
3				✓			
4				✓			
5			✓	✓			
6			✓	✓	✓		
7			✓	✓	✓		
8			✓	✓	✓		
9			✓	✓	✓		
10			✓	✓	✓		
11			✓	✓	✓		
12			✓	✓	✓		
13			✓	✓	✓		
14			✓	✓	✓		
15			✓	✓	✓		
16		✓	✓	✓	✓		
17		✓	✓	✓	✓		
18		✓	✓	✓	✓		
19		✓	✓	✓	✓		
20		✓	✓	✓	✓	✓	
21		✓	✓	✓	✓	✓	
22		✓	✓	✓	✓	✓	
23	✓	✓	✓	✓	✓	✓	
24	✓	✓	✓	✓	✓	✓	✓
25	✓	✓	✓	✓	✓	✓	✓
26	✓	✓	✓	✓	✓	✓	✓
27	✓	✓	✓	✓	✓	✓	✓
28	75 to 84	85 to 94	95 to 104	105 to 114	115 to 124	125 to 134	135 to 144

Figure 1.15 This approach helps clarify the process, but there are quicker and easier ways.

The approach shown in Figure 1.15 uses a *grouped* frequency distribution, and tallying by hand into groups was the only practical option as recently as the 1980s, before personal computers came into truly widespread use. But using an Excel function named FREQUENCY(), you can get the benefits of grouping individual observations without the tedium of manually assigning individual records to groups.

Grouping with FREQUENCY()

If you assemble a frequency distribution as just described, you have to count up all the records that belong to each of the groups that you define. Excel has a function, FREQUENCY(), that will do the heavy lifting for you. All you have to do is decide on the boundaries for the groups and then point the FREQUENCY() function at those boundaries and at the raw data.

Figure 1.16 shows one way to lay out the data.

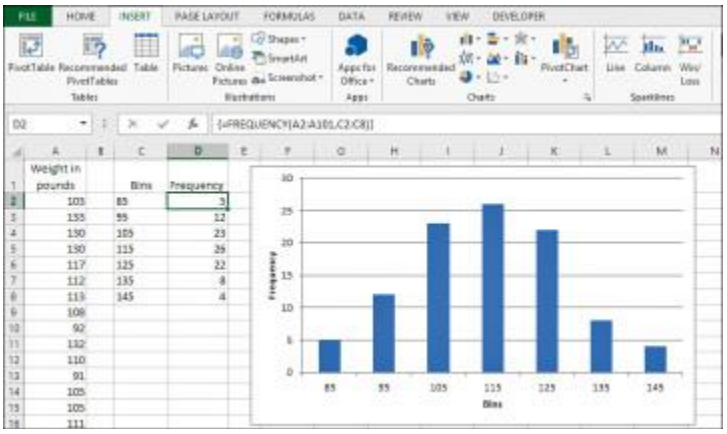


Figure 1.16 The groups are defined by the numbers in cells C2:C8.

In Figure 1.16, the weight of each person in your sample is recorded in column A. The numbers in cells C2:C8 define the upper boundaries of what this section has called *groups*, and what Excel calls *bins*. Up to 85 pounds defines one bin; from 86 to 95 defines another; from 96 to 105 defines another, and so on.

NOTE

There's no special need to use the column headers shown in Figure 1.16, cells A1, C1, and D1. In fact, if you're creating a standard Excel chart as described here, there's no great need to supply column headers at all. If you don't include the headers, Excel names the data Series1 and Series2. If you use the pivot chart instead of a standard chart, though, you will need to supply a column header for the data shown in column A in Figure 1.16.

The count of records within each bin appears in D2:D8. You don't count them yourself—you call on Excel to do that for you, and you do that by means of a special kind of Excel formula, called an *array formula*. You'll read more about array formulas in Chapter 2, as well as in later chapters, but for now here are the steps needed to get the bin counts shown in Figure 1.16:

1. Select the range of cells that the results will occupy. In this case, that's the range of cells D2:D8.
2. Type, but don't yet enter, the following formula:

   =FREQUENCY(A2:A101,C2:C8)

   which tells Excel to count the number of records in A2:A101 that are in each bin defined by the numeric boundaries in C2:C8.

3. After you have typed the formula, hold down the Ctrl and Shift keys simultaneously and press Enter. Then release all three keys. This keyboard sequence notifies Excel that you want it to interpret the formula as an array formula.

NOTE

When Excel interprets a formula as an array formula, it places curly brackets around the formula in the formula box.

TIP

You can use the same range for the Data argument and the Bins argument in the FREQUENCY() function: for example, =FREQUENCY(A1:A101,A1:A101). Don't forget to enter it as an array formula. This is a convenient way to get Excel to treat every recorded value as its own bin, and you get the count for every unique value in the range A1:A101.

The results appear very much like those in cells D2:D8 of Figure 1.16, of course depending on the actual values in A2:A101 and the bins defined in C2:C8. You now have the frequency distribution but you still should create the chart.

Compared to earlier versions, Excel 2013 makes it quicker and easier to create certain basic charts such as the one shown in Figure 1.16. Assuming the data layout used in that figure, here are the steps you might use in Excel 2013 to create the chart:

1. Select the data you want to chart—that is, the range C1:D8. (If the relevant data range is surrounded by empty cells or worksheet boundaries, all you need to select is a single cell in the range you want to chart.)
2. Click the Insert tab, and then click the Recommended Charts button in the Charts group.
3. Click the Clustered Column chart example in the Insert Chart window, and then click OK.

You can get other variations on chart types in Excel 2013 by clicking, for example, the Insert Column Chart button (in the Charts group on the Insert tab). Click More Chart Types at the bottom of the drop-down to see various ways of structuring Bar charts, Line charts, and so on given the layout of your underlying data.

Things weren't as simple in earlier versions of Excel. For example, here are the steps in Excel 2010, again assuming the data is located as in Figure 1.16:

1. Select the data you want to chart—that is, the range C1:D8.
2. Click the Insert tab, and then click the Insert Column Chart button in the Charts group.
3. Choose the Clustered Column chart type from the 2-D charts. A new chart appears, as shown in Figure 1.17. Because columns C and D on the worksheet both contain numeric values, Excel initially thinks that there are two data series to chart: one named Bins and one named Frequency.

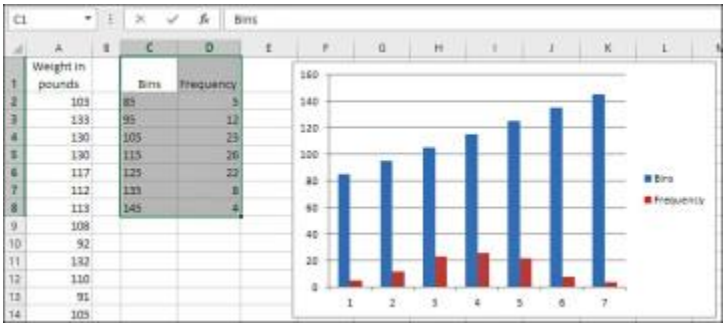


Figure 1.17 Values from both columns are charted as data series at first because they're all numeric.

4. Fix the chart by clicking Select Data in the Design tab that appears when a chart is active. The dialog box shown in Figure 1.18 appears.

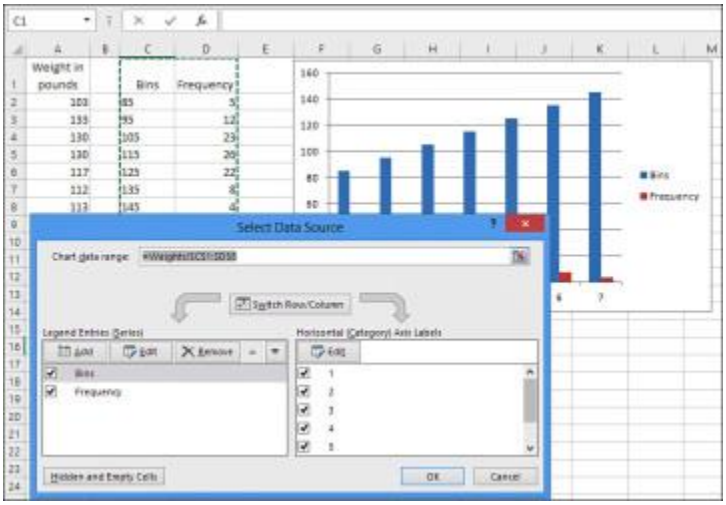


Figure 1.18 You can also use the Select Data dialog box to add another data series to the chart.

5. Click the Edit button under Horizontal (Category) Axis Labels. A new Axis Labels dialog box appears; drag through cells C2:C8 to establish that range as the basis for the horizontal axis. Click OK.
6. Click the Bins label in the left list box shown in Figure 1.18. Click the Remove button to delete it as a charted series. Click OK to return to the chart.
7. Remove the chart title and series legend, if you want, by clicking each and pressing Delete.

At this point, you will have a normal Excel chart that looks much like the one shown in Figure 1.16.

Using Numeric Values as Categories

The differences between how Excel 2010 and Excel 2013 handle charts present a good illustration of the problems created by the use of numeric values as categories. The "Charting Two Variables" section earlier in this chapter alludes to the ambiguity involved when you want Excel to treat numeric values as categories.

In the example shown in Figure 1.16, you present two numeric variables—Bins and Frequency—to Excel's charting facility. Because both variables are numeric (and their values are stored as numbers rather than text), there are various ways that Excel can treat them in charts:

- Treat each *column*—the Bins variable and the Frequency variable—as data series to be charted. This is the approach you might take if you wanted to chart both Income and Expenses over time: you would have Excel treat each variable as a data series, and the different rows in the underlying data range would represent different time periods. You get this chart if you choose Clustered Chart in the Insert Column Chart drop-down.
- Treat each *row* in the underlying data range as a data series. Then, the columns are treated as different categories on the column chart's horizontal axis. You get this result if you click More Column Charts at the bottom of the Insert Column Chart drop-down—it's the third example chart in the Insert Chart window.
- Treat one of the variables—Bins or Frequency—as a category variable for use on the horizontal axis. This is the column chart you see in Figure 1.16 and is the first of the recommended charts.

Excel 2013, at least in the area of charting, recognizes the possibility that you will want to use numeric values as nominal categories. It lets you express an opinion without forcing you to take all the extra steps required by Excel 2010. Still, if you're to participate effectively, you need to recognize the differences between, say, interval and nominal variables. You also need to recognize the ambiguities that crop up when you want to use a number as a category.

Grouping with Pivot Tables

Another approach to constructing the frequency distribution is to use a pivot table. A related tool, the pivot chart, is based on the analysis that the pivot table provides. I prefer this method to using an array formula that employs FREQUENCY(). With a pivot table, once the initial groundwork is done, I can use the same pivot table to do analyses that go beyond the basic frequency distribution. But if all I want is a quick group count, FREQUENCY() is usually the faster way.

Again, there's more on pivot tables and pivot charts in Chapter 2 and later chapters, but this section shows you how to use them to establish the frequency distribution.

Building the pivot table (and the pivot chart) requires you to specify bins, just as the use of FREQUENCY() does, but that happens a little further on.

NOTE

A reminder: When you use the FREQUENCY() method described in the prior section, a header at the top of the column of raw data can be helpful but is not required. When you use the pivot table method discussed in this section, the header is required.

Begin with your sample data in A1:A101 of Figure 1.16, just as before. Select any one of the cells in that range and then follow these steps:

1. Click the Insert tab. Click the PivotChart button in the Charts group. (Prior to Excel 2013, click the PivotTable drop-down in the Tables group and choose PivotChart from the drop-down list.) When you choose a pivot chart, you automatically get a pivot table along with it. The dialog box in Figure 1.19 appears.

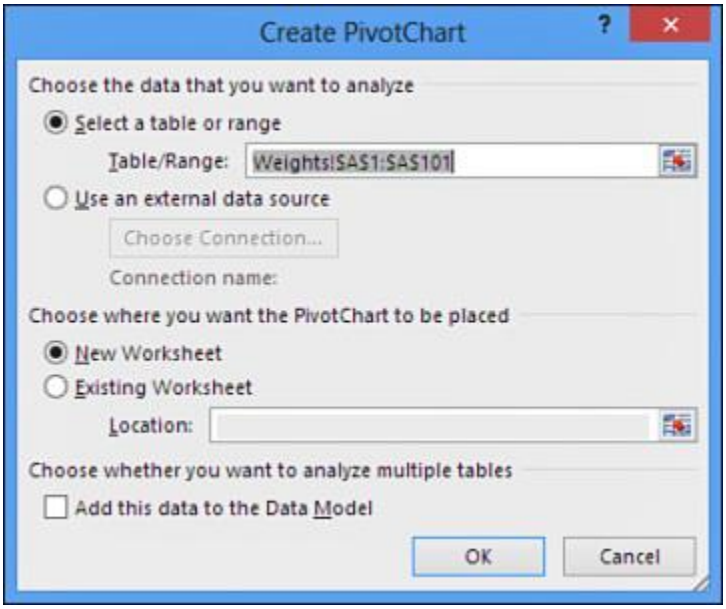


Figure 1.19 If you begin by selecting a single cell in the range containing your input data, Excel automatically proposes the range of adjacent cells that contain data.

2. Click the Existing Worksheet option button. Click in the Location range edit box. Then, to avoid overwriting valuable data, click some blank cell in the worksheet that has other empty cells to its right and below it.
3. Click OK. The worksheet now appears as shown in Figure 1.20.

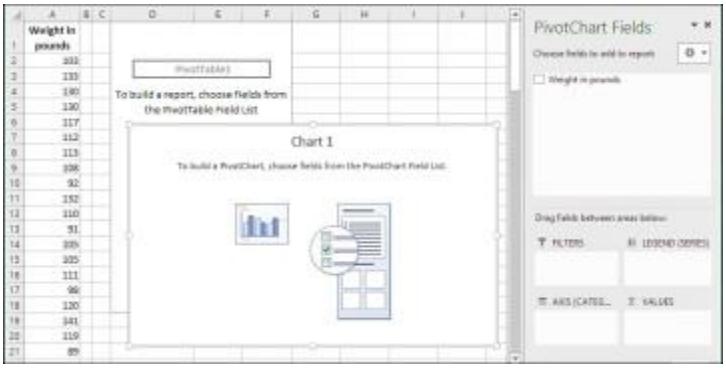


Figure 1.20 With one field only, you normally use it for both Axis Fields (Categories) and Summary Values.

4. Click the Weight In Pounds field in the PivotTable Fields list and drag it into the Axis (Categories) area.
5. Click the Weight In Pounds field again and drag it into the **Σ Values** area. Despite the uppercase Greek sigma, which is a summation symbol, the **Σ Values** in a pivot table can show averages, counts, standard deviations, and a variety of statistics other than the sum. However, Sum is the default statistic for a field that contains numeric values only.
6. The pivot table and pivot chart are both populated as shown in Figure 1.21. Right-click any cell that contains a row label, such as C2. Choose Group from the shortcut menu.

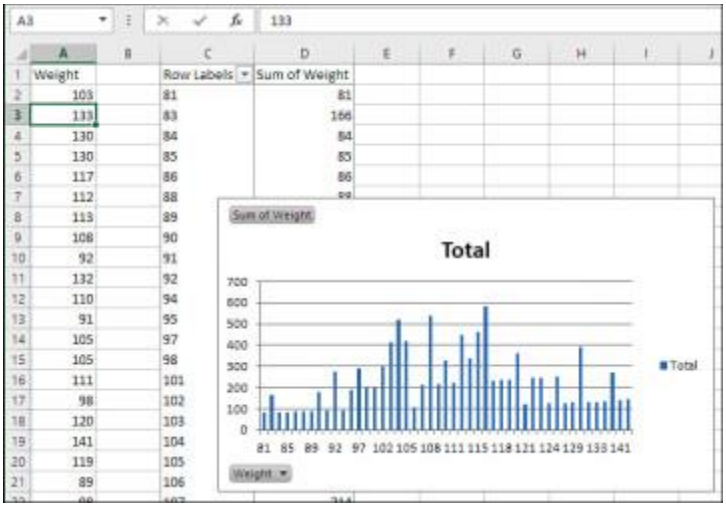


Figure 1.21 The Weight field contains numeric values only, so the pivot table defaults to Sum as the summary statistic.

The Grouping dialog box shown in Figure 1.22 appears.

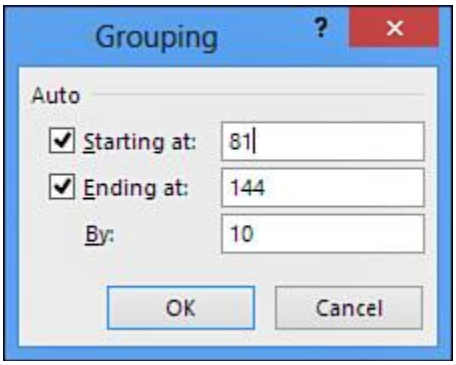


Figure 1.22 This step establishes the groups that the FREQUENCY() function refers to as *bins*.

7. In the Grouping dialog box, set the Starting At value to **81** and enter **10** in the By box. Click OK.
8. Right-click a cell in the pivot table under the header Sum of Weight. Choose Value Field Settings from the shortcut menu. Select Count in the Summarize Value Field By list box, and then click OK.
9. The pivot table and chart reconfigure themselves to appear as in Figure 1.23. To remove the field buttons in the upper- and lower-left corners of the pivot chart, select the chart, click the Analyze tab, click Field Buttons, and select Hide All.

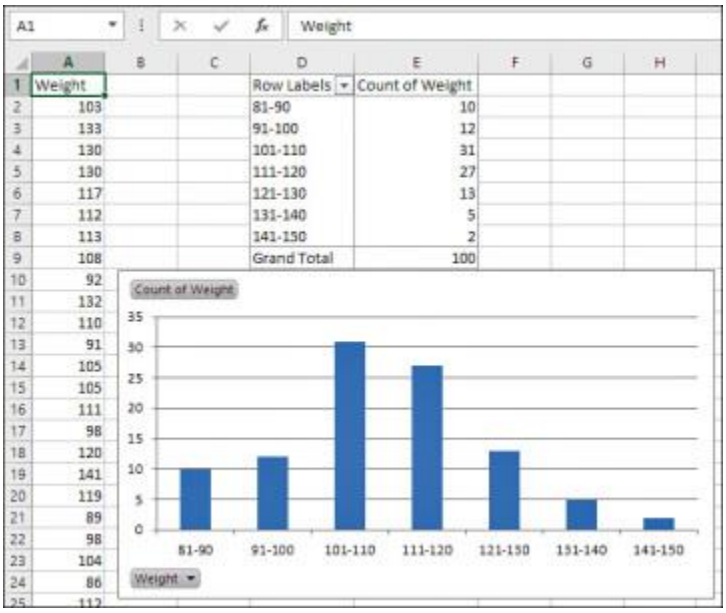


Figure 1.23 This sample's frequency distribution has a slight right skew but is reasonably close to a normal curve.

Building Simulated Frequency Distributions

It can be helpful to see how a frequency distribution assumes a particular shape as the number of underlying records increases. *Statistical Analysis: Excel 2013* has a variety of worksheets and workbooks for you to download from this book's website (www.quepublishing.com/title/9780789753113). The workbook for Chapter 1 has a worksheet named Figure 1.24 that samples records at random from a population of values that follows a normal distribution. The following figure, as well as the worksheet on which it's based, shows how a frequency distribution comes closer and closer to the population distribution as the number of sampled records increases.

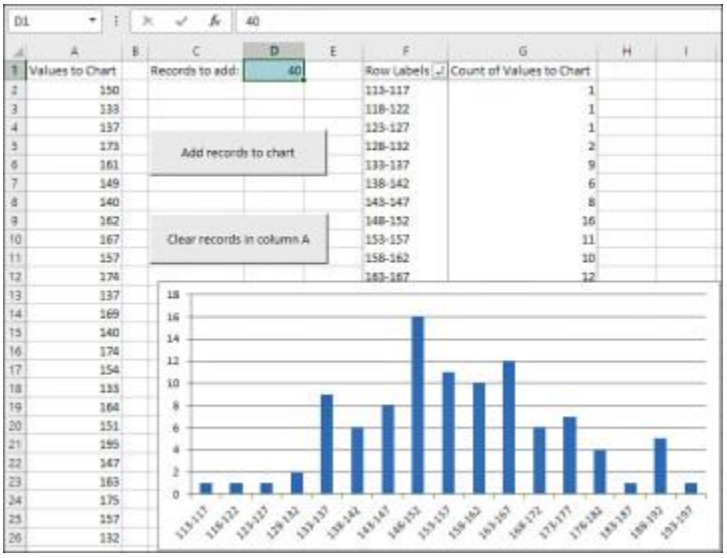


Figure 1.24 This frequency distribution is based on a population of records that follow a normal distribution.

Begin by clicking the button labeled Clear Records in column A. All the numbers will be deleted from column A, leaving only the header value in cell A1. (The pivot table and pivot chart will remain as they were: It's a characteristic of pivot tables and pivot charts that they do not respond immediately to changes in their underlying data sources.)

Decide how many records you'd like to add, and then enter that number in cell D1. You can always change it to another number.

Click the button labeled Add Records to Chart. When you do so, several events take place, all driven by Visual Basic procedures that are stored in the workbook:

- A sample is taken from the underlying normal distribution. The sample has as many records as specified in cell D1. (The underlying, normally distributed population is stored in a separate, hidden worksheet named Random Normal Values; you can display the worksheet by right-clicking a worksheet tab and selecting Unhide from the shortcut menu.)

- The sample of records is added to column A. If there were no records in column A, the new sample is written starting in cell A2. If there were already, say, 100 records in column A, the new sample would begin in cell A102.
- The pivot table and pivot chart are updated (or, in Excel terms, *refreshed*). As you click the Add Records to Chart button repeatedly, more and more records are used in the chart. The greater the number of records, the more nearly the chart comes to resemble the underlying normal distribution.

In effect, this is what happens in an experiment when you increase the sample size. Larger samples resemble more closely the population from which you draw them than do smaller samples. That greater resemblance isn't limited to the shape of the distribution: It includes the average value and measures of how the values vary around the average. Other things being equal, you would prefer a larger sample to a smaller one because it's likely to represent the population more closely.

But this effect creates a cost-benefit problem. It is usually the case that the larger the sample, the more accurate the experimental findings—and the more expensive the experiment. Many issues are involved here (and this book discusses them), but at some point the incremental accuracy of adding, say, ten more experimental subjects no longer justifies the incremental expense of adding them. One of the bits of advice that statistical analysis provides is to tell you when you're reaching the point when the returns begin to diminish.

With the material in this chapter—scales of measurement, the nature of axes on Excel charts, and frequency distributions—in hand, Chapter 2 moves on to the beginnings of practical statistical analysis, the measurement of central tendency.